Big Data and Cloud Security

Sakshi Dewangan*
Rajesh Singh Negi**

Abstract

In today's face paced technologically mingled life, everybody has so much information that they do not know how to properly access it or the fact whether it is even worth keeping. These calls for a proper analytics mechanism that helps to keep all data sorted be it from multiple sources. The heap of information present today is most sitting in raw format as the proper means and tools to analyze and use the information efficiently had been missing. This is where Bigdata comes into action. It helps in data-driven decision making. Big Data are high volume, high velocity and high variety information asset that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. It is an intelligent analyses tool that can process data that exceeds our current capability. Cloud computing in other hand provides services on demand rather than product. Big data integration with cloud is the future of IT sector.

In this paper we will discuss we will discuss Big data its advantages with cloud. Issues related to cloud security and Hadoop. Some possible approaches to reduce the risks that can make these technologies more efficient in terms of practical usage.

Keywords: Bigdata, analysis, file system, database, risk, map reduce, Hadoop, internet, cloud computing, security, privacy, encryption.

I. Introduction

The traditional system of storing data used flat files which are maintained by the file system and regulated by the Operating System. File systems were an attempt to computerize the manual filing system. When a user wants to store some data in the computer, he must place the data in files. Files are further placed in directories that are located in specific areas of the hard disk. A number of directories can be created with provision of a directory contained in another directory. User has a number of options such as renaming, deleting files, etc...Application programs use these files to access the data stored in them. This allowed a user even with little knowledge to perform operations such as deletion, insertion, manipulation and retrieval of data as required. These flat files store data in the form of records. A delimiter is used to separate the records such as a comma, space, pipe or any other special character. A special predefined character is used to mark the end of the file. Despite its ease, file system had various problems such as redundancy, security breaches; high cost of storing the data, data inconsistency, more memory was required due to redundancy of data. Also, data in file systems was scattered around multiple files. This made access and analysis of data a troublesome task. Changing one part of the data meant changing the application to suit the modified file format since applications were specific to the file system. Thus, a need for more systematic and structured arrangement of data aroused.

To solve the problem of unstructured data, the concept of databases emerged. These databases are built on top of the data storage services provided by file systems. A database is a collection of a large number of file systems and other sources of data that stores the data in a structured format for use by the user. A database is a computer database that is an up to date repository of information that can belong to either a specific organization or may be placed on the World Wide Web. The typical definition of a database is that it is an organized collection of data that are modeled in such a way that the data can be used in real aspects of life. For example, a user has data about the number of owners in a specific locality who own a dog, now he wants to find the breed of dog that is preferred by most owners. So he can analyze the data present in the database and retrieve results on the basis of his query.

Sakshi Dewangan*
Management Education & Research Institute
New Delhi

Rajesh Singh Negi**
Management Education & Research Institute
New Delhi

This query analyses purpose is served by the Database Management System also known as DBMS. The general definition of DBMS holds that Database management systems (DBMSs) are computer software applications that interact with the user, other applications, and the database itself to capture and analyze data. A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases.[1] Databases are of multiple types out of which relational DBMS are most commonly used. It stored data in the form of tables allowing the user to form various relationships among various tables .i.e. a logical connection exists between the tables in the database. It is the collection of schemas, tables, queries, reports, views and other objects. Some of the well known databases include MySQL, PostgreSQL, Microsoft SQL Server, Oracle, Sybase and IBM DB2. Databases eliminate the shortcomings that are faced in the traditional file system. It provides a higher level of service as compared to the traditional file systems. Databases use sophisticated protocols and algorithms to implement reliable data storage on top of unreliable file systems. These algorithms make database storage a little expensive than file system storage, but databases provide far greater security than file systems. Relational database systems are based on Codd's Rules which are a set of thirteen rules beginning from zero that define the basic standard for a relational database. Some of them are logical data independence, physical data independence, data integrity, information rule, etc...

When the amount of data is a database exceeds our current capability of processing data, it is termed as Big Data. As stated, Big Data are high volume, high velocity and high variety information asset that require new forms of processing to enable enhanced decision making, insight discovery and process optimization[2].The size of big data is measured in terabytes or petabytes or in higher memory units.

II. What is Big Data and Importance

When the data in an organization becomes so large and complex that it cannot be handled by traditional tools we say it as Big data. It basically concerns large-volume of data which is very complex [7]. We can define Big data with its for V's-

- Velocity
- Volume
- Variety

Volume

Data storage in now days is growing exponentially as data now is not just text data.

