IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-11 E-ISSN: 2347-2693

A Proposed Big Data as a Service (BDaaS) Model

Mazin S. Al-Hakeem

Department of Information Technology, Lebanese French University (LFU), Erbil-Iraq

e-mail: mazin_ictc@yahoo.com

Available online at: www.ijcseonline.org

Received: Oct/19/2016 Revised: Oct/26/2016 Accepted: Nov/20/2016 Published: Nov/30/2016

***Abstract*— Big Data can bring big benefits for all sectors of our life via smarter moves, for examples, by analyzing huge dataset immediately and allowing for making decisions based on what they have learned, by gauging customer needs immediately and analyzing customer satisfaction in a timely manner, or by providing many diagnosis or treatment options quickly. These can driving business and economy growth. Until recently, it was hard to get benefits of Big Data without heavy infrastructure investments; for that, the enterprises suffered from many challenges which related to the lack of capacity to process and store the huge dataset adequately, and inability to manage and extract value from these huge dataset; but times have changed. The technology of cloud computing was evolved rapidly to bridge the storage and processing gap and opened up a lot of options for using Big Data by both individuals and organizations without having to invest in massive on-site storage and data processing facilities. This paper presents the concept, advantages, characteristics, processing and applications of Big Data. Then proposes a model to integrate Big Data and cloud computing technology based on three basic cloud service layers to present a new model of Big Data as a Service (BDaaS). The proposed BDaaS model allows enterprise to implement various Big Data functions using variety outsourcing (like Hadoop, Altiscale and Qubole) clearly, easily and moving them out of the expensive whirlpool of updating and maintaining their infrastructure.**

Keywords- Big Data; BDaaS; Cloud Computing; Hadoop; Altiscale; Qubole.

I. INTRODUCTION

Big Data, generally, refers to the ever-growing amount of information that created and stored by everything around us at all times, and analyzed using new analytics technologies in order to drive business and economy growth [1]. Big Data is a field dedicated to the analysis, processing and storage of huge collections of data that frequently originate from disparate sources [2]. These huge collections of data generated by everything around us at all times, every digital process and social media exchange produces it; systems, sensors, IoT applications and mobile devices transmit it [1].

Nowadays, ninety percent of data has been created in just the last two years; analysts see that every day users create 2.5 quintillion bytes of data, for that the amount of data will increase 30 times by 2020 [2].

Unfortunately, Big Data facing a big challenge, because the traditional data analysis, processing and storage technologies and techniques are insufficient to deal with the huge collections of data that increase rapidly [1].

To face these problems and to extract meaningful value from these huge collections of data, Big Data adds newer techniques that leverage computational resources as well as dependence on cloud computing and high-speed data networks which decrease storage costs and provides a better storage and processing solutions [3].

Public and hybrid cloud that evolved rapidly offers cheap storage, fast processing, ubiquitous data collection and availability of third party data. Enterprises can expand their current data strategy by turn to Big Data as a Service (BDaaS) solutions to bridge the storage and processing gap and get advantages of existing technologies and infrastructures of popular cloud [4].

At the moment, Big Data as a Service (BDaaS) is a somewhat nebulous term often used to describe a wide variety of outsourcing of various Big Data functions on the cloud [4].

The significant ideas behind this paper are to produce a new proposed model for BDaaS and to open the door for research community to perform newer research work on the BDaaS Model.

And the original scientific contribution is the propose shifting cloud computing layers paradigm to cover the variety outsourcing for various Big Data functions.

For that, this paper presented in the following structure. In the next section, the literature review will be presented. The concept, advantages and characteristics of Big Data sections will be followed. This is followed by the Big Data processing. The Big Data applications will be followed. This is followed by proposed model of Big Data as a Service (BDaaS). The paper ends with a conclusion section.

II. LITERATURE REVIEW

There are a few researchers who focused on develop BDaaS model. Christian Prokopp, discussed the advantages and differences between the four possible combinations for BDaaS (PaaS only, IaaS and PaaS, PaaS and SaaS and IaaS and PaaS), then classified them into one of four types (Core BDaas, Performance BDaaS, Feature BDaaS and Integrated

BDaaS) [5]. In his article, Bernard Marr presents the clear term of BDaaS and explain the useful of it [6]. Beyond that, Poornima Sharma and others presented the data mining analysis algorithm on MapReduce as a services in cloud environment to find the most frequently occurred pair of products in baskets at a store [7]. Nowadays, many companies' runs different types of commercial BDaaS to fulfillment they different needs.

