Microsoft

MicroHack: Microsoft Fabric Lakehouse End-to-End

Erik Eckhardt | Filip Slánička

Agenda

- Introduction to the MicroHack
- Introduction to the Microsoft Fabric Solution
- MicroHack: Microsoft Fabric Lakehouse End 2 End
- Let’s Start Hacking

Introducing to the MicroHack

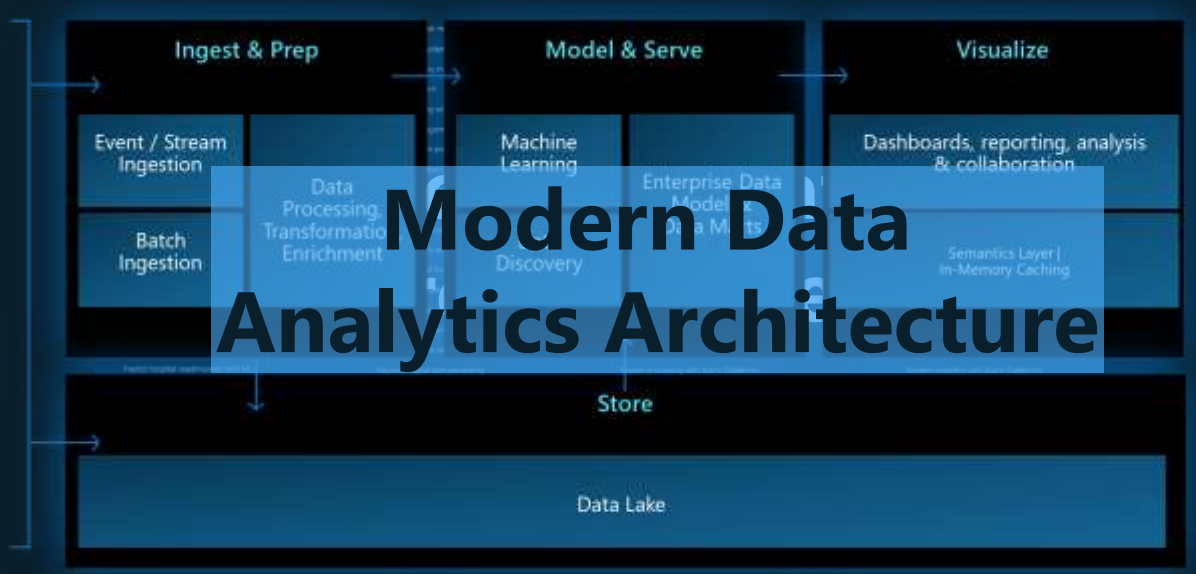
Introducing to the MicroHack

Concept

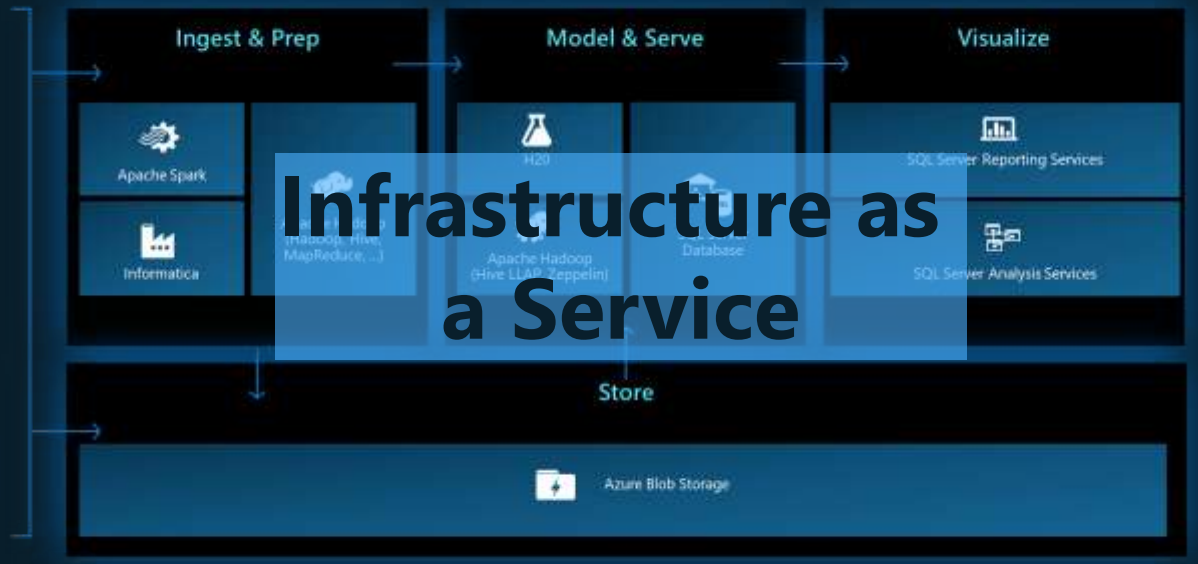
- Challenge-Based Tutorial Format
  - You'll be actively involved in coding, implementation, and problem-solving.
- What to Expect
  - Real-world scenarios and exercises with direct application of concepts in live environments.
- Hands-On Approach
  - Please read the detailed descriptions provided for each challenge.
  - Carefully review the goals, required actions, and success criteria.
- Support and Collaboration
  - Instructors are here to guide you. Feel free to ask questions and support your fellow participants.
  - A complete end-to-end solution guide with screenshots is also available.
- Outcome
  - By the end of this MicroHack, you'll gain practical experience and new skills that you can directly apply to your projects.

