Object Storage: Minio, PySpark Programming

1. Connect to the Minio client. Create an alias to your Minio server, named **ms**. Create a bucket **labd**. Inside the bucket, create folders **iplookup** and **logs**. Copy the three log files from **/datasets/clickstream** to the Minio **logs** folder. Copy **iplookup.json** to the **iplookup** folder.

When you are finished you should have the following structure. Provide a list of commands necessary to complete this task. And include screenshots to show the files are there.

ms

|

|\_\_ labd

|

|\_\_ iplookup

| |

| |\_\_ iplookup.json

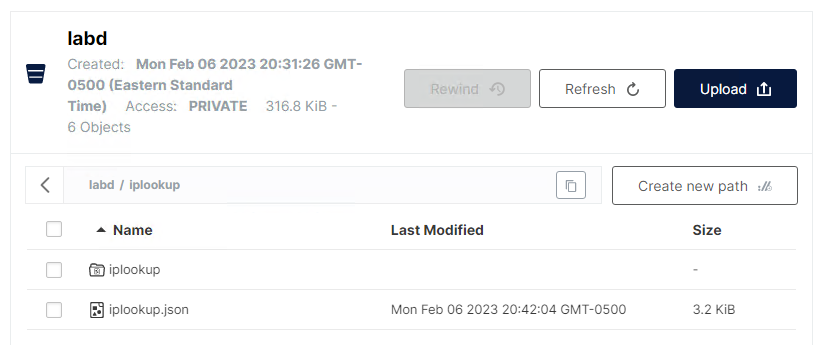
|

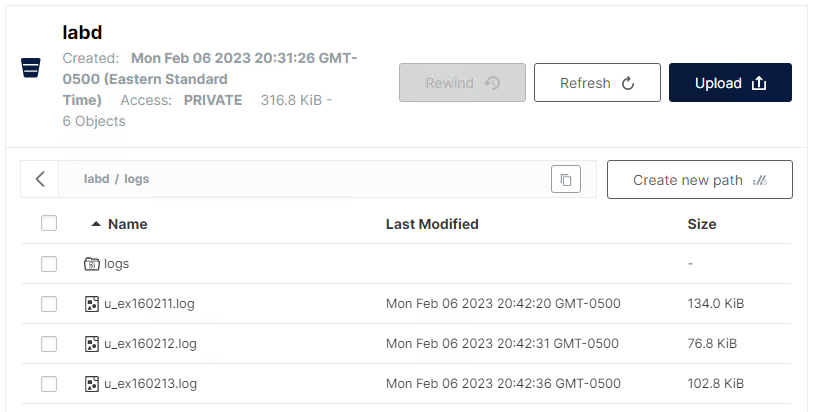
|\_\_logs

|

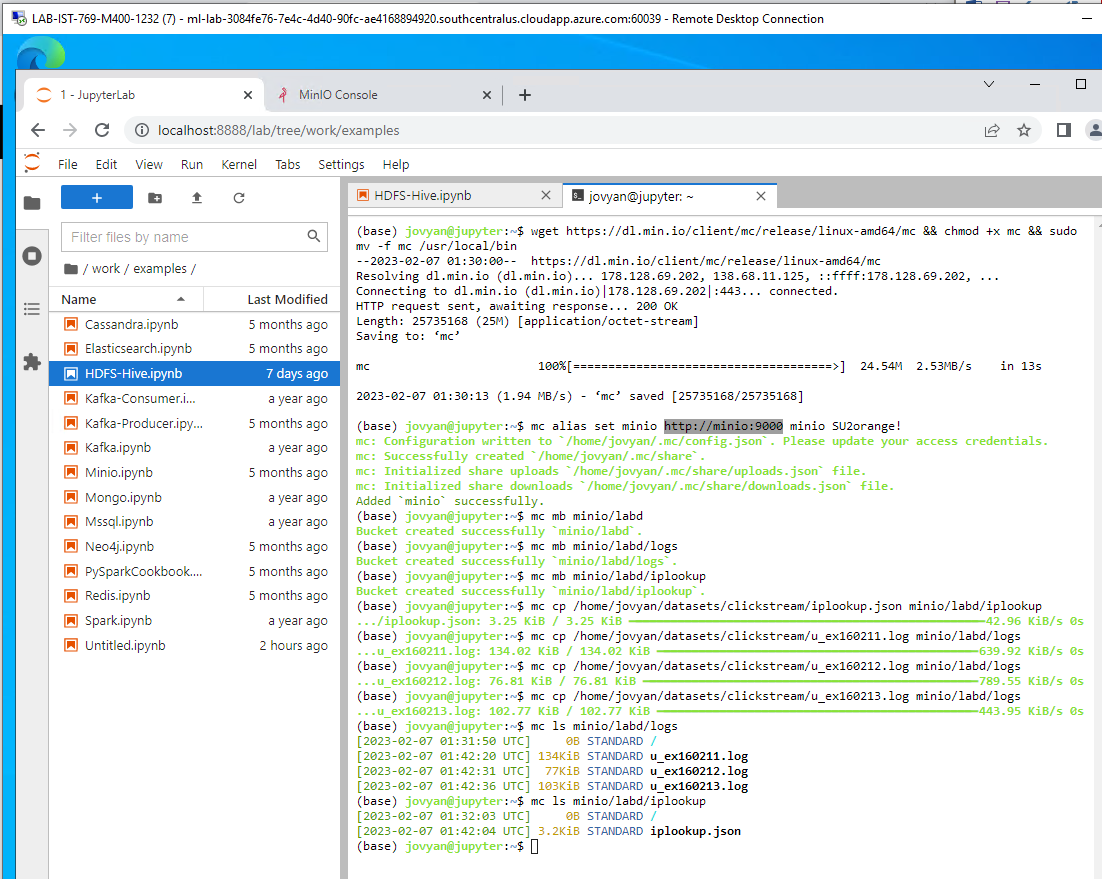
|\_\_ u\_ex160211.log, u\_ex160212.log, u\_ex160213.log

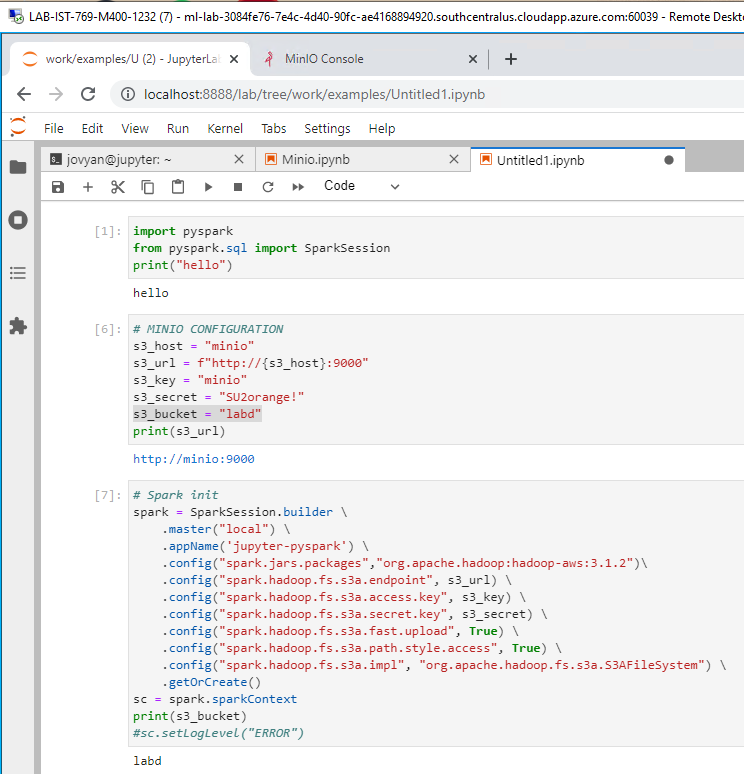
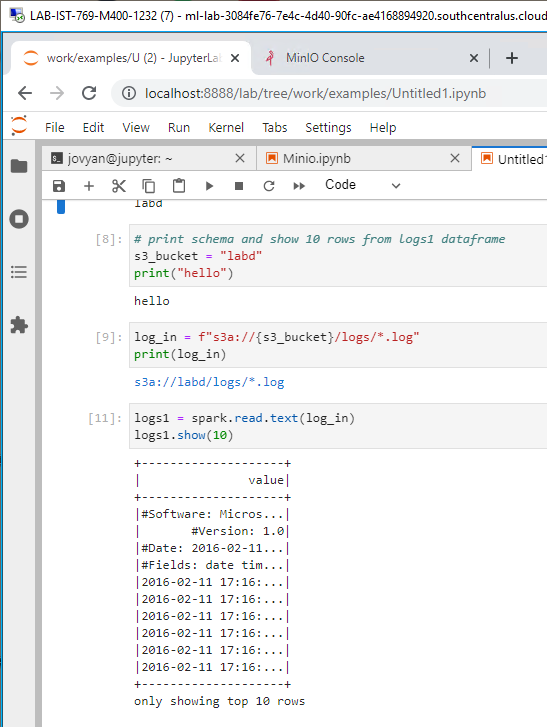
Files:

