Big Data Quality Challenges in the Context of Business Analytics

Mirva Toivonen

Master's Thesis
Helsinki 10.5.2015

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion –Faculty/Section	Laitos – Institution – Department
Faculty of Science	Department of Computer Science

Tekijä– Författare – Author

Mirva Toivonen

Työn nimi – Arbetets titel – Title

Big Data Quality Challenges in the Context of Business Analytics

Oppiaine – Läroämne – Subject

Tietojenkäsittelytiede

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's Thesis	10.5.2014	57 pages

Tiivistelmä – Referat – Abstract

Big data creates variety of business possibilities and helps to gain competitive advantage through predictions, optimization and adaptability. Impact of errors or inconsistencies across the different sources, from where the data is originated and how frequently data is acquired is not considered in much of the big data analysis.

This thesis examines big data quality challenges in the context of business analytics. The intent of the thesis is to improve the knowledge of big data quality issues and testing big data

Most of the quality challenges are related to understanding the data, coping with messy source data and interpreting analytical results. Producing analytics requires subjective decisions along the analysis pipeline and analytical results may not lead to objective truth. Errors in big data are not corrected like in traditional data, instead the focus of testing is moved towards process oriented validation.

ACM Computing Classification System (CCS):
Information systems → Information systems applications → Decision support systems
Information systems → Data management systems

Avainsanat – Nyckelord – Keywords

big data, business analytics, data quality, data quality control

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Contents

1 Introduction

Organizations can improve efficiency, growth and competitive advantage with business analytics. A clear positive correlation between analytics and business success has been found [LaValle et al 11]. Increasing variety of data types and data sources, volume and velocity of the data help organizations to gain more information and to make more informed decisions. Big data complements traditional analytics like reports and dashboards and helps to gain competitive advantage through predictions, optimization and adaptability. However, managing data quality is becoming more challenging as data variety and the number of data sources increase. Big data increases the amount of data (volume), speed of data in and out (velocity) and range of data types and sources (variety). The poor data quality is a growing problem. Impact of errors or inconsistencies across the different sources, from where the data has originated and how frequently data is acquired is not considered in much of the big data analysis [Loshin 13].

This thesis is a literature review of big data quality challenges in the context of business analytics. Data quality difficulties are approached mainly from information quality perspective, because the purpose of big data is in decision making and gaining information through analysis of data. The intent of the thesis is to improve the knowledge of big data quality issues. This thesis describes quality attributes and quality challenges through five basic data warehouse processes which are similar to analysis pipeline processes: design and administration, software implementation and/or evaluation, data loading, data usage and data quality.

Most of the quality challenges are related to understanding the data, coping with messy source data and interpreting analytical results. The data quality focus is moved from correcting the data towards process oriented validation. Understanding the properties of the data sets requires knowledge of the lineage of the data. The person who analyses the data set should be aware of where the data set is collected, when it is collected, how it is prepared and what are the limitations of the datasets. Selection bias issues and drawing inaccurate conclusions from data are challenges. Big data analytics has realistic focus on processing the data. It is understood that big data is messy and analytical methods try to cope with the messiness. The processing of big data is resource intensive so the data

is usually not corrected in the same way than data in data warehouses. It is the data consumer's responsibility to understand data quality issues and to decide if the data is good enough for the analysis.

The rest of the thesis is structured as follows. Section two presents differences between big and small data, and briefly represents the characteristics of big data analytics. Section three describes data quality attributes through data warehouse related processes. Hadoop based big bata architecture is described in section four. Data quality governance related issues are described in section five. Section six describes big data testing related issues. Section seven contains the conclusions of the thesis.

2 Differences Between Big Data and Small Data

This section introduces the concept of big data and big data analytics. Subsection 2.1 defines the concept of big data through volume, variety and velocity. The downsides and benefits of volume, variety and velocity are briefly introduced from data quality perspective. The most used big data sources and data structures are introduced as well the mechanisms that bring big data into existence. Subsection 2.2 compares the differences between small and big data and divides the differences into 11 categories: data sets, location of data, data structure and content, longevity, measurements, data preparation, reproducibility, data access, project costs, introspection and data processing. Subsection 2.3 helps to understand the characteristics of big data analytics and presents some big data use case examples.

2.1 Big Data

The data used in business analytics can be small or big. Term small data is a synonym for traditional data. Traditional data is defined as electronic data that is stored in databases, data warehouses or legacy systems [Batini et Scannapieco 06]. This thesis follows the definition of data used by Tien: the definition of data includes digital data measurements, raw digital values, processed digital values and metavalues [Tien 13].

Big data is electronic data for which management challenges have exploded in three dimensions: volume, velocity and variety [Laney 01]. When big data term is used it can mean that the amount of data is large (volume) or speed of data in and out is fast

(velocity) or the range of data types and sources is wide (variety). The term big data can mean all three of them. Big data literature has adopted Laney's volume, velocity and variety dimensions into the definition of big data [Russom 11]. The volume, variety and velocity of data make more precise analytics possible. However, each big data dimension has its own data processing and data quality difficulties.

Volume: An example of volume is multi-sensory data that can obtain terabytes of data from video surveillance cameras [Tien 13]. Today a data set is considered big if it contains a few terabytes to many petabytes of data, however as the software tools become more powerful, the definition of big is shifting [Tien 13]. Volume is a relative concept because technical improvements in handling large amount of data make today's big data smaller. One thing that makes big data voluminous is that big data cannot be processed with traditional computing methods. Processing volumes of data needs improved scalability and performance. It may be time consuming to apply data quality activities for large amounts of data because even simple operations can lead to major delays in runtime and responsiveness when data volume increases [Parker 12]. Large amount of data makes it more difficult to find relevant and meaningful information.

Variety: Big data comes in many formats and from many new sources like smart sensors in mobile devices, web applications (clickstream behavior), pictures, audio, video, blogs, news, microblogs (Tweets) and social media. The wide spectrum of possible data representations is distinguished into three types of data: structured, semi-structured and unstructured data [Batini et Scannapieco 06]. Structured data elements are associated with fixed structure, for example relational tables. Structured data can be machine-generated like sensor and web log data or human-input generated data like input and click-stream data. Semistructured data elements have some flexibility associated with the fixed structure, for example XML-documents where same kind of data may be presented in multiple ways. Unstructured data is expressed in natural language with no specific structure or domain defined. Unstructured data can be the content of wikis, blogs, power point representations, e-mails, word documents and PDFs. Unstructured human-generated data is naturally messy, it represents real data in day-to-day life. For example, e-mails are indented to deliver information between people, not to be analyzed by computers. Unstructured data is filled with nuances, variation, and double meanings. This make finding relevant and meaningful information

difficult.

Velocity: Data velocity means that data comes in with continuous stream. The data needs to be used immediately as it flows in to the system [Laney01]. Real-time big data comes from sources such as positioning data (e.g. GPS -data, Global Positioning System) and sensors like motion- or picture sensors [Tien 13]. The speed of the process from data acquisition to decision-making is increasing. The speed of data enables faster reaction time in business. That may mean more fine-grained customer segmentation based on day-to-day situation rather than segmentation based on historical data.

Big data is created digitally and collected automatically. Data is produced passively as a product of our daily lives or interaction with digital services [Letouzé 12]. The data can include temporal traceability like call duration or geographical data from mobile phone location data. Big data can be found through different mechanisms [Berman 13]:

1. The data is already collected in the course of normal activities and is waiting to be used. The data owner does not want to discover or to do anything new, but to do better what it has always been doing.
2. The data is already collected but new activities are supported by the data.
3. A business model is planned based on a big data resource. An example of this mechanism is data intensive services like Amazon.
4. A group of entities that have large data resources federate their data resources, for example hospital databases.
5. Large amounts of data are collected and organized to benefit an organization and their user-clients. These projects require skills and vision.
6. Big data resources are built from scratch. No data and no big data technologies exist before big data project.

Structured operational data, human generated documents and transactional data are the three most used data sources in big data projects [Devlin et al. 12]. The following data sources are being used or planned for use in big data projects: 50% of cases mentioned structured operational data (e.g. point of sale, customer care, supply chain), 40% of cases mentioned human generated documents (e.g. email, application form documents), 33% mentioned deep operational transaction (e.g. audit log information or network probe), 32% mentioned image content (pictures, video), 31% mentioned external structured augmentation data (demographic or psychographic), 30% mentioned

machine generated operational data (click-stream, sensor or geo-location), 28% mentioned external social data (Twitter, Facebook) and 19% mentioned audio (streaming audio, call center voice logs).

Both structured and unstructured data is used in big data environments. The respondents were asked to describe the primary data structure within their organization's big data environment. The most used data structure in an organization's big data environment is schematic (24%) and programmatic (22%) where the structure of data is defined by applications creating the data. Compound (XML) and multiplex data structure (image, audio or video) was used by 18% of respondents. Textual data structure (data from documents, JSON) is used by 16% of respondents.

2.2 Differences of Small and Big Data

Berman, Tien and Loshin compared differences between small data and big data [Berman 13], [Tien 13], [Loshin 13]. The result was that there are differences between the two and the differences can be divided into the following categories: data sets, location of data, data structure and content, longevity, measurements, data preparation, reproducibility, data access, project costs, introspection and data processing.

Data sets: It is possible to specify the content of small data resource, how the data will be organized, connected to other data resources or usefully analyzed [Berman 13]. Big data does not have a small set of rules, known sources and moderately sized data sets [Loshin 13]. Datasets were created for one functional purpose like sales or marketing but are used multiple times in different context, especially in reporting and analysis. Sparse data sets are common in many big data use cases [Letouzé 12]. Big data applications take data from within and outside the organization, use a variety of social networking streams, public or open-sourced datasets and sensor networks [Loshin 13].

Location of data: Small data is located within one institution, maybe on one computer or even in one file whereas big data is spread onto multiple Internet servers, located anywhere on earth. Big data is distributed across thousands of processors [Berman 13], [Kimball 13].

Data access: Big data is accessed on-demand and real-time compared to the traditional on-supply and over-time access [Tien 13].

Data structure and content: Small data contains highly structured data and often comes in the form of uniform records whereas most of big data is unstructured data. Big data resource may cross multiple disciplines. The individual data object in the resource may link to data contained in other big data resources [Berman 13].

Longevity: Small data is kept for a limited time whereas big data projects contain data that must be stored for a long time [Berman 13]. Many big data projects extend into the future and the past acquiring data prospectively and retrospectively. Big data is kept for a long time so that past and future concerns are available for discovery. Because the original data sources are available for a long time, most of the data used in analytics is probably thrown away after analysis.

Measurements: Measurements may be obtained by many different protocols because of the variety of data types. Small data can be presented using one set of standard units and measured using one experimental protocol [Berman 13].

Data preparation: In small data it is possible for a user to prepare his/her own data for the user's own purpose. Because big data comes from many diverse sources and it is prepared by many people, people who use the data are seldom the people who have prepared the data [Berman 13]. Preprocessing of the data is common and the data should not be used under impression that the data received is raw. Big data is analyzed in incremental steps. The data is extracted, reviewed, reduced, normalized, transformed, visualized, interpreted and reanalyzed with different methods [Berman 13]. For example, CERN's Large Hadron Collider produces petabytes of data every day and researches filter this data to produce much smaller data sets for analysis. Unstructured data is turned it into structured data that can be stored, accessed and analyzed along with other structured data [Dayal et al. 09].

Reproducibility: small data projects are repeatable. If, for example, the validity of the conclusions drawn from the data is questioned, the entire project can be repeated. Big data projects are seldom repeatable and project users have to hope that data quality issues will be found and noticed [Berman 13].

Project costs: Small data project costs are limited, whereas big data projects are very expensive. A failed big data project can lead to bankruptcy [Berman 13].

Introspection: Individual small data points are identified by their row and column

location within spreadsheet or database table. The content of the big data resource can be intangible.

Data processing: Big data uses cloud computing whereas small data is computed locally [Tien 13]. Big data is stored in the original capture formats. Query and analysis applications are supported without converting or moving data. Big data supports data variety, arbitrarily hierarchical data structures and collections of name-value pairs. Data is loaded into the database before exploring its structure. Big data is integrated from multiple sources at GB/sec [Kimball 13].

2.3 Big Data Analytics

The purpose of analytics is to extract useful information from massive data repositories [Cuzzocrea et al 13]. Information can be extracted using qualitative analysis, where a phenomenon is studied by making connections and conclusions about variables that measure the phenomenon. The purpose of business analytics is to produce business value faster and to find essential changes [LaValle et al 11] and to make decisions [Tien 13], [Michalewicz et al. 07]. Senior executives want to run businesses on data-driven decisions. Also scenarios and simulations are wanted that provide instant guidance on the best actions to take when disruptions occur.

Big data is used to discover new insights for developing customer relationships, for identifying new areas of business opportunities and for supply chain management [Tien 13]. Traditional data is used for reporting what has happened and why [Kimball 12]. For example, mobile phone data gives researchers an ability to quantify human movement and an opportunity to discover new insights [Letouzé 12]. By using the information from mobile phones, researchers were able to give 93% accurate prediction where a person was physically located at any time based on their past movements. Another example is using big data for solutions that help to react on surprising events like earthquakes in supply chain area or customer showing first signs in changing the producer [LaValle et al 11]. Retailers can use analytics to boost competitive advantage on displays, marketing, customer service and customer experience management. The purpose of big data is to bring nuances and depth to the traditional reporting, not to replace small data analytics.

