Int. J. Advance Soft Compu. Appl, Vol. 7, No. 1, March 2015
ISSN 2074-8523

On a Hadoop-based Analytics Service System

Mikyoung Lee, Hanmin Jung, and Minhee Cho

Korea Institute of Science and Technology Information (KISTI)
e-mail: jerryis@kisti.re.kr, jhm@kisti.re.kr, mini@kisti.re.kr

Abstract

In this study, we discuss the development of an analysis service that uses Hadoop for improving the performance of a database management system (DBMS)-based analysis service system that processes big data. DBMS-based systems are not suitable for processing big data and providing service because of their disadvantage in consuming more time for processing and analyzing. We introduced a distributed parallel platform, Hadoop ecosystem, for improving the performance of the system by minimizing the processing time in analyzing big data. In addition, we also carried out a method of optimizing an existing analysis module to minimize the processing time. SPARK and Hive were implemented in the Hadoop platform to optimize the distributed parallel infrastructure. The developed analysis service system showed 127.87 times higher performance than that of other existing systems.

Keywords: *Big data, Hadoop ecosystem, Analytics service, Distributed object-oriented platform, Prescriptive analytics.*

1 Introduction

Nowadays, with increasing interest, big data analysis processing technology, which rapidly processes big data, has become a key technology in big data analysis. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, creation, storage, search, sharing, transfer, analysis and visualization[1]. Hadoop is a distributed process technology that is widely used in current big data applications. Evolving to a high level, it can store many types of data and perform complex data analysis. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Apache Hadoop(High-Availability Distributed Object-Oriented Platform) is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware[2]. It is an open-source data processing technique that provides results rapidly by distributive processing vast amounts of data. Hadoop is the most preferred distributed processing core technology for processing big data, and it consists of the Hadoop distributed file system (HDFS), HBase, MapReduce, etc. MapReduce is a programming model for data processing in Hadoop. In addition, Hive supports the query language HiveQL, which is similar to SQL, and when a query is issued from HiveQL, it is converted to an optimized map and reduced by a parser. Spark was introduced to overcome the disadvantages of MapReduce, and it can support Iterative algorithms. This paper describes the distributed parallel environment design method that uses the Hadoop ecosystem to process massive amounts of data from the big data-based prescriptive analytics service system. In this study, the analysis model was parallelized and optimized using Hive and Spark.

2 Big data-based Analytics Service System

2.1 Data resources

The data in the big data-based analytics service system is comprised of 17 billion linked data items, including theses, patents, web data, technical terms, product names, names of persons, organizations, and regional information in the science technology field. These data are stored in an RDBMS, and the results are analyzed through multiple processing stages provided by the service. The data used in the service are listed in Table 1.

