

A Design for Big Data Analysis System

Jin MA*, Jong-Suk Ruth LEE*, Kumwon CHO* and Minjae PARK **

* National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information,
245 Daehak-ro, Yuseong-gu Daejeon, South Korea

**BISTel, Inc., BISTel Tower, 128, Baumoe-ro, Seocho-gu, Seoul, South Korea

majin@kisti.re.kr, jsruthlee@kisti.re.kr, ckw@kisti.re.kr, mjpark@bistel.com

Abstract— Nowadays, the importance of big data processing has been mentioned. So, we propose a data analysis system for big data processing. The proposed system is considered the processing of GPDB and HDFS, and also supports the processing of files, and a relational database. This paper describes a conceptual analysis of the big data processing system, and provides the processed data and display the results.

Keywords— Analysis System, Big Data, Manufacturing Data

I. INTRODUCTION

Recently, the importance of Big Data processing [1] is emphasized. As a result, many companies are trying to big data processing, and provide many solutions. Besides, the analysis system is a field that is closely related to the big data. In addition, data analysis system [2] has been used in many industrial groups; in particular, we expect an improvement in manufacturing productivity through data analysis.

In this paper, to handle efficiently a large-scale data, and provides an analysis system that can be applied to big data platform [3, 4].

II. ANALYSIS SYSTEM DESIGN

The proposed system will be divided into three types, as shown in Figure 1. There exist a Client, Server, and Data Sources.

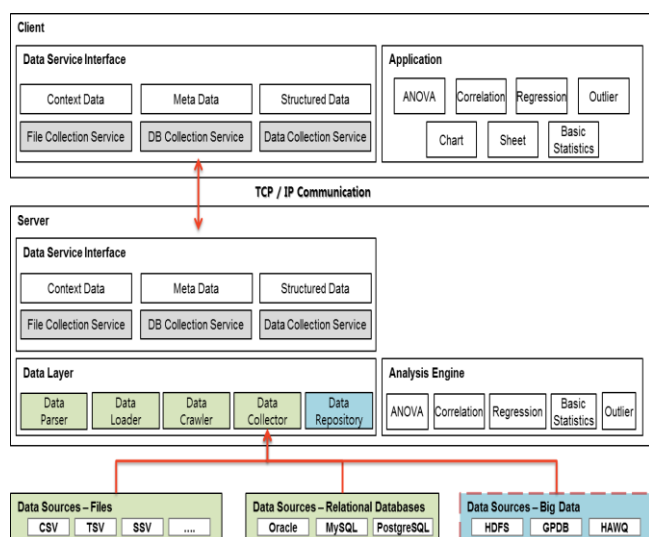


Figure 1. Design of the Analysis System

A. Client

The client provides a user interface for analysis and data services interface Element is provided as follows:

- **File Collection Service:** It provides the interface associated with the files. File Collection service module that provides an interface between any files such as CSV, TSV, SSV using the XML configuration.
- **DB Collection Service:** It provides the interface associated with the Databases. DB Collection service module that provides an interface between any storages such as rdb, odb using the XML configuration.
- **Data Collection Service:** It provides the interface associated with the big data. Data Collection Service module that provides an interface between HDFS, GPDB, HAWQ using the XML configuration.
- **Meta Data:** Maintains data access profiles for collecting data.

B. Server

The server provides Relational data Bases, Files and Big data analysis system. Important elements of the server are as follows:

- **Analysis Engine:** Plug-in system that provides analytical capability for data processing. (Data Mining, Basic Statistics, Regression, Correlation, ANOVA, Outlier etc...).
- **Data Layer:** Data layer takes a role of providing Storage for the proposed system. And manages data and solves heterogeneity problems. It manages information that was collected from distributed heterogeneous source, and notifies to data service interface.

