# POLITECNICO DI MILANO

Corso di Laurea Magistrale in Computer Science and Engineering

Scuola di Ingegneria Industriale e dell'Informazione

# A Flexible Approach to Data Quality Exploration for Big Data Analysis Results

Relatrice: Dr. Cinzia Cappiello

Tesi di Laurea di:
Pavel Gichevski
Matricola: 838627

# Acknowledgement

*I would like to acknowledge everyone who took part in the process of my academic accomplishments. It has been a long but incredible journey followed by lifelong experiences.*

*First of all, I would like to express my gratitude towards Professor Dr. Cinzia Cappiello for her support and guidance while working together on this thesis.*

*Special thanks go to all my friends, especially the ones that came with me in Milan. Without you for sure it wouldn't be the same, and I must say I am happy how we handled everything together.*

*Last, but not least, my deepest appreciation goes to my family for their continuous support and belief in me. You all kept me going and I will always be grateful for that.*

# SOMMARIO

Questo lavoro è realizzato per proporre un'implementazione di uno strumento di esplorazione della qualità dei dati che consente la visualizzazione e l'esplorazione dei risultati ottenuti da un modulo Data Quality as a Service. Il modulo è inserito nell'architettura del progetto EuBRA-BIGSEA dove il Politecnico di Milano entra come partner di collaborazione. In generale, il progetto mira a sviluppare una serie di servizi cloud che potenziano i Big Data Analytics per facilitare lo sviluppo di enormi applicazioni di elaborazione dati.

La qualità dei dati nei Big Data ha un enorme impatto nel processo decisionale e quindi richiede la comprensione dei risultati dell'analisi di qualità. Per comprendere i risultati di un'analisi di qualità, dobbiamo esplorarli e catturare il loro reale valore di qualità. Generare valore da Big Data è probabilmente la parte più importante di un'analisi di qualità perché l'utente deve prendere una decisione definitiva se i punteggi di qualità dei dati sono sufficienti per essere utilizzati nei prossimi calcoli e software di statistica, oppure i dati dovrebbero avere una nuova qualità analisi con impostazioni modificate.

I risultati di un'analisi della qualità dei dati possono variare in base a due fattori: l'oggetto dei dati di interesse e le misure che eseguono la valutazione. La ridefinizione dell'oggetto dati porta a una variazione dei risultati di qualità poiché ogni aspetto dell'origine dati può essere valutato su un insieme di misure e viceversa, ciascuna configurazione delle misure fornisce punteggi diversi. Al fine di aiutare gli utenti nell'esplorazione dei risultati e nel processo decisionale, questa tesi è stata proposta. Trova il suo posto come componente di supporto del servizio di qualità nel progetto BIGSEA. L'idea è di consentire ai professionisti e ai neofiti nel campo dell'analisi della qualità nei Big Data un modo di esplorare e apprendere il processo di analisi attraverso un'interfaccia utente guidata. Inoltre, l'interfaccia utente fornirà una serie di funzionalità che consentiranno l'ordinamento, la ricerca, il filtraggio personalizzato e persino l'esportazione dei dati. Ciò salverà l'uso iterativo dell'utente del servizio di qualità, che significa meno potenza di elaborazione.

# ABSTRACT

This work is realized to propose an implementation of a Data Quality Exploration Tool that allows visualization and exploration over results obtained of a Data Quality as a Service module. The module is placed in the architecture of the EuBRA-BIGSEA project where Politecnico di Milano enters as a collaboration partner. In general, the project aims to develop a set of cloud services impowering Big Data Analytics to ease the development of massive data processing applications.

Data quality in big data has a huge impact in decision making and so it requires understanding of the quality analysis results. To understand the results of a quality analysis we need to explore them and capture their real quality value. Generating value from big data is probably the most important part of a quality analysis because the user needs to make a final decision if the data quality scores are good enough to be used in next computations and statistics software's, or the data should pertain a new quality analysis with modified settings.

The results of a data quality analysis can change by varying two things: the data object of interest, and the measures that perform the evaluation. The re-definition of the data object leads to variation of the quality results since each aspect of the data source can be evaluated over set of measures, and vice versa, each configuration of the measures gives different scores. In order to help the users in results exploration and decision-making this thesis was proposed. It finds its place as a support component of the Quality Service in the BIGSEA project. The idea is to allow professionals and newcomers in the field of quality analysis in Big Data a way of exploring and learning the process of analysis by a guided User Interface. Besides, the UI will provide a set of features that will allow sorting, searching, custom filtering, and even exporting data. This will save the user iterative usage of the quality service, meaning less processing power.

# TABLE OF CONTENTS

# LIST OF FIGURES

# Chapter 1

# INTRODUCTION

## 1.1 GENERAL

Since the beginning of the 21st century and the massive internet coverage that the world now has, we are witnesses of big changes in the technologies. Not so long ago it was possible to read the news on our phones, to search for job opportunities, or even to chat with someone over the internet and be “online” without interruption. With this change every IT company oriented their services and adapted to the new era. This lead to the Internet of Things, to Cloud computing, and AI. As the technologies started being accepted by the world a lot of information started to be accumulated, processed, and even bargained on the market. Just the fact that we carry over 100GB of data on over phone tells us how the measurements and processing power changed with the years.

All these changes in the field of data and the IOT attracted the attention of a lot of leading companies and educational institutions to invest in further processing of the data they accumulate. Based on this we saw a lot of discoveries in terms of medicine, finance, statistics, weather, etc. Later, it became a trend to use Big data in order to understand the customers, to predict their logic, to sell them the things they exactly need.

It became possible to collect data from various sources and contents in instants of seconds and that lead to even bigger sets of information. However, it’s not about quantity but the quality, and so it is in the data. It must have a high-quality in order to bring value, which can later be used in decision-making. This requires an assessment framework and assessment process oriented on Big Data because once the size changes we cannot apply the same analytical processes we did on small sets of data. In order to propose a solution for this, Politecnico Di Milano as a collaboration partner of the EU-Bra BIGSEA project proposed an assessment framework as a Quality Service. The service can be invoked by users/applications by using API’s. Later, when the results are ready the user should be able to explore them and understand the level of quality of the data over different quality measures and aspects. This could be an issue since the results being evaluated on different aspects and over different quality measures, all together in one place as a raw data can only cause a confusion and no pattern can be extracted. This is where this thesis finds its place, as a support component of the QS. The idea is to provide guided visualization and exploration over the results set so that the user can make a final decision if the data has quality to be used in further processing software's, or it should pertain new quality analysis with modified settings.

## 1.2 EU-BRA BIGSEA PROJECT

EUBra-BIGSEA is a project funded in the third coordinate call Europe-Brazil and it is focused on the development of advanced QoS services for Big Data applications. The idea is to provide an architecture that will predict resource allocation and processing power when it comes to Big Data analysis. The project is based on API's in order to serve request on different applications. In particular, the platform is able to manage and store different types of data and offers a set of Big Data services. The services are:

- Data ingestion service that is used to load the data from sensors and devices distributed over the city.
- Query service that is used to make selections on the stored data source in order to retrieve information.
- Data quality service that is used to evaluate the quality aspects of the stored data sources. The produced quality metadata from the quality analysis will be used by data mining applications, predictive models and descriptive models.

As stated, this thesis will relate to exploration of quality analysis over Big Data, so, in our interest will be the DQaaS (Data Quality-as-a-Service). The service was built in a collaboration between Politecnico Di Milano and the BIGSEA project. Its main goal is to provide information about the quality of requested Big Data sources. Also, the architecture behind the service allows parallel processing on a set of nodes in order to address volume and velocity requirements.

It is tested in a real scenario, the public transportation system of the city of Curitiba, in Brazil. The analyzed data sources are a heterogeneous data sources collection that includes: the registration of the ticket validation of the users, the records related to the localization of the buses in the city, the weather conditions, the events that are extracted from the social networks posts.

With a set of provided API's the users/applications can request analysis in order to understand different aspects of the data set as being evaluated over different quality measures. The evaluation of the data is done over two different modules of the service:

Profiling and Assessment modules. The Profiling Module provides results related to the data in general, such as value ranges, uniqueness degree of each attribute, and number of represented objects. The second module is the one being responsible for the evaluation of the data set over set of defined quality measures [1].

After setting up the analysis and obtaining the result, the user should recognize if the quality of the data is good enough, or the testing didn't cover all necessary aspects and it must be repeated. As the topic of data quality in Big Data is a new one we recognized that it would be hard for newcomers in this field to grasp all the details.

To help the user this thesis proposes a method that will act as a support component of the quality service in result exploration. The idea is to provide a graphical interface for visualization of the results in such a manner that every user would get more familiar and understand the most important aspects of the analysis. Besides that, it will offer detailed exploration and custom filtering in order to save the user from many repeated revocations of the quality service since it requires allocation of processing power.

## 1.3 STRUCTURE OF THE THESIS

The thesis is organized in six chapters. Chapter 2 introduces the fundamentals in the literature regarding Big Data and Data Quality in Big Data. An accent is also given on the selection of quality measures regarding this topic. We also go through technologies that allow the growth of quality assessment in Big Data. Last, but not least, we go through challenges and techniques regarding Big Data Visualization and analytics. Chapter 3 makes a brief description on the Data Quality-as-a-Service offered by the EU-Bra BIGSEA project. As stated, the service will be of huge importance of this thesis because the results exploration process is being implemented on top of its evaluations. In chapter 4 we go through the methodological aspect of the solution proposed. First, the main goal of the model is described, and it follows the working logic and complexity behind. Chapter 5 provides aspects of the model implementation and technologies used. Also, the user interface and each of its features is explored in detail. Chapter 6 gives an overview of the work presented in this thesis together with a future development possibilities and usage of the exploration tool.

# Chapter 2

# STATE OF THE ART

## 2.1 BIG DATA FUNDAMENTALS

As much as we want to refer to the Big Data as a new term, we must agree that it has been around for a while. Thirty years ago, one gigabyte was taken as a Big Data and it required some special processing, and yet now we manage to download the same amount on our phones in a minute. The same thing occurs right now, but measures are much bigger, and of course the technology went rising and bringing higher processing power.

We can think of Big Data as a field that covers analysis, processing, and storage of large collections of data that most frequently arrive from different sources. The issues that regard problems with Big Data are arising just in the moment when the standard analysis and processing techniques are not enough. This is happening not just because of the amount of data that is being collected but also because the data is arriving from unrelated datasets in large amounts of unstructured sets. The whole process gets worsened since it is not possible to have errorless data while the records are of enormous size [2].

The analysis done over sets of Big Data can be explained as multiple distinct processes in the areas of mathematics, statistics, and computer science. Anyway, we should always consider the level of expertise of the one performing the analysis. Also, it should be clear of what it counts as a Big Data problem, since the times are changing, and the new advanced software and hardware technologies are stepping up.

Irrelevant how the data is collected in the data warehouse, if it is errorless, there are a lot of benefits such as [3]:

- Decision-making
- Optimizations
- Predictions

- Customer satisfaction
- New discoveries

But, as said before, to be able to adopt all the analytics done over Big Data we need to be prepared for numerous issues. These can lead to wrong and costly decisions. So, in order to invest and adopt Big Data analytics we need to know more about our data source, the way we obtain the data, how we save it, and which are possible factors that can give wrongful information. That's why we first must weigh the issues against the benefits in order to be sure that we can proceed with informed decisions [3].

## 2.1.1 Big Data Characteristics

A dataset can be considered as Big Data, if it possesses various characteristics that provide the difference between a "Big Data" and normal data, which are also known as the Five Vs and include volume, velocity, variety, veracity and value.

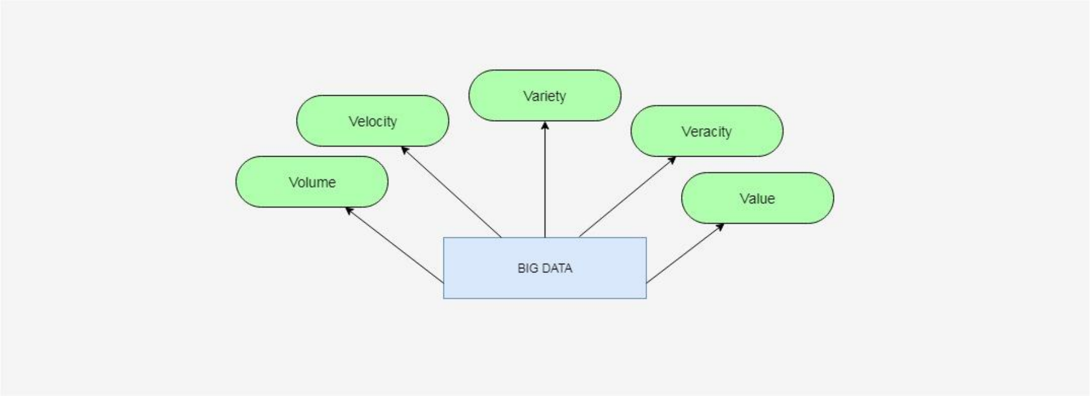


*Figure 1: The Five Vs of Big Data*

These characteristics were initially recognized by Douglas Laney (analytics scientist) in an article in which he described the effects on enterprise data warehouses. Moreover, these characteristics prove to be essential for the solution design and for the architecture of the analytic environment. Out of these five, veracity is often overlooked since it deals with uncertain and imprecise data. The main objective is to reach results that would respect the aspects of quality data within a deadline, therefore, achieve optimal value for the enterprise [4].

#### *2.1.1.1 Volume*

The volume of data is constantly on the rise as its being processed by Big Data solutions. Big data volumes request extremely high amount of data storage and processing, as well as other data properties on a projectual level.

Data sources that can lead to high data volumes are: online transactions, scientific research (Large Hadron Collider and Atacama Large Millimeter/Submillimeter Array telescope), GPS sensors, RFIDs and nowadays the overused social media networks such as Facebook and Twitter [4].

#### *2.1.1.2 Velocity*

Represents usually the speed at which data transfer and accumulation can arrive, and this is essential because it can generate huge amount of data on a short period of time. On an enterprise level, this represents the time of data translation as soon as it enters it's perimeter. In order to achieve this, an enterprise must comply with high and efficient requirements. This may not always be the case though, as the velocity depends on the type of data source. Data velocity is not put into perspective about MRI scans as these are not often generated from high-traffic webservers. On the other side, data volume can be easily computed in a given minute for the following aspects: 350,000 tweets, 300 hours of video footage uploaded on YouTube, and 330GBs of sensor data from a jet engine [4].

#### *2.1.1.3 Variety*

Refers to different formats and types of data that are supported by Big Data solutions. These characteristics brings into focus data integration, transformation, processing and storage. Examples of data variety can be financial transactions, semi-structured data in the form of emails and unstructured data such as images [4].

### *2.1.1.4 Value*

It is one of the most important aspects of Big Data. And it's intuitively related with veracity in terms of data quality, which holds the key value of an enterprise. Volume is dependent on processing time, because the data analytics can have a shelf-life; for example, a 20-minute delayed stock quote has almost no value in comparison to a quote that is 20 milliseconds old. Overall, value and time are inversely relating i.e. the longer it takes for data to be transformed into useful information, the less value it has for an enterprise [4].

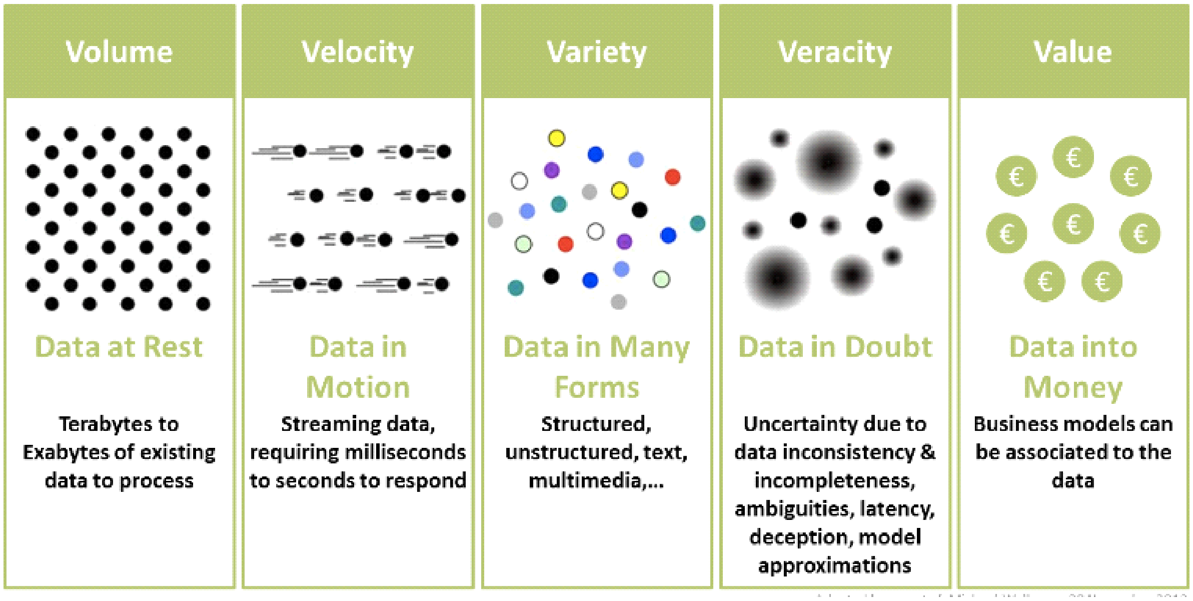

Volume
Velocity
Variety
Veracity
Value
€
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Data into Money
Terabytes to Exabytes of existing data to process
Streaming data, requiring milliseconds to seconds to respond
Structured, unstructured, text, multimedia,...
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
Business models can be associated to the data

*Figure 2: Illustration of the Vs of big data (Source: [5] Information Catalyst for Enterprise (ICE))*

## 2.2 DATA QUALITY IN BIG DATA

When we say data quality we usually think of data that is errorless, but that is only one part of the equation. In general, the data quality depends on many factors that collectively give evaluation.
In order to be able to understand how each aspect evaluates we must distinguish data quality issues in different categories. Figure 3 shows attributes that describe the data in different aspects in terms of quality.

The first five attributes (Accuracy, Integrity, Consistency, Completeness, and Validity) are intrinsic to the content of the data. That covers data entry errors, misapplied business rules, duplicate records, and missing or incorrect data values. The remaining two attributes (Timeliness and Accessibility) are giving the user understanding of the current data in a timely manner. Of course, this data will be useful only to the ones that are familiar with the business in question [6].

The combination of both categories of attributes can increase the data quality value we can extract out of the data set. The following list describes characteristics of high quality data [6]:

- Complete: All relevant data such as accounts, addresses and relationships for a given customer is linked.
- Accurate: No misspellings, typos, and random abbreviations.
- Available: Required data are accessible on demand; users do not need to search manually for the information.
- Timely: Up-to-date information available to support decisions.

