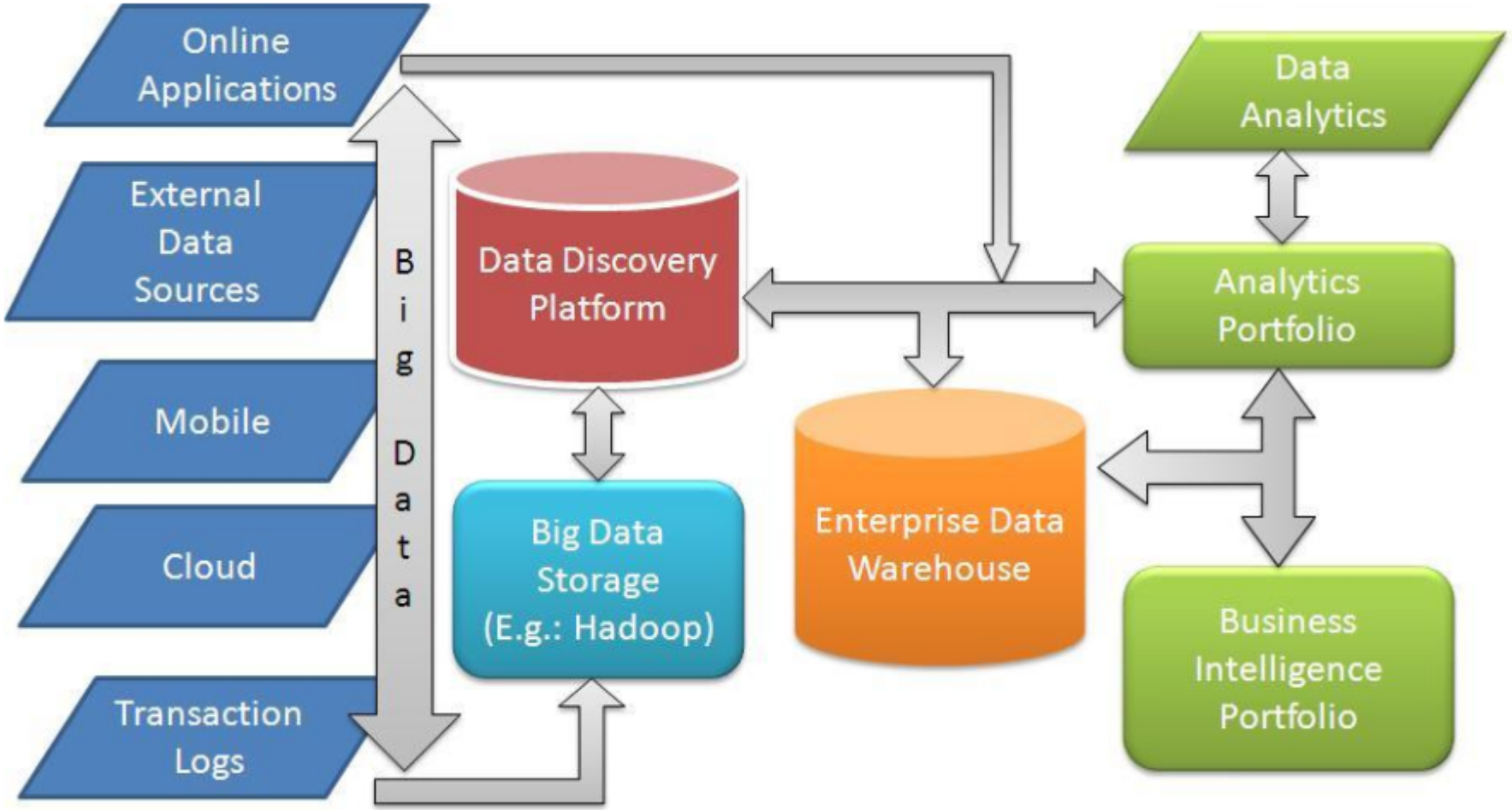
Hadoop Environment

Ahmad Rio Adriansyah S.Si. M.Si

Ekosistem Big Data

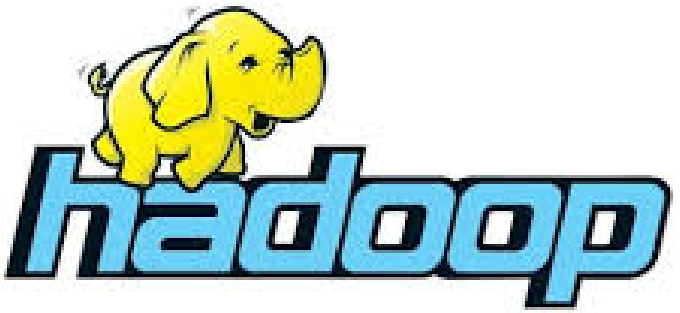


Hadoop

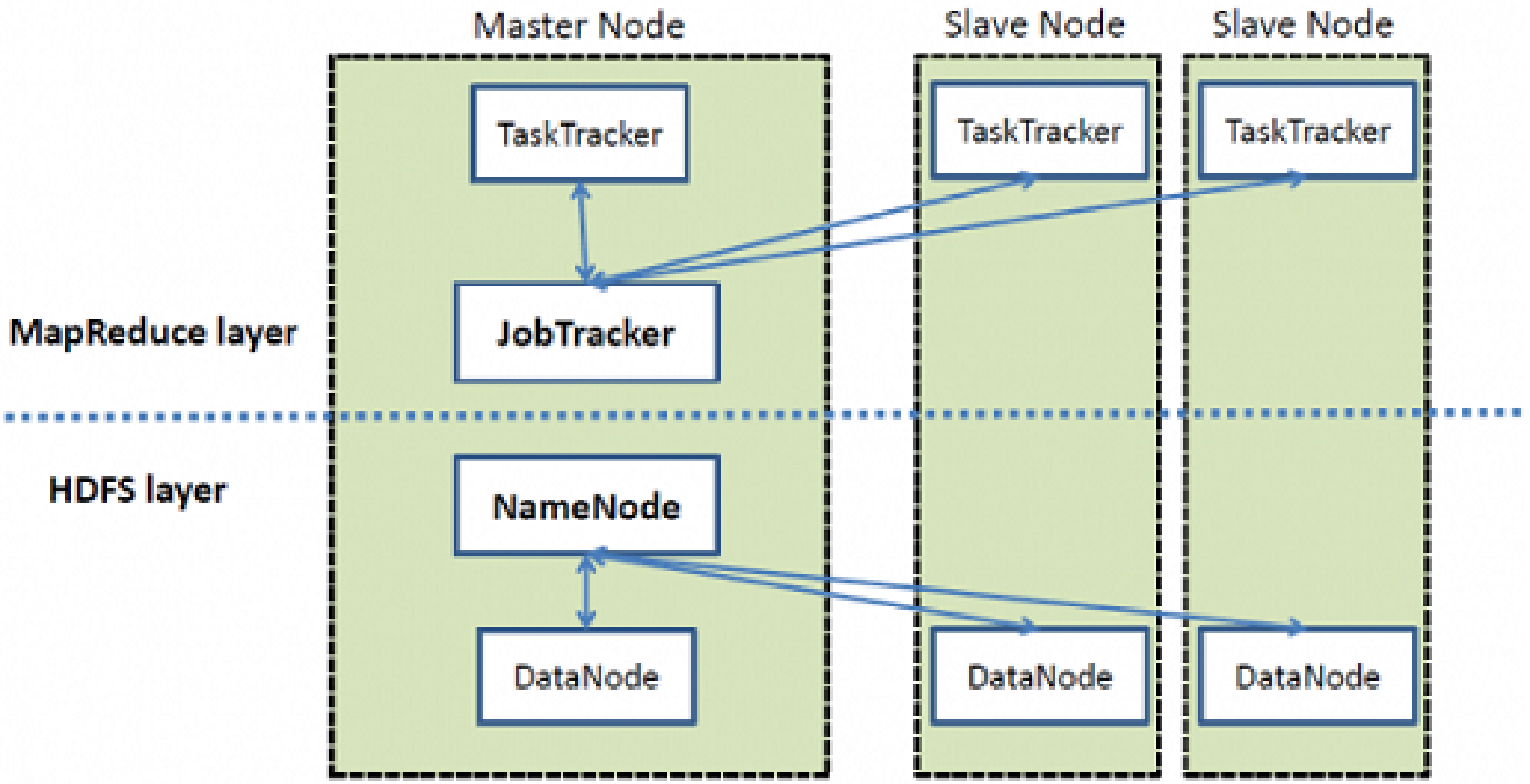
- Apa itu Hadoop?

  Sekumpulan program dan prosedur open source untuk operasi pada dataset yang besar

- 4 Modul Utama Hadoop:
  - Distributed File System (HDFS)
  - Map Reduce
  - Hadoop Common
  - Resource Manager (YARN)

