MODUL 3 – DISTRUBUTED FILE SYSTEM DENGAN HDFS

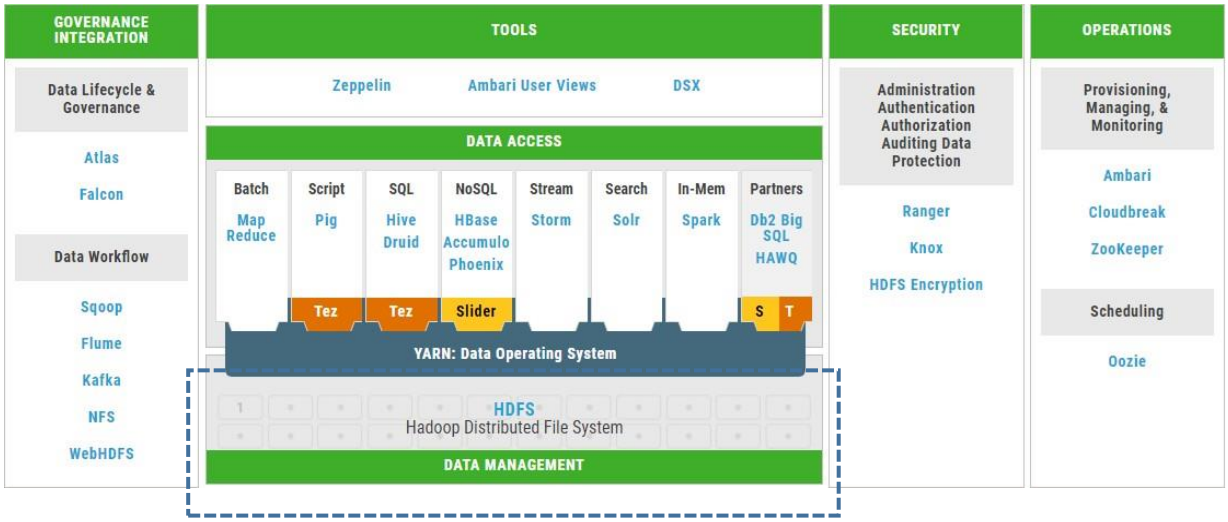
1.1. Deskripsi Singkat

Sistem File Terdistribusi Hadoop (HDFS) adalah sistem file terdistribusi yang dirancang untuk berjalan pada perangkat keras. Ini memiliki banyak kesamaan dengan sistem file terdistribusi yang ada. Namun, perbedaan dari sistem file terdistribusi lainnya sangat signifikan. HDFS sangat toleran terhadap kesalahan dan dirancang untuk digunakan pada perangkat keras berbiaya rendah. HDFS menyediakan akses throughput tinggi ke data aplikasi dan cocok untuk aplikasi yang memiliki kumpulan data besar. HDFS mengaktifkan akses streaming ke data sistem file. HDFS awalnya dibangun sebagai infrastruktur untuk proyek mesin pencari web Apache Nutch. HDFS adalah bagian dari proyek Apache Hadoop Core.

Pada pertemuan 1 dan 2 sebelumnya, telah dipelajari instalasi dan konfigurasi penggunaan Hortonwork Data Platform sebagai salah satu platform untuk mempelajari Hadoop. Salah satu komponen dalam HDP adalah HDFS. HDFS merupakan komponen untuk menyimpan dan manajemen data. Pada praktikum ini, akan dilakukan manajemen file pada HDFS.

1.2. Tujuan Praktikum

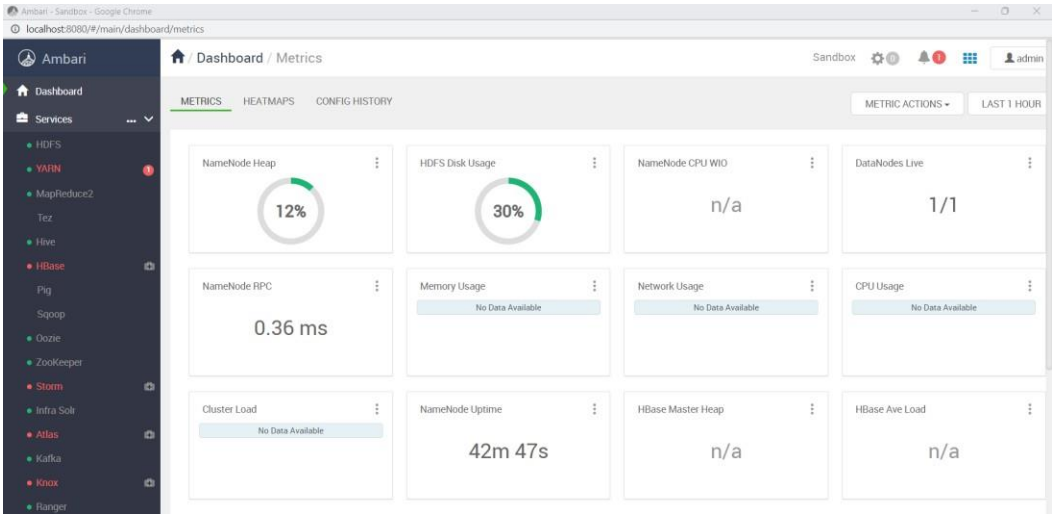
Sebelum melakukan praktikum, mahasiswa diasumsikan telah melakukan instalasi HDP dan memiliki pengetahuan tentang Hadoop. Setelah praktikum pada modul 6 ini, diharapkan mahasiswa mempunyai kompetensi untuk melakukan manejemen file pada HDFS. Adapun komponen yang dimaksud dalam keseluruhan HDP dapat dilihat pada gambar berikut.



