Big Data and Cloud Computing: Trends and Challenges

https://doi.org/10.3991/ijim.v11i2.6561

Samir A. El-Seoud
British University in Egypt-BUE, Cairo, Egypt

Hosam F. El-Sofany
Cairo Higher Institute for Engineering, Computer Science, and Management, Cairo

Mohamed Abdelfattah
British University in Egypt-BUE, Cairo, Egypt
mohamed.abdelfattah@bue.edu.eg

Reham Mohamed
British University in Egypt-BUE, Cairo, Egypt

Abstract—Big data is currently one of the most critical emerging technologies. Big Data are used as a concept that refers to the inability of traditional data architectures to efficiently handle the new data sets. The 4V's of big data – volume, velocity, variety and veracity makes the data management and analytics challenging for the traditional data warehouses. It is important to think of big data and analytics together. Big data is the term used to describe the recent explosion of different types of data from disparate sources. Analytics is about examining data to derive interesting and relevant trends and patterns, which can be used to inform decisions, optimize processes, and even drive new business models. Cloud computing seems to be a perfect vehicle for hosting big data workloads. However, working on big data in the cloud brings its own challenge of reconciling two contradictory design principles. Cloud computing is based on the concepts of consolidation and resource pooling, but big data systems (such as Hadoop) are built on the shared nothing principle, where each node is independent and self-sufficient. The integrating big data with cloud computing technologies, businesses and education institutes can have a better direction to the future. The capability to store large amounts of data in different forms and process it all at very large speeds will result in data that can guide businesses and education institutes in developing fast. Nevertheless, there is a large concern regarding privacy and security issues when moving to the cloud which is the main causes as to why businesses and educational institutes will not move to the cloud. This paper introduces the characteristics, trends and challenges of big data. In addition to that, it investigates the benefits and the risks that may rise out of the integration between big data and cloud computing.

Keywords—Big data, Cloud computing, Analytics.

1 Introduction

Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis. Data sources extend beyond the traditional corporate database to include email, mobile device output, sensor-generated data, and social media output [1]. Data are no longer restricted to structured database records but include unstructured data – data having no standard formatting [2].

Big Data requires huge amounts of storage space. While the price of storage continued to decline, the resources needed to leverage big data can still pose financial difficulties for small to medium sized businesses. A typical big data storage and analysis infrastructure will be based on clustered network-attached storage (NAS). Clustered NAS infrastructure requires configuration of several NAS "pods" with each NAS "pod" comprised of several storage devices connected to an NAS device. The series of NAS devices are then interconnected to allow massive sharing and searching of data [3].

Cloud computing is an extremely successful paradigm of service oriented computing, and has revolutionized the way computing infrastructure is abstracted and used. Three most popular cloud paradigms include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The concept however can also be extended to Database as a Service or Storage as a Service. Elasticity, pay-per-use, low upfront investment, low time to market, and transfer of risks are some of the major enabling features that make cloud computing a universal paradigm for deploying novel applications which were not economically feasible in a traditional enterprise infrastructure settings. Scalable database management systems, both for update intensive application workloads, as well as decision support systems—are thus a critical part of the cloud infrastructure. Scalable and distributed data management has been the vision of the database research community for more than three decades. Much research has focused on designing scalable systems for both update intensive workloads as well as ad-hoc analysis workloads. Initial designs include distributed databases for update intensive workloads [5], and parallel database systems for analytical workloads [6]. Parallel databases grew beyond prototype systems to large commercial systems, but distributed database systems were not very successful and were never commercialized – rather various ad-hoc approaches to scaling were used.

Changes in the data access patterns of applications and the need to scale out to thousands of commodity machines led to the birth of a new class of systems referred to as Key-Value stores which are now being widely adopted by various enterprises. In the domain of data analysis, the MapReduce paradigm [7] and its open-source implementation Hadoop [9] has also seen widespread adoption in industry and academia alike. Solutions have also been proposed to improve Hadoop based systems in terms of usability and performance [10].

On the other hand, data warehouses have been used to manage the large amount of data. The warehouses and solutions built around them are unable to provide reasonable response times in handling expanding data volumes. One can either perform analytics on big volume once in days or one can perform transactions on small amounts

of data in seconds. With the new requirements, one needs to ensure the real-time or near real time response for huge amount of data. The 4V's of big data – volume, velocity, variety and veracity—makes the data management and analytics challenging for the traditional data warehouses. Big data can be defined as data that exceeds the processing capacity of conventional database systems. It implies that the data count is too large, and/or data values change too fast, and/or it does not follow the rules of conventional database management systems (e.g., consistency). One requires new expertise in the areas of data management and systems management who understands how to model the data and prepare them for analysis, and understand the problem deeply enough to perform the analytics. As data is massive and/or fast changing we need comparatively many more CPU and memory resources, which are provided by distributed processors and storage in cloud settings [11].

Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques. Cloud computing is on-demand network access to computing resources which are often provided by an outside entity and require little management effort by the business. A number of architectures and deployment models exist for cloud computing, and these architectures and models are able to be used with other technologies and design approaches. Owners of small to medium sized businesses who are unable to afford adoption of clustered NAS technology can consider a number of cloud computing models to meet their big data needs. Small to medium sized business owners need to consider the correct cloud computing in order to remain both competitive and profitable [4

The paper is organized as follows: in section two, we present the literature review done in the field. In section three, we present a general concepts and definition of Big Data. In section four, we present the movement of data management to cloud computing. In section five we introduce the management of Big Data in cloud computing environments. In section six, we introduce some solutions and methodologies for data storage. In section seven, we present the "Map Reduce" and "Hadoop" a free programming framework supports the processing of large sets of data in a distributed computing environment. In section eight, we introduce the advantages of Big Data applications. In section nine, we introduce the advantages and risks and challenges to integrating big data with cloud computing. The paper finally concluded in section ten.

2 Literature Review

Big Data and Cloud computing are a major trend that are rapidly growing and new challenges and solutions are being published every day.

In 2014, [35] a publication was made to define the best methods to deploy big data analytics applications to the cloud. The series of steps are defined in 6 steps: the first stage, we develop the business use case by focusing on how the deep business values will be achieved by moving to the cloud and identify the drivers to achieve this. Following this is aligning the stakeholder's requirements with the case in order to achieve their support. Finally the case needs to be feasible by identifying the key advantages that out weight other solutions available; the second stage, is to access you

application workload. Depending on the functional requirements set and business case, the cloud service should have to the ability to support the workload with the ability to rapidly optimize as the new workloads come online; The third stage, is to develop a technical approach to the big data platform by focusing on the both the topologies and data platforms; The fourth stage is to address governance, security, privacy, risk and accountability requirements. The big data platform should adapt tools to maintain security of the data and architecture whilst operating under the policies provided by the organization; finally is to deploy, integrate, and operationalize the infrastructure. Here the cloud platform has been set up and the data has been provided and integrated. The platform is then deployed and the organization starts taking advantage of it by going about the normal day to day tasks.

Another paper in 2014, [33] was published on the integration of big data and cloud computing technologies. The integration of big data and cloud computing has long term benefits of both insights and performance. Due to the large amounts of data collected, they need to analyzed otherwise the data is useless; hence the cloud services can handle these extensive amounts of data with rapid response times and real time processing of the data. There are currently a few integrated cloud environments for big data analytics; Canpaas is an environment that was developed by Vrije Universiteit Amsterdam, the University of Rennes 1, Zuse-Institut Berlin and XLAB to simplify the process of creating scalable cloud applications without the need to put into consideration the complexity of these applications. The environment also provides a handful of resources for hosting web applications written in PHP or java, as well as, managing different database management systems including SQL or NoSQL. Another environment includes MapReduce provided by apache to aid in parallel computing by enabling distributed file storage systems and processing power. Task Framing is another environment used mainly for batch processing, which allows the execution of large number of non-related tasks simultaneously. The benefits of cloud computing include parallel computing, scalability, elasticity, and being inexpensive. In terms of security and privacy, data is encrypted using advanced encryption techniques to secure data but this causes high processing overheads hence other solutions are available. In regards to the performance, some challenges do arise which include data transfer limitations, data retention, isolation management, and disaster recovery.

In 2015, [34] shared a paper on the efficiency of big data and cloud computing and why both technologies complement each other. Big data and cloud computing complement each other and are both the fastest growing technologies emerging today. The cloud seems to provide large computing power by aggregating resources together and offering a single system view to manage these resources and applications, so why big data should be placed on the cloud? The following reasons provide a sufficient answer to the question in hand: cost reduction, where organizations can make use of the pay per use model instead of doing a major investment to setup servers and clusters to manage the big data that can become obsolete and require upgrades; reduce overheads, by acquiring any new components needed automatically; rapid provisioning/time to market, where organizations can easily change the scale of the environments easily based on the processing requirements; flexibility/scalability, environments can be set up at any time at any scale in just a few minutes.

3 Big Data: General Concept and Definition

Big Data are used as a concept that refers to the inability of traditional data architectures to efficiently handle the new data sets. Characteristics that force a new architecture to achieve efficiencies are the data set-at-rest characteristics ***volume***, and ***variety*** of data from multiple domains or types; and from the data-in-motion characteristics of ***velocity***, or rate of flow, and ***variability*** (principally referring to a change in velocity). Each of these characteristics results in different architectures or different data lifecycle process orderings to achieve needed efficiencies. A number of other terms (often starting with the letter 'V') are also used, but a number of these refer to the analytics and not big data architectures, as shown in Figure 1, [12].