The aim and challenge of business analytics is to get value out of data, to solve how to utilize information to get commercial value out of it. Top challenges in adopting analytics in business are managerial and cultural rather than related to data and technology. Almost 40% of respondents lacked the understanding of how to use analytics to improve the business [LaValle et al 11]. Analytics derives information from data, knowledge from information and wisdom from knowledge.

Figure 1 illustrates the relationship between data to information, from information to knowledge and from knowledge to decisions. Data is digital and collected in the form of bits, numbers, symbols, flat files, JSON objects, etc. Data is organized into information by preprocessing, cleaning, arranging it into structures and removing redundancy. Knowledge is mined from information. Knowledge can be seen as facts and relationships that are perceived, discovered or learned [Michalewicz et al. 07, Ch. 1]. Data mining includes probabilities, statistics, fuzzy logic, multivariable testing and pattern analysis. Knowledge is transformed into decisions. Decisions are based on optimization and predictions that recommend near-optimal decisions. Adaptability module between knowledge and decision improves future recommendations and adapts to changes in marketplace.

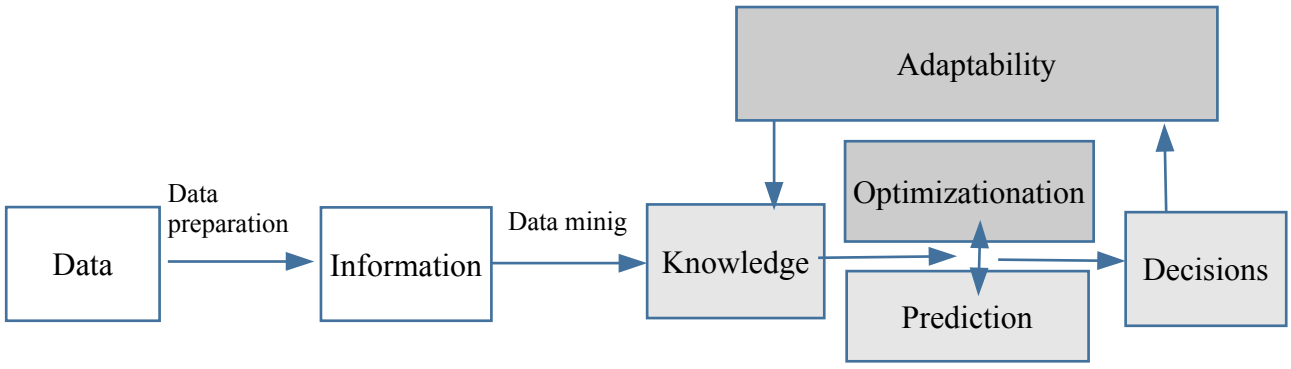

Figure 1: Relationship between data to decisions by [Michalewicz et al. 07, Ch. 1]

Some business questions are answered in an ad hoc manner. Microsoft presented fictional Blue Yonder Airline example in TechEd North America event [Bice 13]. Let us

imagine that an airline wanted to analyze customer satisfaction and to create better frequent flight passenger program. This use case example demonstrates the integration of data sources available to the airline: flight data, tweets, sentiment data and mobile app log data from frequent flyer app. Frequent flyer app may be used to manage frequent flyer miles, manage travel budgets and navigate terminals. Sentiment data helps to know how satisfied or dissatisfied customers are. Information about day, tweet, airport and sentiment is collected from Twitter. Airport related data can be found from Tweets by finding words that contained the name of an airport and words like "I was in this airport" and "my experience sucks", or has-tags like "#failed", "#flight delayed". Sentiments are loaded by using sentiment dictionary that contains words that describe dissatisfaction. Sentiment score can be counted by incrementing the count of a sentiment every time someone said something negative about the subject. The higher the sentiment score the greater the dissatisfaction.

After combining relational flight data and sentiment score we can find out which airports and on which day had the highest sentiment score. The aim was to find airports that people had most to say about in Twitter. By analyzing tweets an investigation can be made what people were actually saying. Selected top five words were: airport, delay, weather, app, thanksgiving. Unhappy sentiment was connected with words like "app", "airport", "delay" and "weather". The maybe surprising word app gives a clue that perhaps there is something going on with airlines' frequent flyer app. By investigating app log data, a peak of average processing time delay was found. The frequent flyer app could not scale to many simultaneous users which lead to unhappy customers. The action resulting from analytics was that app processing scalability was improved. Blue Yonder example is fictional, but demonstrates how different data sources and text analytics are combined in order to answer new types of business questions.

This example case answers the question how customers feel about the airline, *why* customers were dissatisfied and helped to react on scalability problems in the frequent flyer app. Analyzing unstructured data like blogs and wikis helped to understand how customers feel about the products or company. Better customer understanding helps to increase the quality of services and helps to create added value to the products or services. Blue Yonder case used tweets to understand customers' satisfaction.

Dayal et al. introduce an example scenario of integrating structured and unstructured

data in web site advertisement [Dayal et al. 09]. A web-site displays an advertisement with a discount on a selected product. Monitoring the sales of that product and evaluating the utility of the ad needs fresh data so that the campaign can be dynamically evaluated and adjusted. Nightly refresh cycle is not adequate. In another example a discount offered to the user is based on ”up-to-the-minute profile” which includes current inventory and active marketing promotions and the current actions taken by the user in this transaction. Traditional offers are based on a historical customer segmentation model and last week’s inventory. To implement these examples a low latency is required in the extract, transform, load (ETL) pipeline to capture and transport the information to the data warehouse within seconds. ETL phases of an integration process have to deal with streaming data.

There are many big data analytics use cases [Kimbal 13]: New insights can be found using search ranking, ad tracking, causal factor discovery, social CRM, document similarity testing, genomics analysis, cohort group discovery, in-flight aircraft status, smart utility meters, building sensors, satellite image comparison, /computerized axial tomography (CAT) scan comparison, financial account fraud detection and intervention, online game gesture tracking, big science data analysis, loan risk analysis and insurance policy underwriting and customer churn analysis.

Big data analytics use cases can be divided into three levels based on their analytical capabilities. There are three levels of analytical capabilities which are aspirational, experienced and transformed. Each of them have distinct opportunities [LaValle et al. 11]. Aspirational level is searching for new ways to cut costs, experienced are looking for optimization of their organization, transformed organizations use analytics as a competitive differentiator.

Regardless of analytical capabilities, most of the big data use cases represented earlier follow big data analysis pipeline. Figure 2 represent an interpretation of big data analysis pipeline that consists of five phases which are data acquisition/recording, extraction and cleaning, integration, analysis and interpretation.

Figure 2: Big Data Analysis pipeline [Agrawal et al. 11].

The data is first sampled and recorded from data sources. Data recording refers to automatically generating metadata to describe what data is recorded and how it is recorded and measured [Agrawal et al. 11]. *Extraction and cleaning* refers to preparing the data ready for analysis. Required information is extracted from original sources and it is transformed into structured format suitable for analysis. This phase is highly application dependent [Agrawal et al. 11]. Preparation includes data transformation, normalization, creation of derived attributes, variable selection, elimination of noisy data, supplying missing values, and data cleaning. Preparation is done in preliminary data analysis where the most relevant variables are identified and the complexity of underlying problem is determined [Michalewicz et al. 07]. Third, the data is integrated and represented because of the heterogeneity of the data. *Integration and representation* includes expressing differences in data structure and semantics in a way that is computer resolvable with algorithms [Agrawal et al. 11]. For effective large-scale analysis locating, identifying, understanding and citing data has to happen in a automated manner.

Most of the data is not interesting and it can be filtered and compressed (e.g. CERN data). Raw data needs to be processed so that it is more usable to analysts. Small data analysis can analyze all of the data at once whereas big data is usually “right-sized”. Right-sizing refers to transforming big data into small, understandable units in order to answer big data questions. Data sampling is a way to reduce the size of the data. In fact, data sampling gains significance in big data implementation [Gudipati et al. 13]. The real labor of producing small data out of big data is to collect and organize complex data so that the resource is ready for queries. For example, a restaurant locating smartphone app can locate five nearest restaurants. The app reduces the number of all possible restaurants down to five from a big and complex database that uses a map database, a collection of all the restaurants in the world, their longitudes and latitudes, their street addresses, and a set of ratings provided by patrons, updated continuously.

Big data analytics drill down to extract key pattern, trend and root causes. This generally includes a fair amount of mining, slicing and dicing [Deutsch 12]. Producing small or right-sized data out of big data is increasingly important also from a large scale machine learning perspective. When the range of algorithms that are practical for big data processing decreases it becomes important to right-size the data [Parker 12]. The appropriate size is dependent on the objective being learned.

Data analysis and modeling can be conducted on the resulting integrated and cleaned big data. Suggestions and solutions to a problem are needed. To solve a business problem, one way is to build a model of the problem. The model can be used for generating a solution [Michalewicz et al. 07, Ch. 2]. The solution is based on the model, so the solution is only as meaningful as the model is accurate. If the model is based on wrong assumptions, the solution is meaningless.

Data interpretation and visualization make data more understandable. *Interpretation* refers to the interpretation of analysis results. Interpretation involves verifying and understanding the results produced by a computer as well as examining all the assumptions made and tracing the analysis [Agrawal et al. 11]. There may be assumptions made in every part of the analysis pipeline. Since big data sources may be prepared by many different people, the person who analyses the data set should be aware of previous steps. Query provenance provides supplementary information about how each result was derived and what input results are based upon. Visualization is an effective way to support interpretation of analysis since it can represent large amount of information in a compact way.

3 Big Data Quality Attributes

Data quality issues include the presence of noise and outliers, missing, inconsistent, or duplicate data. Bad data is defined as biased or unrepresentative description of the phenomenon or population that the data is supposed to describe [Tan et al. 06]. From business perspective bad data includes issues that might have negative business impact.

Problems in the data quality can be random or systematic. Some errors come from flaws in the data collection process [Tan et al. 06]. A data collection error is defined as excluding data objects or attribute values from analysis or including them to analysis

inappropriately. Wrong or inaccurate data can be collected from broken sensors or because limitations in measuring devices. Untested applications may produce faulty data, in which case quality problems are caused by errors in the code. Outdated, conflicting, intentionally or accidentally wrong or misleading data (e.g. spam) can be collected from blogs, news and social media [Letouzé 12], [Bizer et al. 12]. When dealing with big data, one can expect missing values, missing records, noisy data, huge variations in the quality of records and any and all of the inadequacies found in traditional data resources [Berman 13, Ch. 10]. Traditional data quality issues can be typos, for example, misspelled names or wrong values like incorrect birth date. Some cases, like misspellings, data quality problems can be easily detected but the others are more difficult to find, like cases where admissible but not correct values are provided [Batini et Scannapieco 06].

Data quality is described by quality attributes. Quality attributes are objective measures that help to evaluate data quality in relation to given user application requirements. There are many data quality metrics that help to categorize data quality problems. There is a tendency to make distinction between data quality difficulties that refers to technical problems and information quality that refers to nontechnical problems [Madnick et al. 09]. Information quality problems are semantic challenges of locating and integrating meaningful data. Information quality can also refer to the acceptance and use of the analytic product. Technical problems are related to how to efficiently manage and process large data sets. This includes choosing right technologies and tools for analytics. Well implemented technology may be unnoticed by the business but low quality is visible and affects overall acceptance, usage, trust, value realization, and sustainability.

Quality attributes are defined differently depending on which viewpoint is taken e.g. data type, information system type and organizational level viewpoint [Batini et Scannapieco 06]. This chapter describes data quality attributes from data type, information system type and data usage viewpoint. Quality assessment can be task-independent or task-dependent, subjective or objective [Pipino et al. 02]. Subjective data quality can be measured with surveys that reflect the needs and experiences of stakeholders. Task-dependent attributes include business rules, company and government regulations that are developed in specific application contexts. Task-independent attributes do not require contextual knowledge.

Vassiliadis researched data warehouse quality issues through process quality. He categorized quality issues according to five basic data warehouse processes: design and administration, software implementation and/or evaluation, data loading, data usage and data quality [Vassidialis 00, p. 2.29]. Since analysis pipeline has similar processes than data warehouse, Vassiliadis categorization is used in following subsections. Subsection 3.1 describes design and administration quality attributes, subsection 3.2 combines data loading quality and software implementation or evaluation quality attributes, subsection 3.3 describes data quality attributes and subsection 3.4 describes data usage quality.

3.1 Design and Administration Quality Attributes

Data warehouse administration includes how the data is represented in the system. Schema quality and metadata evolution are main categories in design and administration quality issues. The *schema quality* refers to the ability of a schema to represent adequately and efficiently the information [Vassiliadis 00, p. 2:22]. *Metadata evolution quality* refers to schema evolution for example versioning and time stamping of metadata. Metadata evolution can be measured as the number of not documented changes in the metadata [Vassiliadis 00, p. 2.23].

Figure 3 illustrates how design and administration quality consist of schema quality and metadata evolution and how quality attributes like correctness, completeness, minimality, traceability and interpretability have been classified under schema quality. Quality attributes are classified in relation to data warehouses. Since big data sources may be schemaless, this thesis interprets schema quality in a broader sense, in the context of how well *data sets* represent information adequately and efficiently.

