American Journal of Engineering Research (AJER)
e-ISSN : 2320-0847 p-ISSN : 2320-0936
Volume-03, Issue-05, pp-266-270
www.ajer.org

Research Paper Open Access

Convergence of Big Data and Cloud

Sreevani.Y.V.[1], Sharat Vikas Jogavajjula[2]
[1](Computer Science and Engineering , Hyderabad Institute of Technology and Management(JNTUH), AP,India)
[2](Computer Science and Engineeringt, Hyderabad Institute of Technology and Management(JNTUH),AP, India)

Abstract: - Big Data that is data created in enormous amounts from various heterogeneous sources that are difficult to process and handle by means of general methods of storage management. Majority of organizations are moving to the cloud because of proven benefits like cost cutting, better flexibility and more importantly data security. In this paper it is proposed the way how Cloud technologies are gaining momentum every day in the largest data centers and the smallest businesses, offering services that encompass everything from the infrastructure down to the software level. The convergence of Big Data with Cloud Computing will have an edge over the traditional computing methods which is far more accessible at a cheaper rate. Converging of the Big Data & Cloud Computing is a perfect fit, but delivering on the potential of infinitely scalable analytics is much easier said than done.

Keywords: - *Big Data, Cloud Computing, Data Security, Hybrid Cloud, Scalable.*

I. INTRODUCTION

Big Data that is the Data created from hundreds of sources, including e-mails, documents, apps, pictures, videos tweets, credit card data is unstructured, can't be easily compartmentalized. Data is huge so it is highly a difficult task to analyze and needs special bigdata analytic engines for greater efficiency & reduced cost. It will be necessary for organizations to coordinate activities effectively across the value chains. This can be done only if they have integrated platform that manages all the data from the device to the data center[1]. Organizations are moving to the cloud because of certain benefits like cost cutting, quality of service (QoS), scalability, flexibility and more importantly data security. There is huge demand for closed services in the market which overcome the issues such as Security, reduced infrastructure and operational cost. For example, Relational databases rely on the notion of logical data independence. The present techniques can be applied as how they are or with some extensions, to at least about what they want to compute, while the system determines how to compute it efficiently.

Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query requirements and, in principle, facilitates customers to choose between various vendors thus increasing the competition. The challenge is to combine these existing features of prior systems as we develop novel solutions to the many new challenges faced in the area of Big Data[1][2]. For example, in the early days of the cloud, one tended to view the network as a separate technology from the data centers and clients it linked to.

Today, Lamport's[2] theories of distributed computing are widely recognized as foundational steps towards solving problems of consistency, fault-tolerance and coordination. It will be necessary for organizations to coordinate activities effectively across the value chains. This can be done only if they have integrated platform as shown in Fig1 that manages all the data from the device to the data center which simplifies the whole process and thus maximizes the investments and lower the application development costs. Further, Organizations have started looking for mobility aware solutions that simplify application development process and allow developers to securely create and deliver more compelling user experiences. Private clouds are meant for specific group of organization that limits access to only that group. But the biggest challenge is the speed at which they process a user query which force ubiquitous network access, resource pooling consumption based billing are the features driving public cloud adoption among the enterprises.

Hybrid cloud concept combines the best of both worlds allowing the business to combine on premises private clouds with public closed services wherever appropriate. This trend is likely to increase owing to the reduction in infrastructure cost and its ability to protect mission critical workloads. Business increases especially in Infrastructure as a Service. Hybrid is essentially combination of at least two clouds. Hybrid clouds based on Azure.

