Tutorial - Entorno de Trabajo
Curso BigData
Programa Ciencia de Datos
Fundatec
2010

Alonso Nuñez Sanchez

Servicio/Aplicación: **Entorno de Trabajo / BigData**
Tipo de Documento: **Funcional**

Información de Control del Documento.

Información Documento.

Nombre Documento	Tutorial Entorno de Trabajo Curso BigData
Nombre proyecto	BigData / Ciencia de Datos
Autor	Alonso Nuñez Sanchez
Versión	1.0

Historial de Modificaciones.

Versión	Fecha	Agregados/Modificaciones	Editor
1.0	23/11/2019	Creación Inicial del Documento	Alonso Nuñez Sanchez

Contenido

Información General. 5
Objetivo 5
Limitaciones 5
Aspectos generales importantes. 5

MiniConda y entorno de trabajo. 6
Descripción 6
Descarga 6
Configuración 6
Consideraciones. 9

JDK 10
Descripción 10
Descarga 10
Configuración 12
Consideraciones. 12

Java. 13
Descripción 13
Descarga 13
Configuración 15
Consideraciones. 15

Spark. 16
Descripción 16
Descarga 16
Configuración 17
Consideraciones. 17

FindSpark – buscar recursos de Spark para ser utilizados en Pyton 18
Descripción 18
Descarga 18
Configuración 19
Consideraciones. 19

PySpark - ejecutar Spark en Python 20
Descripción 20
Descarga 20
Configuración 20
Consideraciones. 20

Información General.

Objetivo

Documentar los pasos a seguir para preparar el entorno de trabajo necesario para el curso BigData del programa Ciencia de Datos

Necesidades

Conocer los prerrequisitos de software necesarios para ejecutar las tareas propias del curso
Brindar las herramientas y el conocimiento necesario para la puesta a punto del ambiente

Limitaciones

El presente documento se limita a la preparación e instalación de los componentes de software necesarios para funcionar en el Sistema Operativo Windows, específicamente Windows 10

Aspectos generales importantes.

El alcance del documento es únicamente para cubrir los aspectos académicos del curso en cuestión.

MiniConda y entorno de trabajo.

Descripción

Conda es un sistema open source para la gestión de entornos de trabajo.
Aunque fue creado para programas en Python, permite la ejecución de múltiples lenguajes de programación.
MiniConda es una instalación reducida de Conda. Para el alcance del curso, los recursos obtenidos con miniconda son suficientes

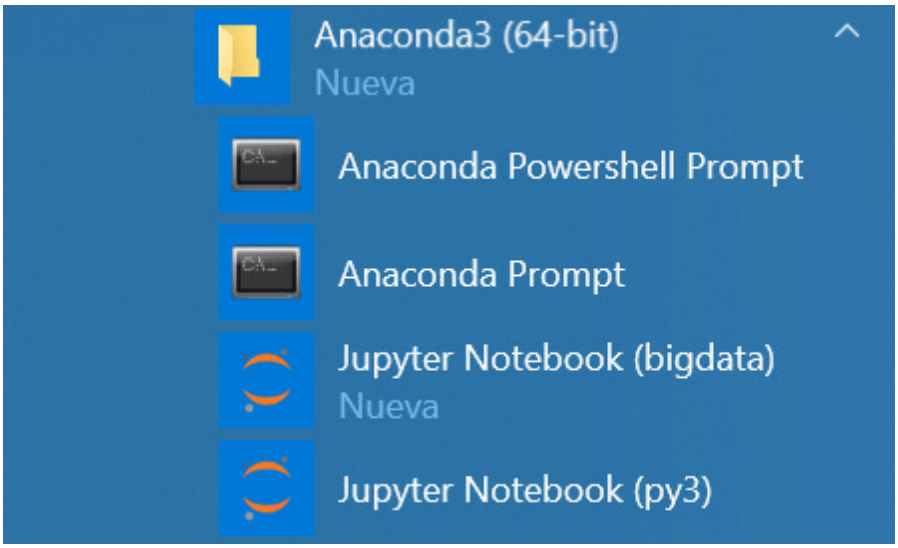
Descarga

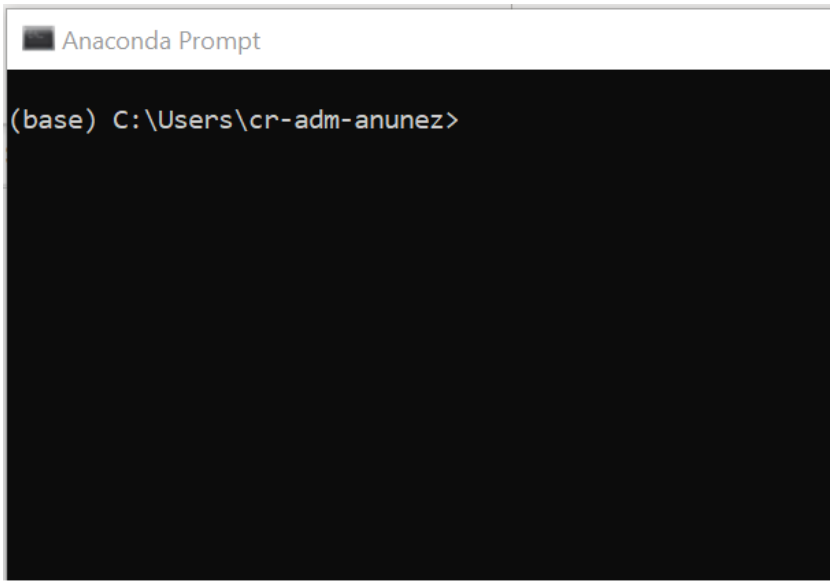
https://conda.io/miniconda.html

Instalación

Configuración

Una vez instalado, lo que tendremos es una consola que se puede encontrar en el explorador de aplicaciones, como “Anaconda Prompt”, dentro de la carpeta “Anaconda3”





