

RESPONSI PRAKTIKUM

INFRASTRUKTUR BIG DATA



Disusun Oleh :

- 1. Husef Sholikhul Ibad (185410110)**
- 2. Annisa Salsabila (185410070)**

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
AKAKOM YOGYAKARTA

2020/2021

Apache Spark adalah teknologi komputasi clustering yang sangat cepat dan dirancang untuk kebutuhan yang memerlukan penanganan data secara cepat seperti big data dan machine learning.

Fitur andalan Apache spark adalah kumpulan memori yang dapat meningkatkan kecepatan pemrosesan aplikasi. Spark dirancang untuk menutupi berbagai beban kerja, seperti proses aplikasi, algoritma berulang-ulang, query interaktif, dan transmisi. Selain mendukung semua beban kerja pada setiap sistem, fitur apache spark ini juga dapat mengurangi beban maintenance management.

Apache Spark akan mengontrol semua metode data dari berbagai repository, seperti dari Hadoop Distributed classification system (HDFS), NoSQL Database dan penyimpanan data relatif, seperti Apache Hive.

Spark akan mengelola memori pendukung untuk membantu proses yang sedang berjalan, contohnya saat sedang menganalisis data.spark akan membagi semua proses ke dalam memori pendukung sehingga dapat memaksimalkan kinerja sistem.

Spark sendiri terdiri dari Spark Core dan beberapa Library pendukung. inti dari Spark engine adalah distributed execution engine, dan API Java, Scala maupun Python yang kemudian Library tambahan akan berjalan diatas Spark Core untuk melakukan berbagai proses seperti Streaming, SQL, machine learning

Kelemahan Hadoop

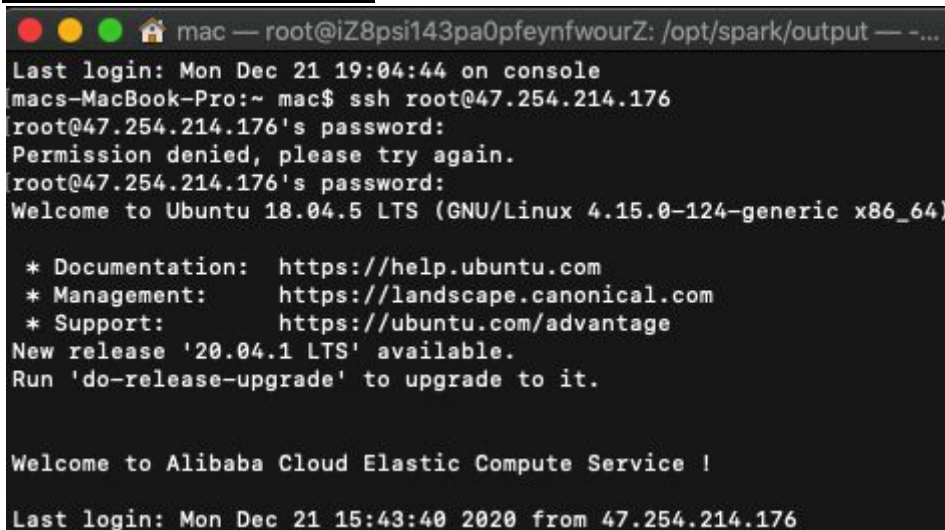
- Kecepatan pemrosesan rendah: di Hadoop, algoritma MapReduce, yang merupakan algoritma paralel dan terdistribusi, memproses kumpulan data yang sangat besar.
- Pemrosesan batch: Hadoop mengimplementasikan pemrosesan batch, yang mengumpulkan data dan kemudian memprosesnya secara massal. Meskipun pemrosesan batch efisien untuk memproses volume data yang besar, ia tidak memproses data transmisi. Akibatnya, kinerjanya menjadi lebih lambat.
- Tidak memiliki Pipeline: Hadoop tidak mendukung pipeline (yaitu, urutan tahapan di mana ID keluaran dari tahap sebelumnya adalah input dari tahap berikutnya).
- Sulit untuk digunakan: Pengembang MapReduce perlu menulis kode mereka sendiri untuk setiap operasi, yang membuat pekerjaan menjadi sangat sulit. Selain itu, MapReduce tidak memiliki mode interaktif.
- Latency: Di Hadoop, struktur MapReduce lebih lambat karena mendukung berbagai format, struktur, dan data yang besar.
- Longline kode: karena Hadoop ditulis dalam Java, kode ini luas. Dan itu membutuhkan waktu lebih lama untuk menjalankan program.