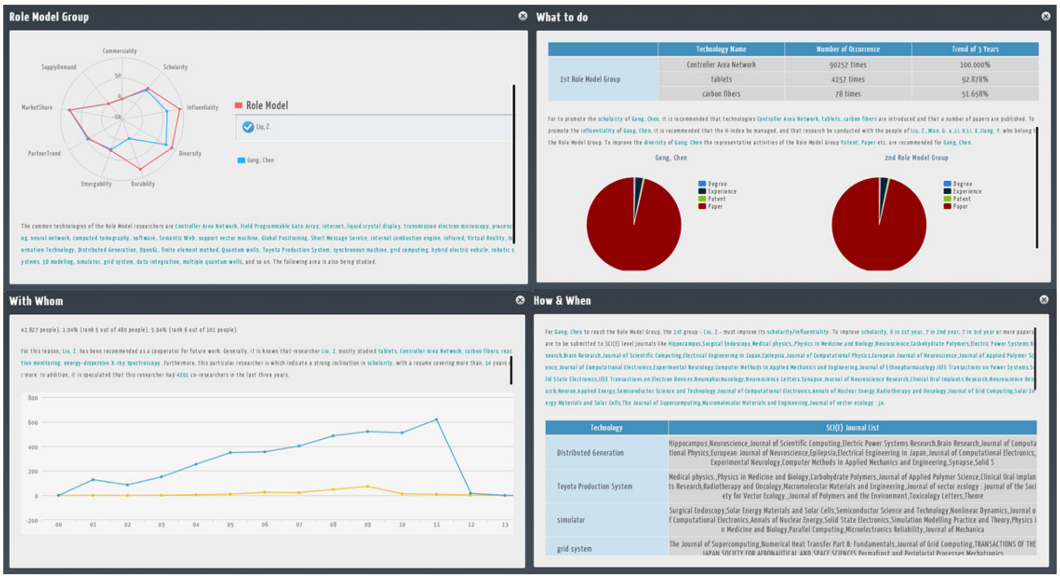
Table 1: Data resources

Item	Resource type	Number
Literature	Web documents	5,268,696
	Scientific papers	9,765,199
	US/EU/PCT patents	7,615,819
Semantic resources	Technical terms	43,201,941
	Products	62,327,156
	Persons	25,110,360
	Organizations	40,993,708
	Locations	50,350,884
Linked data	60 types such as Freebase, DBLP, PubMed, Yago	17,101,084,445

2.2 RDBMS-based analytics service system

The prescriptive analytics service system developed for enhancing a researcher's research competitiveness is composed of the descriptive analytics service and the prescriptive analytics service[3]. The analytics service system can be described as a researcher's research activity analysis system that uses prescriptive analytics.

The descriptive analytics service provides the researcher power index and the research activity history by analyzing the researcher's research information through a big data analysis of academic literature. The prescriptive analytics service presents specific research activity plans recommended for researchers, based on the future forecast results obtained through the analysis by using the 5W1H model. The system can be described as a researcher's research activity analysis system that uses prescriptive analytics[4][5].



The common technologies of the Role Model researchers are Controller Area Network, Field Programmable Gate Array, internet, liquid crystal display, transmission electron microscopy, process
ng, neural network, computed tomography, software, Semantic Web, support vector machine, Global Positioning, Short Message Service, internal combustion engine, infrared, Virtual Reality, In
ormation Technology, Distributed Generation, OpenGL, finite element method, Quantum wells, Toyota Production System, synchronous machine, grid computing, hybrid electric vehicle, robotic s
ystems, 3D modeling, simulator, grid system, data integration, multiple quantum wells, and so on. The following area is also being studied.

	Technology Name	Number of Occurrence	Trend of 3 Years
1st Role Model Group	Controller Area Network	90257 times	100.000%
	tablets	4157 times	92.878%
	carbon fibers	78 times	51.658%

For to promote the scholarity of Gang, Chen, it is recommended that technologies Controller Area Network, tablets, carbon fibers are introduced and that a number of papers are published. To
promote the influentiality of Gang, Chen, it is recommended that the H-index be managed, and that research be conducted with the people of Liu, Z.,Wan, G. x.,Li, H.,Li, X.,Jiang, Y. who belong t
the Role Model Group. To improve the diversity of Gang, Chen the representative activities of the Role Model Group Patent, Paper etc. are recommended for Gang, Chen.

42,827 people), 1.04% (rank 5 out of 480 people), 5.94% (rank 6 out of 101 people)

For this reason, Liu, Z. has been recommended as a cooperator for future work. Generally, it is known that researcher Liu, Z. mostly studied tablets, Controller Area Network, carbon fibers, reac
tion monitoring, energy-dispersive X-ray spectroscopy. Furthermore, this particular researcher is which indicate a strong inclination in scholarity, with a resume covering more than, 14 years
r more. In addition, it is speculated that this researcher had 4201 co-researchers in the last three years.