We all are familiar with Facebook and other social networking websites where data is not just a text there data is in different formats like videos, music etc. Therefore the data storage of an organization is way

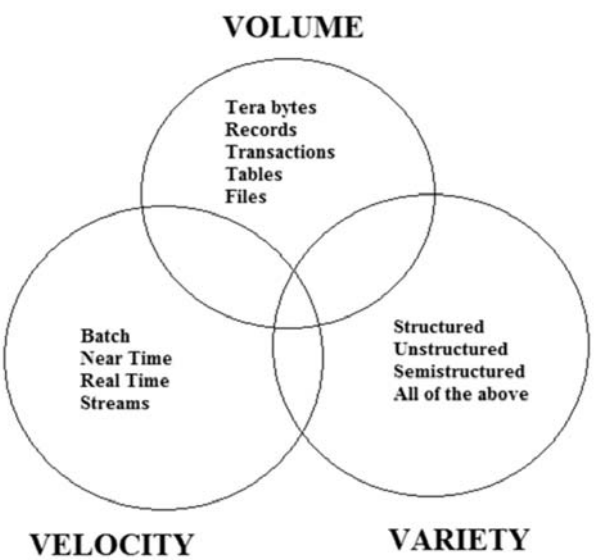

Figure 1: three V's of Big data.

beyond gigabytes and terabytes. Data here not only generated daily but the historical data is also included for analysis. In 2012- 2.5 Exabyte's of data is created each day.
And this is doubling every year.

Velocity
This means how fast the data is coming from different resources. This concept is used for real time data where the data creation is very high speed which includes each click from the button, broadcasting the real time information.

Variety
Data can be stored in different formats .for example we can data in different formats like excel file, word file, text file which are the traditional formats. It is not necessary that data is present in these traditional formats it may also be in the forms like pdf, video, music or on something new formats. As we are in the world of social networks we may have data in the form of events. Now it is necessary for the organization to organize it in a meaningful way.

A. Sources of big data

- Sensor data- It is high velocity data which comes from sensors distributed over a geographical location like geo-location, temperature, noise, pollution, biometrics, red-lights etc.
- Machine log data-Machine data consist of records of all the activity and behavior of the user using the machine. It is used for identifying the third party services.

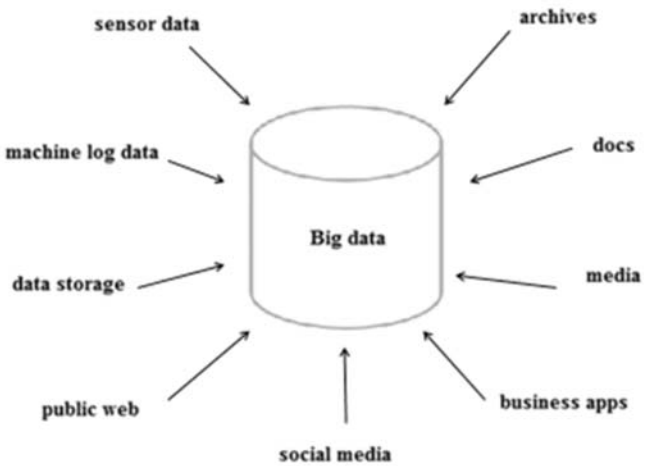

Figure 2: sectors where Bigdata is used

- Data storage- data from different storage location SQL, NoSQL, file systems.
- Public web-it is external where other people can use it on regular basis. Here we can integrate web applications which can generate lots of data.
- Social web-High velocity and high volume data that can be used to detect trends and analyse those trends for a specific brand. It is also used to target a set of customers to social accounts.
- Business Applications- These are the apps that uses API's which can pull data from both inside and outside your organization.
- Media- is connected in and out from the organization, connected with API's.

B. Importance of big data

Data is very important for an organization it serves as a backbone for them. In today's era there is an explosion in data rate and we are trying to store all of this.

Why there is a need to store this much of data?

After capturing the big data the data is analyzed and processed. Companies after analyzing it gains complete understanding of their business , their customers and their competitors. By analyzing such information the companies can check their strategy and improvement in it or in case if analysis is not so good for a particular product then it may result in a new marketing strategy which eventually helps in near future.

All the analysis results in-

- Efficiency improvements
- Increased sales
- Lower costs
- Better customer service
- Improved products and services.

