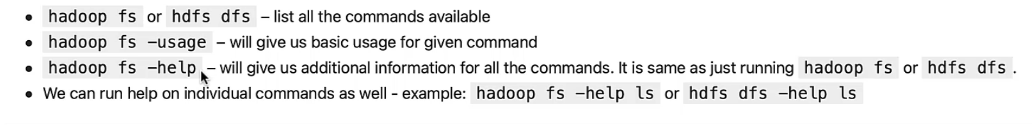
Spark Ultimate

Basic Hadoop commands: -



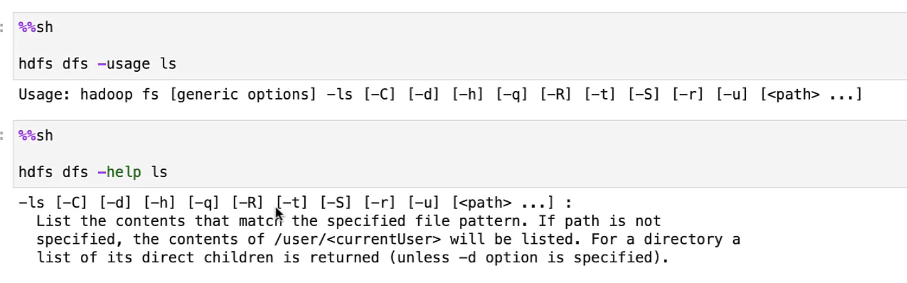
List all the available commands with their details: -



