www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume - 3 Issue -8 August, 2014 Page No. 7519-7524

AN OVERVIEW OF BIG-DATA AND CLOUD COMPUTING

K. Kala Bharathi

Department of Computer Science,
St. Pious Degree & P. G. College for Women,
Nacharam, Hyderabad, India-500 076
kudikalakalabharathi@gmail.com

Abstract:

Big data and Cloud computing both are most popular topics of recent days for organizations across the globe. Big data is a rapidly expanding research area spanning the fields of computer science, information management and it has become a ubiquitous term in understanding and solving complex problems in different disciplinary fields such as engineering, applied mathematics, medicine, computational biology, healthcare, social networks, finance, business, education, transportation and telecommunications. Cloud computing is a service delivered over the internet for computation, data accessing & storage by creating scalability, flexibility and minimum cost. It is a next generation platform for computation which offers various services and applications to the user without physically acquiring them.

Introduction:

Big organizations have grown the data associated with them also grew exponentially. Data never gets old and it is going to stay there forever. Most of the big organizations have data in multiple applications and in different formats. Indeed they are facing challenges to keep all the data on a platform which give them a single consistent view of their data. A research survey by Gartner in 2011 showed how there is a dramatic rise in data generation as compared to data available for analysis, which means that more and more data is flowing out unchecked without analysis. This yielded a revolution to new development named Big Data. Big data traveling around to find, visualize and understand large data to improve decision making. Big data refers to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze the information".[1]

Big data for development is about turning imperfect, complex, often unstructured data into actionable information. For this they used advanced computational tools which have developed in other fields, to reveal trends and association within and across large data sets that would otherwise remain undiscovered. It is difficult to work with using most relational database management systems or traditional data processing applications and desktop statistics and visualization packages, requiring instead massively parallel software running on tens, hundreds, or even thousands of servers and they may take tens or hundreds of terabytes before data size becomes a significant consideration. Above all, it requires human skill and outlook. The term big data varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain.

"Cloud computing is a model for on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction".[2] Before cloud computing traditional business applications have always been very complicated and expensive. The amount and variety of hardware and software required to run them are scary. We need a whole team of experts to install, configure, test, run, secure, and update them. With cloud computing, we can eliminate those headaches because we are not managing hardware and software-that's the responsibility of an experienced vendor. Present is the age of information technology. The aspect of work and personal life are moving towards the concept of availability of everything online. Understanding this trend, the big and massive web based companies like Google, Amazon, Salesforce.com came with a model named "Cloud Computing" the sharing of web infrastructure to deal with the internet data storage, scalability and computation (Kambil, 2009). The shared infrastructure means it works like a utility. We only pay for what you need, upgrades are automatic, and scaling up or down is easy. Cloud-based apps can be up and running in days or weeks, and they cost less. With a cloud app, we just open a browser, log in, customize the app, and start using it. Anything from basic word processing to collaboration to e-mail to multimedia processing can be accomplished more efficiently using cloud computing than using one's personal computer. As cloud computing is Internet-based social networking services, it offers companies an easy method of promoting their product or peoples an easy way of promoting and publicizing themselves without building an infrastructure of their own and it as a potential to enhance business agility while enabling greater efficiency & reducing costs. Now-a-days most of the world's largest companies are simply moving their non-cloud products and services to cloud computing and running all kinds of apps in the cloud, like customer relationship management (CRM), HR, accounting, and much more.

Big data Technologies:

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times.

Column-oriented databases

Traditional, row-oriented databases are excellent for OLTP (Online Transaction

Processing) with high update speeds, but less on query performance as the data volumes grow and as data become more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.

Map Reduce

This is a programming model that allows for massive data processing, scalability against thousands of servers or clusters of servers. Map Reduce implementation consists of two tasks:

- The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
- The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

Hadoop:

Hadoop is by far the most popular implementation of MapReduce. This is developed by Google but Apache in developed a generalized software framework today called as Hadoop, being an entirely open source platform for handling Big Data. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

Hive:

Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization and supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3filesystem. It provides an SQL-like language called **Hive QL** while maintaining full support for map/reduce. To accelerate queries, it provides index. It is initially developed by Face book.

PIG:

SQL is a data flow language, it is not suitable for big data so developed PIG. PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, it can be written in Java and other languages also. PIG was developed by Yahoo!

Cloud computing providers offer their services according to several fundamental model:

Software as a service (SaaS):

It provides user with facility to run applications on a cloud infrastructure. The applications are accessible through a front end portal. Clients no need to install any software on their device. They just need a web browser and network connection. SaaS applications are designed for end-users, delivered over the web .The consumer does not manage or control the underlying cloud infrastructure. User has to pay only for services which he used. For example, web-based emails, Microsoft Office365 etc., With, Microsoft Office 365 you can use services of word, without installing it on your computer. You just need to pay monthly fee. You can use this software anytime from anywhere.[3,4]

Platform as a service (PaaS):

PaaS is the set of tools and services designed to make coding and deploying those applications quick and efficient. It provides platform for developers to create their applications on provider's platform over the internet. PaaS provides physical server for software. Consumer does not have control over cloud infrastructure, but can control the deployed applications. PaaS is mostly used by software development teams. Ex. Google Apps.