For Gang, Chen to reach the Role Model Group, the 1st group - Liu, Z.- must improve its scholarity/influentiality. To improve scholarity, 6 in 1st year, 7 in 2nd year, 7 in 3rd year or more papers
are to be submitted to SCI(E) level journals like Hippocampus,Surgical Endoscopy,Medical physics.,Physics in Medicine and Biology,Neuroscience,Carbohydrate Polymers,Electric Power Systems R
search,Brain Research,Journal of Scientific Computing,Electrical Engineering in Japan,Epilepsia,Journal of Computational Physics,European Journal of Neuroscience,Journal of Applied Polymer Sc
ence,Journal of Computational Electronics,Experimental Neurology,Computer Methods in Applied Mechanics and Engineering,Journal of Ethnopharmacology,IEEE Transactions on Power Systems,S
lid State Electronics,IEEE Transactions on Electron Devices,Neuropharmacology,Neuroscience Letters,Synapse,Journal of Neuroscience Research,Clinical Oral Implants Research,Neuroscience Res
arch,Neuron,Applied Energy,Semiconductor Science and Technology,Journal of Computational Electronics,Annals of Nuclear Energy,Radiotherapy and Oncology,Journal of Grid Computing,Solar En
ergy Materials and Solar Cells,The Journal of Supercomputing,Macromolecular Materials and Engineering,Journal of vector ecology : jo.

Technology	SCI(E) Journal List
Distributed Generation	Hippocampus,Neuroscience,Journal of Scientific Computing,Electric Power Systems Research,Brain Research,Journal of Computational Physics,European Journal of Neuroscience,Epilepsia,Electrical Engineering in Japan,Journal of Computational Electronics,Experimental Neurology,Computer Methods in Applied Mechanics and Engineering,Synapse,Solid S
Toyota Production System	Medical physics.,Physics in Medicine and Biology,Carbohydrate Polymers,Journal of Applied Polymer Science,Clinical Oral Implants Research,Radiotherapy and Oncology,Macromolecular Materials and Engineering,Journal of vector ecology : journal of the Society for Vector Ecology.,Journal of Polymers and the Environment,Toxicology Letters,Theore
simulator	Surgical Endoscopy,Solar Energy Materials and Solar Cells,Semiconductor Science and Technology,Nonlinear Dynamics,Journal of Computational Electronics,Annals of Nuclear Energy,Solid State Electronics,Simulation Modelling Practice and Theory,Physics in Medicine and Biology,Parallel Computing,Microelectronics Reliability,Journal of Mechanica
grid system	The Journal of Supercomputing,Numerical Heat Transfer Part B: Fundamentals,Journal of Grid Computing,TRANSACTIONS OF THE JAPAN SOCIETY FOR AERONAUTICAL AND SPACE SCIENCES,Permafrost and Periglacial Processes,Mechatronics

Fig. 1 Analytics service screenshot

The system is constructed using an RDBMS, and has a structure as displayed in Fig. 2. The data resources store in the Raw DB in Fig. 2. Source DB store extracted entities such as technologies, products, and persons and diverse relationships with each other. Also it is storing statistical data by using SQL commands. The various analysis components in the analytics modules perform multiple analysis operations using the data from the Source DB and Service DB. Lastly, the analyzed information is stored in the Service DB and provided to the users through the service.

The analytics service system, which uses the RDBMS, requires substantial time to process the vast amounts of data, before the analyzed results are available in the service. In particular, its processing time increases linearly as the size of data become very large and the analysis computation complexity increases. Because of the processing speed problem, the system cannot process in real-time; therefore, all services use the batch processing method which processes and analyzes the information in advance. Furthermore, if many users access the analytics service and request service results at the same time, the service response speed decreases exponentially because of the increased processing load.