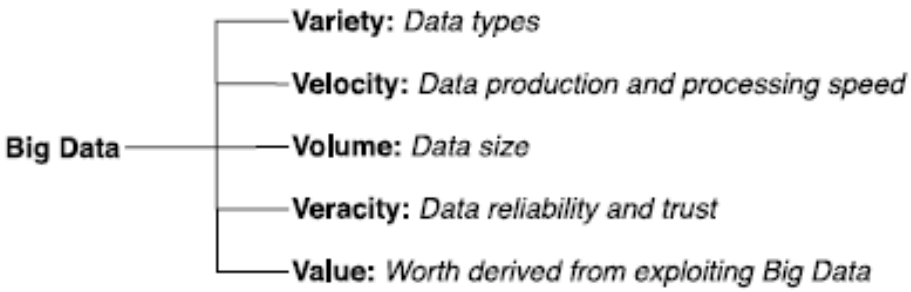

Fig. 1. Some Vs'of Big Data.

The new big data paradigm occurs when the scale of the data at rest or in motion forces the management of the data to be a significant driver in the design of the system architecture. Fundamentally the big data paradigm represents a shift in data system architectures from monolithic systems with vertical scaling (faster processors or disks) into a horizontally scaled system that integrates a loosely coupled set of resources. This shift occurred 20-some years ago in the simulation community when the scientific simulations began using massively parallel processing (MPP) systems. In different combinations of splitting the code and data across independent processors, computational scientists were able to greatly extend their simulation capabilities. This of course introduced a number of complications in such areas as message passing, data movement, and latency in the consistency across resources, load balancing, and system inefficiencies while waiting on other resources to complete their tasks. In the same way, the big data paradigm represents this same shift, again using different mechanisms to distribute code and data across loosely-coupled resources in order to provide the scaling in data handling that is needed to match the scaling in the data.

The purpose of storing and retrieving large amounts of data is to perform analysis that produces additional knowledge about the data. In the past, the analysis was generally accomplished on a random sample of the data.

3.1 Big Data Engineering

Big Data Engineering is the storage and data manipulation technologies that leverage a collection of horizontally coupled resources to achieve a nearly linear scalability in performance.

New engineering techniques in the data layer have been driven by the growing prominence of data types that cannot be handled efficiently in a traditional relational model. The need for scalable access in structured and unstructured data has led to software built on name-value/key-value pairs or columnar (big table), document-oriented, and graph (including triple-store) paradigms.

3.2 Non-Relational Model

Non-Relational Models refers to logical data models such as document, graph, key value and others that are used to provide more efficient storage and access to non-tabular data sets.

3.3 NoSQL

NoSQL (alternately called “no SQL” or “not only SQL”) refers to data stores and interfaces that are not tied to strict relational approaches.

3.4 Big Data Models

Big Data Models refers to logical data models (relational and non-relational) and processing/computation models (batch, streaming, and transactional) for the storage and manipulation of data across horizontally scaled resources.

3.5 Schema-on-read

Schema-on-read big data are often stored in a raw form based on its production, with the schema, needed for organizing (and often cleansing) the data, is discovered and transformed as the data are queried. This is critical since in order for many analytics to run efficiently the data must be structured to support the specific algorithms or processing frameworks involved.

3.6 Big Data Analytics

Big Data Analytics is rapidly evolving both in terms of functionality and the underlying programming model. Such analytical functions support the integration of results derived in parallel across distributed pieces of one or more data sources.

3.7 Big Data Paradigm

The big data paradigm consists of the distribution of data systems across horizontally-coupled independent resources to achieve the scalability needed for the efficient processing of extensive data sets.

With the new Big Data Paradigm, analytical functions can be executed against the entire data set or even in real-time on a continuous stream of data. Analysis may even integrate multiple data sources from different organizations. For example, consider the question "What is the correlation between insect borne diseases, temperature, precipitation, and changes in foliage". To answer this question an analysis would need to integrate data about incidence and location of diseases, weather data, and aerial photography.

While we certainly expect a continued evolution in the methods to achieve efficient scalability across resources, this paradigm shift (in analogy to the prior shift in the simulation community) is a one-time occurrence; at least until a new paradigm shift occurs beyond this "crowdsourcing" of processing or data system across multiple horizontally-coupled resources.

The Big Data paradigm has other implications from these technical innovations. The changes are not only in the logical data storage, but in the parallel distribution of data and code in the physical file system and direct queries against this storage.

The shift in thinking causes changes in the traditional data lifecycle. One description of the end-to-end data lifecycle categorizes the steps as collection, preparation, analysis and action. Different big data use cases can be characterized in terms of the data set characteristics at-rest or in-motion, and in terms of the time window for the end-to-end data lifecycle. Data set characteristics change the data lifecycle processes in different ways, for example in the point of a lifecycle at which the data are placed in persistent storage. In a traditional relational model, the data are stored after preparation (for example after the extract-transform-load and cleansing processes). In a high velocity use case, the data are prepared and analysed for alerting, and only then is the data (or aggregates of the data) given a persistent storage. In a volume use case the data are often stored in the raw state in which it was produced, prior to the application of the preparation processes to cleanse and organize the data. The consequence of persistence of data in its raw state is that a schema or model for the data are only applied when the data are retrieved, known as schema on read.

