ANATOMY OF BIG DATA LAKE HOUSES

Vishwanadham Mandala[1], Mahindra Sai Mandala[2]

[1]Enterprise Architect, B.Tech in CSE, M.Tech in CSE, MS in Data Science,

Indiana University, Bloomington, IN, USA

[2]Sophomore, Center Grove High School, Greenwood, Indiana, USA.

Corresponding author email ID: *Vishwanadh.mandala@gmail.com*

Abstract-

Big data lakes are now important components of emerging big data architectures. Big data lakes are essentially the natural evolution of data warehousing systems in the big data context, and they address a number of requirements arising from the well-known 3V nature of big data. Several issues arose alongside the development of the big data lake research initiative including: i) big data lake models; (ii) big data lake frameworks; and (iii) big data lake techniques. In line with this exciting research perspective, this paper proposes an overview of cutting-edge approaches that are at the foundations of big data lake research, as well as innovative open problems and issues that drive future research directions on advancing the big data lake research trend.

Keywords- Big Data Analytics , Big Data Representation , Big Data Processing , Big Data; Big Data Lakes.

DOI Number: 10.14704/nq.2022.20.9.NQ44752 **Neuro Quantology 2022; 20(9):6413-6418**

I. INTRODUCTION

Big data lakes [1-5] are the natural evolution of data warehousing systems [6] and they have emerged in the emerging trend known as big data [7-9]. Big data lakes essentially deal with the requirements imposed by the well-known 3V model of big data [10] which stands for volume, velocity and variety. According to basic definitions big data lakes are intended to be a type of shared environment containing multiple big data repositories that are represented processed and mined using already-consolidated big data technologies [11-13]. A big data lake's ultimate goal is to support advanced big data analytics procedures [14-16] such as:

- big data exploration [17]
- knowledge discovery over big data [18] and ad-hoc big data analytics [19]
- sophisticated big data analytics [20]
- complicated tools for big data OLAP reports, dashboards etc. Dealing with the variety of heterogeneity of big data is a key feature of the big data lake model [22]. Today's large-scale organizations include a variety of big data repositories, ranging from text data to spreadsheet data, Web data to relational data, social data to censor data, streaming data to real-time data and so on. Because of this heterogeneity as well as the high volume and rate of modern big data designing implementing and running an effective and efficient big data late is a critical challenge for next-generation big data research.

Several technological solutions that support big data lakes already exist. A classic architecture, for example, sees the big data lake built on a standard Hadoop infrastructure [23] with big data from the target global repository stored in a proprietary (data) cluster and Hadoop software providing all the necessary basic as well as advanced routines for representing processing managing and supporting advanced analytics over such data.

It is impossible to address fundamental big data lake research issues without mentioning an important property of big data lakes namely what data scientists can do with modern big data lakes. Essentially, big data lakes must also support the monetization of big data within a large-scale organization by enabling meaningful analytics procedures over big data stored in big data lakes for decision support purposes e.g., business,

social trends, epidemiologic evolutions, and so forth. This unique feature is now supported within big data lakes by sophisticated techniques inspired primarily by machine learning and artificial intelligence solutions [24-26].

Several issues arose alongside the emergence of the big data lake research initiative. The following are some of the most noticeable:

• Big data lake models: how are large amounts of data represented in a big data lake

• Big data lake frameworks: what are the most effective architectural solutions for building a productive big data lake

• Big data lake techniques: how to integrate multiple, heterogeneous big data repositories into a big data lake in an effective and efficient manner. How do you query big data in a big data lake effectively and efficiently

This paper was inspired by these difficult questions. To address these latter, we present in this paper I an overview of state-of-the-art approaches that are at the foundations of big data lake research particularly those that attempt to answer these challenging questions, and (ii) innovative open problems and issues that drive future research directions on advancing the big data lake research trend.

