Big Data Analysis and Storage

Khalid Adam Ismail Hammad
Faculty of Computer System and Software Engineering
University Malaysia Pahang
Kuantan, Malaysia
Khalidwsn15@gmail.com

Mohammed Adam Ibrahim Fakharaldien
Faculty of Computer System and Software Engineering
University Malaysia Pahang
Kuantan, Malaysia
adamibrahim@ump.edu.my

Jasni Mohamed Zain
Faculty of Computer System and Software Engineering
University Malaysia Pahang
Kuantan, Malaysia
jasni@ump.edu.my

Mazlina Abdul Majid
Faculty of Computer System and Software Engineering
University Malaysia Pahang
Kuantan, Malaysia
mazlina@ump.edu.my

Abstract

While technologies to build and run big data projects have started to mature and proliferate over the last couple of years, exploiting all potentials of big data is still at a relatively early stage. In fact, Big data is term refer to huge data sets, have high Velocity , high Volume and high Variety and complex structure with the difficulties of management , analyzing, storing and processing .Due to characteristic of big data it becomes very difficult to Management, analysis, Storage, Transport and processing the data using the existing traditional techniques. This paper introduces Big Data Analysis and storage. First we presents the Big data technology alongside it's the significance of big data in the modern world and venture existing which are successful and essential in changing the idea of science into huge science and society as well. Following that, we present How Fast Data is Increasing and The Importance of Big Data. In addition, we discuss Big Data Technologies include (Big Data Frameworks and Platforms and Databases for Big Data). Moreover, we discuss Data Storage and Big Data Management and Storage. Then, we present Big Data Analysis and Management include (Big Data with Data Mining, Big Data over Cloud Computing and Hadoop Distributed File System (HDFS) and MapReduce). Furthermore, we also discuss big data modeling and big data security issues. Finally Conclusion and Future work.

Keywords
Big data, Hadoop, Data Mining, cloud computing

1. Introduction
Nowadays, we live in a more and more interconnected world that generates a great volume of information every day, starting from the logging files of the users of social networks, search engines, e-mail clients to machine generated data as from the real-time monitoring of sensor networks for dams or bridges, and various vehicles such as

airplanes, cars or ships. According to an info graphic made by Intel, 90% of the data today was made in the most recent two years, and the growth continues. It is evaluated that all the worldwide data generated from the earliest starting point of time until 2003 represented about 5 exaBytes (1 exaByte equals 1 million gigaBytes), the measure of data produced until 2012 is 2.7 zettaBytes (1 zettaBytes equals 1000 exaBytes) also, it is perspective to grow 3 times bigger than that until 2015 (Big Data Infographic 2012). For example, the number of RFID tags sold globally is projected to rise from 12 million in 2012 to 209 billion in (Big Data Infographic 2012). All this volume represents a great amount of data that rise challenges when talking about acquiring, organizing and analyzing it. Big Data is an umbrella term describing all these types of information mentioned above. As the name suggests, Big Data refers to a great volume of data, but this is not enough to describe the meaning of the concept. The data presents a great variety, it is usually unsuitable for typical relational databases treatment, being raw, semistructured or unstructured. Also, the data will be processed in different ways, depending on the analysis that needs to be done or on the information that must be found in the initial data. Usually, this big amount of data is created with great velocity and must be caught and processed rapidly (as in the case of real time monitoring). Often, the meaningful and useful information comprised represents a small percent of the initial big volume of data – this means that the data has a low value density.

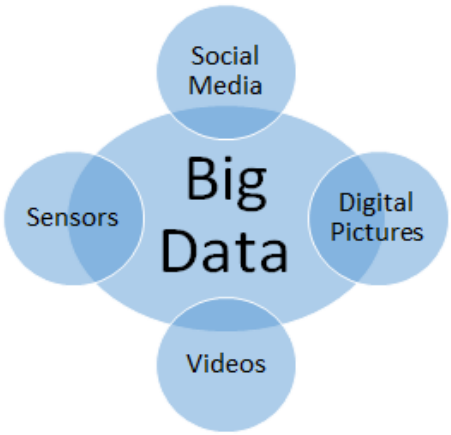

Figure 1: Big data sources

Data is the key factor today. It includes personal, professional, social data and more. Digitalization and interconnectivity lead to an unexpected growth of data. The increased use of media and physical networking through sensor networks for business & private purposes generates an enormous amount of data. This in response changes business processes & open up new opportunities worldwide. The internet is a key driver for data growth. The worldwide generated data already exceed the available storage (Google 2013). Since 2011 interest in an area known as big data has increased exponentially (Nibedita and Sandeep 2014). Unlike the vast majority of computer science research, big data has received significant public and media interest. The era of "big data" has opened several doors of opportunities to upgrade science, boost health care services, improve economic growth, reconstruct our educational system, and prepare new types of social interaction and entertainment services. The area of big data is fast-evolving, and is likely to be subject to improvements and amendments in the future.

