Assignment

Scenario: As a Senior Data Engineer at "DataCo," an e-commerce company, your objective is to design and implement a robust ETL pipeline to handle data ingestion, transformation, and loading from diverse social media and news sources. The data will be delivered as JSON format through APIs from platforms such as Twitter, Facebook, Instagram, TikTok, YouTube, blogs, forums, LinkedIn, etc.

Requirements:

1. **Data Sources (APIs):**
   - Identify and integrate APIs from the specified social media and news platforms to simulate data sources.
   - Extract sample JSON data representing various aspects, such as posts, comments, likes, shares, articles, etc.
2. **ETL Processing:**
   - Design an ETL process that dynamically accommodates varying JSON structures from different APIs.
   - Implement the ETL process using a language or framework of your choice, emphasizing modularity and extensibility.
3. **Data Pipeline:**
   - Construct a data pipeline architecture that automates the ETL process for seamless integration of data from diverse sources.
   - Implement mechanisms for handling API rate limits, retries, and failures.
4. **Data Ingestion:**
   - Develop a data ingestion mechanism capable of ingesting real-time data from social media platforms and batch data from news sources.
   - Consider scenarios where data arrives in bursts or steadily throughout the day.
   - The data has to be ingested in AWS open search
5. **Scalability and Performance:**
   - Optimize the ETL process and data pipeline for scalability and performance given the dynamic nature of social media data.
   - Discuss strategies to handle varying data volumes and spikes in activity.
6. **Documentation:**
   - Provide comprehensive documentation detailing the integration of each API, the ETL process, and the overall data pipeline.
   - Include clear instructions for maintaining and updating API connections.
7. **Testing:**
   - Develop test cases to ensure the accuracy and reliability of the ETL process and data pipeline.
   - Include tests for different JSON structures and variations across platforms.
8. **Security Considerations:**
   - Address security considerations in the design and implementation, especially when handling sensitive data from social media platforms.

Submission:

- Share the codebase via a version control system (e.g., GitHub).
- Include a README file explaining the architecture, setup instructions, and any additional information.

Evaluation Criteria: Candidates will be evaluated based on their ability to:

- Effectively integrate and handle JSON data from diverse APIs.
- Design and implement a flexible and scalable ETL process.
- Build a resilient data pipeline considering API dynamics.
- Efficiently handle real-time and batch data ingestion.
- Optimize for scalability and performance in a dynamic data environment.
- Provide thorough documentation and testing.