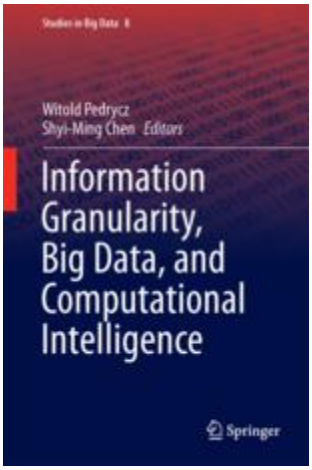

Book Title:

Information Granularity, Big Data, and Computational Intelligence, W. Pedrycz and S.-M. Chen (eds.), Vol. 8, 2015

Chapter 13: **The Role of Cloud Computing Architecture in Big Data**

Authors: Mehdi Bahrami and Mukesh Singhal

Cloud Lab, University of California, Merced

http://Cloudlab.UCMerced.edu

Publisher: Springer International Publishing Switzerland 2015

DOI: 10.1007/978-3-319-08254-7_13

URL:

http://goo.gl/0LxxlH

Or

http://link.springer.com/chapter/10.1007%2F978-3-319-08254-7_13#

Citation:

Mehdi Bahrami and Mukesh Singhal, "The Role of Cloud Computing Architecture in Big Data", Information Granularity, Big Data, and Computational Intelligence, Vol. 8, pp. 275-295, Chapter 13, Pedrycz and S.-M. Chen (eds.), Springer, 2015 http://goo.gl/0LxxlH

The Role of Cloud Computing Architecture in Big Data

Mehdi Bahrami[1] and Mukesh Singhal

Abstract In this data-driven society, we are collecting a massive amount of data from people, actions, sensors, algorithms and the web; handling "Big Data" has become a major challenge. A question still exists regarding when data may be called big data. How large is big data? What is the correlation between big data and business intelligence? What is the optimal solution for storing, editing, retrieving, analyzing, maintaining, and recovering big data? How can cloud computing help in handling big data issues? What is the role of a cloud architecture in handling big data? How important is big data in business intelligence? This chapter attempts to answer these questions. First, we review a definition of big data. Second, we describe the important challenges of storing, analyzing, maintaining, recovering and retrieving a big data. Third, we address the role of Cloud Computing Architecture as a solution for these important issues that deal with big data. We also discuss the definition and major features of cloud computing systems. Then we explain how cloud computing can provide a solution for big data with cloud services and open-source cloud software tools for handling big data issues. Finally, we explain the role of cloud architecture in big data, the role of major cloud service layers in big data, and the role of cloud computing systems in handling big data in business intelligence models.

Keywords Big Data · Cloud Computing · Cloud Architecture · Business Intelligence

1 Introduction

Capturing data from different sources allows a business to use Business Intelligence (BI) [1] capabilities. These sources could be consumer information, service information, products, advertising logs, and related information such as the history of product sales or customer transactions. When an organization uses BI

[1] Mehdi Bahrami . Mukesh Singhal,
Cloud Lab, Electrical Engineering and Computer Science Department, University of California, Merced, USA
IEEE Senior Member, email: MBahrami@UCMerced.edu

Mukesh Singhal
Chancellor's Professor, Email: MSinghal@UCMerced.edu

technology to improve services, we characterize it as a "smart organization*"* [1]. The smart features of these organizations have different levels which depend on the accuracy of decisions; greater accuracy of data analysis provides "smarter" organizations.

For this reason, we are collecting a massive amount of data from people, actions, sensors, algorithms, and the web which forms "Big Data." This digital data collection grows exponentially each year. According to [2], big data refers to datasets whose size is beyond the ability of typical database software tools and applications to capture, store, manage and analyze.

An important task of any organization is to analyze data. Analysis could change a large volume of data to a smaller amount of valuable data, but we still require collecting a massive amount of data.

Big data has become a complex issue in all disciplines of science. In scientific big data, several solutions have been proposed to overcoming big data issues in the field of life sciences [3, 4], education systems [5], material sciences [6], social networks [7, 8] and.

Some examples of the significance of big data for generating, collecting and computing are listed as follows:

Big data generation and collection:

- It is predicated that data production will be 44 times greater in 2020 than it was in 2009 [9]. This data could be collected from variety resources, such as traditional databases, videos, images, binary files (applications) and text files;
- It is estimated 235 Terabytes of data were collected by the U.S. Library of Congress in April 2011 [10];
- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data [11] which includes a variety of data, such as images, videos and texts.

Computing on big data:

- In 2008, Google was processing 20,000 Terabytes of data (20 petabytes) per day [12].
- Decoding the human genome originally took 10 years to process; now it can be achieved in one week [13] with distributing computing on big data.
- IDC[1] estimates that by 2020, business-to-business and business-to-consumer transactions on the Internet will reach 450 billion per day [14].
- Big data is a top business priority and drives enormous opportunities for business improvement. Wikibon's own study projects that big data will be a $50 billion business by

[1] International Data Corporation (IDC) is an American market research, analysis and advisory firm specializing in information technology, telecommunications, and consumer technology.

2017 [15].

- Macy's Inc. provides a real-time pricing. The retailer adjusts pricing in near real-time for 73 million items for sale based on demand and inventory [16].
- The major VISA process more than 172,800,000 card transactions each day [17].

The most public resource data are available on the Internet, such as multimedia steam data, social media data and text. This variety of data shows we are not facing only structured data, but also unstructured data, such as multimedia files (including video, audio and images), and Twitter and Facebook comments. Unstructured data causes complexity and difficulty in analyzing big data. For example, a corporation analyzes user comments and user shared data on social media that could recognize customer favorites and provide best offers.

To collect and process big data, we can use Cloud Computing Technology. Cloud computing is a new paradigm for hosting clusters of data and delivering different services over a network or the Internet. Hosting clusters of data allows customers to store and compute a massive amount of data on the cloud. This paradigm allows customers to pay on pay-per-use basis and enables them to grow (or shrink) their computing and storage needs on demand. These features allow customers to pay the infrastructure for storing and computing based on their current capacity of big data and transactions.

