Flyber Data Strategy MVP

Introduction

Flyber has been massively successful. Results have beaten expectations and projections! This is good news for Flyber, but now it's time to plan for what's next. With success came some challenges. While we were able to grow, the original data pipelines to receive and process data are unable to keep up with the current and future growth.

As a Data Product Manager, working with multiple teams and stakeholders is imperative to success. To understand what our needs are, what scale we are growing at, and how we can build for the future, we need to consider all relevant stakeholders. In this proposal, present your findings along with the analysis and reasoning behind the choices made in order to help Flyber continue its success.

Section 1: Data Customers & Needs

Flyber is a two-sided platform. You have customers who are riders, and you have partners who are drivers/pilots (think Uber: riders and drivers). For the Minimum Viable Product, you will be focusing on the Riders side of the business. To build an end to end data pipeline the very first step is to understand who needs data and why they need that data. Within Flyber, identify who your primary data customers/stakeholders are, why they are your primary data stakeholders and how they want to use the data (primary use-cases).

Identify your primary internal stakeholders and their use-cases:
(You may add more rows if necessary.)

Stakeholder	**Why are they primary stakeholders?**	**Use-Case**
Sales & Marketing	Sell the service and make estimates for the future + Acquiring new customer and retaining existing users	Targeted advertising/promotions Reactivations after lull Customer profile and preferences

Accounting	Recording expenses and earnings on the go	Monitoring expenses and earnings
Finance	Managing the companies' financial health and P&L growth	Monitoring P&L BI & Visualization Tool ML engine for future finance prediction
Engineering	Design and creates the service, product	Monitoring site and app performance Data generated by site and app Dashboard & Alerts
Product Management	Manages the life cycle of the project, prioritise features and roadmap.	Monitoring feature roll-outs and adoption Identify opportunities from user behavior Customer interaction data for site and app Identifying customer pain points
Security and Risk	Secure customer payment information, their personal data and physical security	Fraud detection Ensuring PII, PCI DSS compliance
Customer Care	First point of contact for customers to the company	Monitoring issues around bugs, customer satisfaction. Provide personalized responses Customer360 reports

Section 2: Data Collection and Data Modelling

To support our primary stakeholders's use-cases we need following data:
(You may add more rows if necessary.)

Stakeholder	**Use-Case**	**Data**	**Why is this the primary use-case?**
Sales & Marketing	Targeted advertising	Entity Data: 1. Customer name, 2. Email, 3. phone, 4. address, 5. customer ride history	Helps acquire new customers and retaining existing users
Finance	Monitoring P&L	Entity Data: 1. Aggregated Transactional Data 2. Number of Rides 3. commissions, 4. Ride Costs, 5. Taxes, 6. Tips, if any.	Managing burn-rate, cash flow is important to any fast growing start-up
Engineering	Monitoring site and app performance	Event Data: 1. Event ID, 2. Timestamp, 3. Event type	Managing app performance and outages
Product Management	To understand conversion funnel	Entity Data: 1. Customer name 2. Customer Email	Team can redesign UI/UX to improve conversion leading to increase bookings

		3. No of bookings made by customer Event data: 1. Customer ID 2. Page Type 3. Timestamp	
Customer Care	Provide personalized responses to the customer.	Entity Data: 1. Customer name, 2. Email Event Data: 1. Customer ID 2. Case number 3. Case log 4. Case summary	Addressing customer grievances

The tables we need are:

Note: As a best practice, we should establish these relationships between tables from the very beginning. To complete this exercise we will focus on fundamental concepts of relational databases - tables, normalization and unique keys. Please provide the table header row for each table, tables might be different lengths. Make sure you include the following for each table. You can create as many tables as you feel are necessary (copy and paste from one of the table sections):

Data Modeling

Table 1:

Customer_Table

Customer_ID	*First_name*	*Last_name*	*Address*	*Email*	*Contact*

Primary Key: Customer_ID

Table 2:

Customer_Demographics

Customer_ID	*Age*	*DOB*	*Gender*	*Income Bracket*	*Marital Status*

Primary Key: Customer_ID

Table 3:

Rides

Ride_ID	*Customer_ID*	*Trip_Est*	*Trip_Cost*	*Taxes*	*Tips*	*Copter_ID*

Primary Key: Ride_ID

Foreign Key: Customer_ID, Copter_ID

Table 4:

Copter

Copter_ID	*Pilot_ID*	*Pax*	*Model*	*Type*

Primary Key: Copter_ID

Table 5:

Event

event_uuid	*User_ID*	*Event_time*	*Device_type*	*Session_uuid*	*user_neighborhood*	*event_page*	*event_type*

Primary Key: event_uuid

Foreign Key: User_ID

Section 3: Extraction and Transformation

Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are **currently collecting in the pipelines** and they provide you with a section_3_event_logs template (which you can download from the classroom) generated by rider's activities on the Flyber App. Also provided in the Project Resources.

