MENARINI
ASIA-PACIFIC

Case Study for Senior Data & AI Engineer

Company A has rapidly expanded into the Chinese market, generating a significant volume of diverse data sources beyond its traditional ERP system. The Regional Business Intelligence team is tasked with providing timely and accurate daily and weekly reports to support strategic decision-making.

As the newly appointed Senior Data & AI Engineer, you are responsible for designing and implementing a robust data pipeline to efficiently collect, process, and store this critical data.

Key Tasks:

1. **Automated Data Ingestion:**
   - Develop an automated workflow to extract data from various sources, including portals, emails, and other non-API endpoints.
   - Implement error handling and retry mechanisms to ensure data reliability and completeness.
   - Consider scheduling and orchestration tools to automate the data pipeline.

2. **Data Transformation and Enrichment:**
   - Design and implement a comprehensive ETL process to clean, transform, and enrich the extracted data.
   - Address data quality issues such as missing values, inconsistencies, and outliers.
   - Harmonize data from multiple sources, ensuring data consistency and integrity.
   - Create a data flow diagram to visualize the ETL process and its dependencies.

3. **Data Modeling and Storage:**
   - Develop a data model that effectively represents the business domain and supports reporting and analytics needs.
   - Design and implement the data lake architecture to store the processed data in a scalable and accessible manner.
   - Ensure data security and privacy by implementing appropriate access controls and encryption mechanisms.

4. **Data Governance and Documentation:**
   - Establish data governance practices to maintain data quality, consistency, and compliance.
   - Document the data pipeline, ETL processes, and data models for future reference and audit purposes.
   - Create clear and concise documentation to facilitate knowledge sharing and collaboration.

Deliverables:

- A detailed technical design document outlining the data pipeline architecture, ETL processes, and data model.
- A working prototype of the data pipeline, demonstrating the extraction, transformation, and loading of data into the data lake.
- A presentation summarizing the technical approach, key challenges, and solutions.
- A comprehensive data dictionary and lineage documentation.

Constraints:

- You have **72 hours** to complete this task.
- The provided dataset contains a variety of data formats and sources.