Cloud Computing: BDaaS and HDaaS (Big Data as a Service and Hadoop as a Service)

Ujjwal Agarwal

Salalah College of Technology, Salalah, Sultanate of Oman, Oman

*Corresponding Author: drujjwalagarwal@gmail.com Tel.: +968-99651617

Abstract: Two technologies currently change the world and most important for organizations: Cloud Computing and Big Data. Data is everywhere; this data is generated by the organization, people and machine. Insight analysis of this data is most important. Cloud computing provide the services for storage, process and analysis of this large and complex data sets that can create competitive advantage, spark new innovations. The main objective of this paper to describe the concepts of big data, cloud computing, Big Data as a Service (BDaaS) and Hadoop as a Service (HDaaS) on cloud platform. . Apache Hadoop is an open source software framework to store and analysis of Big Data. Demo IBM Ambari Console is used to demonstrate the working architecture of Big Data Hadoop on Cloud platform. Focusing on Infrastructure as a service, platform as a Service and Software as a Service in terms of Big Data model.

Index Terms—BDaaS, HDaaS, Big Data, IBM Ambari Console, Cloud Computing, SaaS, IaaS, PaaS

I. INTRODUCTION

A big Data is the greatest challenges for digital world to store, transport, and process and insight analysis of data. Cloud computing provides fundamental support to address the challenges with shared computing resources including storage, networking and analytical software. Cloud computing offers access to data storage, processing, and analytics on a more scalable, flexible, cost-effective, and even secure basis than can be achieved with an on-premise deployment [1]. These characteristics are essential for customers when data volumes are growing exponentially-to make storage and processing resources available as needed, as well as to get value from that data.

1.1 Big Data: In today's world data is everywhere, either in shop, office, hospital, colleges, universities, mobile application, websites, digital devices, etc. Technologies are changing with extremely high rate and producing a huge amount of data, which comes in a variety of form along with a high velocity. Big data is a growing term that describes any capacious amount of structured , semi structured and un-structured data that has the potential to be mined for information [2]. There are three sources of big data:

a)   Organization generated data
b)   People generated data
c)   Machine generated data

Organization generated data: is the most structured data because all the data is stored in different databases either in MS ACCESS, Oracle, MS SQL etc. in form of tables, rows and column. To search any information in database is an easy task as the data already in particular shape.

People generated data: is in semi-structured or un-structured form like Facebook messages, twitters tweet, Google search query, pictures on Instagram, Picasa. Social media is main platform where people will generate terabyte/petabyte of data every day.

Machine generated data: This is the major source of big data about 85% - 90% of the total data is generated by machine. This data is generated by various real time sensors, camera, mobile; devices like sleeping monitor, heart beat monitoring devices. The widespread availability of smart devices like smart phone, smart cars, smart homes, smart camera, now sea, oceans, and sky also connected to this devises and all together producing a huge amount of data every minute, hours and days.
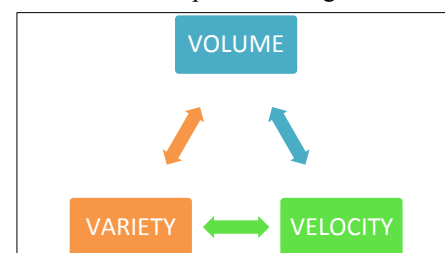
3V's are most important in big data i.e.:-

Figure: 1.1

In big data the quantity or volume of data is very high, every minute total amount of data is generated may be in terabyte or petabyte. This data is generated at a very high velocity for example every minute Facebook generated about 350GB of data, 72 hours video uploaded on YouTube. Data can be Batch, near time, real time or stream data. As data is coming from various sources therefore the structure of big data varies it can be structure, semi-structure or un-structure.

