Big IoT Data Analytics: Literature Review, Opportunity and its Research Challenges

[1]Sandeep Bhargava, [2]Dr. Dinesh Goyal

Research Scholar, Suresh Gyan Vihar University, Jaipur
Professor, Center for Cloud Infrastructure & Security, Jaipur.

Abstract-*IoT data has increased exponentially due to new developments in the field of Information Technology .IoT data which is generated from various devices and sensors such as temperature, motion or sound etc have significant gaps, noisy, highly unstructured & generated at large volume of different variety. Thus it is difficult for Business Intelligence & general analytics tools those available or designed to process big IoT data. The solution to the above problem can only be achieved through a Big IoT Data Analytics. This paper reviewed proprietary and open sourced Big IoT data analytics plateforms and bring forward a comparative analysis with objective to help for further research in the area of Big IoT Data Analytics.*

Keywords: IoT, Big Data, Big IoT Data,Big IoT Data Analytics, IoT analytics challenges. Review literature of IoT Data Analytics.

I. Introduction:

Internet of Things (IoT) data has grown exponentially in large numbers and at high speeds, and it affects all areas of technology and business by increasing the interests of organizations and individuals. The Statistical report [5] shows that by 2020, 12.86 billion IoT sensors and devices will be used by the consumer community and as per CAGR 2017 it is having 34.89% annual growth . By 2020, the global market value of the Internet of Things is expected to reach 7.1 trillion US dollars [8] records shows that the number of sensors will increase by 1 trillion in 2030. This growth will affect the growth of large IoT data. The size of the Internet of Things -analytics market [6] is expected to grow from $7.19 billion in 2017 to $277.8 billion in 2022, with a compound annual growth rate (CAGR) of 31.0%.

However IoT Analytics is now one of the most demanding & emerging markets globally and increasing pressure to find reliable Big IoT Data Analytics tool find actionable value from IoT data.

1. IoT: IoT(Internet of things) refers to the network of things/devices which are connected to internet and embedded with sensors, software and necessary electronics that enables them to collect and exchange data. In 1999, the word "The Internet of Things" was initially coined by Kevin Ashton and later regarded as a father of Internet of Thing. Internet of things is now a new source of data that's initially aims to make decision on resource utilization and increase the efficiency of resource.

2. Big IoT Data: Massive amount of data which is generated from sensors at high speed of different variety is termed a Big IoT data. Sensors appeals across almost every sector of industry, the internet of things is going to trigger a massive influx of big data is termed as big IoT data. IoT(Internet of Things) & Big Data, even though two different domains , are closely intertwined, to talk about IoT without Big Data is not so much relevant for the end users and technocrats.

3. Big IoT Data Analytics: The rapid equipment of IoT devices, will contributing Big IoT Data but without analytics, the Big IoT Data would be like trying to hear a single voice in a crowd of millions. Also continuous need of organization, pressure to have insight value from Big IoT Data to take competitive advantage and for the betterment of life.

This is where big data analysis enters the picture. Big IoT data analytics can handle large amounts of data generated from IoT devices, creating a continuous stream of information. Big data analysis has been applied in many fields and domains, some of these applications include medical research, solutions for the transportation and logistics sectors, global security, and prediction and management of issues related to the socio-economic and environmental sectors.

3.1 How IoT analytics is different from other analytics? Analytics is a tool that realize value from the huge volumes of data generated by connected Internet of Things devices and transforming it into actionable intelligence. The domain of analytics is depend on the type of data and and type of knowledge extract it. The IoT Data analytics is never similar to mobile analytics nor it is similar to Web analytics or Log analytics.

