ASSESSMENT - 18 | PySpark - 03

*“PYSPARK COMMANDS & FUNCTIONS”*



Submitted By

VAIBHAV PATIDAR

IPS ACADEMY, INDORE (M.P.)

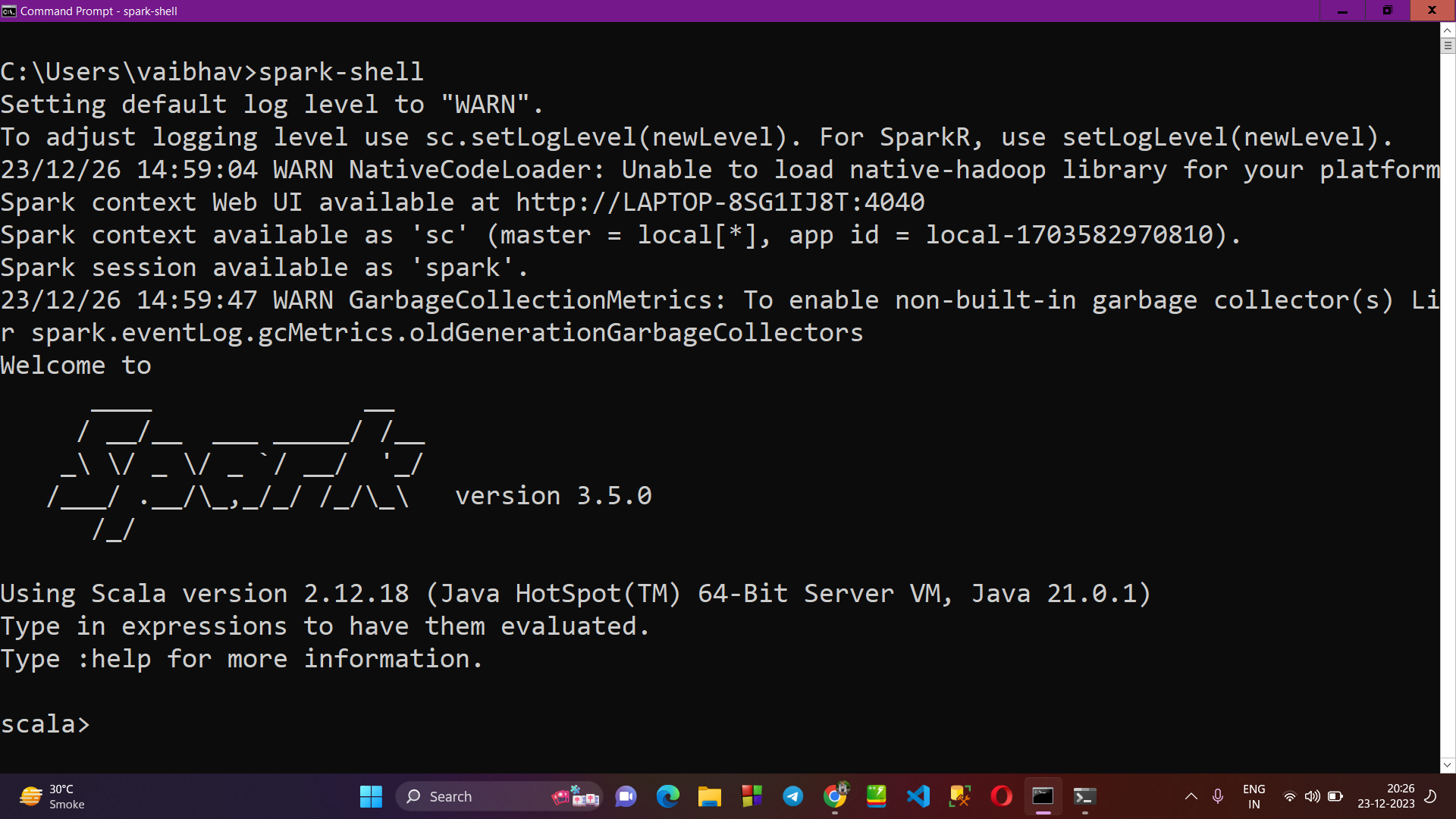
**Date :** 23-12-2023 | **Day 18**

**Week 3** | **Day - 06**

**→ Basic Spark Commands :**

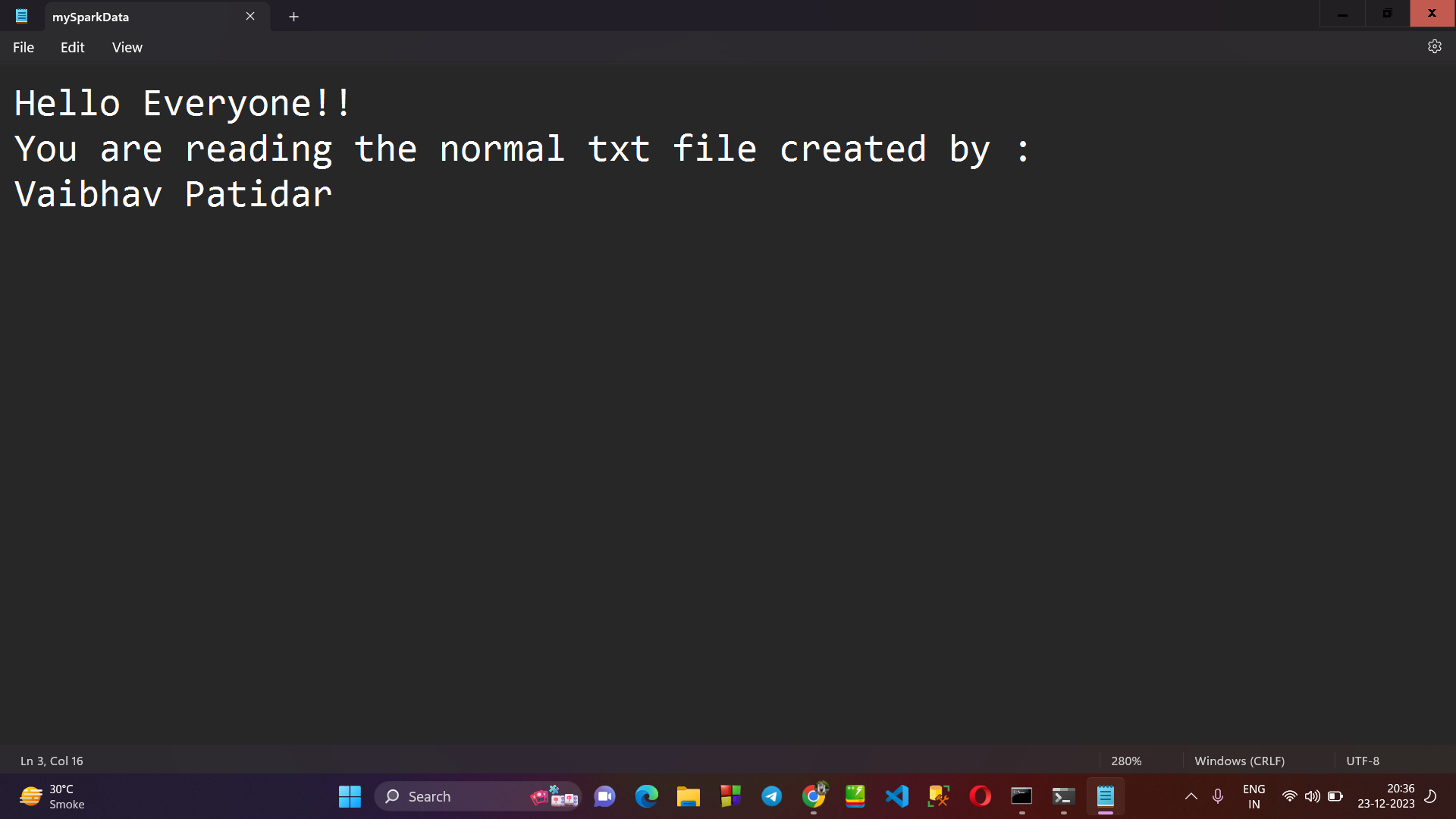
1. **Start the spark shell :**

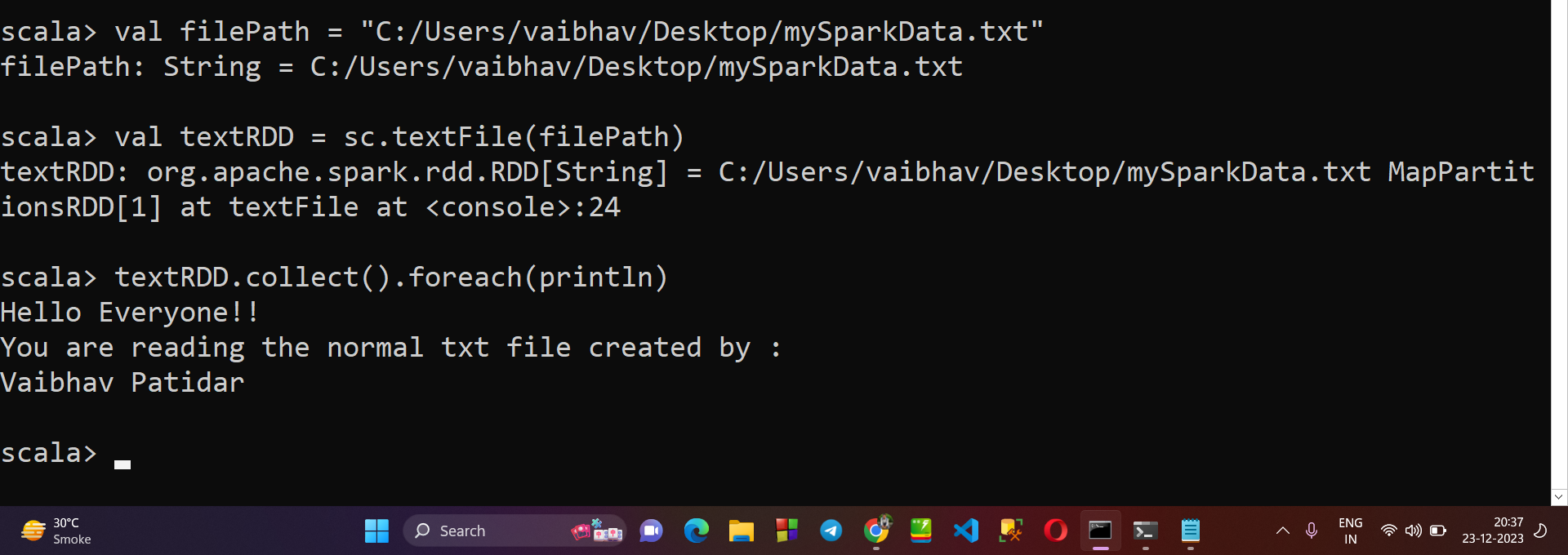
* For starting a spark shell open command prompt and type ***spark-shell*** and press ***enter***.

****

1. **Read the data of a file :**

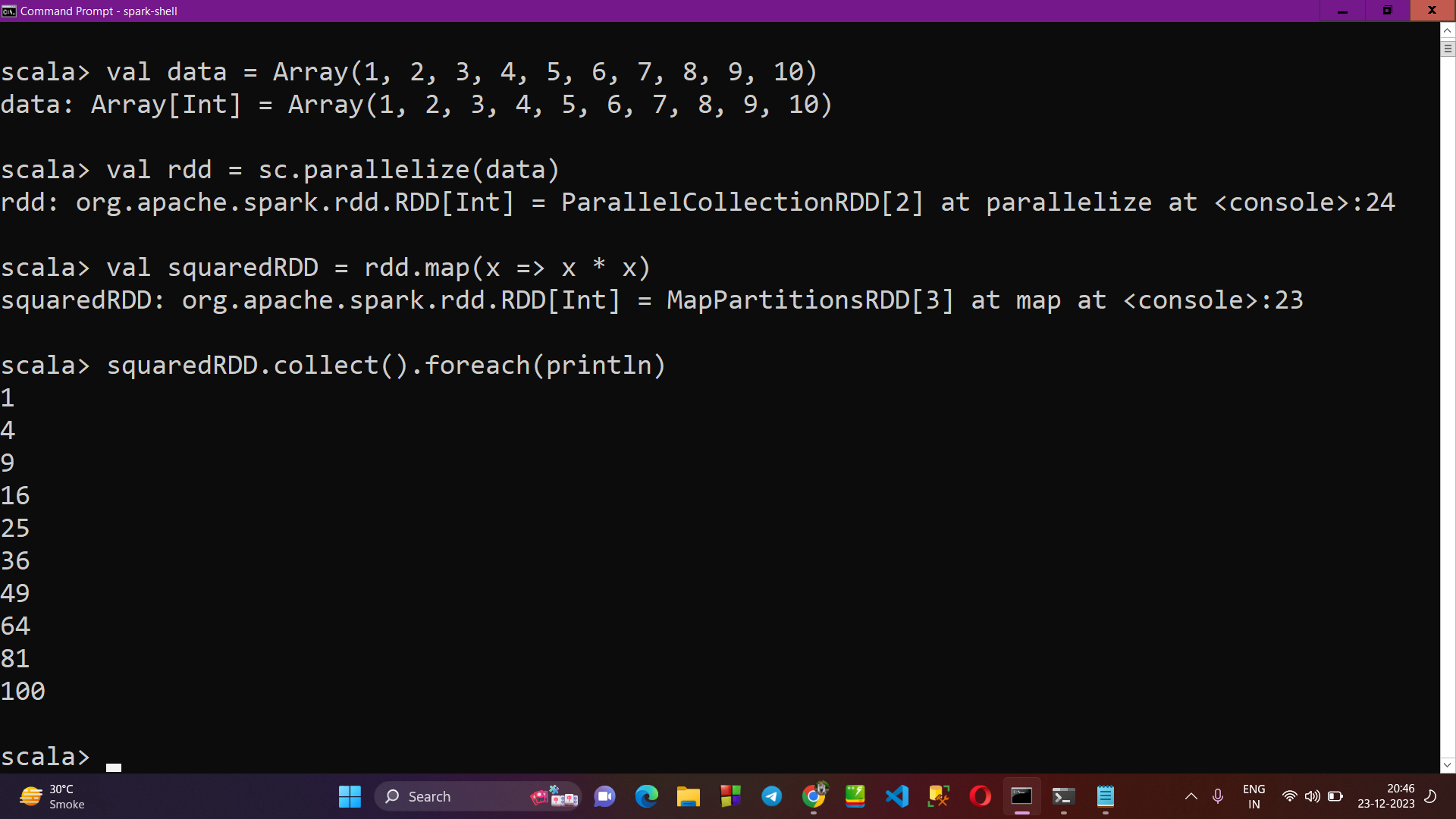
* First create a txt file , add some content in it and save it anywhere.