1.3. Material Praktikum

Persyaratan yang dibutuhkan untuk melakukan praktikum ini yaitu aplikasi hortonwork yang sudah di install pada modul praktikum 1 dan 2.

1. HDP yang telah terinstal pada virtualbox.
2. Ambari dapat diakses



1.4. Kegiatan Praktikum

Melakukan manajemen HDFS dengan menggunakan command (CLI).

1. Akses ke HDP dengan menggunakan akun yang telah didaftarkan sebelumnya. Akses http://localhost:4200 pada browser.
2. Lakukan login terlebih dahulu.

   Username: **root**

   Existing Password: **Tp4stis**
3. Hadoop menyertakan berbagai perintah mirip shell yang berinteraksi langsung dengan HDFS dan sistem file lain yang didukung Hadoop. Perintah bin/hdfs dfs -help mencantumkan perintah yang didukung oleh Hadoop shell. Selanjutnya, perintah bin/hdfs dfs -help command-name menampilkan bantuan yang lebih detail untuk sebuah perintah. Perintah-perintah ini mendukung sebagian besar operasi sistem file normal seperti menyalin file, mengubah izin file, dll. Ini juga mendukung beberapa operasi khusus HDFS seperti mengubah replikasi file.
   a. Mengubah role menjadi superuser

   ```
   [root@sandbox-hdp ~]# su - hdfs
   Last login: Sat Mar 11 06:48:24 UTC 2023
   ```

   b. -report : melaporkan statistik dasar HDFS. Beberapa informasi ini juga tersedia di halaman depan NameNode

```
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -report
Configured Capacity: 112248641024 (104.54 GB)
Present Capacity: 83152613376 (77.44 GB)
DFS Remaining: 78978715648 (73.55 GB)
DFS Used: 4173897728 (3.89 GB)
DFS Used%: 5.02%
Replicated Blocks:
        Under replicated blocks: 202
        Blocks with corrupt replicas: 4
        Missing blocks: 4
        Missing blocks (with replication factor 1): 4
        Low redundancy blocks with highest priority to recover: 202
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 172.18.0.2:50010 (sandbox-hdp.hortonworks.com)
Hostname: sandbox-hdp.hortonworks.com
Decommission Status : Normal
Configured Capacity: 112248641024 (104.54 GB)
DFS Used: 4173897728 (3.89 GB)
Non DFS Used: 23079822848 (21.49 GB)
```

c. -safemode : meskipun biasanya tidak diperlukan, administrator dapat masuk atau keluar dari Safemode secara manual.

```
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -safemode
Usage: hdfs dfsadmin [-safemode enter | leave | get | wait | forceExit]
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -safemode enter
Safe mode is ON
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -safemode leave
Safe mode is OFF
```

d. -finalizeUpgrade: menghapus cadangan cluster sebelumnya yang dibuat selama pemutakhiran terakhir.

```
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -finalizeUpgrade
Finalize upgrade successful
```

e. -refreshNodes: Memperbarui namenode dengan sekumpulan datanode yang diizinkan untuk terhubung ke namenode. Secara default, Namenodes membaca ulang nama host datanode dalam file yang ditentukan oleh dfs.hosts, dfs.hosts.exclude Host yang ditentukan dalam dfs.hosts adalah datanodes yang merupakan bagian dari cluster. Jika ada entri di dfs.hosts, hanya host di dalamnya yang diizinkan mendaftar ke namenode. Entri di dfs.hosts.exclude adalah datanode yang perlu dinonaktifkan. Atau jika dfs.namenode.hosts.provider.classname diatur ke org.apache.hadoop.hdfs.server.blockmanagement. CombinedHostFileManager, semua host yang disertakan dan dikecualikan ditentukan dalam file JSON yang ditentukan oleh dfs.hosts. Datanodes menyelesaikan penonaktifan ketika semua replika dari mereka direplikasi ke datanodes lain. Node yang dinonaktifkan tidak dimatikan secara otomatis dan tidak dipilih untuk menulis replika baru.

```
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -refreshNodes
Refresh nodes successful
```

f. -printTopology : Cetak topologi cluster. Menampilkan letak rak dan datanode yang melekat pada trek seperti yang dilihat oleh NameNode.

```
[hdfs@sandbox-hdp ~]$ hdfs dfsadmin -printTopology
Rack: /default-rack
   172.18.0.2:50010 (sandbox-hdp.hortonworks.com)
```

