SAVONIA

OPINNÄYTETYÖ - **AMMATTIKORKEAKOULUTUTKINTO**
TEKNIIKAN JA LIIKENTEEN ALA

REALIZATION OF BIG DATA ANALYTICS TOOL FOR OPTIMIZATION PROCESSES WITHIN THE FINNISH ENGINEERING COMPANY

Author/s: Karapetyan Karina

SAVONIA

SAVONIA UNIVERSITY OF APPLIED SCIENCES

THESIS

Abstract

Field of Study Technology, Communication and Transport			
Degree Programme Degree Programme in Information Technology			
Author(s) Karapetyan Karina			
Title of Thesis Realization of Big Data Analytics Tool for optimization processes within the Finnish engineering company			
Date	23.05.2016	Pages/Appendices	54
Supervisor(s) Mr. Arto Toppinen, Principal Lecturer at Savonia University of Applied Sciences, Mr. Anssi Suhonen, Lecturer at Savonia University of Applied Sciences			
Client Organisation /Partners Hydroline Oy			

Abstract

Big Data Analytics Tool offers an entire business picture for making both operational and strategic decisions from selecting the product price to establishing the priorities for the further vendor's enhancement.

The purpose of the thesis was to explore the industrial system of Hydroline Oy and provide a software solution for the elaboration of the manufacture, due to the internal analyzing within the company.

For the development of Big Data Analytics Tool, several software programs and tools were employed. Java-written server controls all components in the project and visualizes the processed data via a user-friendly client web application. The SQL Server maintains data, observed from the ERP system. Moreover, it is responsible for the login and registration procedure to enforce the information security. In the Hadoop environment, two research methods were implemented. The Overall Equipment Effectiveness model investigated the production data to obtain daily, monthly and annual efficiency indices of equipment utilization, employees' workload, resource management, quality degree, among others. The Machine Processing sample indicated the amount of machines in each working state. The server executes Hive queries via Secure Shell and forwards the analyzed data in the graphical form to the client.

The objectives set in the thesis were completed as the research progressed: as a result of the accomplished work, a fully functional Big Data Analytics Tool was carried out for Hydroline Oy. The work can be used as a RDI project to forecast possible fault rates, diminish the manufacturing loss and, most importantly, gain greater business value by modernizing the whole venture's structure.

Keywords
Big Data, Big Data Analytics, Hadoop, Overall Equipment Effectiveness

ACKNOWLEDGEMENTS

I would like to express my gratitude to the DigiBoost project, coordinated by Savonia University of Applied Sciences, for giving me an opportunity to write this thesis with the subject of Big Data Analytics Tool.

My special thanks go to my thesis supervisors, Arto Toppinen and Anssi Suhonen, for their valuable guidance, immense knowledge, motivation, enthusiasm, understanding and continuous encouragement.

I wish to express my sincere thanks to Jani Lehikoinen, vice president of business development and marketing at Hydroline Oy, for assistance, support and providing the background information on the case studying.

I am heartily thankful to Matti Tiihonen, System Specialist at Hydroline Oy, for the technical support and guidance.

CONTENTS

1 INTRODUCTION 6
1.1 Background and motivation 6
1.2 Objectives 7
1.3 Structure 7
2 BIG DATA 8
2.1 Definition 8
2.2 The Five V’s of Big Data 9
2.3 Applications 10
2.4 Big Data Analytics 12
2.5 Key Technologies of Big Data Analytics 14
2.6 Big Data for the Enterprises 15
3 BIG DATA ANALYTICS TOOL: THEORETICAL BASIS 16
3.1 Idea and concept 16
3.2 Generation of Big Data in the Case Company 17
3.3 The Principle of Work 18
3.4 Big Data Analytics Tool Components 19
3.5 Technical realization of Big Data Analytics Tool 23
3.6 Investment Scenario 24
4 SET UP FOR BIG DATA ANALYTICS TOOL 25
4.1 Installing OS and all components of the Big Data Analytics Tool 25
4.1.1 Windows Server 2012 25
4.1.2 Microsoft SQL Server Express 2012 25
4.1.3 NetBeans IDE 8.1. and GlassFish 4.1.1. 25
4.1.4 VirtualBox 5.0.16 27
4.1.5 Hortonworks Sandbox 2.3.0.0. 27
4.2 Establishing interactivity among elements in the system 30
4.2.1 Microsoft SQL Server Express 2012 – NetBeans IDE 8.1. Web Server Application 30
4.2.2 Microsoft SQL Server Express 2012 - Hortonworks Sandbox 2.3.0.0 31
4.2.3 NetBeans 8.1.Web Server Application – Hortonworks Sandbox 2.3.0.0. 31
5 IMPLEMENTATION OF BIG DATA ANALYTICS TOOL 33
5.1 Client-Server Network 33

5.2 Login and Registration system 35

5.3 Introduction Page 36

5.4 Infrastructure Process Analysis 37

5.5 Machine Processing Analysis 43

5.6 Data Management 45

6 CONCLUSION 47

7 REFERENCES 48

1 INTRODUCTION

1.1 Background and motivation

The exploration of Big Data and Internet of Things has elaborated from writing the research publication "Big Data and its opportunities for the Finnish engineering companies" as part of the Research and Development activities of the DigiBoost project (2015 - 2016), directed by Savonia University of Applied Sciences. The DigiBoost project focuses on scrutiny of novel opportunities for industrial service business cultivation in conjunction with the local industry. The idea of devising Big Data Analytics Tool for a Finnish engineering venture, Hydroline Oy, arose throughout carrying out of the scientific publication. This company collaborated with Savonia University of Applied Sciences before and was employed for a case studying in the specified research paper with an eye to investigate it in the field of Big Data.

In the thesis, Hydroline Oy is denoted as a Case Company that represents itself as a forward-looking engineering company with a long history. The enterprise is engaged in designing and manufacturing standard and customized intelligent engineering products, e.g., smart hydraulic cylinders. Hydroline Oy is a leader in its industry across Finland and is deemed to be one of the most high-performance vendors in its area. The Case Company embodies diverse attributes, peculiar to the employment of the concept of Big Data, thereby it is considered to be the "Big Data" company:

- Effective usage of the material and energy during the production
- Fabrication control in the real-time
- Strategy of innovation and enhancement within the venture
- Customer-oriented policy
- Three-dimensional space design and superb calculation basis
- Minimum impact on the environment

The purpose of this thesis is to explore the company's production system and provide a software application solution on the devising of manufacturing proceeding and maintenance of the service operations, owing to the internal analyzing in Hydroline Oy. Big Data Analytics Tool uncovers hidden patterns, invisible to the limited human perception, and implements real-time analyzing regarding machinery behavior, human resource management, material usage, performance, quality and availably indices of the working processes on a daily, monthly and annual basis. Big Data Analytics Tool gives a complete picture of the business to make both operational and strategic decisions from selecting the product price to establishing the priorities of the vendor's progress.

In order to implement this thesis work, several methodological approaches were deployed:

- Theoretical methods:
  - Literature reviews: collecting data from books, publications, articles and Internet resources as well as analyzing and organizing this data
  - Case studies

  - Interviews in the Case Company and carrying out the research about Big Data application in the company
- Empirical methods:
  - Observation of manufacturing data from the business' ERP[1]
  - Cultivation of Big Data Analytics Tool
  - Establishing the Overall Equipment Effectiveness model and Machine Processing sample of the Case Company and subsequent analyzing
  - Testing the Big Data Analytics Tool
  - Evaluation of the tests' outcome

1.2 Objectives

The corresponding core objectives were set for the thesis:

- Broaden horizons in Big Data and IoT areas, expand technical background and advance occupational skills.
- Devise and cultivate real-time Big Data Analytics Tool for the local engineering enterprise - this thesis brings the Case Company to the new level, since it assists in optimizing the working operations and making the further steps towards the globalization of the business. Big Data Analytics Tool improves the quality of the intelligent engineering production by indicating the fabrication failures and the reasons causing them. The thesis project is highly applicable for Hydroline Oy, since it takes into account the features of the company and its production.
- Obtain the business value for the venture by analyzing the production data and establishing the corporate overall equipment efficiency model and machinery processing sample.
- Provide the Case Company with an opportunity to track, manage and bring alteration in the fabrication procedure by virtue of the analyzed data, gained from the Big Data Analytics Tool.

1.3 Structure

This thesis is divided into six chapters:

- The first chapter introduces the research with an aspect to its key components, employed methodology and assigned goals.
- The second chapter outlines Big Data in different spheres as a definition, characteristics, applications, principal technologies, its significance for the ventures and analytics methods.
- The third chapter discusses the idea, concept, structure, operating principle and theoretical realization of Big Data Analytics Tool.
- The fourth chapter explores the setting up of the environment for the development process of the software system.

[1] Enterprise resource planning (ERP) – a business management software tool that contains a wide range of the integrated applications with an eye to obtain, store and maintain data from various business activities such as production planning, fabrication, stock management, marketing and sales, etc.

- The most significant chapter of this thesis is "Implementation of Big Data Analytics Tool" that describes the technical realization of the software project.
- The last chapter is concerned with the summary of the performed research and determination of further project's proceeding.

2 BIG DATA

2.1 Definition

In the past few years, the topic of Big Data has gained incredible popularity in mass media and scientific world. It is one of the most notable technical buzzwords nowadays (Pries & Dunnigan 2015, 2). The phenomenon of Big Data received a great response in all spheres of human interactivities. Big Data is not only the direction of the industry, but it represents an entire science, capable of forecasting the future prospects. Moreover, Big Data is considered to be the revolution of the digital era, compared with "novel petrol" in the significance level for the society (Rotella 2012, 1). As well as raw materials, the massive data quantity in a pure form involves a much lower key insight value in contrast with the production, derived from data management and analyzing. Big Data is a source of knowledge that brings a superb impact into the routine life.

Needless to say, Big Data is a complex phenomenon. On one hand, this concept includes a massive amount of unstructured data, on the other hand – a set of methodologies and technologies for operating with Big Data (Hopkins 2011, 1).

Enterprises around the world allocate budgets and plan resources to employ data specialists for gathering, scrutinizing and analyzing vast volumes of corporate information. Books, magazines, journals, Internet resources contain plenty of stories concerning how Big Data analysis assists the diversity of companies in magnifying profits, enhancing productivity and efficiency of working processes, optimization of the operating structure or solving different problems. With regard to e-commerce, Big Data is a trump card, which forecasts a customer demand, determines and discovers the target audience, implements unique trade proposals and builds prosperous businesses (Lee 2016, 880-885).

The diverse sources of Big Data surround us everywhere which cover mobile and GPS comprised devices, computers, browsers, social media, search engines, sensors, radio channels, television, and machinery equipment, along with others (Finlay 2014, 17). The amount of data, generated by humankind, is continuously expanding. In the near future, data will be produced by almost each existing object. Nowadays, the quantity of data deployed on a daily basis is greater as opposed to the data volume, utilized for the whole lifespan by our ancestors from the fifteenth century (Woodie 2016, 1).

According to the name, Big Data simply refers to the handling and exploring the massive loads of data. Formerly, "Big Data" stood for operating with the large data quantities, produced by the digital world. A broad range of enterprises and organizations offers dozens of sundry definitions for Big Data. These definitions fully reveal the essence of the "Big Data" term:

- *"Big Data is a process to deliver decision-making insights. The process uses people and technology to quickly analyze large amounts of data of different types (traditional table structured data and unstructured data, such as pictures, video, email, transaction data, and social media interactions) from a variety of sources to produce a stream of actionable knowledge."* (Kalyvas & Overly 2015, 1)
- *"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."* (Gartner IT Glossary 2012, 1)
- *"To define big data in competitive terms, you must think about what it takes to compete in the business world. Big data is traditionally characterized as a rushing river: large amounts of data flowing at a rapid pace. To be competitive with customers, big data creates products which are valuable and unique. To be competitive with suppliers, big data is freely available with no obligations or constraints. To be competitive with new entrants, big data is difficult for newcomers to try. To be competitive with substitutes, big data creates products which preclude other products from satisfying the same need."* (Weathington 2012, 1)

The proposed explanations are descriptive and practical at the same time. Big Data is a substantial technical process that results into valuable key insights, applied for decision-making within the companies.

2.2 The Five V's of Big Data

Douglas Laney, Gartner analyst, proposed the first expository layout of Big Data in MetaGroup issue "3D data management: Controlling data volume, variety and velocity" in 2001. He characterized the traditional data operations as storing, scrutinizing and visualizing in the 3V's model, named after its three primary attributes: (Doug Laney 2012, 1)

1. ***Volume*** - the massive quantity of data produced each second by appliances, human intercommunion and networks in sundry resources as social nets, media, search engines, science and research interactivities, telecommunication systems, healthcare and public health services, airline and tracking applications, e-commerce, etc. (Sarma, Rai & Borah 2014, 112-115). Ninety percent of existing global data is generated during the past 2 years. (Big Data, for better or worse: 90% of world's data generated over last two years 2013, 1)
2. ***Velocity*** - the rate at which vast and continuous data flow is generated, transferred and researched. The New York Stock Exchange collects nearly one terabyte of data concerning the trades on the daily basis. (NYSE Euronext 2014, 1)
3. ***Variety*** - the number of resources and different types of data: structured, semi-structured and unstructured. The data comes in a wide range of formats: photos, videos, sensor readings, au-

dio records, text files, numeric data in the relational databases, transaction entries, etc. Due to the extreme popularity and utilization of social networks, the data expands tremendously. Hence, 80-90 percent of all data, existing in the world, is considered to be unstructured. (North & Riniker 2014, 430) The diversity of unstructured data causes challenges for accumulating, mining and investigating data.

With the time, the 3V's model was not able to fully depict the constantly increasing data volume, the diversity of resources and contents of Big Data. Therefore, the International Business Machines Corporation (IBM) presented the fourth parameter to define Big Data – ***veracity***. (The Four V's of Big Data 2013, 1)

Veracity applies to the indeterminacy in the information, on the grounds of its inconsistency and deficiency, causing the complications for keeping data structured. The regular companies with the annual profit of approximately a billion of dollars and higher lose around 130 million of dollars per year due to scarce data maintenance. (The Cost of Mismanaged Data 2015, 1)

The 4V's model was not able to entirely cover the functionality for the standard analytics procedures to operate efficiently and opportunely. Heretofore, the fifth attribute was appended to the current designation of Big Data – ***value***. (Bienh 2013, 1)

Owing to the dynamic data analyzing, the expedient exploitation of the massive amount of information, gained from the vendor's lifespan, outcomes into supreme income and a broad range of business opportunities. Processing the enormous data volumes retrieves the significant business insights, implicated in the structured, semi-structured and unstructured data flows of the company. These business values are employed in the enhancement of the supply chains, monitoring of sales and trading activities, gauging the performance of the systems and turn into the on-demand occupation. The strategy of Big Data is to offer a business an opportunity to rev up the expansion of the profits by analyzing its data. (Knilans 2014, 1)

2.3 Applications

The world leaders of industry, science and innovation claim that Big Data has turned out into being a technology revolutionary and a game changer in the majority of the business directions during the past few years. Since Big Data became the actual part of the everyday life, the attitude towards it evolved from the derivable agiotage to discovering the authentic value, attained from the proper employment of data. Not only retrieving the insights of Big Data is an invocation, but also the practical issues appear, embracing funding, technological capabilities that persist on foreground for plenty of diverse industries, assimilating Big Data. (Bilbao-Osorio, Dutta & Lanvin 2014, 3-8)

According to the Gartner Survey, more than three-quarters of the enterprises worldwide are in the process of devoting money into Big Data technology or planning to do it in a couple of years. The results of the survey showed that the number of vendors increased by 17 percent, comparing to a similar survey, carried out in 2012. (Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years 2015, 1)

Maryanne Gaitho designated several industrial directions, which employed Big Data in the concept, the problems, faced by ventures, and the applications of Big Data in resolving the challenges. (Gaitho 2015, 1)

1. ***Finance sector and Security*** - A research of sixteen projects in ten leading corporate and retail banks delineated the issues that the industry was struggling: preliminary notification about fraudulence with the tradable financial assets, real-time effectuation of lending analyzing methods, revealing swindle with payment cards, the archive of audits, documentation of business credit risk, customer records, analytics of public commerce and information technology procedures, etc. (Costley & Lankford 2014, 3-14)

   For instance, the Securities Exchange Commission is employing Big Data in order to track the processes, occurring in the economic arena. Moreover, SEC applies network analytics as well as innate language processors to capture illicit commercial activities in the market. (Jackson 2015, 1)

   The enterprises in the financial sector utilize Big Data for: (Gross 2014, 1)

   - Business analytics, deployed in the sentiment scaling, analytics in the decision-making before trading, financial forecasting, HTF[2], along with others.
   - Risk examination, occupied in striving against money laundering, internal company's risk control, client behavior and fraudulent schemes or actions.

2. ***Mass media, gaiety and utilities areas*** - this industry direction is carried out in various forms. For example, a user can access social networks, read an article by a simple query in Google search engine by utilizing computer, cell phone or tablet. Therefore, mass media, gaiety and utilities industry incurs the multiplicity of challenges:

   - Congregating, analyzing and sustaining data, related to customer demand and user profile.
   - Making more sufficient, engaging and efficient maintenance of the mobile services, social networks, etc.
   - Revealing modern patterns in media, entertainment and utilities applications.

   The companies in this sector contemporaneously investigate client records in conjunction with the observable data to establish consumer portraits, conducted for producing the on-demand services and applications for different target audiences and measuring the upkeep capacity. (Wolkowitz & Parker 2015, 3-9)

   Spotify is an illustration of an entertainment company that adopts Hadoop Analytics for deriving the information about millions of customers, located throughout the world, and further data processing to procure corresponding music choices to the individual clients. On the other hand, there is a company, Amazon Prime, which intensively employs Big Data to offer the analyzed selection of products, according to the user profile and previous shopping experience. (Gaitho et al. 2015, 1)

3. ***Public health services*** - This industrial sector operates with Big Data in the decision-making process during the medical treatment, establishing the international patients' databases, handling the readings from the wide range of sensors, and so on.

[2] HTF- an abbreviation for Hybrid Tensor Factorization, a freeware ganged tensor factorization set of services, designed to execute algorithms in a reliable and high-performance manner.

Beth Israel medical institution utilizes the mobile application, where the data about the numerous amount of patients is accumulated, to permit physicians to implement based on the evidence treatment. In addition, the University of Florida embeds available healthcare data and Google Maps to construct visual records for quicker detection and efficacious scrutinizing of information, concerning the development of chronic ailments. (Gaitho et al. 2015, 1)

4. ***Education*** - Big Data brings substantial impact into higher education. Specifically, the University of Tasmania grants education for more than 26000 students. This university introduced "Learning and Management System" for monitoring students' login into the service, the amount of time, contributed on different applications. (Connolly 2014, 1)

   Furthermore, Big Data assists in the evaluation of the lecturers' performance to provide valuable experience for either education workers or students. The efficiency is evaluated according to the number of students, scrutiny subject, goals of the students, determination of the behavior and other attributes.
5. ***Transport facilities*** - In past few years, the extensive data volume has been engaged in the location-associated social networks and fast data transmission. Administration institutions apply the benefits of Big Data in traffic control, smart transportation, route blueprint, overload maintenance via forecasting the transportation state. Enterprises employ Big Data in transportation services for: technological improvements, profit handling, optimization of freight traffic to achieve the strategies for competitive assets. A single user implies Big Data for a blueprint of itinerary to preserve fuel and time consumption, adjust navigation routes, etc. (Thakuriah & Geers 2013, 5-7)
6. ***Government*** - Public facilities contain a broad spectrum of Big Data utilization: an exploration of energy resources, commercial arena analysis, the reveal of swindle, healthcare investigation and environmental security. For instance, the Food and Drug Administration embodies Big Data to discover and contemplate samples of foodborne illnesses and sicknesses. Thence, the more effective treatment is discovered and mortality causes are eliminated. (Beaudoin 2015, 1)
7. ***Energy industry*** - In this type of industry, Big Data is employed for superior resource and workforce maintenance by preliminary identification of the system's failures. In addition, the smart meter readings are terminated to estimate the consumption of energy, deployed for the analyzing of the demand.