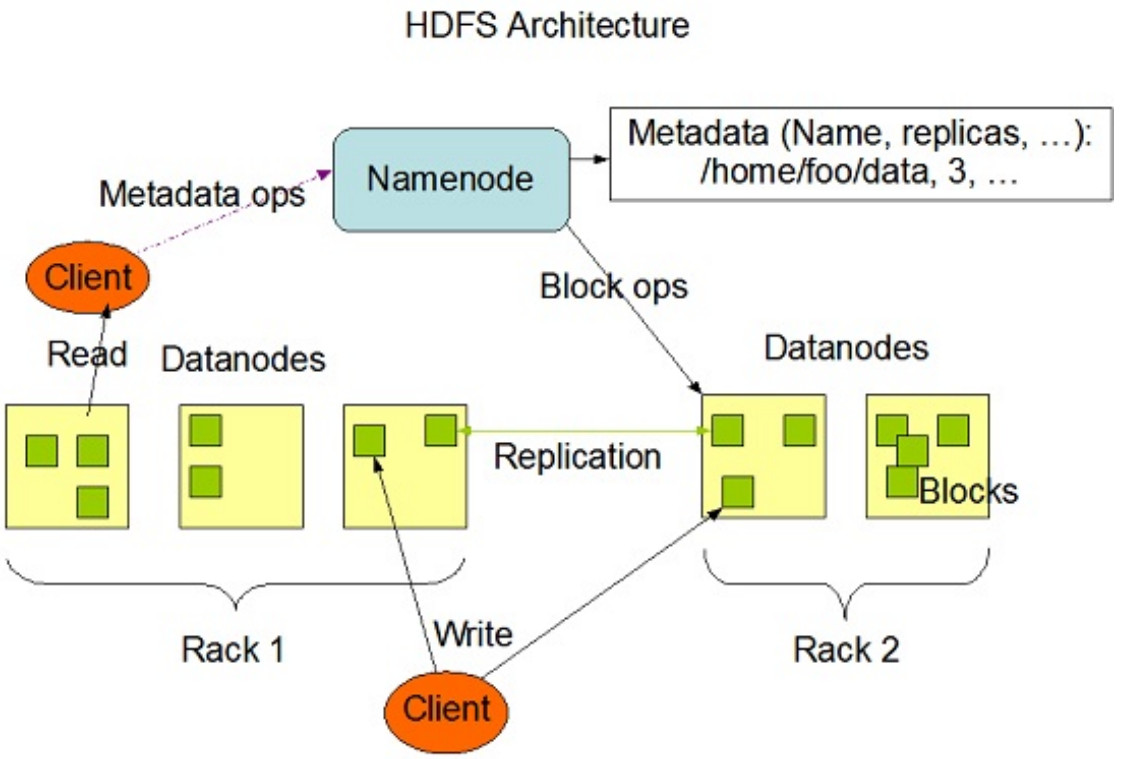


Arsitektur Hadoop

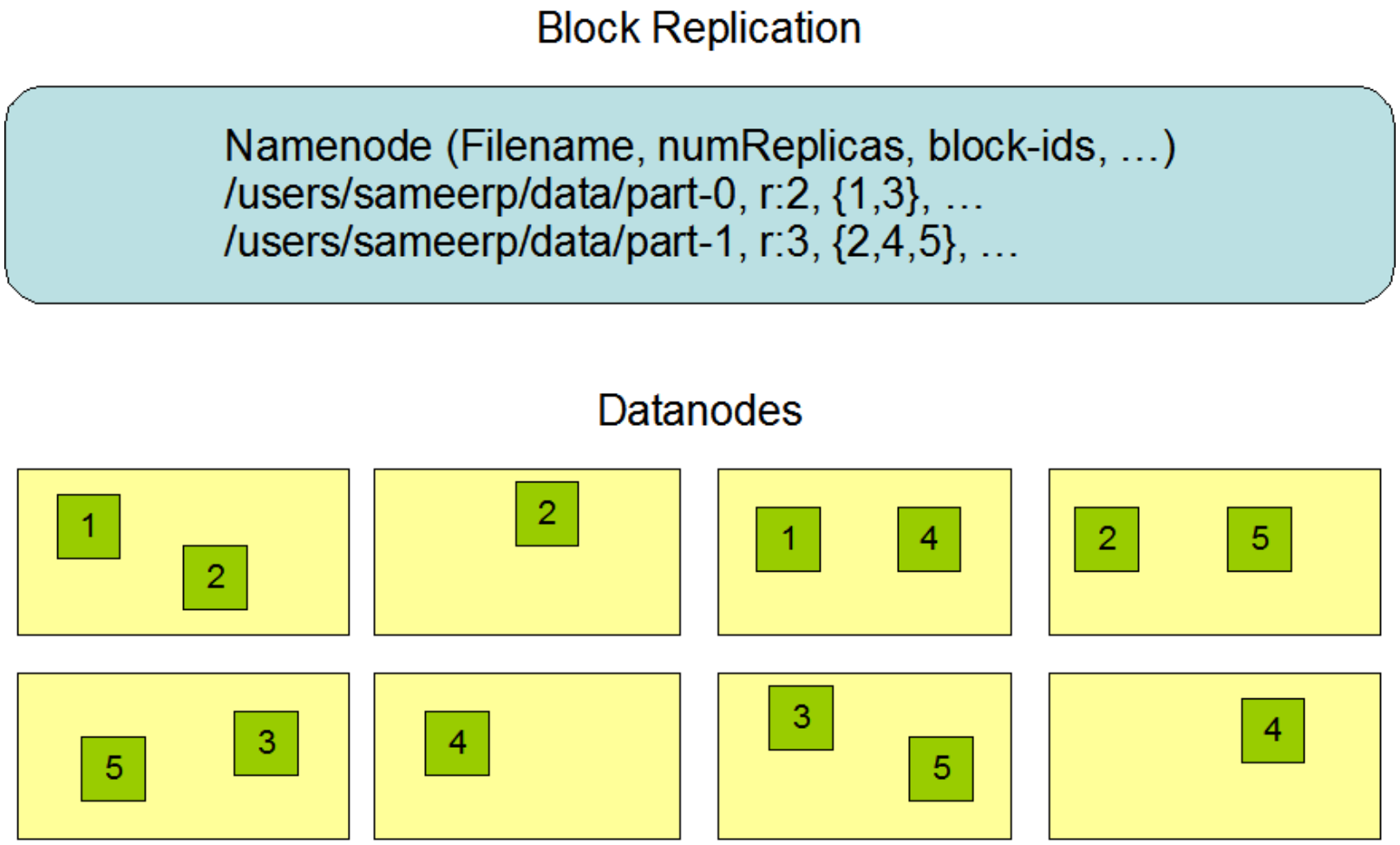


HDFS

