Earthquake Data Engineering Pipeline with Microsoft Fabric

1. Introduction

Objective

This project aims to build a robust data engineering pipeline that leverages Microsoft Fabric's Data Factory, Data Engineering, and Power BI capabilities. The pipeline processes earthquake data sourced from the USGS API, implementing the Medallion architecture for efficient data management and insightful visualizations.

2. Project Structure

Bronze Layer

- **Description**: The raw data ingestion layer stores earthquake data in its original form directly from the USGS API.
- **Data Structure**:
  - **Type**: FeatureCollection
  - **Features**: List of earthquake events, each containing properties and geometry.
  - **Properties Fields**:
    - `mag`: Magnitude of the earthquake
    - `place`: Location description
    - `time`: Time of occurrence
    - `updated`: Last update time
    - `url`: Link to more details
    - `felt`: Number of reports from people who felt the earthquake
    - `cdi`: Community Internet Intensity Map value
    - `mmi`: Modified Mercalli Intensity value
    - `alert`: Alert level
    - `status`: Status of the event
    - `tsunami`: Tsunami warning
    - `sig`: Significance of the event
    - `net`, `code`, `ids`, `sources`, `types`, `nst`, `dmin`, `rms`, `gap`, `magType`, `type`, `title`
  - **Geometry Fields**:
    - `type`: Type of geometry
    - `coordinates`: Latitude, longitude, and depth

Silver Layer

- **Description**: This layer transforms and cleans the raw data, adding contextual information to make it structured and useful for analysis.
- **Data Transformations**:

- **Data Cleaning**: Handle missing values, replace nulls with defaults, or drop incomplete records.
- **Type Conversion**: Convert `time` and `updated` fields from epoch to human-readable datetime format.
- **Derived Columns**:
  - Extract latitude, longitude, and depth from coordinates.
  - Categorize magnitude into bins (e.g., minor, light, moderate, strong, major, great).
- **Data Enrichment**: Enrich data with the nearest city or country using geospatial coordinates.

Gold Layer

- **Description**: The final layer for business-ready, aggregated data. Optimized for queries and visualization in Power BI.
- **Schema Design**:
  - **Fact Table**: `Earthquake_Fact`
  - **Dimension Tables**:
    - `Place_Dimension`: Contains place details with `place_id` as the primary key.
    - `Magnitude_Category_Dimension`: Contains magnitude categories with `magCategory_id` as the primary key.
    - `Time_Dimension`: Contains time details with `full_date` as the primary key.

3. Pipeline Implementation

Tools Used

- **Microsoft Fabric's Data Factory**: Orchestration of data pipelines.
- **PySpark**: Data transformations, enrichment, and business logic application.
- **Power BI**: Visualization and analysis of data trends and insights.

Key Pipeline Steps

1. **Data Ingestion**: Ingest raw earthquake data from the USGS API to the Bronze layer.
2. **Data Transformation**: Clean and enrich data in the Silver layer.
3. **Data Loading**: Load the transformed data into the Gold layer for business analytics.

4. Data Model

Star Schema Design

- **Fact Table**: `Earthquake_Fact`
  - **Columns**:
    - `earthquake_id`: Unique identifier for each earthquake event.
    - `magnitude`: Magnitude of the earthquake.
    - `latitude`: Latitude of the earthquake's epicenter.

        - `longitude`: Longitude of the earthquake's epicenter.
        - `depth`: Depth of the earthquake.
        - `sig`: Significance of the earthquake.
        - `updated`: Last update time of the earthquake event.
        - `place_id`: Foreign key linking to `Place_Dimension`.
        - `magCategory_id`: Foreign key linking to `Magnitude_Category_Dimension`.
        - `full_date`: Foreign key linking to `Time_Dimension`.
- **Dimension Tables**:
    - **Place_Dimension**:
        - **Columns**:
            - `place_id`: Primary key.
            - `place`: Description of the earthquake location.
            - `country_code`: Country code where the earthquake occurred.
    - **Magnitude_Category_Dimension**:
        - **Columns**:
            - `magCategory_id`: Primary key.
            - `magCategory`: Category of magnitude (e.g., minor, light, moderate).
            - `sig_class`: Classification of significance.
            - `magType`: Type of magnitude measurement.
    - **Time_Dimension**:
        - **Columns**:
            - `full_date`: Primary key.
            - `year`: Year of the earthquake event.
            - `month`: Month of the earthquake event.
            - `day`: Day of the earthquake event.
            - `month_name`: Name of the month.
            - `week_name`: Name of the day in the week.
            - `quarter`: Quarter of the year.
- **Fact Table**: `Earthquake_Fact`
    - Contains core earthquake event data linked to dimensions by foreign keys.
- **Dimension Tables**:
    - `Place_Dimension`: Stores unique places.
    - `Magnitude_Category_Dimension`: Stores magnitude categories.
    - `Time_Dimension`: Stores time-related details such as year, month, and day.

