Taller introductorio a Spark

Contenido General

El lenguaje a utilizar es Python 3.x, principalmente en ambientes Windows y Linux (y puede aplicarse a macOS).

- Visión general de Spark
- Ambiente de trabajo e instalación de software
- Modelo de ejecución
- Spark SQL
- MLib
- Aplicación en clasificación de Texto

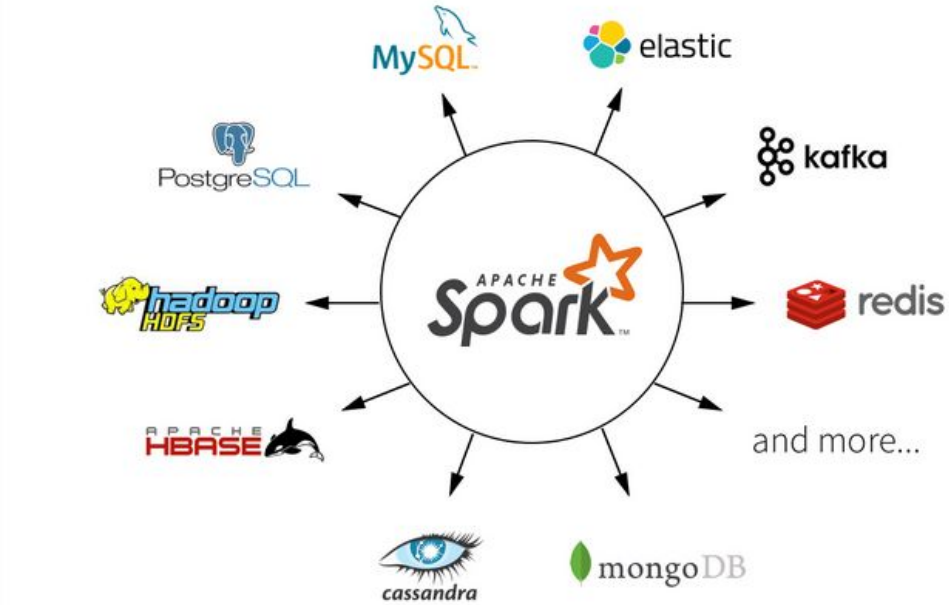
¿Que es Spark?

Apache Spark

Apache Spark se puede considerar un sistema de computación en clúster de propósito general y orientado a la velocidad. Es ampliamente considerado como el sucesor de MapReduce para el procesamiento de datos en clústeres Apache Hadoop.

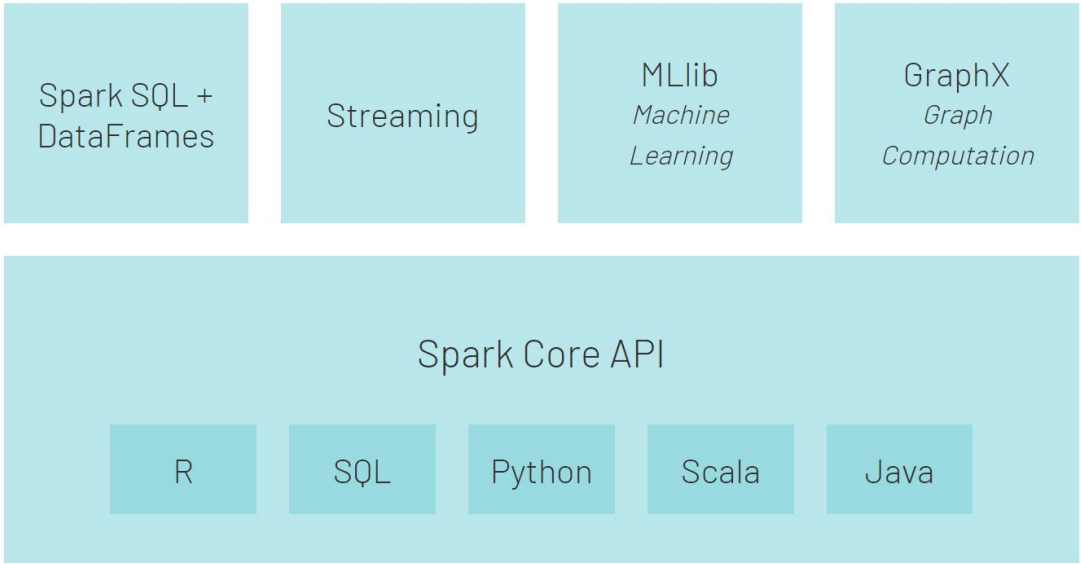
Apache Spark Ecosystem

Spark puede correr en modo cluster standalone, EC2, Hadoop YARN, Mesos, o Kubernetes. Puede acceder datos HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, entre otros.



