Advanced Science and Technology Letters
Vol.51 (CES-CUBE 2014), pp.42-45
http://dx.doi.org/10.14257/astl.2014.51.10

Review on Present Situations of Big data

Byung-Tae Chun[1], Seong-Hoon Lee[2]

[1]Dept. of Computer Web Information Engineering, Hankyong National University, 327, Chungang-no, Anseong-si, Kyonggi-do, Korea
chunbt@hknu.ac.kr
[2]Division of Information&Communication, Baekseok University, 115, Anseo-dong, Cheonan, Choongnam, Korea
shlee@bu.ac.kr

Abstract.A major property in our society is a acceleration on convergence byIT technology. Because of these properties, various data types are produced through different many devices. In this paper, we described the reviews on current state of Big Data.

Keywords: Big Data, Hadoop, Convergence, Usage.

1 Introduction

Today, the hot issues in IT industry include big data, cloud computing, and convergence. Gartner, a research consultant agency, released the 10 major technologies and trends such as war of mobile devices and strategic big data that companies should cope with in 2013. Gartner predicted that in 2013, mobile phones would overtake PCs as a web access device most widely used all over the world, and that by 2015, smart phones would account for more than 80% of all mobile phones sold in advanced countries[1]. When the term, 'big data,' first appeared, the meaning was interpreted differently. One group defines it as "terabyte data," and another defines it as "architecture of processing a large quantity of data." Since the meaning of the term, "big," itself is relative, however, it would not be appropriate to define an absolute standard for the data capacity.

Mckinsey, one of the global consulting agencies, defined in one report released in 2011[2] big data as "a dataset that exceeds the capacity of existing database management tools in data collection, storage, management, and analysis," stating that "the definition is subjective and will continue to change."

Element technology of big data includes media-related data volume, data input/output velocity, and data variety. Fig. 1 shows three element technologies. The term, 'volume,' means a data attribute of tens of terabytes or tens of peta bytes in general. 'Velocity' is an attribute referred to in fast processing and analysis of large capacity data.

In a convergence environment, digital data is produced at a high speed, and thus the system should be capable of saving, distributing, collecting, and analyzing it real time. 'Variety' indicates that there are various types of data, and they could be classified to structured, semi-structured, and unstructured data sets depending on the

ISSN: 2287-1233 ASTL

sort of structure.

It is reported that the quantity of data handled around the world today is doubled every two years [3][5][6]. As IT is converged with other industry sectors and a tremendous amount of data is being generated, the utilization of big data has become a great issue in addressing desires and demands regarding the quality of life in this changing society.

The most important factor in big data processing is the storage technology to collect various types of gigantic data as mentioned above and the data analysis technology to analyze it for a meaningful use. In this era of big data, new technologies such as hadoop have emerged and provided functions to process and analyze data that the existing technologies did not have. Figure.1 shows the overview of Hadoop.

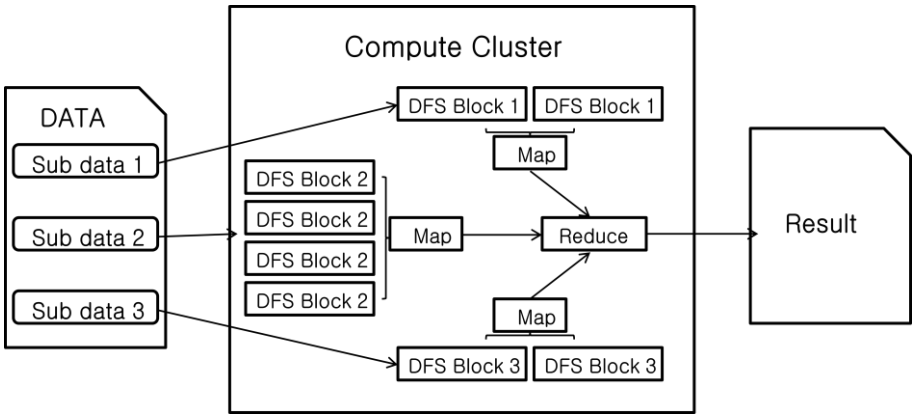

Fig. 1. Structure of Hadoop

The utilization of big data now goes beyond the area of 'big data management' led by business entities and expands into the area of public service for general peoples. Big data is made use of for improvement of national competitiveness, not merely for corporate competitiveness.

