

Big Data model and warehouse for Pestron Foods

Gergely Somogyi
Student Number: 1814098
Big Data (ACUA6002_LON)
Cloud Computing BSC
University of Wales Trinity Saint David
London, United Kingdom
1814098@student.uwtsd.ac.uk

Abstract—Companies nowadays face a data mass. Trends are shown this amount of data cannot be handle by former tools and solutions. Recently more and more firms started to use Big Data methods, approaching the data issues different way or using more effective the available data's. Based on literature review these tools are not an ease to implement in a business's life, but rather worth it. The result makes company workflow and decision making quicker and easier, evolve the effectiveness in all levels in the corporation and open new paths what were inconceivable before. My aim to brief Big Data and give some samples what it could done with data's.

Keywords—*Big Data, Data Warehouse, Machine Learning*

I. INTRODUCTION

My goals with this report are dual. First, I show Big Data concept in broad lines. What it is mean, how it started, what component and changes makes Big Data born? I try to make clear how Big Data could help business, life's, and sciences. What benefits they can earn by using Big Data tools. Then, some of the methods will discussed in detail indeed. Origin of these methods are diversified. It is a mash and the ingredients coming from Mathematics and from Computer Science.

One of the reasons for this is that the quantity and quality of data traffic on the Internet has also improved, more and more and more data can be known about, for example, the visitors of our website, digital services. Countless companies, from banks through the energy sector to the automotive industry, have been able to obtain previously unknown amounts of data about their own activities. For so much data that has already been a qualitative leap from previous eras. Above a certain amount of data, it is possible to get to know a given industrial process, digital service, or even human behaviour patterns so thoroughly that it is already possible to formulate predictions and predictions with good efficiency. That's why Big Data has opened a whole

new era for design, medicine, software development and even marketing.

My second goal to represent what Big Data approach could give a company such Pestron Foods. The discussed methods in previous section, demonstrate in this section in practice. These ways give a lot of opportunity and benefits for Pestron. But it could make some difficulties and issues for a business. If there is a lot of data, it is much harder to work with. These problems from everyday life: it needs a lot of storage space, it takes longer to evaluate, a search runs slowly, difficult to share and copy, complex task of any editing. General arrangement that, it is difficult to make changes and to maintain data security.

Big Data is very often not a fixed database, but a series of data that is continuously produced from a certain source. So, from this database not for taking samples, but observing for continuously to gain real results and values. By the time this result getting more and more reliable provide better information to understand better the company's life cycles, lighten up connection between events, make decision making easier in all levels of workflow.

New times bring new challenges what needs new approaching with better tools. Big Data do not answer all questions yet, do not solve all questions and issues, but a well-design model could deliver a better functional business flow, where all data's have relevance to gain better performance.

II. LITERATURE REVIEW

Before we discuss any methods, see what Big Data is meaning. As it is a relatively new era of IT – and as many IT related concept-, there is not exist an all-round accepted definition.

A. Definition of Big Data

According to Gartner IT Glossary: “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.” [1].

Now look how is defined the Big Data by Oracle, one of the main providers:” The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

- Volume.
- Velocity
- Variety

Two more Vs have emerged over the past few years: value and veracity. Data has intrinsic value. But it’s of no use until that value is discovered. Equally important: How truthful is your data—and how much can you rely on it?

Today, big data has become capital. Think of some of the world’s biggest tech companies. A large part of the value they offer comes from their data, which they’re constantly analysing to produce more efficiency and develop new products.

Recent technological breakthroughs have exponentially reduced the cost of data storage and compute, making it easier and less expensive to store more data than ever before. With an increased volume of big data now cheaper and more accessible, you can make more accurate and precise business decisions.

Finding value in big data isn’t only about analysing it (which is a whole other benefit). It’s an entire discovery process that requires insightful analysts, business users, and executives who ask the right questions, recognize patterns, make informed assumptions, and predict behaviour.” [2].

Looking these two definitions above, Big Data is very difficult to define. It is rather a process than a term. But every process has a beginning.

B. History of Big Data

Since about 2012, we have seen the concept of “big data” appear more and more in both the Internet and the traditional press. There is no doubt that this has become one of the most fashionable IT buzzwords these days, like Cloud Computing (2009-), which appeared in previous years. (See Appendix Fig. 1. and Fig. 2. to watch Google Trends data about search term comparing “Cloud Computing” and “Big Data” in Computer and Electronics topic in the UK from 2004 to today and the last 5 years.)