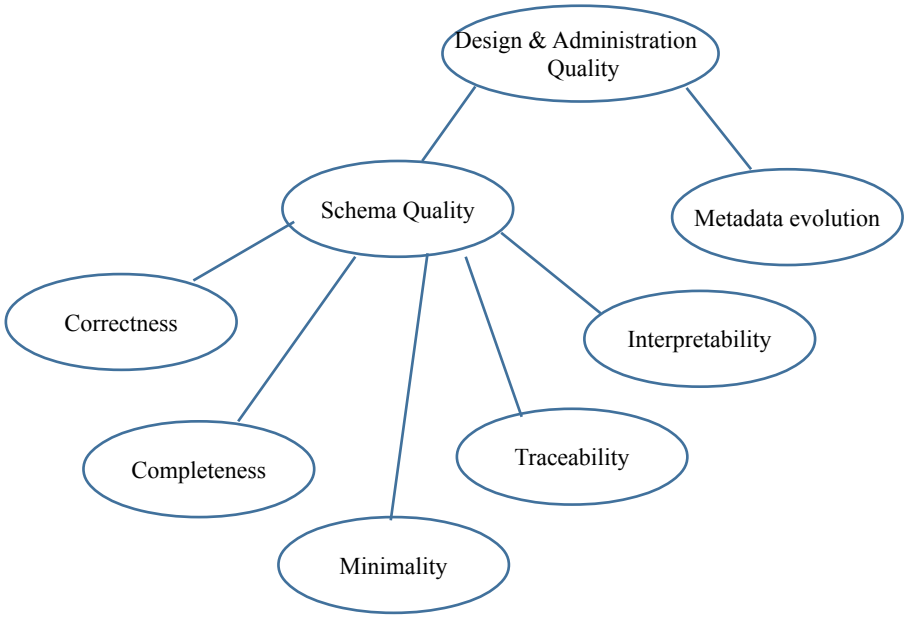

Figure 3: Design and administration quality attributes by Vassiliadis [Vassidialis 00, p. 2.29]

Correctness in data design and administration quality context refers to proper comprehension of the real-world entities [Vassidialis 00, p. 2.23]. This includes the schemata of the sources. Correctness refers to the extent to which values are valid and reliable. Correctness can be described with a term of free-or-error [Pipino et al. 02]. The severity of errors varies depending on the context. One error amongst thousands data units can be more tolerable than erroneous data in mission critical cases. For example, if one lose one geo-location record or one web click out of thousands the cost may not be as important as in the case of money transfer transaction.

Measuring the correctness requires a set of clearly defined criteria where the degree of precision must be specified. However, correct outcomes are not always known, especially in big data analytics.

Schema completeness in the data design and administration quality context refers to the preservation of all the crucial knowledge for the data warehouse schema [Vassiliadis 00, p. 2:22]. Schema completeness describes the degree to which entities and attributes are not missing from the schema. Completeness is covered in the context of sample completeness in big data. Data quality completeness is described in more details in subsection 3.3.

Minimality refers to avoiding undesired redundancy during the source integration process [Vassiliadis 00, p. 2:22]. *Uniqueness* has somewhat similar definition. Uniqueness specifies that each real-world item is represented once and only once within the dataset [Losin 13]. *Unique identifiability* refers to an ability to uniquely identify entities within datasets and data streams [Loshin 13]. Unique identifiability include linking entities to known system of record information by using unique keys.

Traceability refers to the fact that all kinds of requirements and decisions of users, designers, administrators and managers should be traceable in the data warehouse schema [Vassiliadis 00, p. 2:22]. Traceability refers to the ability to trace data back to its origin [Draper 12]. Term *provenance* is also used as describing data lineage. Data lineage includes information about data's origin and where it moves over time.

If there is a processing error at one part of the analysis pipeline, the subsequent analysis may become useless. Data provenance identifies all subsequent processing that is dependent on a step. Provenance of the data and its metadata needs to be carried through the data analysis pipeline [Agrawal et al. 11]. The source of the data, the capturing time and exact copy of the source need to be captured. Database columns or keys would then have an extra field for a timestamp. Data traceability can be implemented with saving the source information in a field of schema [Draper 12]. When the data is processed and transformed the information of the data's original source can be traced. Traceability can be measured as the number of user requirements not covered in the data warehouse schema [Vassiliadis 00, p. 2:22].

Traceability becomes important in the big data context. The same data can be analyzed over and over again. Any acquired dataset may be used for any potential purpose at any time in the future [Loshin 13]. Repeated copying and repurposing of a dataset leads to a greater degree of separation between a data producer and a data consumer. Inherent semantics associated with the original datasets fade away with each reinterpretation of what the data means. Lineage of how different data sources are integrated for analytics needs to be discoverable and reproducible. Who has done the analysis, when the analysis was made, where and how the data was received, cannot be lost. The source, actors and participants of the big data need to be defined consistently with the rest of the data. Big data loses its meaning if taken out of its context.

Immutability is other quality attribute that is connected to the traceability. *Immutability*

refers to the ability to remain unchanged over time. Immutability is in the key role in the data traceability [Draper 12]. Keeping the analyzed data unchanged is important because the original data sources can change inconsistently [Draper 12]. Some of the data sources are updated frequently and some sources like web pages are updated inconsistently. Even if the original data source is tracked, it may be very different compared to the time the data was crawled or processed.

Schema interpretability refers to how well the data model is explained, which makes querying easier [Vassiliadis 00, p. 2.26]. *Ambiguity* is an attribute that is related to the big data interpretability. Ambiguity refers to the quality of being open to more than one interpretation. Synonym for ambiguity is inexactness. Ambiguity is created by the lack of metadata in big data [Krishnan 13].

3.2 System Quality Attributes

System quality attributes are divided into data loading quality issues and software evaluation quality issues. Software evaluation includes many quality attributes such as interoperability, reliability, maturity, recoverability and usability amongst many. Scalability, performance and efficiency of a system are defined in this chapter. These attributes are gaining importance because of increasing data volumes and velocity of data. Figure 4 illustrates how data loading quality consists of analyzability and transactional availability.

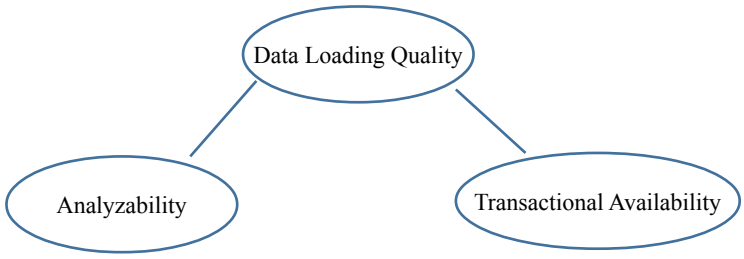

Figure 4: Data loading quality attributes by Vassiliadis [Vassiliadis 00, p. 2.29]

Analyzability refers to the validation of each process and its ability to handle errors and self-report when errors occur [Vassiliadis 00, p. 2.26]. Analyzability should be tested for

self-reporting and error handling. Testing the data warehouse processes for self-reporting in erroneous situations can be done by counting the number of processes which do not self-report.

Transactional availability refers to the time when information is not available due to update operations [Vassiliadis 00, p. 2.26]. Transactional availability can be measured as the percentage of time, when relevant information is not available due to update operations.

Scalability is represented by the amount of data being queried and the number of concurrent users simultaneously running the queries [Gupta et al. 12]. Hadoop have scale-out architecture that divides workloads across multiple nodes. Flexible file system eliminates ETL bottlenecks.

Performance refers to how well the data warehouse is capable of handling large volume of data [Gupta et al. 12]. Evaluation of the performance of a classification model is based on the counts of records that are predicted correctly or incorrectly [Tan et al. 06, p. 149]

	Predicted Class	
	Fraudulent (Positive)	Legitimate (Negative)
Fraudulent	TP	FN
Legitimate	FP	TN

Table 1 Classification results in a confusion matrix. TP = true positives, FN = false negatives, FP = false positives, TN = true negatives.

In the context of classification models, performance metrics measure how well the classification works. Performance metrics for a classification model are computed from a confusion matrix. Confusion matrix visualizes the numbers of true positives, false negatives, false positives and true negatives in a table format. False positive means that the system detects a failure that was not truly a failure. Table 1 represents a confusion matrix where rows represent actual values and columns represent predicted values. Below are some performance metrics from confusion matrix:

- *Accuracy* = (TP+TN)/total (How often is the classifier correct?)
- *Error rate* = (FP+FN)/total or 1 – accuracy (How often is the classifier wrong?)
- *Sensitivity* = TP/actual yes (When the prediction is actually yes, how often does the model predict yes?)
- *False Positive Rate* = FP/actual no (When the prediction is actually no, how often does the model predict yes?)
- *Specificity* = TN/actual or 1 – false positive rate (When the prediction is actually no, how often does the model predict no?)
- *Precision* = TP/predicted yes (When the model predicts yes, how often is it correct?)
- *Prevalence* = actual yes/total (How often does the yes condition actually occur in the data sample?)

Performance metrics can be used in classification cases when the correct values are known, for example, in fraud detection use cases. Fraud detection system classifies transactions as fraudulent and normal. The efficiency of a fraud detection system can be measured by the number of correct classifications [Michalewicz et al. 07, Ch. 12]. There is a tradeoff between the false negatives and false positives measures [Michalewicz et al. 07, Ch. 12]. Making a system more suspicious by flagging more transactions as fraudulent, will increase the number of false positives in fraud detection case. There is a little cost to classify legitimate transaction as fraudulent (false-positive), however, false negative classification carries a higher cost if transaction is significant. Because false negatives have higher costs in fraud detection system, there is a difference between error rate a and error rate b. Error rate a has a higher significance than b.

error rate a = 10 false negatives + 0 false positives/ 100 predictions = 10%

error rate b = 0 false negatives + 10 false positives / 100 predictions = 10%

False negatives affect system relevance and false positives affect system credibility [Letouzé 12]. Many false positives undermine system credibility and the number of false negatives undermine system relevance.

Sensitivity refers to an ability to detect all the anomaly cases in the system and s*pecificity* refers to the ability to notice only the relevant anomalies [Letouzé 12]. The

failure to notice relevant anomalies (specificity) leads to false positive cases. The failure to notice all the cases (sensitivity) leads to false negative cases where there really is a failure but it is not noticed. False negatives cast a doubt on the systems relevance and false positives undermine the credibility of the system. However one cannot say that false positives are more problematic than false negatives. It depends on what is being monitored and why it is being monitored.

3.3 Data Quality

Data quality is not a process itself, but it is influenced by other processes. Attributes completeness, credibility, accuracy, consistency and data interpretability concern the quality of data. Figure 5 illustrates quality attributes that relate to the data quality.

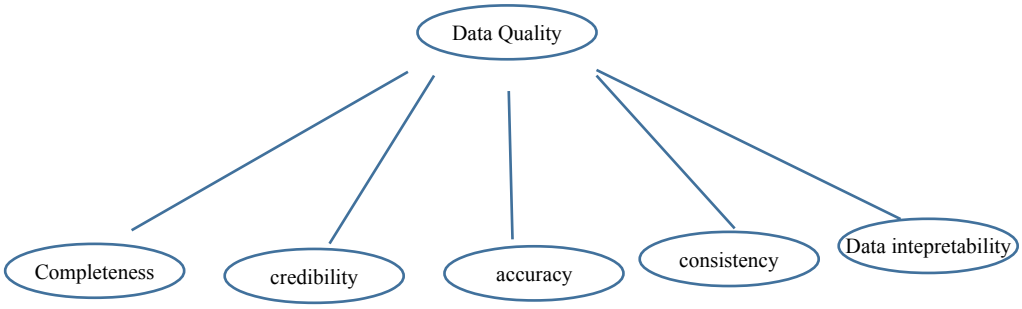

Figure 5: Data Quality attributes by Vassiliadis [Vassidialis 00, p. 2.29]

Completeness refers to the percentage of the interesting real-world information entered into the sources or the data warehouse [Vassiliadis 00, p. 2:29]. Data completeness describes the extent to which data is not missing and it is of sufficient breadth and depth for the task at hand [Pipino et al. 02]. Completeness of a data item can describe if the string describing an address actually fit in the size of the attribute which represents the address [Vassiliadis 00, p. 2:29]. Completeness can be viewed as schema completeness, column completeness and population completeness in relational data [Pipino et al. 02]. Schema completeness is described in subchapter 3.1. Column completeness describes the missing values in a column of a table. Population completeness describes if all the real-world information is entered. If a column should contain at least one occurrence of all 50 states, but it only contains 43 states there is a population incompleteness.

A complete sample includes all data items from a parent population that satisfy a set of a selection criteria. Completeness in big data settles for statistical sampling. If the sampling is made poorly, there may be necessary datasets missing [Gleason et McCallum 13]. Sampling may lose information. The size of the data is meaningless if the sample is not taken into account. Sampling selection may be biased. *Sampling selection bias* refers to a bias that results from an unrepresentative sample. For example, the people who generate real-time digital data from mobile phones or other digital services, are not a representative sample of a larger population [Letouzé 12]. Depending on the data, younger or older or wealthier or poorer individuals can be expected.

