Data mining at cloud

*Submited for SWE 590 fall 2021-2022

Ömer Faruk Çevik
Software Engineering
Boğaziçi University
Istanbul, Turkey
farukcevik@gmail.com

***Abstract*—Big data and analytics software spending has been increasing in last decade [1]. It can be interpreted that companies and people need both hardware computing capability and software to process data. The data mining concept encompasses many interrelated study areas like machine–learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization [3]. Selecting appropriate design for data analysis depends on the institution's needs such as application/algorithm needs, time deadline for results, size of data for processing. Hardware/software requirements depends on size of the data (volume), real-time data collection (velocity), heterogeneous data collection from a diverse range of resources (variety), unpredictable data (veracity), and finally the application of such data in various fields, such as industry and academia (value). In other words, we have requirements and constraints from both data mining and cloud computing aspects. Lastly, adding explainability as a non-functional requirement to data mining at cloud systems can make model more understandable for both engineers and end users. Since software-driven systems' complexity and autonomy increasing everyday, sometimes even domain experts and system engineers struggle to understand certain aspects of a system [21].**

***Index Terms*—Cloud computing, data mining, model design**

I. Introduction

Big data and analytics software spending has been increasing in last decade [1]. It can be interpreted that companies and people need both hardware computing capability and software to process data. Data mining concepts and techniques are tailored to process raw data to reach information, then process information to reach knowledge [2]. There are various data mining methods such as classification, clustering, association and regression [3]. The related past data is gathered, organized, transformed and processed to reach knowledge in a data mining project. The data mining techniques in computer science started to be used by researchers in the 1960s and recognized as a separate field and named in the 1980s [8,9]. The introduction of data warehouses; a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management's decision-making; by William H. Inmon soared data mining research in early 1990s [7]. The data mining concept encompasses many interrelated study areas like machine–learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization [3]. During the 1980s and 1990s data mining and its subfield artificial intelligence started to be used by researchers to build forecasting systems. Sample studies for forecasting with artificial intelligence area are bankruptcy, business failure, foreign exchange rate prediction, stock prices, electric load consumption, airborne pollen, commodity prices, temperature, helicopter component loads, airline passenger traffic, ozone level, personnel inventory, river flow, student grade point averages, total industry production, trajectory, transportation, water demand and wind pressure [5]. Those studies are still being developed and used in the 2020s. After the 2000s, data mining and forecasting studies extended by being able to work with massive data thanks to huge increase in computation power and speed. There are thousands of studies related with big data processing and even instant processing of streaming data like 3D tracking and movement forecasting with rich maps for autonomous driving [10] and predicting taxi passenger demand from live data of vehicles [11]. What is more, storing electronic medical records (EMRs) of billions of people in blockchain is now possible [12,13]. The MedRec system, developed by MIT, enabled storing medical records in blockchain in a coordinated way without disregarding privacy, authentication, confidentiality, accountability issues [12]. Medical researchers, public health authorities, etc. can participate in the network as blockchain "miners" and they can analyze information from many sources in order to identify public health risks, develop new treatments and cures, and enable precision medicine [12,14]. In fact, data mining (data science, big data) is not new. Santovena [15] researched the history of data storage and processing in his master's science thesis. He proposed that the 1890 census was probably the first big data case in North America. Herman Hollerith had built a punched card tabulating machine that could be read by electrical sensing to count statistics. "The way this technology worked was simple, when a stylus was inserted into a hole on a template; a corresponding hole was punched in a card at the other end. Each card represented one person and each hole a different statistic, such. The cards were sorted and later read electronically by a press containing pins that penetrated the card only at its holes. Each pin that passed through a hole made electrical contact with a small

Identify applicable funding agency here. If none, delete this.

cup of mercury, closing a circuit and advancing a dial counter by one." This technology reduced the time of sorting out the census results from eight years (formerly by hand) to one year. The evolution of data mining throughout history summarized by Santoneva [15] in the list below.