Currently, capturing and processing big data are related to improving the global economy, science, social affair, education and national security; processing of big data allows us to propose accurate decisions and acquire knowledge from raw data.

This chapter aims to show the role of cloud computing in dealing with big data and intelligent computing. This chapter is organized as follows: Section 2 discusses a definition and characteristics of big data. In Section 3, we discuss important opportunities and challenges in handling big data. In Section 4, we discuss cloud computing and key architectural components for dealing with big data. In this section, we review how each service layer of a cloud computing system could handle big data issues. Also, we provide a list of services and tools for dealing with big data. Finally, in Section 5, we review some major cloud computing issues.

2 Big Data Definition

Often big data is characterized by “4 V’s” [18] which stand for:

- “**Volume**” which indicates a very large volume of data;
- “**Velocity**” which indicates the speed for data processing in terms of response time. This response

time could be a batch, real-time or stream response-time;

- "**Variety**" which indicates heterogeneity in data that we have collected for processing and analysis this data variety includes structured, unstructured and semi-structured data;
- "**Veracity**" which indicates level of accuracy in the data. For example, a sensor that generates data can have a wrong value rather than provides an accurate data.

Big data could have one or multiple of the above characteristics. For example, storing and computing on social data could have a very large volume of data (*volume*) and specific response-time for computing (*velocity*) but it may not have *variety* and *veracity* characteristics.

Another example, analyzing public social media data regarding the purchase history of a customer could provide a future favorite purchase list when she searches for a new product. In this case, big data have all characteristics: *volume of data*, because collecting a massive amount of data from public social media networks; *velocity*, because response-time limited to near real-time when a customer search a product; *variety*, because big data may come from different sources (social media and purchase history); *lack of veracity*, because data from customers in social media networks may have uncertainty. For instance, a customer could like a product in a social media network, not because this is the product of her choice, but because of this product is used by her friend.

Another important question in big data is*, "How large is big data?"* We can answer this question based on our current technology. For example, Adam Jacobs [19] states in the late 1980s at Columbia University that they stored 100 GB of data as big data via an IBM 3850 MSS (Mass Storage System), which costs $40K per GB. In 2010, the Large Hadron Collider (LHC) facility at CERN produced 13 *peta*bytes of data [20]. So what we call big data depends on the cost, speed and capacity of existing computing and storage technologies. For example, in the 1980s, 100 GB was big data because the storage technology was expensive at that time and it had low performance. However, by 2010, the LHC processed 13 Petabyte as a big data which has $1.363*10^5$ times more volume than IBM 3850 MSS big data in 1980s.

At this time, we can refer to 13 petabytes at CERN. In addition, we can also refer to a text file with 10 GB size as big data because a regular text editor could not handle this file size. So the definition of big data is not only *a massive amount of data* but also depends on *what the technology and which size of big data that technology could handle*.

3 Big Data Opportunities and Challenges

On one hand, when we collect big data, we have an opportunity to make an accurate decision through BI. BI is a set of theories and technologies that aim to transfer data from raw-data into meaningful and useful information for business processes (BP). BI became popular in the 1990s, and Business Analytics (BA), which is an analytical component in BI, became popular in the 2000s. In the traditional model, the queries are pre-defined to confirm or refuse a query's hypotheses, but Online Analytical Processing (OLAP) analysis emerges as an approach to answer complex analytical queries. For example, in a car accident we can make a decision about the incident based on driver information. However, when we collect GPS information, engine information and driver information, we can make a more accurate decision about an accident. Also, if we collect more information, we can trust our decision more (*veracity*). In a second example, Volvo provided performance and fault monitoring for predictive warranty analysis [22]. In another example, sensor data from a cross-country flight (New York to Los Angeles) generate 2.499 billion Terabyte per year [23] (*volume*) from different sensors (*variety*), which could be provided from reliable sensors (*veracity*) or unreliable sensors (*lack of veracity*). Often the processing of this data is real-time (*velocity*) and this computing could be processed by an aircraft's server or by a ground's servers.

Collection of information cannot only help us to avoid car accidents, but also could help to make an accurate decision in any systems, such as business financial systems [15], education systems [24], and treatment systems, e.g., Clinical Decision Support Systems [25].

Some important opportunities are provided by big data. They are listed as follows:

- Analyze big data to improve business processes and business plans, and to achieve business plan goals for a target organization (The target organization could be a corporation, industry, education system, financial system, government system or global system.)
- Reduce bulk data to a valuable smaller amount of data
- Provide more accurate decisions by analyzing big data
- Prevent future system failures by predicting big data

On the other hand, we have several issues with big data. The challenges of big data happened in various domains including *storing of big data*, *computing on big data* and *transferring of big data*. We discuss these issues below:

- **Storage Issues**

A database is a structured collection of data. In the late 1960s, flat-file models which were expensive

and slow, used for storing data. For these reasons, relational databases emerged in the 1970s. Relational Database Management Systems (RDBMS) employ Structured Query Language (SQL) to store, edit and retrieve data.

Lack of support for unstructured data led to the emergence of new technologies, such as BLOB (Binary Large Object) in the 2000s. Unstructured data may refer to multimedia data. Also unstructured data may refer to irregularly or randomly repeated column patterns that vary from row to row within each file or document. BLOB could store all data types in most RDBMS.

In addition, a massive amount of data could not use SQL databases because retrieving data and analyzing data takes more time for processing. So "NOSQL", which stands for "Not Only SQL" and "Not Relational", was designed to overcome this issue. NOSQL is a scalable partitioned table that could distribute data over many servers. NOSQL is implemented for cloud computing because in the cloud, a data storage server could be added or removed anytime. This capability allows for the addition of unlimited data storage servers to the cloud.

