Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technologies

Bachelor Thesis

Big Data in Banking Industry

Author: Mohammad Ghalamkaran

Supervisor: Ing. Jan Pavlík

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Mohammad Ghalamkaran

Systems Engineering and Informatics
Informatics

Thesis title

Big data in banking industry

Objectives of thesis

The main objective is to propose a big data solution for a case company in the banking industry, leveraging the advantages of big data approach to address problems and needs specific to the financial sector. The partial objectives are:
- Perform literature review focusing on the topics of big data and banking industry
- Overview the current usage of big data in banking, outline common approaches and trends
- Create a case company, establish its properties and needs with regards to data processing and storage
- Propose a big data solution for the case company and evaluate its suitability and benefits

Methodology

Methodology of the theoretical part is based on literature review regarding big data and its methods of implementation and specific needs of the banking sector. Practical part includes creation of a case company within the financial sector and proposal of a custom big data solution. The proposed solution will be evaluated in terms of its suitability, practicality, costs and other relevant factors. Based on the outcomes of both the literature review and the case study, final conclusions will be made.

The proposed extent of the thesis

30-40

Keywords

Big data, banking industry, finance, information technology, data storage, data processing

Recommended information sources

Buyya , R., Calheiros, R., & Dastjerdi, A. (2016). Big Data Principles and Paradigms. MA 02139, USA: Elsevier Inc.

Li, K.-C., Jiang, H., & Zomaya, A. (2017). Big Data Management and Processing. Florida, USA: Taylor & Francis Group, LLC.

Mishkin, F. (2019). THE ECONOMICS OF MONEY, BANKING, AND FINANCIAL MARKETS. London: Pearson Education.

Prabhu, C., Mogadala, A., Livingston, L., Sreevallabh Chivukula, A., & Ghosh, R. (2019). Big Data Analytics: Systems, Algorithms, Applications. Singapore: Springer Nature Singapore Pte Ltd.

Expected date of thesis defence

2020/21 SS – FEM

The Bachelor Thesis Supervisor

Ing. Jan Pavlík

Supervising department

Department of Information Technologies

Electronic approval: 2. 11. 2020

Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 5. 11. 2020

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 25. 02. 2021

Declaration

I declare that I have worked on my bachelor thesis titled "Big Data in Banking Industry" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 25.02.2021

____
Mohammad Ghalamkaran

Acknowledgement

I would like to thank express great gratitude to my supervisor Ing. Jan Pavlík for the support and guidance, he gave me during my work which had a great impact on my thesis and research.

I would like to thank the administrator, professors, and staff of the faculty, who impacted me during my study.

I would like to express gratitude and appreciation to all my friends for the support and encouragement they gave me during my studies.

I extend my warm gratitude to my Mom, Dad, Sister and my Uncle who always supported and encouraging me in my life. They have always been a motivation for me to be better.

Big Data in Banking Industry

Abstract

The aim of this bachelor thesis is to implement Big Data into a case company “Mobank”, which is founded in 04.07.1997 in Frankfurt, Germany. Further details of this bank properties are available in practical part of this thesis. This work is intended for financial sector of the banking industry with the usage of Heuristic Machine Learning Imputation (HMLI), which deals with different sets of missing data. The thesis consists of two major part: theoretical part and practical work. In the first part, the literature review of chosen methodology will be provided. It means that every aspect of Big Data will be precisely described. It includes detailed description of different Big Data definitions, storage and processing, architecture and its impact on financial sectors. In practical section of this thesis, the above-mentioned case company, its issue of missing data and selected solution will be evaluated and calculated for its efficiency and suitability. The data will be presented in the form of tables. Based on the results and the literature review, final conclusion will be made.

Keywords: Big data, banking industry, finance, information technology, data storage, data process

Table of content

List of pictures

List of tables

List of abbreviations

HMLI Heuristic	Machine Learning Imputation
ETL	Extraction, Transformation, and Load
BDA	Big Data Analytics
HDFS	Hadoop Distributed File System
BFSI	Banking Financial Services and Insurance
AI	Artificial Intelligence
MCAR	Missing Completely at Random
MAR	Missing at Random
MNAR	Missing Not at Random
SVM	Support Vector Machine
RMSE	Root Mean Square Error
R&D	Research and Development

1 Introduction

It is 2021 and Big Data is everywhere, on TV, in research papers, in the newspapers etc. Although it is a very catchy name but also its credit lies in the sheer quantity of data available. According to IBM, every day 2,5 quintillion bytes of data are created but what does that mean? What is the big deal and most importantly how does this affect or overlaps with personal data?

The first step is to introduce Big Data and its definitions by major companies in information technology. Furthermore, the way Big Data is structured, handled and stored will be thoroughly described and explained. However, the focus of the thesis is Big Data and its impact in Banking Industry, more specifically in financial sectors.

Secondly, a selected case company is presented, where a Big Data solution is proposed in order to deal with the specific issue of the thesis and established the proper steps towards the result. The solution is HMLI for Missing Data values., which primary cost of investing in the proposed solution is suggested to find if the selected HMLI is the proper method to deal with the issue.

“Mobank” inspired by Commerzbank A.G, the second largest bank in Germany, by total value of its balance sheet, but with lower numbers and margins properties.

2 Objectives and Methodology

2.1 Objectives

The main objective is to propose a big data solution for a case company in the banking industry, leveraging the advantages of big data approach to address problems and needs specific to the financial sector. The partial objectives are:

- Perform literature review focusing on the topics of big data and banking industry
- Overview the current usage of big data in banking, outline common approaches and trends
- Create a case company, establish its properties and needs with regards to data processing and storage
- Propose a big data solution for the case company and evaluate its suitability and benefits

2.2 Methodology

Methodology of the theoretical part is based on literature review regarding big data and its methods of implementation and specific needs of the banking sector. Practical part includes creation of a case company within the financial sector and proposal of a custom big data solution. The proposed solution will be evaluated in terms of its suitability, practicality, costs and other relevant factors. Based on the outcomes of both the literature review and the case study, final conclusions will be made.