5. Power BI Reports

Visualizations

- **Earthquake Frequency**: Frequency of earthquakes visualized by location and magnitude.
- **Geospatial Analysis**: Interactive maps showing earthquake locations and affected areas.

6. Code and Implementation

Bronze Layer Script

- **Functionality**: Ingest raw data from the USGS API and store it in the Bronze layer.
- **Key Steps**:
    1. Fetch data using API.
    2. Save raw data in the lakehouse.

Silver Layer Script

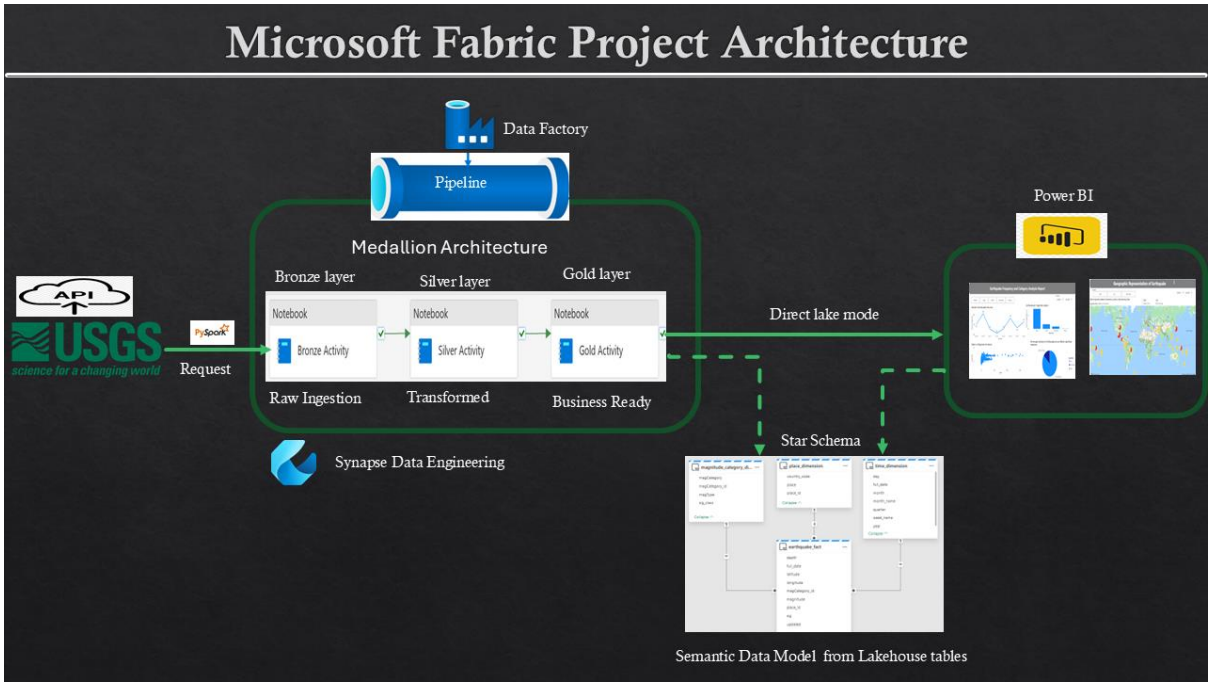
- **Functionality**: Clean and transform the raw data to produce structured, useful datasets.
- **Key Steps**:
    1. Load raw data from the Bronze layer.
    2. Perform data cleaning and type conversions.
    3. Add derived columns and enrich data.
    4. Save processed data to the Silver layer.

Gold Layer Script

- **Functionality**: Aggregate data for business-ready insights and update fact and dimension tables.
- **Key Steps**:
    1. Load cleaned data from the Silver layer.
    2. Perform additional transformations and aggregations.
    3. Save aggregated data to the Gold layer.

7. Project Architecture

Diagram



- **Description**: Diagram illustrating the flow of data from raw ingestion in the Bronze layer to business-ready insights in the Gold layer using the Medallion architecture.

8. Usage

Requirements

- **Environment**: Microsoft Fabric setup.
- **Tools**: PySpark, Power BI.
- **Data**: Earthquake data from the USGS API.

9. Conclusion

This project demonstrates an end-to-end data engineering pipeline using Microsoft Fabric. It showcases how raw data can be transformed into valuable business insights through effective data processing and visualization.