Technical Exercise Part II

This technical exercise requests a concept extract, transform and load (ETL) diagram, and an explanation of how the data will be migrated between the Climate Watch Data database and the Country Climate and Development Reports (CCDRs) database. The proposed ETL Concept Architecture is displayed below and involves three key services, namely the Data Ingest Service, the Data Storage Service, and the Data Access Service. A screenshot of a computer

Description automatically generatedThe data ingest service utilizes the Apache Airflow platform to schedule, orchestrate and monitor ETL data pipelines. In particular, Apache Airflow orchestrates the directed acyclic graphs (DAGs) which are scheduled to run at set times, for example each day at 6am UTC, that (i) extract the data and metadata from the various data sources, for example Climate Watch Data, (ii) clean and transform the ingested data taking into account the mappings between data from various sources and data in the CCRD database, for example mappings to aggregate data on sectoral greenhouse gas emissions across nations to generate World Bank unique country groups, and (iii) utilize Apache Arrow to generate Apache Parquet files of the transformed data and associated metadata for storage at the data storage service.

The data storage service utilizes Azure Data Lake Gen 2 for storage of the Apache Parquet files generated by the data ingest service. The Apache Parquet files are ideally stored in a hierarchical folder structure for ease of navigation and with filenames that detail the data source and time sourced, for example “/<year>/<month>/<source>-<rfc3339-date>”.

The data access service utilizes Apache Iceberg to ingest the data and metadata from the Azure Data Lake at scheduled times and allows SQL queries on the data at specific points in time, for example data may be queried to retrieve the latest version or the version from a specific time – for example the beginning of 2024, i.e. 2024-01-01T00:00:00.000Z. The files stored in the Azure Data Lake are available for ingest into World Bank systems, for example the CCDR database, via rest APIs that retrieve data and metadata from the Azure Data Lake via Apache Iceberg, and optionally also allows visualizations of the data in Apache Iceberg via dashboards created in Apache Superset.