Representative sample refers to a statistical population that accurately reflects the members of the entire population. A big data sample size may be large but it may not be a representative or a complete sample [Madsen 13] [Bizer et al. 12]. For example, if only one mobile phone company's data is available, the resulting sample does not represent the population of mobile-phone holders or of the population of the area. Also, huge data sets collected from Twitter may not be representative [Boyd et Crawford 12]. If Twitter removes tweets containing swear words, tweets composed of nonword character strings, tweets containing highly charged words, or tweets containing certain types of private information, then the resulting data set, no matter how large it may be, is not representative of the population of tweeters. If no identifier for sender is associated with tweets and the tweets are available as a set of messages, tweeters who send hundreds or thousands of tweets will be overrepresented and the one-time tweeters will be underrepresented. Even if there is an identifier associated with the tweet there might be users that have multiple accounts, while some accounts are used by multiple people. Some accounts produce automated content without directly involving a person. Accounts do not represent people.

Non-representative samples are a problem because they lack external validity [Letouzé 12]. External validity refers to the degree to which an internally valid conclusion can be generalized to a different setting. However, non-representative samples are not valueless, they just need to be treated with care. This means that the analyst is fully aware of limitations and keeps the claims and decisions made on the basis of the data [Letouzé 12]. Random sampling is used for avoiding unbiased or unrepresentative samples. Random sampling ensures that there is an equal probability of selecting any

piece of data from a data set.

Credibility describes the trustability and believability of the source that provided the information [Vassiliadis 00, p:2:29]. Believability describes the extent to which the data is regarded as true and credible [Pipino et al. 02]. There are concerns about believability of analytical results because of limited visibility into trustworthiness of the data source. Reliability and validity of unstructured user generated data may be difficult to notice. Unstructured user generated data is spontaneous by its nature and have looser verification steps [Letouzé 12]. It is possible that individuals may alter facts or even publish false information. Especially in web data the main challenge is to assess the quality of web data and to determine the subset of the available data that should be treated as trustworthy [Bizer et al. 12].

Credibility can be measured by examining the documentation of the source which provided the information [Vassiliadis 00, p:2:29]. A measure for credibility is achieved by calculating the percentage of inaccurate information provided by each specific source.

Accuracy refers to how exactly in all details data represents reality or a verifiable source. Accuracy is a term that refers to the degree of measurement error in data [Tan et al. 06]. Accuracy is essential to data but can mean also accuracy of learned extraction models and algorithms. The information is extracted and turned into structured data through learned extraction models that are hardly ever 100 % accurate [Dayal et al. 09]. Accuracy does not only mean how verifiable and trustworthy sources and data samples are but how accurate the predictions are.

Vassiliadis approaches data accuracy from a data entry process point of view. Accuracy describes the correctness of the data entry process which happened at the sources [Vassiliadis 00, p:2:29]. Vassiliadis suggests measuring accuracy as the percentage of stored information detected to be inaccurate with respect to the real world values, due to data entry reasons.

Consistency refers to the logical coherence of the information [Vassiliadis 00, p:2:29]. Consistency can be viewed from a perspective of the same redundant data values across tables [Pipino et al. 12] or with respect to logical rules and constraints [Vassiliadis 00, p:2:29]. For example, data may look fine but there may be inconsistencies: A person

might be 2 meter high but weights only 2 kg.

Precision consistency describes whether each data source share the same precision and if those units are properly harmonized or not [Loshin 13]. Different datasets may not share the same precision, for example, sales per minute versus sales per hour. Semantic inconsistencies complicate analytics. Similarly named attributes of different datasets may not share the same meaning e.g., is a “customer” the person who pays for our products or the person who is entitled to customer support? Or is “M” and “F” for male/female or Monday/Friday? Loshin raises the importance of metadata in semantic consistency [Loshin 13]. Metadata activity may join a glossary of business terms, hierarchies and taxonomies for business concepts. Metadata defines relationships across concept taxonomies for standardizing ways. Entities identified in structured and unstructured data are tagged in preparation for data use.

Consistency can be measured as a percentage of inaccurate information provided by each specific source or as a percentage of violations of a specific consistency type to the total number of consistency checks.

Data interpretability measures the descriptions of data e.g. table description for relational databases, primary and foreign keys, aliases, defaults, domains, explanation of coded values, etc. [Vassiliadis 00, p:2:29]. *Veracity* refers to the biases, noise and abnormality in big data [Normandeau 13] which can make the content of the big data resource intangible. A technique called introspection enables access to data, access to information about data values and to information about the organization of the data [Berman 13]. Introspection refers to an ability of a data object to describe itself when called upon. The term introspection is originally used in object-oriented programming field. Correctness of the introspection and how completely data describes itself could fall into schema quality category. Data interpretability can be measured as a number of pieces of information not fully described [Vassiliadis 00, p:2:29]

3.4 Data Usage Quality

Quality attributes presented in previous subsections address syntactic and semantic correctness but fail to address user requirements. Data usage quality takes the content and context of data into consideration.

Data usage quality consists of accessing the data for analysis and usefulness of data. Usefull data is suitable for its intended use. Data usefulness considers temporal characteristics (timeliness), the responsiveness of the system as well as interpretability of data. Figure 6 illustrates quality attributes that relate to data usage quality. This subchapter desribes also big data related quality attributes validity, volatility, virality and viscosity.

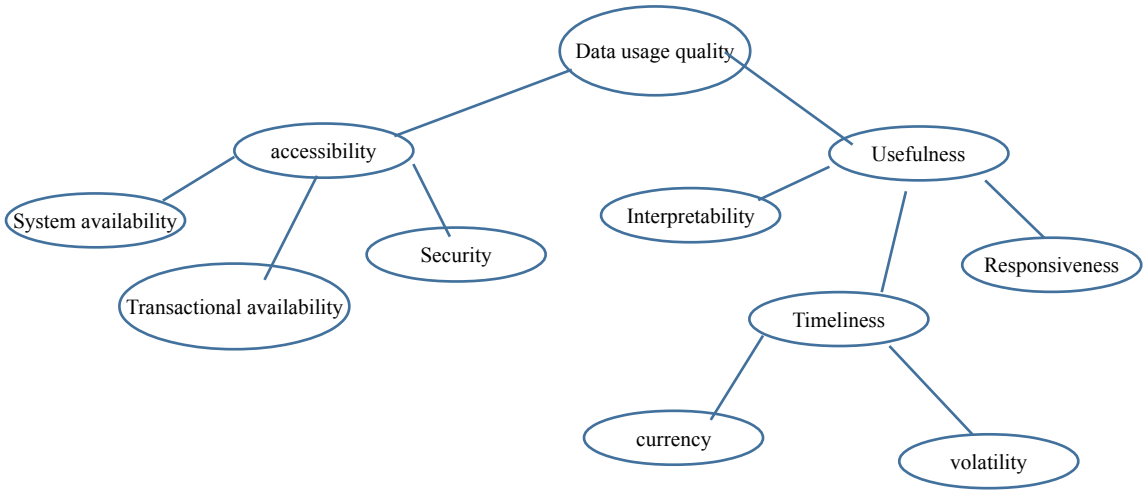

Figure 6: Data usage quality attributes by Vassiliadis [Vassidialis 00, p. 2.27]

Accessibility refers to the extent to which data is available, or easily and quickly retrievable [Pipino et al. 12]. Accessibility relates to the usage of data: is the data easily accessible, understandable, usable and accessible for querying [Eckerson 02], [Vassiliadis 00]. Data is accessible if the analyst is able to read the data, uncompress or otherwise extract the files and convert it into a readable format [Fink 12]. Measuring the system accessibility tracks down the cases where failures, update operations, or other operations, whether in the warehouse or the sources, make information unavailable [Vassiliadis 00]. The access to data may be restricted due to security reasons, e.g. by setting query privilege restrictions.

There may be different data accessing levels based on who tries to access the data. Some companies offer small data sets to university-based researchers for free, some

companies do not share access to their data and some may sell the privilege of access. Researches who can buy access privileges or people inside the company have different data access available. This complicates evaluation of methodological claims, because methodological claims cannot be reproduced nor evaluated if there is no access available to the data [Boyd et Crawford 12].

Legal arrangements are needed to secure reliable access to data streams and to get an access to back up data for retrospective analysis and data training purposes [Letouzé 12]. However, getting a formal access or agreement on licensing issues around data may be problematic. For example, privacy laws like Europe's Data Protection Directive regulates the data collection about residents [Draper 12]. Boyd et Crawford discuss also the ethicality of the data access. Do people know their data is analyzed? What if public blog post is taken out of context and analyzed in a way that the author never imagined? Just because the data is accessible does not mean that researching the data would be ethical [Boyd et Crawford 12].

Security describes the authorization policy and the privileges each user has for the querying of the data [Vassiliadis 00, p. 2:26]. Security can be controlled with preventing unauthorized access. In software implications level security is measured as the number of modules unable to prevent unauthorized access to programs and data. Data usage viewpoint measures security by measuring authorization procedures and their documentation. Measurement is the number of undocumented authorization procedures.

Privacy can be defined as individual's right to control or influence what information related to him may be allowed to be seen [Letouzé 12]. In addition to individual's right to control information related to him, companies may wish to protect their competitiveness by controlling the access to data sources. Companies may need to protect their competitiveness by not sharing data about their clients and users, or data about their own operations. Tweets that are available through API's exclude content that a user chose to make private or ‘protected’.

Privacy has an affect on data acquisition, storage, control, use and presentation. There is a tradeoff between privacy and accessibility. For example, how openly documents are shared in internet may have an effect on the accessibility of the data. The more people want to protect their data, the less data is accessible.

Interpretability describes the extent to which the data warehouse is modeled efficiently in the information repository [Vassiliadis 00, p. 2:26]. The better the data explanation is, the easier it is to write the queries.

Big data offers a lot of possibilities for analytics and potential information. However, big data does not necessarily imply better understanding of the underlying problem [Tien 13]. The volumes of data offer connections in all directions and therefore there is a risk to see patterns of data where none exist [Boyd et Crawford 12]. Massive quantities of data can lead to focusing exclusively on finding patterns or correlations without understanding of the deeper dynamics at play [Letouzé 12]. This may happen because volumes of data offer connections that radiate in all directions. Search for interesting correlations might be interpreted correctly or incorrectly as causal relationships. Correlations may also be affected by a confounding factor.

Jim Fruchterman is a blogger who wrote an example of drawing false conclusions on the basis of crowd -reported information [Fruchterman 11]. There was a catastrophic magnitude 7.0Mw earthquake in Haiti in 2010. A non-profit open source software company Ushahidi found a correlation between building damage and SMS streams: there were more SMS streams in the areas of damaged buildings. However, this correlation was not right because SMS feeds and building damages were correlated with the simple existence of buildings. In the areas with more buildings there are likely to be more people to message about damages. Also, there are likely to be more damaged buildings on the areas where population density is large than in the areas where the population density is small. The existence of buildings confounds the relation between SMS feeds and damaged buildings since the existence of buildings is a cause of both SMS feeds and damaged buildings. Thus the existence of buildings is the confounding factor and the correlation between the SMS feed and the building damage is an artifact or spurious correlation.

When the presence of any buildings were controlled there seemed to be a weak negative correlation with the presence of damaged buildings. Negative correlation means that the presence of text messages suggests there are fewer damaged buildings in a particular area. This seems intuitive because in areas where there are most damages it seems believable that the first thing people do is not to send messages. People may move away from damaged areas before texting. Also, in damaged areas there might be a high

mortality or departure from the zone of interest. This leads to an attrition bias which is a bias caused by loss of participants.

Data dimensionality refers to the number of measured attributes for each data object [Berman 13]. In other words, the number of details in each transactions increases. The attributes for a data object create multidimensional space. As the number of details for a data object increases, the multidimensional space becomes sparsely populated and the distance between any two objects increases. Data dimensionality has an affect to algorithms that compare distances of data objects. Distances are calculated when classifications and predictions are made. Clustering becomes meaningless if the space is too large [Berman 13, Ch. 10].

Sometimes the data can be correct and analysis is still somehow wrong. For example, in case of highly skewed power-law point distributions the typical value does not mean the average [Janert 12]. Highly skewed point distributions need to be diagnosed, otherwise all standard calculations, like calculating averages, are meaningless. Highly skewed point distribution data could be data where a service producer has 2000 accounts and generating a total of 5 million in revenue. If the value of each account is calculated as an average (to be worth 250 dollars) the conclusion is misleading. In reality majority of accounts are worth only a few dollars and a few accounts generate thousands of dollars revenue each. In areas that are related to human behavior variations are so dominant that there is no sense to try to find a typical value. Big data analysis need to be treated on a case-by-case basis. The analyst may find that if an account manager focuses on the top 150 account he or she can still capture 85% of expected revenue.