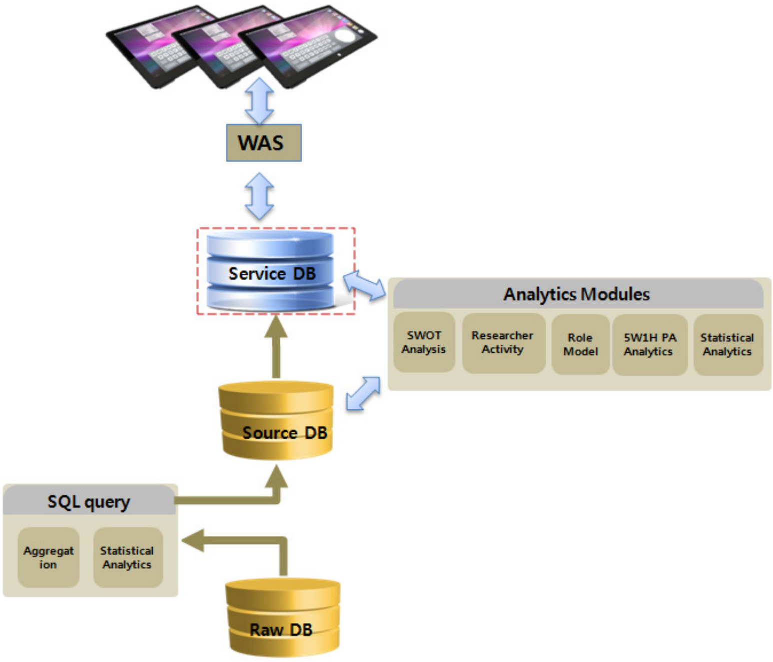


Fig. 2 The existing system architecture

3 Design of hadoop-based analytics service system

To resolve the existing system's problems, we adopted the Hadoop ecosystem. Hadoop provides a distributed parallel processing environment that is appropriate for processing massive amounts of data. We adopted MapReduce framework and Spark framework. MapReduce returns results rapidly by distributing and processing the various calculations to multiple computers. Spark framework run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk, for cretin class of iteration algorithm. But MapReduce does not efficiently support iteration (or equivalently, recursion) or certain key features. We can combine Spark and Mapreduce seamlessly on the basis of the analysis model features. We intend to change the InSciTe advisory, which has the structure displayed in Fig. 2, to the structure of the Hadoop-based system as displayed in Fig. 3. In the proposed design, the science technology literature big data currently stored in the RDBMS are stored in the Hadoop Distributed File System (HDFS), and the existing analysis service modules implemented with Java and SQL are implemented with MapReduce and Spark in Hadoop. This composition makes it easy to express a wide array of computations, including iterative algorithm, streaming data, complex queries, and batch.

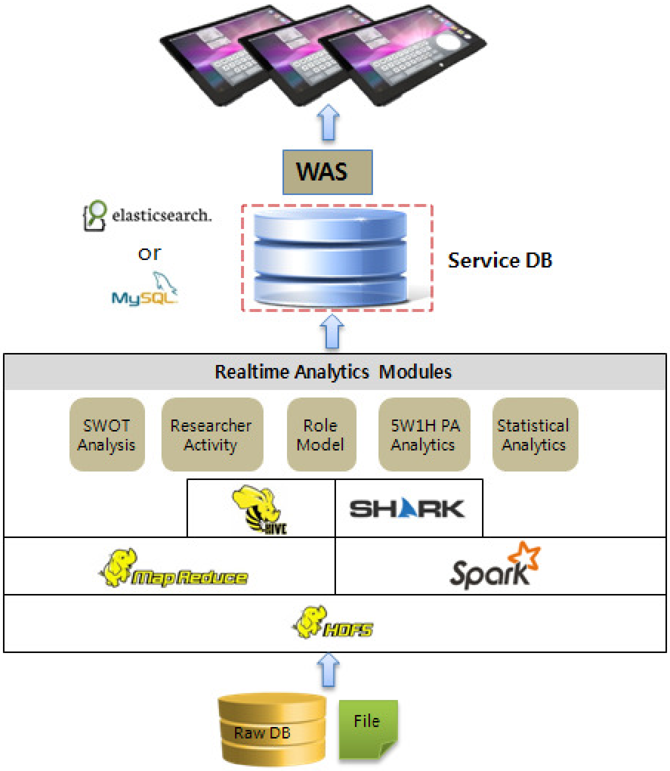


Fig. 3 Analytics service system structure

The analytics service system's new design plan has Real-time analytics modules, which are composed of three layers as described below.

- Distributed filesystem layer: Literature big data like web documents, papers, and patents in the RDBMS or Files are imported to the HDFS.
- Distributed computing layer: This layer anlayzes meaningful relationship such as technologies, products, and persons. For example, number of person related technology each year and number of jnoural papers of all possible person and technology each year. It provides the necessary data from Analytics layer. This layer process by using MapReduce and Spark. Spark is emerging distributed computing layer with in-memory /realtime processing capabilities. Hive, Shark(100% hive compatible) provide SQL layer on Hadoop. Spark will help to implement basic ETL(Extract, Transform, and Load) and statistical analysis.
- Analytics layer: It is composed of five analysis modules – SWOT analytics, researcher activity, role model, 5W1H PA analytics, and statistical

analytics. Analytics routine will be written in HQL(Hive Query language both run on Hive/Shark), Java(for MapReduce), Scala(for spark).