Apache Spark Ecosystem

Proporciona APIs de alto nivel en Java, Scala, Python y R, y un motor optimizado que admite grafos de ejecución general. Es compatible con un amplio conjunto de herramientas de alto nivel que incluyen Spark SQL para SQL y procesamiento de datos estructurados, MLlib para aprendizaje automático, GraphX para procesamiento de grafos y Spark Streaming.



Spark SQL

Spark SQL permite consultar datos estructurados utilizando SQL o la API DataFrame. Utilizable en Java, Scala, Python y R.

Spark SQL puede usar Hive metastores, SerDes (Serializador/Deserializador), y UDFs (Funciones definidas por el usuario). Así mismo se pueden integrar herramientas existentes de BI (business intelligence)

Spark Streaming

Muchas aplicaciones necesitan la capacidad de procesar y analizar no solo datos por lotes, sino también flujos de datos nuevos en tiempo real. Spark Streaming permite potentes aplicaciones analíticas interactivas tanto en stream como en datos históricos. Se integra fácilmente con una amplia variedad de fuentes de datos populares, incluyendo HDFS, Flume, Kafka y Twitter.

Spark Streaming

Muchas aplicaciones necesitan la capacidad de procesar y analizar no solo datos por lotes, sino también flujos de datos nuevos en tiempo real. Spark Streaming permite potentes aplicaciones analíticas interactivas tanto en stream como en datos históricos. Se integra fácilmente con una amplia variedad de fuentes de datos populares, incluyendo HDFS, Flume, Kafka y Twitter.

Spark Machine Learning

MLlib es la biblioteca de aprendizaje automático (ML) de Spark. Su objetivo es hacer que el aprendizaje automático práctico sea escalable y fácil. En un nivel alto, proporciona herramientas como:

- ML Algorithms: algoritmos de aprendizaje comunes como clasificación, regresión, agrupamiento y filtrado colaborativo
- Featurization: extracción de características, transformación, reducción de dimensionalidad y selección
- Pipes: herramientas para construir, evaluar y ajustar pipes de ML
- Persistence: guardar y cargar algoritmos, modelos y pipes
- Utilities: álgebra lineal, estadísticas, manejo de datos, etc.

Spark GraphX

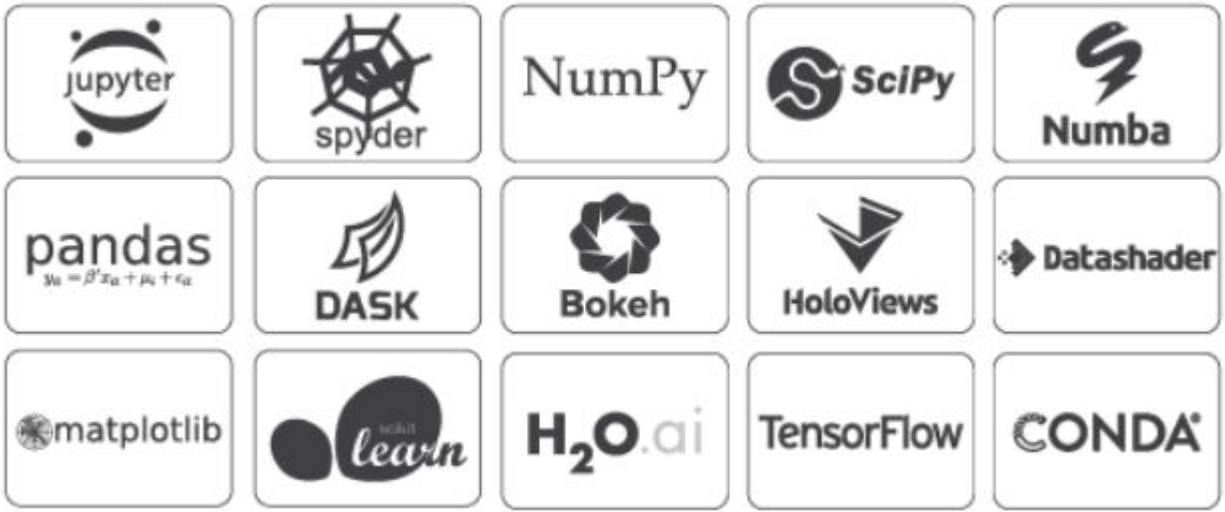
GraphX es un motor de cálculo de grafos que permite construir, transformar y analizar interactivamente sobre datos estructurados. Viene con una biblioteca completa de algoritmos comunes entre los que se encuentran:

- PageRank
- Connected components
- Label propagation
- SVD++
- Strongly connected components
- Triangle count