Data samples are combined and reanalyzed at different stages of the analytical process. However, data is not always additive and conclusions cannot be drawn based on subset comparison because of Simpson's paradox [Berman 13, Ch. 10]. Simpson's paradox refers to reversing of findings that apply to smaller data sets when the data sets are aggregated. There may be a relationship or a correlation for each of smaller data sets, but when the data is aggregated, the correlation that was noticed before may reverse itself. Simpson's paradox has significance in big data research because data sets are combined and reanalyzed in different stages of the analytical process [Berman 13, Ch. 10]. An example of Simpson's paradox is Berkley gender bias example, where men applying to the University of California, Berkeley, were more likely to be admitted than

women [Bickel et al. 75]. The admission figures for the fall of 1973 showed that 44% of men was admitted out of 8442 applicants and 35% of women out of 4321 applicants were admitted to the school. The admission data is aggregated data that combines numbers of admitted applicants from various departments. However, if this data is examined department by department, one can find out that women were being admitted at higher rates than men in almost every department. The nearly 10% difference of admitted percentage exists because women tended to apply to departments which denial percentage is high (popular and oversubscribed department) whereas men applied to departments which were avoided by women at the time (like engineering).

Table 2 presents another example of Simpson's paradox. Letters A and B represent persons who are supposed to improve articles. Person A improves 0 articles and person B improves 1 article during week 1. Person B has higher improving percentage than person A in both weeks 1 and 2. However, person A has higher total percentage.

	Week 1	**Week 2**	**Total**
A	0/3	5/7	**5/10**
B	**1/7**	**3/3**	4/10

Table 2: An example of Simpson's paradox. This paradox happens because the ratio of improved articles were not taken into consideration. Person B has higher percentage than person A in both weeks 1 and 2. However, person A has higher total percentage.

One should be careful about predictions and finding intentions from web-data. Blog posts or online searches about a product and its market prices are based on *expressed intentions*. Online searching and discussions on web may be a poor indicator of actual intentions and ultimate decisions [Letouzé 12]. The line between reported feelings and facts may not be easy to distinguish. Though expressed intentions might give an important insight from the business/marketing point of view. Slang and sarcasm make finding the true intent of a statement more difficult [Letouzé 12]. Besides finding the true intent from slang and sarcasm the true significance of the statement may be hard to notice. For example, there is a difference if a person is loosing “a” job versus loosing

one's only job.

Validity refers to the data correctness and accuracy for the intended use [Normandeau 13]. Validity of the information should be measured with respect to time [Vassiliadis 00, p. 2:27]. The data is valid only for one specific time interval.

Timeliness refers to how current the data is for the task [Batini et Scannapieco 06] and how relevant the stored information is to the real world facts [Vassiliadis 00, p. 2:27]. Timeliness is depended on end-consumer expectations. *Temporal consistency* refers to the timing characteristics of datasets to see whether they are aligned from a temporal perspective [Loshin 13]. Time perspective is out of sync between datasets. For example todays' transaction data is compared to pricing data from yesterday.

Volatility refers to how long data is valid in the real world and how long it should be stored [Vassiliadis 00, p. 2:26]. The question when the data is no longer relevant to the current analysis is more valid in real-time data [Normandeau 13]. Volatility can be measured as the number of pieces of information where valid time is not present, although needed. This requires keeping track of the time period during which the information is valid in the real world [Vassiliadis 00, p. 2:28].

Currency refers to whether the datasets are up to date. Data in internet is updated in different time intervals. Current data may be useless because it is late for specific usage, which refers that the timeliness of the data is bad [Batini et Scannapieco 06].

Currency can be measured with keeping track of the date when the data was entered in the sources and the warehouse [Vassiliadis 00, p. 2:28]. The measurement is the number of pieces of information where transaction time is not present, although needed.

Responsiveness refers to the ability of a system to complete assigned tasks within a given time. Vassiliadis considers the responsiveness through the interaction of a process with the user. Processes are tested for how well they inform the user on their progress. Measurement of responsiveness is the number of processes that do not self-report to the user [Vassiliadis 00, p. 2:28].

Virality refers to the quality of how quickly data is shared in a people-to-people (peer) network [Krishnan 13, Ch. 2]. The rate of spread is measured in time. Counting re-tweets that are shared from original tweet is a good way to follow a topic or a trend. When virality is measured, it does matter what the context of the tweet to the topic is.

Viscosity refers to measuring the resistance to flow in the volume of data [Krishnan 13, Ch. 2]. An organization may refuse to accept the usage of the data, for example social media data, because they cannot understand how it impact their business. Resistance can be shown in data-flows, business rules, and even be a limitation of technology.

4 Hadoop Based Big Data Architecture

The relationship between the data warehouse and big data is merging. Traditional highly structured and optimized operational data remains in controlled data warehouses. Data that is highly distributed is controlled by a Hadoop-based or similar NoSQL infrastructure. Big data architecture integrates data warehousing and Hadoop-based infrastructures into a hybrid model. Big data strategies that corporates are using today are not limited to a single platform or solution [Devlin et al. 12]. Analytical databases, discovery platforms and NoSQL solutions beyond Hadoop can be are used for solving big data requirements.

Figure 7 represents one Hadoop-based big data architecture and big data testing focus areas. Apache Hadoop [http://hadoop.apache.org.] is a popular open source framework that allows distributed and scalable processing of large data sets. Hadoop implements MapReduce paradigm which is a distributed computing paradigm. Hadoop consists of two parts: a file system called Hadoop Distributed File System (HDFS) and a Map Reduce programming paradigm. The components of the hybrid architecture are HDFS, map-reduce paradigm, application development languages Pig and Hive, a NoSQL database called Hbase, enterprise data warehouse and business intelligence (BI) tools for reporting. Data from various sources is extracted based on business requirements and loaded into HDFS before processing the data further for analytics purposes [Gudipati et al. 13].

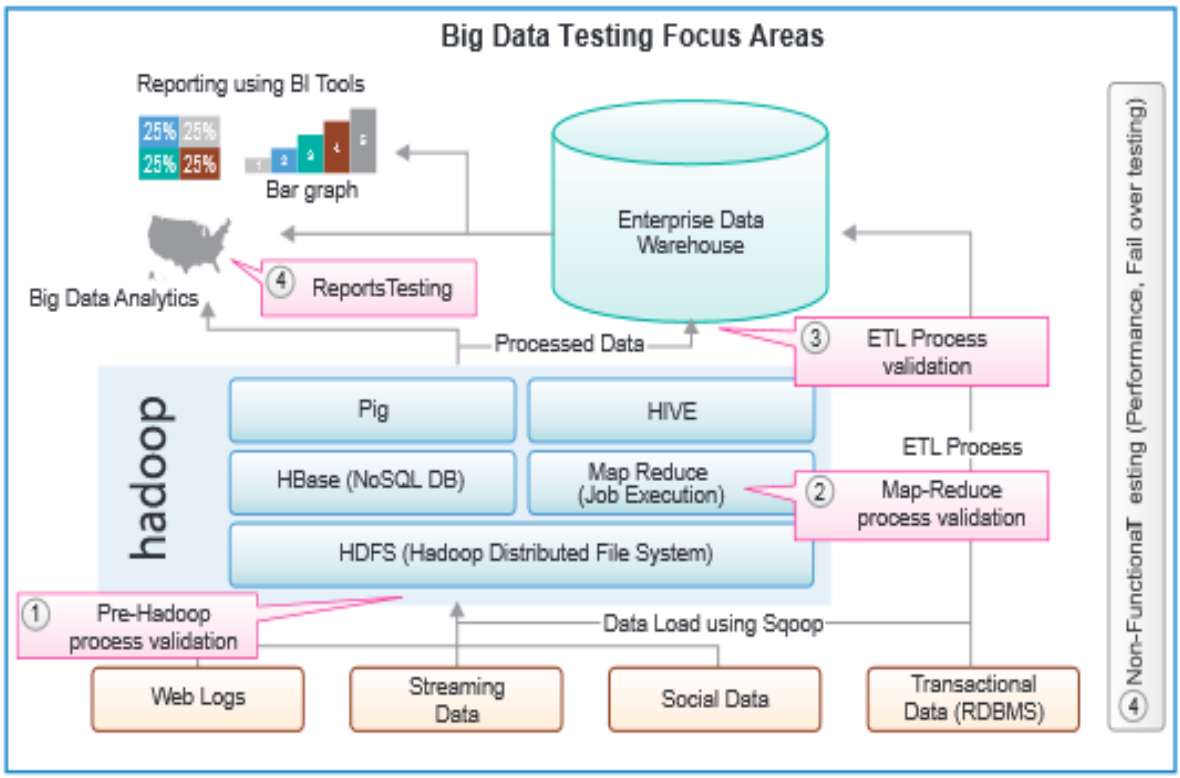

Figure 7. Big data architecture and big data testing focus areas [Gudipati et al. 13].

Hadoop Distributed File System: Data in Hadoop cluster is broken down into blocks and copies of blocks are distributed throughout the cluster [Zikopoulos et al. 11]. One file can be divided into several data blocks, all of them are copied on two additional servers by default. Figure 8 illustrates how files are divided into data blocks. Copying the data allows better failure recovery and availability of the data. Redundancy offers data locality which is critical when working with large data sets. Each server can work on the data at the same time. Hadoop splits up workloads across multiple compute nodes. This is convenient particularly when large unstructured data sets are handled. Hadoop architecture consists of a *NameNode* and hundreds of data nodes hosted on several machines [Gudipati et al. 13]. NameNode server manages data placement logic and keeps track of all the data files in HDFS. When a file is created in Hadoop, the HDFS will automatically communicate with the NameNode and allocate storage on servers [Zikopoulos et al. 11]. A regular backup process is recommended for the NameNode to ensure accessibility and availability of the data.

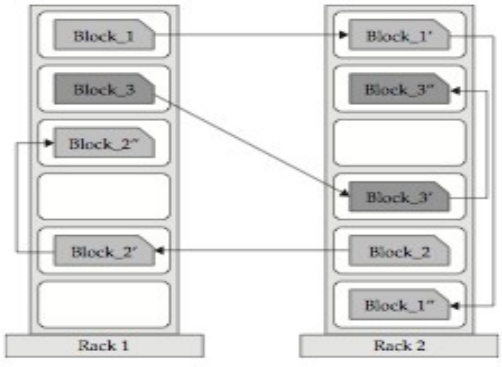

Figure 8: Clusters are prone to failures. Hadoop stores the data redundantly across the clusters to enable better failure recovery. Data block replications are stored in DataNodes. One file can be divided into several data blocks.

Map-reduce consists of two distinct tasks which are called *map* and *reduce*. Map function distributes the computation and reduce task combines the input results. A map function converts a set of data into key/value pairs. Each data node in the map phase reads the input file and does the computation. This output is directed to the appropriate reduce task by key value. All the records which have the same key value are sent to the same reduce task. For example, let us imagine that there are *m* input files and each file contains two columns, a name of a city and the corresponding temperature recorded in that city for the various measurement days. City is the key and temperature is the value. We want to find the minimum temperature for each city across all of the m data files. Same city may be represented multiple times in each file. Each of the m files can be broken down into map tasks where each mapper returns the minimum temperature for each city. All of these output streams are fed into reduce tasks which combine the input results and output a list which contains a single minimum temperature for each city.

Map-reduce program is divided into map and reduce tasks in a daemon called *JobTracker*. Daemon is a program that runs background processes. JobTracker implements the locality principle and attempts to schedule tasks on the nodes where the data is stored. If some tasks fail to complete the JobTracker reschedules that task on another node in the cluster. Pig and Hive are application development languages that run on top of Hadoop. Pig is a programming language that makes it easier to write map and reduce programs. Hive allows developers to write Hive Query Language statements

that are similar to SQL statements [Zikopoulos et al. 11]. Hbase is a NoSQL, column-oriented database management system that runs on top of HDFS.

5 Data Quality Governance

There are three objectives to the data quality governance. The first objective in improving data quality is to understand the value of data within an organization. Instituting the the right levels of control, identifying and prioritizing data issues and correcting data are other objectives of data governance [Loshin 13].

Understand the value of data: There must be understanding on how quality data is expected to improve business processes or how ignoring data problems leads to undesired negative impacts. In other words, the quality of the information must be directly related to the ways the business processes are [Loshin 13]. The relevance and severity of quality challenges are different depending on the questions and decisions made [Letouzé 12]. Understanding business processes and requirements for data quality requires discussion with data consumers. The requirements of the data quality differ depending on the viewpoint and context. There may be different levels of usability and acceptability in acquired datasets by different parties. For example, salespeople may need rough estimates whereas financial analysts need precise data for accurate forecasts [Eckerson 02].

Language gap is a limitation in data quality control. Depending on their role different users may have different opinion on quality attributes. People who collect the data may have a different view of “complete” than people who analyze the data. If business performance is measured, data can be incomplete for analysts if the order date is missing [Gleason et McCallum 12]. There may be a language gap between business stakeholders (people in human resources or finance) and a technology team. Business stakeholders may define data quality in terms of guiding principles like relevancy, timeliness and access, whereas technology team may define data quality in terms of discrete data conditions like accuracy, completeness, consistency [Goetz et al. 13]. IT may define the data quality by the physical nature of data to pass or fail data processing rules.

The value of data can be understood through the costs of poor quality data. Costs due to

low data quality can be categorized in three category [Batini et Scannapieco, Ch. 4]. Process failure costs may be incorrect mailing addresses that cause misdelivered mail. In this case the process does not work properly because of the poor quality data. Redundant data handling, business rework costs and data verification costs are called information scrap and rework costs. If the data source has poor data quality the data has to be collected from another source. This requires time and money. Business rework costs are due to re-performing failed processes. Business rework is done when misdelivered mail is sent again. Data verification costs occur when the data is not trusted and data users have to perform their own quality inspection. Loss and missed opportunity costs may be profits that were lost because of e-mails do not come through. Failed periodic advertising campaigns may have lower revenues because a percentage of customers cannot be reached.

