Haaga-Helia
ammattikorkeakoulu Oy

Leveraging Real-Time Big Data analytics in a Modern Telecom environment

Jyri Hyppönen

Opinnäytetyö
Tietojenkäsittelyn koulutusohjelma
2016

Haaga-Helia
University of Applied Sciences

Abstract

Date: 29.11

Author(s)
Jyri Hyppönen

Degree programme
Heti 09

Report/thesis title
Leveraging Real-Time Big Data analytics in a modern Telecom environment

Number of pages and appendix pages
39 + 0

Big Data is not a new challenge, and nowadays the focus has shifted from getting results to getting results fast.
For analytics, faster is always better because faster reaction time improves many situations, such as detecting network faults. Faster speed of detecting allows for more time to minimize the impact of the incident.

However, reaching real-time analytics is not as simple due to many Big Data technologies just were not designed with speed in mind. Thankfully the value of faster Big Data applications has not gone unnoticed and there are currently multiple interesting applications that can help with faster processing or straight out streaming of data.

Difficulty can be selecting the right technology for the use case. Especially since different enterprises often have different business and technical requirements and platforms that they use.

The challenges were mapped by interviewing few key people in the organization.

For this case company Apache Spark seemed to be most suitable application for real-time analytics as it offers fast processing speed, streaming, is supported by the Hadoop stack they use and uses Java.

Keywords
Big Data. Real-Time Analytics. Apache Hadoop. Apache Spark. Telecom. Apache Kafka.

Table of Contents

1 Intro

I decided that subject of my thesis will be about data already back in 2013 when I got into data analytics and business intelligence because of my, at the time, new job as a chief data analyst in a medium sized startup company. After this, roughly one-year experience, my career and my studies have heavily focused on data.

My perspective towards data has varied depending on my current work position and the subjects I've studied, but I can easily say that I have developed a keen interest towards data, analytics, business intelligence and the future of the field.

I chose this particular topic regarding data due to its current relevance, as it is still a hot topic and due to the fact that my thesis work will be beneficial for my current employer.

Before starting to dig deeper into what Big Data is, it is important to understand what data is. Russell Ackoff (1989, 1.) defines data as symbols which represent properties of an object or an event. When data is processed, which usually means that some kind of calculations are performed on it, it is turned into information which allows for more effective understanding of the same properties.

Essentially data is raw form of information that needs to be processed before it is useful.

The term, Big Data, has been used to describe large amounts of data and the challenges that arise from processing those vast datasets since the year 1998. (Lohr 2014.) However, the technologies, such as Apache Hadoop, designed to solve these issues have not been around for more than 10 years and ready commercial solutions are even more recent. (Hadoop 2007.)

Big Data has been surrounded by lot of hype in the media with much of it being unwarranted due inflated expectations from the capabilities of Big Data technologies and benefits they can create. It was suggested that in order for Big Data to be considered a proper innovation, companies would need to create innovative and solid business models that actually can be proven to create value for their customers. Big Data should be considered as the basis for success, not guarantee of it. (Buhl, Röglinger & Moser 2013, 68.)

Recently Big Data has been again revalidated by the new accessibility and availability to data which allows for new use cases to be created.
For example, servers are now able to send monitoring data to a centralized service that will be able to monitor all logs from all the connected servers and benchmark their performance against each other or predict failures and alert engineers. (Dodson, S. 2014.)

These new opportunities come with new challenges, one of which is particular interest of this thesis work: How can be these vast datasets processed in a real time fashion?

1.1 Goals

The purpose of this thesis work is to examine the current capabilities of Big Data technologies and how they match the increasing needs of the business owners from the point of view of a modern telecom company with emphasis on real time analytics.

The theory part contains general information about Big Data, Big Data technologies, how Big Data can be used, Big Data process, definition of 'real-time', what kind of issues companies have faced with real time analytics and overview of Big Data technologies.

The empirical part will apply the research of the theory part to the case company's situation, which will be briefly explained, and the end result of the thesis work is an implementation plan of which technologies or techniques can be used to reach business requirements.

1.2 Scope

This thesis will focus on retrieving or streaming data from the Data Lake, issues and methods related to saving the data to the Data Lake will be briefly covered, but will not be deeply analyzed by this thesis work.
Legal issues related to the storage and usage of data will not be covered.

1.3 **Target Company**

The target company for which this thesis work is conducted is a large telecom company with over one thousand employees in Finland.

The Big Data platform the company uses is developed by another organization within the company.

The perspective of this thesis work will be from the team that leverages the Big Data platform and develops on top of it to reach business goals.

2 Big Data?

2.1 What is Big Data?

It is important to understand the reasons why Big Data presents a challenge, but also why solving this challenge can be beneficial.

A literature review on Big Data papers by Andrea De Mauro, Marco Greco, and Michele Grimaldi, concludes that Big Data can be formally defined as follows:

> "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" (De Mauro, Greco & Grimaldi, 2015, 8.)

As stated in the quote above, Big Data can be defined with the three terms volume, velocity and variety. Recently, a fourth term, variability, has been suggested to be added alongside the three core terms as a key definition.

2.1.1 Volume

It has been predicted that on 2025 online data growth will reach 180 zettabytes annually, huge growth from the comparably small 44 zettabytes in 2013. (EMC Digital Universe with Research & Analysis by IDC, 2014.) (Kanellos 2016.)
1 zettabyte is 1 trillion gigabytes and one gigabyte is 640 web pages (with 1.6MB average file size)

The rapid growth of online data can be attributed to the following factors: Storage has become cheaper than ever before, online capabilities have improved to allow for increased data collection, more and more people being connected to the internet, being able to use the internet in increasingly varied ways and whole new industry of enterprises generating and collecting data online, appearing in relative short time. (Dumbill 2012.)

There sheer quantity of the data is so massive that the standard ways to process it are unable to cope with the task and so parallel processing is required. (Dumbill 2012.)

As the quantities increase also the need for scalable storage and distributed approach to querying becomes more relevant in order to gain results in a timely fashion. (Dumbill 2012.)

The ability to process and analyze large datasets that could not be analyzed with the usual methods, provides better accuracy for forecasts and analyses as the sample size can be larger and more factors can be taken into account. (Dumbill 2012.)

Many companies already store large volumes of data, but lack the ability to process these datasets, as standard relational database infrastructures cannot handle the operations in reasonable time. (Dumbill 2012.)

2.1.2 Velocity

One of the reasons why datasets have been growing so large, is that more data is being collected more frequently and the data collected is more complex. Faster collection of data is due to faster internet infrastructure, increased usage of online services, mobile devices and sensors. For example, modern cars can contain 100s of sensors. (IBM 2013.)

The increased frequency of data is being saved is enough to cause issues for conventional relational data infrastructures, but velocity also refers to the ability to extract value from the saved data in a manner that is quick enough. Both saving and processing the data have to be fast and accurate enough to fulfill the requirements of the business process trying to benefit from the data. This is called the feedback loop. (Dumbill 2012.)

One relevant example regarding this was in the IBM ad about the use of traffic data:

> “You would not trust 5-minute old traffic data to cross the road, would you?” (IBM Video 2010.)

Specialized industries, such as financial traders, have taken advantage of quickly growing datasets for many years and now the technology has been made available for more general usage, ranging from traffic data to the data generated by the Large Hadron Collider. (Dumbill 2012.)

Most common term for accessing and using stored data is called streaming. Streaming is necessary when data comes in at such speeds that some processing is required before the data is stored, for example data that requires an immediate response such as mobile application and online gaming data. (Dumbill 2012.)

Bottom line is: The faster a company can act on data the better competitive advantage they have. (Dumbill 2012.)

2.1.3 Variety

Taking advantage of data is not a new thing, business intelligence can be traced back to Principles of Scientific Management from 1911 (Taylor, 1911.) and the time management operations that it defined. Accounting and mathematics have been around for even longer.

Since the dawn of commercial computers around the 1960s analytics and business intelligence have taken a huge leap forward, but the data most commonly used has been in a rather strict, structured form. Such as rows in a database or numbers in a specified format. (Nawab 2012.)

Even today, many systems are still bound to strict relational databases and can only handle data in very specific, structured formats, which is why one of the larger issues with data today, is that it can be anything, in any structure and in any format. (Dumbill 2012.)

