**spark shell:**

Spark-shell command starts the spark

A screenshot of a computer screen

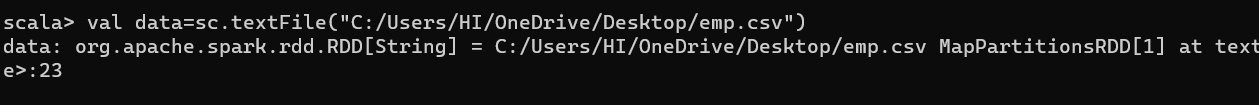
Description automatically generated

Create rdd using parallelize:

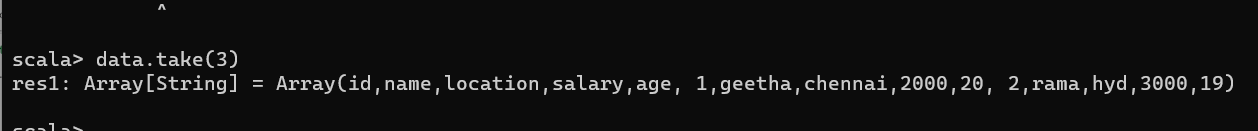
A screen shot of a computer

Description automatically generated

Read:textfile reads the file into the given rdd.



Take:to read first 3 lines



Count:

A black background with white text

Description automatically generated

Filter:

A black screen with white text

Description automatically generated

First:

A black background with white text

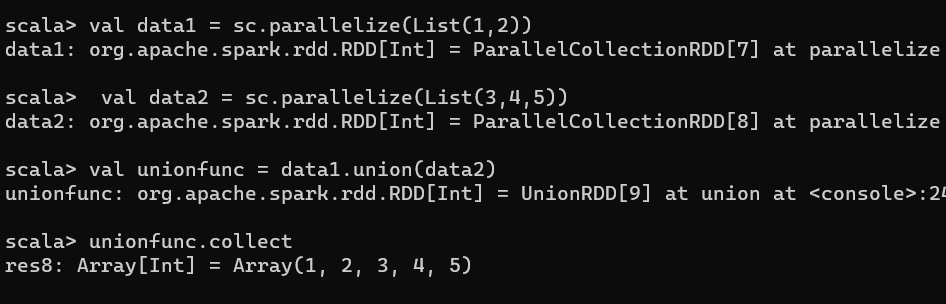
Description automatically generated

Distinct:

A black screen with white text

Description automatically generated

Union:

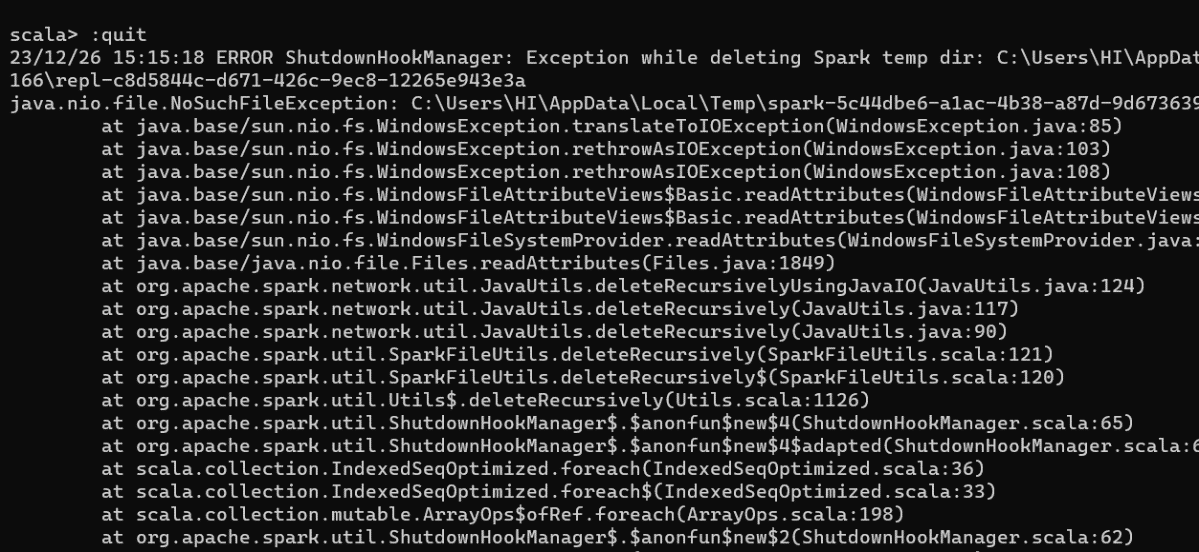


Sort:

A black screen with white text

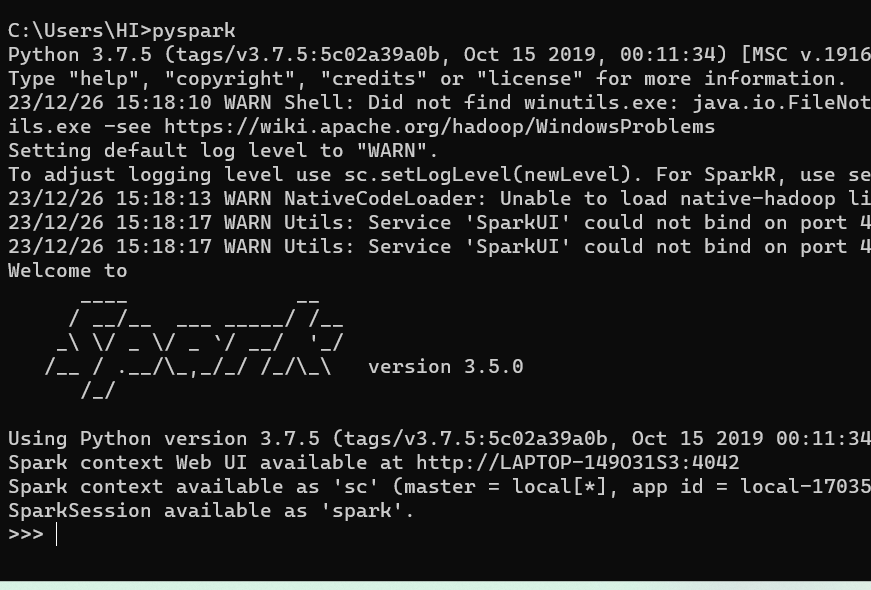
Description automatically generated

Quit command:



**Pyspark:**

Type pyspark command in cmd to start pyspark:



Reading files:

Spark.read.load(): This function is utilized to load a file into the Spark shell, making the data accessible for further processing.

A computer screen shot of a black screen

Description automatically generated

printSchema:

gives the schema of the dataframe:

A screen shot of a computer code

Description automatically generated

Topandas:

Is used to change the dataframe to pandas data frame:

A black screen with white text

Description automatically generated

Limit is used to limit the values upto specified number.

Sort:

A screen shot of a computer code

Description automatically generated

Filter:

A black screen with white text

Description automatically generated

Sql:

* **Var\_name.registerTempTable('table\_name'):** This command registers a DataFrame as a temporary table, facilitating SQL-like operations for enhanced performance.
* **sqlContext.sql('query'):** The sqlContext.sql() function is used to execute SQL queries on the registered temporary tables within PySpark.

**Df.**registerTempTable(‘name’)

A screenshot of a computer program

Description automatically generated