Identify data issues: Data quality assessment is defined as a process for obtaining measurements of the data quality and to determine the current state of the data quality [Woodall et al. 13]. Data quality assessment is also referred as data auditing or profiling in the literature. It is important to distinguish quality dimensions that are only measurable from those that are both measurable and controllable. If there is no control over quality dimensions, the measures can be used to assess usability. Otherwise corrections or updates can be made.

Prioritization: When the value of the data is understood and data issues are identified, quality problems, testing routines and corrections should be prioritized. It is not always possible to perform all the testing routines. Finding equilibrium between a quality of a product and a production cost is vital for organizations [Vassiliadis 00]. Without equilibrium the organization loses by paying too much money for achieving quality or by producing low quality products. Low quality product result in bad reputation and loss of market share. There is a need for a generic testing approach that takes resource limitation into account. A generic testing approach includes using prioritization and differentiation for testing routines according to the importance and impact on the output product. Data warehouse environments lack a generic and well defined data warehouse testing approach that could be used in any project and which takes the dependencies between test routines into consideration [ElGamal et al. 11]. If testers have resource or time limitations they have to be able to decide which testing routines are affecting

highly in the quality of data warehouse. Since big data testing has bigger and more complex resource limitations than data warehouse, the need for a generic testing approach becomes important.

One resource limitation is the movement of data which requires network resources and introduces latency if done on demand. Processing and moving data requires always energy and is prone to errors. Data warehouse testing has resource (employee, project resources) and time limitations. In addition, big data testing has network resource limitations, bandwidth and data processing limitations. For achieving completeness, some time consuming activities like checks needs to be done thus the timeliness is negatively affected [Batini et Scannapieco 06]. In addition to dependencies of the testing routines, there may be correlations and tradeoff between quality dimensions [Batini et Scannapieco 06]. For example, tradeoff can be between timeliness and any one of the three other attributes: accuracy, completeness, and consistency. Information extraction algorithms are usually slow, tradeoffs between accuracy and performance may be important [Dayal et al. 09].

5.1 Understanding the Data Through Preliminary Analysis

The first step in trying to answer the question is to understand what to expect from the data. Headers of rows may provide a clue about what the data contains. Data elements may have a key that hopefully is reasonable descriptive. For example, are the distances in miles, kilometers or meters, are revenue fields in gross or net. The definition of the field plus actual values help to avoid misinterpretations. For example, IP addresses should be integers or dotted quads and currency fields should be decimals with two to four digits after the decimal [Fink 12].

Structured values can be validated using validation scripts that use regular expressions. Scripts can check if the values that are supposed to be numbers are numbers or if the values in enumerable fields fall into the proper set (e.g. months between integers between 1-12 or January – December). The script example below is for validating fixed-format fields like IP addresses [Fink 12]. Example has a text file “sample.txt” that contains two IP addresses for network addresses fink.com and bogus.com. The script validates that bogus.com has invalid IP address.

```
$ cat sample.txt
```

```
 fink.com     127.0.0.1    bogus.com     1.2.3
$ cat sample.txt | \ perl -ape 'warn "Invalid IP!" if $F[1] !~ /^\d+\.\d+\.\d+\.\d+$/'
  fink.com     127.0.0.1
 Invalid IP! at -e line 1, <> line 2.
 bogus.com     1.2.3
```

Simple statistical checks like taking the minimum or maximum value of the field can be used to check if the value makes sense in the context of the field [Fink 12], [Pipino et al. 02], [Barateiro et Galhardas 05]. Minimum value of the counter (e.g. click through rate) should be 0 or greater, and financial values are usually numbers with two digits decimal and should have a reasonable upper bound. Average, mode or medians can be used to check if values of a fields make sense. Statistical checks can be automated. However, in extremely large data sets simple statistics is not adequate [Fink 12].

Visualization is a tool for gaining insight what kind of hypothesis are possible. It helps to give an oversight about the data and helps to understand the relationship between variables in the data. Figure 9 presents the distribution of diamonds over weight (carats) and price (dollars). The diamond data set consists of 10 variables and 53940 observations. From the scatterplot one can see that the price of a diamond varies a lot in the same weight class. There are standard weight classes like 0.5, 0.7, 0.9, 1.0, 1.2, 1.5, 1.7 or 2 and most diamonds sold are small.

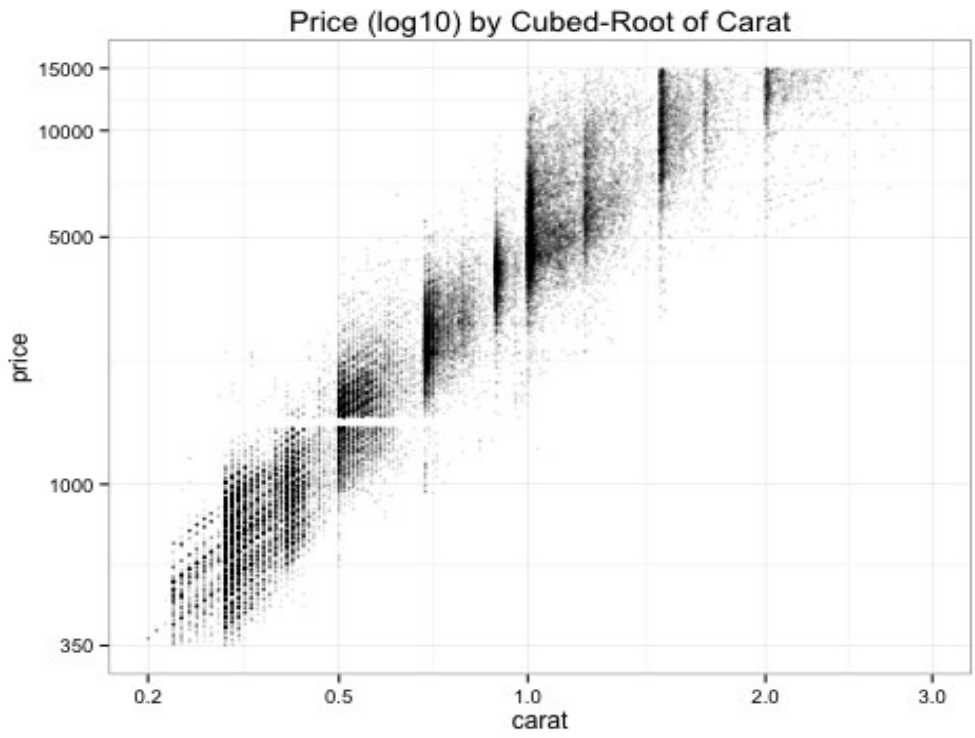

Figure 9: Visualization helps to have an oversight about the data. The picture represents the distribution of diamonds over weight (carats) and price (dollars).The diamond data set comes with the ggplot2 package which is a plotting system for the statistical computing language R.

5.2 Identifying Data Issues

Data quality attributes are measures that help to describe the quality of data. Attributes can be considered from technical, nontechnical, subjective, objective, task-dependent and task-independent viewpoint. Nontechnical quality problems include difficulties in understanding the data and therefore difficulties to gain information from the data. Wrong conclusions can be made if sample representativeness and sample bias issues are not considered. Privacy and different accessibility affect on sample representativeness and completeness. In order to evaluate the accuracy and believability of analytical results data traceability is important. Repurposing the data sets emphasizes the importance of the concept of lineage and traceability. Traceability is connected to the believability and trustworthiness of analytical results. If original data sources can be verified, hypotheses made of the data can be put on test. Big data strategies should keep speed and accuracy in mind because of increasing competitiveness in the business.

Accuracy is critical for unstructured data since information is extracted from unstructured sources through learned extraction models [Dayal et al. 09]. Many of the schema quality attributes could be used for describing the quality of the analytical model as well as describing the system quality attributes. Knowledge is extracted through predictions and algorithms in big data analytics, so the design and administration quality factors are interpreted as the ability of an analytical *model* to represent adequately and efficiently the information.

Findings in big data analytics are subjective compared to objective findings in small data [Tien 13]. The Blue Yonder example combined different data sets for improving frequent flyer customer program. Combining data sets and interpreting the analytical results are based on subjective decisions. Subjective work with big data may not lead to closer claim on objective truth [Boyd et Crawford 12]. Particularly if messages from social media sites are used. Objective truth is questioned because the analytical model may be based on wrong assumptions and subjective decisions are made in every phase of the analysis pipeline. For example, interpretations about the data set are depended on the question asked, context and skills of the analyst. Challenges are noticing selection bias issues and the interpretation of the meaning of human generated text. If the selection bias is not taken into account, then some conclusions of the data may not be accurate.

Data preparation phase includes deciding which data to keep and which to throw away. The decisions included to data preparation and cleaning processes are subjective and limited to the business requirements. It is difficult to understand the data and data's deeper dynamics. There is a risk to see nonexistent patterns in big data and big data may be processed in unfocused, unproductive and shallow manner. Not fully understanding the data and unfocused, shallow data processing are intertwined together. For example, the volume and dimensionality of the data may lead to finding statistical significance without having any significance in reality [Berman 13].

5.3 Correcting the Data

When the quality of data is measured and the measures are below acceptable levels, one can either not use the data at all or use the data anyway but modify end user's

expectations in relation to the quality measures [Loshin 13]. If business application does require trustworthy, accurate and precise results, then one should not use the data. Managing consumers' expectations includes discussion of how quality aspects of the input data that might affect the computed results. For example, things like who the potential end users are, what they want to do with the datasets and what the expectations are, need to be discussed with the consumer [Loshin 13].

In some cases data quantity overcomes some data quality issues. If the data set is very large and have a small number of errors then a minimal percentage of data flaws will not significantly skew the results. For example, the noise in the whole data set is less of a risk than distortion of compressed results from an incorrect or constrained sample. The whole data should be used if that is an option [Deutsch 12].

There is a statement that data samples are complete in big data approach and not representative like in traditional data approach. However, data sampling gains significance in big data implementation [Gudipati et al. 13]. Even if big data samples the data, the whole data should be used if that is an option [Deutsch 12]. Large datasets are recommended because algorithms become more accurate. On the other hand, if data sets are too large, large-scale algorithms become less effective (because of the curse of dimensionality).

The lack of data quality is still problematic in operational level, but larger data acquisition mitigates the problem [Tien 13]. Big data allow messier source data which is still good enough to support informed decisions. However, one should be careful of thinking that large data sets will cancel out bad measurements and overcome data quality issues with data quantity. Berman names this as "cancel-out hypothesis" and continues that it is a belief that is based only on wishful thinking [Berman 13, Ch. 10]. Cancel-out hypothesis implies that huge amount of data cancel out errors in the long run, yielding conclusions that are accurate.

In statistics, the more experiments are done the more accurate the results are because of the law of large numbers. Large numbers of experiments diminish the impact of measures that happens because of change and the average of results should be close to the expected value. For example, more data is better in business analytics, as data mining can produce better results when performed on large data sets. Prediction models

become more accurate [Michalewicz et al. 07]. When human generated text is analyzed, e.g. discussion about the topic, it is easier to manage noisy or bad data when there are more data. The volume of activity takes care of outliers and missing or bad data will probably not cause a misinterpretation of what people mean [Deutsch 12]. Big data enables better observation of rare but important events and better evidence-based decisions that may differ from intuitive decisions [Tien 13]. If the bad measurements are systemic errors, the results are not becoming more accurate. For example, broken sensors produce erroneous data which does not cancel-out in the long run by the amount of data. More data produces more errors.

Third option is to change the data to a more acceptable from. However, this is not as straightforward than in the small data approach. Traditional data is cleaned before analytics tools, for example during ETL processes, and the data is optimal for analysis. Some of the quality metrics like accuracy, completeness, consistency, currency and uniqueness are targeted to moderately sized data sets, from known sources, with structured data [Loshin 13]. These attributes use a relatively small set of rules to validate data, to compare input data to those rules and to correct recognized errors when situation allows.

An approach where all the data is cleaned before using it in analytics does not work in big data [LaValle et al 11]. It leaves too little time, energy and resources for understanding the potential use of the data. Timeliness would be affected if huge data sets are corrected. Because the processing of big data is resource intensive, the focus of big data is usually more realistic than the optimality focus of traditional research methods. It is understood that big data is messy and analytical methods try to cope with the messiness. In fact, the focus of testing is moved from fixing errors towards process-oriented validation, root cause analysis and remediation [Loshin 13]. Traditional data quality tools are used for fixing data. New data quality tools are used for ensuring that the data is valid or correct. Also, the responsibility about the quality of the big data values and their semantics and interpretation is to the data consumer [Loshin 13, Ch. 9]. Small data is cleaned in the data warehouse and data consumer expects to have correct data.

Resource limitations and cleaning all the big data before analytics is not possible because of repurposing the data sets. Small data is designed to answer a specific

question or to serve a particular goal whereas big data analytics is designed with a flexible goal in mind. The aim is to ensure that datasets are fit for the purposes they were originally intended [Loshin 13]. The problem in big data is that the data is used in a way that was not originally intended.