A. Big Data Definitions

Big data has been defined simply as "Big data refers to data volumes in range of exabytes (1018) and beyond" in Kaisler et al. (2013). As per Wikipedia, big data is an accumulation of datasets so huge and complex that it becomes hard to process using database management tools or traditional data processing applications, where the challenges include capture, storage, search, sharing, transfer, analysis, and visualization (Wikipedia 2014)In this definition big data is addressed as a problem. Sam Madden from Massachusetts Institute of Technology (MIT) wrote "Big data means too big, too fast, or too hard for existing tools to process" (Madden 2012). He also explained, the term 'too big' as the amount of data which might be at petabyte-scale and come from various sources, 'too fast' as the data growth, which is fast and must be processed quickly, and 'too hard' as the difficulties of big data that does not fit neatly into an existing processing tool (Madden 2012).

From PC Mag (popular magazine based on latest technology news), "Big data refers to the massive amounts of data that collects over time that are difficult to analyze and handle using common database management tools" (PC

Magazine). John Weathington has defined big data as a competitive key parameter in different dimensions such as customers, suppliers, new entrants and substitutes. According to him, big data creates products which are valuable and unique, and prelude other products from satisfying the same need. He also described, "Big data is traditionally characterized as a rushing river: large amounts of data flowing at a rapid pace" (Weathington 2012) (Doctorow 2008). Philip Hunter in has stated, "Big data embodies an ambition to extract value from data, particularly for sales, marketing, and customer relations" (Hunter 2013). Svetlana Sicular has defined big data as "high-volume, -velocity and –variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Sicular 2013). There are many more big data definitions available describing the different characteristics of it.

B. Big Data Characteristics

The characteristics of big data are well defined in the definition by Gartner (Beyer and Laney 2012). The three Vs (volume, velocity and variety) are known as the main characteristics of big data. The characteristics are described below.

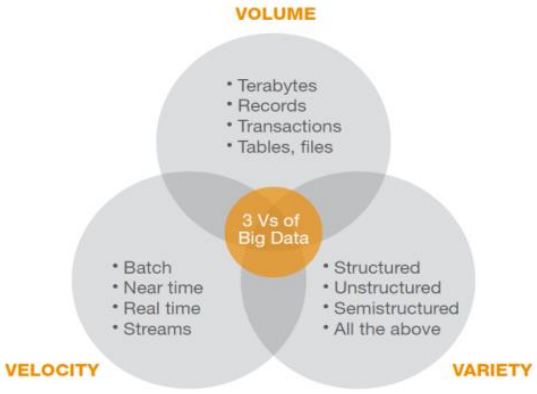

Figure 2:3Vs of big data

Volume: refers to amount of data and there are many factors that can contribute to the volume increase in data It could amount to hundreds of terabytes or even petabytes of information generated for everywhere Avita Katal et al. (2013).The number of sources of data for an organization is growing. More data sources consisting large datasets increase the volume of data, which needs to be analyzed Kaisler et al. (2013). Figure 2 shows that the data volume is growing from megabytes (106) to petabytes (1015) and beyond. Figure 3 indicates that the volume of data stored in the world would be more than 40 zettabytes (1021) by 2020.

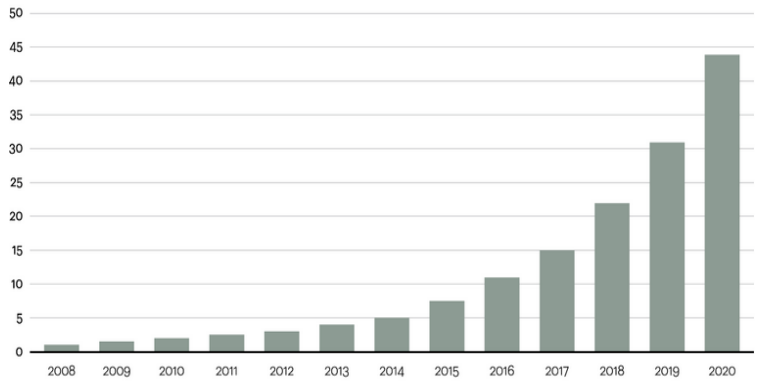

Figure 3: Data volume growth by year in zettabytes

Velocity: refers to data speed measures the velocity of information creation, gushing and collection Kaisler et al. (2013).According to Svetlana Sicular from Gartner, velocity is the most misunderstood big data characteristic (Sicular 2013). She describes that the data velocity is also about the rate changes, and about combining data sets that are coming with different speeds. The velocity of data also describes bursts of activities, rather than the usual steady tempo where velocity frequently equated to only real-time analytics (Sicular 2013).

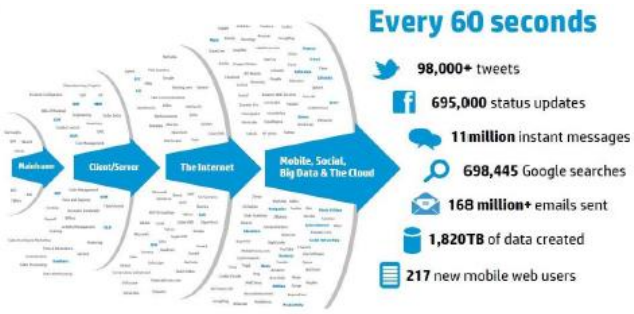

Figure 4: Examples of big data velocity

Figure 4 shows few examples of the pace the data. Data speed administration is significantly more than a bandwidth issue; it is additionally an ingest issue Kaisler et al. (2013).Figure 2 also reflects velocity as a characteristic of big data, showing how it requires near real-time and/or real-time analytics.