Big Data aims to save data as is, without trying to force it into specific form. This data can be numbers, text, pictures, music, video files or even feed from sensors and the saving process will be very similar for all types of data. (Dumbill 2012.)

The Big Data technologies aim to extract value from these different types of data, by processing them at the point when the data will be used, rather than when it is saved. This way there is no loss of source data as everything is saved as is and structure is specified at the point of usage. (Dumbill 2012.)

2.1.4 Variability

Variability is a rather new addition to the definition of Big Data.
It means that the other dimensions of the data can vary depending on time. For example, a set of JSON data with certain structure can be flowing into the data storage, but due to some changes in the source system, the structure of the data can change. This kind of variability in the structure of the data presents a unique challenge that Big Data solutions will have to be able to work around (DeVan, 2016; McNulty 2014; National Institute of Standards and Technology 2015, 4, 15.)

2.1.5 Additional definitions

There are few other terms used to describe Big Data, but they are not as widespread as the four terms discussed above.

These additional definitions include: Veracity, Visualization and Value.

Veracity describes the correctness of the data: If your data is incorrect, it will be difficult to reap benefits from it. (DeVan, 2016; McNulty 2014.)

Visualization addresses the issue of turning data into easily digestible information. Bar charts and the like are a good starting point, but more complex data demands the ability to visualize it in more complex ways. (DeVan, 2016; McNulty 2014.)

Value is used to point out that data itself is not worth anything, if it cannot be used for benefits, but has a great potential for being very valuable. (DeVan, 2016; McNulty 2014.)

2.2 What benefits can be gained from Big Data

According to the study conducted by Jacques Bughin on 273 global telecom companies, of which 80 had started a Big Data project, relatively few telecom companies had fully embraced the Big Data initiative in order to gain significant profits. However, few companies had been able to prove that investing into Big Data pays off. (Bughin 2016, 15-16.)

However, not all telecom companies reported to be successful with their big data initiatives. (Bughin 2016, 8.)

One of the ways the companies reported to have profited from their big data initiative was to use Big Data analytical models to estimate the likely times when their network engineers should be ready to take steps to ease the stress on the network caused by heavy usage of video streaming. (Bughin 2016, 7.)

Another reported way was to combine multiple sources of data, such as sociodemographic, customer touchpoint and network usage data which was then driven through a machine learning algorithm in order to determine, in real time, customers who were most likely to churn, defect or not be able to pay their bills. This allowed the company to improve recovery of payments by 35 percent and reduce churn by 3 percent. (Bughin 2016, 7.)

Cloudera laid out four key categories of opportunities for Telecom companies regarding Big Data: Customer Experience Management (Customer 360), Network Optimization & Analytics, Telco Operational Analytics & Data Monetization. (Cloudera 2015, 1.)

Customer Experience Management is a key factor in keeping customers happy and from switching operators. Competition among operators regarding customer experience management is constant as even a slightest edge can cause a major boost in revenue or drop in churn. (Cloudera 2015, 1-2.)

A 360 view of the company's customer base can allow for better targeted marketing and personalization of service. This can lead to growth of revenue due to ability to better offer personalized offerings and opportunies for upsell and cross-sell.

Micro-segmenting may allow for creation of right opportunities at the right time for the right customer. (Cloudera 2015, 2.) Showing a customer an offer for a new phone will most likely fail, if the customer has purchased a new phone 1 week ago.

Telecom companies are also able to create better proactive care by taking advantage of Big Data solutions and identifying issues before they affect the customer, before the customer has to notify them of the issue or by just reactively contacting the customer and notifying that the issue can be solved. (Cloudera 2015, 2.)

Churn prediction and prevention is a huge opportunity for the company to save costs; preventing a customer leaving is cheaper than acquiring new customers. For example, a campaign can be launched to target at-risk customers and a better offer or a deal be brokered in order to ensure that the customer stays with the current telecom provider. Similarly social media sentiment analysis can be used to identify issues and proactive help issued. (Cloudera 2015, 2.)

Networks are a core component of any telecom provider's business and thus bring interesting opportunities for optimization and analytics by using Big Data, especially related to mobile data. (Cloudera 2015, 2.)

Network capacity planning and optimization decisions can be made easier with Big Data by taking into account data points such as data usage, customer density and combining them with location and traffic data, as high traffic areas can be more easily recognized and load decreased before outages occur. (Cloudera 2015, 2.)

Similarly, network expansion and investment planning can take advantage of Big Data to analyze customer experience, network load data, revenue potential and location data to maximize the impact of the investments into network upgrades. For example, BT is already using Big Data analytics to prioritize when and where to expand their high-speed broadband services. (Cloudera 2015, 2-3.)

Network monitoring can be atomized and with real time analytics and problems can be detected and response handled without the need for human input. The traffic from surrounding cell towers can tell of a possible issue with a cell tower and the system can dispatch a maintenance group to check out in case of an issue. This faster response will no doubt lead to faster issue handling and better customer experience. (Cloudera 2015, 2-3.)

Telecom operational analytics present opportunities for the companies to plug possible revenue leaks, detect fraud and help combat cyber security issues. (Cloudera 2015, 3.)

Big Data also can create new business opportunities by allowing the telecom companies to monetize their data by combining different datasets and selling Data analytics as service or by leveraging IoT-data and M2M analytics. (Cloudera 2015, 3-4.)

Paper by Nader, M. & Jameela, A. (2015.) outlined few interesting use cases for Big Data systems that are able to handle sub one second analytics. These were Intelligent Transportation, Crowd Control, Large-Scale Emergency Response and Early Warning for Natural Disasters.

Intelligent transportation would be achieved by taking advantage of sensors within vehicles and on the roads. Leveraging these datasets would allow the system to determine the shortest or fuel efficient routes to the destination, while taking account into changing variables such as traffic and weather conditions. These datasets could be also used to minimize the time wasted on deliveries for truck drivers as well. (Nader & Jameela 2015, 3.)

Mobiles apps and vehicle sensors can be used to detect crowd movements which can be a useful asset to law enforcement when dealing with events that cause large crowd movements such as protests, concerts and seasonal celebrations such as New Year celebrations in New York. With this data in hand, the law enforcement can make decisions in real time to close streets from traffic, bring additional resources to the field or proactively plan how to react to a possible emergency situation. (Nader & Jameela 2015, 4.)

When a large-scale emergency occurs, the need for a system that can rapidly process all the information regarding the incident becomes increasingly useful as decisions have to be made as fast as possible and resources allocated to maximize their impact. Ability process the data in real time allows for quick changes to the plan as new information becomes available. (Nader & Jameela 2015, 4.)

Real-time Big Data analytics can also be used as early warning for natural disasters by analyzing data from multiple different sensors to determine which areas will be affected by an earthquake or a tsunami and sending out an alarm fast enough. Analytics can be also used to guide people to safety while avoiding traffic jams and too large crowds from forming. (Nader & Jameela 2015, 5.)

2.3 Big Data Process

There are multiple steps and paths to business benefits when extracting business value from Big Data.

A single typical example process of turning raw data into value is called the Data Value Chain. (Miller & Mork 2013, 1-3.)

The goals of the Data Value Chain are as follows:

Managing the data across the enterprise sequence starting from the systems and stakeholders which generate data to the end users of information.
Coordinating data collection from diverse sources and analyzing the extracted data to gain improvements in service performance and improving the quality of decisions.
Consolidate data management to improve the outcomes for all participants of the process. (Miller & Mork 2013, 1-3.)

The Data Value Chain itself can be divided into three parts: Data discovery, Data integration and data exploitation. (Miller & Mork 2013, 1-3.)

Data discovery consists of finding, preparing and organizing all the relevant data resources. The data resources should be organized, their completeness evaluated and commentated. (Miller & Mork 2013, 1-3.)

First step is to bring the data to a secure and shared storage system. This storage system can be massively parallel distributed system, such as Hadoop DFS, Big Table, or MongoDB. Equal amount of attention will have to be paid to access control due to legal and privacy reasons. (Miller & Mork, 2013 1-3.)

Next the syntax, structure, metadata and semantics should be defined and organized. The better the metadata, easier integration up- or downstream will be because of easier information sharing. (Miller & Mork, 2013 1-3.)

