Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 50 (002) 209-214 (2018)

http://doi.org/10.26692/sujo/2018.06.0037

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Big Data Evolution in Distributed Intelligent Systems

S. S. ZIA[++], I. MALA*, M. NASEEM, T. J. A. MUGHAL*, P. AKHTAR*

Department of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi- Pakistan

Received 20th October 2017 and Revised 24th April 2018

Abstract: The massive quantity of data available in digital form is increasing exponentially. Corporations, government agencies, and even medium-sized companies having huge datasets. These datasets are referred as Big Data but they can't be processed and analyzed through traditional techniques. A Big Data can be described through its velocity, volume, variety, value, and veracity. A data processing mechanism should be developed considering all these characteristics. Distributed Artificial Intelligence (DAI) provides an efficient way to process and analyze Big Data. DAI is basically a subset of Artificial Intelligence (AI) but it distributed nodes or agents to draw preliminary results which are then combined to develop a final solution. This paper identify the history and challenges of Big Data and DAI. The challenges of big data includes capturing, storing, searching, updating, privacy, visualizing, transferring, and analyzing data. Several DAI frameworks are available which address these challenges. Some of the most common frameworks are batch-only framework, stream-only framework, and hybrid framework. Finally identify these frameworks to find out the best option for a certain database.

Keywords: Big Data, Distributed Intelligent System, Distributed AI, Advanced Technology

1. INTRODUCTION

Big data is referred to as huge database which is not only huge in volume, but also very complicated so that traditional data processing tools and application could not deal with it efficiently. A big data usually described in three dimensions including velocity, volume and variety. Recently two more dimensions have been added to explain big data more accurately. These two dimensions include value and veracity. In recent times, the bid data is usually used for the user behavior analysis, finding correlation and predictive analysis (Marr 2018). Analysis of big data has several applications in internet search, advertising, business informatics, connectomics, urban informatics, and genomics (Mohammadi and Al-Fuqaha, 2018).

The analysis of big data involves special techniques. One of such technique is distributed intelligent system or distributed artificial intelligence (DAI). DAI is basically a subset of artificial intelligence technology which is used to solve complex decision making and learning problems. The working of DAI is very complex and depends on the nature and size of big data (Avouris and Gasser, 1992).

This paper is organized as follows. Section 2presents the introduction of DAI system and their functionality Big Data. Section 3presents the challenges faced in Big Data and DAI. In Section 4, we have discuss the different models and frameworks to process big data using artificial intelligence approach. Section5provides the conclusions and future directions of the researchers in this area.

2. DISTRIBUTED AI SYSTEM

The advancing in technology has changed lots of aspects of human life. The global information system also revolutionized with technology. Now, more than 90% of data is stored in digital form and the sizes of databases are increasing exponentially as it is becoming easy to create data. This data is then used for various purposes with the help of DAI. A DAI system usually possesses several data processing or learning processing nodes which are also referred as agents. These agents are distributed throughout the data and work autonomously. These nodes act independently add offer a partial solution. The final solution is developed through the communication between all related nodes (Avouris and Gasser, 1992). It is important to explore the history and previous works of these concepts in order to completely understand them.

Organizations are collecting information about their inventories, products, customers, and clients for different business operations and planning. The volume of data increased a lot in the late twentieth century. The concept of big data is very old but this term was popularized by John Mashey in the 1990s. The method of storing data changed with the arrival of worldwide web and volume of data in different databases started to increase enormously because internet provided a more cost-effective way to store and share data. By the end of Twentieth-century Big Data reached to the size of Gigabytes. In 2000 Peter Lyman and Hal Varian calculated the volume of global digital content which was roughly 1.5 billion gigabytes at that time. The arrival of mobile devices changed the scenario and

[++] Correspondence author: Syed Saood Zia, saood_zia@hotmail.com.
*Department of Electrical Engineering, Usman Institute of Technology, Karachi Pakistan
*Faculty of Engineering Sciences & Technology, Hamdard University, Karachi- Pakistan.

issues of privacy and security became more crucial than ever. Now data management is a professional job and it requires data scientists (Marr2018).