A third consequence of big data engineering is often referred to as "*moving the processing to the data, not the data to the processing*". The implication is that the data are too extensive to be queried and transmitted into another resource for analysis, so the analysis program is instead distributed to the data-holding resources; with only the results being aggregated on a different resource. Since I/O bandwidth is frequently the limited resource in moving data, another approach would be to embed query/filter programs within the physical storage medium.

At its heart, Big Data refers to the extension of data repositories and processing across horizontally-scaled resources, much in the same way the compute-intensive simulation community embraced massively parallel processing two decades ago. In the past, classic parallel computing applications utilized a rich set of communi-

cations and synchronization constructs and created diverse communications topologies. In contrast, today, with data sets growing into the Petabyte and Exabyte scales, distributed processing frameworks offering patterns such as map-reduce, offer a reliable high- level, commercially viable compute model based on commodity computing resources, dynamic resource scheduling, and synchronization techniques [12].

Definition: Big Data is a data set(s) with characteristics (e.g. volume, velocity, variety, variability, veracity, etc.) that for a particular problem domain at a given point in time cannot be efficiently processed using current/existing/established/traditional technologies and techniques in order to extract value.

The above definition distinguishes Big Data from business intelligence (BI) and traditional transactional processing while alluding to a broad spectrum of applications that includes them. The ultimate goal of processing Big Data is to derive differentiated value that can be trusted (because the underlying data can be trusted). This is done through the application of advanced analytics against the complete corpus of data regardless of scale. Parsing this goal helps frame the value discussion for Big-Data use cases.

1. Any scale of operations and data: Utilizing the entire corpus of relevant information, rather than just samples or subsets. It's also about unifying all decision-support time-horizons (past, present, and future) through statistically derived insights into deep data sets in all those dimensions.
2. Trustworthy data: Deriving valid insights either from a single-version-of-truth consolidation and cleansing of deep data, or from statistical models that sift haystacks of "dirty" data to find the needles of valid insight.
3. Advanced analytics: Faster insights through a variety of analytic and mining techniques from data patterns, such as "long tail" analyses, micro-segmentations, and others, that are not feasible if you're constrained to smaller volumes, slower velocities, narrower varieties, and undetermined veracities.

A difficult question is what makes "Big Data" big, or how large does a data set have to be in order to be called big data? The answer is an unsatisfying "it depends". Part of this issue is that "Big" is a relative term and with the growing density of computational and storage capabilities (e.g. more power in smaller more efficient form factors) what is considered big today will likely not be considered big tomorrow. Data are considered "big" if the use of the new scalable architectures provides improved business efficiency over other traditional architectures. In other words the functionality cannot be achieved in something like a traditional relational database platform.

Big data essentially focuses on the self-referencing viewpoint that data are big because it requires scalable systems to handle it, and architectures with better scaling have come about because of the need to handle big data [12].

4 Moving Data Management to Cloud

A data management system has various stages of data lifecycle such as data ingestion, extract-transform-load (ETL), data processing, data archival, and deletion. Before moving one or more stages of data lifecycle to the cloud, one has to consider the following factors:

1. Availability Guarantees: Each cloud computing provider can ensure a certain amount of availability guarantees. Transactional data processing requires quick real-time answers whereas for data warehouses long running queries are used to generate reports. Hence, one may not want to put its transactional data over cloud but may be ready to put the analytics infrastructure over the cloud.
2. Reliability of Cloud Services: Before offloading data management to cloud, enterprises want to ensure that the cloud provides required level of reliability for the data services. By creating multiple copies of application components the cloud can deliver the service with the required reliability of service.
3. Security: Data that is bound by strict privacy regulations, such as medical information covered by the Health Insurance Portability and Accountability Act (HIPAA), will require that users log in to be routed to their secure database server.
4. Maintainability: Database administration is a highly skilled activity which involves deciding how data should be organized, which indices and views should be maintained, etc. One needs to carefully evaluate whether all these maintenance operations can be performed over the cloud data.

Cloud has given enterprises the opportunity to fundamentally shift the way data is created, processed and shared. This approach has been shown to be superior in sustaining the performance and growth requirements of analytical applications and, combined with cloud computing, offers significant advantages [13].

5 Manage Big Data in Cloud Computing Environments

Cloud Computing is an environment based on using and providing services. There are different categories in which the service-oriented systems can be clustered. One of the most used criteria to group these systems is the abstraction level that is offered to the system user. In this way, three different levels are often distinguished: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) as shown in Figure 2, [14].