However, our proposed model depend on how to shift the cloud computing layers paradigm to cover the variety outsourcing for various Big Data functions.

III. BIG DATA CONCEPT AND IT'S ADVANTAGES

According to dictionary of Gartner (one of the best world's leading information technology research and advisory company), Big Data is "Assets of large, high speed and/or high diversity information that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" [3].

The Gartner definition based on the concept of "Three V's" of Big Data that shown in "Fig. 1". This concept originally coined by Doug Laney in 2001 [6], for refer to the challenge of huge data management.

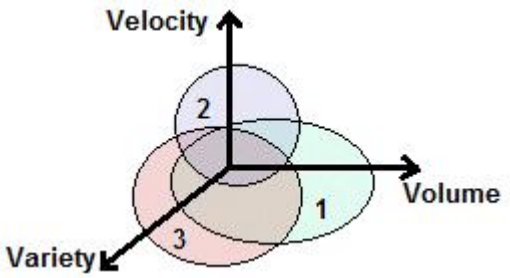

Figure 1. The Three V's Big Data Concept

Essentially, the "Three V's" concept of Big Data measures the following [1] [3]:

- Volume (1stV): Amount of data (or scale of data) and the challenges that related to its storage and analysis.
- Variety & Complexity (2ndV): Range of data types (or diversity of data) and sources and the challenges that related to collects it from diverse sources.
- Velocity (3rdV): Speed of data in and out and the challenges that related to the access and analysis it to fulfilment the business requirements.

Later in 2012, Gartner updated the definition of Big Data as high volume, high velocity and/or high variety information assets [8].

More than that, IBM analytics found that the poor data quality costs US economy more than three trillion dollars a year [9]. Therefore, IBM was developed a new V for Big Data. The Veracity (4thV) of IBM is certainty of data. The IBM Four V's Big Data Concept shown in "Fig 2".

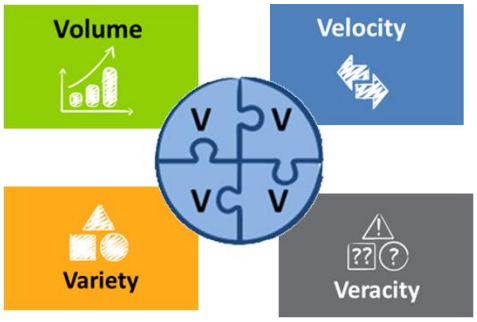

Figure 2. The IBM V's Big Data Concept

Enterprises can get advantages of Big Data by helps Organizations, companies, institutes, etc. harness their data and use it to identify new opportunities. That, in turn, leads to smarter moves, more efficient operations, higher profits and happier customers [3]. New Big Data technologies like Hadoop and in-memory analytics, combined with the ability to analyze new sources of data will helps to get the following significant advantages [3] [9]:

- Cost reduction, by storing large amounts of data, plus the ability to identify more efficient ways of doing business.
- Faster and better decision making, by analyzing information immediately and make decisions based on what they've learned.

New products and services, by gauging customer needs and satisfaction in real time, more companies are creating new products to meet customers' needs.

IV. BIG DATA CHARACTERISTICS

According to the IBM 4V's principle, the characteristics of Big Data includes the following [1] [9]:

- Size: The size of the data is very large (measured by Terabytes and Petabytes), 90% of today's data has been created in just the last 2 years.
- Diversity: The diversity of these data is waving between structured (such as files, databases), unstructured (such as text files) and semi-structured (such as XML) files, 90% of generated data is unstructured, this includes tweets, stock indicators, customer purchase histories and customer service calls.
- Speed: The speed of data occurrence frequency (for example, the speed of tweets differ from the speed of remote sensor scans for climate changes), 50.000

GB/second is the estimated rate of global Internet traffic by 2018.

- Veracity: The certainty of data, 2 in 3 business leaders trust the information they use to make decisions.

V. BIG DATA PROCESSING

Traditional data analysis, processing and storage technologies and techniques are insufficient to deal with the huge diverse data that increase rapidly and contain new types of data that cannot be ignored, such as photographs and videos audio, video, three-dimensional models and geographical data and more [1].

To face this problem, Big Data adds newer techniques that leverage computational resources as well as dependence on cloud computing and high-speed data networks to execute new analytic algorithms for Big Data processing, these techniques includes:

A. Hadoop Framework:

Hadoop is an open-source framework that allows for distributed storage and distributed processing of huge datasets across clusters of computers [11] [12]. Hadoop framework is composed of the following modules [11]:

- Hadoop Common: It contains libraries and utilities needed by other Hadoop modules.
- Hadoop Distributed File System (HDFS): It is a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN: It is a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications.
- Hadoop MapReduce: It is an implementation of the MapReduce programming model for large scale data processing.