Variety: Other than typical structured data, big data contains text, audio, images, videos, and many more unstructured and semi-structured data, which are available in many analog and digital formats. From an analytics perspective, variety of data is the biggest challenge to effectively use it. Some researchers believe that, taming the data variety and volatility is the key of big data analytics (Infosys 2013).Figure 5 shows the comparison between increment of unstructured, semi-structured data and structured data by years. Figure 2 also reflects the increment in verity of data.

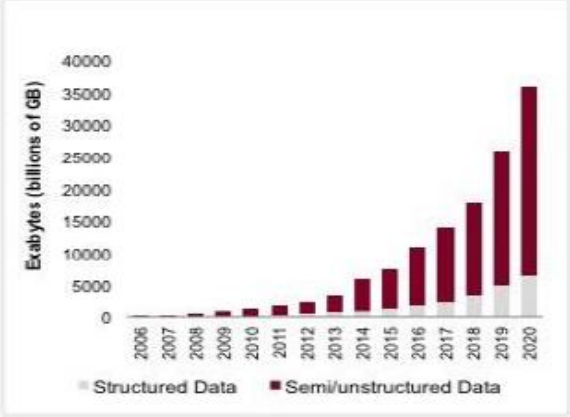

Figure 5: Growth of data variety by years

One of the big data vendors, IBM has coined additional V for the big data characteristics, which is veracity. By veracity, they address the inherent trustworthiness of the data. As big data will be used e.g. for decision making, it is important to make sure that the data can be trusted. Some researchers mentioned 'viability' and 'value' as the fourth and the fifth characteristics leaving 'veracity' out (Biehn 2013).

2. How Fast Data is Increasing

Companies like Google, Facebook ,Twitter, Skype and so on generated data every 60 second ,so by this we can understand how much data being generated in a second, a minute, a day or a year and how exponentially it is generating. As per the analysis by TechNewsDaily we might generate more than 8 Zettabytes of data by 2015.

Figure 6: Big Data generated by companies

3. The Importance of Big Data

We are in a new era in modern information technology - the "Big Data" era. In March, 2012, the U.S. Government announced a "Big Data Research and Development Initaitve" -- a $200 million dollar commitment to improve our ability to "extract knowledge and insights from large and complex collections of digital data." Government agencies such as NSF, NIH, and DOD are investing hundreds of millions of dollars toward the development of systems that can help them extract knowledge from their data. The career potential for our graduates continue to blossom in this field. A recent study released by Gartner projects that in 2013, "big data is forecast to drive $34 billion of IT spending," with a total of $232 billion to be spent through 20162. In another report, they estimate that "by 2015, 4.4 million IT jobs globally will be created to support big data" with the US generating 1.9 million of those jobs. However, as numerous sources have suggested in recent years, despite the rapid increase in opportunities for careers in big data, there is a dearth of talent.

4. Big Data Technologies

Conventional data technologies and methods are most of the time slow, expensive and not suitable to handle the storage and the processing of large growing volumes of heterogeneous data. One challenge is to overcome the complex nature of Big Data (volume, velocity and variety). Another challenge is how to efficiently display the changing"Big insight" for many connected entities depending on their roles? Besides that, Big Data actors are facing other challenges when securing Big Data, such as how to secure huge evolving data sets? How to integrate security layers without affecting the performance of systems? How to integrate privacy policies into common Big Data platforms while providing rapid and granular access to data? Research communities from different sectors have been struggling to develop new, fast, dynamic and user-friendly technologies for Big Data. Nowadays, many open source and proprietary Big Data solutions are available. The goal is to helps decision makers and data scientists to take the next best actions based on discovered patterns, data relations and newly extracted knowledge from Big Data.We presents hereafter some solutions developed to overcome Big Data challenges at different levels.

4.1 Big Data Frameworks and Platforms

Several MapReduce frameworks (e.g., Apache Hadoop, Skynet, Sailfish and FileMap) were developed to handle structured and unstructured massive data. Indeed, they allow to store and process large volumes of immutable data (like logs or large binary objects) as well as incrementally collected data (like web crawls, user comments on social networks, GPS data or sensors events). They are efficient for many use cases (such as log file analysis, scientific simulations or financial predictions). Such frameworks are based on many solid concepts including the following:

- Distributed storage: most of Big Data platform, like Hadoop and Disco, are based on distributed storage of data. Unlike traditional systems, they store blocks of very large files across multiple nodes. They are designed to run on low-cost hardware and provide a high streaming access to data sets (B. Lublinsky and A. Yakubovich 2013) .
- Massive Parallel Processing (MPP): the multiple time consuming tasks, of Big Data applications, are processed in parallel across several servers. MPP helps to avoid copying distant data to execute computations. It executes jobs where data are stored in order to minimize network congestion and to ensure a fast processing.
- Fault tolerant and scalable system: to ensure no point of failure, Big Data systems are usually based on a reliable Master-Slave architecture and data replication (such as Hadoop) or peer recovery concept (such as Skynet (Skynet 2008) .