- Distributed File System
- Run on Commodity Hardware

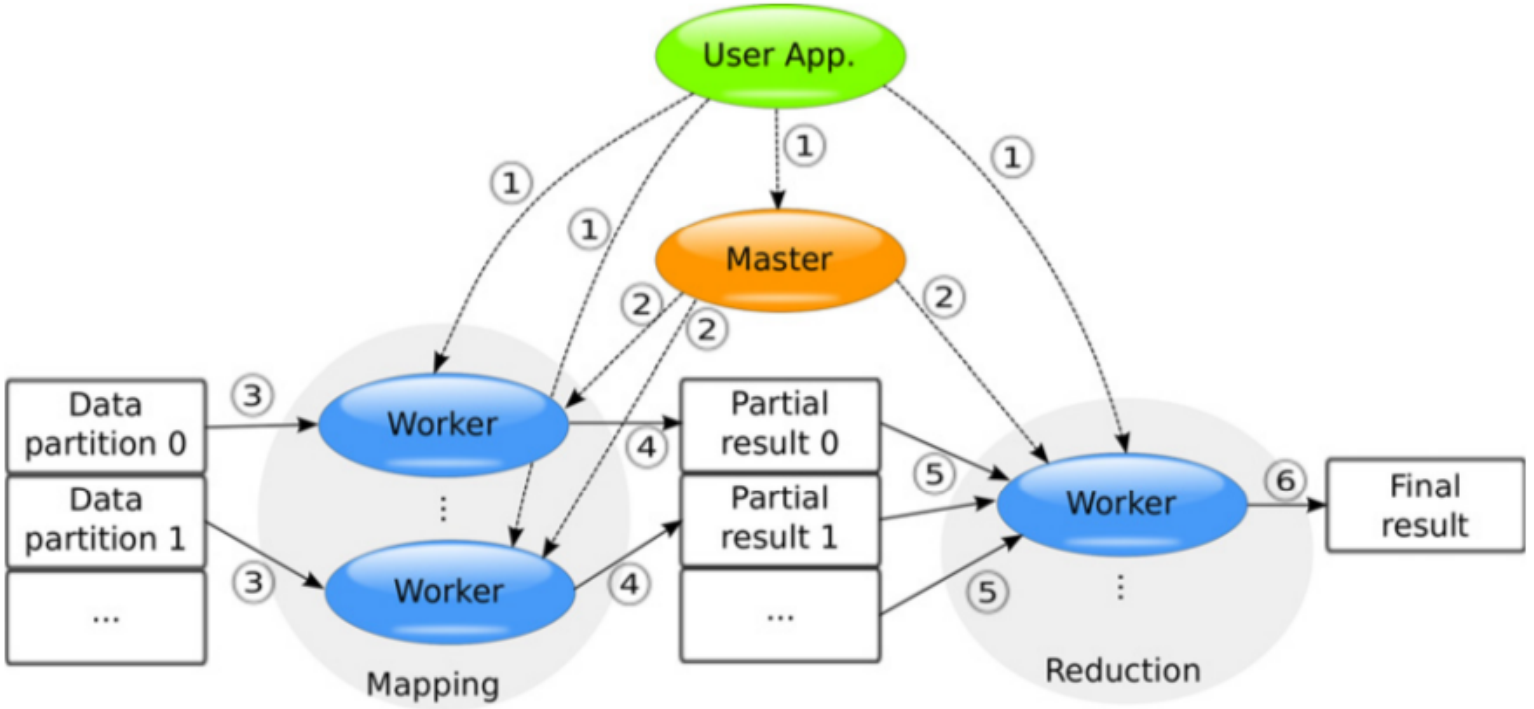


HDFS

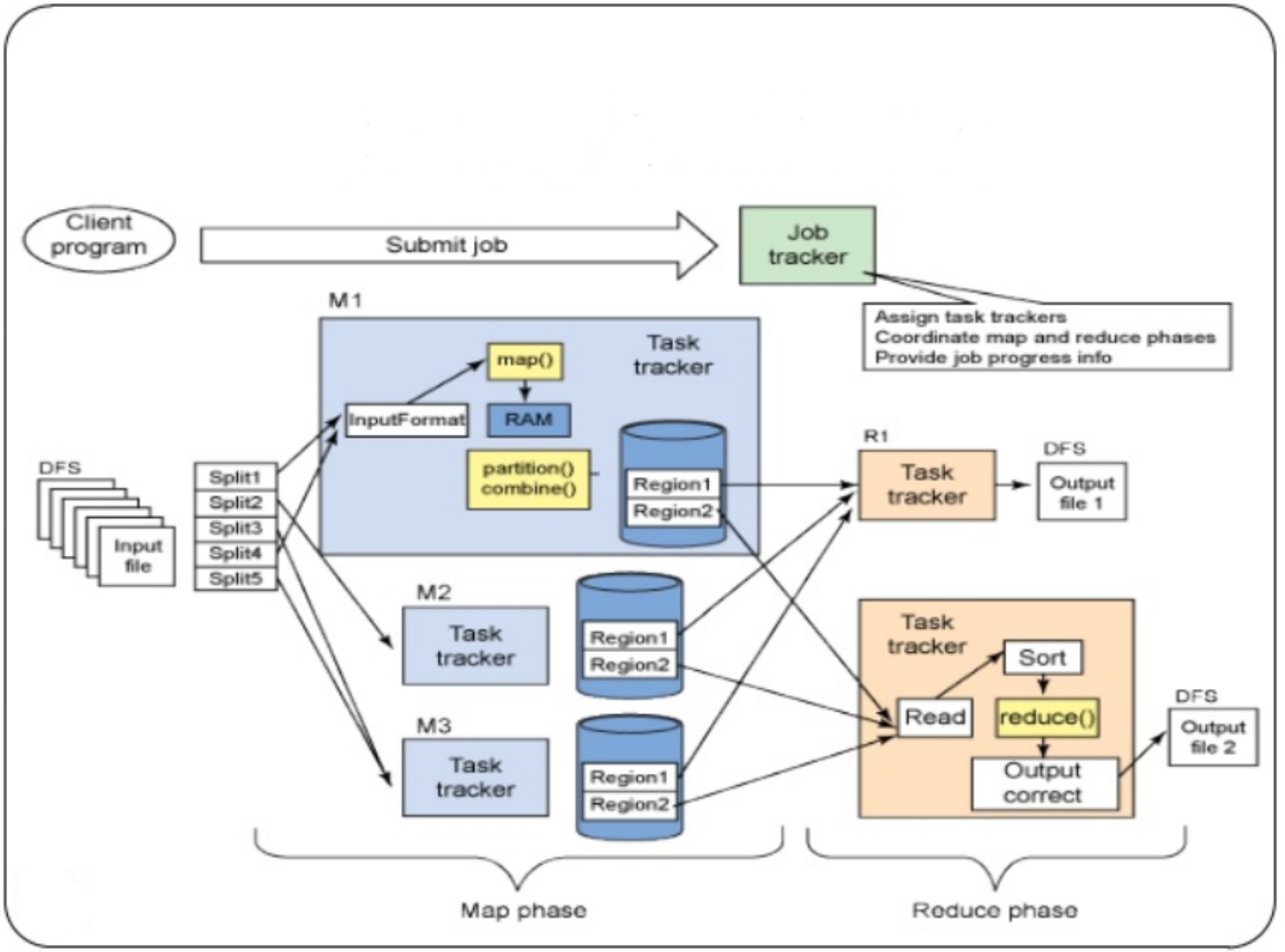

Block Replication
Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …
Datanodes
1
2
2
1
4
2
5
5
3
4
3
5
4

