Creado por:

Isabel Maniega

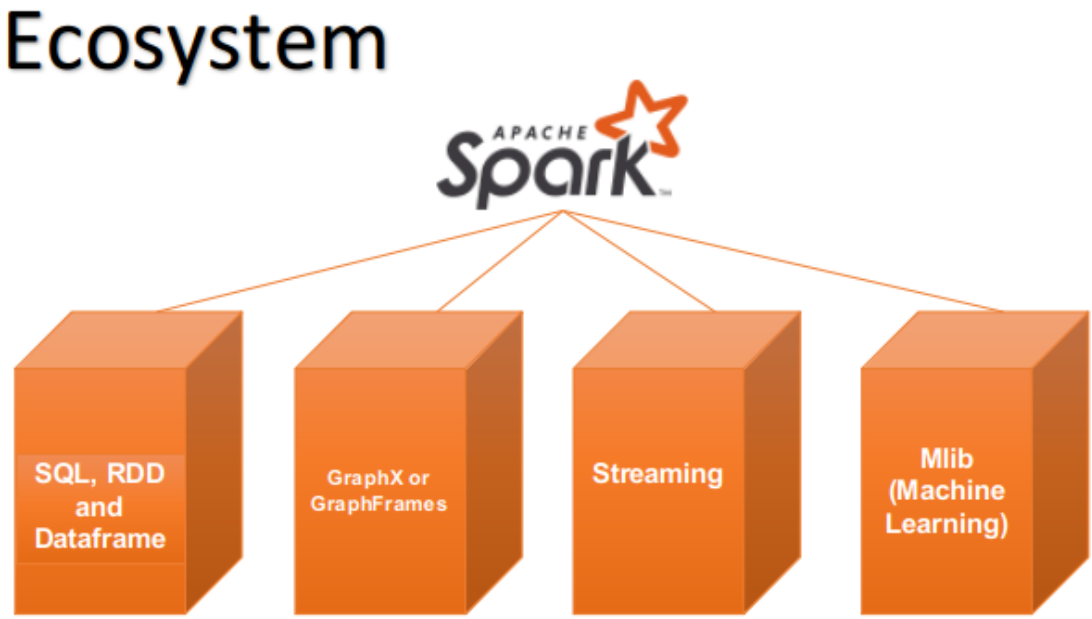
Spark

https://spark.apache.org/ (https://spark.apache.org/)

https://spark.apache.org/docs/latest/ (https://spark.apache.org/docs/latest/)

Framework open source para procesamiento distribuido diseñado para ser rápido. Se basa en Hadoop Map Reduce y permite dividir y paralelizar el trabajo.

Ejemplo dividir un fichero y luego lo vuelve a unir, en procesamiento en memoria.



RDD: Conjunto de Datos Distribuidos Resistentes, estos datos se dividen en particiones lógicas, y se calculan en distintos nodos.

SQL: datos estructurados, brinda información sobre los datos.

Dataframe: colección distribuida de datos organizados en columnas.

GraphX: procesamiento de gráficos distribuido, basada en la teoría de Grafos.

Streaming: procesamiento de datos a tiempo real.

MLlib: librería de aprendizaje automático.

Instalación

Para la instalación usaremos una imagen de docker Hub:

- https://hub.docker.com/r/jupyter/pyspark-notebook (https://hub.docker.com/r/jupyter/pyspark-notebook)

Descargamos la imagen en el pc con:

- docker pull jupyter/pyspark-notebook

Para ejecutar la imagen necesitamos:

- docker run -it --rm -p 8888:8888 -v /home/isabelmaniega/notebooks:/home/jovyan/work jupyter/pyspark-notebook:latest

Si tenemos otro puerto abierto podemos modificar el primero de ellos:

- *docker run -it --rm -p **8890**:8888 -v /home/isabelmaniega/notebooks:/home/jovyan/work jupyter/pyspark-notebook*

Tambien se ha modificado el volume:

- docker run -it --rm -p 8888:8888 -v /home/isabelmaniega/notebooks:**/work** jupyter/pyspark-notebook

- -p: puerto de ejecución 8888
- -v: el volume donde se ubican nuestros proyectos, primera path corresponde a la ubicación en mi pc y la segunda a la ubicación en el contenedor, esta última no se modifica NUNCA.

NOTA:

En el caso de sistemas operativos Windows poner "c" en minúscula: c:/...:/home/jovyan/work

Una vez lanzado nos indicará la url con el token para poder ejecutar en el navegador.

Para ejecutar en segundo plano podemos añadir -d y a la instrucción y para obtener la url realizaremos la consulta al log:

- docker run -d -it --rm -p 8888:8888 --name notebook -v /home/isabelmaniega/notebooks:/work jupyter/pyspark-notebook

--> get the notebook token from the logs

- docker logs --tail 3 notebook

  -->Or copy and paste one of these URLs:

  --> http://878f1a9b4dfa:8888/lab?token=d336fa63c03f064ff15ce7b269cab95b2095786cf9ab2ba3 (http://878f1a9b4dfa:8888/lab?token=d336fa63c03f064ff15ce7b269cab95b2095786cf9ab2ba3)

  --> or http://127.0.0.1:8888/lab?token=d336fa63c03f064ff15ce7b269cab95b2095786cf9ab2ba3 (http://127.0.0.1:8888/lab?token=d336fa63c03f064ff15ce7b269cab95b2095786cf9ab2ba3)

Creado por: