INTISARI

PERBANDINGAN BASIS DATA RELASIONAL DENGAN NoSQL CASSANDRA UNTUK DATA WAREHOUSE YANG MENYIMPAN DATA TWITTER SECARA MENDEKATI *REAL-TIME*

Oleh

Muhammad Rafif Murazza
11/316774/PA/13901

Pesatnya perkembangan aktivitas jejaring sosial saat ini mengakibatkan munculnya ledakan arus data dengan volume yang besar dan menyebabkan beberapa perusahaan mulai memanfaatkan data tersebut untuk mendukung pengambilan keputusan secara mendekati *real-time*. Salah satu media sosial yang sangat populer adalah Twitter. Namun, kebanyakan dari tools yang dikembangkan dari hasil penelitian terhadap Twitter masih bersifat spesifik terhadap suatu jenis analisis. Hal tersebut disebabkan proses penelusuran keseluruhan aktivitas dari database yang sangat besar sangatlah rumit terlebih secara mendekati *real-time*. Oleh karena itu diperlukan teknologi data warehouse yang mampu menyimpan secara mendekati *real-time* hasil pengolahan dan pengubahan data semi-terstruktur menjadi terstruktur sehingga dapat digunakan dalam berbagai jenis analisis.

Pengembangan data warehouse dengan basis data relasional mulai menunjukkan keterbatasannya dalam menyimpan dan mengolah data besar sehingga perhatian publik mulai mengarah pada basis data NoSQL (Not Only SQL). Penelitian ini mengembangkan data warehouse menggunakan salah satu basis data NoSQL yang populer, yaitu Cassandra. Penelitian ini berfokus pada eksplorasi kemampuan Cassandra sebagai media data warehouse untuk menyimpan data Twitter yang mendekati *real-time* dengan membandingkan performa proses penyimpanan dan pemanggilan data dengan basis data relasional MySQL dan PostgreSQL.

Hasil penelitian ini menunjukkan bahwa diantara kedua basis data relasional, PostgreSQL memiliki performa penyimpanan lebih baik dibanding MySQL. Namun, secara keseluruhan Cassandra memiliki performa penyimpanan yang paling cepat diantara kedua basis data lainnya secara signifikan. Dalam aspek pemanggilan data, MySQL lebih unggul dibanding PostgreSQL secara keseluruhan dan hanya unggul pada data kecil ketika dibandingkan dengan Cassandra.

Kata-kata kunci : Data Warehouse, Twitter, NoSQL, Cassandra, *Real-Time*

ABSTRACT

RELATIONAL AND NoSQL CASSANDRA DATABASE COMPARISON FOR NEAR REAL-TIME DATA WAREHOUSE TO STORE TWITTER DATA

By

Muhammad Rafif Murazza
11/316774/PA/13901

The rapid growth of social network activities these days cause the explosion of data stream with big volume which also cause some corporation 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 research based on Twitter are still specifics in some tasks only. Because, exploring every activities from a big database is problematic let alone in real-time. Therefore, a data warehouse which able to store, process, and transform semi-structured data into structured in real-time is needed.

The data warehouse development using relational database start to show its limits on storing big data let alone real-time. Hence, public attention start going toward NoSQL (Not Only SQL) technology. In this research, a near real-time data warehouse using one of NoSQL databases, Cassandra, will be developed. This research focuses on exploring Cassandra ability as the solution of near real-time Twitter data warehouse by comparing its storing and querying performance with relational database, MySQL and PostgreSQL.

Results show that between the two relational database, the storing performance of PostgreSQL exceeds that of MySQL statistically. Although, Cassandra has the best storing performance out of the three databases in general. On the querying performance aspects, MySQL is faster that PostgreSQL. The same goes when comparing it to Cassandra, but only in relatively small data. When a large data is queried, the query performance of Cassandra surpass that of the other two relational database.

Keywords : Data Warehouse, Twitter, NoSQL, Cassandra, Real-time