Introducing the Microsoft Fabric solution

Microsoft Data Analytics Evolution

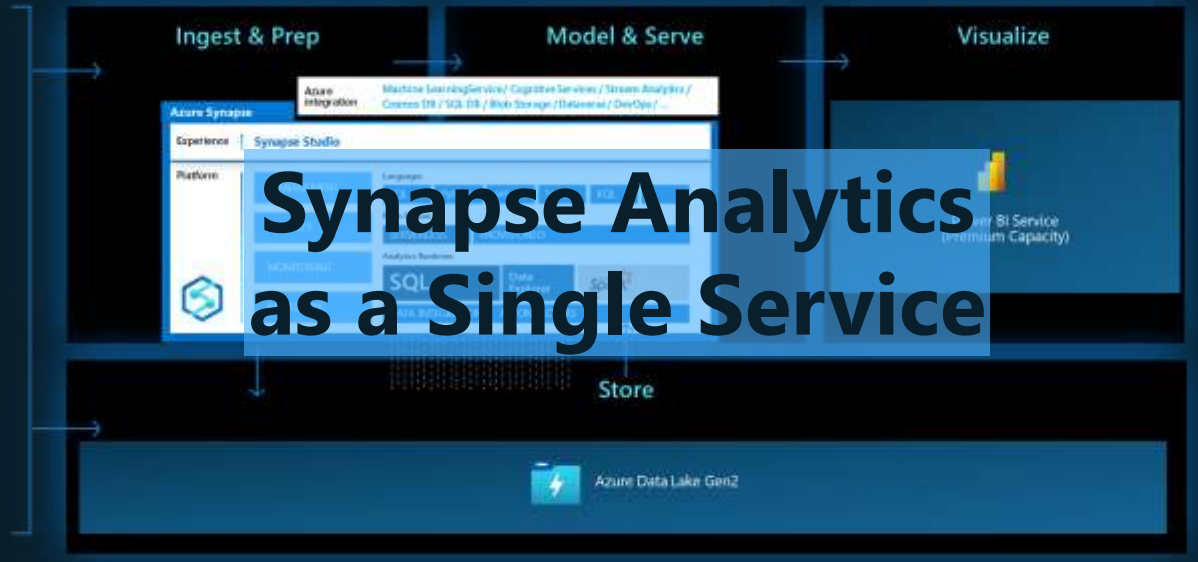


Modern Data Analytics Architecture



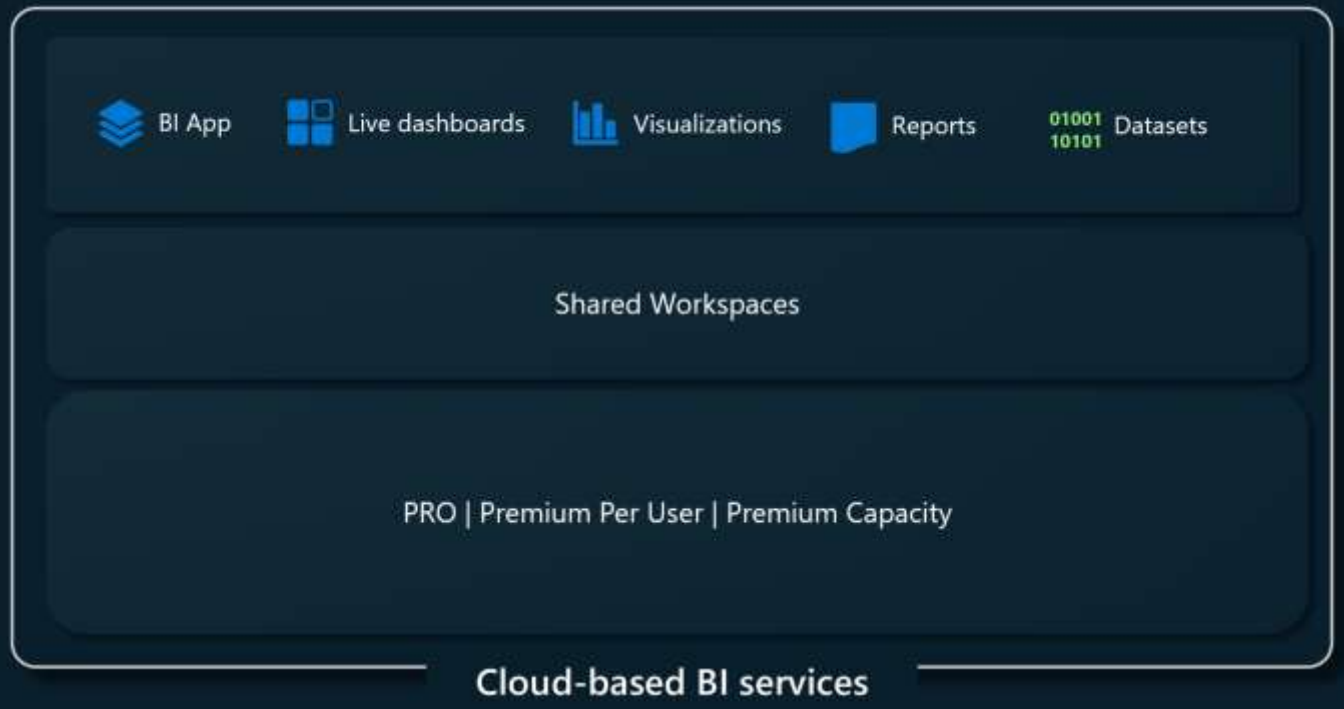
Infrastructure as a Service

Platform as a Service

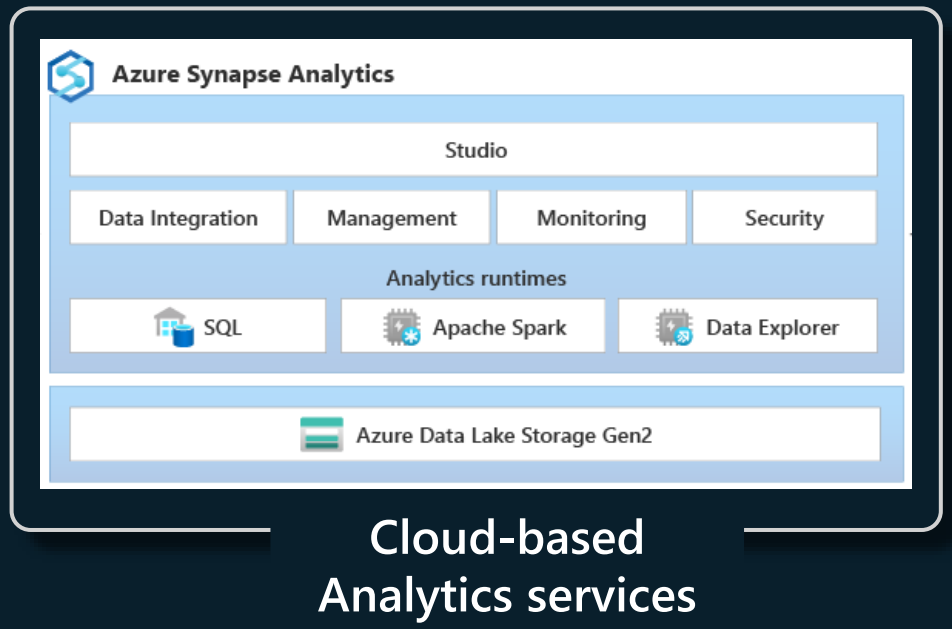


Synapse Analytics as a Single Service

Microsoft Power BI + Azure Synapse Analytics

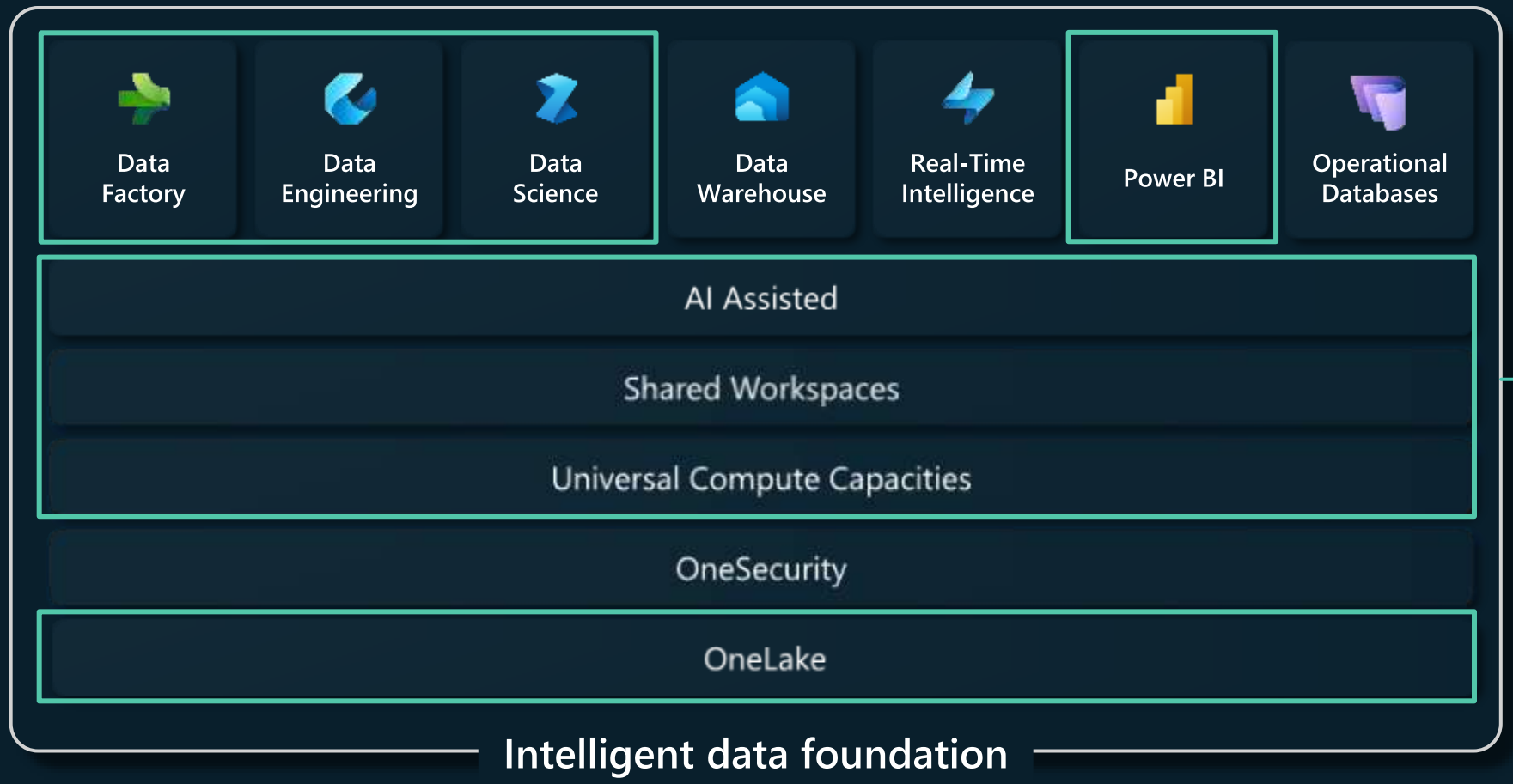


+



Microsoft Fabric

The unified data platform for the era of AI

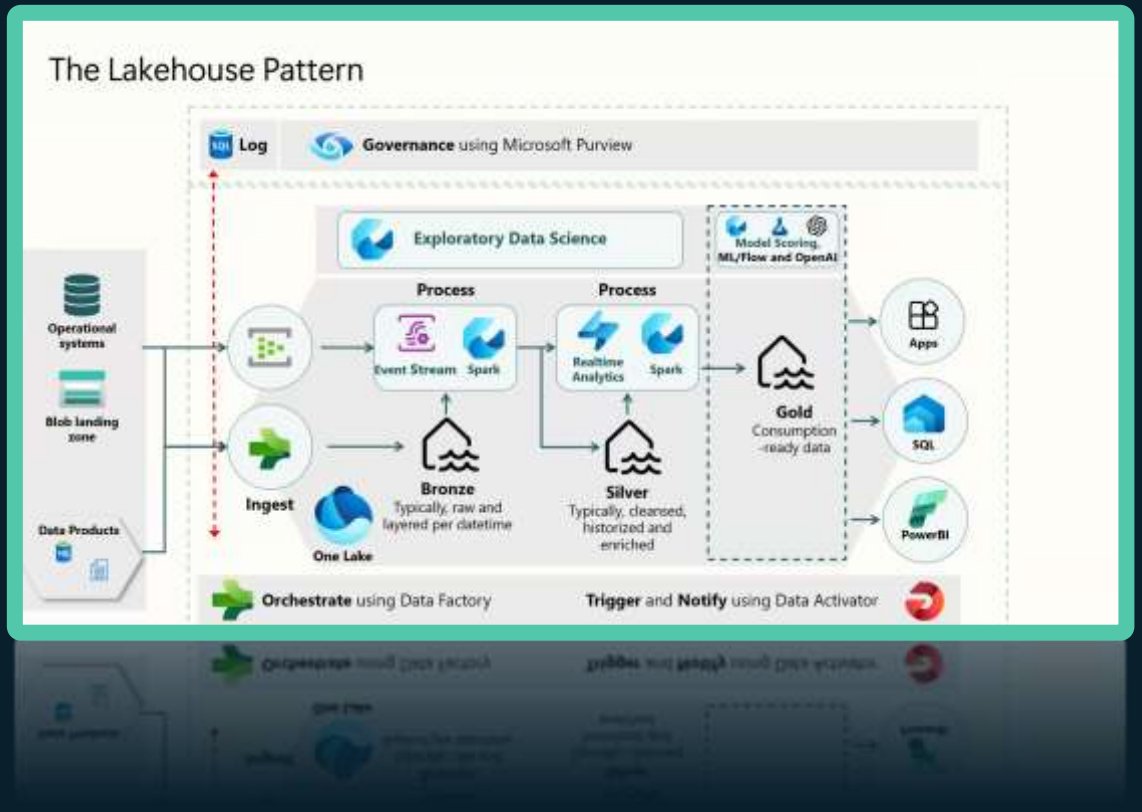


SaaS with single...

- Onboarding and trials
- Sign-on
- Navigation model
- UX model
- Workspace organization
- Collaboration experience
- Data Lake
- Storage format
- Data copy for all engines
- Security model
- CI/CD
- Monitoring hub
- OneLake Catalog
- Governance & compliance
- Domains

Microsoft Fabric

Data Architecture Patterns



Microsoft Fabric

A lakehouse and a medallion architecture

- A lakehouse is an open architecture that combines the best elements of data lakes and data warehouses.
- Lakehouses are enabled by a new system design: implementing similar data structures and data management features to those in a data warehouse directly on top of low-cost cloud storage in open formats with decoupled computes.

- A medallion architecture is a data design pattern used to logically organize data in a lakehouse.
- The goal is to incrementally and progressively improve the structure and quality of data as it flows through each layer of the architecture.
- From a data modeling perspective the Silver Layer has more 3NF write-performant data models, the Gold Layer is more de-normalized and read-optimized data models Kimball style star schema-based or Inmon style Data marts.