Membangun Infrastruktur Apache Spark

Sebelum memasang perangkat lunak baru, ada baiknya untuk menyegarkan basis data paket perangkat lunak lokal Anda untuk memastikan Anda mengakses versi terbaru.

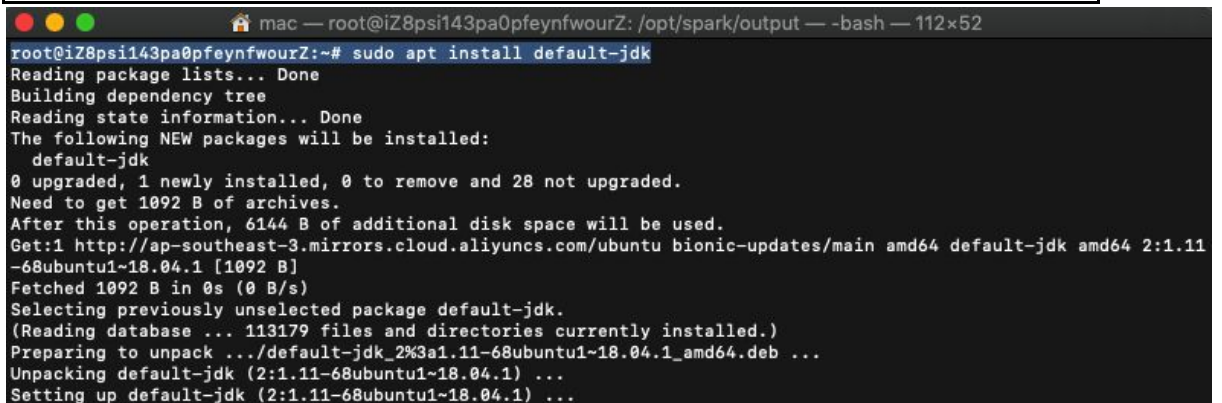
1. Pastikan kalian memiliki vm, buka terminal lalu masukkan kode ssh berikut :

```
ssh root@47.254.214.176
```

A terminal window on a Mac showing an SSH session. The prompt is 'mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -...'. The user enters 'ssh root@47.254.214.176'. The terminal shows the login process: 'Last login: Mon Dec 21 19:04:44 on console', 'macs-MacBook-Pro:~ mac\$ ssh root@47.254.214.176', 'root@47.254.214.176's password:', 'Permission denied, please try again.', 'root@47.254.214.176's password:', 'Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 4.15.0-124-generic x86_64)'. It then displays links for documentation, management, and support, followed by a message about a new release '20.04.1 LTS' and a 'Welcome to Alibaba Cloud Elastic Compute Service !' message. The final line is 'Last login: Mon Dec 21 15:43:40 2020 from 47.254.214.176'.

