BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Hadoop Distributed file System (HDFS) adalah sebuah media penyimpanan data utama dengan kapasitas data besar yang digunakan oleh *Hadoop* yang memiliki kehandalan dalam perhitungan data yang sangat cepat. *Hadoop* menggunakan arsitektur *scale-out* yang menggunakan server komoditas yang di konfigurasikan sebagai klaster, dimana setiap server memiliki disk internal dan memiliki harga *drive* yang terjangkau. Dalam segi data, *Hadoop* akan memecah tiap data menjadi blok-blok dan tersebar di seluruh claster.

HDFS menjadi perangkat lunak open source yang memberikan solusi atas penyimpanan data yang khusus dan memiliki kemampuan untuk melakukan redudensi dan *failover* secara otomatis dalam menganalisis pencarian data *warehouse*. *Hadoop* juga memberikan struktur data dalam penyimpanan yang memfasilitasi dalam memberikan efisiensi biaya untuk menganalisis dan memproses sejumlah data pada *warehouse*.

Apache Cassandra adalah database *NoSQL* yang besarannya secara *scalable*. *NoSQL* sering digunakan untuk mengelola infrastruktur data yang penting sehingga *Cassandra* dikenal sebagai solusi profesional secara teknis saat pengguna membutuhkan database *NoSQL* secara tepat waktu dalam kinerja yang tinggi pada skala besar yang tidak pernah turun.

Cassandra memiliki sistem kerja *peer to peer* yang arsitekturnya didistribusikan seperti “cincin”, yang mudah dalam melakukan *setup* dan pemeliharaan. Di *Cassandra*, semua node adalah sama, tidak ada konsep dari *node master* dengan semua *node* berkomunikasi dengan satu sama lain menggunakan protokol gossip.

Cassandra built untuk skala arsitektur petabyte mampu menangani banyak informasi dan ribuan pengguna secara bersamaan per-operasi, per-detik di banyak pusat data dengan mudah karena dapat mengelola jumlah yang lebih kecil dari data dan pengguna yang berjalan. *Cassandra File System (CFS)* menyediakan pondasi penyimpanan dalam menjalankan analisis data *Hadoop-gaya* pada *Cassandra* yang bebas dari kerumitan. Sehingga diperlukan konfigurasi penggabungan antara *HDFS* dan *CFS* untuk memaksimalkan sistem kerja dalam penyimpanan data besar.

1.2 Perumusan Masalah

1. Bagaimanakah sistem kerja dari *Hadoop Distributed file System (HDFS)* dengan *Cassandra File System (CFS)* menggunakan jaringan peer to peer?
2. Bagaimanakah konfigurasi penggabungan *Hadoop Distributed file System (HDFS)* dengan *Cassandra File System (CFS)* dalam penyimpanan data besar (*big data*)?

1.3 Batasan Masalah

Beberapa batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Platform *Hadoop CFS* yang digunakan di OS Linux Ubuntu
2. *Hadoop HDFS* hanya sebagai penyimpanan Data dan *CFS* sebagai sistem Pencarian.
3. Sebatas pembangunan konfigurasi dari *Hadoop Distributed file System (HDFS)* dengan *Cassandra File System (CFS)*.
4. Tidak menganalisa peforma dan sistem kerja aplikasi (sebatas konfigurasi dan visualisasi).

1.4 Tujuan Penelitian

1. Menganalisis proses konfigurasi dari penggabungan *HDFS* dan *CFS*.
2. Menyederhanakan *Overhead Operasional Hadoop* dengan menghilangkan titik tunggal kegagalan dalam *Hadoop Namenode*.

1.5 Manfaat Penelitian

1. Memberikan gambaran tentang bagaimana menganalisis *Hadoop-gaya (Map-Reduce dan Hive)* dapat dijalankan pada data yang terdapat dalam *Apache Cassandra*.
2. Memberikan gambaran tentang bagaimana hasil dari proses konfigurasi penggabungan *Hadoop Distributed file System (HDFS)* dengan *Cassandra File System (CFS)* dalam penyimpanan data besar (*Big Data*).
3. Memberikan gambaran tentang penangan ribuan informasi pengguna secara bersamaan per-operasi dan per-detik di banyak pusat data dengan mudah.
4. Memberikan gambaran tentang kemampuan menganalisis alur proses, manfaat dan perbandingan dari sistem *HDFS* dengan *CFS*.