This technology allows organizations to collect a variety of data but still increasing the volume of data increases cost investment. For this reason, capturing high-quality data that could be more useful for an organization rather than collecting a bulk of data.

- **Computing Issues**

When we store big data, we need to retrieve, analyze and modify it. The important part of collecting data is analyzing big data and converting raw data into valuable information that could improve a business process or decision making. This challenge can be addressed by employing a cluster of CPUs and RAMs in cloud computing technology.

High-Performance Computing (HPC) is another technology that provides a distributed solutions by different computing models, such as traditional (e.g. Grid Computing) or cloud computing for scientific and engineering problems. Most of these problems could not process data in a polynomial time-complexity.

- **Transfer Issues**

Transfer of big data is another issue. In this challenge, we are faced with several sub-issues: Transfer Speed, which indicates how fast we can transfer data from one location/site to another location/site. For example, transferring of DNA, which is a type of big data, from China to the United States has some delay in the backbone of the Internet, which causes a problem when they receive data in the United States [26]. BGI (one of the largest producers of genomic data, Beijing Genomics Institute in Shenzen, China)

could transfer 50 DNAs with an average size of 0.4 terabyte through the Internet in 20 days, which is not an acceptable performance [26].

Traffic Jam: transfer of big data could happened between two local sites, cities or worldwide via the Internet but between any locations this transfer will result in a very large traffic jam.

Accuracy and Privacy: Often we transfer big data through unsecured networks, such as the Internet. Data transfers through the Internet must be kept secure from unauthorized access. Accuracy aims to transfer data without missing any bits.

4 Dealing with Big Data

Several *traditional* solutions have emerged for dealing with big data such as Supercomputing, Distributed Computing, Parallel Computing, and Grid Computing. However, elastic scalability is important in big data which could be supported by cloud computing services which are described in Section 4.2. Cloud computing has several capabilities for supporting big data which are related to handling of big data. Cloud computing could support two major issues of big data, which are described in Section 3 including storing of big data and computing of big data. Cloud computing provides a cluster of resources (storage and computing) that could be added anytime. These features allow cloud computing to become an emerging technology for dealing with big data.

In this section, we will first review important features of cloud computing systems and a correlation of each of them to big data. Second, we discuss a cloud architecture and the role of each service layer in handling big data. Finally, we review implementation models of cloud computing systems as they relate to handling big data.

4.1. Cloud Computing System Features

The major characteristics of cloud computing as defined by the U.S. National Institute of Standards and Technology (NIST) [27] are as follows:

- **On-demand Elastic Service**

This characteristics show the following features: (i) an economical model of cloud computing which enables consumers to order required services (computing machines and/or storage devices). The service requested service could scale rapidly upward or downward on demand; (ii) it is a machine responsibility that does not require any human to control the requested services. The cloud architecture manages on-demand requests (increase or decrease in service requests), availability, allocation, subscription and the customer's bill.

This feature is interesting for a start-up businesses, because this feature of cloud computing systems allows a business to start with traditional data or normal datasets (in particular start-up business) and increase their datasets to big data as they receive requests from customers or their data grows during the business progress.

- **Resource pooling**

A cloud vendor provides a pool of resources (e.g., computing machines, storage devices and network) to customers. The cloud architecture manages all available resources via global and local managers for different sites and local sites, respectively.

This feature allows big data to be distributed on different servers which is not possible by traditional models, such as supercomputing systems.

- **Service Accessibility**

A cloud vendor provides all services through broadband networks (often via the Internet). The offered services are available via web-based model or heterogeneous client applications [28]. The web-based model could be an Application Programming Interface (API), web-services, such as Web Service Description Language (WSDL). Also heterogeneous client applications are provided by the vendors. Customers could run applications on heterogeneous client systems, such as Windows, Android and Linux. This feature enables partners to contribute to big data. These partners could provide cloud software applications, infrastructure or data. For example, several applications from different sites could connect to a single-data or transparent multiple-data warehouse for capturing, analyzing or processing of big data.

- **Measured Service**

Cloud vendors charge customers by a metering capability that provides billing for a subscriber, based on pay-per-use model. This service of cloud architecture manages all cloud service pricing, subscriptions and metering of used services. This capability of cloud computing system allows an organization to pay for the current size of datasets and then pay more when dataset size increases. This service allows customers to start with a low investment.

4.2. Cloud Architecture

Cloud computing technology could provide by a vendor that enables IT departments to focus on their software development rather than hardware maintenance, security maintenance, recovery maintenance, operating systems and software upgrades. Also, if an IT department establishes a cloud computing system in their organization, could help them to handle big data.

The Architecture of a cloud computing system is specific to the overall system and requirements of

each component and sub-components. Cloud architecture allows cloud vendors to analyze, design, develop and implement big data.

Cloud vendors provide services through service layers in cloud computing systems. The major categories are divided into four service layers: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS) and Business Intelligence (BI) and other service layers assigned to the major service layers as shown in Figure 1, such as Data-as-a-Service(DaaS) assigned to IaaS layer. Description of each service discussed in Section 4.2.5.