Ambiente de Trabajo

Anaconda Distribution

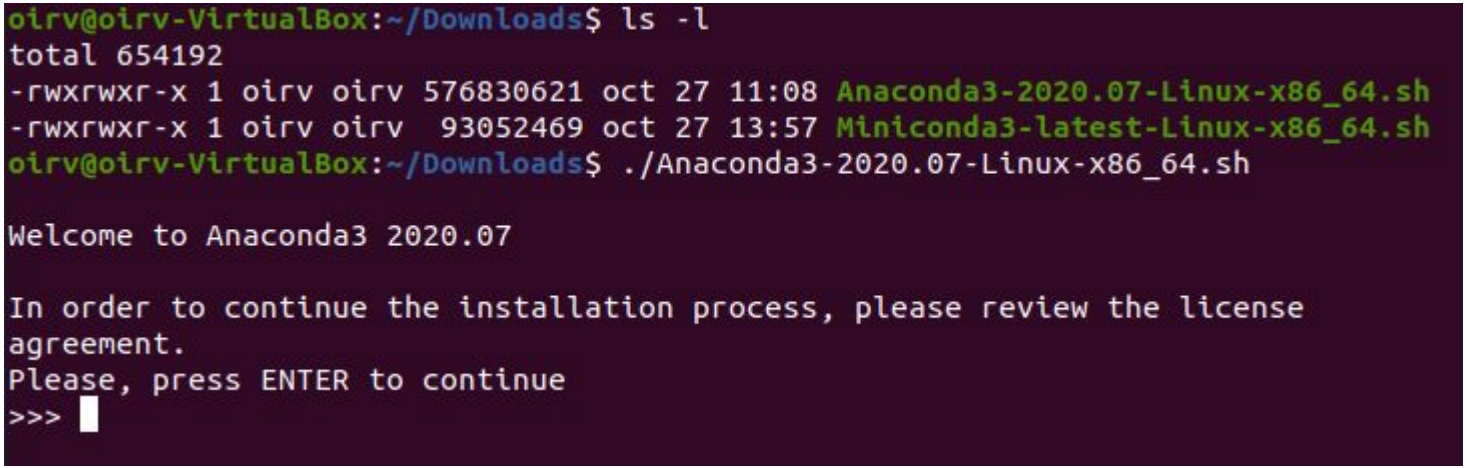
La distribución de código abierto Anaconda es la forma más fácil de realizar ciencia de datos Python / R y aprendizaje automático en Linux, Windows y Mac OS X.



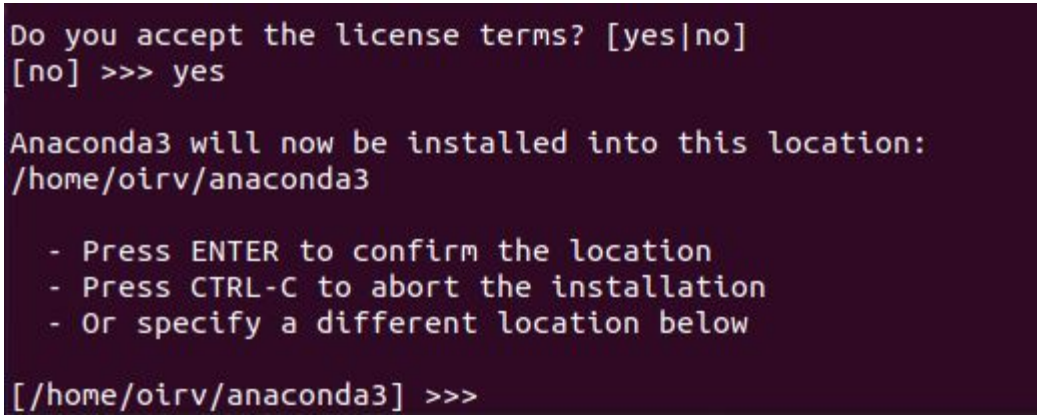
Instalación

Instalación

bash Anaconda3-2019.07-Linux-x86_64.sh



Instalación



Instalación

```
installation finished.

Do you wish the installer to prepend the Anaconda3 install
location

to PATH in your /home/user/.bashrc ? [yes|no]
```

Si se elige “no”, se puede activar anaconda en linux de la siguiente manera:

```
source /home/user/anaconda3/bin/activate
```

Instalación

Use el comando ***conda*** para probar la instalación:

```
conda info
```

Para actualizar

```
conda update conda
```

Crear y activar entornos de anaconda

Cree un entorno de Python 3 llamado ***spark*** haciendo lo siguiente:

```
conda create --name spark python=3.8
```

