ISSN:0975-9646

Emphasizing Big Data Engineering and Software Architecture on Cloud Computing

Mohammed Faisal[1] ,Tamil Selvi Madeswaran[2], Sanjay Gupta[3]
[1]faisal31621@gmail.com,[2]tamilselvimurali@gmail.com,[3]guptasanjay3@gmail.com

Abstract: Changes is the natural phenomenon which we have to accept everywhere in our life. Let's emphasize what changes occurred to maintain the huge amount of data. To do the engineering on the big data we have to find out the appropriate tools, techniques and the architecture where the big data is lying and getting processed. In this research paper concept of big data in term of where it's arrived what are the tools are available will be discussed. In the later section big data software architecture on cloud computing will be also discussed. Now the complete scenario is changed, new technologies has overlapped the existing technologies like instead of client server, cloud computing is better option and the requirements of big data is often demanding on cloud computing. It's a challenge for the software engineering discipline to find out the appropriate, efficient and fast software architecture to serve big data on cloud computing.in this paper it's an effort to show the dependability and requirements to use in efficient way between the three concepts which is big data, efficient software architecture and clod computing, how to use big data on cloud computing , the problems and the proposed solution and what changes is required to serve better in term of an efficient software architecture.

Keywords: Big Data, Cloud Computing, Hadoop, , Software Architecture, Smart Grid

1.0 CLOUD COMPUTING :

Cloud computing is a method of technology where it's used to run different kinds of application and ability to store related data into the central system and deliver the services as required by the customers and other kinds of user to access them[13]. In general, unlike the traditional model the cloud computing provides the service to access the data remotely instead from your home computer, mobile or Organization's networked computer through real time communication network.

The advantage of cloud computing is cost and flexibility. There are main two deployment models of cloud computing whereas public and private cloud. Public cloud is free for all and it's applicable to all kinds of users where as private is meant for single organization and chargeable. Cloud Computing provides three kinds of services such as infrastructure, platform and software service [13]. The major goal of the cloud computing is to share resources which include the above said three types of services. Different architectures are used for the cloud computing. In terms of data architecture the cloud computing was drive by various kinds of applications and use huge amount of data. To manage and implements the need and varieties of applications which is required to store different kinds of data like video ,audio text , images ,biomedical data etc. we emerged the new concept called big data.

2.0 WHAT IS BIG DATA?

"Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making"(Source – says futurist Thornton May, SAS webinar in 2011).

2.1 What we do with the Big Data?

Big data represents a revolution in the data discovery, utilization and data analysis[3] .It's not the issue of acquiring the massive amount of data but the actual vision is what we do with that data[14]. Big Data analysis will analyze any form of data from transparent domain crossing platform to solve complex optimization problems in networks, Oil and Gas, Business, Government, Healthcare system, Biotechnology, Technology, Industrial automation, and agriculture in terms of cost reduction and time reduction with emerging technologies [1]. For example in Oil and gas the outputs taken from the drilling machine is used to take an efficient and safer drilling decision.

