1)What is Spark?

Spark is a unified analytics engine for large-scale data processing including built-in modules for SQL, streaming, machine learning and graph processing.it consists of processing RDD’s and Dataframes

2)What is advantages of Spark and its core components

Advantages of spark are its 1.speed and performance

2.developer friendly tools

3.libraries

4.structured streaming

Spark streaming

Mlib

Spark sql

are core components

3)What are the Transformatin and actions?

When we look at the Spark API, we can easily spot the difference between transformations and actions. If a function returns a DataFrame , Dataset , or RDD , it is a transformation. If it returns anything else or does not return a value at all (or returns Unit in the case of Scala API), it is an action.

4)Camparison between Spark & MR

The primary difference between Spark and MR(MapReduce) is that Spark processes and retains data in memory for next steps, whereas MapReduce processes data on disk. As a result, for smaller workloads, Spark's data processing speeds are up to 100x faster than MapReduce.

Spark is a Hadoop enhancement to MapReduce.

5)Comparison between Scala & Python

6)What is RDD and It types & RDD Creation types with examples

Rdds are resilient distributed datasets it has the collection of data which can be used for datapreprocessing



7)What is the RDD creation syntax for CSV,TXT and Json file.

val rddltxt=sc.textFile("file:/// viv1/patient.txt")

val rddlcsv=sc.textFile("file:///home/ viv1/patient.csv")

val rddljson=sc.textFile("file:///home/viv1/patient.json")

8)What is partitions and how it control through syntax

9)What is Immutable and mutable?

Immutables are the items which cannot be changed

Mutables are items which can be changed as we like

Ex:val and var

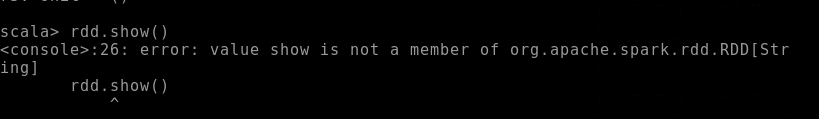
10)What is Lazy evaluation and fault tolrenace in spark

2 question

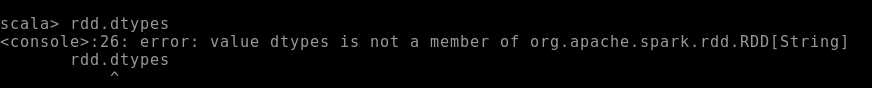
To create rdd



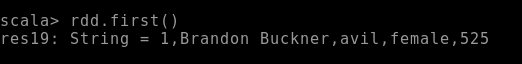
For showing the rdd rdd.show()



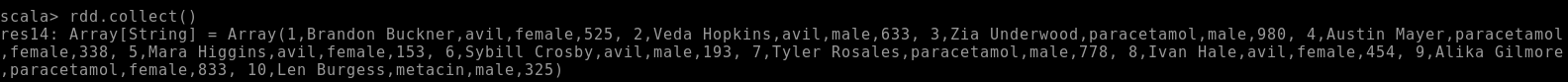
To see datatypes



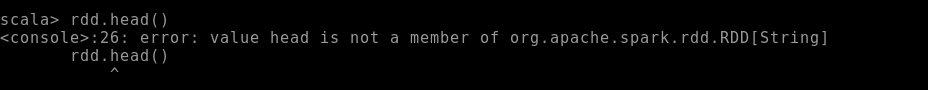
To see first row data



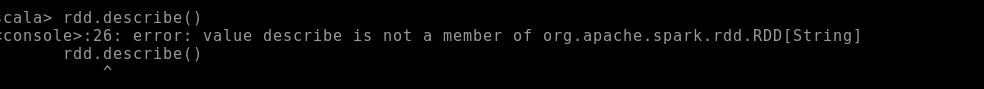
To see data in rdd



To see head of rdd



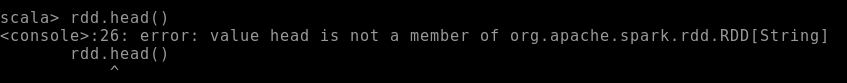
Describing rdd



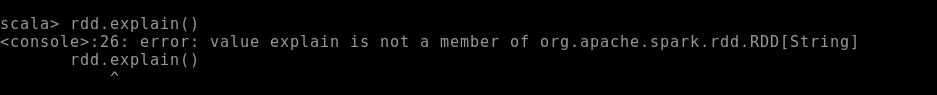
Show(n)



Rdd.head()



Explaining rdd

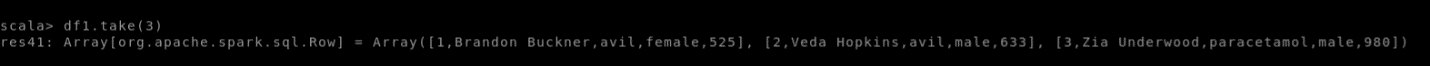


Counting rdd

Graphical user interface

Description automatically generated with medium confidence

Take 3



3 question

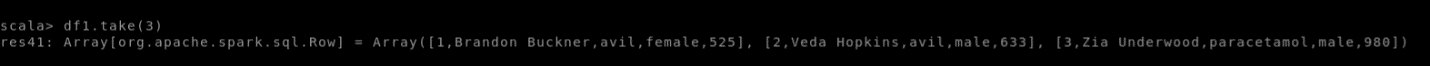
Describing the dataframe



Explain with df1.explain()



Df1.take()

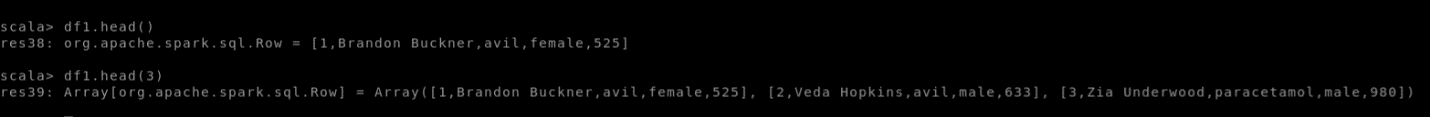


Df1.limitshow()

Text

Description automatically generated

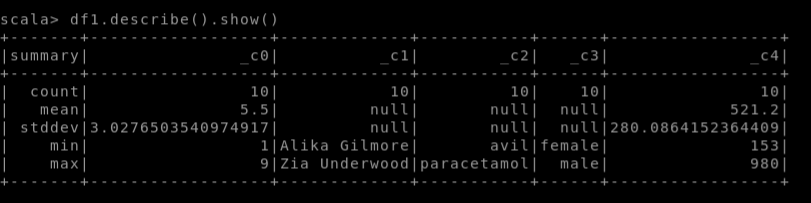
Df1.head()



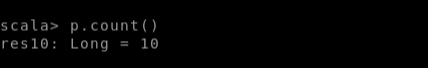
p.first()



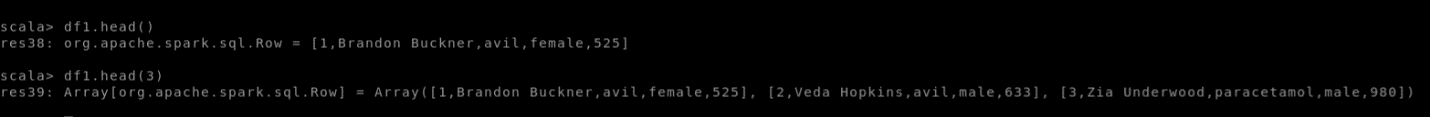
df1.describe().show()



p.count()

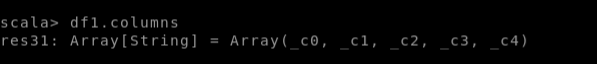


Df1.head()

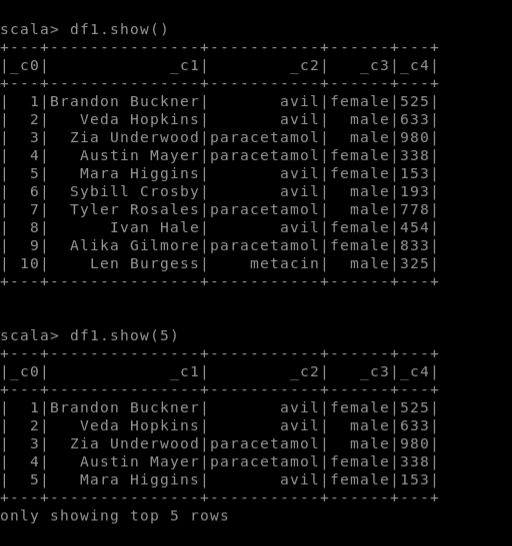


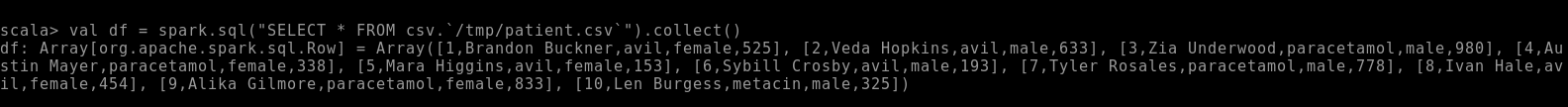
Df1.dtypes

Df1.columns



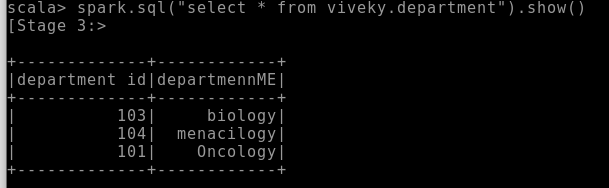
Df1.show()





4 question

Created a department table using spark sql



Created employee table using spark we created the tables and sent to hive

Text

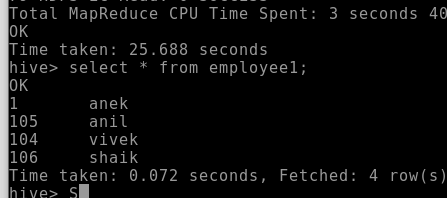
Description automatically generated

Those tables are sent to hive

Text

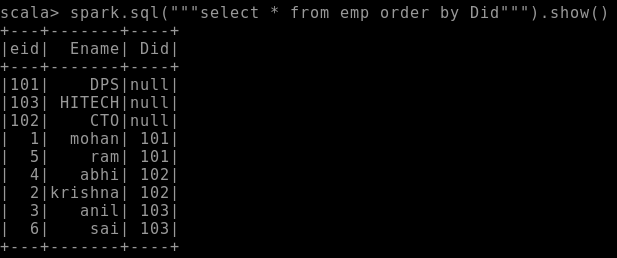
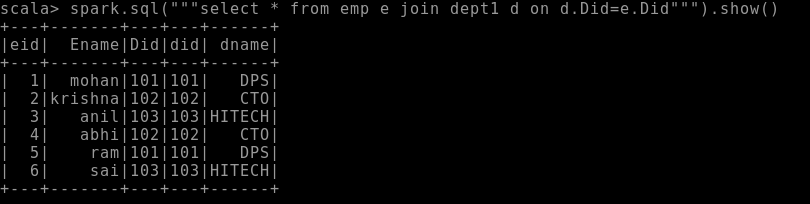
Description automatically generated

Inserted data records using hive



Text

Description automatically generated

Text

Description automatically generated with low confidence