+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  
Cloud Scale Analytics with Microsoft Fabric  
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

++++++++++++

Introduction

++++++++++++

This lab is an 8-hour hands-on experience that teaches you how to use **Microsoft Fabric** to handle large-scale data analytics. You’ll learn how to:

* Set up a collaborative **workspace** in Fabric
* Use **data pipelines** to ingest and prepare data
* Analyze big datasets using a **data warehouse**
* Run **real-time analytics** for immediate insights
* Train **machine learning models** using **notebooks**
* Use **Apache Spark** for high-performance distributed data processing
* Design advanced **data transformations** with **Dataflow Gen2**

The lab simulates a real-world scenario where a global e-commerce company uses Microsoft Fabric to manage and analyze massive volumes of customer transaction data in real-time.

**What It’s Saying:**

The lab is structured to give you end-to-end exposure to Microsoft Fabric's cloud-scale analytics platform. You’re not just clicking through tools—you’re learning how everything fits together in a real solution:

1. **Fabric Workspace:**  
   You begin by setting up a central hub to manage all your data projects. Think of it as your operations center.
2. **Data Ingestion with Pipelines:**  
   You automate the flow of data from various sources into your system. No manual importing—everything flows in automatically.
3. **Data Warehousing:**  
   Once your data is in, you run SQL queries on large datasets to extract insights. This is where you turn raw data into useful answers.
4. **Real-Time Analytics:**  
   You tap into live data streams to make instant decisions—helpful for scenarios like detecting fraud or tracking user behavior live.
5. **Machine Learning with Notebooks:**  
   You build and test ML models interactively using Python/R in notebooks. This is useful for predictions or automation.
6. **Apache Spark for Big Data:**  
   For really heavy processing, you use Spark—designed to handle massive data volumes efficiently across multiple machines.
7. **Dataflow Gen2:**  
   You apply complex data transformations using a visual and code-based interface to clean, shape, and prepare your data pipeline-ready.

**The Big Picture:**

This lab helps you understand how Microsoft Fabric ties together real-time data ingestion, analysis, machine learning, and transformation in one ecosystem. If you're working with big, fast-moving, or complex data, this is the kind of architecture you'd want to implement in a real company.

+++++

Lab 1

+++++

**Goal of the Exercise**

Set up the full pipeline to ingest, transform, and query data in a **Fabric Lakehouse**, ending with a basic **Power BI report**. You work with **pipelines**, **Apache Spark**, **SQL**, and **visual queries**.

**What You Did – Step by Step**

**✅ Task 1: Sign Up for Microsoft Fabric Trial**

* Activated a 60-day free Fabric trial from the Power BI portal.
* Verified trial status was active.

**✅ Task 2: Create a Workspace**

* Created a new Fabric-enabled workspace.
* This workspace holds everything: lakehouse, pipeline, notebook, Power BI assets.

**✅ Task 3: Create a Lakehouse**

* Switched to the **Data Engineering** experience.
* Created a Lakehouse called Lakehouse\_.
* Created a subfolder named new\_data under the Files section.

**✅ Task 4: Create a Data Pipeline**

* Created a pipeline named Ingest Sales Data Pipeline.
* Used the **Copy Data assistant** to pull sales.csv from a public URL.
* Copied the file to the new\_data folder in the lakehouse.

**✅ Task 5: Create a Notebook**

* Created a notebook named Load Sales Notebook.
* Wrote **PySpark** code to:
  + Load sales.csv
  + Add Year, Month, FirstName, LastName columns
  + Reorder and filter columns
  + Write to a **managed table** called sales
* Ran all cells. Verified the sales table appeared under Tables.

**✅ Task 6: Query with SQL**

* Switched to the **SQL Analytics Endpoint**.
* Wrote a SQL query to get **total revenue by item**:
* Executed the query and reviewed results.

**✅ Task 7: Create a Visual Query**

* Created a **Visual Query** using Power Query.
* Grouped the sales data to count **distinct line items per order**.
* Used a simple drag-and-drop UI to perform the transformation.

**✅ Task 8: Create a Power BI Report**

* Opened the **Reporting tab**, started a new report.
* Added Item and Quantity fields from the sales table.
* Switched to a **Clustered Bar Chart**.
* Saved the report as Item Sales Report.

**What You Learned**

* How to build an **end-to-end data ingestion pipeline** in Fabric.
* How to store and transform data using Spark notebooks.
* How to query managed tables using both SQL and visual tools.
* How to create and save Power BI reports from lakehouse data.

