Haaga-Helia
University of Applied Sciences

# Big Data solutions for market intelligence for a B2B company

Vivien Holecz

16 November 2017

# Abstract

Haaga-Helia

16.11.2017

**Author(s)**
Vivien Holecz

**Degree programme**
Business Information Technology

**Report/thesis title**

Big Data solutions for market intelligence for a B2B company

**Number of pages and appendix pages**
50 + 15

Being part of the Big Data – Big Business project of HAAGA-Helia, University of Eastern Finland, VTT Technological Research Center and a handful of companies, the aim of the report was to create suggestions for improving one of the participating firms' market information collection. The objectives were complex. In addition to gaining better understanding of Big Data and relevant technological solutions, the target company's development areas were also researched. Furthermore, the feasibility of a custom solution was investigated too.

The research was conducted as constructive research. The three different parts used separate methodologies. First, a questionnaire was conducted to learn about the company's practices and needs, while also examining the current market intelligence solution they are using. Second, a market research was carried out, aiming at finding relevant open source and commercial software in the areas of lead generation, social media and competitor intelligence. Third, a proof of concept was created to gain experiences and insights about open source Big Data applications and offer suggestions related to acquiring custom built or commercial market intelligence solutions.

The research revealed, that the most promising improvements can be reached in the area of lead generation, considering both commercial and open source software. The findings of the proof of concept are two-fold: it showed that a whole application can be built exclusively from open source tools, such as Python libraries and Apache Spark, to extract information from Twitter, a freely available data source. Furthermore, twitter proved to be interesting, because there is relevant industry data shared there.

The report concludes that there is business potential in seeking out Big Data solutions to solve the problem of collecting timely market insights. However, it is important to keep in mind that it takes a lot of effort to extract relevant information due to the complex challenges of analysing vast amounts of unstructured data.

**Keywords**
Big Data, open source, market intelligence, text mining, social media

## Table of contents

## Glossary of terms

**bag-of-words:** a text analysis approach where the document is treated as a collection of tokens in no particular order during the analysis

**Big Data:** a term referring to the phenomenon of exponentially growing volumes of data and the methods to handle them

**Business Intelligence (BI):** software that summarizes structured business data to report on past events

**case folding:** a practice of converting the text to lowercase during pre-processing for text analysis. The intention is to decrease the number of different words.

**categorical data:** a data type where the values represent a belonging to a group or category

**character encoding:** a character map that contains the key to decode the stored bytes as characters

**classification:** one of the predictive analytics types used to find similarities between elements of a dataset and group together similar ones. The groups are formulated based on a manually tagged test dataset.

**clustering:** one of the predictive analytics types used to find similarities between elements of a dataset and group together similar ones. The groups are formulated automatically by finding naturally occurring similarities.

**continuous data:** a numerical data type which takes any number as a value

**community cloud:** cloud deployment method where the location of the servers is a private location known by the customers who are a group of companies that trust each other

**data analytics:** a computational method which aims at finding patterns in historical data. The drivers are business problems and decisions.

**data mining:** same as data analytics

**data science:** same as data analytics

**descriptive analytics:** analytics type consisting of aggregates of operational business data, often covered by a Business Intelligence system

**diagnostic analytics:** analytics type investigating the reasons behind the findings of descriptive analytics

**DIKW pyramid:** a model to explain the meaning of data as the basis of a hierarchy consisting of data, information, knowledge and wisdom

**discrete data:** a type of numerical data which takes only specific numbers as values

**ETL (Export, Transform and Load):** the process traditionally involved in creating Business Intelligence systems. The data is first exported from the source, then transformed into the desired format, and finally, loaded into the data warehouse for the end users to see.

**external data:** data originating from outside of the company

**hybrid cloud:** cloud deployment method which involves using both public and private cloud servers depending on the sensitivity of the data.

**IaaS (Infrastructure as a Service):** one of the cloud service types. Maintenance responsibilities are shared between service provider and customer. The service provider's responsibilities are limited to maintaining the physical infrastructure and the virtualization, and the customer takes care of the installation of an operating system and everything beyond.

**internal data:** data originating from within the company

**lemmatization:** reducing all the words to their dictionary base to eliminate the differentiation of the same word with different endings, a step of pre-processing for text analysis

**Magic Quadrant:** a graphical representation of IT market segments by Gartner categorizing the players into one of four groups: leaders, challengers, visionaries and niche players

**Market Intelligence:** software that collects and digests data to provide market related information for decision-makers

**middleware:** software meant to handle the communication between different systems. In parallelization, it is responsible for tasks related to the management of parallelization and maintains a level of replication for fault tolerance

**modelling:** part of predictive analytics. It means the creation of a model with the purpose of gaining understanding from it about the entity it is derived from

**nominal data:** a type of categorical data that does not have a natural order of the categories

**numerical data:** a data type which takes numbers as values

**ordinal data:** a type of categorical data that has a natural order of the categories

**PaaS (Platform as a Service):** one of the cloud service types. Maintenance responsibilities are shared between service provider and customer. Most tasks are taken care of by the service provider, only the installation of applications and uploading of data are tasks of the customer.

**parallel processing:** processing multiple things at the same time in a distributed system. To manage the concurrency of the processes, middleware is needed.

**Part-Of-Speech tagging (POS tagging):** adding the acronym of the part of speech a word represents to the end of the word, a step of pre-processing for text analysis

**predictive analytics:** pattern discovery in historical data in order to find out about interdependencies and make fairly reliable predictions

**pre-processing:** part of text analysis, prerequisite of text mining. The purpose is transforming the natural text into a form that a predictive algorithm can be applied on it.

**prescriptive analytics:** optimizing business outcomes based on the findings of predictive analytics, determining what options are the most beneficial

**private cloud:** cloud deployment method where the location of the servers is a private location known by the customer

**public cloud:** cloud deployment method which enables changes in the location of the used hardware, because the service provider allocates resources based on customers' demand.

**reputation management:** the efforts spent by a company to measure and influence their reputation in a positive way

**SaaS (Software as a Service):** one of the cloud service types. The system is completely managed by the service provider, and the cloud service customer is only a user of the services without any IT administration tasks.

**semi-structured data:** data that has some kind of inherent structure that can be understood by a computer, but it is not as well-defined as a relational database

**sentiment analysis:** a form of sophisticated text analysis aiming to determine the emotional impression as negative, positive or neutral

**stemming:** removing anything from the end of the words that resembles a suffix to eliminate the differentiation of the same word with different endings, a step of pre-processing for text analysis

**stop words:** words that occur very often in the text but don't have significant meaning, removing them is a step of pre-processing for text analysis

**structured data:** data that has a clearly defined structure, like in a relational database

**term frequency:** calculations of how many times each term appeared in a document, one of the simplest forms of text analysis

**text analysis:** a special type of analytics aiming at extracting valuable information form natural human language

**text mining:** text analysis involving algorithms used in predictive analytics to discover patterns

three V's

**tokenization:** breaking up sentences into words during text analysis, so words become the smallest unit of operations, a step of pre-processing for text analysis

**topic recognition:** a form of more sophisticated text analysis aiming at identifying underlying topics in the text

**unstructured data:** data that does not have inherent structure that can be understood by a computer

# 1 Introduction

The motive behind the first data collection was already to provide information for better decision making when writing was invented. Recent years of technological development have resulted in opportunities that have never been possible earlier in history. With the rise of cloud computing, social media, and devices that keep us connected at all times, new ways of gathering and utilizing data and new sources of valuable information have been evolving at high speed.

At the same time, the consumer economy and global competition poses great demands on businesses. Competitive advantages diminish shortly unless a non-stop effort is made to seek out new opportunities to keep ahead of competition and up-to-date about customer needs.

These trends have lead to the emergence of a new phenomenon. On one side, there are all the technological advances in processing power, the social changes triggered by innovative products, and the appearance of open source software (Minelli, Chambers & Dhiraj 2012, 1-2). On the other side, there is a pressing need for a competitive edge that companies face. The result is the new kind of data analysis: Big Data.

This thesis project is part of the Big Data – Big Business project that aims at finding new business opportunities for Finnish companies by utilizing Big Data solutions[1]. The focus of this report is on one of the participating companies of this project. This company is operating on the business-to-business market in the construction and energy industries, with majority of sales abroad.

The research question is related to the challenges that the target company is facing in the area of market intelligence. The initial interests are to increase the quality and quantity of international, market related information while also gaining understanding of Big Data to become more informed customers. There are three main areas where more market information is needed: generation of new leads, customer feedback and opinions and competitor activity.

[1] http://www.haaga-helia.fi/fi/big%20data

The scope is complex. To narrow it down, initial exploration is needed, so the first, more general research question is broken up into parts and based on the findings the investigation continues with a more concrete case. In the first stage, the three interest areas of the company are examined. The goal at this point is to choose the one area that is estimated to have the highest return of actionable results. In the second stage, the chosen topic is explored in more detail.

The initial research question is: *"How can Big Data help a B2B company to get more and better market information?"*

On a general level, the type of the research is constructive research, since the main goal is to provide suggestions for a company's business problems. First, the company needs are clarified through the analysis of a questionnaire and the examination of their currently used tool. Second, available software are investigated in order to find possible solutions to all three of the initial research areas. The one which promises the biggest improvements is chosen at the end. Finally, the research question is refined and an open source demo solution is built aiming to fulfil the need specified in the refined research question. Besides trying to provide actual market information, the goal of the proof of concept is to collect experience about the challenges related to Big Data analytics.

Due to the complexity of the research, the structure of the report needs clarification. First, the Overview of the Big Data phenomenon lays down the theoretical background for business understanding. The chapter continues with the explanation of Big Data and analytics, with special focus on unstructured, external, especially textual data due the company case. In addition, the role of cloud computing in data science is clarified. Discussion of the theoretical background ends with the definition of market intelligence to provide a basis for finding and building solutions. After the theory the methods used during each part of the research are explained. The results are broken up into three parts: the company case, market intelligence software research and the proof of concept. Each part contains the related results and the suggestions based on them. Finally, the individual analyses are wrapped up in the Suggestions section to give a global answer to the initial research question and discuss the value of the results. The conclusion closes the report by summing up the findings and suggesting next steps.

The author wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

# 2 Overview of the Big Data phenomenon

In order to understand what Big Data is about, at least two aspects need to be considered: the demand for better data utilization and the supply of solutions enabled by scientific advancement. This chapter begins with discussing the needs that can be fulfilled by Big Data. The second section goes into detail about the technical side, including the definition of Big Data itself and the related methodologies. Besides the basics of analytics, especially predictive and text analytics, the role of cloud computing is also covered. Finally, a main application area, market intelligence is clarified.

## 2.1 Business needs

Recent trends in business have brought significant changes to many areas, which resulted in information playing a key role in success. According to Loshin (2013, 27), the competitiveness of a business depends greatly on having practical information available as quickly as possible for decision making. This section will discuss **business trends** that shape the demand for analytics and Big Data.

Due to **globalization**, markets are becoming more and more complicated and it is very difficult to act in a collection of different local markets without competent information sources. According to executives, the local competitors are most of the time more informed about the environment and market demand than a multinational company operating in the country (Dewhurst, Harris & Heywood, 2012). Competition is escalating, since buyers can now purchase from around the world, and this is true not only for consumers but business buyers as well (Cavusgil, Knight & Riesenberger 2014, 61). There are new risks involved, which are not entirely understood, and together with an overly standardized, rigid risk assessment process, the lack of input for decision making can result in lost opportunities (Håkansson & Nelke 2015, chapter 1; Dewhurst, Harris & Heywood, 2012)

**Economical uncertainty** of recent years has had a major impact on both buying and selling – marketing departments are hard pressed to prove their effectiveness by quantifying return on investment numbers, while the main goal of purchasing is to maximize spending efficiency. The unpredictable market has made buyers more averse to risks and they often chose to keep traditional approaches and tools instead of investing in radical projects. Decisions are made more often by a committee than individuals. (Roetzer 2014, 17; Eades & Sullivan 2014, 31-40)

**Technological innovations** have also been shaping business in several ways. First, the **explosion of data** is proving to be both a problem and an opportunity in itself. According to IBM, humankind produces 2,5 quintillion bytes of data every day. Furthermore, the rate of growth is increasing too – the sum of all existing data is expected to double every two years (EMC, 2014). Not just the amount and growth rate of data are remarkable, but also the variety of formats and data types (Loshin 2013, 5). All of this results in "information paralysis"(Hedin & Irmeli Varnaas 2014, 8), meaning that decision -making is flooded with way too much input, which is extensive, unstructured and noisy, and, therefore, meaningless.

Another aspect of technological advancement is the emergence of special new data sources, like websites and **social media**. They are some of the fastest growing types. Every day about 200 million tweets are posted on twitter (TwitterEng, 2011), 4 petabytes of data uploaded to Facebook (Bronson & Wiener, 2014), and, as of 2014, 70 million photos shared on Instagram (Systrom, 2014). Social platforms offer great benefits to buyers. They have the ability to share and gather information about the products they are interested in.

From companies' perspective, this shift in buying behaviour is a great challenge due to the changes in the role of sales and its effect on reputation management. Reputation is a key asset of a company, since it plays an important role in customer decision-making (Riel & Fombrun 2007, 48). For this reason, corporate **reputation management** became a practice, which means the efforts spent to measure and influence reputation in a positive way (Doorley & Garcia 2015, 20). Before the rise of social media, the main three information sources that formed the reputation of a company were first-hand experiences, experiences of relatives and acquaintances and mass media (Riel & Fombrun 2007, 46). Social media takes everybody's first-hand experiences and makes them publicly available, as if it was in mass media. To businesses, this means full time exposure, the possibility of losing their reputation in any moment.

The abundance of online information results in buyers going through the **purchasing journey** by themselves, and only contacting a sales professional when their decision is made (Roetzer 2014, 10). For this reason, marketing has an increased role in the success of sales by producing online presence that provides value to possible customers and support for existing ones. At the same time, marketing doesn't have the means to control messages and company image as they used to due to the aforementioned effect of social media. Under the new circumstances, salespeople's personal interactions with possible

buyers online is a way of building and maintaining reputation – traditionally a task of marketing. The bottom line: marketing and sales tasks are becoming harder to distinguish, and both are pressed for results in a fast changing environment. (Roetzer 2014, 11; Eades & Sullivan 2014, 30-31)

One more aspect of the effect of technology on business is the **fast pace of innovation**. The rate of advancement leads to a situation where the opportunities provided by IT companies can not be fully utilized, because businesses can not keep up with the changes. There are a plethora of different solutions aimed at making marketing more efficient – according to a current summary of all marketing software on the IT market (Brinker 10.5.2017), there are 4 891 companies offering 5 381 solutions in 49 different categories. At the same time, businesses are using legacy systems, CMOs are not entirely prepared to dive into the software industry to this depth and users are tired of learning how to use yet another tool that will pass in a few years (Roetzer 2014, 6-9, 23). The risk avoidance is true for the software buyers as well – often the legacy system remains because of compatibility and integration difficulties, company politics and financial issues. However, there are also promising trends, like cloud computing and Software as a Service solutions that help to remove the commitment and financial obstacles, and the emergence of "marketing middleware", which is meant to make the communication between many different systems easier (Roetzer 2014, 107-111).

## 2.2 Big Data solutions

In the previous section current business trends were discussed to identify the demand for information systems and to clarify information needs. The following chapter focuses on the technology side, showing the possibilities that the Big Data phenomenon carries. The chapter starts with the definition and discussion of Big Data, data, and analytics. The role of cloud computing in the Big Data domain is also explained. The section ends with the definition of IT market terms like market intelligence, customer intelligence and competitor intelligence.

### Big Data

First and foremost the idea "**Big Data**" needs some explanation. The name comes from the high volume of data, but there are other aspects that differentiate the era of Big Data from the previous decades. Often mentioned as the **three V's**, Big Data has unprecedented big Volume and Variety, and high Velocity (Minelli & al. 2012, 37).

However, over time the list of adjectives starting with a v to describe the phenomenon have grown uncontrollably, the longest list containing as many as 42 “V” words (Shafer, 2017). The most important ones can be seen on Figure 1 (Shafer, 2017).

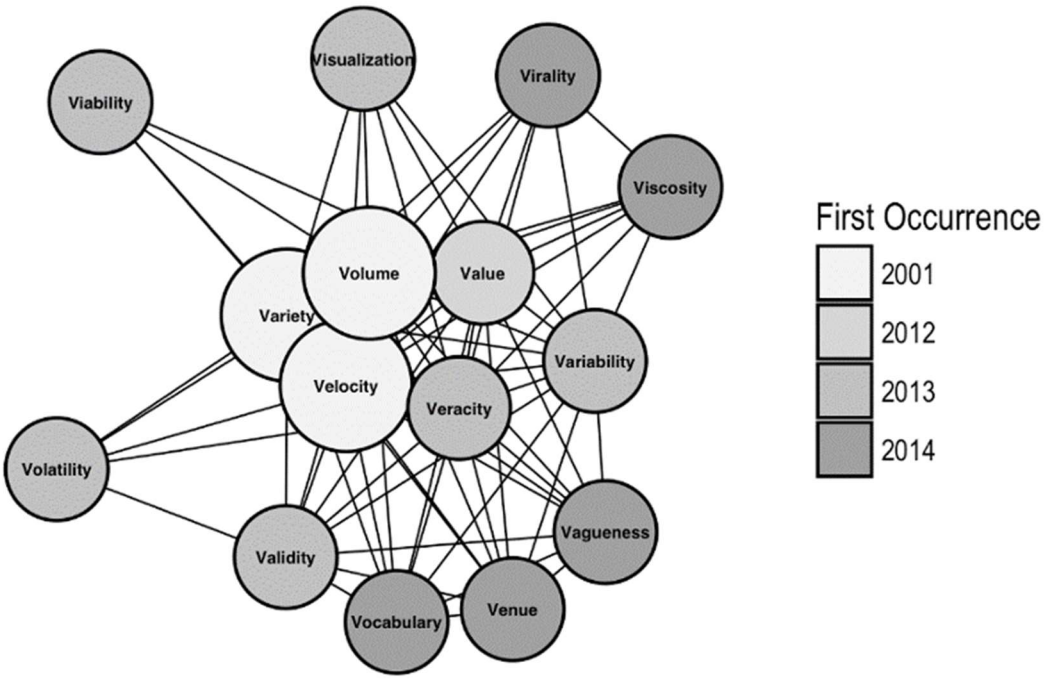



Figure 1: The V’s of Big Data (Shafer, 2017)

Starting with the original three characteristics, **Volume** seems to be the most self-explanatory in the group. Figure 2 (EMC 2015, 16) shows that as time passes, more and more data is created worldwide.

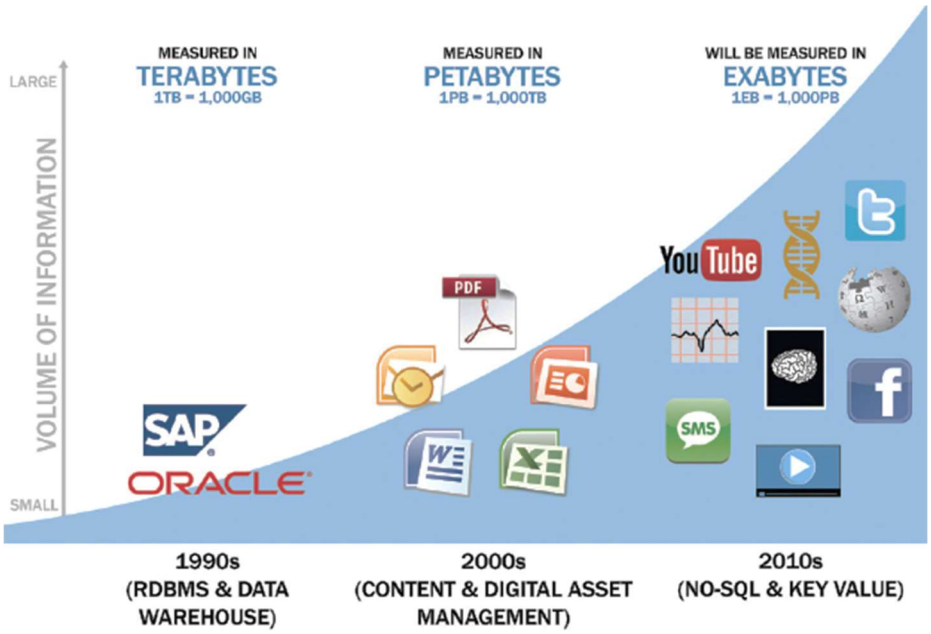



Figure 2: Changes in data volumes over time (EMC 2015, 16)

This poses a great demand to invent techniques and to build infrastructures that can handle the huge load (EMC 2015, 16). It also demonstrates that the answer to the question 'how big is "Big"?' is changing over time. Furthermore, the relative size depends on the company as well – some dataset might be huge and unmanageable for one company, while posing no challenge for another (Devlin 2013, 167).