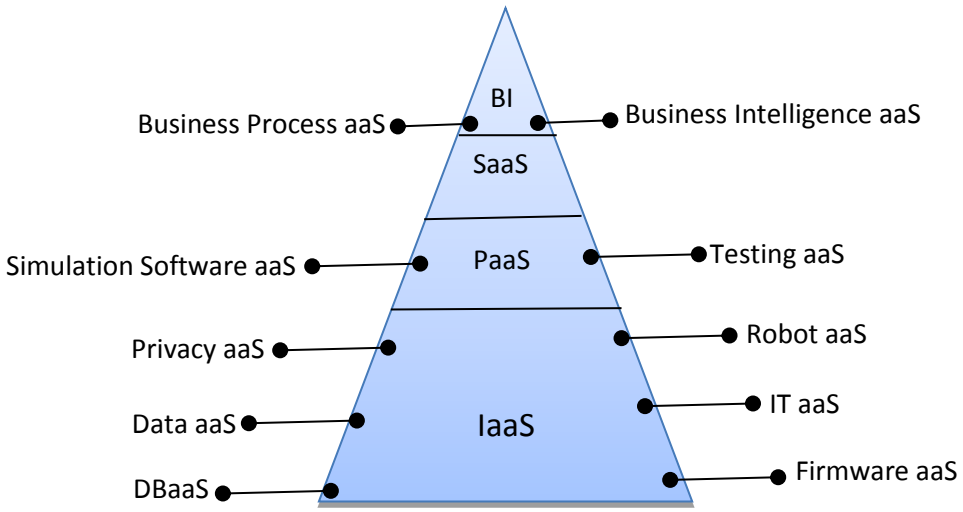

Figure 1. Cloud Services

4.2.1. The Role of Infrastructure as a Service (IaaS)

The IaaS model offers storage, processors and fundamental hardware to the cloud customers. This model covers several services, such as firmware, hardware, utilities, data, databases, resources and infrastructure. This model allows clients to install operating systems, receive quoted infrastructure, and develop and deploy required software applications. This model is often implemented via Virtualization, which enables multi users/tenants work on share machines with his own privacy.

The IaaS model provides several opportunities for big data: ***(i) storage data:*** this feature allows customers to store big data. Storage on the cloud computing system enables customers to store, retrieve and edit big data by employing a cluster of storage devices. These clusters could be added or removed dynamically; ***(ii) hardware:*** this feature enables customers have an access to a resource pool of hardware for big data. This feature could be used for capture data, such as through sensors, Radio-Frequency Identifications (RFIDs) or Communication-as-a-Service (CaaS). The CaaS is responsible for the required hardware and software for delivering Voice-over-IP (VoIP), audio and video conferencing. The hardware

feature also provides network access and network traffic control that could to transfer big data.

Amazon Elastic Compute Cloud (Amazon EC2) provides virtual and scalable computing systems at the IaaS. Amazon EC2 customers could define instances of a variety of operating systems (OSs). Each OS and required hardware, such as CPUs and RAMs could be customized by a customer on the fly. Customers should create an Amazon Machine Image (AMI) in order to use Amazon EC2. The AMI contains the required applications, operating systems (the customer could select various operating systems such as Windows or Linux versions), libraries, data and system configuration. Amazon EC2 uses Amazon S3, which is a cloud storage service and stores data and uploads AMI into S3.

The impact of big data in this service layer is higher than other service models in cloud computing systems, because IaaS users could access and define the required *data framework*, *computing framework* and *network framework*.

In a *data framework*, users could define structured data, unstructured data and semi-structured data. Structured and semi-structured data could be defined via traditional databases, such as RDBMS and OODBMS. In these models, structured data stored which has a schema before adding data to the databases. All of data frameworks and in particular unstructured data could be defined by cloud databases, such as Hadoop which is based on MapReduce programming model. MapReduce programming language technique allows storing data on a cluster of resources. The implementation model of MapReduce is provided by Hadoop which is provided a category of open-source database, applications and analytics tools.

In *computing framework*, users have full-permission for developing, installing and running new application for computing purposes. Each application could reserve a cluster of CPUs and RAMs. Several tools and databases with analysis tools emerge to provide computing framework on big data. For example, Hive is an open-access “SQL-like” BI tools that allows BI applications to run query on Hadoop data. Other example, Pig is another open-source platform that allows analyzing on big data by a “Perl-language-like” feature.

In *network framework, users have a* significant benefit, because they have access to required network control, such as network cards and the Internet connectivity. For example they could access to regular network transfer infrastructure such as Optical Carrier (OC) 768 backbone [29], which is capable of transferring 39,813.12 Mbit/s.

This accessibility to data, computing and network framework allows the users to control require hardware like an administrator in IT department. However, these users could handle infrastructure without worrying about maintenance.

4.2.2. The Role of Platform as a Service (PaaS)

PaaS is a platform that provided by cloud vendor. The PaaS model does not require users to setup of any software, programming language, environment application, designer, tools or application. Developers use vendor's platform, library and programming language for developing their applications. This model provides a software application for outgrowth of the cloud applications delivery. PaaS allows developer to focus on software application development, without worrying about operating system maintenance like in IaaS. The PaaS provides services for software programmers to develop and deploy their applications with an abstraction on the hardware layer.

The role of PaaS in handling big data is less than IaaS, because some restrictions and limitations are applied to PaaS users in order to work on the data framework, computing framework and transfer frameworks. In this service layer, users are limited to cloud vendor frameworks. For example, Google App Engine provides a platform which supports Python, Java, PHP, Go and MySQL compatible Cloud SQL to develop applications. So, in this service layer, users could not access other languages, such as C# or C++ and server hardware. However, developers still could build, deploy and run their scalable applications on the cloud computing systems. These applications could capture a massive amount of data from anywhere and use a cluster of CPUs for computing and analytics of big data.

4.2.3. The Role of Software as a Service (SaaS)

The traditional model of software is to purchase software applications and install them on the local computer. However, SaaS model provides applications in the cloud though a network and does not require customers to install applications on their local computers.

According to Microsoft, SaaS model could be divided to the following categories (lower-level to higher-level) [30]:

- *Ad-hoc/Custom,* which supports by minimum requirement to migrate traditional and client/server application to this level. Ad-hoc/Custom models allow developer to build their application based on ad-hoc or peer-to-peer technology;
- *Configurability,* which provides more flexibility through configuration metadata and supports peer-to-peer technology;
- *Multi-tenancy,* which adds multi-tenancy to the configuration level, and a single instance of application allows serving all the vendor's consumers;
- and *Scalability,* which supports all other lower-levels. In addition, this level supports scalability through architectural design that adds a capability of dynamic load-balancing for growing or shrinking cloud servers. Most applications in the cloud are developed at this level.

