Architecture Diagram Documentation

Overview

The COVID-19 Prediction and Reporting Solution is designed to automate the ingestion, transformation, storage, and analysis of COVID-19 and population data. It leverages multiple Azure services to create a scalable, secure, and highly automated pipeline, from data ingestion to reporting and visualization.

Key Components

1. Data Sources

-ECDC API: Provides COVID-19-related data including daily/weekly confirmed cases, deaths, hospital admissions, and testing numbers.

-Eurostat: Supplies each country's demographic information (population by age group) to help contextualize the COVID-19 data for per capita calculations.

Data is ingested via Azure Data Factory (ADF) and stored in Azure Blob Storage for processing.

2. Azure Blob Storage (Raw Data Storage)

-Purpose: This is the first stage where raw data (e.g., from ECDC and Eurostat) is ingested and stored.

-Description: Incoming data is stored in its original form before being moved to Azure Data Lake Storage Gen2 for further processing.

3. Azure Data Factory (ADF)

-Role: Orchestrates data ingestion, transformation, and movement between services.

-Data Pipelines:

-Population Data Pipeline: Ingests demographic data (population by age group) from Azure Blob Storage.

- **COVID-19 Data Pipeline**: Retrieves COVID-19 cases, deaths, hospitalizations, and testing data from the ECDC.
- **SQL Copy Pipeline**: Moves aggregated and transformed data from Azure Data Lake Gen2 into Azure SQL Database for querying.

4. **Azure Data Lake Gen2 (Data Repository)**
  - **Purpose**: Stores both intermediate and processed data. The data lake is used for efficient, scalable data storage.
  - **Data Stages**:
    - **Raw Data**: Data from Azure Blob Storage is ingested here in a structured form.
    - **Processed Data**: After transformation in Azure Databricks and ADF pipelines, the cleaned and aggregated data is stored here.

5. **Azure Databricks**
  - **Role**: Executes data transformations, cleaning, and aggregation tasks. Data is prepared for downstream machine learning models and reporting.
  - **Key Transformations**:
    - Cleanses and formats raw data.
    - Aggregates COVID-19 data to weekly values (when necessary) and normalizes by population.

6. **Azure SQL Database**
  - **Purpose**: Stores transformed, structured data for fast querying and reporting. This data is queried by Power BI for trend analysis and predictive modeling.
  - **Data**: Includes aggregated COVID-19 case data, population data, testing positivity rates, hospital and ICU occupancy, and predictive model results.

7. **Power BI (Visualization)**
  - **Role**: Visualizes trends, predictions, and key insights from the structured data stored in Azure SQL Database.
  - **Dashboards**:
    - **COVID-19 Trends**: Daily and weekly reports of new cases, deaths, and hospitalizations.
    - **Prediction Dashboards**: Displays future trends of COVID-19 cases and healthcare resource utilization using machine learning models.

Data Flow Overview:
1. **Ingestion**:

   - ECDC and Eurostat data are ingested through ADF pipelines and stored as raw files in Azure Blob Storage.

2. **Transformation**:
   - Data is cleaned, formatted, and aggregated in Azure Databricks, and then stored in Azure Data Lake Gen2 as structured datasets.

3. **Storage**:
   - Cleaned data is moved from Data Lake Gen2 into Azure SQL Database, where it is ready for analysis and reporting.

4. **Orchestration**:
   - ADF pipelines schedule and automate the entire data flow, from ingestion to reporting.

5. **Visualization**:
   - Power BI dashboards pull data from the Azure SQL Database to visualize COVID-19 trends, predictions, and key public health metrics.

Security Considerations
- **Role-Based Access Control (RBAC)**: Ensures that only authorized users can access or modify sensitive resources such as storage accounts and databases.
- **Encryption**: Data is encrypted both in transit and at rest to ensure security.
- **Managed Identities**: Used for authentication across Azure services, eliminating the need for storing sensitive credentials.
- **Network Security**: Private endpoints are used to limit access to Azure services, ensuring that only internal traffic can interact with storage and databases.

Monitoring and Alerts
- **Azure Monitor**: Provides real-time monitoring and alerts for pipeline health, data ingestion failures, and other performance issues.
- **ADF Monitoring**: Monitors individual pipeline runs and allows re-execution of failed pipelines.
- **Power BI Dashboards**: Displays real-time visualizations of data flow, pipeline status, and any potential anomalies.

CI/CD Integration
- **Continuous Integration/Continuous Deployment (CI/CD)** is implemented using Azure DevOps. Pipelines are version-controlled in a Git repository, and changes to ADF, Databricks, and SQL Database deployments are automatically deployed through ARM templates.

This documentation explains how the Azure components work together to provide an automated solution for COVID-19 prediction and reporting. The architecture can be enhanced with detailed process flow diagrams, pipeline details, and service configurations to complement the high-level overview.