RESEARCH ARTICLE OPEN ACCESS

Big IoT Data Analytics: Architecture, Tools and Cloud Solutions

POONAM

Department of Computer Science & Application, A.S.College - Khanna

ABSTRACT

Internet of Things (IoT), Cloud computing and big data has been identified as emerging technology in information technology. The growth of IoT is continuously expanded in the era of technology. The IoT can be viewed as a gigantic network consisting of network of devices and computers connected through a series of intermediate technologies. The IoT is getting smarter every day through the convergence of digital and physical world. The number of connected devices will be 30.7 billion devices in 2020 and 75.4 billion in 2025. Cloud computing gives the centralized platform to access the data from anywhere in the world with the shared infrastructure. The combination of Cloud Computing and Internet of Things build a robust, maintainable, end-to-end Internet of Things solution on cloud platform. Also with the pervasive use of IoT devices, a massive amount of big data generation needs a high storage process and high processing speed. So the cloud brings the advantage of vast storage space. Cloud enhances the security and privacy of IoT data. Such numerous amounts of data are not meaningful without analytic power. Big data and data analytical are crucial to the effective functioning of IoT. So IoT analytics is the application of data analysis tools and procedures to realize value from huge volume of data generated by connected IoT devices. This paper aims to the relationship between IoT and big data analytics, its architecture and tools. It also defines different cloud IoT solutions such as AWS IoT platform, Google Cloud IoT platform, IBM Watson IoT, Microsoft Azure IoT platform to perform real time analytics for high streaming of IoT devices data. A developer can adopt any adopt any platform depending on its use case

Keywords: IoT, Big data, Big data Analytics, AWS IoT, Amazon IoT, Azure IoT , Google Cloud IoT

I. INTRODUCTION

Internet of Things aims to create world where all the devices and appliances are actually connected to the network and are used collaborate to achieve complex task that require a high degree of intelligence. It is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. Over the past few years, IoT has become one of the most important technologies of the 21st century. It enables the communication between electronic devices and sensors through the internet in order to facilitate our lives.

The term IoT was coined by Kevin Ashton in 1999. The main idea behind the suggestion is based on the advancement in the field of interconnected sensors and their continuous annual growth. These devices are getting smarter as they are able to communicate with each other directly to process their generated data locally for smart decision making based on defined algorithms. Considering the smartness and importance of IoT, multiple organizations like IBM, Cisco and Microsoft are working for the development of devices in this domain [1]. Now that we can connect everyday objects—kitchen appliances, cars, thermostats, baby monitors—to the internet via embedded devices, seamless communication is possible between people, processes, and things.

The rise of IoT as a primary data contributor in big data application has posed new data quality constraint. IoT data is different from traditional data and issues assuring its quality are likewise distinct. IOT are the one of major source of big data. As per the study, around 4.4 trillion GB of data will be generated by the year 2020 through the Internet of Things. So using Big Data we can manage those huge data which we got from IoT devices and other sources. For making real time decision, IoT is driving big data Analytics.

By means of low-cost computing, the cloud, big data, analytics, and mobile technologies, physical things can share and collect data with minimal human intervention. From a storage solution to accessing data remotely IoT and cloud computing builds an integration. Many cloud services such as Amazon Web Services, IBM Watson, Microsoft Azure, and Google Cloud are integrated with IoT which enables to connect process and secure data with cloud application. IoT big data analytics as steps in which variety of IoT data are expanded to reveal trends, unseen pattern, hidden correlations and new information [2]. It is the science of applying statistical technique to large dataset to obtain actionable insight for making smart decisions. It is process to uncover hidden patterns, trends and any other useful information.

II. IoT AND BIG DATA

A. *Components of IoT System*

Internet of Things is the network of physical objects or things embedded with electronics, software, sensors and network connectivity which enables these objects to collect and exchange data. IoT does not work on a single technology, it work with a group of technologies which can address different nodes like identification, sensing, information processing etc. The four major components of IoT that defines how IoT works is depicted in figure 1.