Techniques such as the Dublin Core and Department of Defense Discovery Metadata Specification can be used to help with preparing the data. Dublin Core is a metadata vocabulary that can be used to describe web resources as well as physical assets and DDMS is more for developing a suitable classification system for metadata and applying that to the datasets used in the project. (Miller & Mork, 2013 1-3.)

Data integration step consists of mapping datasets and metadata together into a production that suits the purpose of the case in question. Essentially, during this step the different datasets will be combined together to create the basis for the analysis so that the wanted end result can be achieved. (Miller & Mork 2013, 1-3.)

Data exploitation step consists analyzing, visualizing and decision making.
Once the data has been collected and integrated, business benefits can be then "exploited" from it. (Miller & Mork 2013, 1-3.)

In analysis phase the programming model and implementations for processing and generating large datasets. What this means, is that during analysis, the data will be processed in to a form that information can be extracted from it. (Miller & Mork 2013, 1-3.)
One of the most popular Big Data techniques is called Apache MapReduce.
There are variety of techniques related to big data analytics and it can be considered the most mature based on the amount of different techniques and tools. (Miller & Mork 2013, 1-3.)

Analysis also covers maintaining the origin of the data and the metadata, so that the analysis can be recreated if necessary. (Miller & Mork 2013, 1-3.)

Second to last step is visualize. This step takes the calculations from the last step and portrays them in visually efficient ways so that human brain can interpret the results easier. Visualizations can be anything from simple bar charts and dashboards to virtual reality systems. (Miller & Mork 2013, 1-3.)

The final step of the Data Value Chain is decision making.
In the end, the whole process is useless if no action is taken or if no change is being made. Sometimes the results can be just used to verify facts or gain new ideas for future analyses, but the main point of the whole process is to guide decision making. (Miller & Mork 2013, 1-3.)

2.4 Definition of Real Time

The term “real-time” can mean different things to different people depending on how urgent their needs actually are. If the exact meaning of “real-time” is not specified, confusion regarding the expectations of the goals of a real-time analytics project can rise. (Schulte 2016, 3.)

Most large companies are making thousands of real-time decisions every minute. These decisions can be customer specific, such as showing tailored offers to customers on the website based on customer data or rerouting delivery trucks based on traffic data. (Schulte 2016, 3.)

Determining when to use real-time analytics can be difficult because many things need to be considered and getting to real-time is not always even possible depending on the source systems and used technologies. (Schulte 2016, 3.)

When speaking about real-time analytics in engineering, real time usually means that the systems act in seconds, milliseconds or even microseconds. This is usually requirement for control systems of airplanes, power plants, self-driving cars and other machines. Time-sensitive software applications, like high-frequency trading systems, might also take advantage of engineering real-time concepts, even though they might not actually work in real-time. (Schulte 2016, 4.)

When business people say “real time” they normally just imply fast or close to real time. Business real time is more about situational awareness: It answers the questions, what is currently happening in the business or what will happen in the near future based on historical data. (Schulte 2016, 4.)

Business real time is relevant to ventures that concern people because people are not as rigid as machines in their interactions. It is very difficult separate business real time from recent time. (Schulte 2016, 4.)

2.5 Challenges with real-time analytics

In order to understand the challenged faced with Real-time Big Data analytics, it is necessary to first understand the difference between open loop and closed loop Big Data approaches. (Nader & Jameela 2014.)

Most standard Big Data projects are open loop in nature, which means that a Big Data project is started for a specific business realm with the goal of gaining insights that can be used to improve the realm or related processes. Open loop systems do not integrate the results of their analysis back to the loop itself. (Nader & Jameela 2014.)

Real-time Big Data analytics projects are usually closed loop approaches where the frame of action is very limited and in most cases, cannot wait for human input. In closed loop systems, feedback from the process itself is used to direct the next actions of the system. (Nader & Jameela 2014.)

One of the main challenges of real time Big Data analytics is real-time event transfer. What this means is that once an event is created, it will have to be transferred somewhere to be processed. This is called centralized approach and it is good when the amount of events does not exceed the speed of the network used to transmit them. If the amount of events exceeds the capacity, then Real-time analytics cannot be achieved. (Nader & Jameela 2014.)

A solution to high volume of events is to use a decentralized processing approach, where some processing, such as filtering or aggregation, will be done before the events are transmitted, reducing the amount of data being sent. This filtering and aggregation can be open- or closed-loop in nature: (Nader & Jameela 2014.)

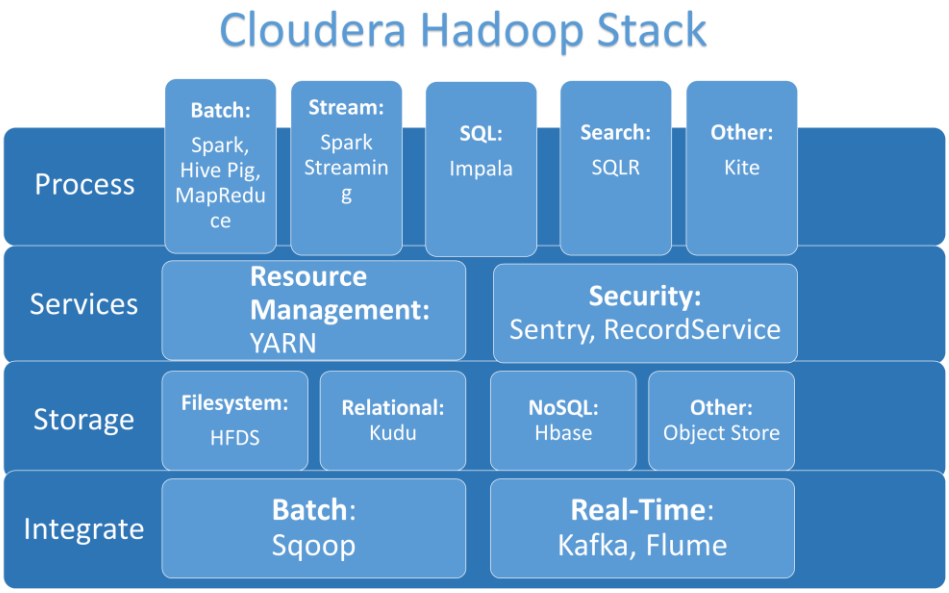
The filtering and aggregation rules can be predefined or be defined on the fly based on recent events or resource usage, which is called Real-time situation discovery. (Nader & Jameela 2014.)

A direct challenge for Real-time analytics is combining stored Big Data with real-time data in order to predict performance and risks for the business domain. Big Data's volume presents a challenge, but the data can be abstracted to smaller pieces so that it can be used within the demands of the business. These analytics processes are usually very resource intensive which is why techniques such as parallel processes are required.
(Nader & Jameela 2014.)

2.6 Big Data Technologies

Big Data is almost synonymous for Hadoop as it was Google's Big Data work on MapReduce and Google File System that was made open source and then molded to Apache Hadoop. (EMC Education Services 2014.)

Big Data platforms can be roughly divided to four different layers: Integration layer, storage layer, services layer and process layer. See Picture 1 how the components are aligned on these layers in Cloudera's stack.



Picture 1. Components of Cloudera Hadoop Stack (Cote 2016)

2.6.1 Big Data Platforms

Big Data platform is a combined IT-solution that provides capabilities of many different Big Data applications in an environment that ensures that all the provided components work together. These platforms can work in cloud or be deployed to a server. (Techopedia.)

The largest benefit of Big Data platforms is the ease of getting started. The users do not have to spend lot of time in order to get started as all the components should work together and integration to other systems is also easier. Of course customization is possible, but not necessary. (Techopedia.)

Biggest differences between vendors are the building blocks that they use, with large vendors such as Microsoft, IBM and Oracle, using many of their own products. (Techopedia.)

Products: Amazon Web Services, Cloudera Enterprise Platform for Big Data, IBM Big Data Platform, Microsoft HDInsight, SAP Hana & Oracle Big Data.

2.6.2 NoSQL databases

NoSQL is a general term of describing new data storage technologies that are aimed to work with applications that need to store massive amounts of new and swiftly changing data in all forms, structured and unstructured. (MongoDB.)