II. BIG DATA LAKES: REVIEW OF LITERATURE

Several proposals have recently surfaced in the context of big data lake research, under the three research axes introduced in Section 1. i.e., models, frameworks, and techniques. This Section provides an overview of the most important ones. examines the definition and popularity of big data lakes in today's big data era More specifically the authors recognize that the concept of a data lake is gaining popularity as a popular way to organize and build the next generation of systems to address new big data challenges. Nonetheless, there are numerous concerns and questions that large enterprises must address before implementing data lakes. The paper discusses the concept of data lakes and shares the author's thoughts and practices on data lakes within this line of research [5]. The emphasis remains on the fundamentals of big data lakes and related research issues. According to the authors, two new concepts, in addition to big data, are emerging today: data lakes and fast data. Are they just new marketing labels for the old big data term or are they truly new As a result the primary goal of this paper is to determine the relationship between these three concepts. considers a real-world application of big data lakes specifically architectures for supporting personalized healthcare service recommendation According to the authors, personalized healthcare services tailor medication recommendations using relational patient data and big data analytics. However, the majority of healthcare data is unstructured and requires a significant amount of time and effort to convert to relational format. This paper proposes a novel data lake architecture to reduce data ingestion time and improve healthcare analytics precision. It also eliminates data silos and improves analytics by allowing connectivity to third-party data providers such as clinical lab results, chemist, insurance company etc.. The Hadoop Distributed File System (HDFS) is used to store both structured and unstructured data in the data lake architecture, and the K-means clustering algorithm is used to find patient clusters with similar health conditions. Finally a Support Vector Machine (SVM) is used to determine the most successful healthcare recommendations for each cluster. Experiment results show that data lakes can reduce the time required to ingest data from various data vendors regardless of format.

Likewise focuses on another intriguing case emerging smart grids. Indeed, the authors acknowledge that advances in smart grids are allowing massive amounts of data to be aggregated and analyzed for a variety of smart grid applications. Traditional smart grid data management systems, on the other hand are incapable of scaling and providing adequate storage and processing capabilities. To address these issues, this paper presents a smart grid big data eco-system based on the cutting-edge Lambda architecture which can perform parallel batch and real-time operations on distributed data. Furthermore, the presented eco-system stores various types of smart grid data, such as smart meter, image and video data in a Hadoop Big Data Lake. The smart grid big data eco-system is implemented on a cloud computing platform [3]. Real-time visualization and data mining applications were performed on real smart grid data to test the capability of the presented eco-system. The outcomes of those applications on

top of the eco-system indicate that it is capable of performing a wide range of smart grid big data analytics. Azure Data Lake Store (ADLS) is a fully managed, elastic, scalable and secure file system that supports HDFS and Microsoft's Cosmos file system semantics. It is specifically designed and optimized for a wide range of Big Data analytics that require a high degree of parallel reads and writes as well as compute and data collocation for high bandwidth and low-latency access. It combines key components and features of the Cosmos file system, which has long been used by internal Microsoft customers, and HDFS, and is a unified file storage solution for analytics on Azure. This unified platform supports both internal and external workloads. The design of ADLS for handling multiple storage tiers exabyte scale and comprehensive security and data sharing features are distinguishing features. To summarize the paper provides an overview of ADLS architecture, design considerations and performance. Instead Hydria is an online data lake that enables users with no prior knowledge of information technology (IT) to harvest, store, organize, analyze and share heterogeneous, multifaceted cultural heritage data. Hydria is a zero-administration zero-cost integrated framework that enables researchers, museum curators and other stakeholders in the cultural heritage domain to easily I deploy data acquisition services(such as social media scrapers, focused web crawlers, dataset imports questionnaire forms (ii) design and manage versatile customizable data stores (iii) share whole datasets or horizontal/vertical data shards with other stakeholders and (iv) search, filter a dataset. According to the authors best knowledge this is the first solution in the literature that focuses on collecting, managing, analyzing and sharing diverse, multi-faceted data in the cultural heritage domain and is aimed at users with no prior IT experience.

Finally considers big data lake maintenance issues. The authors recognize that data lakes (DLs) are large repositories of raw datasets from various sources. As more datasets are ingested into a DL, there is a growing need for efficient techniques to profile them and detect relationships between their schemata, a process known as holistic schema matching. To aid in information discovery and retrieval schema matching detects similarity between the information stored in datasets. With the volume of state-of-the-art DLs available today, this is currently computationally expensive. To address this issue the authors propose a novel early-pruning approach to improve efficiency, in which various types of content metadata and schema metadata about the datasets are collected and then used in early-pruning steps to pre-filter the schema matching comparisons. This includes computing proximities between datasets based on metadata discovering relationships between datasets based on overall proximities and proposing similar dataset pairs for schema matching. The authors improve the task's effectiveness even further by introducing a supervised mining approach for detecting similar datasets that are proposed for further schema matching. Extensive experiments on a real-world DL show that the proposed approach is effective at detecting similar datasets for schema matching with recall rates exceeding 85% and efficiency improvements exceeding 70% [28].