If we take an example of a Company, who manufactures shoes. The company should know who buys his products and who does not. From where the company could get this sort of data here company can use social media and web log files from the ecommerce sites that can help company in two ways-

1. Revaluation of strategy- case who didn't buy the product in this case company can make modifications in their product or strategy to attract them.

2. Target customers- case who buy the product in this company will target particular set of customers to increase their sales

III. Hadoop

Hadoop is a processing engine which can handle large volumes of data in any structure. Hadoop provides two things-

1. Distributive storage of data (HDFS)
2. Distributive processing of large data sets (Map Reduce algorithm)[11]

. HDFS- Hadoop distributive file system

Distributive storage of data is called Hadoop Distributed File System, or HDFS. It provides cheap storage of data with low fault-tolerance. It also sees the hardware failure and in case of failure tried to use different node to save the data by replicating it therefore if one fails other node can be used. HDFS stores files in different servers , files are divided into blocks and saved to more than one servers. This helps in reduction of disk failures and performance. HDFS keeps an eye on every server and the blocks that they maintains this ensures continuous data availability. If we want to read some data we request for a block checksum of block is checked if it is found damaged then other server is looked for the same block. It lets the organization to spend less money for looking the servers; it's the software that looks for the servers now.

Map Reduce

Map reduce consist job scheduler a software component. Job of job scheduler to choose the server for user query (job) it also schedules multiple user jobs on cluster of servers. It takes care of distributed computing [10].

Hadoop map reduce is an open source implementation Google Map reduce algorithm [8].Hadoop Map reduce consist of two functions- map and reduce both are user defined. Input to Hadoop Map reduce is a key value pair (key, value) and then a map function is called for each set of pair. The function produces zero or more intermediate values (key', value')., then these intermediate keys are grouped and reduce functions are called which produces aggregate results. Users have to define only the input and reduce functions rest is on Map Reduce. It uses HDFS to read and write the data[9].However it an reads the data from other sources like local file system, web and databases.

The advantage of using Hadoop MapReduce is job optimization. It allows anon-expert user to efficiently use Big data for analysis. The user need not to know

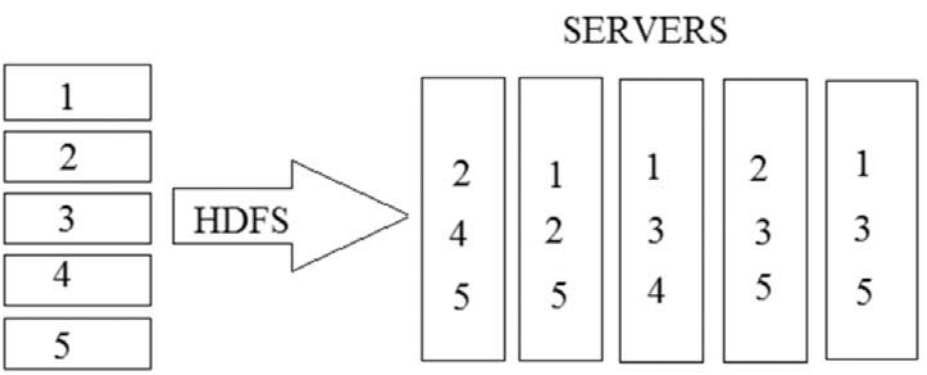

Figure 3: HDFS breaking of files.

SQL for queries only some knowledge of java is required [12].

Advantages-

- Scalability- the data is stored and distributed across number of servers which can operate in parallel.
- Flexibility- organizations can introduce new data sources and operations can be done of any type of data structured or unstructured.
- Fast-Hadoop stores data in distributed manner and MapReduce algorithm uses key value pair as an input which enables faster processing to locate the data.

IV. Cloud Computing

If we talk about cloud computing it is a kind of technology that depends on the sharing of computer resources. It basically delivers the services through INTERNET. Primary goal of cloud computing is to reduce the investment cost for hardware and software, to increase the scalability as it provides everything on demand and the resources on cloud are always available and reliable [14]. Cloud computing consist of computers connected to network that handles the load. The main benefit of cloud computing is to eliminate the cost at users end.

User only required having a computer and simple software to access the cloud services rest is handled by the cloud. The user can put any king of data in the cloud and data in the cloud is safe from any damage and the user can access that data any time any place he or she just needs a INTERNET connection.

