JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

NAMA	**:**	**ZAHROH NISHFUL LAILIYAH**
NRP	**:**	**5110100180**
DOSEN WALI	**:**	**Victor Hariadi, S.Si., M.Kom.**
DOSEN PEMBIMBING	**:**	**1. Umi Laili Yuhana, S.Kom, M.Sc.**
		2. Abdul Munif, S.Kom, M.Sc.

2. JUDUL TUGAS AKHIR

"Implementasi Sistem Penyimpanan Data dan Penggalian Data Terdistribusi menggunakan Hadoop dan Mahout dengan Studi Kasus Dataset Stack Overflow"

3. LATAR BELAKANG

Berkembangnya aplikasi berbasis web yang memerlukan pengolahan data dalam skala besar melahirkan paradigma baru dalam teknologi basis data. Ukuran data yang sangat besar *(big data)* menimbulkan permasalahan dari segi skalabilitas, karena pertambahan data yang terjadi setiap saat. Peningkatan kemampuan server secara vertikal yang dimiliki basis data relasional (RDBMS) terbatas pada penambahan prosesor, memori, dan media penyimpanan yang terbatas. Sedangkan peningkatan kemampuan server secara horizontal yang meliputi penambahan perangkat server baru dalam suatu jaringan memerlukan biaya yang mahal dan sulit dalam pengelolaannya. Salah satu cara yang diterapkan oleh aplikasi web berskala besar untuk mengatasi permasalahan tersebut adalah dengan menggunakan basis data non relasional (NoSQL), sebuah paradigma basis data yang merelaksasikan aturan-aturan konsistensi yang terdapat pada basis data relasional [1].

RDBMS menggunakan aturan *Atomicity*, *Consistency*, *Isolation*, dan *Durability* (ACID) untuk penyimpanan dan pengolahan data, tetapi NoSQL menggunakan paradigma *Basically Available*, *Soft State*, and *Eventually consistent* (BASE) untuk merelaksasikan aturan tersebut. Hasilnya, NoSQL dapat mengolah data dalam jumlah besar dengan memartisi data ke dalam beberapa server secara lebih mudah.

NoSQL menyimpan data dengan metode yang berbeda, salah satunya adalah dengan metode *key values* dan juga mempunyai teknik *MapReduce. MapReduce* adalah model pemrograman untuk memproses data yang sangat besar secara paralel dan terdistribusi. Implementasi basis data non relasional dengan teknik ini yang paling populer adalah Apache Hadoop karena bersifat *open source*. Apache Hadoop adalah sebuah kerangka kerja perangkat lunak yang mendukung aplikasi data intensif terdistribusi dan disahkan di bawah lisensi Apache. Hadoop mendukung kerja aplikasi pada *cluster* dengan jumlah besar [2].

4. RUMUSAN MASALAH

Rumusan masalah yang diangkat dalam Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana menyimpan data dalam skala besar secara terdistribusi menggunakan Hadoop?
2. Bagaimana mekanisme sistem terdistribusi dalam menangani pengolahan data dalam skala besar, *real time* dan responsif?
3. Bagaimana mengimplementasikan algoritma penggalian data (*data mining*) serta pengambilan keputusan pada sistem terdistribusi menggunakan Mahout?
4. Bagaimana melakukan *benchmarking* atau perbandingan sistem penyimpanan data NoSQL terdistribusi dengan sistem RDBMS biasa?
5. Bagaimana NoSQL mampu menangani penambahan/pengurangan infrastruktur secara dinamis?

5. BATASAN MASALAH

Adapun batasan ruang lingkup permasalahan dari pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Studi kasus yang digunakan adalah *big data* yang berasal dari situs web 1 yakni situs web yang menjadi tempat tanya jawab untuk bermacam – macam topik di bidang pemrograman komputer dan rekayasa perangkat lunak.
2. Basis data non relasional yang dibangun akan menggunakan kerangka kerja Hadoop mode Standalone pada sistem operasi Windows.
3. Teknik yang digunakan untuk memproses *big data* adalah MapReduce.
4. Implementasi penggalian data pada basis data non relasional yang digunakan adalah *library* Mahout yang merupakan bagian dari kerangka kerja Hadoop.

6. TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari pembuatan Tugas Akhir ini adalah sebagai berikut :

1. Dapat menyimpan data dalam skala besar secara terdistribusi menggunakan Hadoop.
2. Dapat memahami mekanisme sistem terdistribusi dalam menangani pengolahan data dalam skala besar, *real time* dan responsif (proses penulisan/pembacaan data yang cepat).
3. Dapat mengimplementasikan algoritma penggalian data (*data mining*) serta pengambilan keputusan pada sistem terdistribusi menggunakan Mahout.
4. Dapat melakukan *benchmarking* atau perbandingan sistem penyimpanan data NoSQL terdistribusi dengan sistem RDBMS biasa.
5. Dapat memahami mekanisme menangani penambahan/pengurangan infrastruktur secara dinamis pada NoSQL (penambahan atau pengurangan server).

7. MANFAAT TUGAS AKHIR

Tugas Akhir ini dikerjakan dengan harapan dapat memberikan manfaat pada bidang informatika dalam penyimpanan data berukuran besar secara efisien dan mengolahnya secara paralel dan terdistribusi yang yang nantinya memudahkan pengguna dalam mengakses data bersifat *real time*.

8. TINJAUAN PUSTAKA

Pada bab ini akan dibahas tinjauan pustaka yang dipergunakan pada Tugas Akhir ini, yaitu: Data Skala Besar, Apache Hadoop, MapReduce, Apache Mahout.

8.1 Data Skala Besar (*Big Data*)

Big Data adalah kumpulan *data set* yang begitu besar dan kompleks sehingga sangat sulit untuk memproses menggunakan alat manajemen basis data atau aplikasi pengolahan data tradisional. Kecenderungan untuk *data set* yang lebih besar adalah karena informasi tambahan diturunkan dari analisis dari data besar tunggal yang terkait, yang memungkinkan korelasi dapat ditemukan untuk melihat tren bisnis dan menentukan kondisi lalu lintas jalan secara *real time* [3].

Sumber data yang digunakan adalah *big data* yang mempunyai format XML dan berasal dari situs web *www.stackoverflow.com* yakni situs web yang menjadi tempat tanya jawab untuk bermacam – macam topik di bidang pemrograman komputer dan rekayasa perangkat lunak. Stack Overflow menempati urutan ke 55 dari 100 situs terbaik di dunia versi Alexa.com karena banyaknya jumlah pengunjung per hari, konten yang baik, dan menjadi tempat diskusi yang bermanfaat. Tak heran jika data *dummy* yang akan digunakan berukuran sebesar puluhan *gigabyte*.

8.2 Apache Hadoop

Google telah mempublikasikan tiga sistem canggihnya dalam hubungannya dengan pengolahan dan pemberdayaan Big Data. Ketiga sistem canggih tersebut adalah Google File System, Google MapReduce, dan Google Bigtable. Semuanya merupakan sistem terdistribusi yang dikenal handal, mampu mengolah data berukuran raksasa dengan efektif dan efisien, serta fleksibel. Ketiga sistem terdistibusi ini adalah sistem yang saling terkait erat, namun memiliki pembagian tugas yang jelas.

Apache Hadoop merupakan kerangka kerja yang dibangun di atas bahasa Java dan merupakan produk yang terinspirasi dari konsep ketiga sistem milik Google. Dari konsep Google File System lahirlah Hadoop Distributed File System (HDFS), dari konsep Google MapReduce lahir Hadoop MapReduce, dan dari spesifikasi Google Bigtable diciptakanlah Hadoop HBase. Ketiga produk Hadoop ini bersifat *open source* yang merupakan teknologi gratis dan boleh dipakai oleh siapa saja, dan memang sudah digunakan oleh banyak perusahaan besar seperti halnya Yahoo!, Facebook, Twitter, IBM, Trend Micro, NTT Docomo, Recrute Japan, Adobe, Amazon, Rakuten Japan, Benipal Technologies dan masih banyak lagi [4].

