LUND UNIVERSITY
School of Economics and Management

Department of Informatics

Data Quality Management

Trade-offs in Data Characteristics to Maintain Data Quality

Master thesis 15 HEC, course INFM10 in Information Systems
Presented in June 2017

Authors: Athul Vijayan Nambiar
Dilip Prasad Nair

Supervisor: Olgerta Tona

Examiners: Paul Pierce
Odd Steen

Data Quality Management: Trade-offs in Data Characteristics to Maintain Data Quality

Authors: Athul Vijayan Nambiar and Dilip Prasad Nair

Publisher: Dept. of Informatics, Lund University School of Economics and Management.

Document: Master Thesis

Number of pages: 99

Keywords: Data Quality, Quality Assessment, Data Characteristics, Data Characteristics Trade-offs

Abstract:

We are living in an age of information in which organizations are crumbling under the pressure of exponentially growing data. Increased data quality ensures better decision making, thereby enabling companies to stay competitive in the market. To improve data quality, it is imperative to identify all the characteristics that describe data. And, building on one characteristic results in compromising another, creating a trade-off. There are many well established and interesting theories regarding data quality and data characteristics. However, we found that there is a lack of research and literature regarding how trade-offs are handled between the different types of data that is stored by an organization. To understand how organisations deal with trade-offs, we chose a framework formulated by Eppler, where various data characteristics trade-offs are discussed. After a pre-study with experts in this field, we narrowed it down to three main data characteristic trade-offs and these were further analysed through interviews. Based on the interviews conducted and the literature review, we could prioritize data types under different data characteristics. This research gives insight to how data characteristics trade-offs should be accomplished in organizations.

Acknowledgements

We would like to express our gratitude to our supervisor Olgerta Tona for her assistance and contributions without which this thesis would not have been possible. We would like to thank our interview candidates, Beth Benzie, Jyothsna Raj, Kiran Kumar, Sandipan Ghosh, Suyash Kumar and Santhosh Meenhallimath for their participation. We also like to thank Ales Popovic who helped us in identifying the problem area of our study. Special thanks to Styliani Zafeiropoulou for providing suggestions for our thesis. Finally, we like to thank everyone who was involved and helped us to complete this research.

Athul Vijayan Nambiar

Dilip Prasad Nair

Content

Figures

Tables

1 Introduction

This chapter gives an insight to our study. Initially, we discuss the background and problem area of our study. Then the purpose, research question and limitations of this thesis are discussed. This chapter concludes with describing research design and structure of the thesis.

1.1 Background

Research has shown that nearly half of all the anticipated values of all business projects is never achieved (Friedman and Smith, 2011). This is attributed to poor data quality in both planning and execution phases of these initiatives as the primary cause of failure. Poor data quality affects efficiency, risk mitigation and agility by compromising the decisions made in each of these areas (Friedman and Smith, 2011). To identify these issues, we need to recognize what is data quality and understand how it affects business. With the advent of big data, volumes on the range of exabytes and more are generated every day. Data, information and knowledge are created, collected and utilized at an extreme rate that is rapidly increasing by the day (Labrinidis and Jagadish, 2012). With its increase, data analysis is considerably more challenging in locating, identifying, understanding and citing data (Labrinidis and Jagadish, 2012).

Organizations that use high quantity of data face numerous challenges in maintaining the data quality and it is also difficult to come up with business decisions based on the inaccurate and unstructured data (Aloysius et al., 2016). High data quality is the pre-condition to any organization's success in guaranteeing the value of data. With the advent of significant technologies like cloud computing, Internet of Things (IoT) and social media, the amount of data generated is growing exponentially (Cai and Zhu, 2015). This makes it important to have better management of the data quality and thereby improve the quality of their decisions and solutions.

Decisions making based on data is a phenomenon that is rapidly growing within the business world. And currently with the success of big data and its associated technologies, a decision maker or a manager can make business decisions more effectively (Chaudhuri et al., 2011). As specified earlier taking decisions can sometimes be risky with the possibility of data being inaccurate or inadequate (Chaudhuri et al., 2011). However, maintaining and governing the data is of vital importance for organizations because of the data being unstructured and too complex sometimes (Blumberg and Atre, 2003). Hence, multiple data quality assessment frameworks are available that can provide guidelines to determine data quality in various industries.

There are many characteristics that determine data quality. According to Eppler (2006), accessibility, traceability, accuracy, timeliness, comprehensiveness, convenience, correctness, security, applicability, and maintainability are some of them. But these factors can be conflicting to each other and there will be some trade-offs required between the characteristics in certain scenarios. For instance, let us consider the characteristics, accuracy and timeliness. When

the information has to be current, the time available to check the accuracy of the information is less and the outcome may not be as expected (Eppler, 2006). Similarly, there exists a trade-off between accessibility and security. Organizations must decide what all data should be accessible and what need to be secured. Therefore, it is up to organizations to decide the characteristics that determine data quality that are relevant to them for making right business decisions.

How good is a company's data quality? Answering this question requires usable data quality metrics. Currently, most data quality measures are developed on an ad hoc basis to solve specific problems, and fundamental principles necessary for developing usable metrics in practice are lacking (Kaisler et al., 2013). In this article, we describe principles that can help organizations categorize data content under usable data quality metrics.

1.2 Problem Area

Even though information quality is critical in every organization, the field lacks a perfect framework or methodology for its evaluation and improvement. Data preparation and analysis become a complex process as the volume, variety, velocity and veracity of the data increase (Chen et al., 2012). As organizations rely on their data to make business decisions, the process of maintaining quality of data got utmost importance. There are number of business intelligence tools to rely on for data transformation and data preparation. Yet, organizations can end up in making wrong decision when the quality of the data prepared is poor. Inaccurate and poor quality data can lead to false bias for the decisions makers and thereby, it leads to wrong business decisions that consequently affect the performance of the organization (Griffith et al., 2008).

A new challenge emerging for organizations is 'quality vs quantity' with companies acquiring more and more data every day. (Kaisler et al., 2013). This can be because companies consider that with more data, they can perfectly explain whatever phenomena they are interested in. But this is not always true. With more data, sometimes there is more uncertainty and confusion. Organizations, both large and small, should ideally be investing in better understanding their data rather than spending enormous amounts of time and money in trying to correct the parts that are unable to provide its worth.

The field of our study is retail industry and it is an environment which is changing more rapidly than ever before. There is a need for retailers to differentiate themselves by exceeding their customer's expectations and thereby exceeding their competitors (Dabholkar et al., 1995). Every industry needs to understand what is data quality and how good quality data can deliver them better results. Particularly in the retail industry, with data generated from multiple sources, organizations must be on their alert to maintain and enhance the quality of data (Aloysius et al., 2016). In the retail domain, huge investment on data quality maintenance processes is worth to have as it can affect the business of the organization.

Retail is involved in many development processes to improve data quality and with the humongous amount of data generated through customer transactions, vendor information, and internal and external feeds, it is imperative to leverage the data for business growth (Aloysius et al., 2016). Currently, retail incorporates both Brick and Mortar and Ecommerce. Brick and Mortar incorporates physical retail stores and Ecommerce comprises of retail in online world.

The data quality measures can have minor variance between these two. We have identified that there lacks an extensive research on data quality characteristics in retail domain. Therefore, our study aims to explore on these lines.

1.3 Purpose of the Study and Research Question

The main purpose of this study is to analyse data quality measures that are followed by companies in retail industry and come up with a metric of key data quality characteristics trade-offs. This study will examine how organizations deal with several conflicting factors or the trade-offs between characteristics which decide the quality of data. We will choose the most principal characteristics that contribute to data quality from multiple data quality assessment frameworks. Based on those characteristics, the data quality measures of some of the organizations in retail domain will be considered.

Therefore, our research question will be formed as below:

How do organizations accomplish trade-offs in data characteristics to maintain data quality?

1.4 Delimitations

Data quality is a very wide spread topic of discussion and is important to many fields such as healthcare, manufacturing, banking and insurance etc. There are many characteristics that contribute to good data quality and they change based on the domain where certain characteristics of data are more important over the others. The scope of this research work concentrates entirely on understanding the characteristics that constitute to data quality in the retail domain. Also, there are multiple frameworks that help in understanding data quality for retail industry. In this research, a few frameworks have been chosen to cater to the needs of the topic with respect to the selected industry.

We have identified a few candidates who will be able to contribute significantly in our research work. These candidates have key role in data quality management and data analysis in the retail domain. This research is based on their insights and knowledge regarding the importance of data quality management in their organizations. Since there is a time limitation for this study, only a few companies have been chosen. The retail organisations selected for this study are either American or Indian. One of the markets leaders, Target that deals with enormous data transactions will be the primary focus for this study. Therefore, generalizability for this research is restricted due to all these delimitations.

1.5 Research design/ Thesis structure

The figure one (1) represents the different stages in this study. The initial literature review results in identifying key characteristics involved in data quality measurements. The interviews will be focused on analysing how organizations are maintaining data quality by meeting the standards of those characteristics. A conclusion will be reached after analysing the interviews and literature.

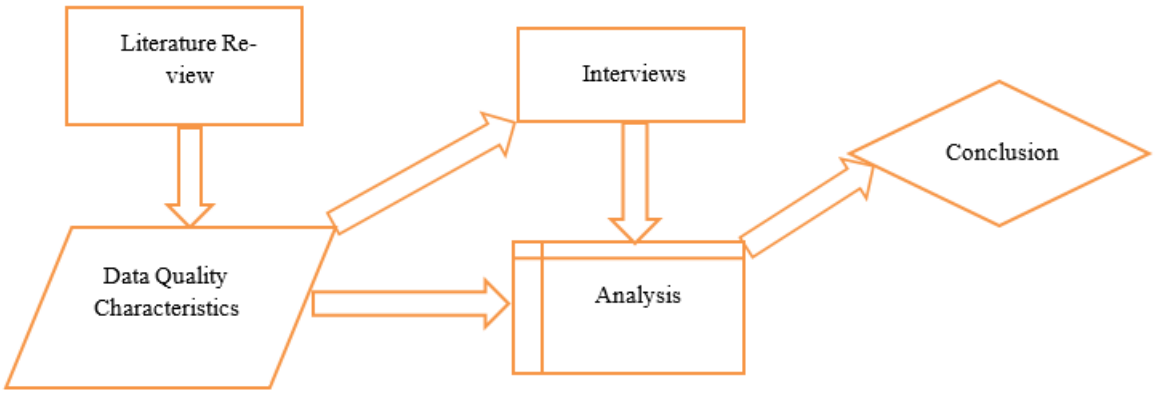

Figure 1: Research Design

This thesis will have the following structure:

Introduction – We will discuss the background of this study and then followed by illustration of problem area. Further, the purpose and research question of the study will be identified. Finally, the delimitations of this thesis will be discussed.

Literature Review – The thesis is focusing on identifying key characteristics that decide data quality in organizations. Hence, the important frameworks on data quality will be found out in this section and we will decide what are the key characteristics or factors that affect data quality.

Research Method – In this section, we will discuss how the study was conducted. Important aspects like data collection methods, selection of informants and data analysis methods will be explained.

Empirical Studies – This chapter includes the analysis of the interviews conducted.

Discussion – The analysis from the chapter, Empirical Studies is examined based on the theories from literature review to answer the research question.

Conclusion – This section concludes our findings of the study, the answer of our research question and contributions to future studies in this field.

2 Literature Review

2.1 Introduction to Data Quality

As the world is moving ahead in big data era, data quality is becoming a concern for every organization (Abbasi et al., 2016). The rich data available in many forms, force organizations to find structured and innovative ideas to deal with data quality issues (Albala, 2011). Since multi source systems dump unstructured data to warehouse, data quality management becomes a complex process. The sources of poor quality data are numerous. Data entry by employees and customers, external data sources, poor data migration processes and system errors are some of them (Eckerson, 2002).

Data can be described as *"primary base of information that describes real world objects in a format that can be stored, retrieved, and processed by a software procedure, and communicated through a network"* (Zahedi Nooghabi and Fathian Dastgerdi, 2016, p.185). Data can be structured, semi-structured and unstructured (Zahedi Nooghabi and Fathian Dastgerdi, 2016). The diagram in figure two (2) shows data categorization based on its quality.

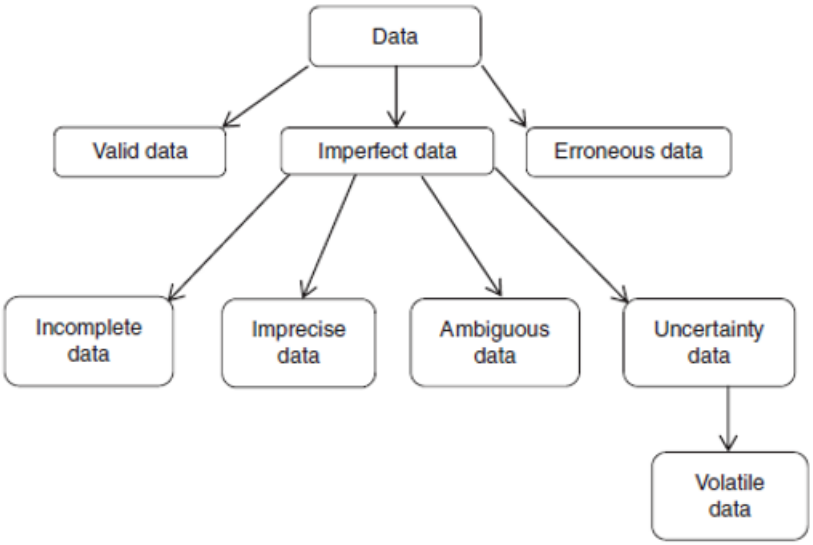

Figure 2: Types of Data by Quality (Laitio, 2011)

Valid data is considered as perfect and erroneous data is marked as false. (Laitio, 2011). The imperfect data can be transformed to valid, but it fails sometimes and thus become erroneous. The imperfect data can be categorized to different types and they are incomplete, imprecise, ambiguous and uncertainty data. Incomplete data is not satisfactory for making decisions even if they are transformed to valid data. Imprecise data is inaccurate and it is unusable in most of the organizational business scenarios. When there are multiple interpretations of the same data, possibly from various sources, then the data is ambiguous. The final category, uncertain data is due to lack of truthfulness. If the data is subjected to rapid changes, then it becomes

volatile data and organizations don't want to depend on these categories as decisions made on these may not be valid long term.

The direct access of information by users have increased the importance of data quality in organizations (Lee et al., 2002). The issues of poor data quality include time loss, delay in deployment, revenue loss, customer dissatisfaction and loss of credibility for the brand (Eckerson, 2002). A few important reasons for data quality issues are (The Interaction Design Foundation, 2017, p.1):

- "Widespread data
- Simplicity of creating, duplicating and sharing of information online
- The exponential increase in channels to receive information by; radio, television, print media, websites, e-mail, mobile telephony, RSS feeds, etc.
- The increasing weight of historical data available
- High volumes of conflicting, contradictory and plain old inaccurate information
- No simple methodologies for quickly processing, comparing and evaluating information sources
- A lack of clear structure in groups of information and poor clues as to the relationships between those groups"

There are instances where data quality becomes important during processes such as customer matching, corporate house-holding and organization fusion. (Batini and Scannapieco, 2016). Since the source is heterogeneous in most of the cases, the data provided by those systems can be overlapping. Same customer profile can be in different source systems. Retrieving unique list of customers can become a complex task in organizations. Corporate house-holding relates to establishment of relationship between members of households to reconstruct relationship and improve marketing strategies. Here also, merging similar data becomes an issue. Another instance is organization fusion which requires integration of organization's legacy systems, be it merging of different organizations or inter department fusion. This integration demands high quality effort and system compatibility to handle the increased volume of data.

The processes like data quality benchmarking, maintenance and assessment gain more importance than ever before as data quality impacts business processes, productivity and business decision making (Friedman and Smith, 2011). The challenge for organizations is to develop a model to assess the quality of the data they are dealing with (Lee et al., 2002). A disciplined approach for data quality management will improve organization's productivity and customer satisfaction (Geiger, 2004).

2.2 Data Quality Dimension

To define and understand data quality we need to understand and identity the definition of data quality dimension. What is a Data Quality dimension? A data quality dimension is a recognized term used by data management professionals to characterize a feature of data that can be evaluated, measured or assessed in comparison with defined standards to establish the quality of data. (Askham et al., 2013). Organisations select the data quality dimensions and associated dimension thresholds based on their business context, requirements, levels of risk etc. Note that each dimension is likely to have a different weighing and to obtain an accurate

measure of the quality of data, the organisation will need to determine how much each dimension contributes to the data quality.

According to Askham et al. (2013, p.5) a typical data quality assessment approach will have the following stages:

- "Identify which data items need to be assessed for data quality, typically this will be data items deemed as critical to business operations and associated management reporting.
- Assess which data quality dimensions to use and their associated importance.
- For each data quality dimension, define values or ranges representing good and bad quality data.
- Apply the assessment criteria to the data items.
- Review the results and determine if data quality is acceptable or not.
- Where necessary, take corrective actions; e.g. clean the data and improve data handling processes to prevent future recurrences.
- Repeat the above on a periodic basis to monitor trends in Data Quality".

For any retail organization, it is important to take consumer viewpoint of quality because ultimately it is the consumer who will judge whether a product is fit for use. Wang and Strong (1996) defined a data quality dimension as a set of data quality attributes that represent a single aspect or construct of data quality. Eppler (2006) identified the important characteristics that determine data quality and those characteristics are shown in the figure three (3). Our study will be focusing on some of these characteristics that are important in the retail domain.

1. Comprehensiveness
2. Accuracy
3. Clarity
4. Applicability
5. Conciseness
6. Consistency
7. Correctness
8. Currency
9. Convenience
10. Timeliness
11. Traceability
12. Interactivity
13. Accessibility
14. Security
15. Maintainability
16. Speed
17. Objectivity
18. Attributability
19. Value-added
20. Reputation (source)
21. Ease-of-use
22. Precision
23. Comprehensibility
24. Trustworthiness (source)
25. Reliability
26. Price
27. Verifiability
28. Testability
29. Provability
30. Performance
31. Ethics/ ethical
32. Privacy
33. Helpfulness
34. Neutrality
35. Ease of Manipulation
36. Validity
37. Relevance
38. Coherence
39. Interpretability
40. Completeness
41. Learnability
42. Exclusivity
43. Right Amount
44. Existence of meta information
45. Appropriateness of meta information
46. Target group orientation
47. Reduction of complexity
48. Response time
49. Believability
50. Availability
51. Consistent Representation
52. Ability to represent null values
53. Semantic Consistency
54. Concise Representation
55. Obtainability
56. Stimulating
57. Attribute granularity
58. Flexibility
59. Reflexivity
60. Robustness
61. Equivalence of redundant or distributed data
62. Concurrency of redundant or distributed data
63. Nonduplication
64. Essentialness
65. Rightness
66. Usability
67. Cost
68. Ordering
69. Browsing
70. Error rate

Figure 3: Data Quality Attributes (Eppler, 2006)

2.3 Data Quality Standards: Frameworks to Assess Data Quality

According to Cai and Zhu (2015) data quality depends on its own features and the business environment of the data being used. To assess data quality there are certain questions need to be answered and they are (Dedeke, 2000, p.126):

> *"What do the data permit consumers to know?*
>
> *What do data permit the consumers to do?*
>
> *What degree of effort and time is needed before the consumer could achieve desired outcome from data?"*

There is not enough specific framework or criterion to measure the quality of the data by answering the above questions. Researchers have come up with many models or frameworks in these years. We examine some of the important frameworks that are relevant to our study. The frameworks that are considered have many commonalities between them and categorized similar data quality dimensions under different taxonomies.

2.3.1 Framework by Wand and Wang (1996)

One of the initial and well accepted frameworks in this field is developed by Wand and Wang (1996). They came up with a set of dimensions to measure the characteristics that define data quality. As shown in figure four (4), the Wand and Wang's framework is divided into two; external and internal view of the data quality dimensions. Internal view is related to design and operation of the information system. (Wand and Wang, 1996). It includes data characteristics like accuracy, timeliness, completeness, consistency, precision and reliability as system characteristic. External view determines the use and value of the data provided by the information system. Some of the data related characteristics are timeliness, content, usefulness, scope, interpretability and understandability. System related characteristics include timeliness, flexibility, format and efficiency. This framework helps to distinguish characteristics in internal and external viewpoints of a data quality management system.

View	Dimensions
Internal view	**Data related:** accuracy, reliability, timeliness, completeness, currency, consistency, precision **System Related:** Reliability
External view	**Data related:** timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, freedom from bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability **System Related:** Timeliness, flexibility, format, efficiency

Figure 4: Division of dimensions (Wand and Wang, 1996)

2.3.2 Conceptual Model for Data Quality Measurement (Stvilia et al., 2007)

Stvilia et al. (2007) introduced a framework which identifies sources of data or information quality problems, types of activities involved in data management process, data quality attributes and context of those attribute's practice. It is shown in figure five (5). The integral part of the framework is the taxonomy of the data quality dimensions. Characteristics are divided into three categories; intrinsic, relational and reputational. Intrinsic characteristics include general and conventional attributes. (Stvilia et al., 2007). Relational refers to immediate context of data quality measures and reputational taxonomy relates to organization or community. The framework connects sources of data quality problems to these characteristics and to types of activities. Thereby, the framework provides a meaningful mechanism to understand the data quality issues and the reasons for it.

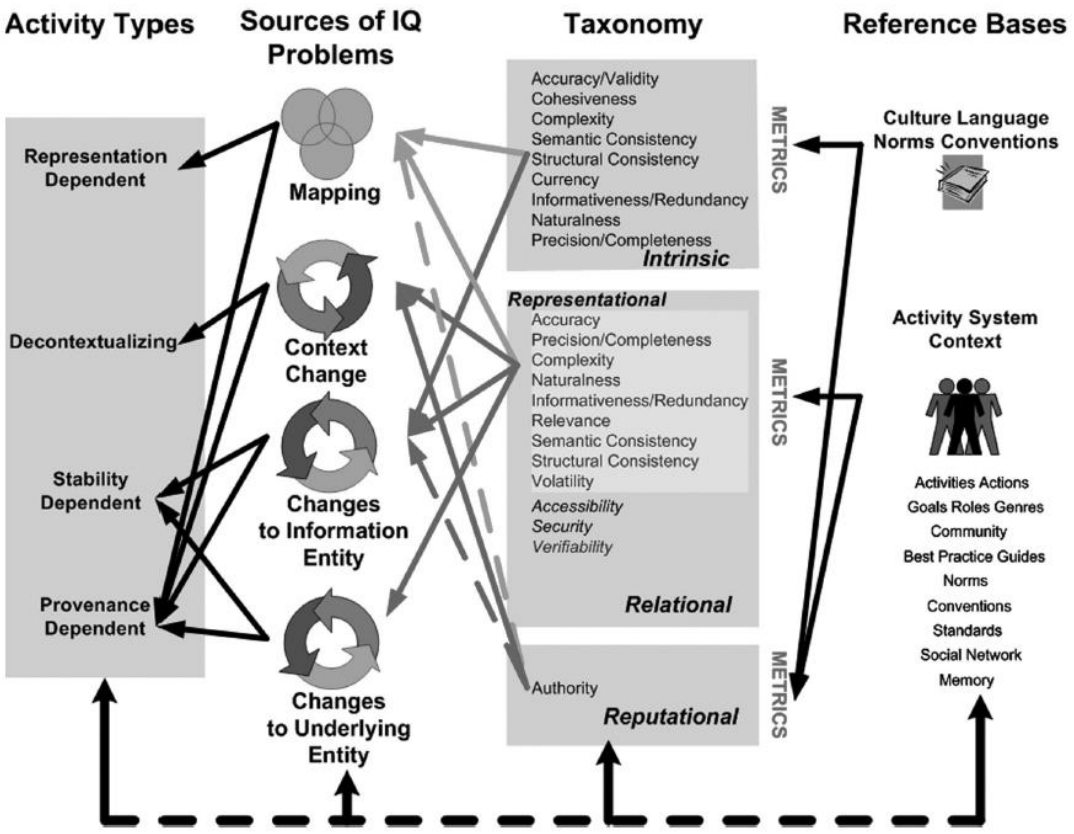

Figure 5: Conceptual Model for Data Quality Measurement (Stvilia et al., 2007)

2.3.3 Cai and Zhu's Framework (2015)

Another important framework which depicts data quality standards is universal two-layer framework developed by Cai and Zhu (2015). Their framework is shown in figure six (6). In this framework, main dimensions are further divided into sub elements. The primary characteristics, availability, usability, reliability, relevance are inherent characteristics of data quality. (Cai and Zhu, 2015). Presentation quality which is further divided into readability and structure, relates to satisfaction of the customers. Availability includes accessibility, timeliness and authorization. They together determine the degree of convenience for data access. Usability indicates level of acceptance among users. It includes definition of the data, credi-

bility and metadata. Reliability which consists of accuracy, integrity, consistency, completeness and auditability, refers to user's trust on the data provided. Another dimension in the framework is the relevance of the data. This framework covers important aspects of data quality and they are well structured in a two-level hierarchical way.