TABLE I
SANTONEVA'S REVIEW OF EVOLUTION OF BIG DATA

Year	Topic
1890	The census - the first technology enabled "snapshot" data collection
1935	The Social Security Act 11
1943	The "Colossus"
1944	A librarian storage and organization problem
1961	The U.S. National Security Agency backlog
1961	The growth of scientific knowledge
1964	The recipe on how to cope with the information explosion
1965	First Data Center is conceived
1967	Data compression
1974	Data Science is born
1977	The International Association for Statistical Computing
1989	The Internet is born
1996	The Data Science evolution
1997	The year for Data Mining and Big Data
2004	Let's play with Hadoop
2011	The Jeopardy champion: Watson

[a]by Santoneva [15]

Selecting appropriate design for data analysis depends on the institution's needs such as application/algorithm needs, time deadline for results, size of data for processing, the fundamental nature of model: iterative or single iteration, increased data processing capability in future, speed of data transfer, type and nature of data, handling hardware failures, etc [18]. Choosing the right hardware and software platforms play a crucial role to handle big data. The definition of the term big data again depends on circumstances [16]. In fact, all the hardware/software requirements and the definition of big data summarized in Vs in the literature. Doug Laney defined 3Vs; Volume, Velocity, Variety. Tsai Chun-Wei et al, added another V to 3Vs, veracity, and described 4Vs [17]. IBM defined 5 Vs volume, velocity, variety, variability, and value [19]. So, a 500 megabyte data can be big data for a company, whereas another institution can regard over 1 exabytes as big data.

II. MOTIVATION

A. Need for Tailored Approach

When I was an undergraduate student in 2000s, I tried to find stock values of Istanbul stock exchange in US Dollar terms. I downloaded raw data of prices of stocks in Turkish Lira terms from the stock exchange's website, then I downloaded US Dollars value in the central bank's website. I used a spreadsheet application to divide TL price of stocks to USD. I don't even mention about how I calculated other indicators such as RSI (Relative Strength Index), and applied data mining algorithms. Today, as the year 2022, anyone able to check economic indicators in a cloud-based charting and social-networking portal such as TradingView or Investing in a few seconds. The cloud-based platforms able to calculate general or predefined parameters , such as The Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), in advance and share it to related parties. Thus, participants, may be numbered in millions, do not need to spend computing power to calculate the statistical terms. In other words, cloud systems can handle data gathering, processing and displaying executions and distribute results to all stakeholders. The cloud or hybrid approach, instead of putting data directly into database cleaning, formatting and analyzing the data while it's being generated, can increase speed and decrease cost of data mining design. Lastly, adding explainability as a non-functional requirement to data mining at cloud systems can make model more understandable for both engineers and end users. Since software-driven systems' complexity and autonomy increasing everyday, sometimes even domain experts and system engineers struggle to understand certain aspects of a system [21]. Explainability tries to add 5Ws, what, when, why, who and where, to make system understandable and configurable by related stakeholder. Descriptions improve usability, help in locating sources of error, and can minimize the chance for human error[21]. A lack of explainability, on the other hand, not only gives rise to various moral, social, and legal problems[21]. It further fuels distrust, diminishes user acceptance and satisfaction, and inhibits the adoption of new technologies. Explainability of overall system can be satisficed, not satisfied. Meaning that minimum requirements of descriptions can be provided. Köhl et al. [22] dig out three definition of explainability.

• Definition 1 (Explanation For): E is an explanation of explanandum X with respect to aspect Y for target group G, in context C, if and only if the processing of E in context C by any representative1 R of G makes R understand X with respect to Y

• Definition 2 (Explainable System): A system S is explainable by means M with respect to aspect Y of an explanandum2 X,for target group G in context C, if and only if M is able to produce an E in context C such that E is an explanation of X with respect to Y , for G in C.

• Definition 3 (Explainability Requirement): A system S must be explainable for target group G in context C with respect to aspect Y of explanandum X.

When we consider adding explainability to a cloud system, detailed information such as which equipment added, when and why added should be written. Why vertical scaling is preferred instead of horizontal scaling or vice versa should be documented. Thus, after a certain time passes, the engineers can modify the system easily. Moreover, maintenance issues can be handled in short time because the designers or engineers can understand why certain design is selected, where to intervene to solve an issue. In a company perspective, which functions of which cloud service provider is used, for example, should be documented. In a case of changing inputs of a function, the engineers can find where to look to modify the functions.

III. Data mining and Cloud

Applying big data analytics at cloud for mining information, extracting knowledge and making predictions/inferences has recently attracted significant attention, especially after COVID-19 outbreak exposed the shortcomings of traditional healthcare system [21]. In other words, health ministers need to process health data, even may be navigation data, of millions even billions of citizens to fight against outbreak efficiently. This is a very specific or rare case. For a generalized approach to design any cloud base data mining solution, engineers and specialists should consider end user needs, expected usage length or life of the system, technological capabilities as well as constraints such as budget, privacy, explainability. Of course, the data mining in cloud should also take account the famous Vs, size of the data (volume), real-time data collection (velocity), heterogeneous data collection from a diverse range of resources (variety), unpredictable data (veracity), and finally the application of such data in various fields, such as industry and academia (value). In other words, we have requirements and constraints from both data mining and cloud computing aspects.

