IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907073 |

A Review on Big Data Storage Analytics with Its Methods and Components Used to Store Huge Data

Jaswinder Kaur

Assistant Professor, Dept. of Computer Science & Engineering, Chandigarh University, Gharuan, Mohali, Punjab, India

ABSTRACT: Big data technology is used in Cloud computing and databases. It is mainly used to process large sized datasets. These large sized datasets are difficult to process through the normal database technologies. Big data technology provides various methods to process the datasets. Big data system consists of components to deal with large sized datasets. This proposal depicts the methods and components used by the Big data technology.

KEYWORDS:- Big data, Warehouse, Hadoop, MapReduce, HBase, Hive, Pig

I. INTRODUCTION

Big data storage is a evaluated and depository architecture which can use to gather and control huge datasets and perform real-time analyses. These analyses can be used to initiate intelligence from metadata.

Generally, big data storage is collection of hard disk drives due to the media's lower cost. However, flash storage is obtaining popularity due to its decreasing cost. When flash is used, systems can be construct purely on flash media or can be construct as meld of flash and disk storage.

Data within big data sets is formless. To contain this, big data storage is usually construct with object and file-based storage. These storage types are not confined to specific dimensions and typically volumes scale to terabyte or petabyte sizes.

II. IMPORTANCE OF BIG DATA STORAGE

The necessity for being able to store and process information has risen rapidly. The rapidly increase of data generation meant that organization needed to scale-up their big data capabilities, including storage. In being, the key conditions of big data storage are that it can grasp very large amounts of data and carry on to scale to keep up with growth. The ideal big data storage system would allow the storage of an actually unlimited amount of data. It would be able to handle both high rates of irregular write and read access, have adaptable and systematic deal with a range of different data models, support both planned and unplanned data, and only work with coded data for solitude protection.

A. Warehouse storage

A data warehouse is an ample building facility which its main function is to store and process data on a concern level. It is an important inventory for big data inquiring. These huge data warehouses carry the various reporting, business intelligence (BI), analytics, data mining, research, cyber monitoring, and other comparable activities. These warehouses are usually develop to keep and process wide amounts of data at all times while sustain them in and out through online servers where users can approach their data without gap.

Data warehouse inventory make it viable to manage data more well planned as it enables being able to find, access, envision and analyse data to make better business settlement and attain more desirable business results. They are built

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907073 |

with the debate of expanding data growth in mind. There is no hazard of the warehouses being mess up by the increasing amount of data that is being stored.

B. Cloud Storage

The other method of storing an ample amount of data is cloud storage, which is something more people are conversant with. If you have ever used iCloud or Google Drive, this means you were utilize cloud storage to store your documents and files. With cloud storage, data and information are kept electronically online where it can be ingress from anywhere, undo the need for direct attached access to a hard drive or computer. With this greet, you can store almost boundless amount of data online and access it where.

The cloud supply not only readily-available framework, but also the power to scale this infrastructure rapidly to manage large increases in traffic or usage.

III. THE COMPONENTS OF BIG DATA STORAGE

Storage system, computation and logic layer, application interaction and analytics databases are the components of big data storage system. The following Figure.1 shows these components.

