ETL Pipeline Planning Questions

2025-02-11

Table of contents

1 Data Source Access

1.0.1 How will you handle different API rate limits across these diverse systems?

Answer

We'll implement a centralized rate limiting solution using DynamoDB to store and manage rate limits for each ERP system. Our approach includes:
A DynamoDB table storing each ERP's configuration including:
Rate limits and burst limits Contact information Alert thresholds Version history Approval status
AWS API Gateway for enforcing these limits, with usage plans dynamically configured based on the DynamoDB data. A Lambda function that syncs DynamoDB configurations with API Gateway usage plans, ensuring rate limits are always up to date. CloudWatch monitoring to track API usage against limits, with automated alerts at configurable thresholds (e.g., 80% warning, 95% critical). An admin API for managing rate limits with an approval workflow for any changes.
This solution provides centralized management, version control, audit trails, and automated monitoring - all essential for managing multiple ERP integrations at scale."

1.0.2 What's your backup plan for ERPs that don't provide API access?

Answer

1.0.3 How will you maintain and secure credentials for 30+ different systems?

Answer

1.0.4 How will you handle API/connection failures for real-time synchronization?

Answer

2 Data Standardization

2.0.1 Have you mapped the schema differences between all these ERPs?

> Answer

2.0.2 How will you handle inconsistent field names/data types across systems?

> Answer

2.0.3 What's your plan for handling different date/time formats from various systems?

> Answer

2.0.4 How will you maintain data lineage tracking across these diverse sources?

> Answer

3 Processing & Storage

3.0.1 What's your strategy for handling late-arriving data?

> Answer

3.0.2 How will you handle schema evolution in your Iceberg tables?

> Answer

3.0.3 What's your plan for data validation before loading into S3?

> Answer

3.0.4 How are you handling data versioning and rollbacks?

> Answer

4 Operational Concerns

4.0.1 What's your monitoring strategy for pipeline failures?

> Answer

4.0.2 How will you handle partial load failures?

> Answer

4.0.3 What's your data retention policy in the drop zone?

> Answer

4.0.4 How will you manage pipeline dependencies between different ERP loads?

> Answer

5 Additional Questions

5.0.1 What are the data sync frequency requirements for each system?

> Answer

5.0.2 Are there any specific SLAs for data freshness that need to be met?

> Answer