4. Big IoT Analytics Platforms: Many solutions can be accessed for big IoT data analytics. Some examples include proprietary solution i.e AWS IoT Analytics, Micorsoft Azure IoT Suit, IBM Watson IoT Analytics, Splunk Big Data Analytics and some open source solution i.e FIWARE3, OpenMTC4, SmartThings5, Hadoop with Map Reduce, Hive and Spark. All the above mentioned solutions allow us to connect different devices, followed by accessing their data and processing the same data, by using the knowledge gained via these steps in-order to create automated sytem to control the devices but due to lack of standardization for IoT, all the IoT platforms and tools using different terminology and different

technology concept of implementation. As a result of the same, the platform and the tools are not homogeneous. Therefore finding the right platform to implement IoT based solution becomes quite time consuming especially when each solution uses different technologies and platforms.

II. Review Literature

Here we review existing literature available on Big IoT data analytics and will find challenges and technological barriers in Big IoT Data Analytics.

1 IBM Watson IoT Platform & Analytics: IBM Watson IoT Platform for IoT analytics allows us to perform powerful device management operations, by connecting a wide variety of devices and gateway devices, and store and access device data.

IBM Watson Analytics (WA) is widely recognized in the Business Intelligence (BI) domain. IBM Watson Analytics (WA) is an advanced data analysis and visualization solution in the IBM cloud for analyzing and discovering hidden values

IBM Watson Analytics enables to write powerful query for a variety of databases including Cloudera Impala, MySQL, Oracle, PostgreSQL, PostgreSQL on Compose, Structured Query Language (SQL) Server, Sybase, Sybase IQ, and Teradata.

Data Science Experience (DSX) is a powerful machine learning tool IBM Watson that along with IBM Watson IoT Platform , is used to visualize and learn about the data that is sent from devices that are connected to the platform.

Limitation of IBM Watson Analytics is that it does not do streaming data analytics

2. AWS IoT Platform & Analytics: AWS IoT Core is a platform that enables us to connect IoT-equipped devices to AWS services, it also supports to protect data and interactions between data, process and device, and provides interaction with devices even when they are offline.

AWS IoT Analytics is a fully managed service that makes it easy to analyze large amounts of IoT data without worrying about the cost and complexity typically required to build an IoT analytics platform. AWS IoT Analytics is a fully managed service that automates the analysis and scaling to support IoT data up to several petabytes

3. Microsoft's Azure IoT Platform & Analytics: Azure IoT Hub securely connects, monitors and manages billions of devices to develop Internet of Things (IoT) applications. IoT Hub is a flexible cloud platform that supports open source SDKs and multiple protocols.

For analytics purpose, Microsoft Azure provides many tools for according to specific purpose e.g Azure Databricks, HDInsight, Data Factory, Machine Learning, Data Lake Analytics, Stream Analytics, Azure Analysis Services, Azure Data Explorer

4. Tableau Big Data Platform & Analytics: Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tool.
Tableau Big Data platform enables to collect, store and manage more data than ever before. Tableau's analytics platform incorporates enterprise-grade security, governance, deployment flexibility and management to empower IT.

Tableau Analytics empowers organization to maximize the value of its data and people.Tableau reduces the need to pull individual reports from multiple software or databases

A wide range of analytics solutions provide by Tableau e.g Tableau Desktop, Tableau Prep, Tableau Server, Tableau Online.

5. Splunk Big Data Platform & Analytics: Splunk is a powerful platform for analyzing machine data.. But now it is becoming more and more important in the technical and commercial fields [5]. Splunk is different from everyone else, except that it generates an index for the data, similar to the index generation of text. It is not exactly an AI routine collection or report generation tool, even though it implements most of the functionality in this process. This kind of indexing is very flexible, thus making Splunk an tunable platform for applications, so they can be understood and absorbed from the log files. Splunk is marketed through a variety of solution packages such as Microsoft Exchange Server Monitoring and Web Attack Detection. This index is very useful for correlating data in many common server-side scenarios. Splunk aims to get a text string and search roughly in the indicator. For example, Splunk gets the URLs or IP addresses in the document and packages them into a timeline that is built around when the data is detected. All other related fields are used to drill down into the data set. Although this is a simple process, it is very effective if the user searches for the correct type of needle in the data source. If the user can find the correct text string, Splunk is very useful for tracking it. The log file is a huge application. Currently, a new Splunk tool called Shep is used for private beta to provide two-way integration between Hadoop and Splunk, allowing users to exchange data between systems and solidify Splunk data from Hadoop [2].

