LUT
Lappeenranta
University of Technology

Lappeenranta University of Technology
School of Business and Management
Degree Program in Industrial Management
Master's Thesis

10.8.2016

Data quality analysis in industrial maintenance; Theory vs. Reality

Miika Rantala

Examiner: Post-Doctoral Researcher Antero Kutvonen

ABSTRACT

Author: Miika Rantala

Subject: Data quality analysis in industrial maintenance; theory vs. reality

Year: 2016 **Place:** Helsinki, Finland

Master's Thesis. Lappeenranta University of Technology, Industrial Engineering and Management, Cost Management

71 pages, 12 figures and 11 Tables

Examiner: Post Doc. Antero Kutvonen

Keywords: data quality, assessment, analytics, industrial maintenance

The use of Big Data, analytics and simulations for supporting decision making in different business areas regardless the field of industry has gained significant interest lately. Firms believe to improve efficiency and thereby gain advantage by exploiting analytics. Service providers' promises about the possibilities that analytics will bring have increased the interest even more. Nevertheless, it is important to realize that the existing data has a significant impact on the potential of analytics. The vast amount of data available make the situation even worse because detecting corruptions in data becomes an extremely difficult task. Using low quality data causes biased understanding state which in turn might result in bad decisions.

Data quality is a relative concept, which is mainly based on fit for use ideology meaning that data is high quality if it is suitable for the intended purpose. It's also possible to determine the most substantial dimensions of data quality to help in measuring. High quality data should be at least accurate, complete, consistent and timeless. The aim of this study is to create a model for measuring data quality for the needs of industrial maintenance. The data used in this thesis is provided by nine different size factories operating in a variety of industries. Therefore, the test data can be considered quite credible and provides great insight of factors affecting the data quality. The results of this study show that it's possible to find significant errors from analytical as well as managerial perspective. The most errors are caused by poor data collection and management process.

TIIVISTELMÄ

Tekijä: Miika Rantala

Työn nimi: Datan laadun analysointi kunnossapitoliiketoiminnassa; teoria ja todellisuus

Vuosi: 2016 **Paikka:** Helsinki, Suomi

Diplomityö. Lappeenrannan teknillinen yliopisto, tuotantotalous, Kustannusjohtamisen koulutusohjelma.

71 sivua, 12 kuvaa ja 11 taulukkoa

Tarkastaja: Tutkijatohtori Antero Kutvonen

Hakusanat: datan laatu, arviointi, analytiikka, teollinen kunnossapito

Big Data:n, analytiikan ja simuloinnin hyödyntäminen päätöksenteon tukena liiketoiminnan eri osa-alueilla on herättänyt viime aikoina suurta mielenkiintoa toimialasta riippumatta. Yritykset uskovat pystyvänsä tehostamaan toimintaansa ja siten saavuttavansa kilpailuetua hyödyntämällä analytiikan eri keinoja, eikä tätä intoa ole laskeneet lukuisien palveluntarjoajien lupaukset analytiikan mahdollisuuksista. On kuitenkin muistettava, että analytiikka pohjautuu lähtökohtaisesti jo olemassa olevaan dataan, mikä vaikuttaa merkittävästi hyödyntämismahdollisuuksiin. Tilannetta pahentaa entisestään saatavilla olevan datan valtava määrä, jolloin virheiden huomaamisesta tulee erittäin haastavaa. Huonolaatuisen datan hyödyntäminen johtaa virheellisiin tulkintoihin ja siten vääriin päätöksiin.

Datan laatu on itsessään suhteellinen käsite, joka pohjautuu lähinnä ajatukseen, että laadukas data soveltuu sille suunniteltuun käyttötarkoitukseen. Datalle voidaan kuitenkin määrittää merkittävimmät laatuun vaikuttavat näkökulmat laadun mittaamiseksi. Hyvälaatuisen datan tulisi olla ainakin paikkansapitävää, kattavaa, johdonmukaista ja ajantasaista. Tässä työssä pyritäänkin luomaan malli datan laadun mittaamiseksi teollisen kunnossapidon tarpeisiin. Työssä on hyödynnetty dataa yhdeksästä erikokoisesta ja vaihtelevilla toimialoilla toimivista tuotantolaitoksista tarjoten varsin kattavan testiaineiston ja siten monipuolisen katsauksen datan laatuun vaikuttavista tekijöistä. Tutkimus osoittaa, että datasta voidaan löytää merkittäviä virheitä niin analytiikan kuin toiminnan johtamisen kannalta. Suurin osa datan virheistä johtuu joko puutteellisista keräysprosesseista tai datan hallinnasta.

ACKNOWLEDGEMENTS

This master's thesis has been one of the most challenging as well as rewarding tasks of my life. During this project I have learned to code and dived into the exciting world of analytics without previous experience. Therefore, I want to thank my supervisor Samuli Kortelainen for his endless motivation and guidance. I would also like to thank my amazing colleagues for numerous interesting conversations and aspects about analytics. The past eight months have been extremely educating and have opened my eyes about the goals that I can achieve if I just try hard enough.

My studies in LUT have not only created unforgettable memories but have prepared me for the work life. The greatest inspirer of my studies has been my late grandfather who always encouraged me to study and apply to a technical university. I'm thankful for his and my parents' moral and financial support which have helped me reach my goals. Besides that, I want to thank my big brother and especially my little sister for creating a competitive environment and thus assisting in setting my future targets.

Lastly, I want to thank all my friends who have been there for me. The last five years in Lappeenranta have gone incredibly fast because of you. We have experienced a lot, but I believe the greatest journeys are still ahead of us.

Helsinki, 10th August 2016

Miika Rantala

TABLE OF CONTENTS

1 INTRODUCTION 8
1.1 Digitalization is revolutionizing businesses 8
1.2 Goals and scope 9
1.3 Research methodology and methods 11
1.4 Structure 12
2 DATA MANAGEMENT AS A KEY OF QUALITY 14
2.1 Information systems 14
2.2 Master data management 16
2.3 Master and Transaction data 17
2.4 General data types 18
3 WHAT IS GOOD DATA? 21
3.1 Review of data quality frameworks 21
3.2 Dimensions of data quality framework 30
3.3 Data quality testing 33
4 MODEL FOR TESTING DATA 38
4.1 DQA Target and Raw data 38
4.2 Framework adaptation for use 40
4.3 Results of data quality assessment 46
5 DISCUSSION 54
5.1 State of information management 54
5.2 The holistic framework 56
5.3 The design of DQA 57
6 CONCLUSIONS 60
6.1 Focus 60
6.2 Theoretical implications 61
6.3 Managerial implications 62
6.4 Future research 64
REFERENCES 66

LIST OF FIGURES

Figure 1 The scope of study 11
Figure 2 Input-output structure of the study 12
Figure 3 Data types adapted from (Marr 2015, 57-64; Batini et al. 2009) 18
Figure 4 Conceptual Framework of Data Quality (Wang & Strong 1996) 21
Figure 5 Evolutional Data Quality (Liu & Chi 2002) 24
Figure 6 The Measurement of Application Quality (Liu & Chi 2002) 25
Figure 7 Holistic Data Quality Framework 30
Figure 8 Database model 40
Figure 9 Data quality scores 47
Figure 10 Population completeness 50
Figure 11 Object accuracy 51
Figure 12 Type inaccuracy and related costs 52

LIST OF TABLES

Table 1 Defining research questions 10
Table 2 Data quality frameworks 22
Table 3 Data quality frameworks presented by practitioners 27
Table 4 The most cited dimensions 29
Table 5 Definitions of Data Quality Dimensions 32
Table 6 Characteristics of the raw data 39
Table 7 Changes to the holistic framework during adaptation 40
Table 8 Dimensions and metrics of DQA model 42
Table 9 Structure of DQA-model 45
Table 10 Properties of raw data 46
Table 11 Research questions and answers 60

LIST OF SYMBOLS AND ABBREVIATIONS

DQ	Data Quality
DQA	Data Quality Assessment
IS	Information System
DQM	Data Quality Management
MDM	Master Data Management
OLTP	Online transaction processing

1 INTRODUCTION

1.1 Digitalization is revolutionizing businesses

The amount of articles written about opportunities of Big Data and analytics has boomed during the past decade and for a reason, while the advantages gained by exploiting analytics are clear. In general, analytics are seen to support decision making and therefore improve firm's performance by making use of existing data (Bose 2009). It is also estimated that we are producing 2,5 quintillion bytes of data each day which is in fact more than 90 percentage of data generated in the past two years (IBM 2016), making the analytics even more interesting. Data analytics can be used to describe the current situation, make forecasts or even simulate possible outcomes of taken actions (Holsapple et al. 2014; Iverson 2014). The analytics are used to solve various kind of business problems and one of those is industrial maintenance. Industrial maintenance is a complicated and difficult business area from the managerial point of view. Maintenance is not often seen as a core business allowing directors to neglect quality of maintenance activities by focusing on cost reductions. At the same time, though, poorly managed maintenance might cause a lot of expenses due to scrap and production losses, which is also the reason why some companies have started to use analytics for improving the reliability and thereby the overall effectivity of their factory. The positive side of industrial maintenance is that the data is often internal and structured making the usage of analytics much easier. The results of analytics depend entirely on the data, though, making the data quality an important factor. The more the data includes errors and corruptions the higher is the probability of skewed results.

In the era of Big Data and analytics it is common that the provided data is from unknown provenance, meaning that there is no information about where it came, how it was collected, what do the fields mean, how reliable it is and so on. In addition to unknown provenance the data has probably gone through many hands and multiple transformations since it was collected. All of these have a significant affect to the quality of data. In literature there are a number of studies related to analytics but only a few of them really focus on data validation in practice, making this research interesting. Huge amount of data brings a lot of errors in data quality and in data usage. The studies have raised several issues regarding data

collection, processing and analysis, which causes information incompleteness and noise of Big Data (Liu et al. 2015). These problems might cause flawed decisions which can be also really costly. It is estimated that data quality problems cost U.S. businesses more than 600 billion USD a year (TDW 2002). Therefore, it is important to validate the data quality before use so that the result can be trusted and interpreted correctly.

The consequences of low data quality are experienced every day but often misunderstood. For example, there is no mandatory spare part in inventory or the welding robot is not maintained yearly as it should be. Such error might be caused by bad data. In the first case the spare parts might not be ordered because the inventory value claims that they exist. In the second example the yearly maintenance wasn't performed because the maintenance plan didn't exist or the interval was set to biyearly. In existing literature data quality is often handled separately from analytical purposes as a part of Data Quality Management (DQM) or even Total Quality Management (TQM) concepts. It does not mean that the same ideology could not be used as a basis of data quality assessment (DQA) for data analytics and simulation purposes as this study proves, though.

1.2 Goals and scope

The demand for this thesis comes from analytics executed to provide useful information for the needs of industrial maintenance operations. The carrying out of analytics have shown that there is clear need for data quality assessment, while important data is often missing or corrupted causing significant errors during the process. The performed analytics are also scalable which sets the most restrictions for this study as well. Therefore, the data quality assessment must be based entirely on the provided data and not to surveys or other time consuming processes such as comparisons of values. In general, this study has two goals. First, it aims to create a holistic framework to measure data quality. Second, in empirical side of the study the holistic framework is adapted in order to analyze the suitability of created framework and the data quality in industrial maintenance. Following research questions are set to help in examining the research problems.

Table 1 Defining research questions

Research question	Detail
RQ1. How to measure data quality from holistic perspective?	Quality is from general perspective a subjective concept. Data quality as well as any kind of quality can be measured in several ways thus attributes of quality are evaluated unequally and might be alternative.
SQ11. What are the dimensions of data quality?	Quality is a multi-dimensional concept where each dimensions represent unique aspect of quality.
SQ12. How can data quality be measured?	Quality is not a physical variable making the measurement more complicated. Measuring quality requires most likely custom designed metrics.
RQ2. How does the data apply to the holistic framework in industrial maintenance?	Each business area has own kind of special features that will affect significantly to the attributes of data and the requirements of assessment.
SQ21. How does the framework need to be adapted for the use?	Holistic frameworks aim to fit all situations, but it is seldom the reality. Number of changes and adaptions are usually needed before the framework can be implemented.
SQ22. How accurate is the measurement of data quality?	Measurements are often inaccurate, especially when they are related abstract objects or metrics.

In Table 1 are presented two research questions as well four sub research questions. Research question 1 and related sub research questions focus on theoretical aspect of measuring data quality. The aim of these research questions is to help create a holistic framework for evaluating data quality from general aspect. The quality is often seen as a subjective matter affecting significantly to the experienced quality. In general quality consists of multiple attributes making it important to define and understand the meaning of different dimensions. Which of them are substantial and required and which of them are less important if even needed. The second sub research question about how the dimensions should be measured will be answered when there is clear consensus of factors affecting to data quality.

Research question 2 and following sub questions focus on the empirical side of study. There would be no use for a holistic framework if it could not be adapted in practice. The research question 2 is more universal, while it is not clear that the quality of all kind of data could be evaluated. The main topics of the empirical part is to diagnose how well a theoretical approach suits the needs of industrial maintenance and what are the benefits gained by data quality assessment.

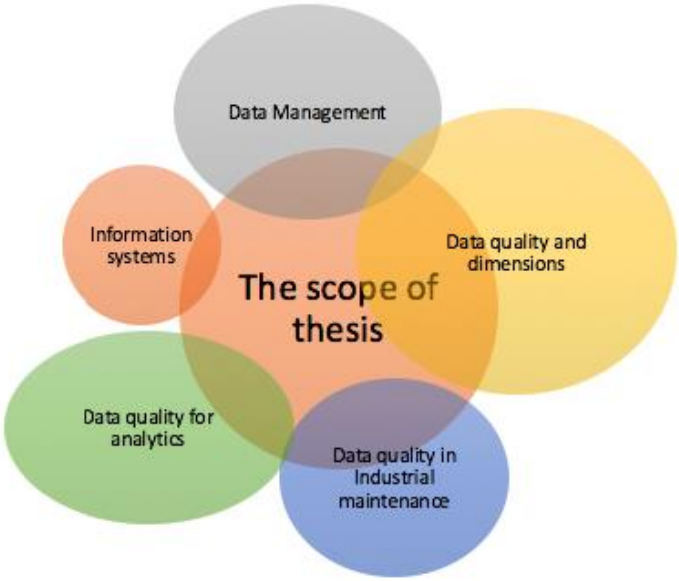

Figure 1 The scope of study

The scope of this study is data quality in industrial maintenance, meaning that the analyzed data is structured internal data that is related to maintenance operations. The study also partially includes information systems and data management concepts, while the data quality is significantly affected by the previous phases. Nevertheless, the empirical part is limited strictly to the data, while the aim of the study is to create a scalable and universal way to evaluate the data quality in a certain context. The analytical tools such as machine learning are excluded from the study while those will be used in later analysis after the data quality assessment. Analytics driven data improvement methods are excluded from the study for the same reason too.