**Velocity** marks the importance of the timeliness of the analysis. Real-time systems need answers to questions instantly, and this makes traditional ETL and data warehousing practices insufficient. **Variety** stands for all the different formats that data comes in, which requires new ways of processing. (Liebowitz 2013, chapter 5)

The additional V's all demonstrate other challenges involved in extracting value from unconventional datasets. **Veracity, Variability, Volatility** and **Viscosity** refer to the problem that Big Data often consists of data that's accuracy is not assured, there are irregularities within the dataset, the dataflow is uneven and navigating between the components is not straightforward. Processing happens in different **Venues,** and **Virality** marks the issues related to spreading the data among the users. **Visualizations** are needed to make sense of the findings. The driver of the whole process is to create **Value**, while the **Validity** of the findings or the **Viability** of utilizing them might be questionable. There is an extensive **Vocabulary** to describe the new ways of analyses and problems, but **Vagueness** of terms, techniques and outcomes are still common. (Shafer, 2017)

Seeing the lack of a clear, unambiguous definition, it is not a surprise that there are several schools of thought about the differences between **Business Intelligence (BI) and Big Data**. Jain, Jayaraman, Sharma (2014, 36) argue that Big Data is an ETL (Extract, Transform and Load) problem, down playing the novelty of the term and the phenomenon. On the other hand, the EMC (2015, 13) emphasizes the differences between Business Intelligence and data science. They draw the line between the two based on the focus areas, the methods used, and the data types involved. According to them, data science investigates hypotheses through modelling based on structured or unstructured data, while BI simply reports on past events by summarizing structured data. Devlin (2013, 166-167) sheds light on the fact that first concerns about Big Data came from sciences, not business, making Big Data a broader term than BI. Minelli & al. (2012, 29) predicts that Big Data will have a great impact on data warehousing and analytics as a whole.

As mentioned in the introduction, one of the key elements of Big Data technologies are open source software. This expression stands for software that's source code is freely

available to anyone. They usually come with a licence that enables free usage, modification and redistribution, with the restriction that the distributed product is launched under the same licence (GNU licence). The freedom of usage, modification and distribution of these software has accelerated innovation to find solutions to the Big Data problems. (Minelli & al. 2012, 67-68)

**Data**

Big Data can not be discussed without clarifying what is data and the many forms it can take. A commonly used model to explain the meaning of data is the **hierarchy of data, information, knowledge and wisdom, or DIKW pyramid** as seen on Figure 3. (Bocij, Greasley & Hickie 2015, 6-7). Rowley (2007) points it out, that it is not a clearly defined model, and the definitions and explanations differ from textbook to textbook.

In this report these terms are going to be defined following the book Business Information Systems (Bocij & al. 2015, 6-7). The two extremes, data and wisdom, are opposites in the sense that data are abundant and meaningless, while wisdom is very rare and valuable. **Data** are observations that are written down. They can come from nature, like a record of temperatures over time, or can be created by computer systems as well. A collection of unrelated, irrelevant data is called “noise”.

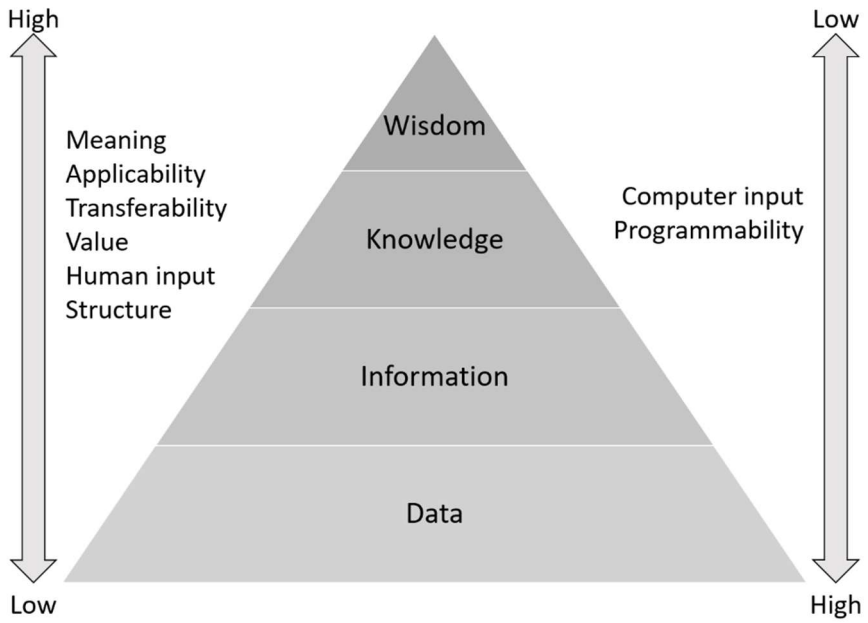



Figure 3: DIKW pyramid (Bocij, Greasley & Hickie 2015, 7)

A more meaningful entity is **information**, which carries the advantage that it is understandable by a human, because it has relevance. The relevance comes from the processing of data with a goal in mind. Information has an impact on actions – actually, influencing decisions is its core purpose. (Bocij & al. 2015, 7-8)

**Knowledge** is the sum of information and the skills and intuition of the person who interprets it. **Wisdom** has the added value of experience, which makes it possible to apply the knowledge to various situations. (Bocij & al. 2015, 15-16)

The **goal of analytics** is to extract actionable information from data. The processes involved in the transformation of data include categorizing, grouping, filtering and sorting. Furthermore, new data points can be created by summarizing existing ones or using them as the basis of calculations. (Bocij & al. 2015, 9)

Data comes in various types and forms. Figure 4 shows Cuesta's (2013, 10) **data categories**. First, data can be categorized as numerical or categorical, based on the values it takes. If the values are numbers, it is numerical, if they represent a belonging to a group or category, they are categorical. An example for numerical data is the temperature on a given day, while the outlook (sunny, partly cloudy, cloudy, rainy, snowy) is categorical data. These two categories can be split further. Within categorical data, there are two options: nominal data represents groupings that have no natural order, while ordinal data has values that can be ordered. Gender is nominal data, while age groups are ordinal. The difference between the two numerical categories is, that discrete data can take a value from a set of values, while continuous data can take any value within an interval.

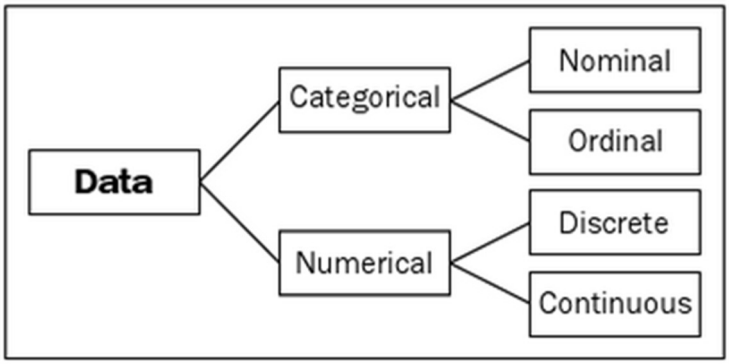



Figure 4: Data categories (Cuesta 2013, 10.)

Cuesta (2013, 10) mentions further categorization based on the structure of the data as well: structured and unstructured.

Marr (2015, 59) describes **structured data** as data in relational databases, or at least in a row-column format. He adds, that this type of data is the one that has a longer history of analysis methods and practices. In comparison, he states that **unstructured and semi-structured data** are the fashionable newcomers in the analytics field, that couldn't be analysed by computers until recent years because of their lack of structure. New Big Data technologies are created with the intention of utilizing data like text documents, emails, sound and video recordings, images, etc. (Marr 2015, 59-79)

Marr (2015, 59-79) continues with further classification based on the source of data – data might come from within the organization or from outside. He describes data types based on these two ways of categorization as follows:

- structured, internal data: ERP, CRM, financial
- unstructured, internal: e-mails, customer service calls, documentations
- structured, external: macroeconomic statistical data
- unstructured, external: social media posts, comments, webpages

**Natural language texts** are a special domain of Big Data analytics with their own unique challenges. Text data is considered unstructured because the computer can not understand them by default, it can only represent it for a human to read and make conclusions.

First of all, text data is stored in bytes just like any other data, and information is needed about the **encoding** to interpret the stored code as characters. Character encodings are character maps that contain the key how to decode the stored bytes as characters. There are many different encodings, Unicode-UTF-8 being the most universal of all. If the text is decoded with a wrong encoding information, any character that is outside of the English alphabet will be displayed incorrectly. (Ishida 16.4.2015)

The **language aspect** also complicates interpretation for a computer. Languages have grammatical logic, but since humans are not very consistent and logical creatures, grammar has a lot of exceptions and irregularities in most languages. Another issue is that there are many different languages on the planet, which means that a universal solution for interpretation does not work. Furthermore, even within the same language there are dialects and slang that makes it even more confusing to separate words that have different meaning and group together the ones that have the same meaning. This is especially a problem in social media, where abbreviations, type-o's and slang words occur very often.

Considering social media, despite being textual data, there is actually some structure due to the data format being JSON. For this reason hashtags, links, usernames and metadata in the tweets can be easily extracted during processing.

**Analytics**

Data analytics has many synonyms or terms with very close meanings, like data science, data mining and knowledge discovery (Baesens 2014, 7-8; Abbott 2014, 3; Jain, Jayaraman & Sharma 2014, 31). According to Baesens (2014, 7), these terms are practically interchangeable, or mean the same thing in most cases. Analysing data is a task that appears in many scientific disciplines and new inventions and good practices emerge from many different areas (Dean 2014, 56). Figure 5 (Dean 2014, 56) shows how statistics, machine learning, database management and other areas overlap with each other.

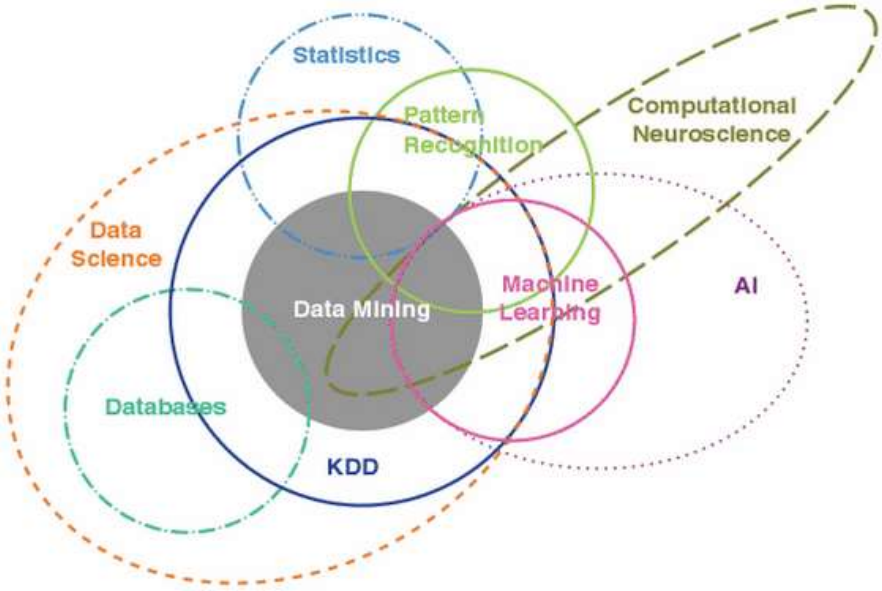



Figure 5: Data mining and related disciplines (Dean 2014, 56)

Baesens (2014, 7-8) also points out that analytics lie in the intersection of several professional and scientific areas, like machine learning, statistics, biology and computer technology.

This report uses the definition of data analytics from Abbott (2014, 3). He defines analytics with the following statements:

- a computational method
- the drivers are business problems and decisions
- the input is historical data
- the aim is to find patterns

Jain & al (2014, 19-20) emphasize the importance of connections with the company's strategy, and the aspect of usability of the insights – if the findings don't trigger actions, they are useless.

Figure 6 shows the different types of analytics and how they compare with each other (Dearborn 2015, 45). The graph suggests, that there is a hierarchy in complexity and usefulness between the **four types of analytics**, descriptive being the least complex and prescriptive the most intricate and useful in guiding business decisions. Dearborn sums up these four stages by specifying what questions they provide answers for. The first two questions, representing descriptive and diagnostic analytics, are about past events:" what happened and why?". Predictive analytics focuses on the future, trying to find out what will happen. Finally, prescriptive analytics provides alternative answers to the question "what to do about it?".

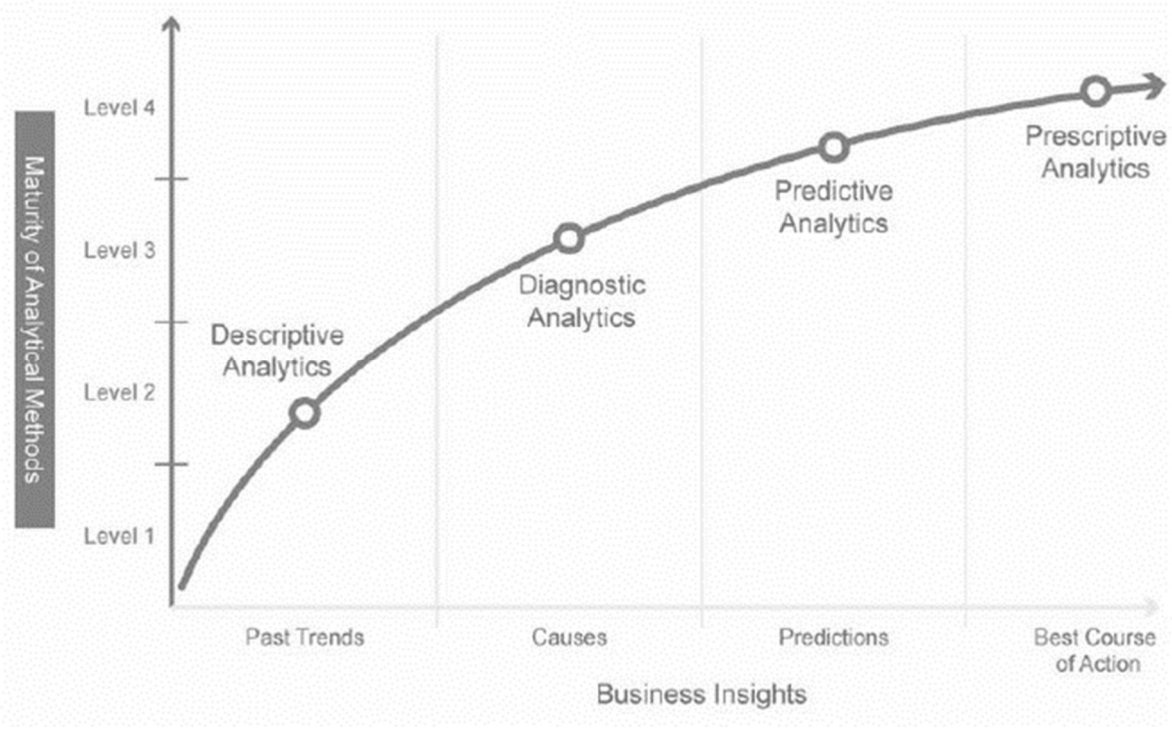



Figure 6: The four types of analytics (Dearborn 2015, 45)

Liebowitz (2014) classifies analytics the same way, except he skips the diagnostic step and discusses only the **descriptive, predictive and prescriptive types**. In his understanding, descriptive analytics are the type of reporting that most companies are familiar with and it consists of aggregates of their operational data. It is mostly related to money and performance and this type of analysis is often covered by a Business Intelligence system. He continues with predictive analytics – this is where the pattern discovery takes place. Patterns in historical data are researched in order to find out interdependencies and

make fairly reliable predictions. Prescriptive analytics takes these findings to the next level by optimizing business outcomes based on the predictions, determining what options are the most beneficial.

Baesens (2014, 35) states that the goal of **predictive analytics** is to predict a characteristic or attribute that has value to the business or research. As mentioned earlier, predictive analytics relies on pattern discovery, which is conducted by building a model based on data. Forte (2015, chapter 1) defines a **model** as follows: “a model is a representation of a state, process, or system that we want to understand and reason about”. This means that first, a model is a representation, a kind of abstraction of something, second, the purpose of it is to gain understanding from it about the entity it is derived from. Abbott (2014, 3) emphasizes, that the difference between traditional statistics and predictive analytics is that while statisticians choose a model to describe the data, predictive analytics takes the data as a base and derives the model from it. This is done automatically by algorithms, which makes the process much faster than human work.

Finding correlations in the dataset is the way of finding patterns. In statistics, it is a common phrase that “**correlation does not equal causation**”. However, in predictive analytics the difference between the two is not taken so seriously. First, because causation is difficult to prove or define, second, because a proven correlation is often enough in business situations to derive value from the findings. (Siegel 2013, 89-93)

On the other hand, Harford (28.3.2014) points out that the pitfalls of statistics are not necessarily avoided just by increased volumes of data or using algorithms to check every possible connection between attributes. He emphasizes the importance of attention to underlying causes, since without understanding the reasons behind the discovered patterns, significant changes in the environment might go unnoticed. These overlooked factors might make the whole prediction inaccurate. He brings Google Flue Trends’ failure of predicting flu outbreaks as an example.

**Text analysis**

A special type of analytics is **text analysis**. The main objective here is to extract valuable information form natural human language through processing and modelling techniques. To take it further, algorithms used in predictive analytics can be applied to pre-processed text to discover patterns, in which case the practice is called **text mining**. (EMC 2015, 256)

There are several **problems** around trying to make a machine understand human text, many of them related to **inconsistencies in languages** themselves. Occurrences of homonyms (different meanings of a single word, like “bear”), synonyms (different words that mean the same thing), the problems of polysemy (related but slightly different meanings of words) and hyponymy (hierarchical relationships between words, like bear - mammal - animal) make automatic interpretations difficult. Depending on the origin, the text might or might not have some kind of **structure**, and if it does, it might or might not be useful in analysis. For example, in case of e-mails, the sender, receiver and topic information can be useful in the analysis, but HTML tags are mostly discarded as unimportant elements. From a modelling point of view, the fact that natural text has a great variety of different words means that there is a huge **number of dimensions** in the model, while a great proportion of them appear rarely. This is undesirable since it makes the model overly complex without adding much value to the analysis. (EMC 2015, 256; Abbott 2014, 331)

Abbott (2014, 332) gives examples for **practical applications** of text analysis: finding information using a search utility, discovering similarities between documents through clustering, automatically categorizing documents through classification, finding out the overall emotional impression of documents, getting out useful information from text. Clustering and classification are predictive analytics types used to find similarities between elements of a dataset and group together similar ones, in this context they are types of text mining.

All these applications need **pre-processing**, which includes many different tasks. Not every analysis process will need every method described below, each situation has to be judged individually depending on the source text and the purpose of analysis (EMC 2015, 258-271).

In order to be able to work with the text, it needs to be tokenized. **Tokenization** means the break up of sentences into words, so words become the smallest unit of operations. The significance of punctuation, entities with names that consist of more than one word, and contractions (I’ll, won’t, etc.) need to be considered before splitting the text into tokens. Without attention to these elements, “I’ll” will be broken up into tokens of “I” and “ll”, or the name of “Wall Street” into “Wall” and “Street“, resulting in important meanings being lost. (EMC 2015, 264; Abbott 2014, 336)

The **multidimensionality problem** can be handled by decreasing the number of unique words. This can be achieved by methods like lemmatization, stemming, case folding, POS tagging, stop word filtering and restricting the analysis to the most frequent terms.

