# Instalacion de Apache Spark

**Instalar Java**

Ingresar a la web de java e instalar la última version estable recomendable

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

**Instalar Apache Spark**

Ingresar a la web https://spark.apache.org/downloads.html

Seleccionar el relees: 3.3.2 (Feb 17 2023)

Seleccionar el paquete:

* Pre-built for Apache Hadoop 3.3 and later
* Download Spark: spark-3.3.2-bin-hadoop3.tgz

**Configuración de archivos Spark**

Descomprimir el archivo descargado en un directorio seguro:

En el directorio conf del Spark realizar los siguientes pasos

Hacer una copia del archivo **log4j2.properties.template** con el nombre **log4j2.properties**

Abrir el archivo y cambiar **info** por **warn** en la línea #19

Adicionar al final del archivo las siguientes lineas:

log4j2.logger.org.apache.spark.util.ShutdownHookManager=OFF

log4j2.logger.org.apache.spark.SparkEnv=ERROR

Hacer una copia del archivo **spark-defaults.conf.template** con el nombre **spark-defaults.conf**

Adicionar al final las siguinetes lineas:

spark.executor.processTreeMetrics.enabled false

spark.jars.packages org.apache.spark:spark-sql-kafka-0-10\_2.12:3.3.1,org.apache.spark:spark-avro\_2.12:3.3.1

**hadoop-3.3.0**

Ingresar a la web https://github.com/kontext-tech/winutils

Descargar los archivos de winutils en forma de zip

Descomprimir el archivo zip descargado y seleccionar el directorio hadoop-3.3.0, colocarlo en una zona segura.

Ingresar al **Edit the System Environment** de Windows

En las variables de ambiente declarar las siguientes:

* SPARK\_HOME: como valor colocar el directorio Spark con toda su ruta completa
* HADOOP\_HOME: como valor colocar el directorio Hadoop con toda su ruta completa

En la variable **path** ingresar los siguientes valores:

* %SPARK\_HOME%\bin
* %HADOOP\_HOME%\bin

**Apache Kafka**

Descargar desde la web https://kafka.apache.org/

Version 3.3.1 para Scala 2.12

Descomprimir el archivo y colocarlo en un directorio seguro

En el directorio config:

En el archivo **zookeeper.properties** actualizar la línea 17 con el siguiente texto:

dataDir=%KAFKA\_HOME%\data\zookeeper

En el archivo **server.properties** descomentar la línea 34

#listeners=PLAINTEXT://:9092

actualizar la línea 62 con el siguiente texto:

log.dirs=%KAFKA\_HOME%\data\kafka\kafka-logs

Entorno de variables de ambiente configurar la variable KAFKA\_HOME con la ruta donde tiene los archivos del Kafka

En el path adicionar el %KAFKA\_HOME%\bin\windows

**Activar Apache Kafka**

Ejecutar cmd en modo administrador

zookeeper-server-start.bat %KAFKA\_HOME%/config/zookeeper.properties

Ejecutar cmd en modo administrador

kafka-server-start.bat %KAFKA\_HOME%/config/server.properties

crear topic

kafka-topics.bat --create --topic quickstart-events --bootstrap-server localhost:9092

crear consumer

kafka-console-consumer.bat --topic quickstart-events --from-beginning --bootstrap-server localhost:9092

ingresar data

kafka-console-producer.bat --topic quickstart-events --bootstrap-server localhost:9092

**Installacion de Pyspark**

Ingresar a la URL <https://www.python.org/downloads/>

Descargar el Python correspondiente e instalarlo

Ejecutar en el cmd la sentencia: pip install pyspark

**Configuración de Jupyter**

Ejecutar en el cmd la sentencia: **pip install notebook**

Declarar las siguientes variables de entorno

* PYSPARK\_DRIVER\_PYTHON: jupyter
* PYSPARK\_DRIVER\_PYTHON\_OPTS: notebook --no-browser --port=8889 --notebook-dir=’D:/bigdata/jupyter’
* PYSPARK\_PYTHON=python

Ejecutar en una consola de comandos la sentencia **pyspark** o **python -m notebook**