1.3 Research methodology and methods

Qualitative case study is used to study complex phenomena within their context (Baxter & Jack 2008). In this study it would be the method for measuring data quality in industrial maintenance. This thesis attempts to define and explain the factors affecting the quality by analyzing multiple data sets. Previous theory of data quality assessment and information management are used to produce generalizations of the subject matter.

As in most case studies (Scapens 1990) the objective of this thesis is to determine whether the theories based on previous literature in this field of research provide good explanations for the phenomenon's observed or whether alternative explanations need to be developed. This thesis will provide a single observation of a phenomenon observed in data quality research. As the phenomenon of data quality concept has already been largely observed by theoretical and survey studies, it is well justified that qualitative case study is an appropriate way to attain new understanding of this phenomenon.

1.4 Structure

Chapters 2 and 3 focus on theoretical side of data process and data quality assessment. In chapter 2 is introduced the data process which includes information systems, data management ideologies as well as concept of data quality management. The purpose of chapter 2 is to provide general understanding about factors effecting the data quality. Chapter 3 begins with a review of commonly known and acknowledged practices for determining data quality. In that part several studies and practitioners' solutions are analyzed in order to create the holistic framework. The section 3 ends with best practices for designing metrics.

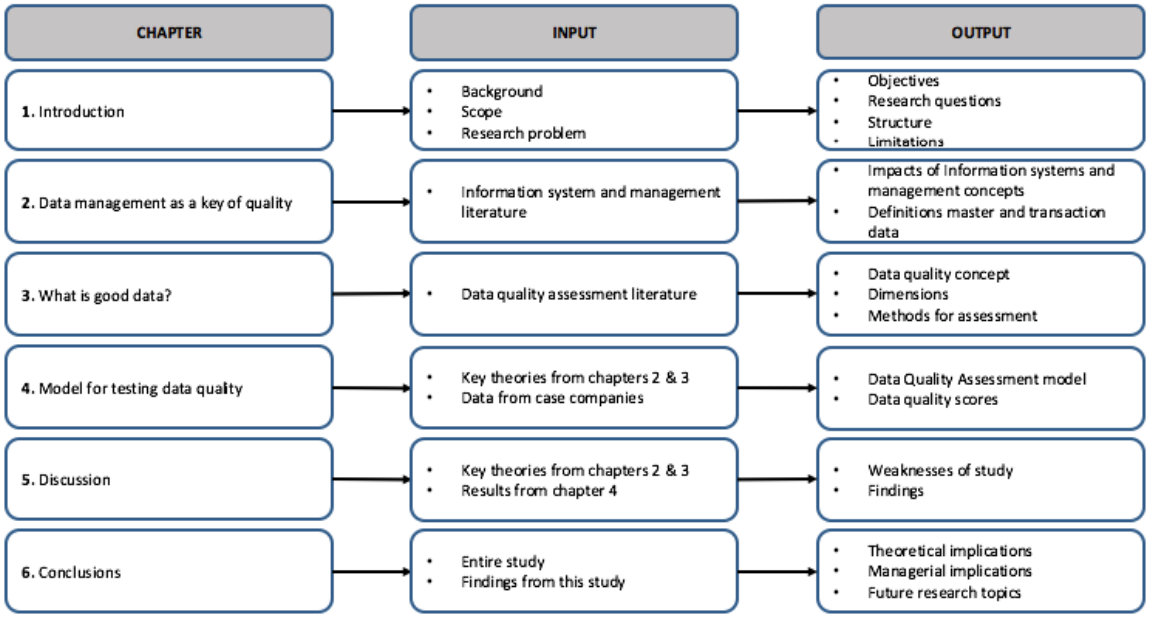

Figure 2 Input-output structure of the study

The empirical side of the study begins in chapter 4 where the case situation and the data are presented. The analyzed data is from nine manufacturing companies and therefore provides

quite credible setting for empirical study. In later parts of chapter 4 the holistic framework introduced in chapter 3 is adapted and implemented. The last part of chapter 4 is introducing the result of the assessment. The empirical part is based on empirical analysis on results and earlier introduced theoretical frameworks. After that follows chapter 5, which is general discussion about introduced holistic framework, implementation of the model and result. The study is ended by chapter 6 where the research questions are answered and theoretical as well as managerial implications are introduced with interesting future research topics.

2 DATA MANAGEMENT AS A KEY OF QUALITY

2.1 Information systems

Database management systems (DBMS) were created in the early 1960s to assist in maintaining and gathering large amounts of data. One of the first systems was designed by IBM but now there are already numerous providers and softwares to meet the growing demand. The need for these kind of systems comes from the intent to consolidate the decision making process and mine the data repositories for important business related information. The early database management systems have developed from simple network data models through enterprise resource planning (ERP) and management resource planning (MPR) systems to the web accessible DMBS's of internet age with access to all relevant business related information. In general, DMBS is an alternative to storing and managing data in files with ad hoc based approaches, which won't carry over time. (Ramakrishnan & Gerkhe 2000, 3-7)

Managing data efficiently in time has become almost a liability to companies because of the vast amount of data available. Data has changed from being an asset into a distraction and a mandatory duty. (Ramakrishnan & Gerkhe 2000, 3-7) DBMS is basically an information system (IS) which includes collecting, storing, elaborating, retrieving and exchanging of data to provide business services for all inside the company. Different types of information systems and their architectures can be classified by following three criteria: distribution, heterogeneity and autonomy. Distribution deals with the possibility to distribute data and applications over network of computers and heterogeneity is about the sematic and technological diversities among systems how the data is modelled and physically presented. The last criteria, autonomy, is determined by the degree of hierarchy and rules of coordination in the company using information systems. Based on these three criteria, five main types of information systems can be described. The main types are Monolithic, Distributed, Data Warehouses, Cooperative and Peer-to-Peer information systems. (Batini & Scannapieca 2009, 9-12)

The type of information system is not the key of this study but it is important to understand the role and the effect of an information systems on the data. A database is simplistically just a collection of data that describes the activities related to the company or organization. The most dominant type to store date is relational data model. The relational model consists of relations, which can be thought of as a set of records. Each relation has a schema which specifies its name, field names (attribute) and the type of field. As an example customer information in a company database might have four fields, which are company Id, name, invoicing address and country of origin. Each record then describes the customer in that customer relation. Also every row follows the predetermined schema of the customer relation. Integrity constrains are conditions that the records in a relation must fulfill. One of the basics is that a record must have a unique Id value, which increases significantly the accuracy with which the data can be described. Other important data models are the hierarchical model, the network model, the object oriented model and the object-relational model. (Ramakrishnan Gerkhe 2000, 3-12)

Since the increased ability to collect and store huge amount of data, companies are facing new challenges in relation to data quality (Haug et al. 2013). An information system might consist of thousands of above described entities, making the whole system exceedingly complex and hard to maintain. It is also claimed that the value of information varies at each point of its life cycle. That is why it is important to understand how to best protect the information from loss and corruption (Tallon & Scannell 2007). Even though information systems include several applications to protect and maintain data are new concepts of data quality management (DQM) and master data management (MDM) generally acknowledged to be useful for ensuring the overall data quality (Otto et al. 2012). DQM is part of Total Quality Management (TQM) concept and practices. The key aim of DQM is to improve data quality by setting data quality policies and guidelines. The DQM doesn't just concentrate on measuring and analyzing data quality but it also includes processes for cleansing and correcting data. (Lucas 2010). The Master data management in turn is a similar concept to DQM but it is based on the ideology that data quality process should begin with the key business objectives.

2.2 Master data management

The concept of Master data management (MDM) has gained significant interest in past years, even though its definition is not clear (Otto 2012). However, it aims to solve very clear problem of bad data. MDM promises to bring together all key information, regardless where that information is collected and thereby provide possibility to exploit the value of key data (Tuck 2008). It is well known that the data in most companies is a huge chaos caused by years of development of IT and information systems. In addition, Smith & McKeen (2008) claim that poor management of data leads to "data silos" which prevents the access to the company's key data. Most companies have a multitude of inconsistencies in data classification, formats and structures, making it nearly impossible to understand the information (Smith & McKeen 2008). Nowadays company's data must be managed on centralized manner and MDM aims to solve that issue. MDM tries to tackle data related issues on many areas, which includes business processes, data quality as well as standardization and integration of information systems (Silvola et al. 2011). The MDM relies on the fact that the master data is key to good data quality (Smith & McKeen 2008).

Smith & McKeen (2008) define four prerequisites for MDM which are also in agreement with the most of the requirements of data quality management concept. The first thing to do is to develop an Enterprise information policy because managing data is overall a highly political exercise at the end. It is particularly important to determinate the number of principles around corporate data management issues such as data ownership, accountability, privacy, security and risk management. The second prerequisite is the business ownership, while it is extremely important that each piece of data has a primary business owner (Smith & McKeen 2008). That is the only way to ensure the consistency. According to Haug, et al. (2013) the lack of ownership and clear roles in relation to data creation, use, and maintenance is one of the biggest reasons for low data quality. The third prerequisite is Governance which is all about making difficult decisions. Changing core data often requires modification in business processes which in turn raises conflicts at all levels. Thus it is important to get all stakeholders into an agreement. The last and most important prerequisite is the role of IT. Issues with Information systems are often considered to be IT problems, which is exceedingly wrong. Data management is entirely a business problem, because it is all about

understanding what is the core data and what kind of data will help us to do better business decisions. After that comes IT, whose role is to help the managers to identify the needed applications and figure out how everything fits together. (Smith & McKeen 2008) In addition to these prerequisites Haug et al. (2013) underlines the meaning of training and education at all phases of data process.

2.3 Master and Transaction data

Master data represents company's most important business objectives constituting the foundation of all data inside the organization. This is the reason why the key objectives should be used unequivocally around the company. (Otto 2012) The master data can be divided into entities such as customer master data, supplier master data, employee master data, product master data and asset master data (Smith & McKeen 2009). A master data object then represents a concrete business object and specified characteristics of this business object. The business object of a manufacturing company could be a welding robot and the attributes of machine id, location, capacity and weight. Furthermore, attributes are selected for representation of a predetermined class of business objects which would be in this case the machinery (Otto et al. 2012). What makes master data different from transaction data is that master data usually remains largely unaltered since characteristic features of a product, an asset or material etc. are always the same. Therefore, there is no need to update or change the values in database frequently. The volume of master data is also quite constant especially when compared to transaction data (Silvola et al. 2011).

Transaction data is generated often from online transaction processing (OLTP), and it consists of retail scanner records, business transactions, hotel reservations and other event related records. It is common that transaction data tend to get quickly high volumes, making it hard to handle with traditional tools. The attributes of transaction data can be divided in two types. First type of attributes are those that describe the identity of a record like name, customer id, transaction id or social security number. These have often unique value for each record but do not contain business related information. Second type are those that describe the properties or behavior of a record. These can be cost, type or object of the record. (Li & Jacob 2008) The important relation between transaction data and master data is that the

master data is needed for creating transaction data but not the other way around. This is because master data always describes the characteristics of real world objects. In practice, master data establishes the reference for transaction data while customer order always involves the product master data and the customer master data. (Silvola et al. 2011)

2.4 General data types

As mentioned, the data is a substantial part of data sciences, but not just the amount is important. Wang et al. (1995a) claim that data manufacturing process is similar to any other product manufacturing process. In data process a single number, a record, a file or a report will be used to produce output data or data products. These processes can also be one after the other, meaning that data product of one process is raw material for the following one. This kind of structure is quite common and it highlights the importance of data quality. A simple incorrectness during the first phase may multiply and corrupt the entire data. Hellerstein (2008) identified four main sources of error which are data entry, measurement, data distillation and integration.

The changes in data collection processes during the past decades have had significant impact on the data. This can be seen not only in the amount of data but also in the form of data. From general point of view Big Data and analytics include various data types such as internal and external data types (Marr 2015, 57-64). The basic data types are presented In Figure 3.

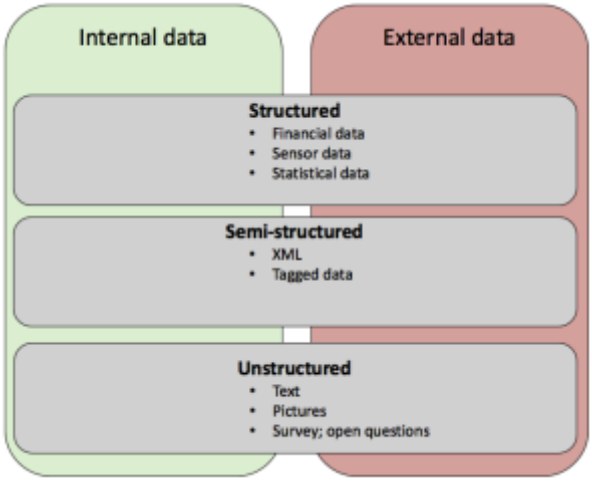

Figure 3 Data types adapted from (Marr 2015, 57-64; Batini et al. 2009)

The data types can be classified in three general categories, which are structured, semi-structured and unstructured data (Marr 2015, 57-64, Batini et al. 2009). The boundary between semi-structured and unstructured data is blurry. The semi-structured data might have partially some structure like date. The actual form of date is not defined, though, meaning that record can be text or numbers in one or several fields in a single data set. Semi-structured data is also commonly defined by XML-file which doesn't have associated XML schema file (Batini et al. 2009). Unstructured data is commonly text, but it might also include numbers or other marks. The basic character of unstructured data is that it is not so easy to put in categories or columns making analyzing with traditional computer softwares very difficult. (Marr 2015, 57-64) Other possible forms of unstructured data are voice, pictures and videos. The conversations don't just consist of unstructured text but also sentimental and perspectival aspects (Zikopoulos et al. 2015).