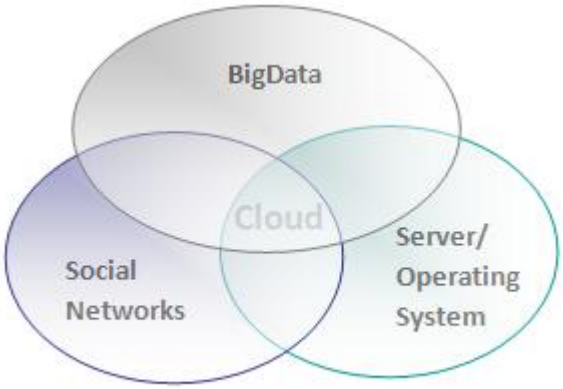

Figure 1: Next Generation Convergence Technologies

The cloud technologies enable the Big Data processing to take place in a much simpler manner because each cloud type is specific and uses different API tools to process the data. For example[2] the virtual machine capacity of Amazon EC2 is 1to20 EC2 processors and 1.7 to 15 GB of memory and 160 to 1.69 TB of storage[2], and it uses CLI or Web Service portal (WS). Cloud toolkits presently do not use virtual infrastructure managers instead manage virtual machines directly because consumers demand a flexible platform. If we look forward, there is a need to restructure the internet service environment, because people are preferring personalized and social services. So there is a need for personalized and social services in the whole service area, such as telecommunication, game, music, search, shopping, etc. This is why there is a need for cloud based service oriented architecture which is becoming more important to process Big Data.

II. BIG DATA

Big data means voluminous amount of data which is unstructured that require too many processes that is difficult to collect, store, manage, and analyze via general database software and it requires special purpose data analysis tools for processing the data.

2.1Challenges in Big Data

Big Data Integration is multi domain. in nature; that is only less than 10% of Big Data world are genuinely relational and meaningful data Integration in the real, schemaless and complex Big Data world of database and using multidisciplinary and multi technology methods. The challenges include not just the obvious issues of heterogenity, lack of structure, but also scalability, error handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. For example Web of cloud data contain 31 billion RDF triples, that 446 million of them are RDF links, 13 Billion government data, 6 Billion geographic data, 4.6 Billion Publication and Media data, 3 Billion life science data. Big Data efficiently engineers data at a large scale and searching and bringing together information that is relevant to the user and is highly challenging but mainly the engineering challenges include Prioritizing the data, Recognition of the links, extracting the Refined version of high quality data etc. that make it interesting for Big Data processing .

2.2Technologies for Big Data

Big data requires technologies to process large quantities of data more efficiently within the time limit. A 2011 McKinsey report suggests suitable technologies which include data fusion and integration, genetic algorithms, machine learning, natural language processing, crow sourcing, A/B testing, data mining grids. And includes cloud infrastructure and the internet. Rough set theory introduced by Pawlak and Fuzzy Set Theory are the most efficient statistical methodologies that can be applied to multidisciplinary Big Data, especially for clustering and characterization.

2.3Cost-Effectiveness in Cloud Computing vs. Datacenter Utilization

Storing, maintaining, and delivering massive amounts of data objects is an essential part where you can cost effectively store an infinite number of digital assets and serve them quickly and reliably to the users around the world. Data once uploaded has no limit on the cloud based storage as in Fig 2. The convergence service distributes data across multiple storage locations.

Most cloud data computing facilities have low utilization effecting both cost effectiveness and scaling the data centers[2]. To improve resource utilization, some datacenter users have turned to open source cluster management frameworks running on a single cluster as opposed to dedicating a cluster to a single workload. This optimizes the performance per watt metric, which is of vital importance to distributed computing.

The following equation is taken from the book distributed and cloud computing[2] .

$UserSpentHours_{cloud} \times (Revenue\text{-}Costof_{cloud}) \geq Userhours_{Datacenter} \times (Revenue - Cost_{Datacenter} / Utalization)$

III. ARCHITECTURE

The BigData processing can be implemented with multilayer architechture. The distributed multi layer parallel architechture distributes data across multiple processing units and parallel processing units process data much faster. The RDF can be used to maintain the sementics related to the data. The MapReduce[3] framework provides a parallel processing model and associated implementation for processing huge amount of data. With MapReduce, it is possible to split the user query and distributed across the parallel nodes in the cloud and processed in parallel (mapping phase). The results are then ranked and prioritised and delivered(reduce phase).the overall integration is done using again using DBMS technology.