The impact of SaaS is less than PaaS, because in this service layer, users could use provided applications and resources. This service layer is limited to developers. However, users still could work on

big data that could be added before or captured by provided infrastructure. For example, Google Apps, such as Gmail, provides services on the web and users could not add or manipulate capturing data from server. Users are limited to web-based interface for email processes such as sending an email.

4.2.4. The Role of Business Intelligence (BI)

The BIaaS layer sits on the top of cloud architecture service layers and aims to provide the required analytic models for cloud customers.

Information granularity as Pedrycz defined [31] is a structure which plays a key role in human cognitive and decision-making computing. The BI service layer could provide a platform for information granularity on the cloud computing and in particular granular computing, which is a processing of complex information entities. Unlike the traditional computing, cloud computing by granular computing on big data may provide a significant result. For example, Bessis at al. [32] propose a big picture by collecting big data and using cloud computing for managing disasters.

Cloud computing could provide the following information granularity and granular computing infrastructures [31]:

- A granular description of data and pattern classification by non-SQL databases, such as SciDB[33];
- A representation of information granules by migrating traditional applications to the cloud;
- Different granular architecture and development by collecting information from different sources and computing with high quality rather than traditional models which were working with a limited computing resource;
- Collaborative and linguistic models of decision-making by collecting information from different sources at the cloud storages.

The information-processing level [34], which is encountering a number of conceptual and algorithmic layers indexed by the size of information granular, could be high if a cloud application provides a computing model. However, if a cloud application provides only a storage model, this impact and granular computing will be low. For example, when an application provides a service for collecting data from financial consumers and running an analytical model on this data to make a decision about investment, cost and profit, this application has a high-level BIaaS impact. For instance, Xu et al. [35] present “Big Cloud based Parallel Data miner (BC-PDM)” which is a framework for integrating data mining applications on MapReduce and HDFS (Hadoop File System) platforms.

Cloud based BI could reduce the total development cost, because cloud computing systems provide environment for agile development and reduce the maintenance cost. Also, the BI could not be

implemented on a traditional system, because the current volume of data for analysis is massive. BI-as-a-Service [36] is other example that shows how the BI could migrate to the cloud computing systems as a software application in the SaaS layer.

One of the major challenges with traditional computing is analysis of big data. Cloud computing at BIaaS layer could handle this issue by employing a cluster of computing resources. For example, SciDB[33] is an open-source and cloud-based database management system (NOSQL DBMS) for scientific application with several functions for analyzing of big data, such as astronomy, remote sensing and climate modeling.

4.2.5. Other Service Layers

The major service models of cloud computing are BIaaS, IaaS, PaaS and SaaS. As shown in Table 1, we assigned each service to the major service models.

Table 1. Other service layers in Cloud Architecture

Service name	Related to	Service Description and Offers	Role of Service in Big Data
Business-Process-as-a-Service (BPaaS) [37]	BIaaS	Automated tool support	Analysis of big data
Business-Intelligence-as-a-Service (BIaaS) [38]	BIaaS	Integrated approaches to management support	Analysis of big data
Simulation Software-as-a-Service (SimSaaS) [39]	SaaS	Simulation service with a MTA configuration model	Analysis of big data
Testing-as-a-Service (TaaS) [40]	SaaS	Software testing environments	Test big data tools
Robot-as-a-Service (RaaS)[41]	PaaS	Service-oriented robotics computing	Action on big data
Privacy-as-a-Service (PaaS) [42]	PaaS	A framework for privacy preserving data sharing with a view of practical application	Big data privacy
IT-as-a-Service (ITaaS) [43]	IaaS	Outsource IT department's resource (on Grid infrastructure that time)	Maintaining of big data
Hardware-as- a Service (HaaS) [44]	IaaS	A transparent integration of remote hardware that is distributed over multiple geographical locations into an operating system.	Capturing and maintaining of big data
Database-as-a-Service (DBaaS) [45]	IaaS	(1) a workload-aware approach to multi-tenancy (2) a graph-based data partitioning algorithm (3) an adjustable security scheme	Storing big data
Data-as-a-Service (Daas) [46]	IaaS	Analyzing major concerns for data as a service	Storing big data
Big-Data-as-a-Service [47]	All layers	Service-generate for big data	Generate big data

4.2.6. Big Data Tools

The Table 2 shows some big data open-source tools which are provided through cloud computing

infrastructures. Most of the tools are provided by Apache[1] and released under the Apache License. We categorized each tool based on those applications of big data.

Table 2. Big Data Tools

Big Data Tools	Description[2]
Data Analysis Tools	
Ambari[3]	A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.
Avro[4]	A data serialization system.
Chukwa[5]	A data collection system for managing large distributed systems.
Hive[6]	A data warehouse infrastructure that provides data summarization and ad hoc querying.
Pig[7]	A high-level data-flow language and execution framework for parallel computation.
Spark[8]	A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
ZooKeeper[9]	A high-performance coordination service for distributed applications
Actian[10]	An Analytics Platform which accelerates the analytics value chain from connecting to massive amounts of raw big data all the way to delivering actionable business value.
HPCC[11]	Provide high-performance, data-parallel processing for applications utilizing big data.
Data Mining Tools	
Orange[12]	A data visualization and analysis for novice and experts.
Mahout[13]	A scalable machine learning and data mining library.
KEEL[14]	An assess-evolutionary algorithm for data mining problems.
Social Network Tools	
Apache Kafka	A unified, high-throughput, low-latency platform for handling real-time data feeds.
BI Tools	
Talend[15]	A data integration, data management, enterprise application integration and big data software tools and services.
Jedox[16]	An analyzing, reporting and planning functions.
Pentaho[17]	A data integration, business analytics, data visualization and predictive analytics.
rasdaman[18]	A multi-dimensional raster data (arrays) of unlimited size through an SQL-style query language.