Extraction and Transformation-1

ETL is performed on the provided Event Logs Template and results will be transferred to the proposal template. The project's ETL should be created inside of your copy of the Event Logs template in the tab titled, ETL. Clicking on the link above will create a copy of the Event Logs for you

After being provided with a CSV log file, use extraction techniques to be able to get the data into a usable form. Because this needs to be a repeatable process we need to document it in order to assess its feasibility. Below,

1. Write the steps you took to extract the data and provide reasoning for why you used this method *Note: Don't forget to include any file type changes*:
2. Perform cleaning and transformation of the data in the ETL tab and document.
3. Document and provide rationale for all of your steps below as well.

Steps for Extraction:

1. *Data Collection*
    - *collect data and organise them in tables for a visualization.*
2. *Data Format Verification*
    - *check our current data type, we check for file extensions and probably the size of our dataset.*
3. *Assimilate both Records and Source*
    - *Confirm that the records we have corresponds to what was recorded initially by the source.*
4. *Search the duplications*
    - *Check our records to look for duplicates and delete them, since our dataset is large we could use tools such as tableau Public.*

Transformation-2

Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?

Day of Event Time						
5/10/19	6/10/19	7/10/19	8/10/19	9/10/19	10/10/19	11/10/19
9,891	18,056	18,202	17,963	17,600	17,694	17,595

2. How many events of each event type per day?

	Day of Event Time						
Event Type	5/10/19	6/10/19	7/10/19	8/10/19	9/10/19	10/10/19	11/10/19
choose_car	1,498	2,843	2,953	2,769	2,725	2,801	2,804
search	1,484	2,891	2,824	2,899	2,749	2,904	2,821
open	6,594	11,733	11,767	11,662	11,531	11,325	11,371
begin_ride	38	49	62	86	57	57	78
request_car	277	540	596	547	538	607	521

3. How many events per device type per day?

	Day of Event Time						
Device Type	5/10/19	6/10/19	7/10/19	8/10/19	9/10/19	10/10/19	11/10/19
ios	2,384	4,337	4,217	4,373	4,380	4,482	4,500
android	1,463	2,870	2,854	2,729	2,744	2,562	2,672
desktop_web	895	2,007	1,600	1,958	1,712	1,866	1,777
mobile_web	5,149	8,842	9,531	8,903	8,764	8,784	8,646

4. How many events per page type per day?

	Day of Event Time						
Event Page	5/10/19	6/10/19	7/10/19	8/10/19	9/10/19	10/10/19	11/10/19
search_page	3,995	7,219	7,307	7,221	6,979	7,201	7,137
book_page	1,977	3,548	3,576	3,572	3,586	3,424	3,506
driver_page	965	1,823	1,871	1,794	1,755	1,689	1,768
splash_page	2,954	5,466	5,448	5,376	5,280	5,380	5,184

5. How many events for each location per day?

	Day of Event Time						
User Neigh..	5/10/19	6/10/19	7/10/19	8/10/19	9/10/19	10/10/19	11/10/19
Manhattan	6,869	12,591	12,807	12,180	12,270	12,371	12,201
Brooklyn	2,009	3,737	3,590	4,025	3,440	3,400	3,556
Bronx	250	533	507	469	510	394	558
Queens	595	842	905	893	1,026	1,069	936
Staten Island	168	353	393	396	354	460	344

ETL Automation and Scalability:

Provide an analysis about this ETL process.

- This is a basic outline of the ETL process, collection, verification, assimilation, duplication removal, these steps are made to ensure data integrity from ingestion to consumption.

Address and provide rationale for manually extracting, loading and transforming the data from the raw logs.

- A rationale for manually extracting, loading and transforming the data from the raw logs could be for quick analysis around a certain time period, for example, a feature bug that was rolled out in an app update. It could quickly help the teams understand if the issue is focused on certain platforms or specific workflows.

Also address potential preliminary recommendations on improving this process:

- Manually Extracting, Loading and Transforming data from raw logs is not efficient.
- Manual ETL processes are prone to more errors versus automated. Manual ETLs are not scalable.
- Yes automation is needed, to reduce error and to best use human resources in more engaging and non-redundant work.