* Now copy the path of that file, and come to command prompt
* Now set the path of file as ***val filePath = “ …. Path …. “***
* Then read the text file into an RDD (Resilient Distributed Dataset).
* At last collect and print the contents of the RDD.





1. **Creating a RDD through Parallelizing :**

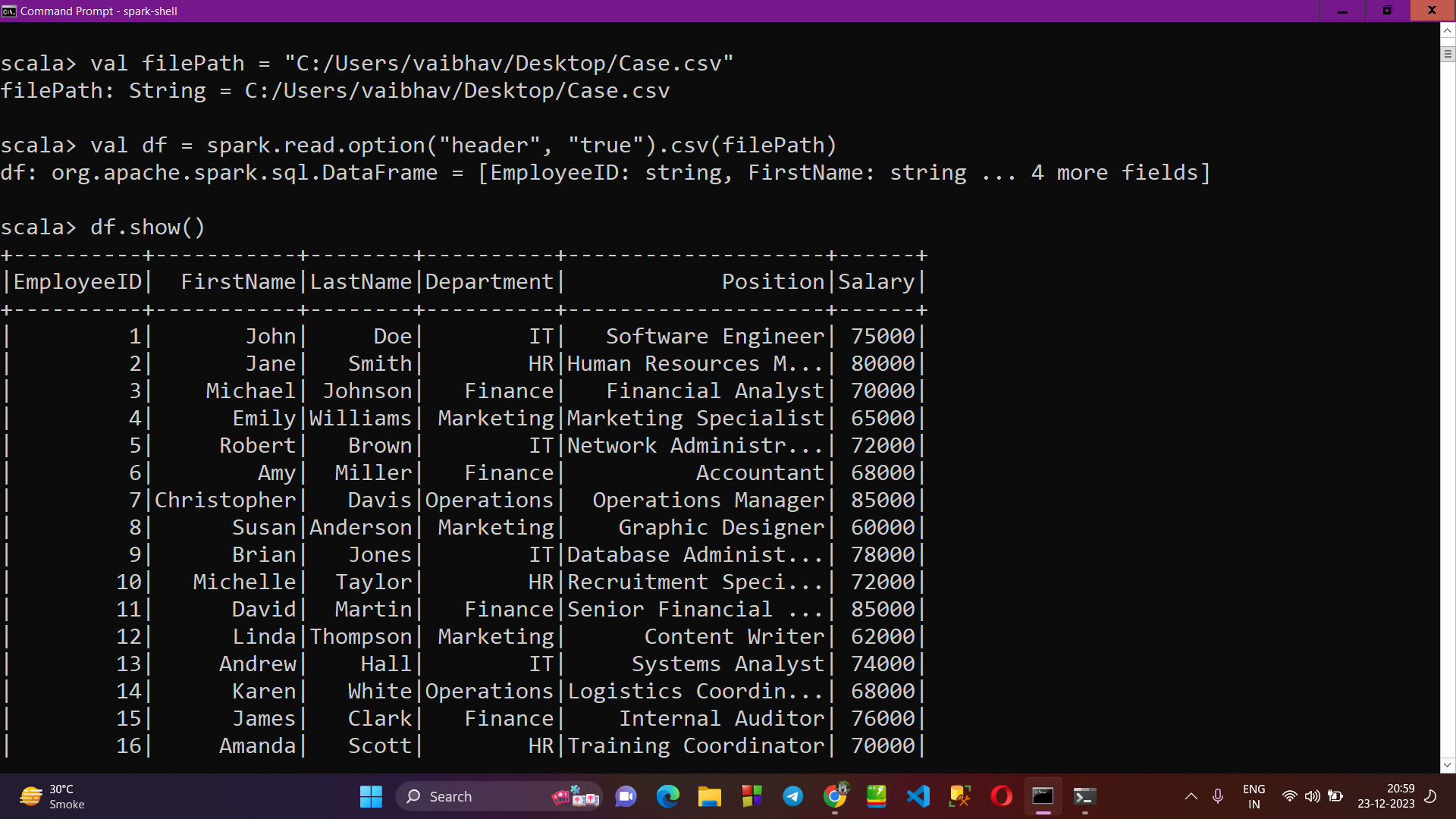
* Lets say we create a RDD for an array, so first create an array.
* Then by passin it into sc.parallelize(arr), a RDD is created.
* Now we further creat a RDD for squares of all elements of array.
* Then at last, we print the the array having squares of each element of the main array.



