UNIVERSIDAD DE VALLADOLID

MÁSTER UNIVERSITARIO

Ingeniería Informática

TRABAJO FIN DE MÁSTER

Adapting a Quality Model for a Big Data Application: The Case of a Feature Prediction System

Realizado por OSBEL MONTERO PEREZ

Universidad de Valladolid

9 de julio de 2021

Tutor: YANIA CRESPO GONZÁLEZ-CARVAJAL

Universidad de Valladolid

Máster universitario en Ingeniería Informática

YANIA CRESPO GONZÁLES-CARVAJAL, profesora del DEPARTAMENTO DE INFORMÁTICA de la UNIVERSIDAD DE VALLADOLID.

Expone:

Que el alumno OSBEL MONTERO PEREZ, ha realizado el Trabajo final de Máster en Ingeniería Informática titulado "ADAPTING A QUALITY MODEL FOR A BIG DATA APPLICATION: THE CASE OF A FEATURE PREDICTION SYSTEM".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección de quien suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Valladolid, 9 de julio de 2021

Vº. Bº. del Tutor:

D. nombre tutor: Yania Crespo González-Carvajal

Resumen

En la última década hemos sido testigos del considerable incremento de proyectos basados en aplicaciones de Big Data. Algunos de los tipos más populares de esas aplicaciones han sido: los sistemas de recomendaciones, la predicción de características y la toma de decisiones. En este nuevo auge han surgido propuestas de implementación de modelos de calidad para las aplicaciones de Big data que por su gran heterogeneidad se hace difícil la selección del modelo de calidad ideal para el desarrollo de un tipo específico de aplicación de Big Data.

En el presente Trabajo de Fin de Máster se realiza un estudio de mapeo sistemático (SMS, por sus siglas en inglés) que parte de dos preguntas clave de investigación. La primera trata sobre cuál es el estado en la identificación de riesgos, problemas o desafíos en las aplicaciones de Big Data. La segunda, trata sobre qué modelos de calidad se han aplicado hasta la fecha a las aplicaciones de Big Data, específicamente a los sistemas de predicción de características. El objetivo final es analizar los modelos de calidad disponibles y adaptar un modelo de calidad a partir de los existentes que se puedan aplicar a un tipo específico de aplicación de Big Data: los sistemas de predicción de características. El modelo definido comprende un conjunto de características de calidad definidas como parte del modelo y métricas de calidad para evaluarlas.

Finalmente, se realiza una aproximación a un caso de estudio donde se aplica el modelo y se evalúan las características de calidad definidas a través de sus métricas de calidad presentándose los resultados obtenidos.

Descriptores

Big Data, Modelos de Calidad, Sistemas de Predicción, Características de Calidad, Métricas de Calidad.

Abstract

In the last decade, we have been witnesses of the considerable increment of projects based on big data applications. Some of the most popular types of those applications have been: Recommendations, Feature Predictions, and Decision making. In this new context, several proposals have arisen for the implementation of quality models applied to Big Data applications.

As part of the current Master thesis, a Systematic Mapping Study (SMS) is conducted which starts from two key research questions. The first one is about what is the state of the art about the identification of risks, issues, problems, or challenges in big data applications. The second one, is about which quality models have been applied up to date to big data applications, specifically to feature prediction systems. The main objective is to analyze the available quality models and adapt a quality model from the existing ones that can be applied to a specific type of Big Data application: The Feature Prediction Systems. The defined model comprises a set of quality characteristics defined as part of the model and a set of quality metrics to evaluate them.

Finally, an approach is made to a case study where the model is applied, and the quality characteristics defined through its quality metrics are evaluated. The results are presented and discussed.

Keywords

Big Data, Quality Models, Feature Prediction Systems, Quality Characteristics, Quality Metrics.

Chapter Index

Tables Index

Figures Index

1. Introduction

The International Data Corporation (IDC) revealed that by the end of year 2020 the total amount of ecommerce transactions would be approximately 450 billion per day over the internet. The volume of data generated by every human being would be 1.7 megabytes per second and the total data volume generated might be doubled every 2 years (Indrakumari Ranganathan, 2020) creating data sets that are beyond the human ability to handle them manually because the extreme size in the order of terabytes, petabytes, exabytes, zettabytes, yottabytes or brontobytes as our coming digital universe (K. Radha, 2015).

At present, it is very common the use of Big Data term, defined by Tim O'Reilly in 2005 as we know today, and that has become part of the Oxford dictionary by the year of 2013 (Abdallah, 2019), where it is expressed that:

> "Big data are a set of information that are too large or too complex to handle, analyze or use with standard methods" (Oxford Learner's Dictionaries, n.d.)

However, according to other authors it has been introduced for the first time by the Gartner Group. The definition presented as follows has a greater scope:

> "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." (Gartner IT Glossary, n.d.)

The term Big Data describes the massive set of data collected from business or systems in a day-to-day basis from multiple and such diverse sources like: social media activities, telecommunications systems, financial and business transactions, surveillance and monitoring devices, sensor systems, and Internet of Things (IoT) networks. Those sources at the same time are generating data massively with a high volume in a fast way that have never seen before and with a clear tendency to increase.

The description above is resumed in the three main V's regarding big data issues or challenges, called: Volume, Velocity and Variety. This approach is widely used by practitioners in technical literature (K. Radha, 2015), (Laranjeiro, Soydemir, & Bernardino, 2015), (Auer & Felderer, 2019).

However, some authors describe other two characteristics which also categorize Big Data: Value, as the actionable information extracted from the data and Veracity, where the trustworthiness and provenance of the data is observed (Rao, Gudivada, & Raghavan, 2015) and (Patel, 2019).

Those five Vs are identified by most researchers as the main big issues or challenges for data quality management in big data applications. At this point, it is necessary to clarify that several researchers are investigating about similar or related topics in: big data analytics, big data systems or big data applications. In the present master thesis, the big data applications are the focus of the current investigation and which quality models and metrics have been proposed for those big data applications.

From the premise that data is worthless in the vacuum, the Fig. 1 shows elements that complete a big data application: **techniques** and algorithms for analyzing data (machine learning, natural language processing), **technologies** for data access and collect (databases, tables, non-structured formats) and data processing or creating (business intelligence, cloud computing or Internet of Things), and **visualization** using charts, graphs, tables, and others. The techniques should be carefully selected regarding the data being analyzed.

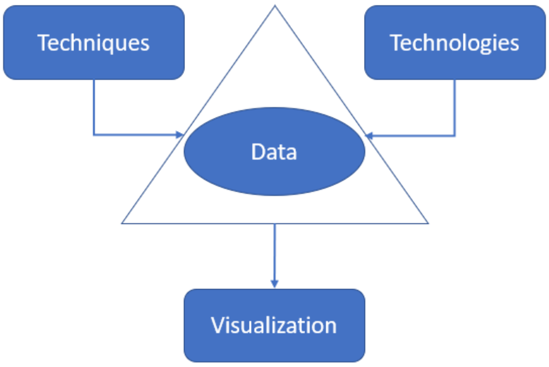


Fig. 1 High level elements that complete a big data application.

Currently, different kinds of Big Data applications can be identified such as Recommendations, Feature Prediction, and Decision Making (Zhang, Zhou, Li, & Gao, 2017). The real-life domains of big data applications include smart cities, smart carts, healthcare systems, financial, business intelligence, environmental control and so on.

Although several researchers have detailed the advantages obtained from Big Data applications, there is a reasonable number of recent publications claiming for the definition of good quality models applied to those Big Data applications. For example, the lack of research on the adequate test modeling and coverage analysis for big data applications, and the clear practitioners demand for having stablished a well-defined test coverage criteria is presented as an issue in (Tao & Gao, 2016).

(Wani & Jabin, 2018) have presented some challenges for big data such as a lack of big data professionals to handle the available tools and algorithms for big data processing, and the need to conduct a rigorous effort by researchers to deal with new challenges arising in both hardware and software.

Recommendation and feature prediction systems are successful examples of big data applications, however there is a clear demand to apply a systematic research in data quality issues presented in these systems (Pengcheng Zhang, 2017) and resolve one of the biggest challenges for these big data applications: automated methods for resolving data quality issues, since an important manual cleansing work for input data is required in data analysis.

According to ISO/IEC 25024 (ISO/IEC25024, 2015), data quality is the degree to which the characteristics of data satisfy stated and implied needs when used under specific conditions.

According to (Garvin, 1996) , quality comprises eight dimensions that are related to three main components: product quality, service quality and people-based quality.

- The performance related to product quality.
- Features related to the characteristics that complement the basic functioning of product and services.
- Reliability, as the probability of a product to malfunctioning or failing within a period of time.
- Conformance, as the degree to which a product's design and characteristics meet stablished standards.
- Durability, as a measure of product life.
- Serviceability, concerning about the time before a service is restored in a not planned break down, and other related features.
- Perceived quality, where the perception about the quality of a product or service is people-based and may vary from one person to another.

In this investigation the focus is on product quality and specifically those aspects related with data quality. The object of study are the data quality models applied in the context of Big Data applications.

It has been identified from a systematic mapping study that how to effectively ensure the quality of big data applications is a hot research issue. In this sense, the main challenge is how to deal with the big data 5 Vs, and also with functional or non-functional factors/dimensions to ensure quality expected on big data applications. On the other side is which quality metrics should be applied to measure some quality characteristics as part of Feature Prediction Systems.

1.1 Motivation

The importance and relevance that Big Data is acquiring these days and the promising future we can expect on this knowledge area has been discussed widely. Many researchers have been investigating about quality assurance techniques and quality models applied to big data applications. Although there has been conducted scientific research aimed at understanding, defining, classifying, and communicating quality assurance methods for big data applications, there is not enough clarity about what characteristics and sub-characteristics should be part of a quality model for a specific kind of big data application such as Feature Prediction Systems. Should be interesting to find if there is a difference validating the different kinds of big data applications by applying a specific quality model with specific quality characteristics and quality metrics associated.

To predict future behavior, feature prediction systems use a statistical technique which works by creating a model from analyzing historical and current data to predict future outcomes related to customers and products while identify potential opportunities and risks. If the prediction is properly made some good advantages are obtained by companies and organizations (Bejarano, 2018).

1.2 Objectives

Once the scope and the motivation of this research are set, the following main objective is defined:

To analyze the available data quality models in the context of Big Data applications and adapt a quality model from the existing ones which can be applied to specific Big Data application such as Feature Prediction Systems.

To support this main objective and to guide the investigation, the following sub-objectives are defined:

- To conduct a systematic mapping study from accredited sources collecting information about which are the main Big Data issues and challenges detected in Big Data applications.
- Review and analyze the available quality models for Big Data applications, select the main characteristics and sub-characteristics including quality metrics, that best adjust to evaluate the quality of Feature Prediction Systems as a major kind of Big Data applications, and obtain an adjusted quality model.
- To apply the adjusted quality model to a real Big Data application, specifically to a case of Feature Prediction System.

1.3 Research questions

In this sense, to clarify more about this topic some research questions have been identified:

- Which quality models have been applied to evaluate the quality of big data applications?
- Which metrics have been defined as indicators?
- Are there different approaches to implement or adjust a quality model to a specific kind of big data application?
- Is it possible to adjust a quality model to a specific kind of big data application?

1.4 Document structure

This work is presented as follows. An introduction to the theme with definition of scope, motivation, objectives, and the research questions is presented here in Chapter 1. In Chapter 2, a background related with the key topics where some concepts and findings are revealed for a better understanding is described. Chapter 3 which presents a Systematic Mapping Study as the methodology followed to conduct the review of the state of the art, results and partial conclusions are also presented. The Chapter 4 discuss the peculiarities of Data Quality Models, presenting the quality characteristics and quality metrics as indicators. In this chapter is also presented a proposed and adjusted quality model adapted to a specific kind of Big Data application: the feature prediction systems. In Chapter 5 are presented the results of quality assessment conducted in an experiment where the quality model is applied in a real example. Finally, the conclusions and future work are presented in Chapter 6.

2. Background

2.1 Introduction to Big Data

Before starting a deep analysis on key topics is useful to clarify some concepts related to them.

Big data involves the interpretation of large datasets that due to its size and structure the capabilities of traditional programming tools are exceeded for collecting, storing, and processing data in a reasonable time. In the present TFM is considered that Big Data has five main characteristics which are: Volume, Velocity, Variety, Value and Veracity as presented in Fig. 2. In some places this is known as a common framework (Laranjeiro, Soydemir, & Bernardino, 2015) and (Kushal Patel, 2017).

- **Volume**: refers to the size of the dataset and is probable the first characteristic associated with big data. The data volume changes over the time with new technological developments and the overall increase in data size.
- **Velocity**: refers to the speed that data is generated, processed, analyzed, and presented for a Big Data application. Some businesses have high requirements regarding their system's velocity because of the high number of transactions per hour, the data is available in a short period of time, and the necessity to make critical decisions on time.
- **Variety**: refers to the various data types within a dataset. The data can be structured (spreadsheets), semi-structured (web pages, .xml files, social media sites, e-mails, and sensor devices data) or unstructured (images, videos, and audios). Because a small portion of data is structured, there is an extra challenge for data analysis, interpretation, cataloging and integration, which results in the reduction of period time for quality assessment.
- **Veracity**: refers to the integrity, authenticity, and genuine origin of data. This characteristic includes two aspects: data consistency and data trustworthiness.

Some other aspects should be addressed to ensure data veracity such as storage trustworthiness, accessibility, accountability, and reputation.

- **Value**: refers to the usefulness of data to make decisions. The value of the data will depend on the outcome of the processes it represents. The process would be stochastic, probabilistic, regular, or random (Y. Demchenko, 2013).

Value can be marked as a constraint for big data because the challenge to derive a real benefit from analytics at the same time to deal with a time pressure, a huge amount of data generated at a high speed from different sources and in different formats. Even worst, should be considered an extra effort to assure the integrity and authenticity of data.

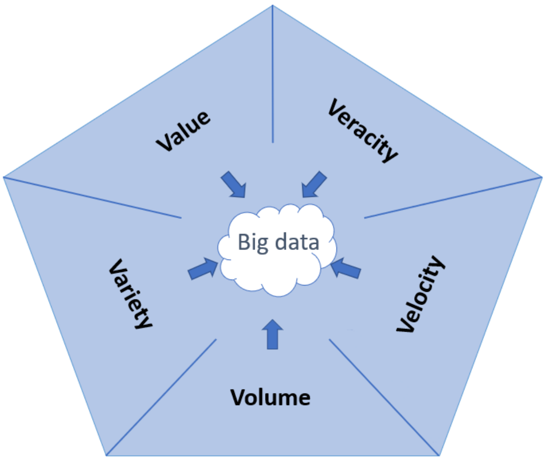


Fig. 2 Pentagon of the five main issues / challenges / characteristics in Big Data.

A big data process is also presented with four main stages:

- Create or collect data from different resources.
- Store the data.
- Compute and analyze the data.
- Visualize the results.

Each one of these stage of the Big Data "process" must be measured against defined quality rules, probably through a Quality Management approach. This process presented here is closely related with that shown in Fig 3 (Zhang, Xiong, Gao, & Wang, 2018).

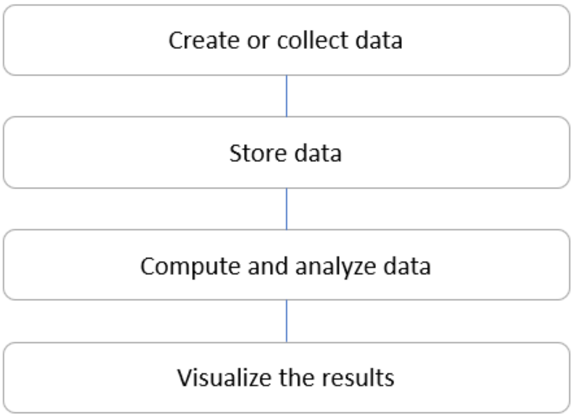


Fig. 3. The Big Data lifecycle at a quick look has four main steps. Source: (Zhang, Xiong, Gao, & Wang, 2018)

Create or collect data: The data collection is the base of the whole process. With the new technologies and the Internet of Things, the range of data producers is increasing, and the output of data is augmenting as well. This process mainly includes system log and network data collection methods.

Store data: Data storing, and processing is not only at big scale, but also required to transmit and process the response quickly. The traditional database storage technology may face problems to adapt to Big Data storage and the system must be compatible with all data types, hardware, and software platforms.

Compute and analyze data: After the previous steps, the next one is to analyze the data collected and stored. Through this process, intelligent, in-depth, and valuable information can be obtained. This information will be particularly useful to make decision processes, predictions, and analytics.

Visualize the results: Is particularly important to apply an adequate tool to report the results obtained in such a way that clients and other stakeholders could interpret the results easily and effectively.

When searching for Big Data applications there are three typical Big Data systems: Recommendations, Feature Prediction Systems, and Decision Making.

2.1.1 Recommendation Systems

A recommendation system will recommend the interests and information to the user according to the user's interest characteristics and network behavior. This kind of systems is widely used by Amazon, Twitter, and Google. According to the user information and location, the user would search restaurants, hotels, cinemas, pharmacies, hospitals, or tourist attractions following the instructions of mobile applications and web systems. In the context of Big Data applications, the Recommendation systems are an extension of traditional recommendation systems. Because the complexity of Big Data environments, extracting and predicting user's preferences in this context can produce more accurate recommendations.

2.1.2 Feature Prediction Systems

The Feature Prediction Systems are based on existing data analysis, according to the development and dissemination of data to assess future trends and conditions. This kind of systems are the core of Big Data applications in a new era where traditional predictive tools cannot handle the size, speed, and complexity of actual data. The weather prediction is one of the major beneficiaries of this kind of systems. Other good examples are predicting the match or game results, the number of expected patients in a global pandemic or the stock market fluctuations.

The accuracy is an ultimate goal of a Feature Prediction System. In the prediction, if a specific variable has a decisive impact on the results and it is difficult to collect, then will be also difficult to predict. The system should be able to accurately capture the data for each variable and adjust the prediction in real time. The problem arises when data is elusive which leads to the fact that prediction accuracy is a big issue.

2.1.3 Decision Making

The ultimate goal of data science should be improving decision making process. Decision Making process, is the process of making choices through the identification of a decision, gathering all necessary information, and finally assessing alternative solutions. It is typically reliant upon data quality and data accuracy, the information and knowledge. In (Alkatheeri, y otros, 2019) is explored a good approach to this kind of quality models.

This kind of Big Data application refers to the practice of basing decisions on the analysis of the data instead on the human intuition or user experience.

2.2 Understanding big data architecture

The big data architecture is recognized as a value chain with four main phases: Generation, Acquisition, Storage and Processing (Ghorbanian, Dolatabadi, & Siano, 2019). The raw data generated by several data sources are analyzed, processed, and changed into useful information.

In data generation the main big data sources are users, applications, services, systems, sensor devices, technological devices, among others (Y. Demchenko, 2013). People and AI contribute to big data in the form of documents, images, videos, software, files with a multi-diverse format styles and in some cases without adequate metadata to describe it. Data collection and preprocessing are key aspects of data acquisition where data is obtained and classified for further phases. At preprocessing stage, the possible errors are integrated and cleansed, data redundancy is eliminated and finally resulted data are compressed.

In data storage phase, the collected data is stored and managed for further uses and applications. Finally, the processing phase is the most important phase where analytic approaches are executed by using inspection and modeling methods to prepare, classified and extract information from the collected data for further processes.

The hot research topics in Big Data since its born have been related with technology such as data mining, cloud computing, machine learning, electronic data processing, and others (Strang, 2020).