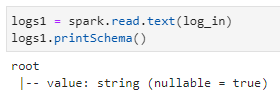


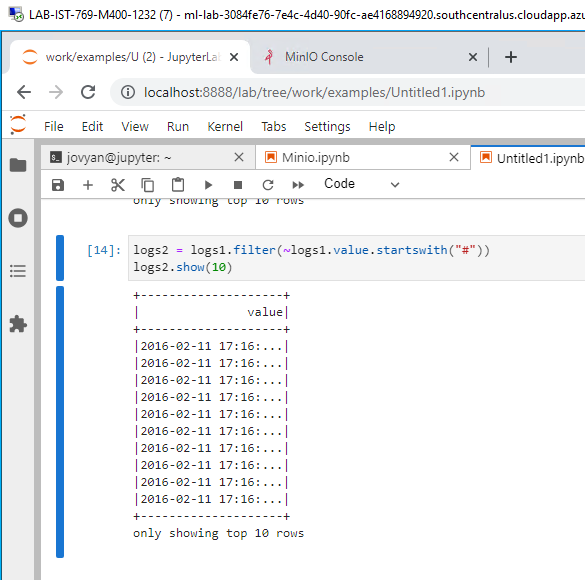
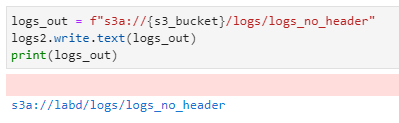


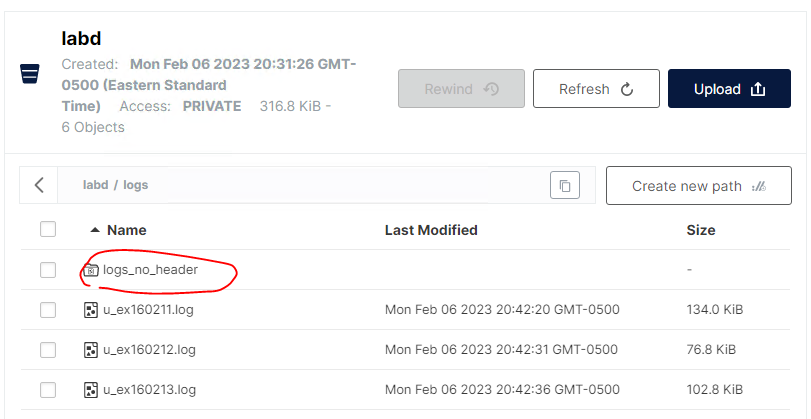
Commands:



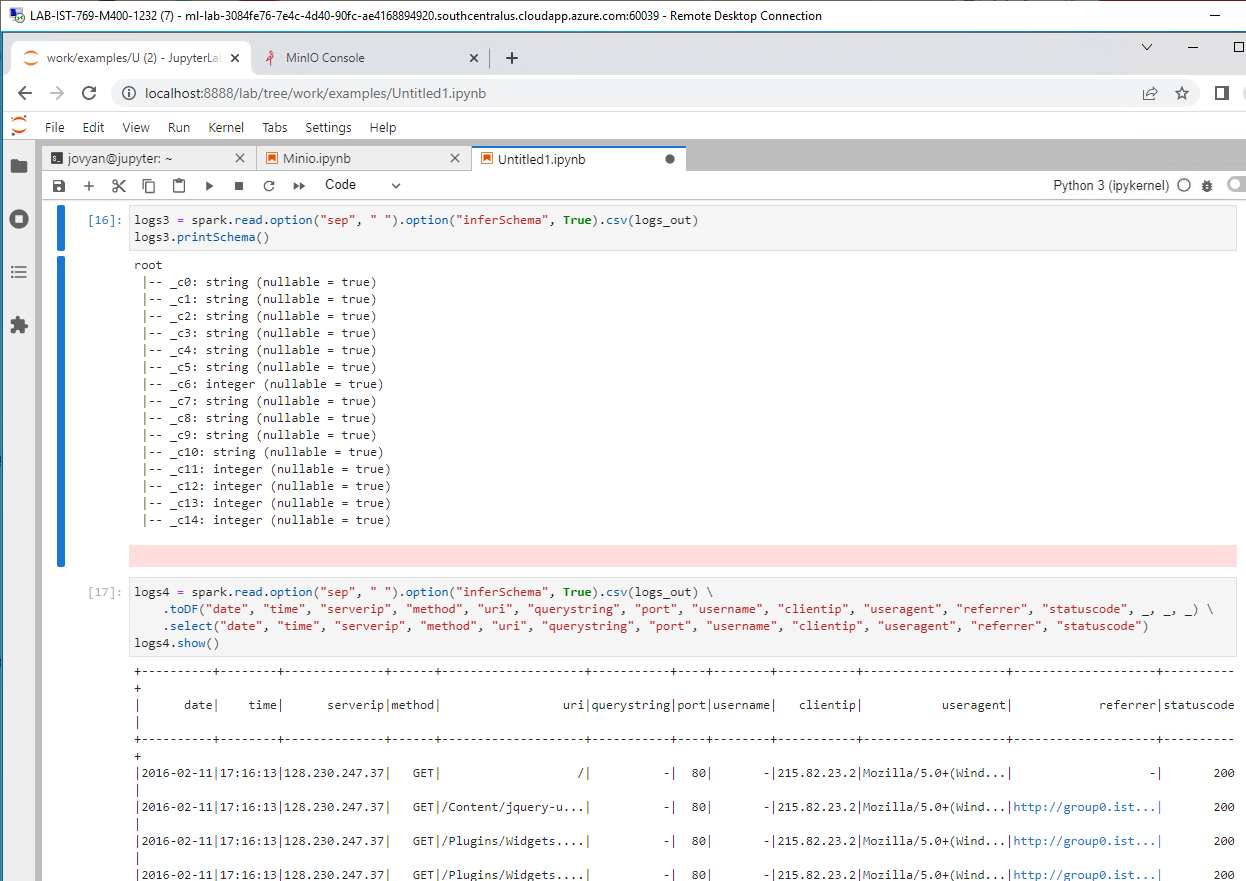
1. Create a new Spark notebook called **labd.ipynb**. Write (or copy and edit) Spark code to set up the Spark session. Make sure your Spark session supports Minio access and include the **hadoop-aws** Spark Jar package. Provide a screenshot of your code and the output.   
   NOTE: You do not need Hive support.  
   
2. Write Spark code to load logs from Minio **labd/logs** into a dataframe **logs1** using **spark.read.text**. Print the schema and show 10 rows from the DataFrame. Screenshot the code and output.   
   



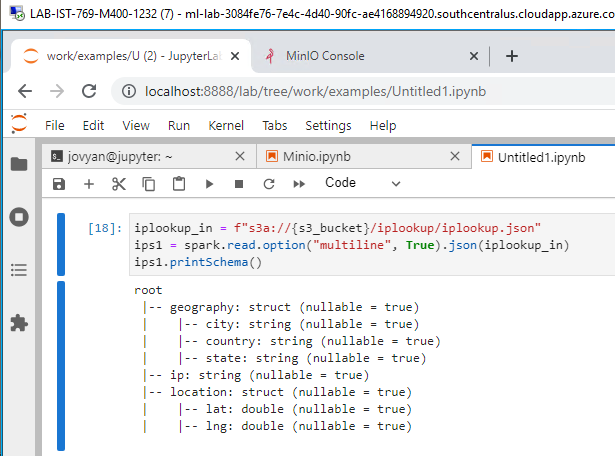
1. We need to remove the rows with **#** in front of them, as these are comments in the web server log files. Use the **filter()** function to do this, and save the results into DataFrame **logs2**. Show the code and output in your screenshot.  
   