Infrastructure as a service (Iaas):

Customer can rent a data center environment without worrying about maintenance i.e. it provides users with virtual servers. IaaS generally includes multiple users on a single piece of hardware [2]. It provides individual servers, disk drives, computing resources, private networks, messaging systems etc. All services are provided to user by paying some fee. User has to pay only for services that he/she uses. i.e pay for what you use model is applied. An organization can build a complete infrastructure using IaaS. The consumer does not have control over cloud infrastructure but can control deployed applications, operating systems, storage devices etc. Ex: Amazon Web Services.

The early driver of widespread adoption of cloud computing was the SaaS delivery model. It offered the user scalability and customization based on their needs and goals. Since then, a combination of technologies has emerged which have further increase the demand for cloud computing.

- **Server virtualization:** Various hardware and software resources are pooled together and users are offered access. It gives the same appearance and capabilities of a dedicated server, but without the cost.
- **Service-oriented architecture (SOA):** Organizes software code so that one set of data, and the code written to process it, can be reused by other applications in the organization.
- **Open source software:** A product's source code is made available to the public with little to no copyright restrictions.
- **Web development:** Basic website development services have driven down the cost and made updating possible for less technically skilled workers.

Big Data Characteristics:

The **3Vs** that define *Big Data* are **V**ariety, **V**elocity and **V**olume.

Variety:

Data can be stored in multiple formats. For example database, excel access or in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something. It is the need of the organization to arrange it and make it meaningful. It will be easy if we have data in the same format, but it may not be in all the time. This kind of challenges we can overcome with the *Big Data*. These varieties of the data represent **Big Data**.

Velocity:

In simple words velocity means "The speed of data in and out". The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. On social media sometimes a few seconds old messages is not something interests users. They often remove old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent **Big Data**.

Volume:

We are currently seeing very fast growth in data storage not only Text data but data in different formats like Videos, music and large images. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. The big volume indeed represents **Big Data**.

Cloud Computing Characteristics:

On demand self services:

Computer services such as email, applications, network or server service can be provided without requiring human interaction with each service provider. On demand self services include Amazon Web Services (AWS), Google.

Broad network access: Cloud Capabilities are available over the network and accessed through standard mechanisms such as mobile phones, laptops and PDAs.

Resource pooling: The provider's computing resources are pooled together to serve multiple consumers using multiple-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The resources include among others storage, processing, memory, network bandwidth, virtual machines and email services. The pooling together of the resource builds economies of scale (Gartner).

Rapid elasticity: Cloud services can be rapidly and elastically unrestricted, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

Measured service: Cloud computing resource usage can be measured, controlled, and reported providing transparency for both the provider and consumer of the utilized service. Cloud computing services use a metering capability which enables to control and optimize resource use. This implies that just like electricity or municipality water services are charged per usage metrics-pay per use. The more you utilize the higher the bill.

Multi Tenacity: It is the 6th characteristics of cloud computing advocated by the Cloud Security Alliance. It refers to the need for policy-driven enforcement, segmentation, isolation, governance, service levels, and chargeback/billing models for different consumer constituencies. Consumers might utilize a public cloud provider’s service offerings or actually be from the same organization, such as different business units rather than distinct organizational entities, but would still share infrastructure.

Combined working of Big data and Cloud Computing - benefits:

Two IT initiatives are currently top of mind for organizations across the globe. They are big data and Cloud.

Most cloud vendors are already offering hosted Hadoop clusters that can be scaled on demand according to their user's needs. Also, many of the products and platforms are either entirely cloud-based or have cloud versions themselves. As cloud computing continuous to mature, a growing number of enterprises are building efficient and agile cloud environments i.e. cloud computing has a structure to support their big data projects. Large to Medium sized companies are getting more value from their data than ever before with the Cloud computing, by enabling intense fast analytics at a fraction of previous costs. Thus, currently Big data & cloud technologies are using as a pair organizations to make big data analytics in clouds a reasonable option, since, data is becoming more valuable. This, in turn drives companies to acquire and store even more data, creating more need for processing and retrieving. Thus, Big Data and cloud computing go hand-in-hand.

Conclusion:

In my view this paper presents an outlines of cloud computing and big data. Cloud computing has become major discussion thread in the IT world. It is reduces the cost of purchasing physical infrastructures like email servers and software. Big data is best for fast query performance, massive data processing and scalability. Hence, it is not suitable for small business i.e. where fewer amounts of data are capture, manage and process.

Acknowledgment:

I thankful to The Principal and The Head (Department of Computer Science), St. Pious X Degree & P.G. College for Women for providing literature facilities and also I thankful to my colleagues for encouraging me in this work.

References:

1. White, Tom (10 May 2012). *Hadoop: The Definitive Guide*. O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.

2. PS Ryan, S Falvey, Sarah and R Merchant (2013-12-19)."When the Cloud Goes Local: The Global Problem with Data Localization". IEEE Computer. Retrieved 2013-02-17.

3. M. Sharma, H. Bansal, A.K.Sharma, Cloud Computing: Different Approach & Security Challenge, IJSCE,2012

4. T. B. Winans, J. S. Brown, Cloud Computing, A collection of working papers.