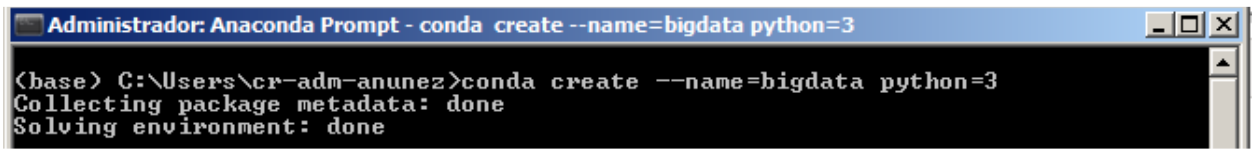
Desde esta consola se ejecutan todos los comandos necesarios para la instalación y ejecución de paquetes, librerías y entornos de trabajo. Para efectos del presente documento, nos limitaremos a los comandos específicos necesarios

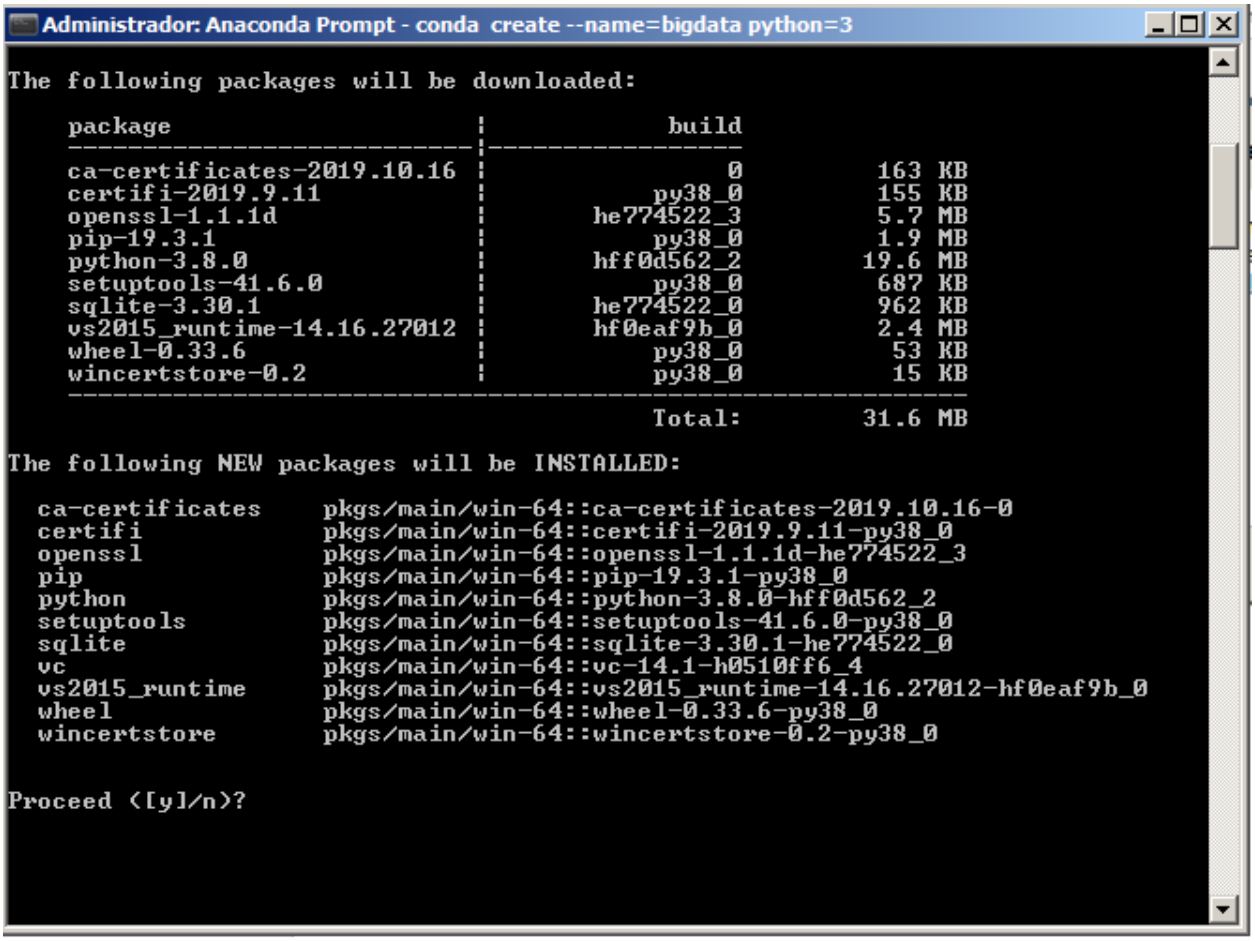
Lo primero será crear el ambiente de trabajo a utilizar, para esto se utiliza el comando ***create***, de la siguiente forma:
>> conda create --name bigdata python=3
En este caso creamos un ambiente de trabajo de tipo python3 llamado "bigdata"