Most solutions offered the possibility to add clusters and components to handle more data and massive processing. Most of the time, free and proprietary components are built on the top of open source Big Data platforms such as the case of Hortonworks Data Platform (HDP) (docs.hortonworks.com) and IBM InfoSphere BigInsights (Ibm infosphere biginsights) . The goal is to offer complete Big Data ecosystem solutions. Indeed, they aim to leverage MapReduce frameworks, to simplify their use and to extend their capabilities. Many Big Data solutions and tools are available as presented hereafter:

- Tools for Big data integration (such Sqoop, Flume and DataLoader);
- Tools to manage resources, workflows and services (such Cloudera Manager, Yarn, Oozie and Zookeeper);
- Tools to handle metadata services (such Hcatalog);
- Tools for data analytics (such R programming language, Mahout, Chukwa and Teradata Analytical Ecosystem);
- Tools for interactive search and native querying (such Cloudera search, Sphinx Search Server and Facebook Unicorn M. Curtiss et al. (2013) that is an online in-memory indexing system designed to search trillions of edges between tens of billions of entities on thousands of commodity servers);
- Tools for data visualization (such Advisor, Visual Analytics and Centrifuge). With those multiple technologies, managers can rely on rapid, cost-effective, fault-tolerance, scalable and user friendly solutions (A. Cardenas and S. Rajan 2013) Managers can choose a complete solution or add as needed components to the existing infrastructures. For example, Oracle Big Data Apliance (Oracle white paper 2013) combines, in one system, the power of optimized industry-standards hardware, Oracle software experience as well as the advantages of Apache Hadoops open source components.

4.2 Databases for Big Data

A. NoSQL

RDMS (Relational Database Management System) requires to structure data in a defined formats, which is not adequate in Big Data context where rapid and huge volumes of unstructured data are generated. To face this challenge, several NoSQL databases have been developed to handle non-relational and unstructured data like HBase, Cassandra, DynamoDB, MongoDB, Riak, Redis, Accumulo, Couchbase and so on. They support one or more data models including: key-value pairs, document oriented (such JSON, BSON, XML, HTML documents), graphs (designed for highly connected data), wide-columns and geospatial data. The NoSQL databases provide a cheaper way (than RDMS) to handle the storage and the management of Big Data in distributed environment. Such databases offer different levels of fault-tolerance and data availability.

B. NewSQL

NoSQL databases have many downsides that pushed the development of NewSQL databases. In fact, they usually do not support indexing and SQL querying. They are also often slow in handling large queries. In addition, unlike RDBMS, they do not ensure ACID principles for reliable transactions. To address these limitations, the NewSQL was developed for Big Data applications. It constitutes a new relational database management systems based on a distributed architecture (Marijana 2013).NewSQL databases merge the best of both precedent technologies the RDBMS and the NoSQL. NewSQL provide ACID properties and SQL querying. It enables also to ensure good data availability and performance of online transaction processing.

C. Searching and Indexing

Traditional searching methods are not adapted to distributed environment and Big Data complexity. Enterprises need to run extensive real-time queries through huge volumes of unstructured and structured data sets. This demand have

led to the development of scalable Search Engines based on appropriate searching and indexing technologies such Lucene and Splunk Processing Language .

5. Data Storage

Relational database management systems (RDBMSs) are traditional storage systems designed for structured data and accessed by means of SQL. RDBMSs are facing challenges in handling Big Data and providing horizontal scalability, availability and performance required by Big Data applications. In contrast to relational databases, MapReduce provides computational scalability, but it relies on data storage in a distributed file system such as Google File System (GFS) or Hadoop Distributed File System (HDFS).NoSQL and NewSQL data stores have emerged as alternatives to Big Data storage. NoSQL refers to "Not Only SQL", highlighting that SQL is not a crucial objective of those systems. Their main defining characteristics include schema flexibility and effective scaling over a large number of commodity machines. NoSQL horizontal scalability includes data storage scaling as well as scaling of read/write operations K. Grolinger et al. (2013). Analyze features driving the NoSQL systems ability to scale such as partitioning, replication, consistency, and concurrency control. NoSQL systems typically adopt the MapReduce paradigm and push processing to the nodes where data is located to efficiently scale read operations. Consequently, data analysis is performed via MapReduce jobs. MapReduce itself is schema-free and index-free; this provides great flexibility and enables MapReduce to work with semi-structured and unstructured data. Moreover, MapReduce can run as soon as data is loaded. However, the lack of indexes on standard MapReduce may result in poor performance in comparison to relational databases. This may be outweighed by MapReduce scalability and parallelization. Database vendors, such as Oracle, provide in-database MapReduce (X. Su and G. Swart 2012), taking advantage of database parallelization. Another example of providing analytics capabilities indatabase is the MAD Skills project J. Cohen et al. (2009) which implements MapReduce within the database using an SQL runtime execution engine. Map and Reduce functions are written in Python, Perl, or R, and passed to the database for execution. NoSQL systems from column-family and document categories adopt the MapReduce paradigm while providing support for various indexing methods. In this approach MapReduce jobs can access data using the index, therefore query performance is significantly improved. For example Cassandra supports primary and secondary indexes (Apache Cassandra) . In CouchDB (J. C. Anderson and N. Slater 2010) the primary way of querying and reporting is through views which use the MapReduce paradigm with JavaScript as a query language. A view consists of a Map function and an optional Reduce function. Data emitted by Map function is used to construct an index and consequently, queries against that view run quickly. Another challenge related to MapReduce and data storage is the lack of a standardized SQL-like language. Therefore one direction of research is concerned with providing SQL on top of MapReduce. An example of this category is Apache Hive A. Thusoo et al. (2009) which provides an SQL-like language on top of Hadoop. Another Apache effort, Mahout (Apache Mahout), aims to build scalable machine learning libraries on top of MapReduce. Although those efforts provide powerful data processing capabilities, they lack data management features such as advanced indexing and a sophisticated optimizer. NoSQL solutions choose different approaches for providing querying abilities K. Grolinger et al. (2013): Cassandra and MongoDB provide proprietary SQL-like querying while HBase uses Hive. It is important to point out the efforts on integration between traditional databases, MapReduce, and Hadoop. For example, the Oracle SQL connector for HDFS (Oracle 2014) provides ability to query data in Hadoop within the database using SQL. The Oracle Data Integrator for Hadoop generates Hivelike queries which are transformed into native MapReduce and executed on Hadoop clusters. Even though the presented efforts advanced the state of the art for Data Storage and MapReduce, a number of challenges remain, such as:

- The lack of a standardized SQL-like query language,
- limited optimization of MapReduce jobs,
- Integration among MapReduce, distributed file system, RDBMSs and NoSQL stores.

6. Big Data Management and Storage

In Big Data Big means the size of data is growing continuously, on the other hand increasing speed of storage capacity is much less than the rising amount of Data. The reconstruction of available information framework is needed to form a hierarchical framework because Researchers has come up with the conclusion that available DBMSs are not adequate to process the large amount of data Changqing et al. (2012) .Architecture commonly used for processing of data uses the database server, Database server has constraint of scalability and cost which are prime goals of Big Data. A different business model has been suggested by the providers of database but basically those are application specific forget. Google seems to be more interested in small applications . Big Data Storage is another big issue in Big Data management as available computer algorithms are sufficient to store homogeneous

data but not able to smartly store data comes in real time because of its heterogeneous behaviour Avita Katal et al. (2013) .So how to rearrange Data is another big problem in context of Big Data Management. Virtual server technology can sharpen the problem reason is it raises the issue of overcommitted resources specially when there is lack of communication between the application server and storage administrator. Also need to solve the problem of concurrent I/O and a single node master /slave architecture.

7. Big Data Analysis and Management

Big data analytics is differences from traditional analytics Because of the big increase in the volume of data and that led to Many researchers have suggested commercial DBMS and this not suitable with size of data. This type of data is impossible to handle using traditional relational database management systems. New innovative technologies were needed and Google found the solution by using a processing model called MapReduce. There are more solutions to handle Big Data, but the most widely-used one is Hadoop, an open source project based on Google's MapReduce and Google File System. Hadoop was founded by the Apache Software Foundation. The main contributors of the project are Yahoo, Facebook, Citrix, Google, Microsoft, IBM, HP, Cloudera and many others. Hadoop is a distributed batch processing infrastructure which consists of the Hadoop kernel, Hadoop Distributed File System (HDFS), MapReduce and several related projects.

7.1 Big Data with Data Mining

Data Mining is commonly defined as the technique to extract useful knowledge from database M. Chen et al. (1996). It is almost impossible to derive the value directly from each data. For this reason, data mining needs pre-processing and analytic method for finding the value. Indeed, data mining is closely related with artificial intelligence and machine learning and so on. Scale of data management in data mining and big data is significantly different in size. However, the basic method to extract the value is very similar. In case of data mining, the process of extracting knowledge needs data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation et.. Big data came out after solving the requirements and challenges of data mining. Requirements and challenges are 'handling of different types of data', 'efficiency and scalability of data mining algorithms', 'mining information from different source of data'.

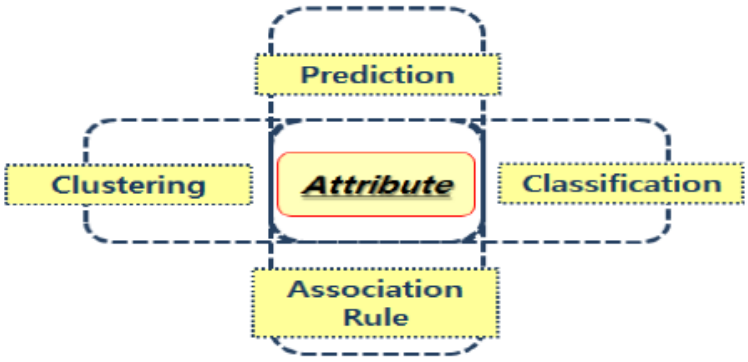

Figure 7: Data mining techniques

7.2 Big Data over Cloud Computing

Cloud computing is usually defined as a type of computing that relies on sharing pooling computing resources rather than having local servers or personal devices to handle applications Divyakant et al. (2011) . The present advances like cloud computing stage and network ,have all proposed to get to immense measures of processing assets (programming ,equipment, application) and that offering in a single framework view. Among these technologies, cloud computing is turning into a capable structural planning to perform substantial scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. Besides , the principle objective of cloud computing is to deliver computing as an solution for handling enormous information, as high dimensional information sets , vast size and multi-media CHANGQING et al. (2012) . There are several leading Information Technology arrangement suppliers that offer these services to the clients.Presently, logically after the idea of the big data came up, cloud computing service model is by degrees moving into big data service model, which are AaaS (Analysis as a Service) DBaaS (Big data as a Service) and DaaS (Database as a Service).

