JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNLOGI 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

"Eksplorasi Kerangka Kerja Hadoop dengan Teknik MapReduce dan Implementasi Penggalian Data dengan Mahout untuk Studi Kasus Big Data 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* [2]. *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 yang bersifat *open source*.

Apache Hadoop adalah sebuah kerangka kerja perangkat lunak *open source* yang mendukung aplikasi data intensif terdistribusi dan disahkan di bawah lisensi Apache. Hadoop berasal dari *Google MapReduce* dan *Google File System* (GFS). Hadoop mendukung kerja aplikasi pada *cluster* dengan jumlah besar [3].

4. RUMUSAN MASALAH

Detail permasalahan yang diangkat dalam tugas akhir ini adalah sebagai berikut:
1. Bagaimana membangun basis data non relasional dengan menggunakan Hadoop?
2. Bagaimana mengolah *big data* dengan menggunakan teknik MapReduce?
3. Bagaimana menggali data pada basis data non relasional dengan mahout?

5. BATASAN MASALAH

Masalah yang dibahas pada tugas akhir ini dibatasi lingkupnya pada:
1. Basis data non relasional yang digunakan adalah Hadoop mode Standalone pada sistem operasi Windows.
2. Teknik yang digunakan untuk memproses *big data* adalah MapReduce.
3. Implementasi penggalian data pada basis data non relasional dengan menggunakan Mahout.
4. Studi kasus yang digunakan adalah *big data* yang 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.

6. TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari penyusunan tugas akhir ini adalah:
1. Mengeksplorasi kerangka kerja Hadoop untuk membangun basis data non relasional (NoSQL). Dalam hal ini basis data non relasional sangat efisien untuk menyimpan data dengan volume besar yang membutuhkan pemrosesan secara *real time*.

2. Mengeksplorasi teknik pengolahan *big data* menggunakan MapReduce.
3. Mengeksplorasi penggalian data dari data yang sudah diubah ke struktur Hadoop menggunakan *tools* Mahout (komplemen dari Hadoop).

7. MANFAAT TUGAS AKHIR

Manfaat dari penyusunan tugas akhir ini adalah:

1. Pengguna mampu menyimpan data berukuran besar secara efisien.
2. Pengguna mampu mengolah data berukuran besar secara paralel dan terdistribusi.
3. Pengguna mampu menampilkan laporan data secara *real time*.
4. Sistem mampu menyelesaikan persoalan komputasi yang lebih kompleks.

8. TINJAUAN PUSTAKA

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. Tantangannya meliputi penangkapan, kurasi, penyimpanan, pencarian, berbagi, transfer, analisis, dan visualisasi. Kecenderungan untuk *data set* yang lebih besar adalah karena informasi tambahan diturunkan dari analisis dari data besar tunggal yang terkait, dibandingkan dengan memisahkan *data set* yang lebih kecil dengan jumlah data yang sama, yang memungkinkan korelasi dapat ditemukan untuk melihat tren bisnis, menentukan kualitas penelitian, mencegah penyakit, hubungan kutipan hukum, perlawanan terhadap kejahatan, dan menentukan kondisi lalu lintas jalan secara *real time* [4].

8.2 Apache Hadoop

Apache Hadoop merupakan *framework*, yang dibangun di atas bahasa Java, untuk komputasi dan pemrosesan dataset yang besar (bahkan sangat besar) secara terdistribusi. *Framework* Hadoop terdiri dari tiga yaitu Hadoop Common, Hadoop *Distributed File System* (HDFS), dan Hadoop *Map Reduce*.

HDFS adalah media penyimpanan dari file yang telah dibagi – bagi berdasarkan blok dan blok – blok ini bisa terdapat pada lokasi yang berbeda dan dilakukan replikasi dengan urutan blok yang mungkin tidak sama per node. HDFS bisa bersifat *single node* atau *multiple node*. HDFS bukan *native file system* seperti layaknya EXT3, EXT4, FAT atau NTFS. HDFS ada pada layer di atasnya. Kemudian terdapat database yang menggunakan framework Hadoop yaitu HBase [4].

HBase adalah basis data terdistribusi yang berorientasi pada kolom dan berjalan diatas HDFS yang mampu memproses data dalam skala besar secara interaktif. HBase merupakan implementasi dari konsep Google Bigtable. HBase inilah yang nantinya digunakan sebagai basis data non relasional.

Basis data non relasional (NoSQL) adalah basis data yang dianggap memiliki kemampuan yang lebih baik dan performa yang signifikan dalam mengolah data yang besar dan aplikasi web yang *real Time.* Basis data ini menyimpan data secara berbeda dimana basis data ini tidak menyimpan data berdasar relasinya tetapi menggunakan empat metode antara lain:

* *Key Values oriented*: basis data NoSQL menggunakan *Key Values* ini untuk menyimpan *unique key* sebagai penanda indeks. Penggunaanya boleh terstruktur dan tidak terstruktur.

* *Document oriented*: basis data NoSQL menggunakan *Document Oriented* sebagai struktur penyimpanannya sehingga bisa ditambahkan field dengan panjang *value* tertentu jadi lebih mudah dan fleksibel yakni tidak terlalu terikat dengan ukuran dari struktur tabel.

* *Table oriented*: basis data NoSQL menggunakan tabel untuk menyimpan data.

* *Graph oriented*: basis data NoSQL menggunakan konsep Graf untuk penyimpanan datanya. Di antara ketiga yang lain, cara graf masih terbilang baru di dalam implementasinya.