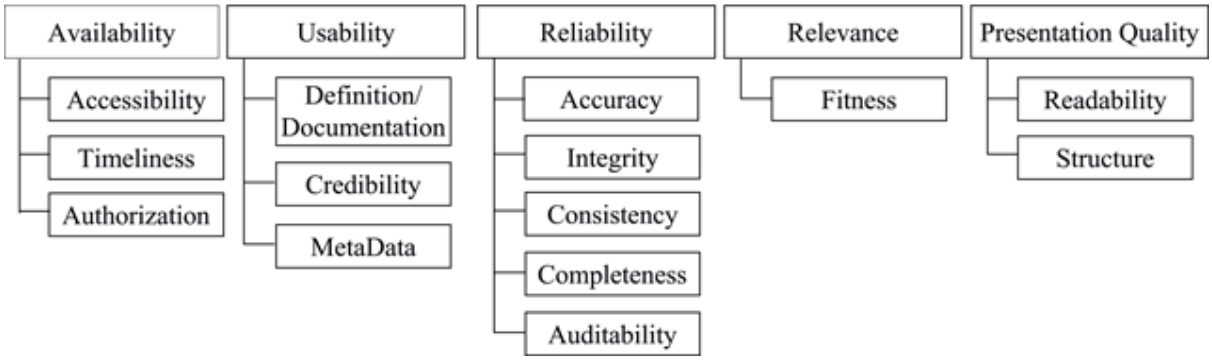

Figure 6: Framework on data quality standards (Cai and Zhu, 2015)

2.3.4 Eppler's Framework (2006)

The framework by Eppler (2006) considers several factors affecting data quality and their potential conflicts. The diagram in figure seven (7) shows the Eppler's framework. The four-level framework splits the characteristics as relevant information, sound information, optimized process and reliable infrastructure.

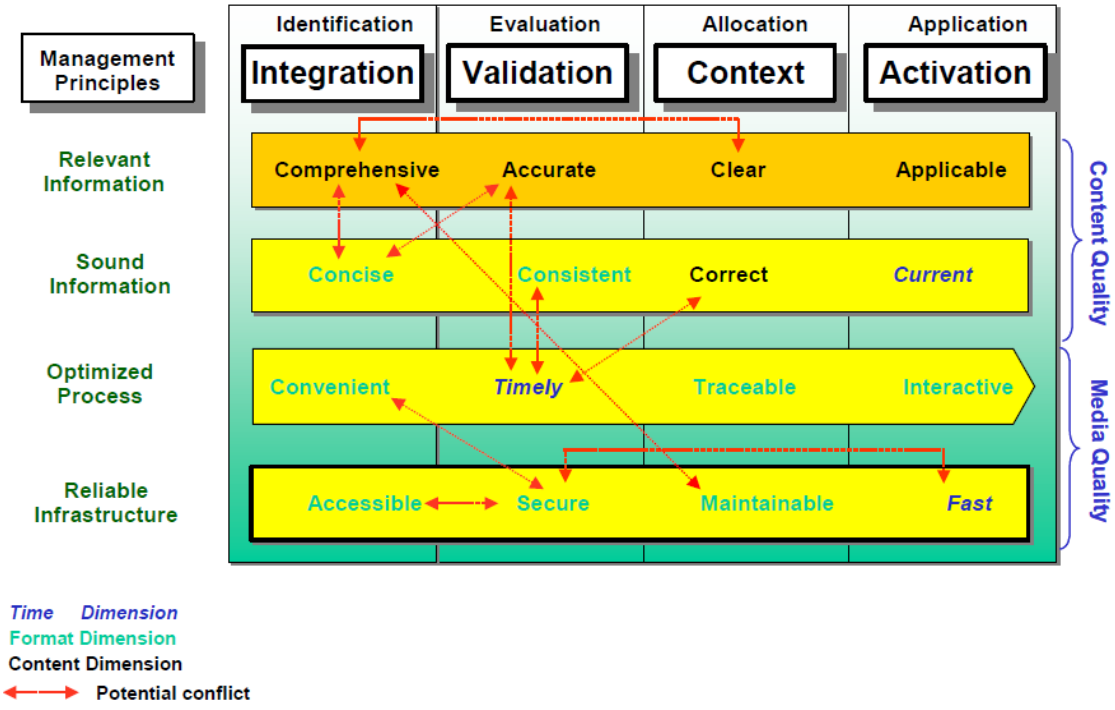

Figure 7: Eppler's framework on data quality (Eppler, 2006)

The upper two layers, relevant information and sound information, are related to the actual information or the content itself and hence called as content quality. (Eppler, 2006). Relevant information refers to comprehensiveness, accuracy, clarity and applicability of data and sound information determine whether information or data is concise, consistent, correct and current. The next two layers, optimized process and reliable infrastructure are about management of data and therefore, they are referred as media quality. Process evaluation consist of measuring convenience, timeliness, traceability and interactivity of the data and infrastructure reliability is measured in terms of the accessibility, secureness, maintainability and retrieval speed of the data.

The potential conflicts or trade-offs between characteristics are shown in the framework. For example, consistency and timeliness of data conflict each other. (Eppler, 2006). Data should be consistent across the process irrespective of the time taken for the process completion. Similarly, there is a trade-off between timeliness and accuracy. When data is provided in a short span of time, the accuracy of the data cannot be guaranteed. Proper testing of the data may not be completed within the time limit. An infrastructure level conflict occurs between accessibility and security. Organizations decide what data should be made accessible and what should be protected. They also need to determine the level of accessibility in each category of data so that it doesn't affect the security system.

2.4 Data Characteristics Trade-offs

In the previous section, we discussed the important data quality characteristics that are relevant to our study. As mentioned earlier, there exist some conflicts and hence, trade-off or balancing is required between those characteristics in different business scenarios in an organization. In this chapter, we are examining the significant trade-offs mentioned by Eppler (2006) between data characteristics and they are explained below.

2.4.1 Accuracy and Timeliness

One of the important characteristics trade-off that organizations deal with is between accuracy and timeliness. Eppler (2006, p.53) explains this trade-off as *"the more current a piece of information has to be, the less time is available to check on its accuracy"*. To handle the accuracy-timeliness trade-off in organizations, decisions needed to be taken on using current but inaccurate or delayed/outdated but accurate information (Ballou and Pazer, 1995). The timely and accurate information delivery depends on the process and infrastructure used and organizations are influenced by the cost involved to balance these factors (Eppler, 2006). Since the inaccurate or outdated information will not be accepted by customer, organizations can't use the information for decision making if desired balance/trade-off is not met between these characteristics (Eppler, 2006).

There are business scenarios where timeliness plays an important role in organizations. For retail organizations, Monday morning report is crucial as it covers key information on their business of the previous week (Mondayreport.ca, 2017). A small variance in data measures is acceptable in this case as organizations are concerned more on getting an overall trend of their KPIs (Key Performance Indicators) over the previous week. When it comes to information on customer profiles, finance data and sales measures, accuracy becomes the key factor than

timeliness. Therefore, organizations need to follow different trade-off rules and mechanisms in different business situations.

2.4.2 Consistency and Timeliness

During the development of a Real-Time Database Management System, we need to understand the trade-off between data consistency and currency or timeliness of data. To achieve this, a very dynamic information system needs to be developed that can identify the behaviours and properties of a real-time system (Cai, 2015). According to Nelson et al. (2005, p.204), data currency is,

"the degree to which information is up-to-date, or the degree to which the information precisely reflects the current state of the world that it represents."

In the current data rich environment, every data warehouse system has data flowing in from various sources including internal systems, external sources and social media sites. The concept of consistency states that data that enters a data warehouse must be normalized to help reduce data redundancy (Date, 2006). To illustrate data consistency, an example is provided below.

Genus	Species	Infraspecies
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

Inconsistency in the Infraspecies field

Genus	Species	Infrasp_rank	Infraspecies
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

Consistency in the Infraspecies field by addition of second field

Figure 8: Data Consistency (Chapman, 2005)

In the figure eight (8), the first table has inconsistent data in the field, Infraspecies. In some records there exists a prefix along with the name (example: subp.bicostata). To make this field consistent across the table, another field, Infrasp_rank is added to the table which is shown in the second table in the figure eight (8). The prefix part of the record is separated and populated in another field and there by data consistency is achieved in the table.

The data architect creates complex scripts that help in Extracting, Transforming and Loading (ETL) data. (El-Sappagh,2015). And, as the sources of data increases, the ETL scripts become more and more complex resulting in time consuming and longer data load time resulting in the conflict of interest between data consistency and currency.

2.4.3 Accessible and Secure

Accessibility and security are opposite poles when it comes to data characteristics and in the trade-off between the two, compromises need to be made. When an information system is more secure, the overall convenience to access data reduces and vice versa (Eppler, 2006). Constructing such a system which is both accessible and secure raises many crucial questions when it comes to balancing these data characteristics and finding the right trade-off between the two data quality attributes is not an easy endeavour. (Braz et al., 2007). Both accessibility and security of an information system can vary based on the context of use which includes users, tasks, software, hardware and environments.

As far as accessibility and security of data is concerned in retail domain, Financial, HR and Customer data are considered vital for an organization. Access is not provided to all employees and layered security is maintained so that the data isn't hacked or tampered with by malicious users. (Herath and Rao, 2009). On the contrast, Merchandising data which includes Item data, Sales data and Inventory data is considered accessible and less secure. This data isn't protected as much and every employee has access to it. Accessibility for this data is crucial because this contributes to the business.

2.4.4 Convenience and Secure

The convenience of data is often closely related to interactivity (Eppler, 2006). *"Convenience designates the ease of- use or seamlessness by which information is produced, administered and most importantly acquired"* (Eppler, 2006, p.78). Convenience is about perceptions of customers on usefulness and ease of use of the organization's products and interacting channels (Aloysius et al., 2016). Eppler (2006) mentions that, too much security can reduce the convenience and organizations need to make choices to achieve the balance between secured and convenient information. The main concept that caters to data warehouse security is CIA; confidentiality, integrity and availability. (Batra and Arora, 2016). Confidentiality means that only the authorized users can access the data from the data warehouse. Integrity refers to the originality of the data which is if the data has been received from authentic resources and availability refers to whether data is available always to the designated users. All these three factors have significant role in determining convenience of information usage.

2.4.5 Accuracy and Conciseness

The term accuracy is self-explanatory, where it can be used to designate to the notions of precision, level of detail and/or correctness of data (Eppler, 2006). Another form to represent accuracy is to define the closeness between value 'a' and value 'b' wherein value 'a' being the value in the source and the value 'b' is its representation in the data warehouse. (Batini et al., 2009). In such situations, we need to consider the value 'a' from the source to be correct and precise in nature.

The characteristic, Concise indicates whether the data is to the point and devoid of unwanted elements (Eppler, 2006). According to Wang and Strong (1996) conciseness designates whether the data is compactly presented, well formatted and aesthetically pleasing. The trade-off between conciseness and accuracy occurs in terms of the level of detail. When more detail

is provided, then the less concise it becomes (Eppler, 2006). Organizations should make sure that the data is presented concisely without flouting accuracy.

2.4.6 Secure and Fast/ Speed

The characteristic, fast or speed relates to the response time of the infrastructure or the server of the organization (Eppler, 2006). It is the most important criterion related to information infrastructure as it impacts the applicability of the information (Eppler, 2006). The response time should match with the user's expectation. The users can be the employees or the customers of the organization.

The speed of information retrieval is in turn related to the security of the information. Data security in any domain is essential primarily due to the reasons mentioned below (Thota et al., 2017):

- To protect and prevent huge size of confidential government, business or governing data from malicious intruders and to predetermine threat.
- Lack of awareness and standards about how data providers maintain their data warehouses and protect confidential data.
- Lack of knowledge or standards about auditing and reporting of data that is shared on cloud.
- Users who might not even work for the organization, but may have control and visibility into the history and archival data stored.

Since there are multiple processes and checks involved in the data security assurance activities, the response time of the infrastructure to user can be affected. There exists a trade-off between security and speed which eventually reflects as response time of the infrastructure.

2.4.7 Comprehensive and Concise

Data being comprehensible and concise is a contradiction. (Eppler, 2006). If data is stored in data warehouse with high level of detail, the less concise that data is going to be and this holds in the vice versa situation as well. As the scope for the information being stored gets higher, it gets even more difficult to present it in a concise format and the pursuit of comprehensive data may also lead to decrease in the clear distinction between central and peripheral information thereby making the data unusable for business decisions.

According to Wang and Strong (1996), data quality aspects incorporates the format of data which needs to a combination of both concise and comprehensiveness. The two characteristics suggest that data consumers need data that is not only well represented but also require it in a concise and non-repetitive manner which is easy to understand (Wang and Strong, 1996).

2.4.8 Comprehensive and Maintainable

Maintainability refers to the characteristics of data where it can be stored or accessed easily, thereby enabling fast interaction between the stored information and the users of data (Eppler, 2006). The factors that contribute to maintainable data are costs, data volume, frequency, quality and infrastructure (Eppler, 2006). Maintainability of data can be expressed in terms of

the probability of that a data system repaired and running within a specified period. Increased maintainability implies shorter repair times (Asq, 2017). With the increase of comprehensive data in a data system, the maintainability of data becomes more difficult. Therefore, maintainability of data is in direct conflict with comprehensiveness of data. The more information we have, it makes data more comprehensive but less maintainable (Eppler, 2006).

2.4.9 Comprehensive and Clear

Eppler (2006) states that data clarity of information is achieved when it is understandable to its target audience. Clear and concise data is cost effective and values the time of the audience by reducing information overload and misunderstandings (Eppler, 2006). But, with the advent of big data, the amount of data in every domain has been exploding and organizations require trillions of bytes of information about their customer, suppliers and operations and there is also data fed from sensors and social media (Manyika et al., 2011). These factors thereby contribute in reducing data clarity but increasing comprehensibility.

2.4.10 Timeliness and Correctness

As per the definition provided by Cappiello et al. (2004, p.71) timeliness of data or data currency is usually defined as:

"the time interval between the time instant in which data are updated and the time instant in which data are being used"

Data currency is a crucial factor in contextual information quality and it can be referred to the degree to which information is up to date or the degree to which the information accurately reflects the current state of the world that it represents (Nelson et al., 2005). The more current and real time the data is, the more business usable it becomes. But timeliness of data can sometimes be at conflict with the correctness of data. According to Eppler (2006, p.84), correctness of data can be interpreted by the question: *"Is the information free of distortions, bias and/or error"*. As stated by Pipino et al. (2002), the degree of correctness of data can be varied based on the user requirement. It can be acceptable to have misspelled string in one circumstance but not in another (Pipino et al., 2002). Therefore, to get data perfectly correct it takes time, resulting in a conflict.

An in-depth analysis of the important data characteristics trade-offs in various contexts will be covered in the chapter 4 of this research paper.

2.5 Result of Literature Review

In this chapter, we discussed the concept of data quality and how data quality management is handled in organizations. We also presented a few frameworks used to assess data quality. The frameworks, Wand and Wang (1996), Stvilia et al. (2007) and Cai and Zhu (2015) have characteristics in common and they are presented in diverse types of dimensions. Whereas, Eppler (2006) could illustrate the trade-offs between data quality characteristics along with the multi-layer dimension model. The trade-offs between the characteristics are shown in the

figure nine (9). We will be assessing how organizations deal with the important and conflicting characteristics and how trade-offs are achieved while maintaining data quality.

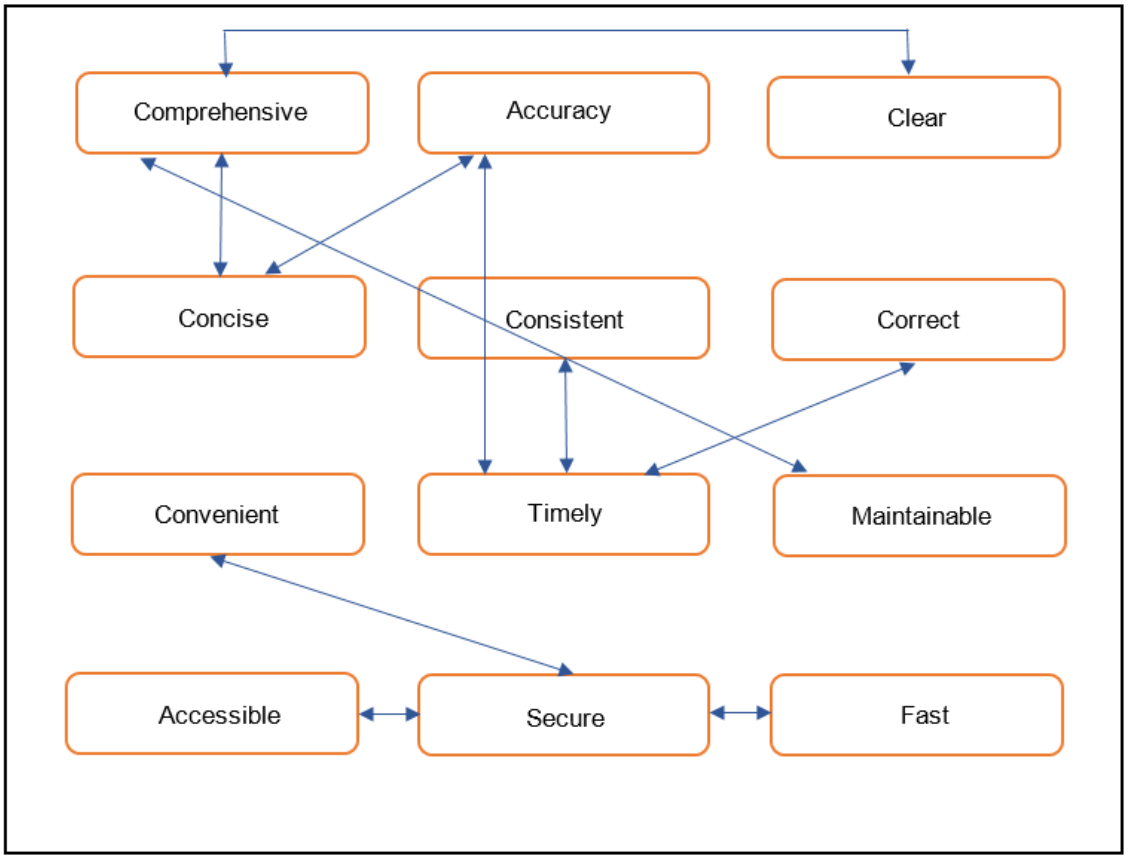

Figure 9: Data Characteristics Trade-offs (Eppler, 2006)

3 Research Method

In this chapter, we discuss the research method of our study. We give the readers an insight to our approach and data collection methods. Later, we present how we conducted the study and how the quality aspects like reliability, validity, ethics and research bias are maintained in the process.

3.1 Method Selection

As per Bhattacherjee (2012), exploration is the start of any research study. Once the research question was identified, we focussed on accumulating literature for the topic of discussion. We used Google scholar and Lund university library portal to get the materials related to our area study. Since we are focusing data quality management, we used keywords like data quality, data characteristics, and data characteristics trade-offs to find the relevant articles. Literature review has helped to find theories and information related to field of study. As suggested by Recker (2013, p.40), we have taken into consideration the below pointers to enhance our literature review process.

- *"Findings and insights into a specific problem domain*
- *Theories that are available and/or in use to examine the problem or phenomena of interest*
- *The current state of the methodologies appropriate and applicable to the study"*

This research covers the topic of data quality, data characteristics and data characteristics trade-offs. It is imperative to follow a research strategy that is suitable for the nature of study. There are multitudes of pre-existing research frameworks that help in defining data quality. But we, in this research, are focusing on using these frameworks in defining data quality for the retail industry and identifying which characteristics of data is more important over others in the retail domain. There is scope for an inductive study in this field and we proceeded with qualitative research.

"Qualitative methods depend on text which captures records of what people have said, done, believed or experienced about a phenomenon, topic, or event" (Recker, 2013, p.88). This approach has helped us to gather insightful information from interviews, observation and documentation, for data analysis we have used techniques like coding. The interviews have helped us in identifying the most important characteristics of data defining, data quality in the retail domain.

3.2 Data Collection

The data collection for our thesis was done by conducting qualitative interviews. Interview is a powerful way to understand a subject based on an individual's life experience (Seidman, 2013). The questionnaire was formed after the completion of literature review. Literature review provided us valuable insights to answer our research question. Since our study is on current practices and issues in data quality maintenance in organizations, we interviewed data

quality management experts using the questionnaire formed. Interviews provide more personalized data (Bhattacherjee, 2012). The interviews that we conducted were semi-structured in nature. This kind of interviews gives flexibility to interviewer as new questions can be brought up during the interview process (Recker, 2013). Semi-structured interviews encourage two-way communication and it provides opportunities for learning as the interviews provide more information apart from the answers to the pre-defined questions (Recker, 2013).

All the interviews we conducted were telephonic interviews as the interviewees were from various parts of the world. Before interviews, we prepared an interview guide which explains the interview structure and the topics of our research. We have sent out this guide to interviewees to make them aware of our objectives with the interview. This has helped us to keep the discussions in the interviews within our area of study. While interviewing we recorded the audio to examine what exactly the interviewee said during the interview. We transcribed the interviews using the recorded audio. In qualitative analysis, interview transcripts are used by researchers to analyse the qualitative data such as text data (Bhattacherjee, 2012). After completing the data collection process, we combined both the findings. It helped us in getting a clearer picture of the issue and thereby, arriving at the answer of our research question.

3.3 Informant Selection

Data gathering is very crucial in any research and for a better understanding on how to incorporate a theoretical framework in a real-time situation. Therefore, it becomes imperative in selecting the manner of obtaining data and from where the data acquired will make the most sense. It is mentioned by Bhattacherjee (2012) that, it is of utmost importance to choose informants who have in-depth knowledge of the topic that is being researched and they need to have a non-biased opinion on the subject as well. Since the research is being done on data quality and ranking the importance of data characteristics in the retail domain, we have decided to choose candidates who are well versed in the above topics.

The respondents or the interviewees are from the retail domain and they are knowledgeable about the data in the retail industry. They have experience in extracting, manipulating, mining, transforming, testing and analysing data in different fields within the retail industry. They also have a thorough understanding about these industries and its functioning as well. This thesis primarily focuses on organizations in retail domain that handle enormous quantities and varieties of data and uses business intelligence applications as tools to make decisions. The main targeted companies of this work are Target, Quotient Technology, Fashionara, Supervalu and Bluestem. The details of the contact persons and the interviews conducted are given in the table one (1).

Table 1: Interview Details

Name	Organization	Designation	Years of Experience	Interview details Date and Duration
Kiran Kumar	Target Corporation	Lead Quality Engineer	10 years	23rd April 2017 44 minutes

Sandipan Ghosh	Target Corporation	Senior Data Engineer	12 years	24th April 2017 38 minutes
Suyash Kumar	Quotient Technology	Senior Data Analyst	9 years	26th April 2017 48 minutes
Jyothsna Raj	Fashionara	Senior Data Analyst	9 years	1st May 2017 45 minutes
Santhosh Meenhallimath	Bluestem Brands	Lead Data Quality	10 years	2nd May 2017 34 minutes
Beth Benzie	Supervalu	Data Quality Manager	20 years	4th May 2017 47 minutes

3.4 Pre-Study

The data quality analyst job has become vital in many organizations because of the implementation of data warehouses, particularly in the retail industry which has a lot of data to process. Most organizations employ engineers who are experts in data extracting, mapping, transformation and loading to improve the data quality situation in their data warehouses. Therefore, a new role called the data quality analyst was created to help ensure the correctness of the data loaded in data warehouses (Pierce, 2003).

So, to understanding how trade-offs works within distinctive characteristics of data in the retail industry, it is imperative to identify the most important characteristics that define data in these industries. For this, we decided to conduct a pre-study to analyse the key factors that define data. In this pre-study, we identified 10 candidates with years of work experience as a data analyst in the retail industry. Their job profile includes:

1. Reviewing data loaded into the data warehouses.
2. Recommending maintenance to improve the accuracy of the data in the warehouses.
3. Analysing data trends to understand purchase patterns of consumers in the retail industry.
4. Educating and suggesting data hygiene techniques to the downstream users such as the reporting and analysis team.
5. Normalizing data from various sources to reduce or eliminate data redundancy in the warehouses.

The details of the candidates involved in the pre-study are provided below:

Table 2: Candidates of Pre-study

Name	Organization	Designation	Years of Experience
Swathi Kadoor	Target Corporation	Data Manager	12 years

Prem Tiwari	Target Corporation	Data Scientist	9 years
Kiran Kumar	Target Corporation	Lead Quality Engineer	10 years
Sandipan Ghosh	Target Corporation	Senior Data Engineer	12 years
Suyash Kumar	Quotient Technology	Senior Data Analyst	9 years
Jyothsna Raj	Fashionara	Senior Data Analyst	9 years
Kavin Raj	Consultant at INTTRA	Senior Tester	7 years
Karthika Panicker	Consultant at D&B	Data Engineer	6 years
Santhosh Meenhallimath	Bluestem Brands	Lead Data Quality	10 years
Devika Ghosh	Target Corporation	Data Quality Analyst	7 years

Based on the study conducted by Eppler (2006), 70 data quality attributes were found to be important and they are shown in figure three (3). These data characteristics identified were presented to data analysts listed in table two (2) and they were asked to choose the most key features that define data quality in the retail sector. The following factors were most commonly selected by the candidates.

- Accuracy
- Timeliness
- Secure
- Consistency
- Accessibility

The above listed 5 characteristics are defined below:

Accuracy: Precision, level of detail and correctness (Eppler, 2006).

Timeliness: Time from creation to publication: Is the information processed and delivered rapidly without delays (Eppler, 2006).

Accessibility: Level of Access: Is there a continuous and unobstructed way to get to the information (Eppler, 2006).

Secure: Is the information protected against loss or unauthorized access (Eppler, 2006).

Consistency: Is the information free of contradictions and uniform across data processing levels (Eppler, 2006).

From the Eppler's framework, we identified three trade-offs between these five characteristics and they are shown in the figure ten (10). We will be evaluating how organizations in retail domain deal with these important and conflicting factors while maintaining data quality.