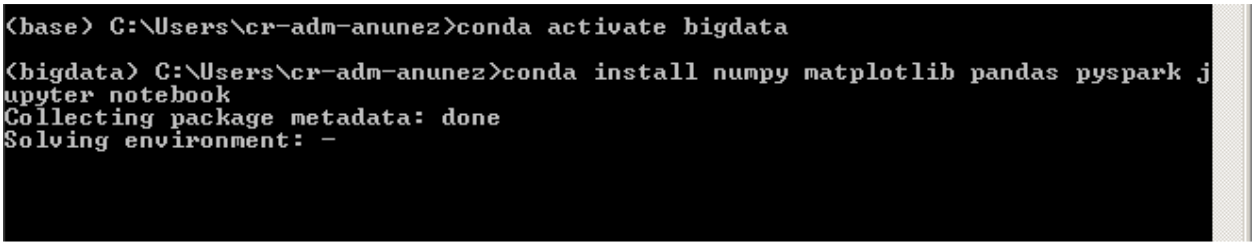
Posteriormente, se activa el ambiente recién creado, con el comando ***activate***
>> conda activate bigdata

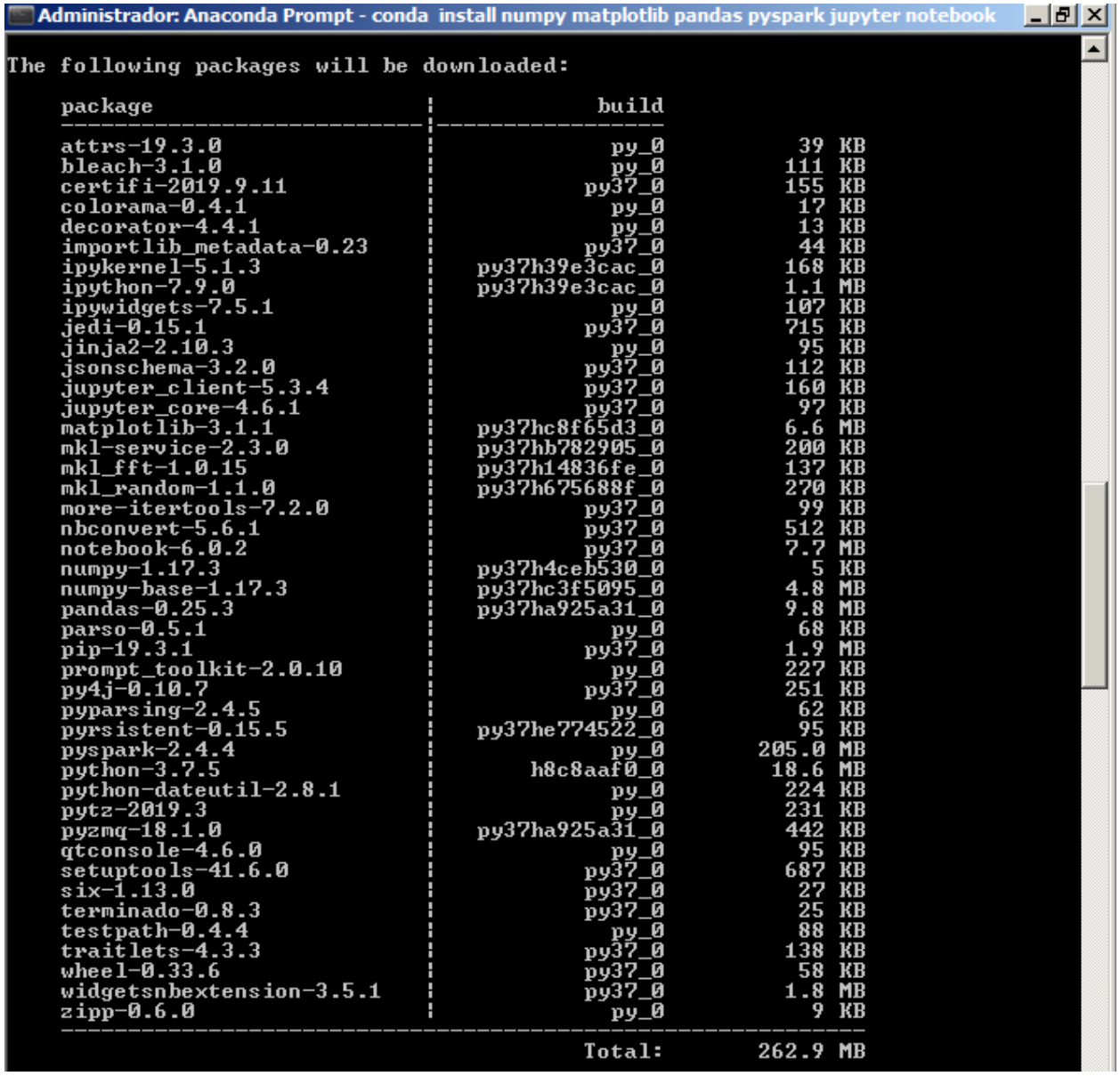
Activado el ambiente, el siguiente paso es instalar las librerías necesarias, comando ***install***
>> conda install numpy
>> conda install matplotlib
>> conda install pandas
>> conda install jupyter notebook
>> conda install pyspark
>> conda install findspark