The potential root cause of errors can be found and corrected in internal data sets but there may be little control over the quality of original data sources. Big data analysts have a little control over who created the data sets outside the organization and there may be a lack of an oversight over data creation. Therefore there are limited opportunities to engage process owners to influence modifications or corrections to the source [Loshin 13]. Correcting the data may not be possible because correct outcomes are not known.

Big data analysis has several iterations of asking questions, trying to answer questions and then verifying the results [Gleason et McCallum 12]. The data is first used trying to answer the question. After the results are verified one knows whether the data was sufficient. Initial rounds of analysis should be treated as checks of completeness. Any findings should be treated as preliminary. It is not enough to assume that the data that was received is the data that was needed. Preliminary data analysis helps to identify the most relevant variables and to determine the complexity of the underlying problem. Often the most relevant information and questions are found in the end of a big data project. In the Blue Yonder example the analyst was interested in what people were saying and the reason why people were dissatisfied.

If one decides to change the data to a more acceptable form, it can be done with enhancing the data with identity profiles. Changing the data may mean linking extracted entities to known identity profiles in the context of big data [Loshin 13]. Linked entities share profile information which enhances the analysis. Data can be linked with metadata hierarchies and taxonomies. This helps to treat cars, automobiles, vans, minivans, SUVs, trucks, and RVs as vehicles. Again the consumers expectations need to be managed.

6 Big Data Testing in Hadoop Environment

A Hadoop-based environment is one architectural solution to big data systems. This chapter describes the testing of a Hadoop-based architecture. Subsection 6.1 describes big data testing areas and subsection 6.2 describes a few testing limitations and solutions.

6.1 Big Data Testing Areas

Figure 7 in chapter 4 presents four big data testing areas: 1) pre-Hadoop process validation which refers to loading source data files into HDFS, 2) map-reduce process validation, 3) ETL-process validation which makes sure that the output results from HDFS are extracted correctly and 4) reports testing. Testing areas are based on the example architecture in the figure 7. This section is mainly based on the Gudipati et al. 13 article.

1) *Pre-Hadoop process validation*

Some of the issues in extracting data from the source system are incorrect data, incorrect storage of data, incomplete or incorrect replication. Therefore, it is important to ensure that the data is extracted correctly, to ensure that the data is the right data and files are loaded into HDFS correctly. The pre-Hadoop process validation ensures that the input files are split, moved and replicated in different data nodes. Ensuring that the data is extracted correctly can be done by comparing the input data file against the source system data. Validating the data requirements ensures that the right data is extracted. Gudipati suggests that compare tools are used for extracting the differences between source data and files that are loaded into HDFS. A complete comparison will take a lot of time. To reduce comparison time, data can be sampled so that most of the scenarios are covered. Also, the comparison scripts can be run in parallel on multiple nodes. Comparison does not ensure that the data is correct for the task at hand.

2) *Map-Reduce process validation*

Issues during map-reduce jobs may be jobs that are working correctly when run in a standalone node, but working incorrectly when run on multiple nodes. Other issues during map-reduce job are incorrect aggregations, node configurations, and incorrect output format. Tests confirm that the data processing is completed and the output file is

generated in map-reduce process validation. The business logic should work on a standalone node and after running against multiple nodes. Tests confirm that key value pairs are generated correctly in the map-reduce process and validate the aggregation and consolidation of data after reduce process. The output data needs to be validated against the source files to ensure that the data processing is completed correctly. The output file can be validated to ensure that the format is per the requirement.

3) *ETL into data warehouse process validation*

After data output files are generated the processed data is loaded with ETL procedures from Hadoop into data warehouse. Incorrectly applied transformation rules, an incorrect load of HDFS files into the data warehouse and an incomplete data extraction from the HDFS are issues in this phase. It is important to ensure the functional correctness in ETL into data warehouse process validation. The data warehouse needs to be correctly populated with the data. Comparing the target table data against HDFS files data validates that there is no data corruption. Performance, reliability, maintainability, freshness, scalability, availability, flexibility, robustness, affordability, audibility, and traceability are other quality objectives in ETL [Dayal et al. 09]. Not all of these quality objectives are possible to implement because there is a tradeoff between quality objectives and business needs. For example, a tradeoff between information accuracy and performance may be important in the information extraction phase.

4) *Reports testing*

Reports testing includes report data issues, layout and format issues. Verifying if the data is extracted correctly to the reports is done by queries. Queries are written to verify that the right data is used in reports. The data in reports are tested against databases. This requires strong query language skills.

Non-functional testing tests the way a system operates and may not be related to user action. Non-functional testing like performance and fail-over testing is done through the big data architecture stack. The HDFS architecture is designed to detect failures like name node failures, data node failure, network failure and automatically recover to proceed with the processing [Gudipati et al. 13]. Hadoop detects and handles failures at the data application layer.

Pre-Hadoop, map-reduce and ETL process validations make sure that the data is

extracted correctly. Most of the tests recommended in the data acquisition phase are system triggered scenarios. There are similarities between big data testing and data warehouse testing. Both the big data architecture and the data warehouse have ETL-process validation and front end testing, e.g. testing reports. Items to be tested in data warehouses are multidimensional schema, ETL procedures, physical schema and front end [Gupta et al. 12].

Test plan activities are designed based on business understanding, in fact all the data warehouse tests are focused on the business logic and data content [Gupta et al. 12]. Because testing the data warehouse is directed at data and information, knowing the data and the answers to user queries are the key to data warehouse testing [Golfarelli et Rizzi 09]. It is important to gather requirements to test the data warehouse and the test plan is created on the basis of the requirements [Gupta et al. 12]. However, big data projects seldom have precise requirements for the project. The value of the data is discovered along with analysis projects.

Big data processing requires data scalability. Big data processing requirements are an ability to support the processing of petabytes of data and an ability to process geographically disperse and potentially heterogeneous distributed data across thousands of processors are some big data processing requirements [Kimbal 13]. The data is loaded at very high rates (gigabytes per second) to be ready for analysis. Other processing requirements are sub-second response time for highly constrained standard SQL queries, embedding arbitrarily complex user-defined functions in processing requests, an ability to implement user-defined functions in a wide variety of industry-standard procedural languages. User-defined functions need to be able to be executed as relation scans over petabyte-sized data sets in a few minutes.

Test infrastructure requirements consist of big data processing requirements and of the number of data nodes in quality assurance environment. Data privacy requirements need to be understood in order to evaluate private or public cloud. It should be noted that privacy issues need special attention because data privacy cannot be recovered once compromised. Software inventory is required to evaluate which softwares need to be setup on test environment (Hadoop, File system to be used, No SQL DBs, etc.). Cloud can offer flexibility that is needed to overcome challenges in data variety, velocity and

volume. Setting up a test environment on cloud will give the flexibility to setup and maintain the environment during test execution [Gudipati et al. 13]. After the big data test infrastructure requirements are assessed and designed, one can implement and maintain the test infrastructure.

Big data test infrastructure design consists of documenting the high level cloud test infrastructure design (Disk space, RAM required for each node, etc.). Cloud infrastructure service provider is identified and service level agreements (SLA), communication plan, maintenance plan, environment refresh plan and the data security plan need to be documented. Big data test infrastructure implementation and maintenance consist of creating a cloud instance of the big data test environment and installing Hadoop, HDFS, MapReduce and other software as per the infrastructure design. Smoke tests are performed on a sample map reduce, Pig/Hive jobs.

6.2 Testing Limitations and Solutions

RTTS (Real-Time Technology Solutions) made a survey which revealed that 60% of organizations executed data quality tests manually in 2013 [Hayduk 13]. Manual testing refers to comparing data sets extracted from databases and data warehouses by eye. In addition to a manual inspection of the data, analysis programs should be used to gain metadata about data properties and to detect data quality problems. When big data is used, there is even more reason to automatize testing routines. There is a need for automated testing routines but the level of automation may be small because of the variety of the data.

The speed of data needs to be considered when performance problems need to be overcome [Gudipati et al. 13]. Problems in velocity can be overcome with good performance testing. Performance testing identifies bottlenecks in the system. After bottlenecks are identified and corrected the system can handle high velocity streaming data. A Hadoop performance monitoring tool can capture the performance metrics like job completion time and throughput. System level metrics like memory utilization are part of performance testing.

Testing unstructured data is very time consuming and complex [Gudipati et al. 13]. Variety of data resources can be validated after the data is transformed into a structured

format by using custom build scripts. The first step is to transform the data into the structured format. Unstructured data can be transformed into the structured format by using a scripting language like PIG. Semi-structured data can be transformed if there are identified patterns. A pattern outline can be used to convert the incoming data into a structured format. After conversion the validations can be performed by using compare tools. The level of automating the structure conversion is low because the input data can change every time a new test is performed.

If data volumes are large, then a good way is to sample the data for tests. The data classification is one solution for enhancing the processing of huge volumes of data. However, the challenge is to locate meaningful data and to decide if the data is relevant or an anomaly. Anomalies may be interesting since they can point out broken sensor for example. Understanding business rules, company and government regulations which are developed in specific application context, is important in assessing the task-dependent quality attributes. Discussion with domain experts may help to recognize common types of data errors that are typical within particular domain.

Selecting a subset of representative cases and using dimension reduction techniques might improve classification efficiency [Berman 13]. However, the selection bias may lead to wrong conclusions about the data. To make the representative cases as good as possible a new set of representative cases can be selected from the current representative cases [Berman 13, Ch. 10]. If subsets are used a random sampling is recommended or the use of all available data. Increasing the accuracy of an analytical model may lead to overfitting the model. In fact, the bigger the data set, the easier it is to overfit the model. Overfitting is discovered by testing the predictor or model on one or several new sets of data. If the data is overfitted, the model does not work well with other data sets. The overfitted model describes the data well but does not predict the behavior of other data sets. Classification algorithms seek models that attain the highest accuracy or the lowest error rate when applied to the test set [Tan et al. 06, p. 149]. However, null error rate is not necessarily the best one. Some classifiers have higher than null error rate and still be better for a particular application.

7 Conclusions

Traditional data is used for reporting what has happened and why. Big data supplements traditional reporting by making predictions and optimizing business processes. Business decisions are made based on optimization and predictions that recommend near-optimal decisions.

Big data projects are expensive and seldom repeatable. If conclusions drawn from the data are questioned, there are limited possibilities to verify results. There is no control over the quality and validity of the big data outside the organization. There are limited opportunities to make modifications or corrections if external data sets are used. The focus of big data is usually more realistic than the optimality focus of traditional research methods. The focus of testing is moved from correcting errors towards process oriented validation, root cause analysis and remediation. Traditional data quality tools correct the data, new data quality tools ensure that the data is valid or correct. Because the data is not corrected, it is in data consumers responsibility to understand the quality of the big data values and their semantics and interpretation.

Because of diversity of data quality problems, there is no singular approach to address the data quality. High level steps in the data governance includes understanding the value of data within an organization, instituting the the right levels of control, identifying and prioritizing data issues and correcting data. Understanding the value of data requires dialog with data users about requirements. Data quality involves consistently meeting business analytics and end customer expectations. There might be a different opinions and definitions on data quality between technically oriented persons and business stakeholders. Since there are many big data use cases, big data quality assessment is usually task dependent.

It is important to be aware of the data and its limitations and biases. Knowing the data includes knowing which questions can be asked and what interpretations are appropriate. This is difficult because data sets are used multiple times. Reusing datasets fades away inherent semantics associated with the original datasets and limits visibility into data creation. Though, reusing datasets makes the data easily available. Big data loses its meaning if taken out of its context. Describing data lineage back to its origin increases the visibility into data creation. The source, actors and participants of big data

need to be defined consistently in order to verify analytical results. Another important quality attribute is immutability of the data. Original data sources are updated inconsistently and even if the original data source is tracked, it may be very different compared to the time the data was crawled or processed. The lack of metadata leads to ambiguity of big data. The quality of being open to more than one interpretation complicates interpretability.

The relevance of quality attributes and quality challenges are depended on the context. Prioritizing data issues gains significance, since equilibrium between quality and production cost is vital for organizations. It is usually not possible to perform all the test routines. Testing approaches should take resource limitations and dependencies between test routines into consideration.

Different decisions made along the analysis pipeline are subjective. Subjective work with big data may not lead to closer claim on objective truth. Interpretations about the data set are depended on the question asked, context and skills of the analyst. Analysis pipeline requires metadata generation to describe what data is recorded and how it is recorded and measured. Automated expression of differences in data structure and semantics requires effective large scale analysis. Locating, identifying, understanding and citing the data has to happen in a automated manner.

Data analysis and modeling phase requires accuracy of the model since it affects the meaningfulness of the solution. Data interpretation includes verifying and understanding the results produced by an analytical model. Verification includes an ability to trace the analysis. Visualization helps to interpret the analysis.

The task independent quality assessment is related to data loading from sources to the big data architecture. Testing areas in the Hadoop environment ensure that data is extracted from sources to the Hadoop system correctly. Validation areas in the Hadoop environment consists of extracting data form the source to the Hadoop system, of the map-reduce process, of the ETL into enterprise data warehouse process, and of reports testing. The level of automation is good since the Hadoop system has a built-in ability to self-report when errors occur. The HDFS architecture is designed to detect failures like name node failures, data node failure, network failure and to automatically recover to proceed with the processing. Copying the data into data blocks allows better failure

recovery and availability of the data in the Hadoop system. Redundancy offers data locality which is critical when working with large data sets. A backup process for the NameNode ensures accessibility and availability of the data.