C. Data Sources

The data source is configured as follows:

- **Files:** Comma-Separated Values (CSV), Tab-Separated Values (TSV), space-separated values' (SSV).
- **Relational Databases:** Oracle DB, MySQL, MS-SQL, PostgreSQL, etc.
- **Big data:** Hadoop Distributed File System (HDFS), GreenplumDatabase (GPDB), SQL on Hadoop analytics engine (HAWQ).

III.CONCEPTUAL PROCESS FOR BIG DATA ANALYSIS SYSTEM

We describe briefly the analysis system flow. For ease of understanding, Figure 2 compares our system flow with that of Google.

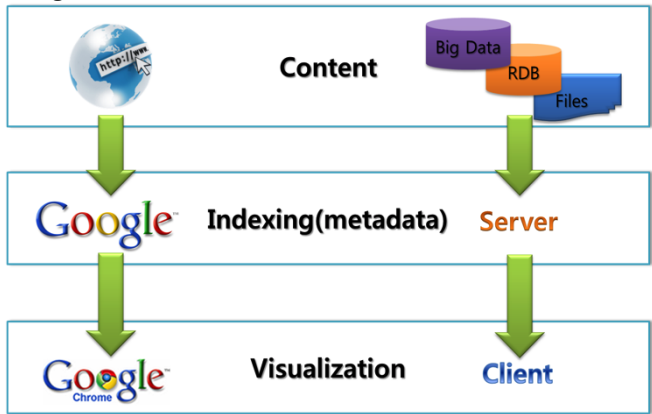


Figure 2. Comparison of Google and the proposed system

Figure 3 shows the analysis system’s indexing operation. Different data sources exist at a various locations. A data crawler is an automated service that methodically scans data sources for new or updated content and indexes them in a fast storage system for later asynchronous access. Client that can be accessed in second search, the user inputs their criteria for data searching. The second search returns a list of contexts that match the criteria and data from all related sources. The user can trigger a task to obtain the original source of the content for more detailed information.

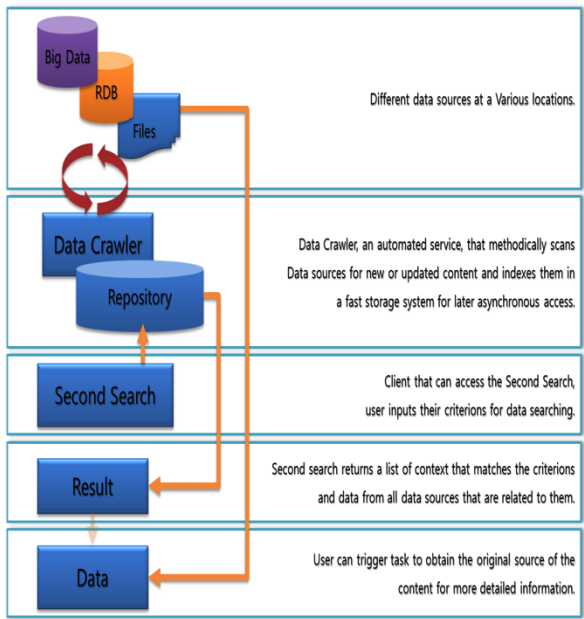


Figure 3. Analysis system flow during an indexing operation

After metadata generation, analysis is performed by the analysis engine. Within the data server component, a module called the analysis engine turns data into meaningful results. The role of the analysis engine is as follows:

- Data mining
- Common statistical analysis
 - Correlation
 - Regression
 - ANOVA
 - Summarization (max, min, mean, etc.)
 - Outlier detection

When the analysis is completed the results can be visualized.

A. Scenario of analysis system indexing

Table 1 shows applicable domains and context data sources with use, case example.

In this paper, we applied to semiconductor domains.