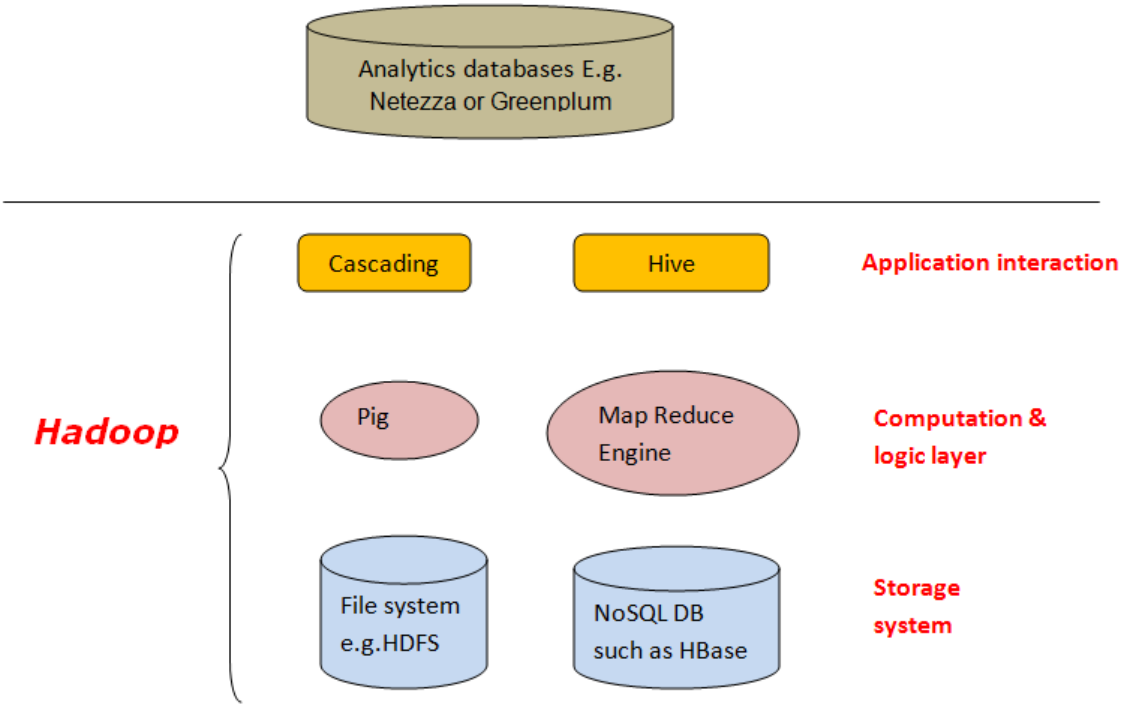

Figure.1. Components of Big data storage system

A. Storage system:-

Hadoop distributed file system is also known as storage layer because it can handle storage of data and also the metadata which is needed to complete the computation. NoSQL database store data in the form of tables for example HBase or key value based columnar Cassandra.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907073 |

B. Computation and logic layer:-

MapReduce consists of two individual processes such as mapper and reducer. Mapper runs and takes input as a raw dataset and converts it to another data structure which is key-value. Then reducer works, it takes data produced by the mapper as an input and processes it and transforms it into a smaller dataset.

Pig is another framework which works at top of Hadoop and processes data and it can be used as a substitute or conjunction with MapReduce. It is a HLL (High level language) and mostly used for making components of processing for analysis of very large datasets. One main aspect is that its structure is corrigible to different degrees of resemblance. It act as compiler which converts Pig scripts to MapReduce jobs.

C. Application interaction

Hive is a data warehousing coating that's construct on top of the Hadoop platform. In simple terms, Hive supply a facility to link with, process, and evaluate HDFS data with Hive doubt, which are very much like SQL. This makes the shift from the RDBMS world to Hadoop obvious.

Cascading is a framework that uncover a set of data handle APIs and other components that expound, share, and carry out the data processing over the Hadoop/Big Data stack. It's above all an abstracted API layer over Hadoop. It's hugely used for application development because of its facility of development, creation of jobs, and job scheduling.

D. Analytics databases

Databases such as Netezza or Greenplum have the potential for go up and are known for a very fast data gulp and refresh, which is a required requirement for analytics models.

IV. CONCLUSION

Big data technology is used to manage, manipulate, store the large sized datasets. To deal with huge data, it is a big challenge for researchers. This paper describes the concepts of Hadoop and NoSQL which deals with data storage, MapReduce and Pig which are required for computation, Hive and Cascading frameworks for application programming interface and analytics databases such as Netezza, Greenplum which is required for analysis of datasets.

REFERENCES

[1.] R. Buyya, S. Pandey, “Data Intensive Distributed computing: Challenges and solutions for large-scale information system”, Information science reference, 2012.
[2.] S. Pandey, “Scheduling and Management of Data Intensive application workflows in Grid and cloud computing environments”, Dec. 2010.
[3.] M. Mustafa Rafique a, Ali R. Butt a, Dimitrios S. Nikolopoulos b,c, A capabilities-aware framework for using computational accelerators in data-intensive computing, J. Parallel Distrib. Comput. 71 (2011) 185–197.
[4.] S. Pandey, A. Barker, K. K. Gupta, R. Buyya ,Minimizing Execution costs when using globally distributed cloud services,2010 24th IEEE International Conference on Advanced Information Networking and Applications .
[5.] Y. C. Lee, A. Y. Zomaya, Rescheduling for reliable job completion with the support of clouds, Future Generation Computer Systems 26 (2010) 1192_1199.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details