To process huge data, Hadoop takes advantages of data locality paradigm and the existing high-speed network, it splits files into large blocks and distributes them across nodes in a cluster to process in parallel based on the data that needs to be processed [11]. The dataset will be processed faster and more efficiently with no matter how huge [13].

In 2006, Hadoop was created by Doug Cutting when he was working in Yahoo!. By 2007, Yahoo started using Hadoop on a 1000 computer cluster. After that, Hadoop was being used by many other companies besides Yahoo!, such as Facebook, Amazon, Twitter, eBay, Microsoft, Apple and more [11] [13].

B. MapReduce Algorithm:

MapReduce is a computing paradigm for processing data that resides on hundreds of computers to provide scalability and easy huge data processing solutions [13] [14].

MapReduce algorithm runs in the background of Hadoop to divides a task into small parts and assigns them to many computers. Later, the results are collected at one place and integrated to form the result dataset [15].

In 2004, MapReduce was released by Google. By 2006, Yahoo! created Hadoop based on Google file system and MapReduce [14].

C. Big Data Processing Software Packages:

Additional software packages that can be installed on top of or alongside Hadoop, such as Pig, Spark, Splunk, Hive, HBase and Altiscale. However, there are many other software packages Phoenix, ZooKeeper, Cloudera Impala, Flume, Sqoop, Oozie, Storm and more.

1) PIG:

Pig is a platform for creating programs that run on Hadoop. It uses for quickly gather, sample, view and perform simple analysis on HDFS data. It is a dataflow engine for Hadoop. In 2006, Pig was developed at Yahoo! for researchers to have an ad-hoc way of creating and executing MapReduce jobs on very large data sets. [1] [16]

2) Spark:

Spark is an open source cluster computing framework to provides a higher level interface to process data (data exploration, cleaning and plotting) and write more expressive code. It can access diverse data sources including HDFS, HBase, Hive and any Hadoop data source. In 2014, Spark was developed at the University of California, Berkeley's AMPLab by Matei Zaharia [1] [17].

3) Splunk:

It is a platform to collect, analyze and deliver real-time insights from machine-generated big data. Also, it is a Log analytics on large dataset and gain valuable insight. In 2003, Michael Baum, Rob Das and Erik Swan was developed as a Splunk Enterprise product [1] [18].

4) Qubole

Qubole is a platform developer that allows users to prepare, integrate, and analyze big data in the cloud. it includes Qubole Data Service, manages Hadoop infrastructure. In 2012, Qubole launched, it was founded by Ashish Thusoo and Joydeep Sen Sarma when they worked as a senior big data engineers at Facebook [1] [19].

5) Hive (Hadoop-based data warehouse):

Hive is a data warehouse infrastructure based on Hadoop for providing data summarization, query, and analysis. It gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. The stable version of Hive was release in 2015 [1] [20]

6) *HBase (Hadoop's database):*
HBase is an open source, non-relational, distributed database. In 2010, HBased was developed as part of Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System) [1] [21].

7) *Altiscale*
Altiscale runs Hadoop on hardware that is purpose-built and tuned for Hadoop. It configures the kernel and network parameters on top of the hardware for Big Data performance. In 2012, Altiscale was designed by ex-Yahoo CTO Raymie Stata to free users from the complexities of deploying, managing, and scaling a big data platform [1] [22].

VI. BIG DATA APPLICATIONS

There are many different sectors and industries might benefits from Big Data technologies, such as the following [10] [23]:

- Travel and Hospitality: Big data gives travel and hotel industry the ability to collect customer data, apply analytics and immediately gauge customer satisfaction or identify potential problems in a timely manner.
- Health care: Big data gives health care industry the ability to analyze patient records, health plans and other information related to patient; and providing diagnosis or treatment options quickly.
- Government: Big data gives government agencies, i.e. law enforcement agencies, the ability to streamlines operations to keep crime rates down and giving a more holistic view of criminal activity.
- Retail: Big Data gives retailers the ability to understand of their customers' needs, predict trends, and recommend new products when their customers need it.

VII. BIG DATA AS A SERVICE TERM

Big Data as a Service (BDaaS) is an outgrowth of two essential IT trends, Big Data and Cloud Computing. Till now, BDaaS is a somewhat nebulous term often used to describe a wide variety of outsourcing of various Big Data functions on the cloud [22].