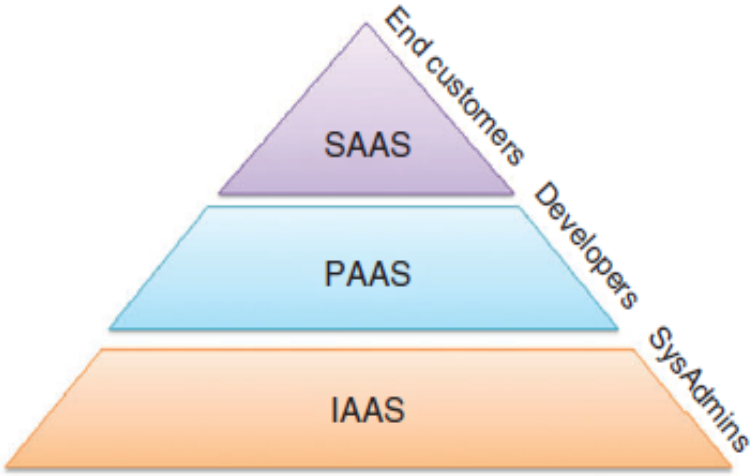

Fig. 2. Illustration of the layers for the Service-Oriented Computing

Cloud Computing offers scalability with respect to the use of resources, low administration effort, flexibility in the pricing model and mobility for the software user. Under these assumptions, it is obvious that the Cloud Computing paradigm benefits large projects, such as the ones related with Big Data and BI [15].

In particular, a common Big Data *analytics framework* is depicted in Figure 3. Focusing on the structure of the data management sector we may define, as the most suitable management organization architecture, one based on a four-layer architecture, which includes the following components [16]:

1. A file system for the storage of Big Data, i.e., a wide amount of archives of large size. This layer is implemented within the IaaS level as it defines the basic architecture organization for the remaining tiers.
2. A DBMS for organizing the data and access them in an efficient way. It can be viewed in between the IaaS and PaaS as it shares common characteristics from both schemes. Developers used it to access the data, but its implementation lies on a hardware level. Indeed, a PaaS acts as an interface where, at the upper side offers its functionality, and at the bottom side, it has the implementation for a particular IaaS. This feature allows applications to be deployed on different IaaS without rewriting them.
3. An execution tool to distribute the computational load among the computers of the cloud. This layer is clearly related with PaaS, as it is kind of a 'software API' for the codification of the Big Data and BI applications.
4. A query system for the knowledge and information extraction required by the system's users, which is in between the PaaS and SaaS layers.

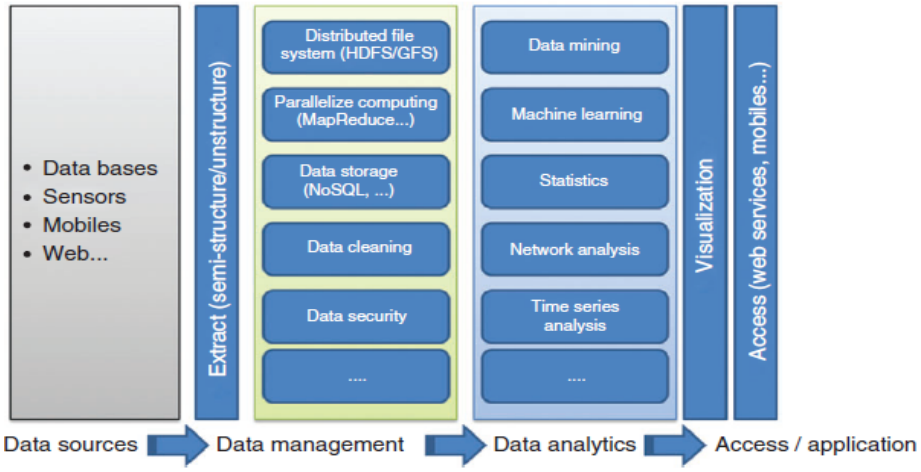

Fig. 3. Big Data framework [14].

6 Solutions and Methodologies for Data Storage

Several solutions were proposed to store and retrieve large amounts of data demanded by Big Data, some of which are currently used in Clouds. Internet-scale file systems such as the Google File System (GFS) attempt to provide the robustness, scalability, and reliability that certain Internet services need [17]. Other solutions provide object-store capabilities where files can be replicated across multiple geographical sites to improve redundancy, scalability, and data availability. Examples include Amazon Simple Storage Service (S3), Nirvanix Cloud Storage, OpenStack Swift and Windows Azure Binary Large Object (Blob) storage. Although these solutions provide the scalability and redundancy that many Cloud applications require, they sometimes do not meet the con- currency and performance needs of certain analytics applications.

One key aspect in providing performance for Big Data analytics applications is the data locality. This is because the volume of data involved in the analytics makes it prohibitive to transfer the data to process it. This was the preferred option in typical high performance computing systems: in such systems, that typically concern performing CPU-intensive calculations over a moderate to medium volume of data, it is feasible to transfer data to the computing units, because the ratio of data transfer to processing time is small. Nevertheless, in the context of Big Data, this approach of moving data to computation nodes would generate large ratio of data transfer time to processing time. Thus, a different approach is preferred, where computation is moved to where the data is. The same approach of exploring data locality was explored previously in scientific workflows [18] and in Data Grids [19].

In the context of Big Data analytics, MapReduce presents an interesting model where data locality is explored to improve the performance of applications. Hadoop, an open source MapReduce implementation, allows for the creation of clusters that

use the Hadoop Distributed File System (HDFS) to partition and replicate data sets to nodes where they are more likely to be consumed by mappers. In addition to exploiting concurrency of large numbers of nodes, HDFS minimizes the impact of failures by replicating data sets to a configurable number of nodes. It has been used by [20] to develop an analytics platform to process Face- book's large data sets. The platform uses Scribe to aggregate logs from Web servers and then exports them to HDFS files and uses a Hive–Hadoop cluster to execute analytics jobs. The platform includes replication and compression techniques and columnar compression of Hive7 to store large amounts of data.