8.3 MapReduce

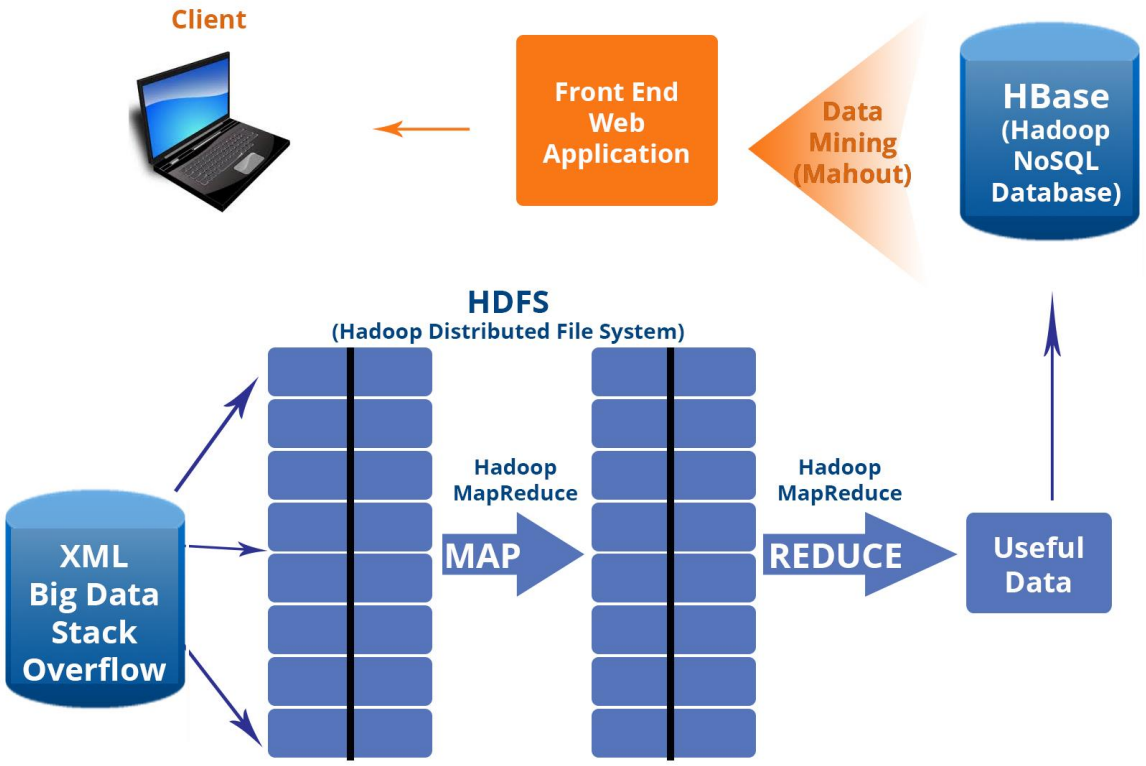
MapReduce adalah teknik pemrosesan data berukuran besar. *MapReduce* dapat dibagi dalam dua proses yaitu proses *Map* dan proses *Reduce*. Kedua jenis proses ini didistribusikan atau dibagi-bagikan ke setiap komputer dalam suatu *cluster* (kelompok komputer yang saling terhubung) dan berjalan secara paralel tanpa saling bergantung satu dengan yang lainnya. Proses *Map* bertugas untuk membaca *input* dalam bentuk pasangan *Key/Value* dari potongan-potongan data yang terdistribusi dalam tiap komputer dalam *cluster*. Hasilnya diserahkan kepada proses *Reduce* untuk dibaca dan digabungkan atau dikelompokkannya berdasarkan *Key/Value*. Kemudian hasil proses *Reduce* merupakan hasil akhir yang dikirim ke pengguna [5]. *MapReduce* yang digunakan pada Tugas Akhir ini adalah *MapReduce* milik Hadoop.

