UDACITY

‹ Return to Classroom

Build a Scalable Data Strategy

REVIEW

HISTORY

Meets Specifications

Udacity student,

your project shows how committed you are to the course. You probably spent hours or days on this project and really should be proud of your work here and I'm proud of being part of your journey!

Keep the great work on the next sections of this amazing Udacity course!!! I share with you some extra links from Medium:

What is Minimum Viable (Data) Product?

MVP in Agile Projects: What, Why and How

This last one is not for only a job interview, but it contains a lot of useful information about some great topics for a data professional:

Data Science and Machine Learning Interview Questions

😄

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?

What makes your primary internal stakeholders and their use-cases important

Great job listing the three stakeholders as requested.

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.

What makes the data collection to your primary stakeholders great

Awesome use-cases for each team. You have shown a complete understand about the core business. Having the right data for each stakeholder is crucial for a business success.

A great article will help you the further understand the data needs for each area and why they are really important.

WHY IS DATA IMPORTANT FOR YOUR BUSINESS?

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.

What makes the data collection to your primary stakeholders great

Great, your 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.

The importance of your tables in your project

You have provided an example of tables that are relevant for the MVP.

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.

The importance to set Primary and Foreign keys in your tables

Great, your project has Primary and Foreign keys identified for each of the table.

I would like to share an article regarding to database keys:

The Purpose and Use of Primary and Foreign Keys in Databases

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.

Data Extraction and Transformation

Great job! 😉

You have exemplified all the steps you need to extract and transform the data in order to initiate all necessary analysis.

As a further information, you can read more about ETL Process

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?**

After ETL process, we need to analyse the data

Great job answering these questions. Depending on the way you evaluated you took a long time or more work to do the calculations.

Using a jupyter notebook could make things easier or even a BI tool.

Data Analysis - Example

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?

ETL Automation Process

Awesome answer based on ETL Automation and Scalability.

ETL processes can be very complex and sometimes really time consuming. But in the following article you can read some tips of how to build ETL process with examples.

3 Ways to Build An ETL Process with Examples

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?
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.

Choosing the right data to answer the question

Great job finding and choosing the right dataset.

✓ How many events of each event type per day?

You could also use:

✓ How many events per page type per day?

Both questions might lead you to answer the question related to the number of events.

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.

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.

Business Insights - Growth

Great job! You have addressed all important points in your answer.

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.

Business Insights - Data Types Rate

Great! Your answer meets the specifications. You have created a nice outline to explain your conclusions.

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

Business Insights - Marketing Campaign

Great job! 😄

The marketing campaigns impact data, more data is generated. There should always be additional capacity available around campaign days. Hence it is important that proper capacity planning is done considering marketing campaigns.

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

Data Infrastructure Strategy - Cloud vs On-Premise

Good choice. I would also go with Cloud. It is faster, reliable and a cost effective solution.

You will find a nice introduction in this article.

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.

Data Infrastructure Strategy - Cloud vs On-Premise

Awesome! You provided a complete argument for your choice. The three aspects below are the most important and your addressed them correctly.

- Pricing;
- Scalability
- Maintenance

DOWNLOAD PROJECT

RETURN TO PATH