

Tugas Big Data Architecture and Infrastructure

Pertemuan 6

Hadoop Storage Layer

Nama : Lukas Febrian Laufra
Kelas/Nim : TI22J/20220040076
Dosen Pengajar : Pak Ir. Somantri, S.T, M.Kom
Waktu Pengerjaan : 03.52/22 November 2024

Soal

1. Jelaskan secara singkat konsep dasar Hadoop Storage Layer (HDFS) dan sebutkan peran utama NameNode dan DataNode dalam sistem penyimpanan HDFS?
2. Buatlah skenario fiktif di mana terjadi kegagalan pada salah satu DataNode dalam kluster Hadoop yang menggunakan HDFS?

Jawaban

1. Konsep Dasar Hadoop Storage Layer (HDFS)

Hadoop Distributed File System (HDFS) adalah sistem file terdistribusi yang dirancang untuk menyimpan dan mengelola data besar pada kluster besar. HDFS memecah file menjadi beberapa blok dan menyimpan salinan blok tersebut pada beberapa mesin untuk memastikan toleransi kesalahan. HDFS dioptimalkan untuk akses data yang cepat dan pola streaming, sehingga cocok untuk pemrosesan data besar.

Peran Utama NameNode dan DataNode

- **NameNode:**
 - Mengelola namespace sistem file dan menyimpan semua metadata sistem file, termasuk informasi tentang lokasi blok data. NameNode bertanggung jawab untuk operasi baca dan tulis, serta mengoordinasikan penyimpanan dan pengambilan data.
- **DataNode:**
 - Menyimpan blok data yang sebenarnya dan melayani permintaan baca dan tulis dari klien. DataNode secara berkala mengirimkan pesan heartbeat dan laporan blok ke NameNode untuk memastikan bahwa statusnya terpantau.

2. Skenario Fiktif: Kegagalan pada Salah Satu DataNode

Dalam sebuah perusahaan yang menggunakan HDFS untuk menyimpan data analitik, terdapat kluster Hadoop dengan tiga DataNode: DataNode1, DataNode2, dan DataNode3. Setiap blok data direplikasi tiga kali untuk menjamin ketersediaan dan keandalannya.

Suatu hari, DataNode2 mengalami kegagalan akibat kerusakan hardware. Karena HDFS telah mengimplementasikan replikasi, data yang disimpan di DataNode2 juga tersedia di DataNode1 dan DataNode3.

Setelah mendeteksi kegagalan, NameNode menerima laporan heartbeat yang hilang dari DataNode2. NameNode kemudian memperbarui statusnya, mencatat bahwa DataNode2 tidak tersedia. Meskipun DataNode2 tidak dapat melayani permintaan, pengguna masih dapat mengakses data yang sama melalui DataNode1 dan DataNode3 tanpa mengalami gangguan.

Untuk mengatasi kegagalan ini, tim IT segera mengganti DataNode2 dan mengkonfigurasi ulang sistem. Setelah DataNode2 diperbaiki dan online kembali, NameNode mengatur ulang replikasi blok data untuk memastikan bahwa semua data tetap memiliki tiga salinan yang aktif di kluster.

Dengan cara ini, HDFS tetap dapat memberikan ketersediaan yang tinggi dan toleransi kesalahan meskipun salah satu DataNode mengalami kegagalan.