**Lemmatization and stemming** are similar in a way, because they both try to eliminate the differentiation of the same words with different endings. For example, very often it is logical to interpret learn, learns, learned, learning, etc. as the same thing. Lemmatization achieves this by reducing all the words to their dictionary base, while stemming simply removes the endings of words. Lemmatization is a more complicated task than stemming, because it needs to find the base form for each word. Stemming just removes anything from the end of the words that resembles a suffix. Both are dependent on language, because grammar rules are very different in different languages. For this reason each language have to have their own stemmer and lemmatizer. Stemmers differ based on how harsh they are on the words when removing endings – Abbott and Komp illustrate this by showing the example that “generate” can be reduced to “generat”, “gener” or “gen”. The appropriate level of stemming depends on the purpose of the analysis. (EMC 2015, 258; Abbott 2014, 337)

**Case folding** is the practice of converting the text to lowercase. While it is a straight forward task to execute, capital letters might hold meaning that are not supposed to be discarded. Acronyms and company names might turn into common words when written in lowercase, which can lead to confusion and skew the model. (EMC 2015, 264)

**Part of speech (POS) tagging** means adding the acronym of the part of speech a word represents to the end of the word. This can help differentiate between homonyms that belong to a different part of speech (for example “tear” as a noun or as a verb). Furthermore, the number of dimensions in the model can be reduced significantly, if the analysis is focused only on one part of speech group, like nouns, verbs or adjectives. Since it is dependent on the sentence structure, it needs to be conducted before the tokenization. (EMC 2015, 258; Abbott 2014, 334)

**Stop words** are words that occur very often in the text but don’t have significant meaning, like articles and other short words demanded by grammar. They don’t only bloat the model without adding value, but disturb the most basic analysis methods as well, like word frequencies (they are discussed later). For this reason they need to be filtered out. (EMC 2015, 270; Abbott 2014, 336)

A very simplistic approach to analysis is **bag-of-words**. It means that the document is treated as a collection of tokens in no particular order during the analysis. This has some drawbacks, since the order of the words holds significant meaning, but it is a good starting point and in some cases even sufficient for the whole analysis. (EMC 2015, 265)

The first and most basic analysis method is **term frequency**. It means calculating how many times a certain term appeared in a document and comparing the documents based on the most frequent words in them. As mentioned earlier, stop words need to be removed, otherwise the first few dozen words will be stop words in all documents, which doesn't provide any insight into the contents. There are also words that appear frequently in all documents but are not stop words. They might be used often for example because of the domain or the topic. A way to handle them is calculating **inverse document frequency**, where the term frequency in the whole collection of documents is looked at. It is inverse, because the fewer documents the term appears in, the more specific the term is for the document it is found in, which makes it more valuable for comparisons. To get the full picture, i.e. the terms that appear often in one document but rarely in others, the **term frequency is multiplied by the inverse document frequency**. This is a proper measure for modelling, so algorithms used for predictive analytics can be used with it. (EMC 2015, 271-274; Abbott 2014, 342-345)

More sophisticated techniques are used for **topic recognition** and **sentiment analysis**. The former means identifying underlying topics in the text, while the latter aims to determine the emotional impression as negative, positive or neutral, and its typical application domain is analyzing reviews and comments online. (EMC 2015, 274-283)

**Cloud computing**

The first connection between Big Data and cloud computing is that both require **parallel processing**. The processing time would be extremely long for large datasets if they were to be processed on one machine. Cloud applications also often need great processing power while the user is not aware of the hardware background of the service they are using. The idea of parallel processing has been around since the early days of computing. While processing multiple things at the same time was a tempting capability to reach, there were several challenges in the way. When trying to compute interrelated things at the same time, it is difficult to ensure that correct information is delivered at the right time to the CPUs about the processes computed elsewhere. Locking systems need to be implemented for preventing the different processes to interrupt or corrupt each other. It is crucial that the same outcome is reached every time the execution takes place, no matter the order or distribution of the tasks. (Marinescu 2013, 21-24)

These problems make it necessary for distributed systems to have a component called **middleware** which is responsible for tasks related to management of parallelization. These tasks include making remote systems available as if they were local while conceal-

ing their actual location, possible other users and system failures. Furthermore, middleware ensures that there is a level of replication for fault tolerance and it is possible to easily configure and scale the system based on demands. (Marinescu 2013, 27)

The other connection between cloud computing and Big Data is that **analytics often utilize cloud environments** for various reasons. As companies start to involve more and more data types in their analytics, many of them originate from the network, like Internet of Things or social media data. It is a natural choice to process the data where it originates from. A cloud environment is also useful in case of testing or proving a concept – if it's successful, the decision can be made if it should be moved in-premises or to remain in the cloud. The relative volume of data also plays a big role in the matter – companies find that a few years ago their infrastructure was sufficient, but recently it has exploded to such a high volume that infrastructure investments can't keep up with the growing demand. The elasticity of a cloud based system is very useful in this situation. For the same reason it is also beneficial for new adapters of Big Data, since they can try analytics before making big investment decisions. Analytics is usually a team effort and cloud solutions are naturally suitable for collaboration. (Asay 6.4.2015; Guiliano 25.5.2016)

There are three main **cloud service types** described, Software-as-a-Service, Platform-as-a-Service and Infrastructure-as-a-Service. The level of service is illustrated on Figure 7 (Vacca 2016, chapter 2.2). The two extremes are the in-premises information systems administered by the company and the Software-as-a-Service cloud solution. In the former case, all tasks related to the maintenance of the IT system belongs to the using company. This offers a high level of security but also high investment and maintenance cost. On the other end of scale, SaaS is completely managed by the service provider, and the cloud service customer is only a user of the services without any IT administration tasks. The intermediate solutions, IaaS (Infrastructure-as-a-Service) and PaaS (Platform-as-a-Service) are both characterized by shared responsibilities between service provider and customer. In the case of IaaS, the service provider's responsibilities are limited to maintaining the physical infrastructure and the virtualization, and the customer takes care of the installation of an operating system and everything beyond. PaaS leaves even more tasks to the service provider, and reserves only the installation of applications and uploading of data to the customer. This service is mostly utilized by software development and testing. (Vacca 2016, chapter 2.4.1; Marinescu 2013, 13; Finn & al. 2012, 7; Mell & Grance, 2011)

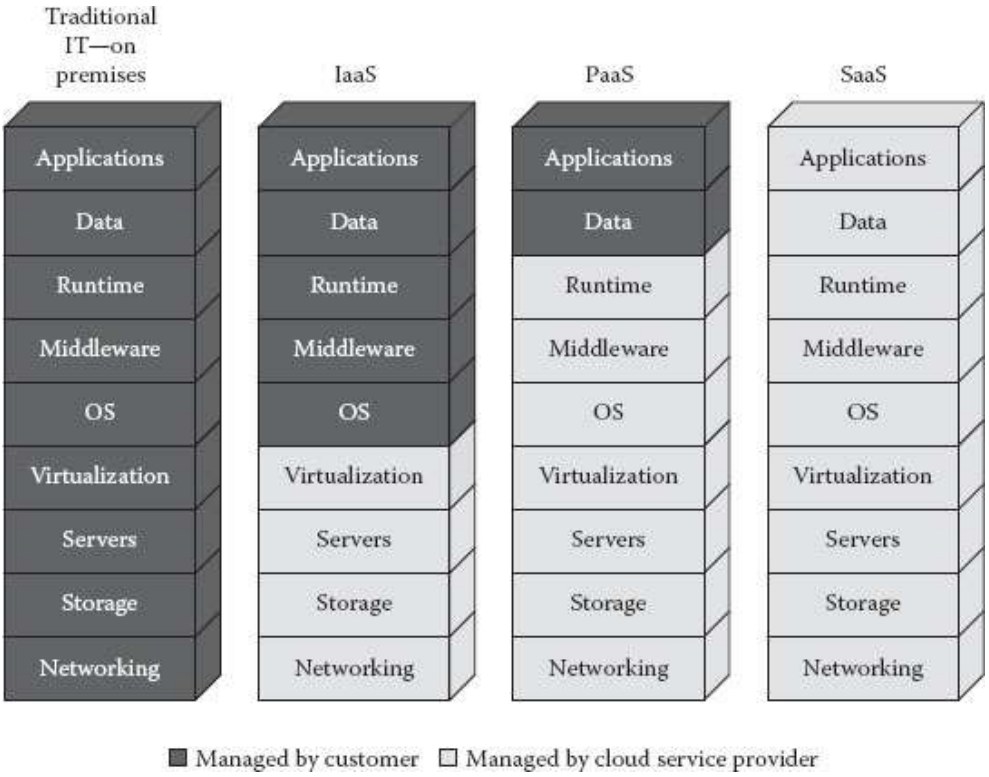



Figure 7: Cloud service types (Vacca 2016, chapter 2.2)

There are four **deployment methods** based on the location and security of the cloud service. The private cloud is the most secure, because in this case the physical infrastructure is either in-premises of the customer company or outsourced to a specific company. On the other hand, public cloud is cheaper but raises security concerns, since the location of the used hardware is not clear and changes all the time, because the service provider allocates resources based on customers' demand. The hybrid cloud is a combined solution that keeps the crucial, most sensitive data in-premises, but for the less confidential information utilizes the cost savings of the public or community cloud. In case of the community cloud, the physical infrastructure is shared by only a limited number of customers who decide to share the resources to bring down costs and also because they have some common interests. (Vacca 2016, chapter 2.5; Mell & Grance, 2011)

**Market intelligence**

Market intelligence is an application area of all the theory discussed in this chapter so far. As mentioned earlier, there is an abundance of marketing software and it is hard to grasp the actual meaning of words that describe the functionalities. For the purpose of this report, The Handbook of Market Intelligence (Hedin & al. 2014, 9) definition of **market intelligence** (MI) is taken as a basis: "As a program, MI collects information about market players and strategically relevant topics, and processes it into insights that support deci-

sion-making." They also clarify that market intelligence is different from business intelligence, but can be used as a synonym of competitive intelligence. To be more specific, **competitive intelligence** is defined in this report as systematic information collection about the competitors and is considered one of the components of market intelligence. Two other areas are considered under the umbrella of the term market intelligence by this report are lead generation and customer intelligence. **Customer intelligence** is meant to collect customer insights and feedback in a structured and automatic manner. **Lead generation** is the functionality of collecting contact information with the purpose of gaining new customers.

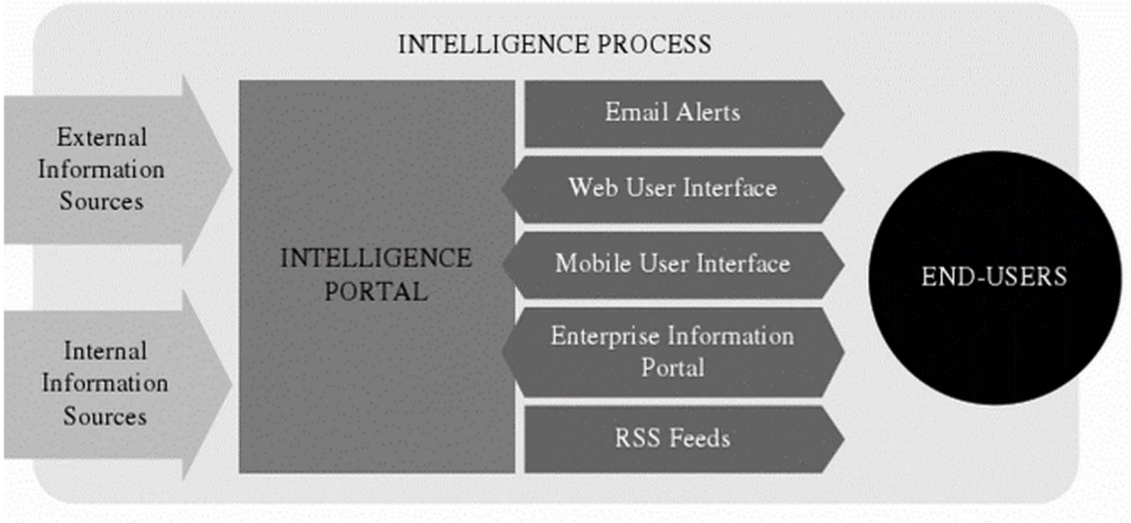



Figure 8: Intelligence portal (Hedin & al. 2014, 88)

The most important tool related to market intelligence is an **intelligence portal**. The role of an intelligence portal is seen on Figure 8. It collects and combines information from external and internal sources and makes it available for end users. The advantage is that end users can find all the market insights in one place. It offers different ways to consume the information: push and pull methods. The "push" methods include e-mail alerts and RSS feeds. Push refers to the system pushing the information on the user instead of the user searching for it. "Pull" methods are methods where the user is actively retrieving the information for example through the web or mobile user interface. (Hedin & al. 2014, 87-88)

Technically, the intelligence portal consists of a few different components. Data needs to be collected from various sources, like web crawling, manual input and other company systems. Data storage has related tasks like indexing, categorization and usage statistics. The "push" functionalities include e-mail alerts and newsletter creation and sending information to other systems. In addition, there are also "pull" functionalities, like customizable

dashboards, search and analytics tools. One of the main advantages of a website is that collaboration is easy. Collaboration functionalities include forums and commenting and networking capabilities. A natural way of providing an intelligence portal is from the cloud using the SaaS model. (Hedin & al. 2014, 91-92)

# 3 Research approach

This section clarifies the different components of the research question and the methods used for the investigation of each aspect.

## 3.1 Research question

As mentioned in the introduction, the general research question is as follows: *"How can Big Data help a B2B company to get more and better market information?"*

The first component of the research scope is the **business context**. On a general level, the domain is business-to-business. On a more specific level, it consists of the energy and construction industries. Keeping the level more general initially helps with interpreting the findings in a broader sense when discussing the outcomes of the research, but the more specific approach is needed in the proof of concept phase. To understand the needs of the company better, the following question is investigated first: *"How does the current and the desired situation look like in the company related to market information gathering?"*.

The **market information** collection problem consists of three areas as mentioned before: lead generation, customer intelligence and competitor intelligence. All three are going to be investigated in the software research to find the most interesting idea with the easiest implementation. For this reason, the findings in the software research category are going to be discussed separately based on the topic.

The element of **Big Data solutions** can be broken up into two groups: out of the box commercial solutions and open source software. The two need to be separated, because the implementation and the benefits differ greatly between them. Furthermore, the Proof of concept section needs information about the two categories for different reasons. Input about the commercial solutions is used to find an idea to implement, while the open source tool findings are utilized when designing and carrying out the implementation. To sum it up, the main purpose of the software research is to answer the question *"Which improvement area has the greatest potential for developing an open source solution?"*. To

arrive to this conclusion, the questions to investigate are *"What kind of commercial software is currently on the market that uses Big Data tools to gather market intelligence?"* and *"What open source software is available to build Big Data tools for gathering market intelligence?".*

After finding out the possibilities and choosing the most promising development area, the question to be answered in the Proof of concept section is formulated as *"How to build a market intelligence solution using only open source tools?"*. It is clarified later what type of market intelligence solution is in question.

### 3.2 Methods

The thesis project is carried out as constructive research. The definitions and guidelines for this research type are used from the book Designs, methods and practices for research of project management (Pasian 2015, 95-107). Constructive research aims at creating a solution to a problem, while also relating to the theory. The steps involved are outlined on Figure 9. The arrows indicate the difference in logic over the course of the research. In the beginning, when learning about the problem domain and creating possible solutions, deductive logic is used, which means that ideas are applied from the general level to the specific case. Afterwards, when proving the feasibility and observing the relationship and contributions to the theory, the logic turns around: generalizations are made from the specific case to a broader scale.

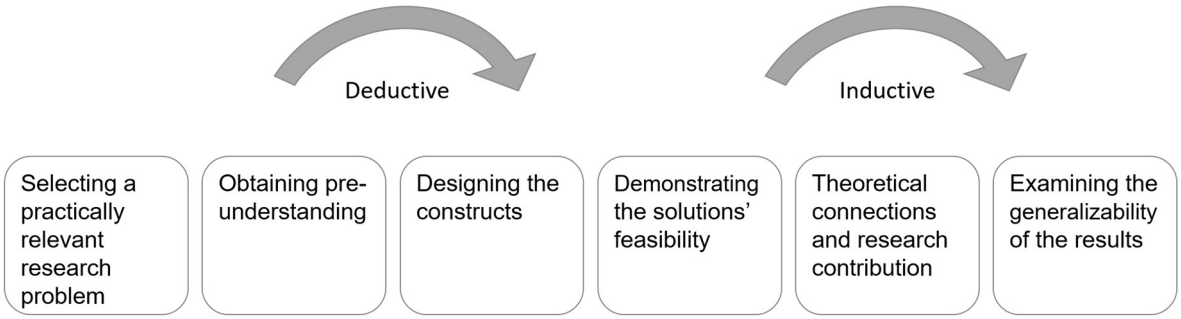



Figure 9: The constructive research process (Pasian 2015, 98)

Constructive research usually involves choosing the steps of the research based on the problem at hand, and a pluralistic approach is advised, which means using multiple research methodologies. (Pasian 2015, 99)

The research problem was selected when the thesis project started. The report starts with the secondary research in the theoretical part, which serves as obtaining pre-understanding. To add the business understanding, a questionnaire is conducted with the target com-

pany and their current market intelligence solution is studied. Finally, the pre-understanding stage is completed by researching software solutions on the market that might be relevant to the target company's business problems. Designing and demonstrating the feasibility of the demo solution takes place in the proof of concept section. The discussion chapter contains ponderings about the connections with theory and generalizability of the findings. The methods to carry out the separate parts of the research are discussed in more detail in the rest of this section.

**Company questionnaire**

To find out the current and desired situation at the company regarding market data collection, a questionnaire is sent containing 29 mostly open questions by e-mail. The questions are related to Big Data usage, customer feedback collections, competitors, lead generation and the currently used market intelligence tool. The answers are summarized in the Company case section.

**Software research**

The main objective of the software research is to decide which focus area is worth further examination, and this goal is reached by looking at the commercial and open source software currently available on the market. Due to the abundance of company information on official websites, the research is carried out as a qualitative online research described by Kananen (2015, 114-129). The **research process** consists of the collection of materials, segmentation, coding and categorization.

To find commercial software and services in the company's interest areas, **sources** needed to be investigated first. There are plenty of IT companies around the world offering analytics services, some are small start ups, some are well-known international companies like Microsoft or SAP. Market analysts like Gartner and Forrester were considered first. However, these companies sell their reports for a fairly high price so their insights weren't accessible for this thesis project. On the other hand, they offer a viable source of information for companies.

For the purpose of this report, emerging crowdsourced IT review sites were chosen as a resource. These companies challenge the role of Gartner and Forrester being the only providers of technology related market information (Dunn, 9 May 2013; Bradley, 15 July 2015). According to Levine (29 July 2016), the firm G2 Crowd even has its own "Magic Quadrant-like" product to determine the highest performing companies in different market segments.

The “Magic Quadrant” is a graphical representation of IT market segments by Gartner categorizing the players into one of four groups: leaders (complete vision and high ability to execute), challengers (less complete vision but high ability to execute), visionaries (lower ability to execute but complete vision) and niche players (both lower ability to execute and less complete vision). (Gartner)

G2 Crowd follows a similar logic but they base their findings on information for example from on-site reviews and social media activity. Zinck (24 February 2016) points out the risk involved in this type of classification: the validity of the reviews written by users is questionable. However, Tim Handorf (Levine 29 July 2016), CEO of G2 Crowd explains the process of controlling the quality of user inputs. These new review companies have opened a window into market research that enables decision-makers to have real-time insight into other customer’s experiences with an IT product, which was unprecedented earlier (Bradley 15 July 2015; Dunn 9 May 2013; Levine 29 July 2016). Besides G2 Crowd there are other IT review companies as well: getapp, TrustRadius, CloudsWave and IT Central Station. They were considered, but the so called “Grid” on G2 Crowd’s page, which resembles the “Magic Quadrant” was the most appealing solution to take a look at relevant companies. In addition to the information on the G2 Crowd site, the software solutions’ own homepages are also used for the research.

Regarding open source software, G2 Crowd is not a sufficient source, because it is mostly targeted at businesses with an intention of buying ready made IT solutions. Information about open source tools is gathered from books about analytics and Big Data. Furthermore, the software’s own project webpages are also looked at.

Segmentation, coding and categorization are all conducted with the research question in mind. As the sources differ, the techniques used for analysis are also slightly different for commercial and open-source tools.

**Segmentation** in the case of commercial tools takes place in two steps. First, the G2 Crowd software categories are studied and the completely irrelevant ones are discarded. Second, the category descriptions of the remaining categories are downloaded and examined to identify which of the three research topics it is related to. Irrelevant categories are discarded once more. During **coding**, the companies in the categories deemed interesting are looked at in more detail, including the inspection of their own website. The methods used by the company and their value propositions are extracted from the materials.

Companies and solutions deemed relevant have to fulfil the following criteria:

- solution for at least one of the three development areas
- utilizes Big Data technology
- is not industry-specific or provides information about the target company's industry

The short summary of the relevant companies is attached in Appendix 1. in the format of a table containing the following:

- the company's name
- website
- suspected tools and methods that it utilizes to fulfil its promise
- the category based on the purpose for the project (lead generation, customer intelligence or competitive intelligence)

A similar table is created for the open source tools. The main source of open source software related to Big Data is books written about analytics and then the software projects' own website. **Segmentation** is created by breaking up the websites based on the functionalities.