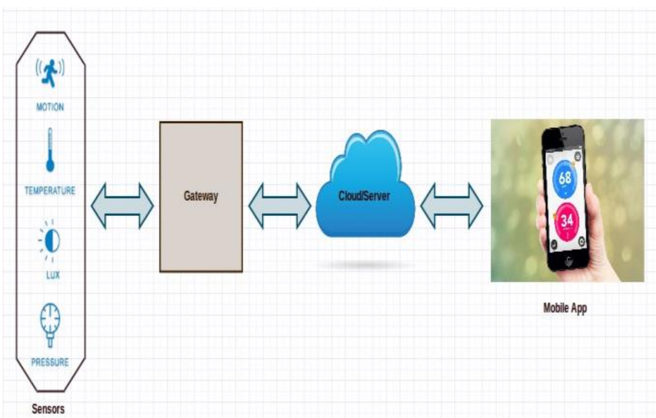

Figure 1: Internal Working of Internet of Things [3]

1) **Sensors/ Devices:** Sensors or devices are a key component that helps you to collect live data from the surrounding environment. All this data may have various levels of complexities.
2) **IoT Gateways: I**oT gateways are the gateways to the internet for all the devices that wants to interact with. These communication mediums include mobile or satellite networks, Bluetooth, WI-FI, WAN, etc. Gateways help to bridge the internal network of sensor nodes with the external Internet or World Wide Web. They do this by collecting the data from sensor nodes. All these collected data is then sent to a cloud infrastructure. The sensors should be connected to the cloud using various mediums of communications.
3) **Cloud/server infrastructure:** The data transmitted through gateway is stored & processed securely within the cloud infrastructure using Big Data analytics engine. Once that data is collected, and it gets to the cloud, the software performs processing on the gathered data.
4) **User Interface:** The information needs to be available to the end-user in some way which can be achieved by triggering alarms on their phones or sending them notification through email or text message. The user sometimes might need an interface which actively checks their IoT system.

B. *Big Data*

Big data is a combination of structured, semi structured and unstructured data collected by organizations that can be mined for information. It is the massive amount of data that cannot be stored, processed and analyzed using traditional tools. Big data and IoT are complementary to each other. Hence the need for big data in IoT is compelling. Hadoop is at the core of many today's big data implementation. Big data has three important characteristics:

1) **Volume**: Volume represents enormous amount of data that is produced. Today data is generated from various sources in different formats. Volume of data is generated exponentially. New big data tools used distributed system so that we store analyze data across databases that are dotted around anywhere in the world.
2) **2. Velocity:** Velocity refers to the speed at which the data is generated, collected and analyzed. This is mainly due to IOTs, mobile data, social media etc. Technology allows us now to analyze the data while it is being generated, without ever putting it into databases. Smart objects can generate machine and sensor data at a very fast rate.
3) **Variety: -** Variety refers to nature of data that is structured, semi-structured and unstructured data. In past we focus only on structured data but in fact 80% of data is unstructured which is in the form of text, video, image etc. With big data technology we can now analyze and brings together data of different types.

C. Integration of IoT and Big Data

Over 50 billion devices ranging from smart phones, laptops, sensors, and game consoles are anticipated to be connected to the Internet through several heterogeneous access networks enabled by technologies, such as radio frequency identification (RFID) and wireless sensor networks. In the model of IoT, sensors are used to collect and transmit data all around the world. These sensors generate increasingly growing data, which tends to form a vast heterogeneous dataset. As IoT devices generate the massive amount of heterogeneous data on a real time basis. There are multiple storage techniques like Microsoft Azure that can handle big data. Nowadays, IoT is used in everything starting from home automation, agriculture, healthcare, transportation etc. The data produced by the IoT is growing rapidly because of the large scale development of various applications [4]. So Big data is a solution of this problem. Big data is a term for collection of dataset so large and complex that it becomes difficult to process using traditional database system. The figure 2 depicts the integration of big data through various IoT devices and other sources.

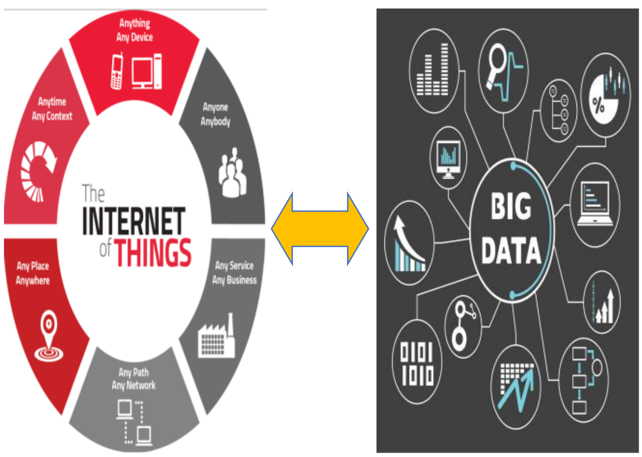

Figure 2: Relationship between IoT and Big Data

Managing the data and extracting information from it is a very vital task associated with IoT. An appropriate analytical platform is required to enable to derive knowledge from IoT data. This data needs to be stored and processed such that the quality of data does not compromise. To maintain the quantity and mutual relations of such extensive data, existing enterprises of IT have to improve their architectures and infrastructures [5].

