Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015), Palembang, Indonesia, 19 -20 August 2015

BIG DATA MANAGEMENT PROTOTYPE DEVELOPMENT for ANALYSIS VARIOUS of DATA

Sulistyo Heripracoyo
Department of Information Systems, School of Information Systems, Binus University
Jl. K.H. Syahdan No. 9, Palmerah, Jakarta
hpracoyo@binus.edu

Abstract—The phenomenon of big data is currently a growing topic in the world of information technology. From some of the literature mentioned that manage big data can create significant value for the world economy, improving productivity and competitiveness of enterprises and the public sector as well as creating a large economic surplus for consumers. However, from some of the information obtained, big data is still not widely applied in the company or organization. This study aimed to explore more information about the big data and proceed with making an application prototype big data management. To experiment with big data storage that is database, this research use NoSQL database technology that can map the needs of both structured and unstructured. And this research will be carried out migration of Relational Database (RDBMS) into the database MongoDB. Prototype will be create with the object of study is structured and unstructured data. The expected result of this research is a model or prototype of big data management that can help organizations and companies (especially education) to make decisions based on various types of data.

Keywords—Information Technology, Big Data, data analytic, NoSQL, MongoDB

INTRODUCTION

New research from the McKinsey Global Institute [9] found that collect, store, and explore massive data (big data) for an insight can create significant value for the world economy, improving productivity and competitiveness of enterprises and the public sector as well as creating a large economic surplus for consumers. There are five ways to leverage big data, namely: Creating large data more accessible and timely, use of data and experiments to expose variability and improve performance, segment of the population that can be adjusted, use the automatic algorithm to replace and support decision making by humans and innovate the business model, products, and new services.

In addition, companies that use IT will have the data to be accumulated and be a big volume. Beginning in 2000, when a sharp rise in the volume of data, CPU and storage technology (storage) are faced with a large number of terabytes of big data volumes, when faced with a crisis that IT data scalability. Storage and CPU not only developed the capacity, speed and greater intelligence, but also the price falls. The Company cannot buy or manage big data related to budget abundant collection and analysis [6]. With the cloud computing, the company can buy services on the type of infrastructure services (storage).

However, due to large volumes of data, companies that have such data are not much use and manage. Companies need to manage large amounts of data (big data) to explore and analyze information for the needs of the company, the company needs to manage data [2][3]. In addition, a survey of InformationWeek (September, 2012) gives 6 things about big data with different views, namely big data is not needed and required.

According to the disagreement about the benefits of big data, then the phenomenon of big data need to be explored further, how big data management in the company and its implications. Early studies of this research is structured and unstructured data is available at Bina Nusantara University. The data will be processed and processed to be able to support a prototype in use and analyze the information generated.

Research Purposes

This study was conducted to determine various information related to big data is used as reference. Furthermore, in this study will be made a prototype of big data management, which can be used to assist in the analysis to support decisions in organizations and companies. In this study, the prototype will be made for this type of education organization, which is specifically the case study will be conducted universities (in this study is a case study of Bina Nusantara University).

Usefulness of Research

a. Provide a model / prototype data management within an organization or company. Where data is maintained throughout the data types do not differentiate between structured data and unstructured data.

b. Provide a data model that can be used to assist the analysis with reference to a variety of information generated from structured and unstructured data.

LITERATURE STUDY

Not many big data describe term certainty. Nevertheless, the term "Big Data" is often used by companies to describe the huge amount of data. This is not referring to a specific amount of data, but to describe a set of data that cannot be stored or processed using traditional database software. Examples of big data include Google Search Index, Facebook database (user profile). Big data is often distributed over many storage

devices, can be in several different locations. There are several different types of software solutions of different big data, including data storage platforms and data analysis programs. The most common product of the software includes Apache Hadoop big data, IBM's Big Data Platform, Oracle NoSQL database, Microsoft and EMC HDInsight Pivotal One [8].

Other definitions, Big Data is a data overflow in pace never seen before - has doubled every 18 months - as a result of access to a larger customer data derived from public sources, exclusively, as well as new information gathered from the web community deployed in a way new. [3]

Identify the key elements of big data is, first companies today can collect data across business units, increasing the data, even data from partners and customers (large and complex). Second, a flexible infrastructure that can integrate information and effectively to meet the increasing wave of large data. Third, experimental and analytic algorithm can make sense of all the information of big data. Big Data has also become a core element of the strategy. [3].