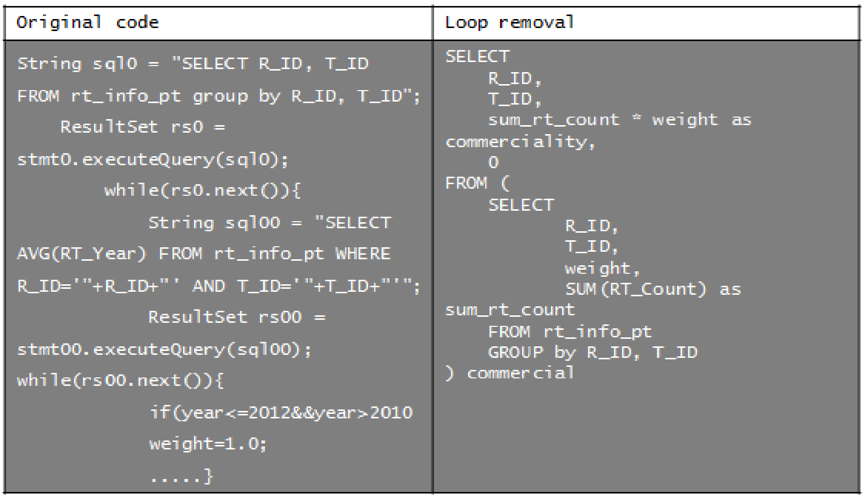
The Service DB in Fig. 3 lengthens the service response time when multiple users access the service, or massive data must be processed. To resolve this problem, we are considering two plans: using an indexing engine and optimizing the structure of the existing DBMS. Whereas the use of the indexing engine facilitates distributed processing of massive data, complex computational functions used in the service cannot be supported.

4 Implementation

We created a parallel distributed analysis environment for analyzing big data, and using this environment we developed a Hadoop-based analysis service system that improves the performance of the existing analysis module and UI. The implementation method is as follows: Firstly, we created a distributed parallel analysis environment, wherein existing text and RDB data can be processed. Secondly, we optimized the module after re-implementing the algorithm of the analysis model to improve speed. Thirdly, we measured the performance in comparison with existing systems to confirm the improvement in the speed of the distributed parallel environment-based analysis model.

We optimized the existing analysis module and edited the source code to a form that can facilitate the implementation of MapReduce. Based on the analysis results of existing modules for parallelizing and optimizing the analysis module, we performed the following: (1) Removed the Loop, (2) Removed the unnecessary Insert and Update query, (3) Removed Union query, (4) Used Hive Transform script, and (5) Wrote with the Spark code.

In detail, whenever a query is executed, query parsing, logical plan generation, physical plan generation, execution, and result fetch steps are repeated. The iteration method has a very weak structure in MapReduce and Hive, which have the advantage of parallel processing; thus, to resolve this, an unnecessary loop was removed. Then, a whole table, which calculates the results by removing the insert query and update query that are repeatedly applied to each row in the existing module, was inserted. In addition, the method was changed to a single query method instead of using a Union query, which executes the query that queries the table by dividing it several times. Coding was performed by using a transformation script to facilitate the implementation by Hive Query, and the analysis module was optimized by using SPARK code in the part that was not optimized by Hive. Parallelization was performed using SPARK application or HiveQL script according to the characteristics of each module.