Section 4: Choosing Relevant Dataset

The previous exercise gave you a sneak peek into the Extraction and Loading aspects of ETLs in data pipelines. For making business decisions, a data consumer would like to have all the data they want. However, for any ecosystem, it is impossible to collect or provide everything that the customers need. In this exercise, you will get a taste of real world scenarios wherein:

- All the resources are not always available to get what you need.
- You have to get creative and get the most insights with a minimal data set.

Oftentimes your stakeholders/customers will “ask for the moon”, but you’ll have to push them to work with the small amount of information you have and get creative.

Note: As you learned in the course, being a Data Project Manager involves an extraordinary amount of collaboration. Complete the next sections based on the following scenario.

After the analysis in section 3, we made sense of the numbers, and realized the total number of events seems to be too small (this was a week's worth of data, but you need at least a month). Further investigation reveals that this was a subset of logs, but the actual data that is being collected is much bigger. Working through this small data set was tedious, and repeating this exercise on a much bigger data set manually won’t be feasible. Considering the time constraints of this project, engineering is willing to help with some automation. They also have limited bandwidth and are busy scaling systems up.

Engineering is willing to provide some data, but they have asked for the criterion that is most important.

First **provide your business question and provide a rationale for why this is the most important.**

Question:

Where in the user journey are we experiencing the lowest conversion rates?

- *Understand at which point potential users drop-out of the funnel*

Rationale:

Losing customers that flyber have "acquired", essentially people who have downloaded the app or logged on to the website, interacted on it but failed to book a ride on Flyber.

It meant that the user journey has some friction that is causing users to not complete their booking. It is probable that after a failed experience that such users may not return to the app again to use it.

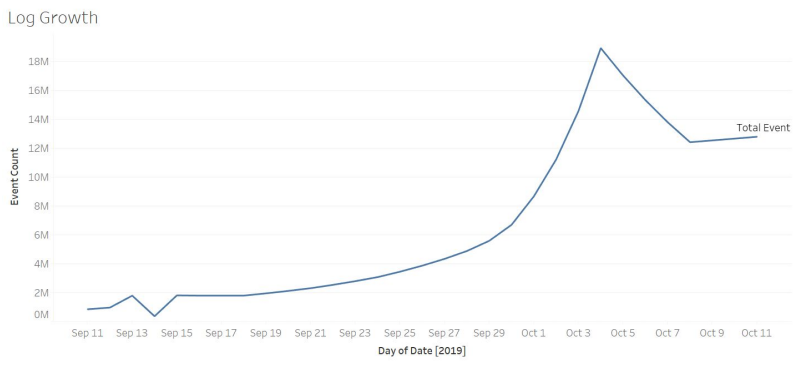
Thus it is important to understand where in the user journey flyber is experiencing the lowest conversion rates in order to improve the loss percentage of potential users

Choose one of the following prompts that you think can get you the most relevant information to proceed further.

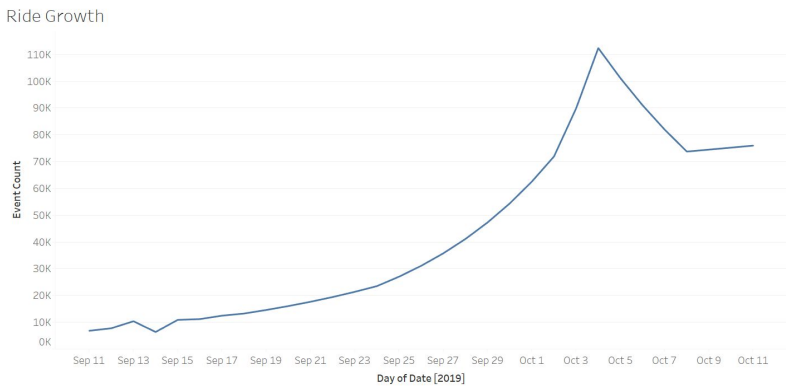
1. How many events are being recorded per day?
2. **How many events of each event type per day? (SELECTED)**
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

For your chosen question also answer the following using the data from section 3 to support your answer:

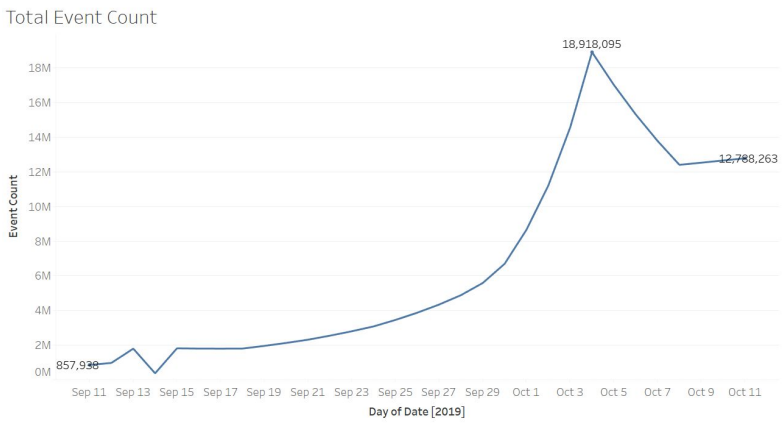
1. How much is the customer data increasing?