2. Sebelum menginstall apache spark, kita membutuhkan sebuah Packages / menginstall dependensi yang diperlukan seperti JDK .

```
root@iZ8psi143pa0pfeynfwourZ:~# sudo apt install  
default-jdk
```

A terminal window on a Mac showing the installation of default-jdk. The prompt is 'mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 112x52'. The user enters 'root@iZ8psi143pa0pfeynfwourZ:~# sudo apt install default-jdk'. The terminal shows the installation process: 'Reading package lists... Done', 'Building dependency tree', 'Reading state information... Done', 'The following NEW packages will be installed: default-jdk', '0 upgraded, 1 newly installed, 0 to remove and 28 not upgraded.', 'Need to get 1092 B of archives.', 'After this operation, 6144 B of additional disk space will be used.', 'Get:1 http://ap-southeast-3-mirrors.cloud.aliyuncs.com/ubuntu bionic-updates/main amd64 default-jdk amd64 2:1.11-68ubuntu1-18.04.1 [1092 B]', 'Fetched 1092 B in 0s (0 B/s)', 'Selecting previously unselected package default-jdk.', '(Reading database ... 113179 files and directories currently installed.)', 'Preparing to unpack .../default-jdk_2%3a1.11-68ubuntu1-18.04.1_amd64.deb ...', 'Unpacking default-jdk (2:1.11-68ubuntu1-18.04.1) ...', 'Setting up default-jdk (2:1.11-68ubuntu1-18.04.1) ...'.

3. Cek java Version, sekaligus mengecek apakah JDK sudah terinstall seperti yang kita inginkan.

```
root@iZ8psi143pa0pfeynfwourZ:~# java  
--version
```

```

mac — root@iZ8psi143pa0pfeynfourZ: /opt/spark/output — -bash — 11
root@iZ8psi143pa0pfeynfourZ:~# java --version
openjdk 11.0.9.1 2020-11-04
OpenJDK Runtime Environment (build 11.0.9.1+1-Ubuntu-0ubuntu1.18.04)
OpenJDK 64-Bit Server VM (build 11.0.9.1+1-Ubuntu-0ubuntu1.18.04, mixed mode, sharing)

```

4. Unduh Spark dari situs web dengan menggunakan perintah wget/curl dan tautan langsung untuk mengunduh arsip :

```

root@iZ8psi143pa0pfeynfourZ:~# curl -O
https://downloads.apache.org/spark/spark-2.4.7/spark-2.4.7-
bin-hadoop2.7.tgz

```

```

mac — root@iZ8psi143pa0pfeynfourZ: /opt/spark/output — -bash — 87x52
root@iZ8psi143pa0pfeynfourZ:~# curl -O https://downloads.apache.org/spark/spark-2.4.7/
spark-2.4.7-bin-hadoop2.7.tgz

```

% Total	% Received	% Xferd	Average Speed	Time	Time	Time	Current
			Dload Upload	Total	Spent	Left	Speed
100	222M	100	222M	0	0	10.8M	0
				0:00:20	0:00:20	--:--:--	11.5M

Diatas merupakan perintah untuk mendownload paket dari Apache Spark.

5. Selanjutnya Extract file spark yang kita download tadi dengan perintah tar. Disini dilakukan proses kompresi file apache yang sudah didownload sebelumnya dengan menggunakan perintah dibawah ini :

```

root@iZ8psi143pa0pfeynfourZ:~# tar xvzf
spark-2.4.7-bin-hadoop2.7.tgz

```

```

mac — root@iZ8psi143pa0pfeynfourZ: /opt/spark/output — -bash — 87x52
root@iZ8psi143pa0pfeynfourZ:~# tar xvzf spark-2.4.7-bin-hadoop2.7.tgz
spark-2.4.7-bin-hadoop2.7/
spark-2.4.7-bin-hadoop2.7/kubernetes/
spark-2.4.7-bin-hadoop2.7/kubernetes/tests/
spark-2.4.7-bin-hadoop2.7/kubernetes/tests/py_container_checks.py

```

6. setelah file ter extract semua, pindahkan spark-2.4.7-bin-hadoop2.7 ke directory /opt/spark, menggunakan perintah mv.

```

root@iZ8psi143pa0pfeynfourZ:~# sudo
mv spark-2.4.7-bin-hadoop2.7 /opt/spark