**Proof of concept**

A small demo is built to investigate the feasibility and challenges of a custom built market intelligence solution and to demonstrate how a Big Data application works. The design and implementation of the development follows the CRISP-DM (CRoss-Industry Standard Process for Data Mining) framework. The framework was first developed in the middle of the 90s, when data mining was in its infancy (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth 2000, 1). Since then it has become the most often utilized framework for a wide variety of analytics problems (Abbott 2014, 20). The framework consists of four layers: phases, general tasks, specific tasks and process instances. These layers provide a guideline for planning. Each phase is described, together with the corresponding general tasks. More detailed planning is left to the user of the framework, keeping in mind the project-specific actions that need to be taken. (Chapman & al. 2000, 6)

In order to help clarifying the boundaries of the analytics effort, first the context is determined. The "data mining context" consists of four elements . (Chapman & al. 2000, 7):

- application domain
- problem type
- technical aspect
- tool and technique

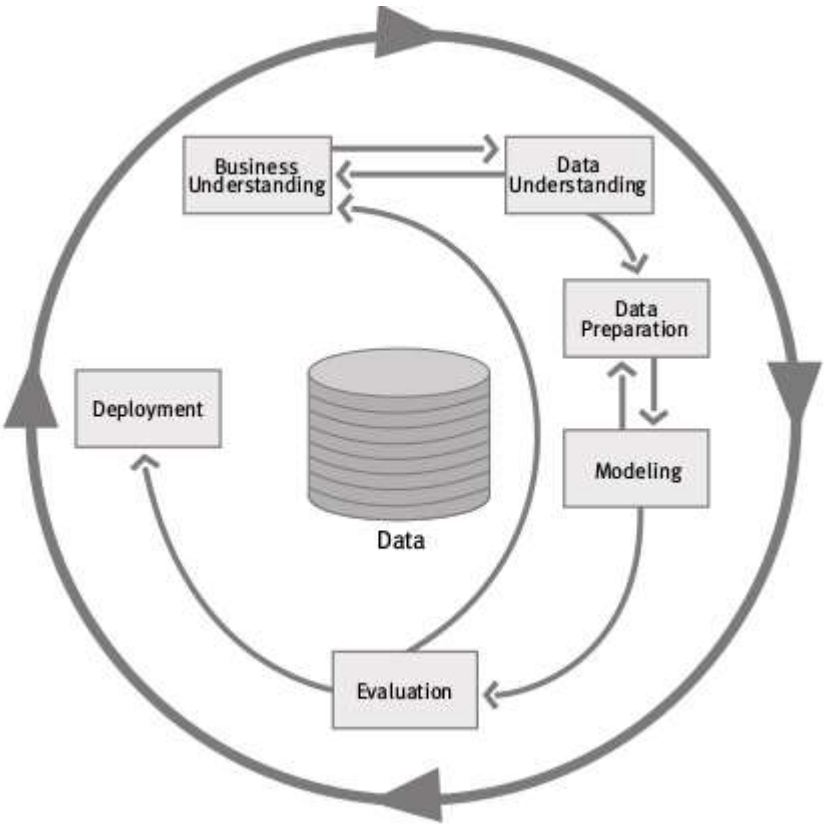



Figure 10: CRISP-DM phases (Abbott 2014, 21)

Second, the tasks can be determined by following the phases laid out on Figure 10. The returning arrows have great significance – they ensure that the business needs and goals are understood properly and prevent wasted effort on unwanted or irrelevant elements. (Chapman & al. 2000, 10-11)

The first phase is the business understanding. In this phase the main tasks are comprehension and definition of the business problems, evaluating the situation and setting goals for the data mining project, and creating a project plan. Since the project plan is a separate document during the thesis project, this step is skipped here together with the risk assessment part as well. (Chapman & al. 2000, 12)

In the second phase the relevant data is explored. The general tasks include gathering the data, describing it and then exploring it with simple tools for initial hypotheses and experiences. Data quality is also evaluated. Again, in case of misunderstandings or lack of clarity, the business phase can be revisited as needed. (Chapman et. al, 2000, 12.)

The third phase prepares the data for modelling, or, in general, for analysis. In this phase the data are filtered, cleaned, calculated attributes are created and the correct format is reached. The outcome is a dataset ready for analytics. (Chapman & al. 2000, 12)

The modelling phase includes the choosing, testing, building and assessment of a model. This phase links back to pre-processing, since unanticipated changes in the modelling phase might require a different format or some other changes in preparation. Due to the complexity of modelling, this phase is out of the scope of the thesis project. (Chapman & al. 2000, 12)

For the sake of completeness, the last phases are also explained here. When the model is built and reported, evaluation of the results is the next step. Further inconsistencies with the business goals might surface at this point, so a link back to the business understanding phase might be necessary. Otherwise the next steps are determined and the lastly, the solution is deployed. (Chapman & al. 2000, 12)

# 4 Investigation of the company case

This part contains the results and analysis from original research conducted for the thesis report. The practical findings are divided into three sections: company case, market research and proof of concept. The goal of breaking it up into three parts is to first introduce the business problems, then investigate existing software on the market to solve them, and finally, to create a demo solution to illustrate the structure of a Big Data solution and to collect experiences about an analytics development project. The proof of concept section contains the design and implementation results of the solution based on the research question. In the end, the general suggestions are formulated for the target company, summing up the findings of all parts.

## 4.1 Description of the company case

This section contains the results of the questionnaire and the inspection of the demo of the market intelligence tool used by the target company. In addition, the first conclusions are made about the case.

**Results**

The target company is a Finnish industrial part producer with most of their sales abroad. Their customers include both distributors and end users, like energy companies and construction contractors. Attaining new customers happens in many ways: participating at

embassy events, utilizing existing connections, receiving hints from the market intelligence system and researching online sources. Being in a business-to-business (B2B) setting, there are various channels used for communication with customers: the afore mentioned events, meetings, e-mail and phone calls, social media and advertisements. The company maintains three social media accounts: LinkedIn, YouTube and Twitter. They are updated regularly with news, demos of products and articles. The main motivation for them to join the Big Data project is to find ways to gain new leads, incorporate customer feedback in the R&D processes and to get information about the competition. These three topics will be discussed further in this section.

As mentioned earlier, there are various methods used to gain new customers, however the efficiency of **lead generation** could be improved. Being a multinational market, the challenges explained in the “Business needs” section fully apply – they need to act in a very diverse environment, selling a niche B2B product. The language barrier is remarkable, together with the fact that networking is more difficult with far away countries. Germany is a special focus area. Common practice in the industry complicates the situation even more – timing is key for advertising. Contractors are not willing to change blueprints and plans to incorporate a different product, so learning about a project and contacting the contractor as soon as possible is crucial.

The main motivation for collecting **customer intelligence** is to aid the product development process. New products are developed continuously, with about 1-5 new products per year. Customer satisfaction assessment is carried out once every two years, but increasing the frequency is planned. Feedback comes from the salespeople, resellers and customers themselves on an irregular basis. Including customers in the brainstorming phase has been practiced with success. However, there is interest in a more structured and automated process too.

**Competitive intelligence** is the third major focus area. There are about 10 relevant competitors worldwide, excluding cheap Chinese manufacturers. Counting them in makes the number much higher. Information about the competition is coming from distributors, salespeople, customers and also from the market intelligence tool. Information about the competition is important for research and development decisions and also because of the timing issues mentioned earlier in the lead generation process.

The company is currently using m-brain's **market intelligence tool** called Intelligence Plaza[2], which has been implemented for about one year. According to the classification in The Handbook of Market Intelligence (Hedin, Hirvensalo & Varnaas 2014, 144), this product can be categorized as a form of "intelligence portal". Push & pull functionalities are also present in it, which means there is opportunity to pull information form the software, and also the software pushes relevant information in the form of notifications and alerts. Extracting information can happen in three ways: using search, browsing by topic categories (competitors, customers, trends, etc.) or browsing relevant rss feeds and articles. To get important information as soon as possible, e-mail alerts can be set up based on search keywords, and they can be appended to the dashboard. The software offers a high level of customization – the contents of the dashboard completely depends on user preferences. Furthermore, there is a possibility to save interesting insights for later or for sharing with others via e-mail newsletters. From the infrastructure point of view, it is a Software-as-a-Service solution, with the possibility of self-hosting. The purpose of implementation was lead generation and informing R&D decisions. Users of the tool include sales, customer service, quality management and sourcing. Mostly used feature is the newsfeed that contains information about possible leads and competitor activity in a country-specific manner. Information is retrieved on a weekly basis. One disadvantage of the software is that it is private, so business partners don't have access to it, but this is partly solvable by creating and sending out newsletters containing the most relevant news.

**Discussion**

Currently market intelligence is collected from many sources including the m-brain tool. At this stage, lead generation and competitor intelligence seem to be the more problematic areas. Lead generation is facing serious challenges because of industry practice and geography. Customer intelligence looks like a more positive area where good practice can be expanded. Regarding the MI tool, information broken down by country is an important feature. Finding German leads is a priority. It would be beneficial if customer feedback could be collected in a more efficient, automatic way. The afore mentioned technology fatigue also causes some problems in using the MI tool – even the current tool has a lot of potential if the users can obtain a higher level of awareness and activity.

[2] https://www.intelligenceplaza.com/

## 4.2 Market intelligence software market review

This section contains the results and suggestions derived from researching the IT market of market intelligence software. Based on the outcomes suggestions will be made in the three development areas and a topic for the demo implementation will be chosen.

### Results

Discussion of the **commercial software** results starts with some quantitative data about the G2 Crowd website. After this second filter 21 categories remained containing 1409 companies, which are all parts of the CRM & Related category.

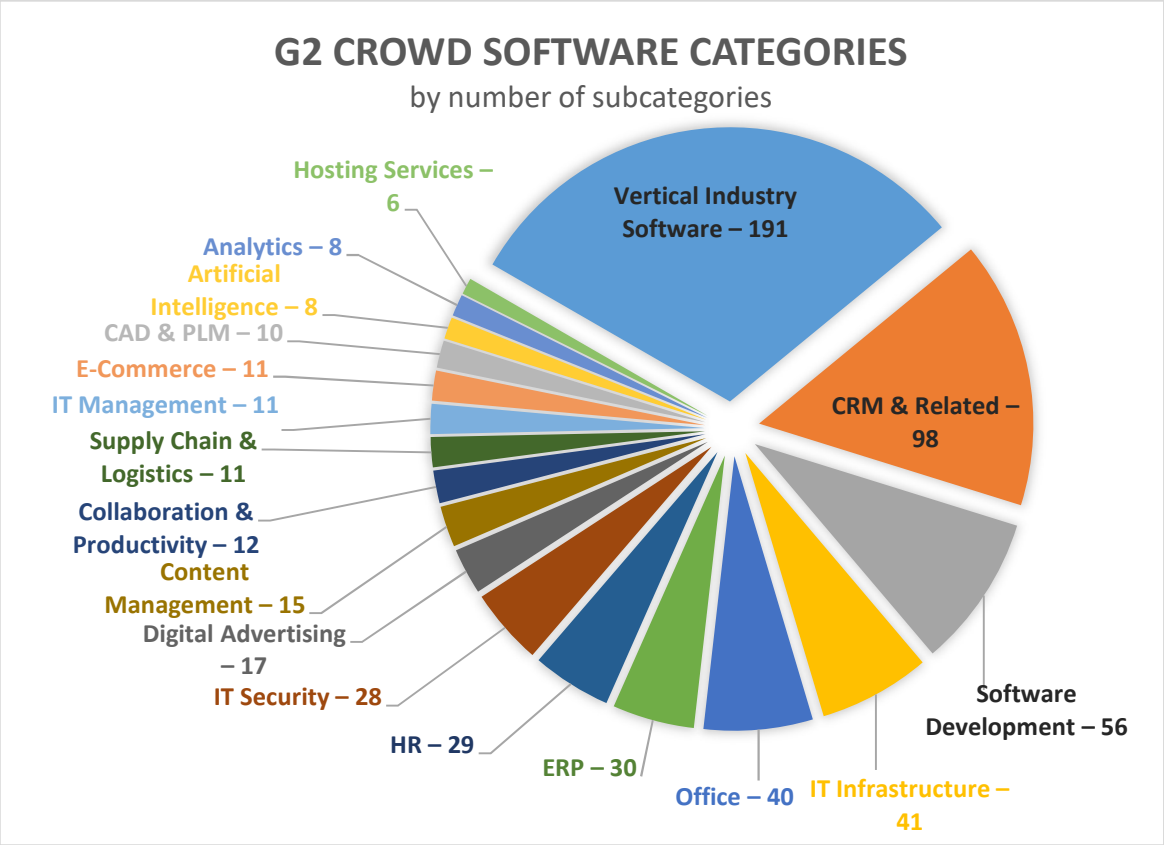



Figure 11: G2 Crowd software categories by number of subcategories

The relevant software subcategories are shown on Figure 12. 14 categories were found relevant to lead generation, 8 categories for customer intelligence and 2 for competitor intelligence. The total of these numbers is over 21 because the “market intelligence” category was labelled both as lead generation and competitor intelligence and the “other social media” category might be relevant to all three topics.

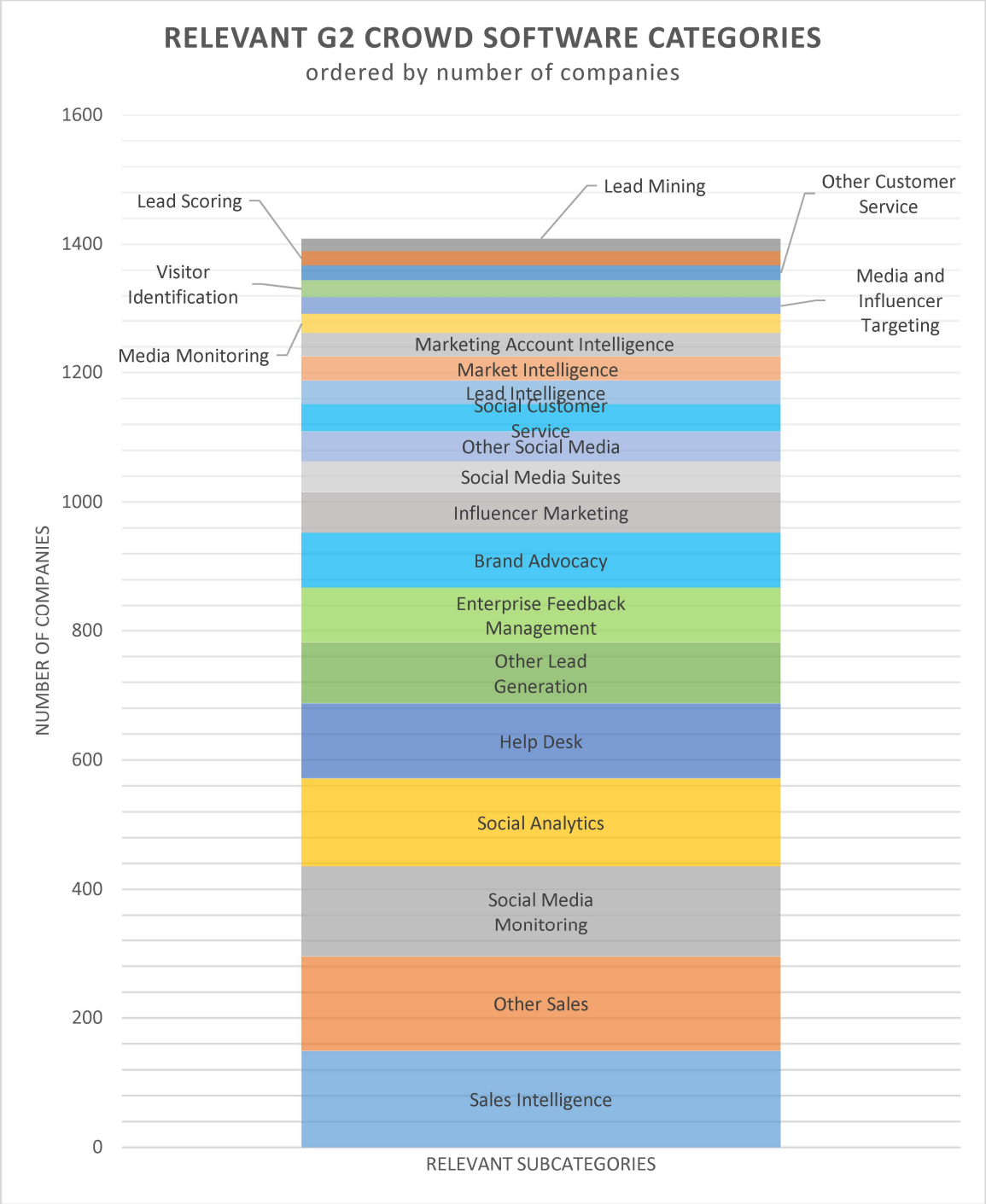



Figure 12: Relevant G2 Crowd software subcategories ordered by number of companies

However, this level of filtering still wasn't sufficient. Four subcategories were chosen from the previous 21: sales intelligence, lead intelligence, lead mining and social media monitoring. These categories were found to be the perfect matches to the company focus area, but this choice is in no way exclusive – other categories, which are out of scope of this analysis, probably still carry valuable information. The final decision to narrow the scope was to include only the companies on the "Grid", not the full list of companies in each category.

The first subcategory to discuss is **sales intelligence**. There are 43 companies on the "Grid" (Figure 13). Individual description of them can be found in Appendix 1.

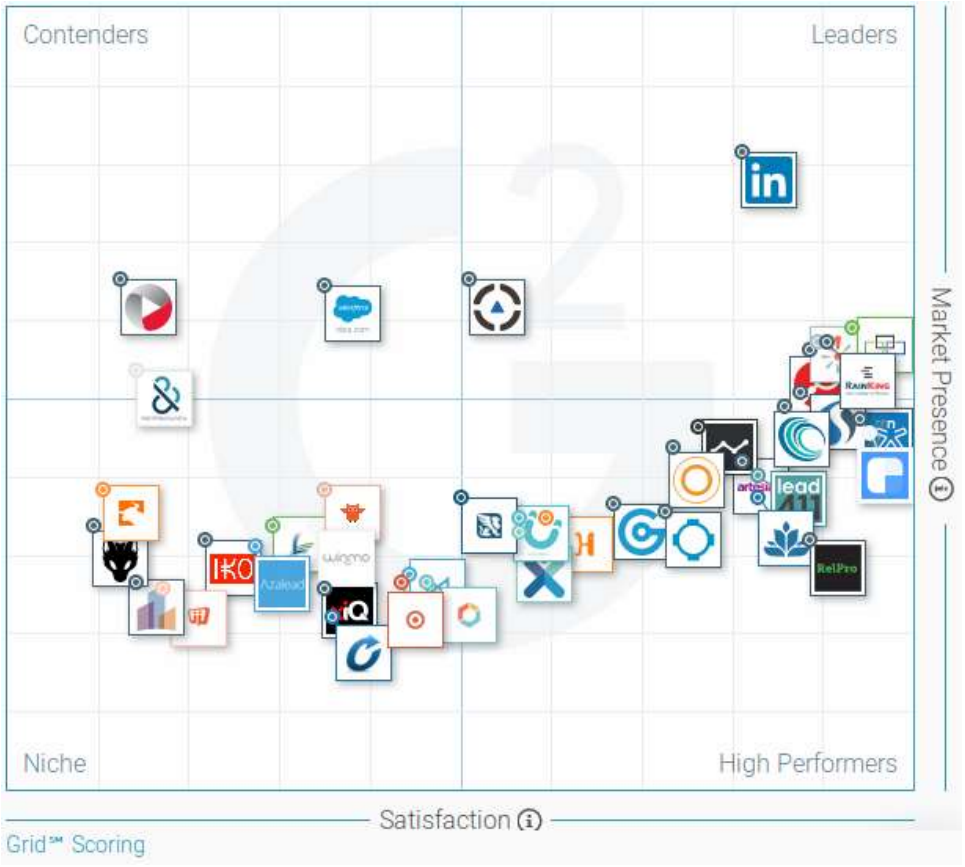



Figure 13: G2 Crowd Grid℠ for Sales Intelligence (G2 Crowd, 2017c)

The companies in this category were found to be more versatile that initially expected. Some solutions were relevant not only in the lead generation, but also in the competitor and customer intelligence domains.

The breakdown of the 44 companies by purpose for this project is illustrated on Figure 14. It is important to clarify the meaning of None and Irrelevant columns. Companies that didn't fulfil the criteria of providing value in the development areas belong in the "None" category. On the other hand, the ones that don't work with Big Data or target a specific industry different from the target company's are grouped together in the "Irrelevant" category. Some companies provide solutions for more than one of the focus areas, in this case they are counted in all relevant columns.

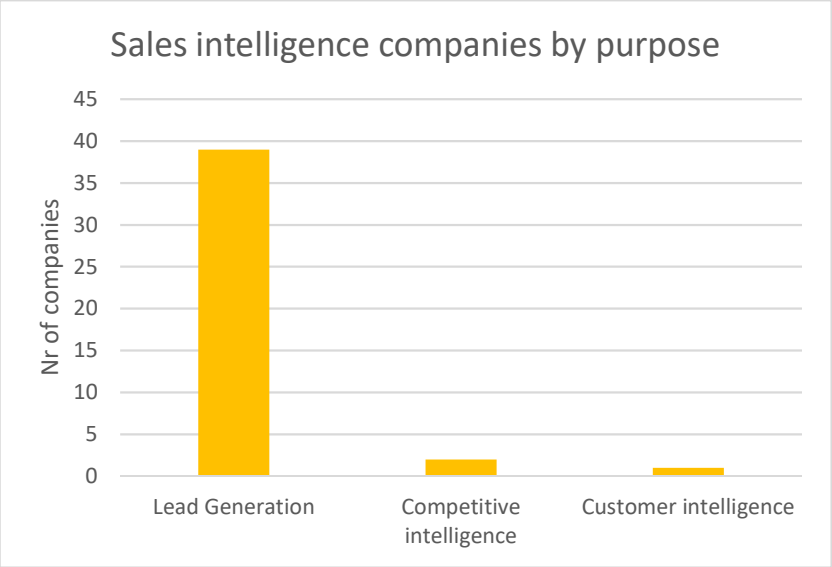



Figure 14. Sales intelligence companies by purpose

The next chart (Figure 15) shows the suspected methodologies behind the solutions. Again, some companies utilize more than one technique to deliver their results, and they are all accounted for in the different columns. The most common method used in this category is a company and contact database maintained by web-scraping, social media monitoring and following news sources to offer contact information to clients.