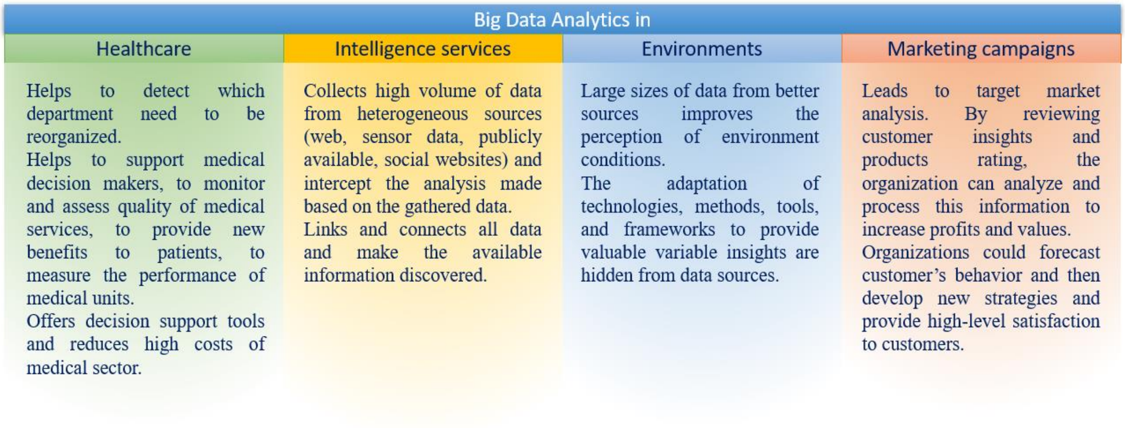
2.2.1 Big Data Analytics

Big Data Analytics (BDA) is a method of logical analysis on a large dataset. This is a process used to collect, store, and analyze heterogeneous data at a high velocity from a high-volume of data. Through this process, hidden information such as useful patterns and meaningful insights can be extracted. The result is used for different kind of Big Data applications which helps in business grow and online business, online browsing, social media, weather forecast, among others (Rani & Sagar, 2018).

According to (Patgiri, 2018) , the BDA can be divided into three main categories:

- predictive analytics: that uses statistical data.
- descriptive analytics: that contains historical data which is related with business intelligence.
- prescriptive analytics: that is used to find out the optimized solutions for the concerned problem.

BDA has evolved to assist and guide the decision-making processes where the data-driven inside decision-making is identified as the most crucial and critical part of an organization to make a move and take a decision. The most prominent research in BDA is about supporting healthcare specifically in decision support systems for a better outcome, prevention, low-cost and early detection of an event. Fig 4 exposes examples of BDA in some areas.



Big Data Analytics in			
Healthcare	Intelligence services	Environments	Marketing campaigns
Helps to detect which department need to be reorganized. Helps to support medical decision makers, to monitor and assess quality of medical services, to provide new benefits to patients, to measure the performance of medical units. Offers decision support tools and reduces high costs of medical sector.	Collects high volume of data from heterogeneous sources (web, sensor data, publicly available, social websites) and intercept the analysis made based on the gathered data. Links and connects all data and make the available information discovered.	Large sizes of data from better sources improves the perception of environment conditions. The adaptation of technologies, methods, tools, and frameworks to provide valuable variable insights are hidden from data sources.	Leads to target market analysis. By reviewing customer insights and products rating, the organization can analyze and process this information to increase profits and values. Organizations could forecast customer's behavior and then develop new strategies and provide high-level satisfaction to customers.

Fig. 4 The Big Data Analytics in some crucial areas

One of the first key challenges of BDA is real-time events monitoring with high data volume. The other one is about the prediction of variables in future based on Big Data using machine learning algorithms. The main challenging questions are:

- How to manage BDA if data volumes gets so large and varied that makes not sure how to deal with it?
- Does all data need to be stored?
- Does all data need to be analyzed?
- How to find out which data values/aspects are really important?
- How can the data be used to obtain the best advantage?

2.2.1.1 Predictive Analytics

These BDA predict future trends based on statistical techniques and its nature is "probabilistic". A characteristic is that predictive analytics can only predict and forecast in future probabilities by using techniques such as logistic and linear regression to predict the future trends and outcomes. These techniques extract patterns from big datasets and by using data mining, statistical methods, and machine learning, make the expected predictions.

2.2.1.2 Descriptive Analytics

This is the simplest method of BDA. The purpose is to summarizes historical data about which topics are involved. This type of analytic refers to what happened in past and looks historical data to understand the reason behind success or failure. Some examples that use descriptive analytics are marketing operations and sales.

2.2.1.3 Prescriptive Analytics

This BDA summarizes the data, business rules and computational science obtained and then present the results to business analytics which is useful for augmenting efficiency in business process. Prescriptive analytics make use of new current data every time for improving accuracy and providing better decisions. The use of hybrid datasets as input allows prescribing how to take advantages of the prediction in the future.

2.2.2 Examples of Frameworks for Big Data Analytics

Several types of Frameworks are required to run the types of data analytics. Some of these examples are described below.

MapReduce for batch Analytics applied in historical data: In this example, MapReduce is suitable for batch analysis and allows to write applications that process huge amount of data. In this framework, every data processing is divided into Map and Reduce steps in a parallel model.

MapReduce works with a master node (the job tracker) and several slave nodes (task trackers). The framework split large datasets into smaller ones and execute parallel processing which are thrown on slave nodes. The nodes are computed independently.

This is a fault tolerance framework that plays an important role in BDA working with sequential data, but it is not suitable to works with random data.

Batch processing is adopted because the faster response of the framework to real-time applications.

Stream Processing: In this example, current online data is processed in streams. The process takes input in the form of stream data and is suitable for online analytics. Only a few passes over the stream are necessary to find approximation results. The results are provided as quickly are needed, for example in milliseconds.

2.3 Big data issues and challenges

A survey on data quality was presented by (Laranjeiro, Soydemir, & Bernardino, 2015) where the quality of data is analyzed from multiple "dimensions", identifying a dimension as a measurable data quality property which represents some data aspects like accuracy, consistency, currency, etc. The authors identify also quality problems inside the dimensions like "format problems under the accuracy dimension". The dimensions are grouped into four categories: **Intrinsic**, representing the natural quality of the data; **Contextual**, expressing the fact that data quality must be considered in a specific context; **Representational**, related with the format and the meaning of the data; and **Accessibility**, expressing how accessible the data is for users.

This study concludes that there is a subset of dimensions which are the most cited among the papers analyzed and also coincides with the definitions presented in ISO/IEC 25012 standard but with the particularity that the standard named those as characteristics. The dimensions are: Accessibility, Accuracy, Completeness, Consistency and Currency.

The authors in (Tao & Gao, 2016) presented a work on big data system validation and quality assurance with a well proposed test process for big data-based applications. In this process the authors remark the steps in a bottom-up order as follows: big data

System Function Testing, big data System Testing, big data System Feature Testing and big data System Timeliness Testing. A sample scope of validation for quality assurance should include factors like: correctness, performance, robustness, reliability, data security, consistency, accuracy, and accountability. Additionally, for a prediction service, some quality factors are added including: usability, duration, a deviation analysis, and data pattern style.

In (Zhang, Zhou, Li, & Gao, 2017), have published a survey on quality assurance techniques of big data applications introducing big data properties and quality attributes. The paper provides contributions in aspects like: 1. the quality assurance approaches for big data applications such as testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and prediction as main ways that we can use to ensure the quality of big data applications; 2. combining quality assurance techniques with big data characteristics; and 3. they have collected quality parameters like Data accuracy, Data correctness, Data consistency and Data security from other published papers and a set of quality factors such as performance, reliability, correctness, scalability and security.

(Abdallah, 2019) presents a list of quality factors from different perspectives including data, management, processing and service, and users. Several new factors come to light, but others are the same than previous studies like: accuracy, correctness, consistency, completeness, accountability, and scalability. The author is arguing about quality factors and issues to measure the big data application, these factors are spread into the four perspectives mentioned above.

According to the author, the Quality Management approach has five facets, there are: the people, as key players; data profiling where the data is reviewed, compared, and analyzed to check its accuracy; data quality where quality rules for the date are created; data reporting; and data repair where should be finding the best way to remediate and change corrupted data.

Finally, a partial conclusion can be obtained here, and it is nothing more than big data quality characteristics (as dimensions) are used to produce different big data quality models, those characteristics are going to be used to measure the quality from different aspects (or layers, or perspectives) regarding the big data applications.

2.4 Big Data and Good Practices

Some Big Data good practices where analyzed from (Katal, Wazid, & Goudar, 2013), (Ciancarini, Poggi, & Russo, 2016), (Tepandi, y otros, 2017), and (Castillo, y otros, 2018). These are applied depending upon the context of application, such examples are:

- Creating dimensions and facts of all data being stored in datasets.
- All those dimensions should have durable keys that cannot be changed by any business rule. These keys should be generated by hashing algorithms ensuring uniqueness and be assigned in sequence.
- Trying to integrate structured semi-structured and unstructured data as the same kind of data.
- Building technology around key values is needed to deal with different data formats.
- Analyzing data sets including the information about individuals or organizations privacy.
- Applying different tasks at the earliest point possible such as: filtering, cleansing, pruning, conforming, matching, joining, and diagnosing, to obtain better input data and best results.
- Identifying necessary limits on the scalability of data stored.
- Investing in data quality and metadata to reduce the processing time.

These good practices could be analyzed when developing or adjusting a quality model approach.

2.5 Big Data Tools

A huge amount of Big Data tools are available in the market, both to buy or for free to be applicable at data extraction, data storage, data cleaning, data mining, data visualization, data analysis and integration (Rahman, Begum, & Ahmed, 2016). Table 1 shows some of the most popular Big Data tools by areas.

Big Data area	Tools implicated
Data Storage and Management	Hadoop, Cloudera, MongoDB, Talend
Data Cleaning	OpenRefine, DataCleaner
Data Mining	RapidMiner, Teradata, FramedData, Kaggle
Data Analysis	Qubole, BigML, Statwing
Data Visualization	Tableau, Silk, CartoDB, Chartio, Plot.ly
Data Integration	Blockspring, Pentaho
Data Languages	R, Python, RegEx, XPath
Data Collection	Import.io

Table 1. Some popular Big Data tools. Source: (Rahman, Begum, & Ahmed, 2016).

2.6 Data Quality

The concept of quality could resemble to “fitness for use” and Data Quality is the data that is best suitable for data consumers to be used (R.Y. Wang, 1995). Studies into the multi-interpretations of data quality indicate that fitness for use is a fundamental criterion in data quality evaluation. Data quality is a key factor that determines usefulness of data and user satisfaction, and its measures are developed from ad hoc bases and its evaluation depends principally on the needs and expectations of final users and stakeholders. Users not only want to access available data, but also high-quality information where quality requirements are user specific.

In (Nikiforova, 2020) a variety of surveys regarding the effects of data quality are presented, including losses caused by poor data quality. Some examples are:

- 2011: About 40% of data in companies is of poor quality, according to Gartner studies. It specifies that data quality is closely linked to process quality and business succeed.
- 2015: The US postal service provider USPS had a loss of US $3.4 billion per year due to incorrect address data.
- 2017: An IBM's research found that business decisions taken on the basis of low-quality data have cost to the US economy approximately $3.1 trillion per year.
- 2017: The annual Gartner group research demonstrates that companies with data quality problems may lose $15 million annually because of these problems.
- 2018: The low quality of data was considered "the leading cause of failure for advanced data and new technologies with losses of up to $9.7 million for American businesses each year".

Poor quality of data could negatively impact on social and economic aspects due to wrong decision-making process based on consumer behavior data. Also, leads to hurts employee's morale and its more difficult to put project in order. Its effects at operational level are related with customer satisfaction decreasing, cost increasing and decreasing in employee's job satisfaction and tactical level. Recently, have been increased the references related with poor data quality and its impact have been appeared in different kinds of literature, social media, and other publications (Rudraraju & Boyanapally, 2019).

The reasons to have poor data quality are diverse. Among others, data entry errors, incorrect methods for collect the data, missing data values, incorrect data values, impossibility to update data and make necessary changes over time, misapplying business rules and duplicating data, are good examples to illustrate this issue.

Related to the anomalies presented in data quality, in (Talha, Kalam, & Elmarzouqi, 2019) three types of them are discussed:

- syntactical anomalies
- semantical anomalies
- coverage anomalies

Syntactical anomalies are examples of: lexical errors, which means differences between data structure elements and the specified format of them; domain format errors, which occur when the value for an attribute doesn't conform to the expected domain format; and using data values in a non-uniform manner, for example when using different date formats.

Semantical anomalies are examples of: integrity constraints violations, that is present when values in a tuple or a set of tuples don't satisfy the integrity constraints; contradictions presented in data values in a tuple or between tuples, that violate dependencies between those values such as a difference between age and date of birth; and duplicating entries or having invalid tuples.

Coverage anomalies are examples of: missing values in a tuple and missing tuples.

Regarding the source of data, (Laranjeiro, Soydemir, & Bernardino, 2015), have presented a study about data quality research and the most frequently mentioned data quality problems. It split the quality problems into single and multiple sources:

- **single source problems:** are related to a single source of the data and mainly with integrity constraints (missing data, incorrect data, misspellings, ambiguous data, extraneous data, outdated temporal data, misfielded values, incorrect references, duplicates, domain and functional dependencies violations, wrong data type, referential integrity constraints and uniqueness violation.
- **multi source problems:** are mainly related with the integration of multiple data sources (different units, different representations, structural conflicts, different word orderings, different aggregation levels, temporal mismatch, synonyms and

homonyms wrongly used, using special characters and different encoding formats.

Related to data quality there are main aspects such as data consistency, data deduplication, information completeness, data currency and data accuracy. The study of data quality has been mostly focused on data consistency and data deduplication in relational data (Fan, 2015). Other problems related with data analytics are: data testability, data availability, data scalability and data security (Sangeeta & Sharma, 2016). In next subsections these data quality characteristics are explained with others extracted and analyzed from different sources such as (Cortes, Bonnaire, Marin, & Sens, 2015), (Libes, Shin, & Woo, 2015), (Noorwali, Arruda, & Madhavji, 2016), (Tao & Gao, 2016), (Ardagna, Cappiello, Samà, & Vitali, 2018), (Castillo, y otros, 2018), (Zhang, Xiong, Gao, & Wang, 2018), (Musto & Dahanayake, 2019) and (Talha., Kalam, & Elmarzouqi, 2019).

2.6.1 Data Consistency

Data consistency is understood as the validity and integrity of data which represents real-world entities. The focus is to detect inconsistencies and conflicts in the data which are identified as data dependencies violations (integrity constraints). In this cases the corrupted data could be repaired by fixing the errors.

This property could answer two questions: How reliable is the data? Are data values the same across all systems? If two values are read from separated sources, it is expected to match and align them, regardless of what source the data is collected its value cannot be contradicted.

2.6.2 Data Deduplication

Data deduplication is a problem that raises when identifying tuples from one or more object relations that refer to the same real-world entity. This problem is also known as record matching, record linkage, name matching, database hardening, instance identification, duplicate identification, and object identification. This is the most extensively studied data quality problem.

Data deduplication is important in Big Data context at a large number and heterogeneous data sources particularly for data quality management, data integration and fraud detection. In some cases, it is necessary to accurately identify tuples from different data sources that refers to the same entity so that, data can be fused and enhanced to make practical use of them. The data from different sources can be dirty and even more, when data sources are seemly reliable, the inconsistencies and conflicts could emerge when data is integrated.

2.6.3 Data Completeness

Data Completeness refers to whether the data base has complete information to answer any query by using only the data contained in the database. There are two approaches on this scenario, close or open. For the Close World Assumption (CWA) the database includes all the tuples necessary to represent real-world entities but for some attributes the values may be missing. For the Open World Assumption (OWA) the database could be a subset of the set of tuples which represent real-world entities, in this case, both tuples and values could be missing. Normally, the CWA approach is too strong in the real world, in contrast with OWA approach few queries with correct answers could be found.

2.6.4 Data Currentness

The objective of data currency is to identify the current values of the entities represented by tuples in a database, and to answer queries with the current values. It answers the question: How recent was the data collected or updated? Refers the degree to which the data is current with the world values.

When the data value is renamed, moved, or changed in one source then the data is not current and must be updated at the rest of sources. These updates could be manually or automatically and take place as needed or can be scheduled periodically. However, this task could be expensive, so that 100% of data currency could be neither affordable nor required.

2.6.5 Data Accuracy

Characterizes the degree to which data attributes represent the true value of the intended attributes in the real world, like a concept or event in a specific context of use. The information that data contains corresponds to the reality and answer the question: Is the data free of mistakes and exact? Some possible mistakes could be information outdated, redundancies or typographical errors. Others could be related to sensors that are not capable of capturing data appropriately. The main goal should be to increase the accuracy of the data, even when the source grows in size.

2.6.6 Data Scalability

Characterizes the increasing of data volume at a higher speed than actual computing resources and processor speeds available. It measures the scalability that occur in storages.

2.6.7 Data Accessibility

This characteristic describes to what extent the data is accessible for the data analyst which is dependent on technical system conditions. If data is not accessible for the data analyst, the BDA cannot be performed as required.

2.6.8 Data Timeliness

This characteristic refers to the data that is strongly related with time such as, weather, news, current affairs, and others. The Big data applications requires real-time analysis of the available input data for a precise performance. The characteristic requires that the storage system be able to maintain a high-speed response, because the response delay may lead with the problem of expired content presented to the user and would result in invalid outputs.

2.6.9 Data Availability

This characteristic is related with data accessibility and timeliness. It is necessary to measure whether data access is available or not. Data availability incurs in some features such as:

- the data is easily made public or easy to purchase.
- the data arrive on a given time.
- the data are required to be regularly reviewed and updated.
- the data meets requirements in the interval from data collection and data processing.

2.6.10 Data Confidentiality

This characteristic refers the degree data is accessible only by authorized users in a specific context of use. Due to some privacy regulations, access to data may be restricted pursuant to the procedure stablished and regulated. Data may be confidential or non-confidential independently from other characteristics such as accuracy, completeness, consistency, among others.

2.6.11 Data Traceability

This characteristic covers the tracking of changes made to the date as part of the data collection process. Without data history information, BDA could lead to wrong interpretation of the results.

2.6.12 Data Credibility

This characteristic describes to what extent the signal data is true and believable. Unreliable data should not be used for BDA and therefore deteriorates data quality. It includes the origins of data if the data comes from a valid and believable source or not.

2.6.13 Data Precision

This characteristic refers the degree of how exact the data is. A good example is about location data where this can be represented by exact latitude and longitude values or represented by more broad areas depending upon the data requirement needed. The precision degree needs to be defined at the initial steps and all values are then compared to the predefined precision to get the degree of precision.

2.6.14 Data Security

This characteristic is a major concern in Big Data applications. Two classical attacks to data security are: attacks during execution phase and training phase.

During the execution phase, the attacks can be produced by aggregating the input streams that are used to influence intelligence and actionable analysis and are generated by the Big Data application. During the training phase, the attackers could create data generators that will affect the reliability of the Big Data results.

Other typical attack is denial of service which cause the problem of denying access to system and may be also viewed as a data accessibility issue.

2.6.15 Data Privacy

This characteristic is one of the foremost concern in Big Data applications. Data privacy is essential as there is fear of inappropriate use of personal data which might be revealed when integrating such data from other sources. Data privacy is not only a technical problem, but also a sociological problem. Social websites, consumer and business analytics and governmental surveillance are some areas where privacy issue is crucial. Other important areas are related with location-based services where is require the user to share its location, which leads to leakage in privacy because user's identity may be revealed. A good example of this issue is related with Google Map where the user's habits are tracked by a service which provides useful information to the user such as the better route to arrive at the destination, car traffic information, weather information, etc. However, at the same time, such information reveals privacy issues that must be managed.

2.6.16 Data Compliance

This characteristic describes the fact of strict observance of standards and conventions. A clear example is related with the programming language used where programming rules need to be accomplished. In this regard, the timestamps are more sensitive since the notation varies widely. The data need to be compliant for an adequate usage of the given information and a high data quality.