TABLE 1. APPLICABLE SCENARIO OF ANALYSIS SYSTEM INDEXING

Analysis Domain	Semiconductor	Solar
Use case example	•Yield Loss analysis •Tool control •Recipe management •Inventory control	•Efficiency analysis •Tool control •Inventory control
Contexts	Lot and Wafer	Sheet and Panel
Data Sources	•Tracking •Metrology •Defect •Recipe •Process (Trace) •Alarm •Tool maintenance •WET •Sort Test •Packaging •Final Test	•Tracking •CD •Recipe •Process (Trace) •Defect •IV •Packaging

B. Configuration for Data Integration

In this Section, the description of the configuration files for data integration. Each module that was introduced in Figure 1 has a configuration file. But among them, it will introduce the important settings that are used for data integration. First, DB Configuration files for connection to the information of heterogeneous data sources..

First, DB Configuration files for connection to the information of heterogeneous data sources. Figure 4 is a configuration file for the oracle DB. If you want to add other kinds of Storage are adding a configuration file, enter the appropriate information.

Second, Figure 5 is Information crawling to collect of metadata. Meta information was created based on the information used in the semiconductor manufacturing process.

Third, Collector of Figure 6 is Data gathering and Second Searching Crawler Filter metadata to get data.



Figure 8. Analysis Result : Q-Q plot

IV. CONCLUSIONS

We have applied a data analysis system for big data. The Analysis system can be applied in manufacturing as well as in other areas. Using the analysis system can find the cause of the defects in processes and increase productivity.

In the manufacturing industry, the proposed system unifies and normalizes all outputs and configurations, reducing the learning curve and enhancing productivity.

Since the proposed system is applied to the well-known big data processing method: Further work is required to improve performance.

ACKNOWLEDGMENT

This research was supported by the EDucation-research Integration through Simulation On the Net (EDISON*) program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2011-0020576).

REFERENCES

- [1] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, Vol. 36 No. 4, pp. 1165-1188, Nov. 2012.
- [2] Agresti, A., & Kateri, M., "Categorical data analysis", pp. 206-208. Springer Berlin Heidelberg, 2011.

- [3] AGRAWAL, Divyakant; DAS, Sudipto; EL ABBADI, Amr. "Big data and cloud computing: current state and future opportunities". In: *Proceedings of the 14th International Conference on Extending Database Technology*, p. 530-533, ACM, 2011.
- [4] RUSSOM, Philip, et al. "Big data analytics", TDWI Best Practices Report, Fourth Quarter, 2011.
- [5] M. Purdy, "Dynamic, weight-based sampling algorithm," in *Proc. ISSM Int. Symp. Semicond. Manuf*, pp. 1-4, Oct. 2007.
- [6] Regression analysis, *Encyclopedia of Mathematics*. Springer. 2001. ISBN 978-1-55608-010-4.
- [7] Wang, K., and Tsung, F. "Run-to-run process adjustment using categorical observations", *Journal of Quality Technology*, Vol.39, No.4, October 2007.



Jin Ma is a Researcher at the Supercomputing R&D Center of Korea Institute of Science and Technology Information (KISTI), South Korea. He received B.S., M.S. degrees in computer software and computer science from Kwangwoon University in 2010, 2012, respectively. He had worked as member of research staff at the solution R&D research center of BISTel, Inc., South Korea. His research interests big data, analysis system and visualization, distributed system.



Jongsuk Ruth Lee received her Ph.D. in Computer Science from Univ. of Canterbury, New Zealand. She is a principal researcher at at National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information (KISTI) and an adjunct faculty at Univ. of Science & Technology of Korea. Her research interests are smart learning, parallel/distributed computing & simulation, and big data handling.



Kum Won Cho received his Ph.D. in Mechanical(Aerospace) Engineering from KAIST, Korea. He is a head of Supercomputing R&D center and a director of EDISON (Education-research Integration through Simulation On the Net) Center, National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information (KISTI), Korea.



Minjae Park is a senior member of research staff at the solution R&D research center of BISTel, Inc., South Korea. He received B.S., M.S., and Ph.D. degrees in computer science from Kyonggi University in 2004, 2006, and 2009, respectively. His research interests include groupware, workflow systems, BPM, CSCW, collaboration theory, process warehousing and mining, workflow-supported social networks discovery and analysis, and process-aware factory automation systems.