For a small brokerage firm with 10 employees, for example, can benefit cloud-based charting platforms, and the platforms can serve all the requirements of a small investment advisor firm. All known economic indicators and indices are already available in cloud systems. The cloud systems are specialized to maintain streaming of the raw data. To put in another way, the small company do not need to search for data sources, how to reach the data via web services or data link, is data in JSON format or not, what is the refresh interval and etc. Thus, no data resource and streaming integration afford is needed. Platform as a service (PaaS), approach might be enough for that type of small brokerage firms. On the other hand, if a brokerage firm need to calculate its own values or/and a data source is not available in the cloud service provider, then the company should search for cloud-service providers which provide special data source definition capability. As an example, suppose the brokerage firm needs number of active distinct traders in an exchange. The data is not publicly available on the cloud platform. In this situation, the company will pay for the data, number of active users, to exchanges. The exchanges will share the data in an agreed way. There will be an integration for data to stream from exchanges to company's servers, either in cloud or local servers. There can be privacy issues related with data. On the other hand, imagine a trillion dollar hedge fund which streams data from thousands of sources. They might calculate some parameters which must be kept secret. In that particular case, the company probably need to calculate and store the parameters in local servers in encrypted way. Kim et al.[24], researched encrypted data storage at cloud, querying data mining association data and privacy issues. Then, they proposed technical solutions. When I return to my case, Infrastructure as a Service (IaaS), cloud-based services, pay-as-you-go for services such as storage, networking, and virtualization, approach can be preferred. As second alternative, the company can setup his own servers on local. A local bare metal platform can be established. Again depending on number of technical staff, education and experience level of IT personnel, budget allocated for privacy, budget allocated for hardware. Of course, after the initial investment there will be electricity, cooling, surveillance and other streaming costs for keeping system up and running.

There are various data mining algorithms such as classification, pattern recognition, clustering, association rule, recommender systems and etc. When we look at history of data mining software, there are four generations of algorithms[23]. There are dozens of both open source and proprietary software related with big data or data mining solutions. R, Python and WEKA among the first generation of data mining tolls. Those first generation tools requires algorithms and data should be kept in the same memory for computing. The memory of Registers, L1 (SRAM), L2 (SRAM) and DRAM (main memory) are being used for computation on the memory. The second generation of data mining software can work on multiple servers. The second generation mostly used Hadoop/MapReduace framework together with Hive,HBase and many others. Extensive libraries developed in Java mainly for Apache's Hadoop. The third generation of used MapReduce to perform parallel and distributed programming. With this model image, map and other type of data can be handled in real time. The fourth generation is dealing with hybrid approach and integration of different platforms. The special term "plumbing" is used to refer integrating heterogeneous platform with their diversity of scalability and impedance mismatches. Gadde and Viraj [16] researched evolution of big data with Hadoop throughout 20 years. According to the authors, although Hadoop with its complement products is a powerful solution for handing Big Data, Hadoop is a suitable solution for only OLAP. However, Hadoop doesn't sounds good for frequently changing data. In other words, we can say that at present there is no transaction support in Hadoop. If our data mining requirement requires instant analysis of data from millions of users, the required design for data mining will be much different than pattern recognition from a 20 years accumulated data. So, data mining requirements can determine some Vs and available technology determines remaining Vs. On the other hand, Royon and Benitez[26] criticize cloud providers like Amazon (SageMaker), Google (Google Machine Learning), IBM (Watson) or Microsoft (Azure ML) that they lack of a standardized definition and description. The two authors suggest that after definition and description, appropriate algorithm should be selected to provide a complete data mining platform.

Today, cell phones, smart phones, servers, personal pcs, sensors, cameras, IOT devices and etc. generate huge amount of data. Statistics vary on open sources[27]. Total data produced

in 2020 is estimated as 44 zettabytes. Google has 1.5 billion active accounts. Facebook has 2.8 billion active users. There are 1 billion active Instagram users, and 95 million photos and videos are shared on Instagram per day. On average, every human created at least 1.7 MB of data per second in 2020. All the world created 2.5 quintillion data bytes daily in 2020. On average there are 575,000 tweets per minute. By 2025, the amount of data generated each day is expected to reach 463 exabytes globally. There are 5.7 million Google searches per minute on average. These millions of searches are used by Google's algorithms to constantly update and improve its search functionality. When a governments, municipalities, companies or individual want to perform data mining operations they might prefer big data as a service (BDaaS), data mining as a Service (DMaaS), Blockchain as a service (BaaS) or any other thing as a service. Those as a service probably convenient to use, but a cloud computing specialist should tailor those facilities according to requirements and constraints.