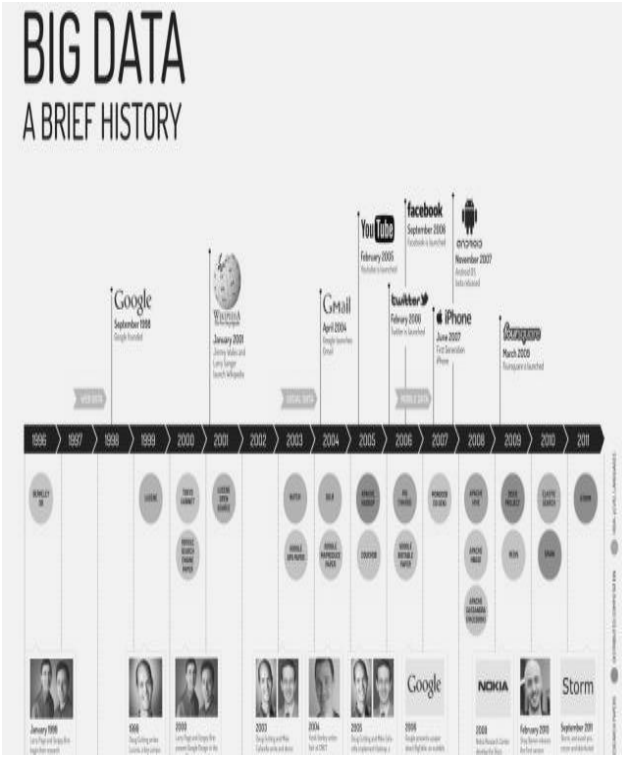

Fig.1. History of Big Data

The history of distributed artificial intelligence is somewhat linked with the history of Big Data but the concept of DAI is way older. In 1975 this concept came into existence when a unique solution was presented to solve a complex problem. This solution included multiple intelligent entities (agents) that cooperate, coexist or compete in the same environment. These entities are still referred as agents. Various techniques have been developed for DAI over the time depending on the coordination approach and activities of agents. The applications of DAI expanded with the arrival of Big Data (Bond and Gasser, 2014).

The architecture of the Big Data has gone through several modifications. In 2000 a distributed file-sharing system was introduced by Seisint Incorporation. Google introduced MapReduce architecture in 2004. This program had some limitations due to which Apache Spark was introduced in 2012. The multiple-layered architecture was also proposed in the same year. The concept of Big Data led to the conception of some other technologies like the Internet of Things (IoT) because one of the biggest reasons of increasing volume of Big Data is the increasing number of devices which can generate data (Marr,2018).

Tremendous work has been done to utilize and improve DAI. Two main approaches have emerged for DAI over the last few years. One approach is to give authority to agents to made decision autonomously and modify the state of the environment around them accordingly. The other approach is to let agents coordinate with each before making a final decision. This technology has been applied to various fields using different techniques. One of the biggest applications of DAI in recent times is for e-commerce. E-commerce websites generate massive quantity of data about the location of visitors and their activities on the website. These companies also get data about payment method, most sold product, buying time, customer's location, and lots of other things. This data can be used to develop business strategies. Telecommunication companies are using this technology to control their cooperative resources, especially for their WLAN networks (Bond and Gasser, 2014).

3. CHALLENGES

There are several challenges of big data including capturing, storing, searching, updating, privacy, visualizing, transferring, and analyzing data. Capturing of data has become very easy with the help of mobile devices and digital tools but it is also important to verify and authenticate the captured data. Various identification mechanisms and techniques like password, digital signature, and fingerprint detection can be used to ensure the authenticity of data. The inclusion of unauthentic data can lead to serious problems in some cases. For instance, the circulation of false news about any sensitive issues can trigger chaos in a community. There should be a mechanism to flag a data as false or less trusted (Heureux. *et. al.*,2017).

The storing of big data requires huge hardware and cold storage to keep hard drives running. It is very costly and requires special skills. Facebook, a leading social media website, need to process more than 500 TB data every day to store in its storage facility in North Carolina. This kind of huge data also bring issues of searching data efficiently. Google, a leading search engine, adds almost 1 Petabyte (100 million Gigabyte) every day and to keep this data searchable it uses extensive programming. The company needs huge resources to keep pace with increasing data. It is also challenging to link new data with the existing data and transferring a small portion of data on demand without compromising the privacy of remaining data. This kind of operations on big data mostly depends on the algorithms. Algorithms are also used to analyze big data in order to extract some meaning out of it. It is difficult to develop a learning algorithm, therefore, organizations are conducting workshops to discuss this type of

challenges for algorithms of Modern Massive Data Sets (MMDS) (Pulse, 2012).

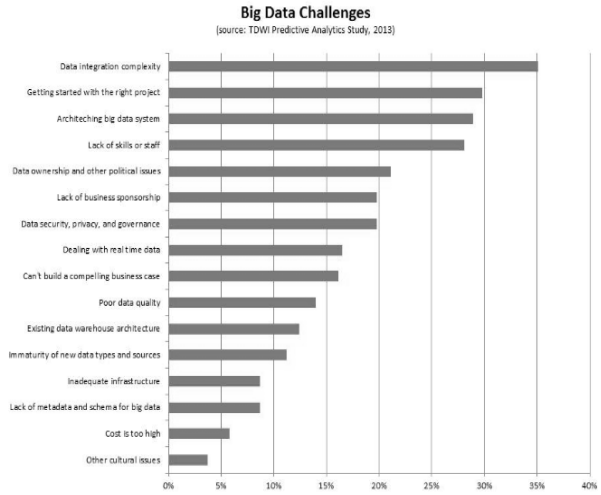

Fig.2.Big Data Challenges (Omair, 2015)