2.6.17 Data Usability

This characteristic represents a measure of the effectiveness and efficiency in achieving the stated goal. Data usability can be measured from two aspects, one of them is the description of the dataset which offer a clear understanding to the user about data quality dimensions that are affecting the entity. The other aspect, the mapping between data and its corresponding analytics would become cumbersome when data are not stored by human-friendly representations.

2.6.18 Data Relevancy

This characteristic is related with addressing customer's needs.

2.6.19 Data Portability

This characteristic is dependent of the technical system that data arises from, such as moving the data between different sources / stakeholders. If it is not possible for the data analyst to move data to a specific tool, the BDA could be compromised.

2.6.20 Data Understandability

This characteristic covers large data content or scope and measures the facility to be understood by final users and technical teams. It is the degree of how easily the data can be interpreted and understood, and whether the data is expressed in appropriate symbols and units. In some cases, the same metrics could obtain different measures depending upon the scenario and country where is applied.

2.6.21 Data Uniqueness

This characteristic assure that nothing will be recorded more than once based upon the data is identified.

2.6.22 Data Validity

This characteristic measures weather data are valid by conforming to syntax rules (format, type, range).

2.6.23 Data Heterogeneity

Normally, the data generated by users are heterogeneous by nature whereas big data applications expects homogeneous data for better processing and analysis. This characteristic refers to the necessity to work with structured data at the beginning of analysis phase because structured data is well organized and manageable. If this is not possible to achieve, should be considered that unstructured data is costly to work with it and also is not feasible to convert unstructured data to structured data.

2.6.24 Data Efficiency

This characteristic represents the degree of how fast the data can be processed and accessed against expected results using a predefined number of resources in a specific scenario that is properly defined and considered system performance limits. An example is about the overall time for accessing certain amount of data from the entire dataset.

2.6.25 Data Recoverability

This characteristic refers to the degree of how well the system where data is stored, could maintain a required and specified operability and quality level when a failure occur.

As have been identified, data quality can be divided into separate characteristics, and most of them have some type of relation with other characteristics. When handling some characteristics, other ones may be considered or dismissed. A good examples is having Availability and Confidentiality working together, while considering data available some confidentiality issues could appear. Another example could be the possible tradeoffs between understandable data and compliant data where in some cases while accomplish with rules, data restrictions and regulations some understandings could be missing.

2.7 Data Quality Models

Data Quality models enhance traditional models trying to have these models as a base to represent a data quality dimension and other quality dimensions which are related to it. This kind of quality model allows to analyze a set of requirements for data quality and helps to represent it in terms of conceptual schema. This kind of model accesses and investigate all data quality dimensions in a logical schema. Through the use of this quality model the selected data can be traced down from its source where all the changes of every data stage are viewed until data reached the final stage. This is especially useful to detect the root cause presented in poor data quality and the necessary actions to solve the problem and improve the quality.

This kind of model provides a method to describe a specific data element, to clarify the stages of expected data and the information involved, and to give all additional attributes needed for identify quality requirements.

The ISO standards introduced data quality assessment in the 2501n family that is focused on software product quality models. The 25012 (International Organization for Standardization, 2008) standard is focused on data and defines fifteen quality characteristics split in two main dimensions: inherent and system dependent. The inherent dimension is where data have the intrinsic characteristic to satisfy the user requirements when is used under specified conditions. The system dependent dimension is where data quality is attained and preserved within a computer system when data is used under specified conditions, at this dimension data quality depends on technological domain.

Achieving data quality as the suitability of a given dataset or several datasets, and its properties for a particular data usage is a goal that is reachable by applying a data quality model. However, the same data in the datasets may be suitable for a unique Big Data application, a set of them or none of them. It may be necessary to define different data quality characteristics for the same data, depending on the usage of these data. Other key information related with data quality improvements can be obtained from (Alarcos Group, 2020).

2.7.1 Data quality characteristics

As part of a data quality model, there may be identified several data quality dimensions or quality characteristics. The quality of data can be analyzed from multiple dimensions. These are the concepts that define the quality of data and also can be interpreted as quality characteristics.

Data quality characteristics have the ability to describe various intrinsic attributes of data quality such as relevancy, accessibility, understandability, uniqueness, recoverability, timeliness, among others presented in sections 2.6.1 to 2.6.25 that represent a single aspect or construct of data quality. Some particular data can be described as of being of high quality if accomplish with some quality characteristic.

2.7.1.1 Measuring quality characteristics

To measure a quality characteristic, some metrics can be associated and applied as quality indicators. A quality metric will define how to evaluate a quality characteristic. When the metric is based on quantitative measures is called “objective” (obtaining a numerical value, e.g., the result of a condition or a mathematical equation). When the metric is based on qualitative evaluations such as user perceptions, needs and expectations is called “subjective” (validating a feedback questionnaire or reviewing user surveys) (Talha., Kalam, & Elmarzouqi, 2019). There are three types of metrics as data quality indicators:

- content-based: the information is used as a quality indicator.
- context-based: the metadata is used as a quality indicator about the circumstances in which the information was created or used.
- rating-based: the information has explicit ratings.

Some strategies could be used to improve quality, among others:

- deleting duplicates
- updating obsolete data values
- correcting data values
- data cleansing
- record linkage

- adding a data format check-up before storing
- adding a validation step for data source reliability

2.7.2 Big Data characteristics for a Big Data application

Some conventional quality parameters can be applicable to any Big Data system, some of them are:

System Performance: Indicates the performance of the system such as: availability, response time, throughput, scalability, and others.

System Data Security: Is used to evaluate the Big Data security systems from different perspectives at different levels.

System Reliability: This characteristic evaluates the system durability when is executed a required function under stated conditions in a specific period of time.

System Robustness: This characteristic evaluates the ability of a system under test to resist any change without adapt the initial proper system configuration.

3. A Systematic Mapping Study

The methodology used to conduct the current research is presented in this section. The review type Systematic Mapping Study (SMS) has been selected for evaluating and interpreting available research relevant to the current key topics as can be seeing in Fig 2.

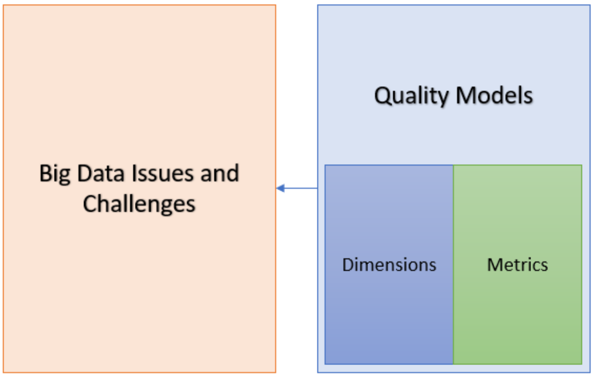


Fig. 5 Key topics presented in this SMS.

By applying a SMS, the main issues and challenges from Big Data applications could be revealed and several quality models proposed to evaluate those Big Data applications in the last decade could be identified by applying a distinction between the different types of quality models used in the context of Big Data. In addition, it is expected to define the principal quality dimensions and metrics proposed in this area.

A part of this investigation has been accepted in QUATIC 2021 (Montero, Crespo, & Piattini, 2021).

The idea to conduct a SMS on these topics rise from previous reading of authors like (Tao & Gao, 2016), (Abdallah, 2019) and (Zhang, Zhou, Li, & Gao, 2017). In their papers, the necessity to conduct a deep research on big data quality challenges and future needs is stated.

3.1 Related work

At the beginning is recommended to present some findings of previous work related with the key topics such as: big data issues and challenges, big data quality models, quality dimensions and quality metrics.

(Ashabi, Sahibuddin, & Haghighi, 2020) has conducted a study about big data characteristics and challenges and have presented an investigation between 2009 and 2018. The paper has exposed a general outline of the characteristics of Big Data and identified the challenges and present limitations in this area. This paper do not expose big data quality dimensions although is representative of issues and challenges in big data.

(Muthukrishnan, Yasin, & Govindasamy, 2018) have conducted a SLR on stablishing gaps that need to be addressed in big data analytics for education landscape where 59 research papers were selected. One of their review objectives was focused on predictive models using Big Data applied to measure the student's performance and concluded with the fact that there are limited papers on this topics. Only a few approaches are presented and do not represent a good contribution to the topics on this SMS.

In (Al-Sai, Abdullah, & Husin, 2020) is presented an approach for a SLR applied to the analysis of Critical Success Factors (CSFs) in Big Data analytics and their categories with the selection of 16 related articles. This SLR is focused on factors that inside in the correct implementation on big data like Key Success Factors and do not contribute substantially to big data quality dimensions or factors.

In (Laranjeiro, Soydemir, & Bernardino, 2015) a survey on data quality is conducted and presented a review of 22 papers where a variety of several data quality dimensions is summarized. This survey is focused on data quality itself; it is not related with terms like big data issues and challenges or big data quality models. They survey about the classification of poor data including the definition of dimensions and specific data problems. They support the idea that the quality of data can be analyzed from multiple dimensions, and after identifying frequently used dimensions, they map data quality problems to those dimensions.

Related with quality models a systematic mapping study is presented by (Yan, Xia, Zhang, Xu, & Yang, 2017) focused to resolve the gap of Quality Assessment Models (QAM) limited investigation. In this sense they have defined some research questions, selected the databases, and defined a search strategy. The quality models presented here are oriented to software products and do not present any relation with big data quality models.

In (Pereira, y otros, 2020), a review on key non-functional requirements in the domain of Big Data systems is presented, finding more than 40 different quality attributes related to these systems and concluding that non-functional requirements play a vital role at software architecture in Big Data systems.

(White, Nallur, & Clarke, 2017), presents another review that evaluates the state-of-the-art of proposed Quality of Services (QoS) approaches on the Internet of Things (IoT) where one of the presented research questions, refers to the quality factors that quality approaches consider when measuring performance.

The research that comes closest to the actual investigation is (Rahman & Reza, 2020). These authors presented a SMS involving concepts such as “quality models”, “quality dimensions” and “machine learning”. A selection of 10 papers is done where some quality models are reviewed and a total of 16 quality attributes are presented that have some effects on machine learning systems. Finally, the review is evaluated by conducting a set of Interviews with other experts.

As a conclusion of reviewing related work, there is a lack of papers which have conducted a study on the big data issues and challenges, examples of big data quality models and big data quality metrics.

3.2 Definition of the Research questions

Considering the aim and objectives formulated, the following research questions are proposed:

RQ1: What are the main issues and challenges detected on Big Data applications from 2010 to 2020?

RQ2: What were the quality characteristics, quality dimensions or quality factors applied to Big Data applications that have been identified by the authors in their publications from 2010 to 2020?

RQ3: What quality models related to Big Data applications have been proposed from 2010 to 2020 and which quality characteristics were defined as part of these models? Which quality metrics have been applied?

RQ4: For which Big Data specific context have these quality models been proposed?

RQ5: Have Big Data quality models been proposed to be applied to any type of Big Data application or by considering the quality characteristics required in specific types of Big Data applications?

3.3 Literature Review

In this section, the procedures followed by implementing a systematic mapping study are described. This methodology can be conducted to get an overview of the research topics.

3.3.1 The search method

The search method selected in the current investigation was the Keyword-based method, which looks for matching documents that contains one or more keywords (as words or phrases) specified by the user. This type of search while using a set of keywords focuses on finding structural information among similar papers indexed in a database. When adding terms such as “AND, OR” to the search string, helps in refining the search that is completed with strings like: “Big Data issues”, “Big Data challenges”, “quality models” and “quality metrics”.

In this sense, the following databases were defined to consult papers and initiate a more in-deep research:

- ieeexplore.ieee.org
- sciencedirect.com
- scopus.com
- acm.org

The search string was refined and obtained as:

- (“Big Data” AND “quality model”) OR (“Big data challenges” OR “Big data issues”).

A database of 1770 initial papers was obtained and prepared for the next revisions. After processing the results of these searches, other papers could be analyzed by applying “snowballing”. Snowballing as systematic reviews refers to using the references included on a consulted paper to reach or identify additional papers related with the current study.

Because the amount of information that is tried to collect and analyze, this investigation will focus the search on scientific databases such as SCOPUS, IEEE, and ACM. Several articles from other databases are indexed in SCOPUS or IEEE. In some cases, papers could be duplicated, and others could not be relevant to the current investigation. For this reason, some inclusion and exclusion criteria are defined.

3.3.2 Inclusion and Exclusion criteria

In this section, pre-defined criteria for inclusion and exclusion of the literature are presented.

Papers included are published between 1st January 2010 and 31st December 2020 whose main contribution is the presentation of new or adapted quality models for Big Data applications or even the discussion of existing ones. Other inclusion criteria are:

- Papers published in a recognized and indexed source.
- Papers which main topics were related to big data issues and challenges; quality models applied to big data applications; and quality metrics presented in those quality models.
- Topics that seemed to be useful to answer the defined research questions.
- Desirable: Peer-reviewed papers.

Papers excluded where those duplicated in different databases or published in Journals and Conferences with the same topic. Papers which could not answer any of the research questions proposed were equally excluded. Other excluded papers were:

- Those that require an additional payment to provide access.
- Those with a low quality in the defined methodology.
- Those focused on the proposal of new algorithms for the implementation of Big Data solutions in a specific area such as: Health, Education, Image Processing, Smart Cities, Aeronautics, etc.
- Those focused on the proposal of new paradigms and models for Big Data application development.
- Those with non-English redaction.

3.3.3 Search and selection process

This search process consist of a manual search in the mentioned databases of specific published articles, conference papers, books, and book chapters since 2010. A primary reading of the abstract is necessary to select those papers which seemed to be relevant to the SMS in a first selection. The papers selected are tabulated for future readings.

The search string was applied to obtain papers which correlate the key topics. After searching in the selected databases additional papers were included using snowballing. From the total of 1770 papers obtained, the inclusion / exclusion criteria were applied and finally a set of 132 papers were selected as the primary studies. The Fig 6 represents the different phases applied during the selection process.

Reviewing the number of citations in the primary studies, it has been found that five papers stand out from the rest. The most cited with 121 citations, is related to measuring the quality of Open Government Data using data quality dimensions (Vetrò, Canova, Torchiano, & Minotas, 2016). With 76 citations, (Merino, Caballero, Rivas, Serrano, & Piattini, 2016) proposes a Quality-in-use-model through their “3As model” which involves Contextual Adequacy, Operational Adequacy and Temporal Adequacy. The third most cited paper has 66 citations and explores the Quality-of-Service (QoS) approaches in the context of Internet of Things (White, Nallur, & Clarke, 2017). In (Immonen, Paakkoneen, & Ovaska, 2015) is reviewed the quality of social media data in big data architecture and has 48 citations. Finally, (Máchová & Lněnička, 2017) with 40 citations, proposes a framework to evaluate the quality of open data portals on a national level.

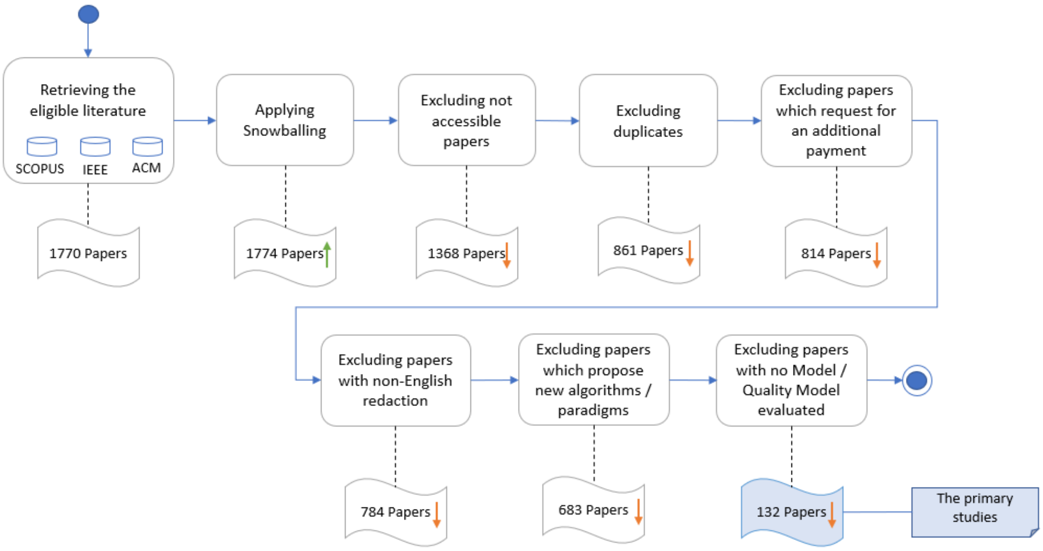

Retrieving the eligible literature
SCOPUS
IEEE
ACM
1770 Papers
Applying Snowballing
1774 Papers
Excluding not accessible papers
1368 Papers
Excluding duplicates
861 Papers
Excluding papers which request for an additional payment
814 Papers
Excluding papers with non-English redaction
784 Papers
Excluding papers which propose new algorithms / paradigms
683 Papers
Excluding papers with no Model / Quality Model evaluated
132 Papers
The primary studies

Fig. 6. Phases inside the search and filtering process.

The data extracted from each paper will be:

- ID: Paper unique identification
- BDD: The source library
- Authors: The authors / contributors of the paper
- Place: International conference, symposium, journal, Lecture Notes, etc., where the paper was published.
- Title: The title of the paper.
- Publisher: The publisher of the paper.
- Number of citations: Number of times the paper was cited.
- Year: The year of publication.
- Topics: The topics addressed in the paper (as a reference).

The data will be tabulated and analyzed to obtain the basic information about each research topic. Finally, the results will be documented and reported.

3.3.4 Quality Assessment

To assess the quality of the chosen literature some parameters were defined as Quality Assessments (QA) such as:

- QA-1: Are the objectives and the scope clearly defined?
- QA-2: Do they proposes/discusses a quality model or related approaches? (if yes, the quality model is applied to a specific Big Data application?)
- QA-3: Do they discuss and present quality dimensions/characteristics for specific purpose?
- QA-4: Do they provide assessment metrics?
- QA-5: Where the results compared to other studies?
- QA-6: Where the results evaluated?
- QA-7: Do they present open themes for further searches?

At this point, the next step was assessing the quality to the selected primary studies which overall results are presented in Table 2. This is a process which complements the inclusion/exclusion and is assigned to answer the quality assessments described above. These primary studies were scored to determine how well the seven quality items defined were satisfied. The punctuation system used was basically a predefined scale with Y-P-N (Y: Yes, P: Partially, N: No), which was weighted as Y: 1 point, P: 0.5 points, N: 0 points.

Quality Assessment	Total Score	Compliance Ratio
QA-1	66	98,51%
QA-2	67	100,00%
QA-3	56	83,58%
QA4-	22,5	33,58%
QA-5	8,5	12,69%

QA-6	31,5	47,01%
QA-7	52,5	78,36%

Table 2. Quality Assessment overall results

3.3.5 Results

After executed the primary review and the papers selected has been tabulated, a complete reading of each paper was necessary to reveal useful information to the current SMS.

Distribution per document type

A distribution per document type exposed in Fig 7, represents that the largest number of documents obtained (94,38%) are distributed as conference papers and journal articles. Other documents reviewed were Books, Book chapter and Reviews.