Listing basic usages and the syntax of all commands: -

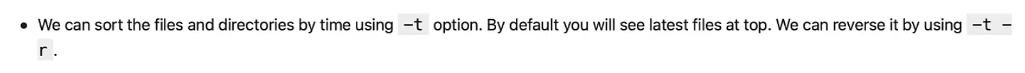


Details for a specific command: -



Reverse the order of sort of the ls command output: -

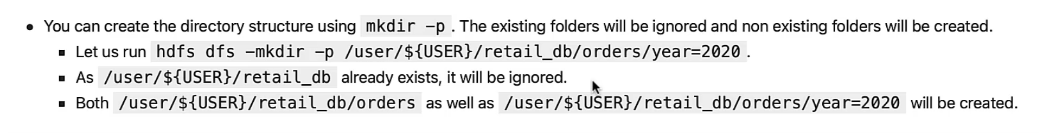




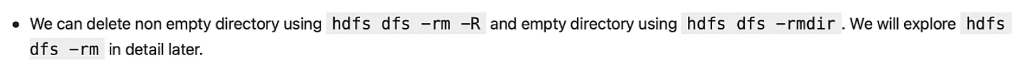




Creating a folder structure: -

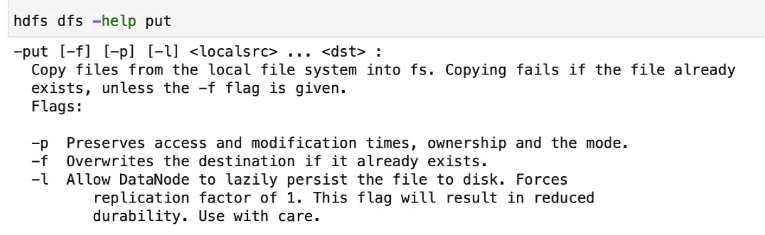


Deleting folder: -



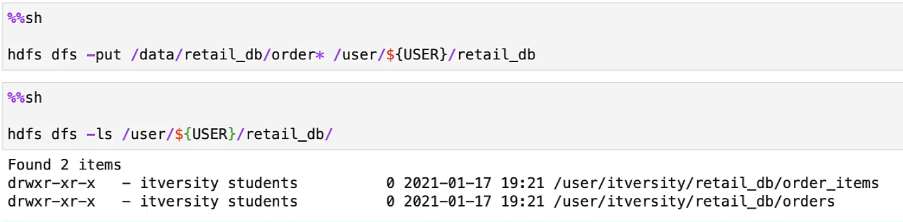


Copy files from Local to HDFS: -





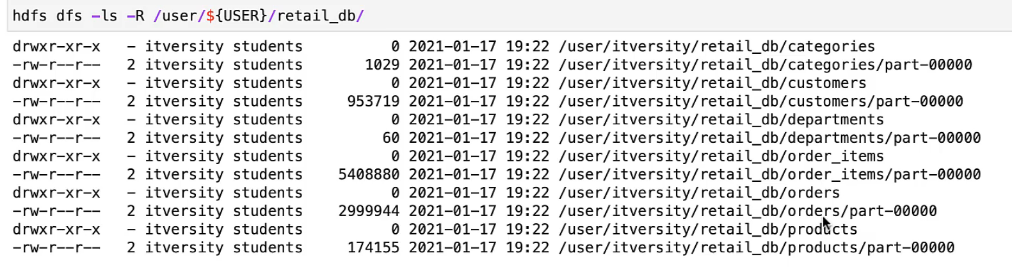
Using wild cards / patterns: -





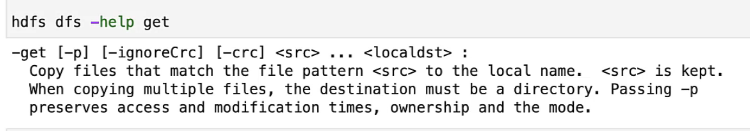
* Already existing folders will be replaced

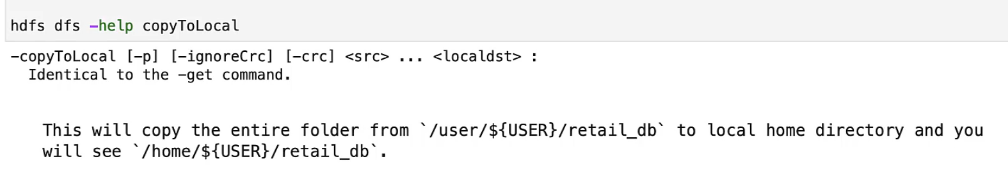
Validation: -





Copy files from HDFS to local: -

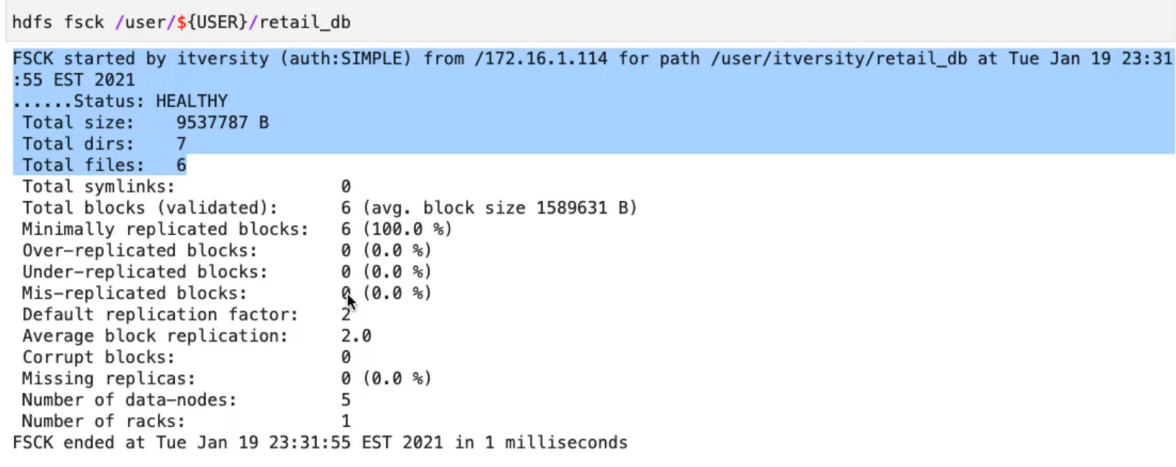


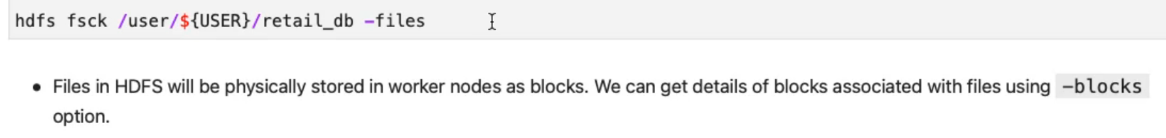




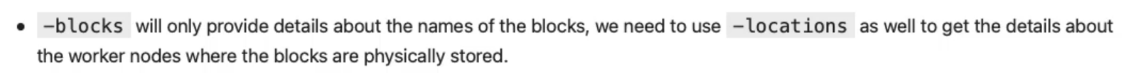
* If some files or folders are already present it will copy only the nonsexist ones