Map Reduce

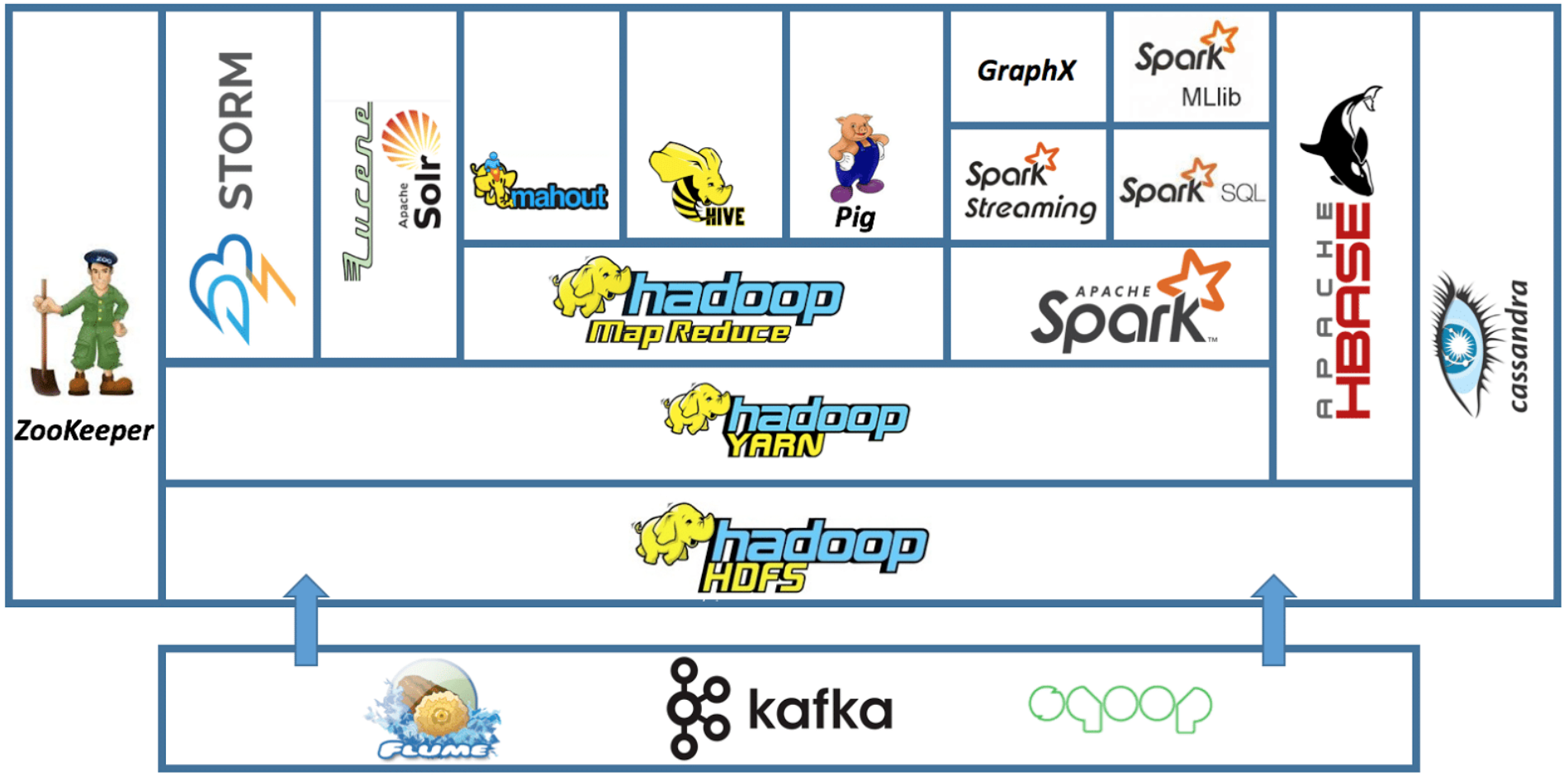
- Distributed Computing Framework
- Parallel Programming Model



Map Reduce


Client program
Submit job
Job tracker
Assign task trackers
Coordinate map and reduce phases
Provide job progress info
M1
Task tracker
map()
InputFormat
RAM
partition()
combine()
Region1
Region2
DFS
Input file
Split1
Split2
Split3
Split4
Split5
R1
Task tracker
DFS
Output file 1
M2
Task tracker
Region1
Region2
M3
Task tracker
Region1
Region2
Task tracker
Sort
Read
reduce()
Output correct
DFS
Output file 2
Map phase
Reduce phase

Ekosistem Hadoop



Instalasi Hadoop

- Single Node Setup
- Cluster Setup
- On Cloud (di luar bahasan)

Hadoop Single Node Setup

- Prasyarat
  - **Java 8**

    https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
  - **Maven (kalau mau build dari source)**

    http://maven.apache.org/download.cgi
  - **Hadoop versi 3.x**

    https://hadoop.apache.org/releases.html

    * **Note** : Hadoop versi 3.x hanya mendukung Java 8

Persiapan

- Setting ip komputer ke dalam hosts
  $ sudo gedit /etc/hosts

```
192.168.45.1    node-master
```

- Setting user group
  $ sudo addgroup hadoop
- Buat user baru dan masukkan ke group
  $ sudo adduser –ingroup hadoop hduser
- Konfigurasi SSH

Konfigurasi SSH

- Install SSH Server
  $ sudo apt install openssh-server
- Aktifkan SSH
  $ sudo service ssh start
  $ sudo systemctl status ssh
- Generate SSH key pada hduser
  $ su - hduser
  $ ssh-keygen -t rsa -P ""
- Beri hak otorisasi ke komputer lokal
  $ cat /home/hduser/.ssh/id_rsa.pub >>
  /home/hduser/.ssh/authorized_keys

Nonaktifkan IPV6

- Masukkan baris berikut pada file **/etc/systctl.conf**

```
# disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

- Reboot
- Periksa apakah IPV6 sudah tidak aktif
  $ cat /proc/sys/net/ipv6/conf/all/disable_ipv6

Periksa instalasi Java

- Periksa dengan

  $ java -version

- Apabila Java 8 sudah terinstall, langkah instalasi bisa diskip

Instalasi Java (JDK 8) - Linux

- Download
- Extract
  $ tar -xvzf jdk-8u201-linux-x64.tar.gz
- Tambahkan PATH untuk JAVA_HOME
  $ sudo gedit .bashrc

