

Design of Educational Big Data Application using Spark

Min-Seon Lee, Earl Kim, Choon-Sung Nam, Dong-Ryeol Shin

Department of Information and Communication Engineering, Sungkyunkwan University

2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do, Korea

mslee01@skku.edu, wise0314@skku.edu, namgun99@gmail.com, drshin@skku.edu

Abstract—Educational institutions utilize various academic systems. Systems related to various tasks such as academic administration, general administration, and subsidiary institution administration are used. These systems provide information with complex and diverse attributes. Students and faculty want to get the information quickly. Also, systems are important for providing services for students and faculty to use their academic information.

The academic information has large datasets, among which lecture information can be used to find students' learning patterns. So we use an FP-Growth algorithm that compresses data of frequent items into a frequent pattern tree and then divides the compressed data into a set of related condition data for one frequent item and mining it separately does.

Keywords—Spark, MLlib, Recommendation Algorithm, Big Data

I. INTRODUCTION

Currently, educational institutions such as universities are building and operating various systems related to work such as academic, administrative, research, and industry-academia cooperation.

These systems are designed to enhance the effectiveness of work efficiency by directly supporting the integrated information system that directly supports the task process, reflecting the characteristics of each task. In terms of system architecture, data is structured in terms of academic information, administrative information, research information, online education, library information, dormitory information, housing information, and employment information, but data management is not managed in the system but is not structured in the context of each system. So, in this thesis, there's a bunch of different kinds of unstructured data that are generated by a variety of devices and systems. Provides a means of integrating the system into a management system and performs analysis based on the machine learning of these data. In addition, the current use of the Big Data platform offers data collection, analysis, and processing methods for analysis, analysis, and processing of Big Data platforms that can not be performed and analyzed for future use. we will build a big data platform for data analysis and use a spark, a framework for improving distributed computing. Spark maintains data in Resilient Distributed Dataset (RDD), an abstraction of the dataset, to In-Memory to improve

performance. Spark also overcomes the disadvantages of disk I/O and network I/O caused by repetition of write and read when performing Iterative Algorithm. We analyze various data attributes of academic information based on Big Data Platform. In addition, we propose architecture design of academic Information System that provides specialized services such as analyzing patterns of students' various learning information using MLlib of the machine learning library in spark and predicting and recommending the courses to be taken in the next semester.

In this paper, Section 2 describes the Spark, FP-Growth algorithm, Academic system Architecture and Section 3 describes Our proposed system design in this paper, and Section 4 presents the conclusion.

II. RELATED WORK

Apache Spark is a distributed computing framework similar to MapReduce but maintains data in Resilient Distributed Dataset (RDD), an abstraction of the data set in In-Memory, to improve performance.[1]

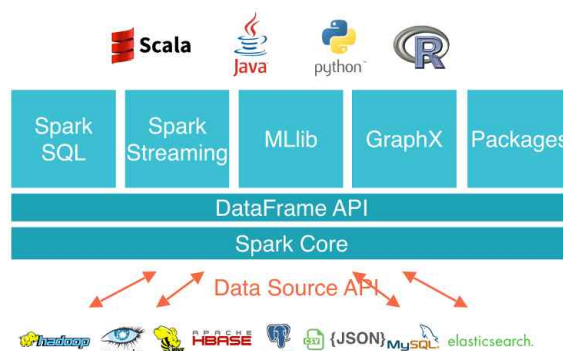


Figure 1. Basic structure of platform

As shown in Figure 1, it is useful for algorithms that perform iterative operations and data flow processing. When Iterative Algorithm is performed in Hadoop, it overcomes the disadvantages of disk I/O and network I/O due to the repetition of writing and reading in HDFS every iteration.

Spark has been developed as Scala and uses Spark-specific Scala Interpreter, and provides a high-level instruction set for parallel operations such as Map, Filter, Join, etc., to facilitate development. Spark provides Shark, an interactive query

analyzer, Bagel, a high-volume graph processing and analyzer, Spark Streaming, a real-time analyzer, and Mllib, a machine learning library, and provides a programming paradigm Easy for applying[2].