Getting file metadata: -

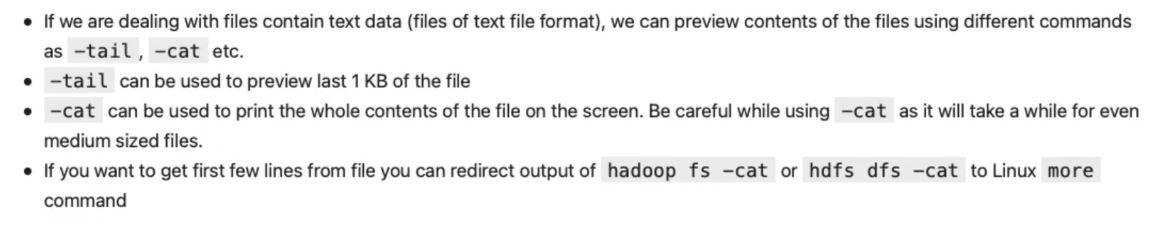


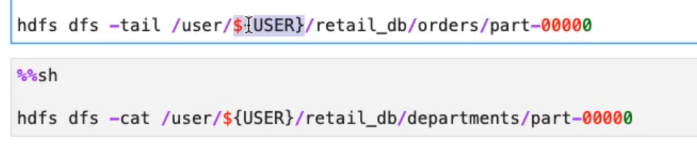






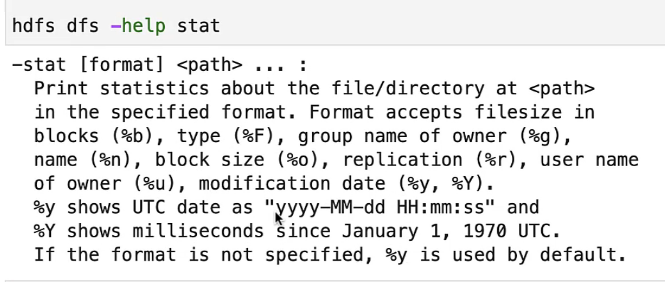
Previewing data in HDFS: -



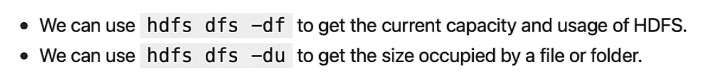


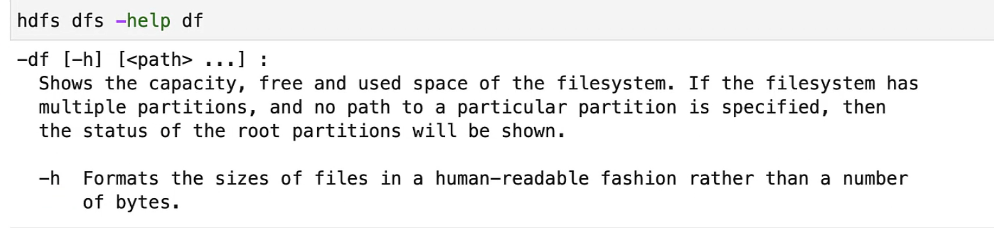
* Cat supports patterns but tail doesn’t