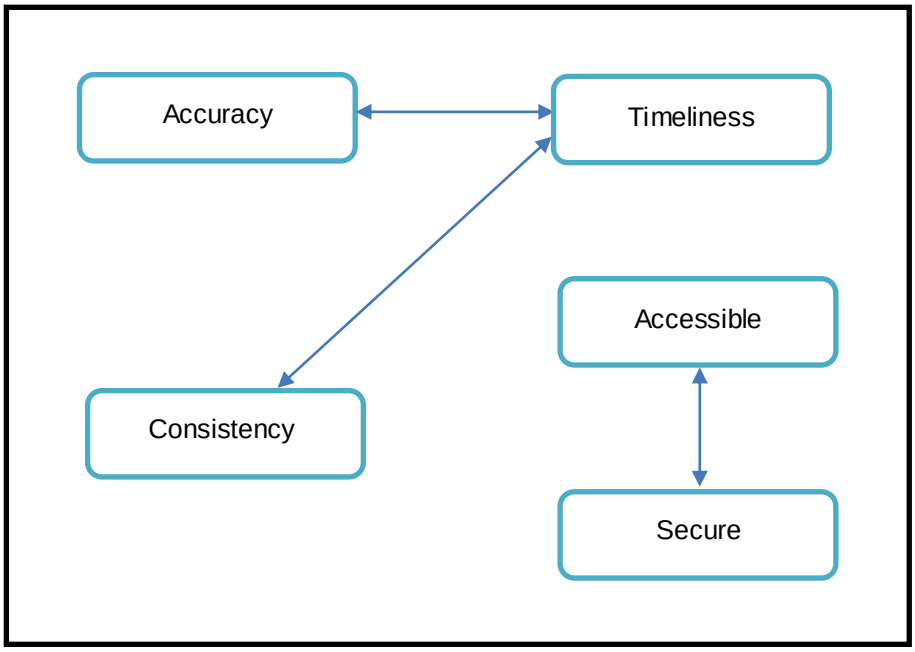

Figure 10: Important Data Characteristics Trade-offs in Retail

3.5 Writing up the Study

It is good to have a thesis structure that is common and familiar to the readers (Recker, 2013). For the manuscript of this thesis, we have followed the structure and content explained by Recker (2013). According to Recker (2013), the general writing process should have the following steps: develop a structure, revise the structure, start populating the sections, revise the sections, revise the paper or thesis, pre-submit to a reviewer and finally submit the paper. The final paper has the important sections such as Abstract, Introduction, Background, Research Model, Literature Review, Research Method, Discussion, Conclusion, Practical Implications and Knowledge contribution.

This research paper will give insights about data quality characteristics trade-offs to the readers. In the introduction of this paper, we discussed about the background, problem area and purpose of this study. The literature review gives better understanding of the research question. Then, research method explains how we conducted this study. The analysis of the interviews will be presented to the readers in the next chapter. We will discuss the findings in the chapter, discussion. Eventually, we present our conclusion and our knowledge contribution to the field of information systems. It helps to give a closing frame for this study. To improve quality and to keep ethics of this study, we have appended the interview transcripts in the appendix after all the chapters. We followed the guidelines from Recker (2013) to meet the standards of a paper for publishing.

3.6 Interview Procedure

3.6.1 Profile of Interviewees

As mentioned by Myers and Newman (2007), a qualitative interview can help a researcher in analysing all kinds of topics, be it, positivist, interpretive or critical. The qualitative interview is the most commonly used method in data gathering for qualitative research. The qualitative interview has been used as an instrument for providing us with description, narratives and texts which will be interpreted and analysed based on the interests of the research topic (Kvale, 2006).

Helping us to contribute in our research topic, we identified a few candidates who have in depth and extensive technical knowledge in the retail domain. For this thesis, a total of six interviews were conducted. The informants were from India and USA and that made it difficult to have face to face interaction. To minimize this disadvantage, all the interviews were done through telephone and audio was recorded for transcribing the interview. We believe the interviewees were open and we could extract the required information from them.

The first interview was conducted with Kiran Kumar, who has approximately ten years of experience in Information Technology. Out of the ten years, he has worked extensively in the retail industry at Target corporation. He is currently employed as a Lead Quality engineer and his job role includes maintaining data quality and assuring the right people get the right data to make business decisions. His knowledge and expertise in data quality and retail assisted us in obtaining clarity in our field of research.

Our second candidate of interest for our research was Sandipan Ghosh. He has a total experience in the Information Technology domain for twelve years and has worked in the retail industry for nine years in multinational companies like Target and Tesco. He has extensive experience as a database administrator, database engineer, quality analyst and Hadoop developer in big data. He is currently working in Target corporation as Senior Data Engineer and Solution Analyst.

The third interviewee who contributed to our research topics was Suyash Kumar. He has a total of eight years of work experience having worked in companies like Infosys, Target and Quotient technologies. He has worked on multiple data warehouse technologies and big data concepts and applications such as Hadoop. He is currently working as a Senior Data Quality analyst at Quotient technologies since April 2017 and was a Data Quality analyst in target for six years.

Jyothsna Raj was our fourth candidate for improving our understanding of the research topic. She has approximately ten years of technical work experience out of which seven plus years she has worked in the technical side of both Ecommerce and brick and mortar companies. Jyothsna Raj has immense experience in data analysis, data reporting, data governance and slicing and dicing data. In target, she worked as a Data Quality analyst for two years followed by 4 years in Fashionara as a Senior Data Analyst.

The fifth interview was conducted with Santhosh Meenhallimath. He has nearly ten plus years of technical experience in Information Technology with close to seven years of experience in the retail domain. He has worked for 6 plus years as a data analyst in Target and is currently

working as a Team Lead in Bluestem since January 2016. Apart from data quality, he is currently playing a lead role wherein he makes sure that data quality process and standards are maintained.

Our sixth and final interview was with Beth Benzie. She has a working experience of 20 years in retail domain and in that, she worked in the field of data quality management for 7 years. She currently works for Supervalu as data quality manager. Previously, she worked for other retail companies like Target and Bluestem. Beth Benzie has the responsibility of setting up the data quality team and building a data quality assessment framework in the current organization she works for.

Based on our interview list, it can be clearly understood that we have obtained a wide range of data experts and data stewards who are well versed and knowledgeable in the field of data management in the retail domain. They are the Subject Matter Experts (SME) or Centre of Excellence (CoE) in their field of work and have both business and technical understanding in the various subsections that contribute to the functioning of a retail industry.

3.6.2 Interview Guide

The interview guide was formed based on the results from literature review and pre-study. Research questions were based on the field of data quality management. Our experience in the field of retail also helped us in forming the questionnaire. The interview guide consists of four sections. They are explained below.

Section 1: Introduction

Initially, we introduced ourselves to interviewees and asked for consent to record the audio. We asked questions to get an overview of candidate's experience in the field of our study. Later, interviewees were asked to describe their role and activities in the organization they work for.

Section 2: Defining Data Quality

The second section has questions on data quality. We asked the interviewees to explain their understanding of data quality and the importance of data quality in their organization. Candidates were also requested to identify the main characteristics defining data quality. This section helps us to give an insight to candidates on the focus area of our thesis.

Section 3: Data Characteristics Trade-offs

In this section, the questions were formed to explain the three trade-offs in data quality characteristics. The trade-offs, accuracy and timeliness, consistency and timeliness, and accessibility and security were discussed. Candidates were asked to explain the importance of data characteristics and trade-offs. Thereafter, we told them to illustrate a scenario where the trade-offs occur in their organizational business processes. This is the most important section in our interview as it gives information regarding the trade-offs in data quality characteristics.

Section 4: Suggestions and other information

This section concludes the interview. We ask candidates to give suggestions for other data qualities and trade-offs. This section helped us in understanding if we missed out on any other crucial data characteristics or trade-offs pertaining to the retail domain. We concluded the conversation with our interviewees by thanking them for their valuable time and consideration.

The complete interview guide is provided in Appendix A.

3.6.3 Transcribing and Data Analysis

Interview transcripts were prepared using the audio recorded from the interviews. We went through the interviews several times to make sure that we did not miss any valuable information. We followed the steps mentioned by Bhattacherjee (2012) and Recker (2013) for transcribing and analysing the interviews. *"Analysing consists of examining, categorising, coding, tabulating, testing, combining and studying the evidence collected to draw empirically based inferences and other conclusions"* (Recker, 2013, p.99). During data analysis phase, the transcripts helped to get sense of the whole and to establish 'units of significance' (Bhattacherjee, 2012). The units of significance were put together to form a meaningful content.

The units or codes are the building blocks to create inferences from qualitative interviews (Graneheim, 2004). The coding process was done using the coding scheme explained in the section 3.6.4. We categorized the relevant information or content of transcript into four. The units of significance we identified are domain, data type and context, data characteristics, and data characteristics trade-offs. Each subunits or keywords were coded separately. Subunits help in detailing of the information and hence lead to a quality data analysis. In the next chapter, Empirical Studies (chapter 5), we are discussing the inferences obtained by the analysis of the interviews.

3.6.4 Coding Scheme

The coding scheme used for our thesis is provided in table three (3).

Table 3: Coding Scheme

Domain/ Channel	Data Type and Context	Data Characteristics	Data Characteristics Trade-offs
Retail Physical Store: REP	Financial Data: FD	Accuracy: AC	Accuracy & Timeliness: ACT
Retail Ecommerce: REC	Customer/ Guest Data: CD	Timeliness: TL	Consistency & Timeliness: CST
Others: OD	Item Data: ID	Consistency: CS	Accessible & Secure: ACS
	Inventory & Inventory Relief: IR	Security: SE	Others: ODT

	Sales data: SD	Accessibility: AS	
	Human Resource: HR	Others: OC	
	Vendor Details: VD		
	Merchandising Data: MD		
	Marketing Data: MKTD		
	Supply Chain Management: SCM		
	Social Media: SM		
	Others: ODC		

We have categorized the text from the transcripts into four segments and they are domain or channel, data type and/or context, data characteristics and the trade-offs between those characteristics. The abbreviations mentioned in the above table are explained below for better understanding of business and technical perspectives.

Domain/ Channel: This research work considers the following channels in the retail domain for its analysis.

1. Retail Physical Store: This encompasses all the brick and mortar (B&M) stores that has a physical presence to sell products.
2. Ecommerce: This refers to all the Business to Customer (B2C) transactions that happen through an online medium. This can refer to website, mobile applications etc.

Data Type and Context: The following are the different data types and contexts that are widely used and vital in retail domain.

1. Financial Data: This refers to all the cash related information stored in the data warehouse. This can include the financials allocated for an initiative to the salary obtained by the company's employee.
2. Customer/ Guest Data: Customer or guest data is the information that is stored regarding the customers who visit and shop at a retail establishment or buy a product online.
3. Item Data: Item data categorizes each product located in a store or an individual product displayed in an online medium.
4. Human Resource (HR): HR data is the personal information about the company's employees.
5. Sales Data: Sales data classifies all the product sales. Every sale of a product is tracked through sales data.
6. Inventory & Inventory relief: This comprises of all the data about the products available in a store and in the nearby warehouse.
7. Vendor Details: This consists of data about all the organizations that have business dealing with the retail or Ecommerce chain being discussed in this research.

8. Merchandising Data: This consists of data that contributes to the entire lifecycle of an item or a product.
9. Marketing Data: The data related to all the advertising and product promotions.
10. Supply Chain Management: This consists of data that caters to the understanding of goods and service flows.
11. Social Media Data: This refers to the data obtained from social networking channels such as Twitter and Facebook.

The detailed explanation regarding data characteristics trade-offs is provided in the chapter 2, section 2.4.

3.7 Assuring Research Quality

3.7.1 Reliability

Recker (2013) states that reliability of a research depends on to what extent a variable or a set of variables is consistent in the process of its measurement. To keep up the reliability of our research work we made sure to uphold a high level of transparency by explicitly describing and defining our purpose and approach. This approach will facilitate future replications and further study on this research topic.

To make this research work highly reliable in nature, we tried to explain the processes we have chosen in doing this research and why we find it appropriate to use them. Prior to the interviews, the interviewees were called and informed about the research topic, giving an in-depth explanation of our research topic and what we were trying to achieve. The questionnaire was also sent to the interviewees a few days before the actual interview, to help them prepare and collect data points that would help them answering the interview questionnaire. During the interview process, we have maintained a neutral objective and have not influenced the informants in any way and the interviews were recorded only after getting consent from the respondents. We have also made sure that all the respondents were provided the same questionnaire and were asked the same questions during the interview process as well.

After the interview was done, the interviewee's replies were considered as source of truth which have been discovered analysing (through coding) and reporting the data (Schultze and Avital, 2011). The interviews were transcribed with precision by listening to the recorded audio multiple times to avoid any possible mistakes. The full interview transcript has been provided in the appendix of this research paper.

3.7.2 Validity

Validity of a research work can be assessed using theoretical and empirical approaches and in ideal circumstances it should be measured by using both (Bhattacherjee, 2012). We have used both theoretical and empirical methodology to ensure validity of our research work. To achieve theoretical validity, we have done extensive research and reading to collect appropriate literatures pertaining to the topic of interest and based on which we have formulated our questionnaire. For the empirical validity, we identified suitable interview candidates who have immense experience in this field of work. Based on their responses we have made suitable

observations and inferences to take us closer to answering the research question in both theoretical and empirical perspective and thereby ensuring overall validity.

Also, to improve the validity of our research work we discussed and questioned every step and measures we took. We even conducted a pre-study to narrow down the factors contributing to our research question based on which we framed our questionnaire. During the interview process, if any respondent had problems in recollecting context pertaining to a question we also provided him/her time to get back through calls or mails later. This approach enriched the validity of our research work furthermore.

3.7.3 Ethics

During our study, we made sure that we could maintain ethics in every process by not deceiving candidates and manipulating any data obtained. Ethics is defined as the normal division between right and wrong (Bhattacherjee, 2012). It evaluates the behaviour of the study based on certain principles and guidelines (Aguinis and Henle, 2002). The terms, ethics and moral can be used interchangeably according to context (Gregory, 2003). The most important ethical principles in a scientific research are provided below. (Gregory, 2003).

1. Voluntary participation and harmlessness
2. Anonymity and confidentiality
3. Disclosure
4. Analysis and reporting

The participants of the study were not forced to collaborate with us. We made sure that candidates had the freedom to withdraw from the study if they were not willing to. In our research, all the informants were willing to disclose their identity as the research does not deal with any sensitive topics. The interviews were scheduled based on the convenience of the participants. The time schedule and medium for the interviews were decided by the participants. We have provided some insights to candidates before the interviews so that they could decide whether to participate in this study or not. We have made sure that while reporting the findings, we did not manipulate or hide any negative findings. Through all these practices, we believe we have maintained ethics in our study.

3.7.4 Research Bias

Intentional or unintentional inclination towards informant selection and evaluation of their responses may create a threat to the validity of a research (Fraenkel et al., 1993). In this research, we have taken the following measures to reduce the researcher bias.

1. Gathering informants across locations: We have a wide range of informants who have worked in both small and big retail companies. The informants have work experience in retail chains across America and Asia. This help us in reducing the bias of having thoughts and ideas that are localized to one region.

2. Gathering informants from multiple companies: Though majority of the informants are currently working in Target Corporation, they had previous work experience in other retail chains. This helped us in getting diverse perspectives of data handling in various companies and how it differed across companies.

3. Gathering informants of diverse industrial work expertise: Even though the informants for this research topic were chosen from the technical background, they have immense business knowledge of the retail domain acquired by years of experience. Also, the respondents are well versed in the data life cycle management process, from the initial stages of data generation at the source to the reporting and analysis of data. This helped us in getting an end to end perspective of the data flow.

According to us, these measures have immensely helped us in reducing research bias and providing an impartial approach to our research work.

4 Empirical Studies

In this chapter, we discuss the findings from the interviews. This chapter has four sections, introduction to data quality in retail, defining data quality, data quality characteristics, and data characteristics trade-offs. We present our ideas in accordance with the interview guide provided in Appendix A. The transcripts of the interviews are provided in Appendix B.

4.1 Introduction to Data Quality in Retail

In retail, data quality applies to both brick and mortar and Ecommerce channels on similar lines (5,18). Basically, data quality is defined by a set of values, quantitative or qualitative, that define variables which needs to be validated before it is business ready and usable for its respective stakeholders. (5,14). So be it in physical store or Ecommerce, business decisions are highly dependent on the quality of data available. The difference between these domains is on the origin of the data that they are dealing with. For example, traditional retail chain will have brick and mortar or physical store for sales whereas Ecommerce organizations do their business online. Hence the business rules applied to them will also be different.

Jyothsna Raj says that in Ecommerce domain, timeliness is a critical factor as it is a fast-paced environment (4,30). In a physical store, one cannot change the entire items as quickly as it can be done in online store (4,30). In e-commerce, companies can change the entire sorting of a page or add new products and promotions instantly to increase the traffic and sales (4,30). If an Ecommerce company identifies that people are logging into their page and moving out immediately as they are not interested to buy anything, then company can instantly reformat that page by sorting it differently, by prize or popularity, to attract the customers and keep them in their portal (4,30). Whereas, in a physical retail store if it is identified that people are not interested to buy the products, then you wait till end of the day to rearrange the products or to add promotions (4, 30). Another difference in the way these domains operate is, in e-commerce if you see that one item is selling fast, you can immediately contact the seller and ask to increase the stock of that item (4,32). But in physical store, if a product sells out, then you need to wait longer time to stock that item again (4,32).

4.2 Defining Data Quality

Kiran Kumar defined quality data as the data that is fit to be reported, analysed, and used for operational decision making (1,14; 1,18). Beth Benzie states that:

"not all businesses understand how important data quality is until somebody points out where it is causing a problem and how it is impacting the reporting and the business decisions" (6,16).

It has been noticed that 40% of all the business decisions where there is lack of quality data, led to wrong decisions and thereby, caused loss to the organizations (1,14; 1,20). It is always the responsibility for a developer or a tester to verify that input data is loading to a warehouse accurately (2,20). Data should be maintained at the most business usable form in the ware-

house (3,16; 4,12). According to Suyash Kumar (3,16) data should follow and maintain the important dimensions of data quality like accuracy, timeliness, accessibility, ability and integrity. Organizations ensure that there is no data loss, data is clean, integrity between the data tables are maintained, and there is no murky data (4,12).

In retail industry, the business depends on the data, be it sales data, inventory data, and unsaleable data (3,20). From each stage of the operation, right from procuring the product in overseas market as well as the local market till it reaches customer and even if it gets returned, the organizations get lots of data that need to be handled (3,20). Organizations should maintain utmost integrity and accuracy in the data and keep up with the time so that correct inference is made of the data available (3,20). An operational delay in the store will affect the overall operational cost of the company (1,20). Business analysts and data scientists use the data not only to see what happened previously but also to forecast what can happen to the business of the organization (5,16).

In retail, the decisions to procure, sell and keep an item would be helped if company knows where the item is, when to keep it, what the procurement cost is, what the selling rate is and whether the inventory can hold it (1,22). Therefore, data quality plays key role to get the accurate data on this information.

4.3 Data Quality Characteristics

4.3.1 Accuracy

Accuracy is nothing but the precision maintained for a particular data (3,26; 6,24). Complete data can't always be accurate (1,40). There can be chances that organization has all the data, but the data is still not accurate. This may be because of the issue with a system or process which is not in accordance with the business operations (1,40). Similarly, there are always negotiations between merchants and the vendors about what quantity should be procured (1,40). Until and unless the information is accurate, the data can't be used for decision making (1,32). Beth Benzie says that:

"Accuracy is really important to make business decisions. If you don't have complete sales, business will be speaking about sales thinking that it is complete and decisions will be taken based on that" (6,22).

One of the scenarios in retail where accuracy has a vital role is when sending campaigns and promotions to customers (5,26). Organizations don't want to send a happy birthday mail or a greeting mail to a person who is diseased as it will be very offensive (5,26). Customers can get upset if they get email promotions even after unsubscribing the promotions (3,18). Similarly, organizations face challenges to send the posts to exact address of the customers as some customers keep moving from one location to other and they do not update their contact details in company's websites (3,18).

4.3.2 Timeliness

Timeliness indicates the time to get the required data (3,26). In some cases, it is not the quality of the data matters but what matters is how fast the data is provided (4,14). In retail, there are campaigns going on most of the time and companies change the campaign within an hour if it is not working (4,14). Organizations need to get the sales data immediately after a campaign is launched to see the response from the customers (4,14). During special days like Black Friday or Christmas, sales will be more and companies look for reports which show response of the promotions, stores that perform well, who is buying what, and product which is selling out quickly (3,26). In these situations, timeliness is a crucial factor and organizations don't expect 100% accurate data (3,26). According to Beth Benzie, since timeliness is a crucial factor, any delay in the operation should be notified at the earliest to the business management team and the issue should be rectified with utmost priority (6,22). To meet the expectations from the business stand point, the timeliness of the data provided is important (6,22).

4.3.3 Consistency

One of the problems organizations face is inconsistency in data due to heterogeneous source systems (2,46). If data is accurate, then it need not be consistent (2,46). Sandipan Ghosh explains a scenario in retail pertaining to data consistency. If one of the stores considers 'OO' as value for out of stock in its records and another as 'OOS', then the issue of data consistency arises (2,46). Organizations may miss details while reporting if the data is inconsistent. For retail companies working with multiple vendors, inconsistency of data becomes a huge problem as they end up in wrong decisions (3,36). Santhosh Meenhallimath talks about the source and target data consistency in the data quality processes. More importantly, sales information must be consistent across the data processes in these industries (5,30).

Suyash Kumar says that *"inventory and customer information are the key areas where consistency should be maintained"* (3,40).

Guest data and customer data is maintained in various places or tables in the database (3,40). For instance, a few of the tables have captured the guest information with the exact email id, phone number and other details and a few have data in some other format (3,40). This lead to inconsistency in the address and other details of a guest and causes lots of issues to the organization (3,40). Similarly, in the inventory section, data needs to consistent when details on sales, inventory stock and area are reported (3,40).

4.3.4 Security

Security is a key data dimension for any retail organization (4,58). Security is to ensure that you don't compromise on important data and anything that needs to be confidential or protected is not leaked out (4,58). According to Sandipan Ghosh finance and HR data should be most secured than any other data (2,50). Security tolerance decides entire landscape of how you design your solution or architecture of the system (2,50). Sandipan Ghosh claims that:

"Retail organizations tend to give access to supply chain information but they secure their financial data" (2,34).

Another instance to show the importance of security is, Ecommerce websites create a layer of security to mask the credit card number of the customers to protect them from fraud (2,34). Kiran Kumar mentioned that anything that organizations deal with external vendors and external people having access to internal systems should have the highest security (1,82).

4.3.5 Accessibility

Organizations should ensure that the data is available or accessible to the right person at the right time in the requested format (4,58). Suyash Kumar opinions that people inside an organization can have access to see general buying pattern of a product, but when it comes to information like customer's name, mail id, gender, and age segmentation, it should be secured (3,46). Similarly, when the data is available for external vendors, organizations need to ensure that each vendor's confidentiality is met and the accessibility of the data is granted only to the person who is supposed to view that data (1,74). Jyothsna Raj says that any aggregate data needs to be more accessible (4,60). For instance, total sales in a region or total sales of a product should be accessible unlike personal information (4,60).

4.3.6 Other Data Characteristics

Along with the data characteristics mentioned in the previous sections, Suyash Kumar suggests integrity, availability, return of investment percentage and completeness as important data quality characteristics (3,22). Sometimes, reliability is also given high importance in this domain (4,18). Depending on the department in the retail organizations, KPIs (Key Performance Indicators) also vary (3,50). For sales team, the indicator would be purchase quantity and for inventory team, it would be stock item and area available to stock the items and the quantity of returned items from customers (3,50). Vendor's information like, their details, which warehouse their products are landing to, and which transportation and logistics have been provided are also important in retail business (3,50). Data Analysts get these KPIs first and therefore, they need not fetch the historical data to come up with matrices that are reported to higher management to make quick decisions (3,50).

4.4 Data Characteristics Trade-offs

4.4.1 Accuracy and Timeliness

All the interviewees had the same opinion that the trade-off between accuracy and timeliness is the most important trade-off in the retail domain. Kiran Kumar says Monday morning report in retail is crucial as it provides details on previous week's sales and it helps in making marketing decisions for the current week (1,46). In this case, 100% accuracy in report is not expected, but the report should be timely delivered (1,44). There might be a situation where a few of the stores have gone completely out of stock. Even if that information is less accurate and there were still some stock left, since it gives a hint that there needs to be a constant supply of that item, organizations can be little low on accuracy but very prompt on the timeliness (3,26). As far as information on unsaleable products concerned organizations require exact figure as it affects profit percentage calculation in the retail industry (3,30).

One of the cases in Ecommerce where accuracy is important than timeliness is, when reports are taken to investigate user's navigation through website (5,22). Ecommerce companies check 'clicks data' which is information regarding the customer clicks in a URL of the website (5,22). Organizations don't need this information to be timely, but they need it accurate to see which all pages the customer navigated to and what products were checked (5,22). But during holiday seasons like Christmas and Thanksgiving, timeliness becomes crucial as the transactions are more and promotions and campaigns need to be changed every hour (4,40; 5,22). Companies want the in-hand data or item data in your store to be more timely than accurate in these scenarios. The competition from other retail companies is more and you need the information on sales quickly in a timely manner (4,40).

Kiran Kumar illustrates an important scenario where inaccurate data report can badly affect the organizations. Accuracy is key when retail companies report for vendors or 3rd parties for two reasons (1,52). Companies are dealing with external agents who are from outside the organization and they are the ones who supply the items (1.52). Kiran Kumar says if a vendor doesn't agree to a policy after accepting the purchase order, then the vendor is legally chargeable 3% of their income (1,52). Similarly, if the retail company is publishing an inaccurate data report after the sales to vendor, the vendor can sue that company. When there is a payment or cash transaction involved with vendors, accuracy is given more importance rather than timeliness (1,52). Instead of urging to charge the vendors, organizations make sure that data is complete and accurate so that they don't lose any profit (3,32).