```

```

mac — root@iZ8psi143pa0pfeynfourZ: /opt/spark/output — -bash — 87x52
root@iZ8psi143pa0pfeynfourZ:~# sudo mv spark-2.4.7-bin-hadoop2.7 /opt/spark

```

7. Kemudian masuklah ke directory /opt/spark dengan menggunakan perintah cd.

```

root@iZ8psi143pa0pfeynfourZ:~# cd
/opt/spark/

```

```

mac — root@iZ8psi143pa0pfeynfourZ: /
root@iZ8psi143pa0pfeynfourZ:~# cd /opt/spark/

```

8. setelah masuk ke directory /opt/spark, run master.sh menggunakan perintah berikut :

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark#  
sudo sbin/start-master.sh
```

```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 87x52  
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# sudo sbin/start-master.sh  
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark-root-  
org.apache.spark.deploy.master.Master-1-iZ8psi143pa0pfeynfwourZ.out
```

Setelah dikonfigurasi, selanjutnya adalah start server-master spark. Perintah sebelumnya menambahkan direktori yang diperlukan ke variabel PATH sistem, jadi perintah ini dapat dijalankan dari direktori manapun ;

9. Gunakan perintah cat untuk mengetahui url spark untuk digunakan menjalankan perintah slave nantinya.

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# cat  
/opt/spark/logs/spark-root-org.apache.spark.deploy.master.M  
aster-1-iZ8psi143pa0pfeynfwourZ.out
```

```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 87x52  
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# cat /opt/spark/logs/spark-root-org.apache.spar  
k.deploy.master.Master-1-iZ8psi143pa0pfeynfwourZ.out  
Spark Command: /usr/lib/jvm/java-11-openjdk-amd64/bin/java -cp /opt/spark/conf:/opt/sp  
ark/jars/* -Xmx1g org.apache.spark.deploy.master.Master --host iZ8psi143pa0pfeynfwourZ  
--port 7077 --webui-port 8080  
=====
```

10. Selanjutnya kita jalankan Spark Slave Server menggunakan perintah :

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# sudo  
sbin/start-slave.sh  
spark://iZ8psi143pa0pfeynfwourZ:7077
```

Perintah diatas sesuai dengan hasil langkah ke 9

```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 87x52  
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# sudo sbin/start-slave.sh spark://iZ8psi143pa0p  
feynfwourZ:7077  
starting org.apache.spark.deploy.worker.Worker, logging to /opt/spark/logs/spark-root-o  
rg.apache.spark.deploy.worker.Worker-1-iZ8psi143pa0pfeynfwourZ.out
```

Perintah ini digunakan untuk mengetahui proses yang sedang berjalan dan lokasi spark yang dapat diakses melalui web.

11. Buat lah file **input.txt** menggunakan perintah pico seperti berikut :

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# pico  
input.txt
```



```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 87x52
root@iZ8psi143pa0pfeynfwourZ:/opt/spark# pico input.txt
```

Lalu isilah dengan konten dibawah ini, kemudian save dengan tekan **CTRL X** , lalu tekan **Y**, kemudian tekan **Enter**.

Pada tanggal 30 Juni 1979 didirikan Yayasan Pendidikan Widya Bakti yang bertujuan mengembangkan dan menyebarluaskan pengetahuan tentang teknologi komputer dan informatika di kalangan masyarakat Indonesia melalui usaha pendidikan yang sistematis dan ilmiah. Yayasan tersebut mengelola sebuah akademi yang bernama Akademi Aplikasi Komputer, yang kemudian disingkat menjadi AKAKOM. Kemudian pada tanggal 31 Maret 1983, Akademi Aplikasi Komputer (AKAKOM), diubah menjadi Akademi Komputer dan Informatika AKAKOM. Pada tanggal 2 Mei 1985, nama Akademi Komputer dan Informatika AKAKOM diubah dan dibakukan menjadi Akademi Manajemen Informatika dan Komputer (AMIK) AKAKOM. AMIK AKAKOM kemudian berubah menjadi Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) AKAKOM pada tanggal 8 Juni 1992, berdasarkan Surat Keputusan Direktorat Jenderal Pendidikan Tinggi Departemen Pendidikan dan Kebudayaan RI Nomor 262/DIKTI/Kep/1992 dengan status terdaftar bagi program Sarjana dan status diakui bagi program diplamanya.

```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark — ssh root@47.254.214.176 — 90x52
GNU nano 2.9.3                               input.txt                               Modified
Pada tanggal 30 Juni 1979 didirikan Yayasan Pendidikan Widya Bakti yang bertujuan mengemb$
```