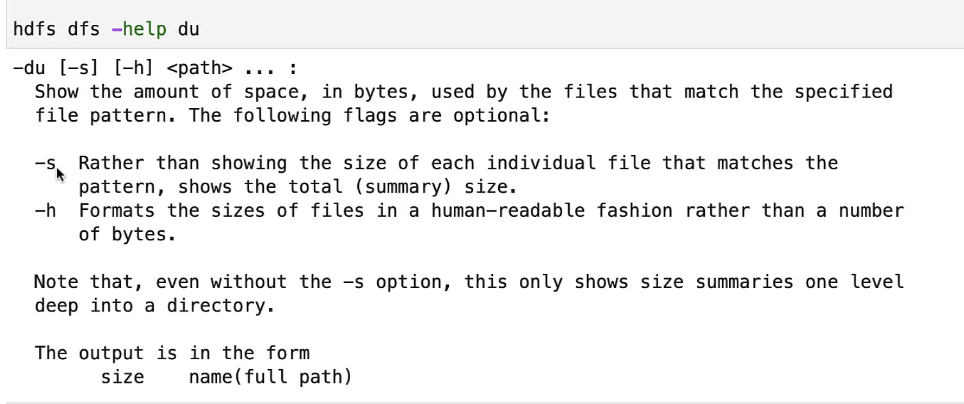
Replication factor: -



Getting HDFS storage usage: -









HDFS file permissions: -



* u 🡪 user
* g 🡪 group
* o 🡪 others

Overriding the properties: -



Spark

Getting help form spark shell: -

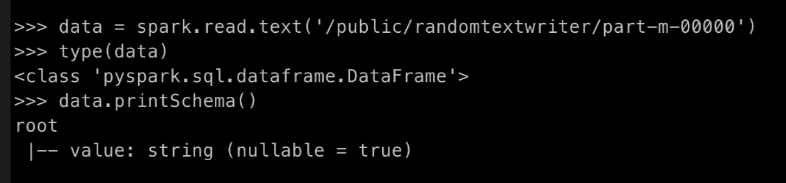


Creating rdd from a text file and minPartitions: -

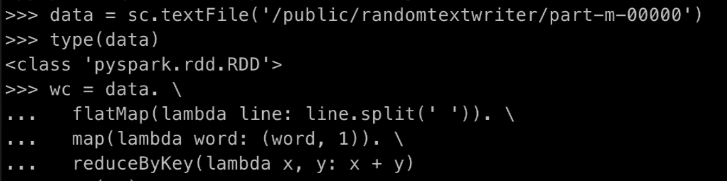


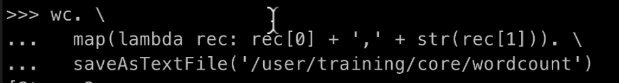
* If you specify less number in ‘minPartitions’ bydefault it will submit actual number of partitions as number of jobs