NoSQL databases are optimized for fast writing as the schema of the data will only be specified at the moment of usage, this also allows for more agile development as schemas can be specified at the moment of reading. (MongoDB.)

Other key features are availability, accessibility and scalability. (MongoDB.)

Key advantage over normal relational databases is that the data can be split on many different servers and similarly the load is also shared between those servers. This means that NoSQL databases can be horizontally scaled relatively easily. (MongoDB.)

The NoSQL technology is however relatively immature and by no means perfect for every situation. (Richards, 2015.)

Products: MongoDB, Redis, Cassandra, CouchDB and HBase.

2.6.3 Hadoop HDSF

Apache Hadoop Distributed File System is way of spreading data on to multiple normal servers by splitting up the potentially huge dataset on to each server on the cluster. Copies of each piece of the data are backed up to different servers as well, in case a single server fails.

This usage of distributed computing power also allows for faster processing of data, because the processing can also be split between machines. (IBM HDFS.)

2.6.4 Hadoop MapReduce

Apache Hadoop MapReduce is a technique which allows for parallel processing of data on separate Hadoop clusters. This allows for huge horizontal scalability and processing power.

It has gotten its name from the two tasks, or jobs, that together create the technique. Mapper jobs. The dataset is divided for the mappers with certain criteria and after all the mappers have completed their jobs, the reducers start working and combine the data from the mappers getting the wanted results. (IBM MapReduce.)

Essentially, MapReduce is a way to perform parallel calculations on a distributed file system. (IBM MapReduce.)

2.6.5 Hive

Apache Hive is a query language built to work on NoSQL databases. It uses a similar syntax to SQL called HiveSQL which Hive then turns into MapReduce jobs. Its main tasks are aggregation, querying and analysis. (Rouse, 2012.)

Hive and Pig are similar languages, but the major difference is that Pig works with data flows while Pig works with data that already exists in the database. When changes are made in Pig, they are not saved, unlike in Hive where the changes persist after each query.
(Hortonworks b.)

2.6.6 Pig

Apache Pig is a high level programming language that allows for easier programming of MapReduce jobs without having to touch MapReduce code. This way the logic can be separated from the MapReduce's own logic. (Rouse, 2014.)

A good use case for Pig would be an ETL process of source data to the database according to rules specified. Pig comes with user defined functions which make it relatively easy programming language to use. (Hortonworks a.)

2.7 Future trends

Gartner's picks for biggest trends related to Big Data, currently and in the near future are AI & machine learning, intelligent apps and intelligent things.

AI & Machine learning is not a new trend, but will be relevant as long as there still are tasks to automate. Machine learning will be more and more relevant for the full automatization of factories as well as improving the performance of individuals via decision support systems. (Panetta 2016.)

Benefits reaped by intelligent apps consists of advanced analytics, autonomous business processes and AI-powered intelligent interfaces. Along with machine learning, AI will help to automate processes, decision making and create better virtual assistants (Panetta 2016.)

IoT will evolve to intelligent things which means that robots and drones may come with some sort of embedded intelligence that will allow them learn to perform their tasks better. This may also lead to intelligent things, such as self-driving cars, to communicate and to work together. (Panetta 2016.) One thing is certain, IoT-data will grow. (Jeevan 2016.)

Another shift is happening in the way Big Data is being thought about. The definition of 3 Vs, volume, velocity and variety is no longer in the focus as there now exists multiple solutions to those challenges. Now the new hot issues are mostly related to fast data, actionable data, relevant data and smart data. (SmartDataCollective 2015.)

Fast data relates to the same need that launched this thesis work: How can be large datasets be processed in a real time fashion? (SmartDataCollective 2015.)

Actionable data points to the challenge of actually getting usable results out of data so that data can be used to guide decision making. A good example is social media data which can be analyzed to find out consumer sentiment regarding new products thus leading to discovering insights before it can be seen from the revenue of the company. (Gambhir.)

Relevant data refers to the difficulty of analyzing data with the context of other datasets. There may have been unseen events that have affected multiple datasets, but the connection is not visible from analyzing just a single dataset. (SmartDataCollective 2015.)

Meaning based computing or smart data is used to indicate the difficulty of creating self-improving models and programs that improve based on the data that they collect from multiple sources. A good example of smart data could be a wind power plant that uses sensor data to detect possible issue and react to it before it causes an outage. (Wolfgang 2015.)

3 Research

3.1 Methods

Five open interviews were performed in addition to literature research. The discussions were open, but help questions were used to guide the direction of the discussion. Emphasis was put on real-time processing, finding challenges in ongoing processes and technologies and Spark.

The product owner of the company's Big Data solution was interviewed so that the business requirements behind the need for real-time analytics were clarified.

Two technical developers were interview to understand what the possible issues are and to identify potential bottle necks in the systems and processes.

The main architect of the Big Data solution was interview to understand what capabilities are currently possible and which will be in the future and also what kind of challenges the developers of the Big Data stack face, especially towards implementing new tools such as Spark.

Last, a person responsible for an IoT-based project was interviewed to better understand what kind of real-time projects the target company is considering.

3.2 Defining the research question:

In order to better understand the business needs behind the research question, an open interview was held with the business owner of the Big Data solution.

The Big Data project for the target company was initiated by the company's business sector, not IT-department, as a need for right time data and enabling real time data became apparent. The two proposed use cases for real time Big Data analytics were a service assurance project and an online customer insight project. (Business Owner 2016.)

The service assurance project aims to use network data to improve the quality of the service by using advance analytics to figure out where and when issues may rise. This is an example of an open loop approach where first the process won't be automated and the

results will be used to react and improve the networks by hand. It is possible that in the future this process will be automated and adjustments to the network will be done on the fly automatically. (Business Owner 2016.)

The online customer insight project tries to leverage already existing customer data in close to real time fashion when the customer accesses the company's online services. For example, if a customer has prior tickets related to issues with a specific product, then the offers shown to the customer will not include upgrades to that service or if customer has looked a product in the past, without purchasing the same offer will be actively shown again to him. (Business Owner 2016.)

This is a project's result will be most likely an automated system which will guide the customer's online experience based on their prior interactions with the company. (Business Owner 2016.)

Due to the low maturity level of Big Data in the target company, there still have not been clearly defined rules on what kind of data can be saved to the data storage and what cannot be. This has been a substantial bottleneck for the development, but efforts are being made that an understanding can be reached with the stakeholders. (Business Owner 2016.)

One of the key issues that the Big Data initiative faces is lack of knowledge of which Big Data tools to use in order to reach the business goals. (Business Owner 2016.)

3.3 Case Company's Big Data solution

The case company's Big Data solution is known by the name Common Data Lake, CDL. CDL is used to mean the whole Big Data project, including systems, data, people and processes.

In order to understand the system, it is first important to understand what exactly Data Lake means and how it compares to the more traditional data warehouses.

When saving data to a warehouse, it has to adhere to the rules, restrictions and structure of the target database's tables, which can require lot of manual mapping and transforming work.

Data Lake's design principles allow for lesser emphasis to be put on structuring the data on saving. This design philosophy allows for fast saving of new data sources, as any structure needed can be specified later on when the data will be used. (Dull 9.2015.)

Traditional data warehouses can be expensive as they are usually proprietary software. Hadoop is open source software and because it is distributed, it will run on multiple nodes which usually are not high end hardware. Power of Hadoop comes from the distributed computing power from multiple nodes. (Dull 9.2015.)

Data Lake can also be relatively agile, as the structure will be defined case by case depending on the project because there is no set structure for the data when it is saved. Data warehouses have prespecified structure, which can prove to be difficult to change if many business processes already have been adapted to it. (Dull 9.2015.)