V. Benefits of Big Data Anayltics

The main importance of big data is only in the field of analysis. Analytics is a broad term for data analysis application. Stored data holds no value if not efficiently utilized. Stored data cannot generate business values. Once the data is stored such as in Hadoop , it can be analyzed and that in turn gives it a tremendous value.Diverse analysis technologies exist in the market today that can be used for big data application such as in-memory analytics, in-database analytics, and appliances.Analysis is a broad term that is used with Big Data in different ways. Analytics can mean either getting data in or getting data out. Proper analysis of data helps in decision making [4]. This was commonly used for business intelligence. Business intelligence is a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions. This term was more popular in the 1990's. Today, it has been replaced with analytics since 2010. Analysis can be predictive, explorative or descriptive depending upon the purpose of data analysis.

Following are the features of Bigdata :

- Easy analysis of information from multiple sources that otherwise have no meaning.
- Bigdata is timely that means that workers are working hard to manage the data and make decisions.
- Big data is trust worthy. Data accumulated from multiple sources help in identification of exact patters. This data is more reliable than the one performed manually by workers.
- Big Data is Secure
- Big Data is Relevant -most of the companies are not happy with the way their filtering applications work. Thus they turn to big data.
- Big Data is Actionable
- Big data provides ample of opportunities for scratch companies to enter into the market.

Some of the benefits of Big data analysis in different sectors are:

- Big data analytics benefits business intelligence, customer relations and many analytic applications. For example, big data is most commonly used to identify the buying patterns of customers. The products that customers view, order or just visit are used by vendors to gain knowledge about the likings of the customer. Based on this, only relevant products or items are shown to the customer.
- The most easy, efficient and fast way to gain review about anything these days is the internet .For example, Starbucks introduced a new coffee with its major concern being that the taste might be too strong for customers. Once the coffee was out, Starbucks monitored all blogs, articles, comments

and post on the internet and discussion forums. Using all this information from the internet and analyzing it, Starbucks discovered that although the new coffee was appreciated by the customers but the price was too high. So by morning, Starbucks lowered the price of the coffee and all negative reviews had disappeared.

- Government uses big data in many of its aspects. For example, police use phone records of criminals to track them or their GPS records to see the places they have been prior to their crime commitment in order to find the guilty. Health ministry uses Bigdata to check infant ratio, percentage of diseases widely spread in a particular area or season.
- Business organizations use Bigdata to analyze their business sales, profit and losses across regions over the year.
- Big data is not just for big companies. Retailers use big data analytics to check current trends among the population. retailers use Bigdata .i.e. collaborating data from web browsing patterns, social media, industry forecasts, existing customer records, etc. and predict trends, prepare for demand, pinpoint customers, optimize pricing and promotions, and monitor real-time analytics and results.
- Role of Social media in big data analysis- social media is one fast growing field. Billions of users from all over the world are active on multiple social sites. Millions of comments on various technologies, software, apps, dressing materials etc are used to analyze the customer likings. Based on this , various vendors make new catalogs of their products.
- Cloud and Bigdata-Big Data stores most of its data in distributed file systems. These files can be accessed from anywhere .i.e. cloud data. Thus extracting and analyzing this data from any part of the world becomes a cake walk.
- Fraud detection- fraud is a billion dollar business that is increasing every year at an unstoppable rate. Traditional mechanisms of detecting fraud have been long used. However, they are complex and time consuming. Big data records the patterns of every customer and their usual ways of depositing, transferring or withdrawing money. Any unusual activity either in the amount withdrawn can indicate the risk of a fraud.
- Data from applications is used by app developers to gain information about the likes and dislikes of the customer. For example, game developers add new features to their games every month. On app purchases that are made by the player helps them earn a lot of money. [3]
- Weather forecasting is solely based on using old data to determine weather patterns. Data from weather sensors, satellite, etc.. is collected and used to determine weather conditions for each region.
- Another use of big data is made by Courier companies. The use of big data helps them to map out more efficient trucking routes. The resulting improvements have allowed UPS to save 30 million miles and 3 million gallons of fuel per year from their routes. The more efficient trucking routes have also led to less traffic accidents.
- Science - scientist use large amount of astronomical data collected over the years to determine constellation positions, meteor showers in a region, birth and death of a star, measuring distances between various space objects and on the basis of these patterns , they model their spacecrafts with ample fuel, food, oxygen and other necessities for sustenance.