Big data quality issues includes all the inadequacies found in the traditional data resource plus huge variations in the quality of the records. Challenges and requirements are summarized through three dimensions of big data: volume, variety and velocity.

Challenges and requirements in data **volume**: Processing volumes of data can lead to major delays in runtime and responsiveness. Even simple operations may be time consuming. Finding the relevant and meaningful information is difficult since most of the data may not be relevant to the task at hand. A challenge of big data is to make a distinction between the complete data set and the representative data set. Huge data sets collected from Twitter may not be representative data sets, even though the whole data is loaded. Also, large data set does not imply to accurate data. In some cases, the bigger the data set is, the more accurate classifications can be made. Large data sets enable better observation of rare but important events.

Huge volumes of data may lead to focusing exclusively on finding patterns or correlations without understanding of the deeper dynamics at play. Non-representative samples can provide internally valid conclusions that cannot be generalized to a different setting. Biased and unrepresentative samples are avoided by using random sampling. Data is not always additive and conclusions cannot be drawn based on subset comparison. Processing large data sets require scalability and performance. Data is usually filtered to produce smaller data sets for analysis. Data usage requires finding relevant and meaningful information, understanding the value of the data and understanding the context and question asked.

Challenges in data **variety**: Different data types are distinguished into structured data, semi-structured data, unstructured data. Unstructured data represents real data in day-to-day life and it is expressed in natural language with no specific structure or domain defined. Human generated unstructured data is filled with nuances, variation, and double meanings. One should be careful in interpreting the content of human generated unstructured data. Semantic inconsistencies complicate analysis. Metadata can improve consistency by joining a glossary of business terms, hierarchies and taxonomies for business concepts. Big data interpretability can be improved with introspection which

refers to an ability of a data object to describe itself when called upon. Finding the true intent from slang and sarcasm and the true significance of the statement may be hard to notice. For example, there is a difference if a person is loosing “a” job versus loosing one’s only job. Reliability and validity of the unstructured user generated data may be difficult to notice.

Variety of available data sources has increased. In the same time analysts should be careful of the ethicality of using the data. The different accessing levels based on who tries to access the data complicates evaluation of methodological claims. Methodological claims cannot be reproduced nor evaluated if there is no access available to the data. Different privacy levels affect on the availability of the data.

Big data is inherently messy, especially internet sources are unreliable. Variations are dominant if data sets are related to human behavior and standard methods may not apply. For example, statistical measures like averages becomes meaningless in sparsely populated data sets. Messiness of big data makes it difficult to understand the properties and limits of a dataset, regardless of its size.

Processing data variety requires converting unstructured and semi-structured data into structured format so that they can be stored, accessed and analyzed along with other structured data. Data usage requires understanding the nuances, variations and double meanings in human generated unstructured data. Other requirements are ethicality of using data sets and privacy preserving analysis.

Challenges in data **velocity**: Big data comes in with continuos streams, which enables more fine-grained customer segmentation based on day-to-day situation rather than segmentation based on historical data. The question when the data is no longer relevant to the current analysis is more valid in real-time data. Velocity related quality attribute is how quickly data is shared in a people-to-people network.

The data is used immediately after it flows into the system. Processing data velocity requires on-demand and real-time accessibility compared to the traditional on-supply and over-time access. Data usage requires faster decision-making and faster reaction time in business.

Table 3 concludes downsides and benefits that volume, variety and velocity bring to big data analytics from data quality point of view.

	Con's (-)	**Pro's (+)**	**Techniques/methods that handle challenges in variety, volume and velocity.**
Volume (relative concept)	Delays on runtime and responsiveness.	Messy source data which is still good enough to support informed decisions.	Compare tools, comparison scripts.
	Difficult to find relevant & meaningful information.		Data (or comparison scripts) can be run in parallel on multiple nodes.
	Difficulties to understand the data (focus on finding patterns and correlations leading to possible misinterpretations)		Data sampling so that most of scenarios are covered (right-sizing the data). Visualization, data integration and data classification.
Variety	Unstructured data is naturally messy.		Transform data into custom build scripts: pattern identification for semi-structured data, language like PIG for unstructured data.
	Integration challenges: - missing information because of privacy issues - different statistical confidence levels - sample collecting frequency - semantic correctness	Helps to create added value to the products and services	Data conversion into structured format.
	Creating ETL procedures is expensive and slow (when structured and unstructured data is integrated)		Unstructured data validation based on business scenarios.
	Unstructured data is not intended to be analyzed by computers.		
	Dimensionality increases distances between data objects and the multidimensional space becomes sparsely populated.		Dimension reduction techniques.
	The number of different possible errors increases.		
	New data sources in social media pose specific challenges in misinterpreting correlations to causal relationships.	Human generated data, e.g. Tweets enable sentimental analysis.	
	If the data is available, it may not be ethical to use it.		
Velocity	Finding relevant information.	Faster reaction time in business.	Hadoop performance monitoring tool: Performance testing to identify bottlenecks in the system, e.g. job completion time and throughput.
		Fain-grained customer segmentation based on day-to day situation vs. segmentation based on historical data.	

References

Agrawal et al. 11

Agrawal, D., et al., Challenges and opportunities with big data, A community white paper developed by leading researchers across the United States, Tech. Rep., 2011.

Apache Hadoop

http://hadoop.apache.org.

Barateiro et Galhardas 05

Barateiro, J., Galhardas, H., A Survey of Data Quality Tools, Datenbank-Spektrum 14, no. (15-21), 48, 2005.

Batini et Scannapieco 06

Batini, C., Scannapieco, M., Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), Springer-Verlag New York, Inc., Secaucus, NJ, USA. 2006.

Berman 13

Berman, J.J., Principles of big data: preparing, sharing, and analyzing complex information, Elsevier, Morgan Kaufmann, Amsterdam, 2013.

Bice 13

http://channel9.msdn.com/Events/TechEd/NorthAmerica/2013/FDN01#fbi , cited 7.11.2014, speaker Shawn Bice June 3-6, New Orleans, LA, 2013.

Bickel et al. 75

Bickel, P. J., Hammel, E. A., O'Connell, J.W., Sex bias in graduate admissions: Data from Berkeley. Science, 187, no. 4175, p. 398-404, 1975.

Bizer et al. 12

Bizer, C., Boncz, P., Brodie, M. L., Erling, O., The meaningful use of big data: four perspectives -- four challenges, ACM SIGMOD Record, 40, no. 4, January, p. 56-60, 2012.

Boyd et Crawford 12

Boyd, D., Crawford, K., Critical Questions for Big Data, Information, Communication and Society, 15, no. 5, p. 662–679, 2012.

Cuzzocrea et al. 13

Cuzzocrea, A., Saccà, D., Ullman, J. D., Big data: a research agenda. In Proceedings of the 17th International Database Engineering & Applications Symposium (IDEAS '13), ACM, New York, NY, USA, p. 198-203, 2013.

Dayal et al. 09

Dayal, U., Castellanos, M., Simitsis, A., Wilkinson, K., Data integration flows for business intelligence, Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, ACM, 2009.

Devlin et al. 12

Devlin, B., Rogers, S., Myers, J., Big data comes of age, Tech. Rep., November, 2012.

Deutsch 12

Deutsch, T., Big Data: Data Quality's Best Friend? Part 1, 8, 2012, http://ibmdatamag.com/2012/08/big-data-data-qualitys-best-friend/ [19.10.2014].

Draper 12

Draper, R., Data Traceability, p. 205 – 212 in McCallum, Q. E., editor. Bad Data Handbook, O'Reilly Media, Inc. USA, 2012.

Eckerson 02

Eckerson, W., Data Quality and the Bottom Line – Achieving Business Success through a Commitment to High Quality Data, The Data Warehousing Institute (TDWI) Report Series, 2002.

ElGamal et al. 11

ElGamal, N., El Bastawissy, A., Galal-Edee, G., Towards a data warehouse testing framework, ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2011 9th International Conference on, IEEE, 2012.

Fink 12

Fink, K., Is It Just Me, or Does This Data Smell Funny?, p. 5 – 29 in McCallum, Q. E., editor, Bad Data Handbook, O'Reilly Media, Inc. USA, 2012.

Fruchterman 11

Fruchterman, J. Issues with Crowdsourced Data Part 2., Beneblog: Technology Meets Society. 28 Mar. 2011.

<http://benetech.blogspot.com/2011/03/issues-with-crowdsourced-data-part-2.html>. Ball, Patrick, Jeff Klingner, and Kristian Lum. “Crowdsourcing Data is Not a Substitute for Real Statistics.” Beneblog. 17 Mar. 2011. <http://benetech.blogspot.com/2011/03/crowdsourced-data-is-not-substitute-for.html> [5.12.2014]

Gleason et McCallum 12

Gleason, K., McCallum, Q. E., Data Quality Analysis Demystified: Knowing When Your Data Is Good Enough, p. 225 - 238 in McCallum, Q. E., editor, Bad Data Handbook, O’Reilly Media, Inc. USA, 2012.

Goetz et al. 13

Goetz, M., Owens, L., Jedinak, E., Build Trusted Data with Data Quality, Forrester Research, Inc. August, 2013.

Golfarelli et Rizzi 11

Golfarelli, M., Rizzi, S., Data warehouse testing: A prototype-based metholdology, on Information and Software Technology Journal, vol. 53, p. 1183-1198, 2011.

Gupta et al. 12

Gupta, S.L., Pahwa, P., Mathur, S., Classification of Data Warehouse Testing Approaches, International Journal of Computer & Technology, 3, no. 3, p. 381-386, December, 2012.

Gudipati et al. 13

Gudipati, M., Rao, S., Mohan, N. D., Gajja, N. K., Big Data: Testing Approach to Overcome Quality Challenges, Infosys Labs Briefings, 11, no. 1, 2013.

Hayduk 13

Hayduk, B., Enterprise Business Intelligence & Data Warehousing: The Data Quality Conundrum, A Study By RTTS (Real-Time Technology Solutions), vol. 3, July, 2013.

Janert 12

Janert, P. K., Will the Bad Data Please Stand Up? p. 95 – 106 in McCallum,Q. E., editor, Bad Data Handbook, O’Reilly Media, Inc. USA,

2012. Solutions), vol. 3, July, 2013.

Kimball 12

Kimball, R., Newly Emerging Best Practices for Big Data, Kimball Consulting Group White Paper, September, 2012.

Kimball 13

Kimball, R., Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling (3rd Edition), John Wiley & Sons, Somerset, NJ, USA, Ch 21, July, 2013.

Krishnan 13

Krishnan, K., Data warehousing in the age of big data, Morgan Kaufmann is an imprint of Elsevier, USA, 2013.

Laney 01

Laney, D., 3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note, February 6, 2001.

LaValle et al. 11

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., Kruschwitz, N., Big data, analytics and the path from insights to value, MIT Sloan Management Review, 52, no. 2, p. 21-32, Winter, 2011.

Letouzé 12

Letouzé, E., Big Data for Development: Opportunities & Challenges, Global Pulse, May, 2012.

Loshin 13

Loshin, D., Big data analytics : from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph, Morgan Kaufmann, 2013.

Madnick et al. 09

Madnick, S. E., Wang, R. Y., Lee, Y. W., Zhu, H., Overview and Framework for Data and Information Quality Research, Data and Information Quality 1, 1, Article 2, June ,2009.

Madsen 13

Madsen, M., The Challenges of Big Data & Approaches to Data Quality: Using big data to examine and discover the value in data for accurate analytics, Technology White paper, Third Nature Inc. and SAP AG, 2013.

Michalewicz et al. 07

Michalewicz, Z., Schmidt, M., Michalewicz, M., Chiriac, C., Adaptive business intelligence, Springer, Berlin Heidelberg, 2007.

Normandeau 13

Normandeau, K, http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ September 12, 2013 [cited 12.1.2015].

Parker 12

Parker, C., Unexpected challenges in large scale machine learning, In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '12), ACM, New York, NY, USA, 1-6, 2012.

Pipino et al. 02

Pipino, L. L., Lee, Y. W., Wang, R. Y, Data quality assessment, Communications of the ACM, 45, no. 4, p. 211-218, 2002.

Russom 11

Russom, P., Big Data Analytics, TDWI Best Practices Report, co-sponsored by IBM, fourth quarter, 2011.

Strong et al. 97

Strong, D. M., Yang, W. L.,Wang, R. Y., Data quality in context, Communications of the ACM, 40, no. 5, pp. 103-110, 1997.

Tan et al. 06

Tan, P-N., Steinbach, M., Kumar, V., Introduction to data mining, Boston : Pearson Addison Wesley, cop. 2006.

Tien 13

Tien, J. M., Big data: unleashing information, Journal of Systems Science and Systems Engineering, 22, no. 2, p. 127-151, June, 2013.

Vassiliadis 00

Vassiliadis, P,. Data warehouse modeling and quality issues, Ph.D Thesis, National Technical University of Athens Zographou, Athens, GREECE, 2000.

Woodall et al. 13

Woodall, P., Borek, A., Parlikad, A. K., Data quality assessment: The Hybrid

Approach, Information & Management, 50, no. 7, p. 369-382, November, 2013.

Zikopoulos et al. 11

Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., Lapis, G., IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Companies,Incorporated, 2011.