# An Experimental Analysis of the Real Time Big Data Platform Performance based on the ParStream

[1]Wonjung Jang and [*2] Geunbong Park
[1] *Goodmorning Information Technology, Big Data Team, wjjang@goodmit.co.kr*
[2] *Goodmorning Information Technology, Big Data Team, kbpark@goodmit.co.kr*

## *Abstract*

*Recently, the amount of data has been increasing rapidly and data forms have also been diversifying. Although data is loaded into Hadoop as a big data technology, it is hard to make practical use of it due to its slow processing speed. To resolve this problem, this study conducted a comparative analysis on the data processing performance of the big data platform based on ParStream available to realtime analysis. When source data was loaded into ParStream and an abnormal pattern was detected, the data could be checked in real time and the complexity could be decreased by performing all the work in a single database. The findings of the comparative analysis reveal a 300 times improvement in real time processing speed, compared to the existing method.*

***Keywords:*** *Real Time Big Data, ParStream,* Experimental Analysis, Big Data Platform Performance

## 1. Introduction

Various technologies based on open source have been developed and applied to deal with a rapid increase of data and an increasing demand for faster data processing in a data processing industry. Hadoop is representative of these technologies. Even though many eco-friendly systems for data collection, loading, analysis and application have been developed, there are still problems to be solved such as technological complexity, difficulty in attracting engineers, increased operating expenses and slow processing speed.

To establish a traditional BI/DW system, the Level 2 Mart is built from the source data and aggregated information is used for visualization. However, a waste of unnecessary time and hardware resources occurs since the Level 2 Mart is built without a direct use of source data due to the processing speed.

The standard API (JDBC, ODBC, etc.) and SQL needs to be provided to link with the existing solutions (BI, ETL, applications, etc.) and a demand for advanced analysis functions linked with data analysis tools (R, etc.) is increasing.

This study conducted a comparative analysis on the processing performance of the existing big data processing technologies and the real time big data platform, the ParStream. Thus, we investigated the processing performance of the big data platform for real time analysis to resolve the processing speed problems and reduce a waste of unnecessary resources by applying a pilot to several industry fields of the ParStream, the big data platform for real-time analysis and analyzing the cases. For a pilot application test, problems of the machine sensor data analysis system developed by S Electricity were confirmed and the SPC Trend Rules applied to the data pattern analysis were applied, so that we could conduct a comparative analysis on the results of the PoC test that the ParStream was applied under the same conditions. A real-time retail data processing system of E-Mart in the distribution field was also analyzed and the results of the PoC test applied the ParStream under the same conditions were analyzed. Lastly, an analysis on a traffic data system for the improvement of the daily batch-processing speed in D City was conducted, and the results of the PoC test applied the ParStream under the same conditions were analyzed.

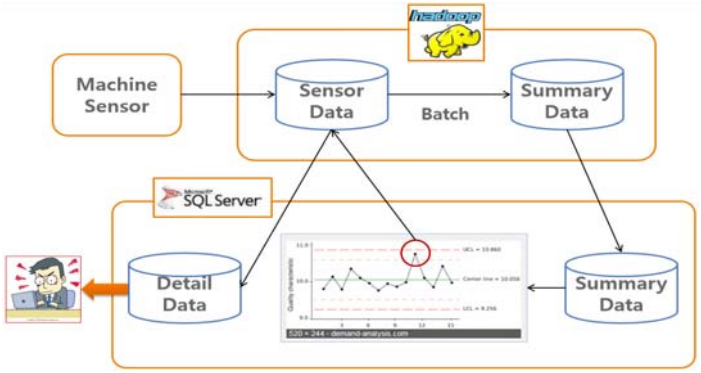
* Corresponding Author
Received: Jan.17, 2015, Revised Feb.12, Accepted: Apr. 02, 2015

## 2. An Experimental Analysis of Real Time Big Data Based on ParStream

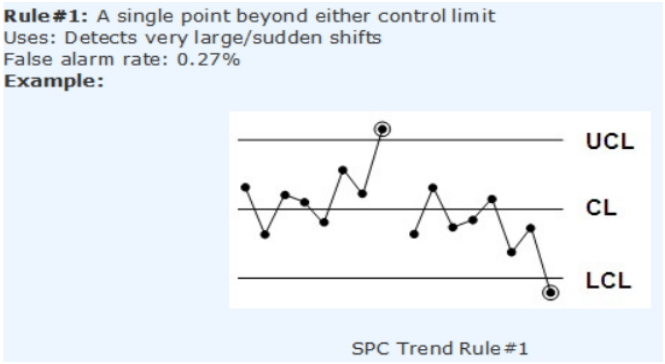
### 2.1. Data processing PoC on machine sensor of S Electricity

For a real-time analysis of machine sensor data, S Electricity loads data into Hadoop, and also into MS-SQL after generating aggregate data in batch, to make use of it for data analysis. Rule#1 and Rule#4 of the SPC Trend Rules are used for a data pattern analysis When an abnormal pattern is detected in an aggregate data of MS-SQL in RDBMS, it takes about 10 minutes to check the loaded data in Hadoop. Additionally, data is subject of a distributed management in Hadoop and MS-SQL in RDBMS, having a complexity for management.