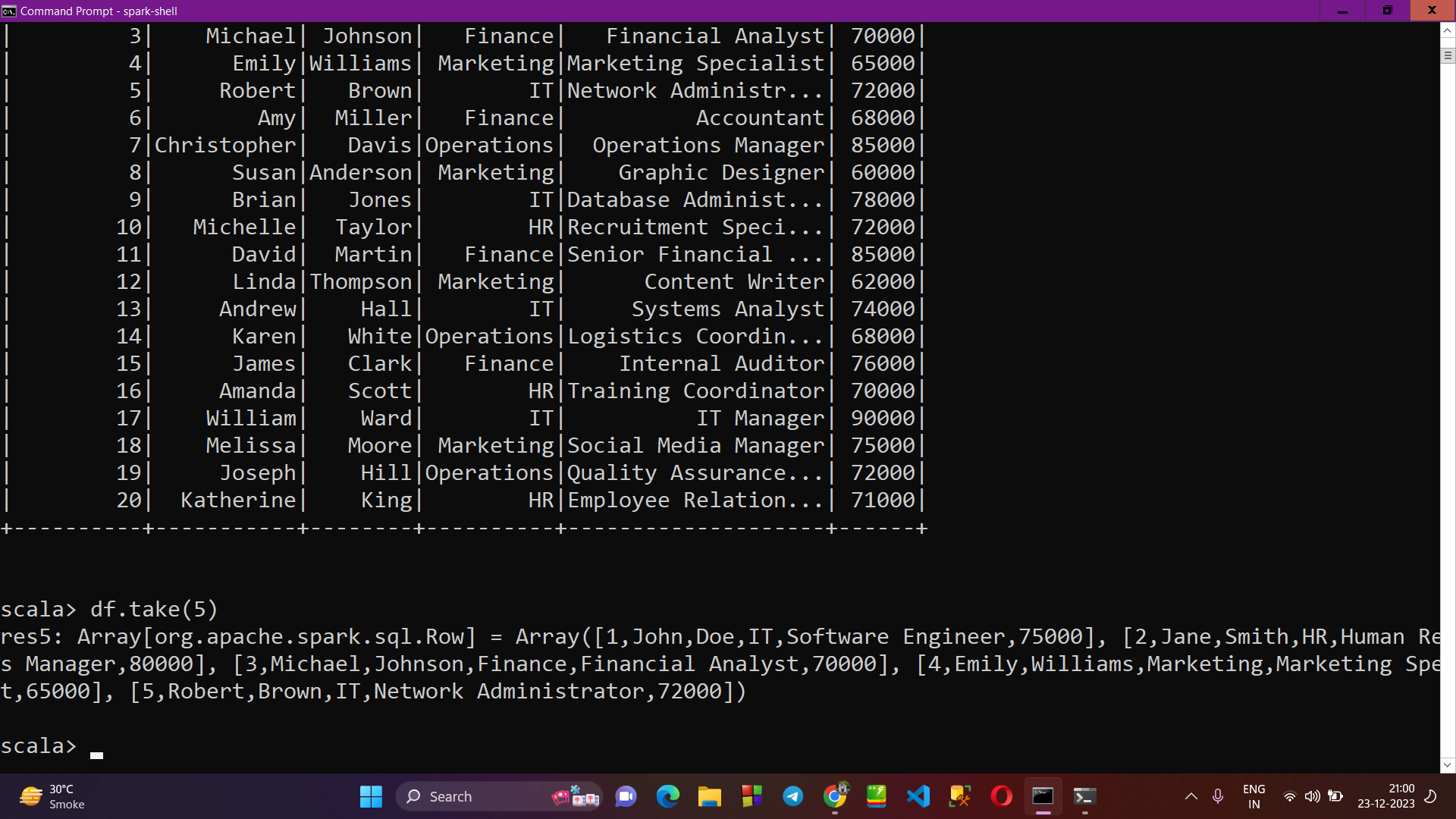
1. **Reading the data of a CSV File in spark-shell :**

* First create a CSV File and adds data into it, now save the file.
* Copy the path of the file, and open command prompt.
* Now type ***val filePath = “ … path … “*** and press enter.
* Now using spark.read.options() to read the data.