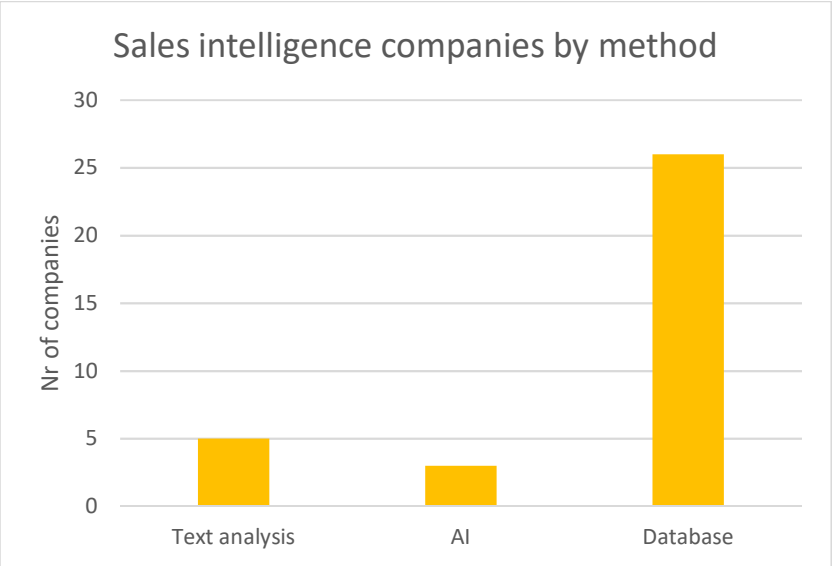



Figure 15. Sales intelligence companies by method

The contact database offer is often enhanced by data cleaning services, which means that the customer's own contact data is verified and holes are patched. In addition, predictive analytics capabilities are also natural extensions of a database. Predictive analytics helps with determining which contacts in the database have the highest chance of turning into sales. An example company is of offering these three services together is Oceanos. The

contact tracking column shows a group of solutions that don't keep a contact database but fetch the contact information from social media when changes are detected. Text analysis is used by a few solutions to process data resulting from web-scraping. Some companies are working with Artificial Intelligence to increase the efficiency of prospecting.

The **lead mining** subcategory is much less populated, containing only 8 companies. All solutions focus on providing leads, so only the applied methodology is discussed in this case (Figure 16). The irrelevant company again means that they offer their software for a different industry.

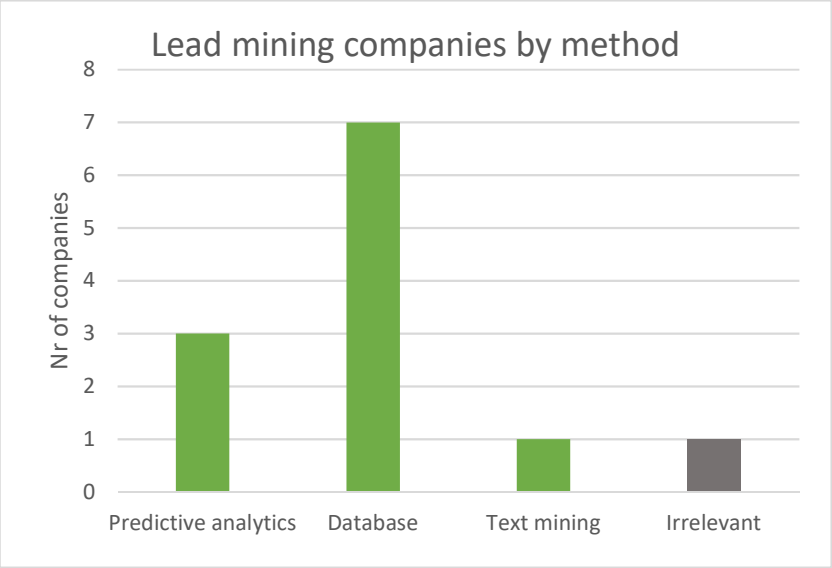



Figure 16. Lead mining companies by method

All these companies keep their database of contacts as explained earlier, and combining it with predictive analytics occurs here as well. It is important to notice that there are some overlaps between the sales intelligence, lead mining and lead intelligence categories, so some of these solutions have appeared earlier already. An interesting product is offered by LeadGnome, which is an exception is this group in the sense that it is the only one that doesn't' utilize it's own contact database but mines e-mail text, especially out-of-office e-mails for new contacts.

The lead intelligence subcategory has exactly 8 companies, just like the previous one. The purpose is clear in this case as well, so only the methods are discussed. Figure 17 shows the techniques applied by these companies.

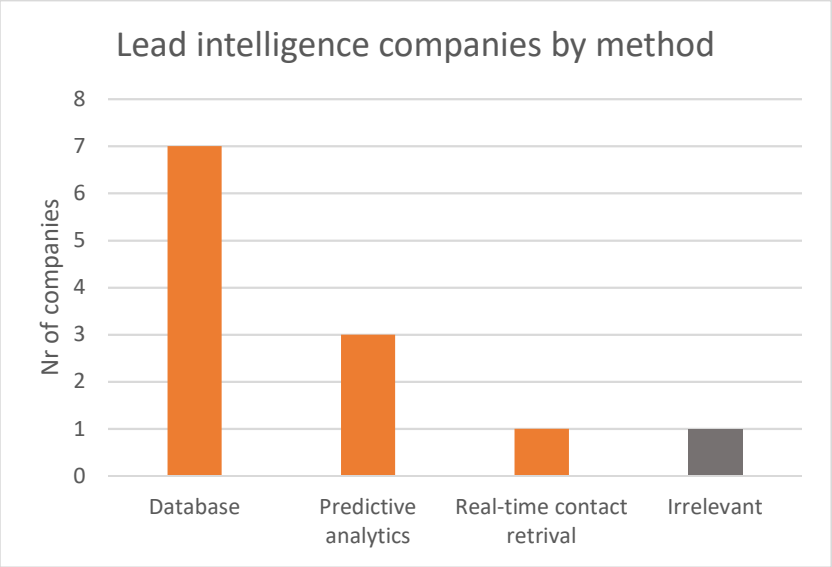



Figure 17. Lead intelligence companies by method

The contacts database combined with predictive analytics is a reoccurring theme here too. In case of real-time contact retrieval, there isn't any contacts database, but the information is looked up as needed.

Finally, the findings from the **social media monitoring** subcategory are clarified. This group is as populated as the first one, sales intelligence was, containing 46 companies.

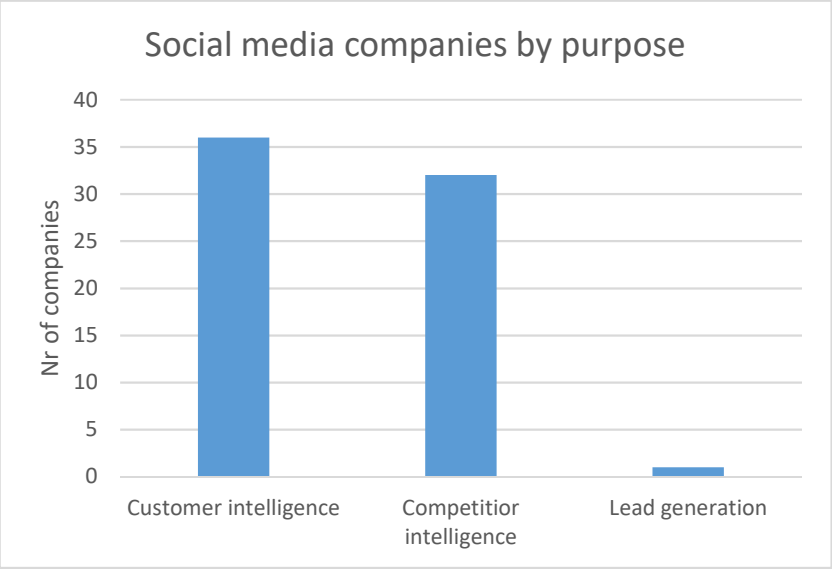



Figure 18. Social media companies by purpose

It is also quite versatile in the sense that there are software for all three focus areas here. The number companies for each development area are illustrated by Figure 18. Most solutions in this category served both the customer intelligence and the competitor intelligence purpose. There was one software even for lead generation. Quite a few couldn't be

tied to any of focus areas (as in “None”) , and there were a few which serve a different industry (as in “Irrelevant”).

Suspected technologies are shown on Figure 19. To find out about interesting topics, probably most use a form of text analysis, topic recognition. However, based on the data from the company homepages, some solutions have a more basic capability, tracking and searching by key words. Sentiment analysis is also occurs rather often, which is not surprising, since many solutions focus on monitoring and maintaining brand image. Machine learning solutions can extend the capabilities beyond text analysis to image analysis as well. A small fraction of the companies uses network analysis to find influencers on social platforms.

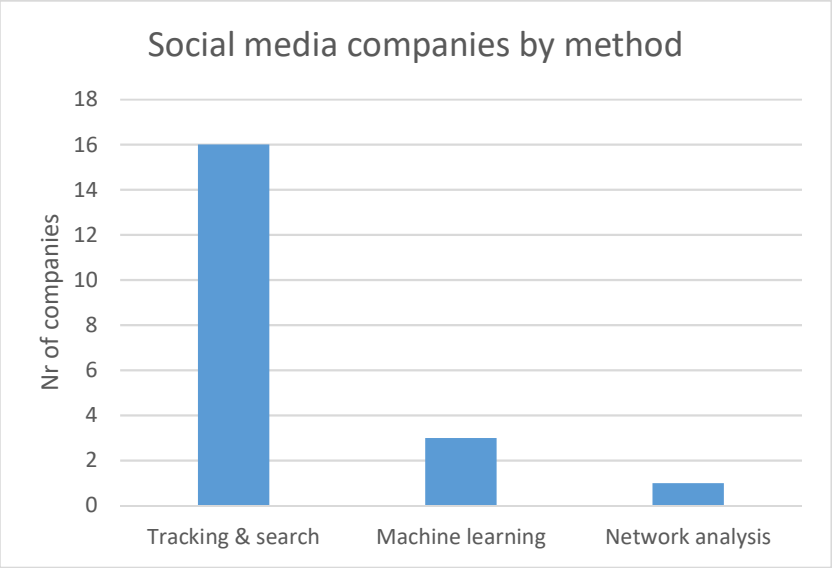



Figure 19. Social media companies by method

Other companies’ case study collections also contained innovative solutions and services. Bombora offers data on companies’ buying intent based on the browsing information they collect from content creators, like business and technology related magazines and newspapers. The idea behind it is that companies who read about a certain topic are likely out on a hunt for products and services in that domain. Bombora collects buying intent data in a wide spectrum of industries around the world. Dstillery is similar in a way that they also utilize online activity tracking, but they do so based on browser information. Their services are more targeted at B2C companies as they focus on context aware advertising for consumers.

There are several advantages of **open source software**. First, there is a community of developers behind these projects so continuous improvements are assured. Second,

open source software means a cheap solution for businesses (and consumers as well), since the program does not cost anything. An example is the free operating system, Linux. While the operating system can be installed for free, Red Hat is a commercial company that takes care of maintenance and support. It happens with other open source tools as well, that a commercial company maintains them. Even in that case, they are cheaper than if the software licences would have their own price as well. It also enables the trial of tools without requesting demos or negotiations with software companies. Another business advantage is the flexibility. This means that the solutions can be customized at will, given that there is a developer available for the task. (Minelli & al. 2012, 67-68)

Hadoop is an element that can not be left out of the discussion about open source Big Data tools. Hadoop was created in 2008 and is a system that provides an opportunity to store and analyse big volumes of unstructured data in a cost effective and fault tolerant way through cluster computing. It means that data files are divided into smaller, duplicated chunks that are stored on separate servers. This way, in case of server failure, the data is still available from another source. Relatively quick processing is possible by executing the code on the server nodes themselves, instead of moving the data to a different location for processing. The stored file format is called HDFS (Hadoop Data File System). The creator of Hadoop founded Cloudera to distribute his invention. This is another example of an open source product provided by a commercial company.

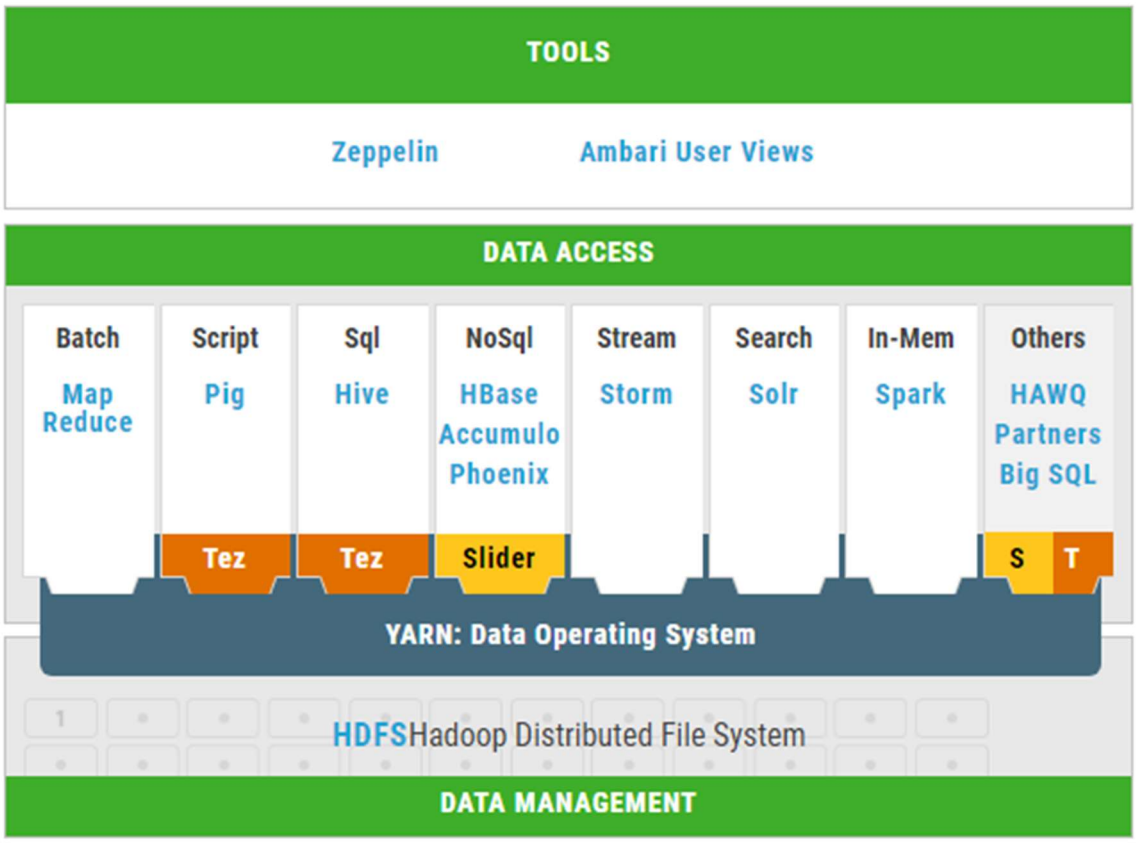



Figure 20: Hortonworks Data Platform (Hortonworks.com)

On the other hand, there is a provider that shares a complete data analysis platform for free – Hortonworks. Figure 20 shows some of the open source components that Hortonworks consists of.

In-Mem stands for in memory, which means that Spark is a data analysis software that computes the whole dataset without accessing the hard disk. The result is a processing speed that is promised to be a hundred times faster than the batch processing with the usual Hadoop method, MapReduce. Even if the dataset doesn't entirely fit in memory, Spark is still ten times faster than the batch approach according to the projects homepage (Apache Software Foundation, 2017). Spark offers four language choices to program it with: Scala, Java, Python and R. Scala is the main language, so there are some functions that are not available in Python or R: There are four libraries that make it possible to solve different tasks: working with SQL databases, machine learning, data streams and graphs. Supported data sources include JSON and csv files, SQL and NoSQL databases and HDFS files. Processed data can be output to the same wide range of storage options.

Python as a programming language for the project is beneficial for two reasons: first, Spark can be operated using its Python API, second, there are dozens of libraries for solving a wide range of development tasks from the interpretation of JSON data sources to data visualization. Naturally, these libraries are free to use to build applications with, and as Python's popularity has been gradually growing, there are plenty of free tutorials online explaining how to use them.

Another great tool for data visualization is the JavaScript library d3.js. Compared to the Python external libraries, it carries the advantage that it doesn't need any installation, only the link to the JavaScript functions needs to be included in the HTML head element to be able to use it right away. The charts and graphs rescale nicely with the page, since they are generated as scalable vector graphics (svg's).

Even when handling Big Data, there might be a need for more traditional databases than the Hadoop HDFS mentioned earlier. Depending on the needs, there are SQL and NoSQL databases which can be handled more easily by web applications than newer storage forms and transferring legacy data into them is straightforward. Examples of NoSQL databases are MongoDB and Apache Cassandra. The most popular open source SQL databases are MySQL, MariaDB, PostgreSQL and SQLite.

**Suggestions**

Based on the collected commercial solutions and the possibilities mapped by investigating open-source solutions we can draw a conclusion on what are the improvements that are feasible and have the best ratio between costs and benefits.

Regarding paid solutions, LeadGnome's e-mail mining software seemed to be the most interesting. The implementation is not very complicated or demanding, and the monthly costs are relatively low, and it promises new leads and the maintenance of existing contacts. The starting price of LeadGnome's services is 50$ per month and this is the package that would be relevant to the target company based on e-mail volume.

However, there is another approach that is worth further investigation for the demo implementation. Twitter and LinkedIn are openly available resources to find information about professionals, and there is an excessive amount of data to conduct research. Furthermore, the commercial solutions that claim to keep lists and offer new leads also utilize online research, so by creating a demo we get insight into how those software companies operate. The next part will discuss the implementation of mining social media data to find leads. The **refined research question** is phrased as follows: *"How to generate new leads for the target company, using only openly available online resources?*"

### 4.3 Proof of concept of a social media monitoring tool

This section contains the results of building a proof of concept for a social media monitoring solution and the suggestions drawn from the experiences. Regarding the constructive research framework, the design and feasibility demonstration of the construct are located here. First, the CRISP-DM methodology is used to go through the process from planning the solution for the concrete business problem to the deployment of the demo. In the end, experiences are summed up and suggestions are made for developing the solution further. More detailed technical information about the implementation can be found in Appendix 3.

**Results**

The design of the demo solution starts by creating the analytics context based on the refined research question and the CRISP-DM guideline. Different aspects of the project can be seen on Table 1.

First, the **application domain** is defined as lead generation for a B2B company. It is important to underline that the user of the system is not a B2C company, because a completely different approach is needed to gather new customers for a business that is selling to other companies, not consumers. In this case the group of possible buyers is much smaller and more specific than if the target company was selling, for example, soft drinks. For this reason the main challenge of the analysis is to find relevant leads and only relevant leads.

| Analytics context | | | |
|---|---|---|---|
| Application domain | Data mining problem type | Technical aspect | Tool and technique |
| Lead generation for a B2B company | Text mining, classification | Veracity of social media data, spelling issues, character encodings | Linux, Python, Spark Streaming, Twitter |

Table 1: Analytics context

The chosen approach is extracting information from social media activity, which means that the **problem type** is text mining, more specifically, classification. As explained in the theoretical part, text mining is conducted by applying algorithms used in predictive analytics to find patterns in the text. For an algorithm to work, the text needs pre-processing. Classification is the appropriate predictive method, because by collecting a test dataset and tagging the tweets in it as relevant and irrelevant, a classification algorithm could predict the relevance of new tweets.

There are some **challenges and risks** involved with the implementation. The veracity of social media data is always questionable, because anyone can tweet anything they want, there isn't any kind of regulation to check the truthfulness of the message. Another problem with this type of data is that people don't pay careful attention to spelling and grammar, which makes it even more difficult for a machine to process the natural text. Finally, debugging issues with character encodings can slow down the implementation.

The main **tool** chosen for the implementation is Apache Spark Streaming due to the streaming capability and the integration with the Twitter API. Spark Streaming is run in the CSC cloud, on a Linux server. The programming language of choice to operate Spark with is Python because of previous experience with Python in a text mining project. Moreover, later elements, like the output homepage and data visualizations are also possible with Python libraries.

**Business understanding**

The company case has been documented with a more generalized and unstructured approach in Part 3. In this stage of the CRISP-DM model the requirements for this specific research question are clarified and concretized.

The **business objectives** related to the demo implementation can be summed up as follows: gaining new customers in the construction and energy industries abroad, first and foremost in Germany. The success criteria of fulfilling these objectives is to find relevant contact information for sales representatives.

The **available resources** at the beginning of the development are the CSC cloud environment with a Linux operating system, internet access and Python and Apache Spark installed. There are two initial **assumptions**: first, the employees of the possible customers are active on Twitter. Second, they share relevant industry related news. There are also **constraints** to make the development more manageable: only tweets posted in English or German are analysed. In addition, web crawling restrictions specified by websites' terms and conditions are to be respected.