III. BIG DATA ANALYTICS

Big data analytics is emerging as a key to analyzing IoT generated data from connected devices which helps to take the initiative to improve decision making. The main goal of big data analytics is to help organization to make better business decision, future prediction, analysis large number of transactions that done in organization. IoT will enable big data, big data needs analytics and analytics will improve processes for more IoT devices. Applying real time analytics is the need in IoT environment [6].

Big data analytics in IoT requires processing of large amount of data and store the data using various storage technologies. Big data analytics helps to make sense of data and information that is gathered by IoT devices. Big data analytics require technologies and tools that can transform a large amount of structured, unstructured, and semi-structured data into a more understandable data and metadata format for analytical processes. After analyzing the data, these tools visualize the findings in tables, graphs, and spatial charts for efficient decision making. Thus, big data analysis is a serious challenge for many applications because of data complexity and the scalability of underlying algorithms that support such processes. Deep learning is very effective to deal with such large information and can provide the results with high accuracy.

A. IoT and Big Data Analytics

IoT data is one of major data resources now and IoT data analytics has become a paradigm in big data era. IoT analytics refers to the analysis of data from multiple IoT data sources, including sensors, actuators, smart devices, and other internet connected objects.

The big data analytics lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches. Big data analytics can consist of different software pieces that collect, store, manipulate and analyze all different data types. For example big data analytics is used to expose unseen patterns, undefined correlations, market trends, customer preferences and additional useful information that can help organizations that make more informed business decisions. The basic relationship between IoT and Big data analytics which is shown in figure 3 is divided into three steps which are explained as below.

Figure 3: Relationship between Big Data Analytics and IoT

- The step 1 is created to manage IoT data sources, where associated sensor devices communicate with each other using applications. For illustration, gadgets like CCTV cameras, smart devices, and smart traffic lights communicate with each other, producing huge data sources in diverse formats. Such data can be saved on cloud in low cost storage of commodities [7]
- In step 2 a large amount of generated data through these IoT devices refers to as big data which is collected in big data file system.
- The last step comprises various analytics tools such as Hadoop, MapReduce, and Spark etc. that can analyze the stored big IoT data sets. After analyzing the data these tools are used to visualize these outcomes in the form of tables, graphs etc.

B. Big IoT Data Analytics Architecture

The architecture concept of IoT has several definition based on IoT domain abstraction and identification. It offers a reference model that defines relationships among various IoT verticals, such as, smart traffic, smart home, smart transportation, and smart health. Smart objects in Io can generate both structured and unstructured data. The architecture for big data analytics offers a design for data abstraction [8]. The proposed architecture, which integrates IoT and big data analytics, is depicted in figure 4.

