

LAPORAN PRAKTIKUM 1 ANALISIS BIG DATA

INSTALASI HADOOP

Dosen Pengampu: Sevi Nurafni ST., M.Si., M.Sc.



IKOPIN

University

Disusun Oleh Ryan Fadhilah Faizal Hakim

Program Studi Sains Data

Fakultas Sains dan Teknologi

©2024 Ryan Hakim All Rights Reserved

Pendahuluan

Hadoop diciptakan oleh Doug Cutting, pencipta Apache Lucene, perpustakaan pencarian teks yang banyak digunakan. Hadoop berasal dari Apache Nutch, mesin pencari web open source, yang merupakan bagian dari proyek Lucene.

Asal Nama “Hadoop”

Nama Hadoop bukanlah akronim; Itu nama yang dibuat-buat. Pencipta proyek, Doug Cutting, menjelaskan bagaimana nama itu muncul:

“Nama yang diberikan oleh anak saya kepada boneka gajah kuning. Pendek, relatif mudah dieja dan diucapkan, tidak berarti, dan tidak digunakan di tempat lain: itulah kriteria penamaan saya. Anak-anak pandai menghasilkan hal seperti itu Googol adalah istilah anak-anak”.

Subproyek dan modul "contrib" di Hadoop juga cenderung memiliki nama yang tidak terkait dengan fungsinya, seringkali dengan tema gajah (Elephant) atau hewan lainnya ("Pig," misalnya). Komponen yang lebih kecil diberi nama yang lebih deskriptif (dan karenanya lebih biasa). Ini adalah prinsip yang baik, karena itu berarti Anda umumnya dapat mengetahui apa yang dilakukan sesuatu dari namanya. Misalnya, jobtracker9 melacak pekerjaan MapReduce.

Apache Hadoop adalah kerangka perangkat lunak sumber terbuka (open source) yang digunakan untuk menyimpan dan memproses kumpulan data besar secara efisien mulai dari ukuran gigabyte hingga petabyte data. Alih-alih menggunakan satu komputer besar untuk menyimpan dan memproses data, Hadoop memungkinkan pengelompokan beberapa komputer untuk menganalisis kumpulan data besar secara paralel dengan lebih cepat.

Hadoop membuatnya lebih mudah untuk menggunakan semua kapasitas penyimpanan dan pemrosesan di server cluster, dan untuk menjalankan proses terdistribusi terhadap sejumlah besar data. Hadoop menyediakan blok bangunan di mana layanan dan aplikasi lain dapat dibangun.

Aplikasi yang mengumpulkan data dalam berbagai format dapat menempatkan data ke dalam kluster Hadoop dengan menggunakan operasi API untuk terhubung ke NameNode. NameNode melacak struktur direktori file dan penempatan "chunk" untuk setiap file, direplikasi di DataNodes. Untuk menjalankan pekerjaan guna mengkueri data, sediakan pekerjaan MapReduce yang terdiri dari banyak peta dan kurangi tugas yang berjalan terhadap data dalam HDFS yang tersebar di DataNodes. Tugas peta berjalan pada setiap node terhadap file input yang disediakan, dan reduksi berjalan untuk menggabungkan dan mengatur output akhir.

Dalam ekosistem Hadoop terdapat beberapa modul diantaranya:

1. Common

Satu set komponen dan antarmuka untuk sistem file terdistribusi dan general I/O (Input/Output)

2. Avro
Sistem serialisasi untuk RPC lintas bahasa yang efisien dan penyimpanan data persisten.
3. MapReduce
Model pemrosesan data terdistribusi dan lingkungan eksekusi yang berjalan pada kelompok besar mesin komoditas.
4. HDFS
Sebuah filesystem terdistribusi yang berjalan pada kelompok besar mesin komoditas.
5. Pig
Bahasa aliran data dan lingkungan eksekusi untuk menjelajahi himpunan data yang sangat besar. Pig berjalan pada kluster HDFS dan MapReduce.
6. Hive
Gudang data terdistribusi. Apache Hive mengelola data yang disimpan dalam HDFS dan menyediakan bahasa kueri berdasarkan SQL (dan yang diterjemahkan oleh mesin runtime ke pekerjaan MapReduce) untuk mengkueri data.
7. Hbase
Database berorientasi kolom yang terdistribusi. HBase menggunakan HDFS untuk penyimpanan yang mendasarinya, dan mendukung komputasi gaya batch menggunakan MapReduce dan kueri titik (pembacaan acak).
8. ZooKeeper
Layanan koordinasi yang terdistribusi dan sangat tersedia. ZooKeeper menyediakan primitif seperti kunci terdistribusi yang dapat digunakan untuk membangun aplikasi terdistribusi.
9. Sqoop
Alat untuk transfer data massal yang efisien antara penyimpanan data terstruktur (seperti database relasional) dan HDFS.
10. Oozie
Layanan untuk menjalankan dan menjadwalkan alur kerja pekerjaan Hadoop (termasuk pekerjaan MapReduce, Pig, Apache Hive, dan Sqoop).

