Spark & Hadoop developer

Spark is not a language but a distributed processing engine.

Understanding HDFS commands

Python

Hive (for SQL queries)

Itversity

There are multiple ways to connect to itversity labs

1. Console in the lab
2. Command prompt using ssh username@

Or we could download it on our machine :

Download Spark

Download Java and Javac

Making everything as envirionement variables

Validate spark by writing a simple command like reading a file

Also it could be integrated with python IDEs , for that we nned to add pyspark to the

IDE , check the course

From pyspark import sparkcontext, sparkconf

Hadoop using

Via

RDD , transformation API are the ones that we can process the data and Actions are those that we can preview the data or save the data

Lazy evaluation : transformation does not perform any job but actions will

DAG : Directed acyclic graph

Every transformation will be collected in Dag and when an action is called , then the DAG runs those transformation

Collect action is a giving list !!! it is super inefficient ..

Rdd 🡺 list using collect ()

How to create a RDD from list . sc.parrelize

Hdfs bash command are Hadoop fs ….

SC , spark context is a gate way to talk to spark

Sqlcontext is another variable created by sc.

Load and read are the methods .

load and read returns dataframe

load (path , fileformat)

read.json

read.text

read.orc,

read.parquet

**Parquet** is a columnar file format that provides optimizations to speed up queries and is a far more efficient file format than CSV or JSON, supported by many data processing systems. **Spark** SQL provides support for both reading and writing **parquet** files that automatically capture the schema of the original data.