Activar el entorno:

```
conda activate spark
```

Instalar anaconda el entorno:

```
Conda install anaconda
```

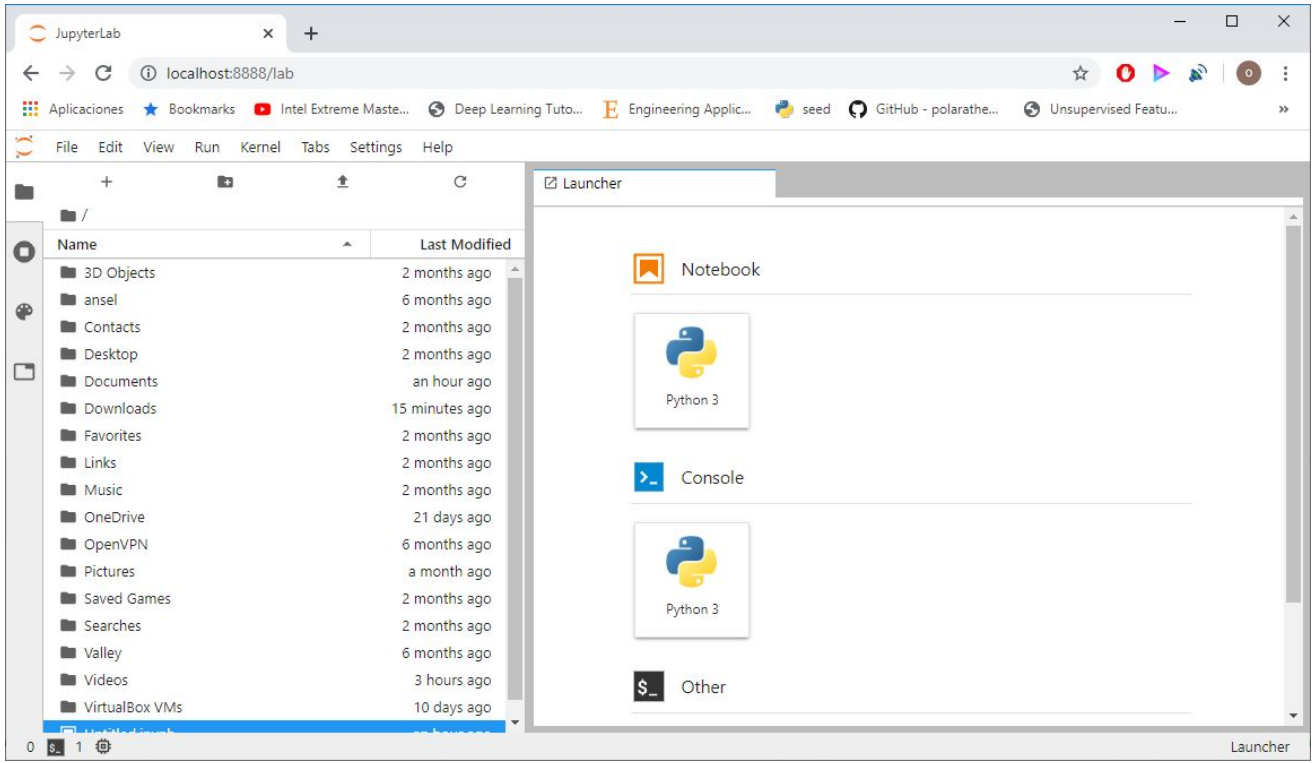
JupyterLab

JupyterLab es un entorno de desarrollo interactivo basado en web para notebooks de Jupyter, código y datos.

JupyterLab

Ejecutar en consola:

```
jupyter lab
```



Instalación Spark

Anaconda

Aunque Spark fue diseñado para correr en clusters, es fácil de ejecutarlo localmente en una máquina; solo se necesita tener Java instalado y la variable de entorno JAVA_HOME que apunta a la instalación de Java.

Para usar en JupyterLab:

- Descargue e instale el JDK de Java versión 8
  - En ubuntu: `sudo apt install openjdk-8-jdk`
  - En ubuntu: `export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64`
- Ejecutar en consola anaconda e instalar Pyspark
  - conda install -c conda-forge pyspark
- Exportar variable de entorno JAVA_HOME

Externa

Si se desea utilizar la versión más nueva disponible en la pagina oficial de spark.

Para usar en JupyterLab:

- Descargue e instale el JDK de Java versión 8
  - En ubuntu: `sudo apt install openjdk-8-jdk`
  - En ubuntu: `export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64`
- Ejecutar en consola anaconda e instalar Py4J
  - conda install -c conda-forge py4j
- Descargar spark https://spark.apache.org/downloads.html
- Descomprimir spark en alguna carpeta
- Agregar ruta de pyspark
- Exportar variable de entorno JAVA_HOME

Rutas

Se mencionó que se debe crear una variable de entorno JAVA_HOME así como agregar la ruta a la instalación de spark (en caso que se desee conectar remotamente), dentro de python se puede hacer fácilmente:

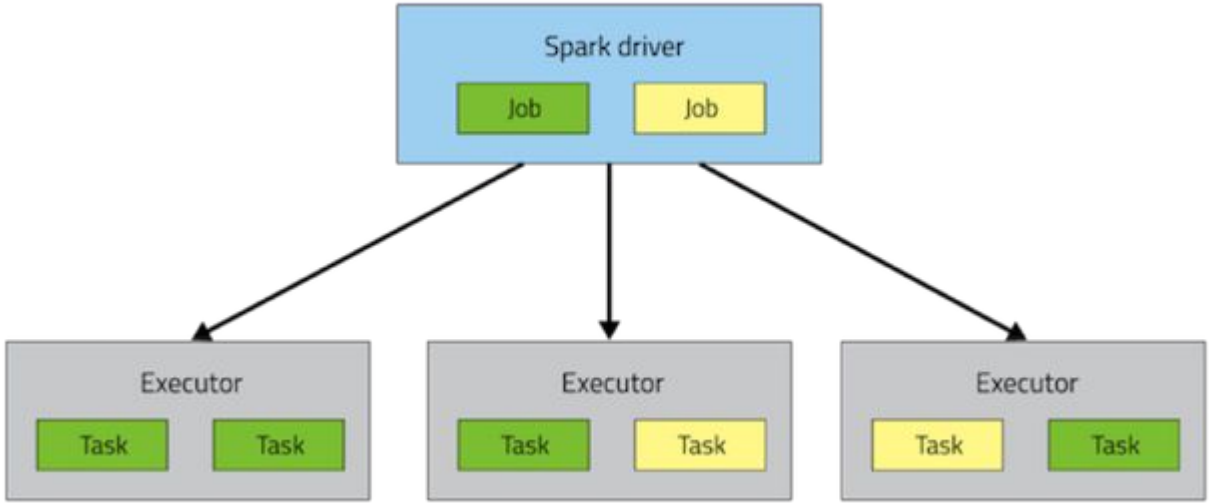
```
import os
# en windows la diagonal es \\ en linux y mac es /
os.environ["JAVA_HOME"] = "C:\\Program Files\\Java\\jdk1.8.0_241"


import sys
# es la ruta donde se descomprime spark
sys.path.append("/home/octavio.renteria/spark/python")
```

Modelo de ejecución

Spark execution model

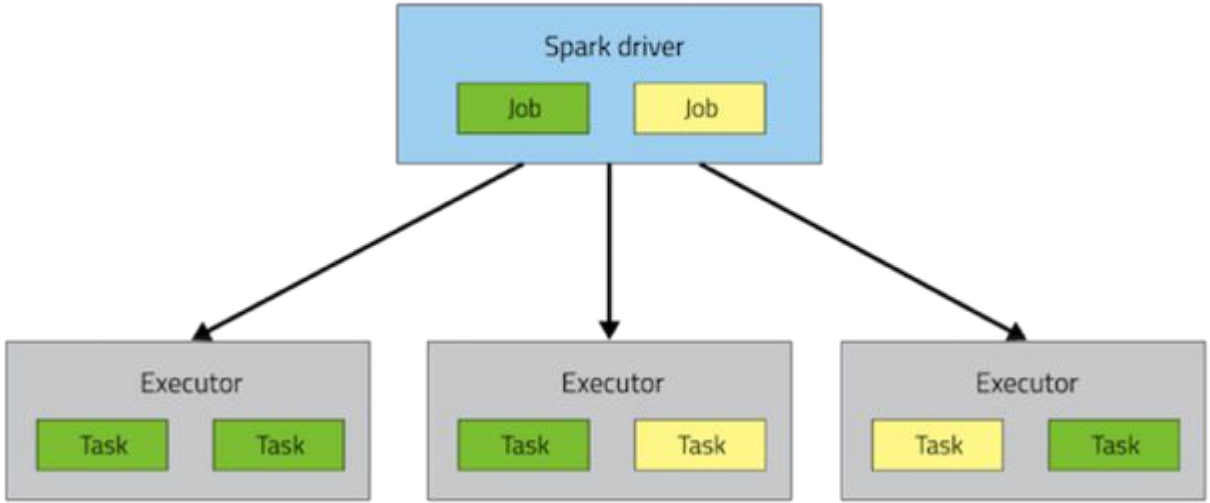
La ejecución de una aplicación Spark implica conceptos como **driver, executor, task, job,** y **stage**.