Dalam praktikum kali ini, dibuatkan suatu proses penginstalan Hadoop di perangkat komputer.

Metode

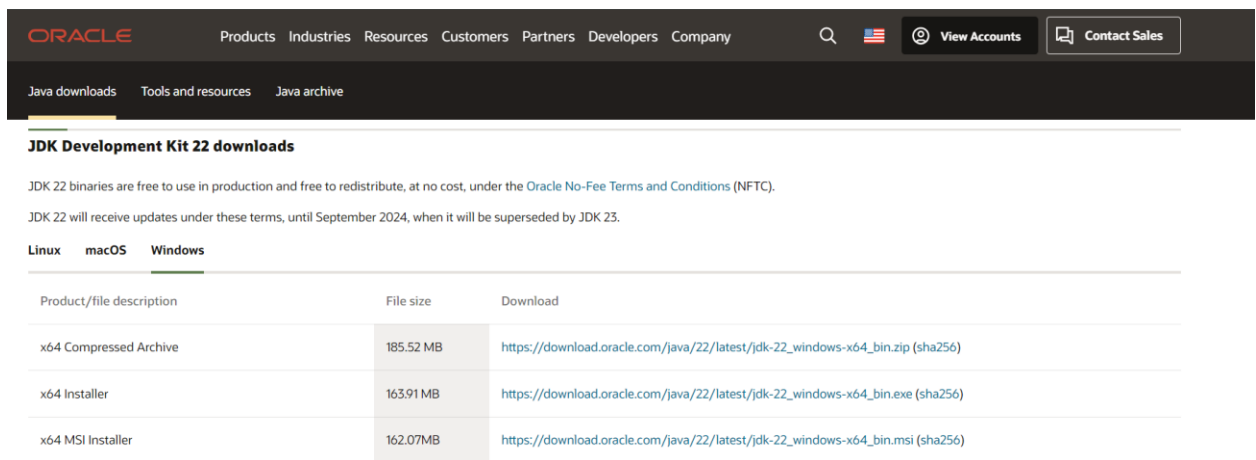
Berikut merupakan langkah penginstalan Hadoop di perangkat komputer berdasarkan yang saya praktekan di laboratorium Sains Data:

Prerequisite (Prasyarat):

- Laptop/PC
- Java Environment
- Apache Hadoop
- Apache Spark

Prosedur penginstalan:

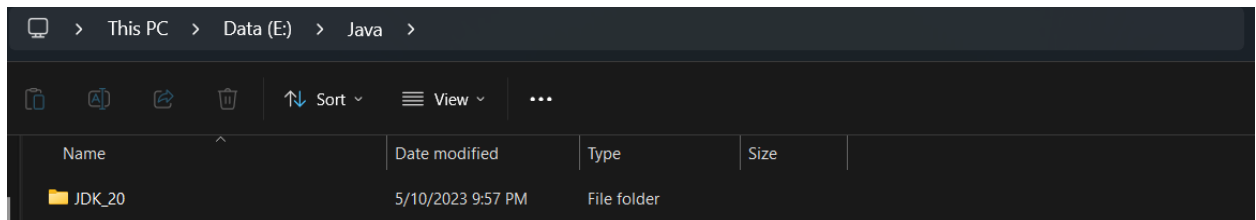
1. Install Java Development Kit (JDK) pada tautan berikut:
<https://www.oracle.com/java/technologies/downloads/#jdk22-windows>



The screenshot shows the Oracle website's 'JDK Development Kit 22 downloads' page. It features a navigation bar with 'ORACLE' and links to Products, Industries, Resources, Customers, Partners, Developers, and Company. Below the navigation bar, there are tabs for 'Java downloads', 'Tools and resources', and 'Java archive'. The main content area is titled 'JDK Development Kit 22 downloads' and includes a disclaimer about the Oracle No-Fee Terms and Conditions (NFTC). A table lists download options for Linux, macOS, and Windows. The Windows section is active, showing three options: 'x64 Compressed Archive' (185.52 MB), 'x64 Installer' (163.91 MB), and 'x64 MSI Installer' (162.07MB). Each option has a corresponding download link.

Product/file description	File size	Download
x64 Compressed Archive	185.52 MB	https://download.oracle.com/java/22/latest/jdk-22_windows-x64_bin.zip (sha256)
x64 Installer	163.91 MB	https://download.oracle.com/java/22/latest/jdk-22_windows-x64_bin.exe (sha256)
x64 MSI Installer	162.07MB	https://download.oracle.com/java/22/latest/jdk-22_windows-x64_bin.msi (sha256)

Dalam tampilan di atas, pilih download file .exe dengan deskripsi “x64 Compressed Archive” atau klik link: https://download.oracle.com/java/22/latest/jdk-22_windows-x64_bin.zip. Setelah melakukan download, lakukan penginstalan dengan mengekstrak file JDK tersebut.



The screenshot shows a Windows File Explorer window. The address bar indicates the path: 'This PC > Data (E:) > Java >'. The main pane displays a folder named 'JDK_20'. The details pane on the right shows the folder's metadata: 'Name: JDK_20', 'Date modified: 5/10/2023 9:57 PM', and 'Type: File folder'.