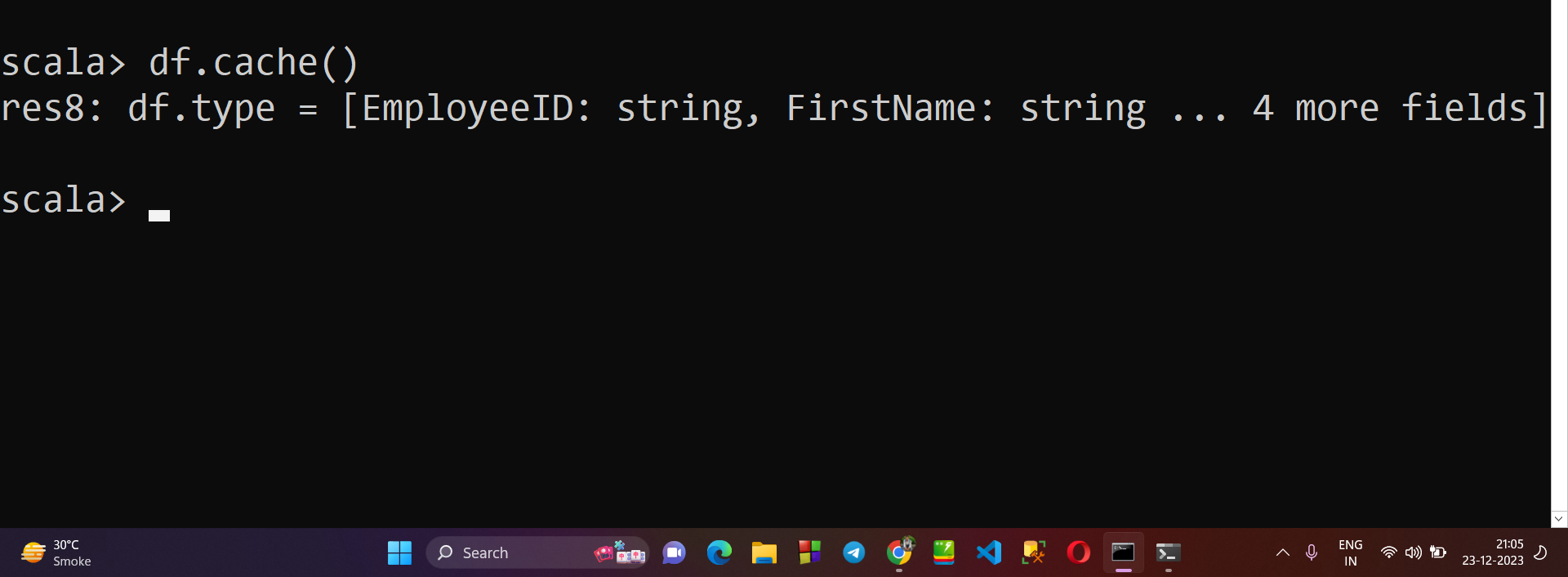


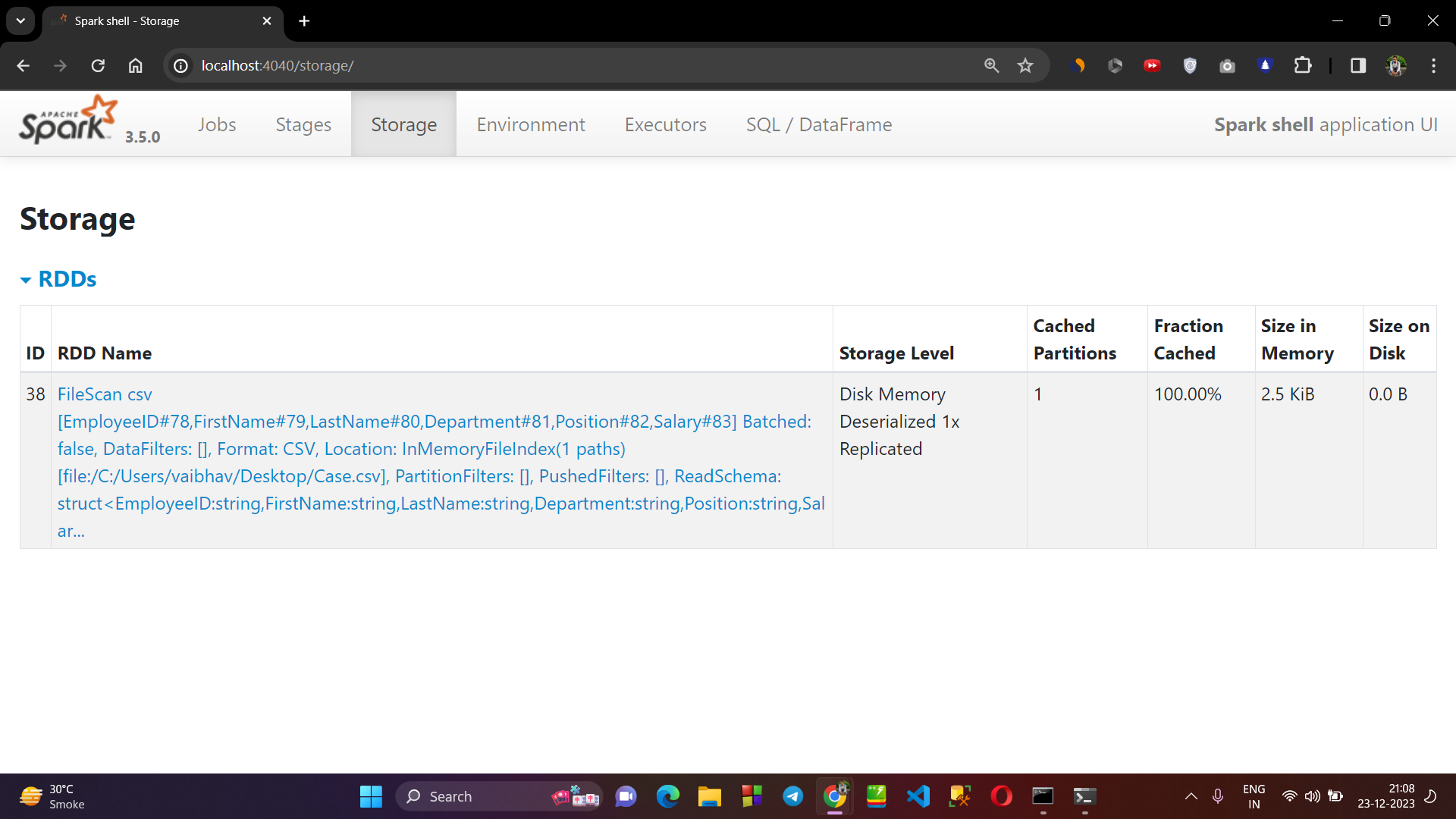
1. **Using Data.take() to print only 5 Enteries :**

****

1. **Using .cache() to cache the data :**

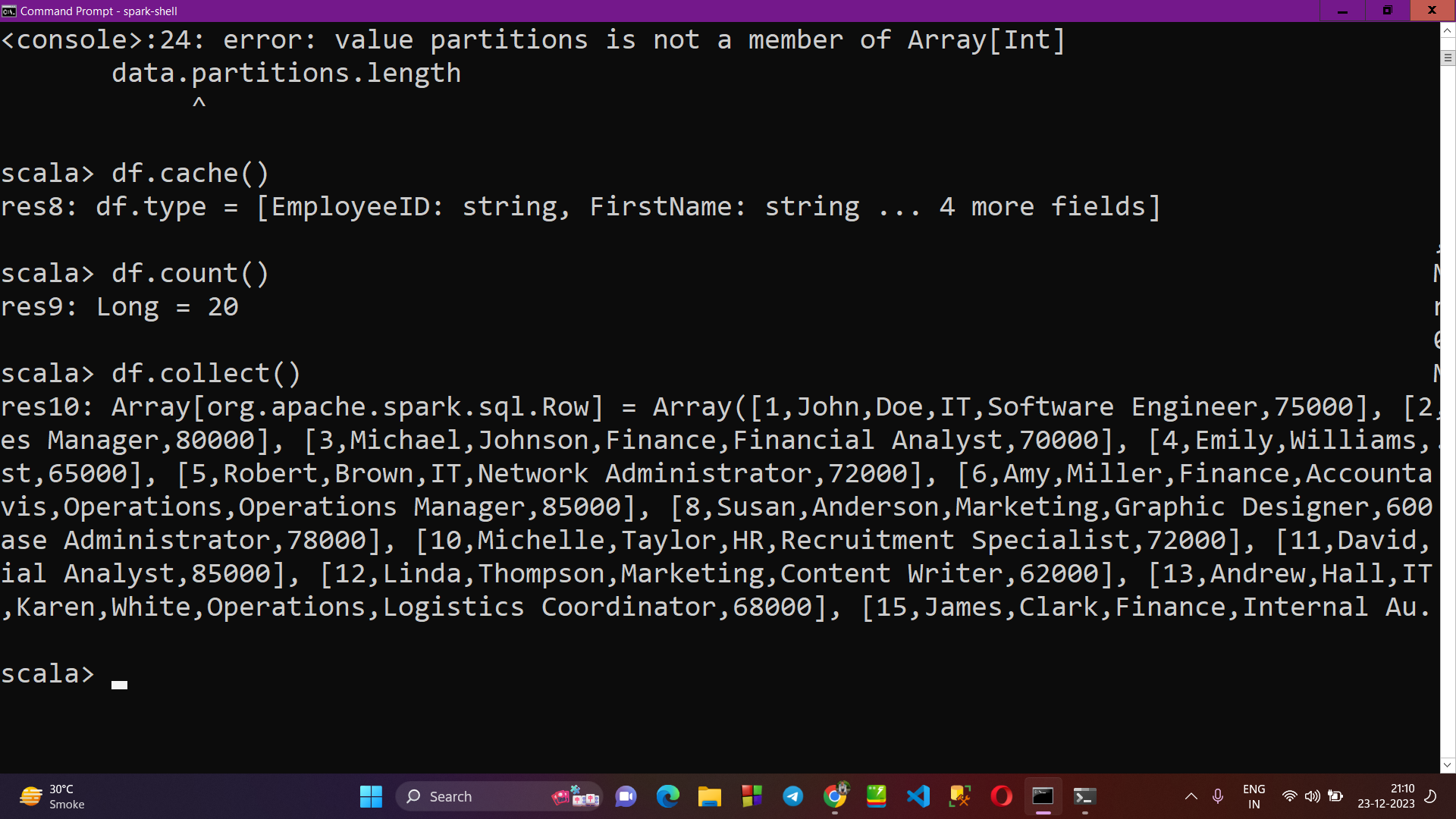
* All the data which is cached is store in storage section, which can be accessed by visiting the url : <http://localhost:4040/storage/>