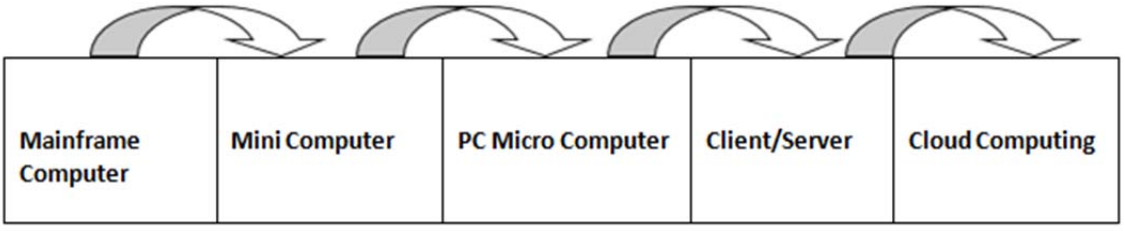

Fig 1: Wave of change

Analysis Type	Processing Methodology	Data Frequency	Data Type	Content Format	Data Sources	Data Consumers	Hardware
-Real Time -Batch	-Productive Analysis -Analytical • Social Network Analysis • Location based Analysis • Features Recognition • Test Analytics • Transcription • Search Analytics -Query and Reporting	-On Demand feed -Continuous feed -Real time feed -Time Series	-Meta Data -Master Data -Historical Data -Transactional Data	-Structured • Images -Unstructured • Text • Video -Semi structured • Documents • Audio	-Web & Social Media -Machine Generated -Human generated -Internal data sources -Transaction data -Biometric data -Via data providers -Via data generators	-Human -Business Process -Other Enterprise Applications -Other Data Repositories	-Commodity Hardware -State of Art Hardware

Table 1: Anatomy of Big Data

In the above mentioned Table 1 [9] its shows the anatomy of big data, where the top row items shows the relevant information about big data in the below mentioned column.

2.2 Technologies behind Big Data

Traditional data analysis methods like cluster analysis, factor analysis, statistical analysis, correlation analysis etc. are used for big data analysis. Since the volume of big data datasets is so large and complex it's difficult to manage by the traditional tools[3]. Some more technologies that can be applied to handling the big data is parallel processing, distributed file system, Hadoop and NOSQL Database Technologies, grid computing, cloud computing and Data Visualization Technologies. With that big data database we can built a new application, improve the effectiveness of the existing application and built a new era in the business intelligence.

3.0 Road Map of Big Data to Cloud Computing: A Case Study

A Smart grid is a modernized electrical grid that uses analog or digital Information and Communication Technology [10]. Smart Grid is very young area for IT researchers to work on the relevant technology, even though it's directly belongs to electrical field but because of involvement of the Information Technology and Communication system which is used to manage the behavior of supplier and consumer in an automated fashion it's also a challenging area for IT researchers. In the **Fig 2** Smart grid providers are providing the electricity supply to their consumers. Where the bill will be generated by smart meter, huge amount of data will be generated in continues fashion from the different electricity consumers, data will be directly stored on the cloud which will be managed by an authorized third party assigned by the smart grid electricity supplier . The third party used big data technology to store and process the data received from the different consumers and provides the final bill to the supplier. At the end electricity supplier will distributes the utility bills to concern consumers.

3.1 Big Data, Hadoop and Cloud Computing used in Smart Grid Case Study

Hadoop is basically a framework used in distributed processing system that access the data from clusters of computer using simple programming model [2]. Hadoop is a platform which helps the cloud computing to provide services to the customers. It develops so many API using JAVA simple code. Hadoop tools improves the confident decision making process in the business intelligence from unsaturated, structured or semi structured data from cloud computing.

In our Case Study we used the three Hadoop components: 1. Hadoop's MapReduce, 2. Hadoop's Distributed File System (HDFS), 3. Hadoop storage system for structured data. Smart Grid use large amount of data for processing the bill amount. MapReduce Tool handles large processing data in parallel way [2]. Each and Every input is supplied to different nodes and converted into the files and processed in the HDFS. Here the feature of the Mapreduce tool handles the parallelism, reliability and fault tolerance.