III. BIG DATA LAKES: RESEARCH TASKS

Under the axiomatic vision proposed in this paper, there are numerous research challenges and directions in the context of big data lakes. Following, we will report on and discuss some notable ones.

Architectures for Big Data Lakes: The problem, as with the ancestor data warehousing systems, is designing and implementing effective and efficient big data lakes. Indeed, the architectural design of every big data lake has a significant impact on the effectiveness and efficiency of the big data lake itself. As a result, architectural paradigms for big data lake design, which must evolve from early data warehousing system design experiences are a first important big data lake research open issue.

Data Population/Alimentation Paradigms for Big Data Lakes in the Modern Era Traditional data warehousing systems rely heavily on data population/alimentation processes. These govern how input data sources are imported from the data source layer to the data storage layer of the target data warehousing system within the general scope of the major ETL (Extraction, Transformation, Loading) process. However, how can this model be replicated with big data lakes. The 3V nature of (input) big data makes it difficult to apply as-is, as recent studies [29] have also highlighted. Innovative solutions are required in this situation. Among other things

sampling-based approaches appear to be a promising initiative.

Paradigms for Query Rewriting in Big Data Lakes Big data lakes typically keep big data repositories ingested within their core layer in a kind of lazy integration in which these repositories are kept in their original form and the "laking" effect is obtained through schema mapping methodologies and most importantly query re-writing techniques. One of the most difficult issues in future big data lake research is designing and developing effective and efficient query re-writing techniques for the core layer of big data lakes.

Metadata Managers of the Future for Big Data Lakes Metadata management in big data lakes is critical, even more so than the corresponding functionality in ancestor data warehousing systems. Indeed, due to the 3V nature of big data, big data exploration is the first and most important procedure to be supported within a big data lake. Because standard big data lake architectures are complex, the key to success for supporting effective (and efficient) big data exploration is simply effective metadata manager design and implementation for big data lakes. Simply put, traditional approaches are ineffective in modern big data application scenarios, necessitating the development of innovative, modern metadata manager design and implementation paradigms [30].

Over Big Data Lakes Machine Learning Tools The support for machine learning tools at the frontier layers of big data lakes is a distinguishing feature. This is intended to support value extraction from core big data repositories, which will then be used for knowledge discovery and decision-making processes. A significant number of machine learning tools have been proposed in the literature over the years. Unfortunately, these use a basic raw (input) data model that is incapable of dealing with modern features of big data repositories. As a result, designing effective and efficient machine learning tools to be implemented at the front-end of big data lakes becomes one of the most relevant challenges to face owing to the outcomes in a plethora of real-life big data applications ranging from healthcare analytics to social evolution analysis tools smart city applications to e-science solutions and so on.

Methodologies for Big Data Governance via Big Data Lakes Data governance refers to a collection of models, methodologies and paradigms for supporting the so-called data-driven digital society i.e., a futuristic vision in which the major societal processes e.g., demography citizen services, public administration, and so on should be driven by historical and current data analysis used as a decision-making pillar. Indeed, a clear case that adheres to this vision is already visible today: the COVID-19 pandemic outbreak and the policies that countries are implementing to limit the phenomenon with these policies being entirely driven by daily data analysis of the outbreak's evolution. Are big data lakes the solution to finally enabling true big data governance for next-generation societal challenges This question identifies a critical issue that future research efforts must address.

Big Data Lakes with Privacy Protection. In contrast to ancestor data warehousing architectures, where the privacy-preservation data management problem was poorly addressed and countered privacy-preservation for big data lakes is critical, because big data applications typically deal with sensitive data e.g., social data, censor data, bio-medical data, healthcare data, and so forth. This clear evidence supports the need to truly achieve so-called privacy-preserving big data lakes i.e., big data lake architectures in which all core procedures e.g., data population, query-rewriting, metadata management, and so on are designed and implemented using privacy-preservation data management paradigms. The latter, in our opinion is one of the more important critical challenges for future-generation big data lake research.