Creating a dataframe: -

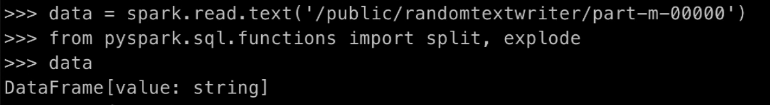


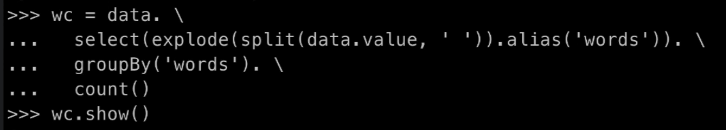
Word count problem RDD: -





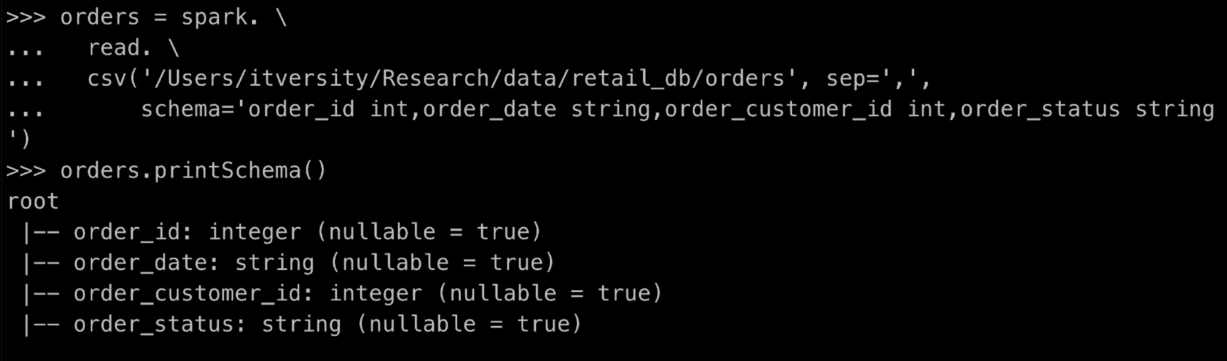
Word count problem Dataframe: -

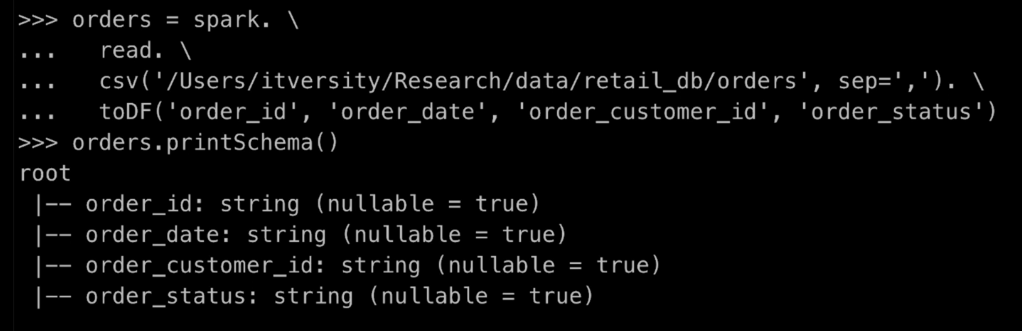




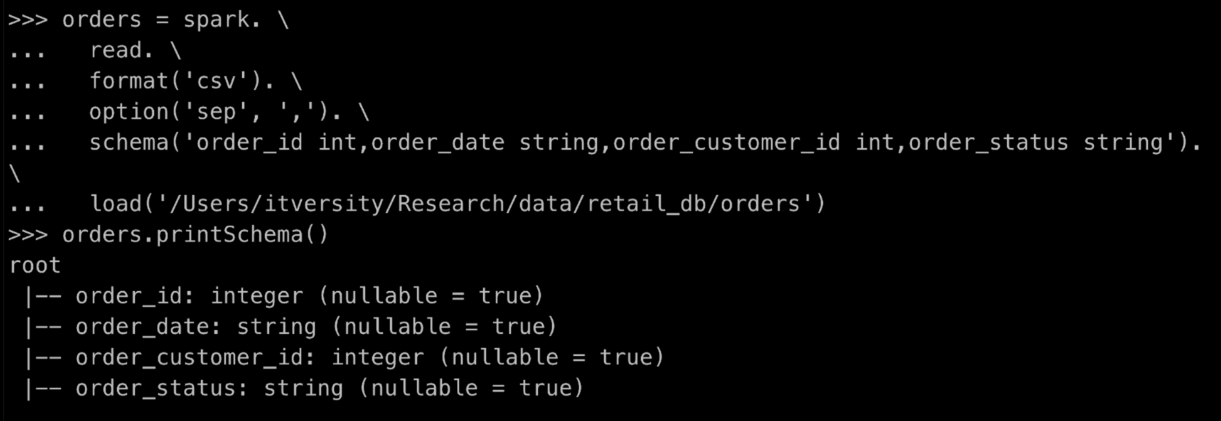


Reading data and attaching schema: -

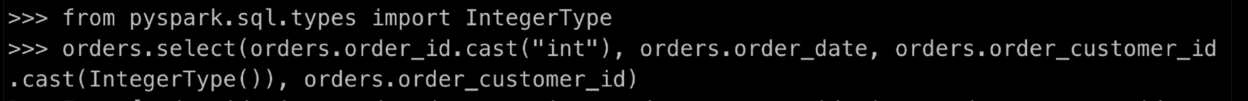


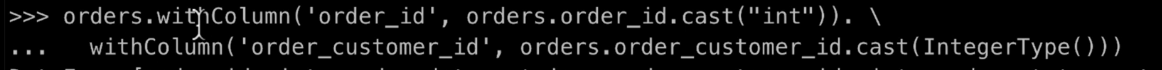


* It will not change the datatype



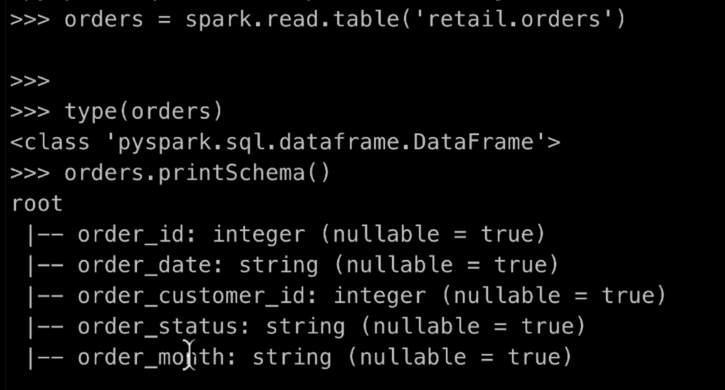
Type Casting: -





* Type casting without selecting all columns

Reading Hive Table from Spark: -

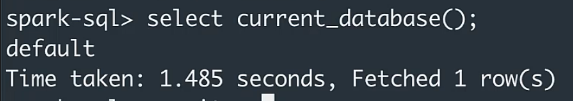


* You can use it if you already have hive and spark connectivity

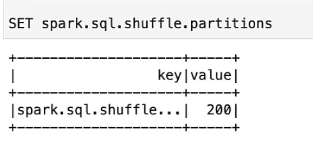
Spark SQL

If we are using Pyspark the command will be spark.sql(“the command”).show()