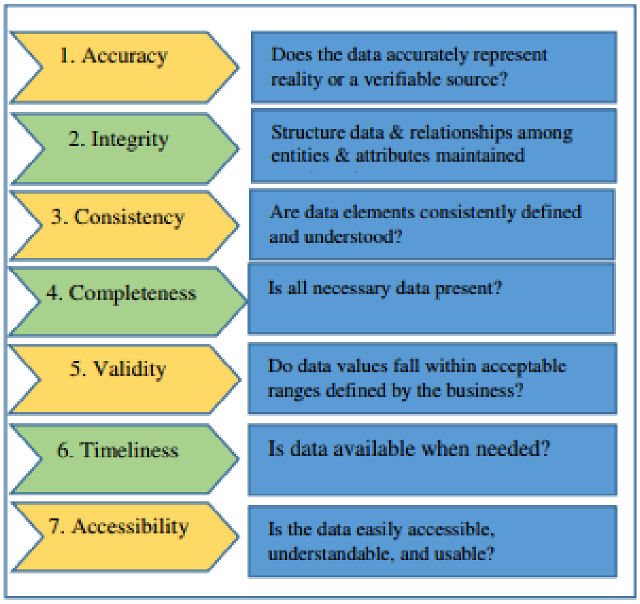

1. Accuracy
Does the data accurately represent reality or a verifiable source?
2. Integrity
Structure data & relationships among entities & attributes maintained
3. Consistency
Are data elements consistently defined and understood?
4. Completeness
Is all necessary data present?
5. Validity
Do data values fall within acceptable ranges defined by the business?
6. Timeliness
Is data available when needed?
7. Accessibility
Is the data easily accessible, understandable, and usable?

*Figure 3: Data Quality Attributes (Source: [6] Data Quality in Big Data: A Review, Int. J. Advance Soft Compu., Vol. 7, No. 3, November 2015)*

There are many ways which can lead to data quality problems such as valid, but not correct data, unexpected changes in the design of the system, lack of integrity checks, etc. Due to the poor quality of the data a lot of companies have suffered losses, problems, or costs for assigning new resources. In the figure bellow we can see which problems are most frequently occurring due to poor quality in the data according to a survey done by The Data Warehousing Institute (TDWI) [6].

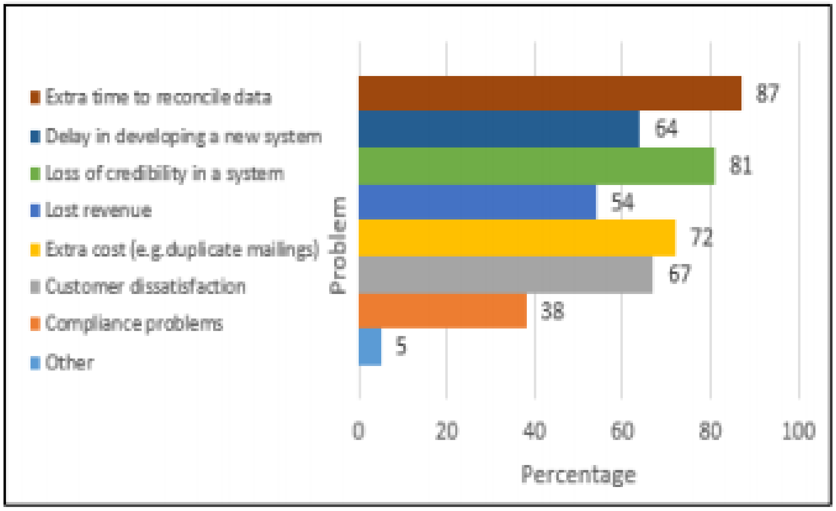


*Figure 4: Problems due data quality (Source: [6] Data Quality in Big Data: A Review, Int. J. Advance Soft Compu., Vol. 7, No. 3, November 2015)*

### 2.2.1 Data Quality and Measurement

When we speak of measurement we usually think of taking notes of size, amount, or degree of something by using an instrument. But, in the case of Data Quality we think of a process that lets us compare values to a well-known standard, or in other words, to assess the information and draw a conclusion.  This process in the world of experts in the field is called data quality assessment. The main goal of data quality assessment is to identify data errors and to measure impact of data-driven processes. The process can be done on general level concerning a data set but also it can go into detail and regard specific aspects [7].

The main idea behind quality assessment is to understand how the data weights against a company's expectations. This implies the need also to understand how effectively the data represents the objects and its properties internally.

Professionals in the field of quality are referring to quality dimension as measures used to identify aspects of data that can be evaluated and through which data's quality can be described and quantified. So, it means that the quality of a data source can be analyzed from different perspectives, and on a different level of detail, also referred as granularity level. In other words, it will depend on the set of quality dimensions that we want to use. So, each quality dimension will give indication how good or bad that data object (subset of records selected from the whole data source) is.

There are 179 quality dimensions, but in the interest of the thesis we will focus only on 9 of them, since they are used in the Quality Service of the EU-BRA project. All Data Quality dimensions can be grouped in four categories [8]:

- Intrinsic: the data has a quality on its own. For example, Accuracy is a quality dimension that is intrinsic to data.
- Contextual: highlights the requirement that must be considered within the data source. It must be relevant, timely, complete, and appropriate in terms of amount.
- Representational: focuses on aspects related to the quality of data representation
- Accessibility: considers the accessibility of the data and its level of security

## 2.2.2 Big Data Architectures

As the world becomes more advanced on information aspects, a major challenge has become how to deal with the amount of data. Fortunately, in parallel the technology of data processing and management has advanced in such a level that can cope with the super-massive scale of information arriving and changing fast. These technologies also demand a new DBAs and developers to manage new type of systems.

It follows an overview of technologies that permit the growth of quality assessment in Big Data.

### 2.2.2.1 NoSql

Relation databases together with SQL programming language are still the only choice of database technologies for organizations. All the possibilities that are covered by these two now can be replaced and maybe a bunch of new ones can be offered by the evolution of various NoSQL software platforms. Databases support dynamic schema design, scalability and customization. This makes them a perfect fit for Web applications, content management systems and other uses that involve large amounts of data and frequent updates. Said so, NoSQL technologies cover all the issues that have been introduced in Big data [9].

NoSQL databases are centered around the concept of distributed databases where unlike relational databases that are highly structured, here the data can be unstructured and stored across multiple processing nodes, and also often across multiple servers. This allows the databases to be horizontally scalable. This means that if there is more data to be kept in the data base than just add more hardware to keep up and there will be no slowdown in performance [9].

The NoSQL distributed database infrastructure has been the solution to handling some of the biggest data warehouses on the planet – i.e. Google, Amazon, and the CIA.A comparison between the scalability of NoSql database vs Traditional Relational Database can be seen in Figure 5.

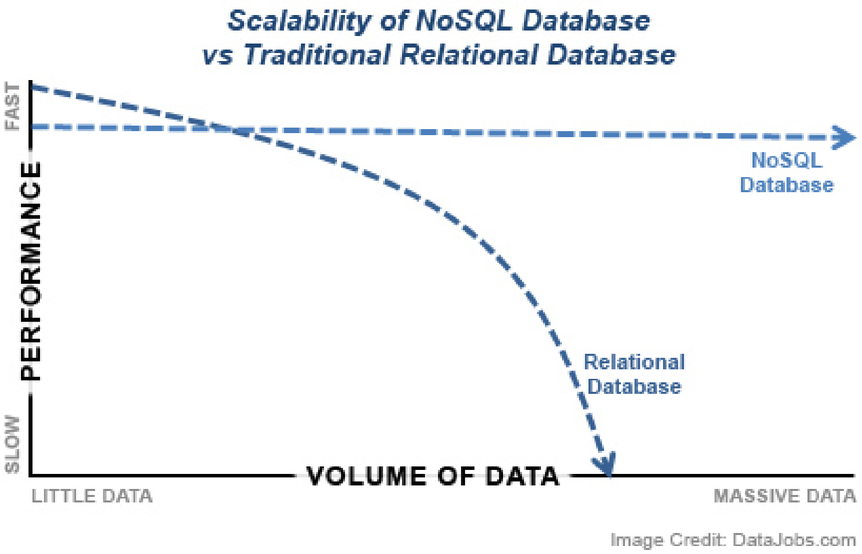


*Figure 5: Scalability of NoSQL vs Traditional RDB (Source: [9] DataJobs.com)*

### 2.2.2.2 Apache Hadoop

The Apache Hadoop library is a framework that can permit processing of large data sets across clusters of multiple nodes using simple programming logics. The idea is to change from the single servers to thousands of nodes so that the framework can detect and handle failures at application layer. This can allow delivering a highly available service on top of a cluster of computers, where each can be prone of failures [10].

The project includes these modules [10]:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

A crucial part of the Hadoop ecosystem is MapReduce, a computational model that basically takes high data processes and spreads the computation across a big number of servers (generally referred to as a Hadoop cluster). It has been a game-changer in supporting the enormous processing needs of big data; a large data procedure which might take 20 hours of processing time on a centralized relational database system, may only take 3 minutes when distributed across a large Hadoop cluster of commodity servers, all processing in parallel [11].

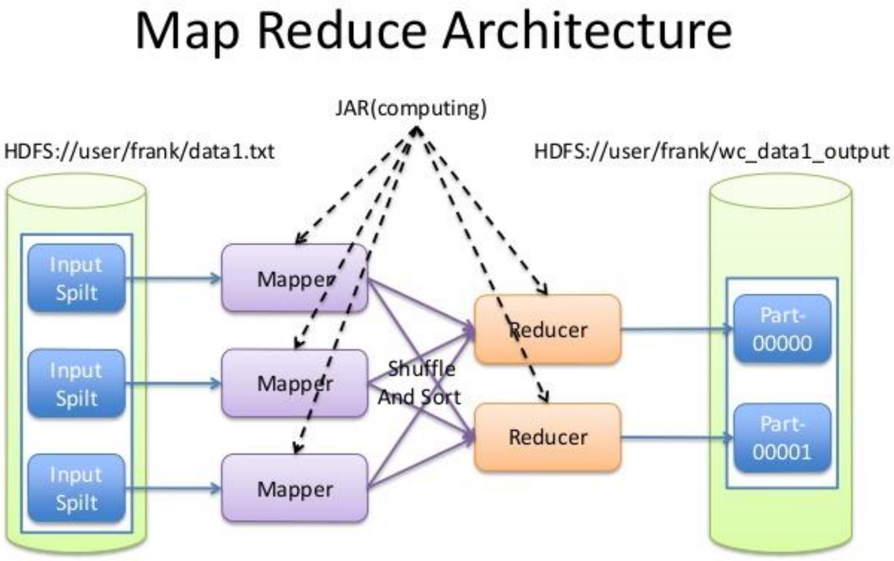


*Figure 6: Map Reduce Architecture (Source: [11] Introduction to the MapReduce Life Cycle)*

### 2.2.2.3 Apache Spark

Apache Spark is a fast, in-memory data processing engine with development APIs to allow workloads that require iterative and fast access to datasets. This sounds like a major coverage of the issues addressed by Big Data. With Spark running on Apache Hadoop YARN, developers can now exploit Spark's power. In general, The Hadoop YARN-based architecture is providing the foundation that enables Spark to share a common cluster and datasets while the level of service it is about normal processing on small data sets [12].

Apache Spark consists of Spark Core and a set of libraries (see Figure 7). The core is the distributed execution engine and the Java, Scala, and Python APIs offer a platform for distributed ETL application development.

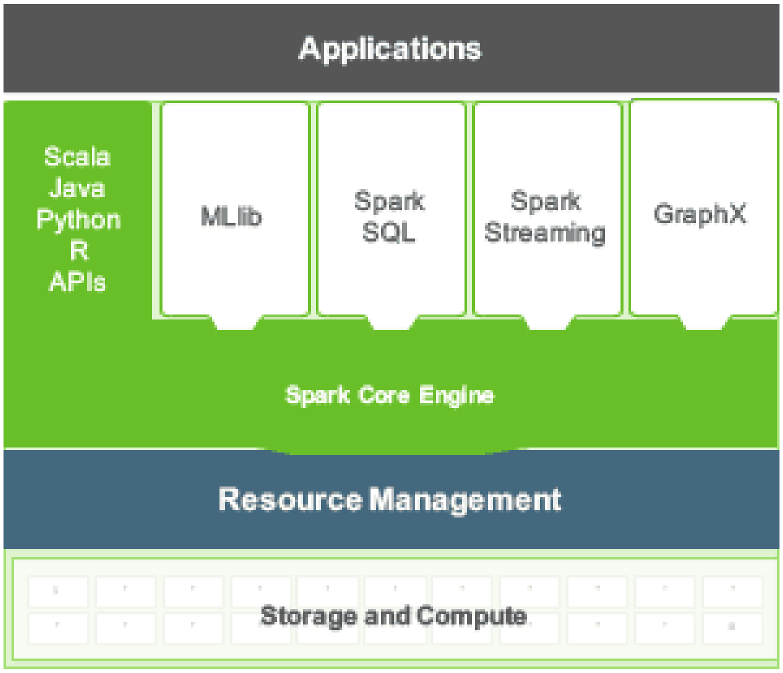


*Figure 7: Apache Spark Libraries (Source: [12] http://hortonworks.com/apache)*

Additional libraries, built on top of the core allow diverse workloads for streaming, SQL, and machine learning.

Spark is designed for data science and its abstraction makes data science easier. Data scientists commonly use machine learning – a set of techniques and algorithms that can learn from data. These algorithms are often iterative, and that is where Spark's ability to cache the dataset in memory comes handy [13].
Here are the biggest benefits of Apache and a comparison with Hadoop as defined by Databricks Inc. [1]:

- *Speed: Engineered from the bottom-up for performance, Spark can be 100x faster than Hadoop for large scale data processing by exploiting in memory computing and other optimizations. Spark is also fast when data is stored on disk, and currently holds the world record for large-scale on-disk sorting.*

[1] © Databricks 2018, https://databricks.com/spark/

- *Ease of Use: Spark has easy-to-use APIs for operating on large datasets. This includes a collection of over 100 operators for transforming data and familiar data frame APIs for manipulating semi-structured data.*

- *A Unified Engine: Spark comes packaged with higher-level libraries, including support for SQL queries, streaming data, machine learning and graph processing. These standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.*

## 2.3 BIG DATA VISUALIZATION AND ANALYTICS

Within few years the topic of Big Data affected a lot of industries not only it the IT business. At the beginning the biggest issue was how to store the big amounts of data. Later, organization as companies, governments, and institutions recognized how it can be useful to explore the data and learn more for different aspects regarding their customers, needs, or even new scientific discoveries. Then, we got familiar for the first time with other problems that are rising with the Big Data. It became clear that it is not only about storing, but first capturing the right data through sensors, sharing, and most importantly analyzing it.

As complex it is the processing and analyzing huge sets of data, the output results should be precise and clear so the user can understand how the data can be used. But in fact, this is not always the case. Usually, the output results of an analysis are giving raw data that for the human eye is hard to read. This is a huge issue concerning the fact that companies started investing huge amounts of money but at the end they didn’t know how to use the collected data. So, in order to help the user, Big Data tools started using an old technique in the needs for results exploration - Visualization. With Data Visualization, companies started rendering results after their software finished the execution of processing. In this way the user could catch complex patterns that are hiding behind [14].

There is a huge importance in visualization of output results since the human brain cannot process huge amount of information given in a single report. It is always better to see the data organized in tables or graphs where we can experiment and form slight adjustments, in fact that’s how we will learn more about the data itself.

### 2.3.1 Challenges in visualization

The main goal of data visualization should be clear and concise patterns that describe a result set. With the evolution of the data amount it became impossible for traditional visualization tools to be used. In order to cope with the new era, the main challenge was to render as fast as possible, since no user would want to wait for visualization the same amount of time as for the data processing. Another huge issue was the fact that the data arrives in structured and unstructured forms and it became challenging to cope with two different modules of visualization in the same user interface.

The fact that the data evaluation is done over different aspects of the data object and with different measures of quality says that the visualization must show all this differences in levels. This became also an issue because exploring the data in high detail means a lot of similar information in one place, causing confusion to the user. On the other side, if the level of detail is too low the user is losing important aspects and patterns for decision making [14].

In order to solve the issues and propose faster visualization, companies started proposing a merged architecture. The idea is to process and render the data in the same time. Of course, this will increase the level of parallelism and allow the user exploration over aspects of the data object while the rest is still being evaluated.

### 2.3.2 Visualization techniques

Visualization approaches should always provide a general overview of the results set. This information can be exported from the software that does the processing, or if needed the visualization tool should provide it to the user with a self-processing of the result set. This is of a huge importance since it is suggested to explore results in iterative manner, from low to high detail. Of course, the visualization service should provide groupings, searching, and custom filtering on the result set as long the results are saved by a recognized standard.

There are several methods for visualization. Follows a list of the most used ones [15]:

- Treemap: Used for visualization of hierarchical data.
- Circle Packing: Different way of representing hierarchy.
- Sunburst: Polar coordinate system. Used to show similarities between data.
- Parallel Coordinates: Allows exploration on multiple data factors
- Streamgraph: Used in representing similarities between data.
- Circular Network Diagram: Line and color used as indicators to show similarities between data

## 2.4 CONCLUSION

As Big data continues in growth, for sure all of the mentioned architectures in the likes of NoSQL databases and Hadoop software (in correlation with Spark) – will be central to allowing companies reach full potential with data. Additionally, this rapid advancement of data technology has raised a demand for a new working positions. Of course, the cost of the technology and the talent may not be cheap, but for all of the value that big data is capable of bringing to table, companies are finding that it is a very worthy investment.

Having in mind all the data quality dimension mentioned before, and all the technologies that can help us perform analysis over Big Data there is still a gap in the literature because there is no appropriate method to assess the Quality of Big Data.

Taking part in the EU-BRA project, Politecnico Di Milano proposed a method that will address this issue and regard the Quality of Big Data. The Quality Dimensions are being reinterpreted and redefined based on the type of considered data and new metrics consider the increased complexity generated by the volume, variety and velocity challenges. The proposed method was implemented as Data Quality Service and was tested on data sources offered on the cloud platform of the EU-BRA project. Also, an application that offers user interface was created based on the API's that the service is providing. All with one purpose, to set up a quality analysis on Big Data which will allow further understandings when it comes to quality analysis as the data grows bigger. The implementation of the proposed model executes the computation of the quality dimensions using the Spark framework that allows to split the computation of the code over several nodes in order to speed up computation time of the quality analysis.

Once the quality service gives the results of the evaluation, in place comes the exploration part. At the current implementation of the Quality Service by BIGSEA the results are exported as a set of JSON files. Here in place comes the goal of this thesis, providing the visualization part in order to help the user in decision making. As we saw in section 2.3, the data visualization in Big Data should be carried with a huge importance. In order to take care of the visualization part this thesis will concentrate on the results exploration with a combination of techniques and user-friendly interface. The idea is to allow iterative exploration starting from lower to higher level in order to prepare the user with general quality results and yet not causing confusion right away with too many details.

In the following chapters we will describe the proposed method for quality analysis and the Quality service in details. This will be of a huge importance to the thesis since all the results are based on its output results. Chapters 4 and 5 will provide detailed information about the ideas and implementations leading to the data exploration tool as a support component of the QS.

# Chapter3

# EU-BRA DATA QUALITY SERVICE

This chapter will go in details through the Data Quality-as-a-Service offered by the EU-Bra BIGSEA project. As stated, the service will be of huge importance of this thesis because the results exploration process is being implemented on top of its evaluations.  A specific accent will be given on the Interface that the Quality Service provides in order to communicate with users/application. This work was proposed by Giacomo Maccagni in the thesis “A method for Raising Quality Awareness in Big Data Analysis” [16].