Kebetulan saya sudah menginstal JDK versi 20 yang saya ekstrak di direktori E:\Java\JDK_20.

2. Install Hadoop pada tautan berikut:

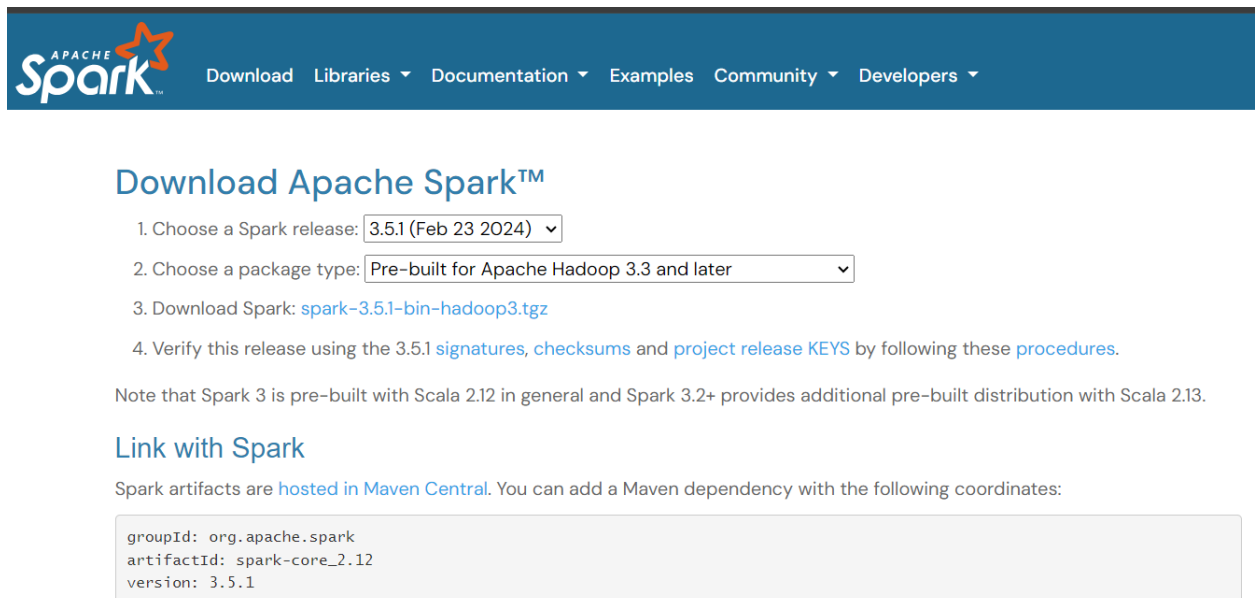
<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-src.tar.gz>



Pada tampilan di atas, klik link “.tar.gz” untuk melakukan download Apache Hadoop atau klik link: <https://d1cdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-src.tar.gz>

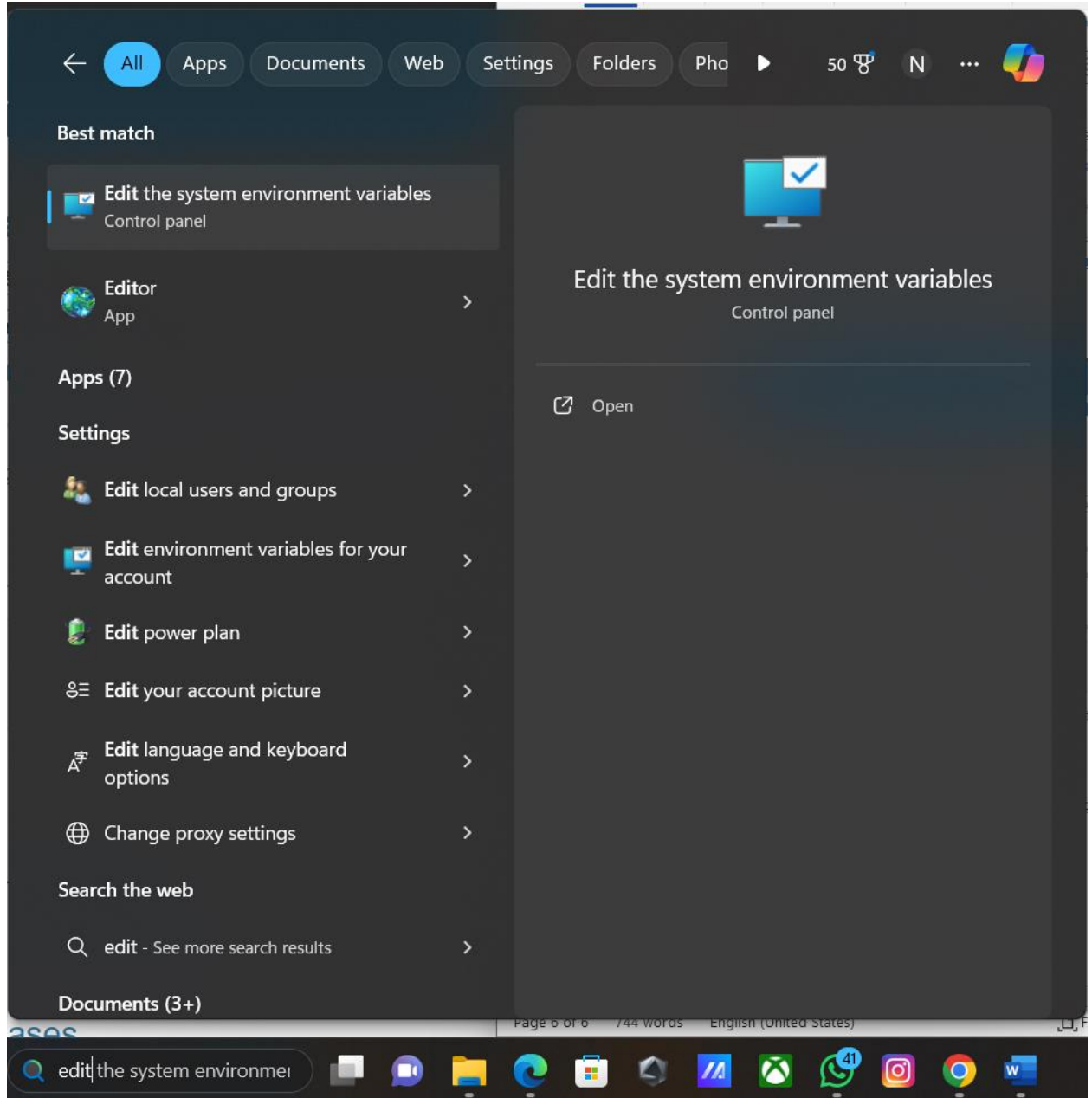
3. Install Apache Spark pada tautan berikut:

