International Journal of Engineering & Technology, 7 (2.32) (2018) 147-149

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Unified Frame Work to Integrate Hadoop and IOT to Resolve the Issues of Storage, Processing with Leveraging Capacity of Analytics

Gudapati Syam Prasad[1], P.Rajesh[2], Sk.Wasim Akram[3]

[1,2]Department Of CSE, K L E F, Vaddeswaram,Guntur-522 502,A.P,India.
[3]Department Of CSE , VVIT, Guntur.
**Corresponding Author E-Mail:Syamprasad.Gudapati@Gmail.Com.*

Abstract

The new trend in the research and real time applications is Internet of Things (IOT). The functional benefits of IOT are ranging from smart house to smart cities. The main purpose of IOT is to integrate various devices logically and interacting between the devices without human intervention. The current discussion mainly focuses on leveraging the capacity of analytics in IOT and resolves the storage issues of the bulk data generated by IOT. The proposed idea gives the usage of Hadoop platform to store the data and from that data performing analytics for the sake of better utilization of IOT communications. The importance is explained with some real time scenarios where there is perfect blend of Hadoop platform and IOT. To store the various categories of the data Hadoop Distributed File System (HDFS) can be used, and to ingest the data from external platforms we can make use of Sqoop or Flume. The data available in HDFS can be used to process with the usage of Map Reduce (MR)technique. Once the data is available in HDFS the analytics can be performed with Hive, Pig or R in the context of Machine learning or data mining techniques. The outcome of the proposed idea is integration of Hadoop and IOT platforms with a unified frame work which accommodates the integration of Hadoop and IOT, storage provisions to handle bulk data, processing of the stored data and applying analytics so as to effectively serve various stake holders.

Keywords: *Hadoop, IOT, Analytics, Storage, Map Reduce.*

1. Introduction

The data is everywhere starting from individual human up to the MNC's all is contributing the data towards the data centres. The best examples we can mention here are Social media data generation (Facebook, Twitter) and flights generating the huge amount of sensor data with in short span of the times. The fact here is in the span of 20minutes 1,587,000 wall posts, 2,716,000 messages sent, 750 million photos were uploaded to Face Book over New Year's weekend, 3.2 billion likes comments are posted every day. Similarly in the span of 30 minutes of the time an airline jet collects 10 TB of the sensor data. The NYSE generates about 1 TB of the data per day. The other domains like transport, medical, banking and retail industries are generating the huge amounts of the data which is very difficult to manage and process by the existing systems. To handle this much amount of the data the best and perfect solution is Hadoop. Hadoop provides a platform to store bulk data along with the support of various formats of the data like structured (RDBMS), semi-structured (XML, E-Mail) and unstructured (Images, human readable text file, audio and video files).The IOT usage is crept into our lives in many ways like establishing a smart home with automatic identification of human body temperature and adjusting the AC, like wise crowd based sensing of the statistics to accommodate the number of persons through a pass over bridge.

According to Gartner, when compared with 2015 the number of interconnected devices raised by 30 % in 2016 i.e., 6.4 billion, by 2020 the prediction is 26 billion devices are going to connect and 2.5 quintillion bytes (2.5 * 10^{18})of data is produced every day.
The main issue with IOT in the current situation is storage of sensor data, processing and analysing the stored data for effective serving of the various stake holders. The paper is organized as follows in section I the usage of Hadoop and the performance aspects are described in section II the IOT usage and architectural aspects, section III describes the issues exists in IOT section IV describes the proposed frame work to integrate Hadoop and IOT, section V describes the conclusion and future scope of the research.

