UNIVERSITAS
GADJAH MADA
PERFORMANCE ANALYSIS BETWEEN HIVE AND CASSANDRA FROM TWITTER DATA WAREHOUSE
GHANI FARADHA, Suprapto, Drs., M.I.Kom., Dr.
Universitas Gadjah Mada, 2018 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

PERFORMANCE ANALYSIS BETWEEN HIVE AND CASSANDRA FROM TWITTER DATA WAREHOUSE

Ghani Faradha
11/315578/PA/13776

Pertumbuhan pesat aktivitas jejaring sosial telah menyebabkan beberapa perusahaan mulai menggunakannya untuk mendukung pengambilan keputusan mereka secara real-time. Twitter adalah salah satu media sosial paling populer. Namun, sebagian besar alat yang dikembangkan dari penelitian berbasis Twitter masih spesifik dalam beberapa tugas saja. Database tradisional belum menyediakan fungsionalitas untuk melakukan analisis data ukuran besar secara real time. Oleh karena itu, kami memerlukan data warehouse yang mendukung penyimpanan, pemrosesan, dan konversi data secara real time untuk data besar.

Penelitian ini menggunakan Apache Hive, salah satu teknologi data warehouse untuk penyimpanan dan pengolahan data besar secara real time. Data yang akan digunakan adalah data Twitter. Penelitian ini kemudian membandingkannya dengan Cassandra.

Studi ini mengeksplorasi kemampuan untuk menyimpan dan query data yang terkandung dalam Apache Hive dan Cassandra. Hasil penelitian ini menunjukkan bahwa Cassandra memiliki kinerja query terbaik, dengan rata-rata 29.68653325 detik untuk Cassandra dan 3890.7075 detik untuk Hive. Dalam pengambilan data, Cassandra lebih unggul dari Hive secara keseluruhan dan hanya mengarah ke data kecil jika dibandingkan dengan Hive. Saat mengumpulkan data streaming, Hive mendapat total data 40946, sementara Cassandra berjumlah total 67365 data.

Keywords : Data Warehouse, Twitter, Apache Hive, Real-time

UNIVERSITAS
GADJAH MADA

PERFORMANCE ANALYSIS BETWEEN HIVE AND CASSANDRA FROM TWITTER DATA WAREHOUSE
GHANI FARADHA, Suprapto, Drs., M.I.Kom., Dr.
Universitas Gadjah Mada, 2018 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Ghani Faradha
11/315578/PA/13776

The rapid growth of social networking activity has led some companies to start using it to support their decision-making in real-time. Twitter is one of the most popular social media. However, most of the tools developed from Twitter-based research are still specific in some tasks only. The traditional database has not provided the functionality to perform large-size data analysis in real time. Therefore, we need a data warehouse that supports storage, processing, and data conversion in real time for large data.

This study uses Apache Hive, one of the data warehouse technology for storage and processing of large data in real time. The data to be used is Twitter data. This study then compares it to Cassandra.

This study explores the ability to store and query data contained in Apache Hive and Cassandra. The results of this study show that Cassandra has the best query performance in Hive, with mean 29.68653325 seconds for Cassandra and 3890.7075 seconds for Hive. In data retrieval, Cassandra is superior to Hive as a whole and only leads to small data when compared to Hive. When collecting streaming data, Hive got a total data of 40946, while Cassandra totaled 67365 data.

Keywords : Data Warehouse, Twitter, Apache Hive, Real-time