The history of Big Data started around 2003. It was before Google published the GFS (Google Distributed File System). An open-source search engine, what handling billions of webpages, has a very high operational cost was due to the huge size of the files generated by the web search and indexing program, which were difficult to manage in traditional file systems or databases. The difficulty and high cost resulted from the fact that the scalability of the technology at that time was already cumbersome and expensive in this size range. The GFS was a good solution for this issue. In 2006 the Apache announced the Hadoop project, what became the foundation of many search engines project later. Basically, Big Data started in response to the problem of technological scaling in the management of Hadoop - as the first big data technology available to a wide range of people -it became so successful because it can be scaled almost linearly up to the order of Petabyte. Then for a long time this was the main field of application of big data.

However, popularity required something else. It turned out, that there are many different analytical companies who have needs for Big Data which they have not been found so far adequate capacity. They not stored data due of high storage cost or although the data were stored but could not be used because computing capacity required for processing was not available. For a medium or small enterprises even a few terabytes of data should be a problem to store and analyse. By Hadoop (as the first Big Data tool) showed up more and more data began to load into the Hadoop cluster, which had previously was out of reach for analysis.

Due of open-source roots and the wide range of software of varying levels, scope, and functionality developed by many smaller companies or development teams. The kind of “do it yourself” associated with big data. An approach where users select the various hardware and software components and integrate them into a solution. That's why early appeared companies who offered

integrated Big Data software solutions to users (e.g., Pivotal, Cloudera, Hortonworks) or integrated Big Data hardware solution (NetApp, EMC).

With the arrivals of increasingly advanced solutions not only internet companies interested in big data, but large companies in more traditional market segments also, such commercial, financial and production or research companies, whose main driving force was a competitive advantage by analysing stored (or not even stored) data assets. These companies typically use the solutions of the largest IT companies, so soon the largest IT companies began to develop their own solutions (Oracle, IBM, SAP, Amazon Web Services), which were developed in addition to hardware and software level integration, providing other functionalities. They are common to other software and progressively required in large enterprise environments: security, manageability, integration [5].

C. Big Data principles

As Big Data's name foreshadow, data has an important role. Beyond the three "V" known in section "A", there are three main types of data:

- **Structured data:** All data that can be stored, viewed, and processed in a fixed format is referred to as structured data. By the time, IT experts have been successful developing techniques for working with such data (where the format is known in advance). It managed to make this profitable for the businesses. Today, data management are facing problems due to the volume of such data sets. A high growth rate sometimes leads to the size of more Zettabytes. Structured data in big data is the most straightforward to work with. Structured data is a type of big data that profoundly coordinated with measurements described by setting parameters.
- **Unstructured data:** Data sets of unknown shape or structure are subject to unstructured data. In addition to being huge in size, unstructured data leads to multiple challenges in data management in terms of gaining data added value. A typical example of unstructured data is a heterogeneous data source that is simple text files, images, videos, and so on. Contains a combination. There are organizations that have large amounts of data but unfortunately do not know how to use that data for them. This data is in raw or unstructured format. We also call for

such data search and repair data mining. An example of unstructured data is the search results provided by "Google Search".

- **Semi-structured data:** Semi-structured data can contain both data formats. Semi-structured data can be considered structured but are not defined, such as the definition of a table in the relational form DBMS. An example of semi-structured data is an XML file.

The name Big Data refers to the size itself. The size of the data plays a crucial role in determining the value derived from the data. The amount of data determines whether the data can really be considered Big Data. Thus, Volume is always a characteristic that we need to consider when dealing with Big Data.

The next aspect of Big Data is diversity. Diversity refers to heterogeneous sources and the nature of data, which can be structured or unstructured. Custom spreadsheets and databases are the only form of storage that most applications consider. We now also store data in emails, photos, videos, management tools, PDFs, audio, and more. In the form. next. These forms of storage need to be considered in analytical applications. The diversity of unstructured data poses certain problems:

- Storage (data management)
- Data mining
- Data analysis

The term velocity refers to the speed at which data is created. Big Data velocity refers to the speed at which data can be retrieved from sources such as business processes, application logs, networks and social media sites, sensors, mobile devices, and more. coming. These streams are huge and constant. The speed of data generation and processing determines the true potential of the data.

Variety indicates how we handle the inconsistencies that sometimes occur in the data, that the process can process, control, and filter the data efficiently.

To use Big Data tools needs data. Usually, these data's produced day-by-day, like recipes, orders, log information, GPS information. The first process what happening with data is data mining. Data mining is a process used by companies to turn raw data into actionable information. They use special software to search for samples in large amounts of data. This will allow companies to learn more about their customers and develop more effective marketing strategies. The goal is to increase sales

and reduce costs. Data mining is possible through efficient data collection, storage, and high-performance computers.