The last data type is structured data like financial records or other statistical data. It is estimated that only 20 percent of existing data is structured, but still it provides most of our business insights nowadays (Marr 2015, 57-64). The majority of research contribution also focuses either to structured or semi-structured data despite the acknowledged relevancy of unstructured data (Batini I et al. 2009). The reason for higher usage of structured data is rational. Structured data is much easier to handle and analyze than semi- or unstructured data. The data that is located in fixed fields in defined document or record is called structured data. Structured data has also predefined data model or it is organized by a predetermined way. A classic example of structured data is customer data. Customer data has usual fields such as first name, surname, address, phone number and Id, which build up the predefined data model. (Marr 2015, 57-64)

Structured data might include text, numbers or other marks, but it is assumed that each field includes only field specific data. It is also quite common that fields have rules such as phone number field accepts only numbers, forcing the data to be at least a bit better. (Batini 2009) Different kinds of information systems often provide structured data which is also the focus in this study. The analyzed raw data is mostly structured even though certain files have text fields. The data collection process in industrial maintenance often uses predetermined data models and lot of field specific rules like drop down menus to limit the choices. Some of the

fields like starting data are created automatically based on performed tasks. It is also claimed that the next generation of reliability data will be much richer in information due to the changes in technology. One of this is Internet of things. It is already possible but not common to install sensors or smart chips in the area of industrial maintenance for producing highly structured and reliable data. (Marr 2015, 57-64; Meeker & Hong 2014)

3 WHAT IS GOOD DATA?

3.1 Review of data quality frameworks

Defining good data is not an unambiguous task. Several research communities have widely studied the concept of data quality (DQ) in earlier literature. Nowadays the definition of data quality comes in most cases from the needs of primary usage of the data, which is also known as "fitness for use" (Chen et al. 2013). According to Juran (1989) data are of high quality if they are fit for their intended uses in operations, decision making and planning. Similarly taking the consumer viewpoint means that the concept of data quality depends on goal and domain. A set of data might be defined to be convenient for one but may not fulfill the needs of another. Therefore, to fully understand the meaning of DQ researchers have defined numerous sets of DQ dimensions. Wang and Strong (1996) introduced their conceptual framework of data quality in 1996 (Figure 4).

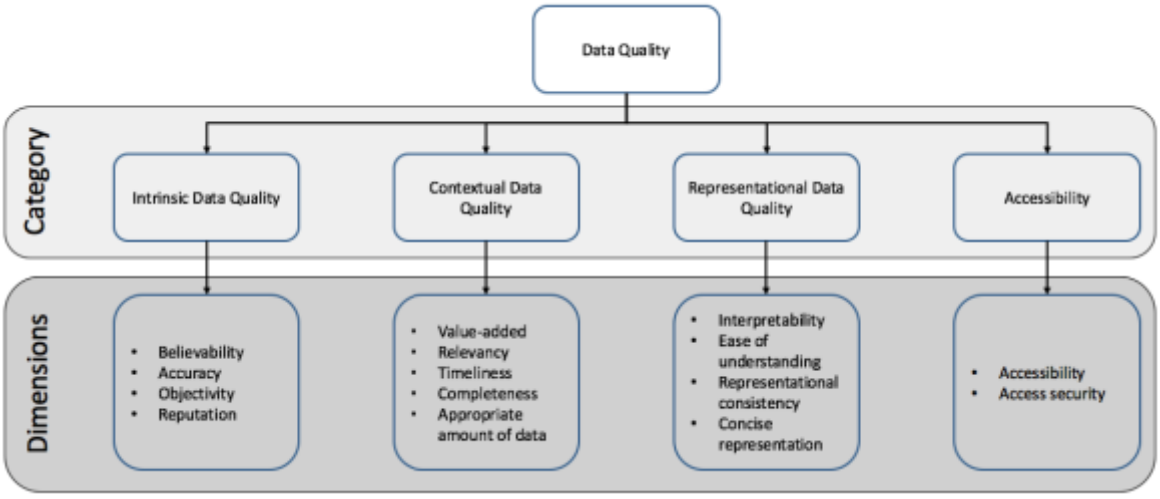

Figure 4 Conceptual Framework of Data Quality (Wang & Strong 1996)

The conceptual framework created by Wand & Strong (1996) is still one of the most significant frameworks related to data quality and it is cited over 1200 times as of April 2016 (Scopus 2016). It is also one of the few studies that is fully focused on the concept of data quality. The conceptual framework of data quality is originally based on consumer viewpoint. Wang & Strong (1996) conducted the study in three phases: (1) an intuitive, (2) a theoretical, and (3) an empirical approach, where the third and last approach is the most

substantial. According to Wang & Strong (1996) the framework has two levels which are the category and the data quality dimensions. The reason for creating hierarchical framework was to make the model more usable. Over fifteen remaining dimensions were just too many for practical evaluation purposes. By grouping the dimension in categories where they support each other makes the model much more simple and balanced (Wang & Strong 1996). They classified the dimensions in four categories, that are supposed to capture the essences of the whole group. Nevertheless, some of the dimensions such as accuracy, completeness and consistency are seen clearly more significant than the others. In Table 2 is presented Wang & Strong's framework with five other studies related to DQ.

Table 2 Data quality frameworks

	Wang et al. (1995b)	**Wang & Strong (1996)**	**Bovee et al. (2003)**	**Liu & Chi (2002)**	**Scannapieco et al. (2005)**	**Huang et al. (2012)**	
Access security		X					1/6
Accessibility	X	X	X			X	4/6
Accuracy	X	X	X	X	X	X	6/6
Appropriate amount of Data		X		X		X	3/6
Available	X						1/6
Believability	X	X				X	3/6
Clarity				X			1/6
Completeness	X	X	X	X	X	X	6/6
Consistency/Consistent representation	X	X	X	X	X	X	6/6
Creditability	X		X				2/6
Currency	X				X		2/6
Ease of manipulation				X		X	2/6
Ease of understanding		X					1/6
Faithfulness				X			1/6
Formality				X			1/6
Interpretability	X	X	X	X		X	5/6
Navigability				X			1/6
Neutrality				X			1/6
Non-fictiousness			X				1/6
Non-volatile/Volatility	X		X		X		3/6
Objectivity		X		X			2/6
Privacy				X			1/6

Relevancy	X	X	X	X		X	5/6
Reliability of Data Clerks				X			1/6
Reputation		X				X	2/6
Retrieval Efficiency				X			1/6
Security				X		X	2/6
Sematic Stability				X			1/6
Sematic	X		X				2/6
Storage Efficiency				X			1/6
Syntax	X		X				2/6
Timeliness	X	X	X	X	X		5/6
Traceability						X	1/6
Trustworthiness of the collector				X			1/6
Unbiased						X	1/6
Understandability						X	1/6
Up-to-date						X	1/6
Useful	X						1/6
Value-added		X				X	2/6

All six studies presented in Table 2 have minor differences, which are either structural, ideological or related to the defined dimensions. Some frameworks consist of sub dimensions, categories or phases which are not specifically presented in Table 2 because it is more essential to understand the entire concept which defines the data quality. The DQ framework created by Wang et al. (1995b) is an attribute-based model. Wang et al. (1995b) see DQ as multidimensional and hierarchical concept, meaning that some of the dimensions must be fulfilled before others can be analyzed. The researchers don't really justify the structure of the model other than it helps user better determine the believability of data which is seen quite important in their framework. The model is generally strongly based on logical analysis. The first category is accessibility because in order to even get the data it must be accessible. Secondly user must understand the syntax and semantics of the data making the data interpretable. Third, data must be useful, meaning that it can be used in decision making process. According to Wang et al. (1995b) usefulness demands that data is relevant and timely. Last category is believability including sub-dimensions: accuracy, creditability, consistency and completeness.

The conceptual framework created by (Bovee et al. 2003) is an intuitive and simplified framework that tries to merge the key features of existing DQ studies. The framework is based on fit for use ideology and the researchers claim that it has only four main criteria:

accessibility, interpretability, relevance and credibility. In reality the creditability consists of accuracy, completeness, consistency and non-fictitiousness making it almost as complex as the other models. According Bovee et al. (2003) the previous studies have failed in distinguishing between intrinsic and extrinsic dimensions because for example the completeness can be classified in both of them depending on the approach. As a specialty researchers name dimension called non-fictitiousness, which means that data is neither false nor redundant. If a database includes records that do not exist or there are fictitious records in existing field, the rule of non-fictitiousness will be violated. Furthermore, the hierarchical structure is exactly the same as in the framework created by Wang et al. (1995b); Accessibility – Interpretability – Relevance and Credibility, but after more specific review there are some discrepancies in the definitions. Bovee et al. (2003) define the interpretability through meaningful and intelligible data. Syntax and semantics can be seen more as a minimum level. Bovee et al. (2003) also highlight the user-specified-criteria in all dimensions and claim that timeliness is just a part of relevancy and not an individual dimension.

Liu & Chi (2002) introduce totally different approach to DQ which resulted a theory-specific and a data evolution based DQ framework. They say that existing frameworks lack a conceptual base, theoretical justification and sematic validity. Existing frameworks are mostly intuitive based and too universal. Liu & Chi (2002) think that data evolves in the process from being collected to being utilized and this evolution is important to take in to account.

Figure 5 Evolutional Data Quality (Liu & Chi 2002)

The model is based in four phases which are: collection quality, organization quality, presentation quality and application quality as presented in Figure 5. Each of the phases has six to eight individual dimensions which are presented in Table 2. According to Liu & Chi (2002) these different models should be used to evaluate data in different stages of life cycle. They claim that the dimensions of previous phases should be included in the assessment but they don't specify how widely or deeply. The idea of evolutional data quality is great and provides a new aspect, but at the same time it is difficult implement. In the scope of this study the Application quality is the most relevant phase. The application quality related dimension of the evolutional DQ framework are presented in Figure 6.

Figure 6 The Measurement of Application Quality (Liu & Chi 2002)

According to Liu & Liu (2002) the first dimension of Application quality is Presentation quality which in turn includes Organization quality and so on meaning that all dimensions will be indirectly measured at the end. This also highlights the problem of the evolutional DQ framework in practice. At the end provided dimensions are very similar to those proposed in existing literature.

Scannapieco et al. (2005) created multi-dimensional framework for DQ which relies entirely on the proposals presented in the research literature. The framework is based on fitness for use ideology and it has only four dimensions which are accuracy, completeness, time-related dimensions and consistency. According to Scannapieco et al. (2005) time-related dimensions include currency, volatility and timeliness and they are all related to each other. They also see that there are correlations among all dimensions. Sometimes the correlation is stronger and sometimes weaker but in most cases they exist. If one dimension is considered more important than the others for specific purpose, it may cause negative consequences on the others (Scannapieco et al. 2005). Unfortunately, they don't specify which of dimensions correlate stronger and what could be the possible effect on the other dimensions.

The newest of the presented data frames is created by (Huang et al. 2012). The DQ framework is designed as part of a study focusing on genome annotation work. In addition to defining the most significant DQ dimensions the researchers wanted to prioritize the DQ skills related to genome annotations. The study was conducted as a survey and there were 158 respondents who work in the area of genome annotations. Based on the findings researchers generated new 5-factor construct, including seventeen dimensions. The model is very similar to the framework created by Wang & Strong (1996) but so was the method of the study too. The most significant differences are naming the categories and the dividing of accessibility category. The reason behind these changes is most likely context-sensitive. Huang et al. (2012) assume that genome community's needs slightly differ from previous studies.

In general, all introduced six frameworks have a lot in common. The amount of dimensions is usually around 15 and the frameworks are divided into categories or phases. Most of them are also based on the fitness for use ideology. What differentiates them is the fact that some of the dimensions are seen to belong below each other and sometimes they are seen as equal, making it hard to get consensus on which of them are more important than the others. The introduced data frames included three main approaches:

1. Empirical (Wang and Strong 1996, Liu & Chi 2002, Huang et al. 2012);
2. Theoretical (Wang et al 1995b, Scannapieco et al. 2005) and
3. Intuitive (Bovee et al. 2003)

According to Scannapieco (2005) and Liu & Chi (2002), despite of the similarities there are neither widely accepted model nor meaning for dimensions. This might be the reason why several practitioners have struggled with the data quality issues too and therefore have provided their tools for defining data quality in practice. In Table 3 five practitioners' solutions for measuring the data quality are presented.

Table 3 Data quality frameworks presented by practitioners

	Kovac et al. (1997)	Mandke & Nayar (1997)	Lucas (2010)	O'Donoghue et al., 2011	Lawton (2012)	
Accuracy	X	X	X	X		4/5
Completeness			X	X	X	3/5
Consistency		X		X	X	3/5
Relevancy			X			1/5
Reliability	X	X				2/5
Timeliness	X			X		2/5
Uniqueness					X	1/5
Validity					X	1/5

Kovac et al. (1997) introduced a data quality framework called TRAQ (timeliness + reliability + accuracy = quality). The model is created for the needs of a database and analytic software provider that wanted to increase their data quality. The framework is originally based on Wang & Strong's (1996) conceptual framework, but in the study only three dimensions were chosen into conceptual TRAQ model. At the end multiple measures were developed only for accuracy and timeliness based on a metric assessment process called RUMBA. RUMBA stands for reasonable, understandable, measurable, believable and achievable, which is the reason for excluding reliability from the measurements. Generally, TRAQ has two main objectives. First, it provides objective and consistent measurement for data quality. Secondly, it provides continuous improvement for data handling process. (Kovac et al. 1997) This is also the way how data quality assessment should be universally utilized.

Mandke & Nayar (1997) claim that there are three intrinsic integrity attributes that all information systems must satisfy. They say that the significance of factors related to data complexity, conversion and corruption has increased due to globalization, changing organizational patterns and strategic partnering causing more and more errors every day. Therefore, Mandke & Nayar (1997) defined accuracy, consistency and reliability to be the most significant DQ dimensions by heuristic analysis. During the analysis approximately eight dimensions were introduced but most of them were defined unneeded as individual

dimensions. For example, accuracy includes completeness and timeliness, while data cannot be accurate if it is not up-to-date (Mandke & Nayar 1997).

The third case is about a Data Quality Management implementation project in Telecommunication sector. The purpose of the whole project was to improve corporate's data quality. Lucas (2010) defined first ten dimensions mostly based on those created by Wang & Strong (1996), but because DQ dimensions should be chosen by the general situation, current goal and the field of application, the count of dimensions was decreased only to the three most important ones; accuracy, completeness and relevancy During the implementation an empirical method based entirely on intuition and common sense was used instead of any formal DQ methodology (Lucas 2010).

Modified early warning scorecard (MEWS) is actually a Patient Assessment-Data Quality Model (PA-DQM) (O'Donoghue et al. 2011). It is created to support decision making processes in patient assessment. Even though MEWS is highly focused on patient assessment and therefore in smaller scale compared to the assessment of huge data warehouses, it is still originally based on well known data quality methodologies and dimensions. Timeliness, accuracy, consistency and completeness were chosen based on questionnaire and workshops where the researchers identified the most significant errors and impacts of poor data quality. According to O'Donoghue et al. (2011), if any of the chosen four dimensions is violated it's clear that the decision will be either wrong or skewed. The results were based on six patient data sets with seven individual variables and therefore the sample is rather low.

