Computer Engineering Program

Monitoring improvement by ETL development and database migration

LAI Tsz Yu, LAW Kan Ping, CHIU Cheuk Man, CHEUNG Kwan Ling
Advised by Prof. Gary CHAN

Introduction:

The project is to develop a well-performed data analyzing system supporting a practical industrial project called Streamphony which is a next-generation technology for providing a much better video streaming quality. The central idea is to replace the MYSQL database system by a column-oriented database Vertica which provides a much greater analyzing power for fast growing data size. Therefore, the improved system is now capable for generating real-time feedbacks and administrative data to better organize the streamphony cloud network such as load balancing. Besides, it can visualize the real-time graphs and historical graphs in the monitor website which shows all the channels' viewing records until now.

Objective:

- Support real-time data analyzing for better management
- Support big data retrieval and applicable in industry
- Develop an automated data flow system
- Develop a generic system with high flexibility

System Design Overview

MYSQL will get all the data from the data source in the Streamphony Enterprise Server and transfer the data to Vertica through the data flow Controller. Then the monitoring system will display the data in graphs retrieved in Vertica.

Data flow Controller Implementation:

This controller is an intermediate layer to transfer data from MYSQL to Vertica in an efficient way. Also, it is responsible for maintaining the data integrity during migration. Basically, it consists of three logic units. The first part is a data selection unit which exports records dynamically. The second component is a file merging unit which selects small files with similar records and combines them in a big file. This optimizes the insertion by utilizing the benefit of bulk loading. The third unit is a data importing unit which chooses the most optimal file to import to Vertica and reports files that fails the import transaction. This error detection model ensures data consistency between two databases. These three components are running in parallel and automatically manage the data flow in the system.

Database implementation:

Vertica

Column-oriented database system
Execute in VSQL commands

Software implementation:

ODBC Driver

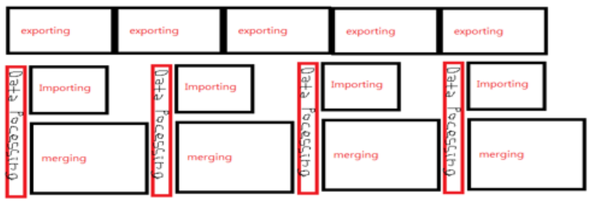
Development platform for data flow controller
Cross-platform, database driver

PHP

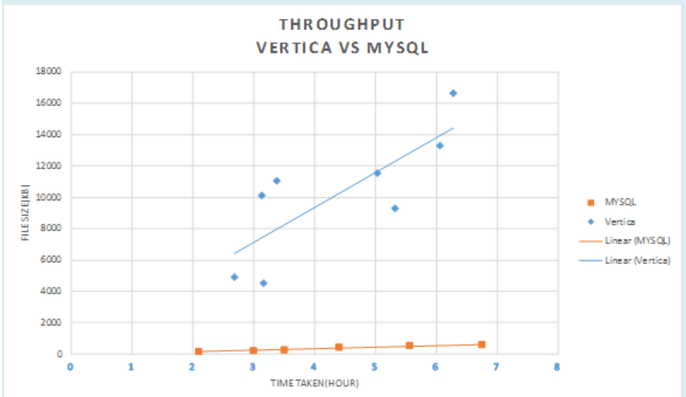
Streamphony monitor Website
Graphs display

Optimization:

Using similar technique with instruction pipelining, the three components are aligned and run in parallel so as to fully utilize the system resources. Therefore, it provides a higher bandwidth for the import link to Vertica.



Testing and Evaluation:



Vertica 30 times faster than MYSQL

scalability increases with multi-nodes

much greater analyzing power for streamphony project

Comparison with the existing system:

Originally, the system cannot visualize the monitoring graphs using MYSQL because of its low analyzing power. After implementing with column-oriented database Vertica, it successfully visualizes the graphs in a reasonable time. It also demonstrates a greater performance in big data analyzing.

Conclusion:

In the project, we design and implement a data flow system to manipulate and handle fast growing size of data to support both real time analyzing and big historical data analyzing. The project successfully migrates the whole system to a column-oriented database Vertica which is a next-generation database management system that has high analyzing power. It demonstrates a method for the streamphony project to enhance its performance in big data analyzing.