# Lab 2 – Data Ingestion & Modeling Using Medallion Architecture

This lab focuses on ingesting and modeling structured and unstructured underwriting data to create an AI-ready data foundation. Learners will work with large-scale structured data from the FEMA Claims dataset and unstructured underwriting reference materials, implementing a Medallion Architecture (Bronze, Silver, Gold) in Microsoft Fabric to support downstream analytics, search, and agentic AI workflows.

**Lab Goals**

- Understand the role of structured and unstructured data in underwriting AI solutions
- Ingest the FEMA Claims dataset into Microsoft Fabric as structured data
- Ingest underwriting manuals and NFIP reference materials into Fabric as unstructured data assets
- Design and implement a Medallion Architecture (Bronze, Silver, Gold) using the FEMA Claims dataset
- Create Bronze layer tables representing raw, immutable claims data
- Transform and cleanse claims data in the Silver layer, including schema normalization, data quality checks, and enrichment
- Create Gold layer tables optimized for analytics, and AI consumption
- Use provided Python notebooks to execute ingestion and transformation logic
- Validate record counts, schemas, and data quality across all medallion layers
- Prepare curated Gold datasets for Fabric Data agent reasoning

**Hands-On Activities**

- Download open source datasets and upload them to Fabric OneLake
- Execute provided Python notebooks to ingest raw FEMA Claims data into the Bronze layer
- Perform data cleansing, normalization, and enrichment in the Silver layer
- Aggregate and optimize claims data into Gold layer tables
- Store underwriting manuals and NFIP materials as reference assets in Fabric / OneLake
- Perform basic data validation and quality checks at each medallion stage

**Dependencies and Prerequisites**

- Microsoft Fabric workspace (Contributor access)
- OneLake storage enabled
- FEMA Claims dataset, underwriting & claims manuals
- Fabric Notebook environment (Python)

**Outputs of This Lab**

- Bronze, Silver, and Gold claims tables implemented in Fabric
- Cleaned and curated datasets ready for analytics and AI Agents
- Unstructured underwriting reference materials stored and cataloged

# Hands-On Activities: Step by step instructions

## Create Fabric Objects to Create

Goal: Set up the Microsoft Fabric environment to host the NFIP dataset and support the multi-layer (Bronze → Silver → Gold) pipeline for the underwriting data agent.

### 1. Download structured and unstructured data for Underwriting solution

Download following datasets onto your laptop from the provided links. We will be uploading them to OneLake in the next steps.

| Name / Source | Description | Link |
|---|---|---|
| **FIMA NFIP Redacted Claims v2 (FEMA)** | Over 2.7 million flood-insurance claim transactions. Ideal for modelling peril-specific property risk and exposure. | FIMA NFIP Redacted Claims - v2 \| FEMA.gov |
| **Texas FAIR Plan Underwriting Manual** | Official underwriting manual for residential property coverage under the Texas FAIR Plan — details eligibility, coverage, inspection, and risk rules. | TFPA-Underwriting-Manual_Edition-Date-04-2023.pdf |
| **NFIP Claims Manual** | FEMA's official claims-handling manual under the National Flood Insurance Program — valuable for flood-risk decision logic and claims-process transparency. | NFIP Claims Manual (June 2025) |

### 2. Create the Lakehouse

A workspace is the logical container for all Fabric items (Lakehouses, pipelines, notebooks, semantic models, etc.). A Lakehouse provides a unified storage and compute layer (OneLake + Delta tables) for structured and unstructured data.

Follow steps below:
1. In the uw-agentic-ai workspace, click New → Lakehouse.
2. Name it lh_nfip.
3. Add a description: Lakehouse for FEMA NFIP dataset and derived underwriting features.
4. Click Create.
5. Verify folders like Files/ and Tables/ exist.

Verify that your workspace now contains lh_nfip.

## 3. Upload Raw FEMA NFIP Parquet/CSV Data

Purpose: Store the raw FEMA data in your Bronze zone.

Instructions:
1. Visit the dataset page: https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2.
2. Download the CSV file. Parquet had issues
3. In Fabric, open lh_nfip → Files.
4. Upload the Parquet file to Files/bronze/nfip_claims/ (create folder if missing).
5. Verify upload completion.

Checkpoint: File visible under Files/bronze/nfip_claims/.

## 4. Create/upload Fabric Notebooks

Purpose: Perform Silver and Gold transformations using notebooks.

Instructions:
1. Create a new notebook named nb_nfip_silver. Or you can import/upload the provided notebook.
2. Attach to lh_nfip Lakehouse.
3. Run each code cell in the nb_nfip_silver.py. Fix any issues using copilot assistance.
4. Repeat above steps for nb_nfip_gold using nb_nfip_gold.py.
5. Run each code cell in the nb_nfip_gold.py. Fix any issues using copilot assistance.

Checkpoint: Both notebooks created and linked in pipeline.

## 5. Validate Workspace Setup

Purpose: Ensure all objects are ready before proceeding.

Validation checklist:
- Workspace uw-agentic-ai created
- Lakehouse lh_nfip with bronze folder
- Notebooks nb_nfip_silver, nb_nfip_gold created & executed successfully
- At least one NFIP parquet file in /Files/bronze/nfip_claims/