****

****

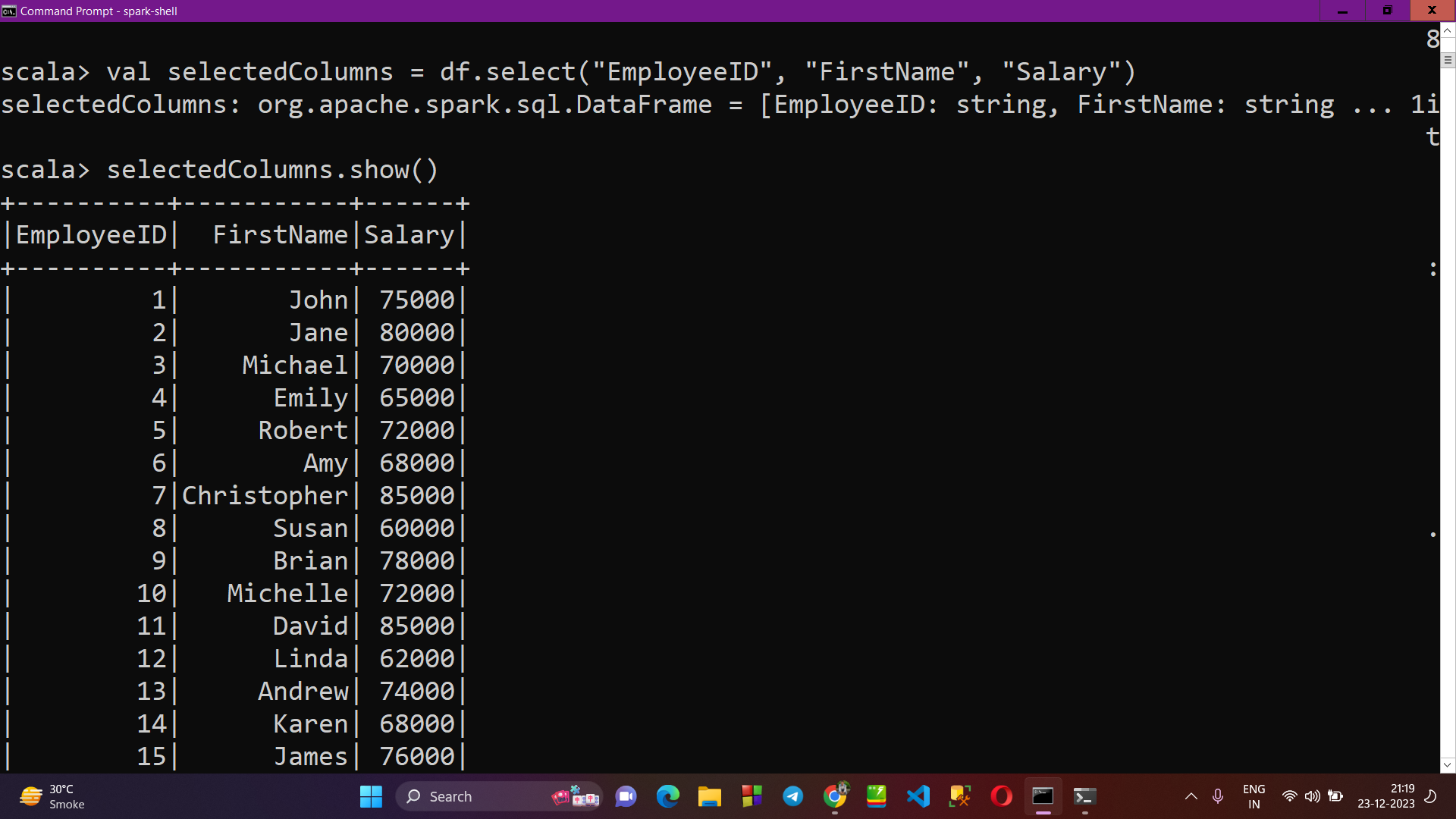
1. **Using .Count() and .Collect() :**

* .count() method is used to display the number of rows in the table or file.
* .collect() method is used to display the data of file.



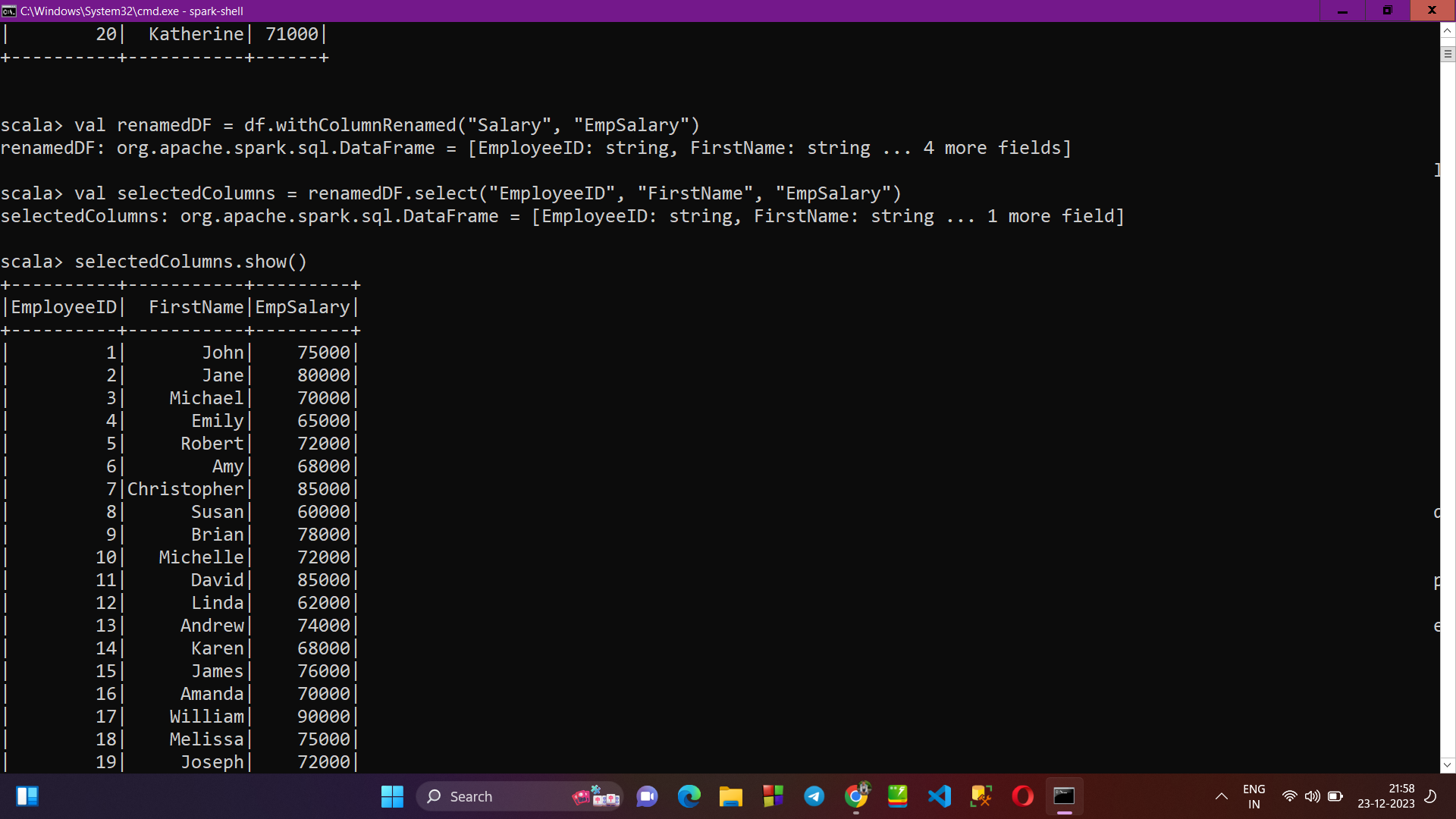
1. **Printing only selected Columns of a CSV file in spark shell :**

* Firstly, we have displayed the complete data of the Case.csv file, having columns as → EmployeeID, FirstName, LastName, Department, Position, Salary.
* But, now we will display the data of only EmployeeId, FirstName and salary columns.