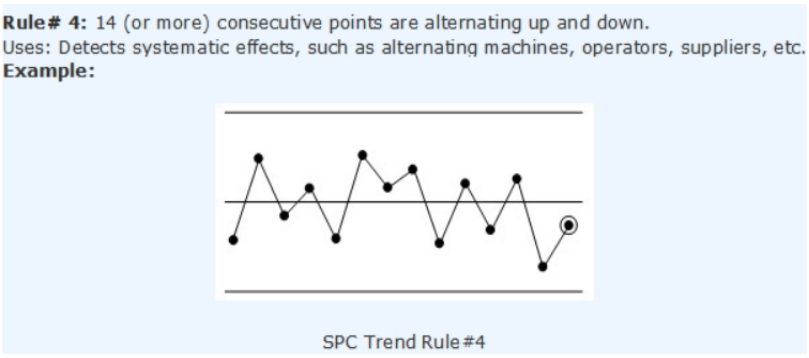
Meanwhile, Rule#1 and Rules#4 of the SPC Trend Rules are applied for analyzing a data pattern in the machine sensor and the Picture below illustrates this.




**Picture 1**. Data processing flow of Hadoop big data platform




**Picture 2**. SPC trend rule # 1




**Picture 3**. SPC trend rule # 4

About 101 billion source data collected for 8 days was loaded for PoC and a processing flow applied the ParStream is as follows.

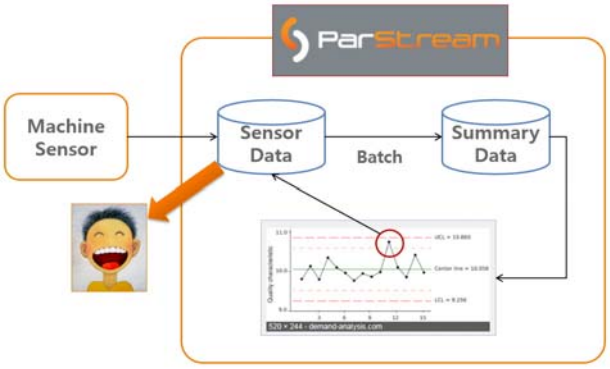
**Table 1**. Daily data generation rate

| *Section* | *The number of generated data* |
|---|---|
| 1 lot | 700 items, 1,800 Mac IDs 1,260,000 (700*1,800) |
| Day | 1,000 lot formation 1,260,000,000 = 1,260,000 * 1,000 |

Unlike the existing Hadoop platform in which source data is loaded into Hadoop and aggregate data into MS-SQL in RDBMS for a data pattern analysis, source data is loaded into the ParStream and aggregate data is generated and used for a data pattern analysis. When detecting an abnormal pattern, it is possible to check data in real time. As all the work can be performed in a single database in this way, the complexity can be reduced.

**Tanle 2**. Data processing flow of the ParStream big data platform

| *Test Item* | *Success/ Fail* | *Duration* | *The Number of Related Data* | *Tool* | *Remarks* |
|---|---|---|---|---|---|
| Measurement data loading | Success | loaded 350,000 data per sec | 10,155,600,000 | SQL, python | loading of 10.1 billion data of 600 GB for 8 hours |
| 1 lot summary | Success | within 2 sec | 112,700 | SQL, Shell | |
| SPC Rule 1 | Success | within 1 sec | | SQL, python | |
| SPC Rule 1 | Success | within 1 sec | | SQL, python | |
| 1 lot conditional search | Success | within 0.1 sec | 1,800 data among 10.1 billion data | SQL | |
| Conditional search 1 | Success | within 2 sec | 1,727,000 data among 10.1 billion data | SQL | |
| Conditional search 2 | Success | within 15 sec | list up of 1,727,000 data among 10.1 billion data | SQL | Searching for daily data in measurement item and channel conditions from each lot id |
| Conditional search 3 | Success | within 3 sec | 57,474,000 data among 10.1 billion data | SQL | Searching for daily data in measurement item and channel conditions |
| Conditional search 4 | Success | within 3 sec | 5,211,000 data among 10.1 billion data | SQL | Searching for daily data in measurement item conditions from each lot id |
| Conditional search 5 | Success | within 5 sec | 121,590,000 data among 10.1 billion data | SQL | Searching for daily data in channel conditions from each lot id |