2. Write back your **logs2** to Minio. Use the **text** format and call the file **logs-no-header**.Include a screenshot of the code and output.  
   

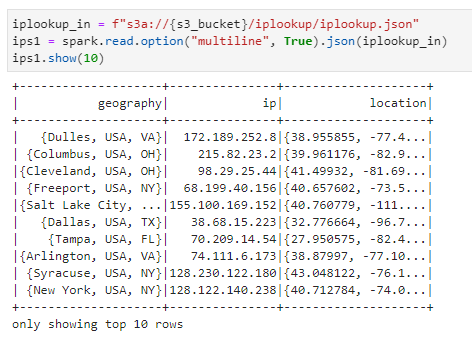


1. Read in the **logs-no-header**, this time using **csv** to delimit on a space into **logs3**. Add headers (date, time, serverip, method, uri, querystring, port, username, clientip, useragent, referrer, statuscode), and provide an output of the schema and the first couple of rows from the dataframe itself in the screenshot.

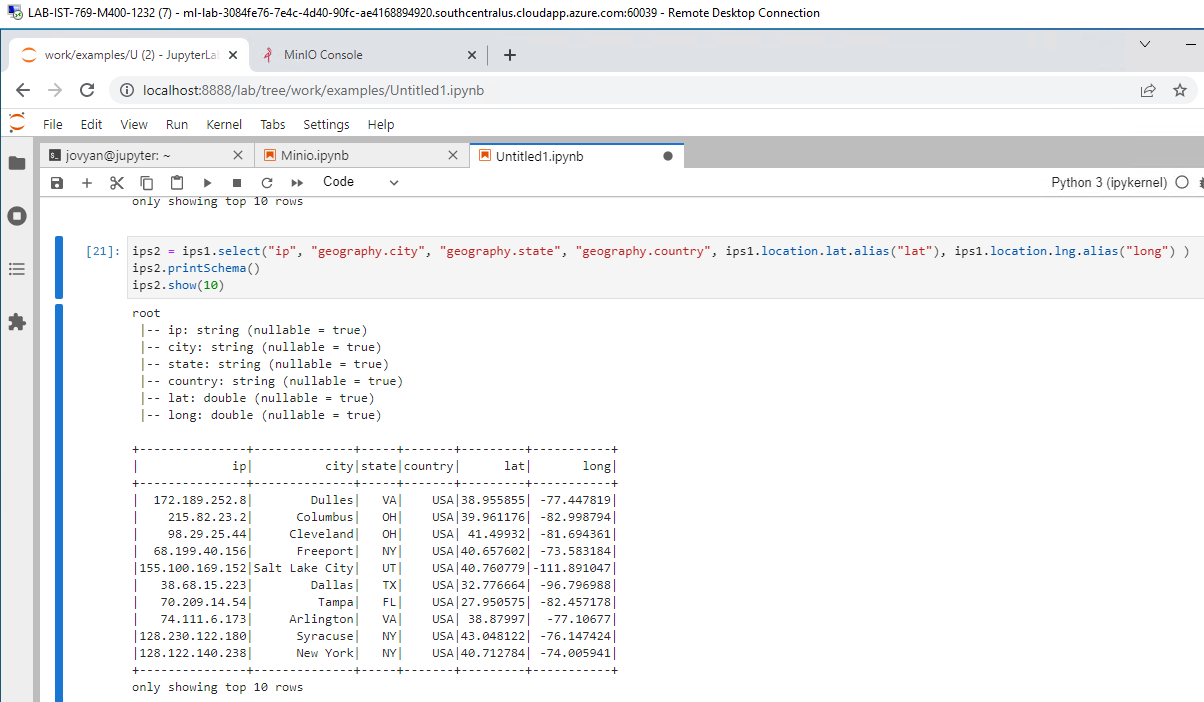


1. Let's handle the IP address lookup data. Write Spark code to load the **iplookup.json** file from Minio into the DataFrame **ips1**. Show the first 10 rows and print the schema for a screenshot to include code and output.