1.2 Cloud Computing: Cloud computing is a complete new technology. It is the development of parallel computing, distributed computing grid computing, and is the combination and evolution of Virtualization, Utility computing. In general cloud provides three types of services: Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) [3]. Cloud gives the opportunity to access the data, software and resources anytime, anywhere regardless of physical location; just we need the internet connection and an individual decision maker finds himself or herself with the technology. Cloud provides the hardware, infrastructure, network, storage, software's application. There are various reasons for organizations to move towards IT solutions that include cloud computing as they are just required to pay for the resources on consumption basis. In addition, organizations can easily meet the needs of rapidly changing markets to ensure that they are always on the leading edge for their consumers [4]. There were many different whitepapers and general introductions to cloud computing, which provide an overview of the cloud computing, [e.g. 5, 6, 7, 8, 9]
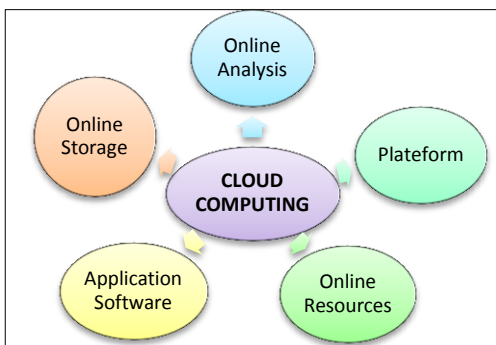

Figure 1.2

Cloud computing provides all different type of services like Online analysis, different Application software, storage, networking and platform to run the applications and software under cloud environment.

II. BIG DATA AS A SERVICE (BDaaS)

Big Data as a Service (BDaaS) are cloud based service for storage, processing and analysis of large or complex data sets that required high scalability.
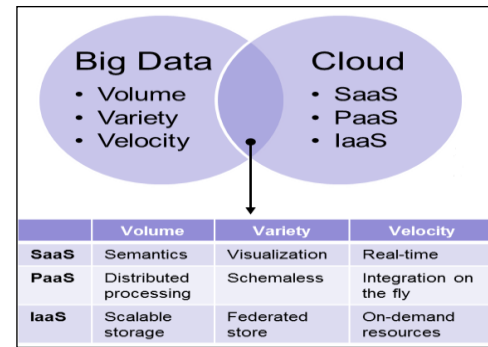

Figure 2.1

Big data become big because of its high volume, velocity and variety of data. Cloud computing provide the services to manage Big Data. The three main services of cloud are:-
Infrastructure as a service (IaaS)
Platform as a service (PaaS)
Software as a service (SaaS)

2.1 Infrastructure-as-a-Service (IaaS): Starting from the base, the first service provided by the host cloud is to provide the all necessary hardware; it provides the network capability, storage, computing, management and support components (virtual servers). This infrastructure accessed by the user via internet, enabling the organization or user to transfer his data. The following is a sample of IaaS solutions from providers in the cloud technology ecosystem.
• Amazon Web Services
• Citrix Cloud Platform
• Windows Azure and Microsoft System Center
• OpenStack Software
• Rackspace
• Savvis
• Verizon Terremark
• VMware vCloud Suite

IaaS in a BDaaS: Using a cloud environment for Big Data needs almost a huge amount of storage memory may be in Petabyte or Exabyte and compute power to store and retrieve the information. IaaS can be utilized by the organization or user to store high volume of data, even if the scale of business is increase they can tap into the cloud resource as and when they need it rather than purchase, install and integrate hardware themselves.

2.2 Platform-as-a-Service (PaaS): PaaS function above the IaaS and below the SaaS. PaaS provides a platform on which the application or software will work. The main work of PaaS is to build, test, deploy, and run applications on cloud infrastructure.
The following is a sample of PaaS solutions from providers in the cloud technology ecosystem.
• Force.com

• Google App Engine
• Red Hat OpenShift
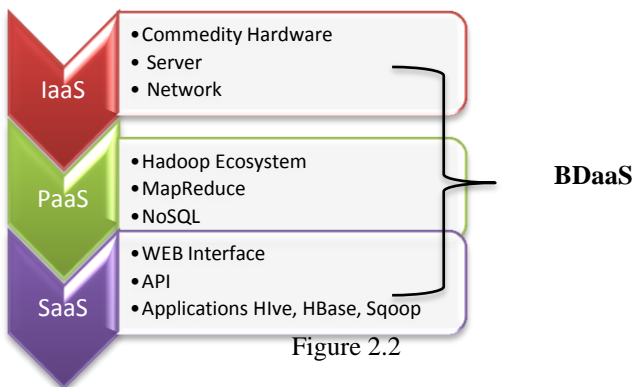• VMware Cloud Foundry
• Windows Azure

PaaS in a BDaaS: When Big data technology incorporate in cloud infrastructure, PaaS vendors are providing the platform such as Hadoop and MapReduce, which eliminate the dealing with the complexities of managing individual software and hardware elements..