Supermarkets are known users of data mining techniques. Many supermarkets offer bonus cards or Airmiles to customers. With these loyalty programs, customers are entitled to discounts that are not available to members. With such a system, stores can easily keep track of who buys what, when, and at what price. Stores can then use this data for several purposes. Consider offering discount coupons tailored to your shopping behaviour. Another use is to decide whether the goods will be sold at sale or at full price. Data mining can provide false information that is not representative of the entire sample group. Note if this information is used to support a particular hypothesis [3], [4], [5].

D. Data warehouse

When companies centralize their data in a single database or software solution, this is called data warehousing. A data warehouse allows an organization to analyse segments of data for specific users. In other cases, however, analysts start with the type of data they are looking for and create a data warehouse based on data analysis methods. Although companies systematize their data, they use it to support management decision-making processes. The results can also be used as input to business cases.

Data Storage and Big Data are both sources for decision making. And in both cases, it requires a lot of data. May it confuse, but data warehouse and big data are not the same. In a data warehouse, stored data is usually stored in a structured way in a database. For Big Data, this is often unstructured. Moreover, Big Data is usually stored in the cloud. These are two factors that do not make data gather easier.

Data mining tools analyse data relationships and patterns based on user requests. For example, Data Mining software can be used to create information classes. Imagine a restaurant using data mining to determine when to add certain promotions to the menu. The software examines the information collected and creates classes based on bookings and orders. In other cases, Data Miners find information clusters based on logical relationships. They also examine associations and sequential patterns to draw conclusions about trends in consumer behaviour.

The data mining process consists of six steps. First, organizations collect data and upload it to their

data warehouse. Then saves the data to internal servers or to the cloud. Business analysts, management teams, and IT professionals have access to the data. They also determine how they want to organize this. The data mining software then classifies the data based on the user's request. Finally, the end user displays the data in an accessible representation, such as a graph or table.

A data warehouse is a large store of all the information in an organization, including all previous data. A data warehouse is a complex process that involves creating a data warehouse in the form of a relational database so that a company can use the data to use, transform, or aggregate useful information. In all cases, organizations gain a competitive advantage through data storage and support decision-making processes through extensive data analysis. Some of the most important components of data warehousing are Decision Support Systems (DSS) and Data Mining (DM). The amount of data in data warehouses can grow exponentially, so you need to find a way to handle this huge increase. The need for storage is one of the things we need to take seriously in a data warehouse. High availability, large amount of data, high performance, scalability, and ease of use play a big role in this.

The data storage is the logical division of data. It contains several databases that are used for an application. For example, different databases (structure, costs) supporting financial applications can be placed in one financial repository. Dividing data into logical or, in some cases, physical Storage can greatly contribute to meeting the requirement for exponential growth in data volumes in the data warehouse. If all the data in the data warehouse is not divided into different data warehouses, the performance and efficiency of the data warehouse will remain below the nominal value.

Another point of attention for large processing operations has always been the central server. If this central server fails, the system shuts down. This was because all the data was in a non-redundant system and if the hardware failed, there was no backup. Today, even a few minutes of downtime in a commercial business can result in a loss of thousands of dollars. The data warehouse can also provide an answer here. Data warehouse should segment into smaller data warehouse, this method distributes the processing to many databases and multiple servers. Instead of a database processing customer data, different databases can handle different aspects of a customer at the same time. This results in redundancy and allows for quick recovery.

Different databases may be active in companies with different locations. Instead of all the clients in a single database, there are different client databases that work with the same storage. This is called a distributed database. The data storage provides easier and faster access as related information is blurred or grouped.

Due to its distributed nature, the storage is a simpler way for database managers to maintain a data warehouse system. If there is a problem with the system, we can easily diagnose the problem without requiring a top-down approach to the entire data warehouse. Most organizations usually designate a database manager to guarantee the reliability of the data for the entire system.

Data Warehouse and Big Data are mismatched. In both cases, there is really a lot of data. To properly search for data, we often set up a shadow system for the data warehouse. Searches do not interfere with daily work. However, the nature of the data is completely different for both concepts. As the above suggests, a warehouse is about data structured in relational databases. Big data is usually unstructured and is used by marketing companies in market analysis. In addition, we often associate the Big Data service with cloud solutions. However, this does not rule out the possibility that large data may include relational data [3], [4], [5].

E. Machine Learning

First, we clarify what machine learning is: There is a set of data whose data content can have certain logical connections. These relationships can be modelled on a static data set, however, if the content of the data set is constantly changing, machine learning solutions enter the picture, the essence of which is that these relationships are constantly reinterpreted in the light of ever-changing data.

Basically, machine learning models can be divided into two major groups: supervised and unsupervised models. The essence of the former (supervision) is to know exactly what is considered good and to examine this using either regression or grading methods. When the examined output is discrete (so it can be decided whether it is good or not good), we use classification algorithms, if the output is continuous, then regression.

