**Online Retail Sales Data Engineering**

**Dataset**: Obtain a [sample online retail sales dataset](https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset/download?datasetVersionNumber=2) with information about products, customers, and sales transactions.

**Tasks:**

**Data Ingestion:**

* Implement a data ingestion process to fetch the dataset from its source (e.g., CSV file, database).
* Load the dataset into a suitable storage system (e.g., relational database, data warehouse).

**Data Cleaning and Validation:**

* Develop a data cleaning pipeline to handle missing values, duplicates, and outliers.
* Validate the integrity and consistency of the data during the cleaning process.

**Data Transformation:**

* Transform the data into a structured format suitable for analytical processing.
* Create a pipeline to convert relevant columns to appropriate data types.
* Implement any necessary data standardization or normalization.

**Data Warehouse Design:**

* Design and create a data warehouse schema to support efficient querying and reporting.
* Define tables for products, customers, and sales transactions.
* Establish appropriate relationships between tables.

**ETL (Extract, Transform, Load) Process:**

* Build an ETL pipeline to automate the extraction, transformation, and loading of data.
* Schedule the ETL process to run at regular intervals (e.g., daily, hourly).

**Partitioning and Indexing:**

* Implement partitioning and indexing strategies to optimize data retrieval performance.
* Evaluate and choose suitable columns for partitioning and indexing.

**Data Quality Monitoring:**

* Set up monitoring mechanisms to track data quality over time.
* Establish alerts for anomalies or issues detected during the ETL process.

**Data Versioning:**

* Implement a versioning system for the dataset to track changes and updates.
* Archive previous versions of the dataset for historical analysis.

**Metadata Management:**

* Create a metadata repository to document the structure and lineage of the dataset.
* Include information on data sources, transformations, and storage locations.

**Performance Optimization:**

* Optimize the performance of data retrieval queries by analysing query execution plans.
* Implement caching mechanisms for frequently accessed data.

**Deliverables:**

* Python or SQL scripts for each step of the ETL process.
* Documentation outlining the data warehouse schema, ETL pipeline, and data quality monitoring.
* Recommendations for further optimizations and improvements.