Based on the business understanding and assessment of the situation, the **analytics goals** can be set. The analytics effort is focused on collecting relevant tweets from Twitter to find new construction projects or employees of possible customers. The success criteria is a list of relevant tweets and users (contact people) output on a website for easy access.

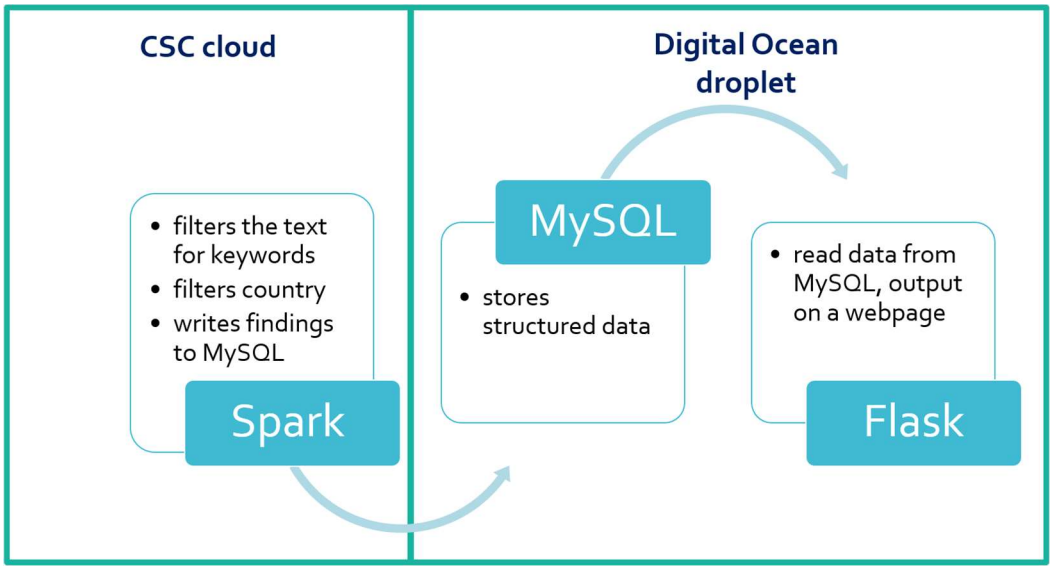



Figure 21: Planned infrastructure of the TwitterWatch application

The structure of the planned solution can be seen on Figure 21. The application is running in the cloud. Spark Streaming is located on Haaga-Helia's CSC environment, while the MySQL database and the web application resides on a private server from Digital Ocean. The reason for splitting the application in two is that the author intends to showcase the outcome of the thesis project on her portfolio page after graduation.

The first step in the process is feeding data coming from **Twitter to Spark Streaming**. Spark Streaming filters the relevant tweets and writes them out to MySQL.

The first question to consider about the database was to use an **SQL or NoSQL** based database. SQL was chosen for two reasons: the author has experience with SQL databases but not with NoSQL databases, and integration with the webpage was going to be more straightforward in the case of an SQL based database.

The second question is **which SQL** database to use. MySQL is chosen for various reasons. First, of course, it is open source. Second, Spark Streaming can use its jdbc connector to write out data into it. Third, it is widely known, which means there are tutorials and documentation about it online and also there are Python libraries for integration with it.

To show the results on a website, **Flask** is used. Flask is a Python based framework to create webpages and it utilizes Jinja2 for templating. The reason to choose this solution was easy integration with MySQL using the Python MySQLdb library and also with the pandas and Bokeh libraries for visualization. In addition, there is a very detailed tutorial[3] on how to set up a website using MySQL with it.

**Data understanding**

This stage contains the first-glance information about the data collected from the Twitter API. The insights laid out here lead to the decisions made in the pre-processing stage.

[3] https://pythonprogramming.net/practical-flask-introduction/

The initial step is to start using the Twitter API. To be able to do this, a Twitter account and an application needs to be set up. The first attempts to connect resulted in the realization that although Spark Streaming has Twitter support, it is only available in the Scala API. A Python workaround[4] was found where the author used a Twitter streaming Python library to create an application that streams and forwards the Tweets to Spark Streaming using a TCP socket. This example code became the basis of the streamer application that feeds Spark Streaming for this project.

The format of the streamed data is JSON. There are roughly 30 attributes and they are specified on the Twitter API documentation page (Twitter 2017a). Most of the attributes are optional. The compulsory attributes include the Tweet ID, a timestamp of the creation, the user who shared it, the message itself, and some retweeting and quoting information. What is interesting in our use case is location. There are two kinds of location data stored, coordinates and place. Both of them are optional, nullable attributes.

Upon investigation it was found that there are indeed energy and construction industry related messages and articles shared on twitter, however their frequency is lower than expected. A simple search using the twitter page suggested that relevant articles about, for example, new construction projects happen once in every few months. The lack of domain knowledge made finding relevant information a challenge.

There are a few issues with the incoming raw data. For the purpose of finding new leads by country, location data is crucial. However, since the coordinates and the place attribute are nullable, most tweets don't have location information. Dealing with text data means encoding issues as well. Furthermore, there are characters within tweets, like single quotes and double quotes, that break Spark Streaming's JSON interpreter, causing the whole program to shut down. Finally, the lang attribute contains a two character identifies for language code, but there are case where there is no language information, and then it is set to "und" for "unidentified".

This initial stage did not result in a saved dataset but in a stream of tweets coming from Twitter and going through the forwarder python application to reach Spark Streaming

[4] http://awesomestats.in/spark-twitter-stream

**Data preparation**

The next step is to work on the raw data collected in the previous stage to prepare it for modelling. Tasks related to selection, cleaning and construction are discussed here. Being a simple project, data integration was not necessary, so that step from the CRISP-DM framework is skipped.

The **selection** of important data is the first step. From the extensive list of attributes the following are needed for this project: id, text, user, created_at, lang, place. Id is chosen to serve as a primary key for the MySQL tables. Text is the actual message of the tweet. User information is important in identifying the people that tweet the most about a topic – they might be working in a position that makes them a possible lead. Created_at contains the date and time of the tweet creation. It is useful for analysing historical data. Extracting the lang attribute enables filtering based on language. The place attribute contains data about the location the tweet is related to. These attributes are extracted by the tweet forwarding application. In addition, it also performs two aspects of filtering on the tweets: initial filtering required by the Twitter API, and language based filtering. The initial filter is required because the volume of tweets posted every second is too huge to analyse everything real time. The initial filter is set to retrieve tweets that were posted in Europe or are talking about it.

Solving the data quality issues mentioned earlier happens in the **data cleaning** phase. Location data is deemed crucial, and basing the initial filter on location ensures that all tweets contain the place attribute. Encoding issues in Python 2.7 are solved by applying a best-practice explained on a company's technical blog[5]. Quotation marks are stripped form the tweet text through simple string operations in the tweet forwarding application. To ensure that there are only English and German tweets in the database, tweets with unidentified language are not forwarded to Spark Streaming.

There were two attempts to filter relevant tweets. The first attempt aimed at adding further attributes to the tweets to mark the presence of a keyword in them. This attempt proved to be prone to errors and ineffective, because the ratio of tweets containing the interesting

[5] https://www.azavea.com/blog/2014/03/24/solving-unicode-problems-in-python-2-7/

keywords compared to all English and German tweets from Europe was very low, resulting in a database mostly containing noise. The second attempt involved writing the tweets that had a keyword in them to a corresponding table in the MySQL database. This way only the filtered tweets are saved and the irrelevant ones are discarded.

To access the tweets found relevant by Spark Streaming, the webpage was set up. The structure of the page is based on the three topics: energy, oil and construction. Each topic contains a bar chart with the top ten users who tweeted the most about the topic, the top ten cities where the most tweets came from, and a table with the actual tweets. Visualizations were attempted with the pandas and Bokeh Python libraries, but the Apache2 web server could not find the dependencies. The d3.js JavaScript library was used for the visualizations using a basic tutorial[6]. The charts can be found in Appendix 4.

**Modelling**

The main focus of the thesis project was to create a functioning solution. This meant that the actual predictive modelling stage was not reached within the timeframe of the project. The solution needs further work, and the creation of a well-optimized model to detect the relevance of tweets is in itself a viable option for a whole other thesis project. The stages following this one are described in the context of this thesis project, not from the viewpoint of building the fully functional solution.

**Evaluation**

This section describes the functionality of the solution in its current form. Figure 22 shows the structure of the implemented solution before handing it over to the Big Data – Big Business project. The website can be found at http://165.227.158.183.

As illustrated, the pipeline has four major components, two of which are residing on Haaga-Helia's CSC cloud environment, and the other two in a private server provided by Digital Ocean. Differing from the plan, an extra step needed to be taken in the very beginning of the data flow: the Twitter stream is accepted by the TwythonReader.py application

[6] https://bost.ocks.org/mike/bar/

using the Twython library to access Twitter. The code is based on the Tweepy workaround mentioned earlier[7]. It is adapted to the Twython library and extended to accommodate the extraction of several attributes, sending them as JSON, filtering by language and removing the quotation marks from the tweet text.

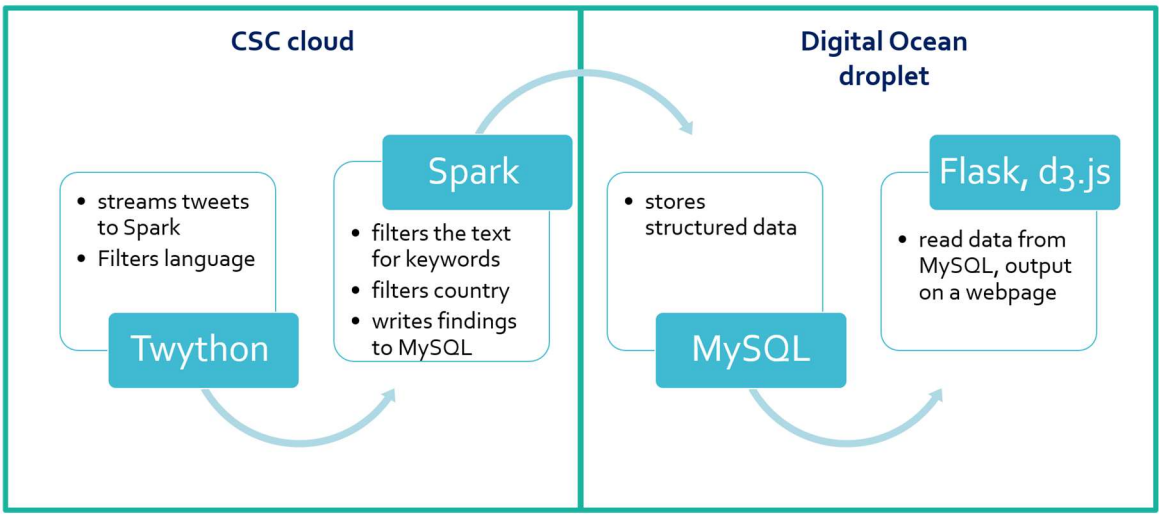



Figure 22: Implemented infrastructure of the TwitterWatch application

The Spark Streaming file is a modified version of one of the example word count files. A regular expression check was used for filtering and the MySQL connection was implemented to output the result. The modifications were based on the Spark Streaming, Spark SQL and Python 2.7 documentations. The MySQL database contains four tables: tweets, oil, construction and energy. Tweets has around 300 generic English tweets left over from testing.

There are no aggregations in these tables, the data for creating visualizations is collected from ad hoc queries. The Flask application consists of python files and HTML templates. The d3.js JavaScript library is used for creating the bar charts from the data. The W3.CSS framework is used for styling[8]. There are two Python files. First, __init__.py, which launches the application, queries the MySQL database and renders the HTML templates

[7] http://awesomestats.in/spark-twitter-stream

[8] https://www.w3schools.com/w3css/

with the retrieved data. Second, dbconnect.py, with the sole purpose of creating a database connection to MySQL. There are separate HTML templates for each page, containing the menus, the JavaScript in case of the visualizations and the HTML for tables in case of the Tweets tables. The Jinja2 templating is used to pass in the data from the SQL queries to the HTML. The Python files are based on the Flask tutorial files[9] mentioned earlier, the HTML menus are taken from a W3.CSS template[10] and the JavaScript is largely based on the bar chart tutorial[11].

From the thesis project's point of view, the next step is to hand over the solution to the Big Data project by duplicating the MySQL database and the homepage in HAAGA-Helia's csc cloud environment.

While the infrastructure is functioning, the keyword search filtering method lets through a lot of noise, which means that at this stage the solution does not provide any useful leads.

**Deployment**

In this section the final tasks related to handing over the solution are clarified and insights are shared related to operating it.

The deployment in this case means the installation of the MySQL database and the homepage in the HH server. This will consist of two main steps: installing and configuring the MySQL server to accommodate connections coming from Spark, and installing Flask and other dependencies for the webpage to function.

Monitoring of the streaming process can be done through the graphical interface online (unavailable as of 15.11.2017). The streamer shuts down unexpectedly after a few hours of running. The suspected reason is lack of RAM, but the solution is outside of the scope of this project.

[9] https://pythonprogramming.net/practical-flask-introduction/

[10] https://www.w3schools.com/w3css/tryw3css_templates_analytics.htm

[11] https://bost.ocks.org/mike/bar/

The main task related to maintenance is to follow the growth of the database and back up old data to keep the cloud environment from expanding indefinitely. The decision needs to be made how much historical data is to be stored for later analysis.

Special attention needs to be paid to the length of operations carried out in Spark Streaming, since the problem of “back pressure” can arise if the processing speed can not keep up with the inflow of tweets. Another possible issue is that the forwarding applications throughput is unknown and it might become a bottleneck if the initial filter is changed from the coordinates of Europe to something else, like a keyword search.

**Suggestions**

This section answers the question what are the components involved in building an open source lead generation solution for the target company. Furthermore, it contains the experiences and opinions formed during implementation and the suggestions for further development.

The technical feasibility of building an all open source solution has been proven, since the implementation is functional. However, to reach the analytics criteria and provide value, more work needs to be done. In spite of the demo not fulfilling the analytics success criteria, the proof of concept yielded valuable insights. The reason behind leaving the solution at this stage is that the main focus of this report was to gain insight into how market intelligence tools are built. This means that the **main priority** was a working infrastructure. The development of a more tailored filtering tool has proven to be a way too complex task for the scope of the thesis. Further work is to be carried out by the Big Data – Big Business project.

Useful experiences were collected during the implementation. While Spark provides the option of choosing the programming language one is the most comfortable with, working with Spark is not straightforward. This is due to **parallelization** and, in this case, **streaming**. The different data abstractions, like RDDs, Datasets, DataFrames and DStreams take a while to understand for someone without experience in parallelization. Creating a streaming application is challenging, because the functions for streams and functions for distributed datasets are not interchangeable. Another complication is the management of the parallelization itself. This project runs Spark locally and ignores the optimization questions and the middleware needs that arise when running it on a cluster. These areas need attention when developing the demo further.

The main improvement area to fulfil the analytics success criteria is the **filter**, which is too simplistic at this point to provide value. Approaches to develop the filtering done by Spark include word frequencies, classification, topic recognition and network analysis.

**Word frequencies** can be used on news articles that contain relevant information about, for example, new construction projects. The words that are common in most articles can be used as new, more specific key words which can then be applied as an initial filter or as a regular expression search like in the current implementation.

The **classification** method can also be used to find relevant news articles. First, a dataset containing both irrelevant and relevant tweets is collected and the tweets are tagged manually as interesting or noise. The next step is pre-processing text to prepare it for modelling. A model is built using a classification algorithm from the machine learning element of Spark based on the tagged dataset. After that the model can be applied to the incoming tweets to classify them as relevant or irrelevant automatically.

The most **advanced methods** that could be utilized in this problem domain are topic recognition and network analysis. Topic recognition is useful for finding the underlying topics in a tweet, making it possible to filter only tweets with a relevant topic. Network analysis doesn't involve text analysis but is used for analysing social connections, which could be applied for finding relevant contact people based on their professional connections on social media. These two methods are more applicable to fixed datasets instead of streaming.

### 4.4 Suggestions

The final answers to the initial research question is covered in this chapter. Moreover, connections with the theory and the value of the research is discussed in accordance with the stages of a constructive research.

First, the question of how Big Data can help with market intelligence is reviewed in a **case-specific** sense. The target company had three main focus areas. In two out of these three cases it was found that Big Data can provide some additional information, and significant ones in the third case.

Especially for **customer intelligence**, the currently practiced personal approach is quite effective. The reason of Big Data's limitations in this context is that the main source of collecting customer feedback is from social media like Twitter and Facebook, and for such a

niche product there isn't much information shared by customers on these platforms. However, processing customer e-mails might reveal strengths and weaknesses related to products or customer service.

Regarding **competitor intelligence**, it was found that it is a complex service involving not only web scraping and online research but also physical surveillance. A worthwhile source of online data about competitors is social media, which is utilized by Intelligence Plaza, the software currently used by the company.

The area of **lead generation** is the one where a great variety of commercial software was found that is utilizing Big Data. The problem of gaining leads from far away countries can be solved by collecting contact data from sources like LinkedIn and using predictive analytics to determine which leads have the highest chance to become customers. Twitter is also a possible source to find news articles which help to overcome the problem of learning about construction project too late. However, a significant amount of effort needs to be invested from the company's side to utilize the full potential of the technology and actually gain useful insights. The demo implementation demonstrated the high need of domain knowledge as the lack of it was a main obstacle when trying to find relevant articles from Twitter. A detailed needs analysis is also an important aspect of purchasing a ready-made software or custom built solution. Regarding monthly fees or development investments, the high costs are the result of the difficulty of tackling Big Data problems. On top of the previously mentioned domain knowledge issue, managing parallel processing is far from straightforward. In case of data streaming, the technological complexity is further increased. Adding text mining to the mix adds another twist, meaning that building such a system needs sufficient hardware resources and also hours and hours of work of experts in the field who are hard to find at this point.

On a more **general level**, the report showed the current software market related to Big Data and market intelligence. Furthermore, it indicated the challenges that a business-to-business company might face when trying to find relevant market information in social media using Big Data tools. The opportunities of open source tools was also demonstrated.

The main **practical value** provided by the thesis project is two-fold: first, the experiences and the code from the proof of concept is handed over to the Big Data – Big Business project for further development. The other practical outcome is the set of insights and suggestions provided for the target company.

# 5 Conclusion

The **main theme** of the thesis project was Big Data. The topic is interesting because it is fashionable, ambiguous and carries great opportunities. The definitions and concepts related to the topic were clarified in the theory part, and the practical applicability was investigated through the company case.

The **company case** was challenging because they operate in a niche business-to-business market. Big Data solutions were found to provide value in all three areas of interest. Lead generation was a field that could benefit the most from the Big Data innovations, like social media monitoring and predictive analytics. In addition to the currently used tool, a custom solution can also be built based on the demo, which also provided information about the risks and challenges involved in a possible project.

Besides the company suggestions, the **outcomes** of the thesis project were the findings of the software market research and the experiences of the demo implementation which can be developed further by the Big Data – Big Business project. The main improvement to increase the usefulness of the program is to upgrade the filtering system by the implementation of text mining.

## References


- Abbott, D. , W. J. 2014. Applied Predictive Analytics. John Wiley & Sons Incorporated. Somerset.
- Apache Software Foundation. 2017. Apache Spark – Lightning-fast cluster computing. URL: http://spark.apache.org. Accessed: 5.10.2017
- Asay, M. 6.4.2016. Big data is all about the cloud. Info World. URL: http://www.infoworld.com/article/2905917/big-data/big-data-is-all-about-the-cloud.html. Accessed: 6.6.2017.
- Baesens, B. 2014. Analytics in a Big Data World. John Wiley & Sons Incorporated, Somerset.
- Bocij, P. Greasley, A. & Hickie, S. 2015. Business information systems: Technology, development and management for the e-business.Pearson. Harlow.
- Bradley, T.15 July 2015. Gartner and its Magic Quadrant challenged by crowdsourced market research. Techspective.net. URL: http://techspective.net/2015/07/15/gartner-and-its-magic-quadrant-challenged-by-crowdsourced-market-research/. Accessed: 11.04.2017.
- Brinker, S. 10.5.2017. Marketing Technology Landscape Supergraphic (2017): Martech 5000. Chiefmartec.com. URL: http://chiefmartec.com/2017/05/marketing-techniology-landscape-supergraphic-2017/. Accessed: 16.8.2017.
- Bronson, N. & Wiener, J. 22.10.2014. Facebook's Top Open Data Problems. Facebook Research. URL: https://research.fb.com/facebook-s-top-open-data-problems/. Accessed: 12.10.2017.
- Cavusgil, S.T., Knight, G. & Riesenberger, J.R. 2014. International Business: the new realities. Pearson. Harlow.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. CRISP-DM 1.0. Step-by step data mining guide. CRISP-DM consortium. URL: https://www.the-modeling-agency.com/crisp-dm.pdf. Accessed: 28.8.2017.
- Cuesta, H. 2013. Practical Data Analysis. Packt Publishing. Olton. Available from: ProQuest Ebook Central. [20 March 2017].
- Dean, J. 2014. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons Incorporated. Somerset.
- Dearborn, J. 2015. Data Driven. John Wiley & Sons Incorporated. Somerset. Available from: ProQuestEbook Central. [26 April 2017].
- Devlin, B. 2013. Business unIntelligence. Technics Publications. New Jersey.