And the data that can be analysed must be properly prepared. Most machine learning (ML) models expect this to be a set of data, which if not properly generated, the ML model will make mistakes and draw erroneous conclusions.

These call the data preparation process data pre-processing. The main steps are:

- Dataset production
- Missing data identification
- Categorical variable
- Training sets and testing sets
- Standardization and normalization

There are three main machine learning type:

- Supervised Learning: For datasets with tags or structure, the data trains the machine so it can make predictions and make decisions more efficiently.
- Unsupervised Learning: By grouping data into clusters, you can more easily handle untagged or unstructured data and identify patterns and relationships.
- Reinforcement Learning: A computer program acting on behalf of human, replaces the operator and helps determine the result based on a feedback loop [6].

III. METHODS

This chapter is about the known and useful methods, what belongs to Big Data. As mentioned before, Big Data is rather a process, so see methods in this approach.

- Data Preparation
 - Feature Selection
 - Selecting the top-ranked features
 - Selecting the top-ranked subset
 - Forward selection
 - Backward elimination
 - Dimensionality Reduction
 - Principal Component Analysis (PCA)
 - Linear discriminant Analysis (LDA)
 - Multidimensional Scaling (MDS)
 - t-distributed Stochastic Neighbour Embedding (t-SNE)
 - Record Selection and Sampling
 - Instance selection

- Support Vector Machines
 - Clustering extension
 - Missing Values Imputation
 - Missing Completely at Random (MCAR)
 - Missing At Random (MAR)
 - Not Missing at Random (NMAR)
 - Outlier Detection
 - Knowledge-based
 - Statistics-based
 - Quantile-based: Box plot
 - Distribution-based: Z-Score
 - Cluster-based: DBSCAN
 - Neural Autoencoder
 - Isolation Forest
 - Data Cleaning and Standardization - From Numerical to Categorical
 - Equi-width discretization
 - Equi-frequency discretization
 - V-optimal discretization
 - Minimal entropy discretization
 - Normalization
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
 - Data Engineering
 - Text data analysis
 - Time series data analysis
 - Image data analysis
 - Graph data analysis
 - Data Integration
 - Vertical Data Integration
 - Horizontal Data Integration
- ETL: Extraction, Transformation Loading
- Data Understanding
- Describing your Data
 - Basic Statistics
 - 1D: Histograms
 - 2D: Scatterplots, Scatter Matrix, Multi-Dimensional Scaling
 - 3D Scatterplots
 - 3D: Parallel Coordinate
- Finding Patterns
- Finding Models
- Finding Predictor
 - Brute Force Predictors
 - Decision Trees
 - (Naïve) Bayes Classifiers
 - Regression
 - (Artificial) Neural Networks
 - Support Vector Machines (Kernel Methods)
- Machine Learning
 - Basic strategies for ML algorithms
 - Close form solutions
 - Gradient descent
 - Search strategies
 - Model validation
 - Types of model errors
 - Strategies for more reliable model evaluation
 - Cross-validation
 - Bootstrapping
 - Coping with unbalanced datasets
 - Measures for model complexity
 - Error measures to validate models
 - Confusion matrix
 - Class statistics measures
 - Accuracy measures
 - ROC curves
 - Cohen 's kappa
 - Numeric error measures
- Visualizing data
 - Methods for One and Two Attributes
 - Barchart and Histogram
 - Boxplot

- Scatter plot and density plot
- Methods for Higher-dimensional Data
 - Principal Component Analysis (PCA)
 - Multidimensional Scaling (MDS)
 - t-distributed Stochastic Neighbour Embedding (t-SNE)
 - Parallel Coordinates
 - Radar and Star Plots
 - Sunburst Chart
 - Correlation Analysis

In case of Pestron, not used all of methods above, due some tools are unnecessary or irrelevant [7].

IV. RESULTS

To represent tools and methods and data warehouse, I used Knime and Orange as data analysis software.

Fig. 3. in Appendix refer to Filtering on Columbia Area in Pestron US Food Stores table.

Fig. 4. in Appendix refer to Clustering of West Area in Pestron US Food Store table.

Fig. 5. in Appendix refer to Distribution of Product by Gender in Pestron Sales table.

Fig. 6. in Appendix refer to Violin Plot of Total Amount by City in Pestron Sales table.

V. CONCLUSION

Big Data may bring evolution in decision making in businesses life. This progress does not depend on the size or sector, because Big Data tools need only data, and they work for the given goals.

Benefits of Big Data is wide range, and significant.

- Apply data beyond traditional analytics cases to support real-time decisions anytime, anywhere.
- Manage all the information that can be used in data-centric decision making.
- Empowers people in all roles to explore and analyse information and provide insight to others.
- Optimize all decision types, whether they are made by individuals or embedded in

automated systems, using analytics-based statistics.