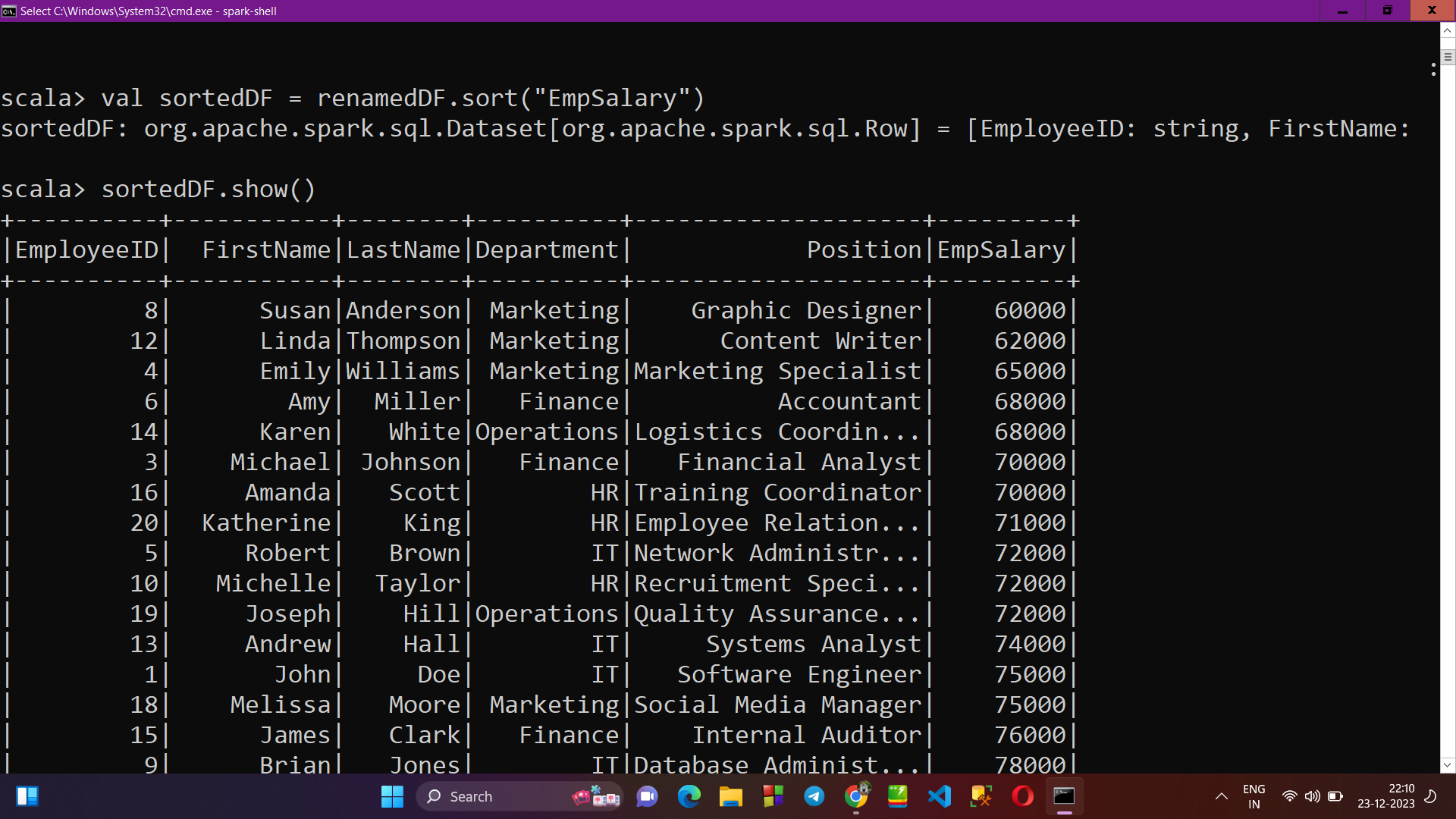


1. **Renaming the column name of table :**

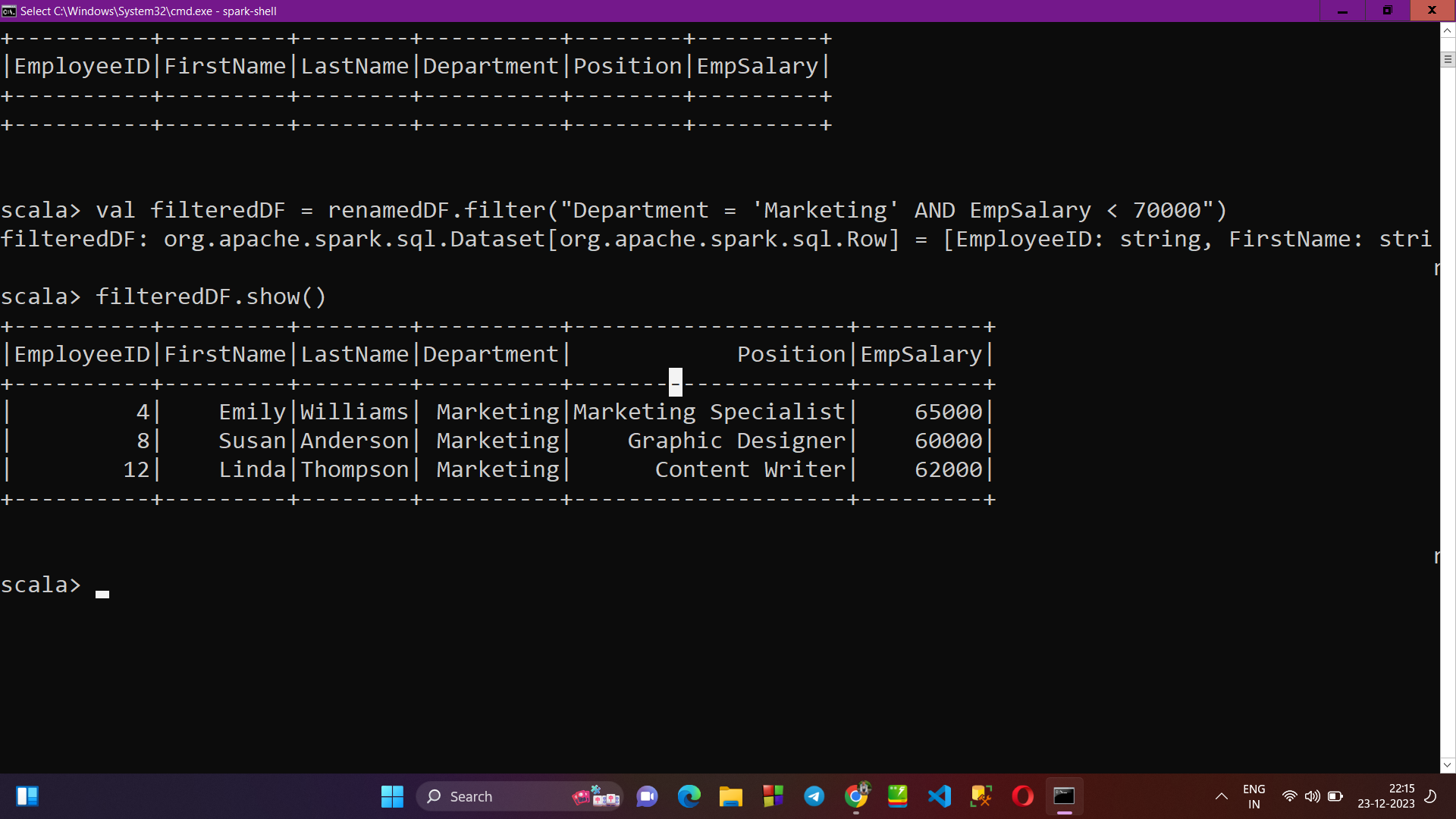
* Now, we’ll rename the salary column of csv table to EmpSalary.



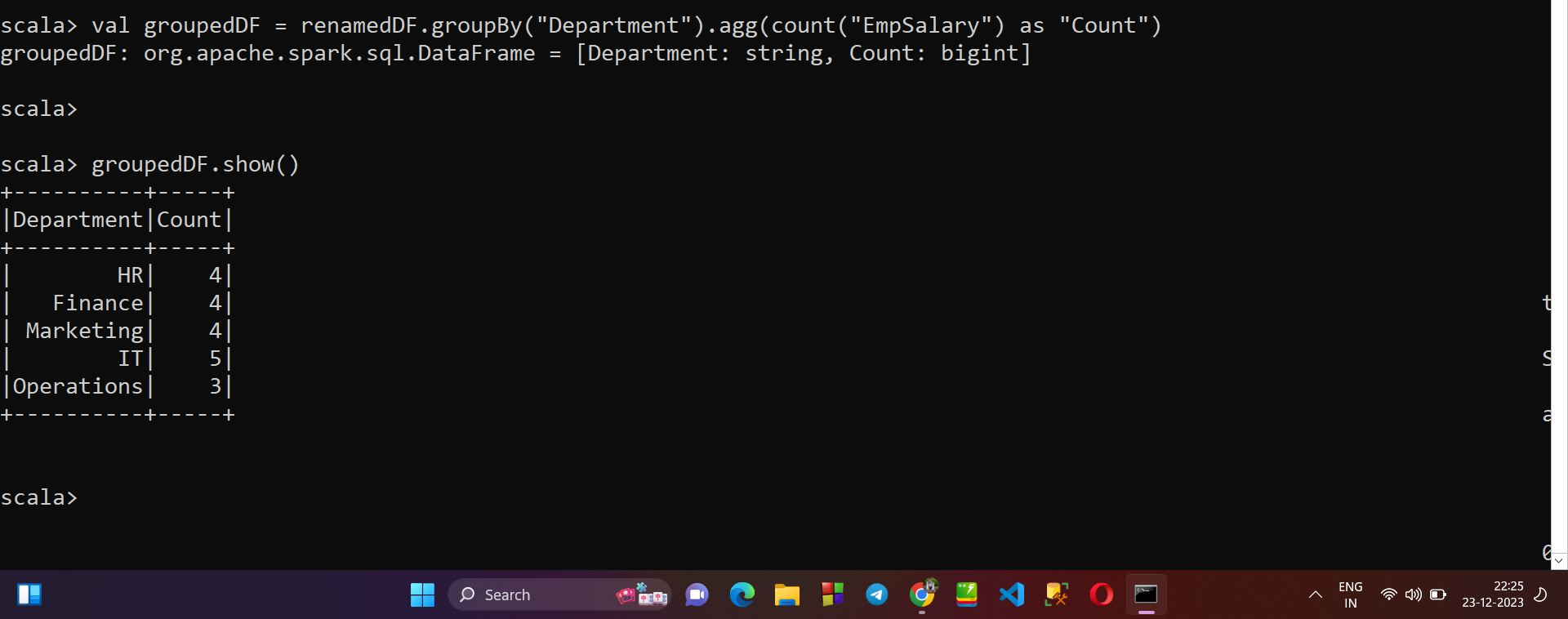
1. **SORTING the csv file data based on salary :**