- Dewhurst, M., Harris, J. & Heywood, S. 2012. The global company's challenge. McKinsey Quarterly. URL: http://www.mckinsey.com/business-functions/organization/our-insights/the-global-companys-challenge. Accessed: 29.8.2017.
- Doorley, J. & Garcia, H. F. 2015. Reputation Management. The key to successful public relations and corporate communication. Routledge. Abingdon.
- Dunn, J. E. 9 May 2013. G2 Crowd challenges Gartner's Magic Quadrant with crowdsourced grids. Techworld. International Data Group. URL: http://www.techworld.com/news/apps/g2-crowd-challenges-gartners-magic-quadrant-with-crowdsourced-grids-3446403/. Accessed: 11.04.2017.
- Eades, K. M. & Sullivan, T:T: 2014. The Collaborative Sale: Solution Selling in a Buyer Driven World. John Wiley and Sons Inc. Hoboken.
- EMC. 2014. Press Release: Digital Universe Invaded by Sensors. URL: https://www.emc.com/about/news/press/2014/20140409-01.htm Accessed: 21.03.2017.
- EMC. 2015. Data Science and Big Data Analytics. John Wiley & Sons Incorporated. Somerset.
- Finn, A., Vredevoort, H., Lownds, P. & Flynn, D. 2012. Microsoft private cloud computing. John Wiley & Sons Inc. Indianapolis.
- Forte, R. M. 2015. Mastering Predictive Analytics with R. Packt Publishing. Birmingham.
- G2 Crowd. 2017a. G2 Scoring Methodologies. URL: https://www.g2crowd.com/static/g2_grid_scores. Accessed: 10.4.2017.
- G2 Crowd. 2017b. All software categories. URL: https://www.g2crowd.com/categories. Accessed: 25.04.2017.
- G2 Crowd. 2017c. Best Sales Intelligence Software. URL: https://www.g2crowd.com/categories/sales-intelligence. Accessed. 8.5.2017.
- Gartner. URL: https://www.gartner.com/technology/research/methodologies/research_mq.jsp. Accessed: 20.10.2017.
- Guiliano, S. 25.5.2016. 5 ways cloud Big Data solves business challenges. Atlantic.net. URL: https://www.atlantic.net/blog/5-ways-cloud-big-data-solves-business-challenges. Accessed: 6.6.2017.
- Harford, T. 28.3.2014. Big Data: are we making a big mistake?. URL: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html. Accessed: 5 April 2016.
- Hedin, H., Hirvensalo, I. & Varnaas, M. 2014. The Handbook of Market Intelligence. John Wiley & Sons Incorporated. Somerset. Available from: ProQuest Ebook Central. [9 March 2017].

- Hortonworks 2016. Data Architecture Optimization. Hortonworks Inc. URL: http://info.hortonworks.com/rs/549-QAL-086/images/hortonworks-data-architecture-optimization.pdf. Accessed: 3.4.2017.
- Hortonworks.com. URL: https://hortonworks.com/products/data-center/hdp/. Accessed: 12.10.2017.
- Håkansson, C. & Nelke, M. 2015. Competitive Intelligence for Information Professionals. Chandos Publishing, Elsevier Ltd. Kidlington.
- IBM. What is Big Data? URL: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. Accessed: 21.03.2017.
- Ishida, R. 16.4.2015. Character encodings for beginners. W3C. URL: https://www.w3.org/International/questions/qa-what-is-encoding. Accessed: 17.8.2017.
- Jain, P., Jayaraman, P. & Sharma, P. 2014. Behind Every Good Decision : How Anyone Can Use Business Analytics to Turn Data into Profitable Insight. AMACOM. New York.
- Liebowitz, J. 2014. Business Analytics: An Introduction. Auerbach Publications. Boca Raton.
- Loshin, D. 2013. Big Data Analytics. Elsevier Science. Saint Louis.
- Marinescu, D. C. 2013. Cloud computing. Elsevier Science. Waltham.
- Marr, B. 2015. Big Data: Using Smart Big Data, Analytics And Metrics To Make Better Decisions And Improve Performance. John Wiley & Sons Incorporated. Somerset. Available from: ProQuest Ebook Central. [20 March 2017].
- Mell, P. & Grance, T. 2011. The NIST definition of cloud computing. National Institute of standards and technology. URL: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf. Accessed: 12.6.2017.
- Minelli, M., Chambers, M. & Dhiraj, A. 2012. Big Data, Big Analytics. John Wiley & Sons Incorporated. Somerset.
- Pasian, B. 2015. Designs, methods and practices for research of project management. Gower Applied Business Research. Farnham.
- Riel, C.B.M.v. & Fombrun, C. J: 2007. Essentials of corporate communication: Implementing practices for effective reputation management. Routledge. Abingdon.
- Roetzer, P. 2014. The Marketing Performance Code: Strategies and Technologies to Build and Measure Business Success. John Wiley and Sons Inc. New York.
- Rowley, J. 2007. The wisdom hierarchy: Representations of the DIKW hierarchy. Journal of Information Science, 33, 2, pp. 163–180.

- Shafer, T. 1.4.2017. The 42 V’s of Big Data and Data Science. Elder Research Data Science and Predictive Analytics. URL: https://www.elderresearch.com/company/blog/42-v-of-big-data. Accessed: 13.10.2017.
- Siegel, E. 2013. Predictive Analytics. John Wiley & Sons Incorporated. Somerset.
- Systrom, K. 10.12.2014. 300 million: sharing real moments. Instagram Press. URL: https://instagram-press.com/blog/2014/12/10/300-million-sharing-real-moments/. Accessed: 12.10.2017.
- Twitter. 2017. Filter real time Tweets – Parameters. Twitter Developers. URL: https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html. Accessed: 3.10.2017.
- TwitterEng. 30.6.2011. 200 million tweets per day. Twitter Blog. URL: https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html. Accessed: 12.10.2017.
- Vacca, J. R. 2016. Cloud computing security. CRC Press. Boca Raton.
- Zinck, B. M. 24 February 2016. The value of peer reviews in the software decision process. Diginomica. URL: http://diginomica.com/2016/02/24/the-value-of-peer-reviews-in-the-software-decision-process/ . Accessed: 11.4.2017.

## Appendix 1. – Commercial software

| Sales Intelligence | | | |
|---|---|---|---|
| Company name | Websites | Methods | Purpose |
| artesian | http://www.artesian.co/ | Text analysis | Lead generation, Competitive intelligence |
| Aviso | https://www.aviso.com/ | Artificial intelligence | None |
| Bullhorn | https://www.bullhorn.com/ | Text analysis | Customer intelligence |
| Buzzboard | https://www.buzzboard.com/ | Database, Predictive analytics | Lead generation |
| charlie app | https://www.detective-labs.com/ | Text analysis | Lead generation |
| Clari | http://www.clari.com/ | Artificial intelligence | Lead generation |
| Clearbit | https://clearbit.com/ | Database | Lead generation |
| D&B Hoovers | http://www.hoovers.com/ | Database, Predictive analytics | Lead generation, Competitive intelligence |
| data.com | https://www.data.com/ | Database | Lead generation |
| datafox | https://www.datafox.com/ | Database, Predictive analytics | Lead generation |
| datahug | https://datahug.com/ | Predictive analytics | Lead generation |
| DiscoverOrg | https://discoverorg.com/ | Database, Predictive analytics | Lead generation |
| FullContact | https://www.fullcontact.com/ | Database | Lead generation |
| Growbots | https://www.growbots.com/ | Contact monitoring, Artificial Intelligence | Lead generation |
| Gryphon Sales Intelligence | https://gryphonsalesintelligence.com/ | Predictive analytics | Lead generation |
| HG Data | https://www.hgdata.com/ | Database | Lead generation |
| hunter | https://hunter.io/ | Database | Lead generation |
| IKO systems | http://www.iko-system.com/ | Database, Predictive analytics | Lead generation |
| InsideView | https://www.insideview.com/ | Database, Predictive analytics | Lead generation |
| Intricately | https://www.intricately.com/ | Database | Lead generation |
| lead leaper | http://leadleaper.com/ | Database | Lead generation |
| lead411 | https://www.lead411.com/ | Contact monitoring | Lead generation |
| LinkedIn Sales Navigator | https://business.linkedin.com/sales-solutions/sales-navigator# | Database | Lead generation |
| LiveHive | https://livehive.com/ | Predictive analytics | Lead generation |
| MediaRadar | https://mediaradar.com/ | Database, Predictive analytics | Lead generation |
| Nimble | https://www.nimble.com/ | Contact monitoring, Predictive analytics | Lead generation |
| Owler | https://www.owler.com/ | Text analysis | Lead generation |
| RainKing | https://www.rainkingonline.com/ | Database, Predictive analytics | Lead generation |

| | | | |
|---|---|---|---|
| Relpro | http://relpro.com/ | Contact monitoring | Lead generation |
| S&P Global | https://www.spglobal.com/ | Text analysis | Lead generation |
| Salesgenie | https://www.salesgenie.com/ | Database, Predictive analytics | Lead generation |
| SalesLoft | https://salesloft.com/ | | Lead generation |
| Salestools.io | https://salestools.io/ | Database | Lead generation |
| ScanBizCards | https://www.scanbizcards.com/dymo/ | Machine learning | None |
| Sendbloom | https://www.sendbloom.com/ | Database | Lead generation |
| tellwise | https://tellwise.com/ | Predictive analytics | None |
| Termscout | http://www.termscout.com/ | Database | Lead generation |
| vainu.io | https://vainu.io/ | Database, Predictive analytics | Lead generation |
| Winmo | https://www.winmo.com/ | Database, Predictive analytics | Lead generation |
| ZoomInfo | https://www.zoominfo.com/ | Database | Lead generation |
| The List Inc | https://www.thelistinc.com/ | Database, Predictive analytics | Lead generation |
| datanzye | https://www.datanyze.com/ | Database | Lead generation |
| Oceanos | https://www.oceanosinc.com/ | Database, Predictive analytics | Lead generation |

| **Lead Intelligence** | | | |
|---|---|---|---|
| Company name | Website | Metods | Purpose |
| datanzye | https://www.datanyze.com/ | Database, predictive analytics | Lead generation |
| The List Inc | https://www.thelistinc.com/ | Database, predictive analytics | Lead generation |
| LeadIQ | https://leadiq.com/ | Database | Lead generation |
| synthio | https://synthio.com/ | Database | Lead generation |
| lead411 | https://www.lead411.com/ | Database | Lead generation |
| Rocket Reach | https://rocketreach.co/ | Real-time contact retrival | Lead generation |
| Lead Fuze | https://www.leadfuze.com/ | Database | Lead generation |
| Growlabs | https://www.growlabs.com/ | Database, predictive analytics | Lead generation |

| **Lead Mining** | | | |
|---|---|---|---|
| Company name | Website | Methods | Purpose |
| Oceanos | https://www.oceanosinc.com/ | Predictive analytics, Database | Lead generation |
| LeadGnome | https://www.leadgnome.com/ | Text mining | Lead generation |
| Kickfire | https://www.kickfire.com | Predictive analytics, Database | Lead generation |
| AeroLeads | https://aeroleads.com/ | Database | Lead generation |
| Socedo | https://www.socedo.com | Database | Lead generation |

| | | | |
|---|---|---|---|
| Builtwith | https://builtwith.com/ | Database | Lead generation |
| VisitorTrack | http://netfactor.com/ | Database | Lead generation |
| Growlabs | https://www.growlabs.com/ | Predictive analytics, Database | Lead generation |

| **Social Media Monitoring** | | | |
|---|---|---|---|
| Company name | Website | Code | Purpose |
| sprout social | https://sproutsocial.com/ | Tracking & search | Customer intelligence |
| Hootsuite | https://hootsuite.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Cision | https://www.cision.com | Sentiment analysis, competitor benchmarking | Competitor intelligence |
| Sysomos | https://sysomos.com/platform/listen/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Tracx | https://www.tracx.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Spredfast | https://www.spredfast.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Brandwatch | https://www.brandwatch.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Crimson Hexagon | https://www.crimsonhexagon.com/ | Natural language processing | Customer intelligence, Competitor intelligence |
| NetBase | https://www.netbase.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Synthesio | https://www.synthesio.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Adobe Social | http://www.adobe.com/hu/marketing-cloud/social.html | Sentiment analysis | Customer intelligence |
| Mention | https://mention.com/en/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Sprinklr | https://www.sprinklr.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Audiense | https://audiense.com/ | Machine learning | Customer intelligence, Competitor intelligence |
| CX Social | https://cxsocial.clarabridge.com/ | Topic recognition | Customer intelligence |
| Radarly | https://linkfluence.com/ | Topic recognition | Customer intelligence |
| Oktopost | https://www.oktopost.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Agora-Pulse | https://www.agorapulse.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Sendible | https://sendible.com/ | | None |
| Socialbakers | https://www.socialbakers.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| digimind social | http://www.digimind.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Percolate | https://percolate.com/ | Topic recognition | None |
| Falcon.io | https://www.falcon.io/ | Sentiment analysis, Topic recognition | Customer intelligence |

| | | | |
|---|---|---|---|
| Infegy | https://infegy.com/ | Sentiment analysis, Topic recognition | Customer intelligence |
| NUVI | https://www.nuvi.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Tailwind | https://www.tailwindapp.com/ | Topic recognition | None |
| BuzzSumo | http://buzzsumo.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Cyfe | https://www.cyfe.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Klear | https://klear.com/ | Network analysis | Customer intelligence, Competitor intelligence |
| Moz | https://moz.com/ | | None |
| Plumlytics Social | https://www.plumlytics.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Talkwalker | https://www.talkwalker.com/ | Sentiment analysis, Machine learning | Customer intelligence |
| Mediatoolkit | https://www.mediatoolkit.com/ | Topic recognition | Competitor intelligence |
| Social Studio | https://www.salesforce.com/products/marketing-cloud/social-media-marketing/ | Topic recognition, Sentiment analysis | Customer intelligence, Competitor intelligence |
| brand24 | https://brand24.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Social Report | https://www.socialreport.com | Tracking & search | Customer intelligence, Competitor intelligence |
| Union Metrics | https://unionmetrics.com/ | Topic recognition | Customer intelligence, Competitor intelligence |
| Zignal Labs | http://zignallabs.com/ | Topic recognition | Competitor intelligence |
| Komfo | https://komfo.com/ | Topic recognition | None |
| BrandsEye | https://www.brandseye.com/ | Sentiment analysis | Customer intelligence, Competitor intelligence |
| Buzzlogix | https://buzzlogix.com/ | Sentiment analysis | Customer intelligence, Competitor intelligence |
| Ubermetrics | https://www.ubermetrics-technologies.com/ | Topic recognition | None |
| Zoho Social | https://www.zoho.com/social/ | Sentiment analysis, Tracking & search | Customer intelligence, Competitor intelligence |
| OutboundEngine | https://www.outboundengine.com/ | Topic recognition | None |
| eClincher | https://eclincher.com/ | Tracking & search | Customer intelligence, Competitor intelligence |
| Mentionlytics | http://www.mentionlytics.com/ | Topic recognition | Customer intelligence, Competitor intelligence |

## Appendix 2. – Open source software

| Software name | Website | Purpose for project | Value proposition | Methods, too |
|---|---|---|---|---|
| Spark | https://spark.apache.org/ | stream processing | faster processing than Hadoop | in-memory da streaming, SQ |
| Python | https://www.python.org/ | operating Spark, creating a website | easy to learn, general purpose and fast programming language | multi-paradigr |
| Hortonworks | https://hortonworks.com/ | Big Data processing | secure and open source Hadoop distribution | cluster proces |
| d3.js | https://d3js.org/ | data visualization | JavaScript library for powerful visualizations | HTML, SVG an |
| MongoDB | https://www.mongodb.com/ | storing processed data | own query language, scalability and flexible data models | NoSQL |
| Cassandra | http://cassandra.apache.org/ | storing processed data | high scalability, availability and fault tolerance | distributed sto gration with s |
| MySQL | https://www.mysql.com/ | storing processed data | cost effective, scalable database | InnoDB storag |
| MariaDB | https://mariadb.org/ | storing processed data | fast and scalable, with rich ecosystem of storage engines | XtraDB, Aria s |
| PostgreSQL | https://www.postgresql.org/ | storing processed data | reliability, data integrity, correctness | object-relatio for data acces |

## Appendix 3. - Technical details of proof of concept

Structure of this appendix is based on the CRISP-DM framework.

| Business Understanding | |
|---|---|
| **Business Objectives** | Objective: Gaining new customers in the construction and energy industries, especially abroad, first and foremost in Germany.<br>Success criteria: Generating leads with a high chance to turn into sales. |
| **Situation Assessment** | Available resources: ssh connection set up to Ubuntu 16.04 cloud environment, Python 2.7, Apache Spark 2.1.1, installed in the folder /usr/local/src/spark/spark-2.1.1-bin-hadoop2.7 with /usr/local/src/spark/spark-2.1.1-bin-hadoop2.7/bin added to the commands path for easy launch, Internet access.<br>Assumptions: employees of possible customers are active on social media, industry related news are shared on social media<br>Constraints: only English and German language, web crawling restrictions are respected |
| **Analytics Goals** | Goal: collecting relevant tweets from Twitter to find new construction projects or employees of possible customers<br>Success criteria: creating a list of relevant news and contact people |

| Data Understanding | |
|---|---|
| **Initial data collection** | Actions:<br>• Twitter account setup: registration, application creation on apps.twitter.com, saving tokens and keys for Oauth2 identification<br>• Spark Streaming setup: Spark Python API does not have Twitter support. Workaround: installation of a Python library for streaming Tweets, creation of Python application to send Tweets to Spark through a TCP socket connection on localhost port 5555. The application code was based on the Tweepy connector on the page awesomestats.in/spark-twitter-stream/. However, Tweepy had unexpected errors which turned out to be bugs, so a switch was made to Twython and the connector application had to be reworked.<br>At this stage a demo file was used to test connectivity. The location of the demo file is /usr/local/src/spark/spark-2.1.1- |

| | |
|---|---|
| | bin-hadoop2.7/examples/src/main/python/streaming/network_wordcount.py. The Twython application is run first from a terminal window, and then the spark application is run from another terminal window. The Twython application waits for Spark Streaming's request for a connection to start streaming the tweets. The command spark-submit is used to launch the spark application. The whole launch command looks like this:<br><br>spark-submit \<br>--master local[4]\<br>/usr/local/src/spark/spark-2.1.1-bin-hadoop2.7/examples/src/main/python/streaming/network_wordcount.py\<br>localhost 5555<br><br>The command is broken up into several lines for better readability and backslashes are used to escape the new line characters. The "–master local[4]" option specifies that the application is going to run locally instead of on a cluster and 4 threads are allocated for the processing. Four is a standard number to launch demo applications with, because it ensures that spark has just enough resources to run properly. If only one thread is allocated, it is used by the receiver to accept the inflow of data and there is no processing power left for the actions and the output. Next parameter is the location of the example file, and finally, the host and port where the input data is coming from, in this case the Twython application.<br><br>Outcome: Functioning workaround to send tweets into Spark Streaming, Spark processing data on the run, no saved dataset. |
| **Data description** | The format of the streamed data coming from Twitter is JSON. There are roughly 30 attributes and they are specified on the twitter API documentation page (https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object). Most of the attributes are optional. The compulsory attributes include: |

| | |
|---|---|
| | - tweet id<br>- timestamp of the creation<br>- user who shared it<br>- message<br>- retweeting and quoting information<br>There are two kinds of location data stored, coordinates and place. Coordinates contains latitude and longitude data provided by the application and it means the location the tweet was created at. However, the place attribute stores data that the user supplies in the location tab when tweeting, which means that this is not necessarily the location the tweeter is at but a location they are tweeting about. Both of these are optional, nullable attributes.<br><br>The Twython connector in its original state interprets the JSON data with the built-in Python module “json”, and only sends the message text to the demo Spark Streaming application. |
| **Data exploration** | Actions:<br>- Twitter in-site search: experimentation with keywords to find relevant data<br>- Examining the JSON data: modification of the Twython app to extract the whole JSON objects not just he messages<br>- Adjusting initial filter: it is impossible to get all tweets real time. The Twitter API forces and initial filter to fulfil requests. This can be based on keywords, location, username, etc., as specified on developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html . Attempted filters: keywords ( “oil”, “energy”, “construction”, “hvac”, “contract”), location parameters ( bounding box for Germany, USA, Europe ). Explanation of bounding box: developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters.<br>Outcomes: There are tweets related to energy and construction industry, but their frequency is lower than expected – one in a month or less. Twython connector sends JSON data to Spark instead of just the message strings. The frequency of German tweets is quite low, which results in occasional empty lines sent by the Twitter API to keep the connection alive. Europe and the USA have a suitable |