## 3.1 EUBRA-BIGSEA PROJECT

*“EUBra-BIGSEA project aims to develop a cloud platform for Big Data Management and exploitation. It is funded in the third coordinated call Europe – Brazil, focused on the development of advanced QoS services for Big Data applications, demonstrated in the scope of the Massive Connected Societies.  The idea of the project is to develop architecture that will predict resource consumption of Big Data Analytics applications in order to pre-allocate and dynamically adjust virtualized infrastructures. It leverages mixed horizontal and vertical elasticity on hybrid container and VM infrastructures to support Data Analytic framework powered by OPHIDIA, COMPSs, and Spark. The integrated technologies should support:*

- *Fast data analysis over continuous streams*
- *Data mining and machine learning*
- *OLAP-based Big Data analytics*

*The project is API oriented, which means it is oriented to service and application developers. In order to test the services and in the same time to offer the services the project is being tested in a real user scenario in Curitiba, Brazil.*

*The Data Quality Service is in charge to provide a quality analysis over a data sources that provide information about GPS location of local buses and ticket usage in the same buses."*[2]

## 3.2 Model Architecture of the Data Quality Service

The data quality service provides additional information about the data sources saved on the platform. We can distinguish the information in two groups [16]:

- Data Quality Profiling results related to data that is in the data source, such as value ranges, and uniqueness degree of each attribute and number of represented objects.
- Data Quality Assessment results based on the performed quality analysis over the data source, obtained by evaluating specific data quality dimensions.

So, the Data Quality Service is composed of two modules: Profiling module and Assessment module. The first one executes periodically over the whole data source in order to produce general information. This information is being used as input in the second module. For example, when setting the Quality Dimensions for the analysis, the Assessment module uses the metadata of the first module, so it can know which data types can be used for each quality dimension. In case of adding a new attribute in the data source it is crucial to re-run the Profiling module, so all the information is up to date.

The Assessment module is run on demand of users or applications that specify the details of the quality analysis which is being run on the service. Also, the results of the assessment module are kept as metadata for a further analysis [16].

Follows a figure illustrating the global architecture of the model.

[2] EU-Bra BIGSEA Project Vision. Retrieved at: http://www.eubra-bigsea.eu

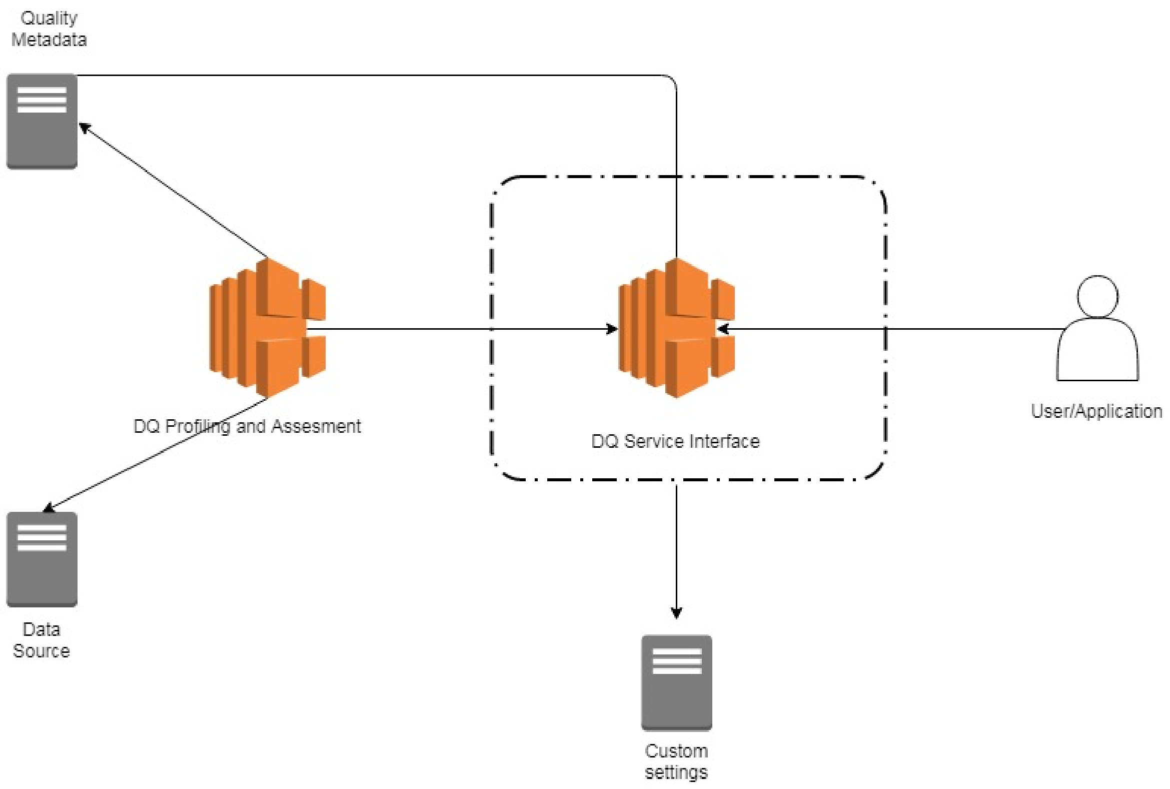


*Figure 8: Architecture of the Quality Service*

The Data Quality Service is in charge to provide a descriptive view of the quality of the sources with the aim to support the analytics applications in understanding which are the relevant and useful data to consider in more advanced analyses. Once the user specifies the sources of interest the Data Quality Service interface visualizes in the Quality Metadata Repository all the general information about the data, which is the number of records contained in the data source, data ranges of values that each attribute can have. After the data source registration phase the user/application sets up a quality analysis over it. The Assessment module also has access to the Quality Metadata Repository, so the service interface will be familiar with any restrictions on the data source (for example, allowed data types for each dimension).

Once the user sets up the quality analysis, all the settings are being saved in the Custom Settings Repository. The purpose is to keep the quality analysis settings for each quality analysis for a possible re computation and modification on the data object or varying values on the dimension set.

The quality request set by the user/application is submitted to the Data Quality Assessment module where all the computations specified in the request take place. Once the results are ready they are saved in the Quality Metadata Repository. The quality information of the selected object can let the user be aware of the quality of the input data, so it can be useful to consider in more advanced software's for statistics and predictions.

To make a brief recap, the Quality Metadata Repository stores metadata both from the results of the Profiling and Assessment Module, and metadata used to define parameters that will be selected using the available interface of the Service.

Parameters are being inserted by the user/app when interacting with the service interface. As far we thought of the user as a professional in the field of data analytics, but the service is also providing an applicable interface for semi-professional users, so it can make the data analysis process more user friendly. For example, the user can select a list of data mining applications in order to set up a quality analysis based on the quality dimensions that the selected data mining application requires. Also, the service is providing an interface where the user can choose from a group of quality dimensions instead of choosing all dimension one by one [16].

## 3.3 DATA QUALITY SERVICE INTERFACE

The user interacts with the Data Quality Service Interface using a set of Application Programming Interfaces(API's). An API allows a user to manipulate resource remotely using communication protocols. The service provides many API's. Each of them is defining a portion of a quality analysis defined by a user/application. The whole set of communication between the Service and the user is saved into a configuration file. The configuration file keeps all the details of the configuration defined by the user/application and it will be saved in the Custom Settings Repository, from where it is passed to the Data Quality Assessment module [16]. From the figure below, we can also see how an exploration tool would be placed in the architecture, right after the results are returned to the user.

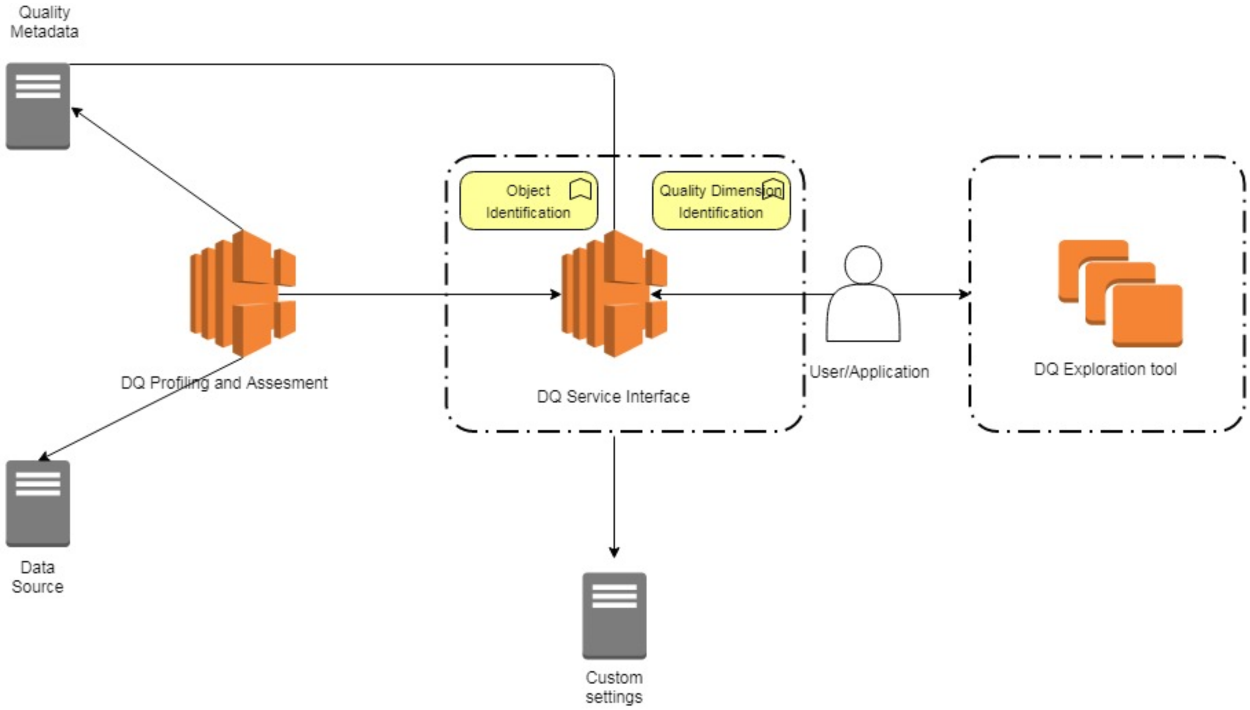


*Figure 9: Service - User/app communication with User/Application and a DQ exploration tool as a support component*

When making a quality analysis over a data source the user/app should enter information about the part of the source that is going to be exploited. This is the process of selecting a Data Object of interest and it forms one of the two main nodes of the service: Data Object Identification Module. Depending on the level of granularity the user is interested to explore, there are three submodules of parameters that are composing this module [16]:

- Data Source Selection Module: selection of a data source (for example it can be the one that keeps track of the BusGPS, or the other with User-ticket information)

- Data Attribute Selection Module: selecting attributes of interest from the data source (for example we can select the code number of the bus only)

- Data Value Selection Module: selecting data values for each of the selected attributes. We can either select a range of values, or specify which values exactly are in our interest.

The module where the service offers the user an interface to select the quality dimensions that will be computed over the previously defined object is Quality Dimension Identification Module. It is composed of four other submodules that will help user either professionals or not to set up a quality analysis. The submodules are the following one [16]:

- Automatic Dimension Identification Interface: it is mainly used by the unexperienced data analysts that want to perform a quality testing on the data source, but they don't have deeper knowledge in the quality dimensions.

- Focus Category Dimension Identification Interface: Used to set up a quality analysis by selecting a group of dimensions. So, the user has to choose one of the predefined groups which the service determines by reading the profiling data from the Quality Metadata Repository.

- Data Mining Dimension Identification Interface: It proposes to the user a set of data mining applications. If the user knows which data mining application is going to use afterwards than it is a good idea to choose this option.

- Custom Dimension Identification Interface: It is used in the case when the user manually enters each quality dimension that will be computed over the data object. This is mostly used by data analysis professionals.

## 3.3.1 Data Object Identification

The process of selecting the data object has a big importance for the data analysts because it helps them understand interesting information about the whole data source, and usually that's why the data quality analysis is being re computed many times. If there is no selected portion of interest the whole data object is selected automatically, but this can be risky since the results of the analysis can return an unexpected value which can cause confusion.

When interacting with the Data Quality Interface, the user/application can define the data object in three granularities [16]:

- Source schema selection - The user is obligated to select the source of interest, because without any source the quality analysis cannot continue. Or, alternatively it is also possible to select more than one data source, in a case the user wants to compare the results over the same quality analysis.

- Attribute schema selection - For each data source there is a set of attributes that can be selected by the user. Also, the user is allowed to select multiple attributes of the data source. This will help to gather more knowledge for the data source. This is a crucial process when building the configuration file because the service can check the metadata of the Profiling module and get to know which dimensions can be computed over the selected attributes.

- Value schema selection - The set of attributes selected in the Attribute Schema Selection allows the user to select values for each of them. Without any attribute selected it is not possible to select values, and the quality analysis will be done only on data source level.
  There are two possible way to select values. One way is to select a set of values which we are familiar with and we want to observe all quality dimension only on them. The second way is to select a range of values.

### 3.3.2 Quality Dimension Identification

After selecting the data object of interest in the first module, the Data Quality Service provides interface for selecting the Quality Dimension Identification node. This is the most important module since it shapes the quality analysis the user wants to perform over the data object. Each of the selected submodules in a combination with the previously selected Data Object, gives a different quality analysis results [16].

The identification of quality dimension and theirs level of granularity can differ from user to user, because some of them are more skilled than the others. So, the interface provides step by step dimension identification. The four submodules of the interface are grouped in two groups (se Figure 10).

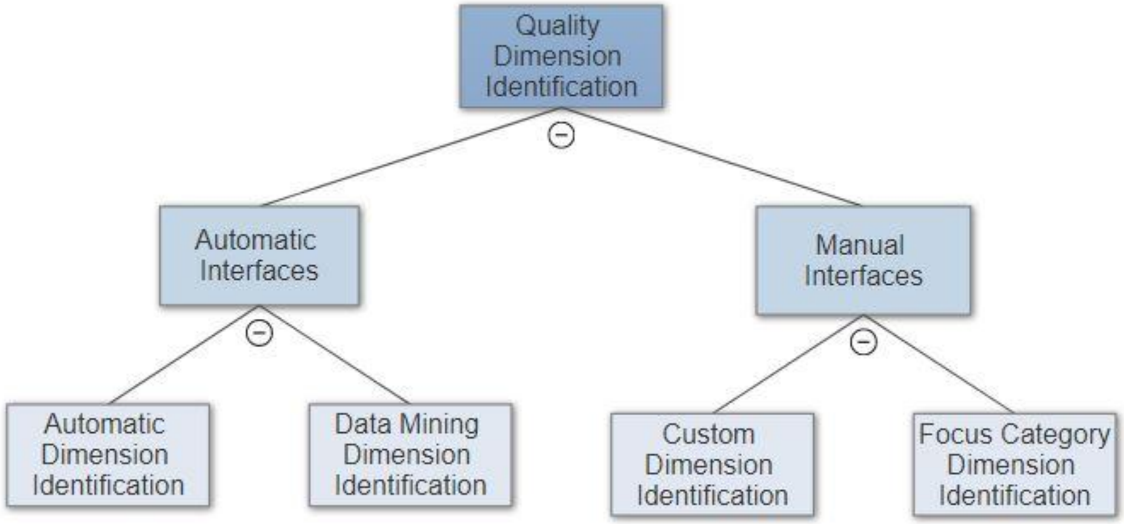

Quality Dimension Identification
Automatic Interfaces
Manual Interfaces
Automatic Dimension Identification
Data Mining Dimension Identification
Custom Dimension Identification
Focus Category Dimension Identification

*Figure 10: Quality Dimension Identification (Source: [16] Giacomo Maccagni, "A method for Raising Quality Awareness in Big Data Analysis", 2016)*

### *3.3.2.1 Automatic Interfaces*

Assuming the user that performs the analysis is not skilled and it is only interested in some generic quality indicators about the data object that was firstly selected. With the selection of the automatic interface the user is helped for the setup of the quality dimensions [16].

- Automatic Dimension Identification: Knowing the attribute that the user previously selected, the data quality service will infer all the available dimensions that can be set to make an analysis. There is need of some metadata parameters of level 2, but in this case, they are already stated by the administration. The level of granularity in which the user can explore the results is the lowest one. It means that the quality dimensions will refer only to the data source.
- Data Mining Dimension Identification: The user chooses a data mining application that will be run over the data object. There are three levels of granularity on which the selected data object can be analyzed, from which the user has to choose only one level of detail on which the quality analysis will be computed.
  If the level of detail is "low" it means that each quality dimension will be computed over the data object without giving any attribute quality indicators.
  If the level of detail is "medium" it means that the service will produce a quality evaluation over each attribute of the selected data object.
  Leve of detail "high" means that the service will produce quality evaluation for each value of the selected attributes from the data object.

#### *3.3.2.2 Manual Interfaces*

Assuming the user that performs the quality is an experienced one. The interface offers two possibilities [16]:

- Focus Category Dimension Identification: it is an interface that offers to simplify the setting up of the quality dimensions. Basically, all the dimension are grouped in categories where they have something in common. Each group contains dimensions that are exploiting the same aspect of a data object. The metadata of level 2 Is automatically set by the model.
- Custom Dimension Identification: In this interface the user specifies all the quality dimension on its own, supposing that is an expert and knows which aspect of the data source will be compute by each dimension. In case that some of the quality dimensions require an additional parameter, the user has to provide it on its own (for example the user has to define volatility when choosing timeliness as a quality dimension).

In both manual interfaces the user is required to specify the level of detail on which each quality dimension will process the analysis (global, attribute, value).

## 3.4 DATA QUALITY SERVICE IMPLEMENTATION

In this section we can see how the architecture of the proposed method for a Data Quality Service was implemented, which technologies were used, and API communication with user/app.

### 3.4.1 Selection of Dimensions

The quality evaluation of a data object can be performed over many quality dimensions.
The Data Quality Service implements 9 dimensions [16]:

### *3.4.1.1 Accuracy*

It is defined as a closeness between a data value V and a data value V0. Basically, it is the ratio between the two values, letting us know how close is the data value of exploration to the one that is considered as a real-life phenomenon. Returns two indicators:

- Static Accuracy: if a value is in the admirable interval or not
- Dynamic Accuracy: Represents how the value is close with respect to the mean

It is represented as a linear value ranging from 0 to 1.
*Granularities*: global, attribute, value
*Allowed data type*: float

### *3.4.1.2 Completeness*

It is used to estimate how many items in the data collection are included in the expected item set. This dimension can be viewed from many perspectives and levels of detail which leads to different dimension metrics:

- Completeness Missing: measuring the number of missing/or present values in a data set
  It is represented as a linear value ranging from 0 to 1.
  *Granularities*: global, attribute, value
  *Allowed data types*: float, string, datetime

- Frequency Completeness: derived as a ratio between effective frequency of a value Vi measured over the data object represented by parameter Fi, and expected frequency of a value Vi over the data object represented by Hi
  It is represented as a linear value ranging from 0 to 1
  *Granularities*: global, attribute, value
  *Allowed data types*: datetime

- Population Completeness: defining how many instances of the data source are included in the data object. If all instances are present, the Completeness population value is 1. Otherwise, the value can be between 0 and 1.
  *Granularity*: attribute
  *Allowed data types*: float, string

#### *3.4.1.3 Consistency*

It is a metric that helps the user to understand if there is a logical association between records of the selected data object. It is calculated with the following ratio:

$$Consistency = 1 - \ ConsistencyViolations/Total\ Consisntecy\ Rules$$

It is represented as a linear value ranging from 0 to 1
*Granularities*: attribute, value
*Allowed data types*: all

#### *3.4.1.4 Precision*

It is defined as the degree to which repeated measurements show the same or similar results, and it means that all the values should be equally distributed and not too distant between each other. Returns three output data related to the analyzed value:

- Precision of the value: how the aggregated values are close to each other
- Association mean of the values
- Standard deviation: In the literature the standard deviation is often used to assess the dispersion of the data, leaving the interpretation of the obtained value to the application or user that requested the analysis, but in order to access the Precision dimension of an entire source, a value between 0 and 1 is returned.