- Provide insight from all perspectives and time horizons, from historical reporting to real-time analysis, predictive modelling
- Improve business results and manage risks, now and in the future.
- Organizations can apply decisions beyond knowledge.
- Companies can access community data from search engines and websites to develop business strategies.
- Traditional customer feedback systems will be replaced by new systems using Big Data technologies. These new systems use Big Data and natural language-based technologies. This is how evaluate consumer feedbacks.
- Early identification of product / service risk.
- Better operational efficiency.
- With new applications, Big Data technologies use to temporarily store new data before knowing which data need to add to the data warehouse. In addition, the integration of business applications with Big Data technologies and data warehousing solutions can help load the real-time data stored there.

What may concern a medium or small sized company is the Big Data requirements. Big Data has few backwards of course, but some of them hit SME's more. First the list of general disadvantages:

- Data quality issues
- Improved security risk
- Cost of resources
- Difficulties with data integrating
- Hardware requirements
- Lack of Data Expert / Data Scientist

From this list most concern of a smaller firm is related to high cost. These businesses sometimes cannot afford to hire a designated Data Expert, cannot afford investing in expensive hardware or software. Better to reckon what is the cost of investment to use Big Data tools, altogether with hardware, software, salary of Data Expert. Meanwhile should investigate the additional revenue using Big Data solution. If the investment costs more than the expected revenue, the project of Big Data is not beneficial, so as all profit-oriented party better to keep existing solution. However, a cloud-based pay-as-you-go Big Data solution generate higher operational cost in beginning, in a

long-term may it be profitable, because a cloud-based solution always means better solution for the whole IT department.

Summed up, to introduce a Big Data solution in a small or medium sized enterprises is a risky and hard decision, but this kind of decision making becoming more straightforward in the future, if company using Big Data tools [8].

VI. REFERENCES

- [1]"Definition of Big Data - Gartner Information Technology Glossary", Gartner, 2021. [Online]. Available:<https://www.gartner.com/en/information-technology/glossary/big-data>. [Accessed: 02- Sep- 2021].
- [2]"What Is Big Data? | Oracle", Oracle.com, 2021. [Online]. Available: <https://www.oracle.com/big-data/what-is-big-data/>. [Accessed: 02- Sep- 2021].
- [3]N. Raheem, Big data : a tutorial-based approach, First edition. Boca Raton: CRC Press, 2019, pp Chapter 1-3.
- [4]B. Marr, Big data : using smart big data, analytics and metrics to make better decisions and improve performance, 1st edition. Chichester, West Sussex, [England: John Wiley & Sons, 2015, pp Chapter 3-5.
- [5]A. Phillips, "A history and timeline of big data", WhatIs.com, 2021. [Online]. Available: <https://whatis.techtarget.com/feature/A-history-and-timeline-of-big-data>. [Accessed: 02- Sep- 2021].
- [6]J. Watt, R. Borhani, and A. K. Katsaggelos, Machine learning refined : foundations, algorithms, and applications. Cambridge: Cambridge University Press, 2016.
- [7]M. Berthold, C. Borgelt, F. Höppner and F. Klawonn, Guide to Intelligent Data Analysis, 2nd ed. London: Springer-Verlag London Limited, 2010, pp. Chapter 3-9.
- [8]C. Harvey, "Big Data Pros and Cons | Datamation", Datamation, 2018. [Online]. Available: <https://www.datamation.com/big-data/big-data-pros-and-cons/>. [Accessed: 02- Sep- 2021].