[1] http://apache.org/
[2] The description retrieved from each tools official website and Wikipedia at http://wikipedia.org
[3] http://ambari.apache.org/
[4] http://avro.apache.org/
[5] http://incubator.apache.org/chukwa/
[6] http://hive.apache.org/
[7] http://pig.apache.org/
[8] http://spark.incubator.apache.org/
[9] http://zookeeper.apache.org/
[10] http://www.actian.com/about-us/#overview
[11] http://hpccsystems.com/
[12] http://orange.biolab.si/
[13] http://mahout.apache.org/
[14] http://keel.es/
[15] http://www.talend.com/
[16] http://www.jedox.com/en/
[17] http://www.pentaho.com/
[18] http://rasdaman.eecs.jacobs-university.de/

Search Tools	
Apache Lucene[1]	An application for full text indexing and searching capabilities.
Apache Solr[2]	A full-text search, hit highlighting, faceted search, near real-time indexing, dynamic clustering, database integration, rich document (e.g., Word, PDF) handling, and geospatial search.
Elasticsearch[3]	A distributed, multitenant-capable full-text search engine with a RESTful web interface and schema-free JSON documents.
MarkLogic[4]	A NOSQL and XML database.
mongoDB[5]	A cross-platform document-oriented database system, JSON-like documents with dynamic schemas.
Cassandra[6]	A scalable multi-master database with no single point of failure.
HBase[7]	A scalable, distributed database that supports structured data storage for large tables.
InfiniteGraph[8]	A distributed graph database.

4.3. Implementation Models of Cloud Computing Systems

A Cloud computing system based on infrastructure location could be implemented as Private, Public or Hybrid cloud.

The ***private model*** is a local implementation of cloud computing system. In this model, hardware is located in local data centers and uses cloud software applications to provide service to local users. This model is the best option for consumers who needs cloud computing capabilities with low-risk in IT departments because this model allows an IT department to migrate from the traditional model to the cloud computing system and does not require data to be migrated to another location (such as cloud vendor location). This model is implemented for local trusted users. This model still allows scalability, on-demanded self-service, and elastic service. However, this model requires high investment in maintenance, recovery, disaster control, security control, and monitoring.

In addition, the private cloud computing model enables an IT department to handle a local organization's big data by its own infrastructure, such as the storage of big data and computing big data. This model provides a flexible resource assignment and could enhance the resource availability.

Several open source applications have been developed for establishing private cloud computing based on IaaS and SaaS service layers. For example, CloudIA is a private cloud computing system at (HFU) [48]. The targeted users of the CloudIA project are HFU staff and students running e-Learning

[1] http://lucene.apache.org/
[2] http://lucene.apache.org/solr/
[3] http://www.elasticsearch.org/
[4] http://developer.marklogic.com/
[5] http://www.mongodb.org/
[6] http://cassandra.apache.org/
[7] http://hbase.apache.org/
[8] http://www.objectivity.com/

applications, and external people for collaboration purposes.

The ***public model*** is a regular model of cloud computing system. This model is provided by cloud vendor who supports billing and a subscription system for public users. This model, unlike a private model, does not require high investment, because consumers could pay on pay-per-use basis for cloud storage or cloud computing services on demand.

The **hybrid model** composes private and public clouds. This model could connect a private cloud to public cloud through network connection, such as the Internet.

This model has several advantages, which are listed below:

- *Collaboration between cloud computing systems:* Often collaboration between two clouds led to emergence of hybrid cloud model. An organization could keep their own cloud security and maintenance, and simultaneously have collaboration with other clouds. This collaboration could be permanent or temporary.
- *Scalability:* This model also is useful for extending the scalability of a private cloud computing system, because in case of limited resources at a peak time, a cluster of new resources could be added temporary from another cloud.

5 Cloud Computing Issues

The cloud computing technology is the best option for dealing with big data. However, cloud computing is still nascent state and we still needed to address some major issues. In this section, we review the major cloud computing issues which are shown in Figure 2 and are based on an IDC Survey in 2009 [49].

When big data costs customers, and a system disaster could cause organizational destruction in the digital age, migration applications and databases from traditional model are difficult to cloud, because:

- migration to the cloud computing system is difficult; Migration requires to redevelop applications, data and sometimes requires to use efficient programming models to save resources as well as resource costs;
- returning data to the IT department is difficult;
- connection is via an unsecured network, such as the Internet;
- cloud vendor administrator users could have an access to users data;
- data warehouse location is transparent to consumers ;
- We do not have a cloud computing standard and standard cloud architecture. It causes some big

issues, such as different architectures, difficulty with migration data and application to another cloud vendors;

- We do not have any customization in cloud computing systems;
- We do not have a strong Service Layer Agreement (SLA) for customer satisfaction.

Cloud customers need to have a contract with one or more cloud vendor(s) -often one cloud vendor- and they should use the provided operating systems, middlewares, APIs and/or interfaces. Data and application are dependent on the platforms or are provided by cloud vendor infrastructure. This dependency in cloud services has several issues. For example, in Figure 2, "*Security*" is the major concern in cloud computing systems. Cloud features, such as a shared resource pool and multi-user/tenancy cause security issue because the resourced pool are shared through users and we could expose users' data and users' privacy to others.

Unsecured connection to the vendor, network access security, Internet access security and cloud vendors' user security emerged as other major security concerns based on accessibility to the cloud via the Internet.