8.4 Apache Mahout

Apache Mahout adalah *library* yang merupakan bagian dari kerangka kerja Hadoop untuk pemrosesan data yang menghasilkan implementasi terdistribusi dari algoritma *machine learning* yang difokuskan di bidang penyaringan kolaboratif , *clustering* dan klasifikasi, tetapi tidak terbatas pada *platform* Hadoop saja. Mahout juga menyediakan *library* Java untuk penggunaan matematika umum (difokuskan pada aljabar linear dan statistik) dan koleksi Java yang primitif. Sementara algoritma inti Mahout dibangun untuk *clustering*, klasifikasi dan *batch* yang berbasis penyaringan kolaboratif diimplementasikan di atas Apache Hadoop menggunakan MapReduce, tetapi hal ini tidak membatasi Mahout untuk hanya digunakan pada Hadoop. Mahout juga dapat bekerja pada cluster non-Hadoop [6].

9. RINGKASAN ISI TUGAS AKHIR

Berikut adalah gambaran arsitektur sistem yang akan dibangun ditunjukkan pada Gambar 1.

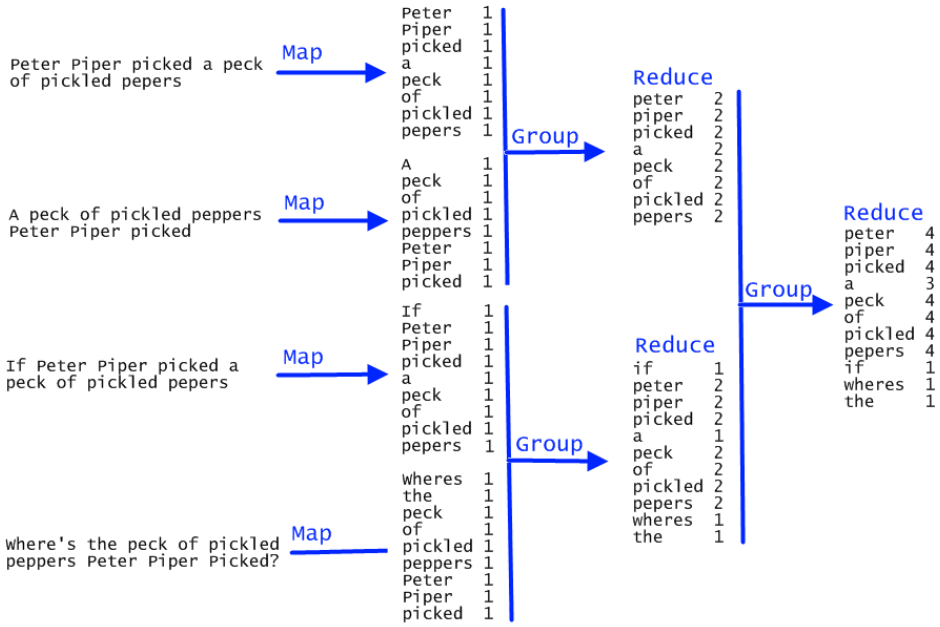


Gambar 1. Arsitektur Sistem

Gambar 1 menjelaskan beberapa tahap pembangunan sistem, antara lain:

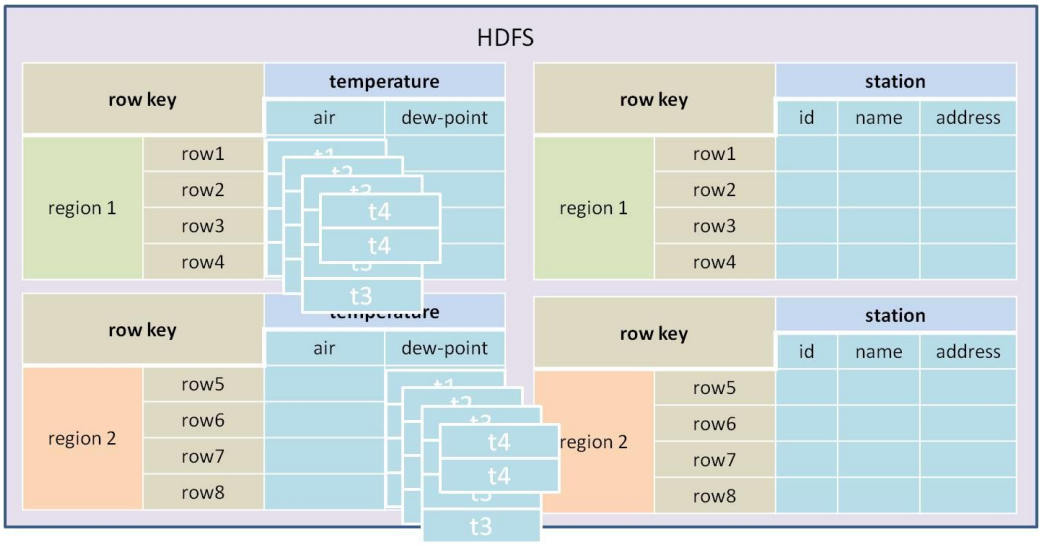
1. Sumber data yang digunakan adalah *big data* yang mempunyai format XML yang berasal dari situs web *www.stackoverflow.com* yakni situs web untuk diskusi di bidang pemrograman komputer dan rekayasa perangkat lunak. Data yang disimpan terdiri dari data pengguna, data lencana pengguna, data pertanyaan yang diajukan, data riwayat daftar pertanyaan yang diajukan, data jawaban yang diajukan, dan data *vote* yakni mengenai penilaian pengguna untuk menentukan pemilihan jawaban yang paling baik.