Spark execution model

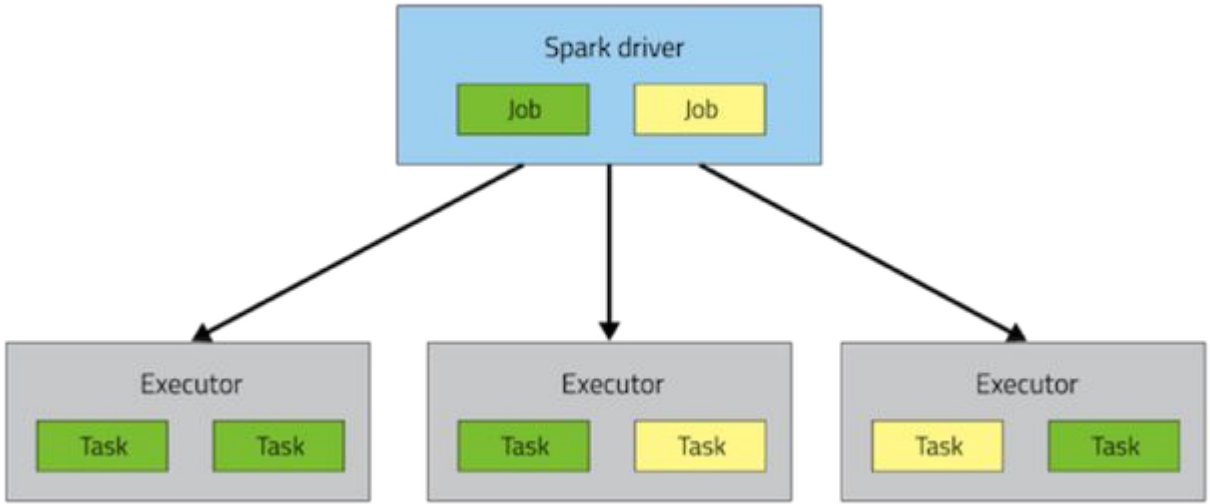
En tiempo de ejecución una aplicación Spark se mapea a un **driver** único y a un conjunto de procesos de ejecución distribuidos entre los nodos de un clúster.

El **driver** de procesos gestiona el flujo de **jobs** y **tasks**, el driver está disponible todo el tiempo que se ejecuta la aplicación.



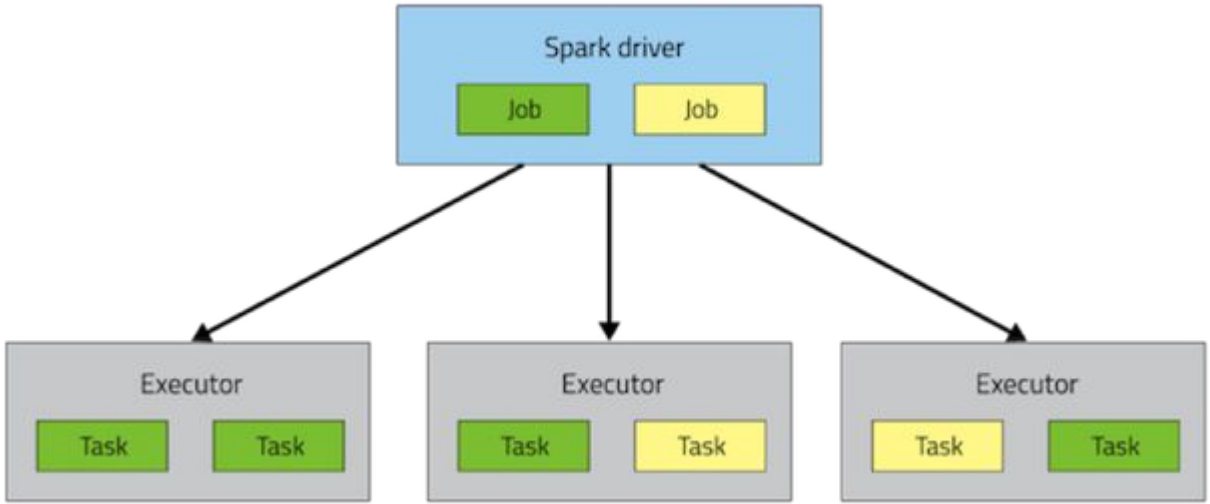
Spark execution model

Los **executors** son responsables de realizar el **job**, en forma de **tasks**, así como de guardar en caché cualquier información generada.