The Data Quality Report Card is the most universal of presented frameworks. It is created for validating financial data quality. Lawton (2012) originally defined seven dimensions which match with previous literature. Based on these seven dimensions a user should create an adapted report card with suitable metrics for his needs. Anyway at least validity, uniqueness, completeness and consistency should be included in the report card. According to Lawton (2012) these four dimensions can be assessed using software and the other three; timeliness, accuracy and preciseness require manual comparisons between the records and

real world values, making the assessment significantly more time consuming and difficult to perform (Lawton 2012).

These five frameworks created by practitioners are just a cross-section of real world applications, but they reflect on reality of data quality assessments in practice. Interestingly practitioners define significantly less dimensions than related theoretical frameworks. The reason might be that they have only reported the dimensions and measures which are used within the organization lowering significantly the amount of possible dimensions. From theoretical perspective data quality has a vast amount of dimensions, and by taking them all into account it might be possible to define the absolute quality of data. But this is just a highly theoretical point of view and won't work in reality as several empirical studies have shown. Most of the mentioned dimensions are impossible to measure objectively and have only minor effect on the results. In Table 4 the most cited dimensions are presented.

Table 4 The most cited dimensions

Dimension	Times cited in introduced frameworks
Accuracy	10
Completeness	9
Consistency	9
Timeliness	7
Relevancy	6
Interpretability	5
Accessibility	4

Even though the amount of defined dimension varies widely it is possible to identify the most common and important ones. By analyzing existing frameworks, we can see that the following dimensions: accuracy, completeness, consistency and timeliness, are the most significant ones when evaluating data quality (Wang & Strong 1996; Jarke & Vassiliou 1997; Mandke & Nayar 1997; O'Donoghue et al. 2011; Lawton 2012; Zaveri et al. 2012; Hazen et al. 2014). The meaning of timeliness is in fact a bit higher as it seems to be because currency and volatility are often seen as a part of timeliness or even mixed with it. The relevancy and interpretability are in turn many times seen either as a category or part of the first four dimensions. The last of the list, accessibility, reflects more on the data systems than

the data itself. Therefore, it can can be assumed that the first four mentioned dimensions will capture the essential data quality as presented in Figure 7.

Figure 7 Holistic Data Quality Framework

The holistic data quality framework consists of only four dimensions which is the most important change to the introduced theoretical frameworks. This should not be an issue, though, while the chosen dimensions represent all clearly different attributes of data quality. The idea of holistic framework is to simplify the previously introduced concepts but still capture the essential factors of DQ. Furthermore, it is acknowledged that there are correlations among the dimensions.

3.2 Dimensions of data quality framework

Accuracy, completeness, consistency and timeliness are defined to be the most important dimensions of data quality, but what do they really mean and include? Many researchers might use different names for similar dimensions making the situation even more confusing than it really is. In the following part each of these dimensions are explained more specifically in order to get consensus on their real meanings.

Accuracy has several definitions in existing literature. Wang & Strong (1996) define that accurate data is certified, error-free, correct, reliable and precise. According to Huang et al.

(2012) and Bovee et al. (2003) accuracy means that the records are just correct and free of error. Accuracy is also an extent where the data in system represent the real world as it is (Wand & Wang 1996). A simple example of accuracy would be a data record such as a customer's address in a customer relationship management system which should correspond to the street address where the customer actually lives. In this case the data is either accurate or inaccurate, because accuracy is entirely self-dependent (Hazen et al. 2014).

Completeness represents an extent where a record should be in the data set if it exists in the real world. According to Wang & Strong (1996), completeness is about breadth, depth and scope of information contained in the data. Zaveri et al. (2015) defined that "completeness refers to the degree to which all required information is present in a particular dataset". Completeness is a complex and subjective measure. Scannapieco et al. (2005) have a similar definition with Wang & Strong (1996) but in addition define that completeness consists of Schema completeness, Column completeness and Population completeness. The simplest way to measure whether the data is complete is to check if a record exists when required. For example, in customer data, all customers should have a name. If name is not defined the data is most likely missing values and is therefore incomplete. A more difficult situation is to ensure that the data includes everything needed for answering the desired question. Total amount of euros, orders etc. should match some external system. Though it still doesn't eliminate the possibility that some records are cumulative or combined. (McCallum 2012. 227-228) Liu and Chi (2002) define completeness through collection theory, which is heavily related to their ideology of data evolution. According to them all data should be collected as per a collection theory they are collected, but simultaneously they also agree with the existing definitions that all existing data must be included as a result.

Consistency belongs in the representational category (Wang & Strong 1996). Identifying the category may not be necessary when defining consistency, but it gives us a hint about the related attributes. According to Laranjeiro et al. (2012) consistency is "the degree to which an information object is presented in the same format, being compatible with other similar information objects". Pipino et al. (2002) didn't define consistency but consistent representation which refers to format as well. Consistency also means that data is free of logical or formal contradictions and data is understandable without particular knowledge

(Liu & Chi 2002; Zaveri et al. 2012). Some researchers claim that consistency has also inter-relational aspects, meaning that one part of data has an effect to another part of data. (Zaveri et al. 2012; Hazen et al. 2014; Batini et al. 2009). An example of consistency issue can be that costs are presented once in euros and second time in dollars. The previous example becomes extremely dangerous if the record itself doesn't include the unit but it is determined in some other field, making it difficult to detect visually.

According to Wang & Strong (1996) **timeliness** is the age of data and belongs to the same category as completeness and is therefore contextual. Pipino et al. (2002) define timeliness as "extent to which the data is sufficiently up-to-date for the task at hand". That means that date must not correspond with the real world all the time if it is not affecting the end results. According to Batini et al. (2009) there is no general agreement for time-related dimension, but currency and timeliness are often used to represent the same concept. Anyhow timeliness is in most cases measured by combining volatility and currency which results in two metrics (Wang et al. 1995; Bovee et al. 2003; Scannapieco et al. 2005). Currency refers to the delay between the real world and information systems and volatility measures the time difference between observation time and the invalid time. (Zaveri et al. 2012).

Definitions of mentioned dimensions are complex which is partially caused by the adopted fit-for-use approach. As an example completeness does mean that all real world records are included, but it doesn't mean that data must include all existing data in the world. Data is complete when it includes all relevant events. Simplified definitions for the dimensions of holistic framework are presented in Table 5.

Table 5 Definitions of Data Quality Dimensions

Dimension	Definition
Accuracy	A record represent values as they are in the real world.
Completeness	Data includes all relevant events, records and values that exist in real world.
Consistency	All records are presented in same format and are therefore understandable.
Timeliness	A record is up-to-date for the intended use.

In this study all dimensions are seen strictly from consumer point of view, meaning that there is no absolute accuracy, completeness, consistency or timeliness. For example,

timeliness doesn't mean that the data must be exactly in time but rather suitable for later usage.

3.3 Data quality testing

Nowadays organizations have more and more information partially due to past trends, where everything must be recorded, and partially because current technology enables producing data by relatively low cost. But data itself is worth nothing if it can't be trusted or used to support decision making. The risk of poor data quality is also becoming remarkably high when larger and more complex information recourses are utilized (Watts et al. 2009). In this section we will go through phases of data quality assessment, and metrics.

Data quality has been the object of active research and practice for decades but still the field lacks generally acknowledged methodology for assessing the quality in practice. This might be caused by the continually evolving concept of data. Data quality assessment methodologies usually contain several phases. The most common three phases of data quality assessments are: State reconstruction, Assessment and Improvement (Batini et al. 2009). State reconstruction focuses on collecting contextual knowledge on organizational processes, data collection and usage of data (Batini et al. 2009). The assessment phase can be derived to several steps, but by the simplest way there are only two steps, which are performing the assessment or measurements and comparing the results (Pipino et al. 2002). In addition, many researchers identify the steps of identification of critical areas and process modelling (Batini et al. 2009). The last phase of assessment is improvement (Pipino et al. 2002; Batini et al. 2009). In general, there are two ideologies for improvement, data-driven and process-driven approaches. Both apply various techniques to improve the data quality. Data-driven is usually used for individual tasks and it concentrates on improving DQ afterwards, while process-driven approach tries to tackle root causes of bad data by enhancing the actual data management processes. (Batini et al. 2009)

Data quality assessments and metrics are usually done by ad-hoc basis. As mentioned earlier, data quality is a multidimensional concept. The data dimensions are either objective or subjective (Pipino et al. 2002; Watts et al. 2009). Dimensions such as accuracy and

completeness are objective, while the quality of information doesn't vary regardless the context of usage. On the other hand, the quality of dimensions such as relevance and believability cannot be measured objectively, making them subjective quality dimensions. For example, believability depends strongly on the user's experience and suppositions. For a student some information looks reasonable and at the same time it appears unconvincing to an expert. (Watts et al. 2009; Parssian et al. 2004)

In many studies metrics are presented either as an example or on an ideological level. Surveys are a quite common part of DQ assessments too but are out of scope in this study while the idea is to create scalable data quality assessment model. According to Pippino et al. (2002) there are three pervasive functional forms that can be used as a basis when creating overall quality metrics. These are simple ratio, min or max operation and weighted average. **Simple ratio** can be used for measuring the ratio of desired outcomes to the total outcome (Pipino et al. 2002). In many cases the favored form is the number of undesirable outcomes divided by the total outcome subtracted from one, because this way the result represents the "goodness". According Pipino et al. (2002) simple ratio suits well for traditional DQ metrics such as free-of-error, completeness and consistency.

Min or max operation is suitable when handling dimensions that require multiple DQ measurements. Min and max operations both work similarly. They compute either the minimum or maximum value of chosen data quality measurements. The Min and max operators fit well for dimensions such as believability, appropriate amount of data, timeliness and accessibility where individual DQ indicators have strong correlations. Timeliness is a classical case where max operation has been used. Timeliness can be measured by volatility and currency as presented earlier. If volatility is defined to be 0,6 and the currency 0,8 then the timeliness will be 0,8. **Weighted average** is an alternative to min and max operation. Weighted average is more sophisticated, but also more challenging to use. It is based on the idea that some variables are more important than the others. Therefore, using weighted average method requires good understanding of the influence of each DQ indicator to the overall assessment of the dimension. (Pipino et al. 2002) Lawton (2012) proposed similar approaches as min or max operations and weighted average for defining the overall quality

of certain data set. The total quality of one dimension can be measured by summing up the count of violating record and then dividing by total amount of all records (Lawton 2012).

But what is then a desired outcome? It relies partly upon the reviewed dimension, but in most cases a desired outcome reflects the real world (Pipino et al. 2002). The issue arises when we don't have correct reference data or it is not possible to compare the data to the real world, which are often the case. This is the reason why assessments are done uniquely almost every time, depending on the data available and desired results of assessment.

Completeness is a difficult dimension to measure because it can be viewed from a number of perspectives. As mentioned earlier, completeness can be divided in schema, column/row and population completeness (Pipino et al. 2002; McCallum 2013, 227-228). This means that each viewpoint of completeness should have separate measurements. Column completeness can be measured by the number of missing values divided by total count subtracted from one. But how can all relevant rows be ensured to be included in the data? In many case it is just assumed that all the data is included, but what if one file was just forgotten to export by human error. Usually measuring schema completeness requires additional information from external systems but in cases where we know the target, for example such as the count of states in U.S. it might still be measured (Pipino et al. 2002). Another challenge related to completeness is the dilemma of missing values. There is a significant difference if the value does not exist or it is not known (Scannapieco et al. 2005). For example, we cannot know if a missing email address of a customer is existing or not resulting totally different outcomes. Therefore, the completeness of certain field cannot be evaluated if no information of real world values are provided from other sources.

Even & Shankaranarayanan (2005) introduced an interesting viewpoint of value based quality measurement. According to Even & Shankaranarayanan (2005), structural completeness [C_i] can be calculated by the *equation 1*:

$$C_{S,a} = \frac{E_a}{M_a} \quad (1)$$

where:

E_a - count of existing values in column "a"

M_a - count of all values in column "a"

The structural completeness is similar to the simple ratio and thus the basic way to compute completeness. Imagine a situation where only one row out of five is missing. Described situation is presented in *equation 2*:

$$C_i = \frac{4}{5} = 0{,}80 \tag{2}$$

As computed in equation 1.2, the completeness still stays quite high. Another way to measure completeness is to compute it based on Amount-Factored Intrinsic Value as presented in *equation 3* (Even & Shankaranarayanan 2005):

$$C_f = {(\sum_{n=1}^{N} M_n C_n)}\Big/{(\sum_{n=1}^{N} M_n)} \tag{3}$$

where:

M_n – Value stored in field

C_n – Completeness of record [n] defined by weighted-average

N – Records in data set, which are indexed = 1 . . N

This approach takes into account the intrinsic value of a missing cell. What would be the completeness if the value of that missing row in column "Amount €" (Missing row 100 000€) is twice as much as the value of other rows together (50 000€)? Based on Amount-Factored Intrinsic Value the completeness should be computed as presented in *equation 4*:

$$C_f = \frac{50\ 000€}{150\ 000€} = 0{,}33 \tag{4}$$

As we can see the results are entirely different depending on the method of calculation. Same ideology can be used to measure Customer-Factored Intrinsic Value meaning that "Customer Id" is used as a scaling factor (Even & Shankaranarayanan 2005).

In addition to these, the dimensions are usually evaluated by using business rules (Woodall et al. 2013). Business rules are actually a set of rules which should not be violated or the record can be handled as an undesired outcome. Correctness itself can be interconnected, meaning that some values have an allowed range or because of one value another value must be something (McCallum 2010, 232). As an example, a machine with theoretical capacity of 2 tons of ice cream per hour won't produce under any circumstances over 48 tons of ice cream per day. If the daily total production is over 48 tons, the value can be defined to be incorrect and therefore undesired outcome. Other similar business rule could be that the age of a machine cannot be negative or repairing doesn't take any time. Creating comprehensive set of business rules requires some degree of domain knowledge, but will enable finding certain undesired outcomes. (McCallum 2010, 232)

4 MODEL FOR TESTING DATA

4.1 DQA Target and Raw data

Data quality assessment always requires at least some domain knowledge, but it does not mean that it would not be possible to create scalable model for data models with certain level of similarity. In this chapter the previously defined holistic data frame will be adapted in order to create scalable model for measuring the quality of industrial maintenance data. The data quality assessment (DQA) model is executed by Python code for maintaining efficiency also with large data sets. The first task in every assessment is defining the goal of assessment. After that we can continue with raw data and its characters. The last step is defining the suitable metrics and building the adapted model.