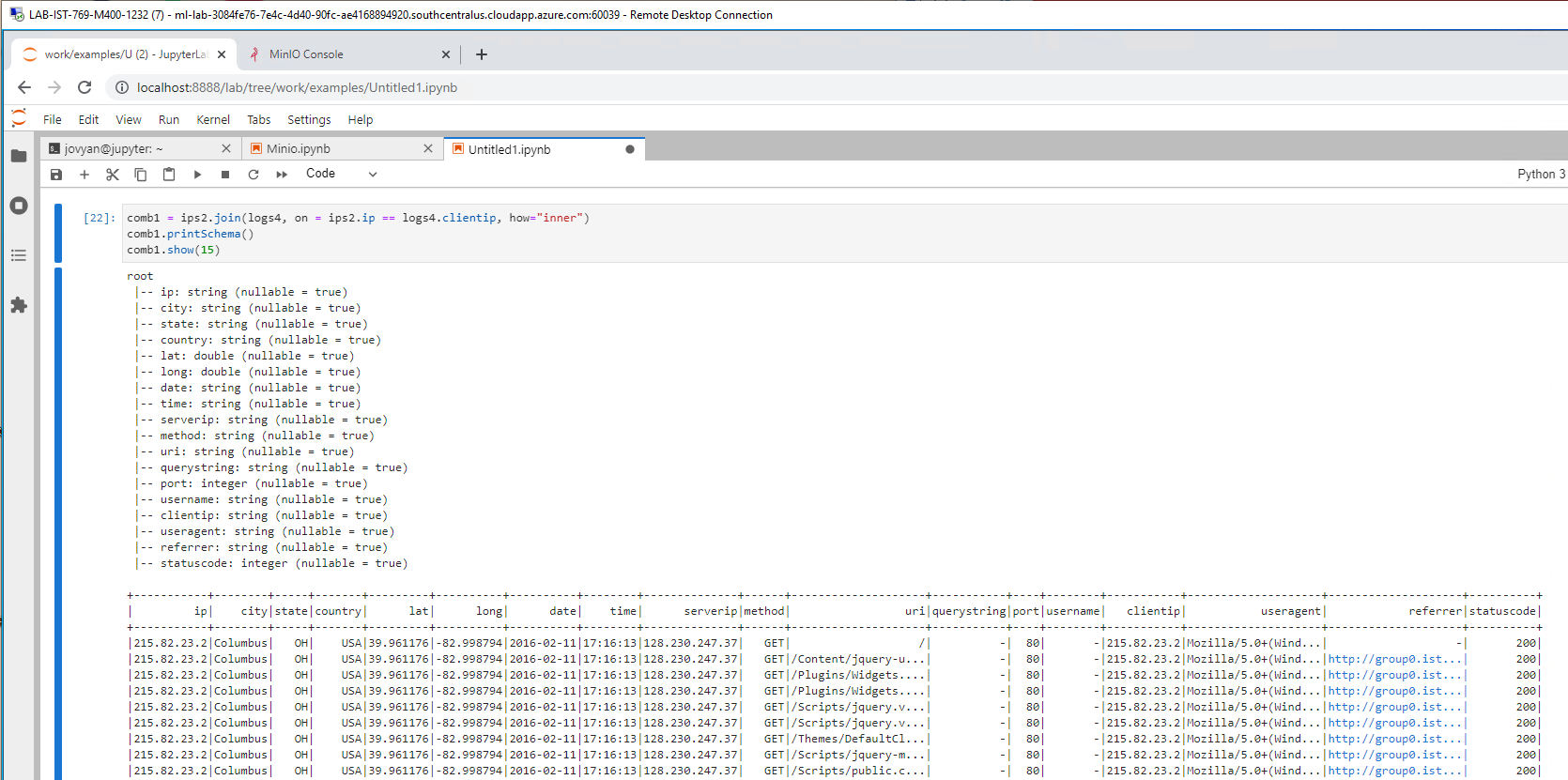




1. We need to flatten the nested JSON data. Use the **select()** function with dot notation to do this, saving the DataFrame as **ips2**. Provide a screenshot of the schema and output of the first few rows.



1. Now join the two DataFrames together on their business key, making the new DataFrame **comb1**. Provide a schema and sample of the first few rows in your screenshot.



1. Write the **comb1** DataFrame in **parquet** format back to Minio in the folder **cleaned-logs**. Again, show evidence the code ran and the file was created.