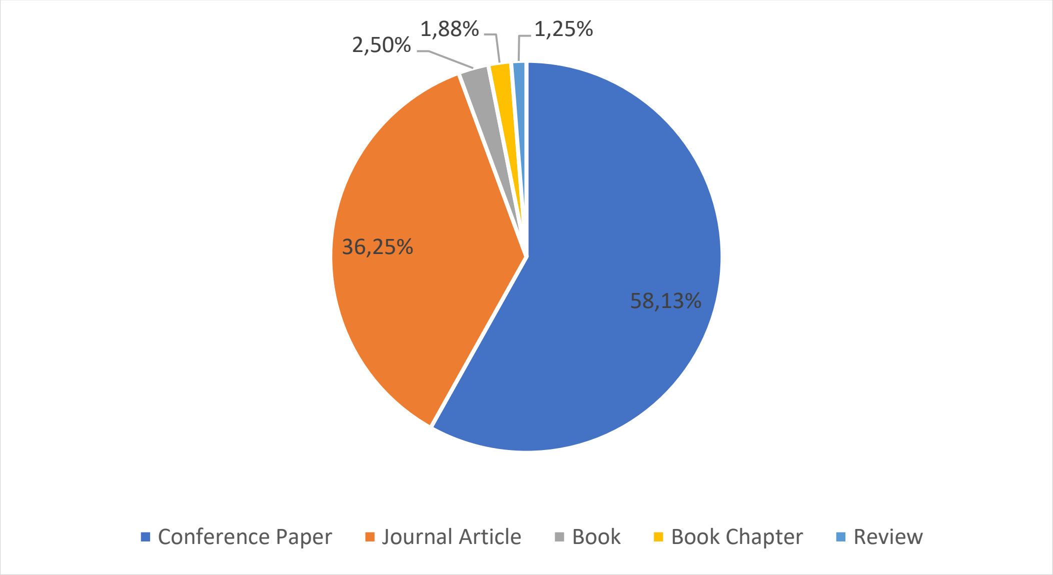


Fig. 7. Distribution per document type

The distribution per source / proceedings is extremely spread because the large of different sources and publications obtained. Regarding the topics in the SMS, is not a surprise to find that majority of publications come from sources related with Big Data domains. The sources from which the largest number of documents were selected is The International Conference on Big Data on IEEE and SCOPUS databases.

Distribution over time period

Regarding the year of publication inside the initial range of 2010 – 2020, Fig 8 shows a clear grouping of publications between 2015 and 2019 with the 77,2% of total papers. From 2010 to 2012 there is no publications selected which accomplish with the inclusion criteria. There is a peak in 2018 with 23 papers where the majority of them were conference papers. The rest of publications are spread between 2013 and 2020 with a gradual increment of the papers published related with these topics.

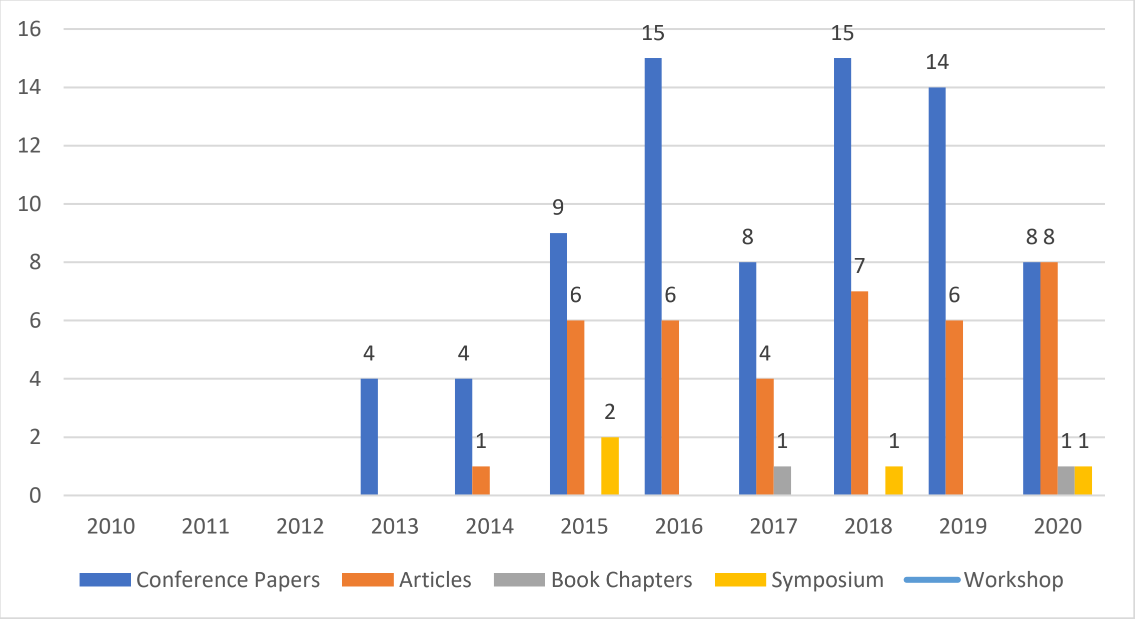


Fig. 8 Distribution over time period per document type (including al key topics).

By separating these findings into two groups: "Big Data issues and challenges" and "Big Data and Quality Models", a more detailed graphic can be obtained. Initially a distribution per time period regarding the topics of Big Data issues and challenges shows a clear grouping of publications between 2015 and 2019 with the 81.5% of selected papers. A peak is observed in 2016 with 16 papers where the majority of them were conference papers as is shown in Fig 9.

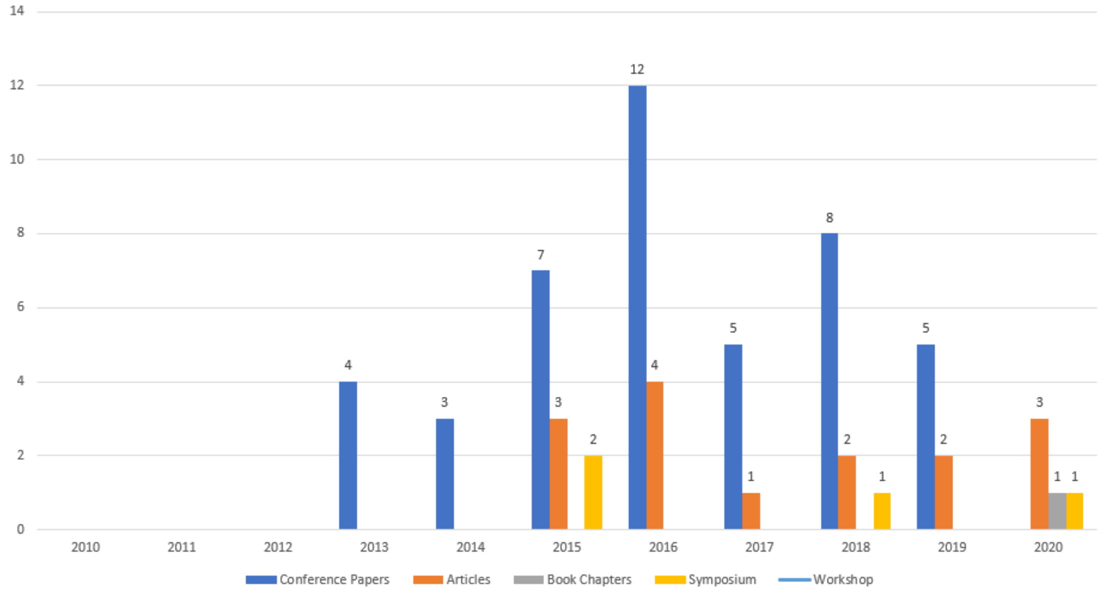


Fig. 9. Distribution over time period per document type (Topics: Big Data issues and challenges).

Regarding the topics of Big Data quality models and quality dimensions, a gradual increase can be seen starting from 2014 in the number of studies published. The 67% of all selected studies were published in the last three years (2018-2020), the 88% of all selected studies were published in the last five years (2016-2020) which is indicating that the issue of quality models in the context of Big Data is receiving more attention among the researchers, and if this trend continues the theme could become in one of the hottest research topics. Fig 10 represents these findings.

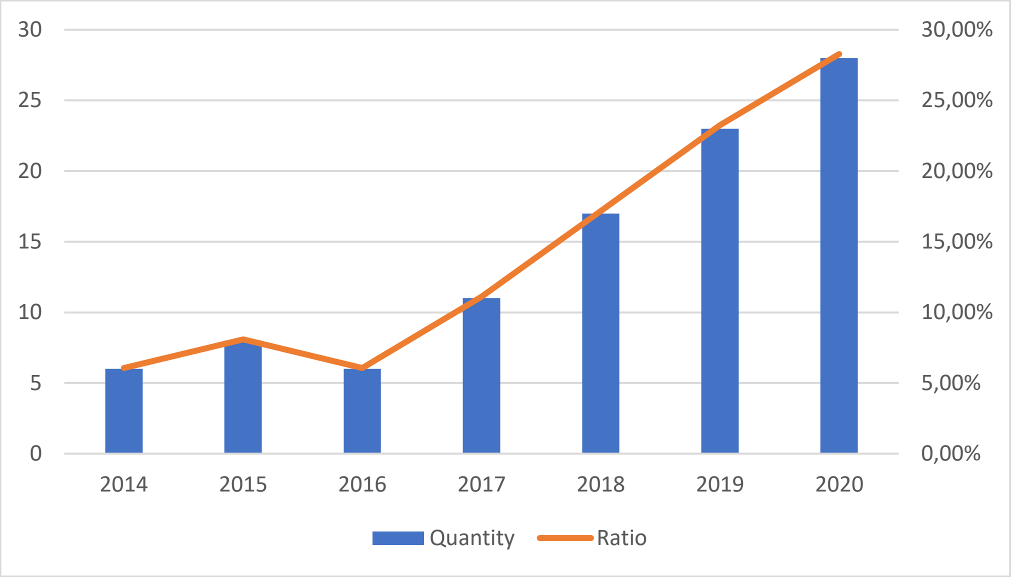


Fig. 10. Distribution over time period (Topic: Big Data quality models)

Distribution per publisher

Regarding the publisher with topic of Big Data and quality models, Fig. 11 shows that most papers were published between Springer (26,87%), IEEE (25,37%), ACM (11,94%), and Elsevier (8,96%), most of them were indexed in SCOPUS, IEEE, and ACM.

Regarding the publisher with topic of Big Data issues and challenges, the distribution per source / proceedings is not so spread and is not a surprise to find that majority of publications come from sources related with Big Data domains as can be seen at the distribution in Fig 12.

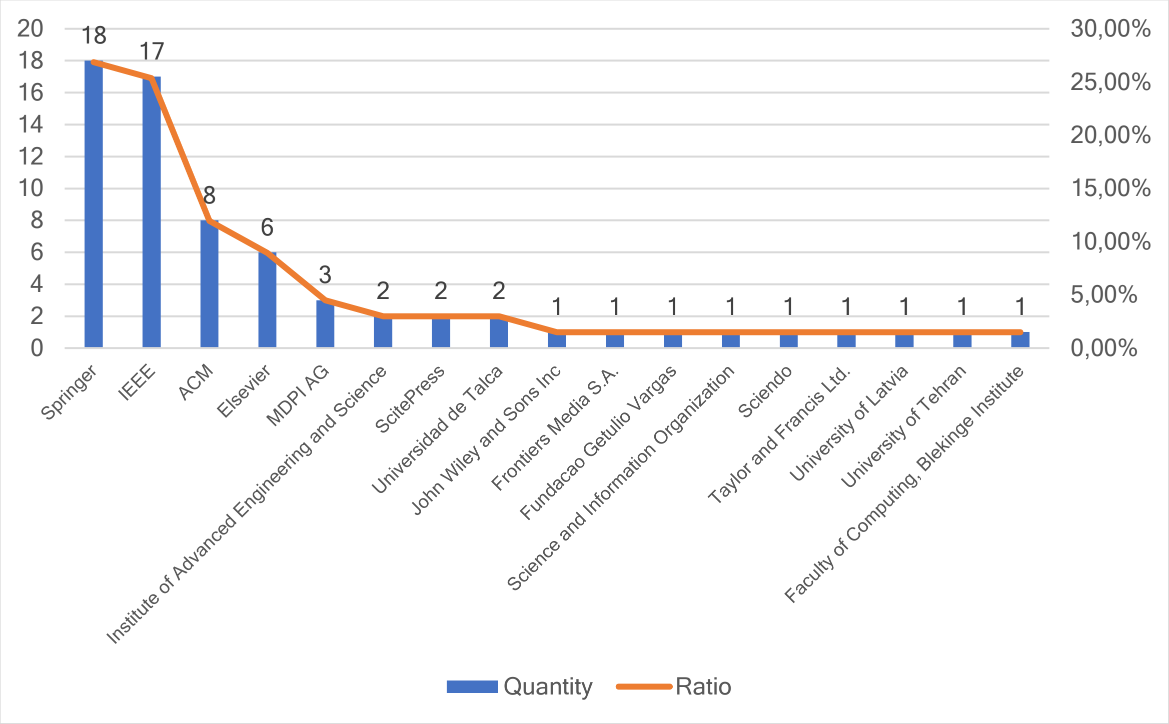


Fig. 11. Paper distribution per publisher (Topic: Big Data and quality models).

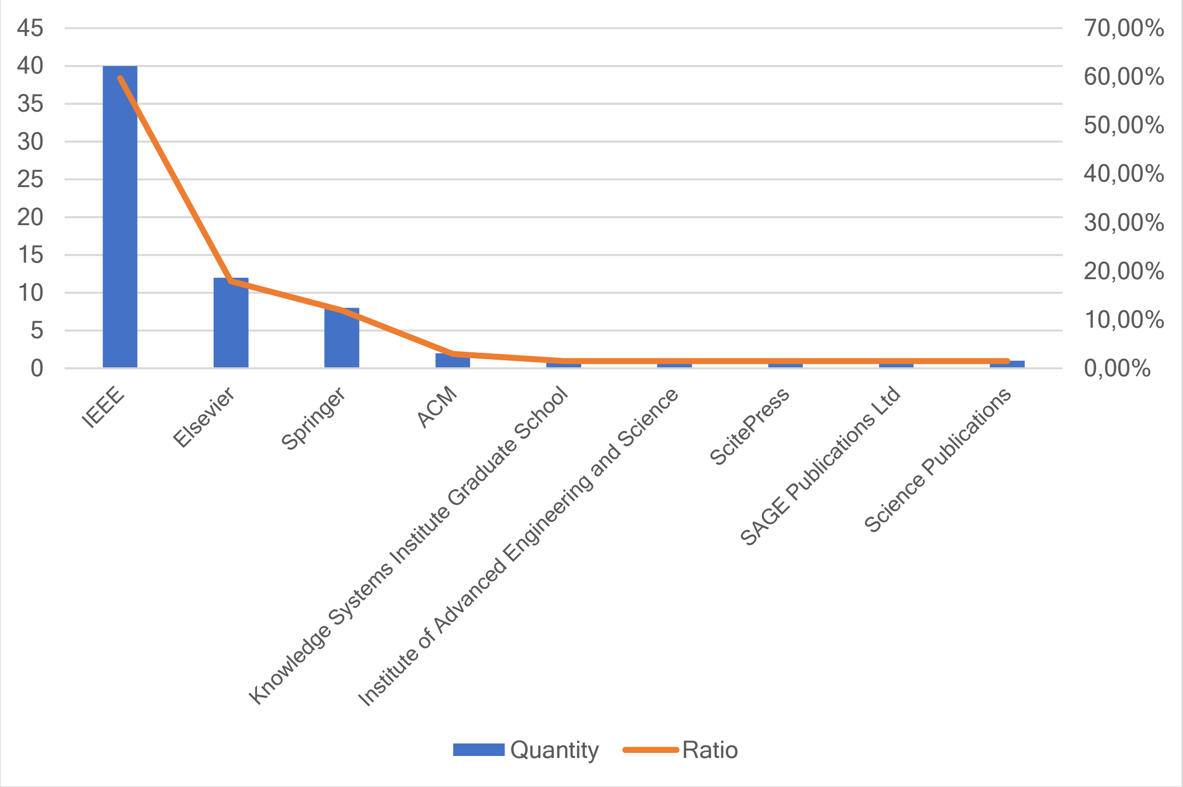


Fig. 12.Paper distribution per publisher (Topic: Big Data issues and challenges).

Some papers were better suitable to answer the SMS questions because their impact on the key topics. In next section, the research questions are answered.

3.4 Answering the research questions

3.4.1 Answering RQ1

What are the main issues and challenges detected on big data applications from 2010 to 2020?

There are two points of view to answer to this question. One of them is related with the V's that characterized Big Data, and it should be considered because those characteristics can be seen as challenges that need to be solved. A total of 15 characteristics has been found from 48 papers consulted in which these terms were addressed. Fig. 13 presents the distribution of Big Data characteristics, where 5 of them stand out from the rest: Volume, Velocity, Variety, Veracity and Value.

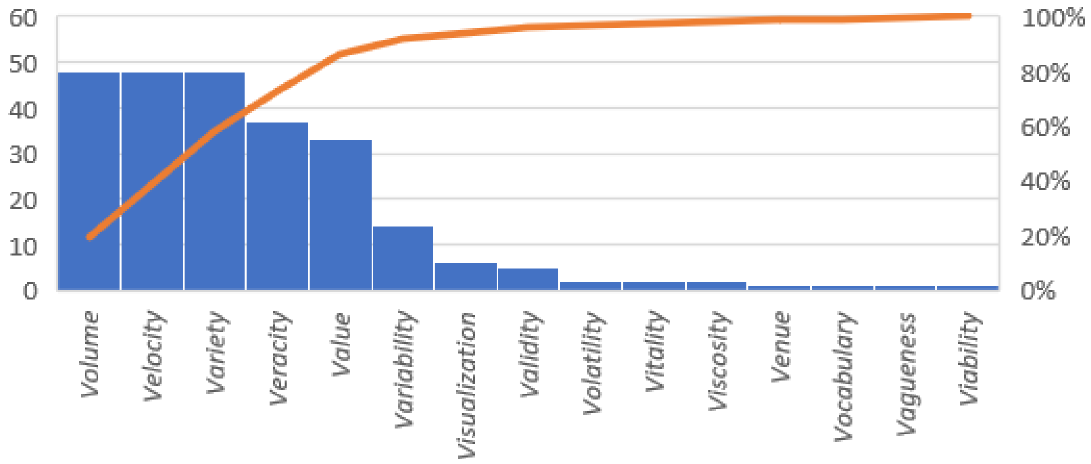


Fig. 13 Distribution of Big Data characteristics

Those characteristics are suitable to understand the nature of Big Data. Some authors like (Katal, Wazid, & Goudar, 2013) have identified them as properties associated with Big Data. In (Chandarana & Vijayalakshmi, 2014) are viewed as commonly used aspects that can characterized Big Data. For (Chang, y otros, 2016) are explained as major factors in Big Data systems, perhaps seen as challenges in an underhanded manner and for (Patel, 2019) are seeing as the 5Vs that often characterized Big Data.

As can be seen the Vs in Big Data are presented with different nomenclatures, but in practice what they represent are challenges for projects trying to implement a Big Data application. For this research they will represent the starting point to identify the challenges and main issues encountered in Big Data applications.

The second point of view is related with what authors call Big Data issues and challenges. To support this point, 30 different papers have been selected which explore one or more approaches related with the issues and challenges being faced in Big Data applications. After reviewing these papers, a total amount of 52 different issues were obtained. Some of them are related only to data quality aspects. Fig 14 represents the distribution of the complete list of Big Data issues compiled from the papers consulted.

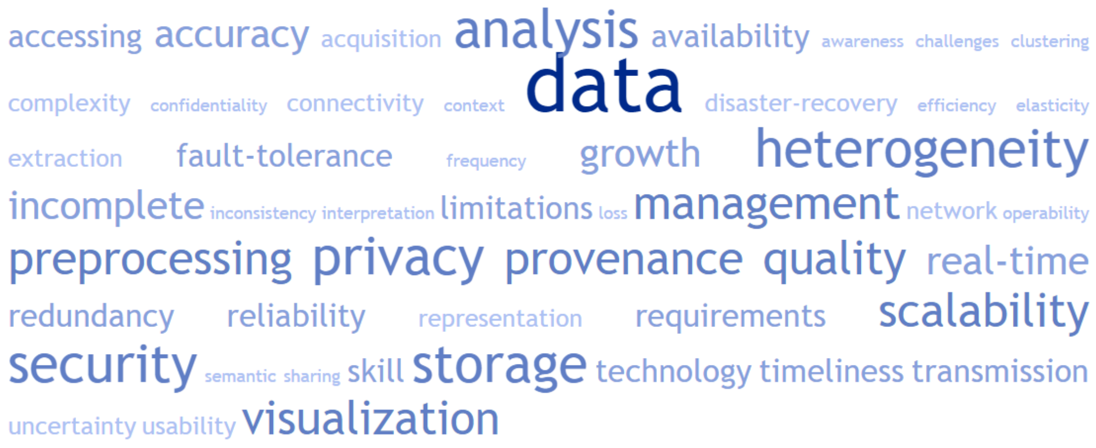


Fig. 14 Cloud of tags produced from joining Big Data issues

Some Big Data issues were considered for some authors in a context-dependent or at the implementation of a Big Data system in a specific organization. Those issues will not be taken into consideration for the present study. Fig 15 represents the distribution of the 24 Big Data issues and challenges most mentioned among the authors. Other 28 issues were dismissed because the low level of occurrence.