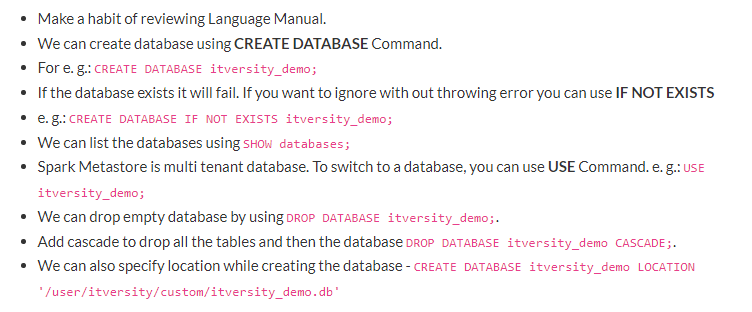
Showing current database: -



Shuffle partitions: -



Spark meta store database: -

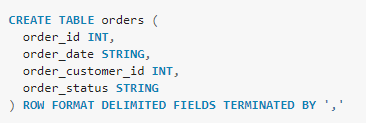


Creating a database: -

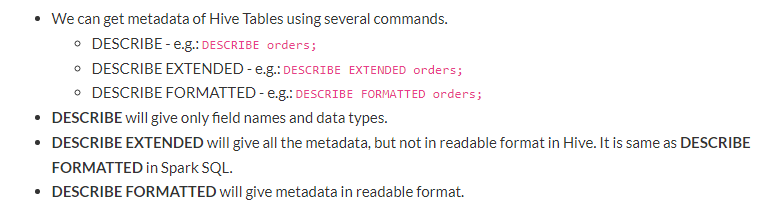




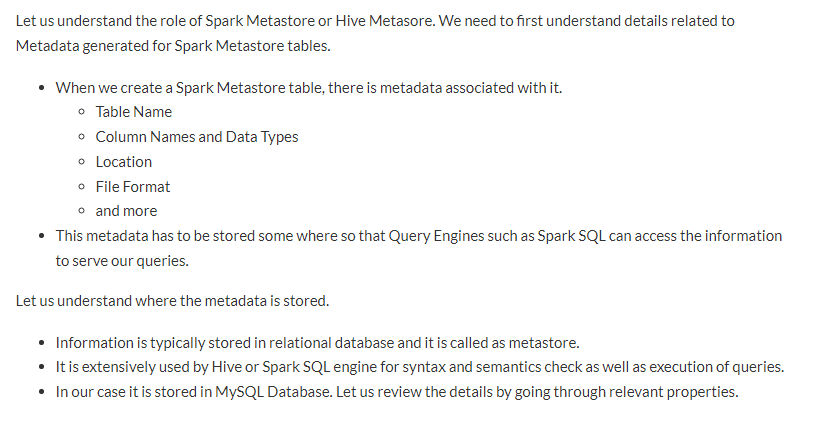
Creating a spark table: -



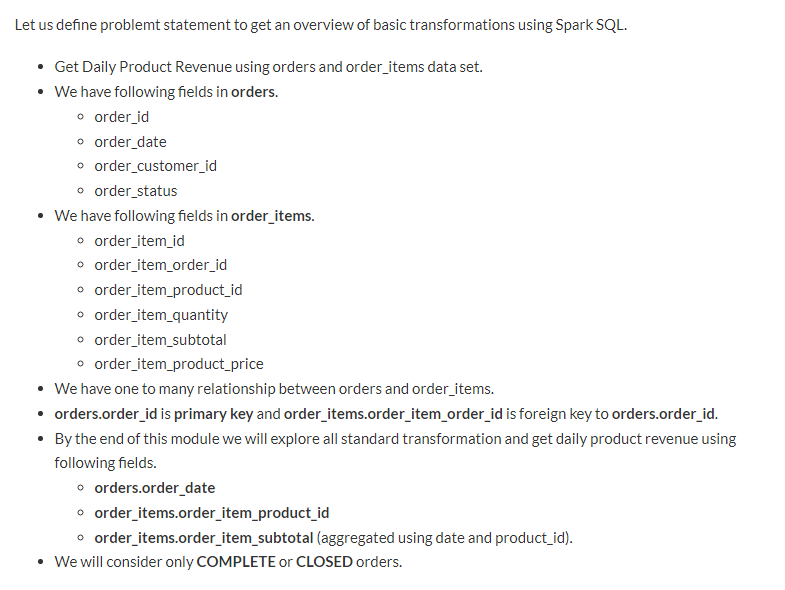
Retrieving metadata of a table: -



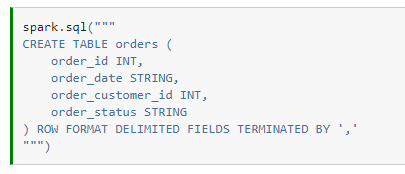
Role of Spark Metastore or Hive Metastore: -