2. Hadoop Provisions of Storage and Processing the bulk Data

Hadoop is a solution to big data problems. A big data problem is one that when data itself becomes a part of the problem in case of storage and processing. The examples of big data scenarios includes many some of them are

- Social Media
- Banking Data
- Insurance Domain
- Medical Domain
- ONGC
- Stock Exchanges

The existing architecture of storage is not suitable to the bulk data storage and Hadoop provides a Distributed File System named as Hadoop Distributed File System (HDFS) so as to store the data with the provision of Reliability and Fault Tolerant.While processing the bulk data due to some issues the data may not be accessible so the File system should be reliable with the provision of Fault Tolerant Mechanism. The inbuilt architecture of Hadoop (version 1.X) contains various daemons to serve the storage and processing of the data. The daemons are

- Name Node
- Data Node
- Job Tracker
- Task Tracker
- Secondary Name Node

The Name Node is Meta data of the storage contains location of the files and replica information and holding the information about the Data Nodes. The Data Node is a machine which actually provides the data storage for the files, so the data can be accessed through the corresponding Data Node location which is available with Name Node Meta data. The Job Tracker is a processing unit where the scheduling of the jobs and splitting the tasks in the context of Task Tracker. The Task Tracker is the actual implementation of the task by getting the data from the Data Node with the support of the Name Node. The secondary Name Node is the backup of the Name Node in case of any failure of the Name Node the SNN can be used as the backup copy. All these daemons should run to store and process the data.

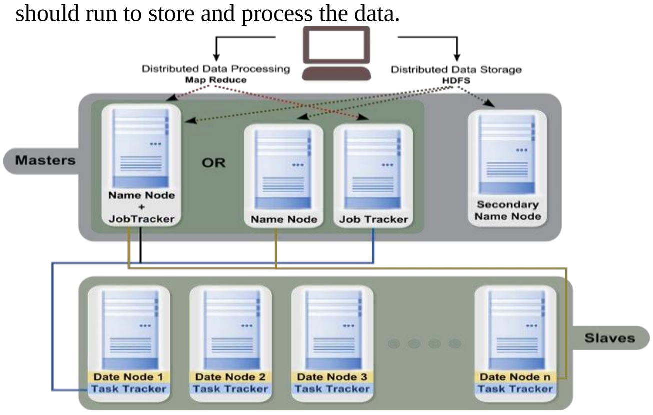

Figure 1: Hadoop Daemons Organization.

The storage of Hadoop is surrounded by HDFS with distributed file system storage and the data can be loaded from local file system or external file system. If the data is available in LFS then with the HDFS commands like –copyFromLocal -copyToLocal and –moveFromLocal where –moveToLocal is not yet implemented.

The other file system commands like Hadoopfs –ls and Hadoopfs –cat serve to perform the directory/file related storage. The processing flow can be implemented with the usage of Map Reduce implementation by creating Mapper Class, Reducer class and Driver class. The Mapper class is used to implement the actual business logic; the reducer logic consists of aggregation of the generated data from the Mapper logic.

The Driver class consists of details of input data location, Mapper reference, Reducer reference, HDFS output directory along with combiner and partitioned logics. The HDFS input provides the data to the Job Tracker and HDFS output directory gives the result from the Reducer class in the form of three files like status of the process execution, log file and actual output file. In the process of exporting we can link the required files such as Hadoop.jar, pig.jar or hive.jar available in the Hadoop environment.

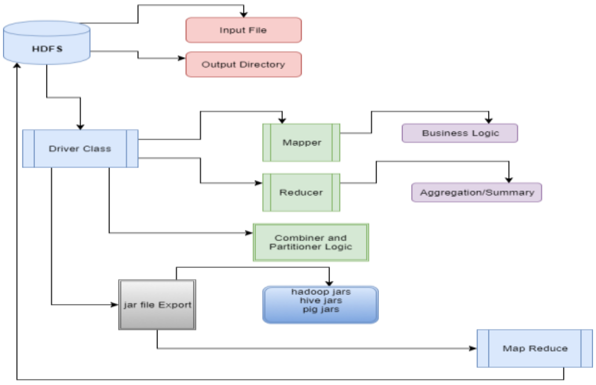

Figure 2: Flow of Data processing in Hadoop

The conceptual benefits of Hadoop are the huge storage capacity and reliable and fault tolerant data processing with efficient data input formats so as to handle text files, audio, images and video files. The another advantage of the architecture is the user can customize the replication factor by editing the hdfs-site.xml, for example if a particular file needs to be available in 1000 machines we can simply mention the replication factor as 1000 in the said file with the property that dfs.replication.