The data is the core of any analytics. Without data it is not possible to perform analytics. On the other hand, the results will be misleading or at least skewed if the data is corrupted. In this case the need for data quality assessment comes from the analytics and the simulations performed later on. The same data is used to perform standardized reports for analyzing the current state of maintenance activities in different kind of factories. The Maintenance Quality Analysis (MQA) uses around half of the provided data, which means that all columns must not be evaluated. This is because of the fit for use approach. The original MQA is created based on quick review of the quality of provided data, which has partially affected to the chosen fields. Therefore, the results of DQA should be fairly good. The exact fields which are evaluated will be defined after introducing the raw data.

The raw data for assessment originally consists of three individual exports from different information systems with unique functions, which has had a major effect in general to the raw data and to the quality of particular fields. The three predetermined data models; Transaction data, Master data 1 and Master data 2 are connected to each other with one or more fields. All data models are based on relational data model structure presented in chapter two. The data is provided by a third-party company meaning that there is no direct access to the data or information systems. Furthermore, the data exports are transferred through an

additional software that should have converted the data in standardized form. In Table 6 the key features of the evaluated data models are presented.

Table 6 Characteristics of the raw data

	Transaction data	Master data 1	Master data 2
Fields	15	31	77
Relevant fields	11	7	6
State	Dynamic	Static	Static

The transaction data is originally from an invoicing system and intended for purchase order follow up instead of analyzing current state of maintenance actions. Therefore, the focus areas of transaction data are probably elsewhere than desired in this study. Transaction data is the only one of data models which includes time series data. Of course this data is also just a representation of past events at a certain point of time and it might change when time passes. Transaction data has fifteen fields and one row represents an individual order which is also the reason why the Transaction data may update over time.

Master data 1 is the simplest of analyzed three files. It represents a static state of machine records at the export date. There are 31 fields in the data model, though only few of them are needed. The last of analyzed files is called Master data 2 and it has 77 fields. Master data 2 represents the static state of preventive maintenance programs. Even though Transaction data has the lowest amount of fields, most of them are relevant for later analysis. At the end these three data models form an entity which is needed to understand the real world situation as presented in Figure 8.

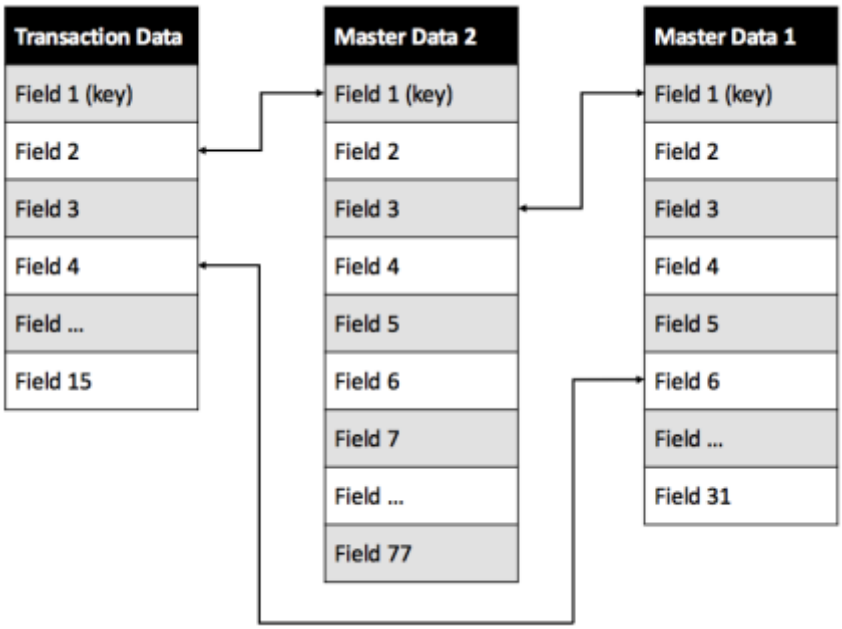

Figure 8 Database model

Figure 8 presents the linkages between the analyzed data models. Each of the data models have at least one link to the other two. Furthermore, in order to effectively analyze transaction data, both master data files are needed. One for enhancing the transaction data and the second as a key for certain values. In all three data models there are fields that must be unique, but it doesn't mean that the unique field would be same for all of them. The unique field of one data model is just one record for another data model.

4.2 Framework adaptation for use

The holistic framework for data quality doesn't fit as it is but it must be adapted in the context. In Table 7 the changes, reasons and expected impacts to the results are presented.

Table 7 Changes to the holistic framework during adaptation

Change framework	Why?	Expected impact to results
Consistency merged with accuracy	The metrics for consistency are similar to the metrics of accuracy	No impact to the results, while inconsistent would be also inaccurate in this case
No implementation of Timeliness	Timeliness is not essential dimension for intended usage of the data	Might have minor affect to the results, but the lack of timeliness will reflect the results of

		accuracy and completeness indirectly

Based on the earlier consensus, accuracy, completeness, consistency and timeliness define the overall quality of analyzed data, but because the data quality is an exceedingly case sensitive concept, few changes are done to the introduced framework. First of them is merging consistency with accuracy while the possible ways to measure accuracy and consistency are extremely similar in practice. In reality it is difficult to make difference between violations of accuracy and consistency rules. At the end it is not even particularly important to know if an undesired record is inaccurate or inconsistent. More important is to detect the records that do not reflect the real world values or are understood incorrectly for one or other reason.

The second change to the original framework is that the timeliness was left outside from the assessment. Timeliness related metrics were not implemented for several reasons. First, the data is used for analyzing state of maintenance during the past few years. It is acknowledged, though, that the latest records of Transaction data might change over time but we can make an assumption that the change will be consistent over time, meaning that the amount of outdated records should be constant despite of the timing of data quality assessment. Furthermore, the conformity of data files should be affected if the data lacs timeliness. Secondly, timeliness can be measured by currency and volatility, meaning that the faster data gets old the more current it should be. In industrial maintenance count of changes is low and they happen seldom, allowing us to ignore the timeliness.

Now when the goal of data quality assessment is defined as well as the basic characteristics of data are known, the designing of data quality assessment (DQA) model can begin. The four most significant dimensions were qualified based on earlier literature and the consensus to choose only two of them for measurements is quite clear. The basic of metrics for each dimension are presented in Table 8.

Table 8 Dimensions and metrics of DQA model

Dimension	Group of metrics	Variable
Accuracy	• Unique values • Group of allowed values • Full agreement in different files • Full agreement in different fields	• $A_{1,n}$ • $A_{2,n}$ • $A_{3,n}$ • $A_{4,n}$
Completeness	• All real world tuples included • All records have a value • All records have a not null value	• $C_{1,n}$ • $C_{2,n}$ • $C_{3,n}$
Consistency	• N/A	• N/A
Timeliness	• N/A	• N/A

All metrics are created by an ad hoc basis and they are based on the individual characteristics of each column. The metrics mostly measure intrinsic quality because there is no possibility to compare the records to the real world values. Metrics for accuracy can be grouped in four categories. Unique value is the key accuracy related metric but it can only be used in few fields. It also detects just the duplicate values. Usually all data models have a field that is used as a key which means that there cannot be duplicate values. A good example of a field where business rule of uniqueness is used is machine ID. All machines should have unique ID because otherwise records cannot be traced correctly.

Group of allowed values is another simple business rule for measuring accuracy. Some fields can only have predefined values such as order status. Order status can be for example, completed, started, etc. but not a single number or a letter. Defining these groups is easy but requires some domain knowledge. Predefined list of allowed values is an effective way to evaluate records but it's also vulnerable for changes. The last two categories are based on restrictive facts what a record can or cannot be. Even if it is not possible to measure the accuracy by comparing data to the real world, it can be done by comparing individual files and fields. Conformity reveals at least clear errors in data, especially if one of the data samples can be considered to be correct or the best possible reflection of real world (Espetvedt et al. 2013). In this case transaction data and master data 2 are compared to master data 1, which is believed to be the best representation of current situation. Similar approach might also be used by comparing values in different fields in a data model. As an example,

the initial preventive maintenance interval of machines on same maintenance route should always be the same. Especially these last couple business rules of accuracy are fairly similar to consistency related metrics which was the main reason for merging consistency. After all a record is accurate even though it is presented in euros instead of dollars but the inconsistent representation might cause undesired results in practice if interpreted incorrectly. At least in this case there aren't additional fields that would help us recognize inconsistent values from inaccurate values.

Completeness is measured at first by the size of data sample in order to respond to the question whether the data sample really includes all relevant tuples. Because measuring schema completeness usually requires additional information from external systems, it was possible to implement schema completeness related metric only in one field. Unfortunately, the metric for schema completeness is in fact able to detect only some of the violations, not all them. The two other metrics for completeness are pretty similar to each other. The only difference between them is that either the zero values are accepted or not. Not all but most of the fields should have a value. As an example all tuples should have status, thus if there is a tuple without status we can make an assumption that the record exists but is missing. A more complicated situation is determining whether a zero is acceptable or not. This becomes an issue especially when analyzing costs and lengths or other numeric fields. In many cases information systems have a zero as default value. Completeness of dates, hours and some costs were partially measured in this case without accepting zero values.

All individual metrics are created based on the ideology of simple ratio introduced in the section 3.3. The metrics and business rules are also designed so that rules don't overlap each other. An example python code is presented below:

```
1for index, rows in WO.iterrows():
2   if pd.notnull(WO.loc[index, 'TYPE'] == True):
3      if WO.loc[index, 'TYPE'] == 'A11':
4         if WO.loc[index, 'ROUTE'] in RouteList:
5            WO.loc[index, 'AccType'] = 0
6         else:
```

```
7            WO.loc[index, 'AccType'] = 1
8      else:
9          WO.loc[index, 'AccType'] = 0
10   else:
11     WO.loc[index, 'AccType'] = 0
```

The previous code is for one of the metrics for computing the accuracy of the Type-field. As seen, there are several preconditions that must be met before the actual rule is violated. Therefore, an undesired record can be detected and counted only once. The second row of the code is the precondition for completeness and the third row excludes the rule of allowed values. By simple modifications this code could also be used for measuring several dimensions at once but it was not seen practical because it would cause loss of transparency. Since the individual rules are designed not to overlap each other, the count of not desired records can be added up. *Equation 5* presents how the quality of certain field is calculated. Variables of different metrics are presented in Table 8.

$$Q_n = 1 - \left(\sum A_{k,n} + \sum C_{k,n}\right) \Big/ T_n \tag{5}$$

where:

$\sum A_{k,n}$ – Sum of records violating Accuracy related business rules in field [n]

$\sum C_{k,n}$ – Sum of records violating Consistency related business rules in field [n]

T_n – Total count of records in field [n]

Total quality of data table is then computed as presented in *equation 6*. In Table 9 is shown an example of computed data quality assessment. The total count of records in the example is 100 tuples.

$$TQ = \sum Q_n \Big/ n \tag{6}$$

where:

$\sum Q_n$ – Total quality of field [n]

n – Count of analyzed columns

Table 9 Structure of DQA-model

Field	Dimensions		Total
	Accuracy	**Completeness**	
Field n	2 records violate metric $A_{1,n}$	6 records violate metric $C_{1,n}$ 4 records violate metric $C_{2,n}$	$Q_n = 1 – (2 + 6 + 4) / 100$ $Q_n = 88\%$
Field n+1	10 records violate metric $A_{2,n}$ 2 records violate metric $A_{4,n}$	N/A	$Q_n = 1 – (10 + 6 + 4) / 100$ $Q_n = 80\%$
Field n+2	0 records violate metric $A_{2,n}$	N/A	$Q_n = 1 – (0) / 100$ $Q_n = 100\%$
Field n+3	7 records violate metric $A_{2,n}$ 8 records violate metric $A_{3,n}$	4 records violate metric $C_{3,n}$	$Q_n = 1 – (7 + 8 + 4) / 100$ $Q_n = 81\%$
Total	**N/A**	**N/A**	**TQ = (0,88+0,8+1+0,81)/4** **TQ = 87,3 %**

Each field is measured individually by as many metrics as possible, in this case meaning one to three individual metrics. In some cases, there is no business rule that could be implemented for a certain dimension and field. The total quality of the field will be calculated by adding up all undesired records and dividing it by the total amount of records evaluated and then subtracting the outcome from one. The overall quality will be computed after that by the average of all columns. The weighted average was not chosen this time because the importance of individual values could not be defined while data is used later for several analyzes. This way the result of data quality of one data table can be presented by one overall score at the end.

The simple ratio method was chosen over other possible methods because it is rational and straightforward. The approach of Amount-Factored Intrinsic Value introduced earlier might be better than simple ration for measuring data quality but it also requires additional information. It cannot be used if the missing tuples or total value of certain field is not known. Furthermore, Amount-Factored Intrinsic Value focuses subjectively to one field and is not therefore practical for computing the total quality of a data model.

4.3 Results of data quality assessment

The designed DQA model was tested with data from nine case companies. The purpose of testing is to find out if the DQA -model can be used for measuring the data quality without any additional knowledge. The case companies are different size and they represent various industries. Case companies 5 and 6 belong to the same corporation, but they aren't identical. Since the data exports are done with similar search terms and methods, they should be comparable. Table 10 shows the row counts of analyzed data.

Table 10 Properties of raw data

	Transaction data	Master data 1	Master data 2
Case 1	3 666	15 478	1 251
Case 2	3 310	10 967	1 621
Case 3	7 566	2 399	1 147
Case 4	1 821	12 794	1 232
Case 5	10 451	2 214	1 069
Case 6	9 190	1 330	608
Case 7	13 301	17 872	7 203
Case 8	20 344	29 740	1 690
Case 9	11 820	22 845	1 813

The amount of analyzed data varies significantly. The biggest variations are in the Master Data 1 but also the size of Transaction data fluctuates considerably. The reasons behind these differences are most likely the size and the type of factory, the time period of transaction data and the data processes in general. However, these factors should not have any impact on the data quality.

The testing began with quick visual checkup that the data exports really are as they should be, all required columns exist and the overall structure is generally consistent with the predetermined data models. The only changes to the data model was the renaming of some columns. This is because the validation process is executed with python code so that it can be upscaled as easily as possible. Furthermore, performing the DQA requires the usage of python or other similar tools in the long run, because eventually the amount of data will be

too big for Excel. The results of data quality assessments were as expected. In order to make results easier to understand, the total scores of master data 1 and 2 were merged together. However, they both represent static state of current situation. The overall quality scores of master data tables were multiplied because incorrectness of one table will reproduce. That's how we get only two scores, one for transaction data and one for master data. The scores are presented in Figure 9.