Big Data focus on the size of the data in storage. The size of the issue, but there are other important attributes of big data is the diversity of data and data rates. Three V big data (volume, variety and velocity) establishes a comprehensive definition, and thus reducing the myth that big data is just about the volume of data [6]. In 3 V big data, related to the size of the volume that is Terrabyte, records, transactions, tables and files. Velocity related to batch, near-time, real time, stream (stream). While Variety relating to structured, unstructured, semi-structured and three. As shown in the figure below.

Figure 1. The Three of V big data [6]

The volume of data is a major attribute of big data, most people define the size of big data-terabyte (TB), sometimes in petabytes (3 to 10 TB). Some organizations find more useful for the quantification of big data in terms of time, such as Act limitations seven years in America, many companies maintain data for seven years for risk analysis, complaints and law. [6].

Big data is a difficult part of the competition for market share. It is important to note that the threats and opportunities associated with large data often has implications for organizations that only the attention of senior executives can handle it. Leaders are too little to understand the potential of big data in their business, data assets and liabilities of the business, or their strategic choice is to be made to begin utilizing large data. By focusing on these issues, senior executives can help organizations build a competitive advantage based on the data. [3]

Technology to use and analyze the information widely available, but many companies are taking a new level data using IT to support appropriately, directing decisions and test new products, business models, and innovation to the customer experience, in some cases this approach helps companies to make decisions in real time. The company sells physical products also use big data to the appropriate experiments. Use the information to analyze new business opportunities, such as the effective promotion of the right segment. Other companies collect data from social networks in real time (Ford Motor, PepsiCo and Southwest Airlines). The use of experimental and big data as an important component in management decision making requires new capabilities, as well as organizational and cultural change, most companies away from accessing all the data available. Generally, the company does not have the right talent and processing for designing experiments and extract business value from big data, which requires a change in the way many executives who currently make a decision: to trust instinct / instinct and experience during the experiment and rigorous analysis. Big data over time will be a new type of company assets, which indicates an important key to the competition. If it is true, the company should start thinking seriously whether they are organized to exploit the potential of large data (big data) and to manage the threats that may arise. Success will require not only new skills, but also a new perspective on how to revolutionize the era of big data-expanding circle of management practices that can influence and foundations represent novelty, potential business models (disruptive). [2]

InformationWeek Survey (2012), advise that before buying storage, warehouse platform upgrade, or adopting Hadoop need to check reality.

Other survey data related to the consideration of InformationWeek Big Data analysis tool, which shows that the consideration for the analysis of big data tools only 15% consider, 41% did not know and 44% do not consider.

Management or management of big data is to organize, administer and governance of large volumes of data both structured and unstructured. The purpose of the management of big data is to ensure a high level of data quality and accessibility for business purposes, including business intelligence and big data analysis applications. Corporations, government agencies and other organizations using a large data management strategies to help them compete with rapidly growing data sets, usually involves a lot of terabytes or even petabytes of information and various types of data.

MGI mentions the potential benefits of big data between private and public sectors, identified five ways big data can create a value[[9]:

- Segmenting the audience to adjust activities: data collection and capacity relative audience segmentation-based databases have become the key driver for use by many arts organizations based on the organization's core strategy of operating data (such as databases).

- create transparency: making big data to be more accessible to stakeholders (stakeholders) that are relevant in a timely manner can create tremendous value. In this case the big

data can be integrated so that the required information can be obtained more efficiently.

- Support / replace human judgment with automated algorithm: a sophisticated analysis with the help of the algorithm can improve decision-making, minimizing risk, and explore valuable insights, which can be done automatically. Although the decision is not necessarily automatic.

- Enables experimentation: for the manufacture and storage of transactions in digital form, the organization can collect data more accurate and more detailed performance.

- Innovate business models, new products and services: Big Data enables companies create new products and services, improve existing ones, and created a business model that is completely new.

Of some industries, MGI demonstrated the benefits of big data like in the picture below [9].

Figure 2. Big data generate financial values [9]

Because of the way to secure a competitive advantage of big data is still evolving, some CEOs believe that big data initiatives (big data) should be the responsibility of the IT department or marketing specialized companies - large-scale functional group wherein the amount of data that is most frequently gathered, analyzed, and implemented. Therefore associated with big data, here are some things that could be the CEO and his team view, a) can be derived from the perfect opportunity core operations to create new business lines - even in the same industry. b) To be useful, the data must traverse between the organization - but it often causes friction. Only the senior team of dedicated and focused can eliminate various purposes / objections, c) whether a company planning initiative, a single large or smaller that much, senior team should actively plan to take advantage of the opportunities generated. Stay aware of the necessary resources (technology and vice versa) quickly shifted into a mode of implementation of the pilot.