Image 1: Log Growth 	From Sept 21 to Sept 27, the customer data doubled (2m - 4m) From Sept 27 to the peak of Oct 4, customer data ~5x. From Oct 9 onwards, it appears like customer data ranged between 12-14m. Data is growing on log scale for the initial stage of the launch

2. How much is the transactional data increasing?

Image 2: Ride Growth 	Using "Begin rides" as a proxy on how much transactional data is growing. Begin ride event generated only when a rider begins a ride (a transaction). From trough to peak (~10k to 110k), it's about 11x growth in transactions. Transactions stabilize at between 70k - 80k/day

3. How much is the event log data increasing?

Image 3: Total Event Count/day 	Lowest point: ~0.86m Highest point: ~19m Latest point: ~13m From trough to peak, it represents close to 20x increase

Which of the following data is ***most*** important to answer this question? Why?

- **Event Log Data**

Event log data is the most important in answering "How many events of each event type per day", as it represents all the events generated (search, book, etc) from the aggregated number of customers using the application everyday. The rest of the data (customer data, transaction) only represents a part of the data generated on the application.

Section 5: [Optional] Loading and Visualization On Your Own

This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

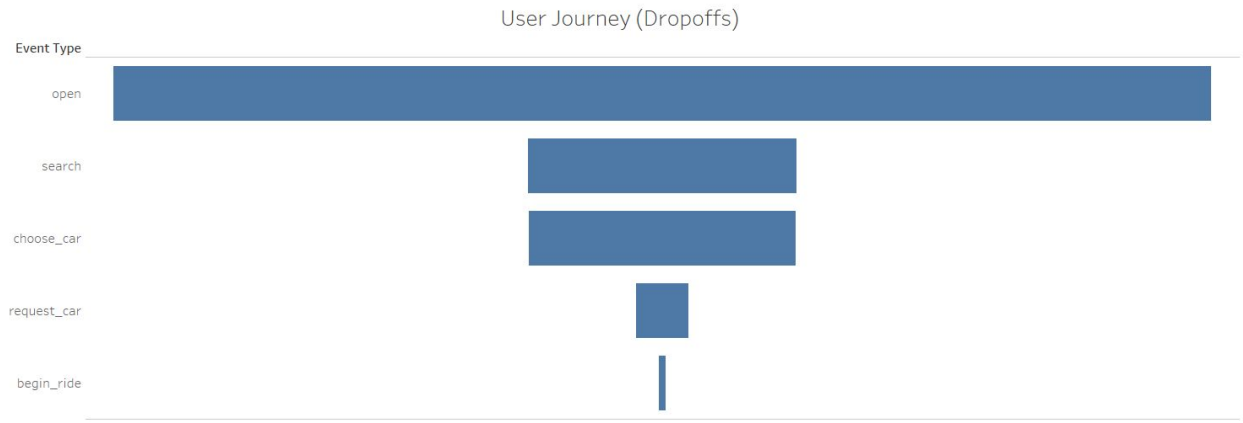
After sharing your criterion with engineering, they give you a new set of data:

Section 5 Event Type Log also available in the classroom resources. Also provided in the project resources section.

Engineering provided you with the data you want, but you still have yet to achieve your ultimate goal as a Data Product Manager. Now, utilize the data to make business decisions. Your executives do not want you to give them a bunch of data tables; instead, they prefer visualizations to help convey the key insights succinctly. Visualizing this data will help you understand the underlying trends and help you determine the story that needs to be told in your proposal to executives.

In this section, you can load and visualize the data into whatever platform you would like. A Python Notebook, Tableau or any other visualization tool you are familiar with. Create two visualizations that might help you to better understand your data trends and place either a screenshot or exported image of your visualizations and the details of each below. Please provide the steps you took to visualize your data and what the visualization tells you about your data.