Sandipan Ghosh who works in supply chain department of the retail organization points out that accuracy is not a concern in supply chain and they accept even if the data is only 80% accurate (2,38). In supply chain where companies are calculating the forecast based on some attributes, there can be 20 % as limit of tolerance whereas in finance it should be 0%, (2,26). Everyone agrees the fact that financial data should be 100% accurate. For instance, if the salaries of the employees are entered wrong in the records, then it badly affects the company (2,32). Usually, the pay slip of an employee is generated a few days later, after the salary has been credited (2,32). This is because of the multiple processes involved to check the accuracy of the information (2,32).

Another important data to be considered in this domain is data related to advertisements (4,34). Companies put the ad in various places in their website or stores. If the advertisement is not working and doesn't fetch any interest from customers, then it should be changed or removed immediately (4,34). The conversion rate of visit to purchase needs to be measured in timely manner in these scenarios (4,34). Jyothsna Raj also pointed out a scenario where accuracy is more important than timeliness. Any information that is sent back to customers should be accurate (4,38). For instance, any suggestion or proposal to customers derived after analysing their buying pattern and cart must be 100% accurate as it can tend customers to unsubscribe their email campaigns (4,38). It is acceptable to send this kind of information delayed, but when it is sent out, it should be accurate (4,38).

4.4.2 Consistency and Timeliness

Kiran Kumar states that consistency of data is paramount in the financial world in the retail domain (1,56). Beth Benzie explains the reason for consistency as:

"when there are many sources and there are inconsistencies within them and if we try to combine them, consistency will become worse" (6,34).

Financial data comprises of capital resources, monetary results of operations, assets, mergers and acquisitions information, stocks, positive or negative trends in market fluctuations, inflation and other uncertainties (1,56). Since financial data is something that affects the functioning of an organization and it is also information that incorporates certain legalities associated with, it is checked thoroughly (1,56). Also, consistency needs to be maintained w.r.t when to publish the financial data, how to publish it and who are the required audience of the data (1,58).

Item data should also be consistent if not complete. According to Kiran Kumar, in Target, item data plays a crucial role as upstream data and are loaded from the source systems before all the other tables. Consistency of item data assures consistency of the other downstream tables. Hence, thorough checks for inconsistencies like duplicates, trends and other checks are incorporated as monitors to prevent incorrect data (1,70). Jyothsna Raj claims that forecasting data or any data that helps retailers in predicting the future, such as sales prediction, need to be consistent (4,50).

Santhosh Meenhallimath emphasises that customer data should always be consistent across different tables. In most retail companies, customer information is stored to send marketing information such as vouchers, sales coupons or catalogues (5,28). If the data across tables are not consistent, this might result in the wrong person getting the marketing information leading to a failed strategy. Extra precautions are taken to make sure that customer data across tables are consistent and complete.

On the contrary, it is unanimously agreed upon that the sales data needs to be timely, particularly in the Monday Morning reports. Kiran Kumar states that sales information needs to be on time for the business users, and they are ready to compromise with a small variance in consistency and accuracy if the sales information is on time (1,70). So, based on the sales figures, the business users decide whether to continue with products or marketing and advertising strategy and for this sales data needs to be timely (4,44). Similarly, inventory and inventory relief data needs to be timely; most importantly for perishable items. If an item is fast depleting in a store, inventory data should get reported on a timely manner for business to replenish it from the closest warehouse (1,70). Timely reporting of perishable items is mandatory to prevent decaying of the items in stock.

Kiran Kumar mentions that sometimes non-timely inventory can lead to not only loss of sales but also loss of reputation to a retailer (1,72). According to a context provided by him, there can be a new product launched and lot of marketing and advertisements associated with its launch. But due to the delay in the inventory data, business is not able to decide whether the product is in the shelf of stores or not. This can also lead to dissatisfied customer, if they come and see empty shelves (1,72).

4.4.3 Accessible and Secure

Kiran Kumar states that all members of the data quality team should have access to data that helps to build business (1,76). He clarifies that every retail establishment runs on item data and hence it is imperative that accessibility to item data must be granted to all users. All the downstream users such as sales, inventory, vendor and supply chain should have access to item data so that they will get to know how the item information will affect their tables (1,78). In simpler terms, all of merchandising data can be permitted to be accessible across the data analysts of the organization.

Sandipan Ghosh states that social media information i.e. data obtained from social media sites such as twitter and Facebook regarding marketing or product response can also be accessible to all (2,50). He worked on a project where feeds from multiple social media channels are taken to analyse consumer sentiments to products and marketing activities. He mentions that sales data should be provided to all but exporting it to other devices or formats must be strictly controlled.

As for secure data, every respondent felt that Financial and HR related data need to be protected with layers of security (2,50). This is to prevent hacking of data from malicious users. Also, Financial and HR data should be available only to those users who work in it (2,50). This is because Financial data consists of crucial information like stocks and merger and acquisitions. If leaked or in the wrong hands might cause unspeakable losses (2,50). Beth Benzie states that *"anything that has personal information type data has to be secure and however needs to access, it needs to prove the identity"* (6.38). Sandipan Ghosh asserts that Human Resource tables that holds employee information such as address and pay should also be secure (2,50). This should be mandatory so that an employee's salary and address details is not shared with other employees (2,50).

Security is also crucial in the case of customer or guest data. Every retail establishment has their own customers and are very private about it. (5,44). Customer data includes data regarding their shopping patterns, credit card details, their physical location like address, phone number, email ID and Social Security Number (SSN). This information is highly crucial for a retail organization and loss of this information to hackers and malicious elements can not only lead to loss of reputation but also cause many legal issues. Santhosh Meenhallimath says that multilayer security is provided to the customer and guest tables and very limited access is also provided to those employees working on it.

Sandipan Ghosh has reaffirmed Santhosh's thoughts regarding customer data and he claims that when handling social media related data, they make sure that their location and ID are kept secure when using other details like shopping patterns and likes and dislikes (2,52). Finally, Santhosh Meenhallimath claims that brand information should be available to all users working in retail establishment, but one vendor should not be able to see the brand information about another vendor (5,38). This can accentuate wrong usage of data.

5 Discussion

In this chapter, we compare our empirical findings with our research questions and the theoretical framework. The discussion will provide a clearer picture about how organizations accomplish data characteristics trade-offs to maintain data quality.

5.1 Data Quality in Retail

Data Quality is a precondition for any business that uses huge volumes of data for guaranteeing the value of the data and currently, extensive analysis and research associated with data characteristics trade-offs is lacking in the retail domain (Cai and Zhu, 2015). And, with the dawn of the big data era, the retail industry faces the need to compare its two modes of retailing i.e. physical and online. (Otto and Chung, 2000). Only once a clear understanding of their differences is established, sound strategies can be applied to improve business.

Most of the data quality attributes for both physical stores (Brick and Mortar) and online stores are identical, but there are a few data characteristics that vary. As stated by Otto and Chung (2000), the primary difference between physical and online retailers are the data collection patterns. Substantial amounts of information can be collected from online shoppers directly (such as name and address) and indirectly (through browsing and buying patterns) and an online retailer is able to personalize the shopping experience of a customer (Otto and Chung, 2000). The same thought has been collaborated by one of the interviewees who stated this as a fundamental difference between data in online and physical retail (4,60). Secondly, the arrangement of the items or products in an online retail varies drastically from physical retail. There is space and economic limit to the product that can displayed in a physical retail in contrast to an online retail where unlimited product information can be displayed (Otto and Chung, 2000). To add on to this, our respondents claim that online retailers can change the entire sorting of a page or add new products and promotions instantly to increase the traffic and sales (4,30).

The time and marketing strategies varies between physical and online retailers. Online retailers can provide their offers and advertisements to anyone who can access the internet, whereas traditional retail establishment have their marketing strategies concentrated on one or few physical stores (Otto and Chung, 2000). This can be attributed to the time factor since, inventory can be stocked faster in online retail whereas in physical retail one need to wait a longer time (4,32). Therefore, timeliness and pace of decisions are key in Ecommerce (4,14).

5.2 Defining Data Quality

With the increase in the number of sources that feed data into warehouses, data quality problems are becoming more common (Daniel et al., 2008). And to understand the importance of data quality we need to define data quality. One respondent defined data quality as ‘getting data fit to be reported, analysed and used for operational decision making’ (1, 14; 1, 18) and it was further simplified by another respondent by defining data quality as ‘getting data ready for business’ (4, 10).

But it has been clearly stated by Beth Benzie (6, 16), that not all the business users understand the importance of data quality until someone (data analysts) points out the problem area and the impact it might have on business decisions. This is clearly iterated by Bulger et al. (2014) who says that it is paramount for a retail business user to have a better understanding of its data by interacting across different channels and sources, including social media, internet, mobile and internal systems. Only when stakeholders understand the importance of data, its quality can be improved.

This gives rise to the importance of data quality analysts in organizations dealing with vast amount of data and the challenge they face is to develop a model to achieve quality of the data that is being sourced from various servers (Lee et al., 2002). Another major impediment which data quality analysts face is in educating the business users and other stakeholders the importance of data quality for business. Once this balance is achieved, data quality management for an organization will improve, there by leading to increased productivity and heightened customer satisfaction (Geiger, 2004).

5.3 Data Quality Characteristics

5.3.1 Accuracy

Accuracy is expressed as the degree of closeness to which the information or data in the data warehouse matches with the values in the real world (Batini et al., 2009). Therefore, when discussing accuracy, organizations consider the quality of data and the number of errors represented in a dataset in contrast to its source (Batini et al., 2009). Our informants also stated similar definitions for accuracy. Accuracy is about data precision (3,26; 6,24) and complete data need not be accurate always (1,40). Literature also indicates that, just because a field has a value in it, doesn't mean it is correct (Geiger, 2004).

An example of how accuracy can affect data is when we consider a single record or data in its entirety (Friedman and Smith, 2011). If the age of a single person is wrongly recorded as 40 instead of 15, then it is a drastic change in the demographics for that individual, but the accuracy of the average age in a population of millions is very minutely affected (Friedman and Smith, 2011). Our findings put forward a few scenarios in retail where accuracy of data plays a crucial role. Organizations should be cautious while sending promotions and mail campaigns to the customers (5,26; 3,18). Sending unrelated promotions or delivering mail to incorrect address can upset the customers and it will harm the brand credibility (5,26; 3,18).

5.3.2 Timeliness

In data quality management the terminology, timeliness of data or data currency is frequently used in relation with the 'use-by' period of data and could be related to when the data was last checked and/or updated (Chapman, 2005). From our study, we could understand timeliness as the time to get the required data (3,26). In some cases, it is not the quality of the data but the currency of the data provided matters (4,14). It is a known fact that the quality of data in degenerates over time and according to Fan et al. (2011), it is estimated that nearly 2% of records in a database become obsolete every month. Extensive investigations reveal that the time factor holds extreme importance in data quality management and it plays a significant

role in determining the validity of data in the database (Fan et al., 2011). There are 6 important factors that contribute to data currency in a database and they are normalization, interval scale, interpretability, aggregation, adaptability and feasibility of data (Heinrich and Klier, 2011).

We could identify that timeliness plays a significant role in campaigns and promotions of organizations. Companies need to change the campaigns immediately if they do not create any impact in customers (4,14) and during festive seasons companies look for frequent sales reports to analyse the trading and effectiveness of promotions (3,26). We identified that the literature and the interviews lead to similar inferences on the importance of timeliness.

5.3.3 Consistency

Literature says consistency of data is, preventing more than one state of information system matching a state of the real-world system and inconsistency of data would translate to represent one to many mapping of data between the real world and the data systems (Batini et al., 2009). Sandipan Ghosh mentioned that inconsistency of data is due to heterogeneous source systems and it can't be said that, data which is accurate need not be consistent (2,46). The consistency dimension can be viewed from another perspective of having consistent or the same data values across tables or sometimes across databases (Pipino et al., 2002). According to Chapman (2005), there are two aspects about data consistency and they are Semantic consistency and Structural consistency. Semantic consistency is one where the view of the data should be clear, unambiguous and consistent and in structural consistency data and metadata need to have the fundamental structure and format (Chapman, 2005).

Most database and data warehouse architects use a concept called the data normalization to organize attributes and relations of a database to reduce data redundancy and thereby improve data integrity and consistency (Date, 2006). Normalization can also be called as a process that aids in simplifying the design of a database for it to achieve its optimum structure by eliminating and removing redundant and ambiguous data (Date, 2006). Most of the interviews discussed about the importance of having the source and target data consistency in the data quality processes. Our findings show that the information such as sales data, customer data and inventory data should be consistent when reported.

5.3.4 Security

Security and privacy settings are paramount in an organization handling data and easy access to information may conflict with requirements of security, privacy and confidentiality (Eppler, 2006). From our study, we could establish the role of security as, to ensure that organizations don't compromise on important data and anything that needs to be confidential or protected is not leaked out (4,58). Developing consistent policies and procedures for security are crucial to any organization (Eppler, 2006). Kaufman (2009) has rightly claimed to ensure data confidentiality, integrity and availability, the data custodians must offer capabilities that at the least include:

- A validated encryption schema to safeguard data in the shared storage environment.
- A stringent access control which is of utmost importance to ensure hacking of data and preventing unauthorized access to the data.

- Consistent data backup to prevent loss of data during outages and system crash and proper security features for the backups as well.

The information is secure when both the information and the users are protected and devoid of any manipulation (Eppler, 2006). We could find that finance and HR data should be most secured than any other data (2,50). It is evident that anything related to the organization that is accessible to external vendors should have the highest security (1,82).

5.3.5 Accessibility

Accessibility describes whether there is any continuous or unobstructed mechanism for accessing the data (Eppler, 2006). Presence of data in the warehouses without access is equal to not having it (Chapman, 2005). Stakeholders consider data accessibility as a pre-condition (Wang, and Strong, 1996). Organizations have the responsibility of ensuring that data is available and accessible to the intended person at the right time in the requested format (4,58). Findings from our study show that in retail, people inside an organization should have access to see buying pattern of customers but the individual contact details should not be accessible (3,46). Also, organizations should keep the confidentiality of the external vendors and accessibility of the data is granted to only the people who are supposed to access the data (1,74).

It has been identified that within data manufacturing systems, there are 3 important categories or roles. (Strong et al., 1997). They are data producers, data custodians and data consumers. Data producers are people, groups or others who source or generate the data. Data can be sourced from other applications as well. Data custodians are people, groups or others who aid in storing, processing, maintaining and protecting data. Data Consumers are people, groups or others who are the stakeholders to whom the data makes meaning in a business perspective. Each role is associated with a process and accessibility of data is determined by the data custodians.

5.4 Data Characteristics Trade-offs

5.4.1 Accuracy and Timeliness

As mentioned earlier, accuracy and timeliness have conflicting characteristics and it is in the best interest of data analysts to identify which data needs to be timely over accurate and vice versa. As stated by Baškarada and Koronios (2014), it is a fine balance between these characteristics that need to be addressed carefully. The more current a piece of information is, there has been less time to check if that data is accurate (Eppler, 2006).

Based on the literature analysed and interviewees responses, firstly, it can be clearly stated that any data with information about people needs to be accurate. Information about people can include Customer/Guest data, Human Resource data and Vendor information. Customer/ Guest data is the most crucial data to any retailer and maintaining accuracy of this data is paramount (4,38). Customer data comprises of information about the customers such as Name, Gender, Social Security Number, Address etc. A mismatch or inaccurate information can lead to loss of trust among the customers (4,38). Similarly, Customer data incorporates all the purchasing information of a customer. This needs to be accurate, since this information is

used in analysing what sort of marketing and promotional vouchers need to be sent to that customer and any wrong campaign done can make a customer lose interest (4,38).

Secondly, Human Resource data or information about company employees need to be accurate. This is crucial information about employee's names, SSN number, Bank details and payroll details. Immense accuracy must be maintained to ensure no mismatches occur in the employee details. Thirdly, Vendor data needs to be maintained with high accuracy. It has been stated by one of the interviewees that incorrect information submitted to the vendor can either cause loss in money or can give rise to legal problems (1,52). So, retailers tend to maintain accuracy in vendor data even if it is delayed slightly. Finally, in the ECommerce world 'Clicks Data' information needs to be highly accurate rather than timely. This helps a retailer identify a user's navigation through a website. Organizations need this information to be accurate to identify what are the customers turning on and offs about a website and what causes maximum conversion rate (5,22)

Timeliness is crucial for items and sales (merchandising) data. According to one of the interviewees, the Monday Morning report that consists of sales and marketing information, need to be timely even if it isn't 100% accurate (1,44). Sales data generally needs to be timely and as per our interviewees and it needs to be timelier in nature during the festive seasons. During the holiday seasons like Christmas and Thanksgiving, when the sales are high, retailers would like to see the sales data in a timely manner to analyse if their promotions and marketing strategies are working (4,40; 5,22). In such cases retailers are ready to compromise on small deviation in accuracy over timeliness. Finally, our findings suggest that advertisement and marketing data need to be timely. This is to constantly check whether a promotional activity is providing its desired results.

5.4.2 Consistency and Timeliness

Data consistency refers to how useable the data is and its replicability with respect to the previous data as well as the real-time data (Wand and Wang, 1996). Organizations want the data to be consistent and constant over time for them to be able to use and show the data in multiple ways without needing to change the structure (Wand and Wang, 1996). In the real-time database management system, the trade-off between consistency and timeliness of the system holds significant importance.

Our findings show that financial data which includes capital resources, monetary results of operations, assets, mergers and acquisitions information, stocks, positive or negative trends in market fluctuations and inflation, should be consistent across the processes than timelier (1,56). Other important information that need to be consistent are item data (1.70) and data used for forecasting or prediction of sales (4,50). Santhosh Meenhallimath pointed out that customer data should always be consistent across different tables as the data is used to send marketing information such as vouchers, sales coupons or catalogues (5,28). If the data is not consistent, then the promotions and advertisements reach the unintended customers.

From our study, we could find that inventory and inventory relief data need to be timely (1,70). Inventory data should get reported on a timely manner to identify the items that are fast moving and to replenish them from the closest warehouse (1,70). Both in brick and mortar and online trading, it is important to know constant inventory details for keeping an item in stock and thereby, satisfy the customer (1,72).

5.4.3 Accessible and Secure

Accessibility and security is an important trade-off that organizations should deal with. Finding a balance between these two characteristics is a difficult endeavour (Braz et al., 2007). Our findings show different scenarios where one characteristic is compromised over other. Our informants stated that financial and HR data should be protected with multiple layers of security measures (2,50). Similarly, data security is critical in the case of customer or guest data. Every retail establishment has their own customer database which includes data regarding customer's shopping patterns, credit card details, physical location like address, phone number, email ID, and Social Security Number (5,44). As this information is highly critical for a retail organization and loss of this information to hackers and malicious elements can lead to loss of reputation and cause legal issues (5,44). It is evident that information regarding a vendor should not be disclosed to other vendors (5,38).

The findings suggest that the merchandising data can be accessible across the data analysts of the organization (1,78). Similarly, the social media information or response on products and advertisements can also be accessible to all (2,50). But the same time, when handling social media data, organizations make sure that customer detail like personal ID is kept secured and other details like shopping patterns and likes and dislikes towards the posts made by the company are made accessible (2,52). Therefore, we could conclude that customer information should be maintained with utmost security and transactional data can be more accessible.

5.5 Summary of Discussion

The tables below show the prioritizations of data characteristics trade-offs in various contexts. The (✓) mark indicates priority of a characteristic over the other in the respective context.

Table 4: Accuracy and Timeliness Trade-off

Data Type and Context	Accuracy	Timeliness
Advertisement		✓
Customer/Guest	✓	
Financial	✓	
Forecasting	✓	
Human Resource	✓	
Inventory & Inventory Relief		✓
Item		✓
Marketing		✓
Merchandising		✓
Sales		✓
Supply Chain Management		✓
Vendor	✓	

Table 5: Consistency and Timeliness Trade-off

Data Type and Context	Consistency	Timeliness
Customer/Guest	✓	
Financial	✓	
Forecasting	✓	
Human Resource	✓	
Inventory & Inventory Relief	✓	
Item		✓
Marketing		✓
Merchandising		✓
Sales		✓
Stock Information		✓
Supply Chain Management		✓
Vendor	✓	

Table 6: Accessible and Secure Trade-off

Data Type and Context	Accessibility	Security
Advertisement	✓	
Customer/Guest		✓
Financial		✓
Forecasting	✓	
Human Resource		✓
Inventory & Inventory Relief	✓	
Item	✓	
Marketing	✓	
Merchandising	✓	
Sales	✓	
Social Media (Comments)	✓	
Social Media (Identity & Location)		✓
Supply Chain Management		✓
Vendor		✓

6 Conclusion

The following section summarizes the thesis, providing an understanding in relation to the research question. In this section, the key findings from our study and the limitations associated with it are discussed. Eventually, this section is concluded by analysing how future research can be done to improve the discussions on our research question.

6.1 Key Findings

The main purpose of our study was to identify how trade-offs between data characteristics are handled by organizations, thereby maintaining data quality. This has been clarified by answering the research question, *'How organizations accomplish trade-offs in data characteristics to maintain data quality?'*. Our literature review, pre-study and interviews helped us in answering the above question.

Firstly, we defined data quality and identified the importance of data quality in organizations, focus being on the retail domain. We categorized how data quality measures can differ between the two channels, physical and online, of retail business. In our literature review, we analysed multiple data quality management frameworks which can explain our field of study. Based on the frameworks and pre-study, we identified three key pairs of data characteristics for which organizations need to find trade-offs while maintaining data quality. The trade-offs we identified for this research are accuracy & timeliness, consistency & timeliness and accessibility & security. Our findings help to recognise how crucial customer information is for every retailer and why highest prioritization of accuracy and consistency is given to that. It is evident that timeliness of sales and merchandising data is critical to make quick decisions to keep the transactions happening. It can be also seen that financial and HR related information need to be secured while sales and merchandising related data can be more accessible for decision making.

After the analysis, we arrived at a summary illustrated in a tabular form where we have shown the prioritization of the data characteristics for various data types and contexts. We believe that this can be used by our intended audience to get insights to identifying trade-offs and prioritizing data content based on these trade-offs. We conclude by suggesting further research in this topic to extend to other domains apart from retail.

6.2 Practical Implications

The primary implication of this research work is that it would assist retail organization to build a guideline to help understand data quality based on the data context. Our findings assist in categorizing data types and content based on data quality characteristics for three different trade-offs. We have identified the following audience who can benefit from our research work:

- Retail organizations: Our research is focussed on the data characteristics and content for retail organization. Using the literature review and our key findings from the inter-

view analysis, we have categorized data content based on data quality characteristics. Physical Retail or Ecommerce organizations can use this to identify where to preserve which data characteristic for what type of data.

- Data quality aspirants and practitioners: This research can be helpful for new aspirants who are interested in the field of data quality and data analysis. They can use this a template to identify the importance of one data characteristic over other in various data contexts.

- Business users and stakeholders: As mentioned earlier, there are a lot of incorrect decision made by business users and stakeholders due to their lack of knowledge in data quality. This research work would help them to understand the importance of data quality in decision making and to emphasise the importance of one data characteristic over other based on the context.

6.3 Limitations and Future Research

Our study is limited to only a few companies in the retail domain. As far as this research is concerned, we have focussed on a few multinational companies and a few start-ups to understand how they deal with data quality. Further researchers can focus more on this shortcoming and identify more retail organizations across the globe to categorize how trade-offs are maintained.

As mentioned earlier, based on our literature review there are ten pairs of data characteristics trade-offs that have been identified. Our pre-study helped us narrow this down to three pairs, but our interview respondents have expressed that there are a few more data characteristics trade-offs that would be interesting and suitable for this field of study. Hence, further research works can be on identifying other data characteristics trade-offs and how data content categorizes into them.

The data types and contexts are based on the response provided by the interviewees and different people have different understanding of data context based on the area they are working in the retail industry. It would be interesting if future researchers of this topic include business users and decision makers in their research to identify data content from their perspective too.

Appendix 1: Interview Guide

Introduction

1. Please provide a brief introduction about yourself and work experience?
2. How many years of experience do you have in the retail industry?
3. Please elaborate about your role in the organization you work for.

Defining Data Quality

4. Could you explain your understanding of data quality?
5. How important is data quality in your organization?
6. What do you think are the main characteristics defining data quality (eg: accuracy, timeliness, accessibility etc.)?

Data Characteristics Trade-offs

7. How would you characterize trade-offs between different data characteristics in your field of work?

Accuracy and Timely:

8. According to you, how important are Accuracy and Timeliness of the data that you handle in your organization.
9. What kind/variety/class of data would you classify as more important to be timely than accurate in nature and vice versa.
10. Could you explain a context where you felt data needs to be more important to be timely than accurate and vice versa?

Consistency and Timely:

11. Similarly, how significant is consistency of data in the retail domain?
12. What kind/variety/class of data would you classify as more important to be consistent than timely in nature and vice versa.
13. Could you explain a context where you felt data needs to be more consistent but less timely and vice versa?

Accessible and Secure:

14. Based on your opinion, what is accessible data and what is secure data?
15. What kind/variety/class of data would you classify as more important to be accessible than secure in nature and vice versa.
16. Could you explain a context where you felt data was more accessible but less secure and vice versa?

17. What data would you or other data analysts in your organization consider most important to retrieve during a system crash.