NoSQL Database provides a mechanism for the storage and retrieval of data being modeled in other ways besides the relationship table used in relational databases. The motivation for this approach include simplicity of design, horizontal scale and better control over availability. Data structures (e.g. trees, graphs, key-value) is different from the RDBMS, and therefore some operations faster in NoSQL and some in the RDBMS. (Wikipedia)

Benefits of NoSQL databases [7] is 1. Scale elastic, 2. Big Data, 3. Little maintenance (slightly administration and tuning), 4. Cheap, 5. flexible data model. Five challenges to NoSQL database is 1. Maturity, 2. Support (support), 3. Analytics and Business Intelligence, 4. and 5. Expertise Administration. NoSQL databases become an important part of the landscape database, and when used properly, can offer real benefits. However, the company should proceed with the full attention of the limitations of legitimacy and problems associated with this database.

Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file system. However, the differences from other distributed file system is significant. HDFS highly fault-tolerant and is designed for use on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have a large data set (big data)

RESEARCH METHOD

This study begins with a literature study further about big data, collecting information to create a model / prototype big data applications. Next is to collect samples of structured data to research object, the data sample is a sample of data from a SQL Server database that is used in the operation. Prototype that will be created using a database that can store different types of data (unstructured) is NoSQL (MongoDB).

Structured data from SQL Server database will be converted into NoSQL databases (MongoDB) Database using the conversion process (with a conversion tool from SQL Server to MongoDB). Furthermore, the converted data will be accessible through the applications that can be data analysis (data warehouse), data mining. The model that will be made are as follows:

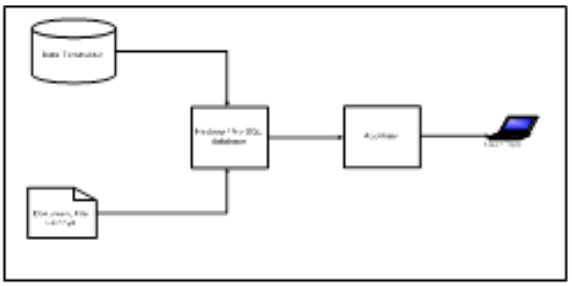

Figure 3. Research Model

RESULT

To make this prototype, the steps are:
- To install the Operating System (Windows Server 2012)
- To install a database NoSQL (MongoDB).

For NoSQL Database, the software used is version 2.6 (64-bit) obtained from http://www.mongodb.org/.

To Setting NoSQL database in Windows Operating System.

a. Determine the location of its log (md "C: \ Program Files \ MongoDB \ log")

b. Determine LogPath (LogPath = C: \ Program Files \ MongoDB \ logs \ mongo.log> "C: \ Program Files \ MongoDB \ mongod.cfg"), for example as shown below.

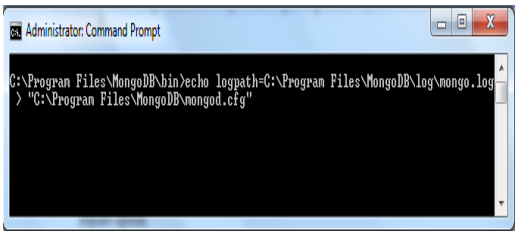

Figure 4. LogPath Setting

c. Install MongoDB Services.(Files\MongoDB\bin\mongod.exe" --config "C:\Program Files\MongoDB\mongod.cfg" –install) , for example in this figure bellow.