MicroHack: Microsoft Fabric Lakehouse End-to-End

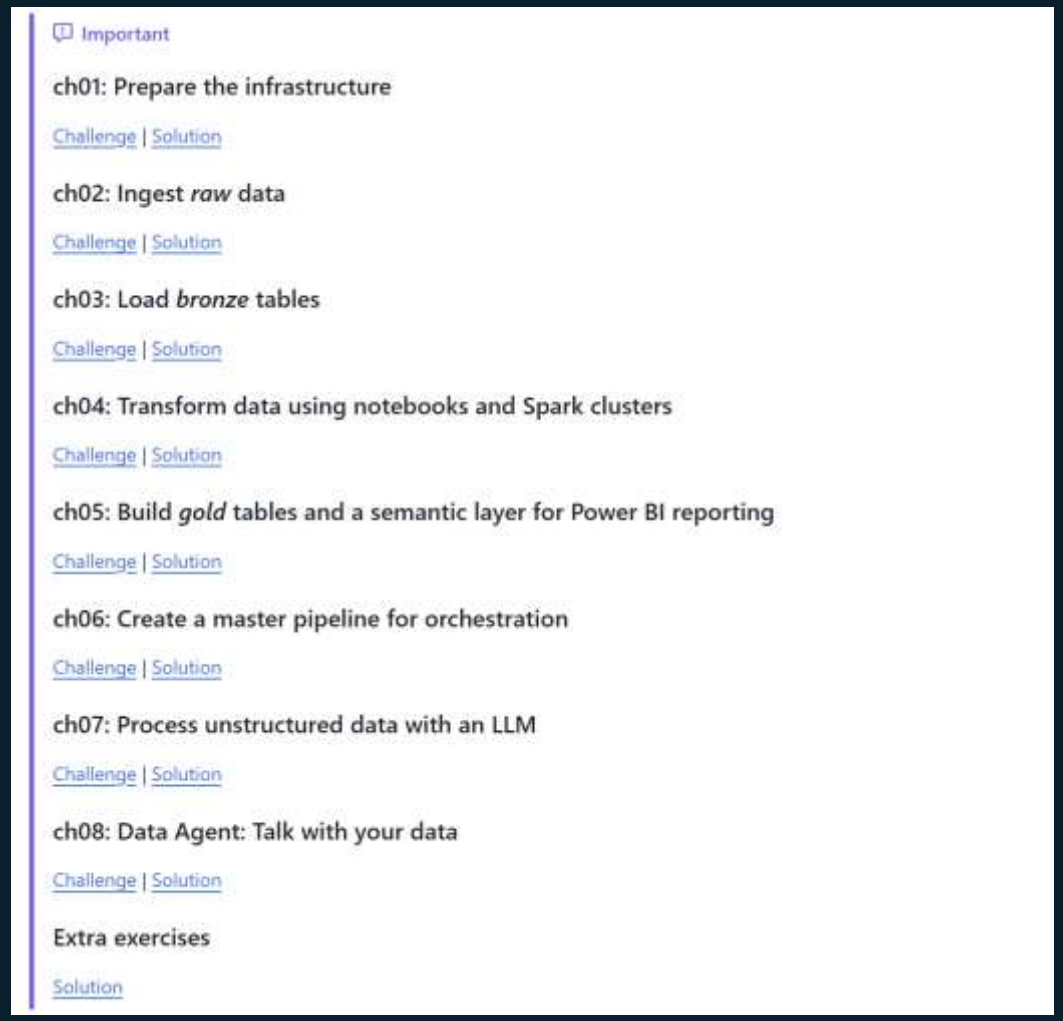
MicroHack: Microsoft Fabric Lakehouse End-to-End

Context & objectives

- Context
  - Learn how to integrate, transform, and manage data in a lakehouse using Fabric workloads such as Data Factory, Data Engineering, Data Science, and Power BI.
- Objectives
  - Develop a complete data workflow: ingestion, preparation, transformation and serving.

MicroHack: Microsoft Fabric Lakehouse End-to-End

Challenges



Important

ch01: Prepare the infrastructure

Challenge | Solution

ch02: Ingest *raw* data

Challenge | Solution

ch03: Load *bronze* tables

Challenge | Solution

ch04: Transform data using notebooks and Spark clusters

Challenge | Solution

ch05: Build *gold* tables and a semantic layer for Power BI reporting

Challenge | Solution

ch06: Create a master pipeline for orchestration

Challenge | Solution

ch07: Process unstructured data with an LLM

Challenge | Solution

ch08: Data Agent: Talk with your data

Challenge | Solution

Extra exercises

Solution

- Focus on implementing at least the first 6 challenges today as a minimum out of the 8 major challenges.

MicroHack: Microsoft Fabric Lakehouse End-to-End

Pre-req. for challenges

> **Caution**
>
> Before starting the challenges, make sure to go through the step-by-step guide for New Azure User Setup.

- You will receive "**XX**" now for your credentials. Please remember it!

> **Credentials**
>
> | Login | Password |
> | --- | --- |
> | `UserXX@fabric.microhack.org` | `dummyHeslo1234` |

MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch01: Prepare the infrastructure for Microsoft Fabric

Goal

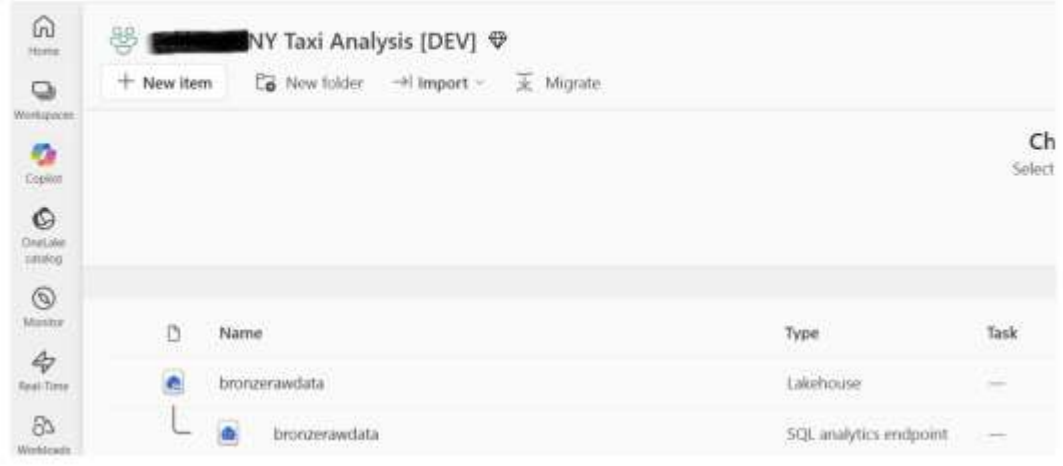
In this challenge, you will prepare the Microsoft Fabric environment for the upcoming challenges. Each participant should get their own Microsoft account for Azure and Office 365. Using your account, create a Fabric capacity and workspace, attach the capacity to your workspace, and create the first lakehouse for the bronze layer of the medallion architecture.

