A close-up of a person smiling

Description automatically generated

# Lab 02: Extending Data Preparation with Python notebook

### 

### Introduction

In this lab, you will use a Python notebook to execute a resource expensive data preparation activity using Spark cluster managed by Fabric.

### 

### Objectives

After completing this lab, you will be better able to:

1. Import the various Python notebooks to the Fabric Environment
2. Run a first notebook to ingest data from the Bronze zone (CSV files) to DELTA tables
3. Run a second notebook to prepare a bigger sales dataset, merging 2 files (using SORT and MERGE join) to the DELTA table

**Estimated time to complete this lab**

60 minutes

Contents

[Lab 02: Extending Data Preparation with Python notebook 1](#_Toc159514380)

[Introduction 1](#_Toc159514381)

[Objectives 1](#_Toc159514382)

[Task 1: Import Python notebooks 3](#_Toc159514383)

[Task 2: Ingest data from Bronze zone (CSV files) to the Gold zone (Delta tables) 6](#_Toc159514384)

[Task 3: Load data from ADLS Gen 2, apply scalable transformation then load DELTA table 13](#_Toc159514385)

**Lab Prerequisites**

* Workspace: Fabric, Power Premium or Fabric trial
* Individual license: Power Pro or Premium Per User account

**Information provided by your training provider**

* Trial tenant (if applicable): login & password, workspace to use for the lab.
* Azure Data Lake Gen2 (containing data sources): account name & shared access signature.

### Task 1: Import Python notebooks

In this task, you will import 2 Python notebooks, to be used later for data preparation and ingestion.

* Choose the Data Science from the Microsoft Fabric menu and go to your Workspace
* Select the **Import > NotebooK** Menu to load the file Lab 02A - Load Tables from CSV.ipynb

A screenshot of a computer

Description automatically generated

* The imported Notebook should appear in the Workspace

A screenshot of a computer

Description automatically generated

* Repeat the same operation for the second notebook Lab 02B - Reading from Private Storage Account.ipynb
* The 2 notebooks should appear in the workspace

A screenshot of a computer

Description automatically generated

You will now configure Spark 3.3 as the default Runtime version when running notebooks.

This setting is accessible through the Workspace settings menu, on the right side of the UI.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* In the next Task, you will have to :
  + Configure the default Lakehouse for each notebook
  + Configure some settings in the code
  + Execute the Python code and understand the logic.

### Task 2: Ingest data from Bronze zone (CSV files) to the Gold zone (Delta tables)

In this task, you will use a Notebook to read data from CSV files stored in the Bronze zone (also named unmanaged zone) and ingest the content to the Gold zone using the DELTA format:

**A screenshot of a computer

Description automatically generated**

* For each CSV files, the script will detect the file structure, create the corresponding table structure, and ingest the file content to the table dynamically.
* From the Lab workspace, open the notebook named “Lab 02A – Load Tables from CSV”

A screenshot of a computer

Description automatically generated

* On the Explorer panel, select the Lakehouses item

**A screenshot of a computer

Description automatically generated**

* The imported Notebook contains a reference to a Lakehouse which does not exist in your environment, that is why the Missing Lakehouse warning appears. You will have to attach the Notebook to the Contoso Lakehouse created previously.

A screenshot of a computer

Description automatically generated

* Click on the Missing Lakehouse warning and select Add Lakehouse.

A screenshot of a search box

Description automatically generated

* Select Existing Lakehouse and click on Add

A screenshot of a computer

Description automatically generated

* On the OneLake data hub, select the Contoso Lakehouse from the Lab Workspace

A screenshot of a computer

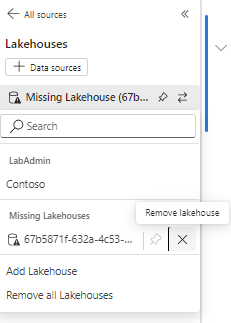
Description automatically generated

* The structure of the selected Lakehouse should now appear in the Notebook interface. Click on the Lakehouse selector.

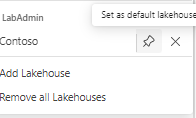
A screenshot of a computer

Description automatically generated

* Remove the missing Lakehouse reference



* Then select the Contoso Lakehouse and set it as the default Lakehouse for your notebook.



* As the default Lakehouse, the Contoso Lakehouse should now be pinned.

A screenshot of a computer

Description automatically generated

* Review the Python script content:
  + Spark session configuration
  + Create and Load tables from CSV files.

A screenshot of a computer

Description automatically generated

* Click on Run all to execute the full Python script

A screenshot of a computer

Description automatically generated

* At the end of the job execution, you can get more details about each Spark job involved during the code execution.

A screenshot of a computer

Description automatically generated

* Close the Opened notebook.

A close up of a computer screen

Description automatically generated

* In the Lakehouse explorer, expand the Tables node to reveal the created tables. Use the Refresh tables if necessary.

A screenshot of a computer

Description automatically generated

* By selecting one of the tables, you can see the table content.

A screenshot of a computer

Description automatically generated

* You can also see the underlying DELTA table structure to display the PARQUET file(s) dans the DELTA log file.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### Task 3: Load data from ADLS Gen 2, apply scalable transformation then load DELTA table

In this task, you'll be working with two files: Sales\_File1.csv and Sales\_File2.csv, each file contains 1M rows. These files are still stored in an external ADLS Gen 2 account, and the Spark Task will load data in a data frame.

**A close-up of a computer code

Description automatically generated**

Your task is to sort the files based on a specified set of columns and then perform a merge join (1 row on file 1 to be joined with 1 rows on file 1) using these columns.

Once you've completed the join, you'll need to write the results back to a Delta table in the Managed Lakehouse.

* From the lab workspace, open the Notebook Lab 02B - Reading from Private Storage Account.

A screenshot of a computer

Description automatically generated

* As you did during the previous task, use the Lakehouse selector to define the Contoso Lakehouse as the default one, and remove the wrong one.
* You also need to update the Python script to specify how to connect to the ADLS Gen 2 account (your trainer will share the information) :
  + Line 3 : the storage account name
  + Line 4 : the container
  + Line 9 : the SAS token

A screenshot of a computer

Description automatically generated

* Once the notebook is configured, execute each cell individually to understand the task performed :

A screenshot of a phone

Description automatically generated

* + Create 2 dataframes from the 2 CSV files (1M rows per file)
  + Show a subset of the tables, the structure and the number of rows
  + Sort each dataframes with the columns SalesOrderNumber and SalesOrderLIneNumber
  + Perform a Merge Join between the 2 files using the sorted columns
  + Drop the Sales table if it already exists in the Lakehouse
  + Load the Dataframe in a new table in the Lakehouse
  + Count the number of rows on the table: 1M
  + Display some rows on the load table

### A screenshot of a computer Description automatically generated

* From the Lab workspace, open the SQL analytics endpoint of the Contoso Lakehouse

A screenshot of a computer

Description automatically generated

* Create a new SQL query using the following snippet to display the number of rows on the fact tables, and the aggregated quantity per store ;

SELECT

[StoreName],Count(\*) as NbRows,SUM([SalesQuantity]) As SalesQuantity

FROM [Contoso].[dbo].[Sales] Sales

JOIN [Contoso].[dbo].[Stores] Stores ON Sales.StoreKey=Stores.StoreKey

GROUP BY [StoreName]

A screenshot of a computer

Description automatically generated

* Use the **Save as View option** to keep the query

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Make sure the newly created SQL view works – it will be used to control data ingestion in the next labs.

**A screenshot of a computer

Description automatically generated**