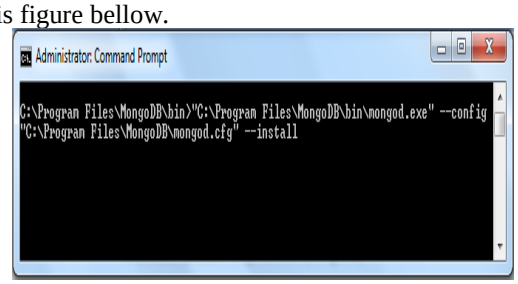

Figure 5. Install MongoDB Service

d. To test Connection to MongoDB, can be seen from the following picture:

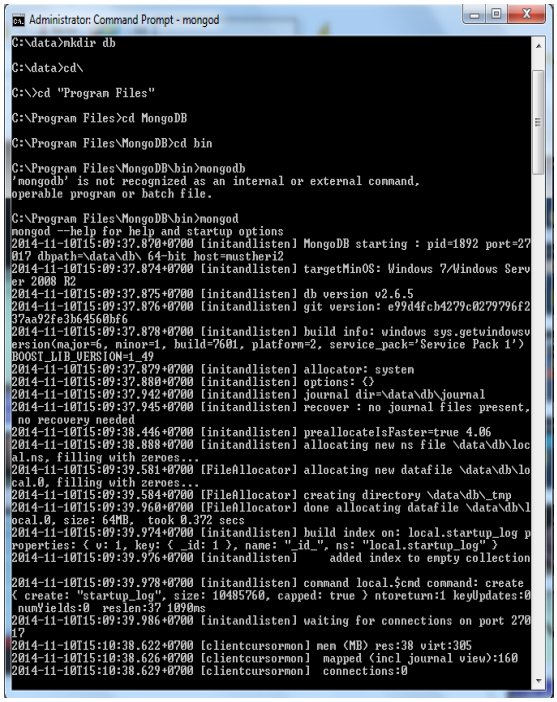

Figure 6. Testing Connection MongoDB

To change the path to the configuration file, change to the path in the file mongod.cfg as required, usually will be located in the directory c: \ program files \ MongoDB \ Mongod.cfg.

If the directory does not exist DBPATH, mongod.exe file will not start. Default value for DBPATH is \ data \ db. If necessary, can also install services (services) for many instant of mongod.exe or mongos.exe. Multiple instances only when appropriate resource existing system and system design requires an instance of the.

Termination Service MongoDB as required

To stop the service (service) MongoDB, the following command can be done.

```
net stop MongoDB
```

As for me-remove Service (Service) MongoDB service, the following command can be used:

```
"C:\Program Files\MongoDB\bin\mongod.exe"
--remove
```

Manually Making Windows Services for MongoDB

If MongoDB has been installed with MSI Installer, the default path is C: \ Program Files \ MongoDB 2.6 Standard. (If installed in another directory, need to set the path accordingly)

Open a Command Prompt Administrator

For the operating system with the version of Windows 7 / Vista / Server 2008 (and R2), the command used is: Press Win + R, then type cmd, and then press Ctrl + Shift + Enter. For Windows 8, press Win + X, then press A.

Create Directory

To create a directory, the same as in create general directory instruction. The directory needs to be made is for the database and log files. Instructions that can be executed is:

```
mkdir c:\data\db
mkdir c:\data\log
```

Create Configuration File

Making a configuration file, the file can include any configuration options for mongod, but must include a proper setting for LogPath. The following configuration file creation, specify the LogPath and DBPATH in the filter:

echo logpath=c:\data\log\mongod.log> "C:\Program Files\MongoDB 2.6 Standard\mongod.cfg"

echo dbpath=c:\data\db>> "C:\Program Files\MongoDB 2.6 Standard\mongod.cfg"

Create the MongoDB service.

Create the MongoDB service.

sc.exe creating MongoDB binPath= "\"C:\Program Files\MongoDB 2.6 Standard\bin\mongod.exe\" --service --config=\"C:\Program Files\MongoDB 2.6 Standard\mongod.cfg\"" DisplayName= "MongoDB 2.6 Standard" start= "auto"

sc.exe need somewhere between "=" and the value of the configuration (eg "binPath ="), and a "" to remove the double quotes.

If successfully created, the following log message will be displayed:

```
[SC] CreateService SUCCESS
```

Starting and Stopping Service MongoDB

Service / Service of MongoDB database applications can be done with the instruction: net start MongoDB, while to stop the service can with the instruction: net stop MongoDB, whereas for me-remove service MongoDB with the instruction: the first stop of his service and the next is with the command: sc.exe delete MongoDB

The process of migration of the SQL Server database to MongoDB

For prototype MongoDB database, the database used is the result of migration of existing database is SQL Server. The tool used is sql2mongodb (MongoDB SQL Server Importer). This tool was developed in C #.