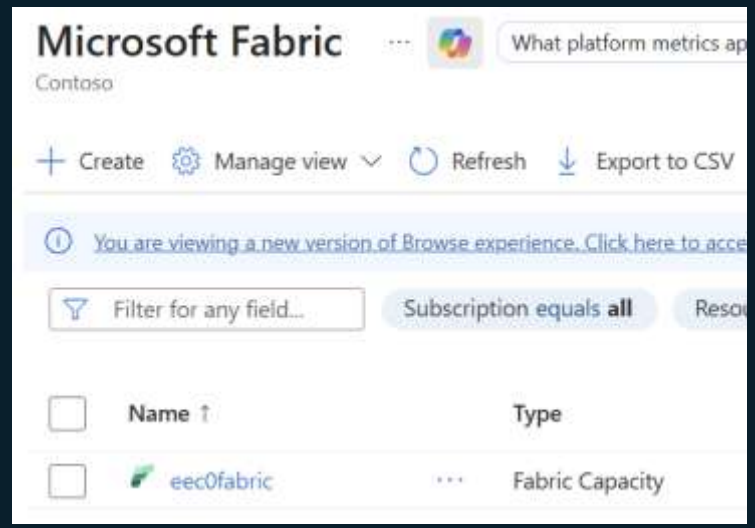
Actions

- Deploy MS Fabric Capacity into pre-defined Resource Group "**rg-MicroHackFabricCapacities**" and **any available region**
  - Name: **yoursurname0fabric** with **size F4**
  - e.g. novak0fabric
- Create Workspace in MS Fabric: **YourSurname NY Taxi Analysis [DEV]**
  - e.g. Novak NY Taxi Analysis [DEV]
- Add **your** Fabric Capacity to **your** Workspace.
- Change Spark settings: **Use 1 node only**
- Create the 1st lakehouse: **bronzerawdata**

Success Criteria

- MS Fabric Capacity is running.
- Workspace is ready to solve your workloads.
- The 1st Lakehouse and SQL analytics endpoint is deployed and you can connect





MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch02: Ingest raw data with data pipelines and shortcuts

Goal

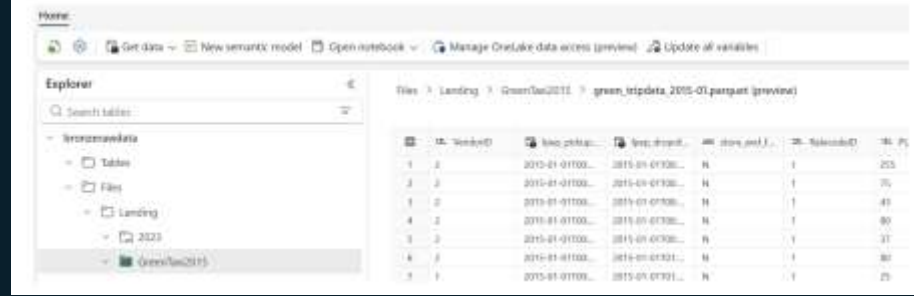
Focus on integrating two data sources from NYC taxi data: one from 2015 and the other from 2023. The table below presents the metaphors we will use to integrate the data:

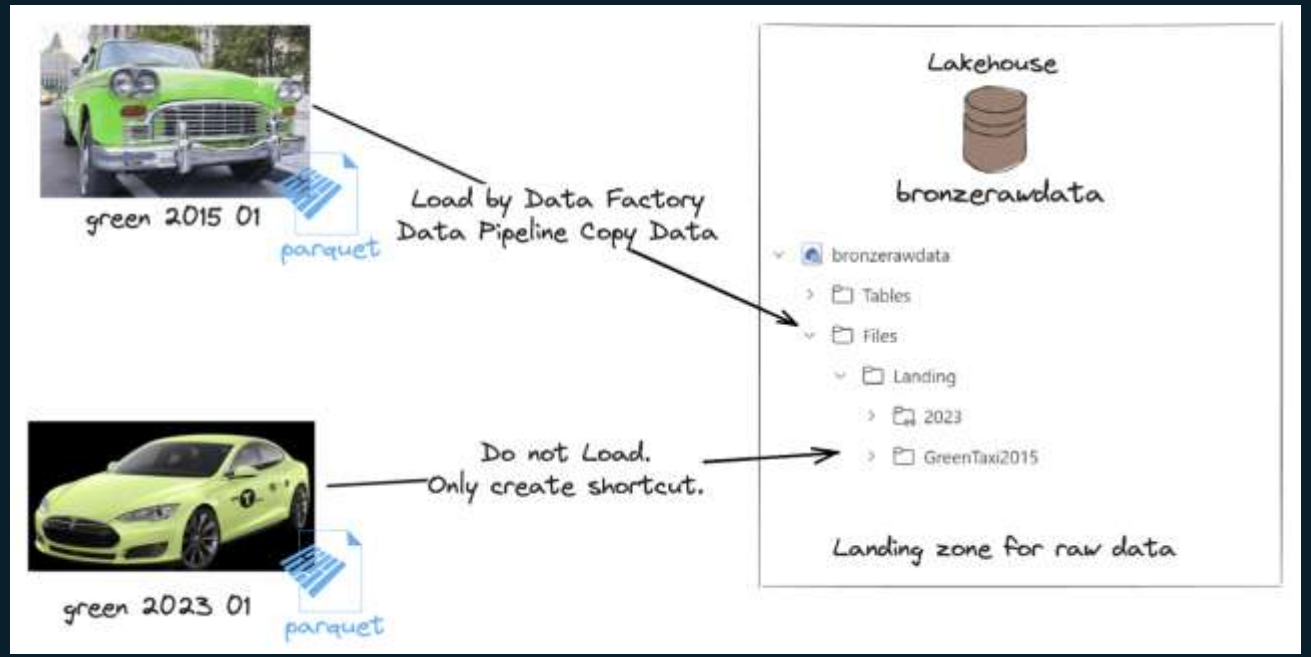
Actions

- Create Pipeline: Ingest Green Taxi 2015
  - With Copy Activity to Ingest Data:
    - Source: Azure Blob Storage: https://stmsfabricmicrohackxw.blob.core.windows.net/
      - Authentication kind: **Organizational account**
      - File path:
        - Container: **taxidata**
        - File: **green_tripdata_2015-01.parquet**
      - File format: **Binary**
    - Destination : Lakehouse **bronzerawdata**
      - File path directory: **Landing/GreenTaxi2015**
      - Format: **Binary**
- Create Shortcut in your Lakehouse into **Files** > **Landing** to **2023** Folder in Azure Data Lake Gen2:
  - https://adlstmsfabricmicrohackxw.dfs.core.windows.net/
    - Authentication kind: **Organizational account**