The distributions like Apache, cloud era and Horton works are providing good support so as to establish the Hadoop environment in Ubuntu/Linux/Unix machines directly or through VMware there is a provision of installing Hadoop on top of windows machine. The Master slave architecture establishment is so simple and flexible at any point of time we can add slaves to the identified master so as to leverage the storage and processing requirements dynamically. The extension of block size like 64 MB, 128 MB or 256 MB as per requirement can be done by changing the **dfs.block.size** dimension in hdfs-site.xml.

3. IOT Architecture and Issues

After the advent of Hadoop and Cloud technologies the next buzz in the common man life is IOT. The whole constellation of inanimate objects is being designed with built-in wireless connectivity, so that they can be monitored, controlled and linked over the Internet via a mobile app.The devices like traffic signals, watches, fridges, televisions, AC or anything will be connected to the Internet via an IP address and can exchange data with other devices. The IOT mainly includes wireless, Big Data,Cloud devices and security.

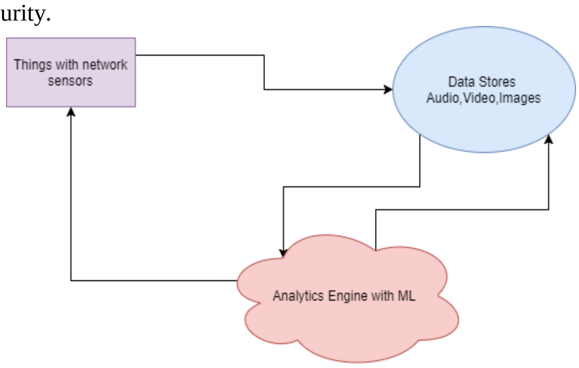

Figure 3: IOT Basic elements.

In the real time the best examples of IOT are Magic Bands and Beacon Technology. The Magic Bands are plastic bracelet that contains RFID radios in use at Walt Disney world resort.

The Magic Band provides a logical connect between park tickets, hotel keys, payments and photo pass information to avoid the ambiguity of maintaining multiple tokens by the customers.

Similarly the Beacon technology is other IOT based implication which is having a signal based identification of person by the security doors and sending the coupon codes and offers to a community of the customers who are visited the malls.

The IOT deals with devices, interconnection and sharing of the data, in the meantime if the system crashes then the entire communication is going to collapse.

The main issues identified in the literature of IOT are Big data sourcing and advanced analytics. The nature of IOT is generation of huge sensor data from the interconnected devices and getting the valuable information from that huge data is a typical task.

The issues can be covered with Hadoop Integration as this is a platform to solve the storage problems of the huge amounts of the data. At the same time applying analytics is another dimension with Hadoop by using Hive,Pig or R language with Machine Learning capabilities.

4. Unified Frame Work to Integrate Hadoop and IOT

To address the issues in the IOT with the context of Hadoop the proposed frame work involves the components of storage, processing logic and analytical engine to serve the needs of the huge amounts of the data management. With the proposed architecture we believe that the storage relevant issues, processing-oriented problems and analytical requirements can be achieved.

The proposed unified architecture will integrate IOT and Hadoop technologies so as to address the issues specified in the previous section. The initial task is the architecture identifies the kind of the source data in the form of structured or semi/unstructured, if it is structured then with sqoop the data can be ingested into HDFS, otherwise flume can be used to ingest the data.

To process the data in the HDFS the Map Reduce will provide the distributed and parallel processing of the data. To impose the basic and simple analytics we can make use of the tools Hive or Pig Latin.

The Hive Query language provides the joins and subset of the operations, Pig Latin provides the analysis of click streams and ranking of the data with the help of customised operation of loading, stream analysis.