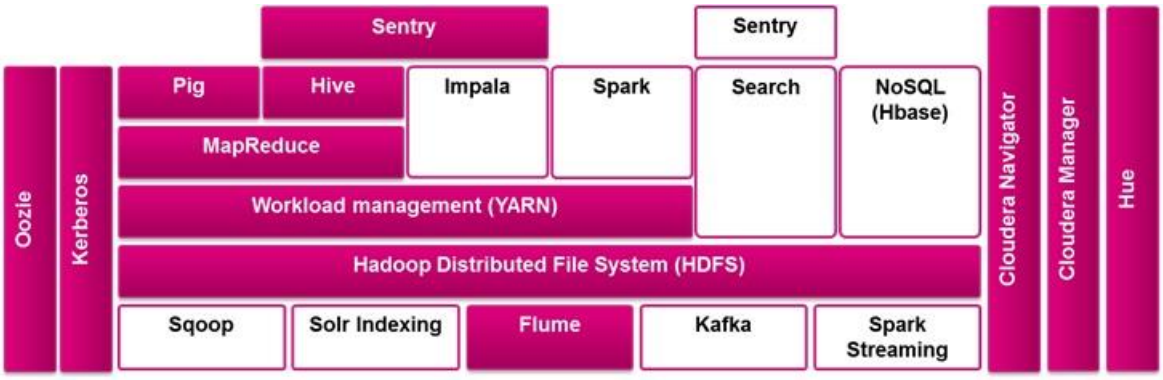
However, Data Lake still lacks the advanced security features of a more mature Data warehouse solution, which limits the data that can be saved to the Data Lake. Similarly, Data Lake is still relatively new concept and has not been widely adopted, which means that the technology and the processes are perhaps not mature enough for all use cases. (Dull 9.2015.)

Bottom line is that Data warehouse and Data Lake are different techniques for saving data. They are not meant to replace each other, but to rather supplement each other. (Dull 9.2015.)

CDL was designed to work alongside with the current data warehouses, data marts and Business Intelligence stack of the company. It was not designed to replace the core data warehousing systems, just to act as a fluid storage between data source systems and analytical data consuming systems.

On top of CDL runs Cloudera's Enterprise Big Data platform that is used for its Data applications and tools. The current version of the Cloudera Apache Hadoop on CDL is 5.4.2, whereas the newest available version is 5.9.0.

As seen on Picture 2, multiple Hadoop components have not yet been implemented as ensuring the maturity, stability and security of them takes time. For the similar reasons the company cannot update to newer versions of Cloudera Apache Hadoop without first making sure that everything works with the updates.



Picture 2. Case company's Cloudera Stack. Red components enabled, white not yet (CDL Solution Architecture)

An interview with a Big Data developer was also organized to gather information on the technical challenges faced by the Big Data initiative.

According to the Big Data developer, the company's current Big Data solution is not ready for real-time analytics. Currently the team is still struggling with getting data to the CDL fast enough so that's where most effort is being channeled and not as much emphasis has been put to the processing side of things. (CDL Developer 10.2016 & 11.2016.)

Another large bottle neck in the Big Data system is the general enterprise bus the company uses to send out messages between systems. The general enterprise service bus is the preferred method of delivering data to the CDL. (CDL Solution Architecture - General Description, 13.) For some systems it works, but it was not designed to work with heavy data transfer in real time and so does not support all use cases. Many legacy systems which are connected to it also cannot output data fast enough to be useful for real-time analytics purposes. (CDL Developer 25.10.2016 & 04.11.2016.)

Development of the CDL platform is handled by a different organization within the same company as the team that develops on top of CDL to answer the needs of the business. The team that handles development of the CDL platform has to ensure that all new software works together without breaking existing applications, compromising the security of the platform. This however slows down the development and introduction of new components. (CDL Main Architect 16.11.2016.)

One notable challenge that was brought up was lack of communication between parties and information sharing. However the different parties were aware of these issues as there were some plans to already improve communication, such as workshops.

3.4 Results

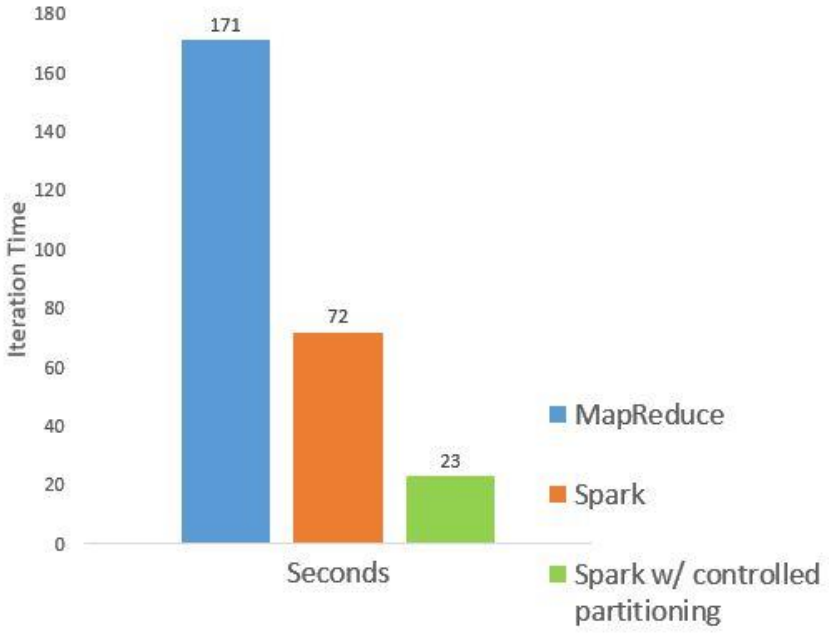
Two of the business demands were to identify Big Data tools that can help with challenges brought on by the real-time or close to real-time demands. From the interviews it came apparent that in some cases MapReduce just cannot provide the processing power for the required speed, so something faster should be considered.

Apache Storm, Apache S4, Apache Apex and Apache Flink were also considered, but Spark was ultimately chosen for the focus of this thesis work as the Cloudera Big Data stack already supports Spark and that there apparently have been some tests with Spark in the target company.

It is worth keeping track of these other technologies and how they evolve as there might be unique use cases for them in the future.

3.4.1 Apache Spark

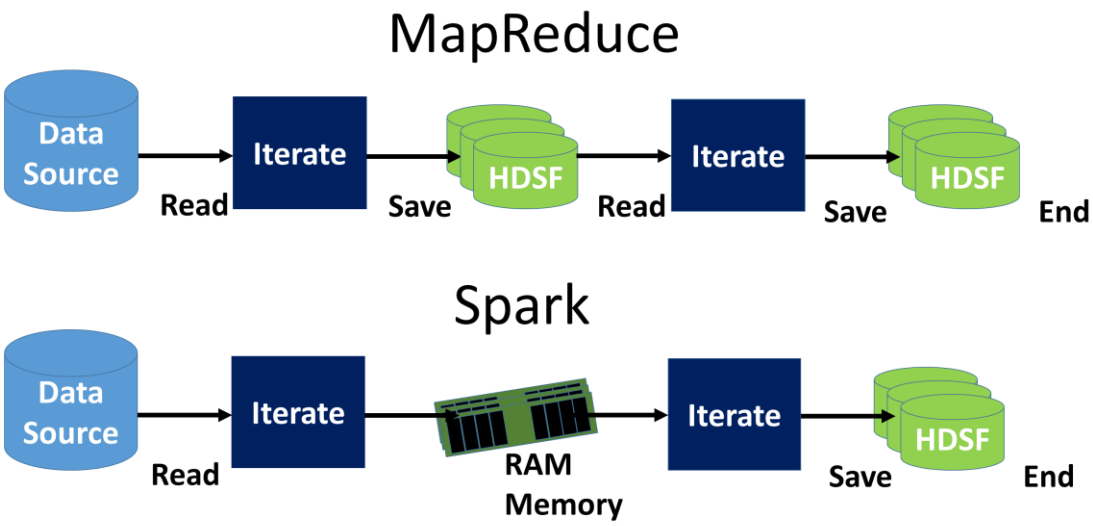
Apache Spark is an open source, large-scale, distributed processing engine that boasts 100x faster in-memory and 10x faster on disk processing compared to MapReduce. This can be attributed to the fact that MapReduce was not designed with speed in mind, its main purpose is to handle and process large datasets. (Vardhan 2015) Spark's faster processing speed is demonstrated in Picture 3.



Picture 3. Processing speed difference between Spark and MapReduce (Vardhan 2015)

Spark core manages task scheduling, memory management, fault recovery and contains the APIs that are used to manipulate the RDDs. (Rathore 2016.)