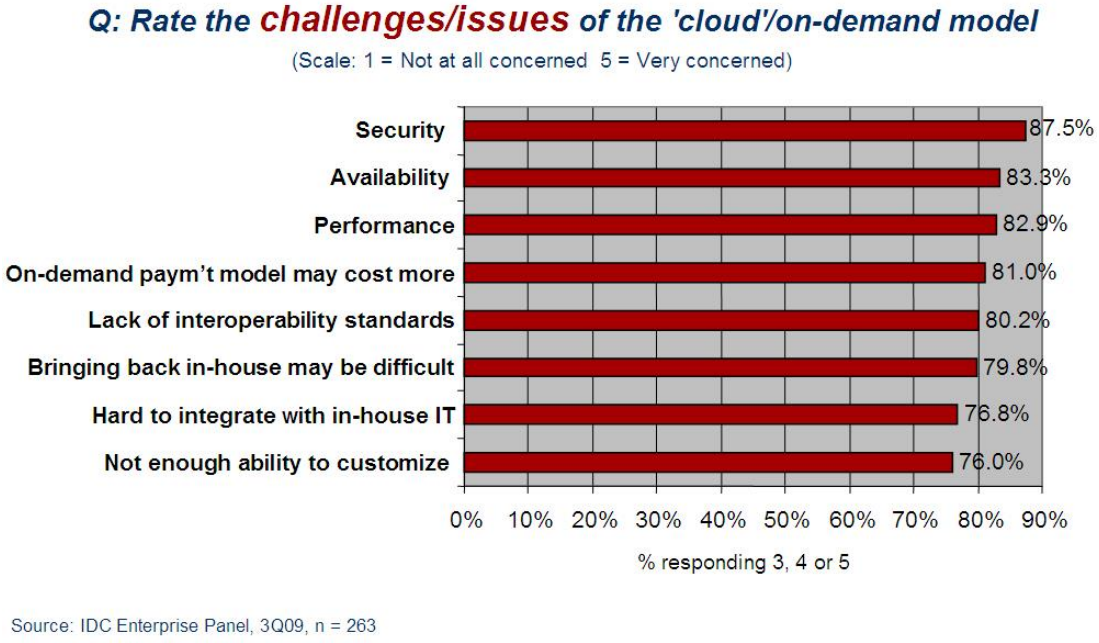

Figure 2. Major Cloud Computing Concerns [49]

"*Bringing back in-house may be difficult*" with 79.8% issue rate and "*Hard to integrate with in-house IT*" with 76.8% issue rate indicates customers are afraid of data and software application migration to the cloud computing systems, because the migration is difficult to integrate with IT departments and it is difficult to return data back to the IT department; "*Lack of interoperability standards*" with 80.2% is

another cloud issue. This issue shows that cloud computing requires higher interoperability with other cloud computing systems; also as indicated in this report, "*Not enough ability to customize*" with a 76.0% issue rates show, the cloud computing system requires dynamic architecture and customization.

Some studies, such as [50] show existing cloud computing systems (Amazon EC2 in this case) could not be responsible with a cost-effective performance for HPC applications over using tightly-couple hardware such as Grid Computing or Parallel Computing systems.

To overcome these issues, some study such as [51, 21] are proposed which introduce "Cloud Template architecture". Especially when we employ the cloud computing system for dealing with big data, this architecture is useful. In this study, we show each template could be organized for each purpose and a template could support several service layers simultaneously.

6 Chapter Summary

In this chapter, we discussed a definition of big data, the importance of big data, and major big data challenges and issues. We understand that, if we analyze big data with business intelligence tools, we may provide a catalyst to change an organization to a smart organization. We discussed the importance of cloud computing technology as a solution to handle big data for both computing and storage. We reviewed the capabilities of cloud computing systems that are important for big data, such as resource scalability, resource shrink-ability, resource pool sharing, on-demanded servicing, elastic servicing, and collaboration with other cloud computing systems. We explained cloud architecture service layers and role of each service layer to handle big data. We discussed how business intelligence could change big data to smaller valuable data by using cloud computing services and tools. Finally, we discussed major cloud computing system issues that need to be addressed for cloud computing to become a viable solution for handling big data.

References

[1] Matheson, David, and James E. Matheson, "The Smart Organization: Creating Value through Strategic", Rand D. Harvard Business Press, 1998.
[2] Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).
[3] Buscema, Massimo, et al. "Auto-Contractive Maps: an artificial adaptive system for data mining. An application to Alzheimer disease" Current Alzheimer Research 5.5 (2008): 481-498.
[4] Howe, Doug, et al. "Big data: The future of biocuration." Nature 455.7209 (2008): 47-50.

[5] Hanna, Margo. "Data mining in the e-learning domain" Campus-wide information systems

21.1 (2004): 29-34.
[6] Wilson, Lori A. "Survey on Big Data gathers input from materials community" MRS Bulletin 38.09 (2013): 751-753.
[7] Tan, Wei, et al. "Social-Network-Sourced Big Data Analytics" Internet Computing, IEEE 17.5 (2013): 62-69.
[8] Jing Huang, Kai Wu, Lok Kei Leong, Seungbeom Ma, and Melody Moh, "A Tunable Workflow Scheduling Algorithm Based on Particle Swarm Optimization for Cloud Computing", International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, pp. 351-358, 2013.
[9] Revisited: The Rapid Growth in Unstructured Data, retrieved on Jan 21, 2014 at http://wikibon.org/blog/unstructured-data
[10] Infographic: The Potential of Big Data, retrieved on Jan 21, 2014 at http://blog.getsatisfaction. com/2011/07/13/big-data/?view=socialstudies
[11] Taming Big Data [A Big Data Infographic], retrieved on Jan 21, 2014 at http://wikibon.org/blog/taming-big-data/
[12] Erick Schonfeld, Google Processing 20,000 Terabytes A Day, And Growing, , retrieved on Jan 21, 2014 at http://techcrunch.com/2008/01/09/google-processing-20000-terabytes-a-day-and-growing/
[13] Data, data everywhere, retrieved on Jan 21, 2014 at http://www.economist.com/node/15557443
[14] The Big List of Big Data Infographics, retrieved on Jan 21, 2014 at http://wikibon.org/blog/big-data-infographics
[15] Josette Rigsby, Studies Confirm Big Data as Key Business Priority, Growth Driver, retrieved on Jan 21, 2014 at http://siliconangle.com/blog/2012/07/13/studies-confirm-big-data-as-key-business-priority-growth-driver
[16] Davenport T H, Dyche J (2013), Big Data in Big Companies, SAS
[17] Fairhurst, Paul. "Big data and HR analytics." IES Perspectives on HR 2014 (2014): 7.
[18] McAfee, Andrew, and Erik Brynjolfsson. "Big data: the management revolution." Harvard business review 90.10 (2012): 60-66.
[19] Adam Jacob, “The Pathologies of Big Data”, Communication of the ACM, Vol.52, No. 8, pp.36-44, 2009.
[20] V. Gewin, “The New Networking Nexus”, Nature, vol.451, no.7181, pp. 1024-1025, 2008.
[21] Mehdi Bahrami, "Cloud Computing Software Architecture and Innovation in the Cloud", International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, pp. 23-24, 2013, Doi: 10.7321/jscse.v3.n3.6.
[22] Mark Young, “Automotive innovation: big data driving the changes”, retrieved Jan 26,2014 at http://www.thebigdatainsightgroup.com/site/article/automotive-innovation-big-data-driving-changes
[23] Jeff Kelly, “Big Data in the Aviation Industry”, Wikibon, Sep 16, 203, retrieved on March 18, 2014 at: http://wikibon.org/wiki/v/Big_Data_in_the_Aviation_Industry
[24] Siegel, Carolyn F. "Introducing marketing students to business intelligence using project-based learning on the world wide web." Journal of Marketing Education 22.2 (2000): 90-98.
[25] Berner, Eta S. Clinical Decision Support Systems. Springer Science+ Business Media, LLC, 2007.
[26] Marx, Vivien. "Biology: The big challenges of big data." Nature 498.7453 (2013): 255-260.

