

Lab 1: Reading a CSV File from DBFS and Displaying It

Author: Dr. Sandeep Kumar Sharma

Learning Objective

In this lab, you will learn how to read a CSV file stored in DBFS using Apache Spark in Azure Databricks, load it into a DataFrame, and display the data.

Learning Outcome

By the end of this lab, you will be able to:

- Access a file stored in DBFS
- Read a CSV file using Spark (Python)
- Understand how Spark loads data into DataFrames
- Display the DataFrame content inside a Databricks notebook

Lab Information

Imagine we have a CSV file named **employee.csv** uploaded to DBFS at the following path:

```
/dbfs/FileStore/tables/employee.csv
```

Your task in this lab is to read this file and display its content using Spark.

Step-by-Step Instructions

Step 1: Set the File Path

First, inside your Databricks notebook, define the DBFS path where the CSV file is stored.

```
file_path = "/FileStore/tables/employee.csv"
```

Step 2: Read the CSV File into a Spark DataFrame

Now, use Spark's `read.csv()` function to load the file into a DataFrame. Enable header reading and schema inference.

```
df = spark.read.csv(file_path, header=True, inferSchema=True)
```

Step 3: Display the DataFrame

Use the Databricks `display()` function to show the CSV data in a tabular format.

```
display(df)
```

Explanation

When you execute these commands: - Spark connects to DBFS - Reads the CSV file - Loads the content into a distributed DataFrame - Databricks `display()` renders the output beautifully in table form

This is the simplest and most common way of exploring raw data inside Databricks.

End of Lab 1

You have successfully read a CSV file from DBFS and displayed its content using Spark. In the next lab, we will learn how to write this DataFrame back to DBFS in CSV format.