*Granularities*: global, attribute, value
*Allowed data types*: float

#### *3.4.1.5 Timeliness*

It is defined as "how current the data in the data source is". Of course, it can be derived if only we have information when the data was created. It is calculated with the following ratio:

$$Timeliness = \max(0.1 - \ Currency/Volatility)$$

Where, the currency is defined as *deliveryTime* – *inputTime* in which the first refers to when the data is delivered to the user, and the second refers to when the data is received by the system. Volatility gives information for how long the data is usable.

It is represented as a linear value ranging from 0 to 1
*Granularities*: global, attribute, value
*Allowed data types*: timestamp

### *3.4.1.6 Distinctness*

Used in defining how much records in the data is unique. It can be calculated over the whole data source, attributes, or for each value. It is calculated by removing duplicate elements.

It is represented as a linear value ranging from 0 to 1
*Granularities*: global, attribute
*Allowed data types*: float, string

### *3.4.1.7 Volume*

It is defined as "number of instances in a selected data object". So, basically it represents count of each record that forms the data object from the data source. In example if our data source keeps names of users, by performing quality analysis with Volume as a measurement dimension we can get the count for each name in the data source.

It is represented as a linear value ranging from 0 to 1
*Granularities*: global, value
*Allowed data types*: float, string

### 3.4.2 Data Quality Service Interface Implementation

The Data Quality Service Interface is implemented as a collection of API's. The service is developed using Jersey RESTful Web Services. The framework itself is open source, production quality, framework for developing RESTful Web Services in Java.

There are two types of resources offered by the service, one to define the data object and another to set up the quality analysis. Each of the resources is defined with an URI. Using HTTP, the request can be GET, POST, PUT, and DELETE. The responses to the API can be in XML, HTML, or JSON format [15].

## 3.5 Conclusion

The proposed Data Quality as a Service Module took place in the EUBra-BIGSEA project. The Profiling and the Assessment module are scalable because are developed using a Spark code. They are being executed over a number of nodes depending on the volume of the object and with there is a big save done concerning processing and time of execution.

It found good acceptance because not only it offers a data quality over Big Data, but proposes a redefinition of quality dimension, and also helps users that are not professionals. It is an adaptive model because for each change of the data object the quality analysis will change.

To capture the characteristic of a data object we must explore the results of the quality analysis. As the final step of the quality analysis it is crucial to understand the results, so we can proceed by performing a preprocessing task before making available the data object, or in the other case to make another analysis quality that is more focused on a specific aspect of the data object. We can improve the results obtained of the analysis by changing the data object (for example we are interested only in one attribute, or in other case we want to limit values on some of the attributes), or set up different quality dimension to evaluate the same data object.

By varying the quality dimensions, we can also even change the level of detail we want the analysis to take place on. This will make the quality of the results to increase.

In the next chapter we will go in detail over a proposed data exploration tool that helps the users examining the produced results as a support component of the EU-Bra Data Quality Service.

# Chapter 4

# A METHODOLOGY FOR QUALITY RESULTS EXPLORATION

The core of the thesis will be a proposition and implementation of a model that will act as a support component (User Interface) of the Quality Service by BIGSEA. The main task on which this model will be concentrated is exploration and guidance over the results that arrive after a quality analysis performed on the Quality Service.

This is probably the most important part of a quality analysis because the user needs to make a final decision if the data quality scores are good enough to be used in next computations and statistics software's, or the data should pertain a new quality analysis with modified settings. In order to help the user with decision making and save future and unnecessary re-processing tasks on the Quality Service we propose our method in this thesis.

In this chapter we will discuss main ideas and decisions that are behind the thesis. First in 4.1 we will briefly introduce the model and approach in its creation. In section 4.2 we will narrow down the need for a data exploration tool as a support component of the Data Quality Service and application tools depending on the service. Then, in 4.3, we will concentrate on the model's architecture.

## 4.1 DATA QUALITY EXPLORATION TOOL AS A SUPPORT COMPONENT OF THE QUALITY SERVICE - MODEL DESCRIPTIOn

The results of a data quality analysis can be improved varying two things: the data object of interest, and the configuration applied to the quality dimensions offered by the service. The group of quality dimension on which a user wants to perform an analysis can always be manipulated, as it is possible to select the level of details on each of these dimensions. As stated in Chapter 3, the configuration of the service allows the user to manually pick quality dimensions and their level of detail. By detail we mean the granularity level with which the quality dimension will produce an indicator of quality about the data object. This change in granularity level can lead to improved or worsened results.

The re-definition of the data object also leads to variation of the quality results since each aspect of the data source can be evaluated over set of dimensions. Having in mind these two factors that are shaping the quality analysis we must be able to understand how they work together and which aspects of the results are affecting. With that the need of a data quality exploration tool emerged.

### 4.1.1 Restrictions

To be able to explore results of different aspects of the data source first we must be sure the configuration of the quality analysis allows that. The selection of the data object will change also the possibilities of exploration that our model can propose to the user. For example, if the analysis was done only on attribute CODLINHA, of course it wouldn't be possible the explore different aspects of the data set.

A core part of the configuration of a quality analysis covers the selection of quality dimensions. Each quality analysis permits a level of detail over which the selected dimensions are computed. As a recap from chapter 3 follows the list with granularity (detail) levels on the dimensions:

- Low – evaluates the dimension globally in respect to the whole data source
- Medium – evaluates the dimension both over the data set and each attribute selected
- High – evaluates the dimension on the data set, each attribute and value selected

Clearly, the explorations that the model could allow would be possible only on the granularity levels selected in the configuration. Note that if the user selected the dimensions automatically the model will allow inspection of the results only on global granularity level.

### 4.1.2 Ideas and discussions

The proposed model will be built over the correlation between data object and quality measures configuration. The idea is to divide the level of detail in a way that the user can see how the quality dimensions are performing for different aspects of the data object. It can be the case where some dimensions are giving low scores when evaluated over the lowest detail level (global), but when observing attribute level of granularity, we see that the reason is that one of the attributes caused this low score in the first place. Also, it can happen that some dimension scores over an attribute are showing low scores, but when inspecting value level of granularity, we can see that the reason is that only few of the records performed bad and that affected the score in the first place.

Beside the division on levels of granularity, in the user interest is to know how a group, or single records of an attribute are performing in comparison with the rest. This taking place in our model can save the user from making another quality analysis just on few records from the data set. Last, but not least, a custom filtering over the records of a data source against selection of quality dimensions can be a huge plus for the user in terms of time and processing power.

As we can see the usefulness of the model finds it place in two categories: guided tool for newcomers and conductive support of the QoS. In the following sections we will give further details about the features and practical usage of the model proposed.

#### *4.1.2.1 Guided tool that will increase the level of knowledge for data analysis*

Having in mind that data quality in Big Data is quite a new thing in the literature we are aware that a lot of newcomers will be brought in contact with quality services like the service offered by BIGSEA. To help the user in navigation of quality results, but also to learn more about the quality measures and levels of granularities this thesis is offering a support component of the Quality Services in a form of a guided exploration tool. We can go through a set of suggestive and informative features that will help the user while exploring data and bringing decisions:

- *While exploring global granularity it would be useful to show all dimensions according to a predefined scale. For example, if the score is bad mark the dimensions with red, otherwise if the score is good mark it green*
- *Let the user group the dimension by a processing task that will be done on the data set afterwards. For example: A user just did the quality analysis manually, choosing all three level of detail and now it is interested to see if the data source would be appropriate for a processing task that requires only correct data. An exploration tool can save the user from another analysis over the data and show the quality dimensions which are the most important for such processing (in this case that would be Precision, Distinctness, Consistency, and Accuracy as dimensions).*
- *Always state the importance of exploring deeper granularity since it can bring more precise information*
- *While exploring attribute level of granularity let the user know which attribute is giving good scores according the dimension*
- *While exploring value level of granularity suggest the user to explore each record independently and see how it performed in all other dimensions*
- *Let the user know how important it is to custom filter the results by bringing examples of its usage*

#### *4.1.2.2 Retaining consumption and preprocessing tasks*

The idea of the BIGSEA project in the first place is to propose an architecture that will be able to predict resource consumption of Big Data Analytics applications. Having this on mind we know that processing Big Data analysis can take a lot of time and energy and with that the cost of "value quality" will increase a lot. To save the user from iterative analysis until the fulfilled results are achieved, our model can be practical. We will go through some scenarios that can illustrate how an exploration tool can be more convenient instead of a new processing task.

**Scenario 1:** *Assume we want to find the bus lines that have the highest frequency and in the same time having completeness frequency greater than 0,6. In order to perform this we need to perform to separate tasks over the quality service, but instead if using an exploration tool, we can filter the set of data and save only the records that are satisfying our condition.*

**Scenario 2:** *A user wants to make statistical analysis over the data source and it is interested only in data records without duplication, and in the same time without missing values. In order to perform this, it is enough to do a custom filtering on which the user will ask to receive all records that have completeness missing set to 1 and distinctness set to 1.*

**Scenario 3:** *Assuming we did the quality analysis and we just remember that we only need to know how the CODLINHA 101 performed in all the dimensions. Making a new analysis would take a lot of time, but with a tool could be calculated in seconds.*

**Scenario 4:** *A user just did the analysis and while exploring the results finds out that the results of completeness frequency over attribute level are bad. To explore further and without making a new analysis, an exploration tool could group the records of each attribute into sets formed by a predefined threshold. Once exploring the sets with low values, the user can see how all the values with low frequency perform as a group comparing to all other attributes results.*

**Scenario 5:** *Extra feature that could be handy to a user would be the possibility to group by name, value, and range, all the records over value level of granularity. Afterwards an export option into excel file could save a lot of time to search and insert each row separately.*

### 4.1.3 Summary on the model features

To proceed further with the model architecture and implementation we will do a brief recap of all the possibilities the model will offer.

- Granularity exploration: All three levels of granularity could be inspected in an iterative way in order to suggest the user possible anomalies if the level of detail is low. Of course, it will be possible to directly go and inspect attribute or value level of granularity if the user prefers so.
  Global level of granularity will show how the data source evaluated in all quality dimensions. Each score will be marked as suggestion to the user (high scores with green color, and low scores with green). Beside the standard presentation of dimensions, the user could select a group of dimensions according category or data mining application.
  Attribute level of granularity can be reached from the global granularity results or also by a direct link. It will show how each of the selected attributes performed in a particular dimension. Similar, as in global granularity, each score will be marked as a suggestion to the user.
  Value level of granularity will show how the records of an attribute evaluated in a particular dimension. Here the user will have the possibility to search and export the data table, and to explore further group and single records.

- Group of values: Value level of granularity will sort all the records from the current attribute in three groups by their scores in the current dimension. The user could inspect each of the groups and understand how this group performed in comparison with the attribute.

- Single value: Value level of granularity will offer possibility to click on a single record and see how the record performed in all quality dimension.

- Custom filtering: This feature will allow the user custom inspection of records of an attribute.

## 4.2 Model Architecture

Many different components are forming the chain when it comes to data quality in Big Data. Since the model proposed with this thesis is placed as the last and support component of the Quality Service in the BIGSEA project, we will first go through the global architecture of the project, and then we will concentrate on the architecture of the Data Quality Exploration tool. Users or applications can use the Quality Platform to perform quality analysis over a data source saved on the same platform. The interaction between the users/applications is handled by a set of APIs'. For each new request that arrives a new configuration file is prepared where all the settings of the quality analysis are saved. Once the file is ready and saved in the custom settings repository, the request is processed by the DQ profiling and assessment module. This module interacts with the data source and all the profiling metadata regarding to the data source. When the results are ready they are saved in the quality metadata repository and are available to the end-user. Once the user is in a possession of the results uses the Data Quality Exploration tool. In the figure bellow we can see the global architecture and all the components.

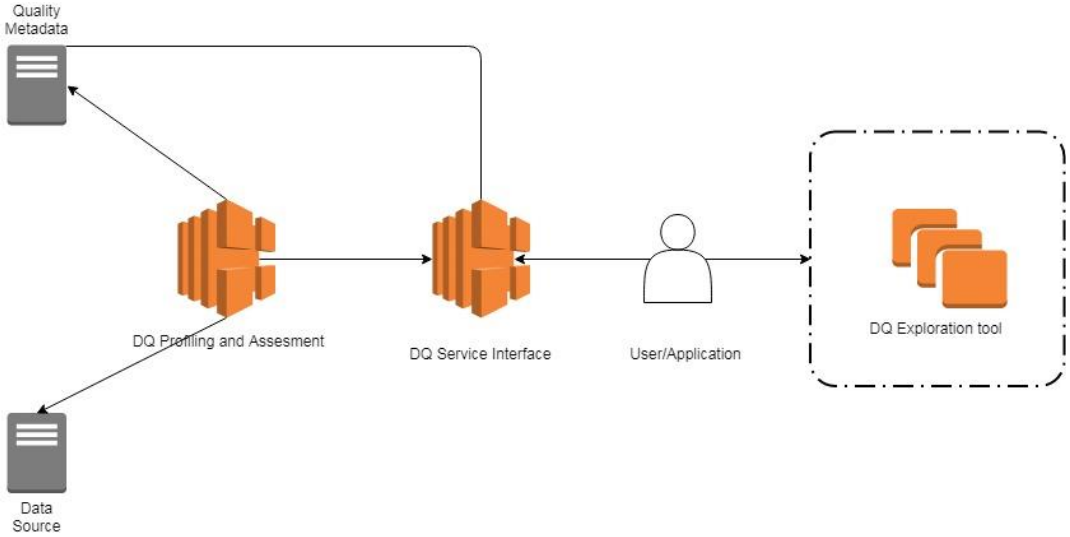


*Figure 11: Global Architecture of the Quality Platform*

The DQ Exploration tool is placed outside of the Quality Platform since at this moment can be only used as a standalone application.

### 4.2.1 DQ Exploration Tool Architecture

The Data Quality Exploration Tool architecture is composed by three modules (Data Source, Granularity Explorer, and Filtering – see Figure 12) that act as a bridge points between the user interface and the quality analysis results. All the results will be saved on a local server, powerful enough to process UI requests.

The Data Source module retrieves information about the data sources on which the user already did quality analysis and it is possible to explore their results. The module first takes information about the data source using the profiling file, and then contacts the server side of the tool to prepare the further exploration.

The Granularity Explorer module retrieves the quality results for all three levels of granularity. It communicates with the server about the user selection and renders the results in a User Interface. It is composed of three submodules: Global, Attribute, and Value submodule. Each of them retrieves quality results evaluated over quality dimensions. The Value submodule contains two additional nodes: Group selection and Single Record selection nodes.

Starting point for both Group and Single Record nodes is the attribute of aggregation and dimension already selected while exploring value level of granularity. Having this information, the Group node is responsible of retrieving information about set of records from multiple results files from the results directory. Similarly, the Single Record node explores multiple results files, but only for a single record.

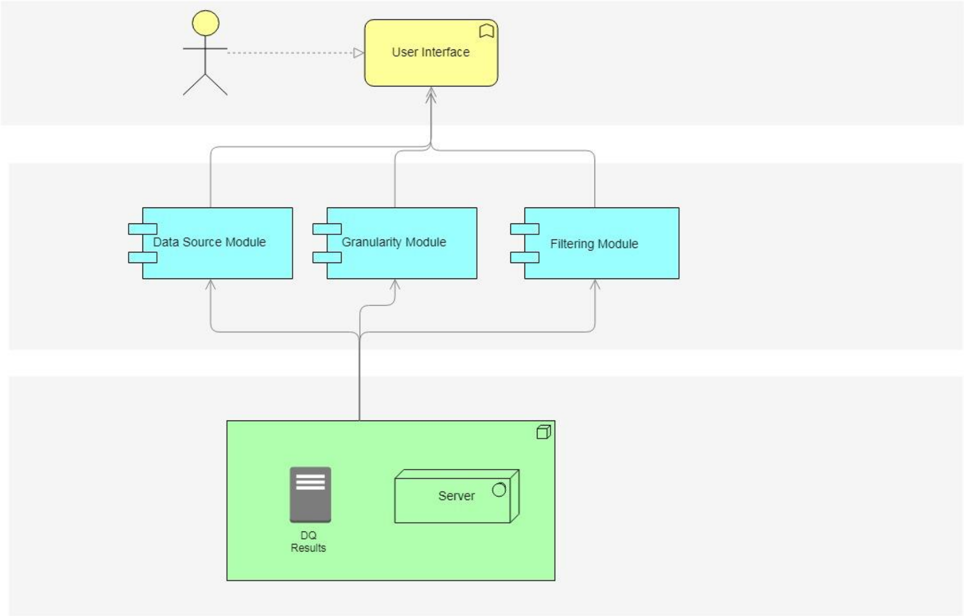


*Figure 12: Model Architecture*

### 4.2.2 Working logic and complexity

For a standalone application to be able to process all the output data it should be able to manipulate the results in a fast manner. All the results saved by the quality service depending of the granularity are kept in JSON files. The size of the files goes from smaller to bigger as the granularity goes from higher(global) to lower(value) level. The only exception that can occur is the file that carries results about consistency as quality measure over value level of granularity. It can happen that the user requested to check for an association rule that as antecedent value takes more than two attributes together.

To be able to show the data on request by the user such a tool must read from the directories and present the results in a user-friendly manner. The tasks that require the less complexity are the ones where the user is interested to see results just as they are saved by the quality service.

In particular, global, attribute, and value level of granularity results are kept in single JSON files which means the data tool has to locate the right directory and read the file. Contrary, the tasks that require the highest complexity and processing are the ones that require reading results from multiple files in the same time. The Filtering module is the most complex one. It passes all custom queries to server side. After processing a query, the server sends the results towards the UI. The complexity of the queries depends on the number of criteria.

## 4.3 CONCLUSION

The method described in this chapter finds its place as a support component of the Quality Service offered by the BIGSEA project. It is a standalone application and can be used only over quality results obtained by the service mentioned. The logic behind the project is based on the JSON files being saved in well-known format. Changes in the QoS export will require future modifications in the model too.

The model is created in such a way that It offers an iterative approach of exploration. As the level of details goes from lower to higher the quality results are more precise and give more information to the user. This can be useful to understand which aspect of the data can be of huge importance in further processing software's.

Variations and re-computations can be computed iteratively on the quality service until the user gets satisfying results. This is also possible on the model with the usage of the filtering module, and that gives even more importance. The user can create many scenarios and choose a set of dimensions in order to obtain more information. Of course, this is only possible if the quality results are including value level of granularity evaluations.