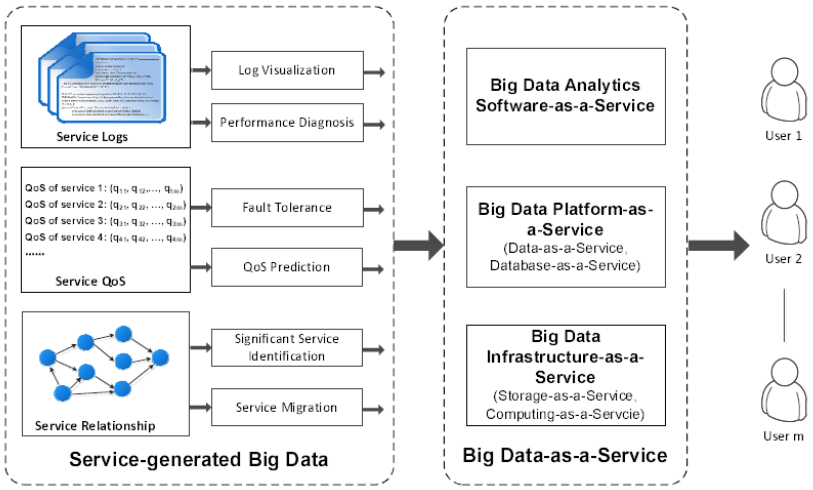

Figure 8: Big Data as Service

7.3 Hadoop Distributed File System (HDFS) and MapReduce

Hadoop comes with its default distributed file system which is Hadoop distributed file system (HDFS) Amrit Pal et al. (2014). It stores file in blocks of 64 MB. It can store files of varying size from 100MB to GB, TB. Hadoop architecture contains the Name node, data nodes, secondary name node, Task tracker and job tracker. Name node maintained the Metadata information about the block stored in the Hadoop distributed file system. Files are stored in blocks in a distributed manner. The Secondary name node does the work of maintaining the validity of the Name Node and updating the Name Node Information time to time. Data node actually stores the data. The Job Tracker actually receives the job from the user and split it into parts. Job Tracker then assigns these split jobs to the Task Tracker. Task Tracker runs on the Data node they fetch the data from the data node and execute the task. They continuously talk to the Job Tracker. Job Tracker coordinates the job submitted by the user. Task Tracker has fixed number of the slots for running the tasks. The Job tracker selects the Task Tracker which has the free available slots. It is useful to choose the Task Tracker on the same rack where the data is stored this is known as rack awareness. With this inter rack bandwidth can be saved. Figure 9 shows the arrangement of the different component of Hadoop on a single node. In this arrangement all the component Name Node, Secondary Name Node, Data Node, Job Tracker, and Task Tracker are on the same system. The User submits its job in the form of MapReduce task. The data Node and the Task Tracker are on the same system so that the best speed for the read and write can be achieved.

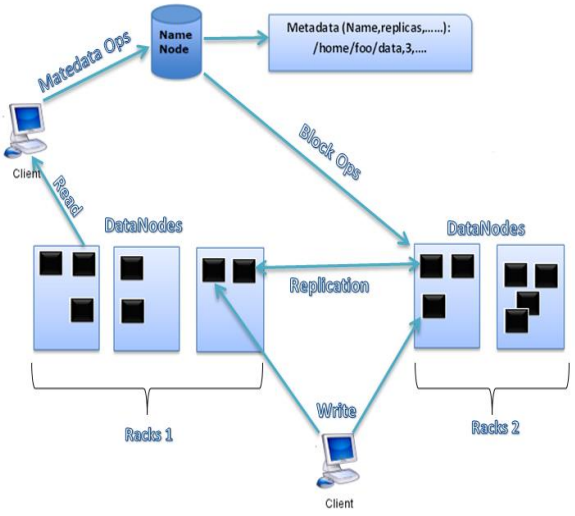

Figure 9: Hadoop Distributed File System (HDFS)

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. It provides a programming paradigm which allows useable and manageable distribution of many computationally intensive tasks. As a result, many programming languages now have Map-Reduce implementations which extend its uptake. On the other hand, Hadoop is a highly popular free Map-Reduce implementation by the Apache Foundation (White T 2012). With the popularity of the Hadoop applications there have been many complementing applications developed by the open source community and packaged up under apache foundation Saecker et al. (2013). Map-Reduce involve two main parts.