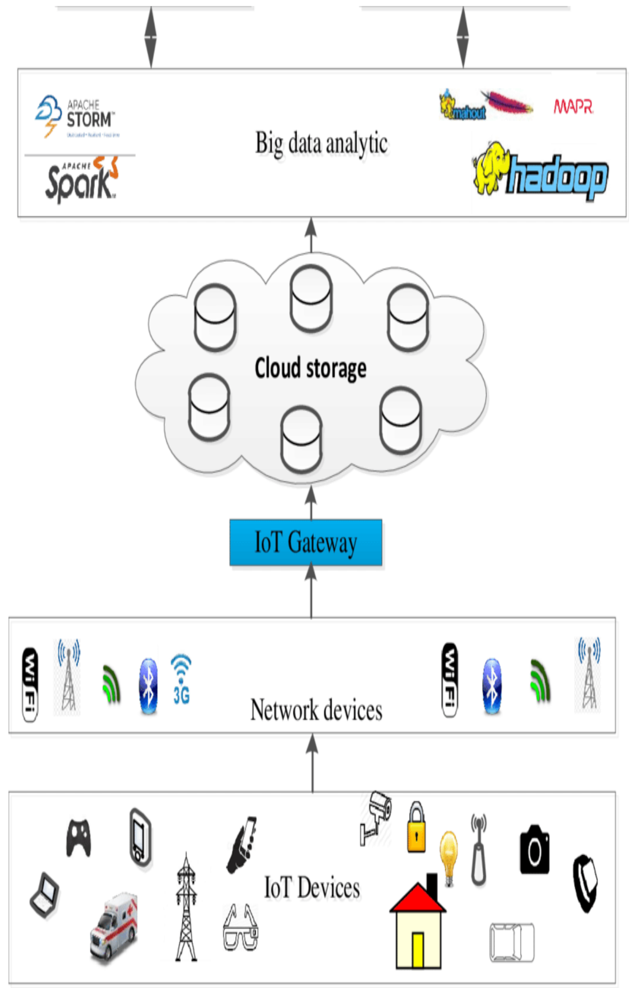

Figure 4: Big IoT Data Analytics Architecture

In this figure sensor layer contains all the sensor devices and the objects which are connected through a wireless network. This wireless network communication can be RFID, Wi-Fi, ultra wideband, ZigBee, and Bluetooth. The IoT gateway will provide communication of the Internet and various webs. The upper layer concerns big data analytics, where a large amount of data received from sensors are stored in the cloud and accessed through big data analytics applications. Big data analytics examines large and different types of data to uncover hidden pattern, correlations and other insights. These applications contain API management and a dashboard to help in the interaction with the processing engine. So Big data analytics is an emerging tool that is used for analyzing the data created by various IoT devices.

IV. BIG IoT DATA ANATYTICAL TOOLS

Big data analytics tools and technologies provide a mean of analyzing big data and sketch conclusions about them which helps organizations to make decision. It uses techniques such as statistics, machine learning, data mining etc. requires many tools to gather insights from data.

A. Hadoop

Hadoop is an open-source framework that is written in Java and it provides cross-platform support. Apache HADOOP is a framework used to develop data processing applications which are executed in a distributed computing environment. It works in environment that provides distributed storage and computing across clusters of computers. There are two core services which Hadoop provides:

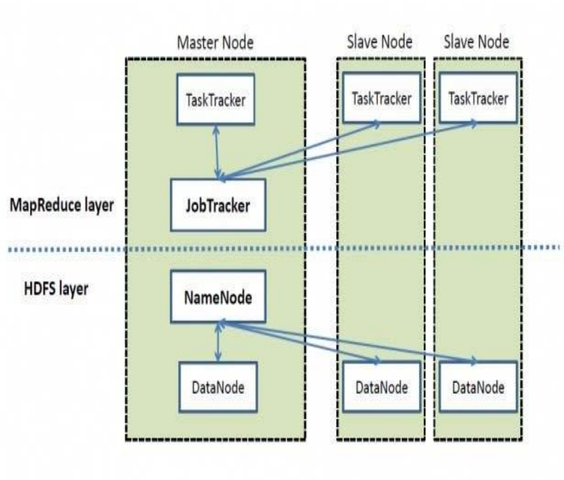

Figure 5: Hadoop Architecture

1) **Hadoop MapReduce:** -MapReduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing massive data in parallel on large clusters of computation nodes [9].
2) **HDFS: -** HDFS is main part of Hadoop as it provides reliable means for managing big data. It was closely related to MapReduce. When HDFS takes the data it breaks the information down into separate blocks and distributes them into different nodes in a cluster. It employs NameNode and DataNode architecture to implement distributed file system that provides high performance access to data highly scalable Hadoop clusters.

B. Apache Spark

Apache Spark is one of the most successful tools by Apache Software Foundation. Apache Spark is a powerful open source processing engine built around speed, ease of use, and sophisticated analytics. Spark APIs are simple, intuitive, and expressive. Spark is designed to cover a wide range of workloads including batch jobs, iterative algorithms, interactive queries, machine learning, and streaming. The main components of Apache are depicted in figure 6.