<https://spark.apache.org/downloads.html>

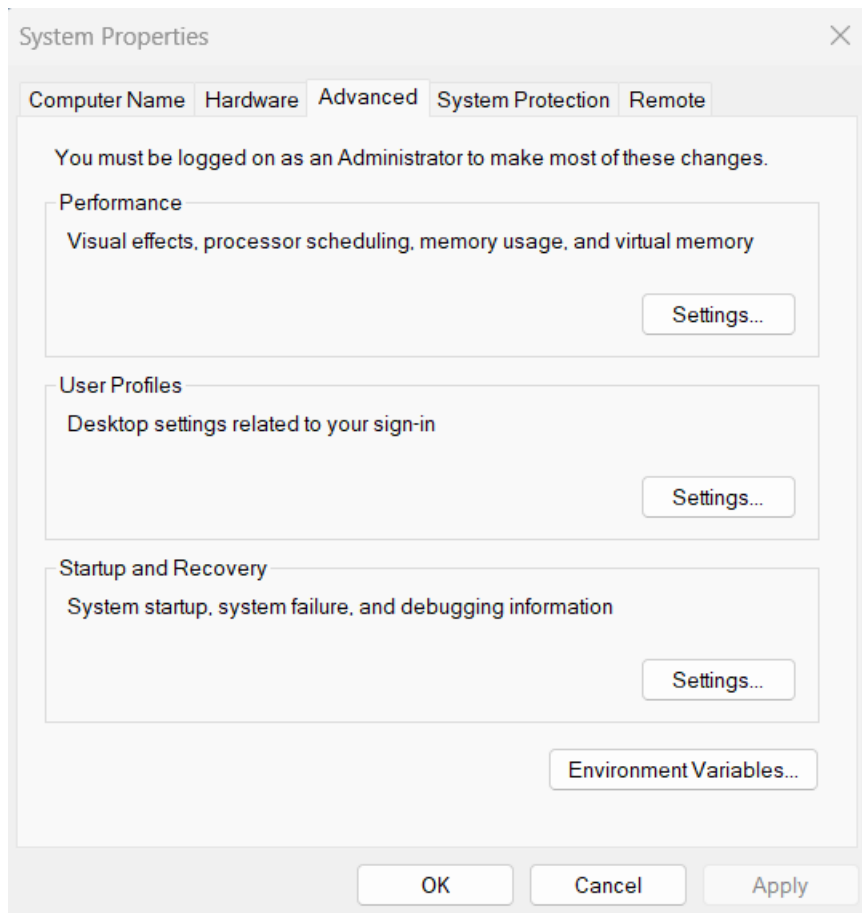


Pada tampilan di atas, pilih versi Spark ingin diinstal pada langkah no 1 di atas. Lalu pilih packages type pada langkah 2 sesuai dengan versi Hadoop yang di install. Klik link yang ada pada langkah 3 di atas.