4. Envvars : menampilkan variabel environtment Hadoop.

```
[hdfs@sandbox-hdp ~]$ hdfs envvars
JAVA_HOME='/usr/lib/jvm/java'
HADOOP_HDFS_HOME='/usr/hdp/3.0.1.0-187/hadoop-hdfs'
HDFS_DIR='./'
HDFS_LIB_JARS_DIR='lib'
HADOOP_CONF_DIR='/usr/hdp/3.0.1.0-187/hadoop/conf'
HADOOP_TOOLS_HOME='/usr/hdp/3.0.1.0-187/hadoop'
HADOOP_TOOLS_DIR='share/hadoop/tools'
HADOOP_TOOLS_LIB_JARS_DIR='share/hadoop/tools/lib'
```

5. Usecase : Melakukan download data csv dari Internet kemudian simpan kedalam virtual mesin. Pindahkan kedalam hdfs, dan cek di disk mana file tersebut disimpan.
    a. Mengambil file dari internet

```
[hdfs@sandbox-hdp ~]$ wget https://stats.govt.nz/assets/Uploads/Annual-enterprise-survey/Annual-enterprise-survey-2021-financial-year-provisional/Download-data/annual-enterprise-survey-2021-financial-year-provisional-csv.csv
--2023-03-11 09:33:35--  https://stats.govt.nz/assets/Uploads/Annual-enterprise-survey/Annual-enterprise-survey-2021-financial-year-provisional/Download-data/annual-enterprise-survey-2021-financial-year-provisional-csv.csv
Resolving stats.govt.nz (stats.govt.nz)... 45.60.11.104, 45.60.15.104
Connecting to stats.govt.nz (stats.govt.nz)|45.60.11.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6617174 (6.3M) [text/csv]
Saving to: 'annual-enterprise-survey-2021-financial-year-provisional-csv.csv'

100%[=======================================================================================>] 6,617,174   1.89MB/s   in 3.5s

2023-03-11 09:33:42 (1.81 MB/s) - 'annual-enterprise-survey-2021-financial-year-provisional-csv.csv' saved [6617174/6617174]
```

    b. Pengecekan File

```
[root@sandbox-hdp ~]# ls                                                    Cek di local mesin
anaconda-ks.cfg  annual-enterprise-survey-2021-financial-year-provisional-csv.csv
[root@sandbox-hdp ~]# hadoop fs -ls
[root@sandbox-hdp ~]# hadoop fs -mkdir practice                             Membuat folder practice
mkdir: Cannot create directory /user/root/practice. Name node is in safe mode.
[root@sandbox-hdp ~]# hdfs dfsadmin -safemode leave
safemode: Access denied for user root. Superuser privilege is required
[root@sandbox-hdp ~]# sudo -u hdfs hdfs dfsadmin -safemode leave
Safe mode is OFF
[root@sandbox-hdp ~]# hadoop fs -mkdir practice
[root@sandbox-hdp ~]# hadoop fs -ls
Found 1 items
drwxr-xr-x   - root hdfs          0 2023-03-11 07:38 practice
[root@sandbox-hdp ~]# mv annual-enterprise-survey-2021-financial-year-provisional-csv.csv sample.csv   Rename file
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg  sample.csv
[root@sandbox-hdp ~]# hadoop fs -copyFromLocal sample.csv practice/sample.csv
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg  sample.csv
[root@sandbox-hdp ~]# hadoop fs -ls                                         Menyimpan file kedalam
Found 1 items
drwxr-xr-x   - root hdfs          0 2023-03-11 07:42 practice
[root@sandbox-hdp ~]# hadoop fs -ls practice
Found 1 items
-rw-r--r--   1 root hdfs    6617174 2023-03-11 07:42 practice/sample.csv
[root@sandbox-hdp ~]# hadoop fs -rm practice/sample.csv
23/03/11 07:44:21 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/user/root/practice/sample.csv' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/root/.Trash/Current/user/root/practice/sample.csv
```

1.5. Penugasan

Kerjakan sesuai dengan yang dijelaskan pada bagian Kegiatan Praktikum dan kerjakan tugas praktikum sebagai berikut:

1. Melakukan load data ke hadoop dengan mekanisme yang telah dijelaskan sebelumnya.
2. Melakukan pengecekan system HDFS dan melakukan pengecekan file berada dalam blok data yang sesuai.