MapReduce uses batch processing, which means that it takes in all the data at once and then process it in multiple jobs and then returns the results at the end of the process, not really useful for real-time use cases. What also slows down MapReduce is the fact that it has to save intermediate results to disk instead of using memory for iteration like Spark as demonstrated by Picture 4. (Siva 2016.)



Picture 4. MapReduce and Spark processing differences (Bosshart 2014)

What makes Spark special is Resilient Distributed Datasets (RDDs) which can be simply explained as distributed collection of elements. This RDD is then split to partitions that can be processed at the same time on different nodes. RDD also contains the information how the dataset was built, which means that RDDs do not have to be loaded into memory all the time, as they can be constructed when needed. (Vardhan 2015.)
Another key advantage of Spark is that it is very fault tolerant. If a node fails, it can be recomputed in seconds as the data can be reproduced by using lineage. (Kharbanda 2015.)

Currently Apache Spark is being supported by multiple larger Big Data vendors such as IBM, MapR, Cloudera, Intel and Hortonworks which means that its development most likely will not cease soon. (Vardhan 2015.)

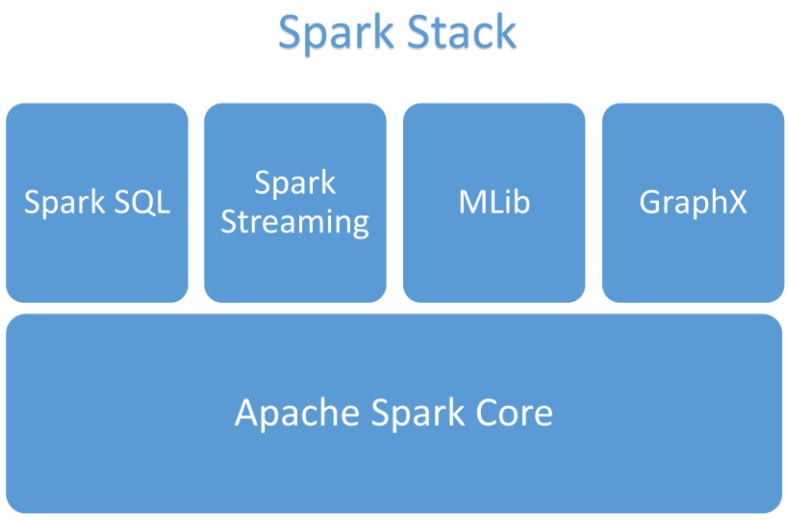
3.4.2 Spark Streaming

Even though Spark may allow for real time exploration of data, it alone is may not be able to answer the real time or close to real time needs posed by the business. Spark Streaming is an extension to Spark which allows for stream processing of data from the data source.

Spark Streaming gets data from a live data streams, divides the data into micro batches which are then in turn processed into the results by the Spark engine. The batch size determines the latency of the system and how fast the data can be streamed, the smaller it is, the faster the data comes in. (Mahanthesh 2016.)

The biggest difference between vanilla Spark and Spark Streaming is that Spark Streaming uses micro batching called D-streaming whereas Spark is use normal batching. Micro batching allows for lower latency as the RDDs are divided into smaller batches. (Mahanthesh 2016.)

Streaming can use services such as Kafka, Flume or HFDS to create D-streams. Creation of simple Spark Streaming apps that listen to a hostname and port for data is fairly straight forward and simple. (Kropp 2015.)



Picture 5. Spark Stack components (Apache Spark)

As seen on picture 5, Spark comes with couple other modules in addition to Spark Streaming. These are: Spark SQL, MLib and GraphX. Although they are not fully relevant

to streaming, it is a good idea to understand what they are about, as all the Spark stack's modules can be combined in a one program. (Apache Spark a.)
Spark SQL is a module that allows for querying of structured or semi-structured data. (Tutorials Point.)

MLib is a machine learning library that contains many common algorithms and utilities allowing for fast machine learning to be performed on top of Spark core. (Apache Spark a.)

GraphX is a graph processing module that allows for processing of graph data sets on distributed systems. (Apache Spark b.)

Installation of Spark requires both Scala and Java Development Kit

3.4.3 Downsides of Apache Spark

Although Spark may seem like an improvement in many ways over MapReduce there still are some downsides to it.

Spark Streaming works with micro batching, which will not be able to reach a level of latency required by engineering real-time standards, latency of milliseconds. However, the batch size can be low as 0.5 seconds which is good enough for most business needs where the latency needs to be in seconds.
(Tathagata 2013.)

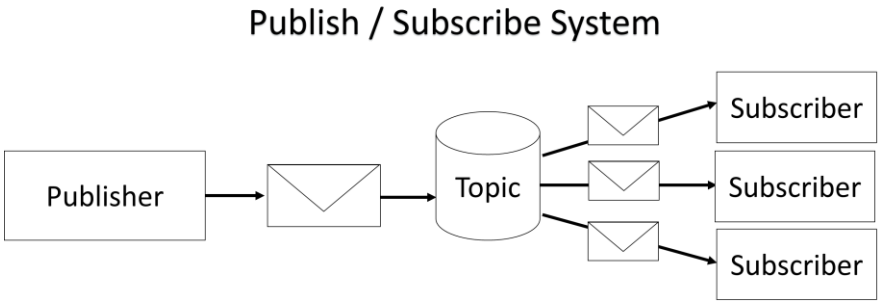
MapReduce is a more mature technology and has been in development for longer which can prove to be significant benefit over Spark, as work on Spark may consume more resources due to Spark being still the more unstable one of the two technologies. (CDL Main Architect 16.11.2016.)

A real downside of any new technology is that the job market initially lacks the necessary skills with the newer technology. This of course makes using MapReduce more appealing as it has been the standard batch processing technology since the start of the Big Data boom and already has a skilled work force. Due to same reasons, MapReduce also has a greater support of other applications. (DeZyre 2014.)

3.4.4 Kafka

One of the recognized bottlenecks for the company's Big Data platform was the limitation of it's publish-subscribe messaging system. (CDL Developer 10.2016 & 11.2016.) Although this thesis work is focused on the processing side of Big Data analytics, it is worth taking a quick, high level look at how Kafka could potentially help with this issue and how it fits to the Big Data ecosystem.

Publish-subscribe system is an event bus that receives messages from publishers and then sends them out via topics. The messages can contain any data. Subscribers then can subscribe only to those topics which are relevant to them. This way the Publish-subscribe systems can relay the same information to many subscribers without having to build separate interfaces for each different subscriber. (Abdelrazzak.) See picture 6 for a simplified example of similar system delivering a message to multiple subscribers.

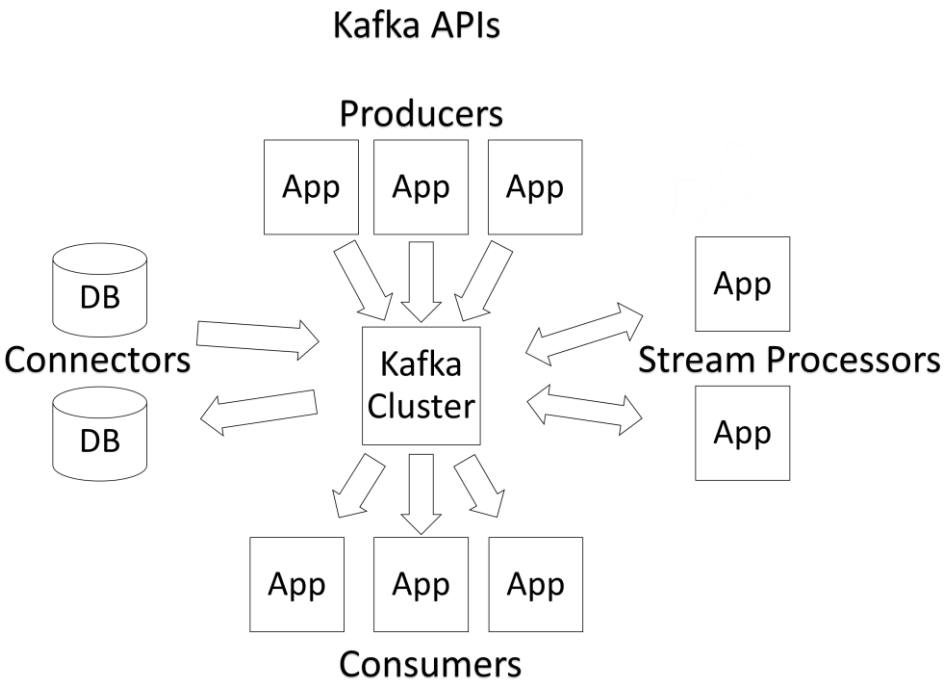


Picture 6. Simple publish-subscribe system (RedHat)

Kafka is a distributed publish-subscribe messaging system that has been designed with speed and size in mind. It was created by LinkedIn as a solution to standardize data pipelines for saving data to Hadoop as creation of separate, custom data pipelines for each system increased complexity and made maintainability difficult. (Shapira & Holoman 2014.)