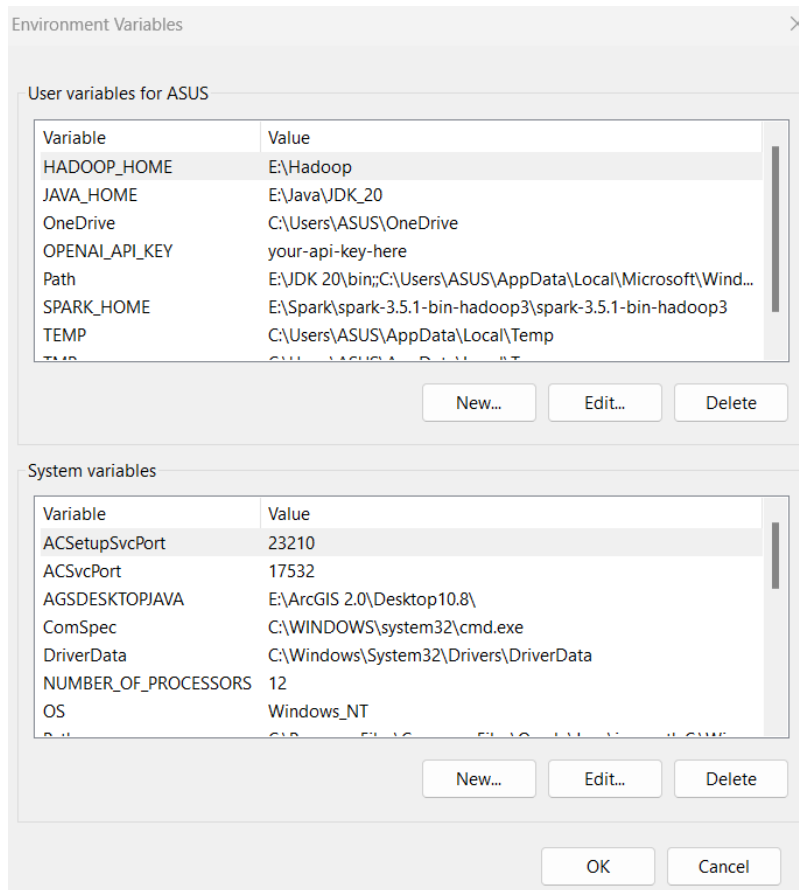
4. Buat environment (lingkungan) untuk Java, Hadoop, dan Spark dengan melakukan search pada dashboard utama dengan kata kunci “edit”



Setelah muncul program di atas, maka klik program “Edit the system environment variables”.



Setelah muncul program di atas, klik “Environment Variables...”.