Dependiendo de la versión de conda, y los canales/repositorios configurados, pueda ser necesario un paso adicional para lograr instalar findspark, ver detalles en la sección "FinSpark" del presente documento




Administrador: Anaconda Prompt - conda create --name=bigdata python=3
The following packages will be downloaded:
package | build
ca-certificates-2019.10.16 | 0 163 KB
certifi-2019.9.11 | py38_0 155 KB
openssl-1.1.1d | he774522_3 5.7 MB
pip-19.3.1 | py38_0 1.9 MB
python-3.8.0 | hff0d562_2 19.6 MB
setuptools-41.6.0 | py38_0 687 KB
sqlite-3.30.1 | he774522_0 962 KB
vs2015_runtime-14.16.27012 | hf0eaf9b_0 2.4 MB
wheel-0.33.6 | py38_0 53 KB
wincertstore-0.2 | py38_0 15 KB
Total: 31.6 MB
The following NEW packages will be INSTALLED:
ca-certificates pkgs/main/win-64::ca-certificates-2019.10.16-0
certifi pkgs/main/win-64::certifi-2019.9.11-py38_0
openssl pkgs/main/win-64::openssl-1.1.1d-he774522_3
pip pkgs/main/win-64::pip-19.3.1-py38_0
python pkgs/main/win-64::python-3.8.0-hff0d562_2
setuptools pkgs/main/win-64::setuptools-41.6.0-py38_0
sqlite pkgs/main/win-64::sqlite-3.30.1-he774522_0
vc pkgs/main/win-64::vc-14.1-h0510ff6_4
vs2015_runtime pkgs/main/win-64::vs2015_runtime-14.16.27012-hf0eaf9b_0
wheel pkgs/main/win-64::wheel-0.33.6-py38_0
wincertstore pkgs/main/win-64::wincertstore-0.2-py38_0
Proceed ([y]/n)?


(base) C:\Users\cr-adm-anunez>conda activate bigdata
(bigdata) C:\Users\cr-adm-anunez>conda install numpy matplotlib pandas pyspark j
upyter notebook
Collecting package metadata: done
Solving environment: -



Consideraciones.

JDK

Descripción

JDK (Java Development Kit) son un conjunto de herramientas que permiten el desarrollo de programas en el lenguaje de programación Java, provee diversas librerías para tal efecto.

Descarga

https://www.oracle.com/technetwork/java/javase/downloads/jdk13-downloads-5672538.html

Java SE Development Kit 13.0.1

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

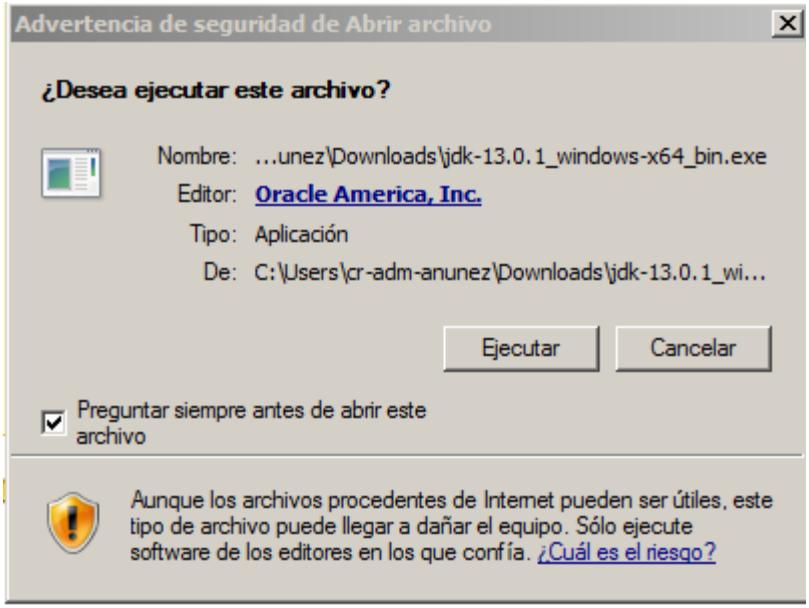
Accept License Agreement **Decline License Agreement**