VI. Need for Security

Big data is now being used by many industries for analysis purpose data which ensures new marketing strategies; such kind of data must be secured. It any security comes it will result in serious legal repercussions and a great damage to reputation. To secure this much of data different mechanisms should be used.

The importance of big data in an organization for fraud detection is quite useful. Detecting threats and malicious intruders should be detected and solved using big data analysis. The main issues with big data is security and privacy.as this concept of big data is increasing day by day different organizations are dealing with the problems with the privacy of data.

VII. Issues and Challenges

Cloud computing delivers the services on demand rather than products but it comes with lots of security issues as it surrounded by different technologies together such as networks, databases, operating systems and memory management therefore the issues related to these technologies are automatically included in cloud computing.

- A network which connects the systems should be secure.
- Mapping between virtual machine –physical machine done carefully.
- Not only encryption should be used appropriate data sharing policies should be enforced.

A. Issues in big data

The amount of data generated every second of the day is unimaginable. Data from multiple sources is flooding devices where its storage lies. Bigdata is data that is greater than terabytes in size. Data from social media, sensors, machines, phones, etc, need to be stored and processed so as to best exploit it. Processing this data is not an easy task. Data that is collected from numerous sources is raw data that is it barely holds any meaning. A series of steps is followed to make this data meaningful and then put it in the warehouse. Fortunately, the technology today has made this painstaking task simpler.

The following steps are followed in the data cleaning process.

1) Data Acquisition - data just does not arise on its own. It is collected, recoded from various sources over time. From the toxins we breath to the number of people dying every day produces terabytes of raw data per day. Data acquisition involves business understanding that means determining business objectives.

2) Information extraction and cleaning- Much of the data collected is useless and can be discarded. This step involves initial data collection (relevant), data description, data exploration, and the verification of data quality. Data exploration includes viewing summary statistics. Various models can be applied to identify patterns such as cluster analysis.

3) Data Integration, Aggregation, and Representation- Once the data resources are identified; they are selected, cleaned and built in to the desired form. Combining the data from multiple sources into a single meaningful format is what this step consists of. For example, we perform multiple scientific experiments but the observations are made on the basis of all those experiments and not just one. Data analysis is considerably more challenging than simply locating, identifying and understanding data. Mathematical techniques are used to identify patterns. On this basis models are assessed and built. Thus data visualization is an important part of the process.

4) Query processing, Data Modeling and Analysis- query analysis in data mining differs from the traditional method. Large volume of data requires efficient algorithms that can analyze data and display desired results. For example, plotting points on a graph for analysis becomes very difficult when extremely large amounts of information or a variety of categories of information is being dealt with. For example, if there are 10 million rows of sales data in 5 years that is to be compared. The user trying to view 10 million plots on the screen will have a hard time seeing so many data points. Thus in such cases, one of the many visualization and analysis technique is used to form a cluster by grouping data together so that it is visible to the user. This technique is called as cluster analysis.

5) Interpretation- The entire purpose of going through all these process is the end purpose of data interpretation. Data interpretation is performed by the user on the basis of the result that is display after Big data analysis is performed by the machine. [13]

Executing these processes is not an easy task. It needs to be performed accurately and timely.

- Timeliness - the larger the data size, the longer it takes to process it. Thus the design of the system must be efficient enough to accommodate a data this vast. A smaller set of data is processed fast and chances of error are also less.

- Privacy- Privacy is a major concern in Big data. Since data lies in distributed file systems, data flows freely from one user to another specially in public servers. This is why most companies prefer having a private storage. In case of health records, some records may be sensitive and unauthorized access much not be permitted. Individuals do not realize how their information is used for data mining purpose by the very companies that provide them with that service. The most prominent threat is to that of a person's location at a particular time or at all times.
- Cost- with in increasing data every second, the cost of storing this data is also increasing. Since Big data required both historical and present data, discarding old data is out of question.
- Skills - the skills required to efficiently use this data in the form of human resource is also less in comparison to the world population.

B. Issues in cloud computing

Issues in cloud computing can be categorized

- Network level- in this level we deals with network that includes network protocols and it security.
- Authentication level- it includes the encryption and decryption techniques also with authenticating services (administrative rights for node, authenticating a node and logging).
- Data level- it deals with data integrity its protection and distribution.