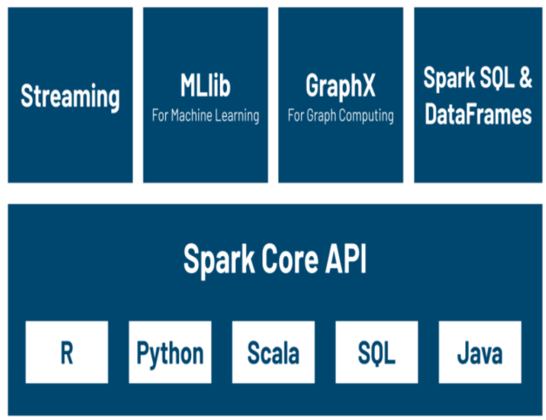

Figure 6: Apache Spark Components

Spark maintains MapReduce's linear scalability and fault tolerance, and extends the processing capabilities in four important areas: in-memory analytics, data federation, iterative analytics, and near real-time analytics [10]. It is an open source cluster computing framework for real time processing. Companies such as Oracle, Horton works, Oracle uses Apache Spark for real time computation of data with ease of use and speed. Recently Apache Spark has released 2.4.5 version with the following new features:

1) Provides high level APIs in Java, Scala, Python and R. Spark code can be written in any of these four languages.
2) Spark ml library that could be machine learning library component is handy when it comes to big data processing.

C. Splunk

Splunk is a platform used to search, analyze and visualize the machine generated data. Many organizations such as Lenovo, Intel are using Splunk in day to day practices to discover the processes and correlate data in real time. According to Splunk official website, Splunk aims to bring data to every part of your organization by helping:

1) To prevent and predict problem with monitoring experience.
2) Detect and diagnose issues with clear visibility.
3) Explore and visualize business process and streamline the entire security stack.

D. Knime

Knime is a free and open source data analytics reporting and integrating platform built for analytics on GUI based workflow. Knime provides two software's Knime Analytics Platform and Knime Server. Knime Analytics Platform is an open source and used to clean and gather data make reusable component accessible to everyone and create data science workflow. Knime Server is a platform used by enterprises for deployment of data science workflows, team collaboration, management and automation. Companies such as Telecom, Continental used Knime to make sense of their data and leverage meaningful insights.

E. RapidMiner

RapidMiner is a platform for data processing, building machine learning models and deployment. Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It means that we don't have to do the coding for data mining tasks. Rapidminer is one of the most popular data science tools. This is the graphical user interface of the blank process in rapidminer. It has the repository that holds our dataset. We can import our own datasets.

F. Talend

Talend is the one of the most powerful data integration ETL tool available in the market. It an open source platform which offers data integration and management. This tool lets you to easily manage all these steps involved in the ETL and aims to deliver complaint accessible and clean data for everyone. Talend is only platform to deliver complete and clean data at the moment you need it.

G. ThingsBoard

ThingsBoard is an open source IoT platform for collecting, processing, analyzing and visualizing IoT sensor data. It is a scalable, fault tolerant tool that is used for high performance computing. The toolkit supports both on-premise and cloud deployments. ThingsBoard node is the toolkit's core service which is responsible for transferring data using rest API calls [11].

H. ThingSpeak

ThingSpeak, offers non-commercial open source solutions that can visualize IoT device data using MATLAB widgets. IoT sensor data transferred to the ThingSpeak cloud using restful APIs and HTTP protocols can be analyzed and visualized for more in-depth insights using MATLAB software. ThingSpeak also has a paid commercial toolkit, but its open source and free-to-use solutions that work alongside MATLAB computational algorithms are more than well-suited for performing the fundamental IoT data analysis and visualizations.

V. BIG IoT CLOUD SOLUTIONS & ANALYTICS

Many solutions can be accessed for big IoT data analytics such as AWS IoT platform, Microsoft Azure IoT suit, IBM Watson IoT and Google Cloud IoT. The description of these platforms and their architecture is explained as below.

A. AWS IoT Platform & Analytics

AWS IoT (Amazon internet of things) is an Amazon Web Services platform that collects and analyzes data from internet connected devices and sensors and connects the data to AWS cloud solutions. AWS IoT can collect data from billion of devices and also connect them to endpoints for other AWS tools and services allow the developer to tie that data into an application. The figure 7 depicts the analytics of data with AWS platform.