To perform the installation Sql2mongodb, can use the following instructions :

```
npm install -d sql2mongodb.
```

To perform the conversion from the RDBMS into MongoDB or other ODBMS, can use the following syntax:

Sql2Mongo.exe <SQL Server Connection String> <MongoDB Connection String>

For example: MongolSqlImport "Data Source = (local); Initial Catalog = AdventureWorks2008; Integrated Security = True" "mongodb: // localhost: 27017"

Would look something like this:

Figure 7. Conversion Process

Term in MongoDB

Here is a variety of terms and terminology of SQL MongoDB.

Table 1: Term in SQL and MongoDB

Concept/SQL Term	Concept/MongoDB Term
Database	Database
Table	Table
Row	Document or BSON Document
Column	Field
Index	Index
Table Join	Embedded Document and Linking
Primary Key Determining Unique each column or combination of columns as the Primary Key	In MongoDB, Primary keys are automatically set to the id field.
Aggregation (eg Group by)	Aggregation pipeline

The following shows some executables MongoDB database and the corresponding executable..

Table 2: Term of Executable database

	MongoDB	**MySQL**	**Oracle**	**Informix**
Database Server	mongod	mysqld	oracle	IDS
Database Client	mongo	mysql	sqlplus	DB-Access

Advantage of MongoDB.

Purposes of MongoDB.
1. Can ensure access speed by User.

MongoDB is the era of Cloud database designed for big data storage and query, and applications such as Facebook Social Network. MongoDB obtain especially the performance with a design based on a key value and easy to add. MongoDB uses the document as the basic storage unit. A document is a simple JSON as an object.

For example, a blog post consists of a Title, content and comments. In the relational model, the comment will be saved as individual table and called the join table Post and table comment. In the model document, all stored in a single document, where the document can be identified by an ID. So getting documents is a key value query, the query is not relational.

Query the value to be faster than relational queries, if using RDBMS then have to perform a process denormalization

2. Document Model

Mongo DB there is no concept of a table, row, SQL, schema, also some queries.

3. Flexible Schema

Actually, no schema in MongoDB, a document can have a field, the field can be added to the existing document at any time dynamically. No ALTER TABLE, no rebuild Indexing. The document exactly as JSON, PHP array, or dictionary Python. Very natural to communicate with MongoDB with dynamic languages such as JavaScript, PHP, or Python.

4. Does not support transactions
5. Does not support JOIN

MongoDB does not support features such as relational databases such Transaction or JOIN, but have the ability to more easily and flexible schema that is easy to manipulate with JSON as the data format.

CONCLUSION

Based on experiments using MongoDB, obtained that the RDBMS relational database (e.g. SQL Server) can be converted into objects (documents) Database Management System (ODBMS). Where data are used in ODBMS be an object / document can be accessed flexibly. With MongoDB database the corresponding data is historical data in large quantities, which do not contain transaction data. So that the database is suitable for use as a data storage that is used as the analysis, because no change again. Especially suitable for use as a decision support data such as data mining or text mining.

REFERENCES

[1] Barar, K., Sunita, B., & Reksheveenay, b. (2013). Future of Cloud Computing. International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol 2 Issue 3 May.
[2] Brown, B., Michael, & Manyika, J. (2011, October). Are you ready for the era of 'Big data' ? McKinsey Global Institute.
[3] Bughin, J., Chui, M., & Manyika, J. (2010, August). Clouds, big data, and Smart assets: Ten Tech-enabled business trends to watch. . McKinsey Quarterly.
[4] Healey, M. (2012, October 31). 6 Lies About Big Data. Information Week.
[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011, June). Big Data: The Next frontier for innovation, competition, and productivity. McKinsey Global Institute.
[6] Russom, P. (2011). Big Data analitics. TDWI Research, TDWI Best Pratices Report, Fourth Quarter.
[7] Harrison, Guy (2010). 10 things you should know about NoSQL databases http://www.techrepublic.com/blog/10-things/10-things-you-should-know-about-nosql-databases/#. Akses data : 26 Februari 2014.
[8] Techterms.com. (2013, August 27). www.techterms.com/definition/big_data. Retrieved November 13, 2013, from www.techterms.com: www.techterms.com/definition/big_data
[9] The Challenge - and opportunity - of 'big data'. (2011, May). Mckinsey Global Institute.
[10] Hadoop. http://hadoop.apache.org/docs/stable1/hdfs_design.html . Retrieve : 28 February 2014.
[11] MongoDB, www.mongodb.org . Retrieve: 28 February 2014.