2.4 Big Data Analytics

Big Data Analytics can be described with the following chain of actions with data to uncover sequences or relationships and reveal beneficial observations: (FIGURE 1)

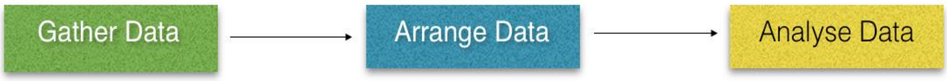


FIGURE 1.The processing steps of Big Data Analytics

Big Data Analytics offers a company an insight view within its structure and brings in superb information for current and future business solutions. The goal for Big Data scientists is to gain knowledge, derived from the data processing.

Big Data Analytics applies Business Intelligence in its concept, where "*Business Intelligence (BI) is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining."* (Bert Brijs 2013, 6).

Marc Schniederjans, Dara Schniederjans & Christopher Starkey defined three categories of Big Data Analytics in the book "Business Analytics Principles, Concepts, and Applications: What, Why, and How" (TABLE 1). (M. Schniederjans, D. Schniederjans & Starkey 2014, 4)

TABLE 1. Types of Big Data Analytics (Marc Schniederjans, Dara Schniederjans & Christopher Starkey 2014, 4)

Type of Big Data Analytics	Objectives	Illustrations of the methodology
Descriptive	Detecting potential patterns in the extensive data storages. The goal is to receive a general view on the current data and theoretical attributes of the information that potentially can discover tendencies or distinguish prospective business behavior.	Descriptive statistics: ○ Dimensions of centric patterns, e.g., median, mean ○ Dispersion classifications, i.e., standard deviation ○ Diagrams ○ Analysing techniques; periodicity, probability allocations - Bernoulli Trials, discrete and continuous distributions, combinatorics ○ Assaying mechanisms
Predictive	Determining and forecasting the future tendencies.	○ Statistical procedures: multiplex linear regression, analysis of variance, etc. ○ Conversant frames: data analysis and classification ○ Exploratory techniques: predictive samples
Prescriptive	Distributing expediently the resources to obtain benefits of forecasted tendencies or potential possibilities.	Exploratory techniques such as linear programming and resolution theorem.

2.5 Key Technologies of Big Data Analytics

Big Data endorses various technologies to fully operate with the information: gather, store, handle, analyze and visualize data. The following key technologies were designated for the interaction with Big Data: (Hu & Kaabouch 2014, 2):

- ***Hadoop*** – a freeware software service that accumulates immense data quantities and executes applications on the aggression of commodity computers, using Java object-oriented language as a default. Since Hadoop's distributed file structure (HDFS) rapidly handles permanently enlarging data volumes and diversities, this technology is commonly utilized by a business sector. (Big Data Analytics: What it is and why it matters 2016, 1)
    - *MapReduce* – a software service for the elaboration of applications, managing immense data sets simultaneously on the group of commodity computers in the secure and efficient way. (MapReduce Tutorial 2008, 2)
    - *HDFS* – a distributed file system for processing and transferring extensive amount of data using MapReduce as a template, whilst the interface is designed on the example of UNIX file service. (Chansler, Kuang, Radia, Shvachko & Srinivas 2016, 1)
    - *Hive* – a data management service that exploits structured data, stored in the HDFS, through terminating the queries via HiveQL language, resembling SQL[3]. (HIVE: hive query language 2015, 2)
    - *Sqoop* – a HDP tool for conveying data betwixt HDFS and relational database management software. (HIVE: hive query language et al. 2015, 2)
    - *Pig* – a platform, based on the procedural programming language, utilized for coding to perform MapReduce jobs. (HIVE: hive query language et al. 2015, 2)
- ***NoSQL*** - a database infrastructure that accomplishes high-efficiency, flexible processing of the vast amount of information. The most popular examples of NoSQL databases are Apache Cassandra, MongoDB and Oracle NoSQL. Relational databases operate with the well-structured data, whereas NoSQL data management tools utilize as a foundation a conception of the distributed storage systems and interact with the non-structured data, accumulated across several analyzing nodes and servers. Due to the distributed structure of NoSQL, the program is flexible - during the magnification of data amount, it is necessary to append more hardware components to retain the efficiency. The world most leading data warehouse enterprises, e.g., Google Inc., Amazon Inc., employ this distributed software for data maintenance. (Lo 2015, 1)
- ***Massive Parallel Processing (MPP)*** – a data management system, cultivated for executing simultaneously several procedures in parallel by numerous amount of the operating blocks, which improves the productivity rate while working with the immense data sets. MPP includes an extensive number of multi-core processors with their operating systems and memory storages, servers and storage devices, capable of parallel cultivation, to process data fragments across diverse operating units contemporaneously to boost the velocity. The majority of the companies

[3] SQL – an abbreviation for structured query language

and organizations apply MPP for maintaining tremendous data volumes. (MPP database (massively parallel processing database) 2015, 1).

- ***In-memory data processing*** – A company can perform more sufficient business decisions, attain significant data comprehension and perform recurrent and interactive analytics scripts through fetching data, located in the system memory, and increasing rate, capacity and reliability when making data requests. (Lo et al. 2015, 1)

2.6 Big Data for the Enterprises

The main users of Big Data are companies that accumulate and process the information concerning their clients to make decisions, impacting business performance.

Big data is a vital topic of the modern technical world, since there is a great amount of sundry data sources, located everywhere. There are search engines and social media, observing users' behavior, a variety of sensors which monitor data in real-time, e-commerce, smart traffic administration, mobile devices, accompanying people around the clock, and it is only a few resources that are specified. The analyzing of data, collected from such resources, can detect new consistent patterns that cannot be distinguished by traditional methods. Big Data provides a wide range of opportunities to the enterprises such as (Petty 2014, 1):

- *Creating novel applications and services* – The processing of Big Data includes the acquisition of real-time records in relation to production, clients and medium to reach business targets, e.g., optimize user experience, significantly minimize expenses and eliminate resource usage. For instance, one of the U.S. cities reduced the crime rate and developed municipal services by utilizing MongoDB that was responsible for handling geospatial data in the real time mode. (Petty et al. 2014,1)
- *Reduction of expenses and raising the efficiency factor* – The expensive and complicated solutions were substituted with freeware Big Data technologies. Tier 1 is a superb example of how the implementation of Big Data can exceedingly decrease company's expenses. This vendor transferred data control system to MongoDB, therefore essentially minimizing outlays on hardware and official licenses. (O'Dowd 2015,1)
- *Carrying out the neoteric sides of competitive edge* – Big Data adds a flexibility feature to organizations throughout adjusting to changes more rapidly than the contenders.
- *Enhancing approach to the client* – Due to the amplification of data volume and its rapid analyzing, Big Data empowers companies to reply in the fast and thorough manner to client inquiries and virtually "listen" to the customers.
- *Defining the ground of production defects in near real-time mode* - One of the primary advantages of Big Data is to identify the causes of the fraudulence by investigation of the operation of all systems within the enterprise.

3 BIG DATA ANALYTICS TOOL: THEORETICAL BASIS

3.1 Idea and concept

Big Data is a technological boost that stands in one line with the Internet and even telegraph. (Mayer-Schonberger & Cukier 2013, 97) Big Data can be described by three world-leading parameters: novelty, productiveness and rivalry. It represents itself a revolutionary science that enables a prediction of the future due to the rapid handling of the vast amounts of data and its instant analyzing. Big Data discovers incredible business ideas, uncovers potential opportunities as wells as possible issues, retrieves latter-day income sources and overcomes obstacles to the realization of new systems. The companies, which employ the benefits of Big Data, obtain high business results to increase competitiveness. In the modern world, the volume of Big Data is always increasing. To demonstrate the scale of Big Data, James Kalyvas and Michael Overly introduced next facts in their book "Big Data: A Business and Legal Guide":

- The amount of information, generated on the Internet by 2004, was estimated as one petabyte. The global television content for a century is approximately the same size.
- In 2011, Big Data achieved one zettabyte or one million of petabytes. This data quantity can be compared with the collection of HD videos with the full duration of 36 million years.
- Big Data reached 7.9 zettabytes or 7.9 million petabytes in 2015.
- As about the future forecast, in three years Big Data will attain pari passu a thousand zettabytes.

Consequently, the traditional data management tools cannot completely maintain the Big Data in processing, analyzing, storing and visualizing the rapidly enlarging information quantity anymore. The novel tools for Big Data handling have been introduced and elaborated to operate with the frequently updated information of massive volumes, diverse contents and different sources. The aim of Big Data Analytics Tools is to increase working efficiency, create new products and services, add novel features to the existing production and enhance competitiveness. In order to receive a full advantage of Big Data, companies use the data maintenance services, establish complex algorithms and carry out large-scale investments in information infrastructure and software. (McAfee & Brynjolfsson 2012, 1)

Big Data Analytics Tool is a system of collaborating specialized software tools, programs and applications for data optimization, data mining and predictive analytics. By utilizing this tool, the venture can rapidly analyze the massive data volume, collected over time, with an eye to drive more advantageous business solutions in the future, taking into consideration the processed data. The concept of Big Data Analytics Tool includes next steps:

1. A collection of the information, generated in the production and infrastructure processes, which is stored in the ERP of Case Company.

2. Real-time exploration of the received data regarding the equipment utilization and working load, human resources operation, production expenses and waste, efficiency, quality and performance attributes and other parameters to gain the business insights.
3. Visualization of the scrutinized information in a user-friendly manner through the website.

Big Data Analytics Tool offers the real-time processing of the production data, owing to the establishment of a daily, monthly and annual Overall Equipment Effectiveness model and machine processing analysis. Hydroline Oy will be able to track the infrastructure process, identify the potential fabrication failures, cultivate novel products or services, optimize the existing working procedures and elaborate current manufacture through completing the superior action plan to decrease costs, production waste and environmental impact.

3.2 Generation of Big Data in the Case Company

When the company starts to collect and utilize data from its suppliers and vendors, the massive amount of information is gathered. This data can be analyzed, visualized and employed with the perspective of enhancing the decision-making. Owing to the combination of data, generated within the company, and the information from the public resources, the volume of data, available for analysis, can grow rapidly and unboundedly. This event is a complementary benefit for Big Data application. Big Data assists in bringing up completely novel breakthrough solutions, i.e., the search of a new market and the expansion of the target audience.

Big Data is not measured only by the volume and its growth rate, but also it is categorized by the diversity of the collected information. The power of Big Data is the combination of various sets of structured and unstructured data that can be a source of discovering novel progressive decisions. There is a wide set of categories, describing the unstructured data: social entries, visual information, documents and presentations, e-mail messages, websites and even voice records. An analysis of several small datasets can result in extraordinary business solutions as well as a processing of the massive data amounts. Hence, the exploration of terabytes or gigabytes of information provides small and medium size companies an opportunity to come to the same level solutions as the global corporations do, backed with petabytes or exabytes of data. (Hassanien, Azar, Snasel, Kacprzyk & Abawajy 2015, 13-20)

Big Data technologies contribute to the progress not only of high-tech companies, but also the members of the traditional industries. Soon, Big Data and Internet of Things will become the essential components of every business sector in the whole world. Therefore, the question arises: How does the Case Company relate to Big Data?

Hydroline Oy is an illustration of Original Equipment Manufacturer that produces constituent parts for the end-products of their clients. Big Data is generated at each step of company's operation (FIGURE 2):

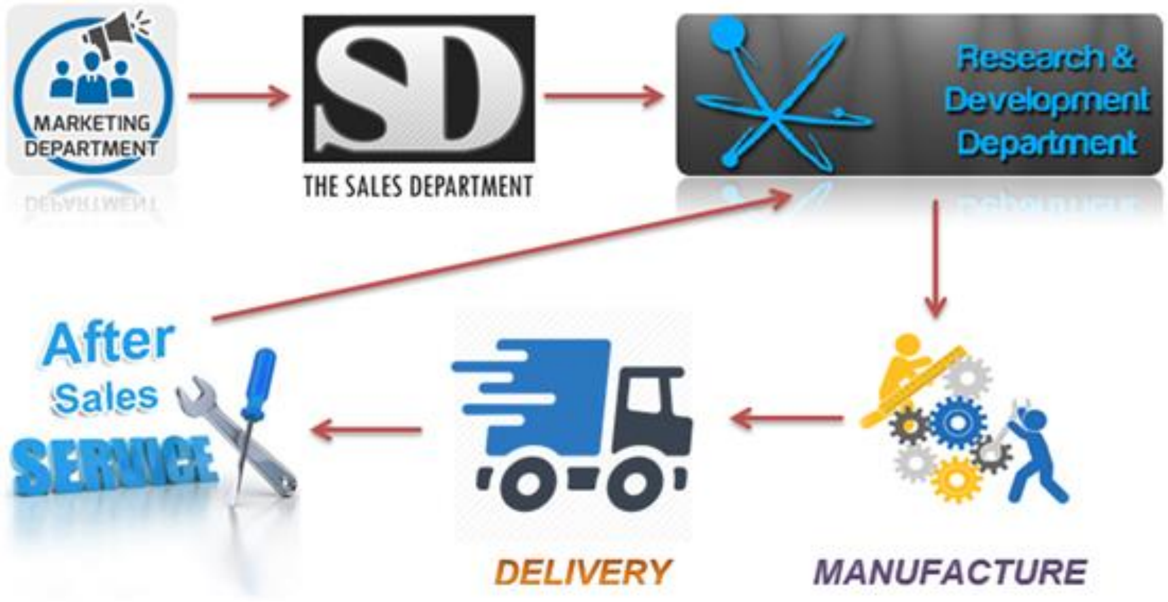


FIGURE 2. Big Data Flow in an OEM Company [UPPM 2016; Edward Robirds 2016; Duct Tape Marketing 2015; Halogen Software 2012; Bright Powertech Ltd 2012; Iconfinder 2015]

1. To begin with, the marketing department gets in touch with a client and obtains significant information about the order like:

- an amount of units;
- specifications and properties of the product;
- fixed or differentiated price of the commodity, including costs of work and services;
- client's contact information;
- timeframes;

2. Furthermore, the sales sector organizes the deal and sends the transaction data to the Research and Development division.
3. According to the received information, R&D team implements the plan of fabricating the product and passes it to the manufacturing department, whereas novel Big Data records are produced.
4. The manufacturing sector fulfills the order, regarding the observed information from the foregoing branches, and spawns more production data, generated during the industrial process.
5. In the long run, the final product is supplied to the customer. Afterward, Hydroline Oy offers a client the technical maintenance. The data, gathered from the after sales service, is employed by the Research and Development team in order to improve the current production and introduce new features, basing on the customers' requests.

3.3 The Principle of Work

FIGURE 3 presents the working principle of Big Data Analytics Tool, applied in the Case Company. According to the scheme, the interplay between Big Data and the company is described by the following steps:

1. The production data, generated during the manufacturing of the Intelligent Engineering Products, is forwarded to the company's ERP where the data is stored in the relational databases. Moreover, the Case Company preserves customer's feedback into the ERP as well.
2. Next, the ERP software automatically transmits raw data to the Big Data Analytics Tool that allocates it in the SQL Server and makes the analysis to present the working efficiency of the vendor in various areas: machinery utilization, time consumption, performance rate, fraudulent coefficient, quality degree, etc.
3. Finally, the Case Company can track and streamline the workflow, manage the business processes and carry out new improvements in the infrastructure process.

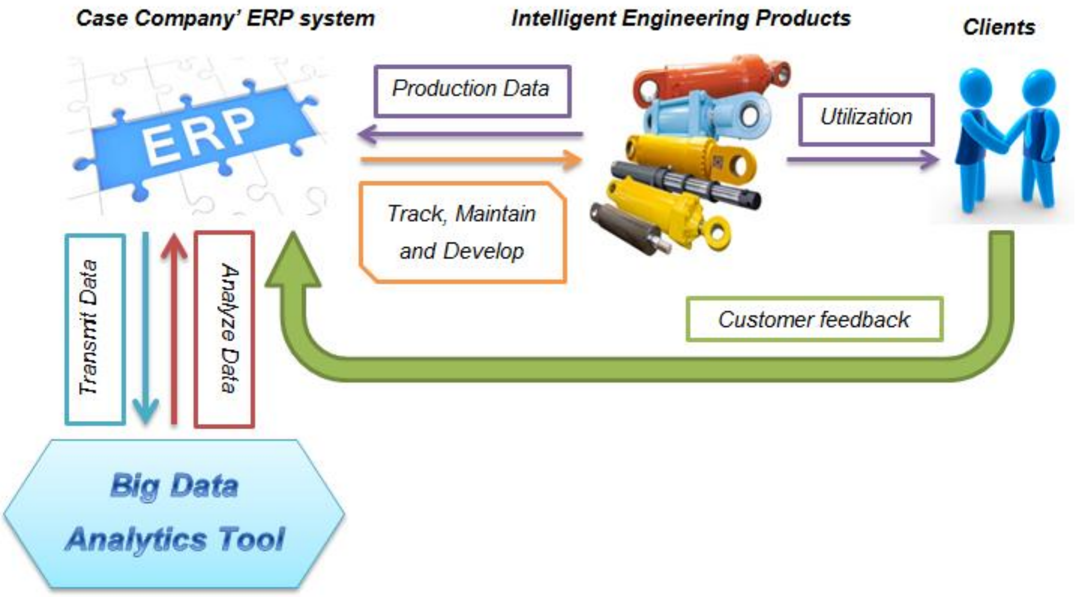


FIGURE 3. The working principle of the Big Data Analytics Tool [Aboard Software 2014; Hydraulex Global 2014; Johnson Pet Trade Consultants BV 2013]

3.4 Big Data Analytics Tool Components

For creation and elaboration of the Big Data Analytics Tool, a number of software programs and tools were deployed as a core of the system:

1. ***SQL Server Express 2012*** – a powerful and secure relational database management system, designed by Microsoft Corporation. This easy-to-use program provides extensive and reliable data storage for web applications and desktop services as well as analysis, integration and notification enclosures. (Microsoft® SQL Server® 2012 Express 2012, 1)

One of the distinguishing features of this database engine is that it is free software tool, available on the official website of Microsoft. The downloaded package includes the kernel SQL server database engine with the following GUI tools for the maintenance of SQL Server instances and databases:

- ***Configuration Manager*** – a software program for operating with the utilities, related to the SQL Server. For instance, this service includes configuration network protocols, employed by the database engine, drivers for the establishment of network connection, etc. (SQL Server Configuration Manager 2016, 1)

- ***Surface Area Configuration*** - a tool for checking what kind of features, services and connections are set up and running on the SQL Server. (SQL Server Surface Area Configuration 2006, 1)
- ***Management Studio Express*** – a software application that sustains, mounts and controls all of the elements in SQL Server through the script editors and graphic tools. (Imran 2014, 1)
- ***Business Intelligence Development Studio*** or ***BIDS*** – an integrated devise environment, designed by Microsoft Corporation, for enhancement of data analysis and business analytics solutions. The functionality was based on the duties, provided by the SQL Server, such as processing, reporting and merger. BIDS is an association of Microsoft Visual Studio IDE with the certain SQL Server's fields: extension tasks, various samples of projects, tools, ETL[4], online analytical processing cubes and designing of data management. (Integration Services in Business Intelligence Development Studio 2014, 1)

2. ***Oracle VM VirtualBox*** – a reliable open-source virtualization software program, designed by Oracle Corporation for 32-bit and 64-bit AMD and Intel processors. (VirtualBox 2015,1)

   This hypervisor's functionality includes creation, maintenance and exploitation of unmodified operating systems in a certain environment, yclept a virtual machine. VM[5] is configured by virtualization software that grants access to the specific hardware constituent elements and specifications and drives above the initially installed OS[6]. Generally, "host" is represented by the physical device, whereas the virtual machine is the "guest". The subordinate operating system or guest OS is carried out on the host computer and considers itself as it is executing on the real physical device. (Virtual machines 2015, 1)

   This free of charge hypervisor can be downloaded from the VirtualBox official website (https://www.virtualbox.org/).