When it comes to processing power, the model doesn’t require extra nodes since all the processing can be done on the user’s machine.

# Chapter 5

# MODEL IMPLEMENTATION

In the following part we will go through the implementation of the data quality exploration tool according the ideas and discussions we present so far in chapter 4. First, we will see how all the results are saved and forwarded towards the user after one quality analysis finishes. Then, in point 5.2 we will go through the technologies used in the implementation. In 5.3 and 5.4 the workflow and all the features of the model will be explained in detail.

## 5.1 ASSESSMENT RESULTS FROM THE EU-BRA DATA QUALITY SERVICE

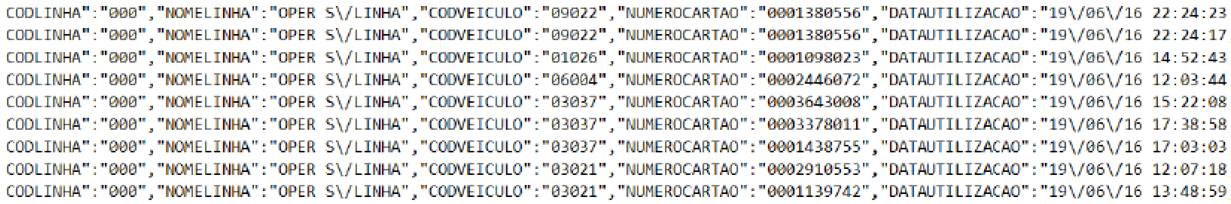
As it was mentioned in the chapter before, when setting up a quality analysis over the Quality Service the user has to choose a data object of interest and to set up quality dimensions that will be computed during the analysis. The APIs exposed by the service allow a user to develop a standalone application, a web or phone application in order to set up the analysis. The idea was to invoke the APIs in such a way that a configuration file will be created to keep the analysis request. Later, according to the configuration file and the profiling metadata that is kept, the quality service can perform the analysis and produce results.

All the experimental results obtained by the service are over data sources available by the Eu -Bra BIGSEA Project. For our interest, in the following section we will concentrate only on experimental results regarding a data source for the public transportation system of the city of Curitiba, Brazil. The source that we will take into inspection in particular in this chapter is related to ticket validations performed by users on the public buses in the city. The available information in the data source is identified by:

- CODLINHA: code of a bus line
- NOMELINJA: name of a bus line
- CODVEICULO: code of the vehicle assigned to the bus line

- NUMEROCARTAO: number of the user card that makes the validation
- DATAUTILIZICAO: date of validation of the ticket

In the figure bellow we can see how the information is saved in the data source:



```
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"09022","NUMEROCARTAO":"0001380556","DATAUTILIZACAO":"19\/06\/16 22:24:23
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"09022","NUMEROCARTAO":"0001380556","DATAUTILIZACAO":"19\/06\/16 22:24:17
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"01026","NUMEROCARTAO":"0001098023","DATAUTILIZACAO":"19\/06\/16 14:52:43
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"06004","NUMEROCARTAO":"0002446072","DATAUTILIZACAO":"19\/06\/16 12:03:44
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"03037","NUMEROCARTAO":"0003643008","DATAUTILIZACAO":"19\/06\/16 15:22:08
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"03037","NUMEROCARTAO":"0003378011","DATAUTILIZACAO":"19\/06\/16 17:38:58
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"03037","NUMEROCARTAO":"0001438755","DATAUTILIZACAO":"19\/06\/16 17:03:03
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"03021","NUMEROCARTAO":"0002910553","DATAUTILIZACAO":"19\/06\/16 12:07:18
CODLINHA":"000","NOMELINHA":"OPER S\/LINHA","CODVEICULO":"03021","NUMEROCARTAO":"0001139742","DATAUTILIZACAO":"19\/06\/16 13:48:59
```

*Figure 13: instance of user validation data source*

Once the Quality Service finishes with the analysis it saves the results in a directory named "DQAssessment_results". The output produced is composed of different objects, where each object represents a quality dimension calculated over an aspect of the data source.

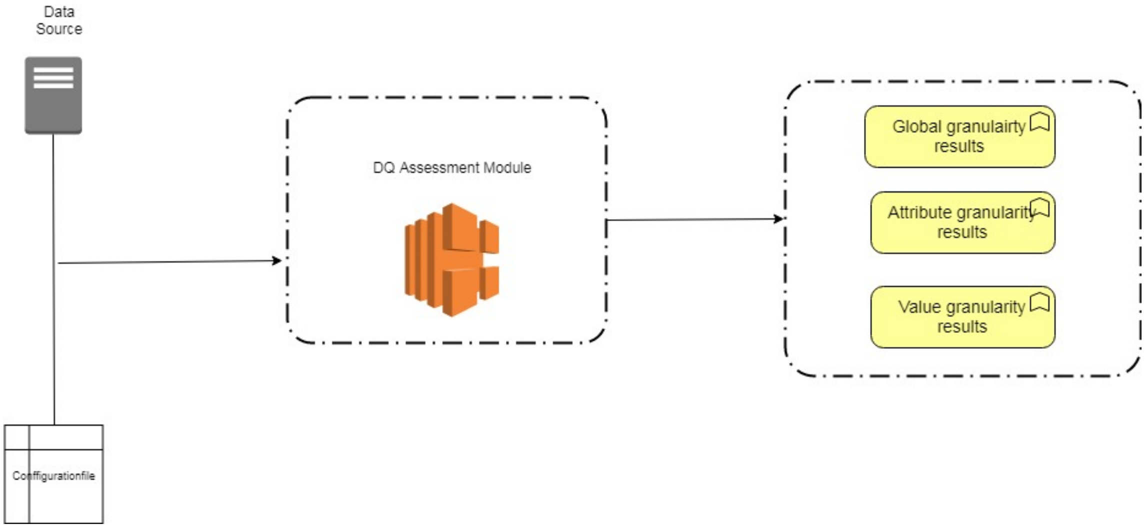

Data Source
DQ Assessment Module
Global granularity results
Attribute granularity results
Value granularity results
Configurationfile

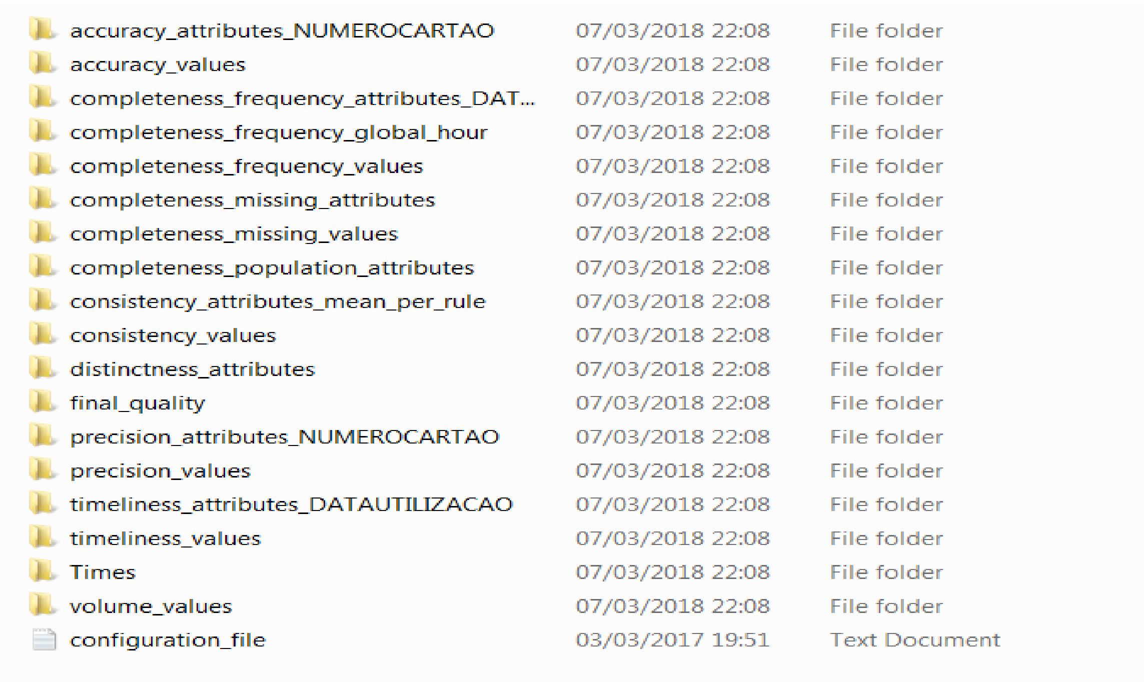
*Figure 14: Output of an analysis*

When exploring the results of a quality dimension the user can see the evaluation of the dimension on different levels of granularity. The level of granularity can be selected when setting up the analysis and of course if the data object allows such inspection. In the implementation of the Eu-Bra BIGSEA project it is possible to inspect three levels of granularity, so the results will be grouped accordingly:

- Global analysis results
- Attribute analysis results
- Value analysis results (for each attribute value of the data source)

Depending the level of granularity selected from the user, the quality service saves the output results in JSON files according quality dimensions. The only exception is the global granularity results file, it is a single file which contains the final results of each quality dimension selected over the whole set of data.

In the figure bellow we can see how the quality service saves the output results in a case where the user requested for all three levels of granularities over all available quality dimensions:



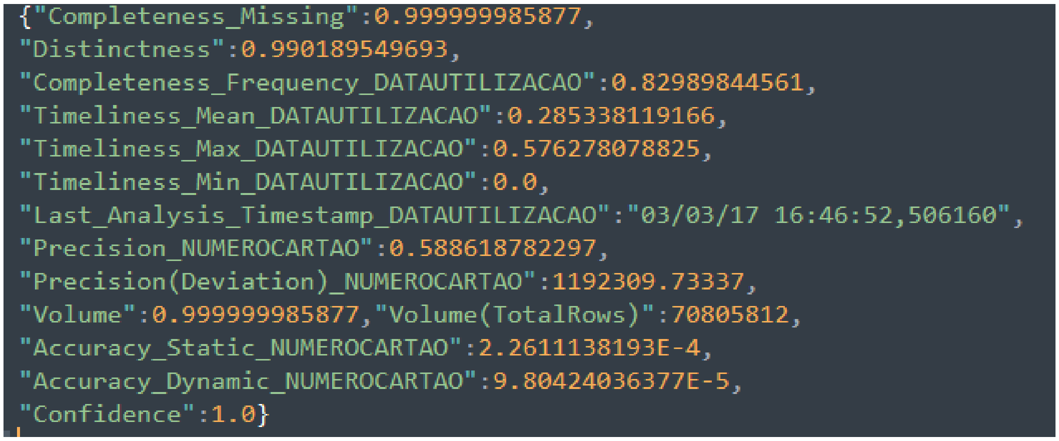
| | | |
|---|---|---|
| accuracy_attributes_NUMEROCARTAO | 07/03/2018 22:08 | File folder |
| accuracy_values | 07/03/2018 22:08 | File folder |
| completeness_frequency_attributes_DAT... | 07/03/2018 22:08 | File folder |
| completeness_frequency_global_hour | 07/03/2018 22:08 | File folder |
| completeness_frequency_values | 07/03/2018 22:08 | File folder |
| completeness_missing_attributes | 07/03/2018 22:08 | File folder |
| completeness_missing_values | 07/03/2018 22:08 | File folder |
| completeness_population_attributes | 07/03/2018 22:08 | File folder |
| consistency_attributes_mean_per_rule | 07/03/2018 22:08 | File folder |
| consistency_values | 07/03/2018 22:08 | File folder |
| distinctness_attributes | 07/03/2018 22:08 | File folder |
| final_quality | 07/03/2018 22:08 | File folder |
| precision_attributes_NUMEROCARTAO | 07/03/2018 22:08 | File folder |
| precision_values | 07/03/2018 22:08 | File folder |
| timeliness_attributes_DATAUTILIZACAO | 07/03/2018 22:08 | File folder |
| timeliness_values | 07/03/2018 22:08 | File folder |
| Times | 07/03/2018 22:08 | File folder |
| volume_values | 07/03/2018 22:08 | File folder |
| configuration_file | 03/03/2017 19:51 | Text Document |

*Figure 15: DQ Assessment results*

From the directory above we can see that the results are saved in three types of folders according the level of granularity.

### 5.1.1 Global granularity results

“final_quality” contains a JSON file with evaluation of all quality dimensions of selection over the data source.



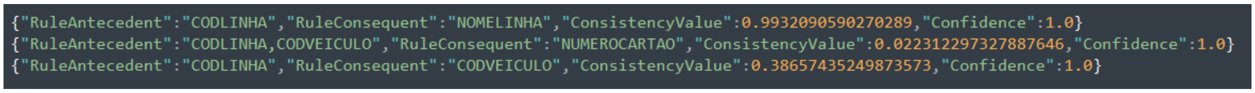
```
{"Completeness_Missing":0.999999985877,
"Distinctness":0.990189549693,
"Completeness_Frequency_DATAUTILIZACAO":0.82989844561,
"Timeliness_Mean_DATAUTILIZACAO":0.285338119166,
"Timeliness_Max_DATAUTILIZACAO":0.576278078825,
"Timeliness_Min_DATAUTILIZACAO":0.0,
"Last_Analysis_Timestamp_DATAUTILIZACAO":"03/03/17 16:46:52,506160",
"Precision_NUMEROCARTAO":0.588618782297,
"Precision(Deviation)_NUMEROCARTAO":1192309.73337,
"Volume":0.999999985877,"Volume(TotalRows)":70805812,
"Accuracy_Static_NUMEROCARTAO":2.2611138193E-4,
"Accuracy_Dynamic_NUMEROCARTAO":9.80424036377E-5,
"Confidence":1.0}
```

*Figure 16: Global granularity evaluation*

### 5.1.2 Attribute granularity results

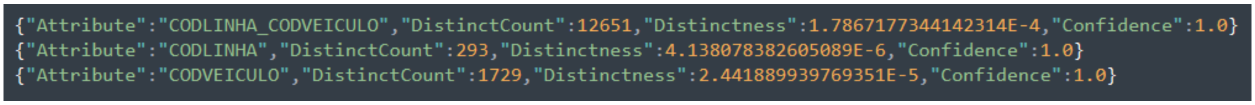
Folders containing “attribute” represent attribute level of granularity. For each dimension the service creates a JSON file with evaluation over each attribute of selection. Of course, not all dimension can be evaluated over attribute level of granularity.

In the examples bellow we can see how consistency and distinctness are evaluated over CODLINHA and CODVEICULO and CODLINHA_CODVEICULO since they were selected as key attributes by the user when setting up the analysis.



```
{"RuleAntecedent":"CODLINHA","RuleConsequent":"NOMELINHA","ConsistencyValue":0.9932090590270289,"Confidence":1.0}
{"RuleAntecedent":"CODLINHA,CODVEICULO","RuleConsequent":"NUMEROCARTAO","ConsistencyValue":0.022312297327887646,"Confidence":1.0}
{"RuleAntecedent":"CODLINHA","RuleConsequent":"CODVEICULO","ConsistencyValue":0.38657435249873573,"Confidence":1.0}
```

*Figure 17: Attribute granularity evaluated over Consistency*

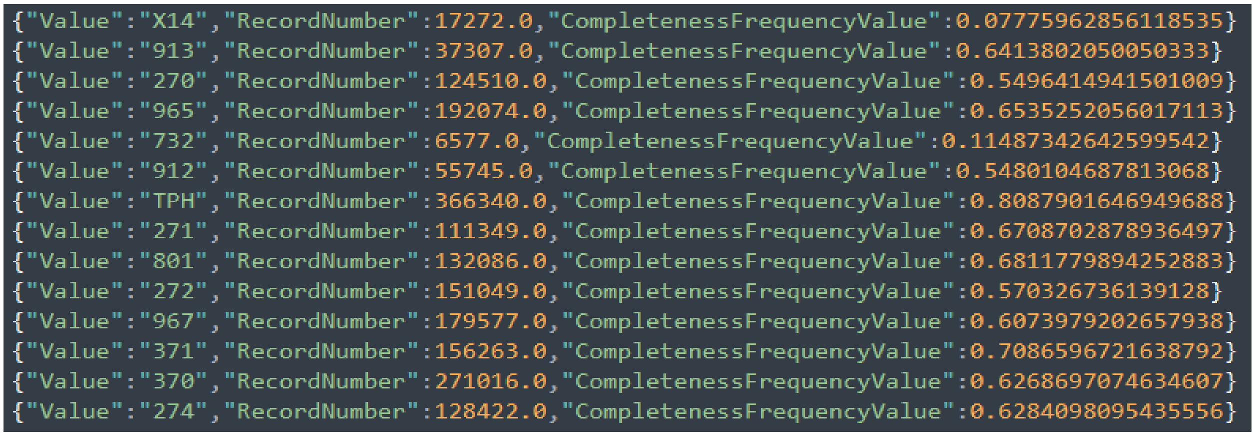


```
{"Attribute":"CODLINHA_CODVEICULO","DistinctCount":12651,"Distinctness":1.7867177344142314E-4,"Confidence":1.0}
{"Attribute":"CODLINHA","DistinctCount":293,"Distinctness":4.138078382605089E-6,"Confidence":1.0}
{"Attribute":"CODVEICULO","DistinctCount":1729,"Distinctness":2.441889939769351E-5,"Confidence":1.0}
```

*Figure 18: Attribute granularity evaluated over Distinctness*

### 5.1.3 Value granularity results

Folders containing "value" represent value level of granularity. For each dimension the service creates a JSON file with evaluation over each value of each aggregation attribute. For example, if the aggregation attributes are CODLINHA and CODVEICULO, the file will contain results for each row of records of that attribute. Of course, not all dimension can be evaluated over value level of granularity. In the set in the figure bellow we can see how completeness frequency is calculated over each CODLINHA record in the data source.



```
{"Value":"X14","RecordNumber":17272.0,"CompletenessFrequencyValue":0.07775962856118535}
{"Value":"913","RecordNumber":37307.0,"CompletenessFrequencyValue":0.6413802050050333}
{"Value":"270","RecordNumber":124510.0,"CompletenessFrequencyValue":0.5496414941501009}
{"Value":"965","RecordNumber":192074.0,"CompletenessFrequencyValue":0.6535252056017113}
{"Value":"732","RecordNumber":6577.0,"CompletenessFrequencyValue":0.11487342642599542}
{"Value":"912","RecordNumber":55745.0,"CompletenessFrequencyValue":0.5480104687813068}
{"Value":"TPH","RecordNumber":366340.0,"CompletenessFrequencyValue":0.8087901646949688}
{"Value":"271","RecordNumber":111349.0,"CompletenessFrequencyValue":0.6708702878936497}
{"Value":"801","RecordNumber":132086.0,"CompletenessFrequencyValue":0.6811779894252883}
{"Value":"272","RecordNumber":151049.0,"CompletenessFrequencyValue":0.570326736139128}
{"Value":"967","RecordNumber":179577.0,"CompletenessFrequencyValue":0.6073979202657938}
{"Value":"371","RecordNumber":156263.0,"CompletenessFrequencyValue":0.7086596721638792}
{"Value":"370","RecordNumber":271016.0,"CompletenessFrequencyValue":0.6268697074634607}
{"Value":"274","RecordNumber":128422.0,"CompletenessFrequencyValue":0.6284098095435556}
```

*Figure 19: Value level of granularity over CODLINHA and completeness frequency as evaluation dimension*

## 5.2 TECHNOLOGIES USED

The user interface of the project is developed in HTML/JavaScript, while the server side is based on Node.js.