6. Apache Hadoop: Apache Hadoop is an open source software framework written in Java for distributed storage and distributed processing of large data sets on clusters of commercial hardware built in a reliable, fault-tolerant manner. The core of Apache Hadoop includes a storage part (Hadoop Distributed File System (HDFS) and a processing part (MapReduce). HDFS is located at the bottom of the Hadoop stack. It is a distributed file system. MapReduce is a programming model for programming. Distribution mode handles large

datasets.map() and reduce() are two functions of the model that apply mapping operations to the data in the partition of the HDFS file, sorting and redistributing the results based on the key values in the output. The data then performs a reduction operation on the output data item group using the matching key from the mapping phase of the job.

7. Apache Hive: Hive is a data warehouse infrastructure for processing structured data in Hadoop. It is built on top of Hadoop to aggregate big data and make querying and analysis simple. Hive provides a SQL-like interface for querying data stored in various databases and file systems integrated with Hadoop. The data is stored in the traditional RDBMS format. Hive uses a query language such as SQL, called HiveQL, which provides a mechanism for querying data. Traditional MapReduce programmers are also allowed in Apache Hive because they are inefficient or inconvenient to express custom mappers and reducers in Apache HiveQL. Map Reduce doesn't have optimization and usability features like UDF, but the Hive framework does. An important component of Hive is the Metastore, which is used to store schema information. This metastore usually resides in a relational database. We can interact with Hive using Web GUI, Java Database Connectivity (JDBC) interface, etc.

8. Apache Spark: Apache Spark is an open source distributed cluster computing framework. It was originally developed at AMPLab at the University of California, Berkeley, and was later donated to the Apache Software Foundation. Apache Spark is an open source, in-memory data processing engine that processes real-time streaming data. Apache Spark uses the most advanced DAG scheduler, query optimizer and physical execution engine to achieve high performance for batch and streaming data.

III. COMPARATIVE REVIEW: This section gives a detail on the comprehensive review of of Proprietary data analytics platforms.

3.1 Proprietary Big Data Analytics Platforms for IoT Data

Big Data Analytics	Commercial/Open Source	Types of Data	Analytics Type Support	Cloud Support or not	Visulaization Support	Storage	Support ML Algorithm	Language Support	Data Collection Protocol	Security
IBM Watson IoT Analytics	Commercial	Both structured and unstructured	Edge Analytics, Offline Analytics	Yes	Yes	Cloudant NoSQL DB	▪ Usage Spark ML Classification & Regression Supported Only ▪ scikit-learn ▪ XGBoost ▪ TensorFlow	R, Python, SQL	HTTP or MQTT protocols.	HTTPS, TLS
AWS IoT Big Data Analytics	Commercial	Both structured and unstructured	Yes	Yes	Yes (AWS IoT Dashboard)	Dynamo Db, S3	▪ Binary and multi-class Classification ▪ Regression	No*(Uses In-Built Tools)	MQTT, HTTP 1.1, or WebSockets protocols	Link Encryption (TLS), Authentication (Sig V4, X.509)
Microsoft Azure IoT Suite	Commercial	Both structured and unstructured	Real Time Analytics In Memory Analytics Offline Analytics Machine Learning	Yes	Yes(PowerBI for visualization.)	SQL Database, SQL Data Warehouse, and Document DB for storage,	▪ Supervised Binary and multiclass classification, Regression, Decision Tree, Forest, Boosted, Bayes, SVM, Neural Network ▪ Unsupervised: K-Means ▪ Anomaly detection ▪ Text Analytics	R, Python, SQL	Yes	Yes