```
JAVA_HOME=/home/ubuntu/BigData/jdk1.8.0_231
export JAVA_HOME
export PATH=$PATH:$JAVA_HOME/bin
```

Instalasi Hadoop

- Download
- Extract
- Tambahkan PATH untuk HADOOP_HOME
- Konfigurasi namenode, datanode
- Konfigurasi map reduce
- Konfigurasi job scheduler (yarn)
- Format HDFS

Instalasi Hadoop

- Download
- Extract
  $ tar -xvzf hadoop-3.2.1.tar.gz
- Tambahkan PATH untuk HADOOP_HOME
  $ sudo gedit .bashrc

```
HADOOP_HOME=/home/ubuntu/BigData/hadoop-3.2.1
export HADOOP_HOME
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

/etc/hadoop/hadoop_env.sh

- Masukkan JAVA_HOME dan HADOOP_HOME ke dalam environment hadoop

```
export JAVA_HOME=/home/ubuntu/BigData/jdk1.8.0_231
export HADOOP_HOME=/home/ubuntu/BigData/hadoop-3.2.1
```

/etc/hadoop/core-site.xml

```
<configuration>
      <property>
            <name>fs.default.name</name>
            <value>hdfs://node-master:9000</value>
      </property>
</configuration>
```

/etc/hadoop/hdfs-site.xml

```
<configuration>

      <property>
            <name>dfs.namenode.name.dir</name>
            <value>/home/hduser/data/nameNode</value>
      </property>

      <property>
            <name>dfs.namenode.data.dir</name>
            <value>/home/hduser/data/dataNode</value>
      </property>

      <property>
            <name>dfs.replication</name>
            <value>1</value>
      </property>

</configuration>
```

/etc/hadoop/mapred-site.xml

```
<configuration>

      <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
      </property>

      <property>
            <name>yarn.app.mapreduce.am.env</name>
            <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
      </property>

      <property>
            <name>mapreduce.map.env</name>
            <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
      </property>

      <property>
            <name>mapreduce.reduce.env</name>
            <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
      </property>


</configuration>
```

/etc/hadoop/yarn-site.xml

```
<configuration>

<!-- Site specific YARN configuration properties -->

      <property>
            <name>yarn.acl.enable</name>
            <value>0</value>
      </property>

      <property>
            <name>yarn.resourcemanager.hostname</name>
            <value>node-master</value>
      </property>

      <property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
      </property>

      <property>
            <name>yarn.nodemanager.aux-
services.mapreduce_shuffle.class</name>
            <value>org.apache.hadoop.mapred.ShuffleHandler</value>
      </property>

</configuration>
```

Format HDFS

- **$ sudo su – hduser**
- **$ hdfs namenode -format**

Memulai Services

- Menjalankan HDFS
  $ start-dfs.sh
- Menjalankan YARN
  $ start-yarn.sh
- Periksa node yang berjalan
  $ jps
- Untuk mengakhiri servisnya bisa menggunakan
  $ stop-dfs.sh
  $ stop-yarn.sh

Monitoring HDFS dan YARN

- Periksa servisnya berjalan di port mana
  $ netstat -tanp | grep java
- Buka pada browser
  - HDFS berjalan pada localhost:<port>
  - YARN berjalan di ip komputer master port 8088
- Atau jalankan perintah
  $ hdfs dfsadmin -report