Kafka has four APIs as shown in picture 7. Producer API publishes messages to a Kafka topic or multiple topics. Consumer API allows an application to subscribe and receive data from one or more topics. Streams API allows for an application to process a stream coming from a topic to an external source or back to Kafka. (KAFKA STREAMS.) Connector

APIs are used to connect to data source such as relational databases and used to read or save large amounts of data into topics. (KAFKA CONNECT.)



Picture 7. Kafka APIs (Apache Kafka)

The main reason Kafka is relevant to this thesis work is because it is widely used way to get data to Hadoop's file system and because its stream API integrates well with Spark Streaming. (Shapira & Holoman 2014.)

Kafka's site lists the following two sentences as the main use cases for it.
"1. Building real-time streaming data pipelines that reliably get data between systems or applications"

"2. Building real-time streaming applications that transform or react to the streams of data" (Apache Kafka.)

3.5 Implementation Plan Example

The following example is based on MAPR Academy Lesson 8 - Create Data Pipelines with Apache Spark.

A possible use case for Spark Streaming might look something like this.

First the data source should be defined. In this example, sensor data from temperature sensors in Json format will be used as it mirrors the kind of use cases the target company actually works with. However, the data could be anything from CSV files, relational database data to sensor streams.

Then the data will have to be ingested via a third party messaging tool such as Kafka, Flume or deposited directly to a file system where it can be read. These systems usually define how big the latency of the whole process will be. If the data is only published every minute, then the fastest latency the whole system can offer is minute plus the latency of processing the data.

Next the data will be processed by Spark streaming or a similar stream processing engine. For this example, a data will be streamed from HDFS system as it is simple to implement. (Poullet 2015.) In this step, the continuous data will be divided to micro batches based on the time when they were received which will be then processed by Spark.

There are two types of operations on D-Streams: Transform and output, which run between time intervals on each step and create the processed output batches. These operations affect batches inside the D-Streams and the data inside of the batches.

After the data has been processed, it should be saved to a No-SQL database that is capable of delivering data fast. This No-SQL database could be possibly Apache Hbase or similar.

Finally the processed data is delivered to dashboard to an application where it will be further used. The data can also be saved back to data storage for future use. In this example the data could be used to send out alarms if a temperature sensor reached a temperature over a baseline fetched from historical records saved in a database indicating an issue with the sensor or a fire.

3.6 Conclusions

It does not seem like the Big Data process for the company is developed enough yet. There are several challenges that need to be solved before Big Data can be truly leveraged to its full potential.

Some of these challenges are already being worked on, but lot of work still need to be done before the Big Data initiative can be considered successful.

Thankfully the technical issues, such as real-time analytics can be solved with the usage of Spark and Spark streaming, but first to be able to use them, they will have to be implemented to the company's Big Data platform.

As discussed with the Big Data platform's lead architect, proper business cases will first have to be created and presented, before investments to new technology can be made. (CDL Main Architect 2016.)

The Big Data team should be more involved with experimenting with the new technologies. It is not always possible to do that on the development environment of CDL, but setting up separate instances of the platform is fairly simple and cheap as the platform is open source.

It may be a good idea for the teams working with the company's Big Data platform to take time and to discover the capabilities of these new technologies and verify potential business cases on their own outside of the main Big Data platform. Thus finding out what ex-

actly needs to be implemented in the main Big Data platform when the business case will be fully resolved.

Work on this thesis work allows for the writer to take part in and ongoing IoT-data project that requires real-time processing of the data. The subject matters addressed in this thesis work were just very small steps in the Big Picture of leveraging data to improve business, but interesting nonetheless.

4 Sources

Abdelrazzak, A. The Publish-Subscribe Pattern on Rails: An Implementation Tutorial. (https://www.toptal.com/ruby-on-rails/the-publish-subscribe-pattern-on-rails) Read: 17.11.2016

Ackoff, R. 1989. From Data to Wisdom. (http://faculty.ung.edu/kmelton/Documents/DataWisdom.pdf) Read: 10.10.2016

Apache Spark a. (http://spark.apache.org/) Read: 20.11.2016

Apache Spark b. Machine Learning Library (MLlib) Programming Guide. (https://spark.apache.org/docs/1.2.0/mllib-guide.html) Read: 20.11.2016.

Apache Spark c. GraphX. (http://spark.apache.org/graphx/) Read: 20.11.2016.

Bughin, J. 2016. Reaping the benefits of Big Data in telecom. Journal of Big Data. 3:14. (https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0048-1) Read: 20.10.2016

Buhl, H. Röglinger, M. & Moser, F. 2013. Big Data A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? (http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1218&context=bise) Read: 11.10.2016

Business Owner. 1.11.2016. Case Company. Open Interview. Helsinki.

CDL Solution Architecture - General Description. Internal document. Read: 10.11.2016

CDL Developer. 25.10.2016 & 04.11.2016. Case Company. Open Interview. Helsinki.

CDL Main Architect. 9.11.2016. Case Company. Open Interview. Online.

Cloudera. 2015. Big Data Use Cases for Telcos. Industry brief. (https://www.cloudera.com/content/dam/cloudera/Resources/PDF/solution-briefs/Industry-Brief-Big-Data-Use-Cases-for-Telcos.pdf) Read: 3.11.2016

De Mauro, A. Greco, M. & Grimaldi, M. 2015. What is big data? A consensual definition and a review of key research topics. (http://cloudtribes.com/docstation/com_docstation/24/what_is_big_data__a_consensual_definition_and_a_review_of_key_research_topics.pdf) Read: 12.10.2016

DeVan, A. 7.4.2016. The 7 V’s of Big Data.
(https://www.impactradius.com/blog/7-vs-big-data/)
Read: 10.10.2016

DeZyre. 2014. Hadoop MapReduce vs. Apache Spark Who Wins the Battle?
(https://www.dezyre.com/article/hadoop-mapreduce-vs-apache-spark-who-wins-the-battle/83)
Read: 19.11.2016

Dodson, S. 2014. Big Data, Big Hype?
(https://www.wired.com/insights/2014/04/big-data-big-hype/)
Read: 11.10.2016

Dumbill, E. 2012. What is big data?
(https://www.oreilly.com/ideas/what-is-big-data)
Read: 15.10.2016

Dull, T. 2015. Data Lake vs Data Warehouse: Key Differences.
(http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html)
Read: 30.10.2016

EMC Digital Universe with Research & Analysis by IDC. 2014. Executive Summary Data Growth, Business Opportunities, and the IT Imperatives.
(http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm)
Read: 14.10.2016

EMC Education Services. 2014. Data Science and Big Data Analytics.
(https://books.google.fi/books?id=axruBQAAQBAJ&pg=PA300&redir_esc=y#v=onepage&q&f=false)
Read: 19.11.2016

Gambhir, A. What is Actionable Data... and How Do You Use It?
(http://hotelexecutive.com/business_review/2722/what-is-actionable-data-and-how-do-you-use-it)
Read: 21.11.2016

Hadoop. 2007. Apache Hadoop Releases.
(http://hadoop.apache.org/releases.html)
Read: 10.10.2016

Hortonworks a. HOW TO PROCESS DATA WITH APACHE PIG.
(http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-pig/)
Read: 11.11.2016

Hortonworks b. HOW TO PROCESS DATA WITH APACHE HIVE.
(http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/)
Read: 11.11.2016