Some authors have contributed elements of greater relevance to this research in Big Data issues and challenges than others. This is the case of (Katal, Wazid, & Goudar, 2013), (Chaudhari & Srivastava, 2016), (Patgiri, 2018), (Wani & Jabin, 2018) and (Ashabi, Sahibuddin, & Haghighi, 2020).

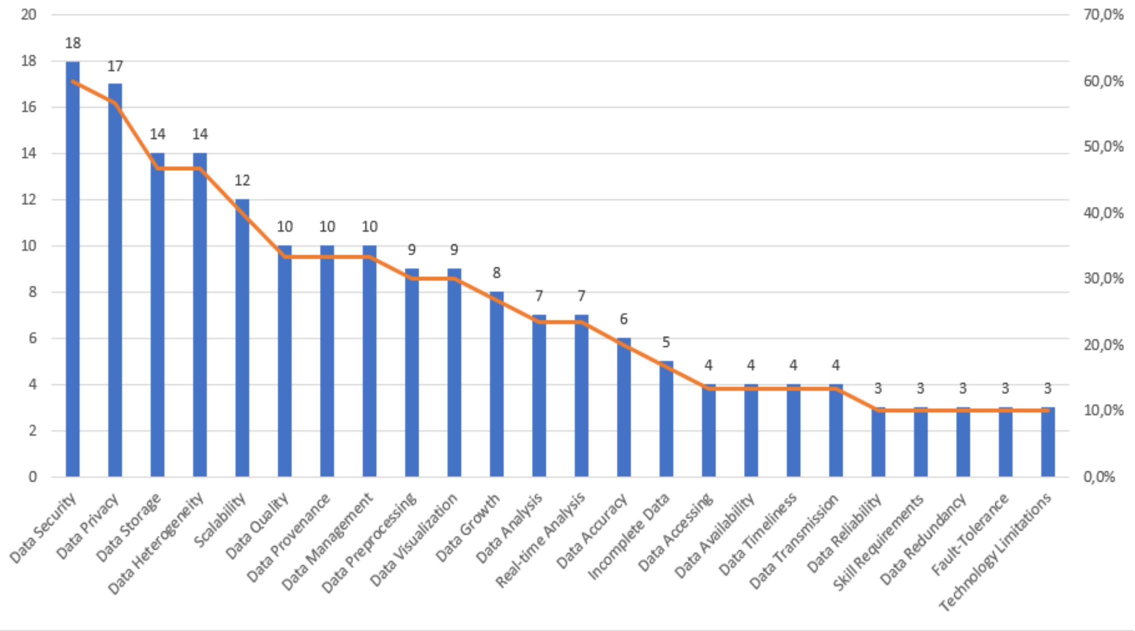


Fig. 15 Distribution of most mentioned Big Data issues

3.4.2 Answering RQ2

What were the quality characteristics, dimensions or quality factors applied to Big Data applications that have been identified by the authors in their publications from 2010 to 2020?

The starting point for analyze the Big Data quality factors or dimensions remain in the ISO/IEC 25012:2008 which split the Data Quality Model characteristics into Inherent and System dependent proposing 15 characteristics. In the present SMS a total of 57 characteristics, factors or dimensions have been found including those that belong to the ISO/IEC 25012:2008. These characteristics can be visualized from the cloud of tags perspective presented in Fig 16.

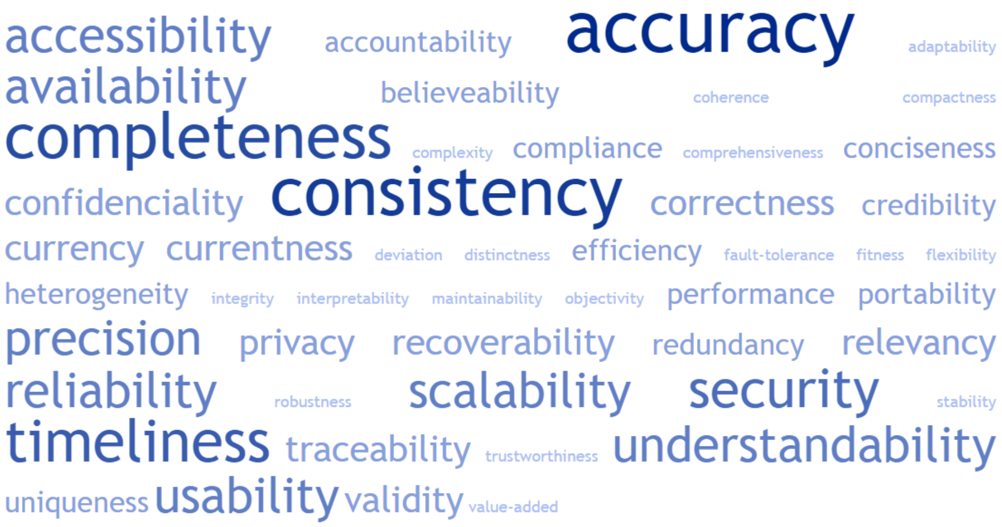


Fig. 16 Cloud of tags produced from joining the total of 57 Data Quality characteristics

Note that there are at least 12 characteristics that clearly stand out from the rest: Accuracy, Consistency, Completeness, Timeliness, Security, Precision, Usability, Scalability, Accessibility, Understandability, Availability and Reliability. However, this number is very short and not so representative so it has been decided to include the 32 most mentioned which can be visualized in Fig 17.

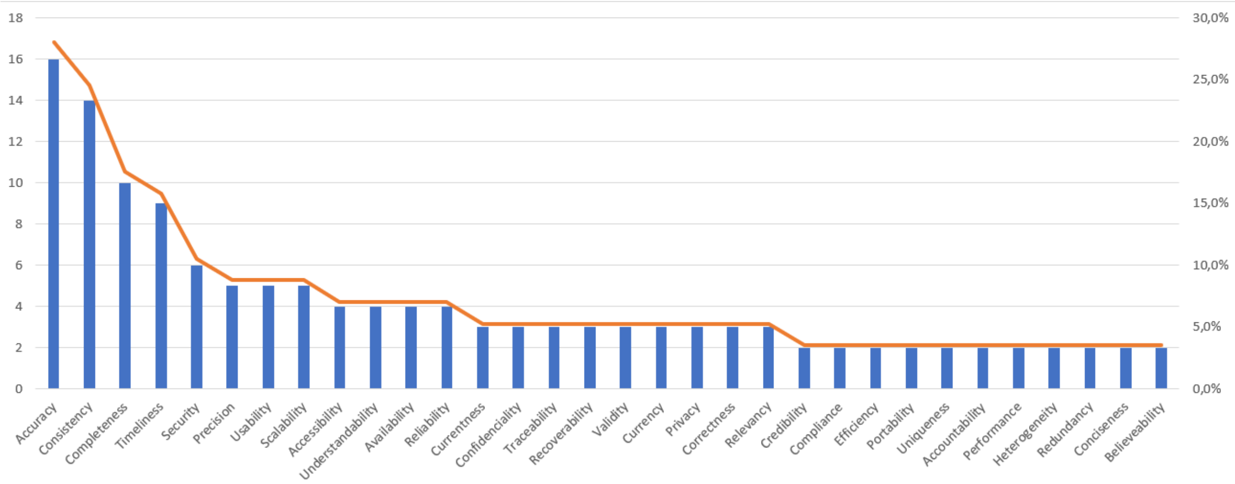


Fig. 17 Distribution of 32 most mentioned Big Data characteristics.

Comparing the characteristics mentioned at starting point (the ISO/IEC 25012:2008) with the findings in the current SMS, some differences are found and shown in Fig 18.

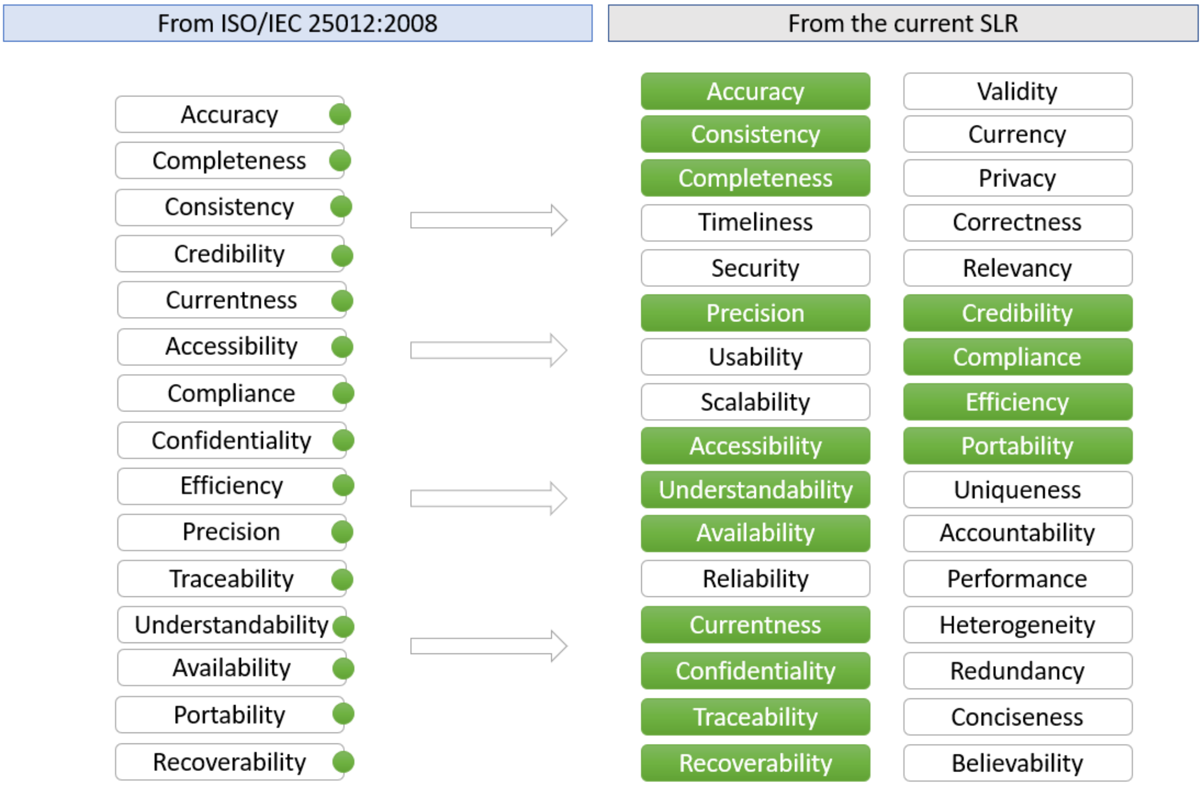


Fig. 18 Differences between the findings of the SMS and the ISO/IEC 25012:200 related to Big Data characteristics

Highlight as a good result that the 15 characteristics mentioned in the ISO are presented in the SMS findings, but not at the order expected. Many authors give more relevance to characteristics such as Timeliness, Security, Usability, Scalability and Reliability than others proposed by the ISO such as Credibility, Compliance, Efficiency and Portability.

3.4.3 Answering RQ3

What quality models related to Big Data applications have been proposed from 2010 to 2020 and which quality characteristics were defined as part of these models? Which metrics have been applied?

The study has revealed that 12 different quality model types has been published in the last 10 years, the most commons are those related with measuring Data Quality, Service Quality, Big Data Quality and Quality-In-Use. A complete distribution of these quality models can be viewed in Fig 19. It is not a surprise that the largest number of quality models proposed are those related to measuring the quality of the data, representing almost half of all models found.

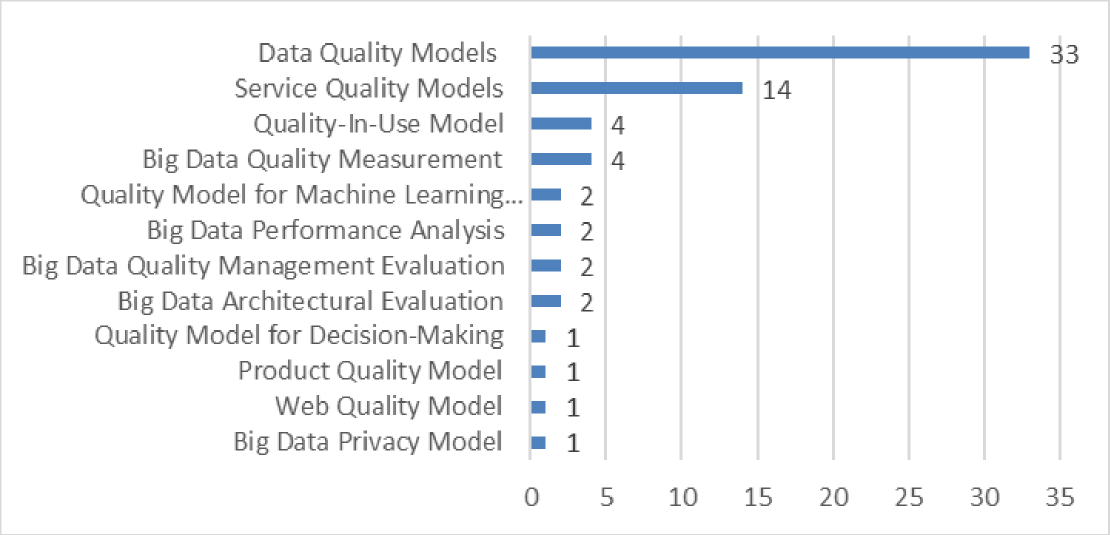


Fig. 19. Big Data quality models distribution per model type

Regarding the quality characteristics for these quality models, the study have revealed that authors have identified different quality characteristics depending upon the focus of the quality model inside Big Data context.

Data Quality Models: are defined as a set of relevant attributes and relationships between them, which provides a framework for specifying data quality requirements and evaluating data quality. Represents data quality dimensions and the association of such dimensions to data. Good examples of those models are presented in (Immonen, Paakkoneen, & Ovaska, 2015), (Vetrò, Canova, Torchiano, & Minotas, 2016), (Fernández, Jedlitschka, Guzmán, & Vollmer, 2018), (Oliveira, Oliveira, Batista, & Lóscio, 2018), (Talha, Kalam, & Elmarzouqi, 2019) and (Jarwar & Chong, 2020). Fig. 20 shows the categories that can be used to group the different data quality dimensions presented in the quality models.

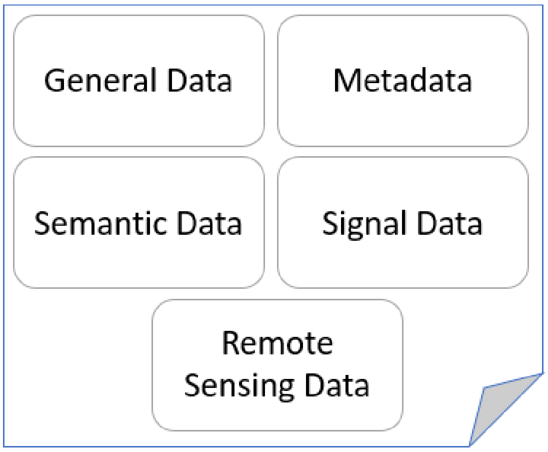


Fig. 20. Categories founded in the SMS that groups the data quality dimensions presented in the quality models.

Quality dimensions are presented in 28 from 33 related papers with data quality models. The most common dimensions for general data quality are:

- Completeness: characterizes the degree to which data have values for all attributes and instances required in a specific context of use. Also, data completeness is independent of other attributes (data may be complete but inaccurate).

- Accuracy: characterizes the degree to which data attributes represent the true value of the intended attributes in the real world, like a concept or event in a specific context of use.
- Consistency: characterizes the degree to which data attributes are not contradicted and are consistent with other data in a context of use.
- Timeliness: characterizes the latest state of a data attribute and its period of use.

In addition, for those quality models where the attention was focused on measuring the quality of metadata, in (Immonen, Paakkoneen, & Ovaska, 2015) the quality dimensions identified are believability, corroboration, coverage, validity, popularity, relevance, and verifiability. Other four dimensions are included apart from existing ones to Semantic Data (Jarwar & Chong, 2020) which are objectivity, reputation, value added and appropriate amount of data. For Signal Data (Kirchen, Schutz, Folmer, & Vogel-Heuser, 2017), other dimensions were identified such as availability, noise, relevance, traceability, variance, and uniqueness. Finally, other two dimensions were included for Remote Sensing Data (Barsi, y otros, 2019) which are resolution and readability.

It should be noted that quality dimensions proposed in each of these quality models refer to quality aspects that need to be verified by them in the specific context of use.

Service Quality Models: are used to describe the way on how to achieve desired quality in services. This model measures the extent to which the service delivered meets the customer’s expectations. Good examples of these models are presented in (Immonen, Paakkoneen, & Ovaska, 2015), (Ali, Hamilton, Thevathayan, & Zhang, 2018), (Basso, Silva, & Moraes, 2019), and (Jarwar & Chong, 2020). The quality dimensions collected from those papers where a service quality model is proposed on different context and domains are presented in Fig. 21.

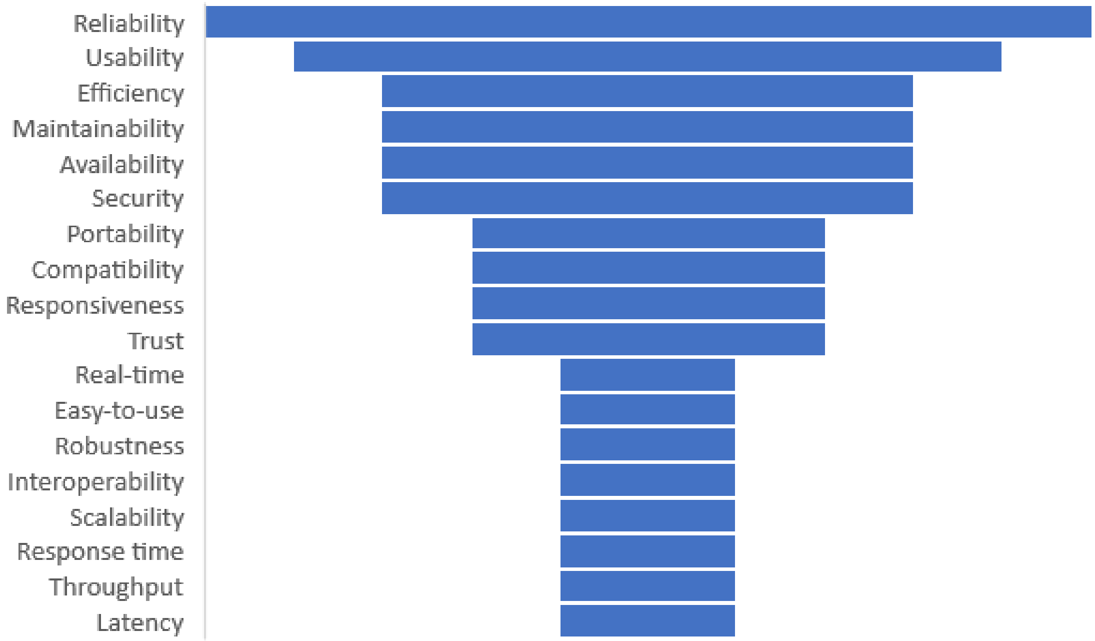


Fig. 21. Quality dimensions for Service Quality Models.