**Final Result**

Your workspace now includes:

* A lakehouse with structured data
* A SQL endpoint
* A managed table (sales)
* A Spark notebook
* A pipeline
* A Power BI report

You’ve completed the full data flow: **ingestion → transformation → analysis → visualization**. Ready to move on to the next exercise.

+++++

Lab 2

+++++

**Goal of the Exercise**

You'll work end-to-end with a **Microsoft Fabric data warehouse**: designing the schema, loading and querying data, building relationships, creating views, running visual queries, and finally generating a Power BI report.

**What You Did – Step by Step**

**✅ Task 1: Create a Data Warehouse**

* Created a new data warehouse in your Fabric workspace.
* This serves as the centralized repository for structured data.

**✅ Task 2: Create Tables and Insert Data**

* Created fact and dimension tables (like DimProduct, FactSalesOrder, etc.) using SQL.
* Loaded sample data into these tables manually and via pre-written .txt files.

**✅ Task 3: Define a Data Model**

* Set up **relationships** between the tables using **Model Layouts**:
  + FactSalesOrder → DimProduct, DimCustomer, DimDate
* Defined these as **many-to-one** relationships to build a star schema for analytics.

**✅ Task 4: Query Data Warehouse**

* Ran SQL queries using **JOINs**, **GROUP BY**, and **aggregation** functions.
* Created insights like **Sales Revenue by Year, Month, and Region**.

**✅ Task 5: Create a View**

* Created a reusable view vSalesByRegion to encapsulate the query logic.
* Queried the view like a table to simplify future reporting and analysis.

**✅ Task 6: Create a Visual Query**

* Used the no-code **Visual Query Designer** to:
  + Drag-and-drop tables
  + Merge queries visually
  + Filter for specific products (like "Cable Lock")
* Explored the results and exported them if needed.

**✅ Task 7: Visualize Your Data**

* Hid unnecessary columns to simplify the model.
* Created a **Power BI report** with visuals (e.g., **Total Sales by Category**).
* Customized chart type, title, and layout.
* Saved the report to your Fabric workspace.

**What You Learned**

* How to build and manage a **data warehouse** in Microsoft Fabric
* How to use SQL for querying and creating views
* How to model data with relationships between fact and dimension tables
* How to explore data visually without writing SQL
* How to generate a business report using Power BI

**Outcome**

You now have:

1. A working data warehouse
2. A defined semantic model
3. A finished Power BI report—all saved in your workspace

You're ready to move on to the next exercise.

++++++

Lab 3

++++++

Learn how to analyze and visualize real-time data in Microsoft Fabric using **KQL**, **Power BI**, and **Spark Structured Streaming with Delta Tables**.

**Key Tasks & Steps**

**Task 1: Create a KQL Database**

* Created an **Eventhouse** and added a **KQL database** named KQL-Database.
* Imported **sales.csv** into a new table called sales.
* Verified and previewed the imported data.

**Task 2: Query Sales Data using KQL**

* Ran KQL queries to:
  + Filter data for a specific product (Road-250 Black, 48).
  + Filter by order date (post-2020).
  + Calculate and sort **total net revenue by product** for 2020.
* Saved the final query as **Revenue by Product** in a **KQL Queryset**.

**Task 3: Create a Power BI Report**

* Used the Revenue by Product queryset to build a **Power BI report**.
* Visualized Item vs TotalNetRevenue as a **clustered bar chart**.
* Saved the report as **Revenue by Item** to the Fabric workspace.

**Task 4: Use Delta Tables for Streaming Data**

* Used a **Spark notebook** to:
  + Create a streaming source from JSON files simulating IoT device data.
  + Store the stream in a **Delta table** called IotDeviceData.
  + Queried the table using SQL to view real-time ingested data.
  + Added more data to simulate continuous streaming.
  + Stopped the stream after validation.

**Outcomes**

* Learned how to:
  + Set up and use **KQL databases**.
  + Run advanced KQL queries for analytics.
  + Build **Power BI visuals** directly from KQL results.
  + Ingest and analyze **real-time streaming data** using Delta tables and Spark.

This exercise connects static and streaming analytics in Microsoft Fabric, bridging data exploration, visualization, and real-time processing in a single workflow.

+++++++

Lab 4

+++++++

**Goal**

Build a machine learning workflow in Microsoft Fabric using **notebooks**, **Scikit-learn**, and **MLflow**—from data ingestion to model training, experiment tracking, and saving the best model.

**Key Tasks & Steps**

**Task 1: Upload Churn Data**

* Created or navigated to a **Lakehouse**.
* Uploaded churn.csv to the **Files** section for use in the ML workflow.

