International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 1|| January 2018 || PP. 91-95

Data Management in Iot Using Big Data Technologies And Tools

*B. Sobhan Babu[1],T. Ramanjaneyulu[2],I. Lakshmi Narayana[3],K. Srikanth[4]
[1](Information Technology, Gudlavalleru Engineering College, Andhra Pradesh, India)
Corresponding Author: *B. Sobhan Babu[1]*

Abstract: *The Internet of Things (IoT)[1] is on its way to becoming the next technological revolution. Given the massive amount of revenue and data that the IoT will generate, its impact will be felt across the entire big data universe, forcing companies to upgrade current tools and processes, and technology to evolve to accommodate this additional data volume. Managing and extracting value from IoT data is the biggest challenge that companies face. Organizations should set up a proper analytics platform/infrastructure to analyze the IoT data. An IoT device generates continuous streams of data in a scalable way, and companies must handle the high volume of stream data and perform actions on that data. The actions can be event correlation, metric calculation, statistics preparation, and analytics. In a normal big data scenario, the data is not always stream data, and the actions are different. Building an analytics solution to manage the scale of IoT[8] data should be done with these differences in mind.From a technology perspective, the most important thing is to receive events from IoT-connected devices. The devices can be connected to the network using Wi-Fi, Bluetooth, or another technology, but must be able to send messages to a broker using some well-defined protocol. Once the data is received, the next consideration is the technology platform to store the IoT data. Many companies use Hadoop[3] and Hive[11] to store big data. But for IoT data, NoSQL document databases like Apache Couch DB[14] are more suitable because they offer high throughput and very low latency.*

Keywords: *CouchDB, Hadoop, Hive, IoT, NoSQL*

Date of Submission: 10-01-2018 Date of acceptance: 23-012-2018

I. Introduction

The Internet of Things (IoT) [1], firstly coined by Kevin Ashton as the title of a presentation in 1999, is a technological revolution that is bringing us into a new ubiquitous connectivity, computing, and communication era. The development of IoT depends on dynamic technical innovations in a number of fields, from wireless sensors to nanotechnology. For these ground-breaking innovations to grow from ideas to specific products or applications, in the past decade, we have witnessed worldwide efforts from academic community, service providers, network operators, and standard development organizations, etc. Generally, current research on IoT mainly focuses on how to enable general objects to see, hear, and smell the physical world for themselves, and make them connected to share the observations. Specifically, the future IoT will be highly populated by large numbers of heterogeneous networked embedded devices, which are generating massive or big data in an explosive fashion. The big data we collect may not have any value unless we analyze, interpret, understand, and properly exploit it. Although there is a consensus among almost everyone on the great importance of big data analytics in IoT, to date, limited results, especially the mathematical foundations, are obtained. IoT is basically a network comprising of physical devices, which are also embedded with sensors, software, and electronics, thereby allowing these devices to interchange data. This ultimately allows better integration between real world physical entities and computer-operated systems.

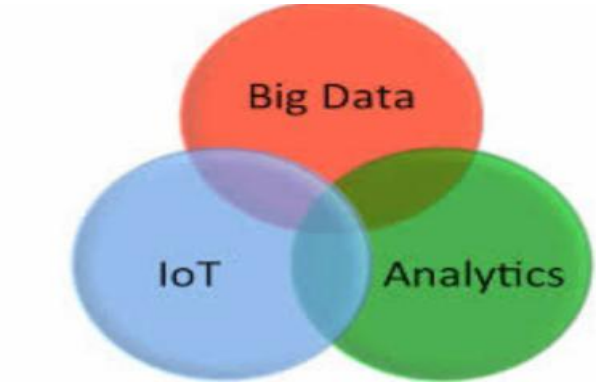

Fig: 1.1

From a technology perspective, the most important thing is to receive events from IoT-connected devices. The devices can be connected to the network using Wi-Fi, Bluetooth, or another technology, but must be able to send messages to a broker using some well-defined protocol. One of the most popular and widely used protocols is Message Queue Telemetry Transport (MQTT). Once the data is received, the next consideration is the technology platform to store the IoT data. Many companies use Hadoop[2], [3] and Hive to store big data. But for IoT data, NoSQL[16] document databases like Apache CouchDB[14] are more suitable because they offer high throughput and very low latency. These types of databases are schema-less, which supports the flexibility to add new event types easily. Other popular IoT tools are Apache Kafka for intermediate message brokering and Apache Storm for real-time stream processing.

II. Capabilities Of Iot Devices

IoT just happens to be something new, with a lot of potential behind it. We are already familiar with the likes of implants for heart monitoring, built-in sensors in automobiles, smart thermostat systems, biochip transponders and more. Also, devices can be tailor-made as per the requirements or needs of any business. A few of the aspects applicable to IoT [8] related services are outlined below.