Although a large part of the data produced nowadays is un- structured, relational databases have been the choice most organizations have made to store data about their customers, sales, and products, among other things. As data managed by traditional DBMS ages, it is moved to data warehouses for analysis and for sporadic retrieval. Models such as MapReduce are generally not the most appropriate to analyze such relational data. Attempts have been made to provide hybrid solutions that incorporate MapReduce to perform some of the queries and data processing required by DBMS's [21].

In [22] the researchers provide a parallel database design for analytics that supports SQL and MapReduce scripting on top of a DBMS to integrate multiple data sources. A few providers of analytics and data mining solutions, by exploring models such as MapReduce, are migrating some of the processing tasks closer to where the data is stored, thus trying to minimize surpluses of data preparation, storage, and processing. Data processing and analytics capabilities are moving towards Enterprise Data Warehouses (EDWs), or are being deployed in data hubs to facilitate reuse across various data sets [23].

Another distinctive trend in Cloud computing is the increasing use of NoSQL databases as the preferred method for storing and retrieving information. NoSQL adopts a non-relational model for data storage. Leavitt argues that non-relational models have been avail- able for more than 50 years in forms such as object-oriented, hierarchical, and graph databases, but recently this paradigm started to attract more attention with models such as key-store, column- oriented, and document-based stores. The causes for such raise in interest, according to Levitt, are better performance, capacity of handling unstructured data, and suitability for distributed environments [24].

In [25] the researchers presented a survey of NoSQL databases with emphasis on their advantages and limitations for Cloud computing. The survey classifies NoSQL systems according to their capacity in addressing different pairs of CAP (consistency, availability, partitioning). The survey also explores the data model that the studied NoSQL systems support.

7 Data Processing and Resource Management

When making an attempt to understand the concept of Big Data, the words such as "Map Reduce" and "Hadoop" cannot be avoided.

7.1 Hadoop

Hadoop, is a free Java-based programming framework supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, and flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (HDFS) [26].

7.2 Map Reduce

Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework [27].

7.3 Hadoop Distributed File System (HDFS)

HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures [28].

8 Advantages of Big Data Applications

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which triggers the development of big data applications. Google's map reduce framework and apache Hadoop are the defacto software systems for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and *Bioinformatics* are the two major areas of big data applications.

Big data provide an infrastructure for transparency in *manufacturing industry*, which has the ability to unravel uncertainties such as inconsistent component performance and availability. In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data. The combination of sensory data and historical data constructs the big data in manufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies such as prognostics and health management [29] and [30].

Another important application for Hadoop is *Bioinformatics* which covers the next generation sequencing and other biological domains. Bioinformatics which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces [31].

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyse the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements.

There are some common characteristics of big data:

1. Big data integrates both structured and unstructured data.
2. Addresses speed and scalability, mobility and security, flexibility and stability.
3. In big data the realization time to information is critical to extract value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies.

All the organizations and business would benefit from *speed*, *capacity*, and *scalability* of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities. Another notable advantage with big-data is, data *analytics*, which allow the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas:

1. Calculate the risks on large portfolios
2. Detect, prevent, and re-audit financial fraud
3. Improve delinquent collections
4. Execute high value marketing campaigns

9 Advantages and Risks and Challenges of Big Data and Cloud computing integration

9.1 Advantages

There are many advantages of integrating cloud computing to big date. Big data raises the need for multiple servers because of the massive data and size it works on, and it demands high velocity and variability. These multiple servers work in parallel to provide the big data high requirements. Cloud computing already uses multiple servers and allow resource allocations. Because of that, it is a great fit to build the big data on these cloud multi-servers and make use of the resource allocation availability provided by the cloud environments which would result in a better efficiency for big data analysis [34]. Using cloud system as a storage for big data would improve the performance for both. As cloud systems are mainly based on remote multi-servers which makes it feasible to handle massive amounts of data simultaneously. This feature allows advanced analytics techniques to enable the big data to deal with massive amounts of data. The integration between cloud computing and big data would result in cost reduction. While big data requires clusters of servers and volumes to support the massive amount of data. Cloud computing systems can work as the structure for all of these servers and volumes instead of creating new ones for big data which also provides more flexibility and scalability and eliminates the huge investments that would be invested on the big data computers and servers [33].

In addition to that, using cloud computing provides faster provisioning to big data as provisioning servers in the cloud is so easy and feasible. So, based on the processing requirements of the big data, the cloud environment used can be scaled accordingly. This fast provisioning is really important for big data as the value of the data rapidly decrease over time. In general, cloud computing complements big data and provides convenient, on-demand and shared computing environment with minimal management effort and reduced overhead. It also makes the environment more robust, automated and provides multi-tenancy. In addition to all of that, the integration between both, makes big data resources more controllable, monitored and reported. Moreover, this integration enables reducing the complexity and improving the productivity. All of these advantages make the cloud based approaches are the right models for deploying big data.

