ISSN:0975-9646

Comparative Study on Big Data Analytic Frameworks in Cloud

D.Avinash Babu[#1], A.Koushik[#2]

[#]*Dept of Computer Science & Engineering, JNTU Kakinada*
Satya Institute of Technology and Management, Vizianagaram, A.P.,India

***Abstract*— Big Data is a challenging research area for everyone to extract massive amount of information which is available worldwide due to vast increase of social media content, transactional data, RFID tags, Internet of Things etc. On the other hand, Cloud Computing is another area in the IT field where different services like Software, Infrastructure, storage etc. are available online and its main benefits are scalability, efficiency and flexibility. In spite of its benefits, big data in cloud is not only used for storage but also for performing in-depth analysis. This paper concentrates upon frameworks available for analysing big data storage in the cloud and comparative study is shown.**

***Keywords*— Big Data, Cloud Computing, SaaS, PaaS, IaaS, IOT.**

I. INTRODUCTION

Storage of data has become major concern for everyone since data grows year by year and is accumulated. Effective strategies must be designed to utilize the data efficiently and must also make predictions on what categories of data will be useful in the future.

The term 'BIG DATA' was first coined by Michael Cox and David Ellsworth. Big data [1] refers to availability, analysing and extracting of different data sets in real time based on its size and complexity. The benefits that can be accomplished with big data are costs can be reduced, diversification of revenue streams, keeps data safe, offers precise customized services and analysis of risks can be performed effectively.

Cloud computing [5], [6] is also an emerging research area where different services are provided online. The different services that can be provided are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In SaaS, software is delivered as a service and its examples are Salesforce, concur, working day, Oracle, Google docs etc. In PaaS, platform for creation of software over the web is delivered as on demand service and its examples are Cloud foundry, Google App engine, Open shift, Windows Azure etc. In IaaS servers, network, storage and operating systems are delivered as on-demand service and its main examples are Amazon EC2, Windows Azure, Rackspace, Google Compute Engine etc.

The different criteria to be considered when storing big data in cloud are building low cost, efficient storage platform and performing in depth analysis on these data sets. This paper concentrates upon different frameworks used for analysing big data in cloud and comparative study is shown on all these frameworks.

II. BIG DATA AND CLOUD COMPUTING

A. *Big Data*

It is defined by four V's-Volume(scale of data), Velocity(analysis of streaming data), Variety(different forms of data), Veracity(uncertainty of data).Its major sources are streaming data, social media data, transactional data, weather related data, RFID tags data, IOT data etc[2][3].Big data reference architecture is shown below[4].

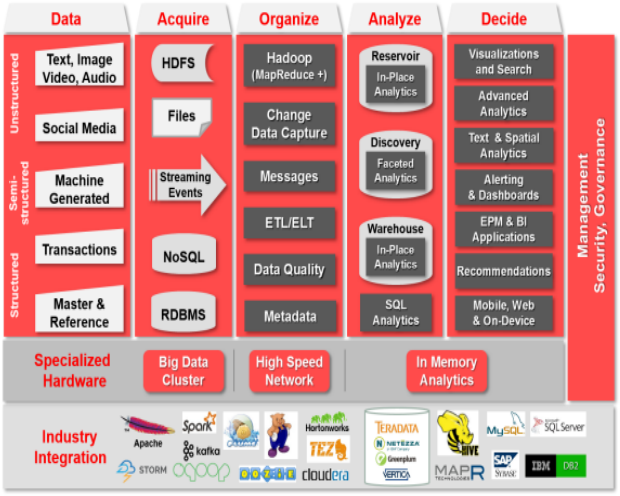

Fig.1. Big Data Unified Architecture

B. *Cloud Computing*

It refers to providing different services on-demand. Its main characteristics are on-demand capabilities, efficient network access, pooling of resources, scalability and measurement of services[5][7].The different deployment models used for hosting different categories of data to cloud are public cloud (providing services accessible to public on commercial basis),private cloud(providing services accessible within an organization), hybrid cloud(combination of public and private clouds). The key drivers for cloud computing are [8]

- Mobility
- Business continuity
- Overcoming resource shortages
- On-demand scalability
- Coping with uncertainty and change
- Savings of cost
- Compliance/Regulatory challenges

III. BIG DATA ANALYTIC FRAMEWORKS IN CLOUD

Analysing of big data in cloud has become major issue so there are different frameworks used for performing in-depth analysis on different categories of data. The different frameworks are outlined below.