**Task 2: Create a Notebook**

* Created a new **notebook** in the workspace.
* Added a **markdown title** for clarity and documentation.

**Task 3: Load Data into a DataFrame**

* Connected to the uploaded churn file.
* Loaded it into a **Pandas DataFrame** using:

**Task 4: Train ML Models**

* Split the data into training and testing sets using train\_test\_split.
* Set up **MLflow experiment tracking**.
* Trained two models:
  + **Logistic Regression**
  + **Decision Tree**
* Used mlflow.autolog() to automatically log parameters, metrics, and artifacts.

**Task 5: View and Compare MLflow Experiments**

* Listed and fetched MLflow experiments using code.
* Retrieved experiment runs by ID.
* Visualized **model accuracy** using matplotlib bar charts comparing estimators.

**Task 6: Explore Experiments in the UI**

* Navigated to the **MLflow UI in Fabric**.
* Compared model runs interactively.
* Switched chart types and axes to visualize accuracy by estimator.

**Task 7: Save the Best Model**

* From the experiment run, saved the best-performing model as **model-churn**.
* Stored it under the **Models** folder for reuse and deployment.

**Task 8: Save Notebook and End Spark Session**

* Renamed the notebook to **Train and compare models notebook**.
* Ended the **Spark session** to clean up resources.

**Outcomes**

* Built an **end-to-end ML pipeline** in Fabric using Python notebooks.
* Tracked model training with **MLflow**.
* Compared models visually and saved the best one.
* Learned how to integrate data science tooling inside the Fabric environment for reproducible workflows

++++++

Lab 5

++++++

**Goal**

Build a simple data pipeline using **Dataflow Gen2** and **Copilot**, leveraging natural language prompts to ingest, clean, and transform sales data without writing manual code.

**Key Steps & Tasks**

**1. Create a Dataflow Gen2**

* Navigated to workspace and created a new **Dataflow Gen2** (without Git/public API).
* Launched the **Power Query editor** for data transformation.

**2. Import Sales Data**

* Imported data from a **CSV file** hosted on GitHub using a public URL.
* Used **anonymous authentication** for connection.

**3. Use Copilot for Data Transformation**

* Accessed Copilot and expanded the UI to view:
  + **Diagram View** for visual steps.
  + **Script View** to see the underlying **M-code**.

**4. Clean and Split Fields**

* Cleaned the Item column by replacing , with a space.
* Split Item into three new fields: **Description**, **Color**, and **Size** using Copilot.

**5. Modify Data with Logic**

* Used Copilot to create a **new Quantity column**:
  + If Color is "Red", multiplied the quantity by 10.
  + Otherwise, kept the original value.
* Replaced the old Quantity column with the updated one.

**Visual vs Script-Based Querying**

**Visual Query**

* Enables quick, code-free data exploration.
* Good for users with minimal technical background.

**M-Code (Script View)**

* Offers advanced transformation and customization.
* Useful for automation, repeatable logic, and complex rules.

**Outcome**

* Built a working **data pipeline** using Dataflow Gen2.
* Transformed raw sales data using **Copilot with natural language**.
* Understood how to move between **visual workflows** and **script-based editing**.
* Showed how AI-assisted transformations speed up and simplify data prep.

This exercise bridges low-code/no-code interfaces with advanced scripting for flexible and intelligent data engineering in Microsoft Fabric.

+++++++

Lab 6

+++++++

**Objective**

Analyze warehouse data using **Fabric Copilot**—explore datasets, understand relationships, run queries, generate reports, and gain insights using natural language prompts.

**Key Activities**

1. **Connect to Data Warehouse**
   * Selected workspace and connected to a data warehouse.
   * Verified presence of four key tables: DimCustomer, DimDate, DimProduct, FactSalesOrder.
2. **Explore Data Model**
   * Viewed relationships between tables in **Model Layout**.
   * Checked relationship properties, especially how FactSalesOrder connects to DimCustomer.
3. **Understand Semantic Model**
   * Recognized how relationships define the semantic layer used by Power BI for building reports.
4. **Generate Reports with Copilot**
   * Created a new Power BI report based on the semantic model.
   * Used Copilot to:
     + Analyze the data structure.
     + Suggest insights.
     + Generate reports using prompts like:
       - "Create a report that shows Total Sales by Product Category"
       - "Suggest content for this report"
       - "Give me an executive summary"
5. **Customize and Save Reports**
   * Explored and edited Copilot-generated content.
   * Saved the final report as **Sample**.

**Outcome**

* Used Copilot to assist with data exploration and reporting.
* Generated sales insights (e.g., total sales, performance by region).
* Understood the structure and value of the semantic model in Power BI.
* Produced a summary-ready report for decision-making.