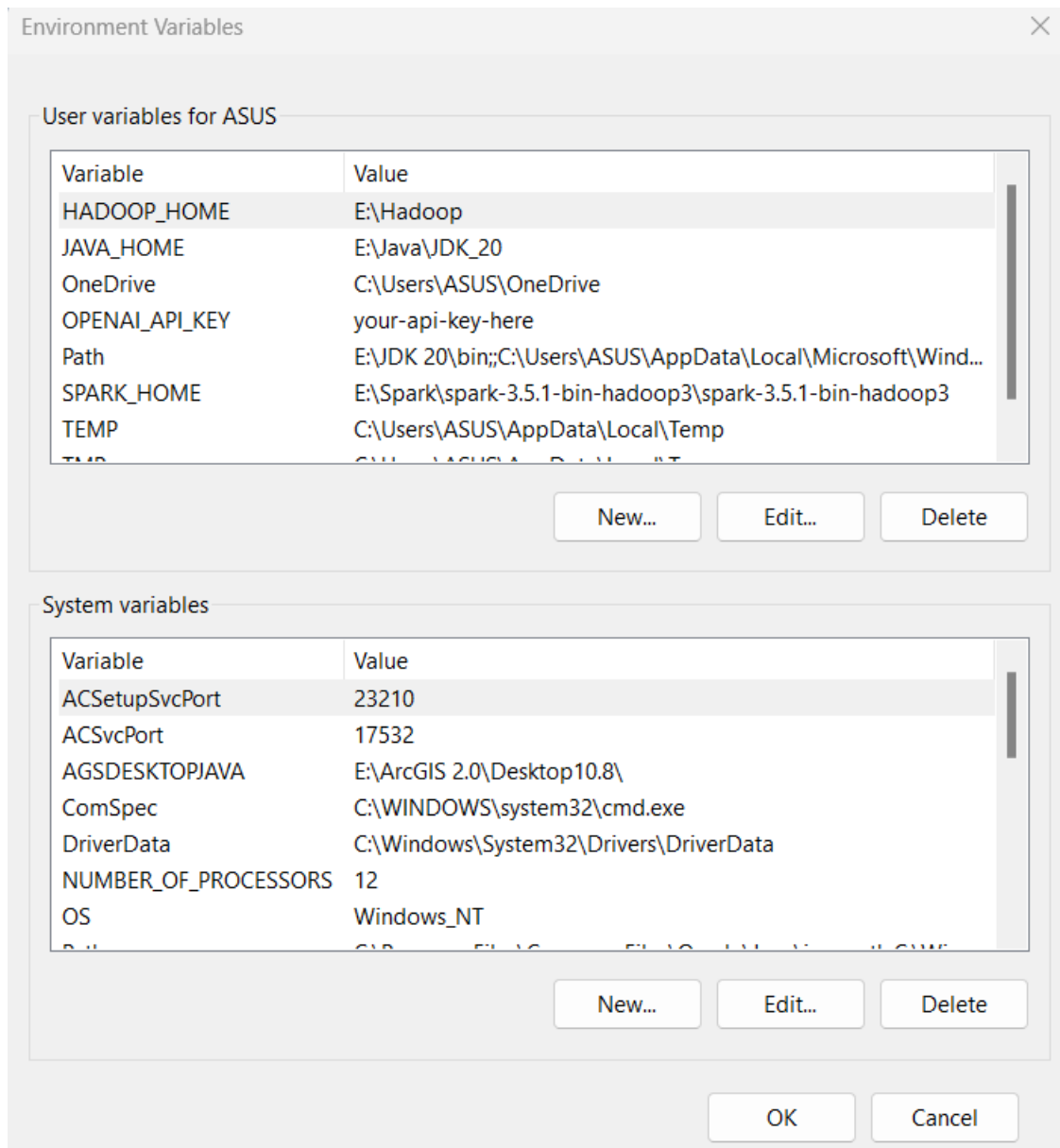


Pada tampilan di atas, pada bagian “User variables for...”, klik “New...”.

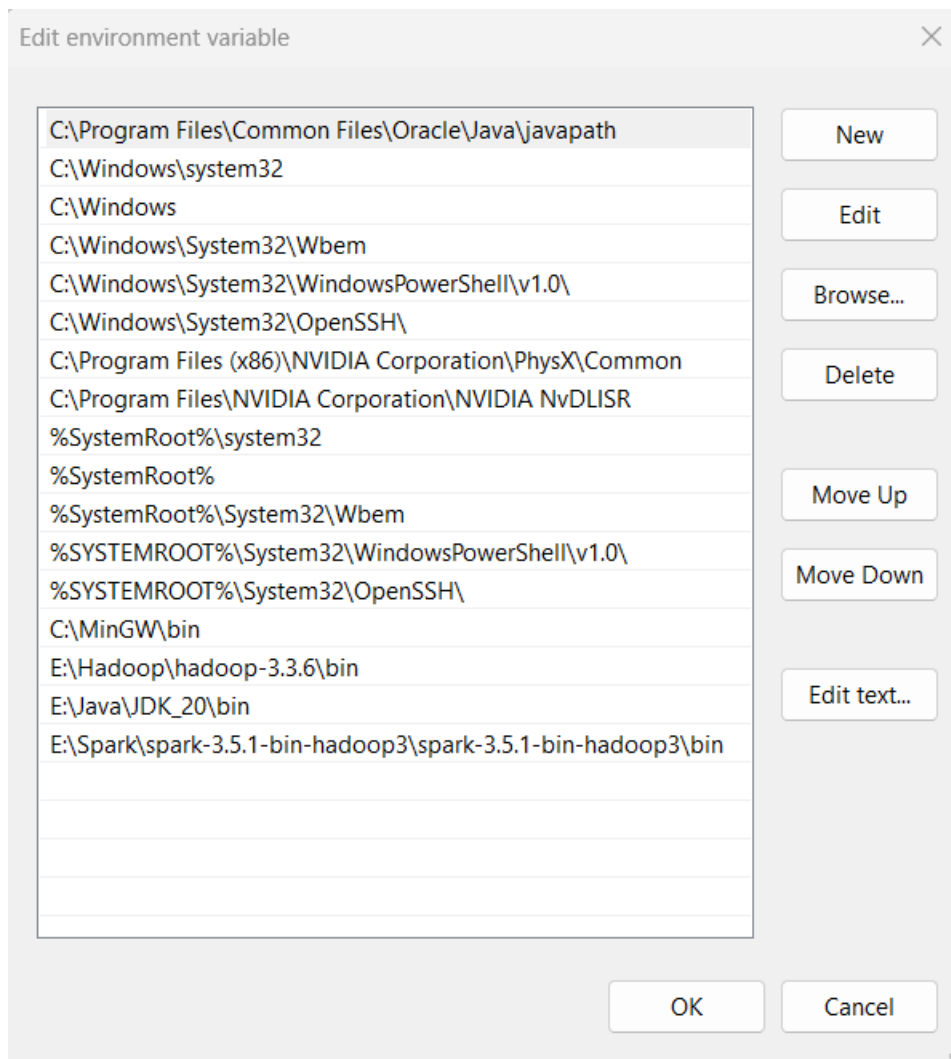


Setelah muncul tampilan di atas, isi variable name dengan “JAVA_HOME” dan isi variable value dengan lokasi ekstrak file JDK yang tadi lalu klik OK.

Setelah itu, buat environment untuk Hadoop dan Spark dengan langkah yang sama yaitu dengan klik “New...” pada bagian “user variables for...” lalu pada bagian variable name diisi “HADOOP_HOME” dan bagian variable value diisi dengan lokasi ekstrak Hadoop yang tadi. Serta setelah itu buat lagi dengan variable name “SPARK_HOME” dan isi variable value dengan lokasi ekstrak Spark yang tadi.