TABLE I
BIG DATA ANALYTICS FRAMEWORKS IN CLOUD

Framework	Features	Advantages	List of companies
Hadoop[10][12]	Provides efficient access by storing redundant data across multiple machines, provides efficient approach for authentication of different machines, provides efficient programming model using Map Reduce framework	Simple model, scalability, robust and fault tolerant	Amazon, Qburst , Hortonworkcloudera etc
Storm[13]	Provides an efficient approach for handling real time data sets in parallel, accessing tens of thousands of data sets per second on cluster	Fault tolerant, flexible, robust, reliable, handle small & large data sets efficiently with low latency	Twitter, Yahoo, WebMD, Yelp etc
Kafka[14][15]	Provides distributed publish-subscribe messaging system, provides online and offline consumption of messages	Reliable, scalable, durable, high performance with no downtime and no data loss	LinkedIn, Oracle, Cloudfare , Wooga, Uber etc
Solr[16]	Provides Restful API's, NoSQL database, full text search, deployed in any kind of systems such as standalone, distributed, cloud etc	Automatic replication of indexes, automated recovery of data, efficient approach for caching	Netflix, Zappos, Stubhub!, AOL, digg, eTrade, Disney, Apple, NASA, MTV etc
Spark[11]	Increasing processing speed of application using in-memory computing, support for multiple languages	Faster job execution, supports machine learning algorithms for future predictions	Yahoo, Intel, Baidu, Trend Micro etc.

A. Hadoop Technologies

The different Hadoop technologies are given below[9]

1) Apache Pig

It is a tool for analyzing larger sets of data by representing them as data flows.

2) Apache HBase

It is similar to Google big table which provides random access to huge amount of data.

3) Apache Hive

It is data warehouse infrastructure tool which queries and analyses massive amount of data in easiest manner.

4) Apache Sqoop

It is tool designed to transfer data between Hadoop and Relational database servers.

5) Apache Flume

It is standard, simple and robust tool for data extraction from different web servers into hadoop.

IV. CONCLUSION

Big data is an emerging research area where a massive amount of information is available worldwide for performing in-depth analysis due to vast increase of various sources of data. Cloud computing is also playing major role by providing different services to users

This paper also presents different frameworks used for analyzing big data in the cloud and also it outlines the features, advantages and companies using big data frameworks. Lastly, quick glance is given to the different technologies used in Hadoop

REFERENCES

[1] D. Assunçãoa, Rodrigo N. Calheiros b ,Big Data computing and clouds: Trends and future directions Marcos, Elseiver, 27 August 2014.

[2] Raghavendra Kune1, Pramod Kumar Konugurthi,Arun Agarwal , Raghavendra Rao Chillarige and Rajkumar Buyya,,The anatomy of big data computing, Software: Practuice and Experience, 2016

[3] Chun-Wei Tsai,Chin-Feng Lai,Han-Chieh Chao, Athanasios V. Vasilakos, Big data analytics: a survey, ‚Journal of Big data, Springer, December 2015

[4] http://www.oracle.com/technetwork/oea-big-data-guide-1522052.pdf

[5] Santosh Kumar, R. H. Goudar,Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey , International Journal of Future Computer and Communication, Vol. 1, No. 4, December 2012

[6] Samiya Khan, Kashish AraShakil, MansafAlam, "Cloud Based Big Data Analytics: A Survey of Current Research and Future Directions", Journal of Contemporary Psychotherapy,2015.

[7] Mohsin Nazir,,Cloud Computing: Overview & Current Research Challenges IOSR Journal of Computer Engineering (IOSR-JCE) ISSN: 2278-0661, ISBN: 2278-8727Volume 8, Issue 1 (Nov. - Dec. 2012), PP 14-22

[8] Mohiuddin Ahmed , Abu Sina Md. Raju Chowdhury , Mustaq Ahmed , Md. Mahmudul Hasan Rafee,An Advanced Survey on Cloud Computing and State-of-the-art Research Issues, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012

[9] V. Srilakshmi, V.Lakshmi Chetana , T.P.Ann Thabitha ,A Study on Big Data Technologies, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2016

[10] SanthoshVoruganti, "Map Reduce a Programming Model for CloudComputing Based On Hadoop Ecosystem", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3794-3799

[11] J. Boehm, K. Liu, C. Alis, "Sideloading - Ingestion of Large Point Clouds Into the Apache Spark Big Data Engine", ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B2, 2016, pp.343-348.

[12] Poonam S. Patil1,Rajesh. N. Phursule, Survey of Big data processing and Hadoop components, International Journal of Science and Research, Volume 3, Issue 10, 2014.

[13] Robert Evans, Apache Storm-A hands on tutorial, IEEE International Conference on Cloud Engineering,2015

[14] Cao Ngoc Nyugen,Jik Soo-kim,Soonwork Hwang, Building a Kafka-Based Distributed Queue System on the Fly in a Hadoop Cluster, 2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS*W)

[15] Jay Kreps, Neha Narkhede, Jun Rao, Kafka: A Distributed Messaging System for Log Processing, ACM, 2011.

[16] Jingyong Wan, Beizhan Wang, Wei Guo, Kang Chen, Jiajun Wang, A distributed search engine based on a re-ranking algorithm model, 2015 10th International Conference on Computer Science & Education (ICCSE)