This exercise showed how AI can simplify analysis and guide report creation—making business intelligence more accessible and collaborative.

++++++

Lab 7

++++++

**Summary: Explore Sales Orders with Apache Spark in Microsoft Fabric**

In this lab, you worked end-to-end on big data analytics using **Apache Spark and PySpark** within Microsoft Fabric. Here's what you accomplished:

1. **Created a Lakehouse**  
   Set up a lakehouse and uploaded CSV sales data for 2019–2021.
2. **Built a Spark Notebook**  
   Created a notebook in the Data Engineering experience and connected it to the lakehouse.
3. **Loaded and Structured Data**  
   Loaded raw CSV files into a Spark DataFrame, corrected missing headers, and applied a defined schema using StructType.
4. **Explored and Filtered Data**  
   Used DataFrame operations to filter rows, select columns, and identify unique customers for specific products.
5. **Aggregated and Grouped Data**  
   Summarized total quantity sold per item and counted orders by year using groupBy and aggregation functions.
6. **Transformed and Saved Data**  
   Added Year, Month, FirstName, and LastName columns, then saved the transformed DataFrame in **Parquet** format and as **partitioned files** by year and month.
7. **Created and Queried Delta Tables**  
   Saved the dataset as a **Delta table**, queried it using both PySpark and SQL, and explored the structure via the Spark metastore.
8. **Visualized Data**  
   Created visualizations using:
   * **Built-in Spark charts**
   * **matplotlib** (bar chart, pie chart, subplots)
   * **seaborn** (bar and line charts with themes)
9. **Finalized Notebook**  
   Renamed the notebook to *Explore Sales Orders Notebook* and stopped the Spark session to release resources.

++++++

Lab 8

++++++

**Summary: Ingest and Transform Data with Dataflow Gen2 in Microsoft Fabric**

In this lab, you created a **Dataflow Gen2** to ingest and transform sales data using **Power Query Online**, then automated its execution using a pipeline. Here's what you did:

1. **Created a Dataflow Gen2**
   * Imported sales data from a CSV file using Power Query.
   * Added a custom column MonthNo by extracting the month from OrderDate.
2. **Configured Data Destination**
   * Connected the dataflow to a **Lakehouse** destination.
   * Mapped columns manually and changed MonthNo to a whole number.
   * Published and renamed the dataflow to **Transform Orders Dataflow**.
3. **Automated with a Pipeline**
   * Created a **Data pipeline** named **Load Orders pipeline**.
   * Added the dataflow as a pipeline activity.
   * Ran the pipeline to load the transformed data into the lakehouse.
4. **Verified the Output**
   * Opened the **fabric\_lakehouse**, located the new **orders** table, and confirmed the data was loaded.

This exercise gave you hands-on experience with ingesting, transforming, and automating data loading in Fabric using **Dataflow Gen2** and **pipelines**—key tools for modern data engineering workflows.

Ask ChatGPT

++++++++

Summery

++++++++

**What these labs were about:**

The labs were a deep dive into how Microsoft Fabric helps handle large-scale data analytics end to end—everything from data ingestion and transformation, to analysis, machine learning, and real-time reporting—all in one platform.

It’s like building a full data analytics system from scratch, but with modern cloud-native tools.

**How it was structured:**

The entire training was broken into **8 labs**, and each one focused on a different piece of the data pipeline. Here’s how it all fits together:

**1. Getting Started (Lab 1)**

You set up your **Fabric workspace**, created a **Lakehouse**, and built your first full pipeline:

* Pulled in sales data using a **Data Pipeline**
* Transformed it with **Spark Notebooks (PySpark)**
* Queried it using **SQL** and **Power Query**
* Built a basic **Power BI report** on top of it

**Takeaway:** You built a working ingestion-to-reporting flow using lakehouse storage and multiple tools inside Fabric.

**2. Data Warehousing (Lab 2)**

You created a structured **Data Warehouse** with proper **fact and dimension tables**, modeled the relationships, and built reports.