Note that in this case, having studied the quality characteristics of the ISO/IEC 25012, other quality dimensions are included for this type of quality models like usability, reliability, efficiency, maintainability, and security.

Quality-In-Use Models: Defines the quality characteristics that the datasets that are used for a specific use must present to adapt to that use. In this research two papers were found that present such type of models (Caballero, Serrano, & Piattini, 2014) and (Merino, Caballero, Rivas, Serrano, & Piattini, 2016), other papers discuss about them. These models are focused mainly in two dimensions: Consistency and Adequacy represented in Fig. 22.

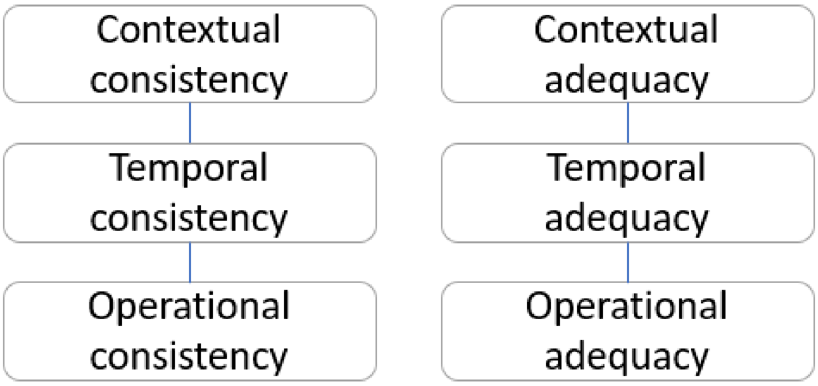


Fig. 22. Quality dimensions in Quality-In-Use Models

It should be noted that, depending upon the quality characteristics that need to be evaluated and the context of use, a different model should be applied. In those models, these two dimensions are presented as:

- Consistency: The capability of data and systems of keeping the uniformity of specific characteristics when datasets are transferred across the networks and shared by the various types of consistencies.
- Adequacy: The state or ability of being good enough or satisfactory for some requirement, purpose or need.

Big Data Systems Quality Models: There isn't a general definition for this types of models because the enormous number of different kinds. In this research will be defined as quality models applied to the context of Big Data viewed at a high level. Good examples of these models are presented in (Serhani, Kassabi, Taleb, & Nujum, 2016), (Kläs, Putz, & Lutz, 2017), (Helfert & Ge, 2018), and (Omidbakhsh & Ormandjieva, 2020). The quality dimensions presented in these quality models are specified in Table 3, those can be separated into three groups:

- Dimensions for Big Data value chain
- Dimensions for Non-Functional requirements in Big Data Systems
- Dimensions for measuring Big Data characteristics.

Big Data Value Chain	Big Data Non-Functional Requirements (NFRs)	Big Data Characteristics
Timeliness	Scalability	Volume
Accuracy	High Performance Computing	Velocity
Completeness	Modularity	Variety
Consistency	Consistency	Veracity
	Security	Valence
	Real-time Operations	Value
	Inter-operability	Volatility
	Availability	Vitality
		Vincularity

Table 3. Quality dimensions for Big Data Systems Quality Models

Regarding the metrics applied, it wasn't expected to find that more than 75% of papers which present or discuss a quality model did not discuss quality metrics associated. In section 4.1.2 the identified metrics are presented for the current investigation.

3.4.4 Answering RQ4

For which Big Data specific context have these quality models been proposed?

The majority of quality models proposed can be applied to any Big Data project without distinguishing between the different types of Big Data applications. A number of 8 approaches have been identified as possible field of application which can be regarded in Table 4.

There is a differentiation between general Big Data projects and Open Data projects mainly because the dimensions presented for those Open Data are related such as free access, always available, data conciseness, data and source reputation, and objectivity among others, are specially required in Open Data projects. For Big Data Analytics, Decision Making and Machine Learning projects there is no such great differentiation with other Big Data projects, only in the case where non-functional requirements must be measured that are specific to the required purpose.

Context	Quantity	Ratio
Any Big Data Project	50	74,63%
Cloud Projects	6	8,96%
Social Information Services (Facebook, Twitter, etc.)	2	2,99%
Big Data Analytics	2	2,99%
Open Data Projects	2	2,99%
Decision Making process	2	2,99%
Machine Learning Projects	2	2,99%
Smart Cities Ecosystem Project	1	1,49%

Table 4. Quality model distribution per Big Data context

As was revealed by this study, there is no information about a quality model which were developed, applied, and measured to a feature prediction system.

3.4.5 Answering RQ5

Have Big Data quality models been proposed to be applied to any type of Big Data application or by considering the quality characteristics required in specific types of Big Data applications?

A Big Data application (BDA) processes a large amount of data by means of integrating platforms, tools, and mechanisms for parallel and distributed processing. As was presented in Table 4, the majority of quality models proposed (74,63%) can be applied to any Big Data project and only a few studies were developed specifically for:

- Big Data Analytics can be seen as the new engine of economic and social value creation. Can be defined as analytics applied on data available with the focus of empowering software development individuals and teams to gain and shared insight from their data to make better decisions. Can be used in scenarios to assess concrete problems like: processing data to predict overall project effort (making project estimations more reasonable), processing security data to identify indicators of software vulnerabilities, among others. A good approach was presented in (Vetrò, Canova, Torchiano, & Minotas, 2016).
- Machine Learning projects are a branch of Artificial Intelligence (AI) focusing on developing applications that can learn from data at the same time of improving their accuracy without being programmed to do so. Papers (Rudraraju & Boyanapally, 2019) and (Santhanam, 2020) should be consulted for further analysis.

This could be means that researchers are not interested on develop a quality model for a specific kind of Big Data application, instead a general quality model is proposed focusing on a general topic like assuring the quality of data, the Quality-in-use, the quality of services involved, etc.

Other interesting finding is about the type of quality model which is intended to use. Because the diversity of quality models found in the context of Big Data it should be noted that before applying a specific quality model to a specific kind of Big Data application some considerations are needed, such as:

1. The focus area which is planned to evaluate.
2. The considerations and restrictions of the selected Big Data application.

For instance, at the current investigation a data quality model is a very good candidate with a great potential to improve the quality of a Feature Prediction System. This is because it has been discussed the importance of assure a good data quality before starting the prediction process.

3.5 Threats to validity

The main threats to validity this mapping study are:

- **Selection of search terms and digital libraries**. The search was done into some digital libraries and to complete our study other libraries should be included such as: Springer, Google scholar. In addition, because Big Data is an industrial issue it is recommended to include gray literature search (Garousi, Felderer, & Mäntylä, 2019) and achieve a Multivocal Literature Review (MLR).
- **Selection of studies**. Could be a better solution to apply other exclusion criteria such as the quality of papers. For example, if the results have been validated and compared to other studies.
- **Quality model categorization**. As a result of the small sample of papers in which quality models are not related to data quality, it is an arduous task to obtain sample quality metrics and quality dimensions for those Big Data quality models. With the amplification of the current study more samples could be obtained to support this task.

4. The Data Quality Model adapted

In section 3.4.3.5 the results of the current SMS have been revealed and in section 3.4 the selected research questions were answered, where different types of quality models in the context of Big Data were obtained. At the current chapter the data quality models are filtered and analyzed to adapt a specific data quality model to be applied in Feature Prediction Systems.

4.1 Exposing Data Quality Models

A set of Data Quality Models have been obtained and analyzed as presented in Table 5. After a depth reading some conclusions can be obtained and discussed to adapt a data quality model with an appropriate structure of data quality characteristics and metrics for data quality evaluation.

Reference	Title	Publish year	Publisher	Place
(Ciancarini, Poggi, & Russo, 2016)	Big Data Quality: A Roadmap for Open Data	2016	Institute of Electrical and Electronics Engineers Inc.	International Conference on Big Data Computing Service and Applications, BigDataService
(Kirchen, Schutz, Folmer, & Vogel-Heuser, 2017)	Metrics for the evaluation of data quality of signal data in industrial processes	2017	Institute of Electrical and Electronics Engineers Inc.	International Conference on Industrial Informatics, INDIN
(Wang, Wen, & Zheng, 2019)	Research on Assessment and Comparison of the Forestry Open Government Data Quality Between China and the United States	2019	Springer	International Conference on Data Science, ICDS

(Tepandi, y otros, 2017)	The data quality framework for the Estonian public sector and its evaluation: Establishing a systematic process-oriented viewpoint on cross-organizational data quality	2017	Springer	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Volume 10680 LNCS
(Cedillo, Valdez, Delgado, & Cabrera, 2020)	A Data as a Service Metamodel for Managing Information of Healthcare and Internet of Things Applications	2020	Springer Science and Business Media Deutschland GmbH	Conference on Information and Communication Technologies of Ecuador, TICEC
(Fagúndez, Fleitas, & Marotta, 2015)	Data streams quality evaluation for the generation of alarms in health domain	2015	Springer	International Workshops on Web Information Systems Engineering, IWCSN
(Ge & Dohnal, Quality Management in Big Data, 2018)	Developing the quality model for collaborative open data	2020	Elsevier	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES
(Jarwar & Chong, 2020)	Web objects based contextual data quality assessment model for semantic data application	2020	MDPI AG	Applied Sciences (Switzerland). Volume 10
(CICHY & RASS, 2019)	An overview of data quality frameworks	2019	Institute of Electrical and Electronics Engineers Inc.	IEEE Access. Volume 7
(Wan, Shi, Gao, Chen, & Hua, 2015)	A general framework for spatial data inspection and assessment	2015	Springer	Earth Science Informatics. Volume 8
(Fernández, Jedlitschka, Guzmán, & Vollmer, 2018)	A quality model for actionable analytics in rapid software development	2018	Institute of Electrical and Electronics Engineers Inc.	Euromicro Conference on Software Engineering and Advanced Applications, SEAA
(Liu, Chen, & Cai, 2018)	Application of requirement-oriented data quality evaluation method	2018	Institute of Electrical and Electronics Engineers Inc.	International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD

(Vostrovsky & Tyrychtr, 2018)	Consistency of Open Data as Prerequisite for Usability in Agriculture	2018	Sciendo	Scientia Agriculturae Bohemica. Volume 49
(Baillie, Edwards, & Pignotti, 2015)	Qual: A provenance-aware quality model	2015	Association for Computing Machinery	Journal of Data and Information Quality. Volume 5
(Nikiforova, 2020)	Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment	2020	University of Latvia	Baltic Journal of Modern Computing. Volume 8
(Musto & Dahanayake, 2019)	Integrating data quality requirements to citizen science application design	2019	Association for Computing Machinery	International Conference on Management of Digital EcoSystems, MEDES
(Parra, Parody, Vaca, Caballero, & López, 2019)	DMN for Data Quality Measurement and Assessment	2019	Springer	International Workshops on AI4BPM, BP-Meet-IoT, BPI, BPMinDIT, BPMS2, DEC2H, MIEL, PM-DiPro, PODS4H, PQ, SPBP, VEnMo
(Castillo, y otros, 2018)	DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data	2018	MDPI AG	Sensors (Switzerland)
(Barsi, y otros, 2019)	Remote sensing data quality model: from data sources to lifecycle phases	2019	Taylor and Francis Ltd.	International Journal of Image and Data Fusion
(Cappiello, y otros, 2020)	Improving Health Monitoring with Adaptive Data Movement in Fog Computing	2020	Frontiers Media S.A.	Frontiers in Robotics and AI
(Davoudian & Liu, 2020)	Big Data Systems: A Software Engineering Perspective	2020	Association for Computing Machinery	ACM Computing Surveys
(Cappiello, Samá, & Vitali, 2018)	Quality awareness for a Successful Big Data Exploitation	2018	Association for Computing Machinery	International Database Engineering & Applications Symposium
(Oliveira, Oliveira, Batista, & Lóscio, 2018)	Towards a meta-model for data ecosystems	2018	Association for Computing Machinery	Annual International Conference on Digital Government Research: Governance in the Data Age

(Taleb, Serhani, & Dssouli, 2019)	Big Data Quality: A Data Quality Profiling Model	2019	Springer	World Congress on Services, SERVICES
(Taleb, Serhani, & Dssouli, 2018)	Big Data Quality Assessment Model for Unstructured Data	2018	Institute of Electrical and Electronics Engineers Inc.	International Conference on Innovations in Information Technology, IIT
(Vetrò, Canova, Torchiano, & Minotas, 2016)	Open data quality measurement framework: Definition and application to open government data	2016	Elsevier	Government Information Quarterly
(Behkamal, Kahani, Bagheri, & Jeremic, 2014)	A metrics-driven approach for quality assessment of linked open data	2014	Universidad de Talca	Journal of Theoretical and Applied Electronic Commerce Research
(Rudraraju & Boyanapally, 2019)	Data Quality Model for Machine learning	2019	Faculty of Computing, Blekinge Institute of Technology	Faculty of Computing, Blekinge Institute of Technology
(Talha, Kalam, & Elmarzouqi, 2019)	Towards a powerful solution for data accuracy assessment in the big data context	2020	Science and Information Organization	International Journal of Advanced Computer Science and Applications
(Immonen, Paakkoneen, & Ovaska, 2015)	Evaluating the Quality of Social Media Data in Big Data Architecture	2015	Institute of Electrical and Electronics Engineers Inc.	IEEE Access

Table 5. Data quality models in the context of Big Data

Some primary rules should be evaluated when analyzing these data quality models:

1. A valid data quality model defines a set of data quality characteristics which are applicable to every data set.
2. The data quality characteristics defined should be also classified by means of their specific roles within the quality model. This could help to determine the required definition of metrics for numerical indicators which will evaluate the data quality objectively.

3. When possible, should be helpful to define aggregation rules to combine the numerical indicators to a single key indicator for data quality. The numerical indicators should be in the same dimension.

4.1.1 Identifying Data Quality Characteristics

In some cases, quality characteristics do influence the data quality but are not qualified to serve as numerical indicators for data quality since their influence is mediate. The data quality characteristics applied as part of the data collection stage are denominated as Upstream. On the other hand, the data quality characteristics within the data processing stage are denominated as Downstream.

Data quality characteristics identified as numerical indicators are: appropriate amount of data, data noise, data variance, data completeness. Their impact on data quality is immediate (Kirchen, Schutz, Folmer, & Vogel-Heuser, 2017). These authors also identified two classifications for data quality characteristics: Upstream and Downstream.

Data quality characteristics denominated as Upstream and classified as influencing factors have a mediate influence on data quality. These characteristics are: availability, accessibility, recoverability, completeness, uniqueness, objectivity, portability, and traceability.

Data quality characteristics denominated as Downstream and classified as influencing factors, appear in data processing and requires measures, which explains their impact on data quality measurement. These characteristics are: data accuracy, data credibility, data consistency, data relevance and data compliance.

4.1.2 Identifying metrics for quality assessment

Choosing the adequate metric would be crucial to apply an efficient quality assessment. In (Kirchen, Schutz, Folmer, & Vogel-Heuser, 2017), (Liu, Chen, & Cai, 2018), (Musto & Dahanayake, 2019), and (Ge & Lewoniewski, 2020), some metrics from the analyzed data quality models are presented. At the current investigations those metrics are analyzed, summarized, and interpreted as follows:

- **Completeness**: Because is required knowledge on the frequency based on metadata as well as user knowledge of what values are incomplete. A metric example to calculate completeness is:

$$Comp = \frac{\text{Total Non Missing Values}}{\text{Total Expected Values}}$$

- **Uniqueness**: Because records need to be not duplicated. A metric example is:

$$Uniq = \frac{\text{Total Unique Values}}{\text{Total Expected Values}}$$

- **Semantic Consistency**: Because some mandatory relationships are identified between tuples. A metric example is:

$$Sem = \frac{\text{Total Unique Values}}{\text{Total Expected Values}}$$

- **Accuracy**: Because data accurate is needed to properly apply a BDA algorithm. Some metrics example are:

$$Outliers = \frac{\text{Data values outliers}}{\text{Total Expected Values}}$$

$$Wrong\_fields = \frac{\text{Wrong values in fields}}{\text{Total Expected fields}}$$

- **Objectivity**: Because data could be collected by using wrong procedures. A metric example is:

$$Object = \frac{\text{Imprecise data values}}{\text{Total Expected values}}$$

- **Believability**: Because data collected has to be accepted as real and credible. A metric example is:

$$Object = \frac{\text{Mismatch data values}}{\text{Total Expected values}}$$

- **Consistency:** Because the same data should be consistent among different sources. A metric example is:

$$Cons = \frac{\text{Corrupted data values}}{\text{Total Expected data values}}$$

  ** Corrupted data values: the number of tuples that have data constraints which means, data value from different sources refers to the different real-world entities.

Other metrics would be designed to be evaluated through the application of questionnaires to a group of designated experts by survey sessions. The results obtained are represented as average points based on the survey results. These metrics are:

- **Source Confidentially**: Because could be necessary to understand data acquisition methods as well as verification methods.
- **Value Traceability**: Because could be necessary to check whether the corresponding function is enabled to record the operation registries of user operation data. On the other hand, to understand whether to record data modification history.
- **Data Understandability**: Because could be necessary to check whether the meaning of data can be understood by the value of data.
- **Periodic Backup (Recoverability)**: Because could be necessary to validate whether regularly data back up and backup strategy are suitable or not.

- **Usability**: Because could be necessary to recover final user feedbacks to evaluate if the product accomplish with user expectations and is fitness for use.

In the current investigation, some data quality metrics could not be applied, these are:

- Consistency: Because is expected to use a unique dataset to evaluate the quality model.
- Source Confidentially, Value Traceability, Data Understandability and Periodic Backups: Because a domain experts are not designated to evaluate the quality model.
- Usability: Because final users or domain experts are not designated to evaluate if the product is usable or not.

4.1.3 Evaluating data quality through business rules

There is another way presented in (Parra, Parody, Vaca, Caballero, & López, 2019) to evaluate data quality assessment by defining some business rules that need to be accomplish. Some examples of business rules describing syntactic and/or semantic data requirement are as follows:

- BR01: The attribute 'Name' contains a string no longer than 256 characters.
- BR02: The attribute 'Year-of-Birth' must be a positive number between 1890 and 2021.
- BR03: The attribute 'citizen-identification' must be only numerical.

Based on these rules, the data quality characteristics are applied and evaluated. Examples:

Completeness: A value is complete when meets the business rules BR01 and BR03.

Accuracy: A value is accurate when meets the business rules BR01, BR02 and BR03.

Consistency: A value is consistent when meets the business rule BR03.

When considering applying a set of quality metrics to measure the model, these kind of metrics are suitable if there are business rules defined previously as part of the project and those business rules can be evaluated.

4.2 Adapting a data quality model for Feature Prediction Systems

For some Big Data applications such as Feature Prediction Systems in (Tao & Gao, 2016) a set of quality characteristics have been identified that can be applicable to this type of Big Data application:

- **System accuracy:** This characteristic is used to evaluate if the system is systematic error free and random error free with consistent results. In the context of Feature Prediction Systems when predicting world entities, in some cases approximated solutions could be obtained. In those cases, the prediction is uncontrollable and may affect actions or behaviors.
- **System consistency:** This is a characteristic used to evaluate different perspectives of the consistency in the system. In some cases, the application do not produce a single and unique correct output for a set of inputs. In this scenario it is hard to properly determine the expected behavior of the software. To avoid that, users with a domain-specific expertise should provide support to validate the consistency.
- **System correctness:** As the name is stated, this characteristic evaluates the correctness of the Big Data application, which is a tedious and difficult task because the characteristics of Big Data. Since Feature Prediction Systems are developed to make predictions about real-world entities, it is not an easy task to obtain the correct output of these systems. In this context, the correctness is related to the prediction model. To evaluate the system correctness, the capability of predictions in the specified conditions and environments should be having in count.