   Oracle VM VirtualBox offers a wide range of guest virtual machines that are running diverse versions and derivations of operating systems like Microsoft Windows, Linux, Solaris, Mac OS X, BSD, etc. (VirtualBox et al. 2015,1)

3. ***Virtual Machine Hortonworks Sandbox 2.3.0.0.*** – a portative virtual environment, based on the Hadoop technology for rapid distributed analyzing of the massive amount of data. The base of Sandbox Hortonworks is a powerful Hortonworks Data Platform where the functionality is enhanced by appending new data management tools and applications. (Learning The Ropes Of The Hortonworks Sandbox 2015, 1) The product was elaborated by American business computer software

[4] ETL – three operations with database: extract, transform and load (Shelley 2012,1)

[5] VM – an abbreviation for Virtual Machine

[6] OS – an abbreviation for Operating System

company Hortonworks and is freeware to utilize (http://hortonworks.com/products/hortonworks-sandbox/#archive).

Hadoop – an open-source, Java-based software framework for keeping data and executing services on the commodity hardware, released by Apache Software Foundation. Hadoop offers a large storage for any type of data, vast analyzing capacity and the capability of carrying out verily a boundless number of the parallel objectives or jobs. (Hadoop Big Data Analysis Framework 2014, 6-7)

Kumar Gauraw defines several Hadoop features in the article "Big Data and Hadoop – Features and Core Architecture": (Kumar Gauraw 2015, 1)

- *Scalable* - the key factor, due to the constant enlargement of data volume and information diversity that is caused majorly by the social nets and Internet of Things. Hadoop provides rapid management of the extensive amount of information of any type. This Big Data technology gives the business an opportunity to accomplish numerous applications on the multiplicity of nodes simultaneously, including hundreds of terabytes in the data flow. (Isaacson 2014, 1)
- *Processing power* – Hadoop employs out-and-outer distributed file arrangement which is founded on the principle of "mapping" data. Usually, the applications for data analysis are settled at the same machines, where the information is kept, to terminate data processing quicker. Hadoop is capable of observing effectively the immense amount of non-structured data. For example, it takes several minutes for this Big Data technology to process terabytes of data, whereas for petabytes - this process lasts for few hours.(Kumar Gauraw et al. 2015, 1)
- *Cost efficient* – In the past, there was a vexed problem with the analysis of data, since the traditional database management tools were not able to provide full processing of continuously enlarging volume and variety of information as well as they were extremely expensive to afford. With an eye to diminish the costs for processing data, the majority of the companies would simplify the data flow by eliminating the information that was potentially less valuable from their point of view. The raw records were erased, since it was too expensive to keep them. In time, business targets have evolved and this kind of approach to the data analysis has lost its power. Contrariwise, the Hadoop technology was settled as a wide-scale system, apt of accumulating the whole enterprise data for the future utilization. The economic benefit is colossal. Instead of spending hundreds or thousands of euros for processing one terabyte, Hadoop proposes storing and analyzing data for freeware. (Kumar Gauraw et al. 2015, 1)
- *Fault-tolerant design* – Hadoop proceeds to operate duly in case of the hardware failure, guaranteeing the preservation of data and running processes. In case the unit stops functioning, all tasks are automatically reoriented to the other units in the cluster to keep the distributed computing procedure up. Moreover, it automatically creates several copies of data, stored therein. (Achari 2015, 15-16)

- *Flexibility* – In advance of retaining data, preliminary processing is not required, in contradistinction to the traditional data management tools. There is no limitation on the volume and the variety of stored data, it can be even non-structured data such as images, emails, web pages, videos, etc. (Achari et al. 2015, 15-16)

FIGURE 4 displays the structure and processes of this virtual machine. The principle of Sandbox Hortonworks' employment can be divided into three steps:

1. *Loading data*: the records are collected from the external data sources, e.g., company's ERP, machinery logs, generated from the production. A structured set of data is transmitted and accumulated in the Hadoop distributed file system through Sqoop technology. For instance, Sqoop is capable of importing tables, loading external information into the datasets in HDFS from the relational databases, stored in Microsoft SQL Server, MySQL, PostgreSQL, Oracle, etc.
2. *Analyzing data*: HCatalog[7] offers users sundry data analyzing tools, for example, Pig, MapReduce, Hive, and creates the relational view of data. Hive and Pig are used for analyzing, investigation and structuration of data to obtain the business insight values.
3. *Visualizing data*: The processed data can be transferred into Microsoft Excel through ODBC[8] connector and visualized by the Power View function in Excel. The explored data can be represented in different ways, e.g., charts, graphs, maps, matrices, cards, tiles, etc.

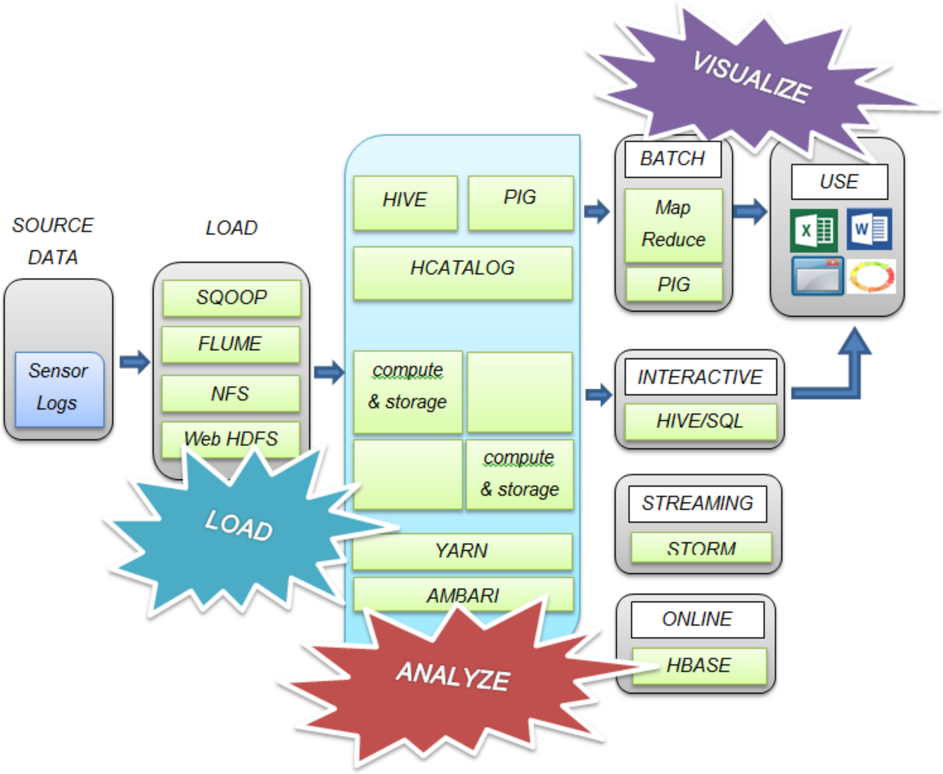


FIGURE 4. The structure and processes of Sandbox Hortonworks VM (Hortonworks 2015)

[7] HCatalog – a data management layer that grants access to Hadoop's services and tools.

[8] ODBC – an abbreviation for Open Database Connectivity

NetBeans IDE 8.1 with GlassFish 4.1.1. – NetBeans is an open-source Java-based integrated development platform that provides a wide range of programming opportunities for software programmers. This environment is foremost designed for coding in Java language; however it also supports the next languages: C, C++, PHP, JavaScript and HTML5. In addition, it is a cross-platform, meaning that you can use this software program on Microsoft Windows, Linux, Mac OS, the number of other operating systems. (NetBeans IDE 8.1 Installation Instructions 2015, 1)

GlassFish is the world's first freeware application server, designed by Sun Microsystems for Java Enterprise Edition platform, and now financially supported by Oracle Corporation (GlassFish Server Open Source Edition 2013, 1). In the book "JavaServer Faces: Introduction by Example", written by Josh Juneau in 2014, GlassFish server is illustrated with three primary authorities: (Juneau 2014, 1-2)

- *Application server* – executing Java Enterprise Edition applications, for instance, servlets, java server pages and enterprise java beans.
- *HTTP server* – operating with the web requests that are usually called from the browsers.
- *Servlet container* – maintaining servlets and java server pages.

3.5 Technical realization of Big Data Analytics Tool

FIGURE 5 represents the scheme for the technical fruition of the Big Data Analytics Tool. Functionality of each component of the system, as well as the performance of the entire software product, are described by the consecutive sequence of operations:

1. NetBeans IDE 8.1 with GlassFish 4.1.1. is a central element of Big Data Analytics Tool, not only it intercommunicates and controls all other components in the project, but it also visualizes the analyzed data through the user-friendly client-server web applications.
2. Microsoft SQL Server Express 2012 intercommunicates with the ERP software of the Case Company, stores and maintains the production data in the tables in the database. Moreover, SQL Server is responsible for the login and registration procedure for the Big Data Analytics Tool, therefore enforcing the information security. In addition, it grants access to the data for all of the components in the project.
3. The Oracle VirtualBox was utilized in the Big Data Analytics Tool to further upload into the environment Hadoop virtual machine and operate with it. Hortonworks Sandbox 2.3.0.0. is one of the major elements of the Big Data Analytics Tool which primary function is real-time data analyzing through the Hadoop technology. By deployment of the Sqoop tool in the NetBeans environment, the structured data is stored in HDFS and Hive software applications in the virtual machine.
4. In the HDP[9], several analysis methods were carried out. One of them was Overall Equipment Effectiveness model that processed manufacture data to obtain daily, monthly and annual efficacy parameters of facilities' operation, employees' workload, working time's consumption, produc-

[9] HDP – an abbreviation for Hadoop Platform

tion's quality and performance. The second research sample was machine processing that indicated a number of machines in every working phase. The analyses were devised by coding in Hive facility. According to the client request, Hive forwards the examined data to the server that converts it into the graphical form, for instance, diagrams and graphs.

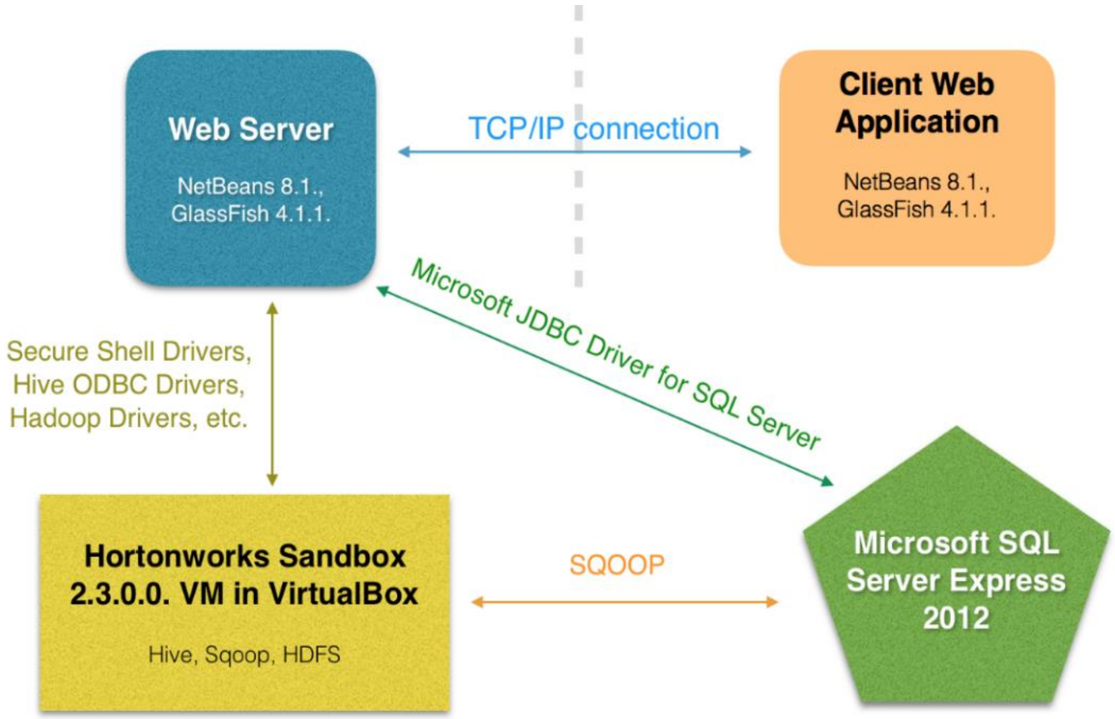

Web Server
NetBeans 8.1., GlassFish 4.1.1.
TCP/IP connection
Client Web Application
NetBeans 8.1., GlassFish 4.1.1.
Microsoft JDBC Driver for SQL Server
Secure Shell Drivers, Hive ODBC Drivers, Hadoop Drivers, etc.
Hortonworks Sandbox 2.3.0.0. VM in VirtualBox
Hive, Sqoop, HDFS
SQOOP
Microsoft SQL Server Express 2012

FIGURE 5. The structure of the Big Data Analytics Tool and interaction among all components

3.6 Investment Scenario

One of the primary criteria for the fruition of the Big Data Analytics Tool in the Case Company was the cost-efficiency. According to the accomplished work, the specified prices should be taken into account:

1. The hardware employed for the Big Data Analytics - the only equipment required in this project is the computer with the available computer data storage (RAM) more than 4 GB and a 64-bit processor. The computer with 8 GB RAM was utilized for the Big Data Analytics Tool, since it made the virtual machine to operate faster and minimized time consumption for implementing the data analyses.
2. The official version of the OS (Windows 7 or later edition) for handling a set of interlinked desktop applications and services like Excel, Word and other programs for the visualization of data analyzing.
3. The software used for the Big Data Analytics - the free of charge software tools and applications were deployed in the project: SQL Server Express 2012, NetBeans IDE 8.1., GlassFish 4.1.1, and Hortonworks Sandbox 2.3.0.0.
4. The cost of programmer's labor and further maintenance of the system.

4 SET UP FOR BIG DATA ANALYTICS TOOL

4.1 Installing OS and all components of the Big Data Analytics Tool

4.1.1 Windows Server 2012

Savonia University of Applied Sciences proposes student accounts in Microsoft DreamSpark where the registered user is able to buy or obtain a free trial version of software programs and tools for learning and exploring information technologies. The DreamSpark website[10] was used for downloading a Standard Windows Server 2012 trial version for a year.

As the first step towards the implementation of Big Data Analytics Tool, 64-bit Microsoft Windows Server 2012 OS was installed on the host computer. Taking into account its wide range of features and enhancements as a user interface, storage spaces, scalability, networking, task management, security, this operating system was chosen to fulfill all of the preliminary needs of the project. (Ferrill 2013, 1-12)

In accordance with the system requirements, Windows Server 2012 operates only with 64-bit CPU. Unlike the previous versions, this release does not support x64 Intel microprocessors that carry out Itanium architecture. (Niccolai 2010, 1)

4.1.2 Microsoft SQL Server Express 2012

To begin with an employment of Microsoft SQL Server Express 2012, this data management tool was downloaded for free from the provider's website[11].

During the setting up process, the first default SQL instance was created, called SQLEXPRESS. The following attributes were added to enhance its capabilities: data management tools, SQL client software development kit and documentation elements. Moreover, the automatic launching for the database engine and browser was defined. In addition, “Mixed Mode” identification was selected, allowing SQL Server and Windows authentication simultaneously. This configuration parameter is important for the further data shift to the Hadoop environment, enabling the transmission of records. Withal, it grants data access to the NetBeans software.

The last step in the configuration of the SQL Server was a specification of the username and password for a system administrator’s account to enforce the security.

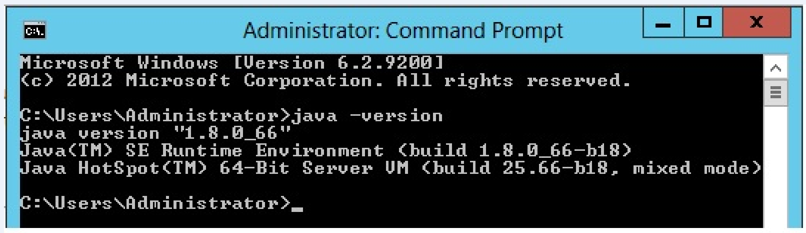
4.1.3 NetBeans IDE 8.1. and GlassFish 4.1.1.

The fundamental requirement for the mounting of NetBeans and GlassFish is the presence of the Java Development Kit on the computer. Thereafter before setting up NetBeans IDE, the version of

[10] DreamSpark website - https://www.dreamspark.com/Product/Product.aspx?productid=42

[11] Microsoft SQL Server 2012 Express official provider’s website - https://www.microsoft.com/en-us/download/details.aspx?id=29062

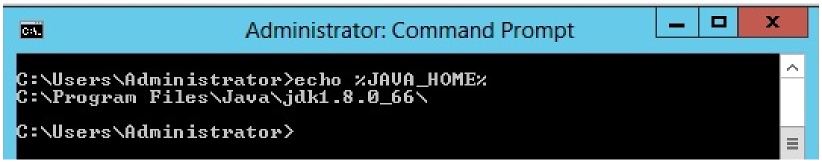
JDK was verified. For instance, the default version can be defined by typing "java -version" command in the prompt window (FIGURE 6). In case Java Development Kit is not settled on the computer, it is available for a free download at the Oracle's website[12].



```
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>java -version
java version "1.8.0_66"
Java(TM) SE Runtime Environment (build 1.8.0_66-b18)
Java HotSpot(TM) 64-Bit Server VM (build 25.66-b18, mixed mode)

C:\Users\Administrator>_
```

FIGURE 6. The verification of the default JDK version

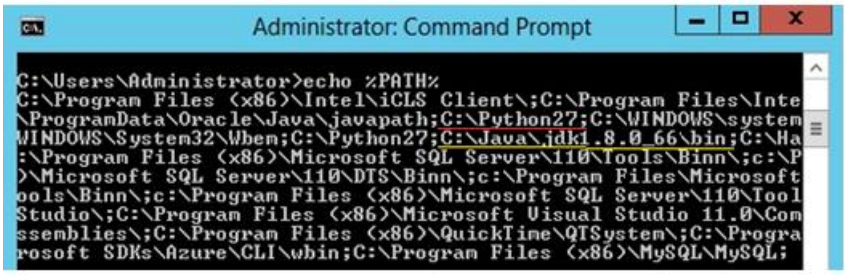
It is convenient for the future maintenance to store the location of the JDK folder to the environment variable in the operating system. Generally this variable's name is "JAVA_HOME". In the computer properties, this variable's value was adjusted to JDK's own folder in "C:\Program Files\Java". To verify the value of "JAVA_HOME", "echo %JAVA_HOME" command was entered in the prompt window (FIGURE 7).



```
C:\Users\Administrator>echo %JAVA_HOME%
C:\Program Files\Java\jdk1.8.0_66\

C:\Users\Administrator>
```

FIGURE 7. "JAVA_HOME" variable

To finish the installation of Java Development Kit, its "bin" folder's location was introduced in the "PATH" environment variable. The procedure of appending the "PATH"'s value is similar to the previous step. In order to check "PATH" value, "echo %PATH%" command was terminated in the prompt (FIGURE 8). Added to the "PATH" variable, JDK's "bin" directory is marked by a yellow line.



```
C:\Users\Administrator>echo %PATH%
C:\Program Files (x86)\Intel\iCLS Client\;C:\Program Files\Inte
\ProgramData\Oracle\Java\javapath;C:\Python27;C:\WINDOWS\system
WINDOWS\System32\Wbem;C:\Python27;C:\Java\jdk1.8.0_66\bin;C:\Ha
:\Program Files (x86)\Microsoft SQL Server\110\Tools\Binn\;c:\P
)\Microsoft SQL Server\110\DTS\Binn\;c:\Program Files\Microsoft
ools\Binn\;c:\Program Files (x86)\Microsoft SQL Server\110\Tool
Studio\;C:\Program Files (x86)\Microsoft Visual Studio 11.0\Com
ssemblies\;C:\Program Files (x86)\QuickTime\QTSystem\;C:\Progra
rosoft SDKs\Azure\CLI\wbin;C:\Program Files (x86)\MySQL\MySQL;
```

FIGURE 8. "PATH" variable

After the required condition was fulfilled, the free of charge NetBeans IDE was downloaded from the official company's website[13]. During the configuration, the default Java Development Kit was selected.