Success Criteria

- You have File "**green_tripdata_2015-01.parquet**" in your Lakehouse: Files > Landing > GreenTaxi2015
  - You can see the content of the file
- You have shortcut to Folder "2023" in your Lakehouse with files "**NYC-Taxi-Discounts-Per-Day.csv**" and "**green_tripdata_2023-01.parquet**"
  - You can see the content of both files





MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch03: Load bronze tables

Goal

In this challenge, you will load three files into three Delta tables in the bronze layer lakehouse. You will use three different load strategies:

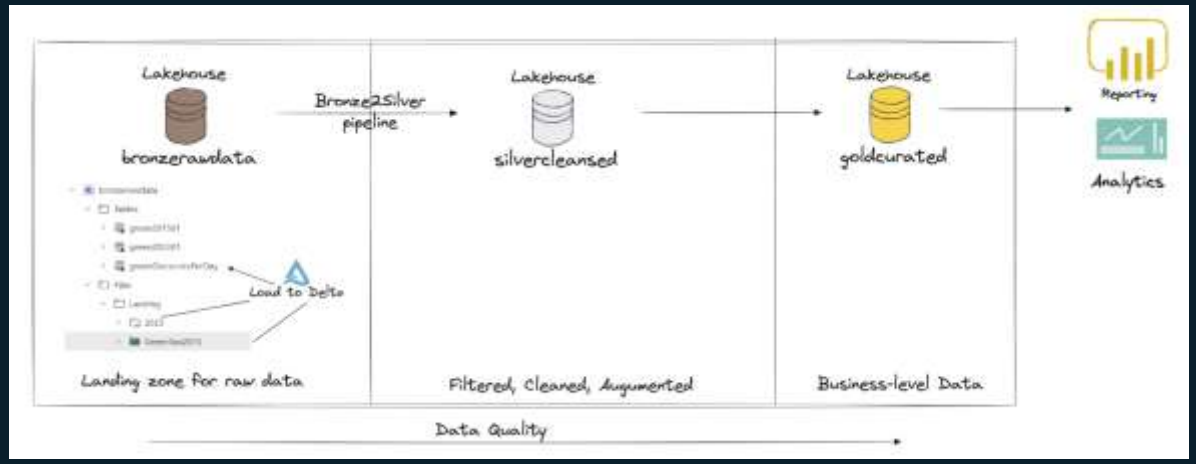
1. ETL Developer: Data Factory Pipeline
2. Citizen Data Developer: Data Factory Dataflow Gen2
3. Data Engineer: Notebook

Actions

- **ETL Developer: Data Factory Pipeline**
  - Create a new pipeline: **Load Green Taxi 2015**
  - Source:
    - Connection: **bronzerawdata** lakehouse
    - Root folder: **Files**
    - File path: **Landing/GreenTaxi2015/green_tripdata_2015-01.parquet**
    - File format: **Parquet**
  - Destination:
    - Connection: **bronzerawdata** lakehouse
    - Root folder: **Tables**
    - Table: **green201501** (new table)
    - Table action: **Overwrite**
- **Citizen Data Developer: Data Factory Dataflow Gen2**
  - Create a new Dataflow Gen2: **Load Green Taxi Discounts**
  - Source: **bronzerawdata** lakehouse
    - File path: **Files > Landing > 2023 > NYC-Taxi-Discounts-Per-Day.csv**
  - Destination: **bronzerawdata** lakehouse
  - Table: greenDiscountsPerDay
- **Data Engineer: Notebook**
  - Create a new Notebook: **Load Green Taxi 2023** or import challenges/ch03/Load Green Taxi 2023 - Ch03.ipynb
    - Remove origin lakehouses from Notebook and Add/attach your bronzerawdata (as default) lakehouse
  - Source: **Files/Landing/2023/green_tripdata_2023-01.parquet**
  - Target: **Tables/green202301**
    - Format: **delta**
    - Mode: **overwrite**
- Check # of rows per delta tables

Success Criteria

- Table **green201501 exists** with **1 508 493 rows**
- Table **green202301 exists** with **68 211 rows**
- Table **greenDiscountsPerDay exists** with **2 rows**



MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch04: Transform data using notebooks and Spark (silver tables)

Goal

In this challenge, you will explore data engineering tasks aimed at transforming raw data into a refined silver layer.

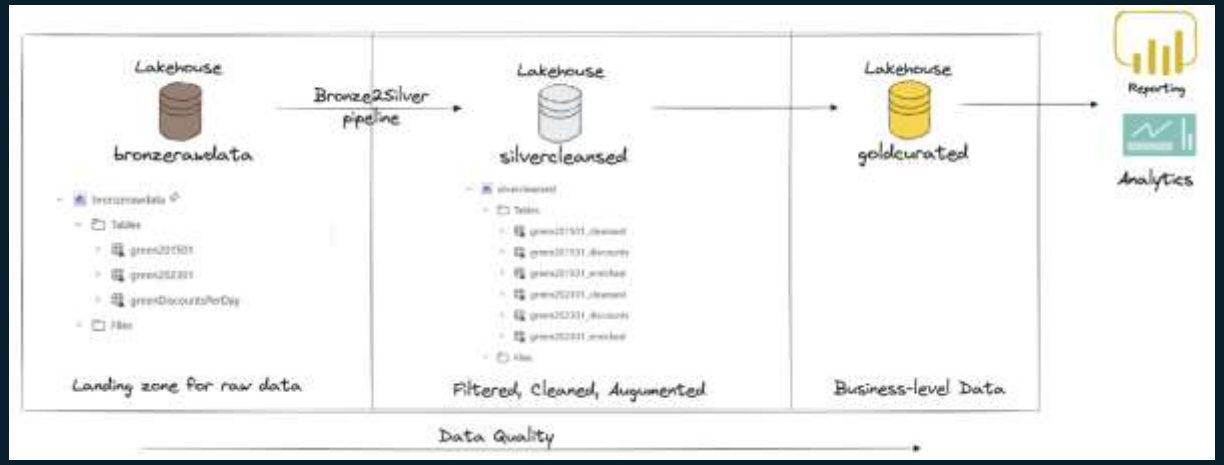
By the end of the challenge, you will have completed the second step, the silver layer, of the medallion architecture:

Actions

- Create the 2nd lakehouse: **silvercleansed**
- Import challenges/ch04/Transform bronze to silver - Ch04.ipynb to you workspace and try to complete all Goals / Actions mentioned in notebook
  - Remove origin lakehouses from Notebook and Add/attach your **bronzerawdata** (as default) lakehouse and **silvercleansed** lakehouse

Success Criteria

- All success criteria mentioned in the notebook are met
- 6 tables with data in **silvercleansed** lakehouse exist



MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch05: Build gold tables and a semantic layer for Power BI reporting

Goal

In this challenge, you will aggregate augmented data from a silver layer into a curated gold layer for advanced business analysts. Next, you will define semantic layer for end users and create the first Power BI report.