Spark execution model

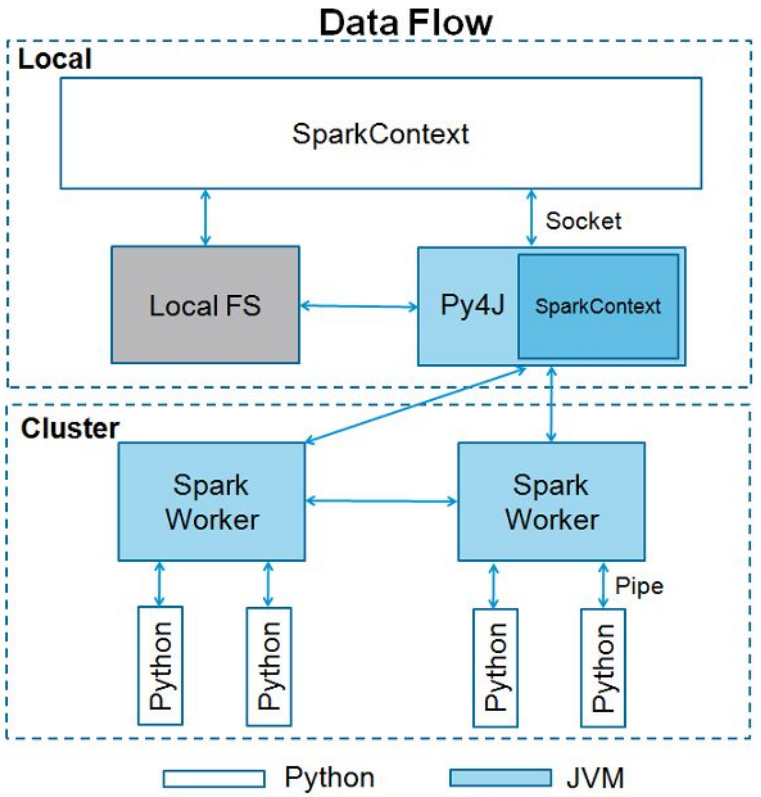
Invocar una acción dentro de una aplicación Spark desencadena el inicio de un **job**. Spark examina el conjunto de datos del que depende esa acción y formula un plan de ejecución. El plan de ejecución ensambla las transformaciones del conjunto de datos en **stages**. Un **stage** es una colección de tareas que ejecutan el mismo código, cada una en un subconjunto diferente de datos.



Spark Context

SparkContext

Es el punto de entrada a cualquier funcionalidad de Spark. Cuando ejecutamos cualquier aplicación, se inicia un programa **driver**, que contiene la función main y el **SparkContext** se inicia aquí.



SparkContext Class

```
class pyspark.SparkContext (
   master = None,
   appName = None,
   sparkHome = None,
   pyFiles = None,
   environment = None,
   batchSize = 0,
   serializer = PickleSerializer(),
   conf = None,
   gateway = None,
   jsc = None,
   profiler_cls = <class 'pyspark.profiler.BasicProfiler'> )
```

SparkContext Class

Algunos de los parámetros más relevantes:

- ***master***: Es la URL del clúster al que se conecta (remota o localmente).
- **appName**: nombre de la aplicación con que se identifica en el cluster
- **conf**: lista de opciones de spark

SparkConf Class

Un objeto Lista {SparkConf} para establecer propiedades de Spark, puede ver la lista completa en https://spark.apache.org/docs/latest/configuration.html

- **spark.executor.cores**: El número de núcleos a usar en cada ejecutor
- **spark.cores.max**: Cantidad máxima de núcleos de CPU que puede solicitar la aplicación al cluster
- **spark.executor.memory**: Cantidad de memoria a utilizar por proceso ejecutor, soporta sufijos de unidad de tamaño ("k", "m", "g" o "t", por ejemplo, 512m, 2g).

SparkConf Class

Ejemplo, se solicitan 4 núcleos y memoria de 4 gigabytes:

```
conf = pyspark.SparkConf()
conf.set('spark.executor.cores', '4')
conf.set('spark.cores.max', '4')
conf.set('spark.executor.memory', '4g')
```

Iniciando un contexto

```
# crear el contexto de spark
conf = pyspark.SparkConf()
conf.set('spark.executor.cores', '4')
conf.set('spark.cores.max', '4')
conf.set('spark.executor.memory', '4g')
# local
sc = pyspark.SparkContext(master="local",appName="MyApp",conf=conf)
# remoto
sc = pyspark.SparkContext(master="spark://10.10.22.162:7077",appName="MyApp",
                          conf=conf)
```

Spark SQL

Spark SQL, DataFrames and Datasets

Spark SQL es un módulo de para el procesamiento de datos estructurados. Trabaja sobre **Datasets**, que son un conjunto de datos distribuidos.

Un **DataFrame** es un **Dataset** organizado en columnas con nombre. Es conceptualmente equivalente a una tabla en una base de datos relacional. Los **DataFrames** se pueden construir a partir de una amplia variedad de fuentes, como archivos de datos estructurados (csv, json, xml, etc), tablas en Hive, bases de datos externas o **RDD** existentes. En este tipo de datos se pueden realizar consultas **SQL**.

Los **RDD (Resilient Distributed Datasets)** son un conjunto de objetos Java o Scala que representan una colección de elementos particionados en los nodos del clúster que se pueden operar en paralelo.