****

1. **FILTER METHOD : displaying the employee details of employee from “Marketing” department having salary less than 70000 :**

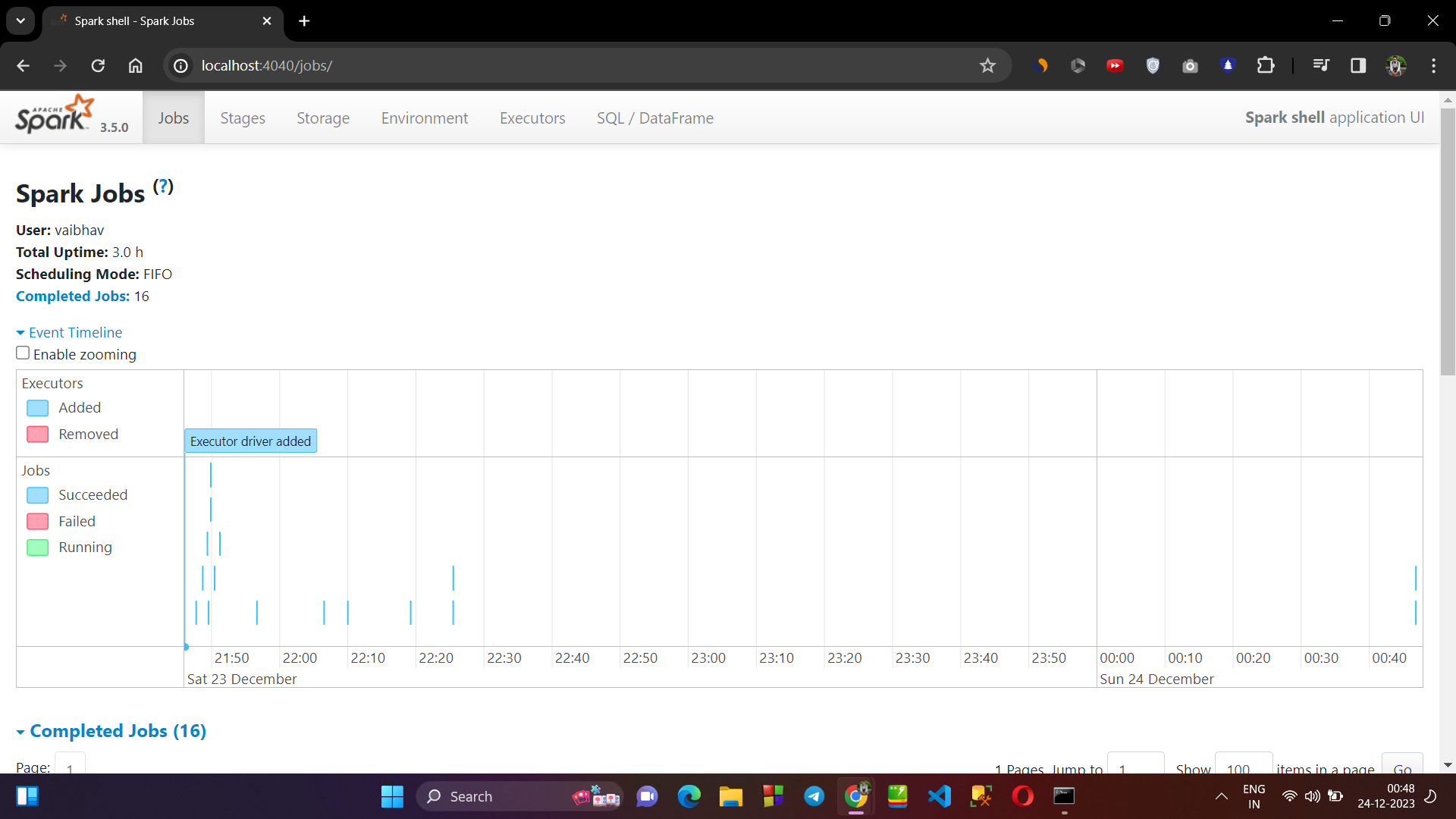
****

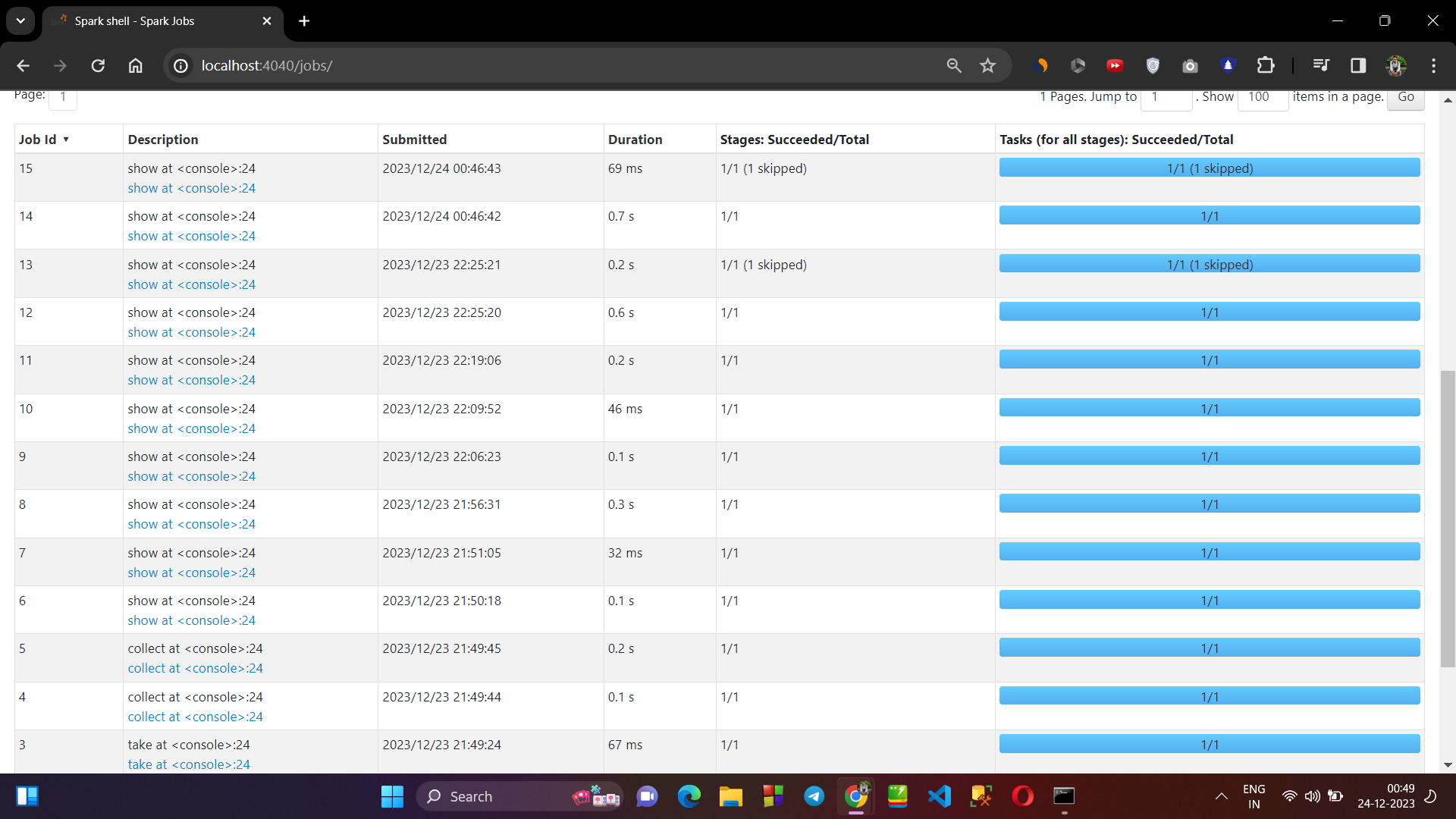
1. **GROUP BY Method : Grouping the data of table based on department :**