3 Literature Review

3.1 Big Data

Data Science has been a discipline that emerged in the last few years, as how did Big Data concept. There have been many different interpretations to what Data Science is. Our adaptation from data science will be the following: Data Science provides significant amount of information based on different sets of complex data or Big Data. Data Science or also known as data-driven science, is the combination of different work fields in statistics and data computation and interpretation for decision making purposes. However, “Big Data” has become exceedingly popular nowadays, but in general there are not a lot of people who know what it really means. Many professional data analysts may imply that Big Data is the process of extraction, transformation, and load, which is also known as ETL, for large datasets. An extremely popular description of Big Data is based on three attributes of Data: volume, velocity, and variety or (3Vs). However, this is not accurately captures all the aspects of Big Data. To do so, we would need to investigate this term historically and see the evolution that Big Data has been through to todays connotation.

3.1.1 History of Big Data

There have been several studies conducted on the historical views in BDA area. Gil Press provided a summary of Big Data from 1944, which was based on Rider’s work. In his studies, the coverage of the history of Big Data’s evolution is between 1944 to 2012 and demonstrated 32 Big Data-related events in current data science history. As its been indicated in Press’s article, the thin line between the growth of data and Big Data has become blurry. Frequently, the growth rate of data has been referred as “information explosion’’; although “data” and “information” are generally used reciprocally, the two terms have different connotations. Press’s study is very extensive and covers BDA events until December 2013. Since then, there have been many other Big Data related events. Nonetheless, its review greatly covered both Big Data and data science events. Thus, the term “Data Science” could be considered as a complementary meaning to BDA.

But today’s definition is based on Gartner, a leading information technology research and advisory company, that defines Big Data as it is previously mentioned, in “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. Simply put, big data is larger, complex data sets, specifically from new data sources. High voluminous of these data sets are more complicated that traditional data processing software can not manage. On the other hand, these huge data volumes can be used to direct business problems that were "untackleable". (Gartner, 2012)

3.2 Definition of Big Data

I have mentioned the 3vs to Big Data before but now I would like to explain this matter in more details.

Gartner

Many attributes have been added to Big Data since 1997. Three of them are the most popular and universally adopted. The first one is Gartner's adaptation or most famously known as 3VS definition.

1. Volume. The amount of data is what matters. These high-volume data of low-density need to be processed. This can be any data of unknown value. For example, Twitter data feeds, sensor-enabled equipment, or any clickstreams on a webpage.
2. Velocity The rate which data is received and follows up to. Generally, memory is where the highest velocity of data streams to in opposition to being written to a local disk. Real time or near-real time act are implemented in internet-enabled smart products and requires high evaluation and action.
3. Variety The signification of many types of data. The structure of traditional data types was stored in a relational database. But with the new rise of Big Data, the new data comes in a unstructured data types, like audio, text and video and they are required additional preprocessing to acquire meaning and support metadata.

IBM

IBM added another attribute or "V" to the list of common attributes, which is known as the "4VS of Big Data". That "V" stands for "Veracity", which involves the uncertainty of data. The reason behind this, is ""in response to the quality and source issues our clients began facing with their Big Data initiatives".

Microsoft

Microsoft extended the list of Vs to six, which is widely know for the sake of maximizing the business value. They have added variability, veracity and visibility. In this case, “Veracity stands for trustworthiness of data sources, “Variability” is referred to the complexity of data sets, which in comparison to “Variety” it is known for the number of variables. And finally, “Visibility” highlights the fact that a clear and full picture of data is needed in order to make informative and instructive decisions.

Other definition of Big Data

In 2013, a 5vs definition has been proposed by Yuri Demchenko which is the added value of dimension along with the IBM’s 4Vs definition.

All these definitions are data-oriented articulation of many different aspects of data. (Rajkumar, 2016)

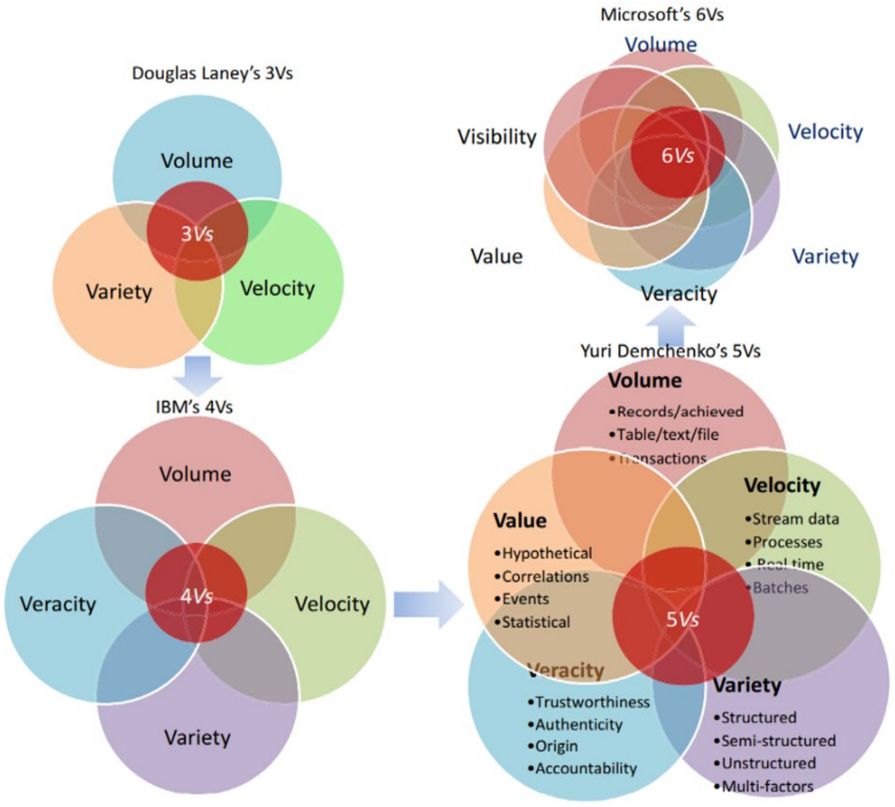


Figure 1: Different Vs Definition (Rajkumar, 2016)

3.3 Big Data Processing and Storage

Big Data Processing has similar steps to processing data in the transactional or data warehouse. The following image shows different steps and stages involving the Big Data process.

