IMPLEMENTATION OF CHANGE DATA CAPTURE USING APACHE HIVE TO IMPROVE ETL PERFORMANCE IN A BIG DATA WAREHOUSE

W. MALITH C. WICKRAMARATNE

MSC BIG DATA ANALYTICS

2022

Abstract

A Data Warehouse acts as a centralized repository for millions/ billions of historical data. In order to provide historical intelligence, the data storage platform and the ETL process play vital roles with regards to the performance of a Data Warehouse. Many organizations tend to use Apache Hadoop as the distribution storage platform for large amounts of data, in other words for ‘Big Data’, however Hadoop has its own limitations when it comes to transactional processing such as inserts or updates or deletes.

This study aims to improve the performance of these transactions using Apache Hive, and thereby develop a logic to capture only the changed data within the ETL process. The experimented test results show that this method would improve the execution time of Hive queries, hence an improvement in the performance of the overall ETL process, which could result in significant lead time improvements to cater historical intelligence for organizations and its stakeholders.