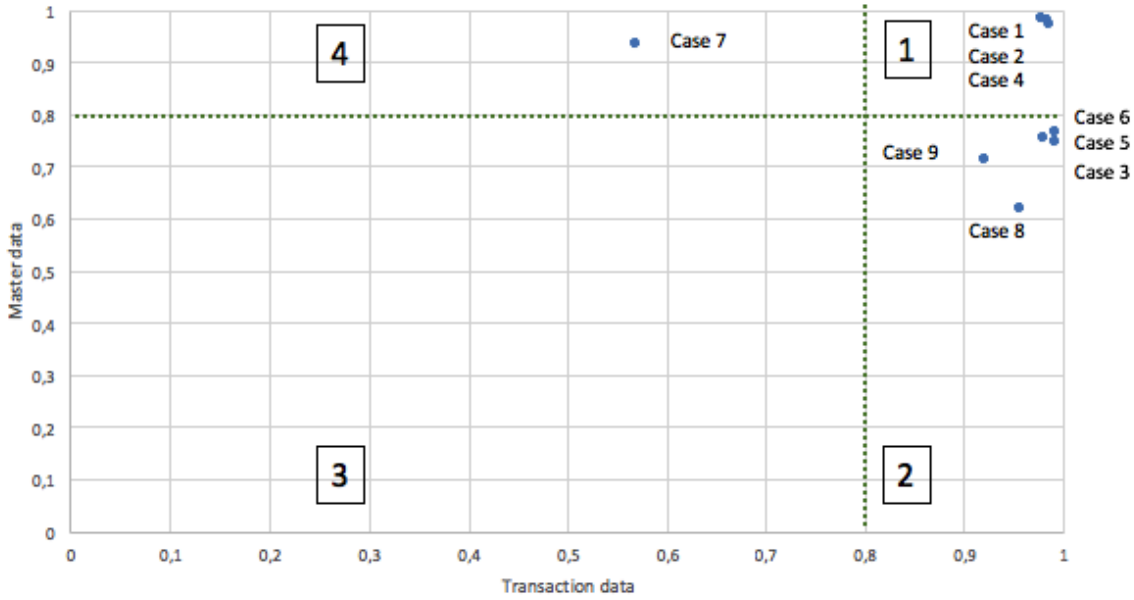

Figure 9 Data quality scores

We can see in the Figure 9 that the biggest variations are in the quality of master data, but the lowest score is caused by bad transaction data. Furthermore, the result chart is divided into four categories in order to better understand the reasons and effects of data quality. Each of the quadrats has its own characteristics. Category one is the desired outcome of the data quality assessment, meaning that the master data and the transaction data are fit for use and include only minor violations. Three of the case companies belong into this group. The quality of master and transaction data must be over 0,8 in order to belong to the category one. More interesting is that all three companies are really close to each other and are almost perfect quality according to the results. A closer examination reveals that the most important common factor of these case companies is the completeness of data. Most of the measured fields are complete and those which are not complete miss only very few values. The overall

quality of category one is really good and there should not be any major issues in later analysis.

Category two means that the transaction data is high or good quality but the quality of master data is poor. The reason for bad master data can be almost anything, but it usually reflects the imperfect knowledge of current situation. The fact that master data is bad is risky because master data is also the foundation of all collected data within a firm. If master data is incorrect it will most likely sooner or later affect the transaction data too. In category two there is clearly seen a group of three case factories and interestingly two of them are from the same corporation. The quality of transaction data of these three firms is as high as in group one but master data is clearly lower quality. The reason for poor quality lies this time on one certain field, which is either inaccurate or incomplete. The field may not be the most important one considering the big picture, but will definitely cause skewed results in later analysis.

The category three is the worst possible case. The data has enormous issues and won't most likely be suitable for the desired purpose. None of the case companies fell in the category three this time. The last category is number four. In the category four the quality of master data is on the acceptable level but the transaction data lacks the desired quality. This is usually caused by bad data collection processes or poor master data. The transaction data of the case company 7 is clearly below the accepted level caused as a result of multiple metrics. The most significant reasons for low quality, though, are few fields that are either missing records or the values are presented in different format as required. The disagreement with desired form of representation, doesn't really mean that values are inaccurate but it will cause problems and wrong conclusions. What is interesting is that the master data as well as the key fields of the transaction data are almost perfect quality, raising the question, what is the actual effect of the missing values. It might be that the tuples with missing values are mistakes or duplicates without no further value for analysis. Making that kind of assumption would presumably skew the MQA reports and simulations later, though.

Generally, the results of data quality assessments seem rational. Few of the tested data samples were almost perfect for use, other few were good but would cause with high

probability a bit skewed results. The last three and worst data samples from companies 7, 8 and 9 would have issues during MQA process, but would still be possible to do after some adjustments and modifications. The results of MQA would be also flawed due to the poor raw data. The results of individual metrics had also substantial variations, meaning that the total score was not just a result of few metrics but all of them. So on we can assume that metrics work as wanted and the metrics are able to catch several kinds of quality violations in both dimensions.

Even though the model seems to is suitable for the designed task it has some weaknesses too. This kind of quality testing only measures the dimensions which have predefined metrics. Unfortunately, in the adapted DQA model there are fields that have only one metric. Moreover, a metric measures sometimes really narrow area of attributes of a certain value. As an example, rule of uniqueness checks only if the value really is unique. The uniqueness is only a certainly small part of accuracy though. A value might be unique but at the same time it is missing one or more letters, the characters are in the wrong order or it just doesn't exist in the real world. All fields are also measured with the same importance, despite of the count of defined metrics or other factors. The field which has more metrics than others is therefore more likely to fail, though that is same for all analyzed data files.

Defining the quality of data is extremely difficult, because data can be wrong in so many different ways. Next are presented few specific and special cases related to the data quality, which are not tested by the created model. As mentioned earlier, completeness can be measured at least by Schema, Column and Population completeness. The transaction data is complete by one definition when all transaction that have actually happened in real world are included, or when all rows have an accepted record. The third methodology is called population completeness and it measures the percentage of objects that are represented in the data. The Figure 10 represents the population completeness of the transaction data.

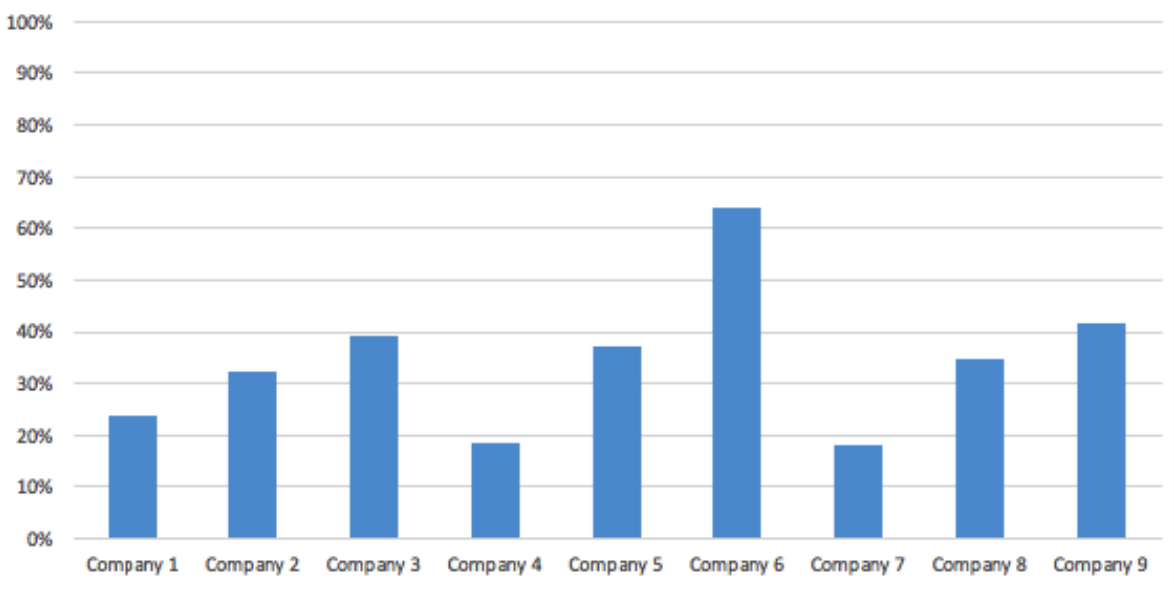

Figure 10 Population completeness

Data from companies 1, 2 and 4 are evaluated to be the highest quality, but unfortunately the population completeness is relatively low. The population completeness of the transaction data is computed by the percentage of machines identified in master data which have at least one record in transaction data. The fact that population completeness is low is not directly related to the bad quality. The issue is more about how useful the data is at the end. For simulation and analytical purposes, the best possible outcome would be complete data with several transactions per machine. It is common to have less relevant data than desired, though (Scannapieco, Missier & Batini 2005), at least in the industrial maintenance. This comes from the fact that transaction data is only created when a breakdown or a failure happens or other maintenance actions are done. And because failures are rare individual events, there is a high probability that no failure has occurred during the observation period. Furthermore, rarely occurring events mean in the scope of industrial maintenance events that happen once in ten year if even then. Second reasonable reason for low population completeness is inaccurate data. The object accuracy is presented in Figure 11.

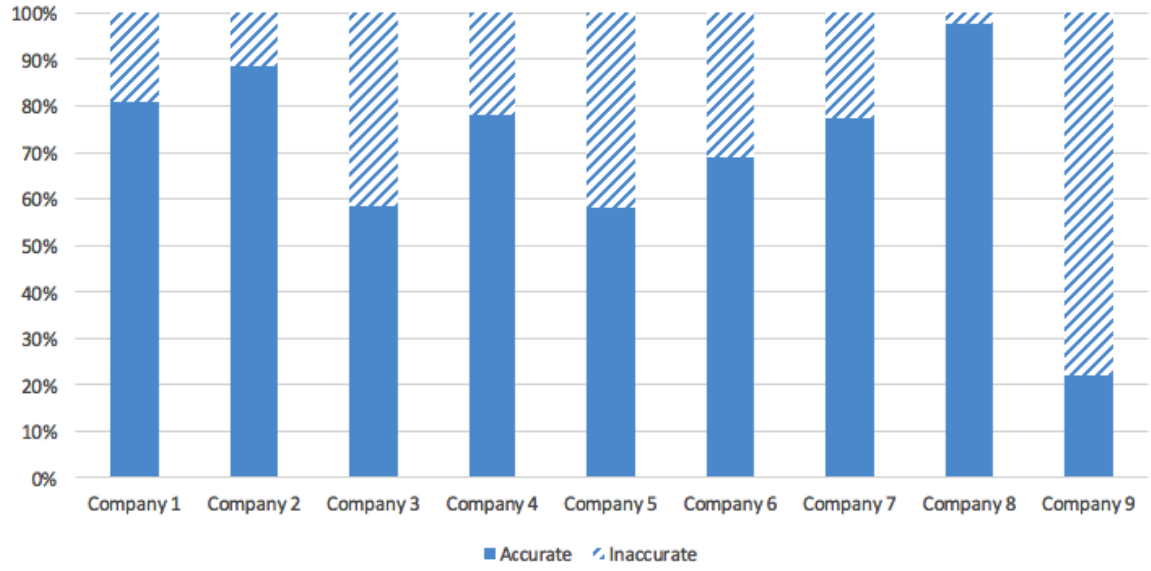

Figure 11 Object accuracy

Machine hierarchy structure is a significant part of industrial maintenance and it is presented in the master data. The machine hierarchy structure usually consists of machines and gates where the last mentioned is more an ideological than a physical object. That is also the reason why all transactions should be allocated to machines rather than gates. The object accuracy is therefore computed by the count of transactions related to machines and gates where the transaction related to machines are considered as accurate and others inaccurate. This time the example companies 1, 2 and 4 are in top class as they should. The object accuracy is debatable as a measure, because it is affected by many factors. A bad machine hierarchy structure may cause some gates and machines to be mixed, affecting directly to the object accuracy. The machine hierarchy structure may also miss some machines that exist in the real world, forcing the transactions to the wrong object in cases where the actual object is not defined. These two are in fact caused by bad master data, but because of the third and maybe the most common reason for low object accuracy, it is not fair to deprecate the master data by low object accuracy. The third reason for low object accuracy is a poorly designed data collection process. The data might be as it was collected and no discrepancies exist but the data is still wrong. The data collection process is consequential for data quality while the data might be collected unintentionally wrong. This often becomes an issue when the education for data collection is neglected. The collectors don't even know how the data should be entered in the information system, resulting in bad data. It is also possible that the

collectors do not similarly understand the machine hierarchy structure in the information system than in the real world, causing inaccurate data.

The last significant observation concerning DQA results focus on the weight of different metrics. Results show that the overall quality scores of transaction data are from 92,1% to 99,2% (except the company 7 with transaction quality of 56,8 %), being pretty high. Despite of high quality it must be noticed that even small inaccuracies might be significant from a different perspective. In Figure 12 are represented inaccuracy of type field and the percentage of total cost which is affected by the violated rows.