2.1 It is largely possible that IoT devices will communicate with humans, just as with other devices.
2.2 Sensors on IoT devices will be used to capture data such as temperature of the body, pulse rate, etc. and further transmit such data.
2.3 Have control over computation and also make decisions.
2.4 Switching devices can be operated by means of controllers.
2.5 Storage of data would be possible on these devices.

III. Iot Impacts In Big Data

One of the very first things that come to mind when one talks about IoT is that of a large volume of data on your data storage. Also, such data is usually big data, of a different type and format and thus, a datacenter responsible for storing such data must be able to handle the load in its varying forms. Obviously IoT has a direct impact on the storage infrastructure of big data [4], [5].IoT brings an entirely fresh proposition to the table. The technology chosen to process and store big data must be appropriate and one that is effective. A number of technologies make up the big data platform, such as the likes of Hadoop, Map Reduce, HDFS [7] and more. Thus, what organizations need to do is to ensure that these technologies can be adapted to IoT data and also their processing. As IoT data is another source of big data, the processing steps will remain the same. So the same big data platform can be used to process IoT data.

The future IoT will be highly populated by large numbers of heterogeneous networked embedded devices, which are generating massive or big data in an explosive fashion. Although there is a consensus among almost everyone on the great importance of big data analytics in IoT, to date, limited results, especially the mathematical foundations, are obtained. This practical need impels us to propose a systematic tutorial on the development of effective algorithms for big data analytics [9] in future IoT, which are grouped into four classes:Heterogeneous data processing, nonlinear data processing, High-dimensional data processing and Distributed and parallel data processing.

Fig: 2

The aggregation of data from a large number of IoT ecosystems, can lead to large data sets for analytic purposes. Consider, for example, the ensemble data from 300 million automobile IoT ecosystems, or 300 million household IoT ecosystems, or the composition of both. Furthermore, IoT applications could connect to one or more big data systems outside the ecosystem, thus creating an aggregate of big data system orders of magnitude larger than any of the constituents. In this sense, every IoT system[3], even a small, local IoT ecosystem, is a potential big data system.The internet of things will challenge conventional approaches to data storage and analysis:

3.1 Challenges for Enterprise Data Management & Analysis

IoT marks a shift in the domain of Big Data management and analysis. IoT will create challenges in all aspects of data management. While traditionally centralized database (data warehousing), data management and analytics architectures will continue to play a valuable role, they are not well suited to handle the large volumes of raw and intermediate IoT data flows spun-off from sensors and mobile devices. The problem is no longer the absence of enough data – it's making sure only important data is moved or analyzed.

3.2 Raising the bar in terms of volume and analysis

Traditional database management technology has evolved over time to be increasingly optimized for the management of transactional data. This data is typically well structured, predictable in terms of volume and well suited to the traditional relational database engines that lie at the heart of all leading enterprise databases. The IoT will produce data in far greater volumes than ever before and the structure of that data will be much less standardized and predictable than the transactional data most organizations are used to handling. The nature and source of the data will create additional challengesand organizations will have to consider privacy and data protection alongside the challenge posed by its volume.

3.3 Moving petabytes of data can become costly

The scale of the data means that moving it from point to point will be costly in terms of money and time. The ability to push processing and analytics into the network becomes an important priority. The key, of course, lies in ensuring that unimportant data is discarded and important data is retained. This in itself will create challenges for many organizations, as one of the leading scientists from the Square Kilometer Array project noted, "We are always worried that we will decide on Monday that we can discard a certain part of the data, only to be told by another researcher on Tuesday that the information we decided to throw away is now vital for their research". The ability to adapt the algorithms that are used to filter and aggregate data at the edge of the network will be essential for some sensor networks.

IV. Big Data Tools Used With Iot

The most important thing is to receive events from IoT-connected devices. The devices can be connected to the network using Wi-Fi, Bluetooth, or another technology, but must be able to send messages to a broker using some well-defined protocol. One of the most popular and widely used protocols is Message Queue Telemetry Transport (MQTT). Once the data is received, the next consideration is the technology platform to store the IoT data. The following are the tools which are used to store and process IoT data

4.1 Hadoop

Apache Hadoop [13] is an open-source software framework used for distributed storage and processing of dataset of big data using the Map Reduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality[13], where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

4.2 Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop. While initially developed

by Facebook, Apache Hive [11]is used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA). Amazon maintains a software fork of Apache Hive included in Amazon Elastic Map Reduce on Amazon Web Services.