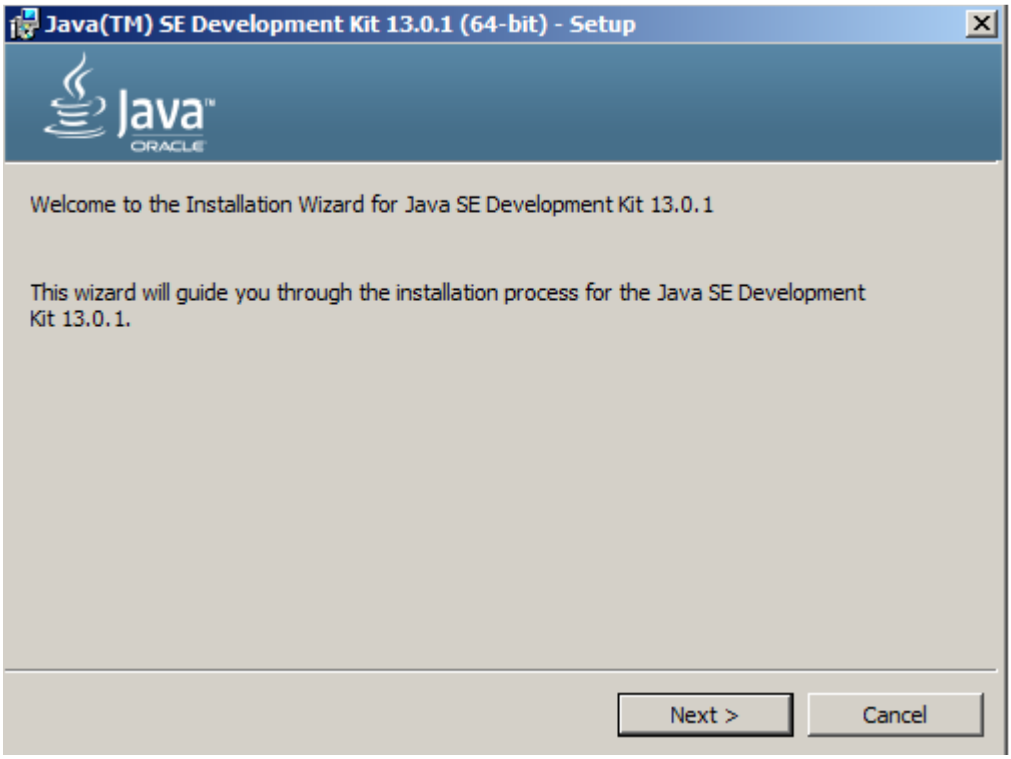
Product / File Description	File Size	Download
Linux	155.88 MB	jdk-13.0.1_linux-x64_bin.deb
Linux	163.17 MB	jdk-13.0.1_linux-x64_bin.rpm
Linux	180 MB	jdk-13.0.1_linux-x64_bin.tar.gz
macOS	172.78 MB	jdk-13.0.1_osx-x64_bin.dmg
macOS	173.11 MB	jdk-13.0.1_osx-x64_bin.tar.gz
Windows	159.84 MB	jdk-13.0.1_windows-x64_bin.exe
Windows	178.99 MB	jdk-13.0.1_windows-x64_bin.zip

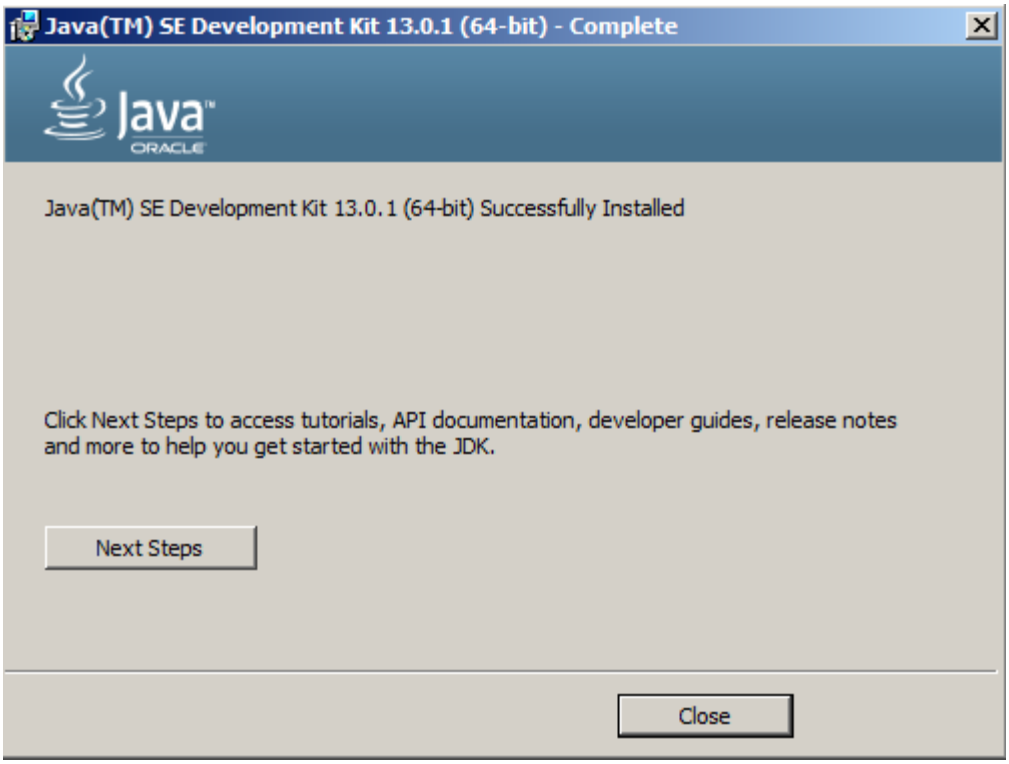
1. Ingresar al link anterior
2. Marcar “Accept Licence Agreement”
3. Seleccionar el Sistema Operativo y arquitectura, en este caso “Windows x64”, **jdk-13.0.1_windows-x64_bin.exe**

Instalación

Ejecutar el instalador y completar el wizard sin cambiar configuraciones







Configuración

No requerida

Consideraciones.

Java.