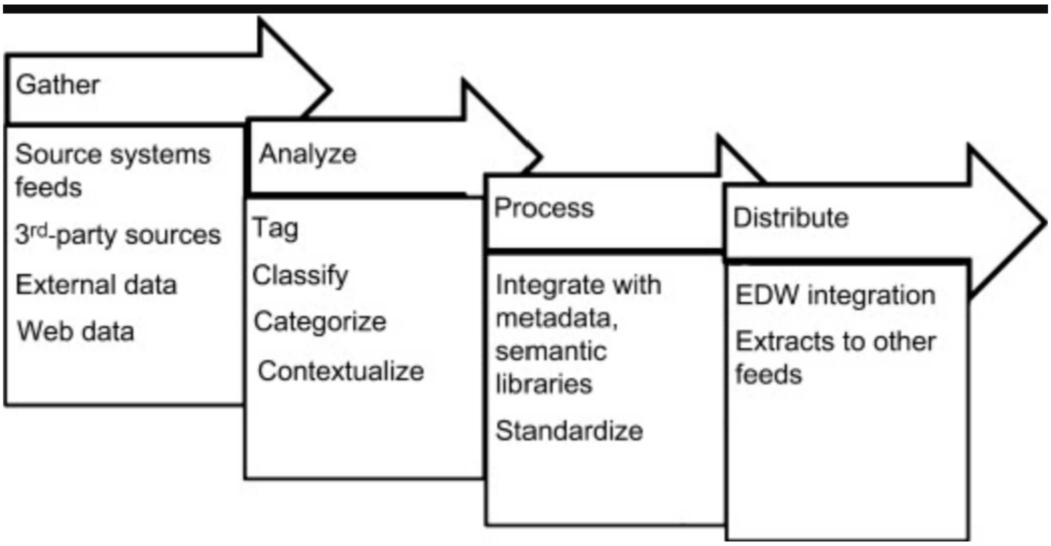


Figure 2: Big Data Processing Steps (Krishnan, 2013)

Important note: While there are similarities in processing Big Data to traditional data processing, there are some key differences such as:

- Data is analyzed and then processed
- Data Standardization happens in the Analyze Stage
- No special emphasis in on data quality where except the usage of metadata, master data, and semantic libraries for data enhancement
- Preparation of data occurs in Analyze stage for further processing and integration

3.3.1 Gather Stage

Data is received from sources including real time systems or near-real time systems. After the data collection, it is loaded to a storage like Hadoop or NoSQL.

3.3.2 Analysis Stage

This stage is also known as data discovery stage for processing Big Data and preparing for data warehouses or integration to the structured analytical platforms. Tagging, classification and categorization of data are the substages in the analysis stage, which closely features the creation data model in the data warehouse.

Tagging: Since 2003, tagging has been a common practice which is frequent on the Internet for data sharing. It is the process of application a term to an unstructured fragment of information that provides metadata-like attribution. It also creates an nonhierarchic type data set that can be used for processing data downstream in the "Process" stage.

Classify: unstructured and unregulated data comes from different multiple sources and is stored in the "Gathering Process"

This classification helps to make a group of data into subject-oriented data sets and eases the data processing.

Categorize: Categorization process is the external organization of data from the data is grouped by both the classification and data type. This happens from a storage perspective to manage the cycle of the data which is stored as a write-once model in the storage layer.

3.3.3 Process Stage

There are several substages to processing Big Data and its significance and transformation at each substages is to produce the correct or incorrect output.

Context Processing: is related to the context of data occurrence within the Big data environment. This relevancy will help with the processing of the proper metadata and master data. The biggest advantage of this processing is the ability to process the same data for multiple and various contexts, aiding in looking for patterns within each result ser for further data exploration and mining.

Metadata, master data, and semantic linkage: the ability to use metadata, semantic libraries, and master data as the integration links is the most important part in creation and integration of Big Data into a data warehouse. This step is started once the data is tagged and geocoding and contextualization are completed. The next part to this process is to connect the data to the enterprise data set. There are different techniques to link the data between structured and unstructured data sets.

Standardize: Standardization of data is required for processing Big Data's integration with the data warehouse, which significantly improves data quality. It also requires the processing of master data components with the data like replacing the master data definitions with the found keys in the data set. For example, taking the data form a social media platform has an exceptionally low chance in finding keys or data attributes that can link to the master data. But if the processed data is owned by the enterprise such as product data, there are higher chances of finding matches with the master data and can be easily integrated. (Krishnan, 2013)

3.3.4 Distribute Stage

This is the stage where the distribution of the Big Data happens. It is distributed to downstream systems by processing its reporting system and analytical applications. By using the data processing outputs from the processing stage, it is loaded into these systems for further process. Web reporting and content management platforms are another distribution technique that involves data exportation as flat files for use. (Krishnan, 2013)

3.3.5 Storage

Big data storage is an infrastructure, designed precisely to store, manage, and retrieve huge amounts of data. This facilitates the storage and its sorting of data in a way that it is easily accessible, usable and processable by applications or services working on big data.

A big data storage system bundles massive number of stock servers attached to high-capacity disk to support analytic software written to crunch huge quantities of data. This system is relied on Massively Parallel Processing databases for analysis of data which were consumed from various sources. However, big data usually lacks structure and comes from various sources, thus making it a inferior fit for processing within a relational database. The Apache Hadoop Distributed File System (HDFS) is the most common analytics engine used for big data, that is generally combined with a slight touch of a NoSQL database.