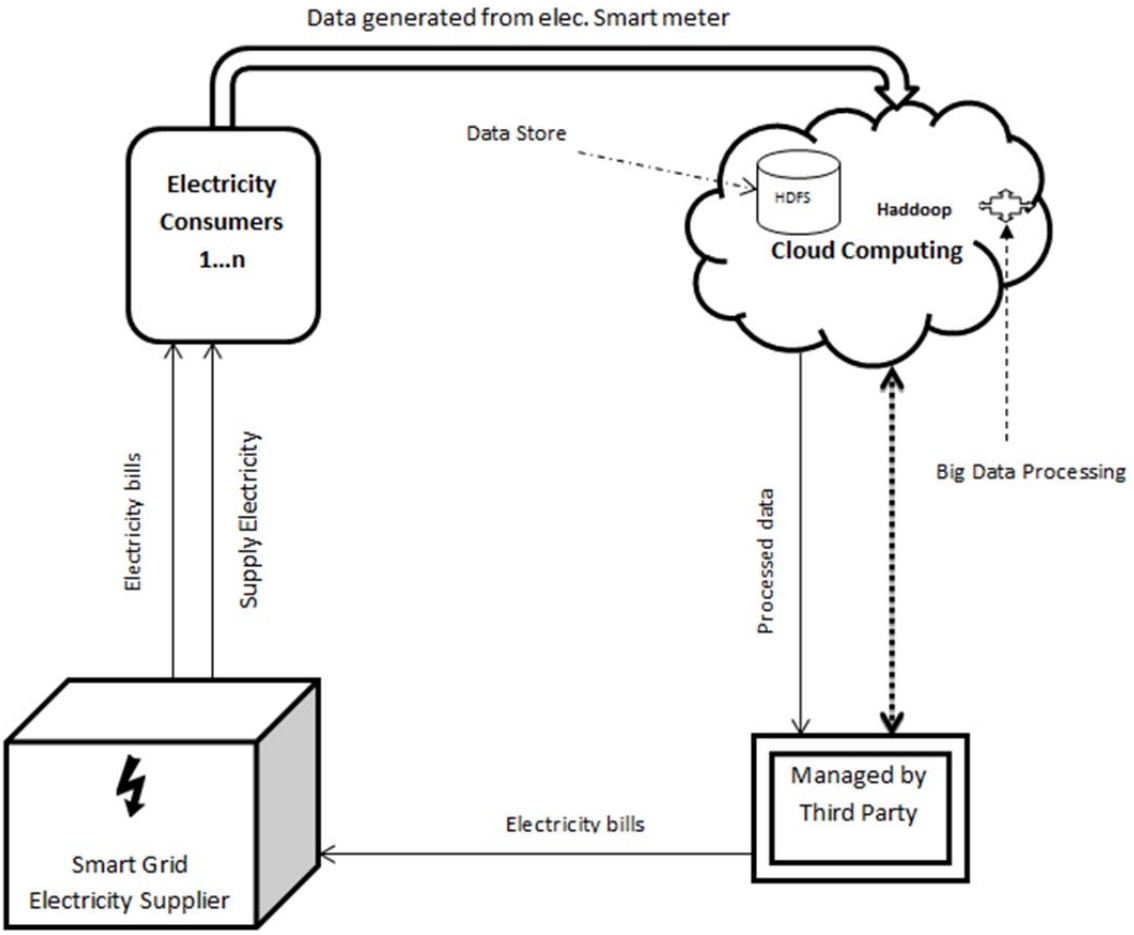

Fig. 2: Smart Grid Electricity bill generation by using big data on cloud Computing

4.0 BIG DATA ARCHITECTURE AND CLOUD COMPUTING

In the above **Fig 3** illustrating the Interaction between the integration middle layer and the external environment, where applications such as M2M and private data vaults utilize the integration middle layer. The integration middle layer enables the selection of the best data-storage type according to the application's needs. This type can be determined by monitoring the application's data-access pattern (such as create/read/update/ delete) [4].

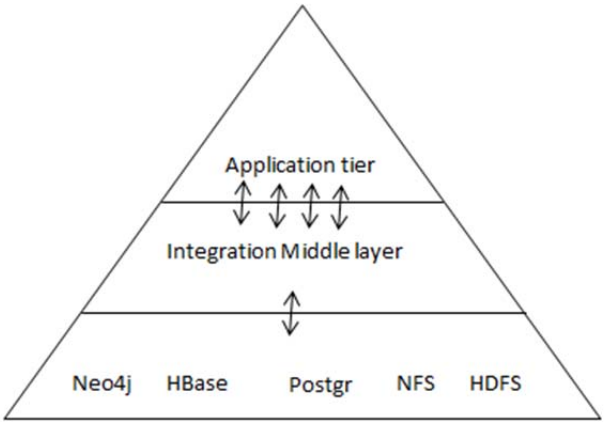

Fig 3: Overview of applications and database integration middle layer

Neo4j: Neo4j is the world's leading Graph Database. It is a high performance graph store with all the features expected of a mature and robust database, like a friendly query language and ACID transactions. The programmer works with a flexible network structure of nodes and relationships rather than static tables [5].