Descripción

Descarga

https://www.java.com/es/download/win10.jsp

Recursos de ayuda

» ¿Qué es Java?
» Eliminar versiones anteriores de Java
» Desactivar Java
» Mensajes de error
» Solucionar problemas de Java
» Otra ayuda

Usuarios de Windows de 64 bits

¿Utiliza exploradores de 32 y 64 bits?
» Preguntas frecuentes sobre Java de 64 bits para Windows

Instalación fuera de línea

¿Problemas al descargar? Intente con el installer fuera de línea.

Descargar Java para Windows

Recomendado Version 8 Update 231 (Tamaño de archivo: 1.97 MB)
Fecha de versión: 15 de octubre de 2019

Actualización importante de la licencia de Oracle Java

La licencia de Oracle Java ha cambiado para las versiones publicadas a partir del 16 de abril de 2019.

El nuevo acuerdo de licencia de Oracle Technology Network para Oracle Java SE es sustancialmente diferente a las licencias de Oracle Java anteriores. La nueva licencia permite ciertos usos, como el uso personal y de desarrollo, sin coste alguno (aunque podría haber otros usos autorizados en licencias de Oracle Java anteriores que ya no estén disponibles). Revise las condiciones con atención antes de descargar y utilizar este producto. Puede consultar las preguntas frecuentes aquí.

La licencia comercial y el soporte están disponibles con una suscripción de Java SE de bajo coste.

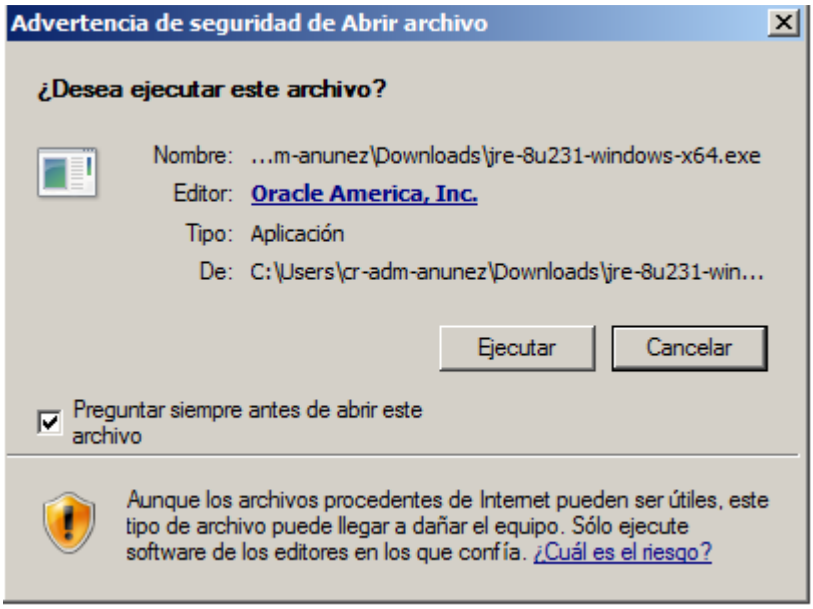
Oracle también ofrece la última versión de OpenJDK con la licencia pública general de código abierto en jdk.java.net.