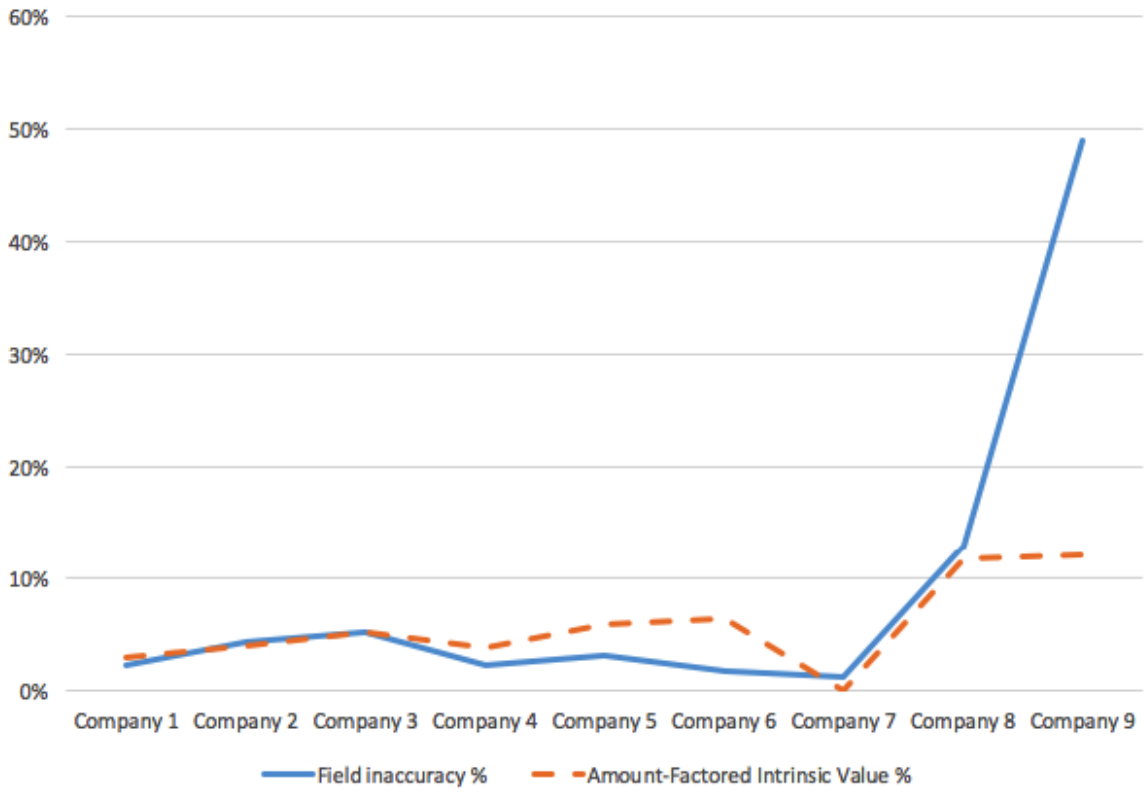

Figure 12 Type inaccuracy and related costs

As seen in Figure 12, the inaccuracy of Type field is between 1,3% and 49,1% and this includes only the rows that exist. The percentage of total cost based on Amount-Factored Intrinsic Value theory introduced by Even & Shankaranarayanan (2005) would be the dashed line varying from 3,0% to 12,1% of total costs. And this is just related to the violations concerning accuracy related metrics of the chosen field. If the violations of completeness are also added, the Amount-Factored-Intrinsic Value would be even higher.

Not to mention violations related to all attributes of certain rows when the Amount-Factored Intrinsic Value would be extremely higher. Though that is not a desirable outcome of the data quality analysis while the purpose is to gain general understanding of data quality not just the effect to the costs. This and the earlier introduced Object accuracy and Population completeness represent the complexity of data quality related issues.

The model does not really measure the absolute quality, if there even is such a concept. The absolute quality includes so many factors that they are just impossible to measure in large scale. And beside the fact that we don't have the possibility to compare values neither to real world values nor to the data that is considered to be the best reflection of real world values. Therefore, the results of the adapted DQA model are subjective. It is not actually believed that any of the case companies really have almost perfect data. The data is just good for intended purpose and it does not have a lot of the clear mistakes. Therefore, this kind of metrics can be used for evaluating a company's data in general. The fact that some company's data is measured to be much worse than the other companies' data refers at least to bad data management.

5 DISCUSSION

5.1 State of information management

The data quality is not just an information system related issue, even though it is often considered to be. Managing data consists of several phases and concepts that all have a significant impact on the data quality. Information systems are just a physical tool for helping with the task at hand, but the real issues are somewhere else. The information systems are originally created to manage increasing amounts of data and to provide insights for decision making. Because of the cheaper data storing and past interest to collect all possible data, the amount of data available has become enormous, making the data management an extremely important topic. Moreover, in many cases the data is not stored in the same system, but in several ones that are not even directly connected, causing inconsistencies and inaccuracies (Smith and McKeen 2009). What is the benefit of data if it is not correct, available or trustworthy? The entire ideology of data sciences lies on the data, how it's available and what is the quality. The whole information process includes data collecting, maintaining, transferring and archiving, meaning that there are several phases where the data might corrupt. Therefore, it is important to understand the complexity of data process.

The modern data management concepts aim to ensure the data quality by comprehensive ideologies that pay attention to the entire data process. An important part of data management is creating the policy and clear methods for maintaining the quality. The data should not be collected without any further intentions and plans; this is because the data does not have any value before it is used for decision making. The modern ideology of Big data and analytics is contradictory with the previous argument. In fact, the whole concept of Big data is based on the fact that the more data is collected the better. The Big data supporters claim that all data should be collected and stored for later usage even though the data might not be useful at the moment. I don't judge the idea entirely because sometimes we don't even know what we really want or need. Therefore, collecting additional data might come in handy at some point of time. But at the same time collecting additional data takes resources from somewhere else as well as makes finding the needed data more and more

difficult. A good example of overwhelming amount of data is the internet. Once you try to find information in the internet you have to find the correct information among several undesired search results. Same situation applies to the company's data if there are number of simultaneously maintained comparable data models. In other words, unneeded data creates noise that overwhelms the wanted data. An even bigger issue are the resources. As shown in this study, if a company is not even able to manage the current data with high quality, why should they collect any additional data. What is the use of knowing someone's address if it is not correct or up-to-date? That is why I strongly argue that companies should focus on the relevant data at first and only after that begin collecting additional data. A solution for these problem is provided by modern data management concepts such as MDM.

The Master Data Management is one of the newest data management concepts that focuses on the company's most important data which is called master data (Otto et al. 2012). The master data and the transaction data have several dissimilarities, such as the master data represents usually a static state of real world whereas transaction data is in most cases dynamic data. The master data is also seen as origin of the transaction data, highlighting the importance of master data. In this study two of the analyzed data models are considered to be master data and therefore key assets within a firm but is this really the situation? Master data 1 is definitely important while it consists of the machinery. How could the maintenance operations be managed effectively if the objects are not known? Unfortunately, this is the reality in many cases. Maintaining machinery lists are often seen as just a mandatory task, whereas real interest of management lies on daily operations and costs. The master data 2 represents the current state of preventive maintenance programs, and therefore more or less determines the predefined maintenance operations. So on it can be assumed that also the master data 2 is important from managerial perspective.

The study doesn't take into account the specific information management practices in industrial maintenance, which is definitely an issue and weakness. But what could be the negative impact of such approach? As in many times in usage of Big data or business analytics, the amount of data sources is simply too high to be analyzed individually. In this case there would have been nine companies located across Finland to investigate more specifically, making the whole analysis an unfeasible task. The possible insight gained

through explicit empirical research would help improve the DQA model, but it's not obligatory. As the results show, the general understanding of information management practices is sufficient background to design a useful DQA model.

5.2 The holistic framework

The data quality can be defined by several ways but no general framework exists that fits all situations (Pipino et al. 2002). The reason for it is most likely that the overall quality of data is seen strongly as a subjective attribute. For example, the last name of a customer should correspond the customer's current last name but if the data is used for analyzing the buying patterns of millions of customers the last name might not be an interesting information. In that case customers are just numbers in data. Previous studies and practitioners have presented various frameworks for analyzing the data quality. Even though the number of dimensions may differ significantly, most of the researches still agree with the fit for use approach as well as the most important dimensions. Accuracy, Completeness, Consistency and Timeliness are the key factors when evaluating the overall data quality (Wang & Strong 1996; Jarke & Vassiliou 1997; Mandke & Nayar 1997; O'Donoghue et al. 2011; Lawton 2012; Zaveri et al. 2012; Hazen et al. 2014).

Nonetheless, few issues remain concerning the dimensions and correlations that exists among them. First, only four dimensions were chosen, which might cause important dimensions of quality to be ignored in the holistic framework. Those kind of dimensions are for example security or accessibility. Both of the dimensions are argued to be substantial, while if you don't have access to data there won't be anything to evaluate. Does it mean that the data is bad? If the data lacks security, it might corrupt, but this should also be detected by accuracy. Therefore, I would argue that accessibility and security are more like perquisites for data and not dimensions of it. Secondly, any of reviewed studies didn't analyze specifically the correlations between dimensions, even though the existence of correlations is recognized by Scannapieco et al. (2005). For example, the accuracy and timeliness have a clear correlation, while the values of records might change over time. But if the metrics are really designed for individual dimensions it should not matter. The correlations are more substantial when data-driven improvement methods are applied, thus

meaning a decision between different dimensions. The quality of one dimension can be improved with the cost of another. The simplest example is deleting the inaccurate records.

5.3 The design of DQA

Despite the fact that Consistency and Timeliness are essential factors of data quality, they were dropped out at the adaption phase for individual reasons. The metrics for measuring consistency are in reality very similar with metrics related to accuracy, making it almost impossible to distinguish them. The reason why metrics for timeliness were not implemented is more case sensitive. The industrial maintenance data is seen to be quite consistent over time, making the timeliness a less important dimension and thus allowing us to leave it out. At the end, data quality depends mostly on the later usage of the data because of the fit for use ideology. The assessment should be done by comparing the records to the real world values but unfortunately it is very rarely possible, forcing us to use alternative methods. The best way remaining to measure the actual quality is to use business rules, or more specifically expressed a set of conditions that must be met. The self-explanatory problem concerning business rules is that the business rules measure only some factors. The more there is domain knowledge, the more business rules can be defined. Other drawback related to business rules is that some dimensions or attributes cannot be measured with business rules. As an example, ensuring the type or object of an old event in transaction data is really difficult. One option would be surveys, but they are subjective and don't really detect the violations but rather provide understanding about how good the data might be in general. The important question is if the designed DQA model is able to reach the intended level of accuracy. More specific analyses would be useful but they require analytical tools, like fitting or outlier detection and are thus beyond the scope of this study.

The adapted DQA model mostly detects major violations, while the usage of different kind of business rules doesn't really allow better assessment. The used methods seem to be sufficient for the task at hand though. The results show that designed metrics are able to detect various kinds of violations as well as define the weaknesses of each company's data. Most of the companies have problems with the master data, which is interesting. The master data represents the most important business objects, which is in this case the machinery and

preventive maintenance plans. There is also a growing number of regulatory and legal provisions that companies must comply with, which should also drive the companies towards better master data. So can the result be correct? As claimed earlier in this chapter, the master data is unfortunately not the key focus of management in many cases. In general, the master data 2 was even worse than the master data 1, which is understandable, while the importance of master data 2 is a bit lower.

The results of transaction and master data analysis differ notably. Looks like the metrics related to transaction data are not as powerful as the metrics for master data sets. Excluding the case 9, the overall scores of transaction data are over 92%, which can be considered as high quality. In comparison, the overall score of master data varies between 62% and 98%. The reason for the level of dissimilarity lies probably on the DQA model, and more precisely on the metrics not on the data itself. In the case of transaction data, it seems like the metrics are not able to discover as many different kind of violations as the metrics defined for master data sets. The current metrics rather measure if the data is error-free and unique than the fact that a record reflects the real world value. Therefore, there is a clear need for improvement. How it should be done is difficult to say. Improving the metrics demand more domain knowledge about the practices and other affecting factors. However, all results of DQA are relative and therefore the transaction data and master data should not be compared together. It is more rational to compare different cases, because then the metrics would be same for all of them resulting more understandable scores.

In an ideal situation the quality of master data and transaction data should correlate, while there is a clear connection among them. That also explains the result of case 7. The overall score is due to few fields which are clearly undesired in this context. Because of the nature of transaction data, it repeats the same error again and again. One single mistake in master data can cause tens of undesired records in transaction data. As an example, incorrect object of preventive maintenance program will cause corrupted transaction data every time when preventive maintenance program is performed. The last mentioned issue is beyond the scope of this study, though, and is therefore excluded from the DQA model. Nevertheless, it explains why certain fields have significantly lower quality caused by a systematic error in the process.

What is the impact of excluding consistency and timeliness from the DQM model? The logic to cut down the count of measured dimensions is empirically justified. It is also acknowledged that it will surely affect the results of DQA. But whether the difference is significant in practice or not is a more substantial question. To make it clear, consistency was not exactly left out, it was merged with the accuracy. So in that sense there should not be any major consequences to the total score. The only thing that is affected is loss of transparency, while now it is impossible to say based on the results which of the violations are caused by inconsistency and which by inaccuracy. When comes to timeliness it is more questionable. We can be sure that not all of the data is high quality concerning the dimension of timeliness. But does it matter, while timeliness measures whether data is up-to-date for the task at hand (Pippino et al. 2002). In the case where time plays primary role, timeliness could be implemented. An example of such situation could be a data quality assessment of boarding passes or table reservations. Situations where it is important that the customer information is available in time. But in a case like this study where the analyzed data is already days old when provided, the value timeliness is low or negligible. Also from more practical point of view I would question the usefulness of timeliness. Let's think again the situation with table reservations. If data is not up-to-date, meaning that the waitress doesn't have information of recent reservation, it means that the data won't be complete. In turn, if recent change of reservation is not updated, the data won't reflect the real world, meaning that the data is inaccurate. Under these circumstances we can assume that the model with two dimensions is reliable and functional as it is at least for the intended use.

6 CONCLUSIONS

6.1 Focus

The need for this thesis arose during the development phase of Maintenance Quality Assessment. MQA aims to analyze the current state of maintenance operations in various kinds of factories and plants. It consists of tens of analyses from several perspectives in order to provide comprehensive understanding of the state of maintenance operations at a certain factory. The results of MQAs has showed that there is clear need for data quality assessments, while the data sets have significant flaws that have caused major errors during the MQA process as well as questionable results. In this case it was highly important that the data quality assessment is scalable and based entirely on the data.

According to previous literature and this study, the data quality is definitely a subjective matter, because it depends highly on context. In this thesis is presented several substantial findings which are interesting from theoretical as well as managerial point of view. The aim of this thesis was to define how data quality can be measured universally and how well the holistic framework suits in the context of industrial maintenance data. In order to solve this research problem two research questions and four sub research questions were determined. In Table 11 are presented the research question and the answers to them. The results will be explained more specifically later in this chapter.

Table 11 Research questions and answers

Research question	Answer
RQ1. How to measure data quality from holistic perspective?	Data quality is a subjective matter, meaning that if data is fit for intended usage it is high quality.
SQ11. What are the dimensions of data quality?	The most important dimensions of data quality are Accuracy, Completeness, Consistency and Timeliness.
SQ12. How can data quality be measured?	Data quality can be measured by a set of business rules, surveys or comparing data to real world values
RQ2. How does the data apply to the holistic framework in industrial maintenance?	Holistic framework suits well into industrial maintenance after a few changes. Results are justified and useful for measuring data quality.

SQ21. How does the framework need to be adapted for the use?	Measuring data quality is highly case dependent which is the reason why only Accuracy and Completeness were implemented in practice.
SQ22. How accurate is the measurement of data quality?	The results of measurement are relative, but they provide great insights of firms' overall data quality.

