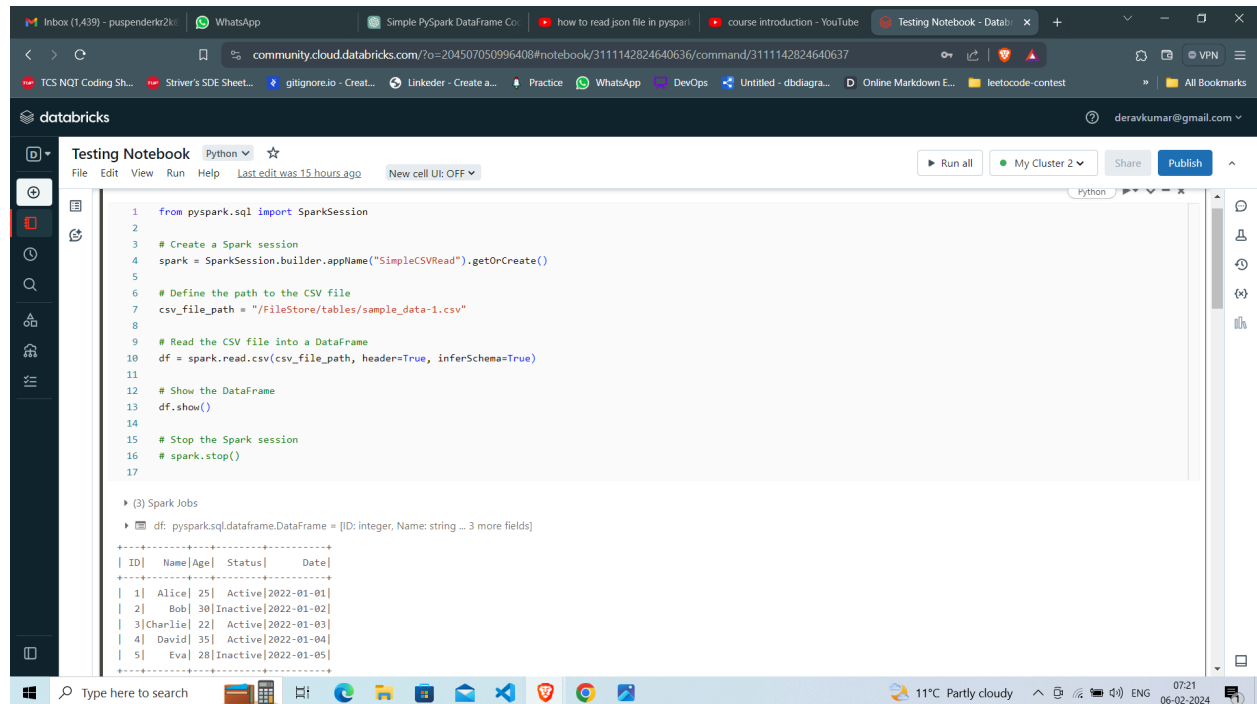


Basic Pyspark Program



The screenshot shows a Databricks Testing Notebook interface. The notebook contains a PySpark program that reads a CSV file from a FileStore and displays the resulting DataFrame. The code is as follows:

```
1 from pyspark.sql import SparkSession
2
3 # Create a Spark session
4 spark = SparkSession.builder.appName("SimpleCSVRead").getOrCreate()
5
6 # Define the path to the CSV file
7 csv_file_path = "/FileStore/tables/sample_data-1.csv"
8
9 # Read the CSV file into a DataFrame
10 df = spark.read.csv(csv_file_path, header=True, inferSchema=True)
11
12 # Show the DataFrame
13 df.show()
14
15 # Stop the Spark session
16 spark.stop()
17
```

The output of the program shows the DataFrame with 5 rows and 5 columns (ID, Name, Age, Status, Date):

ID	Name	Age	Status	Date
1	Alice	25	Active	2022-01-01
2	Bob	30	Inactive	2022-01-02
3	Charlie	22	Active	2022-01-03
4	David	35	Active	2022-01-04
5	Eva	28	Inactive	2022-01-05

RDD's and Transformation in Pyspark

In PySpark, Resilient Distributed Datasets (RDDs) are the fundamental data structures used to perform distributed data processing. RDDs represent a distributed collection of immutable objects, and they can be processed in parallel across a cluster of machines. RDDs provide fault tolerance by automatically recovering lost data partitions due to node failures.

Transformations in PySpark are operations performed on RDDs to create a new RDD. Transformations are typically executed lazily, meaning the computation is not immediately performed when the transformation is called. Instead, transformations are only evaluated when an action is executed.

Transformation of data with Pyspark RDD's

Inbox (1,439) - p...

WhatsApp

Pyspark Day 1, 2

how to read json

course introducti...

Testing Note...

Home - Google I...

Untitled docume...

Assesment Trac...

+

community.cloud.databricks.com/?o=204507050996408#notebook/3111142824640636/command/3355311977861211

TCS NQT Coding Sh... Striver's SDE Sheet... gitignore.io - Creat... Linkeder - Create a... Practice WhatsApp DevOps Untitled - dbdiagra... Online Markdown E... leetcode-contest All Bookmarks

databricks

deravkumar@gmail.com

Testing Notebook Python

File Edit View Run Help Last edit was 7 minutes ago New cell UI: OFF

Run all My Cluster 2 Share Publish

Command took 0.14 seconds -- by deravkumar@gmail.com at 2/5/2024, 3:07:36 PM on my Cluster 1

Cmd 3

```
1
2  dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
3  rdd=spark.sparkContext.parallelize(dataList)
4  result = rdd.collect()
5  ("RDD Contents:", result)
```

▶ (1) Spark Jobs

(('RDD Contents:', [('Java', 20000), ('Python', 100000), ('Scala', 3000)]))

Command took 0.72 seconds -- by deravkumar@gmail.com at 2/5/2024, 3:47:39 PM on my Cluster 1

Cmd 4

```
1  rdd = sc.parallelize([1, 2, 3, 4, 5])
2  squared_rdd = rdd.map(lambda x: x**2)
3  print(squared_rdd)
```

PythonRDD[2] at RDD at PythonRDD.scala:59

Command took 0.22 seconds -- by deravkumar@gmail.com at 2/6/2024, 7:29:32 AM on My Cluster 2

Cmd 5

```
1  rdd = sc.parallelize([1, 2, 3, 4, 5])
2  filtered_rdd = rdd.filter(lambda x: x % 2 == 0)
3  filtered_rdd
```

PythonRDD[4] at RDD at PythonRDD.scala:59

Type here to search

11°C Partly sunny 07:34 06-02-2024