IBM. 2013. Infographics & Animations The Four V's of Big Data. Infographic.
(http://www.ibmbigdatahub.com/infographic/four-vs-big-data)
Read: 14.10.2016

IBM HDFS. About Hadoop Distributed File System (HDFS)™.
(https://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/)
Read: 11.11.2016

IBM MapReduce. What is MapReduce?
(https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/)
Read: 11.11.2016

IBM Video. 2010. IBM Commercial The Road.
(http://www.dailymotion.com/video/xdaoae_ibm-commercial-the-road-intelligent_tech)
Watched: 15.10.2016

Jeevan, M. 20.1.2016. Gartner predicts five big data trends that will dominate 2016.
(http://bigdata-madesimple.com/gartner-predicts-five-big-data-trends-that-will-dominate-2016/)
Read: 26.11.2016

KAFKA CONNECT.
(http://kafka.apache.org/documentation.html#connect)
Read: 17.11.2016

KAFKA STREAMS.
(http://kafka.apache.org/documentation.html#streams)
Read: 17.11.2016

Kanellos, M. 2016. 152,000 Smart Devices Every Minute In 2025: IDC Outlines The Future of Smart Things.
(http://www.forbes.com/sites/michaelkanellos/2016/03/03/152000-smart-devices-every-minute-in-2025-idc-outlines-the-future-of-smart-things/#f1444d669a71)
Read: 19.10.2016

Kharbanda, A. 2015 A guide to Apache's Spark Streaming.
(https://opensource.com/business/15/4/guide-to-apache-spark-streaming)
Read: 19.10.2016

Kropp, H. 22.3.2015. Spark Streaming – A Simple Example.
(http://henning.kropponline.de/2015/03/22/spark-streaming-simple-example/)
Read: 20.11.2016

Lohr, S. 2013. The Origins of 'Big Data': An Etymological Detective Story.
(http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/?_r=0)
Read: 10.10.2016

Mahanthesh, H. 23.3.2016. What is the difference between batch, Dstream and RDD in Spark streaming? Quora post.
(https://www.quora.com/What-is-the-difference-between-batch-Dstream-and-RDD-in-Spark-streaming)
Read: 23.11.2016

MAPR Academy. Lesson 8 - Create Data Pipelines With Apache Spark.
(http://learn.mapr.com/dev-362-create-data-pipelines-using-apache-spark/31407/scorm/fyuy82ilt5r6)

McNulty, E. 22.4.2014. UNDERSTANDING BIG DATA: THE SEVEN V'S
(http://dataconomy.com/seven-vs-big-data/)
Read: 10.10.2016

Miller, H. & Mork, P. 2013. From Data to Decisions: A Value Chain for Big Data.
IT Professional 15, 1, 1-3.
(https://pdfs.semanticscholar.org/a8cb/17a6f21c5c43359f52b07617faa92d3ef1d1.pdf?_ga=1.141610802.1603325735.1478843686)
Read: 10.11.2016

MongoDB. NoSQL Databases Explained.
(https://www.mongodb.com/nosql-explained)
Read: 11.11.2016

Nader, M. & Jameela, A. 2014. Real-Time Big Data Analytics: Applications and Challenges. Conference Paper. The 2014 International Conference on High Performance Computing & Simulation. July 2014. Bologna, Italy.
(https://www.researchgate.net/publication/283212652_Real-Time_Big_Data_Analytics_Applications_and_Challenges)
Read: 6.11.2016

National Institute of Standards and Technology. 2015. NIST Big Data Interoperability Framework: Volume 1, Definitions. NIST Special Publication, 1500, 1, 4 & 15.
(http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf)
Read: 20.10.2016

Nawab, R. 21.8.2012. History of Data Analytics. AnalyticBridge.
(http://www.analyticbridge.com/profiles/blogs/history-of-data-analytics)
Read: 15.10.2016

Panetta, K. 18.10.2016. Gartner's Top 10 Strategic Technology Trends for 2017. Smarter With Gartner.
(http://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017/)
Read: 26.11.2016

Picture 1. Cote, D. Customer Operations Engineer. 16.3.2016. Troubleshooting Hadoop: Distributed Debugging. Cloudera. Great Wide Open 2016 Atlanta, United States. Slide 5. Conference.
(http://www.slideshare.net/GWOcon/troubleshooting-hadoop-distributed-debugging)
Read: 11.11.2016

Picture 2. CDL Solution Architecture - General Description.
Internal document.
Read: 10.11.2016

Picture 3. Vardhan. 18.12.2015. Apache Spark vs Hadoop MapReduce. Edureka. Big Data Analytics.
(http://www.edureka.co/blog/apache-spark-vs-hadoop-mapreduce)
Read: 15.11.2016

Picture 4. Schilder, F. 22.9.2014. Spark meetup TCHUG. Twin Cities Hadoop User Group. St. Paul, United States. Slide 9. Meetup.
(http://www.slideshare.net/RyanBosshart/spark-meetup-tchug)
Read: 15.11.2016

Picture 5. Apache Spark. Simplified for this thesis work.
(http://spark.apache.org/)
Read: 17.11.2016

Picture 6. RedHat. JMS Basics.
(https://access.redhat.com/documentation/en-US/Fuse_Message_Broker/5.4/html/Getting_Started/files/FuseMBStartedKeyJMS.html)
Read: 17.11.2016

Picture 7. Apache Kafka. Introduction.
(http://kafka.apache.org/intro.html)
Read: 17.11.2016

Poullet, J. 4.6.2015. Spark streaming: simple example streaming data from HDFS. Statrgy
(http://statrgy.com/2015/06/04/spark-streaming-simple-example-streaming-data-from-hdfs/)
Read: 09.11.2016

Rathore, D. 2016. SPARK AN UNIFIED STACK.
(https://www.dunebook.com/spark-an-unified-stack/2/)
Read: 18.10.2016:

Richards, J. 2015. Advantages and Disadvantages of NoSQL databases – what you should know. Hadoop360.
(http://www.hadoop360.com/blog/advantages-and-disadvantages-of-nosql-databases-what-you-should-k)
Read: 11.11.2016

Rouse, M. 2012. Apache Hive.
(http://searchdatamanagement.techtarget.com/definition/Apache-Hive)
Read: 11.11.2016

Rouse, M. 2014. Apache Pig.
(http://searchdatamanagement.techtarget.com/definition/Apache-Pig)
Read: 11.11.2016

Schulte, R. 2016. How to Move Analytics to Real Time. Gartner.

Shapira, G & Holoman, J. 12.9.2014. Apache Kafka for Beginners. Cloudera.
(http://blog.cloudera.com/blog/2014/09/apache-kafka-for-beginners/)
Read: 17.11.2016

Siva. 2016. Resilient Distributed Dataset.
(http://hadooptutorial.info/resilient-distributed-dataset/)
Read: 18.10.2016

SmartDataCollective. 2015. Big Data Is Really Dead.
(http://www.smartdatacollective.com/tonyshan/309691/big-data-really-dead)
Read: 21.11.2016

Tathagata, D. Lead developer. 17.6.2013. Deep Dive with Spark Streaming. AMPLab. Sunnyvale, United States. Slide 4. Meetup.
(http://www.slideshare.net/spark-project/deep-divewithsparkstreaming-tathagatadassparkmeetup20130617)
Read: 20.11.2016

Taylor, F. 1911. Principles of Scientific Management. Harper & Brothers Publishers.
(https://www.marxists.org/reference/subject/economics/taylor/principles/ch02.htm)
Read: 15.10.2016

Techopedia. Big Data Platform.
(https://www.techopedia.com/definition/29951/big-data-platform)
Read: 11.11.2016

Tutorials Point. Spark SQL - Introduction.
(https://www.tutorialspoint.com/spark_sql/spark_sql_introduction.htm)
Read: 20.11.2016.

Vardhan. 18.12.2015. Apache Spark vs Hadoop MapReduce. Big Data Analytics. Edureka.
(http://www.edureka.co/blog/apache-spark-vs-hadoop-mapreduce)
Read: 18.10.2016

Wolfgang, H. 2015. Why Big Data Has to Become Smart Data!
(http://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/from-big-data-to-smart-data-why-big-data-has-to-become-smart-data.html)
Read: 21.11.2016