- **System duration:** This characteristic indicates the expected Feature Prediction period and can be measured as how well the input data is up-to-date and whether data remains accurate when changes could impact date and time values.
- **System stability:** This characteristic validates the Feature Prediction System stability when changes are produced in the input data or in the environment. The characteristic measures for example, if the prediction capability of the system remains stable when some changes are produced in statistical data when it is obtained from different timeframes.
- **System usability:** This characteristic refers to how well the Feature Prediction System can be used. The main issue to it is related with the subjectivity that different developers and users can have about user experience.
- **System performance:** This characteristic is used to evaluate how well the data used for Feature Prediction Systems are designed, structured, collected, generated, stored, and managed.
- **System reliability:** This characteristic helps to evaluate the durability of the Feature Prediction System when the required function is performed in a specific period of time and under certain conditions.
- **System scalability:** This characteristic refers to the fact that a Feature Prediction system should be able to support large data sets at present and future.
- **System security:** This characteristic helps to evaluate the security of the Feature Prediction System in various perspectives at different levels.

Actually, in no other paper consulted a set of quality characteristics for Feature Prediction Systems are proposed or even analyzed. This makes a difficult task to adapt a quality model to this kind of Big Data application. For this reason, other aspects of the current investigation need to be in consideration when adapting the quality model for Feature Prediction Systems.

In the current investigation was identified a set of 32 data quality dimensions to be applicable to evaluate the data quality of a Big Data application. Have been stated that different approaches in the Big Data domain need to be managed with the application of different data quality dimension and quality metrics. Finally, considering the limited number of papers related with the evaluation of data quality in Feature Prediction Systems, and the results of the current SMS, an adapted quality model have been identified and is proposed in Fig 23.

The model is proposed to be applied in the initial stages of data processing which are: data collection and data preprocessing. At these stages, data is prepared for the Feature prediction algorithm to be applied and is viewed as crucial to determine the success of the algorithm.

The quality dimensions are split into data quality characteristics and system quality characteristics. Measuring data quality is an important aspect but also when using a Feature Prediction System algorithm, a huge amount of machine resources are consumed to execute all the steps, so consider measuring and watch system performance, duration or stability are also necessary.

Metrics defined will help to evaluate the application of this quality model and are focused to determine whether the model accomplish with data and system quality.

The quality characteristics proposed in (Tao & Gao, 2016) do not cover all of the identified quality characteristics for Feature Prediction Systems. For some characteristics such as: system duration, system usability, system scalability and system security is mandatory to explore and adopt other quality models or refine the existing ones. For other quality characteristics such as: Confidentiality, Traceability, Understandability and Recoverability it is mandatory to elaborate adequate checklists and define domain experts which could evaluate the applicability of those characteristics. Finally, for evaluating the Usability is mandatory to define user's experience checklist and domain experts which would implement and evaluate its efficacy into the system.

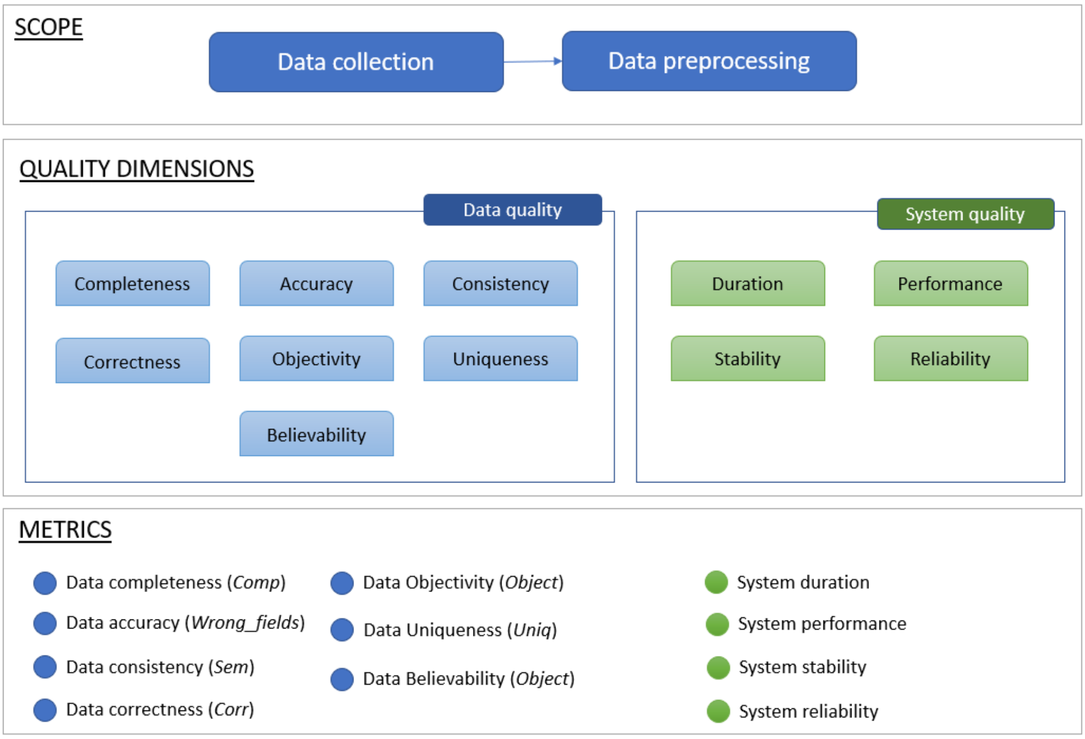


Fig. 23. A data quality model adapted for Feature Prediction Systems

5. Applying the Data Quality Model to a Feature Prediction System

5.1 Presenting the application example: New York City Taxi Trip Duration

The implementation of the quality model was made on the "New York City Taxi Trip Duration" project. This is a Kaggle challenging (Kaggle, 2020) which consist of constructing a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is released by the *NYC Taxi and Limousine Commission* which includes several variables. The datasets corresponds to the records collected from trips made in yellow taxis at the year 2016 in the New York city. In the dataset each row represents a trip made by one of the well-known yellow taxis of NY.

The data is extracted in a .csv file which contains 1458644 trip registries and 11 attributes. With this dataset the training tasks will be carried out by partitioning the dataset randomly to obtain a test dataset which will be used in the evaluation of the models already selected. Table 6 represents the attributes involved in the current example and their description.

Attribute	Description
id	Represents an identifier for each record trip done.
vendor_id	Represents a code indicating the service provider which provided the data.
pickup_datetime	Represents the start date and time in the taximeter registration.
dropoff_datetime	Represents the end date and time in the taximeter registration.
passenger_count	Represents the passenger number inside the vehicle. This value is introduced by the conductor.

pickup_longitude	Represents the longitudinal coordinate from the starting point of the trip.
pickup_latitude	Represents the latitudinal coordinate from the starting point of the trip.
dropoff_longitude	Represents the longitudinal coordinate from the arrival point of the trip.
dropoff_latitude	Represents the latitudinal coordinate from the arrival point of the trip.
store_and_fwd_flag	This variable indicates whether the travel record is previously stored in internal memory of the vehicle or not, before to send it to the server.
trip_duration	Represents the duration of the trip in seconds. This is the focus of the prediction.

Table 6. Description of the training dataset attributes.

This dataset is for public access and is available in one of the challenges of the Kaggle project: “The New York City Taxi Trip Duration” [1]. Data analysts and programmers could find the proposed dataset, other comments, and best solutions while joining the competition. Other datasets are also presented in this initiative which encourage participants to learn and work on better solutions for Big Data applications.

5.2 Applying the GQM methodology to the evaluation of the quality of the application under study

A process of data collection and preprocessing is done where some tasks need to be executed to accommodate the data before executing the algorithm. Some data transformation are:

[1] Source: https://www.kaggle.com/c/nyc-taxi-trip-duration/data

- **Modifying the values in the "trip_duration" column**: Modify this data of the training dataset so that contain the values "long" for a more than 15 minutes of trip duration and "short" for les than 15 minutes of trip duration.
- **Transforming categorical variables into numerical variables:** This is necessary to be able to use the algorithm.
- **Partitioning the dataset into training and validating sub-datasets:** It is recommended to use 70% of the data for training model and 30% for its validation.

At this point, the quality model is applicated. The quality dimensions used and finally evaluated are:

- Completeness
- Uniqueness
- Accuracy
- Objectivity
- Correctness
- Performance
- Stability
- Reliability

To evaluate the quality dimensions, the selected metrics in section 4.1.2 wre applied after the necessary data adequations. The process of quality assessment was done by working directly with the csv data file. The metrics were applied in the corresponding order presented in Table 7 which shows the final measures obtained from each metric.

Metric	Dimension	Result
Data completeness (Comp)	Completeness	95% for numerical attributes.

		98% for alphabetical attributes
Data accuracy (Wrong_fields)	Accuracy	98% for overall attributes
Data accuracy (Outliers)	Accuracy	4.2% of data outliers
Data correctness (Corr)	Correctness	93.1% for overall attributes
Data Objectivity (Object)	Objectivity	95.9% for overall attributes
Data Uniqueness (Uniq)	Uniqueness	99% for overall attributes
Data Believability (Object)	Believability	No sense to measure
System duration	Duration	No sense to measure
System Performance	Performance	2.2 hours for overall execution
System stability	Stability	100%
System Reliability	Reliability	100%

Table 7.The evaluation results of the Quality Model for Feature Prediction System

As can be seen in the results there are two quality characteristics that could not be properly applied. Believability hasn't sense to be included in the validation because was assumed that all rows contained data provided by the NYC Taxi and Limousine Commission, and no extra data was used. It didn't make sense to validate Duration either because the dataset used is up-to-date and is a registry of 2016 without any other dataset from previous years.

It is remarkable that the performance of the system wasn’t the expected and may be caused by the limited conditions of the server used. The stability and reliability of the system was two positive variables because the system was able to execute all tasks properly without interruptions or impediments. Finally, the overall quality assessment has some negative points in Correctness validation, because the huge number of invalid registries founded. Positive points for data completeness (for alphabetical attributes) and data uniqueness. In general, the quality of the dataset was acceptable.

6. Conclusions and Future Work

A SMS have been conducted to analyze and visualize the different quality models that have been proposed in the context of Big Data and the quality characteristics presented on each type of quality model. It has been found that different from what would have been thought, there is a considerable number of papers which do not present or partially discuss the quality metrics to evaluate the quality dimensions proposed in the model. Also, in the majority of studies the results of their research was not analyzed and compared with other similar studies.

This research have revealed that proposing, discussing, and evaluating new quality models in the context of Big Data is a topic that is currently receiving more attention from researchers and with the actual tendency we should expect an increase of papers related with quality models in Big Data context in the coming years.

Despite the minimal amount of available information about the Feature Prediction Systems and the quality characteristics and quality metrics related to them, after finalizing the study, a quality model has been adapted and presented in a case study for further analysis.

As first topic for future work is mandatory to be considered an in-depth review of the analyzed papers where common metrics, quality dimensions and quality models evaluations could be obtained for further analysis on each Big Data quality model type.

In the context of Big Data, most of the proposed quality models are designed for any Big Data application and they are not explicit in evaluating a specific type of Big Data application such as Feature Prediction Systems or Recommenders. Considering their different specificities to assess the expected quality in the final result when using these Big Data applications, we consider this as an open research topic.

Regarding the case application, a new execution of the selected algorithm for Future Prediction Systems should be performed in a server with better characteristics such as

RAM memory, Disk space and CPU because the server used was in a local machine with some limitations about memory, disk, and CPU. This case was executed by ingesting a single available dataset, so that some quality dimensions wasn't validated properly. In a new execution, is recommended to validate the model including different datasets from different sources and with data modified over time to include other quality characteristics into the quality assessment.

Continuing with the current investigation, could be positive to develop a quality framework which automatically ingest different datasets from different sources and reveals the results about the data quality by applying the adapted model. A validation with experts of the quality measures obtained as indicators should be also achieved.

Bibliography

Abdallah, M. (2019). Big data quality challenges. *International Conference on Big Data and Computational Intelligence (ICBDCI)*.

Aftab, U., & Siddiqui, G. F. (2018). Big Data Augmentation with Data Warehouse: A Survey. *IEEE International Conference on Big Data (Big Data)*, p. 2785 - 2794.

Alaoui, I. E., & Gahi, Y. (2020). Network Security Strategies in Big Data Context. *Procedia Computer Science2020Volume 175*, p. 730 - 736.

Alarcos Group. (2020). *Modelo Alarcos de Mejora de Datos v.3.0*. Retrieved from DQTeam: https://www.dqteam.es/index.php/mamd

Ali, K., Hamilton, M., Thevathayan, C., & Zhang, X. (2018). Big social data as a service: A service composition framework for social information service analysis. *International Conference on Web Services, ICWS*, p. 487-503.

Alkatheeri, Y., Ameen, A., Isaac, O., Nusari, M., Duraisamy, B., & Khalifa, G. S. (2019). The Effect of Big Data on the Quality of Decision-Making in Abu Dhabi Government Organizations. *International Conference on Data Management, Analytics and Innovation, ICDMAI*, p. 231-248.

Al-Sai, Z. A., & Abualigah, L. M. (2017). Big data and E-government: A review. *8th International Conference on Information Technology (ICIT). ISBN:978-1-5090-6333-8*, p. 580 - 587.

Al-Sai, Z. A., Abdullah, R., & Husin, M. H. (2020). Critical Success Factors for Big Data: A Systematic Literature Review. *IEEE Access (Volume: 8) ISSN: 2169-3536*, p. 118940 - 118956.

Anand, T., Pal, R., & Dubey, S. K. (2015). Analytical approach of soft computing in big data issues. *International Conference Communication, Control and Intelligent Systems, CCIS*, p. 442-449.

Ardagna, C. A., Ceravolo, P., & Damiani, E. (2016). Big data analytics as-a-service: Issues and challenges. *IEEE International Conference on Big Data*.

Ardagna, D., Cappiello, C., Samà, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems, Volume 89*, p. 548 - 562.

Arfat, Y., Usman, S., Mehmood, R., & Katib, I. (2020). Chapter 20 Big Data for Smart Infrastructure Design: Opportunities and Challenges. In *EAI/Springer Innovations in Communication and Computing* (pp. p. 491-518). Springer Nature Switzerland AG.

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020). Big Data: Current Challenges and Future Scope. *IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE). ISBN:978-1-7281-5034-5*, p. 131 - 143.

Asif, M. (2020). Are QM models aligned with Industry 4.0? A perspective on current practices. *Journal of Cleaner Production*, p. 1-11.

Auer, F., & Felderer, M. (2019). Addressing data quality problems with metamorphic data relations. *IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*, 76-83.

Baillie, C., Edwards, P., & Pignotti, E. (2015). Qual: A provenance-aware quality model. *Journal of Data and Information Quality. Volume 5*, p. 1-22.

Baldassarre, M. T., Caballero, I., Caivano, D., Garcia, B. R., & Piattini, M. (2018). From big data to smart data: a data quality perspective. *ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering*, p. 19–24.

Barbara Kitchenham, S. C. (2007). *Guidelines for performing systematic reviews in software engineering.* Durham, UK: EBSE Technical Report. EBSE-2007-01 Version 2.3.

Barik, R. K., Lenka, R. K., Ali, S. M., Gupta, N., Satpathy, A., & Raj, A. (2017). Investigation into the efficacy of geospatial big data visualization tools. *International Conference on Computing, Communication and Automation (ICCCA)*, p. 88 - 92.

Barsi, Á., Kugler, Z., Juhász, A., Szabó, G., Batini, C., Abdulmuttalib, H., . . . Shen, H. (2019). Remote sensing data quality model: from data sources to lifecycle phases. *International Journal of Image and Data Fusion*, p. 280-299.

Basso, T., Silva, H., & Moraes, R. (2019). On the Use of Quality Models to Characterize Trustworthiness Properties. *International Workshop on Software Engineering for Resilient Systems, SERENE*, p. 147-155.

Behkamal, B., Kahani, M., Bagheri, E., & Jeremic, Z. (2014). A metrics-driven approach for quality assessment of linked open data. *Journal of Theoretical and Applied Electronic Commerce Research*, p. 64-79.

Bejarano, G. (2018). PhD Forum: Deep Learning and Probabilistic Models Applied to Sequential Data. *EEE International Conference on Smart Computing (SMARTCOMP), Taormina, ,* pp. 252-253, doi: 10.1109/SMARTCOMP.2018.00066.

Bhutani, P., Saha, A., & Gosain, A. (2020). Wsemqt: A novel approach for quality-based evaluation of web data sources for a data warehouse. *IET Software. Volume 14*, p. 806-815.

Caballero, I., Serrano, M., & Piattini, M. (2014). A data quality in use model for Big Data. *33rd International Conference on Conceptual Modeling*, p. 65-74.

Canbay, Y., Vural, Y., & Sagiroglu, S. (2018). Privacy Preserving Big Data Publishing. *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, p. 24 - 29.

Cappiello, C., Meroni, G., Pernici, B., Plebani, P., Salnitri, M., Vitali, M., . . . Sanna, A. (2020). Improving Health Monitoring With Adaptive Data Movement in Fog Computing. *Frontiers in Robotics and AI*, p. 1-17.

Cappiello, C., Samá, W., & Vitali, M. (2018). Quality awareness for a Successful Big Data Exploitation. *International Database Engineering & Applications Symposium*, p. 37–44.

Castillo, R. P., Carretero, A. G., Caballero, I., Rodriguez, M., Piattini, M., Mate, A., . . . Lee, D. (2018). DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data. *Sensors (Switzerland)*, p. 1-24.

Cedillo, P., Valdez, W., Delgado, P. C., & Cabrera, D. P. (2020). A Data as a Service Metamodel for Managing Information of Healthcare and Internet of Things Applications. *Conference on Information and Communication Technologies of Ecuador, TICEC*, p. 272-286.

Chandarana, P., & Vijayalakshmi, M. (2014). Big Data analytics frameworks. *International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA). ISBN:978-1-4799-2494-3*, p. 430 - 434.

Chang, V., Ramachandranb, M., Wills, G., Walters, R. J., Li, C.-S., & Watters, P. (2016). Editorial for FGCS special issue: Big Data in the cloud. *Future Generation Computer Systems 65*, p. 73 - 75.

Chaudhari, N., & Srivastava, D. S. (2016). Big data security issues and challenges. *International Conference on Computing, Communication and Automation (ICCCA)*, p. 60-64.

Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: Challenges, issues, and opportunities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, p. 1-15.

Ciancarini, P., Poggi, F., & Russo, D. (2016). Big Data Quality: A Roadmap for Open Data. *International Conference on Big Data Computing Service and Applications, BigDataService*, p. 210-215.

CICHY, C., & RASS, S. (2019). An overview of data quality frameworks. *IEEE Access*, p. 24634-24648.

Conde-Clemente, P., Trivino, G., & Alonso, J. M. (2017). Generating automatic linguistic descriptions with big data. *Information Sciences Volume 380*, p. 12 - 30.

Cortes, R., Bonnaire, X., Marin, O., & Sens, P. (2015). Stream Processing of Healthcare Sensor Data: Studying User Traces to Identify Challenges from a Big Data Perspective. *Procedia Computer Science, Volume 42*, p. 1004 - 1009.

Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. *Robotics and Computer-Integrated Manufacturing April 2020 Volume 62 Article 101861*.

Dave, M., & Gianey, H. K. (2016). Analysis of big data for data-intensive applications. *International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*.

Davoudian, A., & Liu, M. (2020). Big Data Systems: A Software Engineering Perspective. *ACM Computing Surveys*, p. 1-39.

Debattista, J., Scerri, S., Lange, C., & Auer, S. (2015). Linked 'Big' Data: Towards a Manifold Increase in Big Data Value and Veracity. *IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, p. 92 - 98.

Demchenko, Y., Grosso, P., Laat, C. d., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. *International Conference on Collaboration Technologies and Systems, CTS*, p. 48-55.

Desai, P. V. (2018). A survey on big data applications and challenges. *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, p. 737 - 740.