Suggestions and Other Information

18. From your experience, could you suggest any other trade-offs which would be crucial for data quality management and how?

Appendix 2: Interview Transcripts

Interview 1 – Kiran Kumar

Interview Date: 23rd April 2017

Present: Kiran Kumar (K), Athul (A) and Dilip (D)

Duration: 44 minutes

Transcribed By: Athul

Line	Speaking	Text	Code
1.	D, A	Hi Kiran	
2.	K	Hi Dilip and Athul	
3.	D	Currently you know the situation that what our thesis is based on right? We are working on data quality attributes for the retail domain and yours is our first interview. Can you give us a brief introduction about yourself?	
4.	K	Myself Kiran, I have been in retail domain for 10 years. Worked majorly with Target Corporation which is the second largest retail chain in US. Worked majorly into data quality in engineering, worked into different aspects of analysing data and their accessibility and how to improve business decisions based on data. Giving a clarity on what data quality is, maintaining governance and ensuring the right data is captured and right data is reported. Is that good?	OD
5.	D	Yeah, that's perfect. So, you had a chance to go through the initial first page of the questionnaire that we sent you right? In the introduction, we have mentioned about how we are going to do a comparative study of the trade-off between different data characteristics. The things we are focusing here are accuracy & timeliness, consistency & timeliness and accessibility & security. And few brief ideas about each attribute is also mentioned which are accuracy, timely, accessibility, secure and consistency. So, the first question in the introduction, I think you have covered more or less on what you are doing in the company. So, second question is how many hours of experience do you have in the retail domain?	
6.	K	I have 9-10 years of experience in retail domain.	
7.	A	So, most of your experience is in retail domain only?	

8.	K	Yes.	
9.	D	Could you please elaborate about your role in the retail organization that you work for? Like what is your designation and what you do for that?	
10.	K	My designation is lead data quality engineer currently. My roles and responsibilities involve understanding data that is generated from the source system in the retail domain. Analysing and connecting data at different level taking timeliness into account to ensure that there is no delay. If there is any delay reporting it, they do data governance which obviously involves data analysis and reporting. Accessibility is also part of it. We ensure that the right people get access to the right data at the right time to make business decisions. Say for example, if we have a Monday morning report through which I change my sales plan, retailing cost, purchase order cost, we ensure that we capture what is currently reported and accurate.	REC AS AC
11.	A	So that's kind of like a brief introduction about you and the organization that your work for and the role that you play. In the retail domain, what role – we know that you are a lead data analyst - what area do you work on?	
12.	K	I basically work on merchandising area, which involves planning, supply chain, order management, inventory management, sales capturing and pricing.	MD
13.	A	Okay, let's come to the next category, its defining data quality. So, the 4th question that we would like to ask you is about your understanding about data quality. What is data quality according to you?	
14.	K	According to me, the high-level data quality is generally considered as the quality of the data. So, it is like it can be defined as data that is fit to be reported, fit to be analysed, fit to be taken based on – the correct data to take the correct decision at the right time. What we have noticed is 40% of all the business decision where there is no data quality may lead to a wrong decision making and thereby a wrong strategy and thereby a loss to the organization. So what we try to do here is, we try to link data and look at the data at different aspects. Like as we call pillars of data quality – accuracy, timeliness, constancy and so on - to ensure at least it passes the initial layer where we ensure that the data that is presented is as true as possible.	AC TL CS
15.	D	So, what I understood from what you said is like how good the data is and how fit the data is for reporting and analytical purposes.	
16.	K	Data quality is all about generalized consideration of high quality data.	
17.	A	Okay and how good the data is for the company.	
18.	K	yeah, you can always tell that it is used for operational and decision making and planning i.e. data that is used for operational decision making and plan-	

		ning information, like planning strategy or something.	
19.	D	That leads us to the next question. How important is data quality in your organization? Or what this questions means is that – how important is data quality that is the first question I would like to know about, and the second one is about, is there any situation that you have faced where data quality has been compromised and the repercussions of that.	
20.	K	In our organization data quality is given the highest importance. It is similar to a development of an enterprise warehouse. You can have data which has no meaning and make wrong decisions which may lead to losses both credentials and strategically. I can take several cases into account, one case being in grocery department where the purchased fruit - maybe banana - in KGs or in Pounds and sold individually – this is a scenario that can happen in any retail domain. Your purchasing quantity and your selling quantity may not match. The issue here is, I may buy the same item, if I go buy single item by item, what it would result in this scenario is, if a buy a product/grocery item in Pounds, and sell it individually, there is a chance that I may pay a cost of procurement higher than the selling/retail cost. Secondly, due to the fluctuations in product cost, I may buy one item at a huge margin and another item at a lower margin, when I mean margin my difference between what is my procurement cost and my selling cost. If you do not balance it out when I sell this item, it may lead to an issue – issue in the sense loss. Issue in the sense of losses because selling an item at a cheaper rate than market value. Secondly, there can been situations because of some marketing issue what I am trying to tell here is because of item set up or because of some internal issue. Not focussing on external. I miss out on a chunk of all my item from labelling and printing. So, there is an operational delay in the store, which would lead to the overall operational cost of our company.	MD
21.	A	So, if you would take the case of fruits or something like bananas which are perishable that can also be an issue right.	
22.	K	Yes, exactly. When do I sell, when do I keep, when do I procure, these decisions we could help by tracking what is the item, where is it, what is my procurement cost, what is my retail cots, looking into it, clubbing all of it together and ensuring we inform merchants or the sales managers about what is my current inventory and where we can just improve the supply chain.	ID
23.	D	So, these are the crucial factors of data quality in retail domain.	
24.	K	Yeah, this is one example. Second example is related to the same principle of how do you measure. Say for example I have a product which I am importing from some other countries. There are different measuring units for each of the items.	
25.	A	For example, if we take US weight is measure in Pounds and in India it is measured in Kilograms.	

26.	K	Yeah. Meat when I import, I import in KGs and when I sell it generally inside US I sell it in Pounds. If my conversion rates and my costs per conversion is not taken care into picture. We have seen that there is a chance that 40% we are under loss because of conversion issues.	ID
27.	D	So that is the importance of data quality in your organization. So that takes us to the next question. In words how would you describe the main characteristics that define data quality? So, what are all the attributes that define data quality in your organization according to you?	
28.	K	In my organization data quality is defined by accuracy, constancy, timelines and…. We classify data quality into different categorizations, one is validity, completeness, timeliness, accuracy, integrity and consistency. Say for example if I have to go one by one; validity defines, if my data that I have been consuming is valid from the source application standpoint or the transactional system standpoint.	AC CS TL
29.	D	I have one question in that. By validity do you mean the validity of data as the data itself or is it valid because of the time frame?	
30.	K	It valid because of the time frame	
31.	A	Is it old or new data; is it old or current data?	
32.	K	Correct. At inventory levels of an item which am not selling but is still there in the warehouse which is not of my concern, right? I need the current item from the current source. The older transaction may not be captured now which have no impacts. So, completeness is, do I get everything related to that or is this something that is in progress. Say for example, I have a purchase order. There are always negotiations between merchants and the vendors about what quantity has to be procured and so on. Until and unless it is not complete, the transaction is not complete consuming that data is wrong.	IR AC VD
33.	D	Yeah, consuming that data means you are giving half data to the reporting and analysis team, correct?	
34.	K	Correct. I am giving a wrong estimation.	
35.	A	So, as you said you mentioned about the data characteristics now.	
36.	K	Similarly, we can talk about item as well in completeness. Item when it is set up we need to consider the item only if it passes all its criteria.	ID
37.	D	Okay, so we will go to the next category that is data trade-off characteristic trade-off. The next question as you can read from the sheet is, how would you characterize trade-off between different data characteristics trade-offs in your work. So, what we are trying to achieve in this question is, your knowledge about trade-offs between data characteristics. I wouldn't call it as word of conflicting. But there are certain data characteristics as you have mentioned	

		yourself. To be a little compromising on each other. For example, accuracy & timeliness, timeliness & consistency and security and accessibility. So, these are certain trade-offs that we need to do where one needs to be given more priority over another. So, what are your thoughts on it and what are the trade-offs that you would consider that are important in your organization?	
38.	K	There is always interdependency between different pillars, they go hand in hand. In an ideal situation, everything is to be met. But there are cases where we have not seen ideal situations. Like I had a difficulty in setting up 6 items. Because of my 6 items I am not able to proceed completely, 100 percent along with my timeliness. But consistency is good. In such cases, there is always a trade-off between timeliness and consistency with a marginal impact. So, we do accept a 5% variance. So, given a situation for me where constancy is more than timeliness, I go with constancy and ensure that there a minimal impact because of that.	CST ID
39.	A	Okay so is there any other trade-offs that you could think of?	
40.	K	For me accessibility of the data is the top priority, in terms of integrity, accuracy and completeness. I cannot always define by completeness but completeness is always accurate. So, there can be a chance that I have all the data, but it is still not accurate. This may be because of a system issue or probably because of a functionality being introduced which is not according to the business operations. I such cases we ensure that accuracy is maintained and completeness is not taken into account. In which case those transactions which we feel is not accurate, we take it out from completeness to ensure that we only take the right data. For a particular example, I have 4 transactions on the same timestamp. Retail price has fluctuated plus or minus 5 in these 4 transactions. So, if I consider all 4, then I can consider that is the completeness. Now, since it is not aligned according to the data warehousing where I can at least by time identify which is the latest. I go with a sample example and consider that is being completeness, thereby ensure accuracy is maintained. Removing and just take the one which is relevant. So, there is trade-off between completeness and accuracy.	AS OC
41.	D	We here have provided three trade-offs, right? Accuracy, timeliness; consistency, timeliness; and accessibility & security. Apart from that is there anything that comes to your mind regarding trade-off?	
42.	K	No, I think what is documented is good.	
43.	A	Okay now let's go to the next one which is accuracy and timeliness. We here in this section would like to try to get out of you the maximum amount of contextual knowledge that you have and you have faced in your company working with both these data characteristics and where you have given priority over another. So that's what we are trying to get out of you. So according to you, how important are accuracy and timelines of data that you handled in	

		your organization?	
44.	K	Timeliness is a key in my organization because all critical decisions are taken on Monday morning and we ensure whatever we report is up to the mark and accurate to the team. So, we do accept some flaws, my accuracy need not be 100% even I report 95% accuracy. Because the decisions are based on timeliness and not the accuracy. The merchants or the buyers do add in their experience to negate the accuracy issue. Say for example, if I have a situation on my Monday morning reporting in my company, I see some accuracy issues. I try to reach my deadline on timeliness, so that I could at least give reporting view to the merchants and communicate the cautionary accuracy issue that we have faced for that week so that he could cross verify the data against what typically it would have been for last year or previous years and make a useful decision. Thereby ensuring that when it comes to an issue where accuracy and timeliness are not met - one of them, we give importance to timeliness.	TL ACT
45.	D	Okay, the next question that we would like to ask you is about what, what kind, variety or class of data would you classify as more important to be timely than accurate in nature or vice versa. So, what this question implies is that, like for example you work in the retail industry, like you said the situation on Monday morning. You expect the data to be more timely than accurate. Accuracy can dip a little, not too much but it can dip a little, the data needs to be timely. So, what kind of data that are you expecting to be timely and little accuracy fault is not an issue? For example, say item information is there, you wouldn't mind much if for a daily reporting to the managers if the brand or the size or the colour goes wrong right, if the data is not accurate. So, I would like to know from you regarding the kind, class and variety of data that you have worked with to understand what needs to be more important to be timely and accurate and vice versa.	
46.	K	In the retail domain, we have different classifications – business decisions, operational and functionality. When we look at the operational set up which is typically what is support in the merchandising domain. There are changes that happen every week. To give you an example, I have a store, I need to change item based on the selling pattern and the marketing that I have been doing for that week. In which case my sales for that number, if it is inaccurate or not up to 100% I still report what is my requirement and ensure data is available in time for him to make the decisions.	MD ACT OC
47.	D	Okay, this is the situation where timeliness is more important that accuracy.	
48.	K	Yeah.	
49.	A	So, this is for the operational data.	
50.	K	Yes. The other thing was related to?	

51.	D	A vice versa situation where like you said timeliness is more important than accuracy right now right? I need where accuracy is more important, even if data is late it's fine but I need 100% accuracy. What kind of data do you work with?	
52.	K	Coming to the second scenario, where there are financials or financials that needs to be published or reported or viewed by different people who are outside organization. Let's take data reporting that we do for vendors. In this case for me, accuracy plays a very vital role that is because of two reasons. One primarily because we are dealing with someone outside the organization, we are dealing with someone we procure the item from and the typical purchase order or the agreement that we sign with the external agents who supply the item. There is a law of violation charges. If the vendor doesn't agree to a policy after accepting the purchase order he is legally chargeable 3% of his income. Now in this scenario, if I am not able to report current week's data because of an accuracy problem even if accuracy error is 3%, I do not go ahead and publish this data. The reason being the vendor can sue us it's similar to your telephone bill. If I have a facility account, I cannot publish this data to the vendors. Because they can sue and it would lead us to much more problems. When there is a legal issue with finance and vendor's payments and so on, we don't mind re-checking all the data and ensuring accuracy is met rather than timeliness.	FD VD
53.	A	As far as my understanding is concerned, in operations you expect data to be more timely than accurate whereas while handling vendors you expect data to be more accurate than timely. Because, when you handle vendors there are issue of legalities that come into picture so you want data to be more accurate than timely. I think you have more or less answered the 10th question - could you explain the context where data needs to be more important to be timely than accurate or vice versa. So, you have answered that is question 9 as well I guess.	
54.	K	Yeah.	
55.	D	So, we will go to the next one which is constancy and timeliness. So, there is a very fine line that differs between accuracy and constancy, I think you know that, right? We will go ahead with the same procedure that we did previously. Similarly, how significant is consistency of data in the retail domain?	
56.	K	In the financial world in the retail domain, consistency plays a vital role. This is basically we publish our financials to stock market. And any change that we need to record has to be validated before we publish. In that case, we need to be consistently reporting what the organization is doing and ensure the completeness is not taken into account.	FD OC
57.	D	For you and for your organization you would say consistency and completeness go hand in hand.	

58.	K	Yes, they go hand in hand but completeness. We cannot wait for all the data to come to me to publish something. Here consistency in terms of when to publish, how to publish if there is some…Say for example, I opened a new store and this new store data has not flown into my annual revenue. Now, my data that I am publishing is not complete. Because some of the aspects that I have invested are not yet given priority or I have not yet considered it. That's why we have a concept called as mature store's reporting only.	CS OC
59.	A	Is it specific to Target then?	
60.	K	It is specific to target. So, in such cases we wait for a year for it to fully operate and give me a clear picture for me to consider in my financials. In which case I am fine with not complete data. But I am consistently publishing only those stores which are mature.	FD
61.	D	So, we will go to the next one now. Again, on a similar line, what kind, class and variety of data would you consider to be more important to be consistent rather than timely and vice versa? So, the question as previously, what kind of data has to more consistent than timely and what data needs to be more timely than consistent.	
62.	K	Sales data needs to be timely. Inventory data can be consistent. Financial data has to be complete.	SD IR FD
63.	D	Complete as in consistent again? Complete and consistent because complete means all the information needs to be there and consistent means it needs to be consistently there not erratically there.	
64.	K	Yes, I cannot consider a store in one cycle and not in the next cycle. If I am doing a company level financials it has to be against all….	FD
65.	A	Okay, now I get it. So, in the report no column should be blank.	
66.	K	Yes. But it has to be accounted against something. We cannot financially say I don't know what this is. If you do not know something what it is, better not report it.	FD
67.	D	So that leads us to the next question. Could you explain the context where you felt where data needs to be more consistent but less timely and vice versa so on similar lines can you explain a context where consistency plays more importance and another situation where timeliness plays more importance?	
68.	K	As I told the financial data needs to be consistent against time.	FD
69.	A	And what needs to be more timely than consistent, as you said sales data. Can you think of any other context that implies to this situation? You told me about financials and sales. Is there anything else?	SD
70.	K	Item data has to be consistent but not timely whereas inventory management and Inventory Relief. Because based on that I will decide what to be pro-	IR

		cured, what not to be procured, how to supply. When there is a dependency on particular channel to fulfil the sales. Those channels the data needs to be timely.	
71.	D	Okay, based on your report they will make future calls and changes to be done.	
72.	K	Exactly, to redesign my store, to redesign my inventory based on the marketing that I have done. Say for example, there is a new DVD movie released and I do not have inventory at all. So that's a bad situation because I have marketed for it. In such cases timeliness plays an important role to say I do not have data to fulfil all your requirements/marketing needs then we procure and ensure that we are completely available for sales.	ODC MKTD IR
73.	A	I think we are done with consistency and timeliness. So let's go to the next which is accessibility and security. As far as I have understood it's one of the most important thing.	
74.	K	Accessibility is a crucial thing. Say for example, anything that is dealt internally can be accessible for everyone. When the data is available for external vendors/users – say for example there are 2 vendors who are selling the same item. The procurement cost for each of them can be same or different. But we need to ensure each vendor's confidentiality is met and the security of the data is only pertaining to the person who is supposed to view that data.	VD
75.	D	So we will go to the question now. Based on your opinion what is accessible data and what is secure data.	
76.	K	Accessible data, within the organization whatever business decisions and financials can be accessed by everyone within the organization. When I mean within the organization, a team member or an employee of that organization.	AS FD SE
77.	A	What kind, variety or class of data do you consider to be more accessible than secure? So, the first question is what data you would consider to be more accessible that more people can see it than security.	
78.	K	I would say item data in the retail domain needs to be accessible by everyone because retail domain runs on items. So, this is one data that needs to be accessed by each and every one. Be it financials, be it vendor management, and be it supply chain. Everyone needs to have access to item data.	ID FD VD SCM
79.	D	So how about other fields like presentation, inventory and stuffs like that?	
80.	K	They all go hand in hand. I you do not know what item to keep in what location, how can you plan about it, how can you procure it, how can you market it, how can you collect financials, how can you collect margins, how can you identify profitability. How can you report to the outside world this is my	AS

		profit.	
81.	A	So more or less 80-90% of your data in retail domain should be accessible to everybody. So, security is the most important thing. So, what kind of data would you classify as secure data?	
82.	K	Anything that you publish financially needs to be secure and anything that you deal with external vendors, external people having access to internal systems, those will have the highest security.	FD VD
83.	D	And when you mean security, who gets to see those data?	
84.	K	Yeah, who gets to see, who gets to view what decisions they can take and so on. There is dedicated secured teams formed, like vendor operations team in case of vendor, who deal with external vendors itself. Such people look at secure data. To give another example HR of the company and financials team are the only people who look at financial members. Before publishing it probably to the stock market.	VD HR
85.	S	So, 80% of the data like inventory, presentation and item all those data are accessible whereas the financial and the vendor related information are more secure.	
86.	K	80% of all merchandising data that I work is accessible to everyone.	MD
87.	D	Now the last question in that category. What data would you or other data analysts in your company consider most important to retrieve during a system crash. I am thinking this would be a very unlikely situation since you could have all the things backed up, but consider a system crash happening and people are trying to retrieve data back to your data warehouse. What would you consider most crucial? What would be the first thing that you would check?	
88.	K	In a retail domain, every decision can be made out of sales. I would ensure all my sales data being captured accurately and also financials. So, at the time of an event where I cannot report my inventory levels or I cannot report my item set of data info, I am still fine and ok to operate in hazard situation where there is a calamity. I can still live with the financials data and the sales data.	SD FD
89.	S	So, the most important data that you would retrieve would be finance data and sales data. So, we will go to the last question now.	
90.	K	I will add on to the last question. Whatever data that I am obligated to publish to the outside world, be it my vendors, be it my customer or be it my stock market, those data I will retrieve first.	VD
91.	D	No question 18, this is kind of a suggestion from you as you have worked for quite a long in retails. We would like to know from you experience, could you suggest any other trade-offs which would be crucial for data management	

		and how? Like we suggested three right, apart from that is there anything that you would suggest to us? Completeness was one you said, right? So, you would counter completeness with what?	
92.	K	I would counter completeness with a better and smoother operational processes.	ODT
93.	D	So, timely and consistent flow of data?	
94.	K	Yes. Consistency and completeness are all based on the functionality of the retail domain. When I mean functionality the core merchandising operation data. The need of the hour is that there will always be changes but we need to streamline our process to follow a single process and not have multiple process for the same set up. So, in such cases, my consistency and completeness plays a very vital role to tell the operational team that 'look because of these I am not able to consume the data and it is not consistent'. There are variations because of different channels or different streams.	ODT CS
95.	S	So, like you said, multiple processes might result inconsistency of data as well right? So, as you said to improve data quality, you can reduce the number of process also.	
96.	K	Yeah.	
97.	D	I think the questions are over with this. It was quite helpful Kiran and thank you for the interview.	

Interview 2 – Sandipan Ghosh

Interview Date: 24th April 2017

Present: Sandipan Ghosh (S), Athul (A) and Dilip (D)

Duration: 38 minutes

Transcribed By: Dilip

Line	Speaking	Text	Code
1.	D	Hi Sandipan, thanks for the interview you are giving us. Hope you have gone through the document that we sent you. So, you might have got what we are doing for our master thesis. Basically, it's about understanding data quality management in the retail industry. What we are doing is we are trying to understand the trade-offs between different characteristics of data quality. So,	

		for example there are many trade-offs that can happen. Primarily we are focusing on three that are mentioned in the questionnaire which is accuracy & timeliness, consistency & timeliness and accessibility & security. We found these three to be most interesting and without it would most appropriate for the retail domain as well. So, you have an idea what each attribute are right? So, we will go on to the questions now. The first is a very generic question. We separated the questionnaire to different categories. The first three are from the introduction part. The first question is, please provide us a brief introduction about yourself and your work experience.	
2.	S	My name is Sandipan Ghosh. I have been working in the software industry for the past 17 years. Mostly started with database administrator database engineer, spend a lot of time in quality engineer, now into Hadoop development focusing mostly on the data engineering part.	
3.	A	So how many years of experience do you have in the retail domain?	
4.	S	Retail domain probably 9-10 years.	
5.	D	And it is completely in Target, is it?	
6.	S	Not exactly, 2 companies. Majority is target and then in Tesco. I worked as a contractor in Tesco. I was supporting one of their projects.	
7.	A	Can you specify how many years in Tesco as a contractor?	
8.	S	As a contractor, Tesco was one and a half years.	
9.	D	And how many years in Target?	
10	S	6-7 years.	
11	D	What is your designation/role you play?	
12	S	The designation is senior data engineer and solution analyst.	
13	A	Could you please elaborate about your role in the retail organization that you work for? So, you can go ahead because we are concentrating on the retail domain, you can go ahead and tell us what you did in Tesco and what you did in Target as a senior data engineer.	
14	S	Let's start with Tesco. There majority of my work was in the database administrating part. In the phase like design and develop a data warehouse to hold that data. It is a classic data warehouse problem, we followed the dimension modelling. Majority was how you flow your source system to a data warehouse so that you can capture all the attributes and make meaning out of the data. Put into cubes so business can be a bisectional analysis and find a major business outcome from that. The problem what we were finding solution was with classic data warehouse problem. So basically, the solution was how you	REP REC

		are going to store data in a very effective manner so that you can capture all the attributes including sales, time, customer, every other information. Target was quite different. I joined as a data quality engineer. So, the majority of the focus was not developing a data warehouse it was rather supporting the data. Not the technology part but the data part. My major role was in finance. My major role was try to figure it out if the data I am seeing is correct, if the data is loaded perfectly and loaded on time. If tomorrow manager comes and see the report. Are they seeing the correct data and if they drill down are they doing a right drill down? With that comes the automation, so basically instead of doing everything manually we try to automate everything. So, everyday morning and you see a report that tells you how your data is behaving. How data is against a certain limit or certain change, how is the data. You are seeing the health of the data.	
15	A	And so, in Target which domain or which subject area are you working on currently?	
16	S	I started with finance, right now I am in supply chain.	
17	D	So, you would be working with lot of data which you would require certain data characteristics. Like for example certain places you would want the data to be timely, certain places you would be okay to compromise on timeliness but you would want the data to be accurate. Such trade-offs, you would be working on such things, right?	
18	S	Yeah.	
19	D	The introduction part is over, let me go to the next part which is defining data quality. Could you explain according to you what is data quality based on what you worked on in Tesco and Target.	
20	S	When I worked in Tesco, data quality never existed there. Data quality was always a responsibility for a developer or tester to see that input data is flowing correctly and timely fashion to a warehouse. Do I ever check the information and see in the warehouse if correct or not? Not in the source system I am talking about the business perspective. So, if you see a source having 10 dollars, you see 10 dollars flowing.	CS TL
21	A	From the source systems?	
22	S	Exactly, I always do source to destination. But I do not know if the 10 dollar is correct, I do not know if actually sales or not. I do not know or I never care about it.	SD
23	D	Taking the source as a source of truth.	
24	S	Exactly. For me what is important is if I am loading the correct file with all the data. Then figuring out if the meaning of the data is correct or not. So that's for the Tesco part. For Target, it's completely different, I think there	FD

		was a time where data quality was skating up and people slowly realized that quality of data is most important part. There as I told you I was in finance that is the place where you want your data to be correct and you can actually compromise on the timeliness. Obliviously, how much salary you pay to your employee has to be correct to a minute details than if the file is loaded to little late. So that's my feeling of data quality. I think business slowly realized that quality of the data has to be correct, has to be timely, and you are seeing the correct data and you are not missing out anything.	
25	A	One side question to this is, what is the accepted percentage of data quality? Do you know how much is the error percentage that you can allow for quality?	
26	S	I think it varies from company to company based on subject area. For sales and finance it's good to be zero like zero tolerance. You don't want mess up with the data part. Something like inventory or supply chain where you are calculating the forecast base on some other attribute. So probably that you could say 20% is your limit of tolerance. Finance as in 0%, officially it used to be .5-10%, but in supply chain I am seeing the tolerance is around 20-30%.	FD SCM SD
27	D	A bit higher is supply chain and that is acceptable?	
28	S	yeah that is acceptable.	
29	A	So, I think you have covered the next question as well. Just for informing the next question is how important the data quality in your organization. Do you have anything additional to tell in that?	
30	S	I can give you a scenario. What is happening right now targeted to figure it out how they can minimize the last-minute delivery? If I order from target.com, will Target send me in a fedex or will it use its own track or will is use UPS which the government postal service. Say I am calculating a risk whether my data is wrong i.e. I capture a wrong data for customer that how many people have ordered from Target.com and where people have ordered. If people who have ordered from Target.com is higher than any other area right. I do not capture data correctly my entire calculation goes wrong and my entire risk of taking that decision is going for a toss. That is the importance of data quality right now.	ODC
31	D	One thing I would like to say here is, if you could give us examples like you told right, that will be very helpful for us. Because we are trying two scenarios where one attribute is greater over the other. So that is our aim. So, the next question is, what you think are the main characteristics that define data quality? So, what this question implies is, you are working in a specific area right, you worked in finance and you worked in supply chain management. So, the data characteristics in finance will be a little bit different from the data characteristics in supply chain management under the overall umbrella of retail domain. What are the different characteristics that you will use to	

		identify data?	
32	S	I will concentrate from the quality perspective. So, in finance accuracy has to be high. If you load a wrong value for an employee salary will screw up the entire organization maybe more. I think you noticed on every pay slip you will probably get after two days, right? They have multiple processes for the accuracy to be there. Supply chain is the other way around we too bother about timeliness not accuracy that much. Like I was telling 80% is the acceptance level. My data is not accurate on-hand or backroom targeted for the store. That is not exactly accurate. But I need in timely manner because I am calculating for the next two frames. So, let's say I got 10 items and I missed 2 item. So, I still can go ahead and do a calculation for 10 items. But let's say my data came late. I did not load that stores data at all. So, the next week the planning for the store went missing, right? So, the stores stock goes haywire. There your timeliness is more important than accuracy.	FD SCM ACT
33	A	Any other data characteristics that you would like to mention?	
34	S	Probably I can in depth security part a little because finance is a very secure domain. Not everyone can have the access. So that's changed the entire landscape of how you design your solution or architecture of the software. Let's say some other team needs such a report from a finance. You would probably create a layer of views/attributes to mask the credit card number/employee number. So there so many design goes in to mask the data, to expose the report to people. Supply chain do not care about the security that much because this is open data and it's open for everybody. Every one of us in Target have actually the access to the entire supply chain data not everybody in Target has read access to the finance data. Retail organizations tend to give access to supply chain information but they secure their financial data. Security is also quite different.	SE FD SCM
35	D	The next question – data characteristic trade-offs. Apart from the three we have taken – accuracy & timeliness, consistency & timeliness and accessibility & security. Is there any other trade-off o that you can think of in your field of work?	
36	S	No, I think that covers all.	
37	D	Next set of questions is about accuracy and timeliness. According to how important are accuracy and timeliness of the data in the organization that you handle? For example, how accuracy and timeliness plays a picture in supply chain management data.	
38	S	Supply chain accuracy does not matter that much. You will accept the data if it's 80% correct. Timeliness is little important than finance, the example I gave right. Then again that truck movement. Instead of having accurate data you need timely data. It's okay if you miss a carton, that's also fine than missing an entire truck schedule. So that is certain percentage you can expect.	SCM TL FD OC ODC