JavaScript is a high-level, interpreted programming language. It is characterized as dynamic, prototype-based and multi-paradigm. Together with HTML and CSS provides interactive and reliable online programs. On the other side, Node.js represents an open-source, cross-platform JavaScript run-time environment for server-side scripting. The communication between the two parts is crucial since all the information is being first processed on server side and rendered on the front end.

Each page of the tool is dynamically built in such a manner that first instantiates an object of its type (for example Attribute granularity page instantiates attribute class) from which a HTTP AJAX request is sent towards the server for content retrieval. Each HTTP request from the front-end is identified on server-side by a uniform resource locator (URI). Once the request is processed, a response is being forwarded in a JSON format. Follows the list of resources used by the application:

- "/getSources" a URI used for retrieval of the data sources available
- "/results" a URI used for retrieval of the global granularity
- "/getAttributes" a URI used for retrieval of attributes of a dimension
- "/value" a URI used for retrieval of value granularity
- "/custom" a URI used for setting custom filtering

Note that all URIs require query attributes so that it will be feasible for the server to determine the source, dimension, or attribute of interest.

## 5.3 Navigation models in the UI

In this section we will go through the dynamic model of the UI. We will accomplish that by providing various diagrams to describe the interaction between objects and other components of the system. Later we will go through each page and feature in detail. Follows a set of navigations models from the user experience point of view.

### 5.3.1 Homepage

After the data source selection, the server will respond with global granularity details from the selected source. The idea is to start iteratively from the lowest level of detail so the user will not be confused by too many details. This will be on the first navigation tab on the homepage. Of course, if the user prefers to explore directly higher level of granularity, the system offers direct links from the navigation tabs: Attribute, and Value Selection. There is also a link to Custom Filtering. From the Figure bellow we can see components on both pages, and the use cases that can lead to different navigation tabs and features.

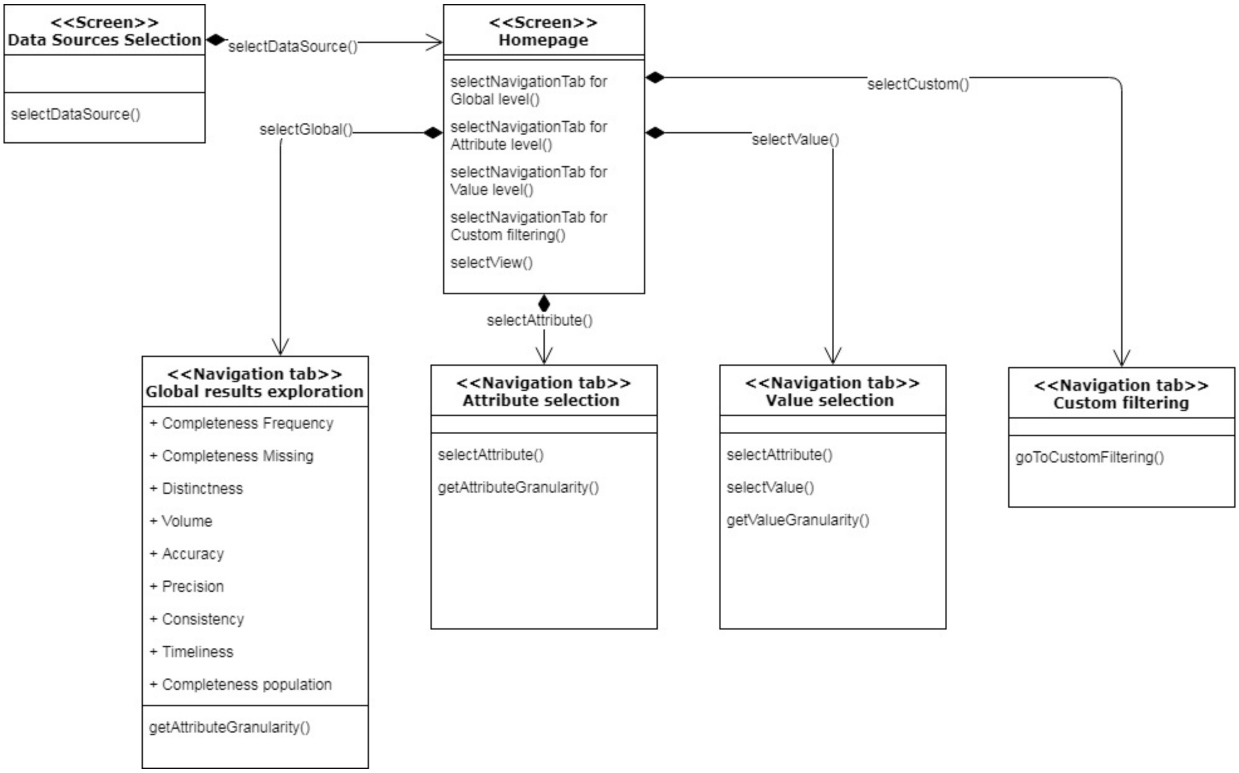
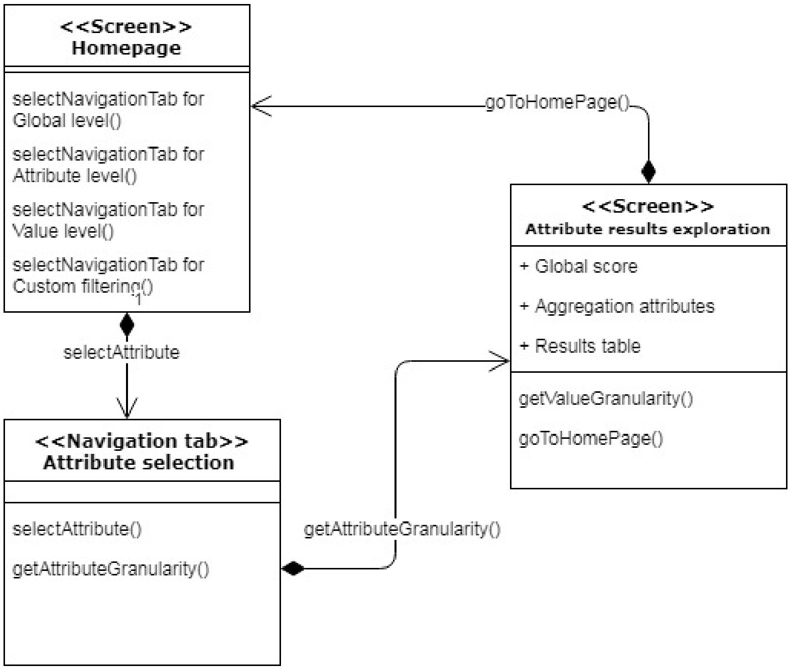


*Figure 20: Homepage of the UI*

### 5.3.2 Attribute page

The diagram below describes the navigation from the home screen to the attribute level of granularity. Note that it is always possible to return to the home screen and select another attribute. From attribute level of granularity, the user can also navigate to value granularity of the selected attribute.



<<Screen>>
Homepage
selectNavigationTab for Global level()
selectNavigationTab for Attribute level()
selectNavigationTab for Value level()
selectNavigationTab for Custom filtering()
1
selectAttribute
goToHomePage()
<<Screen>>
Attribute results exploration
+ Global score
+ Aggregation attributes
+ Results table
getValueGranularity()
goToHomePage()
<<Navigation tab>>
Attribute selection
selectAttribute()
getAttributeGranularity()
getAttributeGranularity()

*Figure 21: Exploration on attribute level*

### 5.3.3 Value granularity and extra features

The diagram bellow presents navigation through value level of granularity. Note that in this case for the sake of simplicity the screen is reached from attribute page. Of course, the screen is also reachable from the homepage. From this screen follows a navigation to the pages of single record explorer and page that does comparison of a selected set over the attribute scores.

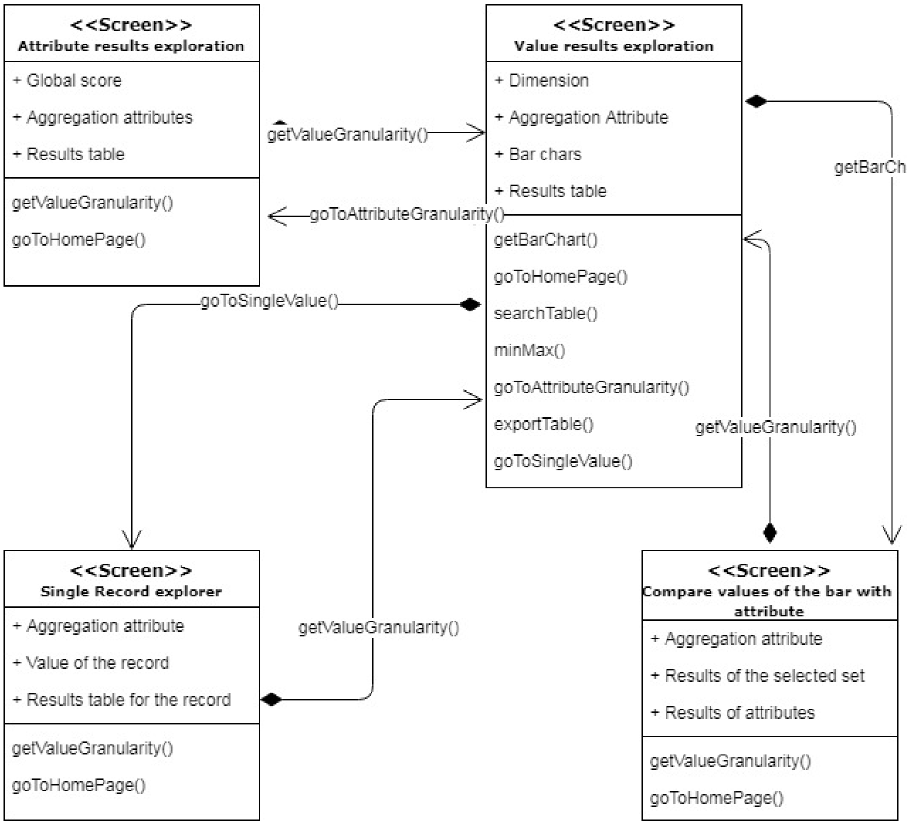


*Figure 22: Exploration on value level*

### 5.3.4 Custom filtering

The page of custom filtering can be reached only form the home page of the UI. Once the user selects attribute and dimensions of interest there is a navigation that leads to the result page.

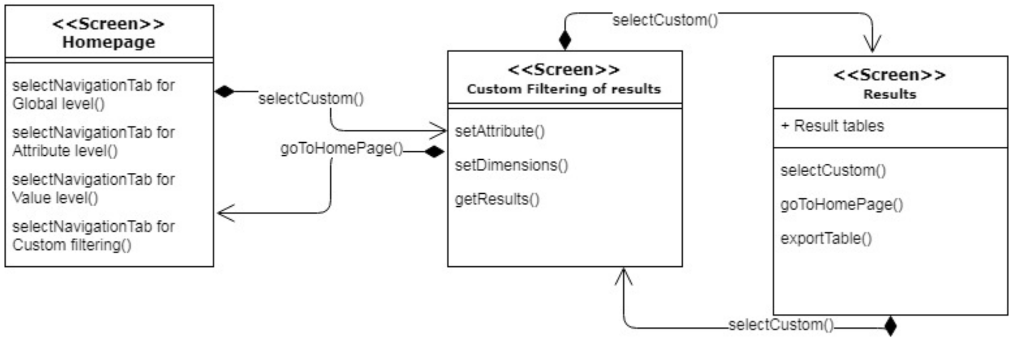


*Figure 23: Custom filtering*

## 5.4 User interface of the DQ Exploration tool

The following section describes the data exploration tool from user's perspective. We will navigate through all of the pages and features available for different use case scenarios.

### 5.4.1 Data source selection

When accessing the exploration tool, the user is asked to select the data source of interest.

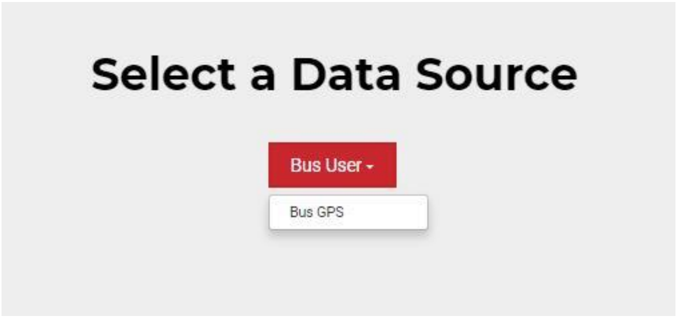


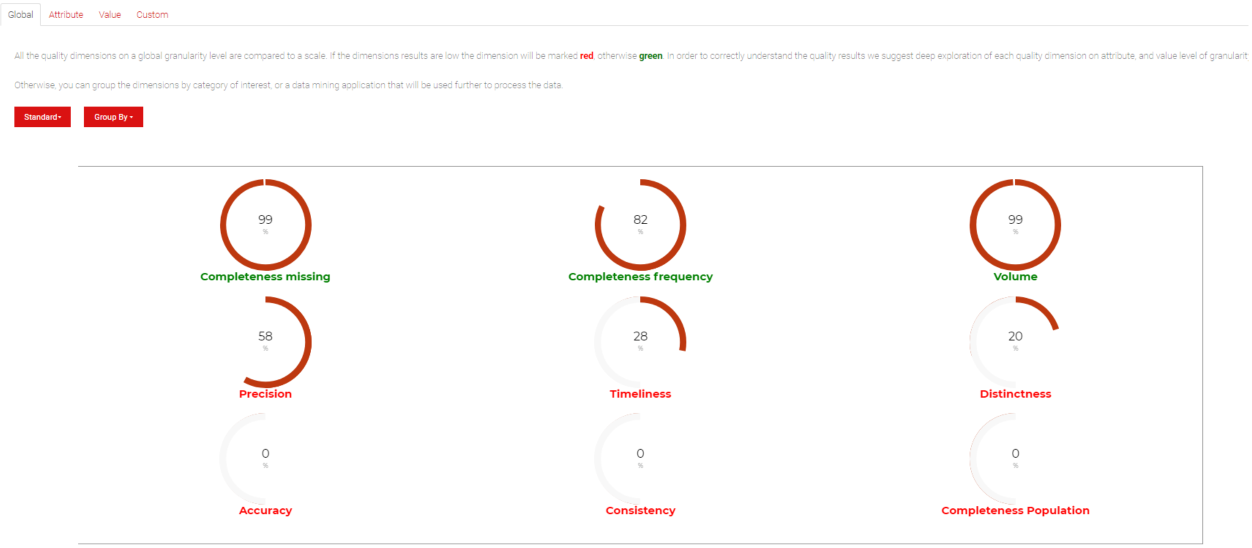
*Figure 24: Selection of a data source*

## 5.4.2 Home page

After selecting the data source on which the user wants to explore the results, the UI makes a request on server side to get access to the results concerning the data source of selection. At first the UI shows to the user the results concerning global level of granularity. The global results are placed in the first navigation tab. There are also navigations tabs that can lead to direct links for attribute and value granularity, and custom filtering.

As it can be seen from the image below, all the quality dimensions are placed in equally sized boxes with the scores accordingly.

In order to help the user, the UI classifies the results for each quality dimension by a set of predefined values that act as a threshold. For instance, if completeness frequency global scores are below 0,35 the UI marks them as bad results and scores above 0,8 as high results. In example, if the global score for completeness frequency in our analysis is 0,30 the UI will mark the box with red border. Otherwise, if it is above 0,8 the box will be marked with green color.



All the quality dimensions on a global granularity level are compared to a scale. If the dimensions results are low the dimension will be marked **red**, otherwise **green**. In order to correctly understand the quality results we suggest deep exploration of each quality dimension on attribute, and value level of granularit

Otherwise, you can group the dimensions by category of interest, or a data mining application that will be used further to process the data.

*Figure 25: Home screen of the UI*

For each quality analysis the selection of dimensions and their configuration can be different. As we said, the user itself chooses when setting the analysis which dimensions will be used to perform the quality analysis:

- Custom selection
- Group dimensions by category

- Group dimensions by data mining application
- Automatic selection

In the UI we want to implement this once more, since it will bring different view of the scores, and it can save time for setting a new analysis.The idea is that the UI will allow the users to manipulate this dimensions and group them by category or data mining application. In this way the user gets a standalone view that will show results as when doing a new analysis and grouping the dimensions by category or data mining application. In the image bellow we can see how the UI offers grouping of the quality results by a data mining application.