2.3 Software-as-a-Service (SaaS): SaaS is top layer service in cloud platform. SaaS provides the application or software to the client or organization.
The following is an example of SaaS from different companies on cloud computing ecosystem.
• Amazon Elastic MapReduce
• *Cetas by VMWare Analytics Solutions*
• Google BigQuery services
• Rackspace Hadoop Service
• Windows Azure HDInsight

SaaS in a BDaaS: SaaS is software provider for user or organization as per there need. For example any organization want the analysis of their customer feedback on twitter or Facebook, SaaS will provide the platform or application for analysis. Office software is the best example of businesses utilizing SaaS. Tasks related to accounting, sales, invoicing and planning can all be performed through SAAS.

Figure 2.2

III. HADOOP AS A SERVICE (HDAAS)

Hadoop is a most major platform for storage, processing and analysis of big data. Hadoop is an apache open source program. Google did all efforts towards developing Hadoop by implementing MapReduce technology and algorithm. Google solved many problem related to the complex structure of Big Data. MapReduce is algorithm divide the problem in to two different phases, first is Map and another is Reduce, between map and reduce we have sort and shuffle. Using this solution by Google, Dough cutting and his team developed Hadoop platform. The first release of Hadoop was launched on 10th December 2011[10].

Hadoop as a Service (HDaaS) in the cloud makes big data project very convenient because organization or user need not worry about the storage machine platform and different application on Hadoop Ecosystem. Some benefits of HDaaS offerings are:
Managed Hadoop – No need to hire a System Admin
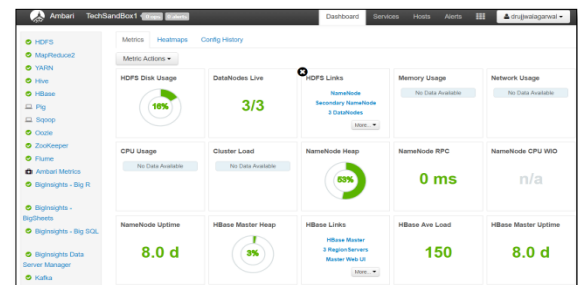Ease of use – Built to get started quickly
No hardware/infrastructure – User or customers no need to purchase any hardware or servers, cloud will provide the necessary hardware.
Support – Each company provides the help desk if any issue will arise.
The following is an example of cloud Hadoop:-
Aleron, Altiscale**,** Amazon EMR**,** CenturyLink**,** CSC**,** Google Cloud Storage connector for Hadoop**,** IBM BigInsights , Microsoft's Hadoop cloud service etc.

The Demo Cloud Apache Ambari Colnos is an example of Hadoop as a Service (HDaaS), this project aimed to provide, managing and monitoring the Apache Hadoop clusters. IBM Ambari provides an easy solution to use Hadoop on cloud platform. Ambari enables System Administrators to: Provision a Hadoop Cluster Ambari provides a step-by-step wizard for installing Hadoop services across any number of hosts. Ambari will handle all management of Hadoop services for the cluster.

Dashboard IBM Ambari Console: Figure 3.1

SaaS: Software as a service, There are number of services provided by the IBM Ambari console including,
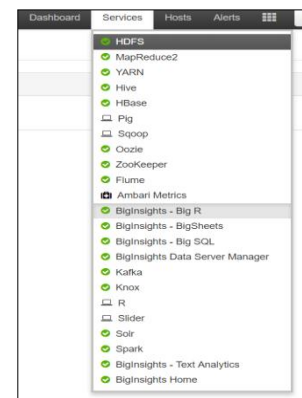

Figure: 3.2

PaaS: Platform as a Service: IBM Ambari provides the platform for the Hadoop and its Ecosystem. To run the MapReduce and Yarn job it gives the platform to execute all different jobs.
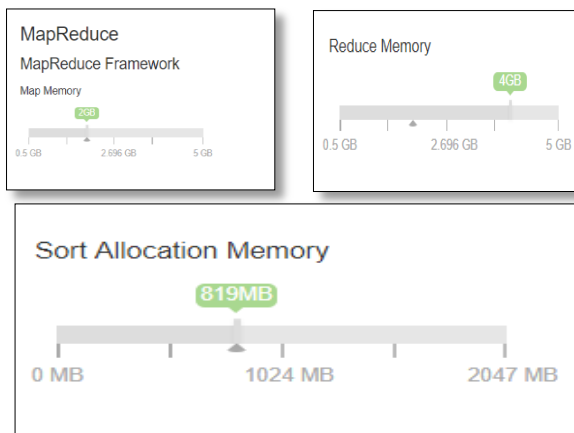