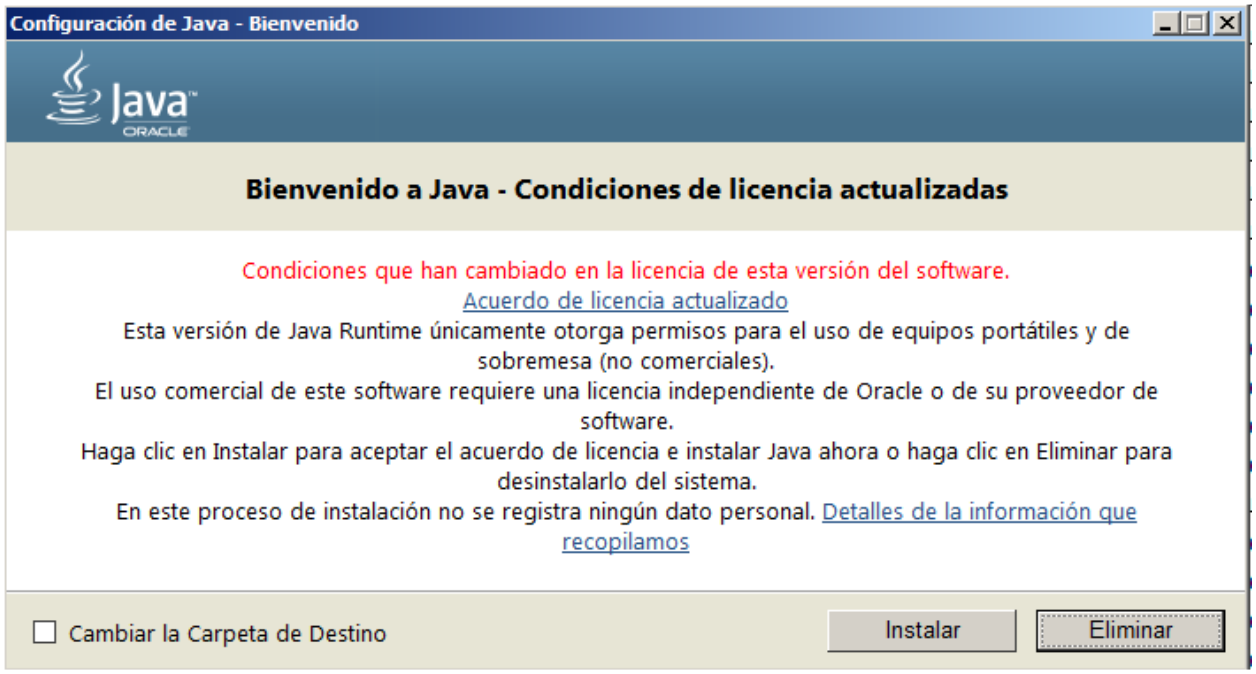
Aceptar e iniciar descarga gratuita

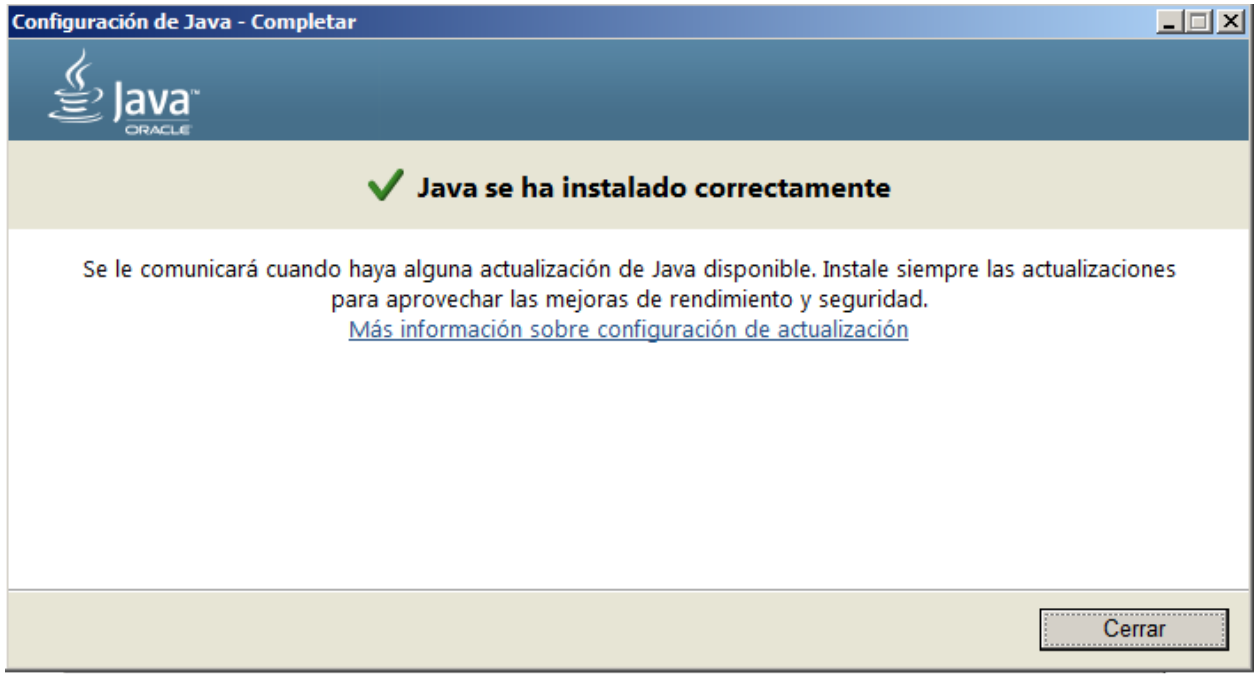
Al descargar Java, confirma que ha leído y acepta las condiciones del acuerdo de licencia de Oracle Technology Network para Oracle Java SE

Instalación

Ejecutar el instalador y completar el wizard sin cambiar configuraciones







Configuración

No requerida

Consideraciones.

Spark.

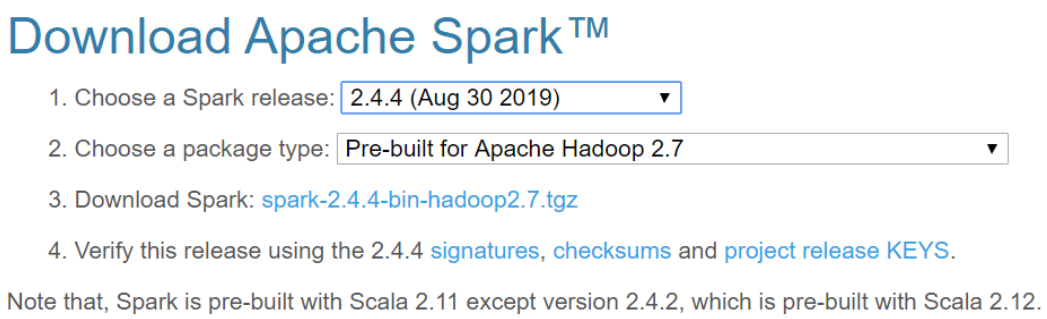
Descripción

Es un framework enfocado en el manejo de grandes volúmenes de datos y la ejecución de cómputo intensivo sobre estos datos.
Es la evolución de lo que fue Hadoop, entre sus ventajas respecto a Hadoop está la reducción en tiempos de ejecución

Descarga

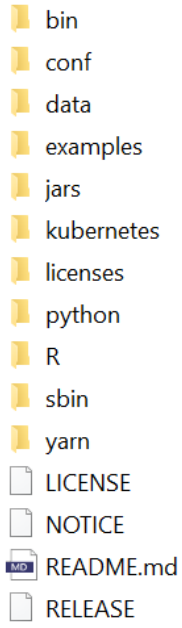
https://spark.apache.org/downloads.html

Usaremos la versión 2.4.4, paquete “Pre-built for Apache Hadoop 2.7”



Instalación

No requiere instalación, solamente copiar el contenido de la carpeta descargada en una ubicación a elegir, en este caso será c:\spark\



Configuración

No requiere ninguna configuración, simplemente la copia del contenido descargado y el directorio elegido

Consideraciones.