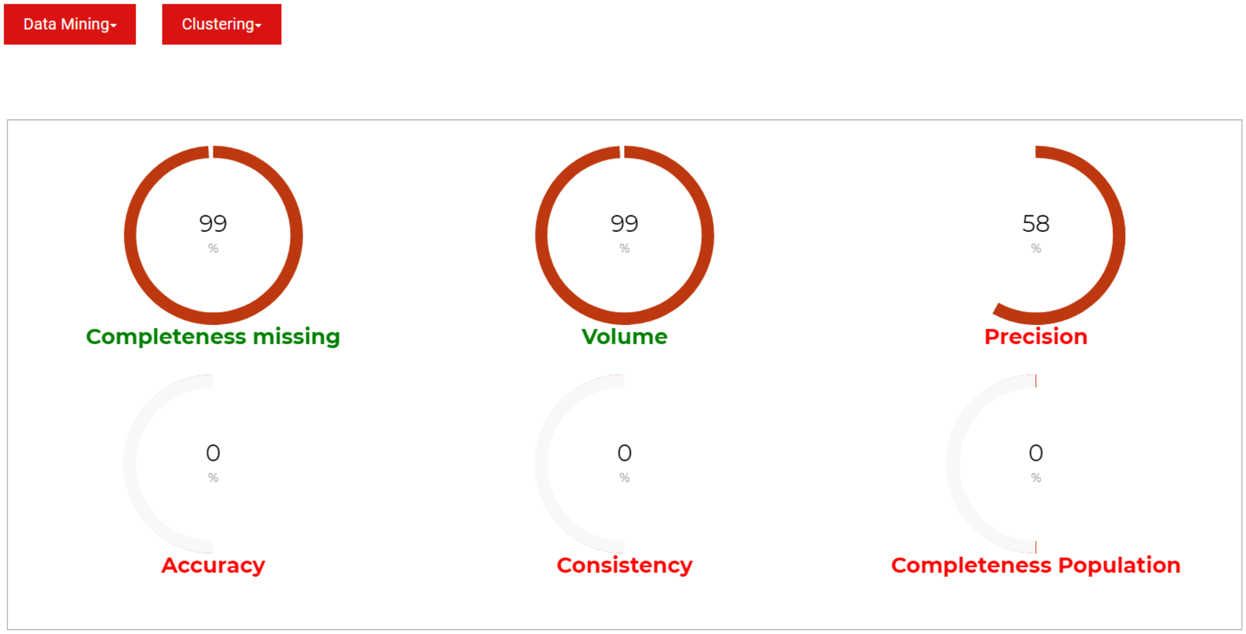


*Figure 26: Data Mining view of dimensions*

This feature is of a high importance because it can help an unexperienced user to determine which quality dimensions should be observed for the further usage of the data. In the case above the user intends to see how the data performed for dimensions that are associated to a data mining classification, since later the data will be used in such type of application for further processing. Beside of this configuration the user can select view within:

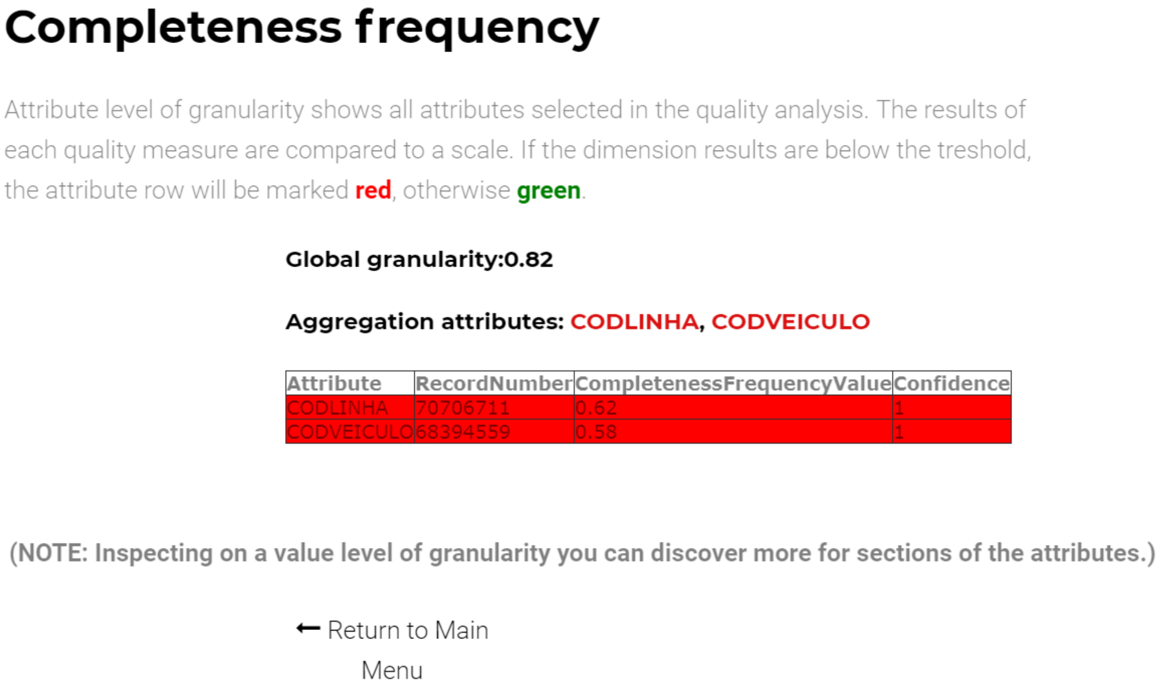
- Data Mining grouping where the application is Clustering
- Data Mining grouping where the application is Association Rule
- Category grouping where we associate Volume
- Category grouping where we associate Time Validity (Timeliness as dimension)
- Category grouping where we associate Correctness (Precision, Distinctness, Consistency, Accuracy as dimensions)
- Category grouping where we associate Completeness (Completeness Missing, Completeness Frequency, Completeness Population)

Of course, the user can always switch back to the first presentation of the quality dimensions. Note that if the quality analysis was performed with custom choice (selecting just some of the dimensions) it can happen that in the views some dimensions will be missing. To escape a confusion, the UI will warn of such a case.

### 5.4.3 Attribute level of granularity

Usually, when an unexperienced user uses the application selects the whole data source as an object of interest. Leaving the whole data source as a data object of interest can give a wrong information for the quality of the data. There can be parts of the source that are inadequate and are giving wrongful data which can create confusing results in a further usage.

The second level of depth regards the selection of attributes from the data source. This means that the user is interested in specific attributes of the source and wants to inspect their quality. This can give better quality results since each selected dimension is being run over each selected attribute. Think of an example where we performed only global analysis over a source and we get low scores for completeness missing. There is no way in this mode of inspection to know from which part of the data source the missing rows are coming. In fact, when selecting attributes of interest, we can get this information correctly.



Completeness frequency

Attribute level of granularity shows all attributes selected in the quality analysis. The results of each quality measure are compared to a scale. If the dimension results are below the treshold, the attribute row will be marked **red**, otherwise **green**.

**Global granularity:0.82**

**Aggregation attributes: CODLINHA, CODVEICULO**

| Attribute | RecordNumber | CompletenessFrequencyValue | Confidence |
|---|---|---|---|
| CODLINHA | 70706711 | 0.62 | 1 |
| CODVEICULO | 68394559 | 0.58 | 1 |

**(NOTE: Inspecting on a value level of granularity you can discover more for sections of the attributes.)**

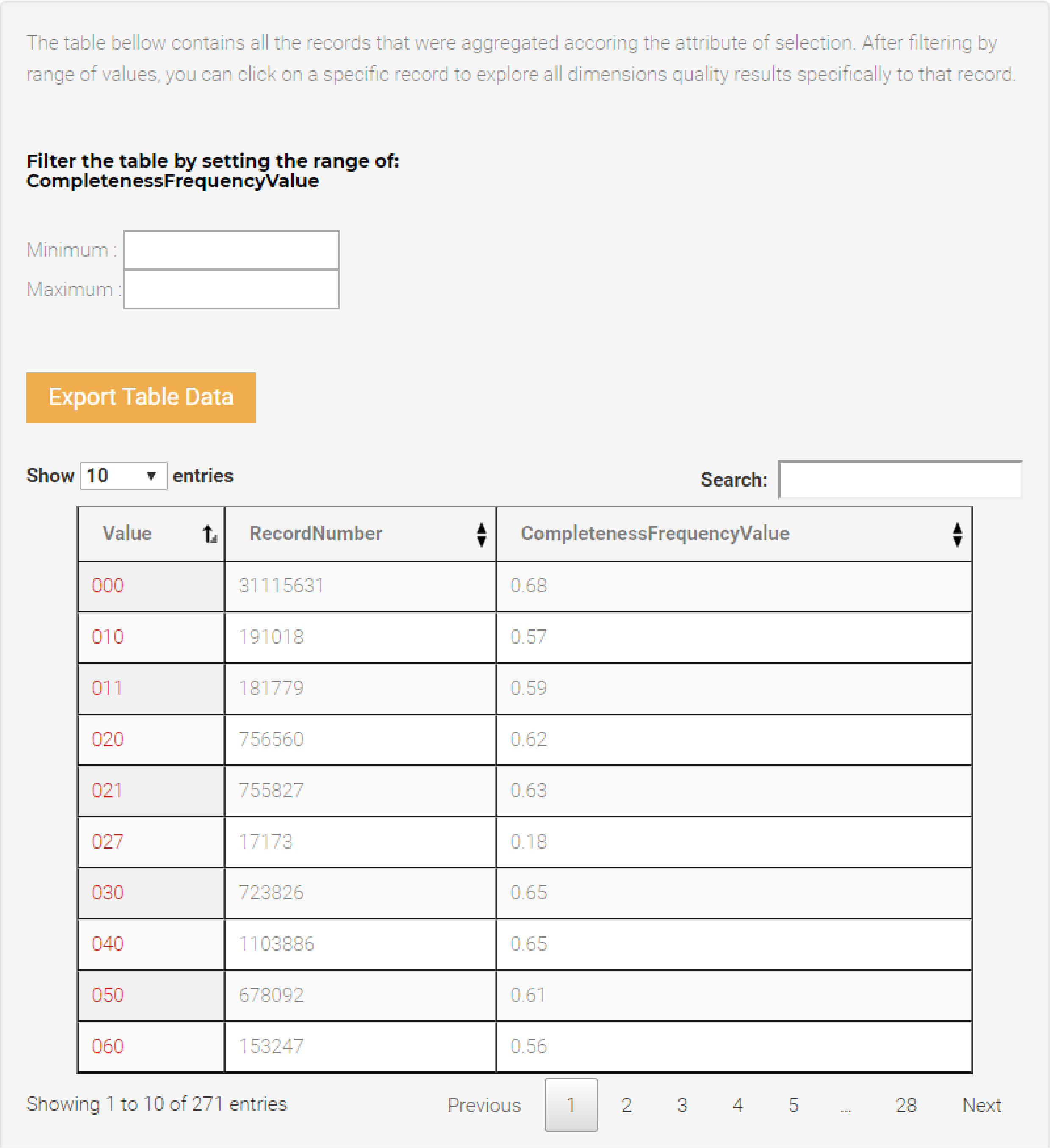
*Figure 27: Attribute level of granularity*

In Figure 27, we can see a screen of the attribute level of granularity of the UI. At the beginning section of the page the user is guided through the logic of marking attribute scores. The classification remains the same as on global depth, all attributes scores are compared to a threshold depending on the quality dimension of interest. This can help the user when setting a further analysis to include/exclude a specific attribute. Further, the user is advised to explore results on deeper level as they can give more details of the selected attribute.

### 5.4.4 Value level of granularity

Last level of depth regarding the quality configuration is value granularity. When setting up the analysis in the QS the users were also allowed to choose values for each attribute included in the data object. There are two types of possible selections for each attribute: set of values selections and an interval of value selection. This can have huge impact on the quality of the results because discovers which exact values are having poor quality results, and so can be omitted from further inspection, or in opposite, they can be very important for further processing and predictions.

In the bottom part of the page we can see values of the attribute with their quality scores accordingly placed in a table. Further, the user can manipulate the data in the table by his/her choice with a filtering section that goes through the scores of the results. Also, it is possible to make a search over all rows of the table. Of course, it is also possible to export the data in xls format, which can find further usage to the user.



The table bellow contains all the records that were aggregated accoring the attribute of selection. After filtering by range of values, you can click on a specific record to explore all dimensions quality results specifically to that record.

**Filter the table by setting the range of: CompletenessFrequencyValue**

Minimum :

Maximum :

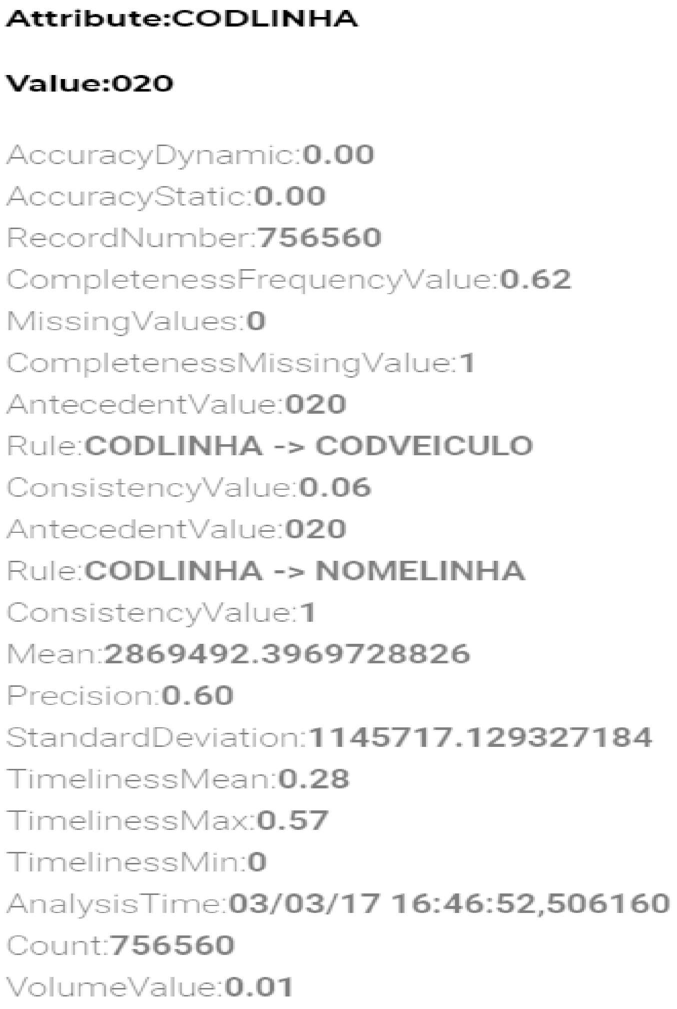
| Value | RecordNumber | CompletenessFrequencyValue |
|---|---|---|
| 000 | 31115631 | 0.68 |
| 010 | 191018 | 0.57 |
| 011 | 181779 | 0.59 |
| 020 | 756560 | 0.62 |
| 021 | 755827 | 0.63 |
| 027 | 17173 | 0.18 |
| 030 | 723826 | 0.65 |
| 040 | 1103886 | 0.65 |
| 050 | 678092 | 0.61 |
| 060 | 153247 | 0.56 |

Showing 1 to 10 of 271 entries

*Figure 28: Value level of granularity (table filtering)*

Small dimensions scores for a value can help us ignore the value in the next testing, and in opposite, high scores for a value can lead us to a new analysis where we can state exact value for an attribute.

Even before starting with a new analysis, the user can click on a value and explore more information concerning it. This leads us to the next feature of the UI (see Figure 29).



```
Attribute:CODLINHA

Value:020

AccuracyDynamic:0.00
AccuracyStatic:0.00
RecordNumber:756560
CompletenessFrequencyValue:0.62
MissingValues:0
CompletenessMissingValue:1
AntecedentValue:020
Rule:CODLINHA -> CODVEICULO
ConsistencyValue:0.06
AntecedentValue:020
Rule:CODLINHA -> NOMELINHA
ConsistencyValue:1
Mean:2869492.3969728826
Precision:0.60
StandardDeviation:1145717.129327184
TimelinessMean:0.28
TimelinessMax:0.57
TimelinessMin:0
AnalysisTime:03/03/17 16:46:52,506160
Count:756560
VolumeValue:0.01
```

*Figure 29: Exploring a single value of an attribute*

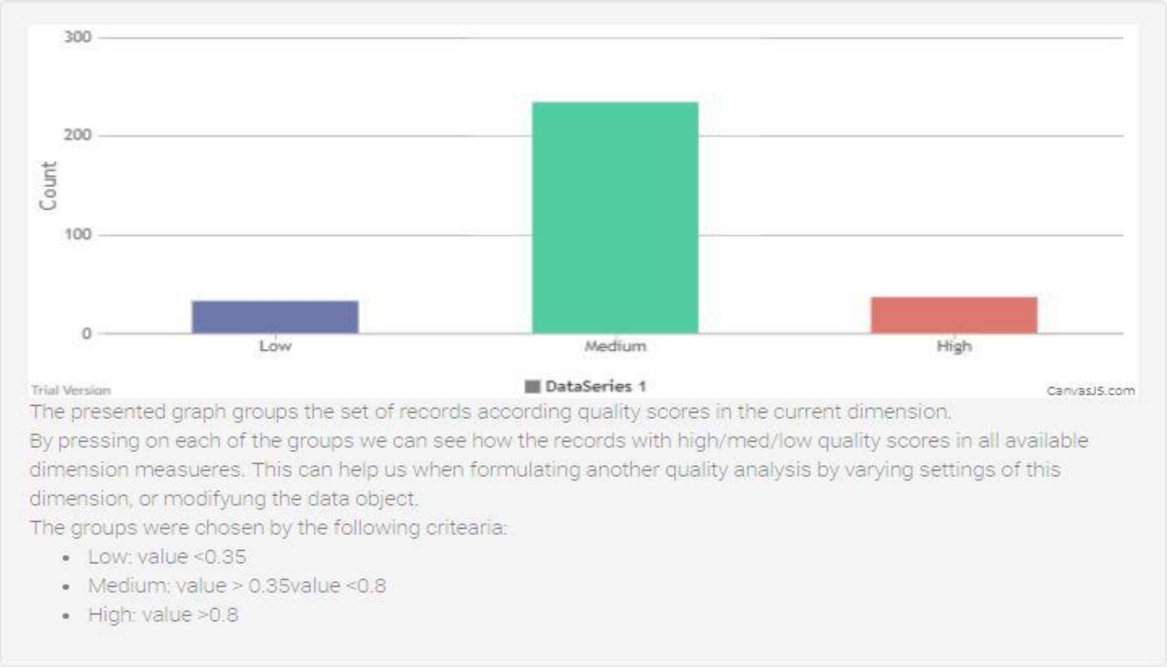
For the sake of clarity, this feature will be explained through the following scenario:

*Assuming the data source on which the analysis is done is the one regarding the BusUsers from the city of Curitiba, Brazil. The data in the source is collected from the ticket validation procedure in each of the buses of the city. The log is composed of the following attributes:*

- *CODLINHA – the code assigned to the bus line in which the ticket validation has been performed*
- *NOMELINHA – name of the bus line*
- *CODVEICULO – code of the vehicle*
- *NUMEROCARTA – code of the ticket card*
- *DATAUTILIZACAO – timestamp of the validation*

*Assuming the user explores value level of granularity for CODLINHA as an attribute and Volume as a quality dimension. Volume as a quality dimension will report which are the values that occur with low or high quality. So, when the user finds which code line is the most frequent one it could be useful to know more about that line. For instance, he could run a new quality analysis where the CODLINHA will have the value of the most frequent one. In order to save the user from doing that, the exploration tool offers a view where we can get to know how the bus line performed on value level of granularity for all of the available dimensions.*

The page that inspects value level of granularity offers another feature that brings in a way a deeper inspection and classification of the whole set of values that one attribute can have. The idea is that for each quality dimension that the user is exploring on value level of granularity the application offers grouping of the set of values (see figure below).



