Data Engineering: Data Pipeline and API Challenge

Overview

Develop a scalable data pipeline that ingests data from an SFTP location, processes it for use in analytics, and provides access through an API with date-based filtering and cursor-based pagination.

Background

You are tasked with creating a data pipeline that operates in an on-premise server environment. This pipeline must:

1. **Ingest data** from an SFTP location.
2. **Process and clean data** for consistency and structure.
3. **Expose an API** for external clients to access the processed data with support for date filtering and cursor-based pagination.

Requirements

1. **Data Ingestion**
    - Retrieve data files from an SFTP location.
    - Handle various data formats (e.g. CSV & JSON).
    - Implement error handling for failed data transfers with alerts.
2. **Data Processing**
    - Flatten and clean the ingested data, as necessary. We want to see curated datasets for usage and consumption by business users at the end of the pipeline process.
    - Apply any necessary data quality measures to get the data into a state ready for business consumption.
3. **API Development**
    - Create an API that allows external clients to access the processed data.
    - Implement filtering by date and cursor-based pagination to handle large datasets.
    - Design the API to ensure basic security and rate limiting.

Deliverables

1. **Working Data Pipeline**
    - Provide a script or automated process that ingests data from the SFTP location and processes it.
    - Ensure the pipeline is robust and can handle data inconsistencies.
    - Curated datasets for usage and consumption by business users

2. **API with Documentation**
    - Develop a functional API that allows data retrieval with date-based filtering and cursor-based pagination.
    - Provide documentation detailing how to use the API, including authentication and rate limiting.

Guidelines

1. Data isn't provided for the exercise. You can use anything that makes sense as a representative data source. Choose and present your source of data for this exercise.
2. An SFTP server/location isn't provided. Use your best judgement and skills to demonstrate how you would handle this in a real-world situation.
3. Where you lack context or where something might be unclear, make assumptions and state them clearly and the exercise will be evaluated accordingly.
4. Submit your solution via GitHub or GitLab or BitBucket (or any other version control platform you prefer). We'll be evaluating your approach to building software iteratively using version control as well as the final solution.
5. Keep it simple.
6. Have fun with it.