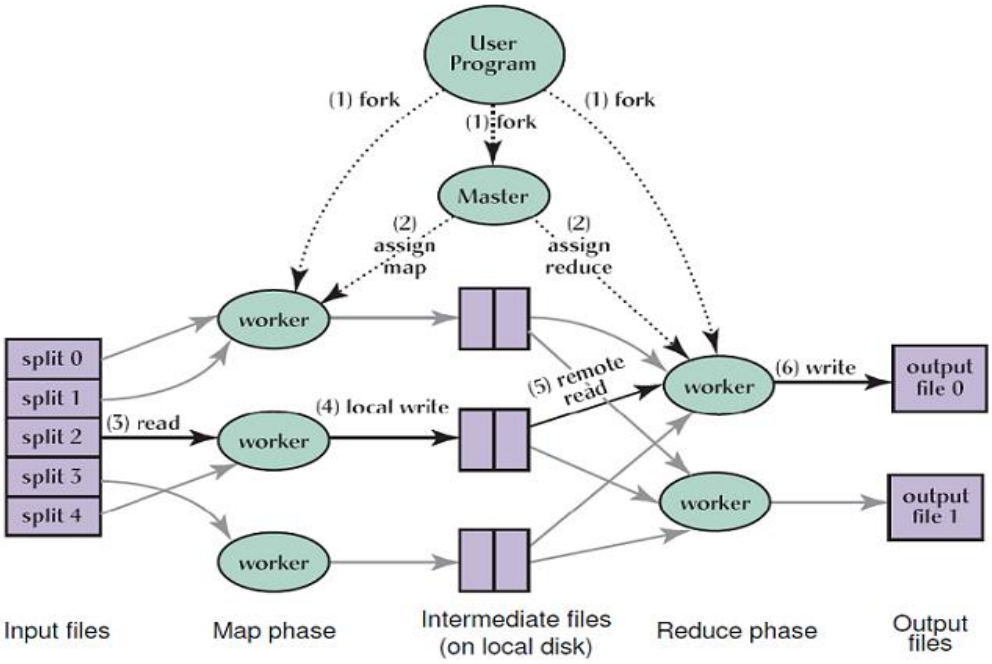

Figure 10: MapReduce Architecture

- Map operation

Where a simple function is used to emit key/value pairs in parallel similar to using primary keys in the relational database world. Once the data to be processed is mapped into key/value groups.

- Reduce operation

Is used to apply the core processing logic to produce results in a timely manner McCreadie et al. (2012). The simple concept of Map-Reduce removes many traditional challenges in HPC to achieve fault tolerance and availability. Therefore, it paves the way for development of highly parallel, highly reliable and distributed applications on large datasets.

8. Big Data Modelling

Although distributed data analysis platform may offer a solution to deal with data of great scale, the data based modelling in big data environment still remains to be a challenge. There are two possible solutions to the big data modelling problem: (a) design a deep learning modelling algorithm, making use of the strong ability of machine learning to process massive high-dimensional data; (b) divide the entire dataset into subsets, based on which submodels can be built, and then obtain the entire model by integrating all the sub-models according to specific strategies Wei chang Kong et al. (2014).

9. Big Data Security Issues

As increasing use of big data and expanding scope of big data, big data security has been considered crucial. There are many security issues about big data. Among them, data protection and access control are recognized as the most important security issue. This is similar to the current information security situation. However, data management

and classification for security are more difficult than current information security issues due to the volume of data Sung-Hwan Kim et al. (2013). For this reason, Management Cost per GB has decreased but security investment for big data has increased. Similarly, access control is more difficult due to huge data scale. As mentioned earlier in the introduction, value is the key deliverable of big data. The data itself is not the subject of protection. In addition, securing the entire data is very inefficient, considering the volume of big data.

10. Conclusion and Future work

Big data provides enterprise with more choices because of its lots of related technologies and tools, which will continue to be developed and become innovative hotspots in the future, such as Hadoop distribution, the next generation of data warehouse, advanced data visualization, etc. In recent years, academia pays more attention to cloud computing. Big data focuses on "data", like data service, data acquisition, analysis and data mining, which pays more attention on ability of data storage. Cloud computing focuses on computing architecture and practices. Big data and cloud computing are two sides of the same issue .It is more accurate to analyze and forecast big data by using cloud computing and release more hidden value of data; in order to meet the service demand of big data, we can find even better practical application to the cloud computing. Nowadays, more and more enterprises hope that they can transfer their own applications and infrastructures to the cloud platform. Cloud computing brings great changes to the big data. First, cloud computing provides a quite cheap storage place for the big data and makes medium-sized and small enterprises complete big data analysis. Second, cloud computing has huge IT resources, distributes widely and becomes an effective way for enterprises which have more heterogeneous system to process data accurately. Although this paper clearly has not resolved the entire subject about this substantial topic, hopefully it has provided some useful discussion and a framework for researchers.

ACKNOWLEDGMENT

The authors would like to thank the University Malaysia Pahang for funding this study through the grant no:RDU 1403163.

References

Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy , " Hive - A Warehousing Solution Over a Map-Reduce Framework", VLDB '09,August 24-28, Lyon, France, , 2009 .

Apache Cassandra, http://www.datastax.com/docs.

Amap-reduce framework," Proc. of the VLDB Endowment, 2(2), pp.1626-1629, 2009.

Apache Mahout, https://mahout.apache.org/.

A. Cardenas, P. Manadhata, and S. Rajan, "Big data analytics for security," Security Privacy, IEEE, vol. 11, no. 6, pp. 74–76, Nov 2013.

AT Kearney,"Big Data and the Creative Destruction of Today's Busniness Model" 2013.

Avita Katal, Mohammad Wazid and R H Goudar," Big Data: Issues, Challenges, Tools and Good Practices", 978-1-4799-0192-0/13, IEEE , 2013.

Avita Katal, Mohammad Wazid and R H Goudar ,"Big Data: Issues, Challenges, Tools and Good Practices" , IEEE , 978-1-4799-0192-0/13 , 2013 .

Amrit Pal et al , "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop" International Conference on Communication Systems and Network Technologies IEEE , 978-1-4799-3070-8/14 , 2014.