Figure: 3.3

IaaS: Infrastructure as a Service: IaaS provides the infrastructure for storage the big data on Hadoop distributed file system. As shown in the figure that in Ambari DFS used is 4.7 terabyte. The replication factor of file in data node is 3. Cloud can store unlimited number of life on its commodity hardware.
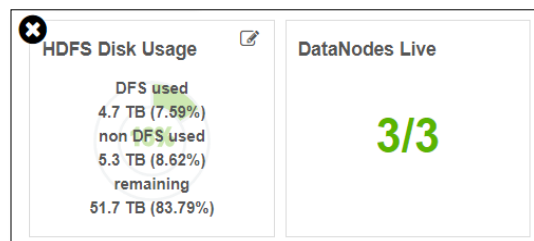

Figure: 3.4

IV. CONCLUSION

This paper presents the fundamental concepts of Big Data; cloud computing and how Big Data as a Service and Hadoop as a Service works on Cloud environment. Cloud computing provides the services to store, process and insight analysis of complex big data by providing resources on-demand with costs proportional to the actual usage. Cloud computing play an important role for Big data not only it provides an infrastructure but also it provides the platform to run the Hadoop Distributed File System commands, MapReduce jobs and Software for Hadoop Ecosystem like Sqoop, Pig, Hive, Flume etc. Three main services on cloud for Big Data as a service namely: Infrastructure as a service (IaaS), Platform as a service (PaaS) Software as a service (SaaS). For Infrastructure it provides the all necessary hardware components like network, storage memory etc. In terms of

platform it provides the operating system or environment to run the different applications like mapreduce jobs. Finally it provides the software for data analysis like R connectors, Hive, NoSQL etc.

Demo IBM Ambari Console, is used to demonstrate the working architecture of Hadoop as a Service on Cloud environment. It provides the platform to store the files and later we can access and do analysis on it.

V. REFERENCE

[1] Harshawardhan S. Bhosale, *Review Paper on Big Data and Hadoop*, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153
[2] Raghavendra Kune, *The anatomy of big data computing*, software: practice and experience Softw. Pract. Exper. 2016; 46:79–105 Published online 9 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2374
[3] Santosh Kumar, *Cloud Computing – Research Issues, Challenges,Architecture, Platforms and Applications: A Survey International Journal of* Future Computer and Communication, Vol. 1, No. 4, December 2012.
[4] A. Kundu, C. D. Banerjee, P. Saha, "*Introducing New Services in Cloud Computing Environmen*t", International Journal of Digital Content Technology and its Applications, AICIT, Vol. 4, No. 5, pp. 143-152, 2010.
[5] R.Piplode, P. Sharma and U.K. Singh, "*Study of Threats, Risk and Challenges in Cloud Computing*", International Journal of Scientific Research in Computer Science and Engineering, Vol.1, Issue.1, pp.26-30, 2013.
[6] Amol C. Adamuthe, *Cloud Computing – A market Perspective and Research Directions*, I.J. Information Technology and Computer Science, 2015, 10, 42-53 Published Online September 2015 in MECS (http://www.mecs-press.org/)
[7] VARIA, J.2009. Cloud Architectures. Amazon Web Services.
[8] CHAPPELL, D.2009. Introducing the Azure Services Platform. David Chappell & Associates.
[9] RAYPORT, J. F. and HEYWARD, A.2009. *Envisioning the Cloud: The Next Computing Paradigm. Marketspace*.
[10] Palaghat Yaswanth Sai, Pabolu Harika, "*Illustration of IOT with Big Data Analytics*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.221-223, 2017.

Authors Profile

Dr. Ujjwal Agarwal completed his PhD(Computer Science), M.Phil. (Computer Science) and MSc(Information Technology). He is currently working in Salalah College of Technology, Salalah, Sultanate of Oman as a Lectirer (IT) alogwith he is member in many International Journals. He has published more than 12 research papers in reputed international journals and conferences including IEEE and it's also available online. His main research work focuses on Big Data Analytics, Data Mining, IoT and Data Sciences. He has more then 12 years of teaching experience and 6 years of Research Experience.