Visualization 1:



Data Story: This graph tells us:

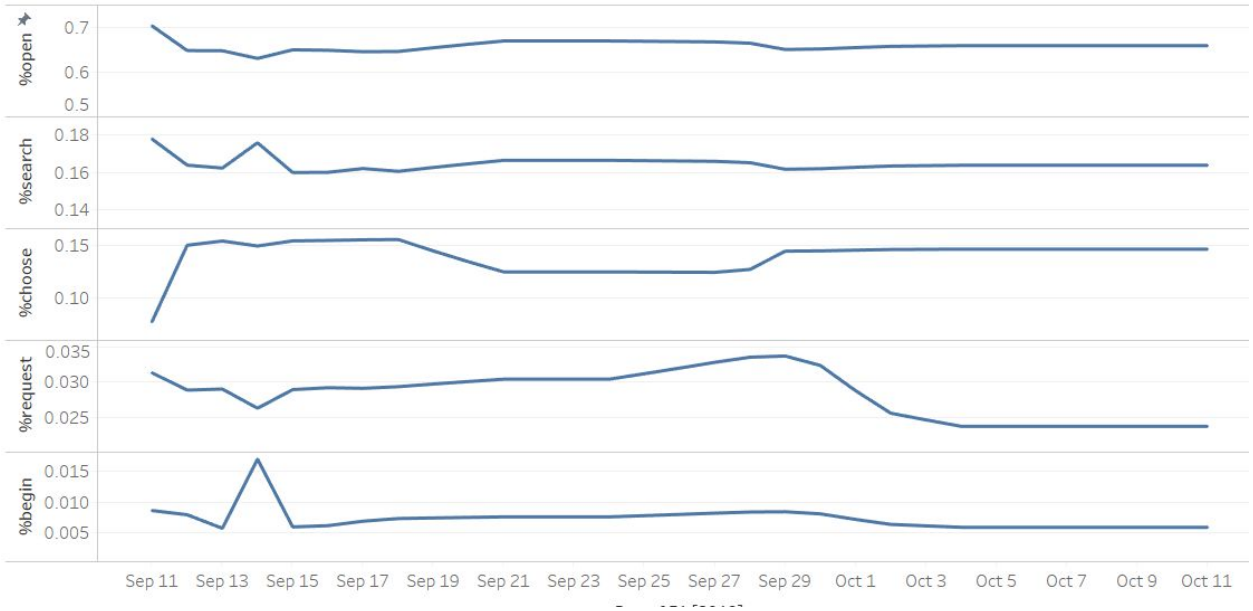
Acquiring customers/users to download the application or exploring the service should not be Flyber's priority at the moment. Instead there are 2 areas where Flyber should be focusing on:

1. *The Open-Search Stage*
2. *The Choose_car-request_car Stage*

This graph was created using the following steps:

1. *Count Distinct Event UUID (X axis)*
2. *Duplicate 1 and reverse axis*
3. *Event Type (Y axis)*

Visualization 2:



Data Story: This graph tells us:
There was a sharp drop in % of users who dropped off in the request car stage after Sept 30. It could be due to a change in the interface or an unsuccessful roll out of a feature within the app.

There was a peak in % of users who started a ride (begin ride stage) on Sept 14. A deep dive could be useful in gaining useful lessons on enticing users to book a ride with Flyber.

This graph was created using the following steps:

1. *Calculated Field of X/total ride for each day*
2. *Plot the Calculated Fields on the Y axis*
3. *Plot Time on the X axis*

Section 6: Business Insights

The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received and considered more robust.

What is the story the data is telling you about Flyber's data growth?

If you created Visualizations, you can use them as well, but they are not required). Include any data and calculations that were made to help tell that story and quantify the data growth.

Data Growth for Last Month (173% increase Month-over-Month)

Data Growth		
	F1 2019	
	September	October
Total Event	54.85M	149.92M
% Difference in Total Event	0.00%	173.30%

Data and calculations used for quantifying of Flyber's Data Growth:

Total Events
vs
Date (Month-Year)

What is the fastest growing data and why?

