UDACITY

‹ Return to Classroom

Build a Scalable Data Strategy

REVIEW

HISTORY

Meets Specifications

Congratulations Student 🏆
Very well performed in this project submission by successfully incorporating all rubric requirements 👏

All the Best for future submissions 👍🏻

Data Collection

The project recognizes at least 3 stakeholders.

The project clearly states why these stakeholders were chosen and identified as primary stakeholders.

The project identifies stakeholders in at least 3 different departments of the business

Hint: Flyber is a startup. The app is live so an engineering team is there. They are one of the primary stakeholders. On similar lines who could be primary stakeholders. What are primary focus areas for a startup?

Already pass✅

The project should have at least 1-2 use-cases for each of the stakeholders identified.

The project clearly states why these use-cases were chosen and identified as primary use-cases for the stakeholders.

Example: Engineering would like to monitor the traffic so that they can respond to scaling needs of app quickly. They need data that can help them with it.

Already pass✅

The project should identify at least 2 data fields required for each of the above mentioned use-cases.

The project clearly states why these data fields were chosen for each use-case.

Optional: Project may identify fields that would be good to include, but not required.

Project has data fields for each of the use-case recognized. These are mostly the must-have data fields for the relevant use-case.

Data Modeling

Project has at least 3 tables defined for the data requirements gathered.

Project clearly states why these tables were chosen and how they connect to Flyber's stakeholders primary use cases.

Hint: These tables are a way to organize data elements identified for the MVP use-cases in the previous exercise.

3 tables defined. Somewhat logical organization of data. Some of the data fields defined in section 1 are not used but this works

Project should have Normalized (no redundancy) tables.

Project has Primary Keys and Foreign Keys identified for each of these tables.

Project states why these Primary and Foreiegn key identifiers were chosen over other potential identifiers.

Tables are normalized, no redundancy.✅

Project has Primary and Foreign keys identified for each of the table.✅

Project addresses why the fields for Primary and Foreign Keys were chosen for each table.✅

A data field that is unique and required for that table is marked as Primary Key.
Primary Key helps to identify a record in the table. Example: In customer table the ID of the customer will be unique, customer can have the same name, multiple customers in a family might use same email or phone number. Where as in Transaction table Transaction ID could be a unique identifier.
Foreign Keys could be any field in a table that is not the Primary key of that table but is primary key of some other table. Foreign keys help in establishing relationships between tables to run complex queries for analytics.

analytics.

Extraction and Transformation

The project includes steps to transform data into a format that can be used for further analysis and visualization.
The project clearly states why this format and tool is chosen.
Hint: This file can be converted using excel to a relevant format.

Already pass✅

The project has details on how data was loaded and transformed to answer the following questions:

1. How many events are being recorded per day?
2. How many events of each event type per day?
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

Already pass✅

Project addresses the following points to summarize data processing needs of Flyber:
Was the process of manually Extracting, Loading and Transforming data from raw logs efficient? Is it scalable?
Is there a need for automated ETL pipeline? Why?

You have answered these questions with reasoning

Choosing Relevant Dataset

The project selects one correct criteria that provides the most relevant information from the following to get data from Engineering:

1. How many events are being recorded per day?
2. How many events of each event type per day?
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

The project analyzes criteria selection considering how it can answer the following questions:

1. How much is the customer data increasing?

1. How much is the customer data increasing?
2. How much is the transactional data increasing?
3. How much is the event log data increasing?

Hint: Only one of the criteria will provide the most information around the above dimensions.

Loading and Visualization

The project contains at least 2 continuous line visualizations for two of the event types. And answer: What do these graphs tell?

The project clearly defines the steps taken to generate continuous line visualizations.
Optional: Project creates other relevant data visualizations to add to the dashboard such as all event types on a logarithmic scale.

Already pass✅

Data Analysis: Telling Stories with Data

Project has some estimation calculation to answer - “By how many times have the event logs grown in the last 1 month?”

Project has a labeled graph to accompany the answer of the above question.

Good job
These graph helps understand that calculations make sense.

Project analyzes data to answer which of the data types is growing at the fastest rate? Analyze around :

- Event logs
- Transactional data
- Customer data

Project has logical reasoning to explain why the data types are growing fastest.

Project has logical reasoning to explain why the data types are growing fastest.

The project has one graph created that shows all the different event types.

The project analyzes the pattern of different event types. To answer the following questions:

1. Are all graphs following the same pattern?
2. Why is this good or bad for the business?
3. There was a marketing campaign run in the first week of October. Did it impact data generation?

4. **What does this tell us about marketing campaigns in terms of impact on data generation?**
5. **Why is it important to know?**

Hint: Use logarithmic scale

Awesome
You have provided some logical reasoning and Yes marketing campaigns impact data.

Data Warehouse

The project proposes either a Cloud or on-premise solution for their Data Warehouse.

The project analysis addresses the following aspects when choosing a warehouse:

1. **Cost**
2. **Scalability**
3. **Expertise needed**
4. **Latency**
5. **Reliability**

The project uses information from the analysis done in previous sections to support the Data Warehouse chosen.

Optional: Project suggests a hybrid strategy and addressing all of the items above

Nice
Now the reasoning and comparison looks good

Example: We want to go with a cloud solution since:

- We can easily monitor and control cost with new cloud technologies. We can scale up and down and do not have to pay for the infrastructure when we scale down.
- In cloud we can easily auto scale, we can scale up and down as per our requirements. As we can clearly see we need to scale up near our marketing campaigns but do not have to maintain that forever. * We can scale down a bit once the campaign is over.
  Currently, we want to focus on our core business and offering. We would like to use the expertise easily available as a service. If we use cloud solutions we will not have to reinvent the wheel. And won't need to hire expensive experts, rather we can use readily available support and services offered by various cloud solution providers.
- How cloud solutions are used will define if these can be used for real-time use cases. If we end up having some data on-prem and some data on cloud, supporting real-time use cases can be tricky. Currently, we do not have use cases that need real-time data. Latency is not the most important factor for us right now. Hence we can go cloud route.
- Cloud solutions are highly available. In cloud, if one region fails we fail back to other. If we do on-premise we will have to invest in building all that infrastructure ourselves.

Project makes suggestion around a specific Data Warehouse solution along with a strong reasoning to

Project makes suggestion around a specific Data Warehouse solution along with a strong reasoning to support its choice.

Pick from following:

Cloud- Amazon Redshift, Google BigQuery, Snowflake or Microsoft Azure

On-prem- Oracle Exadata, Teradata, Vertica, Apache or Hadoop

Project has data/information used from the analysis done in previous sections.

Optional: Project suggests a hybrid data strategy and addresses the relevant products and provides reasoning on the choices made.

Good job here
Project makes suggestion around a specific Data Warehouse solution along with a strong reasoning to support its choice.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START