Big Data functions can range from the supply of huge data, large datasets storage, large datasets processing, large datasets analysis and providing reports [1] [24].

VIII. A PROPOSED BIG DATA AS A SERVICE MODEL

The main ideas behind this work are to produce a new proposed model for BDaaS and to open the door for research community to perform newer research work on the BDaaS Model.

However, the proposed BDaaS Model based on 'how to integrate Big Data with Cloud Computing' to produce a new proposed model for BDaaS.

The cloud computing has three main renowned layers which utilized for offering different cloud services starting by Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). To make it more clear, in general a cloud computing consists of storage, network, and processing.

The proposed BDaaS Model shift cloud computing layers to cover the variety outsourcing (like Hadoop, Altiscale and Qubole) for various Big Data functions. These outsourcing arranged based on cloud layers as shown in "fig. 3".

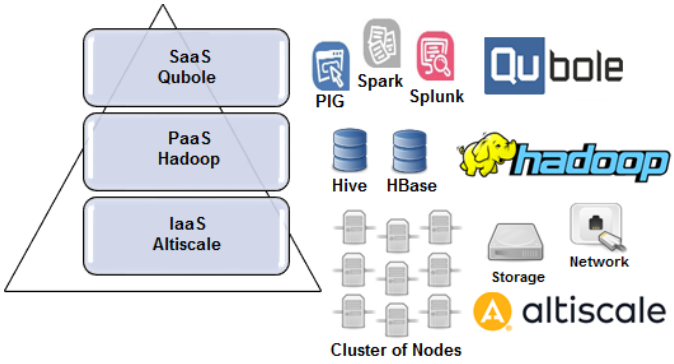

Figure 3. The Proposed BDaaS Model

The proposed BDaaS Model composed from the main three layers as shown below:

1) *Platform as a Service (PaaS):*
'Hadoop Platform' or any other distributed compute and storage technology can used as a PaaS inside BDaaS Model.

It's wrongly to equate BDaaS with Hadoop as a Service.

However, Hadoop stay generic to interact with the rest of the services; for that, it represents a Core Layer in BDaaS Model.

Also, HBase (Hadoop's database) and Hive (Hadoop-based data warehouse) can used as a databases in this layer.

2) *Infrastructure as a Service (IaaS):*
'Altiscale' can used as an IaaS inside BDaaS Model. Altiscale service used to customize infrastructure for run

Hadoop on the Cloud with better performance in order to speed, reliable and easy to use.

Altiscale is used to customize clusters of nodes, storage spaces and network for Hadoop needs.

3) Software as a Service (SaaS):

'Qubole' can used as a SaaS inside BDaaS Model. Qubole service used to manages Hadoop infrastructure and allows users to prepare, int egrate, and analyze Big Data in the cloud.

Also, PIG (A Dataflow Engine for Hadoop), Splunk (Log Analysis) and Spark (Process and Analytics) can used as a service based cloud in this layer.

Table (1) lists the main outsourcing and its purpose which covered by cloud computing layers based the proposed BDaaS.

TABLE I. THE MAIN FUNCTIONS OF THE PROPOSED BDAAS

Cloud Layer	Big Data Outsourcing	Outsourcing Purpose
SaaS	PIG	Hadoop's dataflow engine for creating programs that run on Hadoop.
SaaS	Spark	Big Data's tool for access diverse data sources and process it.
SaaS	Splunk	Big Data's tool for collect, analyze and deliver real-time insights.
SaaS	Qubole	Big Data's tool for prepare, integrate, and analyze big data in the cloud.
PaaS	Hive	Hadoop's data warehouse.
PaaS	HBase	Hadoop's distributed database.
PaaS	Hadoop	Big Data's core for distributed storage and distributed processing of huge datasets across clusters of computers.
IaaS	Altiscale	Big Data's tool for configures the kernel and network parameters to runs Hadoop on hardware.

IX. CONCLUSIONS

The following are some points derived from the proposed model and from the fact that BDaaS is an outgrowth of two essential IT trends, 'Big Data' and 'Cloud Computing':

1) The volume, variety and velocity evolution of Big Data fit naturally with the Cloud Computing paradigm which have ability to scale up when the usage need and simplify delivery of services.

2) The term 'Big Data as a Service' may be unwieldy and somewhat nebulous, but the real fact is the analytical huge data and providing variety reports becomes available "as a Service" just like other cloud services. However, BDaaS allows users to prepare, integrate, and analyze big data in the cloud.