New Big Data Lake Applications Last but not least big data lakes are typical architectures in which real-world applications induce continuous improvement of architectural components through iterative tuning and updates dictated by realistic metrics e.g., temporal complexity of big data access and query procedures accuracy of bit data analytics front-end tools, end-used usability and so forth. As a result modern application scenarios such as smart cities, e-government, healthcare management, energy consumption systems, and so on will become increasingly beneficial to the evolution of big data lake architectures over time. Thus, the research challenge is to learn effective lessons, from authoritative case studies to the foundations of

big data lake architectures from practice to theory.

IV. CONCLUSIONS

In this paper, we explored actual challenges of big data lakes and provided some future research directions in the investigate scientific field inspired by emerging trends that focus on big data lake models, frameworks and techniques. The most obvious conclusion of our work is that there is still a lot of research to be done in this area. We hope that this paper will serve as a useful first step toward that goal. For example, exploiting data semantics appears to be a promising direction for further improving the effectiveness and efficiency of big data lake processing possibly inspired by similar research experiences [31,32].

REFERENCES

[1] H.M. Sabri, A.M. Gamal El-Din, A.A. Amer, and M.B. Senousy. 2020. Comparative Study of Big Data Heterogeneity Solutions. In: *A. Ghalwash, N. El Khameesy, D. Magdi, A. Joshi (eds.) "Internet of Things—Applications and Future. Lecture Notes in Networks and Systems",* vol 114, Springer, pp. 431– 439.

[2] T. White. 2015. *Hadoop: The Definitive Guide,* O'Reilly Media, 4th. ed.

[3] I. Sabek, and M.F. Mokbel. 2020. Machine Learning Meets Big Spatial Data. In: *ICDE 2020,* pp. 1782–1785.

[4] S.V. Oprea, and A. Bâra. 2019. Machine Learning Algorithms for Short-Term Load Forecast in Residential Buildings Using Smart Meters, Sensors and Big Data Solutions. *IEEE Access* 7, pp. 177874–177889.

[5] G. Chatzimilioudis, A. Cuzzocrea, D. Gunopulos, and N. Mamoulis. 2013. A Novel Distributed Framework for Optimizing Query Routing Trees in Wireless Sensor Networks via Optimal Operator Placement. *J. Comput. Syst. Sci*. 79(3), pp. 349–368.

[6] R. Ramakrishnan, B. Sridharan, J.R. Douceur, P. Kasturi, B. Krishnamachari-Sampath, K. Krishnamoorthy, P. Li, M. Manu, S. Michaylov, R. Ramos, N. Sharman, Z. Xu, Y. Barakat, C. Douglas, R. Draves, S.S. Naidu, S. Shastry, A. Sikaria, S. Sun, and R. Venkatesan. 2017. Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics. In: *SIGMOD Conference* 2017, pp. 51–63.

[7] A. Labrinidis, and H.V. Jagadish. 2012. Challenges and Opportunities with Big Data. *Proc. VLDB Endow.* 5(12), pp. 2032–2033.

[8] M. Chen, S. Mao, and Y. Liu. 2014. Big Data: A Survey. *Mob. Networks Appl. 19(2*), pp. 171–209.

[9] D. Laney. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Group Technical Report.*

[10] A. Cuzzocrea, F. Furfaro, E. Masciari, D. Saccà, and C. Sirangelo. 2004. Approximate Query Answering on Sensor Network Data Streams. 2004. In: *A. Stefanidis, S. Nittel, (eds.) "GeoSensor Networks"*, CRC Press, pp. 53–72.

[11] A. Cuzzocrea, R. Moussa, and G. Xu, OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data. 2013. In: *MEDI 2013*, pp. 38–49.

[12] T. Ramalingeswara Rao, Pabitra Mitra, Ravindara Bhatt, and A. Goswami. 2019. The Big Data System, Components, Tools, and Technologies: A Survey. *Knowl. Inf. Syst.* 60(3), pp. 1165–1245.