Las pruebas realizadas y primeros códigos ejecutados funcionan correctamente con la versión indicada. Esto no elimina la posibilidad de probar con la versión más reciente

FindSpark – buscar recursos de Spark para ser utilizados en Pyton

Descripción

FindSpark es una librería de Python para "buscar" y utilizar los recursos de Spark
Básicamente se usa para indicarle a Python la ruta donde se copiaron los archivos de Spark (ver sección "Spark" del presente documento

Descarga

Se descarga desde conda cuando se ejecuta el comando install

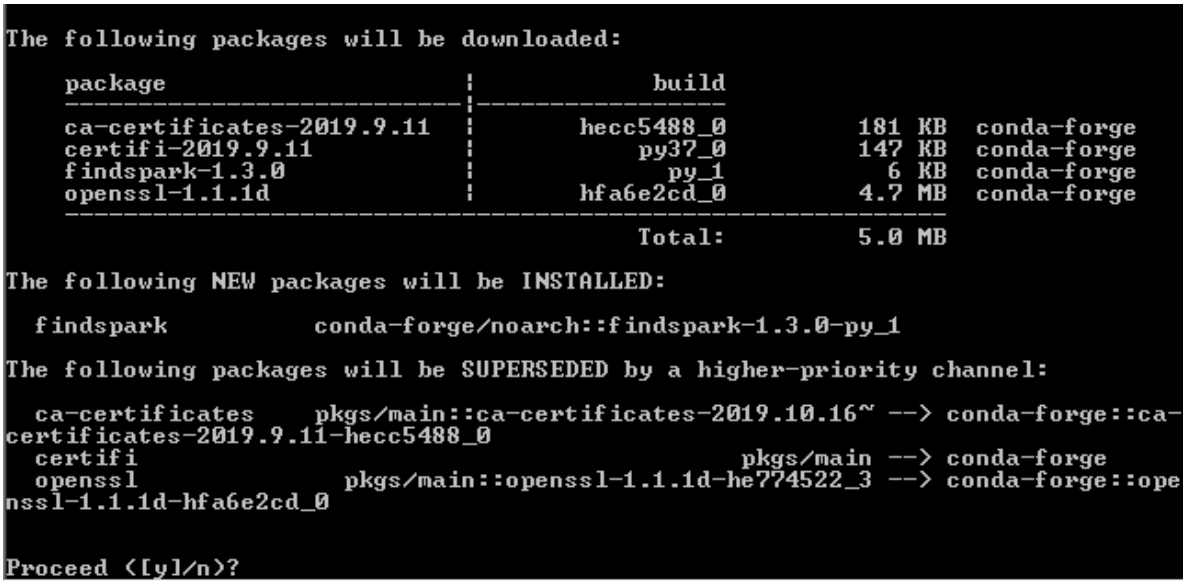
Instalación

Se instala desde conda, en el ambiente de trabajo activo, con el comando install
>> conda install findspark

Dependiendo de la versión de conda, o de los canales/repositorios disponibles, pueda ser necesario indicar un canal distinto, esto se logra con el comando ***config –add channels*** indicando el canal deseado, en este caso será conda-forge, y volviendo a ejecutar el comando ***install***
>> conda config –add channels conda-forge

```
(bigdata) C:\Users\cr-adm-anunez>conda install findspark
Collecting package metadata: done
Solving environment: failed

PackagesNotFoundError: The following packages are not available from current cha
nnels:

  - findspark

Current channels:

  - https://repo.anaconda.com/pkgs/main/win-64
  - https://repo.anaconda.com/pkgs/main/noarch
  - https://repo.anaconda.com/pkgs/free/win-64
  - https://repo.anaconda.com/pkgs/free/noarch
  - https://repo.anaconda.com/pkgs/r/win-64
  - https://repo.anaconda.com/pkgs/r/noarch
  - https://repo.anaconda.com/pkgs/msys2/win-64
  - https://repo.anaconda.com/pkgs/msys2/noarch

To search for alternate channels that may provide the conda package you're
looking for, navigate to

    https://anaconda.org

and use the search bar at the top of the page.
```

```
(bigdata) C:\Users\cr-adm-anunez>conda config --add channels conda-forge

(bigdata) C:\Users\cr-adm-anunez>conda install findspark
Collecting package metadata: done
Solving environment: done
```



Configuración

No requerida

Consideraciones.

PySpark - ejecutar Spark en Python

Descripción

PySpark es un API de Python para poder ejecutar Spark

Descarga

Se descarga desde conda cuando se ejecuta el comando install

Instalación

Se instala desde conda, en el ambiente de trabajo activo, con el comando install
>> conda install pyspark

Configuración

No requerida

Consideraciones.