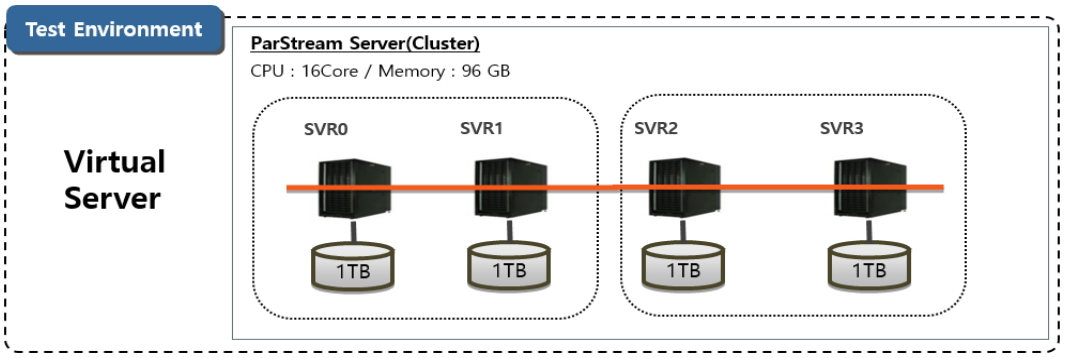



**Picture 4**. PoC results

It is confirmed that the results of manufacturer PoC using the ParStream platform showed a 300-time improvement in speed, compared to the cases using the existing Hadoop platform.

## 2.2. Real-time retail data processing PoC of E-Mart

This study conducted a comparative performance test using the ParStream and a No-SQL solution based on open sources for a real-time analysis of retail data. The test environment is as follows.




**Picture 5**. Performance test environment for retail big data processing

Meanwhile, test scenarios for a performance comparison are as follows.

**Table 3**. Performance test scenarios

| *No* | *Scenario* |
|---|---|
| 1 | 300GB Group By Distinct Count<br>1TB Group By Distinct Count |
| 2 | 300GB Group By Distinct Count<br>1TB Group By Distinct Count |
| 3 | 300GB Group By Distinct Count<br>1TB Group By Distinct Count<br>*Simultaneous performance of 10 |
| 4 | 300GB Group By Distinct Count<br>1TB Group By Distinct Count<br>*Simultaneous performance of 30 random queries |

The PoC results according to test scenarios are as follows.

| | Query | Run First | Run Second |
|---|---|---|---|
| Scenarios1 | 300 Gb Group By Distinct Count | 0.638 | 0.067 |
| Scenarios2 | 300 Gb Group By Distinct Count | 0.703 | 0.063 |
| Scenarios3 | 300 Gb Group By Distinct Count | 6.423 | 2.119 |
| Scenarios4 | 300 Gb Group By Distinct Count | 7.143 | 3.503 |
| Scenarios1 Query | **SELECT** SHIPDATE, RETURNFLAG , LINESTATUS,<br>**COUNT(DISTINCT ORDERKEY)** AS ORDERKEY<br>**FROM** moneymall<br>**WHERE** SHIPDATE BETWEEN date'1992-10-02'<br>AND date'1992-11-01' AND RETURNFLAG='A'<br>AND LINESTATUS='F'<br>**GROUP BY** SHIPDATE, RETURNFLAG , LINESTATUS; | | |
| Scenarios3 | Run 10 scenario1 style query simultaneously with different condition | | |
| Scenarios4 | Run 30 scenario1 style query simultaneously with different condition | | |
| * Import : 50 / 100Gb (6 billion rows) | | | |

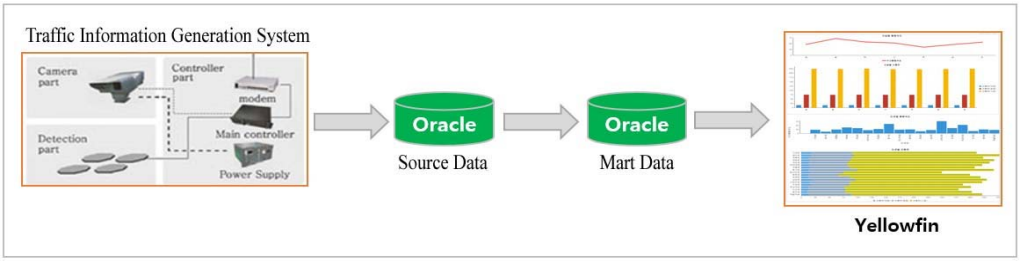
**Picture 6**. PoC results

As a result, 10 times faster speed could be confirmed than a No-SQL solution.