3.1Platforms used for Big Data Analysis

We categorize various platforms into three layers. The bottom most layer[4] is is considered as storage system which is used to actually store the data. The middle layer takes care of handling of the data. And the top most layer comprises of various data processing tools using which the data is actually presented to the end user. The two main platforms used are Parallel DBMS technologies or No SQL and Map Reduce[4].

3.1.1 Storage System

There has been a significant impact on the storage technologies on quantifying the storage. The Storage efficiency techniques are to apply these technologies from one end to the other to avoid re-hydration of the data. Big Data requires storage systems that adapt to meet new requirements. High capacity & bandwidth storage systems are required to understand the management of Metadata. Theoretically, many millions of files and metadata records are centrally handled by a single file system[4]. The facility to share a much more distributed storage resource between processors is a challenge for storage management. The concept of storage being closer to the processor meets the trend making it easier by multi core processors that have additional cycles. Note that there are also the existing storage technologies, namely SAN and NAS, and cloud file storage systems, known as Amazon S3 or OpenStack Swift, as well as distributed file systems, such as GFS and HDFS, which are technologies for storing large data.

The data produced by corporate environment, include customer information, service log, and ERP (Enterprise Resource Planning). There are also CDR (Call Detailed Record) generated from telecom companies, machinery such as planes, oil pipelines, and power cables, or sensor data and metering data of platforms. Web servers or application logs could be a target, and there is user created data from social media blogs, mail, and bulletin boards. In short, everything that can be measured and stored could be a target for analysis.

3.1.2 Handling

Map Reduce is the most widely known technology that helps to handle large data which would be a distribution data process framework of the Map Reduce method, such as Apache Hadoop. Data processing via the Map Reduce is done with a regular computer that uses built in hard disk[5] which is not a special storage. Each computer has extremely weak correlation where expansion can be hundreds and thousands of computers. Since many computers are participating in processing system, errors and hardware errors are assumed as common, rather than exceptional. With a simplified and abstracted basic operation of Map and Reduce[5], one can solve many complicated problems. Programmers who are not familiar with parallel programs can easily perform parallel processing for data. It supports high throughput by using many computers. Data stored in the HDFS [5] storage is divided to available worker and expressed (Map) a value type, and results are stored in a local disk. The data is compiled by reducing worker and generate a result file.