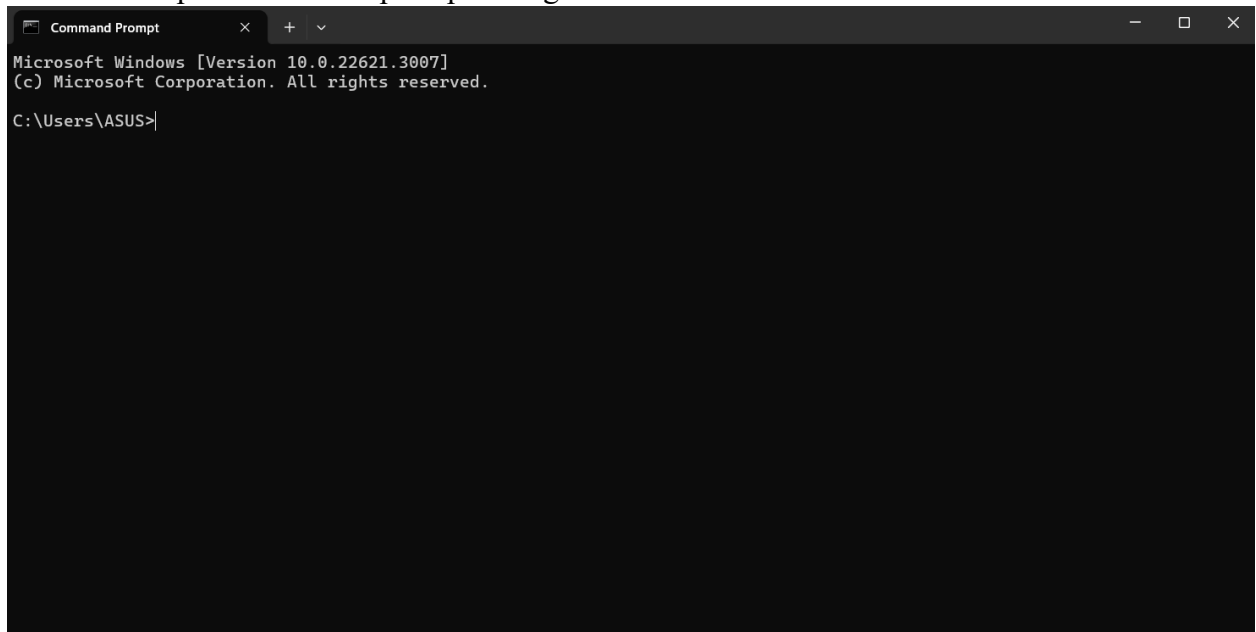


Setelah membuat user variables untuk java, hadoop, dan spark, pada bagian “System variables”, cari variable bernama “path” lalu klik “Edit...”.



Setelah tombol “Edit...” diklik, maka akan muncul tampilan di atas. Selanjutnya Klik “New” lalu klik “Browse...” dan cari lokasi file “bin” dari Java, Hadoop, dan Spark pada masing masing folder yang diekstrak sebelumnya. Setelah itu klik OK dan close semua tab “edit environment variable”.

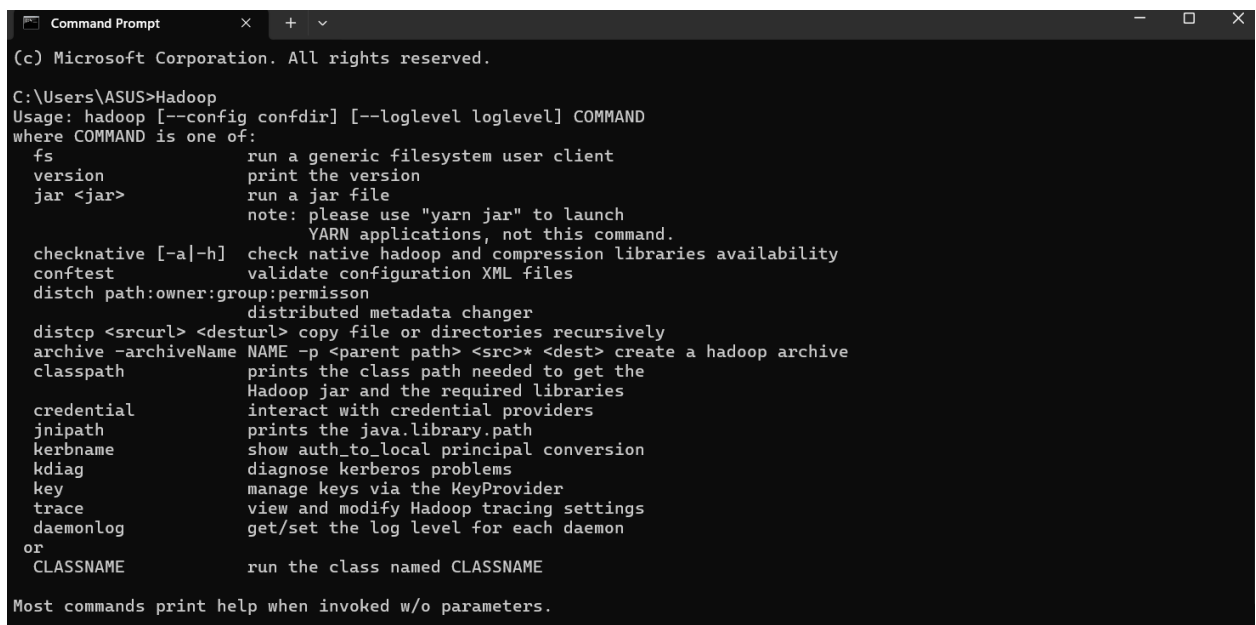
5. Lakukan run pada command prompt di bagian search



```
Command Prompt
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ASUS>
```