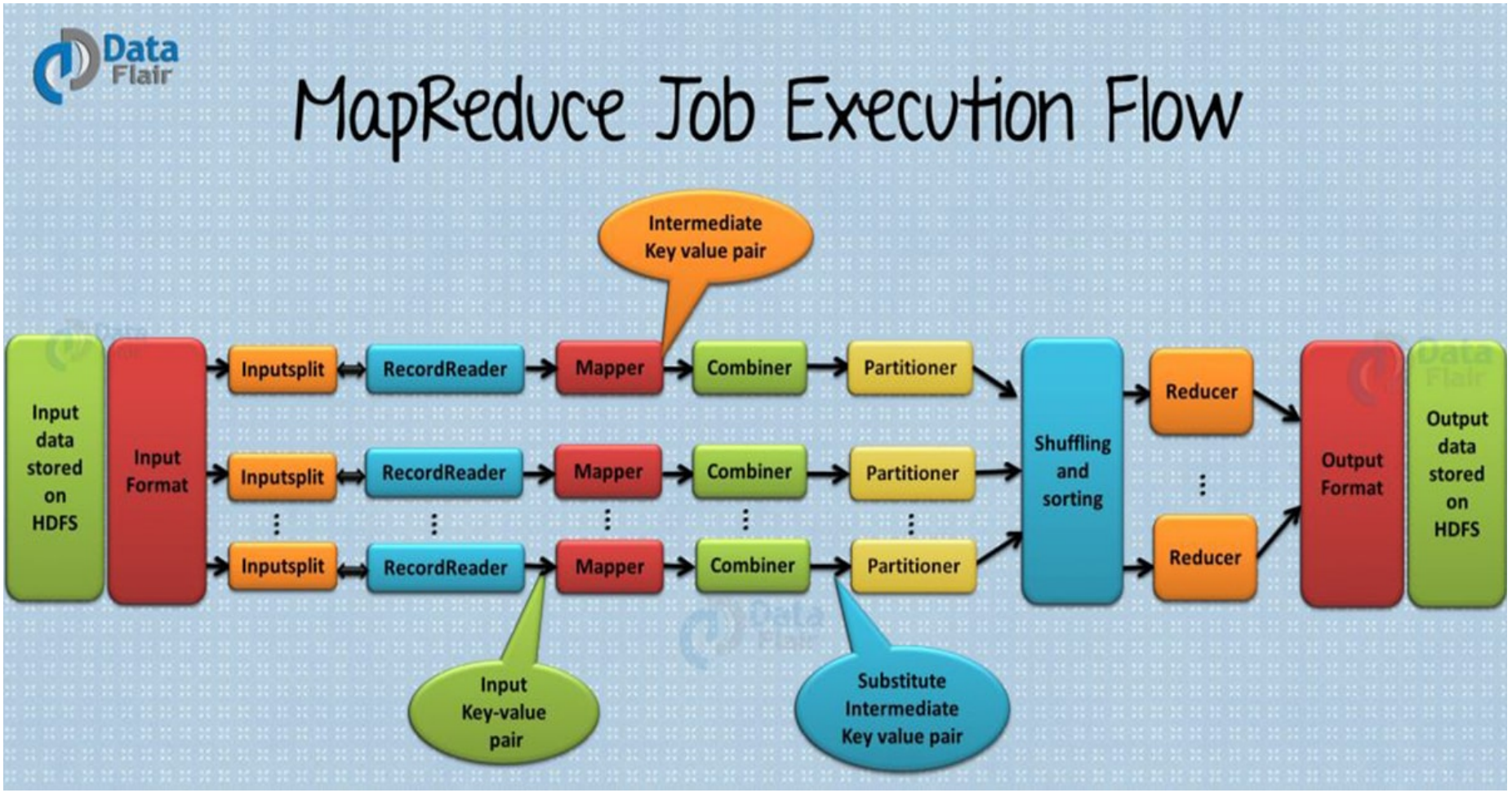
Operasi File Hadoop

- **hdfs** [options] [subcommand] [subcom opts]
- **dfs** : subcommand untuk filesystem
  - cat
  - copyFromLocal
  - copyToLocal
  - get
  - ls
  - mkdir
  - rm

Map Reduce

- Distributed Processing
- Terdiri dari 2 bagian :
  - Mapper = memproses input dan menghasilkan intermediate output berupa key-value pair
  - Reducer = menerima output dari mapper dan memprosesnya ke output final

Alur MapReduce



Hadoop Streaming

- Hadoop dibuat berbasis Java. Tetapi kita dapat memanfaatkan bahasa pemrograman lain untuk menjalankan proses pada Hadoop.
- Salah satunya dengan Hadoop Streaming
- Utilitas ini membaca baris per baris dari stdin dan mengeluarkan outputnya ke stdout

Case : Word Count

- Mapper -> membagi dokumen menjadi kata per kata
- Reducer -> menghitung jumlah kata

Word Count Dengan Hadoop

- Sudah disediakan 3 file dalam folder Sherlock
- Masukkan file ke dalam HDFS

  $ hdfs dfs -copyFromLocal Sherlock/ /user/hduser

- Periksa filenya apakah sudah ada dalam HDFS

  $ hdfs dfs -ls /user/hduser/Sherlock

File Mapper dan Reducer

- Sudah disediakan file **mapper.py** dan **reducer.py** pada folder **MapRed**
- Mapper dan reducernya dituliskan dengan bahasa python untuk word counting. Pastikan python 2.7 terinstall pada sistem
- Pastikan folder MapRed tersebut dapat dieksekusi oleh hduser

  $ sudo chown -R hduser:hadoop MapRed/

  $ sudo chmod 777 MapRed/

Hadoop Streaming

```
$   $HADOOP_HOME/bin/hadoop jar
    $HADOOP_HOME/share/hadoop/tools/
lib/hadoop-streaming*.jar
    -file MapRed/mapper.py
    -mapper MapRed/mapper.py
    -file MapRed/reducer.py
    -reducer MapRed/reducer.py
    -input /user/hduser/Sherlock/*
    -output /user/hduser/output
```

Word Count

- Mapreduce akan berjalan terhadap file pada folder sherlock di HDFS dan akan menghasilkan output di folder outputnya
- Periksa hasil

$ hdfs dfs -ls /user/hduser/output

$ hdfs dfs -cat /user/hduser/output/part-00000

Next?

- Tersedia file berukuran lebih besar, yaitu hasil crawling web berita detik.com tahun 2015
- Ukurannya lebih besar dari 100 MB
- Proses word counting terhadap file tersebut