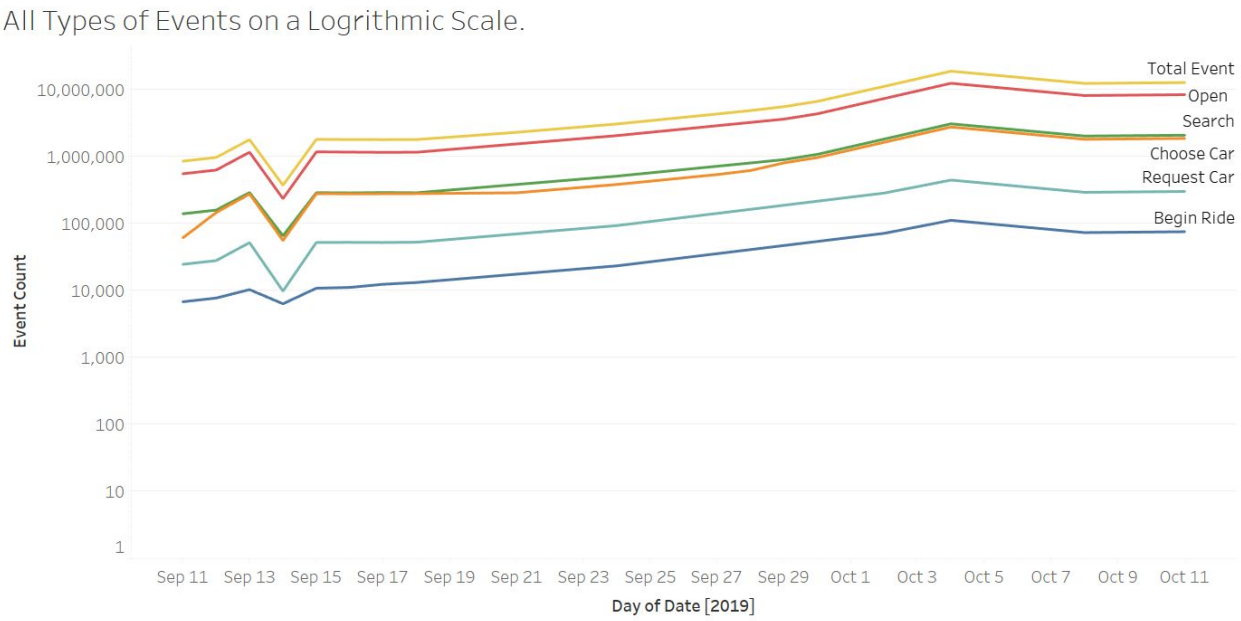
Data Growth		
	F1 2019	
	September	October
Event Logs	54.85M	149.92M
% Change in Event Logs	0.00%	173.30%
Begin Ride	428,490	910,102
% Difference in Begin Ride from the Fi..	0.00%	112.40%

Event logs is the fastest growing data vs Transactional Data, this is explained by the Section 5, Visualization 1, where Flyber is having low conversion rate from its users booking rides on their platform. Even while many users are opening the app and interacting on the application.

The dataset is not actually complete as there is no clear indicator on customer data.

All Event Type Data

Visualization:



What is the Data Story our data tells for each of the following:

- Graph Pattern
    - Trending upwards on log scale after an initial dip on Sept 14
    - All event types are following the same graph pattern

- Good or Bad
    - It is actually a good sign as all events reacted inline with each other (the conversion rate is constant)
    - What could be improved is the gaps between each event type. Ideally, Flyber should look into closing the gaps, which would mean that conversion between the two events is higher.
    - E.g. open/search versus begin ride

- October Marketing Campaign
    - Likely held on between Oct 3-5 and had an impact in data generation for the said period as reflected in the graph pattern

- Marketing Campaign Impact
    - In Image 3: Total Event Count, it shows that marketing campaigns has an exponential effect on Data Generation, likely caused by the widespread publicity the campaign had on users downloading the app and interacting on it

- Importance of Relationship Between Marketing Campaigns and Data Generation
    - There should always be additional capacity available around campaign days. Hence it is important that proper capacity planning is done considering marketing campaigns.

Section 7: Data Infrastructure Strategy

Thus far we have:

- identified data stakeholders and their data needs.
- Identified what data is currently being collected and what data needs to be collected.
- Identified data insights and growth trends.

Now, it's time to tie all the loose threads together and bring this process to its logical conclusion by suggesting which Data Warehouse (DWH) Flyber should invest in and why. Using data warehouse options below, suggest whether Flyber should choose an on-premise or Cloud data warehouse system and which specific data warehouse would best serve Flyber's data needs.

Data Warehouse Options:

Cloud:
- Amazon Redshift
- Google BigQuery
- Snowflake
- Microsoft Azure

On-Premise:
- Oracle Exadata
- Teradata, Vertica
- Apache
- Hadoop

You will address the following factors with a rationale as to why the DWH chosen is the best for Flyber:
- Cost
- Scalability
- In-house Expertise
- Latency/Connectivity
- Reliability

Cloud vs On-Premise

Provide an evidence based solution as to why Flyber would be best served by a Cloud or on-premise DWH. In this response, you don't need to specify *which* specific Cloud or on-premise DWH product you will choose, just if it will be Cloud or on-premise. Remember to address the factors above.