39	A	In supply chain management what variety, kind or class of data would you consider more important to be timely rather than accurate. Is there any separate set of data that you would think needs to be more timely than accurate and vice versa?	
40	S	Probably on hand, in the sense how many items you have in your warehouse that information if I do not get 100% correct, it's fine. But it certainly I want to get it every day. So, we capture the information twice a day but unsure if it's 100% correct but we do capture the information twice a day. Because we really want to know what is happening and if you try to tie that with sales it may not tie. Because there is a difference in data. So, what we are trying to say is timeliness is more important than accuracy.	OC ODC ID
41	D	That is for on hand data, right? Can you define what on hand data is?	
42	S	On hand is basically consider selling an iPhone. So, there is iPhone in the store in the display. But I also have lot of iPhone in the stock in warehouse. So once the iPhone in the store gets sold, I order a new one from the warehouse. The warehouse information or how much I have in the stock is called on-hand.	ID
43	A	Is there any data in supply chain management that you want to be more accurate than timely? It's okay if it if its reaches late in your system but it needs to be 100% accurate.	
44	S	A good example is sales. The sales data gets loaded in a week's delay. So, I probably get the information of sales in a week's delay. There is a delay of a week in the data which is actually acceptable. But I would like to know how much I have sold. I am having my business on sales so I want it to have 100% accuracy. But timeliness is fine. I can relive with a week of delay but I would like to know how much I am selling.	SD TL AC ACT
45	D	As far as I have understood here, you want the in-hand data/item data in your store to be more timely than accurate and you want to your sales data to be more accurate than timely. So now we will go to consistency and timeliness. Accuracy means data is coming accurately and consistency is the information contradiction frame. Like for example there is a column which gender, is it coming as male/female or m/f or 0/1, there is a difference in data. If it's constantly coming as male/female it's consistent. So, the same 3 questions again how important is the consistency of data in the retail domain and what kind of data needs to be more consistent than timely and vice versa and some context where you have felt the same.	
46	S	In the project I am working now, we are facing this same issue that you explained, consistency v/s accuracy. We are getting accurate data but it is not consistent. I am talking about a couple of flags, we are getting the correct data about the flags - yes, but the problem is there are a lot of source systems and they are not consistent with each other. So, one of the store probably	CS AC IR

		talks about out of stock as OO and some others use it as OOS. I am getting the correct data but it is not consistent. So, when I am trying to forecast or try to generate how much item they need next week, I am not able to figure it out which store is saying which item is out of stock. If I consider OO as out of stock, then I am probably serving some of the stores. But the stores which are using OOS for out of stock, am not serving them. That is a classic example.	
47	A	This is data that is accurate but not consistent. Do you have anything to compare consistency with timeliness?	
48	S	I never encountered anything.	
49	D	You have given an example of comparing consistency and accuracy. I think we can add that as well in the questionnaire, it's a good comparison. Consistency and accuracy are sometimes thought to be the same but they are not. The next set of questions is regarding accessibility and security. How important is accessibility and security? As far as data is concerned, which data according to you is accessible and which data needs to be secure in the data that you work with.	
50	S	I think finance and HR has to most secure than any other data. I do not know whether you are aware, where certain people got read access to the HR database and they copied entire data to their local and made everybody know everybody's salary. So, I would say for a company or any other organization HR data is quite important and finance as well. Same thing goes when you are designing it. There were a couple of report where you need data that is masked, you cannot use customer name, customer credit card number. Even that goes in your design. So, accessibility limited there. If the company puts that norm everywhere and thinks they need to secure all the data, not only HR and finance. Let's say they have secures sales information, customer, booking information, twitter information, Facebook information, social media information. If I am working on a project where I do not need the secure information but need the open information, I eventually end up delaying my project. Publicly available details have to be available publicly, do not go and over secure those. I think this is a classic example.	ACS SC AS HR OD OC
51	D	So, taking you to the next question. What variety, kind and class of data would you classify as more important to be accessible than secure and more secure than accessible.	
52	S	yeah HR and finance has to be more secure than accessible and you should have layer after layer security to get that access. And other information like, sales, items like what is the all item sale in Tesco or target, what is my sales, how is my inventory, where I get most of the order. That information has to be open. That is the core business for company like Target. So, if you don't give information to right people at right time. You are eventually delaying the project. One is publicly available data, like people tweeting, that is publicly available, you do not have to secure that data. But let's say you got informa-	HR FD SD SM

		tion about your twitter handling id, your location – those are not publicly available and needs to be secure.	
53	A	So, what you share in twitter is public info less security but things like sharing location on target store is more secure.	
54	S	Yes, Facebooking or tweeting is publicly available and you don't need to secure that data. But If i buy a data from certain user, their username, location those thing is not publicly available.	SM
55	D	So where do you buy data? From third party users?	
56	S	You can buy data from Facebook and from twitter. Target has bought 3 years of data from twitter like anything about Target in twitter is given to Target from twitter. Some information is obviously public- that I talked about Target. But who am I being not public that needs to be more secure. I will give an example, one of the company needed a marketing basket analysis. They are a beauty product company so they want to launch a new eyeliner. So, they need to see whom they should reach to make this eyeliner famous. The eyeliner should be made for customer who goes to beauty parlour and let's say I got 3 years of twitter, Facebook data. I did an analysis and I got the beauty parlours where they have got huge amount of customer who can afford this product. So, will get catch hold of these locations and I would advertise and go and talk to those owners to promote my product. This can be done only if we get certain information like where their actual location is. This is some information which you will get locally, so you have to buy it and there is disclosure that you need to sign that you cannot disclose this information. This information needs to be very secure. But twitter posts like 'I use this eyeliner' is very great, so you leave it publicly.	SM ID
57	A	Now it is widely use right, using twitter information to see how well the product is going right?	
58	S	I will tell you the company name, Tiara beauty. They used twitter information and the eyeliner was pretty successful.	ID
59	D	Through target, is it?	
60	S	No through other company.	
61	A	We can go to the last one. What data would you or other data analyst in your organization consider most important to retrieve after a system crash. So, let financial data go since we know that is very important and no company will let go of that. But apart from that now that you are working in supply chain management, what is the most crucial data that you would consider retrieving during a system crash? You would look for the first to be reconstituted in your data warehouse?	FD SCM

62	S	Any data in the company, any company has multiple backups. Even when there is a system crash I don't think there would be a data loss.	
63	D	It will take some time to put back the data in your system, right? So there will be an order in which you consider to be more crucial to be loaded first?	
64	S	Right now, everybody uses multi node, multi latency systems where when one goes down other will be up, so basically, we do not experience that problem. But definitely HR and finance is most important information in Target. Now if I do not have backup of one month sales data, I can still survive. But if I do not have stock information like how much item I have in stock. So that information get lost then I am in pretty much in pressure because I cannot replenish my store next week right.	HR FD OD
65	D	So, stock, finance, HR is crucial.	
66	S	So, when system crashed you definitely want to do backup of your online information as well.	
67	D	I think this covers everything properly.	

Interview 3 – Suyash Kumar

Interview Date: 26th April 2017

Present: Suyash Kumar (S) and Dilip (D)

Duration: 48 minutes

Transcribed By: Athul

Line	Speaking	Text	Code
1.	D	Hi Suyash. Thanks for the call that we are having right now. I hope you have read through the questionnaire that we had sent you. Athul and I are working on the thesis to understand different trade-offs between in data quality characteristics in the retail industry and we are trying to create a template that will help in understanding the trade-offs. As far as the thesis is concerned the trade-offs that we have considered basically are accuracy and timeliness, consistency and timeliness and accessibility and security. With your experience that you have you would basically already know what these fields are.	
2.	S	Yeah Dilip, I know these are data dimensions and I can surely help you on your thesis.	

3.	D	Thanks. We have divided the questionnaire into different sections. The first part is introduction. Can you provide a brief introduction about yourself and you work experience?	
4.	S	I am Suyash. Right from Infosys I was working in BI and the QA aspect of that. After that I joined Target in 2011 and then I started working for the data quality architect of the warehouse they are maintaining. Their warehouse is actually spending right from IBM to Teradata to Hadoop system so they have vast subject area in all of these systems. I was taking care of the data quality aspect of those all. So, have neatly 7 years of data quality experience all of it in retail industry. One month back I joined Quotient technology which was initially known as Coopers.com based on California and there also my role leads to the same data quality aspect in the warehouse they are maintaining in the Hadoop system. In Infosys, also I was dealing with the same DQ data aspect of the business intelligence like, reporting services or the analysis services, so all in Infosys, Target as well as Quotient Technologies.	REP REC
5.	D	70-75% of your experience is in retail domain, right?	
6.	S	Yeah.	
7.	D	The next question you have already answered, how many years of experience do you have in the retail industry.	
8.	S	I have nearly a? year of experience in retail. In Infosys, it was actually into supply chain and manufacturing. After I joined Target and Quotient, it's all about retail.	
9.	D	In Infosys which client were you working for?	
10.	S	It was a product information engineering, so we were creating a product which as the need of manufacturing industry seeing their key performance indexes and performance of the industry. So that was towards manufacturing. But after 2011 continuously I am working on the retail domain only.	
11.	D	Could you elaborate the role in the retail organization that your work for – what your worked on and which all areas you worked on and your designation.	
12.	S	Target is the 2nd biggest retail industry in US. It is US based and its complete business is in US. Target maintains all relay big warehouse of data that they capture from the sales, inventory, guest, unsaleable and there are different other domains. In the 6+ experience that I have in Target, I have worked in almost all the subject areas. So be it inventory, sales, unsaleable, it is nothing but consumer coming and buying the product. So, I have detailed domain knowledge from all of these subject areas. When I joined Groupon, it is actually into the coupons. So, these are the different domains that I have worked upon and my role was data quality analyst. I have involved right from the requirement gathering of the creation of the warehouse till its delivery in	REP REC

		production and it is getting into support and maintenance phase. So, I have out and out knowledge of how retail domain maintains their warehouse.	
13.	D	Can you mention about your role/designation in Target?	
14.	S	My designation in Target, I joined as a data quality analyst. I then got promoted as senior data quality analyst and in Quotient also I am senior data quality analyst.	
15.	D	The next question is about your understanding or how you would explain what data quality is	
16.	S	I would explain data quality as what we are maintaining our data for any of the organization or any of other domain, whatever data we are maintaining is providing the complete picture and the complete business information that the client wants. It should follow all the 5 dimensions of data quality which is accuracy, timeliness, accessibility, ability, integrity – all should be maintained. Data should be able to take the correct decision that when we are implementing it in the organization we are taking the correct decision based on the correct data. That is what quality means to me – maintain the at most business usable data in the warehouse.	OC
17.	D	Could you give me context where you have faced issues in data quality and how it affected business?	
18.	S	I will give a most recent example. There are a lot of complaints from the guest- we call consumers as guests in Target. So, we have faced a lot of problems where quest is reaching out to us asking that they have already opted out for the email notification or mail coming to their place that they don't want Target circular coming to their place. But Target mails telling about their recent launch or the digital promotion they are doing, they don't want. They have already opted out but it is still getting delivered to their place. Sometimes even if they have changed the address and they are the recent occupiers of the house but they mail is coming in the name of old occupants' in the house. And those people who ate dead, that is also a frustrating situation. This kind of data is still coming and we have a lot of complaints. That is happening because both from the MDM side that is providing the quest data and the data warehouse that we are marinating for the quest. There are some gaps between how we are maintaining. I have worked extensively on those aspects and it is somewhat looking good now.	CD AC
19.	D	Is there anything that you want to add on to say how data quality is important to Target.	
20.	S	It is very important. In retail industry, it is all about data. It is about sales data, inventory data, and unsaleable data. Every stage of your operation right from procuring the product in overseas market as well as your local market till it reaches your guest and even if it is returned and all, you get a lot of data. You need to maintain the utmost integrity and accuracy in your data and	REC REP

		with time factor so that the correct information is being made at every area of your operation. Right from procurement, till it reaches warehouse, till it reaches guests and if it is returned too. You need to maintain data quality at every aspect of your operation and then and there you are a successful organization. You must have seen retail in Australia, America, Canada, UK lot are getting closed because they are taking wrong decision based on the wrong data they have. So, data quality is very important for any organization and Target because it has competition in the market.	AC
21.	D	What do you think are the main characteristics of data that you have handled in the organization that you work?	
22.	S	It is spanned with all the data dimensions that data quality aspect deals with. Let's start with accuracy, integrity, timeliness, availability, maintain the RI, maintain the completeness of the data. So, these are all important aspects.	OC
23.	D	There is this model that we have used called the Eppler's model wherein he talks about conflicting data ideas. For example, if you take accessibility and security, you can't make data accessible and secure at the same time right. If accessibility goes high then security goes low and when security goes high, accessibility goes low. So, we are trying to find out a relationship were which data needs to be more accessible when compared to security and which data needs to be more secure than accessible. You have worked across different areas in Target handling different data. So, you will be able to tell us more about that. The question is how you would characterize trade-offs between data characteristic in your field of work?	
24.	S	You have answered accuracy and timeliness, consistency and timeliness and accessibility and security of the data. There are more like cost of maintaining data and the quality of it, accessibility of the data and the timeliness of it. There can be multiple different aspects. But those three which you have mentioned along with cost and quality because that is also a big trade-off the retail industry goes with, I think that covers most.	
25.	D	You can spend millions of dollars on maintaining a warehouse and its quality but to what extend you can put money and how much ROI you are getting is the question. We will mention these two as well. When we had an interview with Sandipan he was mentioning about accuracy and consistency which was also interesting. So, we will go with the first trade-off considered here which is accuracy and timeliness. According to you how important are accuracy and timeliness of data that you handle in your organization and if you could give me an example as well.	
26.	S	Accuracy is nothing but the data precision like we maintain a particular data, how accurate is it to take business decisions. Timeliness is nothing but when I need the data I should get it. In different subject area, my need of data might be different. So, for the sales data for a particular black Friday or holiday season there will be lot of sales happening. What I want is, every 15/20 min-	AC

		utes every half an hour, I should as a leader should get or report out of that. Which store of mine, what product of mine, what promotion of mine, which are those guests, what are their segmentation, who is buying what. Do I ever cater to all of their needs, all of their want in a proper manner? If there is lot of return happening, I need that information very promptly. So, timeliness is a big factor now. I want to take decisions right away, if the promotion is working, let's increase it, if the promotion is not working let's kill it. If inventory is getting out of stock somehow get the inventory from somewhere because guests are stocking a lot of that so I want it to be provided. So, at that timeliness is very important. During holiday season, out of stock is very important matrix for me. I can be very much interested to understand in all of the stores, what is the current situation of the stock for a particular item if it is selling like a hot cake. There might be a situation where a few of the stores have gone completely out of stock. But that was not an accurate information. There were still some stock left but that is giving me a hint that I need constant supply of that item of I want to catch hold on buyers. I can little low on accuracy not completely but very prompt on the timeliness. But let's see a situation where I want a weekly report of a particular promotion on how it has worked for particular area. But since it is a weekly report I am not very much concerned on time. You give me Saturday morning; Friday evening or Monday morning I am okay. I will take a decision based on the last week complete data that you have provided me. But if it is not accurate, if it is not very precise and if it is not giving me a correct information, that's going be a big pain. Here I cannot trade-off accuracy, can little trade-off on the timeliness. You give me later but you give me accurate.	TL ACT
27.	D	It helps us to understand which data needs to be accurate and which data needs to be timely. Can I ask a question here, what happens to the weekly reports that happen in the festive season?	
28.	S	During the holiday season, that is one quarter where all the US/European companies make a big chunk of their sales/profit. So, lot of things are getting sold. That weekly reporting if you are giving me, some matrix I will be interested, like how the sales is happening, overall organizational perspective, some high-level information like how much return is happening, how much defectives are coming in the store. Those if you give weekly report, will be seeing. But how the sales is happening, how the inventory stocking is happening those things are very critical. I don't want a weekly report out of that, I want an hourly if not in minutes. I want as soon as the sales is happening that sales have happened. That level also people work, so if not that at least hourly I should be aware of how our inventory is turning out to be, how our guests are buying the things and how sales is happening in the store.	ACT
29.	D	I think you have answered the entire set. What kind of data would you consider more required to be more timely than accurate and vice versa. So what kind of matrix or data would you consider to be accurate to the dot but timeliness I can adjust and the other way around as well?	

30.	S	According to me I think, inventory and sales information should be timely. Timeliness is a big factor there because we have to take quick decision based on the inventory stock and the sales happening at different area at the micro level, not at the company level. But as far as accuracy is concerned, I want clear picture on the unsaleable, guest etc., I want accurate than timely. I want actual penetration of digital promotions and unsaleable or guest returns. I want accurate information because that going to lend a lot of profit at the retail industry.	AC TL
31.	D	Can you think of any other context regarding when data needs to be timely and accurate? Any other situation where you felt one is important over another.	
32.	S	Accuracy and timeliness is a good trade-off. When you have items in your warehouse, maybe you have procured that overseas or locally or some of the guest return happening and all. So, when you reach out to the vendor for charging it back or some of the return which is happening that is causing a lot of dent in the actual profit margin you are getting out of the sales. So, you are reaching out to the vendor, so those are the situation where based on the actual item that has returned or the warehouse that you are maintaining with the items and all. You are talking to the vendor, you are doing marketing return, at that time you should have accurate/precise data so that you can charge back, you can make your profit good, and you don't lose out a lot on that. So those here you have to think that whether the data needs to be timely so that you can charge it back fast to the vendor and get your money back fast or you should be completely having accurate data whether it's maintaining the time or not so that when you are reaching out to the vendor, you are not losing anything. So those are the tricky situation which you have to think whether the data should be accurate or it should be timely.	ACT AC TL
33.	D	So as far as I understand from what you said, handling vendors you would need the data to be more accurate than timely, it's okay there is delay of one/two days but if it is inaccurate it is a loss to the retail company.	
34.	S	Yeah, my suggestion would be the data to be more accurate than timely. But that depends how you have signed license with the vendor because if you are very late, maintaining of accuracy actually breach the license since they will have a time period which they will be taking and charging it back. So those are the tricky situation, it should be more accurate rather than the timely factor. But according to the vendor, you have to take a decision.	AC TL
35.	D	Timeliness is not a major trade-off here when compared to the other context that you gave. Now we will go into consistency and timeliness. Consistency – if data coming from different sources have been normalized/ de-normalized. Like if gender has male/female and there is no other value like m/f, 0/1.	
36.	S	In the retail industry when they are maintaining the warehouse, they are getting different kinds of data – sales data, inventory data. Not only that, if some	CS

		retail industry will be actually not selling the product but partially getting it and working with different retailers. Like in Quotient where I am working now we have to deal with lot of retailers – Walmart, Target, Quotient. Now when they are sending the data, if they are maintaining for a particular field the same kind of information like you said the gender code. It should not be coming like yes/no, 0/1, tick something like that. So that when you look at your inventory, you know the information correctly. So, consistency is a very important factor otherwise there will be a big problem in reporting your information from the warehouse.	
37.	D	Is there any inconsistency that you have seen in Target data or in your new company Quotient? Can you tell me in what kind of data you have seen the inconsistency?	
38.	S	In all of the warehouse it is a very big problem, consistency in the data. Both in Target and Quotient, I have seen many inconsistency in data. I will give you some example. In sales information, when you are dealing with a company like Target where there are millions of transactions happening every day across US and in target.com and different digital platforms where the guest is interacting with us. You have different platforms where the company is catering to the guest. We are getting different kind of data, something when they come and buy directly, through target.com and mobile app. So sometimes few of the information misses and when they come to us sometimes it will come yes/no sometimes it will come 0/1. As a data warehouse, we have to maintain consistency around all of the information which we are capturing from any of the sources. What happens is, in the sales we will be getting transactional types which is nothing but the sales information, return information and exchange information. So, what we were seeing in the queries that we were writing for the returning perspective is that, for few the return data was quite high. So, when we drilled down further to find the real reason behind that, what we found is that we were provided the class coded which is nothing but the transactional codes, – Target was maintaining like 0,1,2,3 something like that in which 0 and 1 was for sales, 2 was for return and 3 for dealing with the exchanges – for mobile sales or all the DEV interaction with the guest were coming with some other codes. So, the data was inconsistent from the different platforms we were capturing.	CS SD
39.	D	Okay, but there we some data that needs to be very consistent/regular. For example, when I was working in Target there was an issue where some people used to enter location with some spelling and some others with some other spelling. So, when you pull records those record which are entered with some other spelling doesn't come into the report at all. According to you in the different areas that you have worked on, which area very crucial to be consistent so that your report is consistent and which data is okay/relaxed in case of consistency.	
40.	S	Inventory and customer information are the key areas where consistency should be maintained. For example, guest data which we are maintain in dif-	CS

		ferent tables, those are the real data on which we are interacting with the guest. Few of the tables we have captured the correct guest information with the correct email id, phone numbers whatever we are capturing are very clear and capture exactly the way it is. These are not consistent since somewhere we will be maintaining some address and somewhere else we will have some other address for a single guest and all related information will be inconsistent. Then it is bad and will lead to a lot of issue not only in the organization that we are dealing with but for those the particular guest also. So, consistency is very much required. Same in the inventory area, I want consistently good information. I should not give something like in the web world I have a laptop for sale but in the store, I don't have a single one, I don't want that. I want a uniform warehouse – how much inventory is there in stock and I want the information consistently to make the right decision. But some data like sales, those have already happened, little bit of inconsistency will also be there but I want the timely reports to take the decisions there based on that. I want a timely information coming to me, there I can deal with consistency not with timeliness and in the data like guest inventory I cannot tolerate the inconsistency – it is okay if the mail is delayed but it should reach the right person, a person should not get baby product for digital promotion.	TL CST
41.	D	There is this major news that happened in Target about the baby product that made headlines. Now that you have covered the context as well about consistency and timeliness. The last set of questions is about accessibility and security. You would have worked with data that needs to be accessible and available to all the people and some other data which should be secure and not available to all the people. Based on your opinion, what is accessible data and what is secure data in the areas that you have worked with.	
42.	S	When we are exposing something like what is the sales happening in a store and what kind of item is getting solved without giving any context on that like who is buying, what he is using to buy, where he lives, how we can contact him. Until then, accessibility should be important. Everyone who is doing reporting, who is doing analysis on that data, who is maintain the data quality aspect of that, who is doing data quality validation of that, our managers, quality guys will be accessing the sales data. But as soon as the guest information or finance information like salary comes, those are very secure. Those are not exposed to the people who should not be seeing that. Finance guy can look at your salary but not your colleague. So, that information is very secure while the information related to sales, inventory should be easily accessible as a lot of people are looking at the data.	SD AS
43.	D	How about guest name, mail-ids, age-range and promotion.	
44.	S	Those are also secure data. Their mail ids, or number of males/females in the family, age, phone information, those are critical information. And the financial details like their buying pattern history that is also secure.	SE

45.	D	As far as guest is concerned what data is accessible, like their buying patterns?	
46.	S	I think if we can do a general buying pattern/segmentation and all without the uniquely identifying information about the guest that should be okay. But as soon as it goes to name, mail id, gender, age segmentation, it should be stopped.	SE AS
47.	D	Yeah things need to be separated that way because personal information about guest should not leak out.	
48.	S	Yeah, it's very secure and we do not maintain much. We maintain the surrogate key and we do not maintain the actual details. Guest name was initially there in some warehouse exposed to some people, now it is very much secured and very few people have exposure to that data.	SE
49.	D	Now you have covered accessibility and security as well. We know that data is backed up multiple times and a system crash would not cause any loss of data. But hypothetically assuming a situation of system crash and you are trying to retrieve the data. What data would you concentrate more in the subject area that you worked with.	
50.	S	Yes, we back up everything. The important matrices or important key performance indicators for which we need the data like sales I want the dollar amount or quantity purchased. In inventory, it would be stock. Out of stock items and in guest area it would be segmentation, lifestyle and quantity getting returned. In item area is all about different dimension of the item attributes that we maintain or different items that we have. Vendor information related who are our important vendors, what are things that they are selling to us, which warehouse it is landing to, which transportation and logistics has been provided for the warehouse to be distributed towards all the warehouses and stores. Those are the information/ important matrices and the key performance indicator which is being reported to, I will get that first rather than getting the legacy data of sales for 4 years/inventory for 4 years. It is important to get the matrices that is reported to higher level to make decision needs to be reverted fast.	REP REC IR
51.	D	So much information Suyash, Thank you!	