[27] Liu, Fang, et al. "NIST cloud computing reference architecture." NIST special publication 500 (2011): 292.
[28] Mukesh Singhal, "A Client-centric Approach to Interoperable Clouds", International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, pp. 3-4, 2013.
[29] Cartier C, Paynetitle T (2001-07-30). "Optical Carrier levels (OCx)". Retrieved 01-24-2014.
[30] Rittinghouse, John W., and James F. Ransome. Cloud computing: implementation, management, and security. CRC press, 2009.
[31] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC Press/Francis Taylor, Boca Raton, 2013
[32] Bessis, Nik, et al. "The big picture, from grids and clouds to crowds: a data collective computational intelligence case proposal for managing disasters." P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2010 International Conference on. IEEE, 2010.
[33] Cudré-Mauroux, Philippe, et al. "A demonstration of SciDB: a science-oriented DBMS." Proceedings of the VLDB Endowment 2.2 (2009): 1534-1537.
[34] Bargiela, Andrzej, and Witold Pedrycz. Granular computing: an introduction. Springer, 2003.
[35] Xu, Meng, et al. "Cloud computing boosts business intelligence of telecommunication industry." Cloud Computing. Springer Berlin Heidelberg, 2009. 224-231.
[36] Zorrilla, Marta, and Diego García-Saiz. "A service oriented architecture to provide data mining services for non-expert data miners." Decision Support Systems 55.1 (2013): 399-411.
[37] Accorsi, Rafael. "Business process as a service: Chances for remote auditing."Computer Software and Applications Conference Workshops (COMPSACW), 2011 IEEE 35th Annual. IEEE, 2011.
[38] Hunger, Jens. Business Intelligence as a Service. GRIN Verlag, 2010.
[39] Wei-Tek Tsai, Wu Li, Hessam Sarjoughian, and Qihong Shao. 2011. SimSaaS: simulation software-as-a-service. In Proceedings of the 44th Annual Simulation Symposium (ANSS '11). Society for Computer Simulation International, San Diego, CA, USA, 77-86.
[40] Candea, George, Stefan Bucur, and Cristian Zamfir. "Automated software testing as a service." Proceedings of the 1st ACM symposium on Cloud computing. ACM, 2010.
[41] Chen, Yinong, Zhihui Du, and Marcos García-Acosta. "Robot as a service in cloud computing." Service Oriented System Engineering (SOSE), 2010 Fifth IEEE International Symposium on. IEEE, 2010.
[42] Itani, Wassim, Ayman Kayssi, and Ali Chehab. "Privacy as a service: Privacy-aware data storage and processing in cloud computing architectures." Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on. IEEE, 2009.
[43] Foster, Ian, and Steven Tuecke. "Describing the elephant: The different faces of IT as service." Queue 3.6 (2005): 26-29.
[44] Stanik, Alexander, Matthias Hovestadt, and Odej Kao. "Hardware as a Service (HaaS): The completion of the cloud stack." Computing Technology and Information Management (ICCM), 2012 8th International Conference on. Vol. 2. IEEE, 2012..
[45] Curino, Carlo, et al. "Relational cloud: A database-as-a-service for the cloud", 2011.
[46] Truong, Hong-Linh, and Schahram Dustdar. "On analyzing and specifying concerns for data as a service." Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific. IEEE, 2009.
[47] Zibin Zheng; Jieming Zhu; Lyu, M.R., "Service-Generated Big Data and Big Data-as-a-

Service: An Overview," Big Data (BigData Congress), 2013 IEEE International Congress on, vol., no., pp.403,410, June 27 2013-July 2 2013
[48] Doelitzscher, Frank, et al. "Private cloud for collaboration and e-Learning services: from IaaS to SaaS." Computing 91.1 (2011): 23-42.
[49] IDC Enterprise Panel, 3Q09, retrieved on October 13, 2013 at http://blogs.idc.com/ie/?p=730
[50] Juve, Gideon, et al. "Scientific workflow applications on Amazon EC2." E-Science Workshops, 2009 5th IEEE International Conference on. IEEE, 2009.
[51] Mehdi Bahrami, "Cloud Template, a Big Data Solution", Journal of Soft Computing and Software Engineering , Vol. 3, No. 2, pp.13-17, 2013.