Mllib is a subproject inside Spark, like Spark SQL and Spark Streaming. As the name implies, it was made for a machine learning. There are many algorithms for support and programming is also relatively easy to apply. It is important to select the applicable algorithm when the data and objects to be analyzed are selected. To do so, appropriate algorithms should be applied through algorithm selection methods[3].

The FP-Growth algorithm was an algorithm to improve the Apriori algorithm and required a scan to create a candidate for the Apriori algorithm and to obtain the support. This may occur at least one count of the maximum number of transactions in order to scan the DB for the scan. However, the FP-Growth algorithm has reduced the number of DB to read twice, and another major feature is not creating a pattern candidate. Instead, use a data structure called tree and node links. Create a new Conditional tree with the minsup of the transaction by increasing the number of recurring patterns that occur frequently, resulting in the creation of a new item in the transaction, using the Divide and conquer technique.

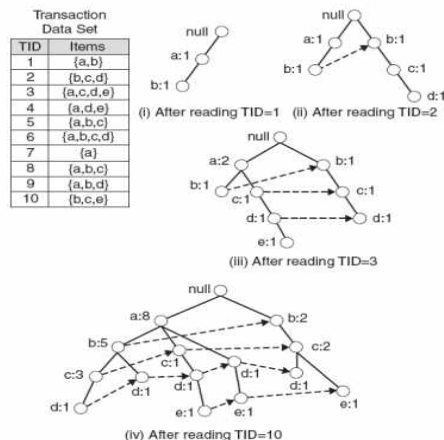


Figure 2. FP-Tree Construction

Figure 2 shows Nodes correspond to items and have a counter. FP-Growth reads 1 transaction at a time and maps it to a path. Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix). In this case, counters are incremented. Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines). The more paths that overlap, the higher the compression. FP-tree may fit in memory. Frequent itemsets extracted from the FP-Tree[4].



Figure 3. Academic system example

As shown in Figure 3, the Academic system is based on a variety of work-related unit systems. It is an integrated information system that manages information related to registration and scholarship, grades, attendance, classes, service, finance, and students and staff. However, in data management aspect, there are problems that unstructured data for each system or task is used in business without being managed in the system.

III. SYSTEM DESIGN PROPOSAL

Table 1 shows the items and properties of the item that can be used to collect structured data. These data are information to design recommendation service through machine learning. In this paper, we design a recommendation model as lecture information.

Table 1. Collection data item

Item	Property value
Student Information	Department, student, year, statement, grade, scholarship, class
Curriculum Information	Class number, lecture id, lecture name, Department of Education, Field
Dormitory Information	Entry period, registration period, Room, Resident
Attendance Information	Access, schedule, classroom, teacher

As you can see in Figure 4, the various data are delivered in HDFS according to each attribute. The structured data is transferred from the RDBMS to HDFS using SQL-to-Hadoop (Sqoop). Sqoop is a tool designed to transfer large amounts of data between structured Data sources, for example, RDBs such as Oracle. Import metadata into Hadoop through Sqoop import process and transfer data to Hadoop Cluster with Map-Only Job[5].

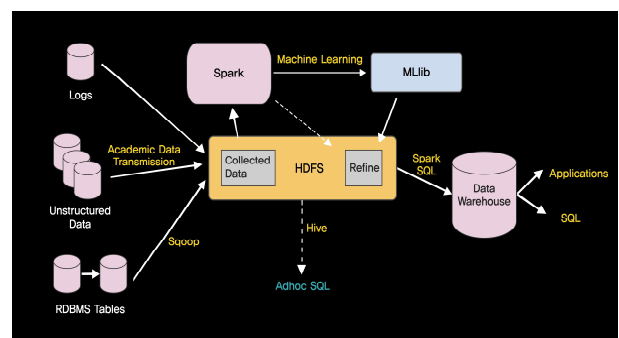


Figure 4. system design suggestion

Among these collected data, lecture data is an important item, and the FP-Growth algorithm is performed using Mllib, a spark machine learning library. The resulting data can be used to identify patterns of lecture data that students have taken for the year and semester. These patterns include pattern information for students' preferred lectures, and based on this,

they can implement a recommendation system that recommends lectures to students who have not yet taken lectures. In addition, by using the data collected from the sensor information of the classroom attendance and the dormitory entrance information, it is possible to determine the population of the students and predict the density of the population and to control the temperature of the classroom and buildings by pattern analysis.