Interview 4 – Jyothsna Raj

Interview Date: 1st May 2017

Present: Jyothsna Raj (J), Athul and Dilip (D)

Duration: 45 minutes

Transcribed By: Athul

Line	Speaking	Text	Code
1.	D	Hi Jyotsna, thank you for the interview that you are giving us. I hope you have gone through the document that we had sent. As you know we are working on our thesis to understand data trade-offs between the characteristics in the retail domain and the characteristics that we have chosen are accuracy and timeliness, consistency and timeliness and accessibility and security. We have mentioned the basic definition of each of these in the questionnaire. The questionnaire has been divided into few sections - the first section is introduction, second is defining data quality, third is data characteristics trade-offs and fourth, fifth and sixth are about the trade-offs itself and finally your thoughts and suggestions. Let's start with the introduction, can you give a brief introduction about yourself and your work experience.	
2.	J	I have approximately 10 years of experience of which I have spent 7 years in retail, both e-com retail and brick and mortar retail. I have worked around data, data warehousing views, slicing and dicing data, understanding data reporting, dashboarding. Every path of my career has been around data - data quality, data analytics, reporting, and business data development.	
3.	A	How many years of work experience do you have in the retail domain and in the retail domain which areas you have expertise and which all companies have you worked for?	
4.	J	I have a little more than 6 years in the retail domain, 2 years in brick and mortar and 4 years in e-com retail. I have worked across huge inventories, sales data. Ecommerce data is simple daily data, it is not huge amount of data whereas brick and mortar data was very huge data.	REP REC
5.	D	The name of the organizations that you have worked for?	
6.	J	The two retail related organizations that I have worked for are Target and Fashionera. Target is brick and mortar retail which I worked around inventory, sales and inventory release. Ecommerce pretty much I did end to end. I was managing the data and the quality.	REP REC
7.	A	By Ecommerce you mean Fashionera, right?	
8.	J	Yes.	
9.	D	Now the 3rd question, can you give me an elaborated view of what role you play, the designation in both Target and Fashionera?	

10.	J	I will start with Target, there I was a senior data quality analyst and I was working in projects in inventory release and inventory. My role there was to ensure that the quality of data is prefect so used to run, write monitors, ensure the data, work with testing team to ensure that there test cases are fine. We used to also prepare reports. Basically, we used to sign-off to business if the data is usable or if it is not usable. We used to understand how the data flows from the foundation tables up to the mart tables and ensure that there is not loss of data or integrity loss of data and that every aspect of the data is good to go and can used by business. We make sure that business can rely on the data. So that was my role as a senior data quality analyst. In Fashionera I was head of QA and head of data quality. I had a team for QA and the data part i used to manage alone. So, I used to work with investors, give them the reports, also to different teams. We had logistics team, studio teams, marketing, sales; so, I used to get on the different reports, different matrices to see how the site is performing, site traffic conversion, sales, demographic, geographic location vs sales, all these data we used to segregate, slice and dice and give different reports to all these people. So that they could use the data to increase the sales or to change the campaign/techniques. Also, investors use the data to decide whether to invest on the company or not.	OD
11.	A	So, we now we will go to the next set of questions which define data quality. Could you explain what your understanding of data quality is?	
12.	J	According to me data quality is ensuring that the data is business ready and reliable. For me step number one would be to understand the data completely, only if you understand the data end to end you can understand what quality means for that data. Quality basically means that the data needs to be reliable and business ready and can be used further for anything. So, you ensure that there is no data lost, the data is clean, integrity between the tables are fine, and there is no murky data. You also ensure that when people are reading the data, they are reading correctly that is also the job of data quality analyst because all the data maybe right but people may be reading it wrongly, with the wrong joins or fetching data from the wrong table.	ODC
13.	D	So, you worked in 2 companies, how important is data quality in the organizations that you have worked for and in the domain.	
14.	J	Target used to consider data quality very seriously because it handles massive amount of data and any miss-information in inventory may lead to wrong forecast. So basically, it is out of control. Wrong inventory will lead to wrong forecast which will lead to wrong sale which will affect the sale if the company, its revenue. There is a huge problem if there is any problem with data quality. There are so many different fields and so many different aspects depending on this to ensure that the data is correct. Data quality was a very integral part and unless the data quality team give sign-off nothing used to go to production or would be given to business or the CEO. And start up was a totally different experience and e-commerce was also totally different. Because in e-commerce is not daily, it's hourly. It not more about quality, it	ODC

		needs to be fast because there are campaigns running and the change campaign within the hour if it is not working. You are competing with Amazon and other e- commerce chains so if it doesn't work in one hour you will lose lots of sales. When you analyse the sales from 12 to 2 and first 15 minutes you realize that something is wrong when you are looking at the sales data. Suppose you are selling t-shirts and you know that black t-shirt is being sold, fast, you know you need to increase the inventory that sales data should come to you in the next 5 minutes. It has to be quick, fast is the keyword in e-commerce.	
15.	A	The major factor in e-commerce you would say is timeliness of data?	
16.	J	Absolutely.	
17.	D	Next question is, what are the main characteristics that you would use to define data? So, we have taken accuracy, timeliness, accessibility, security and consistency, apart from these do you have any adjectives that define data characteristics.	
18.	J	These are the once that are important. The reliability of data as well. These are the key aspects. The other once would be complete.	
19.	D	Completeness and consistency is not the same but can be claimed to be the same. Few others like cost is also one, right? You can have all of these but the company should be able to afford it right?	CS
20.	J	Cost is and all is not very difficult nowadays but yes cost is a factor. Cost of storage, retrieval, and archive. I think cost of archival is the biggest issue in most organizations. Because there would be lot of data and you need a different server to archive.	
21.	D	Now we will go to the next question. How would you characterize trade-offs between different data characteristics in your field of work? Which are the most important trade-off that you have worked on in your field of work.	
22.	J	The biggest trade-off perhaps would be accuracy v/s timeliness. In a start-up cost was a huge issue. Because in a start-up what used to happen is, we used to have a database and we used to have slaves? Because large number of people would be continuously accessing the data. So why it is important for us to continuously read the data, there used to have multiple slaves due to which there would be delay also. While it used to reduce the load on different servers. So basically, we used to have slaves so that everybody can read the data without putting too much load. Slave is a constant application from the master, it didn't know what to prioritize. We needed the slaves so that the load could be reduced, but different slaves caused more replication which in turn increase the load and time to retrieve the data. So that is kind of a unique scenario which most start-ups and ecommerce companies face.	ACT

23.	A	Okay, so we will go to the next set of questions. According to you how important are accuracy and timeliness of data that you handle in your organizations?	
24.	J	Accuracy is absolutely important, maybe not immediately but at least at the end of the day when we pass on the number to the investors and other people. Because unless the data is absolutely accurate there is chance that you miss out something or you miss calculate. If you are running a campaign and you put 'X' amount of money for the marketing and for some reason the data is not accurate. We will end up spending so much money which will be a waste because we are not using the right target. If your data is inaccurate it can lead to huge amount of loss. It need not be immediate accurate but the end of day so that you can plan for the next day - campaigns, marketing, budgets. So basically, wherever money is involved, accuracy is absolutely important whereas for in some demographics like state vise, absolute is okay.	ACT
25.	D	So is money vs demographics, when money is involved, accuracy is important and when a lot of people are involved in it accuracy a little relaxed.	
26.	J	If you want to say how many men are shopping vs how many women are shopping and even if i say 60-40, 69-39 I am okay, approximately you know more men than women or more women than men. General number is okay.	AC
27.	A	Whereas where finance is involved, you want to know the correct amount.	
28.	J	Wherever money is involved, there is no change to that; I can't say 5 million today and say that it was 5.5 million tomorrow.	
29.	D	Now how important is timeliness of data?	
30.	J	In Fashionera, timeliness is critical because it is a fast-based environment. In a store, I cannot change the entire store even if I get the data until the end of the day or until I get people. Whereas in e-commerce I can change the entire sorting of that page to increase traffic and increase conversion. So, for that timeliness is important. If I see that people are logging into the first page and moving out immediate and that they are not interested, if I see that for 15 minutes. At the 16th minute I can reformat that page or change the layout of the page, sort it by prize or popularity. Whereas in a store if I see that people are coming and not interested, I just need that information by end of day before I start rearranging for the next day.	TL
31.	A	Timeliness is more important for the e-commerce than the brick and mortar side.	
32.	J	Yeah, in e-commerce if you're seeing that one item is selling fast, you can contact the seller and tell to increase the stock. But in store, if you don't have the product, you don't have the product you stock out.	

33.	D	So, the next question is what kind, variety and class of data would you classify as more important to be timely than accurate in nature and vice versa. You gave one example that finance/money related data should be more accurate than timely and how e-commerce data needs to be more timely than accurate. Any other data that is important in one or the other accuracy and timeliness?	
34.	J	Ad-related data. You will put on ad that there is sale and the ad is across different places. You don't care about the accuracy there you just need to know if the ad is working or not- conversion rate. You need to know immediately if the ad is working or you need to change that ad. This is not only about e-commerce. If you put a billboard and you know that after putting the board the sales is increasing, you don't need an accurate data that 10 people saw the billboard and 8 people came to store. You just need to know that I am able to see a conversion the moment I put the bill board. So there, timeliness is important.	ACT
35.	D	You need to know it in a timely manner so that you can remove or keep the ad. So, marketing is more on timely scale than on the accuracy level.	
36.	J	Correct.	
37.	A	Now about accuracy.	
38.	J	Any information that you send back to your client/customer. There is something called upselling information where they have bought something previously we say that your buying pattern looks like this so why don't you buy this. Here you need to be accurate. So okay even if it is a little late but when you are sending back any information to the client, it needs to be accurate. Any information like their buying pattern, items in the cart, password reset or any information that goes back to the customer needs to be accurate, it's okay if it goes a little late but it has to be 100% accurate. You should not send incorrect information to your customer.	AC ACT
39.	D	Question number 10 is regarding any context related to accuracy and timeliness. Is there any context where you felt that data needs to be timely and another where data needs to be accurate?	
40.	J	I will give an example. From the US perspective let's say Thanks Giving sales is going on. There timeliness is number one. Data has to come quickly; the CEO wants to know how the sales is going. So, it has to be timely, it can be approximate numbers but it has to be timely. You can say store ABC is doing better that store DEF or electronics is doing better than toys – it doesn't need to be exact but needs to be timely. At the same time for the very famous google organized sale or any online sales day at that time all the ecommerce sites are participating in the sale. There is also a US equivalent day online shopping day, the day after Thanksgiving, it is called Cyber Monday. That day similarly all ecommerce sites are competing to grab attention to ensure	ACT

		that they get the maximum sale. That time when you are competing or when quick sales is happening, you need information quickly in a timely manner. I don't need exact data, I need to know if the product is getting sold more than other products so I need to concentrate more on that.	
41.	A	No, we will go to the next one, consistency and timeliness. How important is consistency of data in retail/e commerce domain?	
42.	J	Consistency is important at a later point of time when you are building the dashboards or analytics point of view.	
43.	D	An information we got from one of the interviews is that of consistency of data is not there, you won't be able to pull reports.	
44.	J	Yes, in dashboard/reporting it is important. So, it is not an immediate thing, later when you look at the data it is essential to build the right report with the right kind of metrics. You have to decide what kind of matrices are important, what are the metrics that define your data to build those reports and slice and dice data consistency is extremely important. If it is not consistent you will not be able to pull data in the right Fashion or build the right reports.	CS
45.	A	What kind of data needs to be more consistent and what needs to be timelier according to you?	
46.	J	Mart data needs to be consistent, and foundation data can be timely.	
47.	D	So, you are talking on a technical level and not on the business level.	
48.	J	Yes, the reporting data has to be consistent whereas the base data that you are getting should be timely. Semantic/mart layer needs to be consistent.	
49.	D	On a business level, what data needs to more consistent that Timely and vice versa?	
50.	J	For example, forecasting data has to be consistent even if it is not timely. You are forecasting for the next year/month/quarter, the forecast has to be consistent with the previous forecast, it has to consistent with the sales, inventory. So, this kind of data needs to be consistent any data that you are using for prediction. Anything that needs to extrapolate, otherwise your extrapolation, projection, and forecast all these things will go wrong.	CS TL CST
51.	D	So, for futuristic data we need consistency and timeliness you have already mentioned. Is there any context that you have worked on where consistency and timeliness plays hand in hand? Or accuracy?	
52.	J	Usually we do forecast quarterly, but at one point in Target they were trying to do daily forecast in inventory project wherein you forecast today for tomorrow. Here it was important that it was timely, it was important that it was accurate and it was important that it was consistent. Because trying some-	CST

		thing called daily forecast.	
53.	A	Since it is such a short span of time you cannot give up on timeliness also.	
54.	J	Yes, because it is a short span it has to be timely, accurate and consistent.	
55.	A	What was the report's name, any idea?	
56.	J	It was called forecast report only.	
57.	D	Now we will go to next one which is accessibility and security. Accessibility and security varies a lot between retail brick and motor and ecommerce and you have worked in both, right? According to what is accessible data and what is secure data in both these areas?	
58.	J	Accessible is to ensure that data is available to the right people at the right time ensure that everything that a person needs to function is available at the right time in the right format for them to assimilate. Whereas security is key for any organization. Security is ensuring that you don't compromise on confidential data at any point. Security is to ensure that anything that needs to be confidential/protected should never be leaked out.	ACS SE
59.	D	So, what kind, variety or class of data would you consider to be accessible than secure and vice versa?	
60.	J	Any aggregate data needs to be more accessible, it is okay if it's not secure. Say the total sale in a region is so much and in another region, is so much such details have to be secure. But if I am giving out details at individual level like person name, e-mail that data have to be secure. Customer information, email, and address all these has to be secure. Retail companies, be it brick & mortar or online, they get customer information through various channels. Customer passwords are always encrypted and stored but even customer emails we give it to logistic companies. But even when we do that we give it in a pdf and not in an excel so that it cannot be exported and shared. The location of stay of the customer, what each customer is buying, all are confidential.	ACS
61.	A	But at a cumulative level like this many people bought this, that is fine?	
62.	J	That needs to be accessible. The merchants need to know that their product is selling and what kind of people are buying it. That is where demographics come into picture. So, they are saying like people from this region/age group are buying the product and so many has been sold. It okay to release that kind of numbers but you cannot say that this person bought this item on this day and he shifted to this address. That information should not be shared with anybody.	ACS
63.	D	That kind of information comes in what tables Jyotsna? For example, we know inventory, item and all those things. Does this come in sales? In sales is	

		it at individual level?	
64.	J	Yes. When you are giving invoice, people swipe their Target red card or whatever and you store that information. So, you know which customer bought what because they swipe their Target card and then we immediately know their details because we have that information.	
65.	A	So, information at individual level should be secure over accessible and what do you say about financial and HR information?	
66.	J	Financial and HR information without doubt has to be secure, password needs to be secure, buying patterns need to be secure, credit card information needs to be secure. They will be encrypted so that even if people know how to access the database they won't be able to access such information.	SE
67.	D	Even if they can access they still won't know how to understand the data.	
68.	J	Exactly. The key will be in one and the hashed key will be in another one. It's not even directly encrypted.	
69.	D	The last question is, is there any context/example where data needs to be more accessible than secure and vice versa?	
70.	J	There will be no scenario where security will be compromised. It is only when aggregate or when we give the details to investors because we trust the investors. We give them accessibility to data even if we compromise security.	ACS
71.	A	So, the stakeholders get accessible data?	
72.	J	Yes.	
73.	A	This is a hypothetical question. We know that there will be so many backups and the retrieval of data will be very simple. But under a hypothetical circumstance that a system crash has happened and you lose most of your data. What is the first data that you would retrieve and why?	
74.	J	Customer data because if we have customer data we can send them reminders and other stiff so that they will come and shop with us. If we lose the customer data, we don't know who our customers are. So, whom will we target, how will you tell them what is happening at the store, how will you send them mailers/campaigns, tell them there is sale in the store. For any retail customers are the key. So, moment you lose customer information we will not be able to target them and every company will pride on their customer information they have. There are lots of people who sell customer information from one company to another.	
75.	D	I have always have had a doubt with that. If individual information is very secure, how is that possible? Isn't that a breach of contract/trust?	

76.	J	It is but people find ways. Because we send mailers out and there are always external agencies that run email campaigns. Nobody has an inbuilt email campaigns. So, when you send these emails out some other organizations get the details and identify the customers.	SE
77.	A	So however secure you make the details, it is still not secure, is it?	
78.	J	There are chances. Because you won't build your own email sending tool. They won't know who this customer is or where they live, they just get their email ids. So that they can use it to campaign their companies.	SE
79.	D	So, you can snatch a customer from one retail to another?	
80.	J	Exactly.	
81.	A	Now to the last question, from your experience could you explain any other trade-offs that will be crucial for data quality management. These were the three that we felt important for retail. Is there anything else that you would like to add on as a trade-off which will be important in retail or ecommerce?	
82.	J	In some cases, completeness over accuracy.	
83.	D	That was suggested to me previously as well, completeness over accuracy. Lot of places we need complete data whereas accuracy can be a little lower.	
84.	J	Yes. Say for example I am sending out a mail which has these different columns. I need to have all the columns' information before I send the email. It is important that the data is complete, it is okay if a few fields are inaccurate but I need every single columns' data to come and the entire information needs to be there.	SE AC
85.	A	It must be mandatory?	
86.	J	Yes.	
87.	D	Thank you, it was quite helpful. Thank you.	

Interview 5 – Santhosh Meenhallimath

Interview Date: 2nd May 2017

Present: Santhosh Meehallimath (S), Athul (A) and Dilip (D)

Duration: 34 minutes

Transcribed By: Dilip

Line	Speaking	Text	Code
1.	D	Hi Santhosh, thank you for the interview that you are going to give us. As an introduction, can you give a brief introduction about yourself and your work experience?	
2.	S	I am Santhosh, currently working for Bluestem Brands which is an online e-commerce company. I have 8+ years in data quality and currently I am working as a data quality analyst and data engineer.	
3.	A	Can you elaborate about your previous experience in the retail domain?	
4.	S	Previous to Bluestem, I was working for Target which is a retail company in US. I have worked for closely 6 years in Target. They have stores and online sales. I have worked in different areas like sales, item, product management and little bit in matric.	
5.	D	Can you give me your total years of experience that you have in the retail industry?	
6.	S	Just in the retail I have 7+ years of data quality experience.	
7.	A	Can you elaborate what your role/designation was in both the companies?	
8.	S	Earlier in Target I was a software engineer/program engineer. There we used to get requirements from the project manager and my responsibility was to go through the requirement/ understand the requirement. We worked in both agile and waterfall model. Create user stories, get tasks, creating testing/data quality scenarios, where and all data quality can be applied, data profiling. Once that is all done, go through the requirements and the finding with the business team. Once everyone is agreed with that, the development starts. Start with data quality, check the functionalities. In case of any issue we communicate with the lead and get it resolved. Apart from the send out weekly updates and participate in the calls. Coming to Bluestem, I was a data lead/data quality lead. So, what I was doing in Target, now I assign the tasks to my repartees. I have to make sure that they are doing the data quality process and standards, creating the right data quality scenarios, doing data quality check on each and every process that's been developed, make sure all the reports are sent to the higher management and talk to the business on what we are doing and whether they are satisfied. 70% coding and 30% management.	REP REC
9.	D	So, what does Bluestem do?	
10.	S	It's an e-commerce company. We actually target the customers who have very low credit score, we give some credit so that they can buy an item from us and monthly pay the amount at very lesser prize. So, over the year of time, they can build the credit score too which helps them to get better credits or	REC

		some other products.	
11.	A	So, this is also on the retail line but on the website, right?	
12.	S	Yes, exactly. We have lot of branch and only couple of them have stores. Other than that, 99% is online.	
13.	D	So, you have worked both in retail and in e-commerce. Can you explain your understanding on data quality and how it has changed while working in Target and Bluestem?	
14.	S	Basically, data quality is a set of values or quantitative or qualitative variables that we go and test the data and make sure that the data that we give to business is business usable and they can make decision out of it. So, whether it is Bluestem or Target they obviously go with giving better data to the business so they can come with the decisions. The only difference is the type of data that we are dealing with. For example, Target, we had store and we used to get sales that happened at store as well as online. So, we had to deal with bluestem.com as well as Target stores. Whereas in Bluestem what happens is 99% of the sales that we get is online. From the data structure perspective that is different and it also applies to the business rules. The business rules that we apply to the .com and Target stores are different compared to the business rules that we apply in Bluestem.	REP REC
15.	D	How important is data quality in your organization?	
16.	S	It is very critical and important. The reason is, if the data that we are providing to the business is incorrect/not good and they make a decision out of it, then definitely you can say that the decision that they have made is not a right decision. So, data is the base for everything. The more the data the business analyst and data scientist have lot of leverage too. We look at the data and come up with good decisions. Not only the decision to see what happened previously but not predict what can happen so that they can have a good forecast.	
17.	A	Is there any variation between the data quality between e-commerce and retail?	
18.	S	The level of data quality in both is the same. The only thing as I said before is the business rules, the way we do it.	
19.	D	According to you what are the main characteristics that define data quality?	
20.	S	We call them as data quality dimensions. Some of them are accuracy, completeness, accessibility, integrity, consistency across the data sources and databases.	
21.	A	Next set of questions are on the trade-offs. According to you how important are accuracy and timeliness of data that you handle in the organization? And	

		if you could give some contextual example it would be helpful.	
22.	S	Definably there are trade-offs when we go to different dimensions. For example, what we do in online is, clicks data, i.e. whenever a customer logs in and he clicks in one URL and goes to the next URL that we call as a clicks data. We don't need it to be timely but we need it accurate, like where the customer navigated, which product he saw. For that what we do is, we don't do an hourly load but we do an end of day load and say this customer went to this site, he accessed through desktop or through iPad or whichever device and what he was looking for. That is the data that we give to the data scientist team, they have algorithms. They say, this particular customer is looking for this particular item. So that's when they call the call centre and say this customer is looking for this, can you try to get some marketing out of it. We need accuracy there. Some other cases timeliness is important. For example, during the holiday season like Thanksgiving and Christmas, we get lots of transactions every minute. There we need timeliness and every minute we need to make sure that the sales is loaded and keep track of that. Yes, we go for accuracy at dollar level but is product code or something is missing we don't care that. We do care but at that timeliness is important. If our website is down that has to be fixed immediately, that is also critical.	ACT
23.	D	So according to you finance/ things that deals with money, sales need to be accurate rather than timely and timely is when dealing with websites and click ratios, right?	
24.	S	Yes.	
25.	A	So, you have answered all the 3 questions. The next question that we would have asked is, what kind, variety or class of data that you would consider to be timelier than accurate and vice versa. If you could also give a context which you have come across.	
26.	S	Sales is something we need timely because every day we send out a report to executives on what was yesterday's sale compared to the forecast - are we doing good or bad or neutral. If we don't send that report by 8AM in the morning, we immediately get mails regarding why it was not sent, was there any issue. On the other hand, suppose we have customer data - who are our existing customer, how long have they been with us, who are our new customers- that is not something critical. Definitely it is not critical to be reported every minute. But they need to be accurate. We need to make sure we have the right information but it need not be updated every minute. One important thing which we note here is whether the person is diseased. We don't want to send a happy b'day mail or a greeting mail to a person who is diseased, it will be very offensive. So, we make sure it is accurate.	ACT
27.	D	The next set of trade-offs is consistency and timeliness, you can also talk about consistency and accuracy as well. How important is consistency of data in the e-commerce and retail domain that you are working for?	