2 Reviews on Big Data

Various types of platforms handling big data basically consist of the three elements: storage system, handling process, and analysis mechanism. This study focuses on the platform technology related to the handling process among the three elements stated earlier. In a big data handling process, the core of parallel processing is 'Divide and Conquer,' that is, to divide data into independent sets and handle them in parallel. Big data processing divides a problem into multiple small operations, collects them, and combines them as one single result. As for operation dependency, however, the advantage of parallel operation is invalid. In reflection of this limitation, the proper data storage and processing method is necessary.

One of the well-known large quantity data processing technologies is the Map-

Reduce distribution data processing framework such as Apache Hadoop.

InMap-Reduce model, A common embedded hard disk drive in a common computer may be used for the operation with no need for special storage means. As each computer has a quite weak correlation to one another, it is possible to expand the link to hundreds or thousands of units.Since a number of computers are involved in the processing, it is assumed that malfunctions of the system including hardware are not exceptional but common.Complicated problems can be solved with the simple and abstract basic operation of Map and Reduce. Even programmers unfamiliar with parallel programs can readily handle data in parallel. It can handle high throughput rates when a number of processors are used simultaneously.

Dryad is a framework to form data channels between programs in graph and handle them as data sets in parallel. Map-Reduce framework developers design Map and Reduce functions while Dryad developers design data processing in graph. Dryad can process data flows in the format of DAG(Direct Acyclic Graph).

Parallel data operation frameworks such as Map-Reduce and Dryad provide sufficient functions to process big data, but they involve some barriers against inexperienced developers, data analyzers, and data minors. It is necessary, therefore, to develop a method of a higher level of abstraction to handle data in an easier manner. Apache Pig and Apache Hive to be explained below are the two examples of such a framework.

Apache Pig provides a high level of large quantity data combining and processing structure. Apache Pig supports the Pig Latin language. A data processing program of Pig Latin is converted into a logic execution plan, which is again converted into a Map-Reduce execution plan. Apache Pig adopts an approach to designing a large quantity data processing program in a format of procedural programming languages such as C and Java. Sawzall of Google adopts a similar approach. Some technologies adopt declarative data processing methods such as SQL instead of specifying data processing procedures as in a programming language. Apache Hive, Google Tenzing, and Microsoft SCOPE are some of the examples.

Apache Hive is a technology to analyze large quantity original data sets such as HDFS and HBase by means of a query language called HiveQL. It can be divided to the Map-Reduce based execution part, metadata information on the data storage, and execution part based on the queries received from users or applications in terms of architecture. To support user expansion, it is possible to designate a user-defined function on the level of scalar value, aggregation, and table.

3 Conclusions

Currently in our society, smart phones are commonly used, and various data producing devices such as tablet PC, camera, and game console recently emerged, which have increased the traffic drastically. In addition, as the volume and types of data become diversified and data increase velocity is rapid, the era of 'big data's seems to be just ahead. Today, it is reported that the amount of digital information handled around the world is doubled every two years. As IT is converged with other industry sectors, a large quantity of data is produced every day, and the issue of utilizing big

data in addressing desires and demands for a better quality of life in the changing society is in a spotlight.

In the future, the interest in and use of big data platforms will continue and expand. The applicable area too will go beyond pure IT and be expanded to every possible sector. When such efforts are consistently expanded and developed, the future society will open the door to a world of infinite possibility.

References

1. Gartner. CEO Advisory: 'Big Data' Equals Big Opportunity. (2011).
2. McKinsey. Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, (2011).
3. Warden, P.: Big data Glossary. O'Reilly Media, (2011).
4. Jung, J. S.: 3 Factors for Successful Big Data Usage: Resource, Technology, Manpower. Big Data Strategy Forum, (2012).
5. IDC. Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO. (2011).
6. IDC. New Analytics Strategies in the Big Data. Era, (2011).