9.2 Risk and Challenges

Despite all of these advantages of the integration between cloud computing and big data, there are some challenges and risks that should be considered while deploying big data on a cloud environment. The fundamental issue that should be considered is the security of the big data cloud environment. There are some security vulnerabilities that rise because of this integration between both and creating a new unfamiliar platform. One of the most known Big Data cloud security vulnerability is platform heterogeneity. There are many Big Data deployments that requires deploying a new plat-

form in the cloud while the existing security tools and practices of the cloud will not work for such platforms and new security tools are needed to be developed to work with these new Big Data platforms. These security tools could include authentication, access control, encryption, intrusion detection, and event logging and monitoring. In addition to the security policies, the Big Data consolidation plans should be taken into consideration while the integration with the cloud environment.

Another challenge is the nature of data and its location, as in Big Data, the data can be in different locations. The cloud environment may include these locations or not. The type of processing that should be applied to the data, the parallelism of the processing, and where the processing should take place either the data is moved to a processing environment or the processing is performed on the location of the data. All of these are also challenges that should be taken into consideration while deploying the Big Data to a cloud system environment. In addition to that, another challenge is the optimization of the Big Data cloud topology as it specifies the configuration, the size of the clouds, clusters and nodes that should be included to reach the optimal Big Data cloud model. [32].

On the other hand these challenges have been overcome, if not in a direct manner, to allow the idea of the cloud and big data integration become the more practical answer instead of investing countless thousands of dollars in creating an environment suitable to operate the large amount of processing required as well as accommodate terabytes of data.

10 Conclusions

Big data is currently one of the most critical emerging technologies. The 4V's of big data – volume, velocity, variety and veracity makes the data management and analytics challenging for the traditional data warehouses. Cloud computing seems to be a perfect vehicle for hosting big data workloads. However, working on big data in the cloud brings its own challenge of reconciling two contradictory design principles. The integrating big data with cloud computing technologies, businesses and education institutes can have a better direction to the future. The capability to store large amounts of data in different forms and process it all at very large speeds will result in data that can guide businesses and education institutes in developing fast. The paper presented the general concepts and definition of Big Data, and illustrated the movement of data management to cloud computing. The paper presented the "Map Reduce" and "Hadoop" as Big Data systems that support the processing of large sets of data in a distributed computing environment. This paper also introduced the characteristics, trends and challenges of big data. In addition to that, it investigates the benefits and the risks that may rise out of the integration between big data and cloud computing. The major advantage with the cloud computing and big data integration is the data storage and processing power availability, the cloud has access to a large pool of resources and various forms of infrastructures that can accommodate this integration in the best suitable way possible; with minimum effort the environment can be set up and managed to allow an excellent work space for all the big data needs i.e.

data analytics. This in turn provides low complexity with high productivity. A couple of challenges cross the path of this integration, but nothing too major that with today's technology and expansion in the field has not already overcome them and give the cloud a much need edge in being the most practical solution to host and process big data environments.

11 References

[1] Villars, R. L., Olofson, C. W., & Eastwood, M (June, 2011). Big data: What it is and why you should care. IDC White Paper. Framingham, MA: IDC.

[2] Coronel, C., Morris, S., & Rob, P. (2013). Database Systems: Design, Implementation, and Management, (10th. Ed.). Boston: Cengage Learning.

[3] White, C. (2011). Data Communications and Computer Networks: A business user's approach, (6th ed.). Boston: Cengage Learning.

[4] IOS Press. (2011). Guidelines on security and privacy in public cloud computing. Journal of EGovernance,34 149-151. DOI: 10.3233/GOV-2011-0271

[5] J. B. Rothnie Jr., P. A. Bernstein, S. Fox, N. Goodman, M. Hammer, T. A. Landers, C. L. Reeve, D. W. Shipman, and E. Wong. Introduction to a System for Distributed Databases (SDD-1). ACM Trans. Database Syst., 5(1):1–17, 1980. https://doi.org/10.1145/320128.320129

[6] J. Dewitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H. I. Hsiao, and R. Rasmussen. The Gamma Database Machine Project. IEEE Trans. on Knowl. and Data Eng., 2(1):44–62, 1990. https://doi.org/10.1109/69.50905

[7] Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. In OSDI, pages205–218, 2006.

[8] Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In OSDI, pages 137–150, 2004.

[9] Apache Hadoop Project. http://hadoop.apache.org/core/, 2009

[10] Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. PVLDB, 2(2):1626–1629, 2009

[11] Rajeev Gupta, Himanshu Gupta, and Mukesh Mohania, "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?". S. Srinivasa and V. Bhatnagar (Eds.): BDA 2012, LNCS 7678, pp. Springer-Verlag Berlin Heidelberg 42–61, 2012.

[12] ISO/IEC JTC 1. Information technology Big data, Preliminary Report 2014. ISO/IEC 2015.