3) The proposed BDaaS model can allows enterprise to implement various Big Data functions (collect huge data, storage, processing, management, analysis, extract meaningful value and more) using variety outsourcing (like Hadoop, Altiscale, Qubole and other Big Data processing software packages) clearly and easily.

4) The proposed BDaaS model allows enterprise to moving them out of the expensive whirlpool of updating and maintaining their infrastructure.

REFERENCES

[1] Thomas Erl, Wajid Khattak, Paul Buhler, "Big Data Fundamentals: Concepts, Drivers & Techniques", Prentice Hall, 2016.

[2] Bernard Marr, "Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results", Wiley, 2016.

[3] Viktor Mayer-Schönberger and Kenneth Cukier, "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Eamon Dolan/Mariner Books, 2014.

[4] Tom Davenport, Jill Dyché, "Big Data in Big Companies Report", SAS, 2013.

[5] Christian Prokopp, "The four types of Big Data as a Service (BDaaS)", Big Data Science and Cloud Computing Blog, 2014-06-23. [Online] Available at http://www.semantikoz.com/blog/big-data-as-a-service-definition-classification

[6] Bernard Marr, "Big Data-As-A-Service Is Next Big Thing", Forbes, APR 2015. [Online] Available at http://www.forbes.com/sites/bernardmarr/2015/04/27/big-data-as-a-service-is-next-big-thing/#6efeb7c93f9a.

[7] Poornima Sharma, Varun Garg, Prof. Randeep Kaur, Prof. Satendra Sonare, "Big Data in Cloud Environment", IntInternational Journal of Computer Sciences and Engineering, Volume-01, Issue-03, pp (15-17), Nov 2013.

[8] Thomas H. Davenport and Jill Dyché, "Big Data in Big Companies", SAS Institute, 2013. [Online] Available at http://www.sas.com/en_us/whitepapers/bigdata-bigcompanies-106461.html .

[9] Doug Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety", META Delta Group, 2001. [Online] Available at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf .

[10] Vishal Kumar Gujare, Pravin Malviya, "A Novel Algorithm for Big Data Clustering", International Journal of Computer Sciences and Engineering, Volume-04, Issue-08, Page No (38-40), Aug -2016, E-ISSN: 2347-2693

[11] "Extracting business value from the 4 V's of big data", [Online] Available at http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data .

[12] "History and evolution of big data analytics", SAS Institute, [Online] Available at http://www.sas.com/en_us/insights/analytics/big-data-analytics.html#dmhistory .

[13] Tom White, "Hadoop: The Definitive Guide", 4th Edition, O'Reilly, 2015.

[14] Hadoop Official Website, http://hadoop.apache.org/ .

[15] Donald Miner, Adam Shook, "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems", 2012.

[16] PIG Official Website, http://pig.apache.org , Last accessed 7th October 2016.

[17] Spark Official Website, http://spark.apache.org , Last accessed 10th October 2016.

[18] Splunk Official Website, http://www.splunk.com , Last accessed 7th October 2016.

[19] Qubole Official Website, http://www.qubole.com , Last accessed 23th October 2016.

[20] Tanuja A, Swetha Ramana D, "Processing and Analyzing Big data using Hadoop", International Journal of Computer Sciences and Engineering, Volume-04, Issue-04, Page No (91-94), Apr -2016, E-ISSN: 2347-2693

[21] HBase Official Website, http://hbase.apache.org , Last accessed 10th October 2016.

[22] P.Kodimalar, "A Study on Big Data and Big Data Analytical Research and Issues", International Journal of Computer Sciences and Engineering, Volume-03, Issue-11, Page No (171-179), Nov -2015, E-ISSN: 2347-2693

[23] Bernard Marr, "Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance", Wiley, 2015.

[24] Mehjabeen Sultana, "An Overview of Emerging Analytics in Big Data: In-Situ", International Journal of Computer Sciences and Engineering, Volume-04, Issue-05, Page No (166-169), May -2016, E-ISSN: 2347-2693.

Author Profile

Mazin S. Al-Hakeem (Associate Professor) received his PhD degree in Computer Science from University of Technology in 2007 (Iraq). He is currently a Head of IT Department in Lebanese French University (Kurdistan Region of Iraq). Founder and previous manager of ICT Center – University of Technology. Previous Head of Coordination and Scientific Marketing Department in Research and Development Directorate (RDD) - Ministry of Higher Education and Research (MoHE)- Iraq. He is editorial and reviewer in many national and international journals. Published several scientific books and many researches in many international conferences and scientific journals. His research interests include network technology, network security, web technology and IoT.