| | |
|---|---|
| | frequency for analysis. Searching by key words results in the most intense inflow. Filtering by language is not working. |
| **Data verification** | Issues:<br>• Location data: when streaming with a key word filter most tweets don't have any location data (guess: about 1 in 20 has it)<br>• Character encoding: problem inherent to Python 2. Occurs not only in German but also in English text due to Unicode emoji characters. The main problem is Python 2's string handling, which doesn't make it specific in string operations if a string is Unicode or byte string. Without special care during programming this results in characters viewed incorrectly or the applications breaking down with encoding errors. More info: www.azavea.com/blog/2014/03/24/solving-unicode-problems-in-python-2-7<br>• Quotes within messages: single and double quote characters occur within tweets and they break Spark Streaming's JSON interpreter<br>• Lang attribute: usually a language code of two characters, but often states "und" for "unidentified"<br>• Date format: the date format in the JSON data is different from the MySQL date format |

| **Data Preparation** | |
|---|---|
| **Selection** | Actions:<br>• Choosing initial filter for Twython connector<br>• Choosing attributes: modification of Twython connector to send JSON data with only the most important attributes to Spark Streaming<br>• Applying the language filter in Twython connector<br><br>Outcome: initial filter is based on location - bounding box of Europe (locations="-25.9,35.5,53.7,69.7"). Twython connector sends six attributes to Spark:<br>• id: string representation of an integer of size between 53-64 bits, extracted for using as primary key later<br>• text: UTF-8 encoded string of the actual tweet message |

| | |
|---|---|
| | • user: a User object with several attributes, only the screen_name is extracted<br>• created_at: string containing the UTC time when the tweet was created<br>• lang: string with the BCP 47 language id of the language the tweet is written in<br>• place: a Place object with more attributes, only the city and country_code are extracted – both are strings<br>Languages other than English and German are not forwarded. |
| **Cleaning** | Actions:<br>• Location data: setting the initial filter based on location ensures that all tweets contain the place attribute<br>• Encoding issues: incoming UTF-8 coded or other text is immediately decoded to form Unicode strings, working only with Unicode strings in the application and encoding them back to UTF-8 on output.<br>• Quotation marks: Twython connector strips quotation marks from tweet messages<br>• Unidentified language: not forwarded to Spark Streaming<br>• Date format: transform the created_at attribute to fit the MySQL DATETIME format using string operations in the Twython connector |
| **Construction** | Actions:<br>• Filtering relevant tweets - Attempt #1: modification of the JSON strings in Spark Streaming by adding three more attributes: “oil”, “energy”, “construction”, and their German equivalents with values: 0 and 1, identifying the presence of the term or the lack of it in the tweet.<br>• Filtering relevant tweets - Attempt #2: Spark Streaming and regular expression search within all attributes of a tweet (including user name, place, etc.) for the key words. When the filter finds a tweet with the keyword in it, Spark Streaming saves it in the corresponding table in a MySQL database. To write data to Spark Streaming, the jdbc connector of MySQL has to be installed and added to the command when launching the Spark application. The jdbc connector was installed in the /home/vivien/thesisFiles directory, so |

the new launch command looks like this:

spark-submit \
--master local[4]\
--driver-class-path /home/vivien/thesisFiles/mysql-connector-java-5.1.44-bin.jar --jars /home/vivien/thesisFiles/mysql-connector-java-5.1.44-bin.jar \
/usr/local/src/spark/spark-2.1.1-bin-hadoop2.7/examples/src/main/python/streaming/network_wordcount.py\
localhost 5555

MySQL also needed to be configured to bind to the external IP address, so that it enables remote connections. The configuration files can be in many locations, in my installation it is the /etc/mysql/mysql.conf.d/mysqld.cnf file that contains the bind-address. A specific user was created for spark with permissions to modify the database. Detailed tutorial on dev.mysql.com/doc/refman/5.7/en/adding-users.html .

There are four tables in the Tweets MySQL database: tweets (early testing), oil, energy and construction. There aren't any relations between the tables. They share the same schema:

- tweet_id: varchar(25), not null, primary key
- date_time: datetime
- lang: char(2)
- text: varchar(250)
- city: varchar(30)
- user: varchar(15)
- country: varchar(2)

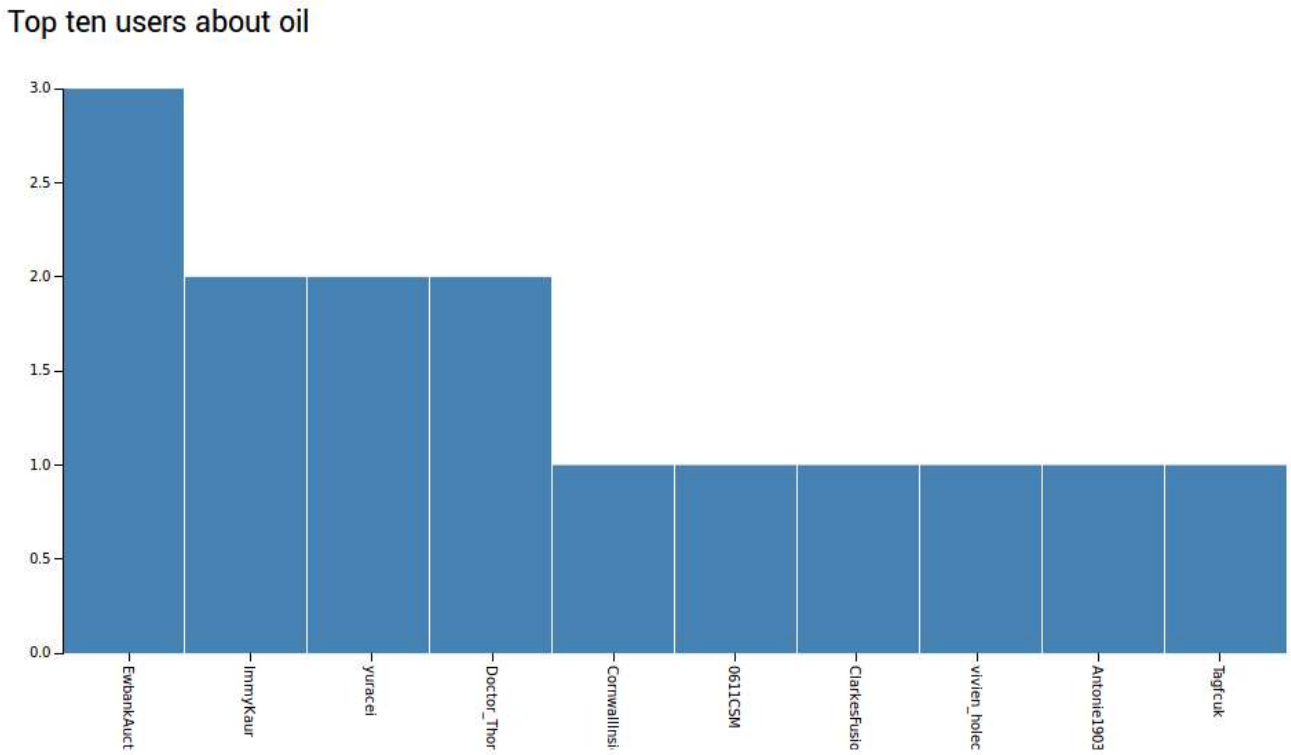
- Accessing information: setting up a homepage to display the gathered data using a private Digital Ocean server. To continue with the Python language while enabling database connection, a webpage framework called Flask is used. Setup of Flask and MySQL are detailed in this tutorial: pythonprogramming.net/creating-first-flask-web-app/. To query the MySQL data in the Flask application, the Python library MySQLdb has to be installed. Default encoding of MySQL is

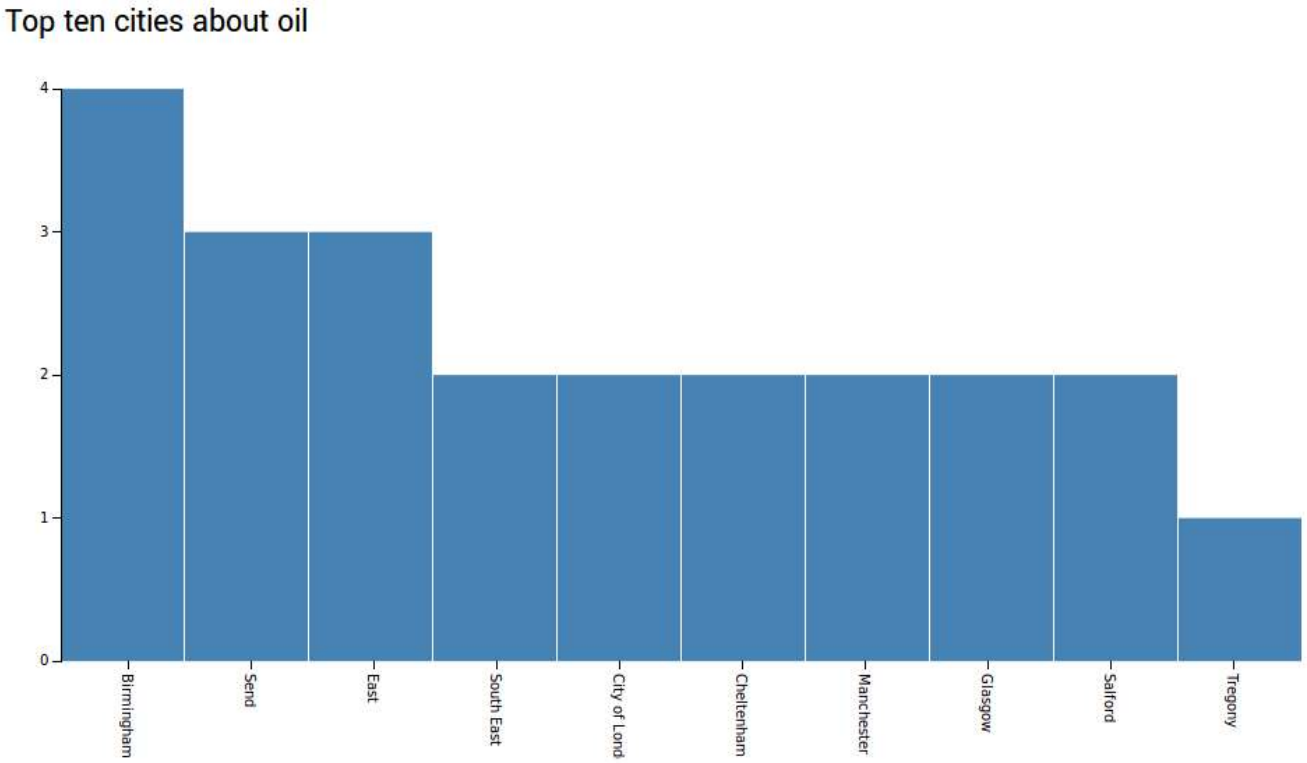
| | |
|---|---|
| | Latin-1 while the whole pipeline uses UTF-8, so query results need to be decoded from Latin-1 instead of UTF-8. Derivatives of the data are created when querying the tables by viewing the output page. Visualizations were attempted with the pandas and Bokeh Python libraries. Another tool for visualization is d3.js - this tutorial: bost.ocks.org/mike/bar is a good introduction.<br><br>Outcomes: first approach for filtering was prone to errors related to string concatenation, and it wasn't effective, since the number of tweets containing one of these key words was very low compared to all the English and German tweets coming from Europe. Second approach results in a working solution but with a lot of noise ending up in the database due to the primitive filter. Flask webpage with database connection is functioning. Visualizations are created with d3.js due to installation problems with pandas (Apache2 web server can't detect it). They can be found in Appendix 4. |

| Evaluation | |
|---|---|
| **Assessment** | Successes: functioning infrastructure including the streaming of tweets from Twitter, filtering them, writing them to a database, showcasing the results on a website including visualization.<br>Defects: The solution at this stage does not provide any useful leads, since the filtering mechanism is very simplistic. |
| **Next steps** | Handing over the solution to the Big Data project: duplicating the MySQL database and the homepage currently in the Digital Ocean server in HAAGA-Helia's csc cloud environment. |

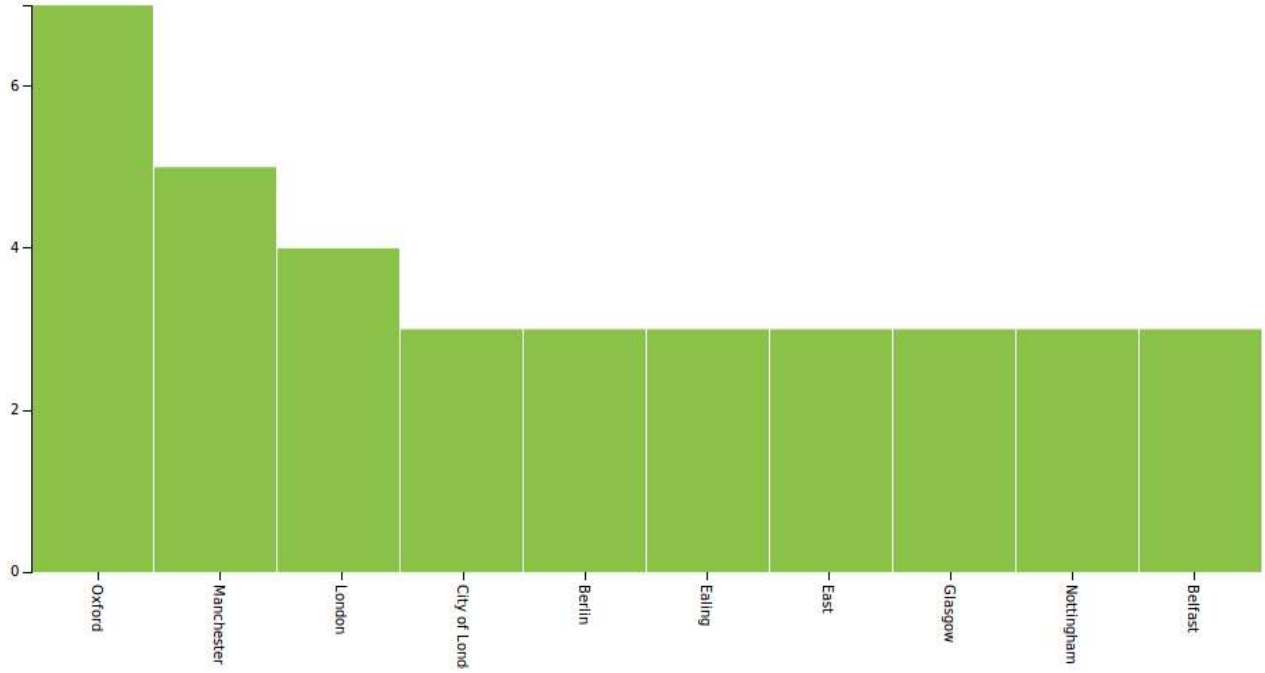
## Appendix 4. - Visualizations

The visualizations created from the database with the d3.js JavaScript library are located here. The findings consist of three topics: energy, construction and oil. For each topic the most frequent tweeters and the cities that have the most tweets on the topic are displayed.

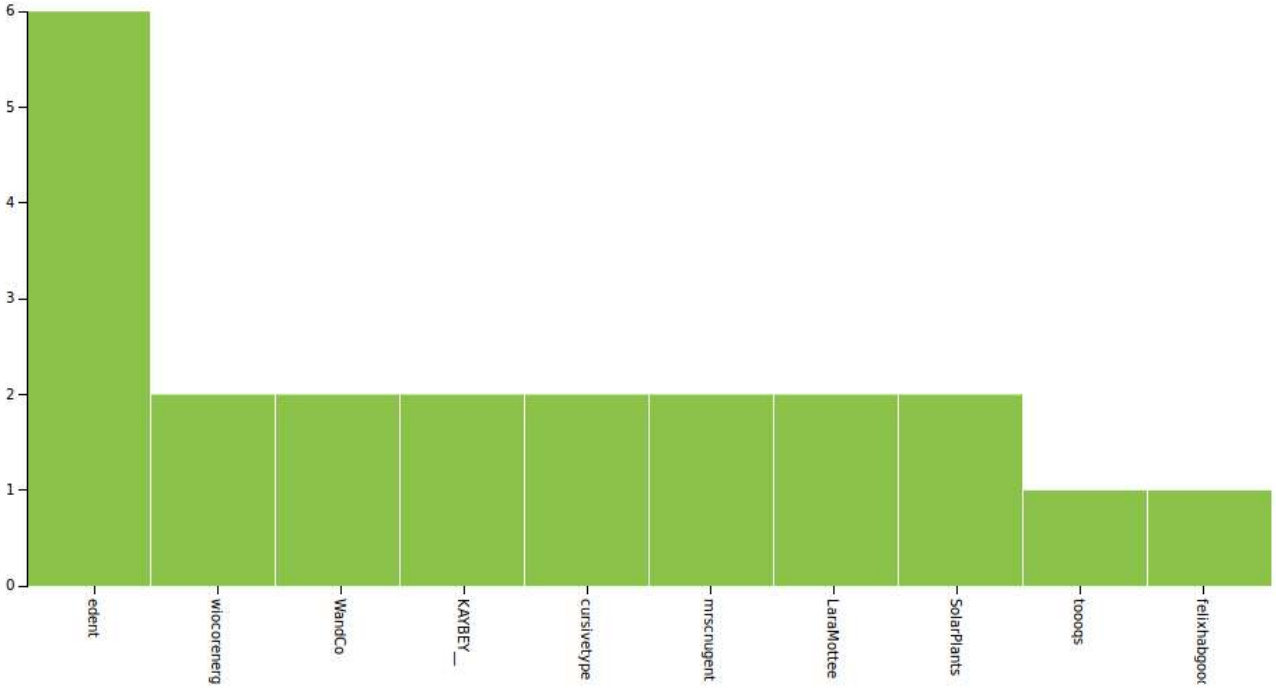


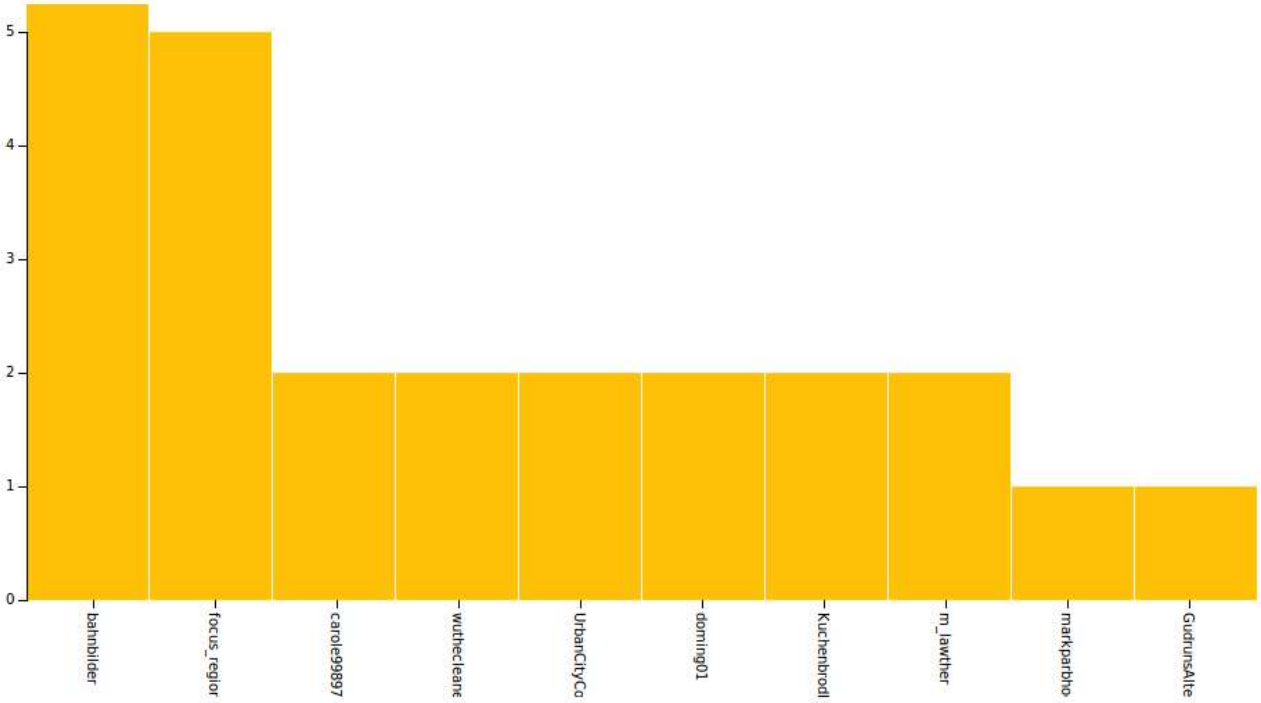
## Top ten cities about energy




## Top ten users about energy



Top ten users about construction




Top ten cities about construction