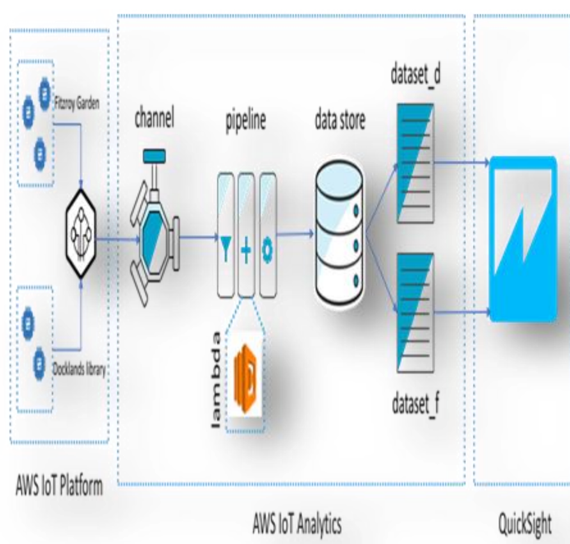

Figure 7: IoT analytics with AWS Platform [12]

AWS IoT Analytics has four components i.e. channel, pipeline, data store and dataset. Channel is used to store raw data and pipeline is used for transformation of data. Data store is used to store the processed data from the pipeline. Once the data is stored in data store this is processed data, transform data and such data is consumed for reporting, analytics by creating small different datasets. Now AWS users accesses AWS IoT with the AWS Management console, Software development kits. AWS IoT Analytics can easily analyze their IoT data. It also integrates seamlessly with Amazon Quicksight for visualization and Amazon SageMaker for hosted machine learning. AWS has broad deep IoT services from the edge to the cloud. AWS IoT is only vendor to bring data management, rich analytics and ease to use services designed especially for noisy IoT data. AWS IoT offers services for all layers of security.

B. Microsoft Azure IoT Platform & Analytics

Azure IoT hub is a collection of Microsoft managed cloud services that connect, monitor and control billion of IoT assests. It is a managed services hosted in client that acts as a central message hub for bidirectional communication between your IoT application and devices it manages. The integration of Azure data explorer with IoT services is used to provide to provide analytical process of various cloud data. It provides a real time analytical solution for a variety of IoT devices, sensors and so on. It provides managed services for bidirectional communication between cloud and IoT devices.

Azure firstly ingests wide variety of high streaming data into sensors and connector devices data into various sources such as Azure Event Hub, IoT Hub or Kafka. After that real time processing is done by using Azure Functions or Azure Stream Analytics. Azure Data Explorer is a big data analytics store for serving near real-time analytics applications and dashboards [13]. There are various methods such as custom analytical apps, Power BI and so on that is used to gain insight from data stored in Azure Data Explorer. This Azure Data Explorer also provides many advanced analytical capabilities such as time series analysis, pattern recognisition, and machine learning and so on.

C. IBM Waston IoT Platform

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. The architecture of IBM Watson IoT also relies on three key general features for IoT platforms: cloud computing, data analytics and artificial intelligence. Such features and functions include registration of devices, connectivity, control, analytics, storage, processing and transmission of data.

IBM has created this platform specifically to make developing and deploying IoT solutions of all sizes easier for developers.

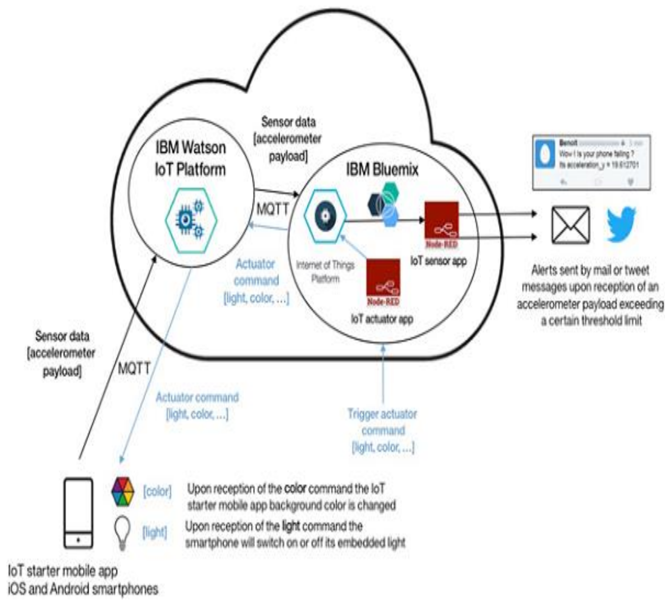

Figure 7: IoT analytics with IBM Waston Platform [14]

IBM Watson IoT architecture, the sensor data is transmitted to this platform using MQTT protocol.IBM Bluemix is an open standard cloud platform for building, managing and running various services and apps. By using Watson IoT Platform, you can collect connected device data and perform analytics on real-time data. IBM created the IBM Watson IoT Platform specifically to make developing and deploying IoT solutions of all sizes easier for developers. 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, Structured Query Language (SQL) Server, Sybase, Sybase IQ, and Teradata. The Waston IoT platform provides different types of analytics such as real time analytics, machine learning & cognitive analytics.