6.2 Theoretical implications

Evaluating data quality in general is a difficult task and requires some domain knowledge of the data processes and the data at hand. The driving force behind data quality is fit for use approach which defines the most of the data quality attributes and requirements. There is no simple or clear way to determine what does fit for use mean and or how it should be utilized in practice, because data quality and data quality metrics are highly case sensitive. A flight passenger list that is updated during flight might be up-to-date for authoritative of arrival country but outdated for flight personnel.

The theory of data quality dimensions is inconclusive. In existing literature there are defined over twenty-five different dimensions which are partially overlapping and conflicting. Also the meanings of different dimensions were questionable, even though some general agreement exists. Most studies are exceedingly theoretical and focus on defining multiple data quality dimensions, which is also their weakness. The researchers have defined a lot of sensible dimensions such as reliability or understandability that are just either impossible or extremely difficult to measure. The highly theoretical approach of data quality might have caused that there is only little empirical evidence on data quality measurement. Despite of the fact that the data quality is case dependent, in this study it is proved that the four most important dimensions from theoretical point of view are accuracy, completeness, consistency and timeliness. They all represent totally different perspectives of data quality and thus provide a solid insight to overall data quality. Other dimensions can be seen either as part of them or as prerequisites.

In this study is introduced a framework to evaluate data quality with just four domain dimensions. The reasoning for four dimensions comes partially from existing literature and

partially from empirical study. Accuracy, completeness, consistency and timeliness are broad concepts and include the most essential factors of data quality. Furthermore, defining strictly dimension specific metrics is an extremely difficult task in practice, while the dimensions are overlapping and affected by each other. In theory, if data is accurate, complete, consistent and timely, it should be also high quality, thus accuracy means that data reflects exactly the real world values whether it is reliable or not. Same applies to completeness. Some researchers have defined dimension called relevancy (Wang et al 1995b; Wang & Strong 1996; Bovee et al. 2003; Huang et al. 2012) which is in fact closely related to completeness while complete data should include all relevant data. These are great examples of the width of defined four domain dimensions.

The last theoretical implication gained in this study is the way to compute overall DQ score for a data set. The aim of introduced DQA model is to provide simple method to evaluate the data, which is the reason why the quality of separate dimension and fields were calculated together by using simple ratio and average scores of measured fields. The method might be a bit crude but it provides easily understandable and comparable results.

6.3 Managerial implications

The data quality is not just an information system related issue but it concerns the entire company. It is acknowledged that information systems are in key role when data is handled but they are just tools for helping with the task. Therefore, management should understand to put enough resources in designing processes and managing data. For usability of firm's data, it is incredibly important to understand the quality of your data. Data is used to support in decision making, which means that the low quality of data will cause implementations of decisions that are based on skewed understating of the current situation.

Fit for use based data quality approach requires understanding on the business problem, which makes data quality once again a managerial problem. There is no use to collect all possible information if it is not useful or required at any phase. Furthermore, if data is collected without clear policy or intention the quality will decrease at some point. In practice data quality can be measured only on a relative level, but accurate measurement of data

quality is not yet possible. This doesn't mean that the scalable DQA model would not be useful, though, but just the other way around. The major errors of data can be detected quite easily as shown in the results. The assessment detected flaws aren't just mistakes in the scope of this study, but also significant indicator for managers in the all nine case companies' data sets. Moreover, several data sets have high amounts of violations in certain fields, clearly indicating the problem areas.

All detected errors of master data will eventually cause corrupted data and results. Therefore, the goal for master data 1 should be that there are no detected violations at all. This is because the designed metrics detect only significant violations like duplicate machine identity numbers or missing parents. The positive side of errors in master data is that they are quite simple to fix, while master data represents the current state of real world. The correction must be done by hand and not with data-driven improvement methods though. The reason for the errors in master data are probably caused by the lack of regular data quality checks or undefined ownership of the data and the data process. The errors detected in master data 1 should have been discovered in routine check and corrected after that. The errors in master data 2 are more likely caused by an undefined ownership. In few companies there is also a case sensitive root cause, which is entirely representational issue and therefore requires data modifications. In that case data is only bad for the task at hand. Of course these kind of errors should not be underestimated while if it is not noticed, it would definitely cause skewed results in later analysis.

Transaction data is more complicated to interpret because in most case companies it was analyzed to be relatively high quality. The Company 7 which had obviously lower transaction data quality was caused by two main reasons. First of them is different representational logic of data than desired. Once again, even though the data would be correct the different form requires additional domain knowledge and manipulation of data. The other issue is either related to missing data in columns or entirely missing columns. Both are definitely substantial problems no matter how the data will be used later. Some analytical methods may enable filling the missing values but in this case where around 70 percentage of values are missing data manipulation would be highly questionable if not impossible. With the missing columns there is usually nothing to do without additional data. The issues

with the transaction data are probably caused by missing policy for data collection and management. It is not fault data collectors if certain fields are not defined when the data model is designed. The enhancing of the transaction data should therefore be done by improving the processes. It is not the only way but by improving the process the quality of data should increase by itself.

As mentioned, the results are relative while there is no possibility to detect all violations in practice. Discovering all possible violations would require extremely high amount of different metrics or different kind of approach such as comparing with real world values. Therefore, the results of DQA should be compared with other cases or assessments performed at different points of time. The transaction data and master data are not comparable together while the defined metrics as well as the nature of data are different. The reason why some transaction data sets have generally higher scores lies on DQA, not on the data. There are unique metrics for each data model as well as each field, even though the basic ideas of individual metrics are similar. Results of data quality assessment could and should be used for improving the data quality or at least the flaws should be acknowledged in decision making.

6.4 Future research

A few interesting future research questions arose from this study, some of them are more theoretical and some more empirical. The theoretical part of this study is agreed with the four most important dimensions, as well as with the measuring methods, but unclear remains the weight of each field and how the total score of data model should be calculated. The overall score of data model is now calculated by average of each field, which is not an incorrect way to do it, but it is a coarse method and might cause skewed results. As an example, few duplicate records were detected in key fields during the assessment. The issue with duplicate values in key fields is that the entire row is then useless. The impact of such error is much bigger than the impact of missing value in some other field, which may not even affect the other fields. Same issue applies with the entirely missing rows. Those are in general extremely hard to detect and they don't mean just one missing record but all of them.

These situations decrease the creditability of data quality assessment and should be studied further.

Another interesting topic is the accuracy of the assessment. The usefulness of DQA model is certain, because it was able to detect vast amount of errors but the universal accuracy demands further studies. Because the metrics don't measure the absolute quality, we cannot be sure what is the percentage of total violations detected. The accuracy of the assessment is also closely linked with the introduced data quality matrix (Figure 9), which consist of two factors: transaction data quality and master data quality. The plot is divided into four general categories which represent the overall data quality and therefore the current situation. Now the limit value that divides the categories is set to 80% on both axels, based on intuition and previous experience. The categories are indicative while some significant violations like missing missing machine ID will require immediate data manipulation before later analysis can be performed, but they are still treated as any other violations. Therefore, it would be interesting to investigate how well the categories reflect the real world and what should be the actual limit value.

REFERENCES

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*. vol. 41, no. 3, pp. 16-16.52.

Baxter, P. & Jack, S. 2008. Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*. vol. 13, no. 4, pp. 544-559.

Bose, R. 2009. Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*. vol. 109, no. 2, pp. 155-172.

Bovee, M., Srivastava, R.P. & Mak, B. 2003. A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*. vol. 18, no. 1, pp. 51-74.

Caballero, I., Serrano, M., & Piattini, M. 2014. A Data Quality in Use Model for Big Data. *In Advances in Conceptual Modeling*. pp. 65-74. Springer International Publishing.

Chen, Y., Zhu, F. & Lee, J. 2013. Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method. *Computers in Industry*. vol. 64, no. 3, pp. 214-225.

Espetvedt, M.N., Reksen, O., Rintakoski, S. & Østerås, O. 2013. Data quality in the Norwegian dairy herd recording system: Agreement between the national database and disease recording on farm. *Journal of dairy science*. vol. 96, no. 4, pp. 2271-2282.

Even, A., & Shankaranarayanan, G. 2005. Value-Driven Data Quality Assessment. *IQ*.

Haug, A., Stentoft Arlbjørn, J., Zachariassen, F., & Schlichter, J. 2013. Master data quality barriers: an empirical investigation. *Industrial Management & Data Systems* vol. 113, no. 2, 234-249.

Hazen, B.T., Boone, C.A., Ezell, J.D. & Jones-Farmer, L. 2014, Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*. vol. 154, pp. 72-80.

Hellerstein. J. 2008. Quantitative data cleaning for large databases. United Nations Economic Commission for Europe. [WWW-document]. [Accessed. 12.5.2016] Available as: http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.

Holsapple, C., Lee-Post, A. & Pakath, R. 2014. A unified foundation for business analytics. *Decision Support Systems*. vol. 64, pp. 130-141.

Huang, H., Stvilia, B., Jörgensen, C. & Bass, H.W. 2012. Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science & Technology*. vol. 63, no. 1, pp. 195-207.

IBM. 2016. [WWW-document]. [Accessed 3.6.2016] Available as: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Iverson, C. 2014. MONETISING BIG CUSTOMER DATA. *Accountancy SA*. pp. 28 28,30,32.

Jarke, M., & Vassiliou, Y. 1997. Data Warehouse Quality: A Review of the DWQ Project. *IQ*. pp. 299-313.

Juran, J.M. 1989. *Juran on Leadership for Quality: An Executive Handbook*. New York: The Free Press.

Kovac, R., Lee, Y.W. & Pipino, L. 1997. Total Data Quality Management: The Case of IRI. *IQ*, pp. 63.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. 2015. A Survey on Data Quality: Classifying Poor Data. *Dependable Computing (PRDC)*, 2015 IEEE 21st Pacific Rim International Symposium on. pp. 179-188). IEEE.

Lawton, J. 2012, The Data Quality Report Card. *Journal of Government Financial Management*. vol. 61, no. 1, pp. 24-30.

Li, X. & Jacob, V.S. 2008. Adaptive data reduction for large-scale transaction data. *European Journal of Operational Research*. vol. 188, no. 3, pp. 910-924.

Liu, L., & Chi, L. 2002. Evolutional Data Quality: A Theory-Specific View. *IQ*. pp. 292-304.

Liu, J., Li, J., Li, W. & Wu, J. 2015. Rethinking big data: A review on the data quality and usage issues, *ISPRS Journal of Photogrammetry and Remote Sensing*.

Lucas, A. 2010. Towards Corporate Data Quality Management. *Portuguese Journal of Management Studies*. vol. 15, no. 2, pp. 173-195.

Mandke, V.V. & Nayar, M.K. 1997. Information Integrity: A Structure for its Definition. *IQ*, pp. 314.

Marr, B. 2015. *Big Data; Using smart big data, analytics and metrics to make better decisions and improve performance*. 1st ed, John Wiley & Sons Ltd, West Sussex, United Kingdom.

Meeker, W.Q. & Hong, Y. 2014. Reliability Meets Big Data: Opportunities and Challenges. *Quality Engineering,* vol. 26, no. 1, pp. 102-116.

O'Donoghue, J., O'Kane, T., Gallagher, J., Courtney, G., Aftab, A., Casey, A., Torres, J. & Angove, P. 2011. Modified Early Warning Scorecard: The Role of Data/Information Quality within the Decision Making Process. *Electronic Journal of Information Systems*

Evaluation, vol. 14, no. 1, pp. 100-109.

Otto, B. 2012. How to design the master data architecture: Findings from a case study at Bosch. *International Journal of Information Management*. vol. 32, no. 4, pp. 337-346.

Otto, B., Hüner, K. & Österle, H. 2012. Toward a functional reference model for master data quality management. *Information Systems & e-Business Management*. vol. 10, no. 3, pp. 395-425.

Parssian, A., Sarkar, S. & Jacob, V.S. 2004. Assessing Data Quality for Information Products: Impact of Selection. Projection, and Cartesian Product. *Management Science,* vol. 50, no. 7, pp. 967-982.

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

Ramakrishnan, R. & Gehrke, J. 2000. *Database management systems.* Osborne/McGraw-Hill.

Scannapieco, M., Missier, P., & Batini, C. 2005. Data Quality at a Glance. *Datenbank-Spektrum*. vol 14 pp. 6-14.

Scapens, R.W. 1990. Researching management accounting practice: The role of case study Methods. *The British Accounting Review*. vol. 22, no. 3, pp. 259-281.
Scopus. 2016. [WWW-document]. [Accessed 20.4.2016] Available as: https://www.scopus.com/

Silvola, R., Jaaskelainen, O., Kropsu-Vehkapera, H. & Haapasalo, H. 2011. Managing one master data-challenges and preconditions. *Industrial Management & Data Systems,* vol. 111, no. 1, pp. 146-162.

Smith, H.A. & McKeen, J.D. 2008. Developments in practice XXX: master data

management: salvation or snake oil?. *Communications of the Association for Information Systems*. vol. 23, no. 1, pp. 4.

Tallon, P.P. & Scannell, R. 2007. Information Life Cycle Management. *Communications of the ACM*. vol. 50, no. 11, pp. 65-69.

TDW 2002, The Data Warehousing Institute. 2002. Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. [WWW-document]. [Accessed. 5.6.2016] Available as: http://download.101com.com/pub/tdwi/Files/DQReport.pdf

Tuck, S. 2008. Is MDM the route to the Holy Grail&quest. *Journal of Database Marketing & Customer Strategy Management*. vol. 15 no.4, pp. 218-220.

Wang, R.Y., Reddy, M.P. & Kon, H.B. 1995 (b). Toward quality data: An attribute-based approach. *Decision Support Systems*. vol. 13, no. 3, pp. 349-372.

Wang, R.Y., Storey, V.C. & Firth, C.P. 1995 (a). A framework for analysis of data quality research. *Knowledge and Data Engineering, IEEE Transactions on*. vol. 7, no. 4, pp. 623-640.

Wang, R.W. & Strong, D.M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*. vol. 12, no. 4, pp. 5-33.

Wand, Y., & Wang, R. Y. 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*. vol 39, no. 11, pp. 86-95.

Watts, S., Shankaranarayanan, G. & Even, A. 2009. Data quality assessment in context: A cognitive perspective. *Decision Support Systems*. vol. 48, no. 1, pp. 202-211.

Woodall, P., Borek, A. & Parlikad, A.K. 2013. Data quality assessment: The Hybrid Approach. *Information & Management*. vol. 50, no. 7, pp. 369-382.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. & Auer, S. 2015. Quality assessment for linked data: A survey. *Semantic Web*. vol. 7, no. 1, pp. 63-93.

Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R. & Andrews, M. 2015. *Big Data Beyond the Hype, A Guide to Conversations for Today's Data Center*. 1st ed, McGraw-Hill Professional, New York, United States.