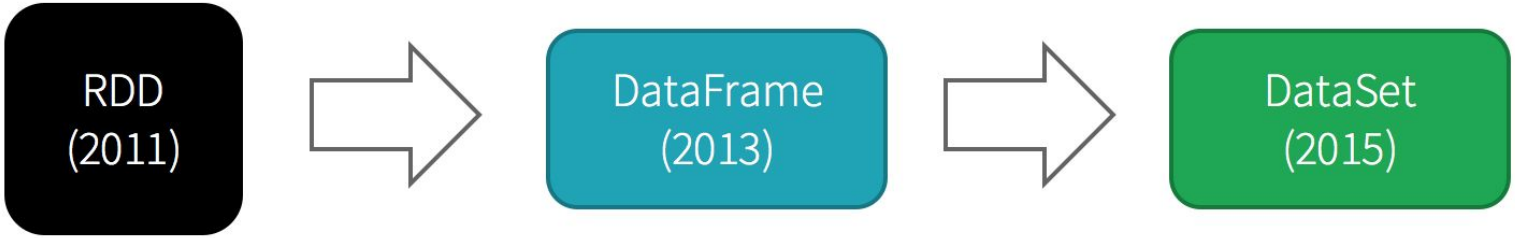
Características de los DataFrame/RDD

- **De naturaleza inmutable**: podemos crear DataFrames/RDD pero no podemos cambiarlos. Se les pueden aplicar transformaciones.
- **Evaluaciones perezosas**: lo que significa que una tarea no se ejecuta hasta que se realiza una acción.
- **Distribuido**: RDD y DataFrame son distribuidos por naturaleza.

Ventajas de DataFrame

- Están diseñados para procesar una gran colección de datos estructurados o semiestructurados.
- Los datos se organizan en columnas con nombre, lo que ayuda a Spark a comprender el esquema de un DataFrame para optimizar el plan de ejecución en sus consultas.
- Tienen la capacidad de manejar petabytes de datos.
- Tiene soporte para una amplia variedad de fuentes y formatos de datos.
- Tiene soporte para APIs de diferentes lenguajes como Python, R, Scala, Java.

Historia de las APIs



Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

Fast/efficient internal representations

Internally rows, externally JVM objects

Almost the “Best of both worlds”: **type safe + fast**

But slower than DF
Not as good for interactive analysis, especially Python

Cargando archivos

sql.SparkSession.read

Interfaz utilizada para cargar un DataFrame desde sistemas de almacenamiento externo (archivos locales o remotos).

Este ejemplo muestra cómo cargar un archivo csv, se pueden exportar archivos de varios orígenes:

```
# archivo local
input_data = sqlContext.read.csv("my_file.csv")
# desde un servidor hadoop de archivos
input_data = sqlContext.read.csv("hdfs://10.10.22.162:9000/user/hadoop/my_file.csv")
# desde Amazon Simple Storage Service S3 (requiere configuración de llaves)
input_data = sqlContext.read.csv("s3n://my_file.csv")
```

sql.SparkSession.read

Interfaz utilizada para cargar un DataFrame desde sistemas de almacenamiento externo (archivos locales o remotos).

Este ejemplo muestra cómo cargar un archivo json, se pueden exportar archivos de varios orígenes:

```
# archivo local
input_data = sqlContext.read.json("my_file.json")
# desde un servidor hadoop de archivos
input_data = sqlContext.read.json("hdfs://10.10.22.162:9000/user/hadoop/my_file.json")
# desde Amazon Simple Storage Service S3 (requiere configuración de llaves)
input_data = sqlContext.read.json("s3n://my_file.json")
```

DataFrame

Detalles del dataframe cargado

```
print( type(input_data) )
input_data.printSchema()
input_data.show()
```

Crear una vista para manejar los datos:

```
input_data.createOrReplaceTempView("zonas")
```

DataFrame SQL

Se pueden hacer consultas con sentencias SQL sobre los DataFrames, la función sql ejecuta la sentencia y regresa un DataFrame con los datos de la consulta:

```
# seleccionar las columnas CVE_ZM, NOM_ZM y POB_2015, en caso de que el nombre de la columna
# Tenga caracteres conflictivos (como espacios o caracteres especiales), se pueden usar
# comillas invertidas ` `
result = sqlContext.sql("SELECT CVE_ZM, NOM_ZM, POB_2015 from zonas")
result.show()
```

El resultado siempre es un DataFrame.

Funciones DataFrame

Algunas funciones relevantes del DataFrame:

- **select**: genera un DataFrame con las columnas requeridas
- **rdd**: convierte el DataFrame a RDD
- **limit**: genera un DataFrame con los primeros “N” renglones especificados.
- **show**: Imprime en pantalla los primeros 20 renglones especificados.
- **filter**: genera un DataFrame con los renglones donde una condición es verdadera.
- **where**: es un alias de **filter**.
- **collect**: Copia al host todo el DataFrame en forma de lista de renglones.
- **withColumn**: Modifica los valores de una columna o agrega una nueva columna.

MLlib