Big Data Infographic: Solve your Big Data Problems?, http://www.intel.in/content/www/in/en/big-data/solving-big-dataproblems-infographic.html.

Beyer and Laney,"The Importance of 'Big Data': A Definition"Gartner, 2012.

B. Lublinsky, K. T. Smith, and A. Yakubovich, Professional Hadoop Solutions. John Wiley & Sons, 2013.

"Skynet." [Online]. Available: http://skynet.rubyforge.org/doc/ 2008.

Biehn, "The Missing V's in Big Data: Viability and Value " http://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/.

Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada and Keqiu Li ,"Big Data Processing in Cloud Computing Environments" International Symposium on Pervasive Systems, Algorithms and Networks ,2012.

CHANGQING JI et al," BIG DATA PROCESSING: BIG CHALLENGES AND OPPORTUNITIES", Journal of Interconnection Networks Vol. 13, Nos. 3 & 4, 2012.

Conference on Management of Data, 2012.

J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, "MAD skills: New analysis practices for Big Data," VLDB.

Doctorow, “Big Data: Welcome to the petacentre. Big Data: Science in the petabytes era “, pp. 16-21. 2008.
Divyakant Agrawal, Sudipto Das and Amr El Abbadi, “Big Data and Cloud Computing: Current State and Future Opportunities” EDBT, Uppsala, Sweden, 978-1-4503-0528-0/11/0003, 2011.
Endowment, 2(2), pp. 1481-1492, 2009.
J. C. Anderson, J. Lehnardt and N. Slater, CouchDB: The Definitive Guide, Sebastopol, CA, USA: O'Reilly Media, 2010.
Google. Google Trends 2013.
Hunter,”Journey to the centre of Big Data. Engineering and Technology,” IEEE Magazine, pp. 56-59,27 April, 2013.
Infosys,”Big Data: Challenges and Opportunities” ,2013.
Ibm infosphere biginsights.[Online]. Available: http://www- 01.ibm.com/.
Kaisler, S., Armour, F., Espinosa, J. A. & Money, W. “Big Data: Issues and Challenges Moving Forward” Wailea, Maui, HI, s.n., pp. 995 - 1004. 2013.
K. Grolinger, W. A. Higashino, A. Tiwari and M. A. Capretz, "Data management in cloud environments: NoSQL and NewSQL data.
M. Chen, J. Han and P.S. Yu, “Data mining: An overview from a database perspective “ , knowledge and data Engineering, IEEE , 1996.
Madden, “From Databases to Big Data”,Internet Computing, IEEE, 16(3), pp. 4 - 6. 2012.
McCreadie R., Macdonald C., Ounis I., "MapReduce indexing strategies: Studying scalability and efficiency", Journal of Information Processing and Management: an International Journal archive, 48 (5). pp. 873-888. ISSN 0306-4573. 2012.
M. Curtiss, I. Becker, T. Bosman, S. Doroshenko, L. Grijincu, T. Jackson,S. Kunnatur, S. Lassen, P. Pronin, S. Sankar, G. Shen, G. Woss, C. Yang, and N. Zhang, “Unicorn: A system for searching the social graph,” Proc. VLDB Endow., vol. 6, no. 11, pp. 1150–1161, Aug. 2013.
Marijana, “Newsql: Handling big data in the enterprise,”. [Online]. Available: http://bizcloudnetwork.com/newsql-forenterprise- big-data/ November 19, 2013 .
Nibedita Chakraborty and Sandeep Gonnade,“Big Data and Big Data Mining: Study of Approaches, Issues and Future scope “International Journal of Engineering Trends and Technology (IJETT) – Volume 18 Number 5 – Dec 2014 .
Oracle white paper: Big data for the enterprise. [Online]. Available: http://www.oracle.com 2013.
Oracle Big Data connectors, http://www.oracle.com/us/ 2014.
PC Magazine, Definition of Big Data. http://www.pcmag.com/encyclopedia/term/62849/big-data.
Product documentation.[Online]. Available: docs.hortonworks.com/.
Sicular,”Gartner's Big Data definition consists of three parts” Forbes, 27 March, 2013.
Saecker M. and Markl V., "Big Data Analytics on Modern Hardware Architectures: A Technology Survey", Springer Lecture Notes in Business Information Processing, Volume 138, pp 125-149. 2013.
Sung-Hwan Kim, Nam-Uk Kim and Tai-Myoung Chung, “Attribute Relationship Evaluation Methodology for Big Data Security”,978-1-4799-2845-3/13/$31.00, IEEE, 2013.
stores," Journal of Cloud Computing: Advances, Systems and Application, 2, 2013.
X. Su and G. Swart, "Oracle in-database hadoop: When MapReduce meets RDBMS," Proc. of the ACM SIGMOD International, 2012.
White T, "Hadoop: The Definitive Guide", Third Edition, O'Reilly, 978-1-449-31152-0, May 2012.
Wei chang Kong, Qidi Wu , Li Li and Fei Qiao ,” Intelligent Data Analysis and Its Challenges in Big Data Environment”, IEEE International Conference on System Science and Engineering (ICSSE) “,978-1-4799-4367-8/14,2014.
Wikipedia, Big Data. http://en.wikipedia.org/wiki/Big_data
Weathington, J., “Big Data Defined. TechRepublic”, 3 September, 2012.