Elia, G., Polimeno, G., Solazzo, G., & Passiante, G. (2020). A multi-dimension framework for value creation through Big Data. *Industrial Marketing ManagementIn press*.

Emmanuel, I., & Stanier, C. (2016). Defining big data. *ACM International Conference Proceeding Series.*

Fagúndez, S., Fleitas, J., & Marotta, A. (2015). Data streams quality evaluation for the generation of alarms in health domain. *International Workshops on Web Information Systems Engineering, IWCSN*, p. 204-210.

Fan, W. (2015). Data quality: From theory to practice. *Sigmond Record. Volume 44, Issue 3*, pp. 7 - 18.

Fernández, S. M., Jedlitschka, A., Guzmán, L., & Vollmer, A. M. (2018). A quality model for actionable analytics in rapid software development. *Euromicro Conference on Software Engineering and Advanced Applications, SEAA*, p. 370-377.

Gandomi A., H. M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management 35*, 137 - 144.

Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2916). A survey on indexing techniques for big data: taxonomy and performance evaluation. *Knowledge and Information Systems*, p. 241-284.

Gao, T., Li, T., Jiang, R., Duan, R., Zhu, R., & Yang, M. (2018). A research about trustworthiness metric method of SaaS services based on AHP. *International Conference on Cloud Computing and Security, ICCCS*, p. 207-218.

Garises, V., & Quenum, J. G. (2019). An evaluation of big data architectures. *8th International Conference on Data Science, Technology and Applications, DATA*, p. 152-159.

Garousi, V., Felderer, M., & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology, Volume 106*, Pages 101-121. ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2018.09.006.

Gartner IT Glossary. (n.d.). *Gartner Group*. Retrieved from https://www.gartner.com/en/information-technology/glossary/big-data. Consulted at 29th october, 2020.

Garvin, D. (1996). Competing on the eight dimensions of quality. *IEEE Engineering Management Review, 24(1)*, 15-23.

Ge, M., & Dohnal, V. (2018). Quality Management in Big Data. *Informatics*.

Ge, M., & Lewoniewski, W. (2020). Developing the quality model for collaborative open data. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES*, p. 1883-1892.

Ghorbanian, M., Dolatabadi, S. H., & Siano, P. (2019). Big Data Issues in Smart Grids: A Survey. *IEEE Systems Journal. Volume 13, Issue 4*, p. 4158-4168.

Gong, X., Yin, C., & Li, X. (2019). A grey correlation based supply–demand matching of machine tools with multiple quality factors in cloud manufacturing environment. *Journal of Ambient Intelligence and Humanized Computing. Volume 10*, p. 1025-1038.

Gyulgyulyan, E., Aligon, J., Ravat, F., & Astsatryan, H. (2019). Data Quality Alerting Model for Big Data Analytics. *23rd European Conference on Advances in Databases and Information Systems, ADBIS*.

Hammou, B. A., Lahcen, A. A., & Mouline, S. (2020). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. *Information Processing & Management, Volume 57*.

Haryadi, A. F., Hulstijn, J., Wahyudi, A., Voort, H. v., & Janssen, M. (2016). Antecedents of big data quality: An empirical examination in financial service organizations. *IEEE International Conference on Big Data (Big Data)*, p. 116 - 121.

Haug, F. S. (2016). Bad big data science. *IEEE International Conference on Big Data (Big Data)*, p. 2863 - 2871.

Helfert, M., & Ge, M. (2018). Perspectives of big data quality in smart service ecosystems (quality of design and quality of conformance). *Journal of Information Technology Management*, p. 72-83.

Hongxun, T., Honggang, W., & Kun, Z. (2018). Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory. *IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, p. 248 - 252.

Immonen, A., Paakkoneen, P., & Ovaska, E. (2015). Evaluating the Quality of Social Media Data in Big Data Architecture. *IEEE Access*, p. 2028-2043.

Indrakumari Ranganathan, P. T. (2020). The growing role of integrated and insightful big and real-time data analytics platforms. In P. E. Pethuru Raj, *The digital twin paradigm for smarter systems and environments: The industry use cases* (pp. 165-186). London: Academic press. An imprint of Elsevier.

International Organization for Standardization. (2008). *International Standard ISO/ IEC 25012:2008*. Retrieved from Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Data quality model.

Ishwarappa, a. J. (2015). A brief introduction on big data 5Vs characteristics and Hadoop technology. *International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015)*, No. 48, pp 319-324.

ISO/IEC25024. (2015). *“Software engineering - software product quality requirements and evaluation (square) - measurement of data quality".*

Jagli, D., Purohit, S., & Chandra, N. S. (2018). Saasqual: A quality model for evaluating SAAS on the cloud computing environment. *Annual Convention of Computer Society of India : Big Data Analytics, CSI*, p. 429-437.

Jarwar, M. A., & Chong, I. (2020). Web objects based contextual data quality assessment model for semantic data application. *Applied Sciences (Switzerland). Volume 10.*, p. 1-33.

Jich-Yan, T., Wen, Y. X., & Chien-Hua, W. (2020). A framework for big data analytics on service quality evaluation of online bookstore. *12th EAI International Conference on Wireless Internets, WiCON*, p. 294-301.

Jung, Y., Hur, C., & Kim, M. (2018). Sustainable situation-aware recommendation services with collective intelligence. *Sustainability (Switzerland)*, p. 1-11.

K. Radha, B. T. (2015). Service level agreements in cloud computing and big data. *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 5, No. 1, February 2015, pp. 158 - 165. ISSN: 2088-8708.

Kaggle. (2020, 07 01). *kaggle*. Retrieved from New York City Taxi Trip Duration: https://www.kaggle.com/c/nyc-taxi-trip-duration

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *46th Hawaii International Conference on System Sciences. ISBN:978-1-4673-5933-7*, p. 995 - 1004.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good practices. *Sixth International Conference on Contemporary Computing (IC3). ISBN:978-1-4799-0192-0*, p. 404 - 409.

Khan, M., Wu, X., Xu, X., & Dou, W. (2017). Big data challenges and opportunities in the hype of Industry 4.0. *IEEE International Conference on Communications (ICC)*.

Khodkari, H., Maghrebi, S. G., Asosheh, A., & Hosseinzadeh, M. (2018). Smart Healthcare and Quality of Service Challenges. *9th International Symposium on Telecommunications (IST)*, p. 253 - 257.

Khurana, R., & Bawa, R. K. (2016). QoS based Cloud Service Selection paradigms. *International Conference on Cloud System and Big Data Engineering, Confluence*, p. 174-179.

Kirchen, I., Schutz, D., Folmer, J., & Vogel-Heuser, B. (2017). Metrics for the evaluation of data quality of signal data in industrial processes. *International Conference on Industrial Informatics, INDIN*, p. 819-826.

Kiruthika, J., & Khaddaj, S. (2015). Software quality issues and challenges of internet of things. *International Symposium on Distributed Computing and Applications for Business, Engineering and Science, DCABES*, p. 176-179.

Kläs, M., Putz, W., & Lutz, T. (2017). Quality evaluation for big data: A scalable assessment approach and first evaluation results. *Joint Conference of the Int'l Workshop on and International Conference on Software Process and Product Measurement Software Measurement*, p. 115-124.

Kushal Patel, B. P. (2017). Privacy issues in big data. *2nd International Conference for Convergence in Technology (I2CT)*, p. 259 - 264.

Lakshen, G. A., & Vraneš, S. (2016). Big Data and Quality: A Literature Review. *24th Telecommunications forum TELFOR. ISBN: 978-1-5090-4086-5*.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: Classifying poor data. *IEEE 21st Pacific Rim International Symposium on Dependable Computing. ISBN: 978-1-4673-9376-8*, p. 179 - 188.

Libes, D., Shin, S., & Woo, J. (2015). Considerations and recommendations for data availability for data analytics for manufacturing. *IEEE International Conference on Big Data (Big Data)*, p. 68 -75.

Liu, Z., Chen, Q., & Cai, L. (2018). Application of requirement-oriented data quality evaluation method. *International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD*, p. 407-412.

Máchová, R., & Lněnička, M. (2017). Evaluating the quality of open data portals on the national level. *Journal of Theoretical and Applied Electronic Commerce Research. Volume 12*, p. 21-41.

Manikam, S., Sahibudin, S., & Kasinathan, V. (2019). Business intelligence addressing service quality for big data analytics in public sector. *Indonesian Journal of Electrical Engineering and Computer Science*, p. 491-499.

Mbonye, V., & Price, C. S. (2019). A model to evaluate the quality of Wi-Fi perfomance: Case study at UKZN Westville campus. *International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD*, p. 1-8.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for Big Data. *Future Generation Computer Systems Volume 63, 1*, p. 123-130.

Micic, N., Neagu, D., Campean, F., & Zadeh, E. H. (2018). Towards a Data Quality Framework for Heterogeneous Data. *Cyber, Physical and Social Computing, IEEE Smart Data, iThings-GreenCom-CPSCom-SmartDat*, p. 155-162.

Montero, O., Crespo, Y., & Piattini, M. (2021). Big Data Quality Models: A Systematic Mapping Studio. *Springer CCIS Series (Communications in Computer and Information Science)*.

Musto, J., & Dahanayake, A. (2019). Integrating data quality requirements to citizen science application design. *International Conference on Management of Digital EcoSystems, MEDES*, p. 166-173.

Muthukrishnan, S. M., Yasin, N. B., & Govindasamy, M. (2018). Big data framework for students' academic performance prediction: A systematic literature review. *IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). ISBN:978-1-5386-3528-5*, p. 376 - 382.

Nadal, S., Herrero, V., Romero, O., Abelló, A., Franch, X., Vansummeren, S., & Valerio, D. (2017). A software reference architecture for semantic-aware Big Data systems. *Information and Software Technology.* , p. 75-92.

Naik, K., & Joshi, D. A. (2017). Role of Big Data in various sectors. *International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC*, p. 117-122.

Nakamichi, K., Ohashi, K., Aoyama, M., Joeckel, L., Siebert, J., & Heidrich, J. (2020). Requirements-driven method to determine quality characteristics and measurements for machine learning software and its evaluation. *International Requirements Engineering Conference, RE*, p. 260-270.

Nersessian, D. (2018). The law and ethics of big data analytics: A new role for international human rights in the search for global standards. *Business Horizons November–December 2018, Volume 61*, p. 845 - 854.

Neves, P. C., Schmerl, B., Cámara, J., & Bernardino, J. (2016). Big data in cloud computing: Features and issues. *IoTBD 2016 - Proceedings of the International Conference on Internet of Things and Big Data*, p. 307-314.

Neves, P., & Bernardino, J. (2015). Big data issues. *ACM International Conference Proceeding Series*, p. 200-201.

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing. Volume 8*, p. 391-432.

Noorwali, I., Arruda, D., & Madhavji, N. H. (2016). Understanding Quality Requirements in the Context of Big Data Systems. *IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE)*, P. 76 - 79.

Oliveira, M. I., Oliveira, L. E., Batista, M. G., & Lóscio, B. F. (2018). Towards a meta-model for data ecosystems. *Annual International Conference on Digital Government Research: Governance in the Data Age*, p. 1–10.

Olsina, L., & Lew, P. (2017). Specifying mobileapp quality characteristics that may influence trust. *Central & Eastern European Software Engineering Conference in Russia, CEE-SECR*, p. 1–9.

Omidbakhsh, M., & Ormandjieva, O. (2020). Toward a new quality measurement model for big data. *9th International Conference on Data Science, Technology and Applications*, p. 193-199.

Oxford Learner's Dictionary. (n.d.). *Big data*. Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/big-data?q=big+data. Consulted at 8th september, 2020.

Pandey, K. K., & shukla, D. (2018). Challenges of Big Data to Big Data Mining with their Processing Framework. *8th International Conference on Communication Systems and Network Technologies (CSNT)*, p. 89 - 94.

Parra, A. V., Parody, L., Vaca, A. J., Caballero, I., & López, M. T. (2019). DMN for Data Quality Measurement and Assessment. *International Workshops on AI4BPM, BP-Meet-IoT, BPI, BPMinDIT, BPMS2, DEC2H, MIEL, PM-DiPro, PODS4H, PQ, SPBP, VEnMo*, p. 362-374.

Paryasto, M., Alamsyah, A., Rahardjo, B., & Kuspriyanto. (2014). Big-data security management issues. *2nd International Conference on Information and Communication Technology (ICoICT)*, p. 59 - 63.

Patel, J. (2019). An effective and scalable data modeling for enterprise big data platform. *IEEE International Conference on Big Data (Big Data)*, 2691-2697.

Patgiri, R. (2018). Issues and Challenges in Big Data: A Survey. *14th International Conference on Distributed Computing and Internet Technology, ICDCIT*, p. 295 - 300.

Pengcheng Zhang, F. X. (2017). Data quality in big data processing: Issues, solutions and open problems. *IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SC*, 1-7.

Pereira, J. D., Silva, R., Antunes, N., Silva, J. L., França, B. d., Moraes, R., & Vieira, M. (2020). A platform to enable self-adaptive cloud applications using trustworthiness properties. *International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS*, p. 71-77.

Pierce, L. (2016). Big Data issues for remote sensing: variety. *International Geoscience and Remote Sensing Symposium (IGARSS)*, p. 7593-7596.

R.Y. Wang, V. S. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering. Volume: 7, Issue: 4*, pp. 623 - 640.

Rahman, H., Begum, S., & Ahmed, M. U. (2016). Ins and outs of big data: A review. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, p. 44-51.

Rahman, M. S., & Reza, H. (2020). Systematic Mapping Study of Non-Functional Requirements in Big Data System. *IEEE International Conference on Electro Information Technology*, p. 25-31.

Rani, K., & Sagar, R. K. (2018). A Rigorous Investigation on Big Data Analytics. *Advances in Intelligent Systems and Computing*, p. 637-650.

Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. *IEEE International Conference on Big Data. ISBN: 978-1-4799-9926-2*, p. 2654 - 2660.

Rudraraju, N. V., & Boyanapally, V. (2019). Data Quality Model for Machine learning. *Faculty of Computing, Blekinge Institute of Technology*, p. 1-107.

Sangeeta, & Sharma, K. (2016). Quality issues with big data analytics. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, p. 3589 - 3591.

Santhanam, P. (2020). Quality Management of Machine Learning Systems. *International Workshop on Engineering Dependable and Secure Machine Learning Systems, EDSMLS*, p. 1-13.

Saraladevi, B., Pazhaniraja, N., Paul, P. V., Basha, M. S., & Dhavachelvan, P. (2015). Big data and Hadoop-A study in security perspective. *Procedia Computer Science*, p. 596-601.

Sen, D., Ozturk, M., & Vayvay, O. (2016). An Overview of Big Data for Growth in SMEs. *Procedia - Social and Behavioral Sciences 24, Volume 235*, p. 159 - 167.

Serhani, M. A., Kassabi, H. T., Taleb, I., & Nujum, A. (2016). An Hybrid Approach to Quality Evaluation Across Big Data Value Chain . *IEEE International Congress on Big Data (BigData Congress)*, p. 418 - 425.

Stegmaier, F., Seifert, C., Kern, R., H¨ofler, P., Bayerl, S., Granitzer, M., . . . Zwicklbauer, S. (2014). Unleashing semantics of research data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, p. 103-110.

Strang, K. S. (2020). Hidden big data analytics issues in the healthcare industry. *Health Informatics Journal. Volume 26, Issue 2*, Pages 981-998.

Surendro, O. K. (2016). Academic Cloud ERP quality assessment model. *International Journal of Electrical and Computer Engineering. Volume 6*, p. 1038-1047.

Swan, M. (2015). Philosophy of Big Data: Expanding the Human-Data Relation with Big Data. *IEEE First International Conference on Big Data Computing Service and Applications*, p. 468 - 477.

Taleb, I., Serhani, M. A., & Dssouli, R. (2018). Big Data Quality Assessment Model for Unstructured Data. *International Conference on Innovations in Information Technology (IIT)*, p. 69 - 74.

Taleb, I., Serhani, M. A., & Dssouli, R. (2018). Big Data Quality: A Survey. *7th IEEE International Congress on Big Data*, p. 166-173.

Taleb, I., Serhani, M. A., & Dssouli, R. (2019). Big Data Quality: A Data Quality Profiling Model. *World Congress on Services, SERVICES*, p. 61-77.

Talha, M., Kalam, A. A., & Elmarzouqi, N. (2019). Big Data: Trade-off between Data Quality and Data Security. *Procedia Computer Science, Volume 151*, p. 916 - 922.

Talha., M., Kalam, A. A., & Elmarzouqi, N. (2019). Big Data: Trade-off between Data Quality and Data Security. *Procedia Computer Science, Volume 151*, p. 916 - 922.

Tao, C., & Gao, J. (2016). Quality assurance for big data application - issues, challenges, and needs. *International Conference on Software Engineering and Knowledge Engineering, SEKE*, p. 375-381.

Tepandi, J., Lauk, M., Linro, J., Raspel, P., Piho, G., Pappel, I., & Draheim, D. (2017). The data quality framework for the Estonian public sector and its evaluation: Establishing a systematic process-oriented viewpoint on cross-organizational data quality. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, p. 1-26.

Vale, L. R., Sincorá, L. A., & Milhomem, L. d. (2018). The moderate effect of analytics capabilities on the service quality. *Journal of Operations and Supply Chain Management. Volume 11*, p. 101-113.

Vetrò, A., Canova, L., Torchiano, M., & Minotas, C. O. (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, p. 325-337.

Villalpando, L. E., April, A., & Abran, A. (2014). Performance analysis model for big data applications in cloud computing. *Journal of Cloud Computing*, p. 1-20.

Vostrovsky, V., & Tyrychtr, J. (2018). Consistency of Open Data as Prerequisite for Usability in Agriculture. *Scientia Agriculturae Bohemica. Volume 49*, p. 333-339.

Wan, Y., Shi, W., Gao, L., Chen, P., & Hua, Y. (2015). A general framework for spatial data inspection and assessment. *Earth Science Informatics*, p. 919-935.

Wang, B., Wen, J., & Zheng, J. (2019). Research on Assessment and Comparison of the Forestry Open Government Data Quality Between China and the United States. *International Conference on Data Science, ICDS*, p. 370-385.

Wang, C., Lu, Z., Wu, Z., Wu, J., & Huang, S. (2017). Optimizing Multi-Cloud CDN Deployment and Scheduling Strategies Using Big Data Analysis. *International Conference on Services Computing, SCC* , p. 273-280.

Wang, L., & Alexander, C. A. (2015). Big data in design and manufacturing engineering. *American Journal of Engineering and Applied Sciences*, p. 223-232.

Wani, M. A., & Jabin, S. (2018). Big data: Issues, challenges, and techniques in Business Intelligence. *Advances in Intelligent Systems and Computing*, 613-628.

White, G., Nallur, V., & Clarke, S. (2017). Quality of service approaches in IoT: A systematic mapping. *Journal of Systems and Software*, p. 186-203.

Y. Demchenko, P. G. (2013). Addressing big data issues in scientific data infrastructure. *IEEE*, p. 48 - 55.

Yan, M., Xia, X., Zhang, X., Xu, L., & Yang, D. (2017). A Systematic Mapping Study of Quality Assessment Models for Software Products. *International Conference on Software Analysis, Testing and Evolution (SATE). ISBN:978-1-5386-3688-6*, p. 63 - 71.

Yu, B., Zhang, C., & Tang, Z. (2018). Verification method of data quality in science and technology cloud in Shaanxi province. *IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, p. 319 - 323.

Zhang, P., Xiong, F., Gao, J., & Wang, J. (2018). Data quality in big data processing: Issues, solutions and open problems. *IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI. ISBN: 978-1*.

Zhang, P., Zhou, X., Li, W., & Gao, J. (2017). A survey on quality assurance techniques for big data applications. *IEEE Third Internationl Conference on Big Data Computing Service and Applications*, p. 313 - 319.