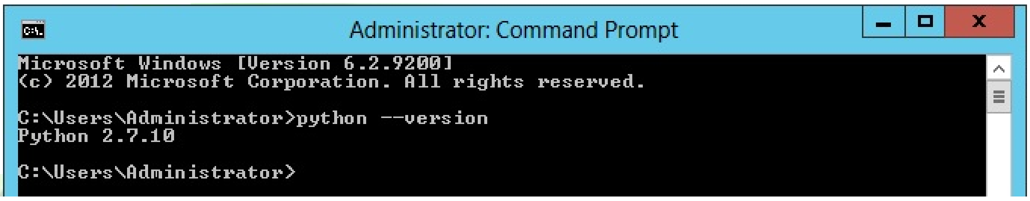
[12] Oracle official website - http://www.oracle.com/technetwork/java/javase/downloads

With regards to the GlassFish arrangement, an open-source software server is in free access at the provider's website[14]. However, there is a method of setting up a GlassFish server inside the Net-Beans environment directly. It is achieved by choosing "Add Server" function in "Servers" window where this application server is available for a free download and is configured straightforward, according to the user demand.

4.1.4 VirtualBox 5.0.16

Before starting to settle the VirtualBox software, these cases were resolved:

- BIOS option "Virtualization Extensions" was enabled in the settings on the host computer.
- Windows Server 2012 OS was deployed in the project. Commonly, Microsoft virtualization product "Hyper-V" is enabled in the system's features. Therefore, this virtualization technology was removed to envisage the VirtualBox with utter access to the hardware.
- The open-source Python software was settled on Windows. To check Python's version, "python -- version" command was implemented (FIGURE 9).



```
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>python --version
Python 2.7.10

C:\Users\Administrator>
```

FIGURE 9. The current Python version on the computer

If the program does not exist in the operating system, the launching package can be downloaded from the Python Software Foundation[15] website. The directory of the default application's folder has to be added to the "PATH" environment variable. Figure 8 shows that the "PATH" variable contains the location of Python version (marked with a red line).

The freeware VirtualBox package was downloaded from the official company's website, taking into account the specifications of the existing operating system. The setup process was completed in the several minutes. The coincident drivers were included with the software: USB support drivers, networking drivers and Python drivers.

4.1.5 Hortonworks Sandbox 2.3.0.0.

Hortonworks Sandbox has several prerequisites: (Installing Hortonworks Sandbox 2.0 – VirtualBox on Windows 2015, 1-11)

- Software:

[13] NetBeans IDE official website - https://netbeans.org/community/releases/81/index.html

[14] GlassFish website address -https://glassfish.java.net/download.html

[15] Python Software Foundation website address -https://www.python.org/downloads/windows/

  - Mounted VirtualBox 4.2 or later version.
  - Supported host operating system.
- Hardware:
  - 64-bit processor with multicore central processing unit
  - The lowest possible value for RAM is 4 GB.

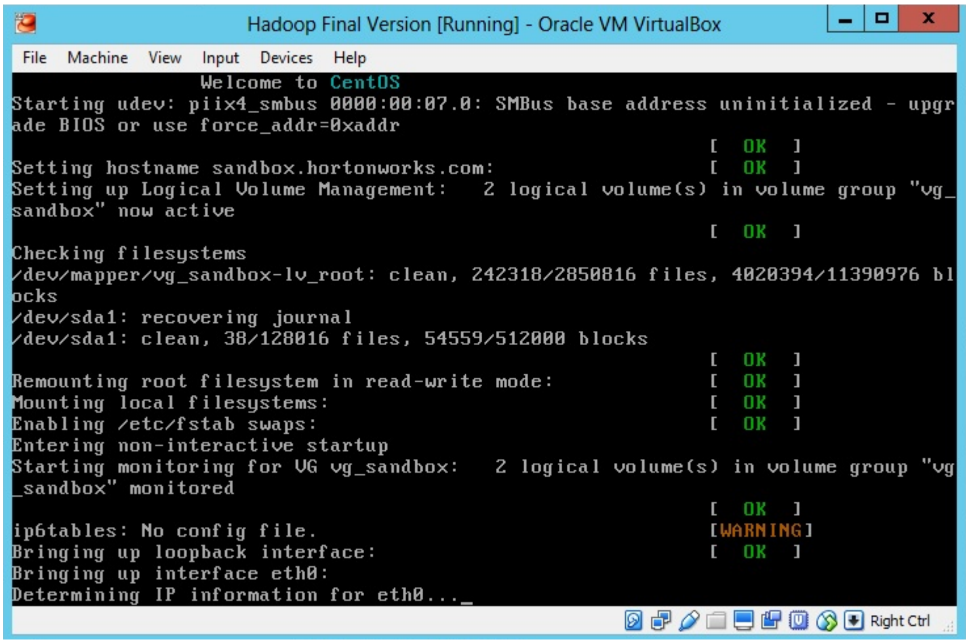
This virtual appliance was elaborated by American computer software enterprise, Hortonworks[16], and is freeware to utilize. After downloading the virtual machine from the website, it was imported to the VirtualBox program. Hortonworks Sandbox VM was configured according to the settings, specified in TABLE 2.

TABLE 2. The settings of Sandbox Hortonworks VM

Parameter	Value
General	
Name	Hadoop Final Version
Operating System	Red Hat (64-bit)
System	
Base Memory	5000 MB
Processors	2
Boot Order	Hard Disk, Optical VT-x/AMD-V, Nested Paging, PAE/NX
Display	
Video Memory	18 MB
Remote Desktop Server	Disabled
Video Capture	Disabled
Storage	
Controller	IDE Controller
IDE Primary Master	Hadoop Final Version-disk1.vmdk (Normal, 48,83 GB)
Audio	
Audio	Disabled
Network	
Adapter 1	PCnet-FAST III (NAT)
USB	
USB Controller	OHCI
Device Filters	0 (0 active)

After the Hortonworks Sandbox VM was imported and launched, the console window appeared and illustrated information about boot up (FIGURE 10).

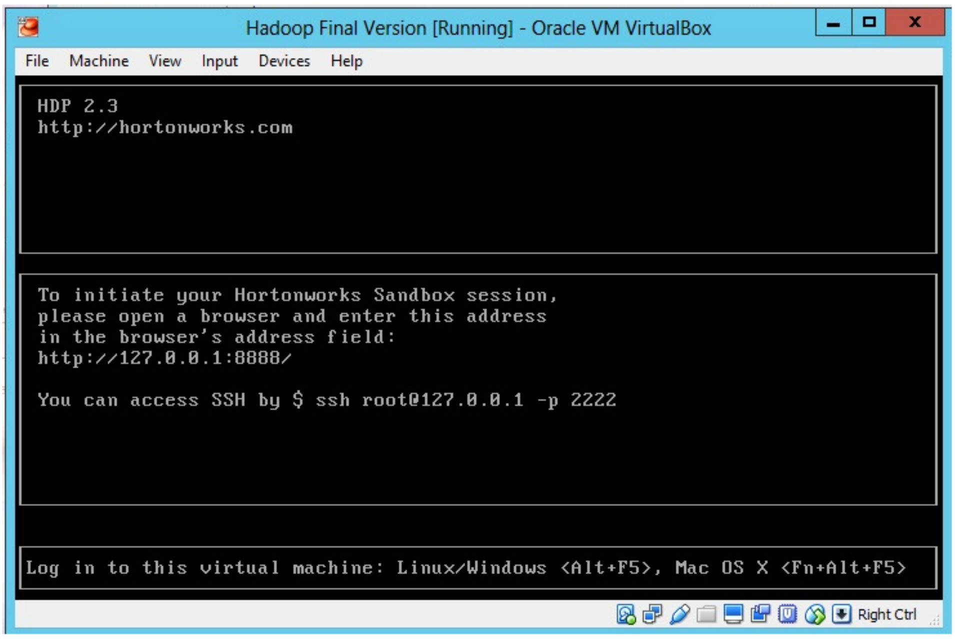
[16] Hortonwoks official page - http://hortonworks.com/


Hadoop Final Version [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Right Ctrl

```
                Welcome to CentOS
Starting udev: piix4_smbus 0000:00:07.0: SMBus base address uninitialized - upgr
ade BIOS or use force_addr=0xaddr
                                                           [  OK  ]
Setting hostname sandbox.hortonworks.com:                  [  OK  ]
Setting up Logical Volume Management:   2 logical volume(s) in volume group "vg_
sandbox" now active
                                                           [  OK  ]
Checking filesystems
/dev/mapper/vg_sandbox-lv_root: clean, 242318/2850816 files, 4020394/11390976 bl
ocks
/dev/sda1: recovering journal
/dev/sda1: clean, 38/128016 files, 54559/512000 blocks
                                                           [  OK  ]
Remounting root filesystem in read-write mode:             [  OK  ]
Mounting local filesystems:                                [  OK  ]
Enabling /etc/fstab swaps:                                 [  OK  ]
Entering non-interactive startup
Starting monitoring for VG vg_sandbox:   2 logical volume(s) in volume group "vg
_sandbox" monitored
                                                           [  OK  ]
ip6tables: No config file.                                 [WARNING]
Bringing up loopback interface:                            [  OK  ]
Bringing up interface eth0:
Determining IP information for eth0..._
```

FIGURE 10. The VM's boot information

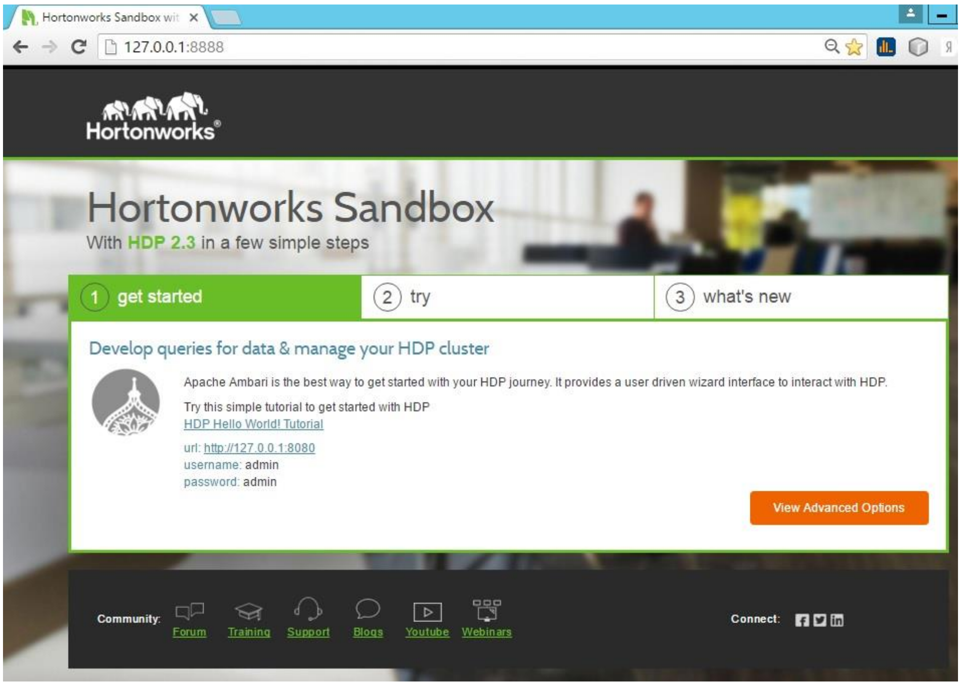
When the boot up was complemented, the console showed login information (FIGURE 11). Initially, the user name is "root" and password is "hadoop". However, the login parameters can be changed through the administrator's rights.


Hadoop Final Version [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Right Ctrl

```
HDP 2.3
http://hortonworks.com

To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://127.0.0.1:8888/

You can access SSH by $ ssh root@127.0.0.1 -p 2222

Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Fn+Alt+F5>
```

FIGURE 11. The login console to the virtual machine

During running time of the virtual machine, the default HDP web page is available at the link: http://127.0.0.1:8888 (FIGURE 12). It grants access to other Hadoop tools and services such as Secure Shell Client, Hue, Ranger, Atlas and Falcon.



Apache Ambari is the best way to get started with your HDP journey. It provides a user driven wizard interface to interact with HDP.

Try this simple tutorial to get started with HDP
HDP Hello World! Tutorial

url: http://127.0.0.1:8080
username: admin
password: admin

FIGURE 12. The entrance web page to the Hadoop platform

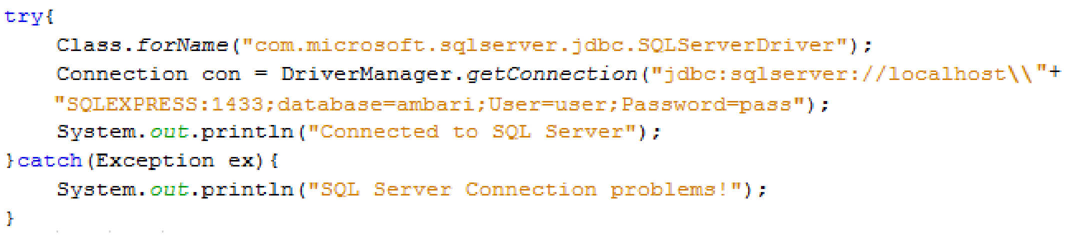
4.2 Establishing interactivity among elements in the system

4.2.1 Microsoft SQL Server Express 2012 – NetBeans IDE 8.1. Web Server Application

First and foremost, the SQL Server was configured for the external connections. The local TCP port of the SQL instance was set to 1433 in the SQL Server Configuration Manger for every IP address. Moreover, the remote connections were allowed to the SQL server through adjusting security settings.

With regards to the NetBeans web server, Microsoft SQL Server JDBC Driver was downloaded from the Internet and added manually to the programming environment, indicating the server's parameters for data management.

By attaining this sequence of actions, the connection between NetBeans and SQL Server was established. Thereby, the web server is capable of executing basic operations to the data, stored in the database engine. FIGURE 13 presents the Java-written code in a server application for mounting connection with the relational database management software.



```
try{
    Class.forName("com.microsoft.sqlserver.jdbc.SQLServerDriver");
    Connection con = DriverManager.getConnection("jdbc:sqlserver://localhost\\"+
    "SQLEXPRESS:1433;database=ambari;User=user;Password=pass");
    System.out.println("Connected to SQL Server");
}catch(Exception ex){
    System.out.println("SQL Server Connection problems!");
}
```

FIGURE 13. Java-written code in the server application for the implementation of the connection with the SQL Server

4.2.2 Microsoft SQL Server Express 2012 - Hortonworks Sandbox 2.3.0.0

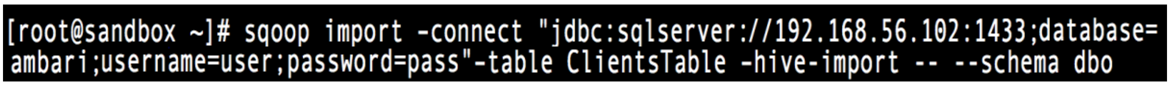
To provide interaction and data flow between Microsoft SQL Server and Hortonworks Sandbox VM, Hadoop offers Apache Sqoop technology. The functionality of this tool includes several operations: (Sqoop Data Transfer Tool 2015, 1-2)

1. Granting complete access to the information, stored in the Hadoop distributed file system.
2. Importing structured data from the external sources:
   - Relational data sources as SQL Server, Oracle.
   - Enterprise data storages and non-relational databases, e.g., Apache Cassandra, Voldemort, etc.

   Sqoop cooperates with a Hive data warehouse and populates tables from reading the imported files row-by-row. Moreover, a copy of the imported file is kept in HDFS.
3. Sqoop allows exporting the HDFS text files to the mentioned above data management services.

To use this Hadoop feature, the external data system driver is added to the Sqoop directory. Microsoft SQL Server was employed in the Big Data Analytics Tool; therefore, SQL Server JDBC Driver was installed to the Sqoop library.

Secure Shell is a HDP service for carrying out Sqoop tasks for data transmission. One of the primary commands, terminated in the project, was importing tables from the SQL Server's database to the Hive facility. (FIGURE 14).



```
[root@sandbox ~]# sqoop import -connect "jdbc:sqlserver://192.168.56.102:1433;database=
ambari;username=user;password=pass"-table ClientsTable -hive-import -- --schema dbo
```

FIGURE 14. Sqoop import command

4.2.3 NetBeans 8.1.Web Server Application – Hortonworks Sandbox 2.3.0.0.

The server application is the central component of Big Data Analytics Tool. One of its key roles is establishing intercommunication with Hortonworks Sandbox environment. The server had to be arranged in a way that it had direct access to the next HDP's programs: Secure Shell, Hive data warehouse and Sqoop.

At the outset, the networking settings of the virtual machine were updated via appending a couple of port forwarding rules to allow the remote connections from hosts (TABLE 3).

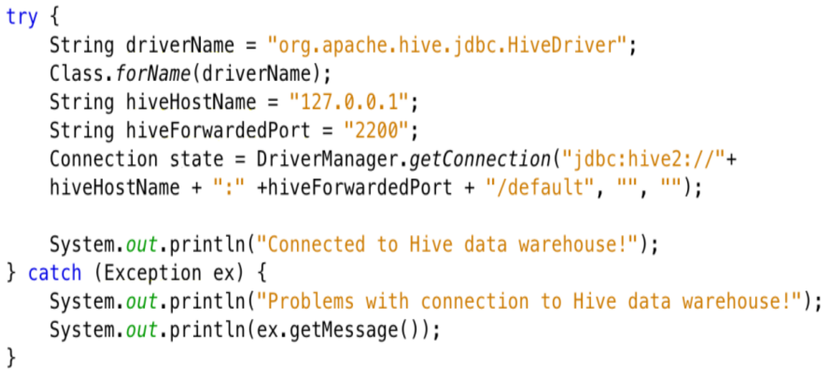
TABLE 3. Complementary port forwarding rules

Name	Protocol	Host IP	Host Port	Guest Port
Hive	TCP	127.0.0.1	2200	10000
SSH	TCP	127.0.0.1	2222	22

Hereafter, Hortonworks Hive ODBC Driver was installed on the computer and configured in accordance with the port forwarding rules of the virtual machine. The successful employment of Hive software library depends on pursuance of a pair Hadoop services:

- *Hive Server* - an auxiliary service that provides access to Hive ODBC driver to execute remotely essential database cooperation and Hive commands.
- *Apache Thrift* - a software framework that establishes interrelations with a Hive Server.