Di atas merupakan tampilan dari Command Prompt.



```
Command Prompt
(c) Microsoft Corporation. All rights reserved.

C:\Users\ASUS>Hadoop
Usage: hadoop [--config confdir] [--loglevel loglevel] COMMAND
where COMMAND is one of:
  fs                run a generic filesystem user client
  version           print the version
  jar <jar>         run a jar file
                   note: please use "yarn jar" to launch
                   YARN applications, not this command.
  checknative [-a|-h] check native hadoop and compression libraries availability
  confest          validate configuration XML files
  distch path:owner:group:permission distributed metadata changer
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath        prints the class path needed to get the
                   Hadoop jar and the required libraries
  credential       interact with credential providers
  jnipath          prints the java.library.path
  kerbname         show auth_to_local principal conversion
  kdiag           diagnose kerberos problems
  key             manage keys via the KeyProvider
  trace           view and modify Hadoop tracing settings
  daemonlog       get/set the log level for each daemon
  or
  CLASSNAME       run the class named CLASSNAME

Most commands print help when invoked w/o parameters.
```

Setelah membuka command prompt, ketik “Hadoop” dan tekan enter pada konsol. Maka akan muncul seperti di atas.

Hasil

Berdasarkan metode yang saya lakukan untuk menginstal Hadoop dan Spark, didapatkan hasil yang baik dalam proses penginstalan hingga Hadoop dan Spark dapat siap untuk digunakan.

Diskusi

Pada bagian diskusi ini terdapat beberapa hal yang perlu diperhatikan:

1. Pengguna dapat menentukan apakah menggunakan Hadoop atau Spark untuk melakukan analisis big data karena keduanya memiliki perbedaan sesuai dengan kebutuhan.
2. Dari banyaknya tutorial untuk menginstall hadoop, mayoritas menggunakan JDK 8. Tetapi dapat juga menggunakan JDK versi terbaru untuk pengalaman yang lebih baik.
3. Belum nampak jelas spesifikasi minimum perangkat laptop/pc seperti apa untuk menggunakan Hadoop atau Spark ini.

Kesimpulan

Hadoop adalah kerangka kerja perangkat lunak open source yang diciptakan oleh Doug Cutting. Nama "Hadoop" berasal dari nama boneka gajah kuning milik anak Doug. Hadoop memungkinkan penggunaan sekelompok komputer untuk menganalisis kumpulan data besar secara paralel. Hadoop memiliki beberapa modul, termasuk Common, Avro, MapReduce, HDFS, Pig, Hive, Hbase, ZooKeeper, Sqoop, dan Oozie.

Penginstalan Hadoop dan Spark pada perangkat komputer melibatkan beberapa langkah, termasuk instalasi Java Development Kit (JDK), Hadoop, dan Apache Spark, serta pembuatan lingkungan untuk Java, Hadoop, dan Spark. Pengguna dapat memilih untuk menggunakan Hadoop atau Spark tergantung pada kebutuhan mereka. Meskipun banyak tutorial menyarankan penggunaan JDK 8 untuk Hadoop, JDK versi terbaru juga dapat digunakan. Spesifikasi minimum perangkat untuk menggunakan Hadoop atau Spark belum jelas.

Setelah instalasi, Hadoop dan Spark siap digunakan untuk analisis big data. Selain itu, pengguna juga perlu mempertimbangkan beberapa hal lainnya, seperti memilih antara Hadoop dan Spark berdasarkan kebutuhan mereka, dan bahwa JDK versi terbaru dapat digunakan untuk pengalaman yang lebih baik. Namun, spesifikasi minimum perangkat untuk menggunakan Hadoop atau Spark masih belum jelas.

Secara keseluruhan, Hadoop dan Spark adalah alat yang kuat untuk analisis big data, dan penginstalan mereka pada perangkat komputer melibatkan beberapa langkah yang harus diikuti dengan hati-hati. Meskipun ada beberapa pertimbangan yang harus dipertimbangkan, manfaat yang diperoleh dari penggunaan alat-alat ini untuk analisis big data dapat sangat berharga.

Referensi

White, T. (2012). Hadoop: The Definitive Guide Third Edition. Sebastopol: O'Reilly Media, Inc.