VII. APPENDIX

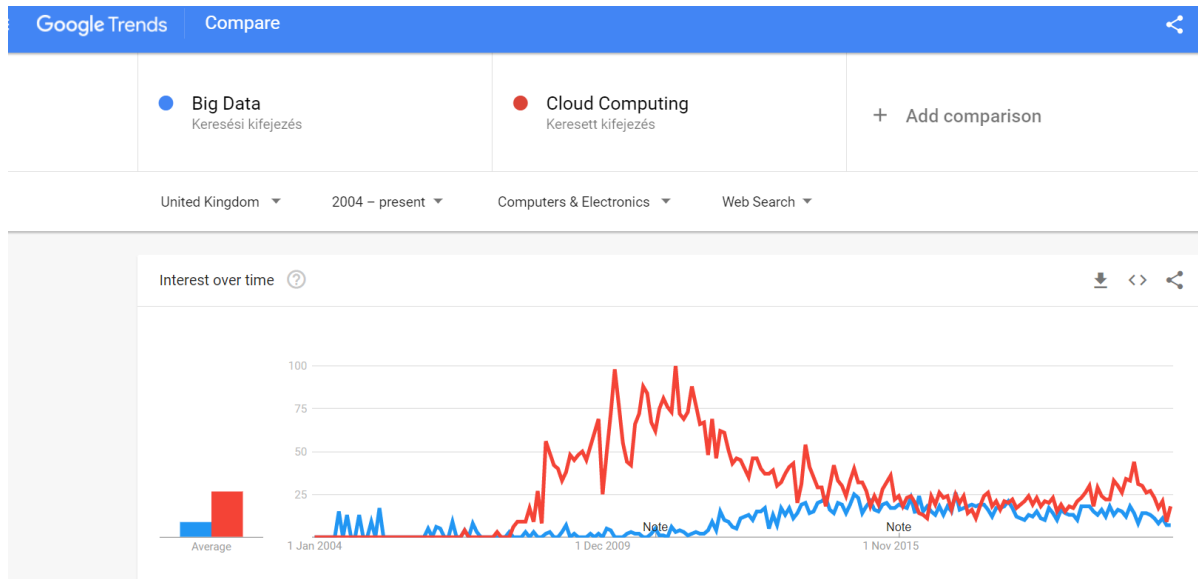


Fig. 1. Comparing Big Data and Cloud Computing search history in Computer and Electronics topic in the UK from 2004-today
(source: Google trends, author)

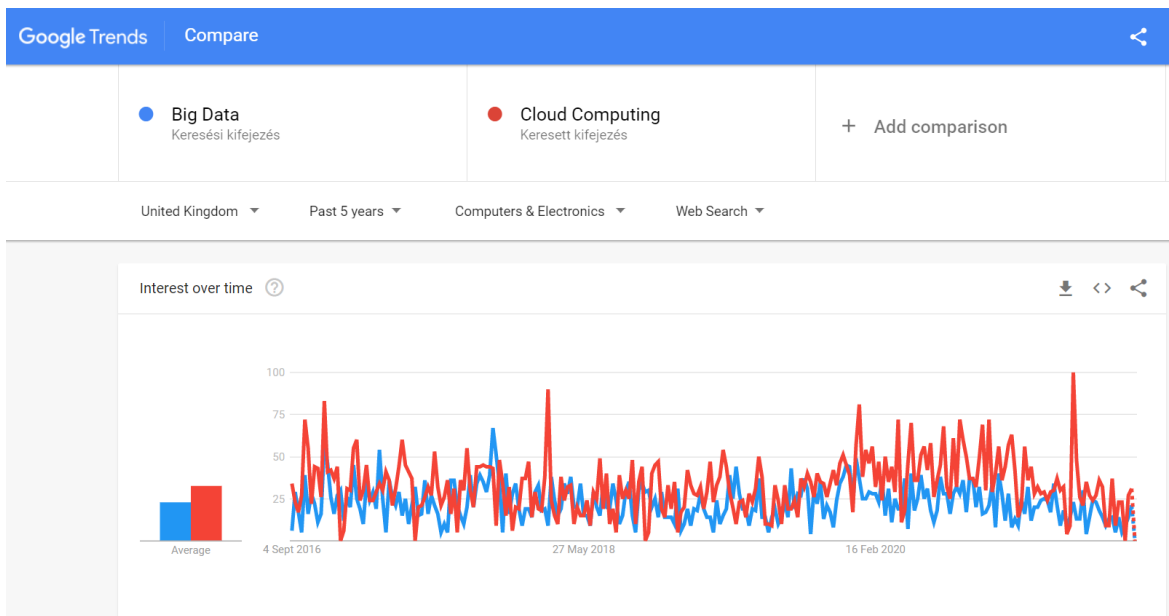


Fig. 2. Comparing Big Data and Cloud Computing search history in Computer and Electronics topic in the UK from 2016-today
(source: Google trends, author)