**Key tools used:**

* SQL for schema design and querying
* Model layouts for defining star schema
* Power BI for visuals
* Visual Query editor for no-code exploration

**Takeaway:** You learned how to model data for performance and reusability inside a warehouse.

**3. Real-Time Analytics (Lab 3)**

You worked with both static and streaming data:

* Created a **KQL database** and queried sales data using **Kusto Query Language**
* Built Power BI visuals directly from KQL
* Set up a **Spark Structured Streaming** job using Delta tables to simulate IoT data

**Takeaway:** You now know how to handle both historical and real-time data using Microsoft Fabric.

**4. Machine Learning (Lab 4)**

Built an end-to-end **ML pipeline** using **Spark notebooks**, **Scikit-learn**, and **MLflow**:

* Trained and compared models (Logistic Regression and Decision Tree)
* Tracked experiments and metrics
* Saved the best model for reuse

**Takeaway:** You can now train, compare, and manage ML models right within the Fabric environment.

**5. No-Code + AI (Lab 5)**

Used **Dataflow Gen2** and **Copilot** to clean and transform data using natural language:

* Loaded a CSV file
* Used Copilot prompts to split and clean columns
* Applied conditional logic for transformations
* Switched between visual and script-based views (Power Query M)

**Takeaway:** You learned how AI tools like Copilot speed up data preparation, great for low-code teams.

**6. Copilot for BI (Lab 6)**

Analyzed data warehouse content using **Copilot** inside Power BI:

* Explored the semantic model (Fact and Dim tables)
* Asked Copilot to generate reports, summaries, and insights
* Edited and saved those reports

**Takeaway:** You saw how Copilot helps non-technical users explore data and build executive-ready reports faster.

**7. Big Data with Spark (Lab 7)**

Used **Apache Spark and PySpark** for heavy data wrangling:

* Loaded multiple years of sales CSVs
* Cleaned, transformed, and partitioned the data
* Saved it as **Delta tables** and ran queries in both PySpark and SQL
* Created charts with Spark’s built-in tools, matplotlib, and seaborn

**Takeaway:** You practiced distributed data processing and large dataset transformation in a scalable way.

**8. Dataflow Gen2 Automation (Lab 8)**

Created a **Dataflow Gen2** to transform sales data and automated it with a **Data Pipeline**:

* Applied transformations like extracting MonthNo from dates
* Loaded the clean data into the Lakehouse
* Set up a pipeline to trigger the dataflow

**Takeaway:** You now understand how to build and schedule repeatable data workflows with no manual intervention.

**The big picture:**

You now know how to:

* Set up a collaborative data platform using Fabric
* Move and transform data using Pipelines, Dataflows, and Spark
* Analyze both structured and streaming data using SQL, KQL, and notebooks
* Build ML models, track experiments, and save models
* Generate insights and reports using Copilot and Power BI

**In short:**  
You’ve worked hands-on with every major piece of the Microsoft Fabric ecosystem—from raw data to real-time dashboards to predictive ML. You could apply this to build a production-grade analytics solution for any data-heavy business.

++++++++

Example

++++++++

**Imagine This:**

You're working for an **online shopping company** like Amazon. Every day, people place orders, you get delivery info, payments, reviews, etc.

Now you want to:

* Know **which products sell the most**
* **Track deliveries in real-time**
* **Predict future sales**
* Show this data in a clear, **easy-to-read dashboard**

**Lab Summary (in very easy terms)**

1. **Bring Data In (Lab 1)**  
   You took raw data files (like Excel sheets of orders) and loaded them into **Fabric Lakehouse** using a pipeline.  
   ➤ Like uploading sales reports into one big digital folder.
2. **Organize the Data (Lab 2)**  
   You turned messy data into clean **tables** (like Product, Orders, Customers).  
   ➤ Think of it like sorting receipts by category.
3. **Real-Time Tracking (Lab 3)**  
   You worked with live-streaming data (like delivery truck GPS or IoT devices).  
   ➤ It’s like tracking packages live on a map.
4. **Machine Learning (Lab 4)**  
   You trained a model to predict something — like who might buy again or cancel orders.  
   ➤ Like how Netflix suggests what to watch next.
5. **Clean with No-Code & AI (Lab 5)**  
   You used a tool (Copilot) that lets you clean data just by typing instructions in plain English.  
   ➤ Like saying “split the full name into first and last” and it just does it.
6. **Generate Reports with AI (Lab 6)**  
   You asked questions like “Which region sells the most?” and Copilot built a chart.  
   ➤ Like asking ChatGPT to make a bar graph of your top-selling items.
7. **Big Data Processing (Lab 7)**  
   You used Apache Spark to work with large files and filter or transform them using code.  
   ➤ Like working with thousands of orders at once instead of manually checking each row.
8. **Automate It All (Lab 8)**  
   You made a setup that runs on its own every time new data comes in.  
   ➤ Like setting up a coffee machine to brew every morning without pressing a button.

**Final Takeaway**

You’ve now learned to:

* Handle big messy data
* Clean and organize it
* Get real-time insights
* Predict trends
* Build dashboards automatically  
  All inside **Microsoft Fabric**.