## 2.3. Speed improvement PoC on traffic data BI/DWin D City

### 2.3.1. The existing traffic data analysis system

Since the existing traffic data analysis system in D City conPictures source data from a traffic information generation system as mart data for aggregation and visualizes data by using Yellowfin's BI solution, there are some problems including an excessive use of time for 720,000 data extraction per day from 268 million data and an excessive use of hardware resources by the aggregate data (about 2TB) than the source data (about 1TB). The existing traffic data processing flow is as follows.

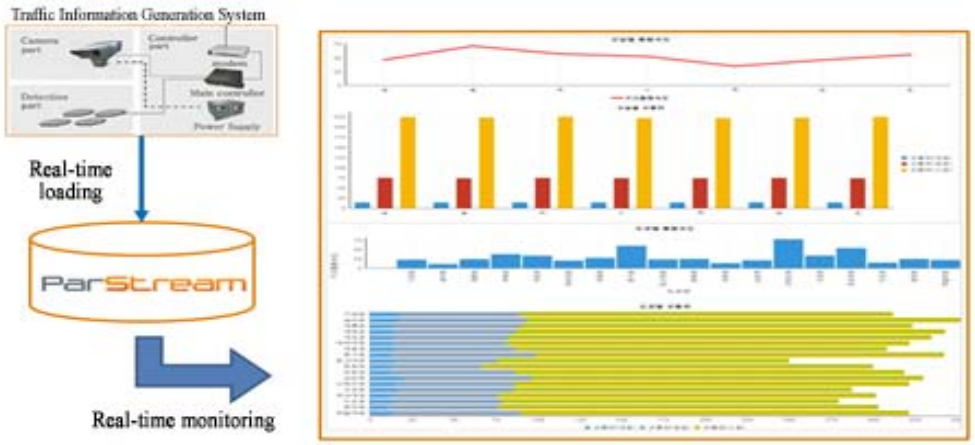



**Picture 7**. Existing data processing flow

### 2.3.2. PoC applied the ParStream big data platform

A data processing flow using the ParStream is as follows. As illustrated in the Picture, the ParStream performs visualization using BI solutions without configuring source data as aggregate mart data.

According to the PoC test results for analyzing the traffic data in D City, the existing daily batch query performance speed can be largely reduced from 1 hour to 2-3 minutes and a linkage between a BI tool (Yellowfin) and real-time queries becomes possible without forming a temporary table (data mart).




**Picture 8**. ParStream data processing flow

## 3. Conclusion

This study conducted a comparative analysis on the data processing performance of the big data platform based on the ParStream available to a real-time big data analysis.

While in the existing Hadoop platform, source data is loaded into Hadoop and aggregate data into MS-

SQL in RDBMS for a data pattern analysis, in the ParStream, all the works such as loading of source data, generation and usage of aggregate data for a data pattern analysis and real-time detection of an abnormal pattern can be performed in a single database, leading to the reduced complexity. The ParStream performs visualization using BI solutions without configuring source data as aggregate mart data. The results of the experimental analysis shows a 300-time improvement in processing speed, compared to the processing performance based on the existing Hadoop.

In conclusion, data governance needs for practical big data applications in various fields such as manufacturing, distribution and transportation. It will also be necessary to have a real-time big data platform apart from a big data platform for batch operation.

## 4. References

[1] D. Lenton, "The small screen [TV to Mobile Devices]", *IEE Rev*., Vol. 49, No. 10, pp. 38-41, Nov. 2003.
[2] Digital Video Broadcasting(DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television, ETSI EN 300 744 V1.5.1, June 2004.
[3] Transmission System for Handheld Terminals (DVB-H), TM3037 DVB-H202r4, June 2004.
[4] http://www.dvb.org.
[5] M. Hsieh and C. Wei, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels", *IEEE Trans. Consumer Electronics*, Vol. 44, No. 1, pp. 217-225, Feb. 1998.
[6] S. Sampei and T. Sunaga, "Rayleigh fading compensation for QAM in land mobile radio communications", *IEEE Trans. Veh. Technol*., Vol. 42, No. 2, pp. 137-147, May 1993.
[7] S. G. Kang, Y. M. Ha, and E. K. Joo, "A comparative investigation on channel estimation algorithms for OFDM in mobile communications", *IEEE Trans. Broadcasting*, Vol. 49, No. 2, pp. 142-149, June 2003.
[8] J. J. Beek, et al, "On channel estimation in OFDM systems", IEEE 45th Veh. Technol. Conf., Chicago, IL, Vol. 2, 25-28 July 1995, pp. 815-819.