Distributed AI also has several challenges including communication, coherency, and synthesis. The biggest challenge is to decide how agents should interact with each other. What communication protocols and language is more suitable as there are several programming languages with dozens of communication protocols. It is difficult to develop a framework which works best for all kinds of databases. Another challenge is to achieve agent coherency. It is important to link the concept of DAI with the big data so that data can be utilized efficiently. Effective DAI approach can help to tackle challenges of the big data. DAI has already been used for managing big data. Google is using dynamic programs and machine learning techniques for its advertising program so that users get to see right ads and promoters get the real value of their money (Bond and Gasser, 2014).

4. MODELS AND FRAMEWORK

There are several frameworks to process big data using artificial intelligence approach. Three most popular frameworks include batch-only framework, stream-only model, and hybrid model.

4.1. Batch-only Framework

Batch-only framework divides the data into batches. It is important to have data in a permanent storage so that batches can be divided among nodes easily. The pre-defined algorithm is then applied on each node and the result from each node is then sent to a file system like HadoopDistributed File System (HDFS). Intermediary results are then combined and summarized to develop a final result which is then transferred back to HDFS. This framework can be used enormous databases but it is a fairly slow model as it requires writing and reading of data for several times. (Ellingwood, 2018).

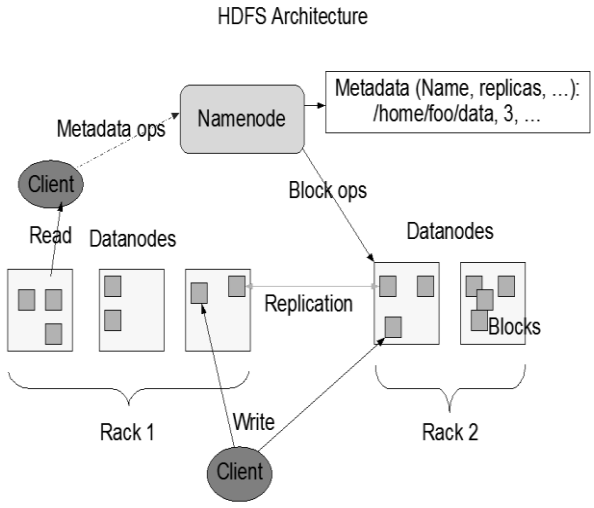

Fig. 3.Apache Hadoop Architecture

Apache Hadoop is an example of batch-only framework because Apache Hadoop provides processing for batches exclusively. It is open source system with immense popularity therefore its several versions have been released over the past few years. Modern versions include several layers. Each layer has its own nodes and work synchronously with other layers for batch processing. Each layer performs its own function. For instance, HDFS layer (distributed file system) ensures the availability of data and it stores intermediary processing results. YARN (Yet Another Resource Negotiator) is another layer which coordinates cluster and jobs in the Hadoop stack. YARN has enable Apache Hadoop to run diverse workloads which was not possible without this layer (Davies, 2017).

4.2. Stream-only Framework

The stream-only framework does not use one defined operation on entire data but uses different operation for each portion of the data. It can even define different operation for each data item whenever it passes through the system. There are lots of benefits of this framework. It can handle unlimited data because each item is processed only once. Results can be acquired almost immediately through this model. It keeps the state of the system to a minimal level unlike other approaches in which system maintain a state for the entire data. It is considered as the best available solution. The only limitation is that it does not work efficiently for already existed huge data as the data need to go through the system before getting stored (Liu 2007).

Apache storm is an example of stream processing system which is very efficient for real time processing. Apache Storm has very low latency and it works by

topologies orchestration. These topologies, also referred as Directed Acyclic Graphs (DAGs), tell what should be done with each piece of data when it enters.These topologies consist of three main components including Streams, Bolts and Spouts. Streams are referred to data which continuously arrive. Spouts are referred to the sources of the streams of data and Bolts are basically a processing step. Bolts perform suitable operation to the new stream and gives output as a new stream. Bolts connect all Spouts with each other to conduct basic operations (Liu 2007).