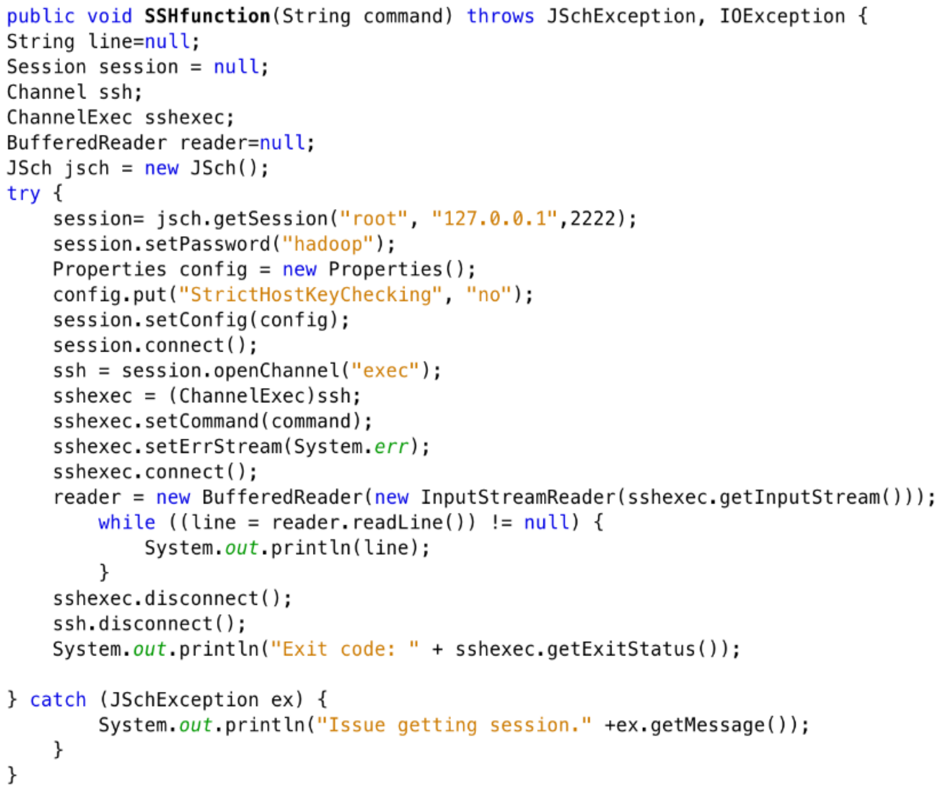
The correspondent Hive and SSH drivers were imported to the server project. The code snippet of NetBeans web server, carrying out the connection with Hive data warehouse, is presented in FIGURE 15. The intercommunication with Hive is maintained through loading Hive driver and observing the attributes of port forwarding rule: host IP and port.



```
try {
    String driverName = "org.apache.hive.jdbc.HiveDriver";
    Class.forName(driverName);
    String hiveHostName = "127.0.0.1";
    String hiveForwardedPort = "2200";
    Connection state = DriverManager.getConnection("jdbc:hive2://"+
    hiveHostName + ":" +hiveForwardedPort + "/default", "", "");

    System.out.println("Connected to Hive data warehouse!");
} catch (Exception ex) {
    System.out.println("Problems with connection to Hive data warehouse!");
    System.out.println(ex.getMessage());
}
```

FIGURE 15. The server code for establishing connection with Hive in Hadoop environment

Moreover, a secure channel was mounted out across the network between the server application and virtual machine to import and update data in Hive and control HDFS. Therefore, the server can approach Hadoop’s secure shell and forward commands for a virtual appliance to devise. FIGURE 16 shows server’s function for executing any command in Hortonworks Sandbox.

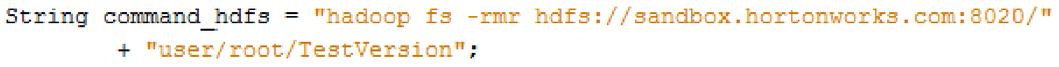


```
public void SSHfunction(String command) throws JSchException, IOException {
String line=null;
Session session = null;
Channel ssh;
ChannelExec sshexec;
BufferedReader reader=null;
JSch jsch = new JSch();
try {
    session= jsch.getSession("root", "127.0.0.1",2222);
    session.setPassword("hadoop");
    Properties config = new Properties();
    config.put("StrictHostKeyChecking", "no");
    session.setConfig(config);
    session.connect();
    ssh = session.openChannel("exec");
    sshexec = (ChannelExec)ssh;
    sshexec.setCommand(command);
    sshexec.setErrStream(System.err);
    sshexec.connect();
    reader = new BufferedReader(new InputStreamReader(sshexec.getInputStream()));
        while ((line = reader.readLine()) != null) {
            System.out.println(line);
        }
    sshexec.disconnect();
    ssh.disconnect();
    System.out.println("Exit code: " + sshexec.getExitStatus());

} catch (JSchException ex) {
        System.out.println("Issue getting session." +ex.getMessage());
    }
}
```

FIGURE 16. The Secure Shell function for mounting the interactivity between the server and the VM’s console

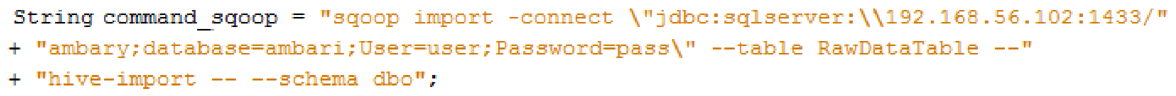
This function uses standard settings of SSH port forwarding rule: Hadoop's login username, password, host IP and port. By utilizing this code, the server accesses and brings changes in the data at Hive databases through Sqoop commands, updates Hadoop's distributed file system, etc.

For instance, the text file "TestVersion" is deleted from HDFS by applying "command_hdfs" string as a parameter to SSH function:



```
String command_hdfs = "hadoop fs -rmr hdfs://sandbox.hortonworks.com:8020/"
        + "user/root/TestVersion";
```

Whilst, Sqoop command imports a new table "RawDataTable" to Hive by terminating SSH function with the "command_sqoop" parameter:



```
 String command_sqoop = "sqoop import -connect \"jdbc:sqlserver:\\192.168.56.102:1433/"
+ "ambary;database=ambari;User=user;Password=pass\" --table RawDataTable --"
+ "hive-import -- --schema dbo";
```

5 IMPLEMENTATION OF BIG DATA ANALYTICS TOOL

5.1 Client-Server Network

Client-server architecture represents itself a communication network where there is a central element, also referred to as "server", which provides maintenance and carries out the requests of end-users (FIGURE 17). (Beal 2015, 1) The client application can be performed on a laptop, desktop computer, smartphone or tablet. It brings up the connection to the server to intercommunicate in a way of requests and responses. Server's duties include verification of end-user's status, renovation and management of the security systems, and delivery of the resources to clients. With the aim of diminishing network traffic, the server shuts down the connection after the service goal was accomplished. (Rouse 2015, 1)

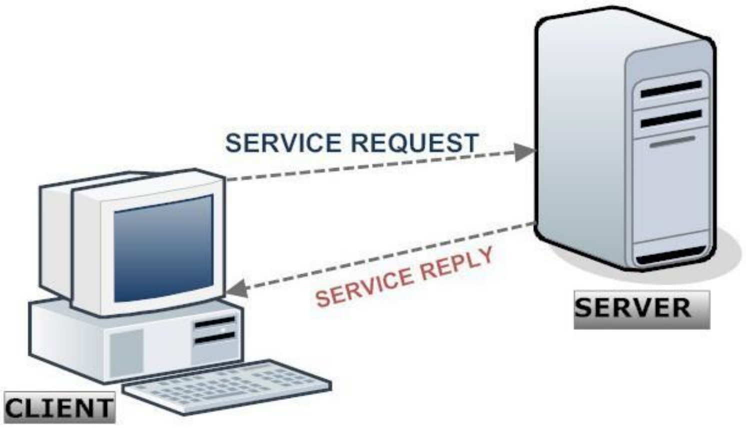


FIGURE 16. The Client-Server model [What is a Client-Server Network 2013, 1]

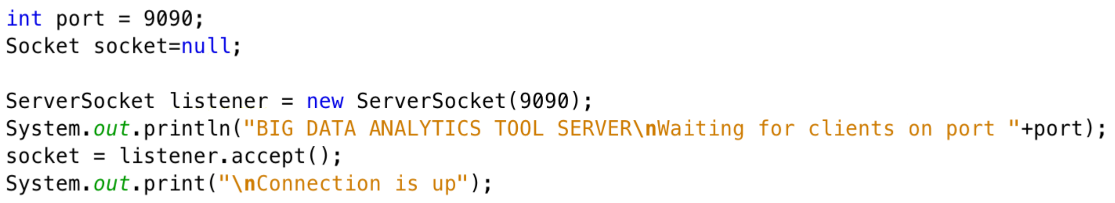
The client and server applications can be terminated on a stand-alone computer. However, the major attribute of this model is networking. The connection between client and server is executed over two telecommunication network types:

- Local Area Network (LAN) – *a communications network that interconnects a variety of data communications devices within a small geographic area and broadcasts data at high data transfer rates.* (What is a Client-Server Network et al. 2013, 1)
- Wide Area Network (WAN) – a long-distance telecommunication network, capable of data transfer, covering extensive geographic territories by concatenating numerous LANs. The supreme example of wide area network is the Internet. (Mitchell 2014, 1)

The majority of enterprise facilities utilize client-server architecture, in particular, TCP/IP protocol for the Internet connection. This communication protocol is devised in a way that every device in the network is indicated by the distinctive IP address. Moreover, each network appliance can establish a connection to another computer at over 65535 different ports. (Abrams 2004, 1) The port exemplifies a peculiar communication line between two end-points and its number is the unique code for the authentication of conjunction.

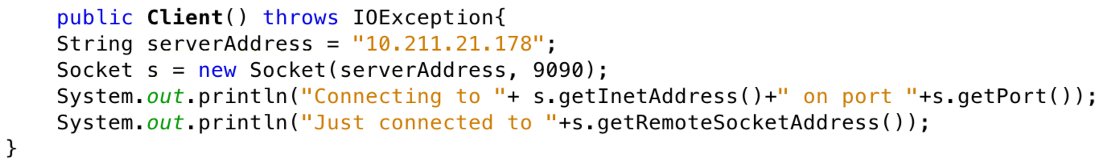
The TCP/IP intercommunication was established in the Big Data Analytics Tool between client and server, taking into consideration several features (FIGURE 18, FIGURE 19):

- The end-user application knows a couple of server's networking parameters: IP address and port number.
- The server is always in the "listening" mode, accepting and rejecting the incoming clients' requests for connection.
- When the client-server intercourse is established through the specified virtual TCP/IP port, the data is transferred in both routes via the network. As long as both of them communicate and do not shut down the broadcasting, the connection continues to be in the "open" state.



```
int port = 9090;
Socket socket=null;

ServerSocket listener = new ServerSocket(9090);
System.out.println("BIG DATA ANALYTICS TOOL SERVER\nWaiting for clients on port "+port);
socket = listener.accept();
System.out.print("\nConnection is up");
```

FIGURE 18. The server-side code snippet for the cooperation with the clients' requests for connection



```
    public Client() throws IOException{
    String serverAddress = "10.211.21.178";
    Socket s = new Socket(serverAddress, 9090);
    System.out.println("Connecting to "+ s.getInetAddress()+" on port "+s.getPort());
    System.out.println("Just connected to "+s.getRemoteSocketAddress());
}
```

FIGURE 19. The client-side code snippet for mounting the intercommunication with the server

One of the primary TCP/IP protocol's advantage is a verification of the forwarded data on errors, thereby ensuring the high quality of data transmission.

5.2 Login and Registration system

To enhance the security of Big Data Analytics Tool, login and registration are a general procedure for entering a client web application. FIGURE 20 displays login web page of Big Data Analytics Tool for Hydroline Oy. Generally, a client is required to fill in standard account information such as username and password in the login form to proceed on operating with the website. In case the client does not have an account in the system, the registration option is available. The registration can be accomplished by completing the accordant form. The SQL Server sustains the data about all of the users of Big Data Analytics Tool.

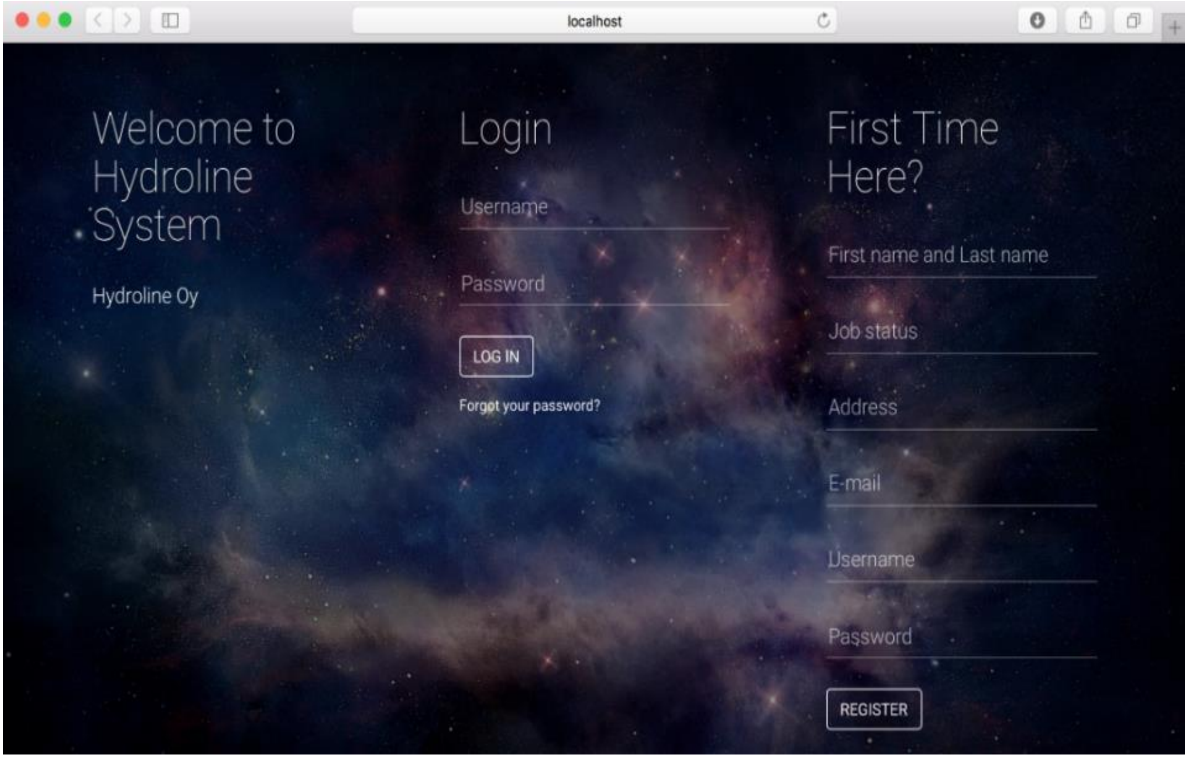


FIGURE 20. The login web page for the Big Data Analytics Tool

According to the elaboration of the entrance web page, programming in the markup language, HTML5, was executed. In addition, the designing and management of the login page were performed through the CSS programming. For instance, several parameters were set up such as positions and dimensions of the elements, selection of the background, fonts' styles and sizes, etc. The JavaScript coding applied special effects to the web page. Flickering background, switching the text color of the selected option and appearing next to it accordant image icon are a few examples of contriving JavaScript code.

The functionality of this web page is maintained by two core HTML elements in the client application:

- *Login form* - a procedure, intended for a user to identify and authenticate own account. To log into the system, the following sequence of actions is committed:
  1. First of all, a user inserts username and password values into the respective input fields and submits the form, calling to the corresponding Java servlet in the client project.
  2. The servlet reads the attributes and forwards the information to the server for the verification of account's existence. Whereas server checks the received client information in the SQL Server's database that contains data about all registered users.
  3. Thereafter, the server sends the response back to the client. If it is positive, the client is allowed to enter the Big Data Analytics Tool's website. In case the server's reply is negative, the error dialogue window will appear, indicating the reasons wherefore login procedure was not terminated successfully.
- *Registration form* - an operation of creating new user account in the Big Data Analytics Tool. During the registration of a novel profile, this chain of events is derived:
  1. To begin with, a user provides personal contact information in the form's input fields: first name and last name, job status, address, e-mail, username and password. After clicking the submit button, the Registration Java servlet is triggered.
  2. This servlet retrieves and transmits to the server information, entered by the client.
  3. Onwards, the server ensures that there is no similar account in the SQL database and adds the client profile to the system.
  4. Finally, the server application transmits the response to the user. If the reply is affirmative, the client is forwarded to the next web page. Otherwise, an alert message box pops up, stating the reason for which the client cannot be registered in the frame.

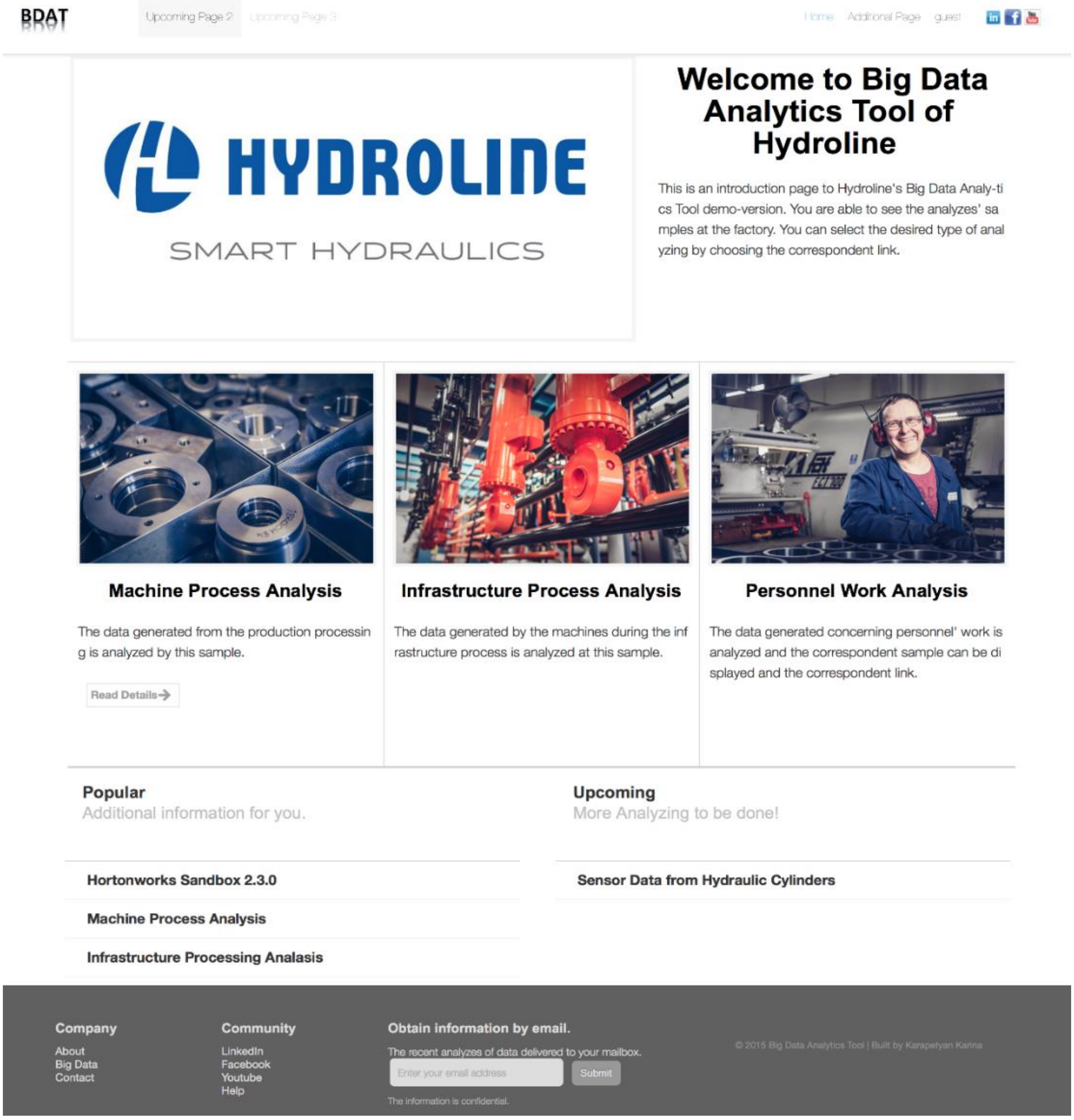
Needless to say, the server's duties include advanced security of Big Data Analytics Tool through user management software. The server tracks users, logged into the system in real-time mode, and has full rights to edit client profiles, e.g., changing the account's parameters, and even removing users from the system.

5.3 Introduction Page

FIGURE 21 displays an introduction web page, which is a fundamental part of the client application in Big Data Analytics Tool. The primary role of the home page is the ability for a user to navigate through the website. Furthermore, the client can get familiarized with the project itself, Case Company and provided services: an analysis of the machine processing and industrial information via linking to the corresponding web pages. In addition, the start page of Big Data Analytics Tools displays the authorized username that is passed from the Login and Registration page.

Concerning the technical point of view, the home web page was carried out through HTML5 coding. The purpose of HTML utilization was a creation of the web page structure, describing its content, and navigation across the client application. To advance the design of the web page, coding in other languages was devised in cooperation with HTML:

- CSS is responsible for webpage's appearance - In order to maintain the layout of the web page, for instance, rendering HTML components on the display, and introduce its styling such as backgrounds, text fonts and colors, CSS coding was integrated.
- Whereas JavaScript proposes the behavior to the web page – To enhance client experience and create interaction with the user, JavaScript offers dynamic functionality to the homepage. For example, JavaScript-written scroll function was utilized to avoid the overflow of the web page.



