Part 1: Data pipeline

Design a data pipeline that ingests data from a MongoDB production database into BigQuery for our data warehouse. The pipeline should be flexible, maintainable, and should clean up data for access by data analysts.

Requirements:

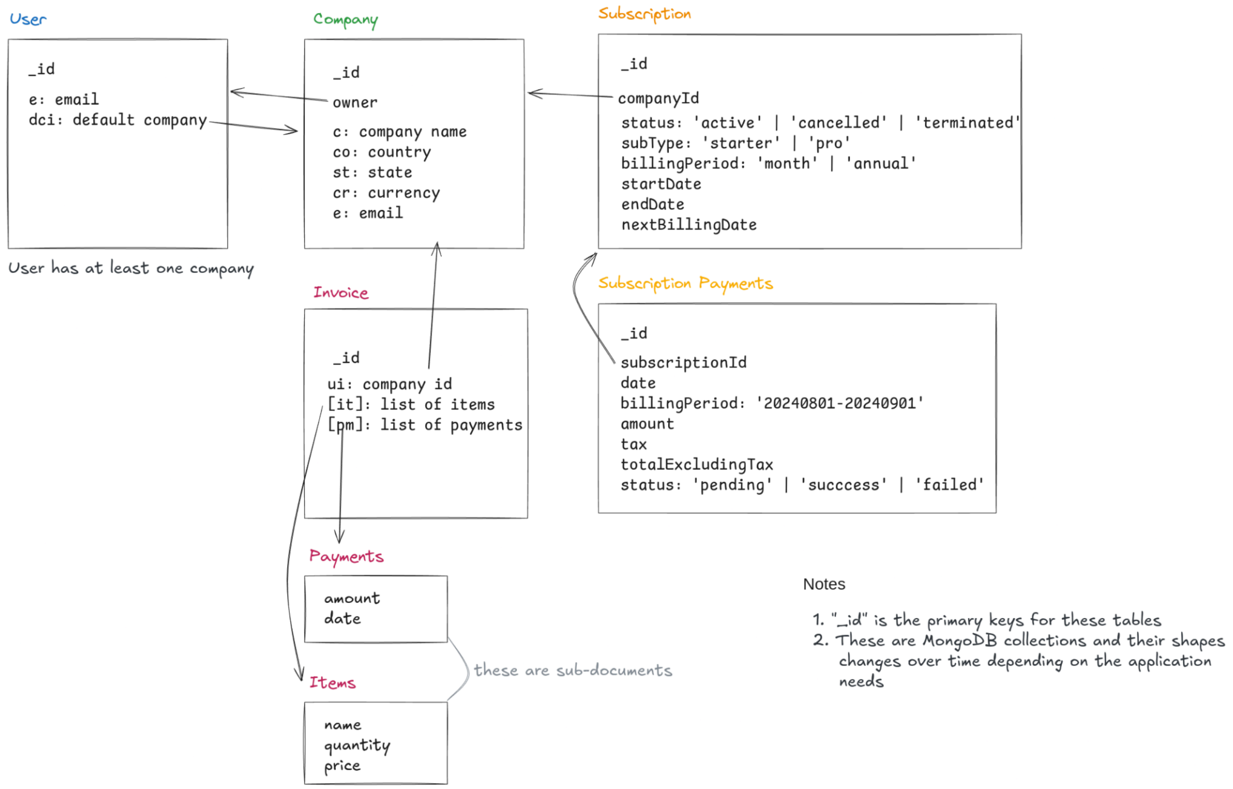
1. Flexible Data Ingestion: Design a data pipeline to ingest and process data into BigQuery.
2. Data Cleanup: Ensure data is systematically cleaned. For instance, replace shorthand notation like "e" for emails with a full column name like "email".
3. Maintenance: The pipeline should be maintainable and adaptable to slow changes in data over time.

Business model

Invoicing Solution: Users can create and issue invoices. They can create up to 4 invoices per month for free but need a subscription for more.

Subscription Status: Subscriptions can be "active", "cancelled", or "terminated".

- **Active**: Unlimited invoices.
- **Cancelled**: Subscription will terminate at the end of the current billing period.
- **Terminated**: No longer entitled to unlimited invoices.



Notes

1. "_id" is the primary keys for these tables
2. These are MongoDB collections and their shapes changes over time depending on the application needs

Part 2: PySpark transformation

Provide a PySpark script to parse the billingPeriod attribute in the "Subscription Payments" table and separate it into date fields suitable for analytical processing. Explain how this would be implemented in the pipeline.

Requirements:

- **Script:** Write a PySpark script to parse the billingPeriod attribute into start and end dates.
- **Implementation:** Illustrate how to integrate this script into the pipeline.

Part 3: Data requests

A data analyst needs a weekly churn analysis report. The key metrics are:

- Number of new subscriptions per week.
- Number of subscriptions cancelled per week.

- Number of subscriptions terminated per week.

Requirements:

- Report Creation: Outline how to build and run this report on a weekly basis.