4.3 Couchdb

Apache CouchDB is open source database software that focuses on ease of use and having a scalable architecture. It has a document-oriented NoSQL database architecture and is implemented in the concurrency-oriented language Erlang; it uses JSON to store data, JavaScript as its query language using MapReduce, and HTTP for an API. Unlike a relational database, a Couch DB[14]database does not store data and relationships in tables. Instead, each database is a collection of independent documents. Each document maintains its own data and self-contained schema. An application may access multiple databases, such as one stored on a user's mobile phone and another on a server. Document metadata contains revision information, making it possible to merge any differences that may have occurred while the databases were disconnected.CouchDB implements a form of multiversion concurrency control(MVCC) so it does not lock the database file during writes. Conflicts are left to the application to resolve. Resolving a conflict generally involves first merging data into one of the documents, then deleting the stale one.

4.4 Nosql

It is an approach to database design that can accomodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL[12], which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.NoSQL is an approach to databases that represents a shift away from traditional relational database management systems[16] (RDBMS).

To define NoSQL, it is helpful to start by describing SQL, which is a query language used by RDBMS. Relational databases rely on tables, columns, rows, or schemas to organize and retrieve data. In contrast, NoSQL databases do not rely on these structures and use more flexible data models. NoSQL can mean “not SQL” or “not only SQL.” As RDBMS have increasingly failed to meet the performance, scalability, and flexibility needs that next-generation, data-intensive applications require, NoSQL databases have been adopted by mainstream enterprises. NoSQL[16] is particularly useful for storing unstructured data, which is growing far more rapidly than structured data and does not fit the relational schemas of RDBMS. Common types of unstructured data include: user and session data; chat, messaging, and log data; time series data such as IoT and device data; and large objects such as video and images.

V. Conclusion

IoT produces more amounts of data in coming days;big data techniques and tools must use to organize those data. Traditional database management technology has evolved over time to be increasingly optimized for the management of transactional data. IoT data is another source of big data; the processing steps will remain the same. So the same big data platform can be used to process IoT data.

References

[1] P. Tiainen, ``New opportunities in electrical engineering as a result of the emergence of the Internet of Things," Tech. Rep., AaltoDoc, Aalto Univ.,2016.

[2] M. Beyer, ``Gartner says solving `Big Data' challenge involves more than just managing volumes of data," Tech. Rep., AaltoDoc, Aalto Univ.,2011.

[3] R. Mital, J. Coughlin, and M. Canaday,`Using big data technologies and analytics to predict sensor anomalies," in Proc. Adv. Maui Opt. Space Surveill. Technol. Conf., Sep. 2014, p. 84. [

[4] N. Golchha, ``Big data-the information revolution," Int. J. Adv. Res., vol. 1, no. 12, pp. 791_794, 2015.

[5] C.-W. Tsai, ``Big data analytics: A survey," J. Big Data, vol. 2, no. 1, pp. 1_32, 2015.

[6] P. Russom, Big Data Analytics. TDWI, 4th Quart., 2011, pp. 1_35.

[7] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, `Big data, analytics and the path from insights to value," MIT Sloan Manag. Rev., vol. 52, no. 2, p. 21, 2011.

[8] L. Atzori, A. Iera, and G. Morabito, ``The Internet of Things: A survey,"Comput. Netw., vol. 54, no. 15, pp. 2787_2805, 2010.

[9] K. Kambatla, ``Trends in big data analytics," J. Parallel Distrib. Comput., vol. 74, no. 7, pp. 2561_2573, 2014.

[10] J. Manyika, Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Inst. Rep., 2011.

[11] Hive Language Manual. Available at http://wiki.apache.org/hadoop/Hive/LanguageManual.

[12] Naseer Ganiee; “NOSQL: The Big Data Solution”; International Journal of advancement in Engineering technology, Management and Applied Science, volume 1 Issues 2 July 2014.

[13] Apache Hadoop, NoSQL and NewSQL Solutions of Big Data (PDF Download Available). Available from:https://www.researchgate.net/publication/268449070_Apache_Hadoop_NoSQL_and_NewSQL_Solutions_of_Big_Data [accessed Jan 09 2018].

[14] Anderson, J. C., Slater N., and Lehnardt J. (2009). CouchDB: The Definitive Guide (1st ed.), O'Reilly Media, p. 300, ISBN 0-596-15816-5.

[15] Dimiduk, N., and Khurana, A. (2012). HBase in Action (1st ed.). Manning Publications. p. 350. ISBN 978- 1617290527.
[16] Tiwari, S. (2011). Using Oracle Berkeley DB as a NoSQL Data Store. Oracle. [Online] http://www.oracle.com/technetwork/articles/cloudcomp/berkeleydb-nosql-323570.html

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

*B. Sobhan Babu, Data Management in Iot Using Big Data Technologies And Tools. ”International Journal of Engineering Science Invention (IJESI), vol. 07, no. 01, 2018, pp. 91-95.