Welcome to Big Data Analytics Tool of Hydroline

This is an introduction page to Hydroline's Big Data Analy-tics Tool demo-version. You are able to see the analyzes' samples at the factory. You can select the desired type of analyzing by choosing the correspondent link.

Machine Process Analysis

The data generated from the production processing is analyzed by this sample.

Infrastructure Process Analysis

The data generated by the machines during the infrastructure process is analyzed at this sample.

Personnel Work Analysis

The data generated concerning personnel' work is analyzed and the correspondent sample can be displayed and the correspondent link.

Popular
Additional information for you.

Hortonworks Sandbox 2.3.0
Machine Process Analysis
Infrastructure Processing Analasis

Upcoming
More Analyzing to be done!

Sensor Data from Hydraulic Cylinders

Company
About
Big Data
Contact

Community
LinkedIn
Facebook
Youtube
Help

Obtain information by email.
The recent analyzes of data delivered to your mailbox.
The information is confidential.

© 2015 Big Data Analytics Tool | Built by Karapetyan Karina

FIGURE 21.The introduction web page

5.4 Infrastructure Process Analysis

Infrastructure Process Analysis is a principal analysis method of Big Data Analytics Tool, devised on the establishment of an Overall Equipment Effectiveness Model for Hydroline Oy. As a new spin of autonomous manufacturing innovation, OEE was introduced in 1960s to determine capacity of machinery and suppliers. To obtain the coefficient of production success, the result of Overall Equipment Effectiveness methodology is utilized as a key metrics in combination with the lean manufacturing exertion. (Hansen 2001, 28-33)

Calculating OEE is a foremost technique of fabrication process. This method enables identification of reasons, causing the industrial losses, and offers significant information of how to systematically advance manufacturing. Overall Equipment Effectiveness is an exclusive method for revealing manufacturing defect rate, benchmarking propulsion, and enhancing the performance of the machinery, e.g., management of waste disposal.

In the book "The OEE Primer Understanding Overall Equipment Effectiveness, Reliability, and Maintainability", Stamatis defined the computation of OEE, based on three attributes:

1. ***Availability*** – a time metrics of OEE model that is calculated as a ratio of operative time to planned production time, measured in percentage. Availability takes into consideration time loss, including interruptions for a respective period of time during the manufacturing. In practice, this OEE parameter is determined as:

$$Availability = \frac{Operating\ Time}{Planned\ Production\ Time} \qquad [1]$$

The operating time represents itself the scheduled manufacturing period, excluding the time when the production flow was eliminated due to the target downtime (e.g., meal break, short, short breaks for employees) or unscheduled lay-up time (i.e., equipment break, preventative maintenance, cleanup, setup, material change, material handling). The formula for estimation of the operating time is introduced by Equation 2.

$$Operating\ Time = Planned\ Production\ Time - Downtime \qquad [2]$$

2. ***Performance*** – a fabrication rate metrics of OEE that is computed as a proportion of total amount of details to the product of operating time and ideal run rate, measured in percentage. Pursuance correlates all factors, associated with the decrease in possible production pace of the equipment. The OEE production tempo is represented by Equation 3.

$$Performance = \frac{Total\ Amount\ of\ Details}{Operating\ Time \times Ideal\ Run\ Rate} \qquad [3]$$

, where Ideal Run Rate is a theoretical minimum amount of time, necessary for manufacturing one unit, and Operating Time is a theoretical minimum quantity of time, needed for the fabrication of all details.

3. ***Quality*** – an OEE parameter for process yield, estimated as a quotient of the amount of the produced sufficient details to the total quantity of manufactured units, whereas the measuring unit is a percentage. This OEE attribute is responsible for the detection of the quality loss such as an interest of units that do not comply with the standard of quality. Equation 4 is utilized to determine quality index:

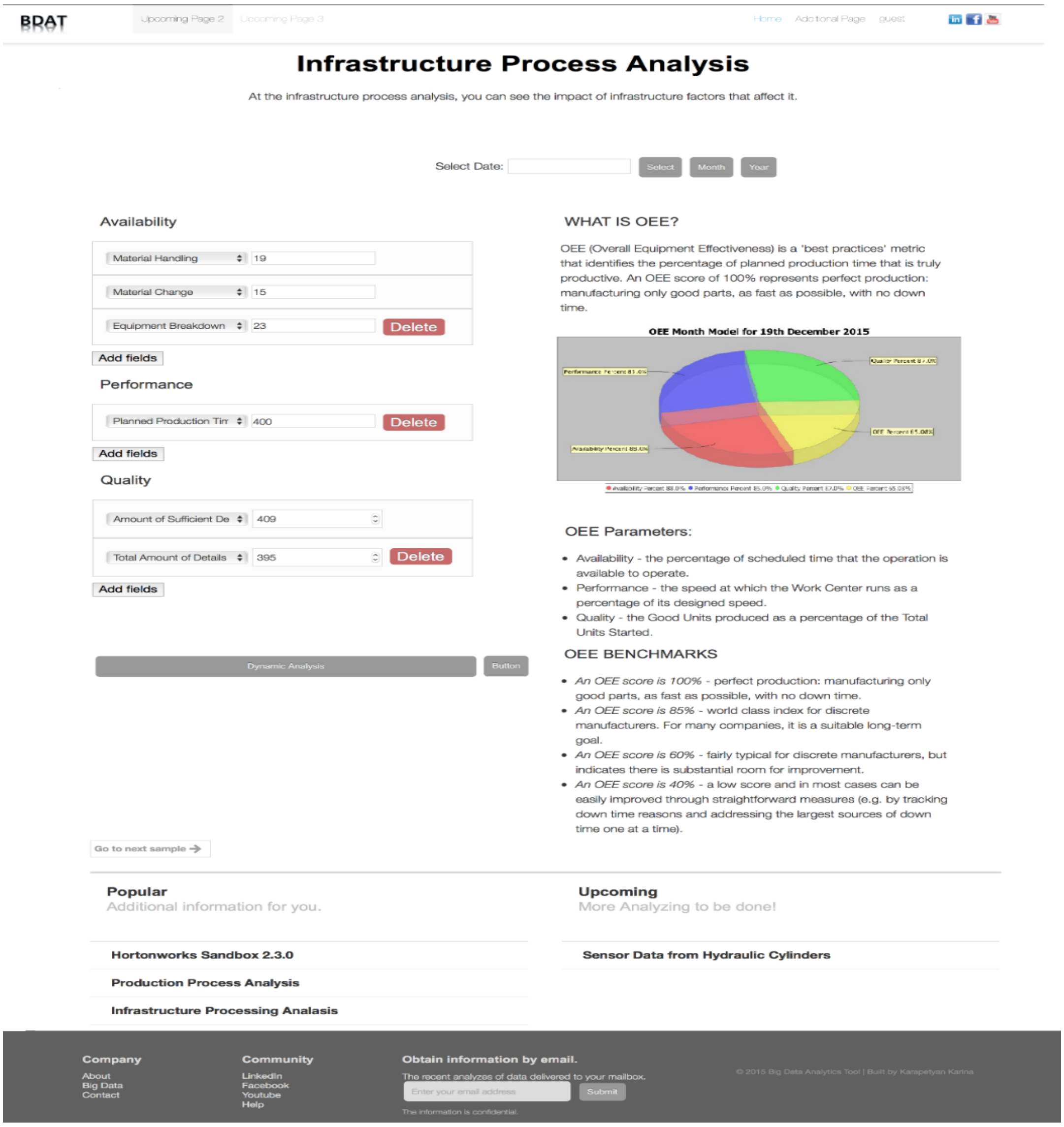
$$Quality = \frac{Amount\ of\ Sufficient\ Details}{Total\ Amount\ of\ Details} \qquad [4]$$

Overall Equipment Effectiveness model contemplates all types of deprivations to identify the actual efficiency index of the machinery production. The OEE calculation is performed by Equation 5 where OEE is a product of availability, performance and quality coefficients:

$$OEE = Availability \times Performance \times Quality \qquad [5]$$

When the cumulative value of all OEE attributes is equal to 100%, the industrial process is classified as “excellent”, designating the maximum fabrication rate, high quality of production and absence of downtime.

Daily, monthly and annual OEE model was carried out for the Case Company in the Big Data Analytics Tool, since Hydroline Oy is an original equipment manufacturer. This model creates an essential comprehension and precise view on how the manufacture is presented within the company. Moreover, OEE emphasizes the areas where it is indispensable to enhance operability and productivity. FIGURE 22 presents the Big Data Analytics Tool web page for OEE analysis.



Infrastructure Process Analysis

At the infrastructure process analysis, you can see the impact of infrastructure factors that affect it.

Select Date:

Availability

Material Handling: 19

Material Change: 15

Equipment Breakdown: 23

Performance

Planned Production Tim: 400

Quality

Amount of Sufficient De: 409

Total Amount of Details: 395

WHAT IS OEE?

OEE (Overall Equipment Effectiveness) is a 'best practices' metric that identifies the percentage of planned production time that is truly productive. An OEE score of 100% represents perfect production: manufacturing only good parts, as fast as possible, with no down time.

OEE Month Model for 19th December 2015

OEE Parameters:

- Availability - the percentage of scheduled time that the operation is available to operate.
- Performance - the speed at which the Work Center runs as a percentage of its designed speed.
- Quality - the Good Units produced as a percentage of the Total Units Started.

OEE BENCHMARKS

- *An OEE score is 100%* - perfect production: manufacturing only good parts, as fast as possible, with no down time.
- *An OEE score is 85%* - world class index for discrete manufacturers. For many companies, it is a suitable long-term goal.
- *An OEE score is 60%* - fairly typical for discrete manufacturers, but indicates there is substantial room for improvement.
- *An OEE score is 40%* - a low score and in most cases can be easily improved through straightforward measures (e.g. by tracking down time reasons and addressing the largest sources of down time one at a time).

Popular
Additional information for you.

Upcoming
More Analyzing to be done!

Hortonworks Sandbox 2.3.0

Production Process Analysis

Infrastructure Processing Analasis

Sensor Data from Hydraulic Cylinders

Company
About
Big Data
Contact

Community
LinkedIn
Facebook
Youtube
Help

Obtain information by email.
The recent analyzes of data delivered to your mailbox.
The information is confidential.

© 2015 Big Data Analytics Tool | Built by Karapetyan Karina

FIGURE 22. The dynamic OEE analysis web page (Provided analyzed data chart is artificial due to non-disclosure agreement with the Case Company)

According to the technical aspect of the implementation of Overall Equipment Effectiveness model for Hydroline Oy, the corresponding HTML page was composed, containing next elements:

1. *Header* – an introduction constituent of the website that includes horizontal menu list for the navigation in the client application as well as the user name of the authorized client, indicated in the specified field. The username is directed from the home web page through Login and Registration Java servlet.
2. *HTML Calendar form* – one of the principal webpage's elements, offering a client to select a date for which to carry out the OEE analyzing. Regarding the functionality of this web form, it is connected to the respective Java servlet with the "POST" query type for reading the date and further transmission of the information to the server. The calendar form has three submit buttons that trigger the Java servlet:
   - *Select button* – reads the entire date for conduction of a daily OEE analysis for the corresponding day.
   - *Month button* – takes into account only the month of the chosen date for monthly OEE analyzing.
   - *Year button* – utilizes only the year of the selected date for an annual OEE analysis.

   When a user clicks one of the buttons, the Java servlet is launched and transmits the respective date parameters to the server, thereunto indicating the type of Overall Equipment Effectiveness analysis: daily, monthly and annual. The design of a calendar is performed with CSS programming to add distinctive features for interactivity with the user, e.g., background adjustments, alteration of the selected context color.
3. *OEE picture* – a graphical representation of an analysis. After the server receives the data, sent by the client application, it communicates with the Hadoop platform by Hive queries via secure shell channel and makes a request for data analysis in obedience to a selected period of time. Hortonworks Sandbox VM stores the information, generated during the production. The original manufacturing data is processed through Hive queries according to the OEE logic. To calculate the OEE indexes, the following industrial parameters are taken into consideration for an analysis:
   - *Availability:* shift length, short break, meal break, setup, equipment breakdown, clean-up, material handling, material change, preventative maintenance and personnel relief*.*
   - *Performance* – ideal run rate, planned production time and amount of sufficient details.
   - *Quality* – total quantity of details and amount of sufficient units.

Further, the HDP returns the processed data: OEE factors and catchall OEE coefficient, indicated in the percentage. Based on the gained data, the server builds OEE 3D pie diagram and transmits it back to the client via a binary array. The example of an OEE pie diagram is illustrated in FIGURE 23.

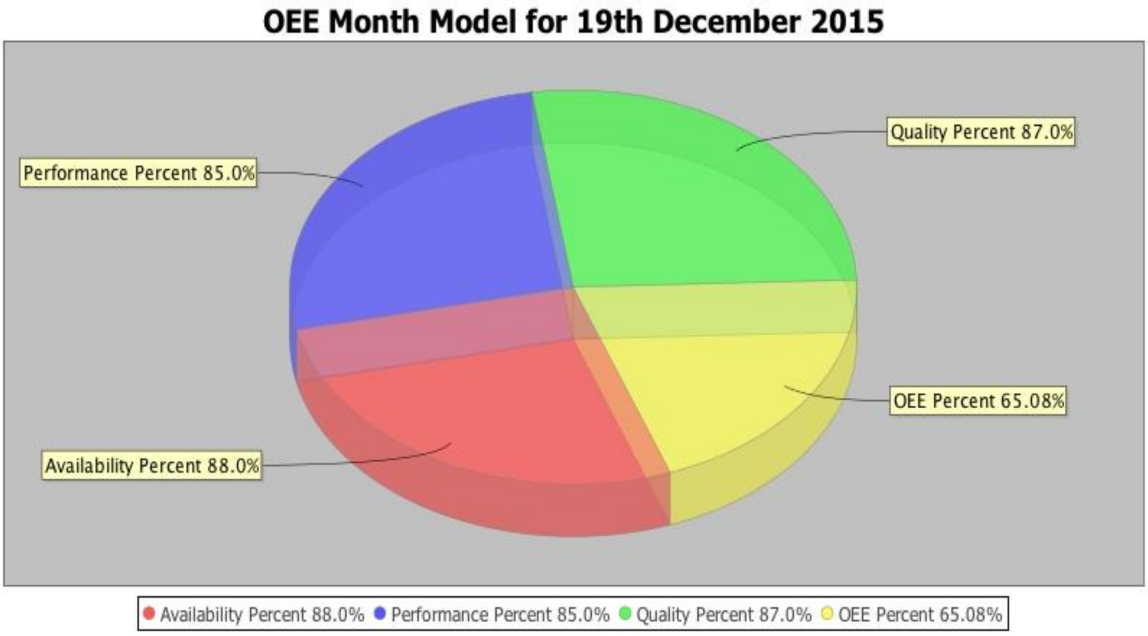
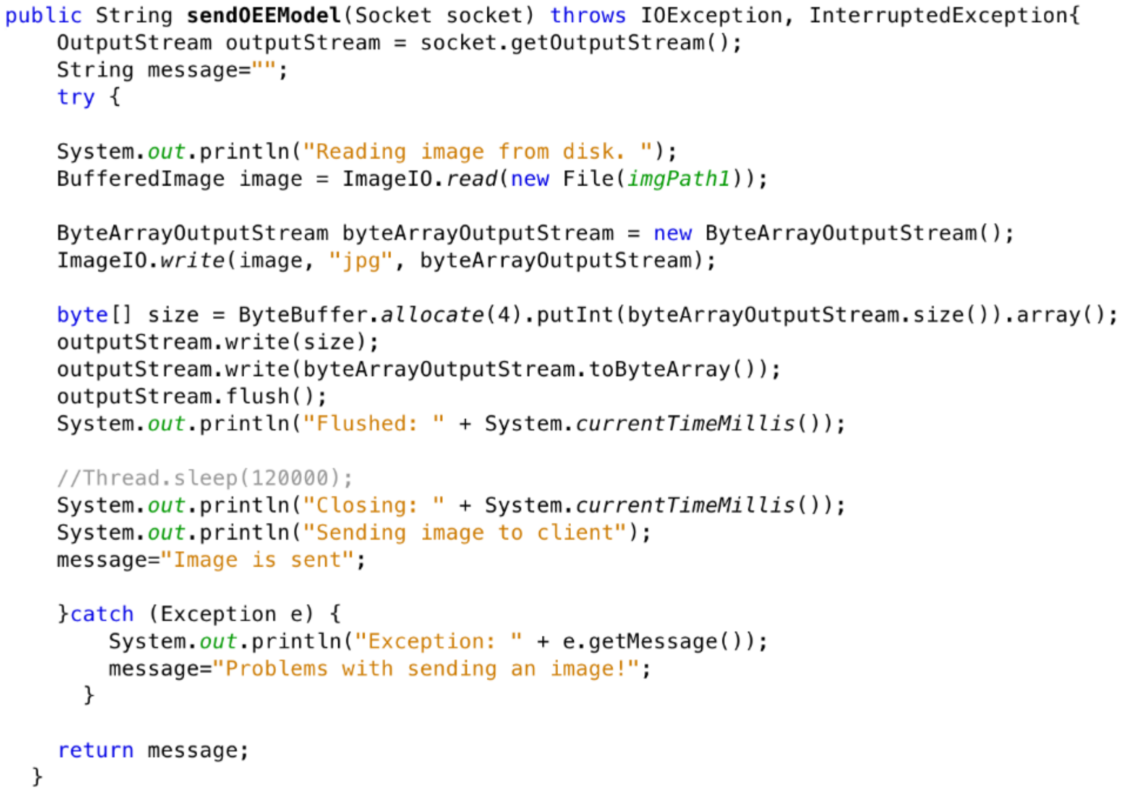


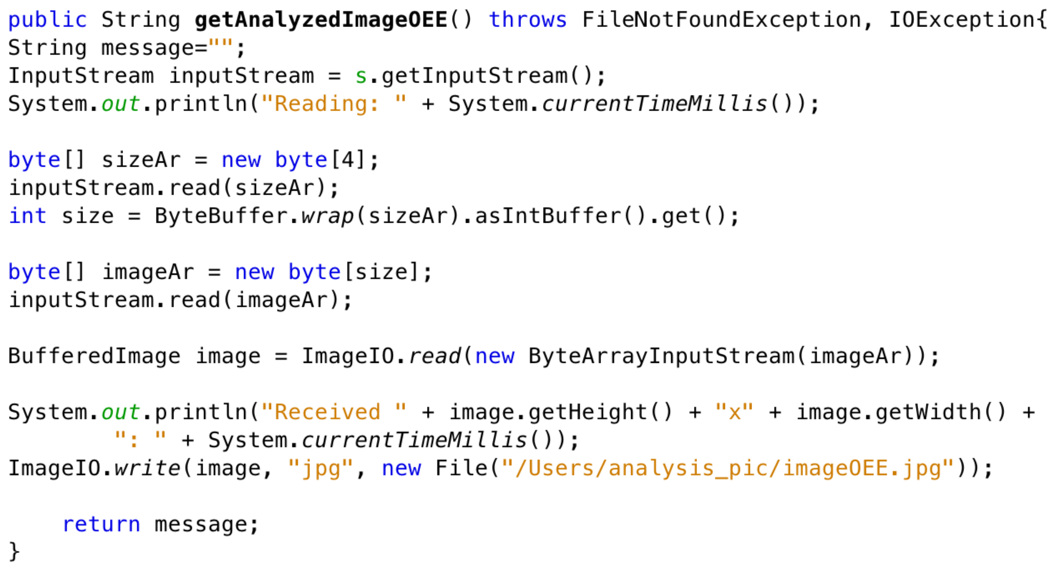
FIGURE 23. The illustration of a pie diagram, obtained from a monthly OEE analysis (Provided analyzed data chart is artificial due to non-disclosure agreement with the Case Company)