Cloud is chosen as the Platform of choice for the DWH.

Reasons:

1. *Cost*
    a. *lower infrastructure cost as it does not require any upfront costs for the setup or procurement of servers. Cost is incurred as a proportion of usage, storage, compute capacities.*
2. *Scalability*
    a. *A highly scalable & agile solution is required as seen from the scenario of event generation per day, from trough to peak, change in data generated was close to 20x. Cloud allows on-demand capacity changes which is a more flexible and cost efficient solution.*

3. *In-house expertise*
    a. *Flyber is a startup and should not invest in hiring expensive resources in the initial stages of launch but instead deploy more resources in focusing on the core business "Flying Taxis"*
4. *Latency/Connectivity*
    a. *MVP use-case might not need data in real-time. Thus, latency does not matter*
5. *Reliability*
    a. *Cloud Solution Providers have high availability architecture and redundancy built-in as part of their offers.*

Suggested DWH

Provide an evidence based solution as to which DWH product is best for Flyber. Remember to address the factors above.

1. Cost: BigQuery cost as much as the rest of the solutions
2. Scalability: Google BigQuery supports a dataset size up to multiple petabytes in an optimal manner.
3. Flexibility: Google BigQuery has storage and computing separated so that can be tuned optimally without disrupting each other
4. In-house expertise: Google BigQuery requires lower maintenance expertise and is a fully managed service
5. Latency & Reliability: High for Google BigQuery