Hadoop is an open-source software written in Java and transmits the data analytics across thousands of server nodes without a drop in performance. Within its MapReduce component, Hadoop administers its processing in a way as a safeguard for slightest possibility of catastrophic failures. The multiples nodes are a platform for data analysis for when a query arrives, MapReduce starts processing directly on the storage node where data is located. Once this is completed, MapReduce gathers the results from each server and scales it down to display a single united response. (Kranz, 2017)

3.4 Big Data Architecture

Big data architecture is important for big data analytics. This staggering system is used to manage massive amounts of data so it can be further analyzed and process for business purposes and present an environment where big data analytics can extort vital business information from raw data. The big data architecture framework can be served as a blueprint for big data solutions, reasonably characterizing how these solutions will work, which components will need to be used.

3.4.1 Big Data Architecture Layers

A big data environment can manage both batch and realtime processing of sources, such as IoT devices, data warehouses or relational database management systems. This layer is called **Big Data Sources.** Furthermore, **management and storage layer** obtain the data from the mentioned sources, convert them into a comprehensible data format for analytic purposes and store them accordingly. From this layer, analytics pull the data in form of business intelligence which is called the **analysis layer**. Lastly, **consumption layer** receives the results and displays them to the pertinent output layer which is also known as business intelligence layer. (OMNI.SCI, 2020)

3.4.2 Big Data Architecture processes

- Connecting to data sources: this process includes adapters capably and efficiently associating any form of data and connecting them to varios storage systems, networks and protocols.
- Data Governance: This step arranges privacy and security for operating from the beginning of the ingestion to processing, analysis storage and lastly deletion stage.
- System Management: this process is highly and largely scalable distributes cluster of data for modern big data architectures, which should be monitored non stop through central management consoles.
- Protecting Quality of Service: this framework backs the definition of data quality, ingestion frequency and supports compliance policies.

To unlock the full potential of big data, it is essential to invest in a infrastructure capable of managing and handling massive quantities of data. There are several benefits to this investment form improving analysis of big data to predicting future trends and consistent implementation technology. Big data infrastructure also poses challenges including management of data quality, which needs considerable analysis; ultimately could be costly and if not sufficient could lead to increasing complexity of big data sets. (OMNI.SCI, 2020)

3.5 Big data and Its Impact on Financial Sectors

Nowadays, Big Data is becoming a powerful game-changer in the financial sectors. Thus, banking financial services and insurance (BFSI) sectors are going full throttle to expand their business opportunities and services, that they provide.

According to an article that was published by **Forbes** “over 2.5 quintillion bytes of data created each day.” (Marr, 2018) This is a huge number, and the rise of customer volume has a humongous pressure on banks and similar organizations to offer more quality services and raise the bar to differentiate themselves from competitors. So, they are looking for new ways every day to not only manage the data better but also to use it to track their customer’s behaviors. This allows them to give their customers the right and exact kind of resources and service they need. Big Data resources have provided these organizations with enhanced customer service and boosted profitability.

This begs the question: “Why BFSI sectors need Big Data?”. The answer can be summarized in 3 main points: 1. Customer Experience 2. Operation Optimization 3. Employee Engagement

3.5.1 Customer Experience

Customers have relatively high expectations from their banks. From the interaction between them to the offered services. And the increase in customer volume has made a headache for banks to fulfil every demands of customers in an orderly way.

Big Data is a big help to help these organizations to have a in detailed picture of their customers. It provides them with constructive personal information, browsing and transaction history etc. helping them in remarkable ways to create customer segmentation. On top of that, Big Data also helps in marketing decisions that banks need to make, in order to bring a notable increase in their marketing productivity. (Lochy, 2019)

3.5.2 Operation Optimization

Big Data can help with improving the predictive power of their risk models. Interestingly, this is one of the main reasons, if not the main reason, that Big Data is gaining increasing popularity in BFSI sectors and ultimately in banking industry. (Lochy, 2019)

Big Data adds a more comprehensive risk coverage and significant cost savings. There are many areas in risk management where Big Data can be used to gain risk intelligence such as:

- Fraud Management
- Credit Management
- Market and Commercial Loans
- Operational Risks
- Integrated Risk Management

3.5.3 Employee Engagement

One of the remarkably interesting Big Data application is in detecting employee performance. This is a hidden potential when it comes to Big Data that to many users is yet to be unlocked. In banks and similar organizations, Big Data can help in identifying top performers and if used in right way, it also can help to enhance and boost employee's performance ratio. This is applicable when using right tools such as analytics and organizations can track their employee's performances, analyze the information and share individual performance, team spirit and overall company culture.

As it is stated, Big Data can help BFSI organizations to a better performance and profibility internally and externally as well as with their customers and clients in providing increased quality services with less operational costs. (Akhter, 2020)

3.6 Big Data in Banking Industry

The global financial services industry makes huge amounts of different types of data every day, caused by processing billions and billions of financial transactions as well as interactions such as email communications, weblogs, call logs and etc.

One compelling driver of this data usage explosion is the increase in payment volumes, caused by ecommerce and mobile payments. Before the global COVID 19 pandemic, which has caused yet unknown downturn in global payment market, it was anticipated that the market will reach to more that 2$ trillion by the end of 2025, with the annual rate of 7,83%. E-commerce's continuous growth has been dramatical due to the fact most of the consumers are doing their shopping online than in person because of the caused global pandemic. The other factors are high usage of peer-to-peer payments through apps, paperless processing documents and ATM usages which are increasing in popularity.

With the rise of technology, becoming a popular platform for information access and commerce, the amount of data that banks are producing is mind-blogging. This has caused different IT and business challenges in managing all this data. However, it has also created opportunities for banking industry to thrive their business and improve their operational efficiencies. (Chalimov, 2019)

Appliance of analytic solutions invigorated by the cloud, AI and more importantly machine learning, banks have the advantage of leveraging their data, which has deemed impossible previously. This level of insight added to every part and sections of their business has enabled them to understand what, why and how of an occurrence that happened in the past. This analysis will lead to next big question: What do the consumers (customers) want?"