The client application converts the acquired array into an image with the JPEG standard. FIGURE 24 and FIGURE 25 display relatively code snippets for transmitting and receiving a binary image, representing scrutinized data chart. By dint of a JavaScript function, an OEE 3D pie chart is broadcasted to the web page without reloading it.



```
public String sendOEEModel(Socket socket) throws IOException, InterruptedException{
    OutputStream outputStream = socket.getOutputStream();
    String message="";
    try {

    System.out.println("Reading image from disk. ");
    BufferedImage image = ImageIO.read(new File(imgPath1));

    ByteArrayOutputStream byteArrayOutputStream = new ByteArrayOutputStream();
    ImageIO.write(image, "jpg", byteArrayOutputStream);

    byte[] size = ByteBuffer.allocate(4).putInt(byteArrayOutputStream.size()).array();
    outputStream.write(size);
    outputStream.write(byteArrayOutputStream.toByteArray());
    outputStream.flush();
    System.out.println("Flushed: " + System.currentTimeMillis());

    //Thread.sleep(120000);
    System.out.println("Closing: " + System.currentTimeMillis());
    System.out.println("Sending image to client");
    message="Image is sent";

    }catch (Exception e) {
        System.out.println("Exception: " + e.getMessage());
        message="Problems with sending an image!";
      }

    return message;
  }
```

FIGURE 24. The code snippet for sending a binary array from the server to a client



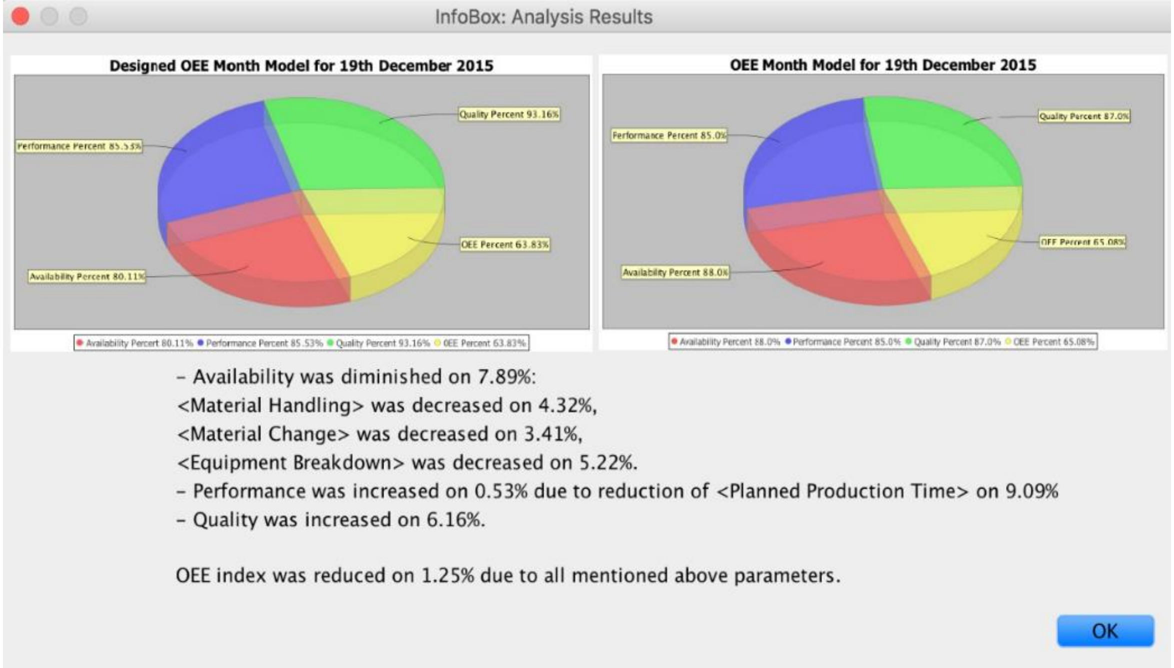
```
public String getAnalyzedImageOEE() throws FileNotFoundException, IOException{
String message="";
InputStream inputStream = s.getInputStream();
System.out.println("Reading: " + System.currentTimeMillis());

byte[] sizeAr = new byte[4];
inputStream.read(sizeAr);
int size = ByteBuffer.wrap(sizeAr).asIntBuffer().get();

byte[] imageAr = new byte[size];
inputStream.read(imageAr);

BufferedImage image = ImageIO.read(new ByteArrayInputStream(imageAr));

System.out.println("Received " + image.getHeight() + "x" + image.getWidth() +
        ": " + System.currentTimeMillis());
ImageIO.write(image, "jpg", new File("/Users/analysis_pic/imageOEE.jpg"));

    return message;
}
```

FIGURE 17. The code snippet for receiving a binary array from a server at the client application

4. *Dynamic HTML form for alteration of OEE properties* – a distinguishing function of Big Data Analytics Tool to simulate the Overall Equipment Effectiveness model, owing to the variation of parameters within OEE factors. JavaScript and jQuery codes were utilized for the original OEE analysis data to be modified by dynamically appending or removing form's input fields. A client chooses what parameters within availability, performance and quality should be altered and determines their values. For instance, a user can set the value of the ideal run rate to be six hundred details per day and the planned production time to be 4 hours in OEE Performance factor and define half of an hour meal break in OEE Availability attribute.

   The Java servlet for dynamic analysis is incorporated into the web form and launched after the "Dynamic Analysis" button is pressed. The servlet reads the input fields of the submitted form and passes this data as well as the information about original OEE results to the server. The server application makes a request to Hadoop environment by implementing Hive queries through SSH to gain the data, concerning the initial record. Taking into account client's adjusted data and, also, the original parameters, the calculation of OEE simulation model is carried out and presented in the graphical form such as 3D pie diagram. The chart is converted into a binary array and transmitted to the user. In addition, the attributes of the original OEE model and the simulated OEE sample are compared and the result is forwarded therewith to the client.

   At the same time, Java servlet pops up a dialog window with a table where the first row includes the received binary array, rendered to the JPEG image, and initial 3D pie diagram, displaying the original OEE model. Moreover, the second row provides the outcome of comparing the two samples. Therefore, the client can see the causes of low OEE indexes and create the plan of action of how to improve the current fabrication. The dialogue window introduces the results of established, according to a customer demand, OEE model and existing OEE sample. (FIGURE 26)


InfoBox: Analysis Results
Designed OEE Month Model for 19th December 2015
Quality Percent 93.16%
Performance Percent 85.53%
OEE Percent 63.83%
Availability Percent 80.11%
● Availability Percent 80.11% ● Performance Percent 85.53% ● Quality Percent 93.16% ● OEE Percent 63.83%
OEE Month Model for 19th December 2015
Quality Percent 87.0%
Performance Percent 85.0%
OEE Percent 65.08%
Availability Percent 88.0%
● Availability Percent 88.0% ● Performance Percent 85.0% ● Quality Percent 87.0% ● OEE Percent 65.08%
OK

- Availability was diminished on 7.89%:
<Material Handling> was decreased on 4.32%,
<Material Change> was decreased on 3.41%,
<Equipment Breakdown> was decreased on 5.22%.
- Performance was increased on 0.53% due to reduction of <Planned Production Time> on 9.09%
- Quality was increased on 6.16%.

OEE index was reduced on 1.25% due to all mentioned above parameters.

FIGURE 26. The dynamic dialogue window with the designed OEE model and current OEE model for 19th of December 2015 (Provided analyzed data chart is artificial due to non-disclosure agreement with the Case Company)

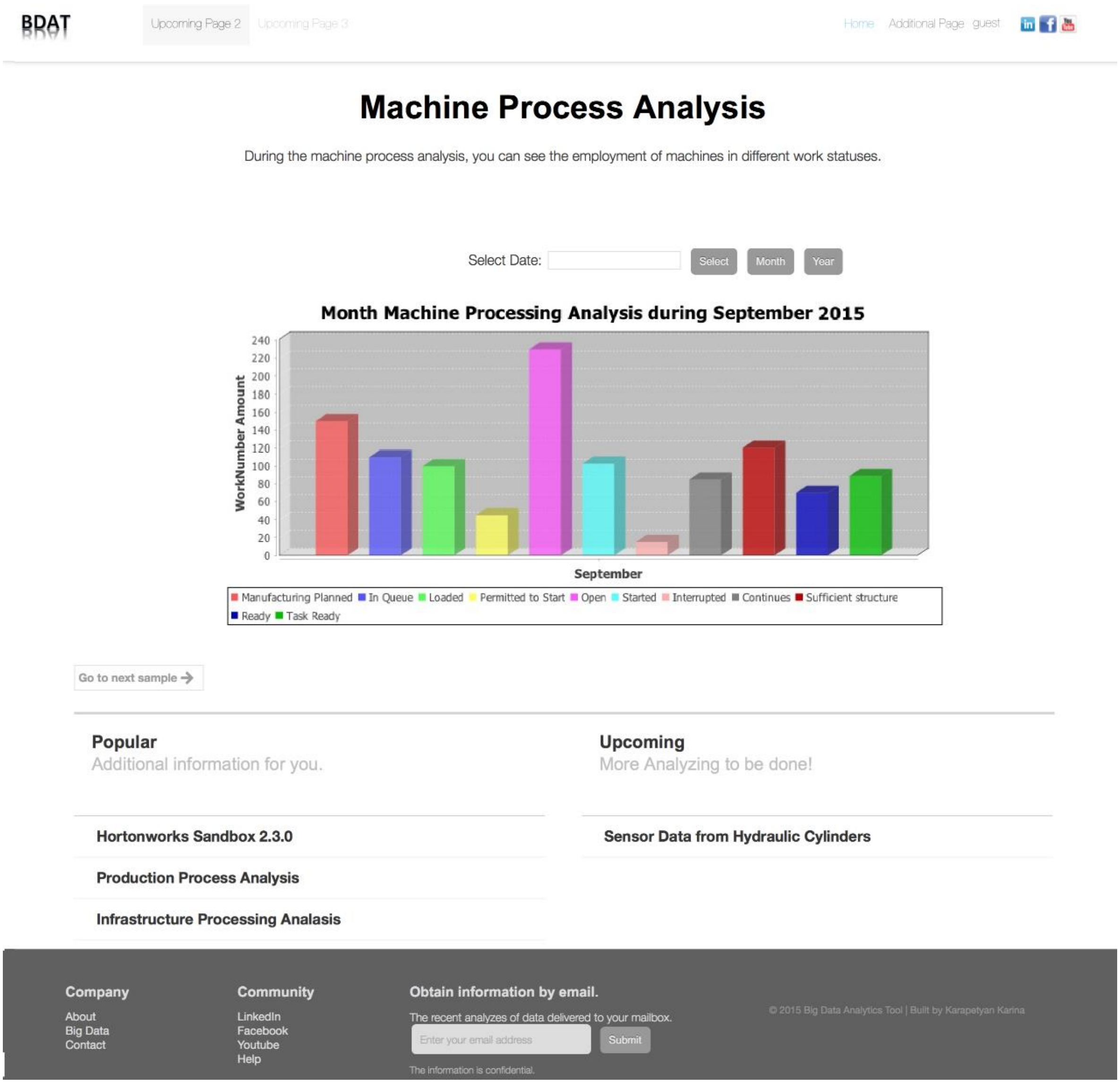
5. *Footer* – the last element of the web page that conveys the contact information of the company, the copyright for the Big Data Analytics Tool, links to the auxiliary resources and the date of the last system's update.

   The HTML code is responsible for the description of the webpage's logical structure, whereas CSS is utilized for the creation of its appearance. To create a user-friendly interface of the web page, CSS programming was employed to define colors, fonts, layout of HTML elements and other aspects of the design.

5.5 Machine Processing Analysis

Machine processing analysis is the second assay method, offered by Big Data Analytics Tool. The data analysis is demonstrated via a web page where a client, by selecting a specific date and type of the time period for data examination, obtains an analyzed information about the number of machines, operating in each working phase, in a form of a bar diagram. (FIGURE 27) The production procedure is divided into 11 categories:

- Manufacturing planned
- In Queue
- Loaded
- Permitted to start
- Open
- Started
- Interrupted
- Continues
- Sufficient structure
- Ready
- Task ready


BDAT
Upcoming Page 2
Upcoming Page 3
Home
Additional Page
guest
Select Date:
Select
Month
Year
Month Machine Processing Analysis during September 2015
WorkNumber Amount
240
220
200
180
160
140
120
100
80
60
40
20
0
September
🟥 Manufacturing Planned 🟦 In Queue 🟩 Loaded 🟨 Permitted to Start 🟪 Open 🟦 Started 🟥 Interrupted ⬛ Continues 🟥 Sufficient structure
🟦 Ready 🟩 Task Ready
Go to next sample →
Popular
Upcoming
Hortonworks Sandbox 2.3.0
Production Process Analysis
Infrastructure Processing Analasis
Sensor Data from Hydraulic Cylinders
Company
About
Big Data
Contact
Community
LinkedIn
Facebook
Youtube
Help
Obtain information by email.
Enter your email address
Submit
© 2015 Big Data Analytics Tool | Built by Karapetyan Karina

Machine Process Analysis

During the machine process analysis, you can see the employment of machines in different work statuses.

Additional information for you.

More Analyzing to be done!

The recent analyzes of data delivered to your mailbox.

The information is confidential.

FIGURE 27. The Machine Processing Analysis web page (Provided analyzed data chart is artificial due to non-disclosure agreement with the Case Company)

The HTML structure is similar to the infrastructure process analysis web page: the header, calendar and footer elements are performed with the same codes and resources. The technical side of this client application's web page is performed as follows:

1. When a user clicks submit button, the Java servlet, joint to the HTML calendar form, is triggered and reads the selected date and type of the button pressed: "Select", "Monthly" or "Yearly". Onwards, the message is forwarded to the server to build an analysis image, according to the chosen time parameters.
2. The server receives the initial data and makes respective Hive queries through SSH channel to Hortonworks Sandbox to observe the information. According to the result dataset, the analysis bar diagram is generated. Furthermore, it is converted into a binary array in order to be transferred to the client side.

3. The Java servlet transforms the received binary array back to the form of a bar chart and, with the help of JavaScript and Ajax coding. The client application uploads the diagram without reloading the web page.

5.6 Data Management

To grant Big Data Analytics Tool access to the new information, generated during production, the SQL Server tables are updated automatically via two methods:

1. ***DML[17] Triggers*** – a certain type of a stored procedure that is implemented automatically when a particular event takes a place in the relational database server: update, insert, delete. (Hutmacher 2013, 1) There are two categories of DML triggers:
    - *After Trigger* – When the DML operation is performed in the SQL Server, the corresponding to the event trigger is called. For instance, the database administrator successfully inserts a new record into the table, whereas the insert trigger is fired. In case the DML operation cannot be terminated, the trigger will not be invoked. (Shehzad 2009, 1)
    - *Instead of Trigger* – An opposite of "After Trigger" procedure that is fired before the DML operation is carried out. For example, instead of update trigger is conducted, whereas the actual update manipulation is not involved. (Shehzad et al. 2009, 1)
2. ***Stored procedure*** – a subprogram, created by a set of SQL commands that performs a certain task such as data verification or access management. This database object combines a logical sequence of SQL queries to avoid the execution of repetitive tasks in an application. Once the stored procedure is compiled, it is accumulated in the server. (Stored Procedures Database Engine 2014, 1)

Hydroline's enterprise resource planning software conveys the latest manufacturing data to the SQL Server database where it is allocated into the table for raw data maintenance. The after triggers are programmed on three data operations with the raw data: insert, update and delete. The triggers are fired to create an interaction and perform corresponding operations with other tables, arranged in the database. This code snippet introduces an after trigger on insert operation:

```
USE [ambari]
ALTER TRIGGER [dbo].[UpdateHydrolineWorksStatus]
ON [dbo].[WorksStatus]
AFTER INSERT, UPDATE, DELETE AS

DROP TABLE [dbo].[HydrolineWorksStatus]
CREATE TABLE [dbo].[HydrolineWorksStatus](
[RowId] [int] IDENTITY(1,1) NOT NULL,
[WorkNumberAccount] [int] NULL DEFAULT ((0)),
[EventValue] [nvarchar](255) NULL,
[RepDate] [date] NULL,
[RepUser] [nvarchar](max) NULL
```

[17] DML- an abbreviation for Data Manipulation Language

```
CONSTRAINT [PK_HydrolineWorksStatusPK] PRIMARY KEY CLUSTERED
([RowId] ASC) WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY])
ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]

INSERT INTO [dbo].[HydrolineWorksStatus]
([RepDate],[EventValue],[WorkNumberAccount],RepUser)
SELECT RepDate, [EventValue],SUM(WorkNumberAccount),STUFF((SELECT (','
+QuoteName(c.RepUser))
FROM WorksStatus c
GROUP BY c.RepUser
ORDER BY c.RepUser
FOR XML PATH(''), TYPE).value('.', 'NVARCHAR(MAX)'),1,1,'')

FROM [dbo].[HydrolineWorksStatus]
GROUP BY RepDate,EventValue
```

To avoid repeating the same codes, i.e., deleting or modifying a certain raw, appending new information, the triggers use stored procedures. This code snippet displays an example of stored procedure on inserting a new record.

```
USE [ambari]
ALTER PROCEDURE [dbo].[Values] AS
Declare @i date, @MaxI date, @TableWorksPhases TABLE (ActualDate date, MyRowCount Int Identity(1,1))
BEGIN

SET @i=(SELECT MIN(ActualDate) FROM WorksPhases);
SET @MaxI=(SELECT MAX(ActualDate) FROM WorksPhases);

WHILE @i<=@MaxI

INSERT INTO [dbo].[HydrolineOEEDay](
[WorkNumber],[ItemCode],[ResourceCode],[Description],[PhaseNumber],[ActualLoad],
[LoadReq],[NeedQty],[SatisfiedQty],[ActualDate],[ActualTime],[RealLastModDate],[RealL
astModTime],[ModDate],[ModTime],[Completness])

Select WorkNumber=(SELECT WorkNumber = REPLACE( (SELECT WorkNumber AS [data()]
FROM WorksPhases WHERE WorksPhases.ActualDate=@i GROUP BY WorkNumber FOR XML
PATH('')), ' ', ',')),

ItemCode=(SELECT ItemCode=REPLACE( (SELECT ItemCode AS [data()] FROM WorksPhases
WHERE WorksPhases.ActualDate=@i GROUP BY ItemCode FOR XML PATH('')), ' ', ',')),

ResourceCode=(Select ResourceCode=REPLACE( (SELECT ResourceCode AS [data()]
FROM WorksPhases
WHERE WorksPhases.ActualDate=@i
GROUP BY ResourceCode FOR XML PATH('')), ' ', ',')),

[Description]=(Select [Description]=REPLACE( (SELECT [Description] AS [data()]
FROM WorksPhases WHERE WorksPhases.ActualDate=@i GROUP BY [Description] FOR XML
PATH('')), ' ', ',')),

[PhaseNumber]=(Select [PhaseNumber]=REPLACE( (SELECT [PhaseNumber] AS [data()]
FROM WorksPhases WHERE WorksPhases.ActualDate=@i GROUP BY [PhaseNumber] FOR XML
PATH('')), ' ', ',')),

ActualLoad=(SELECT SUM(ActualLoad) FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
LoadReq = (SELECT SUM([Load]) FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
NeedQty=(SELECT SUM(NeedQty) FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
SatisfiedQty=(SELECT SUM(SatisfiedQty) FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
ActualDate= (SELECT ActualDate FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),

ActualTime= (Select [ActualTime]=REPLACE( (SELECT [ActualTime] AS [data()] FROM
WorksPhases WHERE WorksPhases.ActualDate=@i GROUP BY [ActualTime] FOR XML
PATH('')), ' ', ',')),

RealLastModDate= (SELECT RealLastModDate FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
RealLastModTime= (Select [RealLastModTime]=REPLACE( (SELECT [RealLastModTime] AS
[data()] FROM WorksPhases
WHERE WorksPhases.ActualDate=@i GROUP BY [RealLastModTime] FOR XML PATH('')), '
', ',')),

ModDate= (Select [ModDate] FROM WorksPhases
WHERE WorksPhases.ActualDate=@i),
ModTime= (Select [ModTime]=REPLACE( (SELECT [ModTime] AS [data()] FROM WorksPha-
ses WHERE WorksPhases.ActualDate=@i GROUP BY [ModTime] FOR XML PATH('')), ' ',
',')),

Completness= (Select AVG(Completeness) FROM WorksPhases
WHERE WorksPhases.ActualDate=@i)

SET @i=DATEADD(day,1,@i)
END
```

6 CONCLUSION

This thesis provides information according to the concept and realization of the real-time Big Data Analytics Tool for the engineering company. The work can be used in wider contexts. For instance, the software system can be utilized as a project and even development program for vendors with an eye to forecast possible halting of the production systems, determining the plan of action in case of eliminating the fabrication loss and, most importantly, obtain greater business value by upgrading the ventures' structures due to the analyzed data, received from Big Data Analytics Tool.