Image Appendix

Image 1: Log Growth

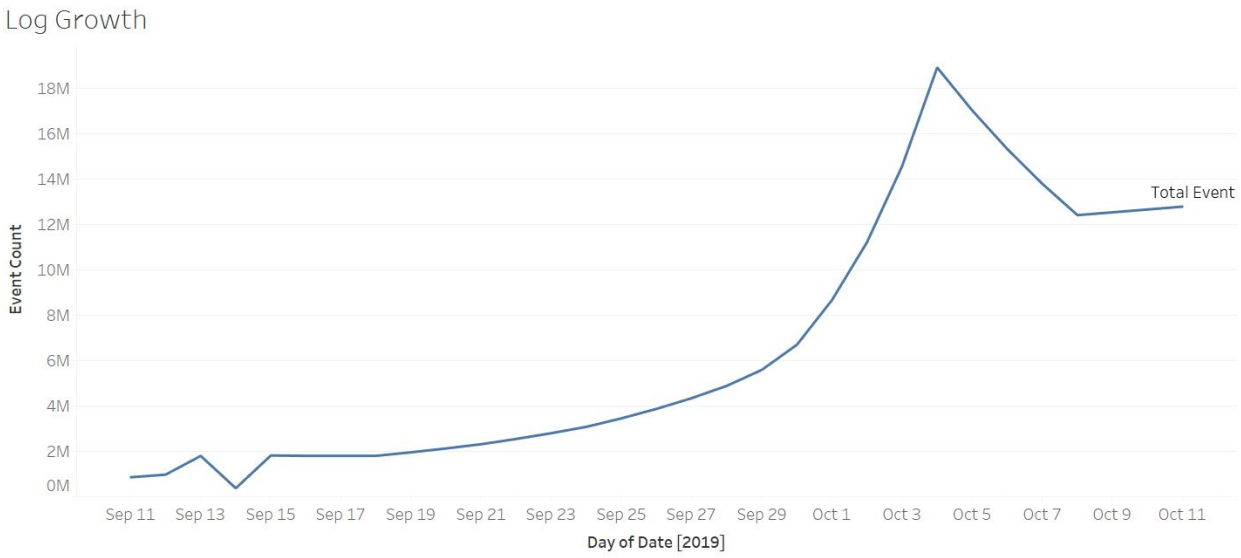


Image 2: Ride Growth

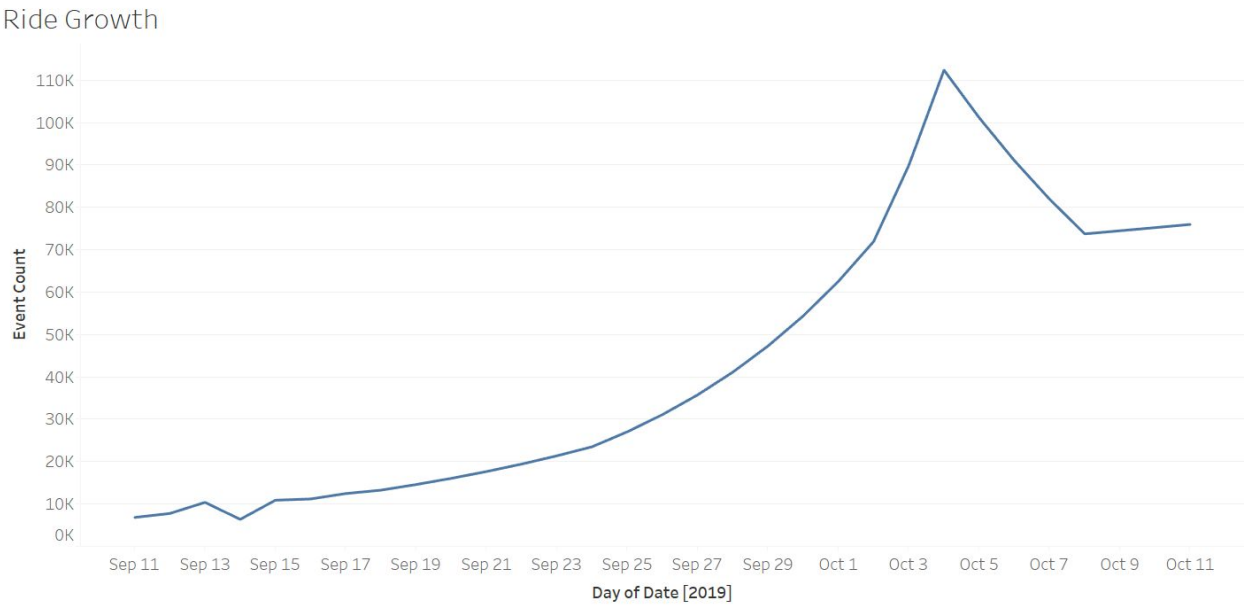


Image 3: Total Event Count

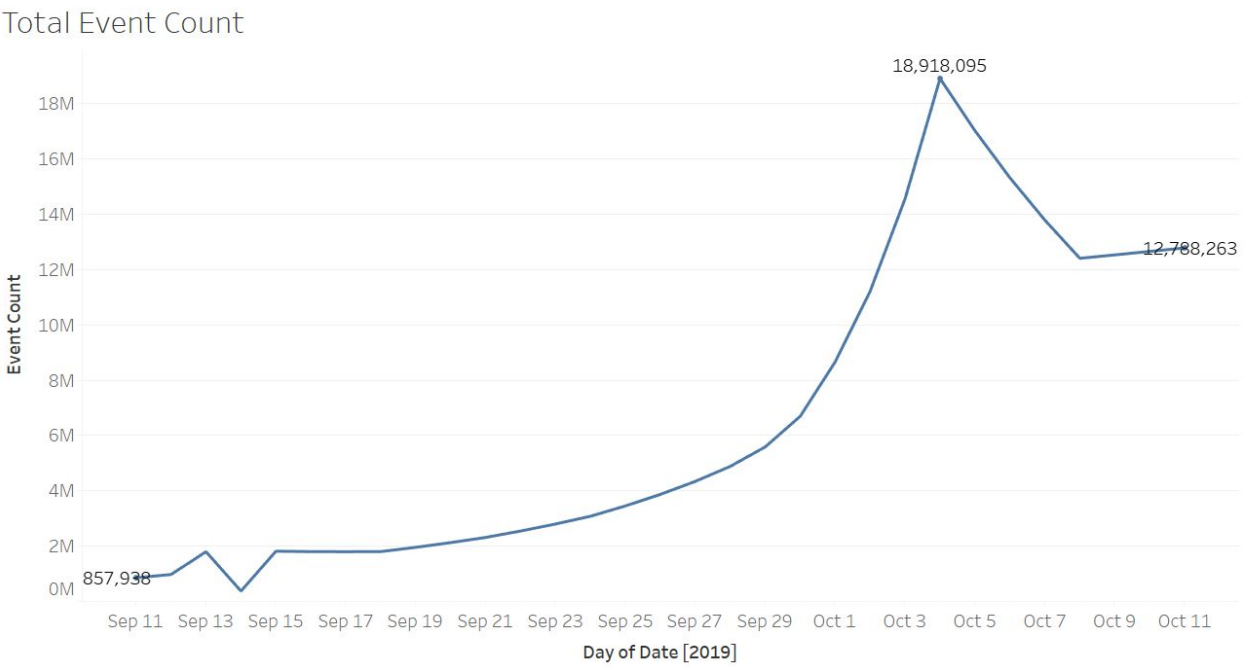


Image 4: All Events Log Scale