Azure Data Lakehouse: Global Air Quality & Health Analysis

Project Overview

This project implements a cloud-native "Lakehouse" architecture on Microsoft Azure to analyze the correlation between global Air Quality Index (AQI) levels and respiratory health statistics. The goal was to process high-volume environmental data without relying on traditional, expensive relational database infrastructure.

Architecture

Raw Data (CSV) -> Azure Data Lake Gen2 -> Azure Synapse Analytics (Serverless SQL) -> Power BI

- **Storage:** Azure Data Lake Storage Gen2 (Hierarchical Namespace enabled).
- **Compute:** Azure Synapse Analytics (Serverless SQL Pool).
- **Visualization:** Microsoft Power BI (Direct Query).

Technical Implementation

1. Data Ingestion & Storage

- Provisioned an Azure Storage Account with **Hierarchical Namespace** to enable file-system semantics.
- Ingested heterogeneous CSV datasets (EPA Air Quality & WHO Health Statistics) into a raw-data container.

2. Data Analysis (Serverless SQL)

- Utilized **T-SQL** and the OPENROWSET function to query data directly from the Data Lake.
- Implemented **Schema-on-Read** to handle data variety and clean formatting issues (UTF-8 encoding) on the fly.

3. Visualization

- Connected **Power BI** directly to the Data Lake using the Azure Gen2 connector.
- Performed data transformation using **Power Query** (M Language) to split text-based statistics into quantifiable metrics.
- Created geospatial visualizations to map mortality rates against high-pollution zones.

Key Insights

- Identified a direct correlation between regions with AQI > 150 and increased respiratory mortality rates.

- Demonstrated that Serverless SQL architectures can reduce standing cloud costs by over 80% compared to dedicated SQL servers for ad-hoc analysis.

Visuals

(Note: Screenshots of the architecture, SQL query results, and Power BI dashboard are located in the /screenshots folder)