Problem statement: -



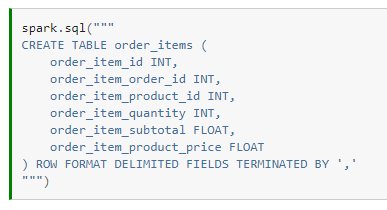
Creating orders table: -



Load data: -



Creating order\_items: -



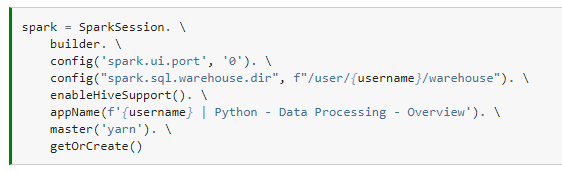
Load data: -



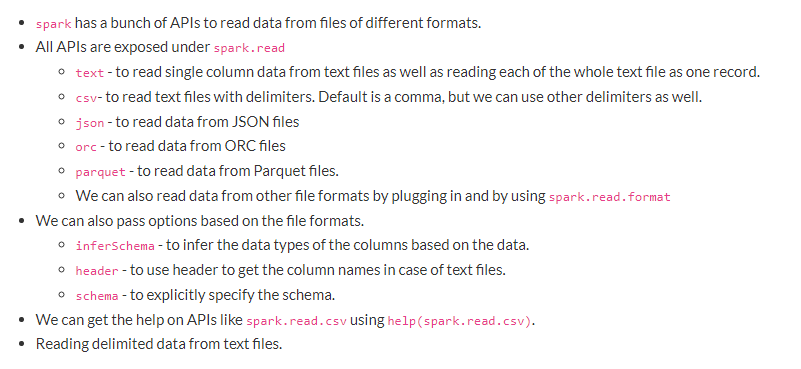
Data processing using Spark – The Pythonic way