****

**→ SPARK WEB UI SCREENSHOTS AFTER EXECUTION:**

1. **JOBS :**

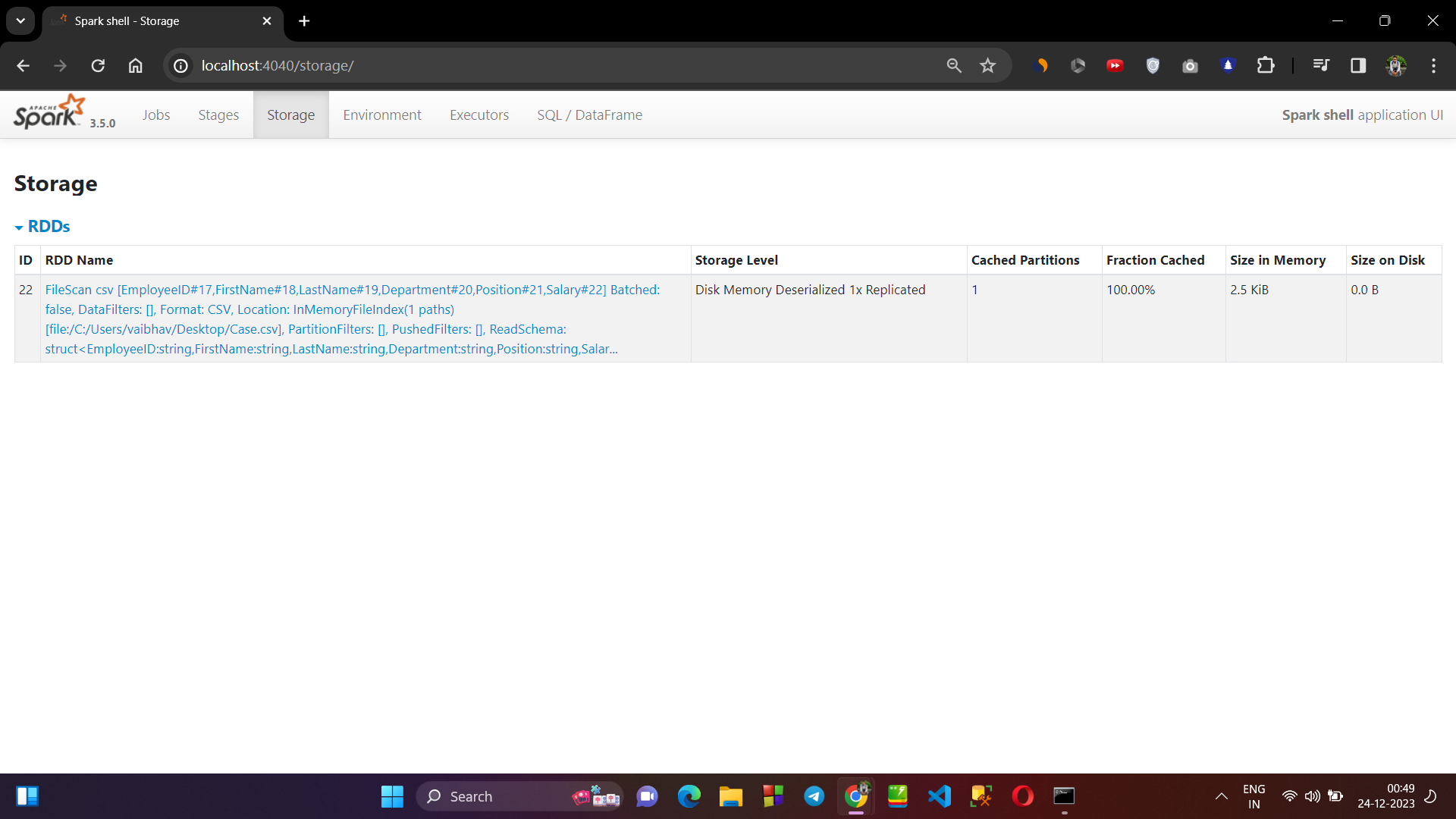
****

****

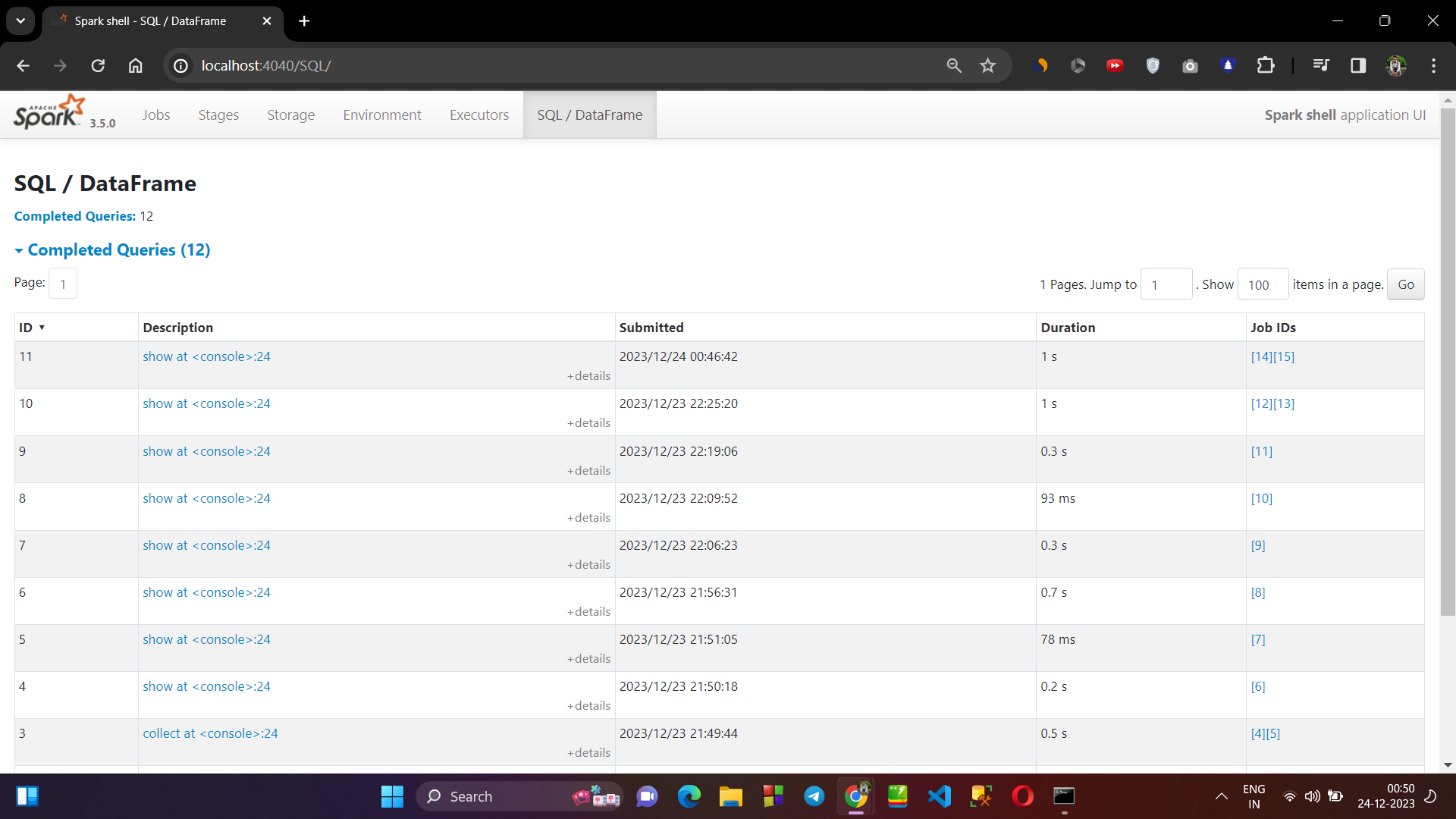
1. **STAGES :**

****

1. **STORAGE :**

****

1. **SQL DATAFRAMES :**

****