C. Approaches

Cloud is a combination of different technologies the approach to follow should be applicable to is integrated technologies. The recommendations are designed in such a way the they don't lower the efficiency and scalability if cloud computing. Some measures are-

- File encryption- data is stored in a cluster therefore any intruder can steal all critical information. Data must be stored in encrypted form and keys should not be publically disclosed. Even if a hacker is successful to enter still he isn't enable to read the data.
- Network encryption- it simply means the network communication should be encrypted such that a hacker won't be able to manipulate the packets that are flowing inside the network.
- Authentication- all the transactions should be authenticated i.e. all those events that include events on data must be logged and the user that is doing such transactions also be logged. This will result in knowing if any user is trying to do any malicious operation.
- Node maintenance- system where the software is running must eliminate the risk of virus.
- Testing of map reduce jobs- map reduce job entered by the user should be properly tested in the distributed environment. [15]
- Nodes authentication- whenever a node joins a cluster it should be authenticated. If in case a node is trying to do some malicious operation it should be disconnected from the cluster. Kerberos can be used to authenticate the node.

VIII. Conclusion

From a historical view point, Bigdata is a huge evolution in the decision making process that provides computer based decision making. It comprises of all benefits of a data warehouse and an added functionality of analyzing data from distributed file systems. It owns the ability to capture, store and analyze high-volume, high velocity, and high-variety data is allowing decisions to be supported in new ways. It is also creating new data management challenges.

From Big data is a special asset that merits leverage. Statistical information, or data, is a potentially rich and a valuable source of knowledge. the most interesting fact about this is the cost of storing this information is getting cheaper and cheaper thus allowing us to keep much more information than was previously possible. Bigdata is cost effective. Sophisticated computer algorithms have been designed to make this process even easier by sometimes revealing interesting relationships that would never have been possible previously.

We also need to resolve legal issues around intellectual property rights, data privacy and integrity, cyber security and Bigdata code of conduct.

Cloud computing is widely used in organization for research purpose and its security is a primary concern and it should be eliminated to minimal in order to provide secure environment for complex operations.

References

1. https://www.google.co.in/search?q=bigdata&oq=BIGDATA&aqs=chrome.0.69i59l2j69i60j0l3.1047j0j4&sourceid= chrome&es_sm=93&ie=UTF8#q=traditional+systems+of+stroring+data
2. Basic Concepts in Big Data ChengXiang ("Cheng") Zhai Department of Computer Science University of Illinois at Urbana— Champaign Bp://www.cs.uiuc.edu/homes/czhai czhai@illinois.edu
3. Chulis, K. (2012) "Big Data Analytics for Video, Mobile, and Social Game Monetization", developerWorks, IBM,
4. http://www.ibm.com/developerworks/library/ba-big-data-gaming/ (current March 7, 2014).
5. Russom, P. (2011) "Big Data Analytics", TDWI Best Practices Report. Seattle: The Data Warehousing Institute,Fourth Quarter, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx (currentMarch 7, 2014).
6. van Groningen, M. (2009) "Introduction to Hadoop", TRIFORK Blog, http://blog.trifork.com/2009/08/04/introductionto-hadoop/ (current March 7, 2014).
7. http://newbooksinbrief.com/2013/03/21/31-a-summary-of-big-data-a-revolution-that-will-transform-how-we-live-work-and-think-by-viktor-mayer-schonberger-and-kenneth-cukier/#31c10
8. http://www.villanovau.com/resources/bi/what-is-big-data/#.VsmRghGqqko
9. J. Dean and S. Ghemawat. MapReduce: A Flexible Data ProcessingTool.CACM, 53(1):72–77, 2010.
10. S. Ghemawat, H. Gobioff, and S.-T.Leung. The Google file system.In SOSP, pages 29–43, 2003.
11. A. Floratou et al. Column-Oriented Storage Techniques forMapReduce.PVLDB, 4(7):419–429, 2011.
12. Hadoop, http://hadoop.apache.org/mapreduce/ F. N. Afrati and J. D. Ullman. Optimizing Joins in a Map-ReduceEnvironment. In EDBT, pages 99–110, 2010.
13. Challenges and Opportunities with Big Data: white paper
14. Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud
15. computing reference architecture: Cloud service management
16. perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.
17. Kilzer, Ann, Emmett Witchel, Indrajit Roy, Vitaly Shmatikov, and Srinath T.V. Setty. "Airavat:Security and Privacy for MapReduce."