Hbase: Apache HBase is a distributed, scalable data store that runs on top of Apache Hadoop's file system, the Hadoop Distributed File System (HDFS) [5].

PostgreSQL: PostgreSQL is a powerful object-relational database management system, provided under a flexible BSD-style license.[1] PostgreSQL contains many advanced features, is very fast and standards compliant[6].

NFS: allows a server to share directories and files with clients over a network. With NFS, users and programs can access files on remote systems as if they were stored locally [7].

HDFS: The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems [11].

4.1 Architectural challenged of Big Data and Cloud Computing:

a) What is the data involved, how big is it and where is it located? One of the challenges of Big Data is that the data may be disparate and start out in a variety of locations.

These locations may or may not be within a cloud service [8].

b) What sort of processing is going to be required on the data? Continuous or infrequent "burst" mode? How much can it be parallelized [8]?

c) Is it practical to move the data to an environment where the processing can be performed efficiently, or is it better to think about moving the processing to the data? With the sorts of data volume involved [8].

5.0 CONCLUSION:

It was an effort to emphasize the Big data, software architecture on cloud computing. Still it is a challenge for IT researchers and Engineers to find appropriate pre-defined software architecture based on the nature of the project. At the same time to store and process the big data in an efficient and secured reliable way need to improve the relevant techniques and tools. Implementation of Big Data and its technology is widely used in the core engineering projects such as smart grid, so again it's a challenge for IT researchers and engineers to fill the gap between core engineering products and software technology to make the drastic changes like smart grid, so based on the novel and critical thinking, researchers and technologists can change the dimension of thinking as well workability of the many engineering products to use in a most efficient and reliable fashion.

REFERENCES:

1. Business- Intelligence And Analytics:From Big Data To Big Impact by Hsinchun Chen,Roger H. L. Chiang ,Veda C. Storey MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012
2. A Big-Picture Look at Enterprise Architectures- Frank J. Armour, Stephen H. Kaisler,and Simon Y. Liu 1520-9202/99/$10.00 © 1999 IEEE January x February 1999 IT Pro
3. The Age of Big Data By STEVE LOHR- http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=1&scp=1&sq=Big%20Data&st=cse
4. http://www.ericsson.com/res/thecompany/docs/publications/ericsson_review/2012/er-big-data-technologies.pdf
5. http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/hbase.html
6. https://help.ubuntu.com/community/PostgreSQL
7. https://www.freebsd.org/doc/handbook/network-nfs.html
8. http://www.cloudcil.org/CSCC_Deploying_Big_Data_Analytics_Applications_to_the_Cloud_FINAL.pdf
9. http://www.ibm.com/developerworks/library/bd-archpatterns1/
10. D. J. Hammerstrom et al. "Pacific Northwest GridWise™ Testbed Demonstration Projects, Part I. Olympic Peninsula Project" (PDF). Retrieved 2014-01-15.
11. HDFS Architecture Guide : https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
12. The Claremont Report on Database Research - Joseph M. Hellerstein - June 2009 | vol. 52 | no. 6 | communications of the ACM
13. Big Data and Cloud Computing: Current State and Future Opportunities_ Divyakant Agrawal Sudipto Das Amr El Abbadi - EDBT 2011, March 22–24, 2011, Uppsala, Sweden.Copyright 2011 ACM 978-1-4503-0528-0/11/0003