Figure 5. UI Implementation

As shown in Figure 5, after the structured data is collected from the user, we have implemented the Academic system user interface(UI) screen to provide services in the actually designed application. The Structured data is stored in the RDBMS and is in a state suitable for use.

IV. CONCLUSIONS

In this paper, we use the Hadoop ecosystem to process Big Data and Spark for machine learning. The data stored in HDFS can be analyzed using Spark's in-memory function. For machine learning, a prediction model can be constructed using MLlib, a Spark library.

The collected data can be used to create a prediction model that recommends the courses that will be taken in the next semester to the students who have taken the data by using Spark's MLlib. This approach is based on the evolution of recommendation systems. It is not just a prediction based on the roadmap of the major, but it provides more personalized information through the subjective inclination of students such as lecture evaluation and processing of various unstructured data. It can serve as a basis for providing and enhancing learning skills.

As a future work, we will use the data collected from the sensors to monitor classroom attendance information and dormitory entrance information to determine the number of student students. In addition, we will perform various comparative data analysis that can control the temperature of the classroom and the buildings of the building through pattern analysis by predicting population density.

ACKNOWLEDGMENT

This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW(R22151610050001002) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

- [1] Xin, Reynold S., et al. "Graphx: A resilient distributed graph system on spark." First International Workshop on Graph Data Management Experiences and Systems. ACM, 2013.
- [2] Mitchell T. Machine learning, McGraw Hill, 1997
- [3] The Apache Foundation, Mahout (2014), Retrieved Oct. 1, 2014, from <http://mahout.apache.org>.
- [4] Qiu, Yong, Yong-Jie Lan, and Qing-Song Xie. "An improved algorithm of mining from FP-tree." Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on. Vol. 3. IEEE, 2004.
- [5] Ting, Kathleen, and Jarek Jarcec Cech. Apache Sqoop Cookbook. " O'Reilly Media, Inc.", 2013.



Min Seon Lee received the B.S degree in the Department of Computer Engineering from Sungkyunkwan University, Korea, in 2008. Currently, he is working for M.S. in Department of Computer Science and Engineering at Sungkyunkwan University, Korea. His research interests include Machine learning and Big data Platform.



Earl Kim received the B.S degree in the Electronic Engineering from Kyung-Hee University, Korea, in 2015. Currently, he is working for M.S. in Department of Computer Science and Engineering at Sungkyunkwan University, Korea. His research interests include Machine learning and Big data.



ChoonSung Nam is currently a researcher in Sungkyunkwan University College of Information & Communication Engineering. He got his Ph.D degree from Sungkyukwan University in August, 2011. his Master's degree from SoongSil University, and his Bachelor's degree from SangMyung Univeristy, Korea. Dr. Nam is the author co-author of over 20 papers in sci(e) journals, domestic journals and conferenees. His interested area is UAVs, Machine-learning and IoT.



Dong Ryeol Shin received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from the Sungkyunkwan University in 1980, the Korea Advanced Institute of Science and Technology (KAIST) in 1982, and the Georgia Institute of Technology in 1992, respectively. During 1992-1994, he had worked for Samsung Data Systems, Korea, where he was involved in the research of Intelligent Transportation Systems. Since 1994, he has been with the Department of Computer Science and Engineering at Sungkyunkwan University where he is currently a Professor in Network Research Group. His current research interests lie in the areas of mobile network, ubiquitous computing, cloud computing, and bioinformatics. And he is actively involved in the security of vehicular area networks, and the implementation and analysis of big data platform, applicable to 3D image processing of robotic arms.