The presented graph groups the set of records according quality scores in the current dimension.
By pressing on each of the groups we can see how the records with high/med/low quality scores in all available dimension measueres. This can help us when formulating another quality analysis by varying settings of this dimension, or modifyung the data object.
The groups were chosen by the following critearia:

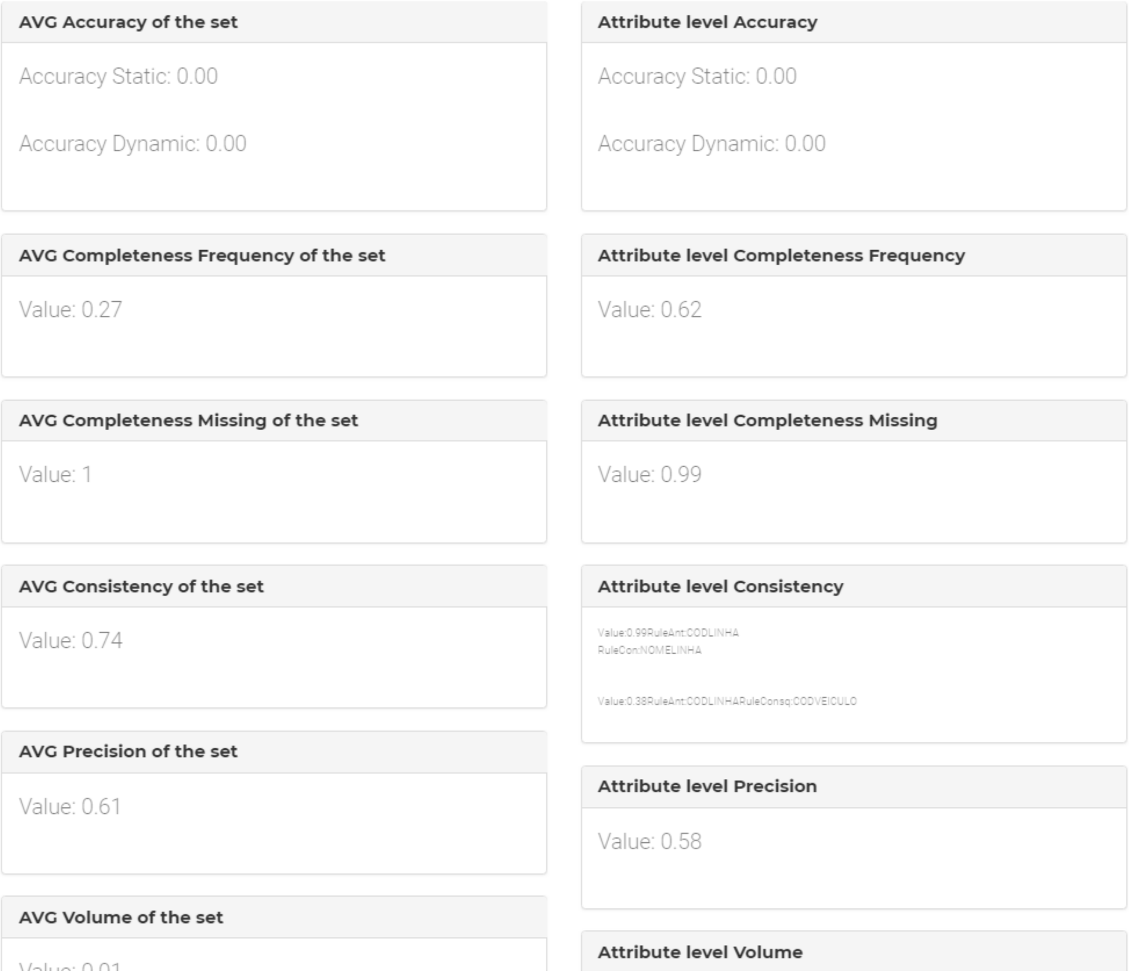
- Low: value <0.35
- Medium: value > 0.35value <0.8
- High: value >0.8

*Figure 30: Values classification*

The classification is done according a predefined quality values that act as reference threshold for low/high scores.

*For example, assuming the user is interested in value level of granularity scores by completeness frequency. The threshold is set at 0.35. This means that every value of the attribute of interest that has a score less than the low threshold will be placed in the set of low values. In opposite all the values that have scores above 0.80 threshold are placed in the set of high scores. All the values in between the thresholds will be placed in the medium set.*

Of course, the presentation of the graph provides a better understanding of the scores, but even a better exploration can be done of each of the sets. By clicking on a set from the graph, the application offers another view (see figure below).



AVG Accuracy of the set

Accuracy Static: 0.00

Accuracy Dynamic: 0.00

AVG Completeness Frequency of the set

Value: 0.27

AVG Completeness Missing of the set

Value: 1

AVG Consistency of the set

Value: 0.74

AVG Precision of the set

Value: 0.61

AVG Volume of the set

Value: 0.01

Attribute level Accuracy

Accuracy Static: 0.00

Accuracy Dynamic: 0.00

Attribute level Completeness Frequency

Value: 0.62

Attribute level Completeness Missing

Value: 0.99

Attribute level Consistency

Value:0.99RuleAnt:CODLINHA
RuleCon:NOMELINHA

Value:0.38RuleAnt:CODLINHARuleConsq:CODVEICULO

Attribute level Precision

Value: 0.58

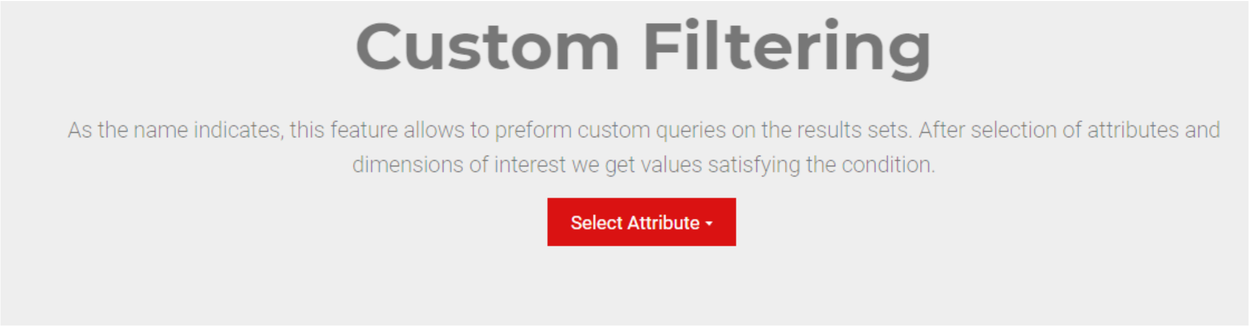
Attribute level Volume

*Figure 31: Comparison between values of the selected set and attribute level*

All the values of the set clicked are forming an average for each quality dimension used in the quality analysis. This average is compared to the quality dimension scores of attribute level of granularity (Note that this is the same attribute on which we inspect value level of granularity). This feature can be of a big interest for a professional in the field of quality analysis since it offers a comparison that couldn't be done when setting up an analysis.

### 5.4.5 Custom filtering

On the home screen the UI is offering an alternative exploration and filtering of the data source. When choosing this option, the UI shows another view where the user can see all available dimensions and filter them by choice. The filtering is performed on value level of granularity, where the user chooses the attribute of interest.



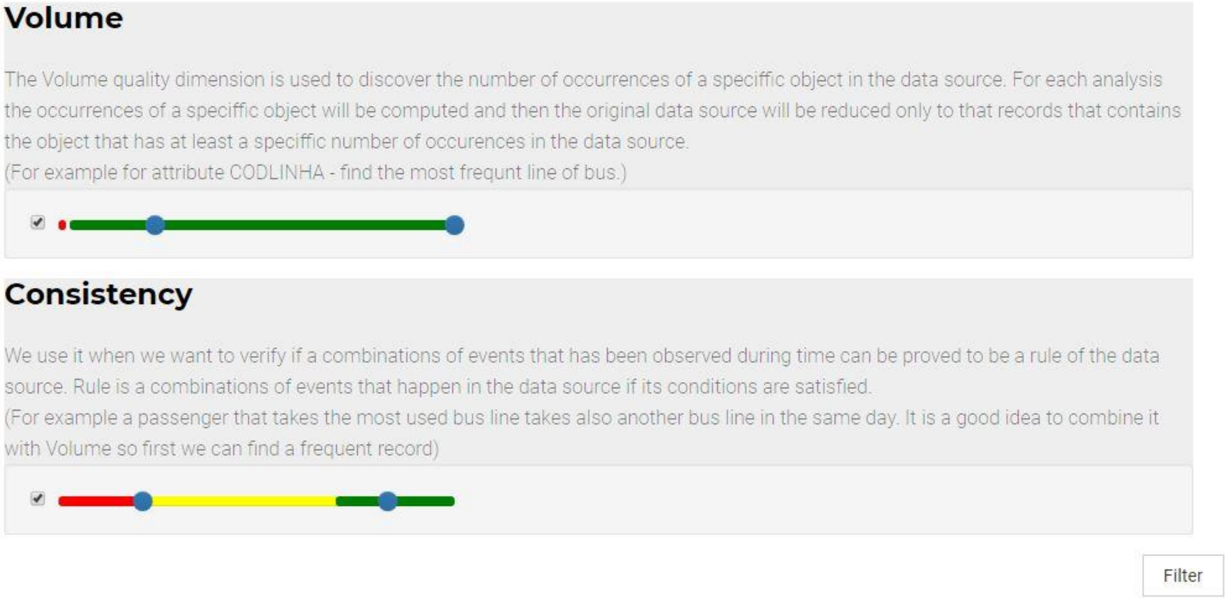
Custom Filtering

As the name indicates, this feature allows to preform custom queries on the results sets. After selection of attributes and dimensions of interest we get values satisfying the condition.

*Figure 32: Selection of attribute in custom filtering*

All dimensions are filtered by choosing a range of values. The default setting for each quality dimensions is on a predefined threshold for which it is assumed that the scores are in a high range (marked with green).

The idea is to combine two or more quality measures over values of a same attribute. For instance, the user would like to know the most frequent bus lines (CODLINHA) and see if there is an association rule associated to them. To perform this, we must first select the ranges of volume and consistency dimensions in the high ranges and exclude all the other dimensions when performing the query.



## Volume

The Volume quality dimension is used to discover the number of occurrences of a specific object in the data source. For each analysis the occurrences of a speciffic object will be computed and then the original data source will be reduced only to that records that contains the object that has at least a speciffic number of occurences in the data source.
(For example for attribute CODLINHA - find the most frequnt line of bus.)

☑

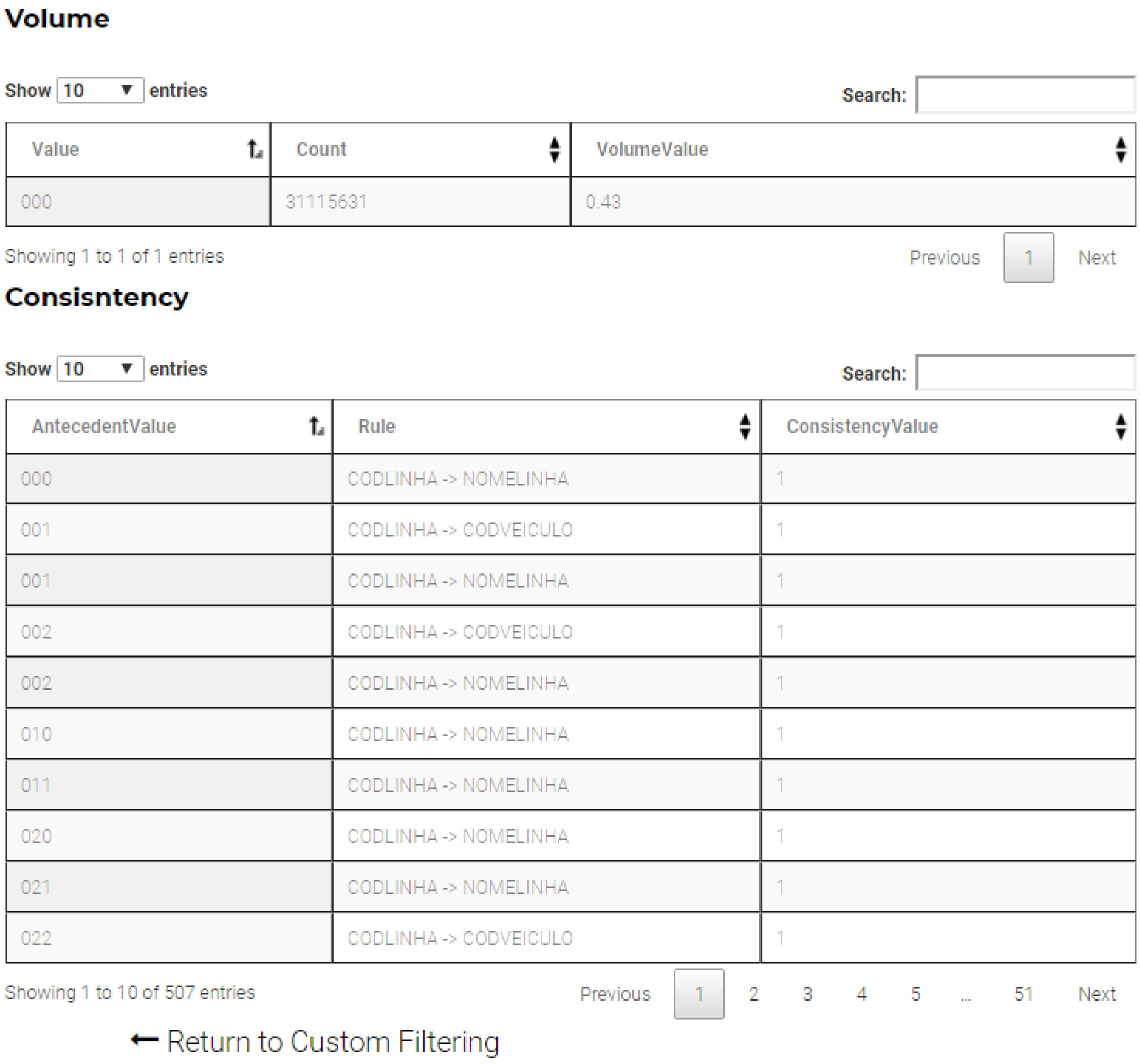
## Consistency

We use it when we want to verify if a combinations of events that has been observed during time can be proved to be a rule of the data source. Rule is a combinations of events that happen in the data source if its conditions are satisfied.
(For example a passenger that takes the most used bus line takes also another bus line in the same day. It is a good idea to combine it with Volume so first we can find a frequent record)

☑

*Figure 33: Filtering over two dimensions by choice*

In the result page the user can see tables related to volume and consistency dimensions with values that match the requirements. Most importantly, there is a merged table that shows values in common (in this case all the CODLINHA lines that are in common for the selected dimensions).



**Volume**

Show 10 ▼ entries

Search:

| Value | Count | VolumeValue |
|---|---|---|
| 000 | 31115631 | 0.43 |

Showing 1 to 1 of 1 entries

**Consisntency**

Show 10 ▼ entries

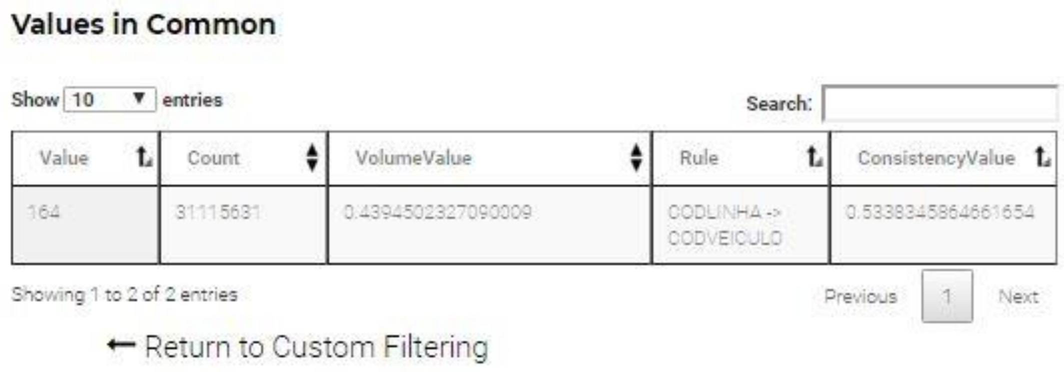
Search:

| AntecedentValue | Rule | ConsistencyValue |
|---|---|---|
| 000 | CODLINHA -> NOMELINHA | 1 |
| 001 | CODLINHA -> CODVEICULO | 1 |
| 001 | CODLINHA -> NOMELINHA | 1 |
| 002 | CODLINHA -> CODVEICULO | 1 |
| 002 | CODLINHA -> NOMELINHA | 1 |
| 010 | CODLINHA -> NOMELINHA | 1 |
| 011 | CODLINHA -> NOMELINHA | 1 |
| 020 | CODLINHA -> NOMELINHA | 1 |
| 021 | CODLINHA -> NOMELINHA | 1 |
| 022 | CODLINHA -> CODVEICULO | 1 |

Showing 1 to 10 of 507 entries

*Figure 34: Results of custom filtering feature*

The settings of a custom filtering could be done on all attributes and with a set of dimensions of our choice. That brings a lot of combinations and possibilities to know more about the data. Having the values in common between the dimension can help the user to define precise analysis.



Values in Common

| Value | Count | VolumeValue | Rule | ConsistencyValue |
|---|---|---|---|---|
| 164 | 31115631 | 0.4394502327090009 | CODLINHA -> CODVEICULO | 0.5338345864661654 |

Showing 1 to 2 of 2 entries

*Figure 35: Values satisfying filtering criteria's*

# Chapter 6

# CONCLUSION

The implemented model in this thesis aims to be a continuation and support of the Data Quality Service on the EU-Bra BIGSEA project. The QS offers an adaptive method with APIs that can be called from a user/application in order to perform a quality analysis over data source managed on the BIGSEA platform. After setting the analysis the quality service responds with evaluation of the data object over quality dimensions as measures. All the quality measures are evaluated in detail as the user requested.

The results of the analysis can help the user in decision-making when it has to be decided if the data is of high quality and can be used in further applications, or to discover that some aspect of the data object can be evaluated in a different way in order to obtain better quality results.

While the Quality Service allows many runs of the same data source with different configurations, it also allocates storage and processing power.

The first aim of the thesis is to help the user in result exploration in understanding how each aspect of the data object performed in the analysis. It can be the case that some aspect gave bad results because the configuration of the analysis was not set correctly, or a set of records are showing much higher scores than the rest. Clearly, it is not easy for the user, either professional or not, to see all the details while exploring huge set of results divided into multiple files. The model provides a guided exploration and gives suggestions to the user whether the results of a particular quality dimensions are high or low. It also suggests further exploration in detail.

Beside exploring all granularities, the model performs processing and manipulation of the results while serving custom filters on different aspects of the data object. This can be useful since it imitates a lot of scenarios and re-processing over the quality service.

Now the model is a standalone application outside the Quality Service of the BIGSEA project and every change on the side of the service will affect also this model. Further considerations and developments can be done related to this thesis in order to propose a merging of the DQ exploration tool as an internal component of the BIGSEA project. This can give benefits both to the user since all the components will be found in one place, and from development aspect.

# BIBLIOGRAPHY

[1] EU-Bra BIGSEA Project. 2017. Retrieved from http://www.eubra-bigsea.eu

[2] Pouria Amirian, Trudie Lang, Francois van Loggerenberg. "Big Data in Healthcare – Extracting Knowledge from Point-of-Care Machines". 2017

[3] Thomas Erl, Wajid Kkhattak, Paul Buhler. "Big Data Fundamentals. Concepts, drivers, and techniques". 2015

[4] "What is Big Data?".2016. Article published on http://www.bigdata.black

[5] "Big Data Value. Data: The elixir of life". 2013. Article published on Information Catalyst for Enterprise (ICE).

[6]" Data Quality in Big Data: A Review". 2015. Article published on Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3.

[7] Laura Sebastian-Coleman. "Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework". 2013.

[8] Richard Y.Wang Diane M. Strong. "Beyond Accuracy: What Data Quality Means to Data Consumers". 2015.

[9] "Big Data Technologies". 2017. Article published on Data Jobs. Retrieved from: https://www.datajobs.com

[10] Apache Hadoop. Retrieved from: http://www.hadoop.apache.org

[11] "Introduction to the MapReduce Life Cycle". Retrieved from: https://www.supinfo.com/articles/single/2807-introduction-to-the-mapreduce-life-cycle

[12] "What Apache Spark Does". Retrieved from: https://www.hortonworks.com/apache/spark/

[13] "What is Apache Spark?" Retrieved from: https://www.databricks.com/spark/

[14] Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, Rakesh Kumar Lenka "Big Data Visualization: Tools and Challenges". Conference paper. 2016.

[15] Lidong Wang, Guanghui Wang, Cheryl Ann Alexander. "Methods, Challenges and Technology Progress". 2015.

[16] Giacomo Maccagni. "A method for Raising Quality Awareness in Big Data Analysis". 2016.