2. Data XML tersebut kemudian diproses dengan menggunakan teknik MapReduce dengan menggunakan Hadoop MapReduce yang kemudian disimpan ke dalam media penyimpanan milik Hadoop yakni Hadoop Distributed File System (HDFS). Adapun ilustrasi singkat dari pengolahan teknik MapReduce ditunjukkan pada Gambar 2.



Gambar 2. Ilustrasi teknik MapReduce

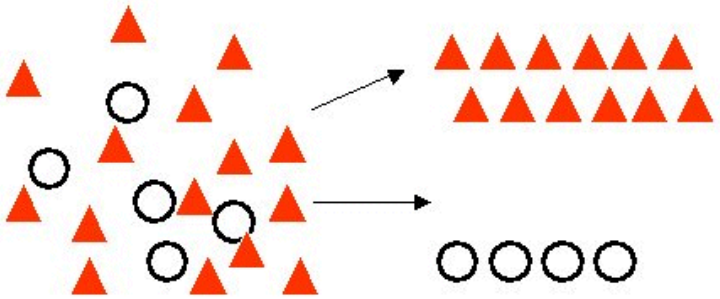
3. Setelah *big data* diproses dengan teknik MapReduce, maka data tersebut menjadi data yang telah siap disimpan ke dalam basis data NoSQL menggunakan HBase yakni basis data yang merupakan bagian dari sistem kerangka kerja Hadoop dan memang dikhususkan untuk data berskala besar. Adapun ilustrasi singkat dari metode pennyimpanan pada HBase ditunjukkan pada Gambar 3.



Gambar 3. Ilustrasi penyimpanan pada HBase

4. Kemudian data akan diproses dengan menggunakan *library* Mahout untuk digali. Jenis metode untuk penggalian data yang digunakan antara lain adalah metode klasifikasi yang digunakan untuk pengelompokan data. Penggalian data ini dapat digunakan

untuk melihat laporan data yang diinginkan. Hasil akhir dari penggalian data ini ditujukan untuk pengelompokan data pertanyaan dan jawaban dari situs web *Stack Overflow* dan juga untuk mendapatkan laporan statistik mengenai beberapa data dari situs web tersebut. Adapun ilustrasi singkat dari metode klasifikasi penggalian data pada HBase ditunjukkan pada Gambar 4.



Gambar 4. Ilustrasi klasifikasi pada HBase

5. Data yang telah digali kemudian disajikan melalui aplikasi web yang dibangun dengan bahasa pemrograman PHP. Aplikasi ini hanya dapat digunakan untuk melihat laporan data – data tersebut tanpa dapat menambahkan, mengurangi atau mengubah data – data yang ada. Aplikasi ini mempunyai beberapa fitur sebagai berikut:

   - Melihat pertanyaan dan jawaban yang telah dikelompokkan berdasarkan topik
   - Melihat statistik tentang situs web *Stack Overflow* yang meliputi popularitas topik, pengguna situs yang paling aktif dan waktu yang paling aktif digunakan untuk mengepos pertanyaan dan jawaban.