By the end of the challenge, you will have completed the implementation of the medallion architecture with semantics layer and reports:

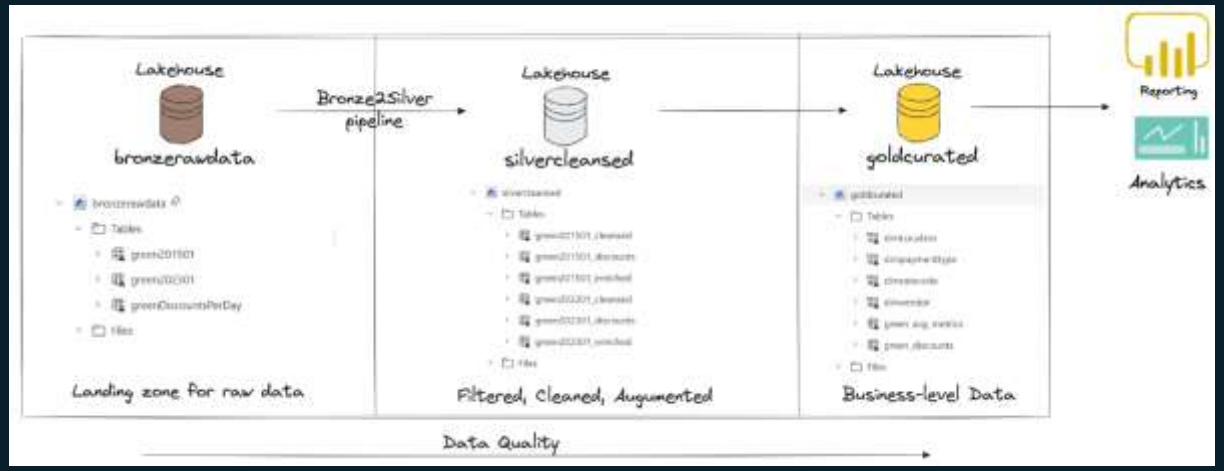
Actions

- Create the 3rd lakehouse: **goldcurated**
- Create Shortcut in your Lakehouse into **Tables** area to pre-created delta tables: **dimLocations, dimpaymenttype, dimratecode, dimvendor** in Azure Data Lake Gen2:
  - https://adlstmsfabricmicrohacksw.dfs.core.windows.net/
    - Authentication kind: **Organizational account**
- Import challenges/ch05/Transform silver to gold - Ch05.ipynb to you workspace and try to complete all Goals / Actions mentioned in notebook
  - Remove origin lakehouses from Notebook and Add/attach your **silvercleansed** (as default) lakehouse and **goldcurated** lakehouse
- Build Semantics model **GreenTaxi** based on 4 dimensions and 2 facts tables with Avg. function and relations
  - dimLocations.LocationID 1:N green_discount.PULocationID
  - dimpaymenttype.payment_type 1:N green_discount.payment_type
  - dimratecode.RatecodeID 1:N green_discount.RateCodeID
  - dimvendor.VendorID 1:N green_discount.VendorID

  - dimLocations.LocationID 1:N green_avg_metrics.PULocationID
  - dimvendor.VendorID 1:N greengreen_avg_metrics.VendorID
- Create sample report based on **GreenTaxi** Semantic model with
  - 1 Clustered bar chart, 1 Line chart, 1 Azure maps, 1 Narrative and 1 Slicer
- Create SQL View **vGetAverageFaresDiscounts** based on "Lakehouse SQL analytics endpoint" for business analysts
  - Get the distinct passenger count, total fares, average fares and discount grouped by the passenger count

Success Criteria

- All success criteria mentioned in the notebook are met
- 6 tables with data in **goldcurated** lakehouse exist
- Semantics model **GreenTaxi** with 2 Fact tables and 4 Dimensions exists
- Report based on **GreenTaxi** Semantics model exist
- SQL View **vGetAverageFaresDiscounts** returns 8 records



MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch06: Create a master pipeline for orchestration

Goal

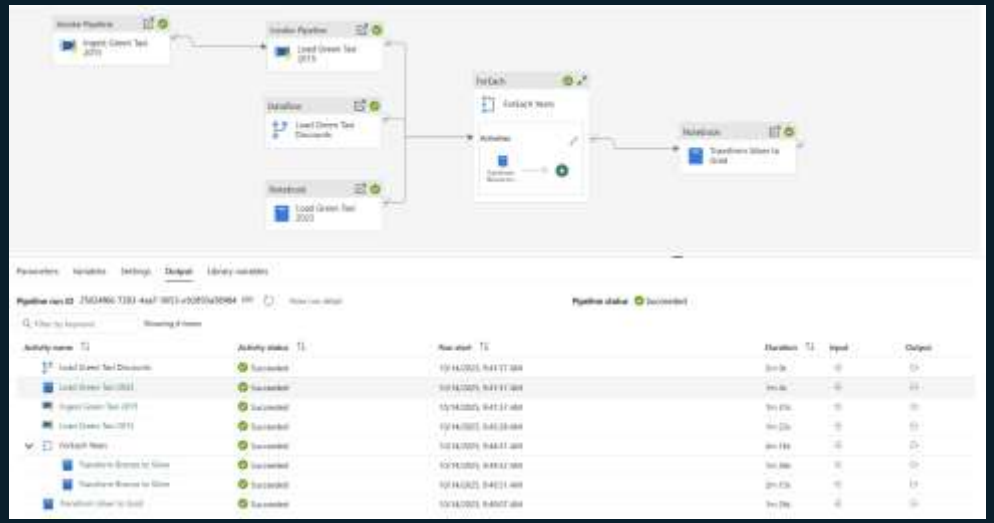
Now, let's focus on orchestration. Considering we have only two tables ingested by two pipelines and processed by three notebooks, imagine a scenario where you need to ingest and process tens or hundreds of tables, files, or data sources and manage dependencies between them. In such cases, the most efficient approach is to build and prioritize a data pipeline for the orchestration workflow.

Actions

- Create master Orchestration Pipeline
  - Add pipeline parameter: Year (Array, Default value = ["2015", "2023"])
  - Build following workflow:
    - Run in parallel:
      - Invoke pipeline: Ingest Green Taxi 2015
        - On Success, run: Invoke Load Green Taxi 2015
      - Run Dataflow: Load Green Taxi Discount
      - Run Notebook: Load Green Taxi 2023
    - On Success, run: ForEach activity: Sequential | Parallel for @pipeline().parameters.Year
      - Run Notebook: Transform Bronze to Silver
        - Add Base parameter p_year binded to ForEach items.
    - On Success: Run Notebook: Transform Silver to Gold
- Run pipeline

Success Criteria

- Pipeline finishes successfully



MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch07: Process unstructured data with an LLM

Goal

In this challenge, you will focus on processing unstructured data from real user reviews of locations in New York.

With Microsoft Fabric, all business professionals—data engineers, data scientists, developers, and analysts—can gain more value from their data through Generative AI.

There are two options in Microsoft Fabric today:

1. Use AI services with prebuilt and customizable APIs and models, but you need to develop code to process the output into the final schema:
   - Language Models: gpt-5, gpt-4.1, gpt-4.1-mini, and gpt-4o-mini
   - APIs: Text Analytics, and Azure AI Translator
2. Use the new set of AI functions for data engineering, which is much easier. Fabric users can take advantage of industry-leading large language models (LLMs) to transform and enrich data with just a single line of code.