When banks first realized and understood the importance of Big Data, the first use cases arose to generate insightful aspects in different sectors like: fraud detection, transaction processing optimization, customer understanding improvement, trade execution and ultimately superior customer satisfaction and experience deliverance. Nonetheless, gathering more data has led to more accurate results and operations. One of the biggest examples of this improvement is when **Western Union** started offering and omnichannel approach exclusively for personalizing customer experiences by processing over 29 transactions in a second and integrating all that generated data into a single platform for statistical analysis and modelling. (Ku, 2020)

Nevertheless, the challenges have altered to more data engineering technology evolution, even though big data use cases in banking have remained the same till today.

There are three challenges to consider:

1. Speed

Altering from traditional data warehousing to Hadoop, with its densely parallel engine on commodity hardware has allowed banks to shorten the length time needed for exctracting information form the data from two months to day or less. Cloud-based data processing adoption has reduced the timeframe needed even more and further. But, banks are still prone to process data in batches monthly. (Ku, 2020)

2. Management

Hadoop or Apache Spark can shift massive volumes and varieties of data in a way than they can be then driven to cloud data warehouse, where business users will have access to it. But this does not mean that data is fit to use. Neither Hadoop and Apache Spark perform or manage data natively, so its useless unless they could understand what it means or how it is used. On top of that, they do not provide data lineage which does not show their data transformation. (Ku, 2020)

3. Modular Technology

Banks may have solved the issues of costs or infrastructure management of big data analytics by altering data processing to the cloud from on-premises hardware, but when a local and a multinational bank have the same access to a managed service provider, the processing of massive volumes of data will no longer be the difference from a competitive perspective. Banks should have the ability to transform their data faster into useful insights and immediately put gathered information into action for customer service improvement or data protection against threats. (Ku, 2020)

3.7 Missing data

Even in a meticulously designed and controlled environment, missing data can occur in almost every aspect. Missing data or also known as missing value is the data value that is not properly stored and is rather common, presenting different problems. First, the absence of data leads to reduction of statistical power, which is assigned to the probability of rejecting the null hypothesis when it is false. Secondly, the lost data can cause distortion in the estimation of parameters. Thirdly, it can cause reduction in representativeness of the sample data. Lastly, it can lead to complications of the analysis. These can greatly harm the validity of the tests and their conclusions.

Primarily, there are three types of missing data:

1. Missing Completely At Random (MCAR)

Missing completely at random is characterized when the probability that the data is missing is not related to either the specific value which is supposed to be obtained or the set of obsevered responses. (Kang, 2013)

MCAR is arbitrary assumption as well as ideal. If the data is missing due to equipment failure or being lost in transit, they can be regarded as MCAR. The advantage to the missing data is statistical for the analysis remains unbiased. (Kang, 2013)

2. Missing At Random (MAR)

Missing at random is a more rational assumption. Data can be categorized MAR when the the probability that the responses are missing depends on the set of observed responses but is not related to the specific missing values which is expected to be obtained. (Kang, 2013)
Randomness could be considered as not producing bias, but we may also assume MAR does not pose any problem. MAR does not suggest ignorance of missing data. (Kang, 2013)

3. Missing Not At Random (MNAR)

If the missing data would not fall under MCAR or MAR category, then missing data can be considered as missing not at random.
MNAR are problematic cases and the only way to gather an unbiased estimation of the parameters is to model the missing data. (Kang, 2013)

4 Practical Part

As for the practical portion of the thesis, a case company will be introduced as a form of evaluating the results of a proposed big data solution. The company will be a commercial bank called ‘’Mobank’’. In this part of the paper, the issue of missing data and one specific solution for this problem will be proposed and evaluated, taking into account the primary costs of not investing in the proposed solution and the costs of putting the proposed solution into practice.

As an addition, the effectiveness of the proposed solution will be taken into consideration in order to fully calculate whether or not this is a solution worth investing in.

The properties below are the information needed to conduct the experiment:

Industry: Financial services

Services: Banking, Financing

Founded: 04, 1997 in Frankfurt, Germany

Revenue: 1.2 billion €

Operating income: 200 million €

Net income: 80 million €

Total assets: 10 billion €

Total equity: 4 billion €

Number of daily transactions: 30,000

Number of Employees in Data management department: apx 500 (80 of them are machine learning eng.)

Average salary for data management employees: € 132,000

4.1 Issue introduction

It is a common issue within Mobank’s Data Analytics Department for data to be inconsistent, that includes many different problems, such as duplicate data, outdated data, inaccurate data and so on. However, the problem to be stressed in the practical part of this paper will be missing data.

Missing data refers to a situation in which no data value is stockpiled for a variable. This a frequent problem to be dealt with which has the capacity to drastically affect the main outcomes in data. The type of missing data taken into account on this analysis is classified as MCAR (Missing Completely at Random), which indicates the lack of connection between the missingness of data and the values.

4.2 Heuristic Machine Learning Imputation (HMLI)

The solution for missing data here will be dealt with by means of automation, through increasing robotic capacity with AI, so as to allow for machines to make value judgments. More specifically through the usage of Heuristic Machine Learning Imputation (HMLI).

HMLI is a form of imputation. It functions through the use of inserting all missing values with zero, mean or median for quantitative variables, or the most repetitive value for categorical value.

Heuristic Machine Learning Imputation is a new imputation program which uses the evolutionary process and machine learning algorithm. HMLI introduces a genetic algorithm, one of the heuristic search techniques, to locate optimally dependent variables. This algorithm mimics the normal evolutionary mechanism, so humans expect it to be able to find one of the better dependent variables set by iteration. Once this algorithm chooses the control variables, the model reverts the dependent variables to the independent variable and estimates the missing values.

The HMLI regression system is the support vector machine (SVM). SVM is one of the most successful machine learning techniques used since its invention in the 1990s, mainly for pattern recognition. The HMLI model has two distinctions with the standard regression model; contingent set of variables and regression line fitting. In the standard regression model, researchers typically choose dependent variables to describe the framework itself statistically and accurately estimate the independent variable. But it is difficult to find the most dependent variables due to a number of factors. If computer algorithms ought to find an optimal mix of dependent variables and estimate independent variables with a low error rate, that can be the ideal solution for imputation.