Creating a Spark session: -



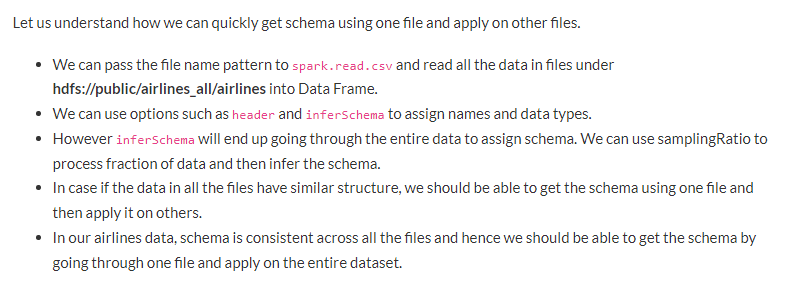


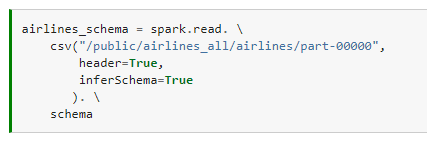
Read APIs: -





Infer schema: -





* If we use it along with the samplingRatio it will be better

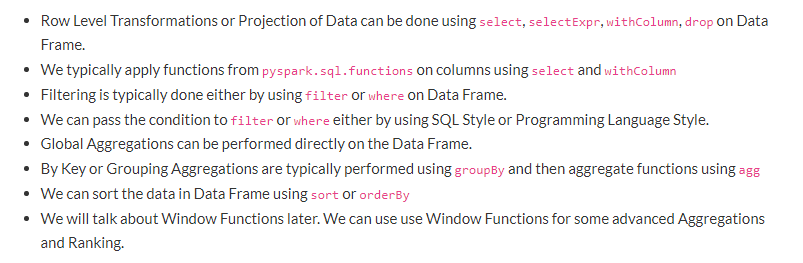


* Airlines\_part\_00000 is a dataframe

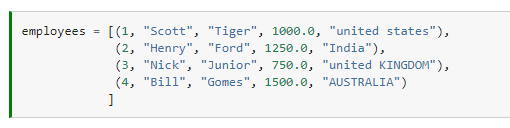
Distinct count: -

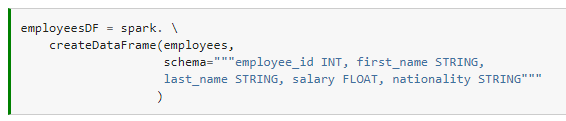


Spark DataFrame API: -

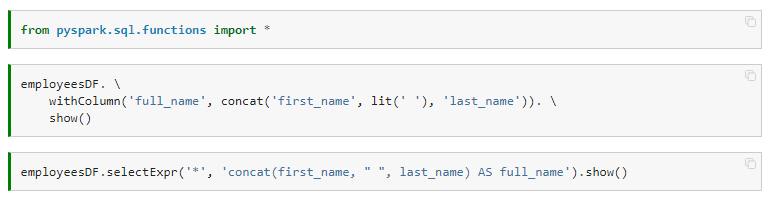


Creating a DataFrame from collection: -



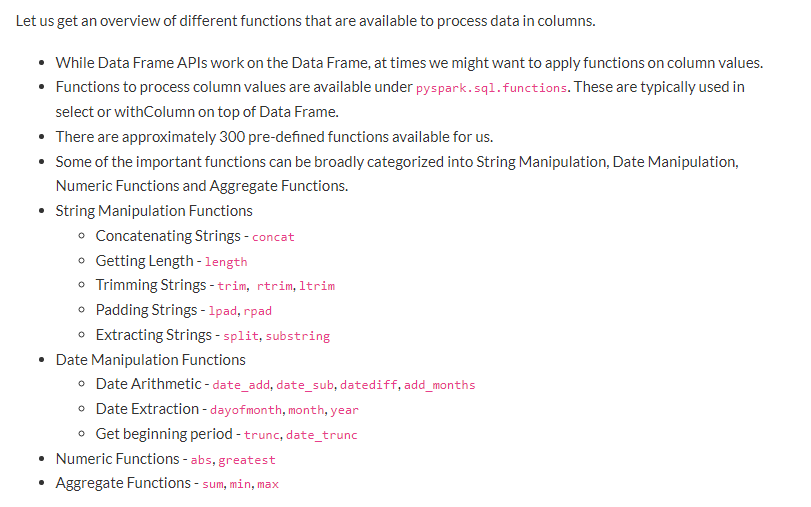


Few operations: -

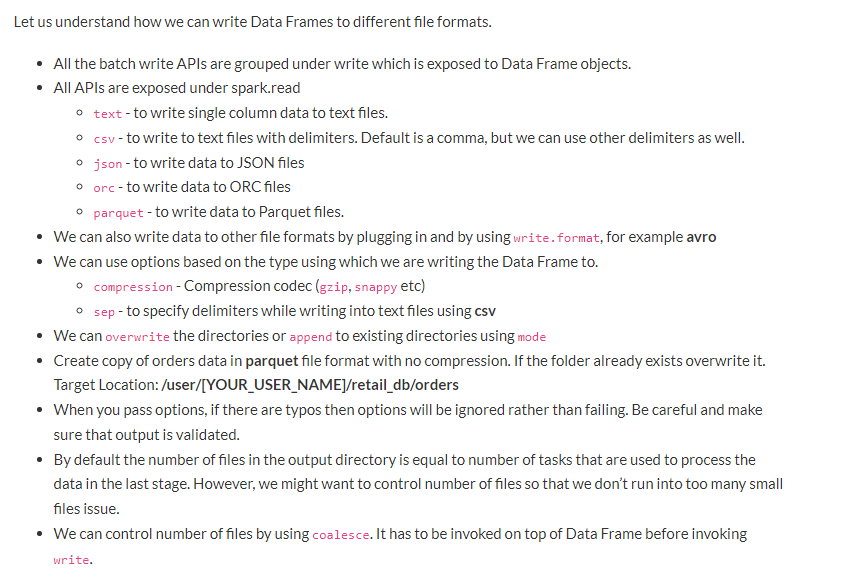


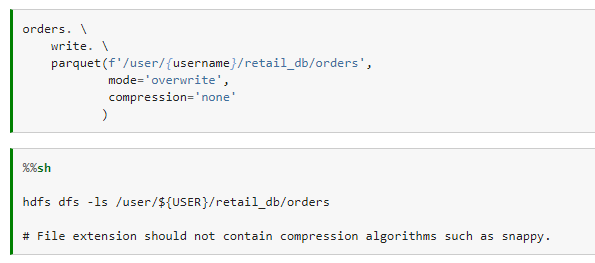
* If you are doing it in programmatic way lit() function is important. Otherwise it will search for a column named ‘ ’ and won’t able to find it and fail

Overview of Functions: -



Overview of Spark write APIs: -









Coalesce: -