By the end of this challenge, you will have a new dataset with sentiment, classification, summarization, translation, key topics, and location extracted from real user reviews.

> **Tip**
>
> It's up to you which option you choose, but this challenge is designed for **AI functions**.

Actions

- You can create new lakehohouse for Sandboxing or use **goldcurated**
- Upload file NY_places_customer_reviews.csv to your lakehouse (Files)
- Import challenges/ch07/User reviews processing with LLM - AI Functions - Ch07.ipynb to your workspace and try to complete all Goals / Actions mentioned in that notebook
  - Remove origin lakehouses from Notebook and Add/attach your **goldcurated** (as default) lakehouse

Success Criteria

- All success criteria mentioned in the notebook are met
- Table goldcurated.reviewLocations exists with 265 rows and AI columns

```
# TRANSLATE input text to Czech language

df["AItranslationsReview"] = df["Review"].ai.translate("Czech")
df["AItranslationsSummarize"] = df["AIsummarize"].ai.translate("Czech")
display(df)
```

```
100%|██████████| 265/265 [00:07<00:00, 36.36it/s]
100%|██████████| 265/265 [00:06<00:00, 39.93it/s]
```

Table view

	LocationID	Place	Review	User	Absentiment	Aiclassify	Aisummarize	AItranslationsReview	AItranslationsSummarize
1	1	New York: ...	Terrible if yo...	Bocodude	negative	historic building	User describes a ...	Hrozné, pokud musíte m...	Uživatel popisuje velké areni...
2	2	New York: ...	Even in centr...	Hobert01	positive	accommodation	In central NYC, yo...	V centru New Yorku si st...	V centru New Yorku můžete ...
3	3	New York: ...	Pelham Gard...	Joe Cosenti...	positive	accommodation	Pelham Gardens i...	Pelham Gardens je ubytn...	Pelham Gardens je centrálně ...
4	4	New York: ...	Alphabet Cit...	Lula N	positive	bars	Alphabet City is a ...	Alphabet City působí jako...	Alphabet City je skutečná čt...
5	5	New York: ...	Arden Heigh...	Anonymous	positive	accommodation	Arden Heights is ...	Arden Heights je domove...	Arden Heights je předměstsk...
6	6	New York: ...	Beautiful vie...	Michael S	mixed	historic building	Beautiful views an...	Krásné výhledy a zajímav...	Krásné výhledy a zajímavá hi...
7	7	New York: ...	Astoria is a v...	Current Re...	positive	restaurants	Astoria is a safe ...	Astoria je velmi pěkná čtv...	Astoria je bezpečná, dobře o...

MicroHack: Microsoft Fabric Lakehouse End-to-End

Ch08: Data Agent: Talk with your data

Goal

In this challenge, you will focus on the Data Agent, which allows you to build your own conversational Q&A systems using generative AI. The Fabric Data Agent uses LLMs to help users interact with their data naturally. It leverages Azure OpenAI Assistant APIs and behaves like an agent: processing user questions, determining the most relevant data source (Lakehouse, Warehouse, Power BI dataset, KQL databases), and invoking the appropriate tool to generate, validate, and execute queries. This way, even people without technical expertise in AI or a deep understanding of the data structure can receive precise and context-rich answers.

During this challenge, you will define a new Data Agent and add specific instructions, examples, and guidance to fine-tune the Fabric Data Agent. This ensures that responses align with your needs and goals, allowing everyone to engage with data more effectively.

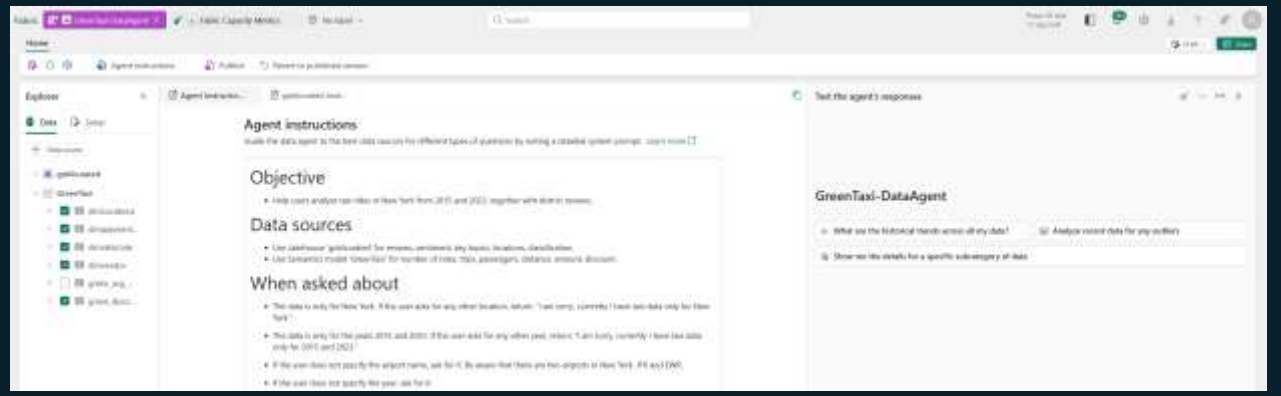
By the end of this challenge, the Data Agent will be ready to be consumed via the Python SDK, Azure AI Foundry, Microft Copilot Studio, and Copilot in Power BI.

Actions

- Create new Data Agent
- Add Data sources:
  - Semantic model: GreenTaxi
    - Tables:
      - dimLocations
      - dimpaymenttype
      - dimratecode
      - dimvendor
      - green_discount
  - Lakehouse: goldcurated
    - Tables:
      - dimLocations
      - reviewLocations
- Test questions:
  - What were the most common pickup locations for green taxi rides in 2023?
  - On which weekday do people travel from the airport the most?"
  - How many rides were in Boston?
  - How many rides were there in 2025?
  - Show me the top 3 districts in New York with the best accommodation.
  - Show me the districts where the most passengers moved from.
  - Show me the districts where most passengers moved from, and what reviews, sentiments, and key topics you know about these districts.
- Tune Data Agent via Agent instructions (System Prompt) and Data source instructions for above questions
- Publish Data Agent
- Consume Data Agent via Copilot (Home) in Power BI or PythonSDK

Success Criteria

- Data Agent exists and is able to answer questions about data in Semantics model and Lakehouse.
- Data Agent is available via Python SDK or Copilot (Home) in Power BI
- Data Agent can answer questions:
  - Question: What were the most common pickup locations for green taxi rides in 2023?
    - Answer: East Harlem North; East Harlem South; Forest Hills; Central Harlem; ...



Let’s Start Hacking

MicroHack-Fabric GitHub Repo:

http://aka.ms/MicroHackFabric
(https://github.com/CZSK-MicroHacks/MicroHack-Fabric)

Microsoft