**Project Flow: Real-Time Customer Interaction Tracking System**

**1. Data Collection and Ingestion**

* **Sources**: Capture real-time customer interactions from various touchpoints (e.g., call centers, web portals, mobile apps, social media).
* **Process**: Use **Azure Data Factory** to ingest data streams from these sources in real time.
* **Output**: Load raw interaction data into a staging area in **ADLS Gen 2** for initial processing.

**2. Real-Time Data Processing**

* **Platform**: Set up **Azure Databricks** for real-time processing.
* **Process**:
  + Use **PySpark** to process incoming data streams.
  + Cleanse and normalize the data (e.g., remove duplicates, standardize formats).
  + Perform transformations such as timestamp adjustments, interaction tagging, and initial aggregation.
  + Enrich data with customer metadata (e.g., demographic or subscription details) using **Python scripts**.

**3. Data Enrichment and Tagging**

* **Process**:
  + Use **Python** scripts within **Databricks** to tag interactions (e.g., issue type, sentiment) and enrich them with context (e.g., service tier, geography).
  + Ensure data tagging aligns with KPIs defined in collaboration with business analysts.
* **Output**: Store enriched and tagged data in a curated layer in **ADLS Gen 2**, partitioned for efficient querying.

**4. Real-Time Analytics and Monitoring**

* **Process**:
  + Enable low-latency querying of enriched data for real-time dashboards or alerts.
  + Integrate with visualization tools or alerting mechanisms to monitor key performance indicators (KPIs) such as call resolution time, interaction sentiment, or customer satisfaction scores.

**5. Performance Tuning and Optimization**

* **Cluster Management**:
  + Optimize **Azure Databricks clusters** for handling real-time processing workloads efficiently.
  + Ensure that cluster configurations minimize latency and processing costs.
* **Job Optimization**:
  + Tune **PySpark** jobs to improve performance, focusing on parallelism, partitioning, and caching.

**6. Collaboration and Validation**

* **Stakeholder Engagement**: Collaborate with business analysts to ensure data pipelines meet reporting and analysis needs.
* **Validation**: Implement automated validation scripts to ensure data integrity and consistency across the pipeline.

**High-Level Tools and Techniques Used**

* **Azure Data Factory**: Real-time data ingestion and orchestration.
* **Azure Databricks**: Real-time processing and enrichment with **PySpark** and **Python scripts**.
* **ADLS Gen 2**: Scalable data storage for processed data.
* **Python**: Data enrichment and tagging.