28.	S	Consistency I will give an example. We have customer databases across. Some use it for marketing the catalogues, some use it for SMS. So, we have customer data at couple of places. They all have you be consistent. So, if we say that we don't want this customer or the customer is out of place, we don't want to send them any mails or anything without their consent. That is up-dated in one database and not in another, it will be an issue for us. So, it has to be consistent.	CST
29.	A	So, data across different database has to be consistent regarding voucher sending and stuff like that?	
30.	S	Marketing or sending the alerts, in that perspective it has be consistent. And even the credit card information and their personal information should also be consistent. We have security over there to ensure that such data is not repli-cated in multiple places but wherever it is, it has to be consistent. The second thing is consistency between source and target. For example, wherever the actual sales happen, they have the sales number. That sales information they send it to us and we load the data finally to data warehouse and create micro strategy report out of it. The sales dollar amount that reflected in the report and the source file should be marginally close. There might be some differ-ence due to timing but it should be marginally close.	CS
31.	A	Is there any inconsistency in data that have had affected the reports or data quality?	
32.	S	Whenever marketing team sends out a mail or catalogue, they always want to track that because of their marketing how much sales the company got. So, we have something called sending the catalogue or mailing the catalogue or somewhere referred that, we all them as communications. So how did we communicate and how was the sales? So, what actually happened is, we were not getting proper communication codes from the vendors. So, most of the dollar amounts that we were getting was send through walkthrough means they just browsed and they brought it. So that was affecting the marketing people because they were doing their work but not getting the credit. When we started looking at it we got to know that the source from the vendors, they were not sending the proper code- communication codes. So, once we fixed it, we could see that the dollar amount showing in communication catalog and mails increased showing that these people got the mails/catalogs before they brought the item. That actually gave positive result on the marketing team.	CS
33.	D	It changed from inconsistent to consistent data?	
34.	S	Yes, we changed it and could see the positive results.	
35.	A	So, is there any example regarding data that needs to be more consistent over timeliness or vice versa?	
36.	S	I might have to think about it.	

37.	D	Now we will go to accessibility and security. What sort of data according to you is accessible and what needs to be secure? Any context that you have in picture.	
38.	S	Any data in the database should be accessible and definitely secure from the outside world but it is at different levels. For example, the technical team will never have access to the HR database which is related to the employees, their medical benefits or any of their credit card details. That is completely secure. Yes, we develop it but once we develop it and deploy it, we won't be able to see the actual production data. We work with the test data but we never get to see the actual production data. So that is secure. Is it accessible, yes only to certain level of people and not to everyone? And they also can't see at database level, can only see the reports. Some of the cases like, we have different brands, as a technical person i can go and see the data. But one brand cannot see the other brands data. We keep a strict line there. Each brand should access only their data.	ACS
39.	A	So, as a technical person you can see all the brands, right?	
40.	S	Yes, we have access because if there is any issue we can work on it. But we cannot share it to the other brands. We maintain the ethics and they cannot ask someone else's information.	
41.	D	How about other information like financial data and guest related data?	
42.	S	We don't have guest concept, we call everyone as customer. Regarding financial, it is same as HR. Financial information some of the case we do have access, for example sales since we do have to generate report. We have a very high level access there but not at company, executive level. For example, how some brand is doing, are they doing bad - that we don't have access to so we don't send a wrong message to the outside world how the company is doing.	FD CD
43.	A	And customer information, what are the things that are accessible and what is secure.	
44.	S	At customer side, I can have access to first name, last name, address, and phone numbers. In some cases, I can access the DOB - only the data and month and not the year - that is only for the leads. I can do it, my reporters can't, and we don't give them access in production data. I cannot go and access their SSN, it is completely restricted, even their credit card numbers.	AS SE
45.	D	How about seeing the buying patterns?	
46.	S	Yes, that is click data. We do check that, not at customer level but in a pattern like most customers buy this item.	
47.	A	But that is not at customer/individual level, right?	
48.	S	We do have that level of access, but we look at the pattern. It's very rare that	

		we look at customer level.	
49.	D	It doesn't make much sense to get at customer level and check, right?	
50.	S	Yes, we have like millions of customers, unless someone has hacked someone's account and we see a change in the regular pattern. At that time only we go at customer level details.	
51.	A	Have you faced any such situations?	
52.	S	We have faced one situation, but it was not hacking. We have person who manually give discounts. We saw that one person gave $787,000 which is not appropriate. So that was a mistake it should have been $78 and he types as $787,000.	
53.	A	So that was a mistake from the testing side?	
54.	S	Not testing side, the person who enters discounts.	
55.	A	At the front end?	
56.	S	Yes.	
57.	D	The last but one question is about system crash. I know every company now has a lot of backups. But in case of a system crash, what is the first data that you would want to retrieve?	
58.	S	The first data we would want to retrieve is financial and customer personal information. That is something we want secure, we won't keep multiple backups but we want to secure that.	FD
59.	D	So, the first information that you would take is the customer information and financial information.	
60.	S	Customer information includes everything, including their credit card information.	
61.	A	Ideally that must be most secure also?	
62.	S	Yes, and we do not have access to take backups. There is a completely different tea, the data governance team and they take the backups. So, which one should be backed up we offer and we do have audits on data that is backed up.	
63.	D	The last question is, apart from the three trade-offs that we have considered, what would you consider as other trade-offs?	
64.	S	Some cases integrity. For example, we get that this customer bought this item and when we look for the customer he will not be in our database. So that's an integrity issue. But we keep the sales record since it is important.	

65.	A	Santhosh, thank you for the interview.	

Interview 6 – Beth Benzie

Interview Date: 4th May 2017

Present: Beth (B) and Dilip (D)

Duration: 47 minutes

Transcribed By: Athul

Line	Speaking	Text	Code
1.	D	Hi Beth, thank you for the interview. Could you start with a brief introduction about yourself and your work experience?	
2.	B	I have been working in data quality for 7 years. But prior to that I was in the business side, I have worked very closely with IT and I really understand what data quality means from a business stand point. I worked at Target for 4.5 years and then I was at Bluestem which is an online company and now I am working at Supervalu where they have both the wholesale and retail side with limited online presence.	OD
3.	D	How many years of experience do you have in retail?	
4.	B	In retail, probably about 20.	
5.	D	In the technical side or with the business?	
6.	B	Most of my experience is in the business side and in technical side it will be around 7 years.	
7.	D	Can you elaborate the roles that you have played in the companies that you have worked for?	
8.	B	In Target, I started as a lead data quality analyst and then got promoted, so when I left I was a data solutions consultant. Bluestem I managed and started up a data quality team on the BI side and at Supervalu I started as a Principle BSA – Business System Analyst. Now I am transitioning into a data quality manager and I am setting up a data quality team which they don't have and they desperately need it.	OD
9.	D	Can you elaborate on the role that you are doing as a data quality manager at	

		Supervalu?	
10	B	I am just building it. We are at the transition phase so I am doing some analysis around all the information and all the tables that they have. They have an old environment and a new environment, basically since quality was not something that they were focused on, they have lot of issue. So, I am trying to find out what data is out there, where we have issue and trying to get it as much fixed as possible. So, from data quality stand point now, I am trying to put the building blocks in place now. I am studying the set up and monitor process so that we can get back going and get the team up and running and understanding well on the technical side, the data quality dimensions and how to analyse tables, how to analyse how things go together across different business/subject areas and across different tables. I started the same thing in Bluestem, so I have the framework in place which I am just using now.	REP REC OD
11	D	What the duration that you worked in Supervalu	
12	B	9 months now	
13	D	Okay, could you explain your understanding of data quality?	
14	B	For me, it is making sure that data is ready for business consumption. Ready and accurate, it accurately reflects what happens from a business stand point for the business consumption.	
15	D	You have worked across retail and e-commerce in 3 companies now. In your opinion, how important is data quality in the organization that you have worked for and how has it been different across different organizations	
16	B	It is very apparent that if somebody isn't watching it, data quality can become very poor. Generally, there is a difference between QA and data quality. Most of the companies that I have been in have either just QA or Supervalu didn't have it either. So, at Supervalu it is extremely important and that is the first thing I started doing. I started finding many financial issue and many issue across all the tables. I think I have been able to portray the impact and the issue out there and how it is impacting their business. Because of that they are well aware and they have decided it is very important. So, they key here is not all businesses understand how important data quality is until somebody points out where it is causing a problem and how it is impacting the reporting and the business decisions.	OD
17	D	What are the characteristics that define data or data quality according to you?	
18	B	Accuracy is kind of a question, I will tell you. I went to a data quality conference and they had many debates about whether accuracy was a true characteristic or not. At a BI side accuracy means if it is accurate from source side but you don't really know if a source is accurate. To me accuracy is critical as long as it needs to accurately reflect what has been provided. Timeliness is	AC CS

		also critical. Consistency is also critical. Completeness, you need to make sure you have everything. Validity you have got to make sure you can tie across where needed. Those are probably the keys for me. I look at trending and duplicability as the first two I look two. Those are the two I consider most important and those are the easiest to check.	TL
19	D	Trending and duplicates come under consistency? What do you say?	
20	B	Kind of, it depends on how you look at consistency. To me consistency is all things being consistent across all the tables. So, you could put them under consistency I would hold them out separately. Duplicates in particular are the most important because if you have those you know there is some problem.	CS
21	D	Now the question is about data characteristics trade-offs. Do you have any example of data characteristics trade-offs in your work?	
22	B	I would call them balancing rather than trade-offs. I usually look at the business impacts to determine where you can find the best balance. So, if I know there is a problem and I know if I try to fix that is going to negatively impact something else in the business, I would look at the impact of both and figure out which one is the least impactful or try to meet them in middle so that is not impacting as much. So, for example, you have timeliness and accuracy. Accuracy is really important to make business decisions. If you don't have complete sales, business will be speaking about sales thinking that it is complete and decisions will be taken based on that. When you run into that situation, you really cannot trade-off completeness over timeliness or accuracy because if you do, you are misinterpreting the information. The only way you can is when you are telling the business that you have not fully loaded everything yet and it is delayed. So, it is key to make sure you are carefully monitoring those, since they are business critical. And if there are any delays you need to actively work on it and update it with the business. So basically, those are the keys to me, making sure you are balancing it. Some of them I won't really think what the trade-off is I am going to take actions over timeliness. In either case where it is not as critical, where the business is not waiting, I will definitely wait for timeliness, 1 week for everything to complete. Sales is always a critical one because you have got the dashboards, every company wants to know the sales of the previous day. That is the first and foremost in their mind and what I found is that the businesses are okay with it being delayed as long as you communicate it. So, it's all about how you manage the expectation from the business stand point as you are working through the problems. Usually you don't have to do the trade-off as much, you basically just wait for the accuracy.	ACT
23	D	So, the business users want sales to be accurate, it is okay to be a little delayed.	
24	B	Yes, completeness and accuracy go hand in hand. Accuracy is making sure	AC

		everything is accurate and completeness is, did you get everything. That's where timeliness and completeness kind of contradict at one another. Because you may in the middle of the load, if you don't have a complete load at the right time, you need to balance that with the business. Business needs to know you are actively working on it, you need to set that expectation. When I was working in Cristopher & Banks and I started a data warehouse there. We had lots of delays as a start-up because we were bringing sales for the first time and again it was all about managing the expectation with the business. From a technical stand point, there will be some give and take that you need to do. Maybe you need to start earlier on the day to monitor this and if there is a problem making sure that we are going to get this resolved as quickly as possible.	TL
25	D	So, one case wherein the sales data needs to be both accurate and timely, is it?	
26	B	Sales data always needs to be accurate, if not your company will be making wrong decisions out of it. Sales is critical to be accurate and if there is an issue it needs to be addressed immediately.	SD
27	D	You have spoken about that needs to be accurate but can be relaxed in timeliness. Is there any data that needs to be timely but accuracy can be tolerated to an extend?	
28	B	Accuracy is always going to win over timeliness. Simply because if you say that you are data can be of 5 or 10 percent of errors, business better not be using it for business decisions. So, if the business is not using it for business decisions, then I don't know why I have the data in the data warehouse. If there is some flow that comes in and basically says it is not business critical information. Its information that the business wants to know what the status is at 10' o clock in the morning. You can somehow say, at 10'o clock in the morning we have 5 million dollars and we are about 80% done with the flow. We can do some maths to figure out that the rest 20%. You can do things like that as long as the business understands that it is not complete and not fully validated. As long as they are okay with that. It is going to break when, with the 20% they say it would be around 6 billion dollars and it actually comes around only 5.1 billion, then it is going to be a problem. You guesstimated but for some reason it was not right. Timeliness usually means they are making decisions at that time. And if they are making them at that ti, you have got to make sure that you have done those estimates and they are inline and they are okay with it. So, the key of all of it is managing the business expectations. They are the once who tell what is timely, what they need, how accurate they need so you can work based on those guidelines.	ACT AC TL
29	D	So, wherever business decisions are made accuracy is more important than timeliness. But one question I have is, how much of a delay can you give the	

		business users?	
30	B	If it is sales you are not allowed have a big window. For example, in Cristopher & Banks we had a special discount day and sales literally up and running by 5 in the morning. It was a weekend and the sales literally ran or 12 hours past because there we lots. In this case, they can tolerate a delay because there is a lot of sales coming in, there is a higher volume than normal. But if there isn't a lot of sales and it is a technical problem, they are not necessarily going to be okay with it. If it's a critical flow the delay has to be top prioritized and figured out as quickly as possible. It is not as critical and it is not affecting any other flows or any other fields that are critical, you can probably let them know, they are usually going to be okay with it.	AC TL
31	D	You have covered the entire set of questions in accuracy and timeliness. Now the next set is on consistency and timeliness. When you want the data to be completely consistent you have to give up on timeliness. Do you have any such contexts/examples/data?	
32	B	Anything with timeliness is kind of the same. If you don't have consistency especially if you are considering trending and duplicates under consistency or if there are problems with the data, the business is again going to make business critical information and hence timeliness is going to be secondary if there is issue in the data. The key is again communicating it with the business as quickly as possible and getting it fixed as quickly as possible. Duplicates from a business stand point, say duplicates in sales they might think that there were great sales and it was just a misinterpretation. So, in the end, the data that you are providing is for business consumption. So, in order for them you have to be aligned with them if there is any issue in the data or consistency, it has to be completed. So, depending on norms you can decide how to balance the monitoring and the fixing of any data flows that have issues. Timeliness for me is somewhat secondary for critical flows.	TL CS CST
33	D	How about consistency and accuracy, your thoughts on that? For example, New Jersey can be given as either NJ or as New Jersey, this is accurate data but not consistent.	
34	B	That is data quality on the source side and it is working with the business team and saying there is too much flexibility in how people can put the information. So, you have drop downs instead of free text. You need to work with the source side to get them to understand the data quality. They are not consistent but they are both accurate. If I look from a business perspective then I would work with the source side, but once the business view this, they will be behind you and getting any other changes made. Data quality to me is you are 100% with the business, 100% of the time and that really is the key. So, in this scenario, source system is allowing to do that, so if you put control on the source side you would be able to bring things in. When there are many sources and there are inconsistencies within them and if we try to combine	CS AC ODT

		them, consistency will become worse. So, in that scenario our resolution would be to create a master out of it, a master data and we will bring together and we will clean up. We will send it back to source but we will also do some cleaning up. Consistency takes time to be cleaned up. At accuracy stand point, you are fine. And always accuracy over consistency, if you have duplicates you need to clean I up.	
35	D	Communicating with the people in source system about issues that you are facing in your data warehouse, you can bring up consistency as high as accuracy as well, over the course of time.	
36	B	Over time, yes.	
37	D	The last set of trade-offs is accessibility and security. What data according to you need to be accessible and what data needs to be secure?	
38	B	Anything that has personal information type data has to be secure and however, if needs to access, then have to prove the identity. So, accessibility I really is not a factor there. To me a highly secure data set shouldn't be accessible. If you have highly secure data, you don't want it to be accessible to people. So, if you are talking about accessibility at function stand point, it is different but if you are talking about who has access and who does not have access at a secure stand point, it is completely different. Security rules are put in there for a reason, you know that everybody should not access at a legal stand point. So, secure is always going to win because of legal ramifications.	SE AC ACS
39	D	What kind of data needs to be accessible across the company?	
40	B	I don't think if there is anything that everybody needs access to. Because if they do not use it, they don't need access. But to me accessibility is just around what is your use for in your job and then from there determine who should or should not have access. All areas will have some level of legality around them that not everybody should have access.	AC
41	D	What about security in shopping trends?	
42	B	Sales is kind of need to know basis. If not PII data its company data, if you share what the company is doing with somebody else. That is a security risk as well. So, sales is another one that not everybody should have access to. If people have access to something that they should not have access to, sales for example, how sales is doing in their company because they have access to it. They can actually give information to other persons, like insider trading. So, accessibility for me is always going to go away over security.	SD ACS
43	D	One last question, assuming that a system crash happens what is the first data that you would want back in your system?	

44	B	It is always going to be how is the company doing, so the sales probably. From the company perspective, you need to get whatever is in front of the customer up and running as fast as possible – websites or whatever it is. You have to look at how it is impacting your business.	
45	D	So, sales and company information?	
46	B	It depends, data quality really runs from source to BI. You can manage it at source side or BI side. But a system crash happens in the company and you are strictly looking from the BI stand point, then your information of how you are doing from a business stand point is what you have to retrieve first, which will be sales. If you are looking from a company stand point then the first what you have to do is what is impacting your business-like websites. Because if those are gone, you are not generating any sales.	OD
47	D	I think we covered everything. Beth, thank you; it was very helpful.	

References

Abbasi, A., Sarker, S. and Chiang, R. H., 2016. Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2), 3.

Aguinis, H. and Henle, C.A., 2002. Ethics in research. Handbook of research methods in industrial and organizational psychology, pp.34-56.

Albala, M., 2011. [online] cognizant.com. Available at: https://www.cognizant.com/InsightsWhitepapers/Making-Sense-of-Big-Data-in-the-Petabyte-Age.pdf [Accessed 4 Apr. 2017].

Aloysius, J.A., Hoehle, H., Goodarzi, S. and Venkatesh, V., 2016. Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. Annals of Operations Research, pp.1-27.

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G. and Schwarzenbach, J., 2013. The six primary dimensions for data quality assessment. Technical report, DAMA UK Working Group.

Asq., 2017. Quality Glossary | ASQ. [online] Available at: https://asq.org/quality-resources/quality-glossary [Accessed 20 Mar. 2017].

Ballou, D.P. and Pazer, H.L., 1995. Designing information systems to optimize the accuracy-timeliness trade-off. Information Systems Research, 6(1), pp.51-72.

Baškarada, S. and Koronios, A., 2014. A critical success factor framework for information quality management. Information Systems Management, 31(4), 276-295.

Batini, C. and Scannapieco, M., 2016. Data and Information Quality: Dimensions, Principles and Techniques. Springer.

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 16.

Batra, M. and Arora, A., 2016. Implementing Data Security in Cloud Computing. International Journal, 4(6).

Bhattacherjee, A., 2012. Social science research: principles, methods, and practices.

Blumberg, R. and Atre, S., 2003. The problem with unstructured data. Dm Review, 13(42-49), 62.

Braz, C., Seffah, A. and M'Raihi, D., 2007. Designing a trade-off between usability and security: a metrics based-model. Human-Computer Interaction–INTERACT 2007, pp.114-126.

Bulger, M., Taylor, G. and Schroeder, R., 2014. Data-driven business models: challenges and opportunities of big data.

Cai, L. and Zhu, Y., 2015. The challenges of data quality and data quality assessment in the big data era. Data Science Journal, 14.

Cai, S., 2015. Modeling real-time transactions in uppaal. Tech. Rep., April 2015. [Online]. Available: http://www. es. mdh. se/publications/3911.

Cappiello, C., Francalanci, C. and Pernici, B., 2004. Data quality assessment from the user's perspective. Paper presented at the Proceedings of the 2004 international workshop on Information quality in information systems.

Chapman, A. D., 2005. Principles of data quality: GBIF.

Chaudhuri, S., Dayal, U. and Narasayya, V., 2011. An overview of business intelligence technology. Communications of the ACM, 54(8), pp.88-98.

Chen, H., Chiang, R. H. and Storey, V. C., 2012. Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

Dabholkar, P. A., Thorpe, D. I., and Rentz, J. O., 1995. A measure of service quality for retail stores: scale development and validation. Journal of the Academy of marketing Science, 24(1), 3-16.

Daniel, F., Casati, F., Palpanas, T., Chayka, O. and Cappiello, C., 2008. Enabling Better Decisions through Quality-Aware Reports in Business Intelligence Applications. In ICIQ (pp. 310-324).

Date, C.J., 2006. An introduction to database systems. Pearson Education India.

Dedeke, A., 2000, October. A Conceptual Framework for Developing Quality Measures for Information Systems. In IQ (pp. 126-128).

Eckerson, W.W., 2002. Data quality and the bottom line: Achieving business success through a commitment to high quality data. The Data Warehousing Institute, pp.1-36.

El-Sappagh, S.H.A., Hendawi, A.M.A. and El Bastawissy, A.H., 2011. A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), pp.91-104.

Eppler, M.J., 2006. Managing information quality: Increasing the value of information in knowledge-intensive products and processes. Springer Science & Business Media.

Fan, W., Geerts, F. and Wijsen, J., 2012. Determining the currency of data. ACM Transactions on Database Systems (TODS), 37(4), p.25.

Fraenkel, J.R., Wallen, N.E. and Hyun, H.H., 1993. How to design and evaluate research in education (Vol. 7). New York: McGraw-Hill.

Friedman, T. and Smith, M., 2011. Measuring the business value of data quality. Gartner, Stamford, 464.

Geiger, J.G., 2004. Data Quality Management, The Most Critical Initiative You Can Implement. Data Warehousing, Management and Quality, Paper, pp.098-29.

Graneheim, U.H. and Lundman, B., 2004. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse education today, 24(2), pp.105-112.

Gregory, I., 2003. Ethics in research. A&C Black.

Griffith, T. L., Northcraft, G. B. and Fuller, M. A., 2008. Borgs in the org? Organizational decision making and technology The Oxford handbook of organizational decision making.

Heinrich, B. and Klier, M., 2011. Assessing data currency—a probabilistic approach. Journal of Information Science, 37(1), pp.86-100.

Herath, T. and Rao, H.R., 2009. Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. Decision Support Systems, 47(2), pp.154-165.

Kaisler, S., Armour, F., Espinosa, J. A. and Money, W., 2013. Big data: Issues and challenges moving forward. Paper presented at the System sciences (HICSS), 2013 46th Hawaii international conference on.

Kaufman, L.M., 2009. Data security in the world of cloud computing. IEEE Security & Privacy, 7(4).

Kvale, S., 2006. Dominance through interviews and dialogues. Qualitative inquiry, 12(3), pp.480-500.

Labrinidis, A. and Jagadish, H. V., 2012. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033.

Laitio, J., 2011. Semantic Web Data Quality Control (Doctoral dissertation, Aalto University).

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y., 2002. AIMQ: a methodology for information quality assessment. Information & management, 40(2), pp.133-146.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity.

Mondayreport.ca., 2017. Monday Report On Retailers. [online] Available at: http://www.mondayreport.ca/mondayreport/monreport.cfm [Accessed 18 Apr. 2017].

Myers, M.D. and Newman, M., 2007. The qualitative interview in IS research: Examining the craft. Information and organization, 17(1), pp.2-26.

Nelson, R.R., Todd, P.A. and Wixom, B.H., 2005. Antecedents of information and system quality: an empirical examination within the context of data warehousing. Journal of management information systems, 21(4), pp.199-235.

Otto, J.R. and Chung, Q.B., 2000. A framework for cyber-enhanced retailing: Integrating e-commerce retailing with brick-and-mortar retailing. Electronic Markets, 10(3), pp.185-191.

Pierce, E.M., 2003. Pursuing a Career in Information Quality: the Job of the Data Quality Analyst. In IQ (pp. 157-165).

Pipino, L.L., Lee, Y.W. and Wang, R.Y., 2002. Data quality assessment. Communications of the ACM, 45(4), pp.211-218.

Recker, J., 2013. Scientific research in information systems: a beginner's guide: Springer Science & Business Media.

Schultze, U. and Avital, M., 2011. Designing interviews to generate rich data for information systems research. Information and Organization, 21(1), pp.1-16.

Seidman, I., 2013. Interviewing as qualitative research: A guide for researchers in education and the social sciences. Teachers college press.

Strong, D.M., Lee, Y.W. and Wang, R.Y., 1997. Data quality in context. Communications of the ACM, 40(5), pp.103-110.

Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C., 2007. A framework for information quality assessment. Journal of the Association for Information Science and Technology, 58(12), pp.1720-1733.

The Interaction Design Foundation., 2017. Information Overload, Why it Matters and How to Combat It. [online] Available at: https://www.interaction-design.org/literature/article/information-overload-why-it-matters-and-how-to-combat-it [Accessed 5 Apr. 2017].

Thota, C., Manogaran, G., Lopez, D. and Vijayakumar, V., 2017. Big Data Security Framework for Distributed Cloud Data Centers. In Cybersecurity Breaches and Issues Surrounding Online Threat Protection (pp. 288-310). IGI Global.

Wand, Y. and Wang, R.Y., 1996. Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), pp.86-95.

Wang, R. Y. and Strong, D. M., 1996. Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.

Zahedi Nooghabi, M. and Fathian Dastgerdi, A., 2016. Proposed metrics for data accessibility in the context of linked open data. Program, 50(2), pp.184-194.