Original code

```
String sql0 = "SELECT R_ID, T_ID
FROM rt_info_pt group by R_ID, T_ID";
    ResultSet rs0 =
stmt0.executeQuery(sql0);
        while(rs0.next()){
            String sql00 = "SELECT
AVG(RT_Year) FROM rt_info_pt WHERE
R_ID='"+R_ID+"' AND T_ID='"+T_ID+"'";
            ResultSet rs00 =
stmt00.executeQuery(sql00);
while(rs00.next()){
            if(year<=2012&&year>2010
            weight=1.0;
            .....}
```

Loop removal

```
SELECT
    R_ID,
    T_ID,
    sum_rt_count * weight as
commerciality,
    0
FROM (
    SELECT
           R_ID,
           T_ID,
           weight,
           SUM(RT_Count) as
sum_rt_count
    FROM rt_info_pt
    GROUP by R_ID, T_ID
) commercial
```

Fig. 4 An example of loop removal

6 Benchmark Test

The performance speeds of the analysis module of an RDBMS-based system and a Hadoop-based system were compared.

The experiment environment is as follows: a Hadoop cluster of Hadoop 2.x (2 Namenodes, 6 Datanodes) was used for the Hadoop-based system. In addition, Hive 0.10 and Spark 0.9.1 versions were used in the development. In the RDBMS-based system, the measurement was performed using MySQL 5.1.73 in an environment of Xeon 6 core, 12-thread CPU, and 24 GB memory.

As shown in Figure 5, the test results indicated that Inner SWOT, EnvSWOT, RoleModel, WhatToDo_Tech, WithWhom, and WhoWhen modules showed a performance improvement by 145 times, 89 times, 2.2 times, 851 times, 272 times, and 580 times, respectively; overall, the execution time was reduced by 127.87 times.

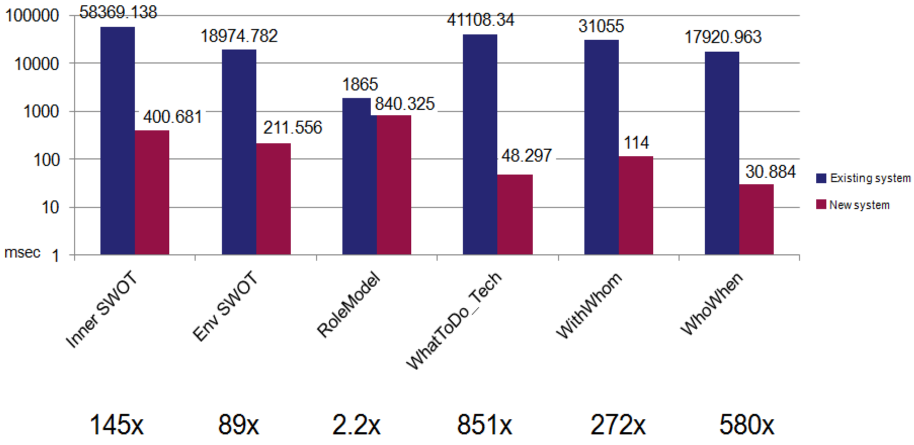


Fig. 5 Result of the benchmarking

6 Conclusion

This study explained the method of changing an RDBMS-based analysis service system to a distributed parallel-based environment system to address the problems encountered during the processing of big data. We optimized the system by using Hadoop ecosystem to improve the performance while processing big data. In addition, as a method to resolve the problem of processing speed, the existing module was implemented with Hive and Spark for optimizing and parallelizing the SQL query. We compared the performance results of RDMS-based analytics service system and Hadoop-based analytics service system and confirmed that the Hadoop-based analytics service system showed a performance improvement by 127.87 times.

ACKNOWLEDGEMENTS

This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers]

References

[1] "Big data". From Wikipedia, http://en.wikipedia.org/wiki/Big_data

[2] "Apache Hadoop". From Wikipedia, http://en.wikipedia.org/wiki/Hadoop

[3] InSciTe Advisory service system, http://inscite-advisory.kisti.re.kr/search

[4] S. Song, D. Jeong, J. Kim, M. Hwang, J. Gim, H. Jung. 2014. Research Advising System based on Prescriptive Analytics. In Proceedings of the International Workshop on Data-Intensive Knowledge and Intelligence in conjunction with the 9th FTRA International Conference on Future Information Technology.

[5] M. Lee, M. Cho, J. Gim, D. Jeong, H. Jung. 2014. Prescriptive Analytics System for Scholar Research Performance Enhancement. HCII 2014. Part1. In Proceedings of the Communications in Computer and Information Science 434. p186~190.