IV. Conclusion

Data mining at cloud enables extensive features for companies or individuals who need data gathering, processing and displaying features. Engineers and business owners need to work together to build a setup according to the data mining needs and business circumstances. 5Vs (volume, velocity, variety, veracity and value), edge computing possibilities, live tracking needs should be analyzed before providing a solution. Despite the fact that cloud computing provide extensive flexibility to supply computational assets on need, it still need time to get mature to be known as Cloud-supported-analysis or Analytics as a Service (AaaS) or Big Data as a Service (BDaaS) [23]. However, business owners working together with IT professionals, such as analysts, infrastructure managers, data base admins, configuration experts, project managers, developers etc., can tailor a cloud base solution suitable for the need of business or governments.

References

[1] www.statista.com/statistics/472934/business-analytics-software-revenue-worldwide/ (reached at December 2021)

[2] Jiawei Han, Michelin Kamber, Data Mining Concepts and Techniques [M], Morgan Kaufmann publishers, USA, 2001 (available at http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf)

[3] Jing He, Advances in Data Mining: History and Future, Third International Symposium on Intelligent Information Technology Application (2009)

[4] https://kommandotech.com/statistics/big-data-statistics/

[5] Zhang, G., Patuwo, B.E. and Hu M.Y. (1998), Forecasting with artificial neural networks: The state of the art, International Journal of Forecasting 14 (1) 35– 62.

[6] Xia Geng, Zhi Yang, Data Mining in Cloud Computing, International Conference on Information Science and Computer Applications (2013

[7] Yan Chen , Ming Yang , Lin Zhang, General Data Mining Model System Based on Sample Data Division, Second International Symposium on Knowledge Acquisition and Modeling (2009)

[8] Frans Coenen, Data Mining: Past, Present and Future, The Knowledge Engineering Review, Vol. 00:0, 1–24. c 2004, Cambridge University Press

[9] wiki Data mining

[10] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Sławomir B ak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays, Argoverse: 3D Tracking and Forecasting with Rich Maps, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[11] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas , Predicting Taxi–Passenger Demand Using Streaming Data, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 14, NO. 3, SEPTEMBER 2013 1393

[12] Asaph Azaria, Ariel Ekblaw, Thiago Vieira and Andrew Lippman, MedRec: Using Blockchain for Medical Data Access and Permission Management, 2016 2nd International Conference on Open and Big Data

[13] Leslie Mertz , (Block) Chain Reaction: A Blockchain Revolution Sweeps into Health Care, Offering the Possibility for a Much-Needed Data Solution, 2018

[14] The Office of the Nat. Coordinator for Health Information Technology, "Report on health information blocking," U.S. Department of HHS, Tech. Rep., 2015.

[15] Alejandro Zarate Santovena , Big Data: Evolution, Components, Challenges and Opportunities , Master of Science thesis (2013)

[16] Ramesh Gadde, Namavaram Vijay, A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP , International Journal of Research In Science Engineering , 2017

[17] Tsai, Chun-Wei, et al. "Big Data Analytics." Big Data Technologies and Applications. Springer International Publishing, 2016. 13-52.

[18] Alka Londhe, PVRD Prasada Rao, Platforms for Big Data Analytics: Trend towards Hybrid Era, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)

[19] https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ (Accessed on 13.12.2021)

[20] Doug Laney," 3d data management: Controlling data volume, velocity and variety", META Group Inc,6 Feb 2001

[21] Wei Li at al. A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System, Mobile Networks and Applications Springer Science (2021)

[22] Maximilian A Köhl, et al., Explainability as a Non-Functional Requirement, IEEE 27th International Requirements Engineering Conference (RE) (2019)

[23] Alka Londhe,Prasada Rao, Platforms for Big Data Analytics: Trend towards Hybrid Era, nternational Conference on Energy, Communication, Data Analytics and Soft Computing (2017)

[24] Hyeong-Jin Kim et al., Privacy-preserving Association Rule Mining Algorithm for Encryped Data in Cloud Computing, 12th International Conference on Cloud Computing (CLOUD) (2019)

[25] Yang Xiao et al., Design of data mining system based on cloud computing, 12th International Conference on E-Commerce and Internet Technology (ECIT) (2020)

[26] Manuel Parra-Royon, Jose M. Benitez, Delivering Data Mining Services in Cloud Computing, IEEE World Congress on Services (SERVICES) (2019)

[27] https://techjury.net/blog/how-much-data-is-created-every-day/ https://seedscientific.com/how-much-data-is-created-every-day/