[13] Curino, C., Jones, E.P.C., Popa, R.A., Malviya, N., Wu, E., Madden, S., Balakrishnan, H., Zeldovich, N.: Realtional Cloud: A Database-as-a-Service for the Cloud. In: Proceedings of Conference on Innovative Data Systems Research, CIDR- 2011.

[14] Alberto Ferandez, Sara del R, Victoria opez, Abdullah Bawakid, Maria J. del Jesus, Jose M. Benitez, and Francisco Herrera. "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks". doi: 10.1002/widm.1134. WIREs Data Mining Knowl Discov, 4:380–409, 2014.

[15] Shim K, Cha SK, Chen L, Han W-S, Srivastava D, Tanaka K, Yu H, Zhou X. Data management challenges and opportunities in cloud computing. In: 17th International Conference on Database Systems for Advanced Applications (DASFAA'2012). Berlin/Heidelberg: Springer 323; 2012. https://doi.org/10.1007/978-3-642-29035-0_30

[16] Kambatla K, Kollias G, Kumar V, Grama A. Trends in big data analytics. J Parallel Distrib, 74:2561–2573, 2014. https://doi.org/10.1016/j.jpdc.2014.01.003
[17] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: Proceedings of the 9th ACM Symposium on Operating Systems Principles (SOSP 2003), ACM, New York, USA, pp. 29–43, 2003. https://doi.org/10.1145/945445.945450
[18] Deelman, A. Chervenak, Data management challenges of data-intensive scientific workflows, in: Proceedings of the 8th IEEE International Sympo- sium on Cluster Computing and the Grid (CCGrid'08), IEEE Computer Society, pp. 687–692, 2008.
[19] Venugopal, R. Buyya, K. Ramamohanarao, A taxonomy of data grids for distributed data sharing, management and processing, ACM Comput. Surv. 38 (1) 1–53, 2006. https://doi.org/10.1145/1132952.1132955
[20] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J.S. Sarma, R. Murthy, H. Liu, Data warehousing and analytics infrastructure at Facebook, in: Proceedings of the 2010 International Conference on Management of Data, ACM, New York, NY, USA, pp. 1013–1020. 2010. https://doi.org/10.1145/1807167.1807278
[21] D.J. Abadi, Data management in the cloud: Limitations and opportunities, IEEE Data Engineering Bulletin 32 (1) 3–12, 2009.
[22] J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, C. Welton, MAD skills: new analysis practices for big data, Proceedings of the VLDB Endow 2 (2) 1481–1492, 2009. https://doi.org/10.14778/1687553.1687576
[23] D. Jensen, K. Konkel, A. Mohindra, F. Naccarati, E. Sam, Business Analytics in the Cloud, White paper IBW03004-USEN-00, IBM (April 2012).
[24] N. Leavitt, Will NoSQL Databases Live Up to Their Promise? Computer 43 (2) 12–14, (2010). https://doi.org/10.1109/MC.2010.58
[25] J. Han, H. E, G. Le, J. Du, Survey on NoSQL database, in: 6th International Conference on Pervasive Computing and Applications (ICPCA 2011), IEEE, Port Elizabeth, South Africa, pp. 363–366., 2011.
[26] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.
[27] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.
[28] K, Chitharanjan, and Kala Karun A. "A review on hadoop - HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
[29] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, p. 32 – 37, 2013.
[30] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr 14-19, p. 35 – 39, 2013,. https://doi.org/10.1109/infcom.2013.6566730
[31] Xu-bin, LI, JIANG Wen-rui, JIANG Yi, ZOU Quan. "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh, Beijing, Jun 19-20, p. 48 – 52, 2012.
[32] Castelino, C., Gandhi, D., Narula, H. G., & Chokshi, N. H. (2014). Integration of Big Data and Cloud Computing. International Journal of Engineering Trends and Technology (IJETT), 100-102.
[33] Chandrashekar, R., Kala, M., & Mane, D. (2015). Integration of Big Data in Cloud computing environments for enhanced data processing capabilities. International Journal of Engineering Research and General Science, 240-245.
[34] James Kobielus, I., & Bob Marcus, E. S. (2014). Deploying Big Data Analytics Applications to the Cloud: Roadmap for Success. Cloud Standards Customer Council.

12 Authors

Samir A. EL-SEOUD is with British University in Egypt-BUE, Cairo, Egypt.

Hosam F. El-SOFANY is with Cairo Higher Institute for Engineering, Computer Science, and Management, Cairo, Egypt.

Mohamed ABDELFATTAH (corresponding author) is with British University in Egypt-BUE, Cairo, Egypt (mohamed.abdelfattah@bue.edu.eg).

Reham MOHAMED is with British University in Egypt-BUE, Cairo, Egypt.

This article is a revised version of a paper presented at the BUE International Conference on Sustainable Vital Technologies in Engineering and Informatics, held Nov 07, 2016 - Nov 09, 2016 , in Cairo, Egypt. Article submitted 21 December 2016. Published as resubmitted by the authors 23 February 2017.