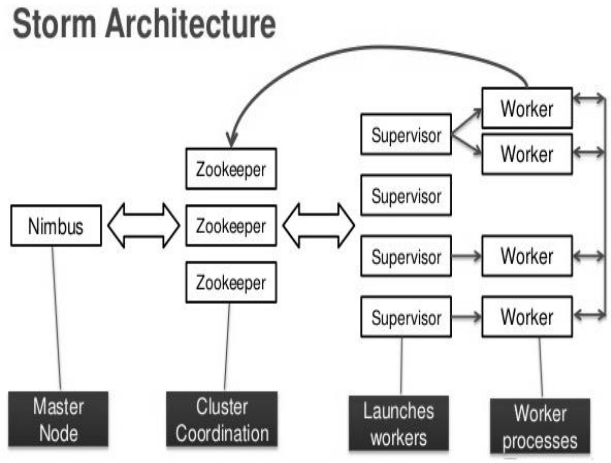

Fig. 1. Apache Storm Architecture

4.3. Hybrid Framework

There is another model which can be used to manage both batch and stream data items. This framework has different processing requirements. In this model, Spark is used for the processing of batches quickly. Spark is efficient because it has big libraries with lots of tools. Its integration is flexible and can be optimized. At the end, the state of data, available resources, and expertise of data management team decides which framework is more effective (Davies. 2017).

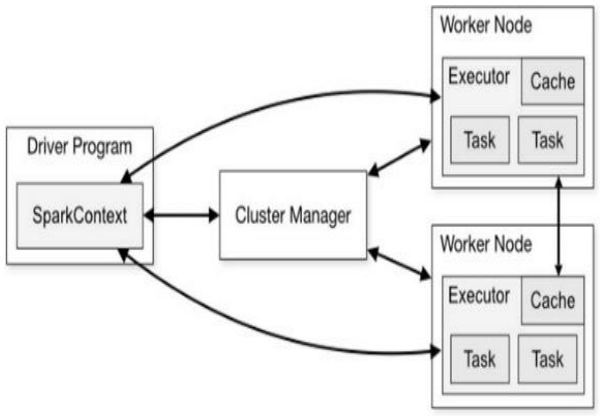

Fig.2. Apache Spark Architecture

5. CONCLUSION

The technology has converted data from analog to digital. Now it is way easier to create, collect and store data through various computing devices. The data available to governments, organizations, and individuals is increasing with the advancement of technology and the boost was stated with the arrival of the internet which gave rise to the Big Data. The creation of Big Data is also very valuable as it can be used for predictive analysis and strategy development. The main challenge is to analyze it effectively without losing its integrity. It is also important to keep data authentic especially in this era where it is super easy to create data. The analysis of Big Data requires special tools as traditional methods can't be implemented to derive valuable and trusted results. Distributed Artificial Intelligence provides an efficient way to analyze Big Data.

DAI has the capability to handle dynamic datasets with complex data items. There are different approaches to process data in this regard and its selection of right model or framework depends on the nature of data, size of database, processing requirements, available resources and expertise of the data processing team.

REFERENCES:

Avouris N. M, L. Gasser (1992) editors. Distributed artificial intelligence: Theory and praxis. Springer Science.

Big Data, (2018) "History of Big Data", Big data.black, 2016. (Online). Available: http://bigdata.black/infrastructure/platforms/history-of-big-data/. [Accessed: 15- Mar- 2018].

Bond, A. and L. Gasser, (2014).*Readings in Distributed Artificial Intelligence*. Burlington: Elsevier Science.

Borthakur. D. (2013). "*HDFS Architecture Guide*", Hadoop, (Online). Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Accessed: 25- Mar- 2018].

Davies. B. (2017). "5th Best Data Processing Frameworks", *Knowledge Hut Blog, 2017*. (Online). Available: https://www.knowledgehut.com/blog/information-technology/5-best-data-processing-frameworks. [Accessed: 24- Mar- 2018].

Ellingwood. J, (2018) "Hadoop, Storm, Samza, Spark, and Flink: Big Data Frameworks Compared Digital Ocean", *Digitalocean.com*, 2018. [Online]. Available: https://www.digitalocean.com/community/tutorials/hado

op-storm-samza-spark-and-flink-big-data-frameworks-compared. [Accessed: 14- Mar- 2018].

Heureux.A. L., K. Grolinger, H. Elyamany and M. Capretz. (2017)."Machine Learning With Big Data: Challenges and Approaches", IEEE Access, vol. 5, 7776-7797.

Liu. Y. (2007). *Query optimization for distributed stream processing*. [Bloomington, Ind.]: Indiana University,

Leonard. A. (2018). "Apache Storm Stream Processing in Azure HD Insight", (Online). Available: https://andyleonard.blog/2018/01/apache-storm-stream-processing/. [Accessed: 25- Mar- 2018].

Mohammadi. M. and A. Al-Fuqaha, (2018). "Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges", *IEEE Communications Magazine*, vol. 56, no. 2. 94-101,

Marr, B. (2018). "A brief history of big data everyone should read", World Economic Forum, 2015. (Online). Available: https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/.

Omair B, A. Emam Towards Big Process Mining. ALLDATA 2015. 19:42.

Pulse UG. Big data for development: Challenges & opportunities. Naciones Unidas, Nueva York, Mayo. 2012 May.