Tableau Big Data Platform Analytics	Commercial	Both structured and unstructured	In-Memory and Disk Analytics	Yes	Yes	MemSQL, Exasol,	All types of Machnine algo	R	Yes	Yes
Splunk Big Data Analytics	Commercial	Both structured and unstructured	Real Time Analytics	Yes	Yes	Couchbase, MongoDB, and Apache Cassandra	All types of Machnine algo	R, Python, SQL	Yes	Yes

3.2 Open Source Big Data Analytics Platforms

Parameter\Big Data Analytics		**Hadoop**	**Hive**	**Apache Spark**
Types of Input Data		Non Structured	Structured & Semi Structured	Non Structured
Storage		HDFS	Used RDBM Model	Used Non-SQL DBMS
Processing	**Processing Engine**	Map Reduce	Hive	Spark
	Types of Processing	batch processing	batch processing	Both Batch & Real Time Processing
Analytics	**Analytics Method**	Java based embedded query	Hive-SQL	1.Spark-SQL 2. Graph Analysis 3. Machine Leaning Algo.
	Types of Analytics	Batch Analysis	Batch Analysis	In Memory Analysis & Batch Analysis Both
Visualization		No, Third party software is required	No, Third party software is required	Yes
Cloud Support or not		both in the cloud and on-premises	both in the cloud and on-premises	both in the cloud and on-premises
Scalability		Manual	Manual	Manual
Language Support		Java	C++, PHP, Java, Python etc.	Java, R, Python & Scala
Access Method		JDBC	ODBC, JDBA & Thrift	ODBC, JDBA
Ability to connect with IoT devices and can do IoT Analytics		Yes	Yes	Yes

IV. Technology Barriers and challenges to IoT Analytics

- Technology that was never designed for IoT analytics, is trying to fit with available IoT analytics are driving a major contribution to the high rate of failure of IoT projects.
- Traditional approaches to data integration struggle to keep up with the volume and velocity of IoT data generated by sensors and devices that needs to be processed and analyzed immediately.
- The characteristics of IoT data, poses different set of challenges that require data storage management solutions, to facilitate rapid decisions, irrespective of how many end points are involved.
- Current IoT data analytics capabilities are insufficient to analysis complex IoT data that take accurate decisions with time lag as performance parameter.

References

[1]. https://www.happiestminds.com/Insights/internet-of-things/
[2]. http://www.faz.net/aktuell/wirtschaft/diginomics/grosse-internationale-allianz-gegen-cyber-attacken-15451953-p2.html?printPagedArticle=true#pageIndex_1
[3]. Nordrum, Amy (18 August 2016). "Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated". *IEEE*.
[4]. Hsu, Chin-Lung; Lin, Judy Chuan-Chuan (2016). "An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives". *Computers in Human Behavior*. **62**: 516–527. doi:10.1016/j.chb.2016.04.023

[5]. https://www.forbes.com/sites/louiscolumbus/2018/06/06/10-charts-that-will-challenge-your-perspective-of-iots-growth/#411c95493ecc
[6]. https://www.marketsandmarkets.com/Market-Reports/iot-analytics-market-52329619.html
[7]. https://azure.microsoft.com/en-in/product-categories/analytics/ Chen, M., et al., Related Technologies, in Big Data. 2014, Springer. p. 11-18.
[8]. Chen, C. L. P. and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275 (2014) 314347.
[9]. O'Driscoll, A., Daugelaite, J. and Sleator, R. D. (2013). 'Big Data', Hadoop and Cloud Computing in Genomics. Journal of Biomedical Informatics. Volume 46, Issue 5, October 2013, pp. 774-781
[10]. https://data-flair.training/blogs/apache-spark-machine-learning-algorithm/
[11]. https://azure.microsoft.com/en-in/overview/iot/
[12]. https://www.tableau.com/solutions/big-data
[13]. https://www.tableau.com/learn/whitepapers/tableaus-vision-big-data
[14]. Pouria Pirzadeh, Michael Carey, Till Westmann," A Performance Study of Big Data Analytics Platforms", 2017 IEEE International Conference on Big Data (BIGDATA)
[15]. https://www.splunk.com/