BAB 6

KESIMPULAN DAN SARAN

Bab ini membahas tentang kesimpulan dari penelitian yang dilakukan dan saran penulis. Saran penulis berupa saran dalam pengembangan perangkat lunak dengan Spark SQL selanjutnya.

6.1 Kesimpulan

Berdasarkan penelitian Spark SQL dan Hive yang dilakukan, berikut kesimpulan yang dapat diambil:

1. Spark SQL adalah *library* Spark yang digunakan untuk bekerja dengan data terstruktur. Struktur utama Spark SQL adalah DataFrame yang merupakan kumpulan RDD dari Row. Aplikasi Spark SQL harus membuat *instance* kelas SQLContext atau HiveContext yang merupakan poin utama Spark SQL.

2. RDD Spark dirancang untuk mendukung penyimpanan data dalam memori, didistribusikan di seluruh *cluster* dengan cara yang terbukti toleran dan efisien. Karakteristik RDD adalah tidak dapat diubah (*immutable*), terpartisi (*partitioned*), dapat menangani kegagalan (*fault tolerant*), dapat diisi dengan tipe data apapun (*strongly typed*), dan komputasi dilakukan di memori (*in memory*).

3. DataFrame Spark SQL merupakan kumpulan data terdistribusi yang disusun menjadi kolom bernama. Perbedaan RDD dengan DataFrame adalah DataFrame memiliki informasi mengenai nama dan tipe kolom dari set data. DataFrame dapat dibuat dari RDD yang sudah ada dan dari sumber data (seperti JSON dan tabel Hive).

4. Hive adalah infrastruktur data *warehouse* berbasis Hadoop. Target penggunaan Hive adalah sebagai sistem untuk mengolah dan kueri data terstruktur. Konsep Hive mirip dengan basis data relasional, seperti tabel, baris, kolom, dan skema. Interaksi dengan Hive dapat menggunakan sintaks yang mirip dengan SQL, yaitu HiveQL. Hive juga dapat berinteraksi dengan MapReduce milik Hadoop untuk melakukan pekerjaan yang membutuhkan pekerjaan MapReduce.

5. Terdapat beberapa pernyataan SQL yang tidak didukung oleh Spark SQL, yaitu SELECT TOP, ROWNUM, INSERT INTO, UPDATE, DELETE, *constraints*, dan *index*.

6. Berbeda dengan Spark SQL, Hive mendukung hampir semua pernyataan SQL kecuali TOP dan ROWNUM.

7. Banyaknya komputer dan ukuran memori komputer sangat berpengaruh terhadap banyak data yang digunakan dan kinerja dalam mengerjakan pekerjaan. Semakin banyak data yang diproses, semakin banyak pula memori komputer yang digunakan.

8. Berdasarkan eksperimen waktu eksekusi yang dilakukan, Spark SQL jauh lebih cepat dalam melakukan kueri data dibandingkan Hive.

9. Waktu eksekusi Hive akan lebih lama bila menggunakan koneksi JDBC. Hal tersebut disebabkan oleh adanya proses-proses lain di JDBC.

6.2 Saran

Berdasarkan eksperimen dan kesimpulan di atas, berikut saran yang dapat penulis berikan:

1. Agar dapat mengolah dan menganalisis *big data* dengan cepat, komputer pada *cluster* Hadoop harus banyak dan ukuran memori masing-masing komputer juga harus besar.

2. Perangkat lunak untuk menganalisis dan mengolah data dengan Spark SQL yang dihasilkan masih menggunakan set data tertentu. Akan lebih baik bila perangkat lunak selanjutnya dibuat agar dapat mengolah set data apapun.

3. Pekerjaan Hive yang dikerjakan dengan koneksi JDBC lebih lama dibandingkan pekerjaan Hive menggunakan Hive CLI. Disarankan agar perangkat lunak selanjutnya dapat menggunakan Hive Client dalam pengerjaan kueri Hive.

DAFTAR REFERENSI

[1] Holmes, A. (2012) *Hadoop in Practice*, pap/psc edition. Manning Publications, Shelter Island.

[2] Sammer, E. (2012) *Hadoop Operations*. O'Reilly Media, Sebastopol.

[3] Lam, C. (2010) *Hadoop in Action*. Manning Publications, Stamford.

[4] Edward Capriolo, J. R., Dean Wampler (2012) *Programming Hive. Data Warehouse and Query Language for Hadoop*. O'Reilly Media, Sebastopol.

[5] Holden Karau, P. W. M. Z., Andy Konwinski (2015) *Learning Spark: Lightning-Fast Big Data Analysis*. O'Reilly Media, Sebastopol.

[6] Guller, M. (2015) *Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis*. Apress, New York.

[7] Rockoff, L. (2016) *The Language of SQL*, 2nd edition edition. Addison-Wesley Professional, U.S.

[8] Du, D. (2015) *Apache Hive Essentials: Immerse yourself on a fantastic journey to discover the attributes of big data by using Hive*. Packt Publishing, Birmingham.

[9] Alexander, A. (2013) *Scala Cookbook*. O'Reilly Media, Sebastopol.

[10] Ben-Gan, I. (2012) *Microsoft SQL Server 2012 T-SQL Fundamentals* Developer Reference. Microsoft Press, Sebastopol.

[11] Nield, T. (2016) *Getting Started with SQL: A Hands-On Approach for Beginners*. O'Reilly Media, Sebastopol.

[12] Frampton, M. (2015) *Mastering Apache Spark: Gain expertise in processing and storing data by using advanced techniques with Apache Spark*. Packt Publishing, Birmingham.

[13] Scott, J. A. (2015) *Getting Started with Apache Spark: From Inception to Production*. MapR Technologies, Inc., San Jose.