D. Google Cloud IoT Platform & Analytics

Google Cloud provides a robust, maintainable, end to end IoT solution on cloud platform. It integrates with a number of cloud services like Cloud IoT Edge, Cloud functions, Cloud Pub/ Sub, BigQuery and many others. Cloud Pub/Sub makes real-time, reliable processing of IoT data easy, and our storage products persist all your data efficiently and economically. IoT on Cloud Platform lets you make extremely fast queries into your business and operating environment, without managing any infrastructure [15].

The two main important components of Google IoT Core are device manager and Protocol Bridge. The device manager helps to register various IoT devices with services whereas Protocol Bridge is a way for devices to access or connect to the Google cloud with the help of various protocols such as HTTP and MQQT. The Google cloud platform provides multiple services that support big data storage and analysis. In Google Cloud IoT architecture devices send data directly to the Google IoT Core that is directly connected with Google Cloud Pub/Sub. Cloud Pub/Sub further sends the data to the cloud Dataflow for further processing. Cloud Dataflow sends the processed data to the BigQuery. BigQuery allows you to store and query the dataset holding amount of massive amount of dataset. Google DataLab is a powerful interactive tool which is created for visualizing, exploring and transforming data with the help of BigQuery. Google Analytics will provide more robust and faster environment for big data processing and analytics.

VI. CONCLUSIONS

IoT, Big data and Cloud are digital solutions that enable better analytics and helps in decision making. This paper describes the relationship between IoT and Big data, how IoT, big data and cloud computing are used in conjunctions to store large amount of data, provides scalable processing and improved real time analysis of data. It also focuses on real time analytics of IoT data with different cloud solutions. Using Big Data, IoT, and the Cloud together means you can have successful communication, connection and transference of data between devices, most effectively and efficiently.

REFERENCES

1. Umar Ahsan and Abdul Bais, “A Review on Big Data Analysis and Internet of Things”, 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems

2. N. Golchha, ‘‘Big data-the information revolution,’’ Int. J. Adv. Res., vol. 1, no. 12, pp. 791–794, 2015.

3. https://www.embitel.com/blog/embedded-blog/how-iot-works-an-overview-of-the-technology-architecture-2

4. Pankaj Mudholkar1 and Megha Mudholkar2, “Internet of Things (IoT) and Big Data: A Review”, International Journal of Management, Technology And Engineering Volume 8, Issue XII, DECEMBER/2018 ISSN NO : 2249-7455

5. Y. Wang, L. Kung, and T. A. Byrd, “Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations,” Technol. Forecast. Soc. Change, vol. 126, pp. 3–13, Jan. 2018.

6. Tulasi, “Blending IOT and Big Data Analytics” , International Journal of Engineering Research& Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICRET - 2016 Conference Proceedings

7. Preeti, A survey of Big Data Analytics on IoT, International Journal of Advance Science and Technology Vol. 29, No. 10S, (2020), pp. 2278-2282

8 M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I.A.T. Hashem, A. Siddiqa, I. Yaqoob “Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges,” IEEE Access, vol. 5, pp. 5247–5261, 2017.

9. Dr. Urmila R. Pol, “Big Data Analysis Using Hadoop Mapreduce”, American Journal of Engineering Research, e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-5, Issue-6, pp-146-151

10. Deploying Spark, HPE Elastic Platform, Big Data Analytics. Spark – a modern data processing framework for cross platform analytics Deploying Spark on HPE Elastic Platform for Big Data (2014)

11. https://www.opensourceforu.com/2020/12/top-5-open-source-tools-for-iot-analytics/

12. https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/iot-azure-data-explorer

13. https://aws.amazon.com/blogs/iot/using-aws-iot-analytics-to-prepare-data-for-quicksight-time-series-visualizations/

14 .https://www.codeguru.com/iot/iot-development-platforms-ibm-watson-iot-overview/

15. Mandeep Kumar , Cloud IoT: A Combination of Cloud Computing and Internet of Things, International Journal of Emerging Trends in Engineering and Development Issue 6, Vol. 6 (November 2016)

16. Satish, Karuturi S R V, and M Swamy Das. "Review of Cloud Computing and Data Security." IJAEMA (The International Journal of Analytical and Experimental Modal Analysis) 10, no. 3 (2018): 1-8, 2018.