12. Lalu bukalah Spark Shell menggunakan perintah :

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark#
bin/spark-shell
```

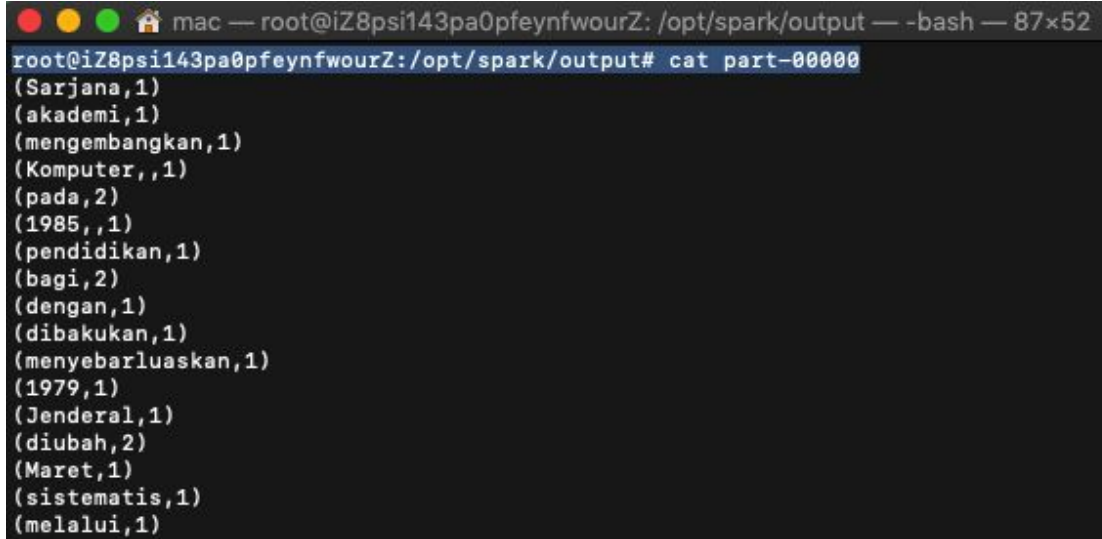

16. Kemudian untuk melihat hasilnya Kita harus masuk ke directory output terlebih dahulu gunakan perintah cd seperti berikut :

```
scala> root@iZ8psi143pa0pfeynfwourZ:/opt/spark# cd output/
```

Lalu gunakan perintah cat untuk melihat hasilnya :

```
root@iZ8psi143pa0pfeynfwourZ:/opt/spark/output# cat part-00000
```

Output dari praktikum diatas adalah sbb ;



```
mac — root@iZ8psi143pa0pfeynfwourZ: /opt/spark/output — -bash — 87x52
root@iZ8psi143pa0pfeynfwourZ:/opt/spark/output# cat part-00000
(Sarjana,1)
(akademi,1)
(mengembangkan,1)
(Komputer,,1)
(pada,2)
(1985,,1)
(pendidikan,1)
(bagi,2)
(dengan,1)
(dibakukan,1)
(menyebarkan,1)
(1979,1)
(Jenderal,1)
(diubah,2)
(Maret,1)
(sistematis,1)
(melalui,1)
```