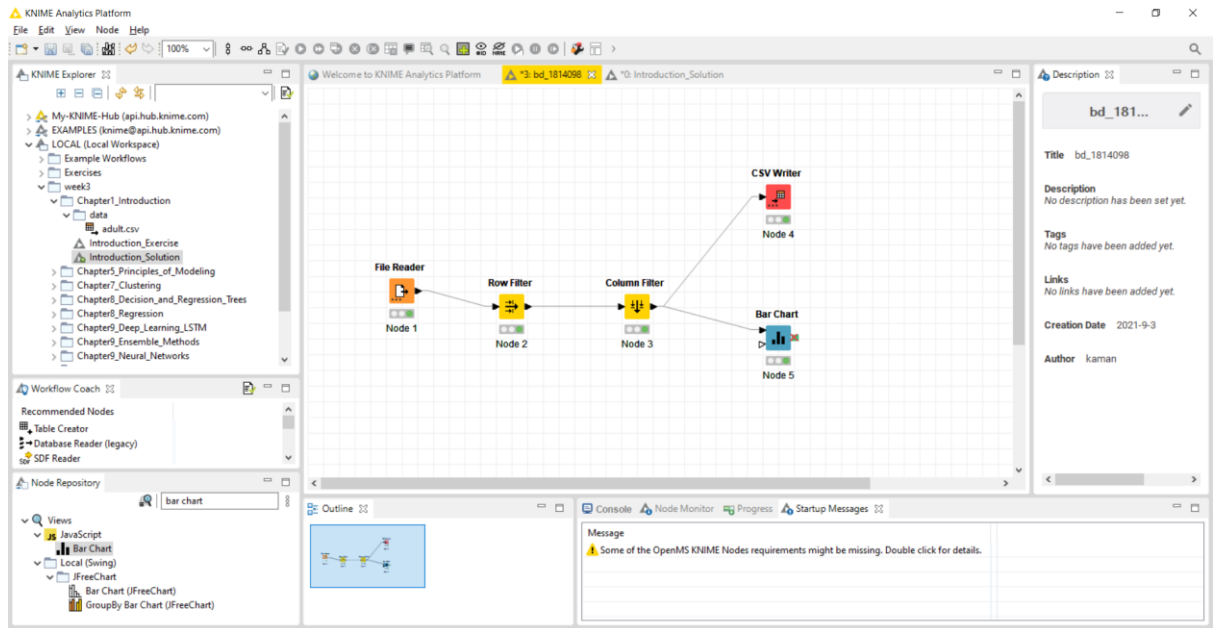


Fig. 3. Filtering on Columbia Area (source: Knime, author)

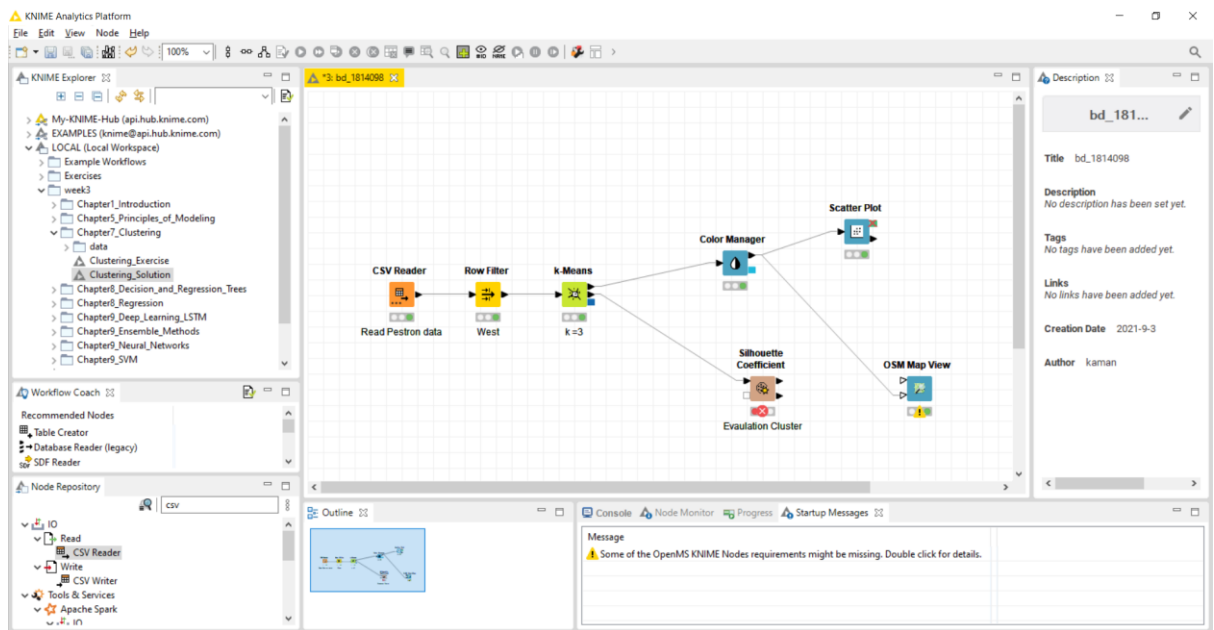


Fig.4. Clustering on West Area (source: Knime, author)