10.METODOLOGI

a. Penyusunan proposal tugas akhir

Tahap awal untuk memulai pengerjaan Tugas Akhir adalah penyusunan proposal Tugas Akhir. Pada proposal ini, penulis mengajukan gagasan mengenai rancang bangun basis data non relasional dengan kerangka kerja Hadoop untuk studi kasus *big data* Stack Overflow beserta penggalian data yang dibutuhkan dengan *library* Mahout.

b. Studi literatur

Pada tahap ini dilakukan pencarian informasi dan studi literatur yang diperlukan untuk pengumpulan data dan desain sistem yang akan dibuat. Referensi tersebut berisikan tentang :

1. *Big data* dan karakteristik, penyimpanan serta pengolahannya.
2. Rancang bangun basis data non relasional (NoSQL) menggunakan Hadoop dan arsitekturnya.
3. Pengolahan data pada basis data non relasional dengan teknik MapReduce.

4. Penggalian data dan algoritma yang digunakan menggunakan Mahout.

c. Analisis dan desain perangkat lunak

Sistem yang akan dibangun pada Tugas Akhir ini adalah:

1. Infrastruktur berupa basis data non relasional.
2. Aplikasi Web berupa laporan dari data yang telah digali.

d. Implementasi perangkat lunak

Rencana pembangunan pada Tugas Akhir ini meliputi:

1. Basis data non relasional dengan menggunakan kerangka kerja Hadoop dengan sistem operasi Windows.
2. Penggalian data menggunakan *library* Mahout.
3. Aplikasi web menggunakan bahasa PHP dengan kerangka kerja CodeIgniter dan IDE Netbeans.

e. Pengujian dan evaluasi

Pengujian pada Tugas Akhir ini akan dilakukan dengan beberapa cara yaitu:

1. Pengujian *blackbox*
   Pengujian *blackbox* adalah pengujian yang berfokus pada spesifikasi fungsional dari perangkat lunak, penguji dapat mendefinisikan kumpulan kondisi *input* dan melakukan pengetesan pada spesifikasi fungsional program.

2. Pengujian berdasarkan waktu
   Pengujian ini dilakukan dengan menganalisa waktu yang digunakan oleh Hadoop dalam memproses *big data* dengan melihat *running time, response time* dan *throughput.*

f. Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam tugas akhir ini serta hasil dari implementasi aplikasi perangkat lunak yang telah dibuat. Sistematika penulisan buku tugas akhir secara garis besar antara lain:

1. Pendahuluan
   a. Latar Belakang
   b. Rumusan Masalah
   c. Batasan Tugas Akhir

    d. Tujuan
    e. Metodologi
    f. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

11.JADWAL KEGIATAN

Tahapan	2013												2014							
	Oktober				Nopember				Desember				Januari					Februari		
Penyusunan Proposal																				
Studi Literatur																				
Perancangan sistem																				
Implementasi																				
Pengujian dan evaluasi																				
Penyusunan buku																				

12.DAFTAR PUSTAKA

[1] F. Firdausillah, E. Y. Hidayat and I. N. Dewi, "NoSQL: Latar Belakang, Konsep, dan Kritik," pp. 1-7, 2012.

[2] Wikipedia, "Wikipedia, the free encyclopedia," 30 September 2013. [Online]. Available: http://en.wikipedia.org/wiki/Apache_Hadoop. [Accessed 30 September 2013].

[3] Wikipedia, "Wikipedia, the free encyclopedia," 03 October 2013. [Online]. Available: http://en.wikipedia.org/wiki/Big_data. [Accessed 03 October 2013].

[4] V. Wijaya, "WJaya Weblog," [Online]. Available: http://vijjam.blogspot.com/2013/03/hbase-hyper-nosql-database.html.

[5] D. Gillick, A. Faria and J. DeNero, "MapReduce: Distributed Computing for Machine Learning," pp. 1-12, 5 November 2006.

[6] Wikipedia, "Wikipedia, the free encyclopedia," 07 Agustus 2013. [Online]. Available: http://en.wikipedia.org/wiki/Apache_Mahout. [Accessed 30 September 2013].