Suppose the data need to be analysed in a high end way like usage of data mining like clustering and classification along with that the Machine Learning tasks like regression analysis, random forest the architecture provides the usage of R/Python based packages or libraries which are built-in to perform the analytics.

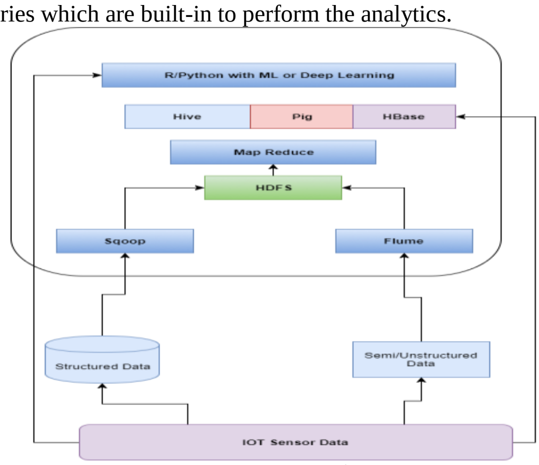

Figure 4: IOT and Hadoop Unified Architecture

5. Conclusion

The proposed unified frame work integrates the Hadoop and IOT platforms to address the problems in the IOT like storage relevant aspects, processing based and analytics related complexities. The architecture covers the usage of various categories of the data and various analytical tools based on the range of analytics.

References

[1] Umapavankumar.K, Dr.B.Lakshmareddy ," Various Computing models in Hadoop eco system along with the perspective of analytics using R and Machine learning" Vol. 14 CIC 2016 Special Issue International Journal of Computer Science and Information Security (IJCSIS) https://sites.google.com/site/ijcsis/ ISSN 1947-5500.

[2] www.cloudera.com

[3] www. https://kontakt.io

[4] S. Lohr, "The age of big data," N. Y. Times, vol. 11, 2012.

[5] S. Madden, "From Databases to Big Data.," IEEE Internet Comput., vol. 16, no. 3, 2012.

[6] P. Zikopoulos, C. Eaton, and others, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[7] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data," Manag. Revolut.Harv. Bus Rev, vol. 90, no. 10, pp. 61–67, 2012.

[8] R. Appuswamy, C. Gkantsidis, D. Narayanan, O. Hodson, and A. Rowstron, "Scale-up vs Scale-out for Hadoop: Time to rethink?," in Proceedings of the 4th annual Symposium on Cloud Computing, 2013, p. 20.

[9] A. S. Tanenbaum and M. Van Steen, Distributed systems.Prentice-Hall, 2007.[7] C. P. Chen and C.-Y. Zhang, "Dataintensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci., vol. 275, pp. 314–347, 2014.

[10] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," Jama, vol. 309, no. 13, pp. 1351– 1352, 2013.

[11] Dr.B.LakshmaReddy,Umapavankumar.K," Big data techniques and analytics in Ecommerce business" International Conference at Pondicherry University, on October 2016.

[12] www.safaribooksonline.com

[13] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[14] J. Y. Monteith, J. D. McGregor, and J. E. Ingram, "Hadoop and its Evolving Ecosystem.," in IWSECO@ ICSOB, 2013, pp. 57–68.

[15] K. Ting and J. J. Cecho, Apache Sqoop Cookbook. O'Reilly Media, Inc., 2013. [14] S. Hoffman, Apache Flume: Distributed Log Collection for Hadoop. Packt Publishing Ltd, 2013.

[16] S. Haloi, Apache ZooKeeper Essentials. Packt Publishing Ltd, 2015.

[17] M. K. Islam and A. Srinivasan, Apache Oozie: The Workflow Scheduler for Hadoop. O'Reilly Media, Inc., 2015.

[18] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig latin: a notsoforeign language for data processing," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1099–1110.

[19] H. Bansal, S. Mehrotra, and S. Chauhan, Apache Hive cookbook.Packt Publ., 2016.

[20] E. Alpaydin, Introduction to machine learning (adaptive computation and machine learning series). The MIT Press Cambridge, 2004.