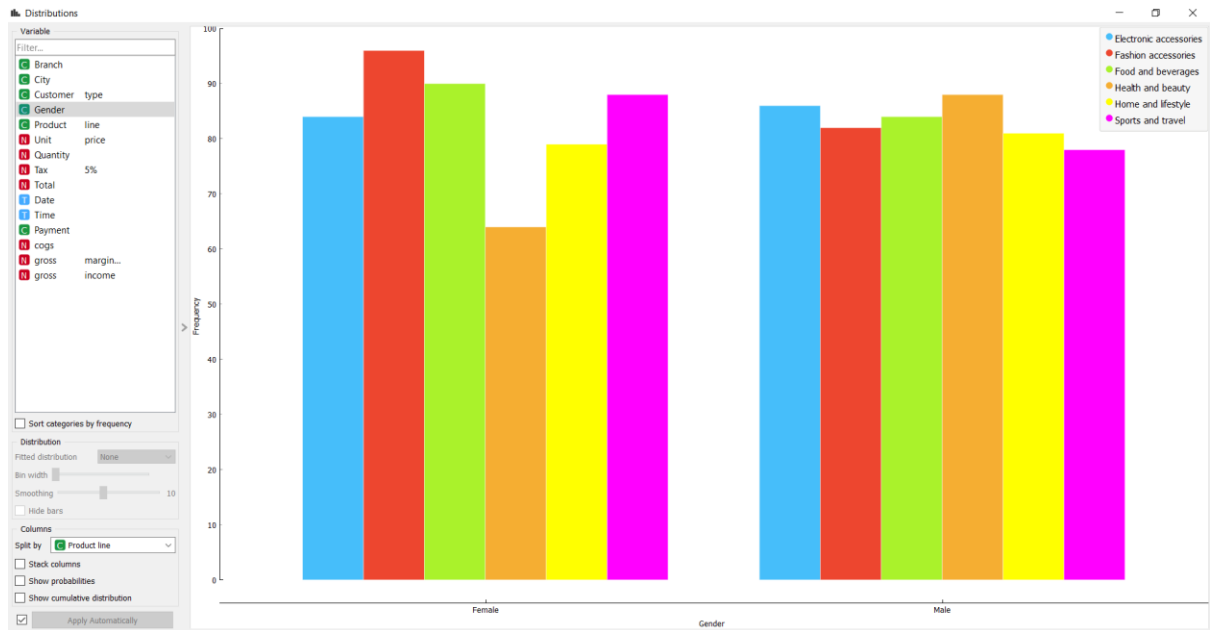


Fig. 5. Distribution of Product by gender (source: Orange, author)

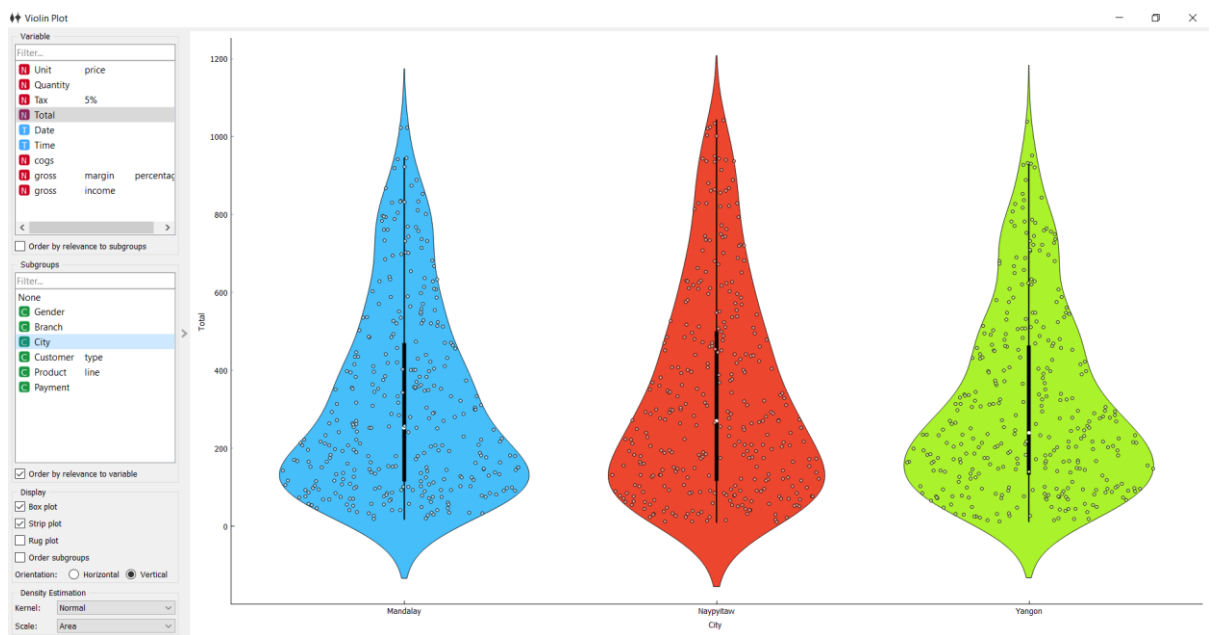


Fig. 6. Violin Plot of Total Amount by City (source: Orange, author)