The purpose of this thesis work was to create and elaborate the software solution for the Finnish engineering company. The objectives set in the thesis were completed as the research progressed: as a result of the accomplished work, a fully operational Big Data Analytics Tool was carried out for Hydroline Oy. Big Data Analytics Tool offers Case Company the real-time data analyzing concerning machinery utilization, working load of employees, resource management, manufacturing expenses and waste, efficiency, availability, quality and performance indices on a daily, monthly and annual basis. The derived data is deployed for tracking the manufacturing process and evolving the current production. Moreover, Hydroline Oy is able to launch new competitive business services by fabricating a progressive version of the Smart Engineering product, taking into account the proposed software scrutiny.

In addition, Big Data is not only a solution for large enterprises. Medium and small companies have a supreme amount of opportunities to gain significantly valuable business insights, basing on the analysis of existing or novel data volumes, collected from internal or external resources. Essentially, ventures have to find creative ways and go beyond the comfort zone in order to operate with Big Data and employ its advantages. When a vendor is willing to invest time, resources and money into the strategy of working with Big Data, the outcome leads to the outstanding production results, optimization of the processes, competitive benefits, reduction of costs and operative time, globalization on the market arena and minimization of the environmental impact.

As for the future advancement of the project, more analysis samples are going to be performed to predict the potential failures of the production system and identify various factors, causing the manufacturing drawback.

7 REFERENCES

[1] Wen-Chen H., Kaabouch N., 2014. *Big Data Management, Technologies, and Applications, Information Science Reference*, an imprint IGI Global, p. 2

[2] Pries K., Dunnigan R., 2015. *Big Data Analytics: A Practical Guide for Managers*, Taylor & Francis Group, LLC, CRC Press, p. 2

[3] Hopkins B., 2011. *Big Data, Brewer, And A Couple Of Webinars* [web publication]. Forrester Inc. [accessed 20 April 2016].

Available from: http://blogs.forrester.com/brian_hopkins/11-08-29-big_data_brewer_and_a_couple_of_webinars

[4] Lee I., 2016. *Encyclopedia of E-Commerce Development, Implementation, and Management*, an imprint of IGI Global, p. 880-885

[5] Finlay S., 2014. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*, Designs and Patents Act, p. 17

[6] Woodie A., 2016. *Documentary Probes the Human Face of Big Data* [web publication]. Datanami Inc. [accessed 20 April 2016].

Available from: http://www.datanami.com/2016/02/24/documentary-probes-the-human-face-of-big-data/

[7] Sarma H., Rai P., Borah S., 2014. *Communication, Cloud and Big Data: Proceedings of CCB 2014,* ACCB Publishing, p.112-115

[8] SAS institute, 2016. *Big Data Analytics: What it is and why it matters?* [web publication]. SAS institute Inc. [accessed 20 April 2016].

Available from: http://www.sas.com/en_us/insights/analytics/big-data-analytics.html

[9] Bilbao-Osorio B., Dutta S., Lanvin B. 2014. *The Global Information Technology Report 2014 Rewards and Risks of Big Data*, World Economic Forum and INSEAD, p. 3-8

[10] Gross M., 2014. *Managing Risk With Big Data & Analytics* [web publication]. DATA & ANALYTICS [accessed 20 April 2016].

Available from: http://www.banktech.com/big-data/managing-risk-with-big-data-and-analytics/a/d-id/1297987

[11] Wolkowitz E., Parker S., 2015. *Big Data, Big Potential: Harnessing Data Technology for the Underserved Market*, The Center for Financial Services Innovation (CFSI), p. 3-9

[12] Thakuriah P., Geers G., 2013. *Transportation and Information: Trends in Technology and Policy*, Springer New York Heidelberg, p. 5-7

[13] The Apache Software Foundation, 2008. *MapReduce Tutorial* [electronic file] The Apache Software Foundation [accessed 20 April 2016].

Available from: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.pdf

[14] Chansler, Kuang, Radia, Shvachko, Srinivas, 2016. *The Hadoop Distributed File System* [web publication]. The Apache Software Foundation [accessed 20 April 2016].

Available from: http://www.aosabook.org/en/hdfs.html

[15] Tutorialspoint simplyeasylearning, 2015. *HIVE: hive query language*, Tutorials Point (I) Pvt. Ltd, p. 1-5

[16] Lo F., 2015*. Big Data Technology: What is Hadoop? What is MapReduce? What is NoSQL?* [web publication]. Datajobs [accessed 22 April 2016].

Available from: https://datajobs.com/what-is-hadoop-and-nosql

[17] Rouse M., 2015. *MPP database (massively parallel processing database)* [web publication]. TechTarget SearchDataManagement Foundation [accessed 23 April 2016].

Available from: http://searchdatamanagement.techtarget.com/definition/MPP-database-massively-parallel-processing-database

[18] Juneau J., 2014, *JavaServer Faces: Introduction by Example*, Springer Science+Business Media New York, p. 1-2

[19] Achari S. 2015, Hadoop Essentials, *Packt Publishing*, p.15-16

[20] Schniederjans M., Schniederjans D., Starkey C., 2014*. Business Analytics Principles, Concepts, and Applications: What, Why, and How*, Pearson Education Inc., p.4

[21] Laney D., 2014. *Deja VVVu: Others Claiming Gartner's Construct for Big Data* [web publication]. Gartner Blog Network [accessed 23 April 2016].

Available from: http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

[22] Dragland, 2013. *Big Data, for better or worse: 90% of world's data generated over last two years Data* [web publication]. ScienceDaily [accessed 23 April 2016].

Available from: https://www.sciencedaily.com/releases/2013/05/130522085217.htm

[22] NYSE Euronext, 2013. *NYSE Euronext Adapting to market changes with near-real-time insight into information* [electronic file] IBM Corporation [accessed 23 April 2016].

Available from: http://www.ibmbigdatahub.com/sites/default/files/document/NYSE-Euronext-IMC14787USEN.PDF

[23] North M., Riniker S., 2014. *Consumer sentiment extraction from unstructured data* [web publication]. Issues in Information Systems, pp. 430-433 [accessed 24 April 2016].

Available from: http://iacis.org/iis/2014/79_iis_2014_430-433.pdf

[24] IBM Big Data & Analytics Hub, 2014. *Infographics & Animations: The Four V's of Big Data* [electronic file]. IBM Corporation [accessed 24 April 2016].

Available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[25] Business Infographics, 2016. *The Cost Of Mismanaged Data* [electronic file]. Designinfographics [accessed 24 April 2016].

Available from: http://www.designinfographics.com/business-infographics/the-cost-of-mismanaged-data

[26] Biehn N., 2016. *The Missing V's in Big Data: Viability and Value* [web publication]. Wired [accessed 25 April 2016].

Available from: http://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/

[27] Knilans E., 2014. *The 5 V's of Big Data* [web publication] Avnet Advantage: The Blog [accessed 25 April 2016].

Available from: http://ats.avnet.com/na/en-us/news/Pages/The-5-Vs-of-Big-Data.aspx

[28] Brijs B., 2013. *Business Analysis for Business Intelligence*, CRC Press, Taylor & Francis Group, p. 6

[29] Gartner Newsroom, 2015. Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years [web publication] Gartner Incorporation [accessed 26 April 2016].

Available from: http://www.gartner.com/newsroom/id/3130817

[30] Costley & Lankford 2014. Big Data Cases in Banking and Securities, IBM Inc., Securities Technology Analysis Center, LLC, STAC , 14 (3-14)

[31] Jackson J., 2015. *Aiming for SEC's big data project, Sungard and Google bet on the cloud* [web publication] Computerworld [accessed 27 April 2016].

Available from: http://www.computerworld.com/article/2935588/big-data/aiming-for-secs-big-data-project-sungard-and-google-bet-on-the-cloud.html

[32] Gaitho M., 2015. *How Applications of Big Data Drive Industries* [web publication] Simplilearn [accessed 27 April 2016].

Available from: http://www.simplilearn.com/big-data-applications-in-industries-article

[33] Connolly B., 2014. *Uni of Tasmania's learning system to identify 'at risk' students* [web publication] CIO [accessed 28 April 2016].

Available from: http://www.cio.com.au/article/553925/uni_tasmania_learning_system_identify_risk_students/

[34] Beaudoin J., 2015. *FDA must make smarter use of big data* [web publication] Healthcare IT News [accessed 28 April 2016].

Available from: http://www.healthcareitnews.com/news/fda-must-make-smarter-use-big-data

[35] Rotella P., 2012. *Is Data The New Oil?* [web publication] Forbes Tech [accessed 28 April 2016].

Available from: http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#7ccae9a577a9

[36] Kalyvas J., Overly M., 2015. *Big Data: A Business and Legal Guide*, CRC Press, Taylor & Francis Group, p. 1-4

[37] Gartner IT Glossary, 2012. *Big Data* [electronic file] Gartner Incorporation [accessed 28 April 2016].

Available from: http://www.gartner.com/it-glossary/big-data/

[38] Weathington J., 2012. *Big Data defined* [electronic file] TechRepublic [accessed 28 April 2016].

Available from: http://www.techrepublic.com/blog/big-data-analytics/big-data-defined/

[39] Mayer-Schönberger V., Cukier K., 2013. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*, Houghton Mifflin Harcourt Publishing Company, p.244 (97)

[40] McAfee A., Erik Brynjolfsson E., 2012*. Big Data: The Management Revolution [web publication] Harvard Business Review* [accessed 29 April 2016].

Available from: https://hbr.org/2012/10/big-data-the-management-revolution/ar

[41] Petty A., 2014. *Enterprise Innovation and Emerging Technologies: Defining Needs* [web publication] Business 2 Community [accessed 29 April 2016].

Available from: http://www.business2community.com/strategy/enterprise-innovation-emerging-technologies-defining-needs-0835857#iqAm3Fj1gC6HTDJe.97

[42] Hassanien A., Azar A., Snasel V., Kacprzyk J., Abawajy J., 2015. *Big Data in Complex Systems: Challenges and Opportunities*, Springer International Publishing Switzerland, p. 13-20

[43] O'Dowd S., 2015. *Top 10 Big Data Trends in 2016 for Financial Services* [web publication] Converge blog Powered by MapR [accessed 29 April 2016].

Available from: https://www.mapr.com/blog/top-10-big-data-trends-2016-financial-services

[44] Shehzad A., 2009. *Using INSTEAD OF triggers in SQL Server for DML operations* [web publication] MSSQLTips.com [accessed 29April 2016].

Available from: https://www.mssqltips.com/sqlservertip/1804/using-instead-of-triggers-in-sql-server-for-dml-operations/

[45] Stalin Z., 2014. *Zora's SQL Tips* [web publication] SQL ServerCentral.com [accessed 29 April 2016].

Available from: http://www.sqlservercentral.com/blogs/zoras-sql-tips/2014/10/10/introduction-of-triggers-in-sql-server/

[46] SQL Server, 2008. *Triggers -- SQL Server* [web publication] Code Project [accessed 29 April 2016].

Available from: http://www.codeproject.com/Articles/25600/Triggers-SQL-Server

[47] Java blog, 2007. *Accessing SQL Server on NetBeans using JDBC, Part 1: Create a connection* [web publication] Linglom.com Just another IT weblog [accessed 30 April 2016].

Available from: http://www.linglom.com/programming/java/accessing-sql-server-on-netbeans-using-jdbc-part-i-create-a-connection/

[48] Ferrill P., 2013. *10 excellent new features in Windows Server 2012 R2* [web publication] InfoWorld [accessed 30 April 2016].

Available from: http://www.infoworld.com/article/2606748/microsoft-windows/108930-10-excellent-new-features-in-Windows-Server-2012-R2.html#slide2

[49] Microsoft Corporation official website, 2014. *Microsoft® SQL Server® 2012 Express 2012* [electronic form] Microsoft Corporation [accessed 30 April 2016].

Available from: https://www.microsoft.com/en-us/download/details.aspx?id=29062

[50] Developer Network, 2014. *SQL Server Configuration Manager 2016* [web publication] Microsoft MSDN website [accessed 30 April 2016].

Available from: https://msdn.microsoft.com/en-us/library/ms174212.aspx

[51] TechNet, 2006. *SQL Server Surface Area Configuration* [web publication] Microsoft TechNet website [accessed 30 April 2016].

Available from: https://technet.microsoft.com/en-us/library/ms173748(v=sql.90).aspx

[52] Imran M., 2014. *SQL Server Management Studio – A step-by-step installation guide* [web publication] SQL-Shack [accessed 30 April 2016].

Available from: http://www.sqlshack.com/sql-server-management-studio-step-step-installation-guide/

[53] Developer Network, 2014. *Integration Services in Business Intelligence Development Studio* [web publication] Microsoft MSDN website [accessed 30 April 2016].

Available from: https://msdn.microsoft.com/en-us/library/ms174181(v=sql.105).aspx

[54] VirtualBox, 2015. *Welcome to VirtualBox.org!* [web publication] VirtualBox official website [accessed 30 April 2016].

Available from: https://www.virtualbox.org/

[55] VirtualBox, 2015. *Virtual Machines* [web publication] VirtualBox official website [accessed 30 April 2016].

Available from: https://www.virtualbox.org/wiki/Virtualization

[56] Tutorialspoint simplyeasylearning, 2014. *Hadoop Big Data Analysis Framework*,Tutorials Point (I) Pvt. Ltd., p. 6-7

[57] Hortonworks, 2015*. Learning The Ropes Of The Hortonworks Sandbox* [web publication] Sandbox Hortonworks Incorporation [accessed 30 April 2016].

Available from: http://hortonworks.com/hadoop-tutorial/learning-the-ropes-of-the-hortonworks-sandbox/#what-is-the-sandbox

[58] Gauraw K., 2015. *Big Data And Hadoop – Features And Core Architecture* [web publication] Data Integration Ninja [accessed 30 April 2016].

Available from: http://www.dataintegration.ninja/big-data-and-hadoop-features-and-core-architecture/

[59] NetBeans Releases & Planning, 2015. *NetBeans IDE 8.1 Installation Instructions* [web publication] NetBeans official website [accessed 30 April 2016].

Available from: https://netbeans.org/community/releases/81/install.html

[60] Oracle Incorporation, 2013. *GlassFish Server Open Source Edition Quick Start Guide Release 4.0*, p. 1

[61] Hortonworks Sandbox Inc., 2013. *Guide, Installing Hortonworks Sandbox 2.0 – VirtualBox on Windows*, p. 1-11

[62] Tutorialspoint simplyeasylearning, 2015. *Sqoop Data Transfer Tool*, Tutorials Point (I) Pvt. Ltd., p. 1-2

[62] Rouse M., 2015. *Client/server (client/server model, client/server architecture)* [electronic file] TechTarget SearchNetworking [accessed 30 April 2016].

Available from: http://searchnetworking.techtarget.com/definition/client-server

[63] Beal V., 2015. *Client-server architecture* [web publication] Webopedia [accessed 1 May 2016].

Available from: http://www.webopedia.com/TERM/C/client_server_architecture.html

[64] Computer Networking Demystified, 2013. *What is a Client-Server Network* [web publication] Computer Networking Demystified [accessed 1 May 2016].

Available from: http://computernetworkingsimplified.com/category-1/classification-of-networks/service-based-classification/what-is-a-client-server-network/

[65] Mitchell B., 2014. *WAN-Wide Area Network* [web publication] About Tech [accessed 1 May 2016].

Available from: http://compnetworking.about.com/cs/lanvlanwan/g/bldef_wan.htm

[66] Abrams L., 2004. *TCP and UDP Ports Explained* [web publication] Bleepingcomputer [accessed 1 May 2016].

Available from: http://www.bleepingcomputer.com/tutorials/tcp-and-udp-ports-explained/

[67] Hansen R., 2001. *Overall Equipment Effectiveness: A Powerful Production/maintenance Tool for Increased Profits*, Industrial Press Inc., p. 8-33

[68] Stamatis D., 2011. *The OEE Primer Understanding Overall Equipment Effectiveness, Reliability, and Maintainability*, CRC Press, Taylor & Francis Group, p. 47-89

[69] Hutmacher H., 2013. *Introduction to DML triggers* [web publication] SQL Sunday [accessed 1 May 2016].

Available from: https://sqlsunday.com/2013/03/03/introduction-to-dml-triggers/

[70] Developer Network, 2016. *Stored Procedures* (Database Engine) [web publication] Microsoft MSDN website [accessed 2 May 2016].

Available from: https://msdn.microsoft.com/en-us/library/ms190782.aspx

[71] Sandbox Hortonworks Guide, 2015. *How to Analyze Machine and Sensor Data* [web publication] Sandbox Hortonworks Incorporation [accessed 3 May 2016].

Available from: http://hortonworks.com/hadoop-tutorial/how-to-analyze-machine-and-sensor-data/

[72] Shelley P., 2012, *Big Data Debate: End Near For ETL?* [web publication] InformationWeek [accessed 3 May 2016].

Available from: http://www.informationweek.com/big-data/big-data-analytics/big-data-debate-end-near-for-etl/d/d-id/1107641

[73] Isaacson C., 2014. *Big Data Scalability: Why Your Database Is Slow, and When You Should Start Scaling* [web publication] InformIT [accessed 3 May 2016].

Available from: http://www.informit.com/articles/article.aspx?p=2258416

[74] UPPM, 2016. *R&D* [electronic file] Uttaranchal Pulp and Paper Mill (Kraft Paper Mill) [accessed 8 May 2016].

Available from: http://www.uppmindia.com/index.php/technology/r-d

[75] Robirds E., 2016. *The sales department logo* [electronic file] Edward Robirds – Art & Design [accessed 8 May 2016].

Available from: http://edwardrobirds.com/work/logos/the-sales-department-logo/

[76] Duct Tape Marketing, 2015. *Marketing and department* [electronic file] Duct Tape Marketing official website [accessed 8 May 2016].

Available from: http://www.ducttapemarketing.com/wp-content/uploads/2015/07/marketing-department.png

[77] Halogen Software, 2012. *Talent management for manufacturer department* [electronic file] Halogen Software official website [accessed 8 May 2016].

Available from: http://www.halogensoftware.com/uploads/learn/centers-of-excellence/talent-management-for-manufacturers/2-coe-manufacturing.png

[78] Bright Powertech Ltd, 2012. *After Sales Service* [electronic file] Bright Powertech Ltd official website [accessed 8 May 2016].

Available from: http://brightpowertech.net/services/

[79] Iconfinder, 2015. *Express Shipping* [electronic file] Iconfinder official website [accessed 8 May 2016].

Available from: https://cdn3.iconfinder.com/data/icons/shopping-1/256/Express_Shipping-512.png

[80] Aboard Software, 2014. *Small Business Consolidators* [electronic file] Aboard Software official website [accessed 20 May 2016].

Available from: http://www.aboardsoftware.com/ProductsAndSolutions/SmallBusinessConsolidator.aspx

[81] Hydraulex Global, 2014. *HYDRAULIC CYLINDERS* [electronic file] Hydraulex Global official website [accessed 20 May 2016].

Available from: http://www.hydraulex.com/hydraulic-cylinders.php

[82] Johnson Pet Trade Consultants BV, 2013. *CUSTOMER RELATIONSHIP MANAGEMENT* [electronic file] Johnson Pet Trade Consultants official website [accessed 20 May 2016].

Available from: http://www.johnsonptc.com/products-and-brands/customer-relationship-management

[83] Niccolai J., 2010. *Microsoft ending support for Itanium* [web publication] Computerworld [accessed 23 May 2016].

Available from: http://www.computerworld.com/article/2516742/computer-hardware/microsoft-ending-support-for-itanium.html