Heuristic search is a rule of thumb strategy to find an answer quicker. When there is a minimal conventional solution. For example, the total number of variations to pick 6 time series of 10,000 time series is approximately 13×1020. It's too large to measure all the combinations in a short time. In this case, heuristic check may be a way to approximate the exact solution.

This heuristic search and machine learning mix repeats dependent variable collection, regression line fitting, incomplete observation estimation, and root mean square error

(RMSE) analysis of real and predictive values. The termination criterion of the recurrence can be either the RMSE satisfaction or the iteration number.

The HMLI computation method consists of nine phases. HMLI implements an evolutionary mechanism by iterating these moves. This programming iteration uses a lot of computational power.

The first step revolves around Data pre-process. If the time series value range is broad, the SVM objective function may not operate properly and the measurement efficiency of the normalization values is much faster. Normalization is required for both independent and dependent variables. When the normalization is completed, the holes in the independent variable tend to be applied. This additional gap eliminates the real values in the time series which can be used as test values. To determine which dates are time series gaps, the model uses the exponential distribution and the λ is a missing rate.

In the second step, called train and test data separation, the real values, replaced with gaps will be utilized to test values. After the HMLI model creates those prediction values, the RMSE between test values and predicted values is a factor about performance of dependent values. Except testing values in dependent variables, other values are used as train data.

As for the third step, wic refers to sampling dependent variables, where a random sampling process chooses a particular number of dependent variables from total variables.

The fourth step (regression) refers to SVM calculating the best fitting curve through train data. By putting test data into this fitting curve, SVM generates predicted values. After that, for the fifth step, RMSE calculation is done, it calculates RMSE between prediction values from SVM fitting curve and actual values replaced with gaps on step 1. Low RMSE means that dependent variables are good to predict gaps.

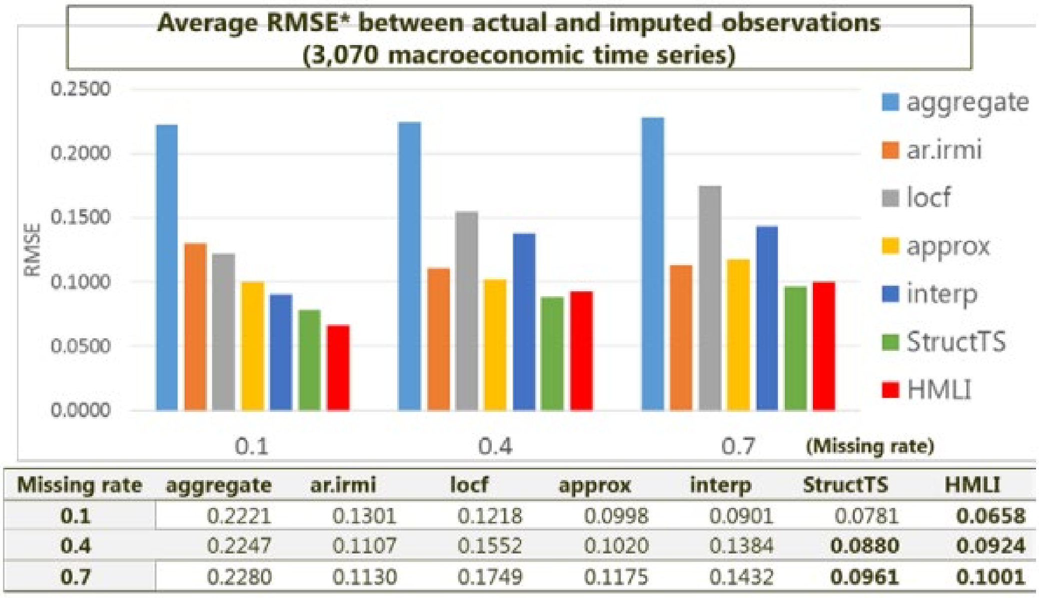
For step six, called the selection process, about 10 sets of dependent variables, this step calculates RMSE and ranks the sets. Based on this rank, it removes 5 lower ranked sets. The next step refers to "dependent variables crossover", which extracts unique variables from top 5 ranked sets and generates 2 variables sets. If it cannot generate new dependent variables, skip this process.

As for step 8, called additional dependent variables sampling: During the next iteration, 10 dependent variables sets are required. Including top 5 ranked sets and 2 crossover sets, then it creates extra sets until it becomes 10 sets.

As for the last step, called repetition, it repeats n times from step 4 to step 8, in the paper used, number n sets up 100. It is mentioned that iteration ends when it satisfies a particular condition. An example given is, if RMSE is lower than a particular value, it halts the iteration. (Kwon,2018)

4.3 Effectiveness

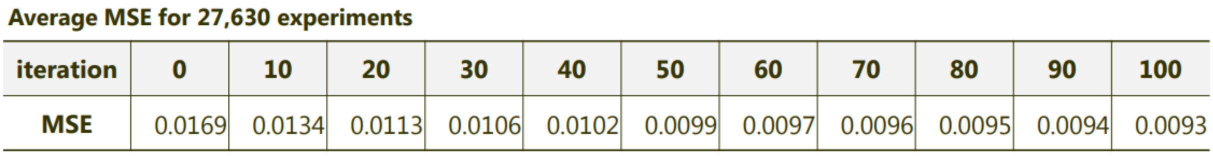
The effectiveness of HMLI was demonstrated by comparison with other 6 imputation methods (locf, approx, interp, StructTS, ar.irmi, aggregate). In the study mentioned, the Root Mean Square Error (RMSE) of all imputation methods mentioned were juxtaposed. (Kwon, 2018). The paper utilized zoo libraries and forecast in R due to the popularity of these libraries to pre-process and allow for time series data analysis. It is possible to demonstrate the comparison by the graphic representation below:



Missing rate	aggregate	ar.irmi	locf	approx	interp	StructTS	HMLI
0.1	0.2221	0.1301	0.1218	0.0998	0.0901	0.0781	**0.0658**
0.4	0.2247	0.1107	0.1552	0.1020	0.1384	**0.0880**	**0.0924**
0.7	0.2280	0.1130	0.1749	0.1175	0.1432	**0.0961**	**0.1001**

Figure 3: Average MSE Plot For Macroeconomics Time Series and 3 Missing Rates (Kwon, 2018)

Below it is also represented the average mean square error for all the 27,630 experiments



Average MSE for 27,630 experiments

iteration	0	10	20	30	40	50	60	70	80	90	100
MSE	0.0169	0.0134	0.0113	0.0106	0.0102	0.0099	0.0097	0.0096	0.0095	0.0094	0.0093

Figure 4: Average MSE For 27,630 Experiments (Kwon, 2018)

A few basic observations can be made: When the missing rate is 0.1, HMLI's RMSE is 0.0658. When the missing rate is 0.4, HMLI's RMSE is 0.0924. When the missing rate is 0.7, HMLI's RMSE is 0.1001. As observed, HMLI presents itself as the most effective solution in terms of dealing with missing data at a 0.1 missing rate. And HMLI ranks second in effectiveness at missing rates 0.4 and 0.7, only not being as efficient as StructTS. In matters of effectiveness, HMLI proves itself to be a top solution.

4.4 Calculation

In order to evaluate the costs of implementing this solution, several pieces of data were gathered. It was necessary to first estimate the company's properties, as it was done on the previous sections. Four costs were taken into consideration, those were divided into costs of not investing in a solution (productivity loss and loss in total revenue) and costs of investing in HMLI (R&D costs and technology costs). Each of those costs will be individually assessed.

First, it was pointed out by Gartner that not tackling the problem of missing data solutions leads to an average 50% productivity loss. (Gartner, 2012) In financial terms, that is translated as half of the hourly wages of those who work in the data management department. Mobank has 500 employees in the department and their average salary is €132,000 per year. The loss in productivity can be quantified by half of the salary of those employees, which had 50% of their time lost in dealing with the problems that arise due to missing data. That is a loss of €33,000,000 in total. €66,000 per employee. This is a constant variable, since the loss in productivity does not vary in this experiment. (Wisconsin, 2020)

Second, the loss in revenue was estimated by MIT Sloan Management Review to be between 15%-25%. Thus both 15% and 25% estimatives will be taken into consideration and calculated. (Redman, 2017, Finextra, 2019)

However, the research utilized in order to obtain the costs, was made on dirty data and its costs, rather than specifically missing data, thus it is necessary to stipulate different situations in which missing data will correspond to 100%, 70%, 30%, 14% and 5% of total dirty data. This was made for two reasons. The first one being the lack of data on the common amount of missing data that is part of total dirty data in companies. The second one being a way to verify if the costs remain plausible regardless of the percentage of missing data that a company has to deal with as part of their total dirty data problems. It is possible to see each cost on the tables below:

X% of dirty data is Missing Data

	Loss in Productivity	**Loss in Revenue**	**Loss in Revenue**	**Total Loss**	**Total Loss**
Percentage Loss		**15**	**25**	**15**	**25**
Monetary Loss (million EUR)	33 Million EUR (Constantvariable:salary doesn't change)	X% of 180 Million EUR	X% of 300 Million EUR	Loss in Productivity + Loss in revenue 15%	Loss in Productivity + Loss in revenue 25%

Table 1: Formulas For Calculations (author)

Losses for different scenarios based on percentage of missing data (million EUR)

Percentage of dirty data	**Productivity Losses**	**Revenue Losses Estimated impact**		**Total Losses Estimated impact**	
		15%	**25%**	**15%**	**25%**
100%	**33**	**180**	**300**	**213**	**333**
70%	**33**	**126**	**210**	**159**	**243**
30%	**33**	**54**	**90**	**87**	**123**
14%	**33**	**25**	**42**	**58**	**75**
5%	**33**	**9**	**15**	**42**	**48**

Table 2: Losses of Missing Data (author)

As for the costs of investing in HMLI, the main two chosen were R&D costs, as well as technological costs, considering both the complexity of HMLI and the fact much computer power is needed in order to run it.

On the determination of R&D costs, it was estimated by the digital marketing solution WebFX that the average value for it lies between €6000 to over €300,000; That is a huge variation, but for the sake of attempting to compare costs between investing and not investing in the HMLI solution, it will estimated that R&D costs correspond to the median value of €153,000, as well as the highest value (€300,000) .

As for technology costs. It is mentioned in Kwons's paper that much computer power is necessary for the use of a solution such as HMLI. It was estimated based on this fact that high performance computers would be necessary. As of January 2021, this would be a model such as the Asus ROG Mothership, and the price for a unit is on average € 6000. Considering the number of employees working with the software, that would be translated as 500 units of high performance computers, which corresponds to € 3 Million. Along with internet costs, that realistically, in a city as Prague, can be estimated as € 23 per person, corresponding to € 11 500 for 500 employees.

5 Results and Discussion

The total cost of investing in HMLI tends to be € 3 311 500, which corresponds to highest R&D costs (€300,000) added to the technology costs (€ 3 000 000) added to the internet costs (€ 11 500).

For the most expensive loss associated with not investing in a solution for missing data, that is, when missing data represents 100% of total dirty data and a company has 25% revenue loss, it's possible to verify that Mobank would have a loss of € 333 000 000. Whereas for the situation with the lowest losses associated with not investing in a solution for missing data, that is, in a situation when only 5% of total dirty data is missing data, and when the company has a 15% loss of revenue, the total loss is € 10 000 000. The cost of investing in the solution (€ 3 311 500) is still less costly than not doing anything, thus, in any of the situations provided, Mobank would always benefit from investing in the HMLI solution.