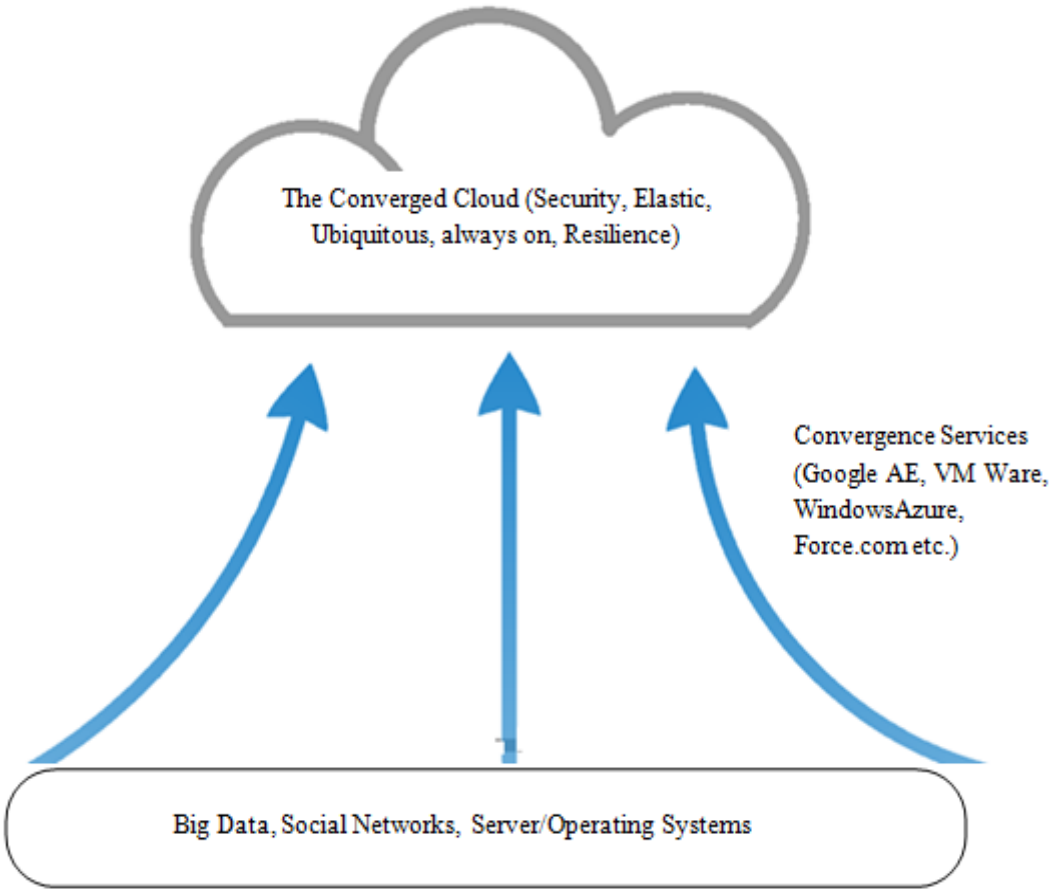

Figure 2: The Convergence Mechanism

3.1.3Analysis aspects

The process of finding meaning in data is called KDD (Knowledge Discovery in Databases). KDD is to store data, process/analyze the whole or part of interested data in order to extract progress or meaning value, or discover facts that were so far unknown and make them into knowledge ultimately. For this, various technologies are comprehensively applied, such as artificial intelligence, machine learning, statistics, and database.

3.2 Programming Big Data in R

R is a free software environment which compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. PbdR basically built on pbdMPI[6][7] depends on several R packages used to analyze Big Data by means of statistical methods. These packages support the SPMD (single program multiple data) programming model and uses most efficient clustering techniques which can be applicable to Big Data. The PbdMPI, PbdSLAP , PbdNCDF4, PbdBASE , PbdDMAT , PbdDEMO , are the set of packages that provide classes and interfaces and methods that implement Single Program and Multiple Data aspect of the parallel programming. The PmClust is a package that implements parallel model based clustering. The PbdPROF[7][8] package provides access to MPI profiling for complicated compiling and linking issues.

IV. CONCLUSION

Better technologies for analyzing the large volumes of data are becoming available. However, there is a requirement for making faster advances in many scientific disciplines for the sustainability and success of many enterprises. Many technical challenges described in this paper can be addressed before full potential is realized. These technical challenges are common across a large variety of application domains. The fundamental research must be explored before addressing these technical challenges if we are to achieve the potential benefits of Big Data.

REFERENCES

Journal Papers:

[1] Satoshi Tsuchiya,Yoshinori Sakumoto,Yuichi Tsuchimoto,Vivian Lee, Big Data Processing in Cloud Environments, Science Technical Journal, *FUJITSU,VOL 48,NO.2,April 2012*.

Books:

[2] K.Hwang, G.Fox, J.Dongarra, *Distributed and Cloud Computing* (University of Southern California, Elsevier Inc.).

[3] Norm Matloff, Davis, *Programming on Parallel Machines*. (University of California)

Others:

[4] Big Data White Paper.

[5] Jeffrey Dean, Sanjay Ghemawat,*Map Reduce :Simplified Data Processing on Large Clusters(Google inc.)*

[6] Jimmy Lin , Chris Dyer, *Data-Intensive Text Processing with MapReduce*

[7] Raim, A.M. (2013). *Introduction to distributed computing with pbdR at the UMBC High Performance Computing Facility* (Technical report). UMBC High Performance Computing Facility, University of Maryland, Baltimore County. HPCF-2013-2.

[8] http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html