[13] P. Russom. 2011. Big data analytics. *TDWI Best Practices Report, fourth quarter.*

[14] P. Zikopoulos, and C. Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data,* McGraw-Hill Osborne Media, 1st. ed.

[15] A. Cuzzocrea, C. De Maio, G. Fenza, V. Loia, and M. Parente. 2016. OLAP Analysis of Multidimensional Tweet Streams for Supporting Advanced Analytics. In: *SAC 2016*, pp. 992–999.

[16] N. Bikakis, G. Papastefanatos, and O. Papaemmanouil. 2019. Big Data Exploration, Visualization and Analytics. *Big Data Res. 18, art.* 100123.

[17] X. Wang, D. Qi, W. Lin, M. Yu, Z. Zheng, N. Zhou, and P. Chen. 2018. A General Framework for Big Data Knowledge Discovery and Integration. *Concurr. Comput. Pract. Exp. 30(13), art.* 100123.

[18] A. Prabhune, H. Ansari, A. Keshav, R. Stotzka, M. Gertz, and J. Hesser. 2016. MetaStore: A Metadata Framework for Scientific Data Repositories. In: *BigData* 2016, pp. 3026–3035.

[19] A. Cuzzocrea. 2006. Combining Multidimensional User Models and Knowledge Representation and Management Techniques for Making Web Services Knowledge-Aware*. Web Intell. Agent Syst.* 4(3), pp. 289–312.

[20] M.E. Coimbra, A.P. Francisco, and L. Veiga. 2019. Distributed Graphs: In Search of Fast, Low-Latency, Resource-Efficient, Semantics-Rich Big-Data Processing. *CoRR*, abs/1911.11624.

[21] S. Rangarajan, H. Liu, H. Wang, and C.-L. Wang. 2017. Scalable Architecture for Personalized Healthcare Service Recommendation Using Big Data Lake. In: *ASSRI* 2017, pp. 65–79.

[22] A.A. Munshi, and Y.A.-R.I. Mohamed. 2018. Data Lake Lambda Architecture for Smart Grids Big Data Analytics. *IEEE Access* 6, pp. 40463– 40471.

[23] H. Fang. 2015. Managing Data Lakes in Big Data Era: What's A Data Lake and Why Has It Became Popular in Data Management Ecosystem. In: *CYBER 2015*, pp. 820–824.

[24] N. Miloslavskaya, and A. Tolstoy. 2016. Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science* 88, pp. 300–305.

[25] S. Chaudhuri, and U. Dayal. 1997. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.* 26(1), pp. 65–74.

[26] A. Cuzzocrea, D. Saccà, and J.D. Ullman. 2003. Big Data: A Research Agenda. In: *IDEAS 2013,* pp. 198–203.

[27] J. Eberius, M. Thiele, and W. Lehner. 2017. Exploratory Ad-Hoc Analytics for Big Data. In: *A.Y. Zomaya, S. Sakr (eds.), "Handbook of Big Data Technologies", Springer,* pp. 365–407.

[28] P. Chopade, and J. Zhan. 2015. Structural and Functional Analytics for Community Detection in Large-Scale Complex Networks. *J. Big Data* 2, art. 11.

[29] A. Cuzzocrea, and I.-Y. Song. 2014. Big Graph Analytics: The State of the Art and Future Research Agenda. In: *DOLAP 2014*, pp. 99–101.

[30] K. Deligiannis, P. Raftopoulou, C. Tryfonopoulos, N. Platis, and C. Vassilakis. 2020. Hydria: An Online Data Lake for Multi-Faceted Analytics in the Cultural Heritage Domain. *Big Data Cogn. Comput.* 4(2), art. 7.

[31] A. Alserafi, A. Abelló, O. Romero, and T. Calders. 2020. Keeping the Data Lake in Form: Proximity Mining for Pre-Filtering Schema Matching. *ACM Trans. Inf. Syst.* 38(3), pp. 26:1–26:30.

[32] M. Bala, O. Boussaid, and Z. Alimazighi. 2017. A Fine-Grained Distribution Approach for ETL Processes in Big Data Environments. *Data Knowl. Eng.* 111, pp. 114–136.