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 mengumpulkan informasi dari potongan-potongan data yang terdistribusi dalam tiap komputer dalam *cluster*. Hasilnya diserahkan kepada proses *Reduce* untuk diproses lebih lanjut. Kemudian hasil proses *Reduce* merupakan hasil akhir yang dikirim ke pengguna.

Fungsi *Map* bertugas untuk membaca *input* dalam bentuk pasangan *Key/Value*, lalu menghasilkan *output* berupa pasangan *Key/Value* juga. Pasangan *Key/Value* hasil fungsi *Map* ini disebut pasangan *Key/Value intermediate*. Kemudian, fungsi *Reduce* akan membaca pasangan *Key/Value intermediate* hasil fungsi *Map*, dan menggabungkan atau mengelompokkannya berdasarkan *Key* tersebut [5].

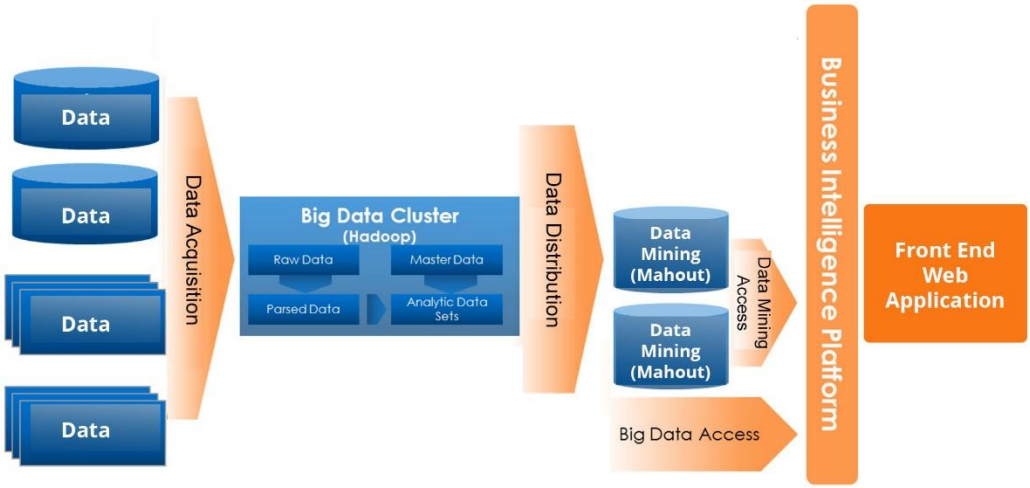
8.4 Apache Mahout

Apache Mahout adalah tools yang merupakan bagian dari Hadoop untuk menghasilkan implementasi terdistribus 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) operasi 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 [5].

9. RINGKASAN ISI TUGAS AKHIR

Berikut adalah gambaran arsitektur basis data non relasional yang akan dibangun.



Gambar 1. Arsitektur Basis Data NoSQL

Gambar 1 menjelaskan tentang penggunaan basis data non relasional yang akan dibangun menggunakan hadoop. Hadoop akan digunakan sebagai *cluster* (kelompok komputer) tempat penyimpanan data berukuran besar. Sumber data utama dapat berasal dari bermacam – macam basis data yang disimpan menjadi satu ke dalam Hadoop. Penyimpanan ini nantinya menggunakan teknik *MapReduce* sehingga lebih efisien.

Penggalian data akan dilakukan menggunakan Mahout. Jenis algoritma untuk penggalian data yang digunakan bisa bermacam – macam, antara lain terdapat algoritma klasifikasi dan pengelompokan (*clustering*). Penggalian data ini dapat digunakan antara lain untuk melihat prediksi atau melihat laporan data yang diinginkan.

Data yang telah digali dapat diakses melalui aplikasi web. Aplikasi ini hanya dapat digunakan untuk melihat laporan data – data tersebut tanpa dapat menambahkan, mengurangi atau mengubah data – data yang ada. Aplikasi ini berperan untuk mempermudah pengguna dalam mendapatkan informasi yang dibutuhkan.

10.METODOLOGI

a. Penyusunan proposal tugas akhir

Proposal tugas akhir ini berisi tentang deskripsi singkat mengenai rancang bangun basis data non relasional untuk studi kasus *big data* beserta penggalian data yang dibutuhkan. Basis data yang digunakan adalah Hadoop dan *tools* penggalian data yang digunakan adalah Mahout.

b. Studi literatur

Studi literatur yang akan dipelajari dalam penyusunan tugas akhir ini antara lain:

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

Hal yang akan dibangun pada tugas akhir ini adalah:

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

d. Implementasi perangkat lunak

Rencana pembangunan pada tugas akhir ini meliputi:

1. Basis data non relasional dengan menggunakan Hadoop *File System* (HFS).
2. Penggalian data menggunakan *tools* Mahout.
3. Aplikasi web menggunakan bahasa PHP dengan kerangka kerja CodeIgniter dan IDE Netbeans.

e. Pengujian dan evaluasi

Pengujian tugas akhir ini meliputi performa basis data dalam menjalankan perintah yang didasarkan pada waktu antara lain *running time*, *response time* dan *throughput*. Evaluasi yang dipakai adalah metode pengujian *black box*.

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			
	September				Oktober				Nopember				Desember				Januari			
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," 26 September 2013. [Online]. Available: http://en.wikipedia.org/wiki/NoSQL. [Accessed 30 September 2013].

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

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

[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].