The other examples which can be given are the situations when missing data is 14% out of all dirty data in the company, which represents a loss of € 75 000 000 in the worst case scenario. As for a situation when missing data counts as 30% of all dirty data, the Company has a potential total loss of € 123 000 000. Whereas for a scenario when missing data corresponds to 70% of dirty data, the maximum loss is € 243 000 000. As previously stated, Mobank benefits immensely from taking the step to fix this issue by means of applying the HMLI solution.

A few objections can be made about this research and its results. On the matter of accuracy, it was difficult to gather an extensive amount of data. The space for variation when it comes to the prices and costs presented on the paper is sizable. It is nonetheless true that the data gathered on the paper does represent mostly an average of those prices, although, even considering averages, there is a margin for error, since those averages were mostly taken in a particular country or city rather than a collection of countries and cities. In order to fix that, it would have been important to gather data which represents a wide variety of countries and towns.

When it comes to comparing the HMLI to the costs and effectiveness, this was not done. It is possible that HMLI is not the most optimal solution, since there hasn't been studies on the comparison between costs and benefits of the missing data solutions discussed in this paper, such as StructTS, interp, locf, approx and others. However, the paper focuses on finding the costs and effectiveness of HMLI by itself, rather than comparing it to other solutions, and as already stated, investing in a HMLI solution is beneficial regardless of the percentage of dirty data which corresponds to missing data (unless the percentage is zero).

This paper quantified the effectiveness and costs of the HMLI solution for Missing Data in banking. It presented how does the HMLI model works, why is it more effective to deal with missing data compared to most other main models, and how investing in this solution barely costs anything in comparison to not investing in it, almost always invariably regardless of the amount of missing data as a percentage of total dirty data in the company.

The main limitations of this research stem from the fact that the quantities estimated are very variable depending on the country, city or region. For example, the prices for the internet costs were estimated based in the city of Prague, but that does not necessarily apply to other cities or countries. Another instance has to do with a few values that are difficult to measure, such as the R&D costs to deal and implement the HMLI solution.

6 Conclusion

This paper quantified the effectiveness and costs of the HMLI solution for Missing Data in banking. It presented how does the HMLI model works, why is it more effective when dealing with missing data compared to most other main models, and how investing in this solution does not cost as much as not investing in it, almost always invariably regardless of the amount of missing data as a percentage of total dirty data in the company.

The main limitations of this research are stem from the fact that the quantities estimated are very variable depending on the country, city or region. For example, the prices for the internet costs were estimated based in the city of Prague, but that does not necessarily apply to other cities or countries. Another instance has to do with a few values that are difficult to measure, such as the R&D costs to deal and implement the HMLI solution.

It is also true that only the financial costs of HMLI were taken into consideration, the other models mentioned and compared to HMLI when it comes to effectiveness, were not compared financially to HMLI. That was due to a personal choice to only consider the investment in HMLI in particular, because it is a relatively new system and it often outperforms other models when it comes to effectiveness.

The main goal was to compare the costs and benefits of applying the HMLI model as a solution to missing data. Considering all calculations made, it is possible to affirm that Mobank would immensely benefit from applying this solution into their system, since investing in HMLI seems to be always more profitable than not doing so.

7 References

RAJKUMAR, J. BUYYA, K. CALHEIROS,R. DASTJERDI, A. *Big Data: Principles and Paradigms* : Todd Green, 2016. page 4-10.

Gartner. Big Data. *gartner.com.* [Online] 2012. http://www.gartner.com/it-glossary/bigdata. Accessed November 2020.

KRISHNAN, Krish. *Data Warehousing in the Age of Big Data* : Morgan Kaufmann, 2013. page 219-240.

Kranz, Garry. Big Data Storage. *techtarget.com.* [Online] December 2017. https://searchstorage.techtarget.com/definition/big-data-storage. Accessed December 2020

Chalimov, Alexey. Big Data in the Banking Industry: The Main Challenges and Use Cases. *easternpeak.com.* [Online] 2019. https://easternpeak.com/blog/big-data-in-the-banking-industry-the-main-challenges-and-use-cases/. Accessed November 2020

Lochy, Joris. Big Data in the Financial Services Industry - From Data to Insights. *finextra.com.* [Online] 2019. https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights. Accessed January 2021

Akhter, Aamir. Will Big Data Be a Game Changer for the Banking and Finance Sector? *readwrite.com.* [Online] 2020. https://readwrite.com/2020/04/28/will-big-data-be-a-game-changer-for-the-banking-and-finance-sector/. Accessed November 2020

Ku, Peter. Big Data in Banking: Use Cases in 2020 and Beyond. *blog.informatica.com.* [Online] 2020. https://blogs.informatica.com/2020/04/13/big-data-banking-use-cases/. Accessed November 2020

Kang, Hyun. The Prevention and Handling of the Missing Data. *ncbi.nlm.nih.gov.* [Online] 2013. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/. Accessed December 2020

OMNI.SCI. Big Data Architecture. *omnisci.com.* [Online] 2020. https://www.omnisci.com/technical-glossary/big-data-architecture. Accessed December 2020

Marr, Bernard. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *forbes.com.* [Online] 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read. Accessed January 2021

Wisconsin, University of. How Much Is a Data Scientist's Salary? *University of Wisconsin Data Science.* [Online] 2020. https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/. Accessed December 2020

KWON, Byeungchun. *Imputation for missing data through artificial intelligence.* Basel : Irving Fisher Committee on Central Bank Statistics, 2018.

Finextra, Research. Assessing the Financial Cost of Poor Data in Banking. *Finextra Research.* [Online] 2019. https://www.finextra.com/blogposting/17366/assessing-the-financial-cost-of-poor-data-in-banking. Accessed December 2020

Redman, Thomas C. Seizing Opportunity in Data Quality. *MIT Sloan Management Review.* [Online] 2017. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/. Accessed December 2020

Redman, Thomas C. Data's Credibility Problem. *Experian study.* [Online] 2013. https://hbr.org/2013/12/datas-credibility-problem/. Accessed January 2021