Lab 4: Reading a JSON File from DBFS and Writing It as CSV

Author: Dr. Sandeep Kumar Sharma

Learning Objective

In this lab, you will learn how to read a JSON file stored in DBFS and write it back as a CSV file using Apache Spark.

Learning Outcome

By the end of this lab, you will be able to: - Read JSON files from DBFS - Convert the JSON data into a structured DataFrame - Write the DataFrame into CSV format - Understand Spark's JSON parsing and CSV writing behavior

---

Lab Information

We are now working with the JSON file created in **Lab 3**. The file is stored at:

```
/FileStore/tables/employee_output_json
```

This directory contains the part JSON files generated by Spark.

Your task is to read those JSON files and write the final output in CSV format.

Step-by-Step Instructions

Step 1: Set the JSON Input Path

Define the DBFS path where the JSON output from Lab 3 is stored.

```
input_json_path = "/FileStore/tables/employee_output_json"
```

Step 2: Read the JSON File

Spark automatically reads all JSON part-files inside the directory.

```
df_json = spark.read.json(input_json_path)
display(df_json)
```

Step 3: Set the Output Path for CSV

Specify the DBFS folder for the CSV output.

```
output_csv_path = "/FileStore/tables/employee_json_to_csv_output"
```

Step 4: Write the DataFrame as CSV

Use Spark to write the JSON DataFrame as CSV.

```
df_json.write.csv(output_csv_path)
```

Once completed, Spark creates a new DBFS directory containing CSV part-files.

Explanation

When reading JSON, Spark automatically detects the structure and converts it into a tabular DataFrame. This makes transforming and exporting JSON very easy.

When writing the output as CSV, Spark produces a set of distributed CSV part-files, following its standard output architecture.

This JSON → DataFrame → CSV workflow is extremely common when working with API outputs, web logs, semi-structured data, or systems that produce nested JSON.

End of Lab 4

You have successfully learned how to read JSON files from DBFS and write them into CSV format. Next, we can continue creating advanced labs such as: - Writing output in a single CSV file - Reading and writing Parquet - Reading and writing Delta tables - Working with schema inference and schema definition Just tell me the next lab you want to generate!