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**
Engineering	They need to be present to ensure that a working product is present.	Monitoring app performance, bugs and features usability.
Product Management	They constantly work on improving the products and identifying areas of development in the app.	Product Managers tend to identify customers'/partners' pain points and needs, and turn them into feasible features.
Marketing	Marketing teams should be focusing on increasing the new customers/partners while retaining the old ones.	They are usually focused on advertising to attract new customers/partners or building awareness.
Finance	They should be able to identify where the profit comes from and what are the expenses.	Finance teams mainly focus on building and monitoring P&Ls.
Customer Care	Customer Care team is the face of the company and they are responsible on handling all different types of clienteles.	They should be able to deal with different types of customers/partners needs while making sure to make their needs heard.

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?
Engineering	Monitoring app performance, bugs and features usability.	EVENT: notification_open, session_open, bug, request, accept, complete	Engineering/Dev team is the implementers of the product design to reach the business objective.
Product Management	Product Managers tend to identify customers'/partners' pain points and needs, and turn them into feasible features.	ENTITY: number of rides, number of successful rides, pickups, dropoffs and locations EVENT: time_stamps, ride_id, price, location, funnels for each cycle	PM teams decide on the roadmap of the product and takes into consideration the available resources, priorities and product growth while relying on data.
Marketing	They are usually focused on advertising to attract new customers/partners or building awareness.	ENTITY: CAC, reach, number of requests, number of installations, retention and churn rates EVENT: first_open, requests, funnels for each cycle	Marketing team decide the campaigns to reach new customers and retaining the old customers.
Finance	Finance teams mainly focus on building and monitoring P&Ls.	ENTITY: costs, P&L	Finance teams decide how much to invest while keeping costs vs profit levels as needed.
Customer Care	They should be able to deal with different types of customers/partners needs while making sure to make their needs heard.	ENTITY: number of tickets, ratings, retention and churn rates	CS teams are primary to maintain an excellent experience for customers/partners and help solve their inquiries and problems.

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):

Table 1:

Customer

(You may add more columns if necessary.)

Customer_ID	*Email*	*Mobile_Number*	*First_Name*	*Last_Name*	*Date_Of_Birth*	*Home_Address*

Rationale for Choosing Primary and Foreign Keys for the Table 1:

***Customer_ID** acts as **a unique identifier and a Primary Key**. It should be the unique identifier between all customers' info and will be used to connect with other different tables.*

Table 2:

Partner

(You may add more columns if necessary.)

Partner_ID	*Email*	*Mobile_Number*	*First_Name*	*Last_Name*	*Date_Of_Birth*	*Home_Address*

Rationale for Choosing Primary and Foreign Keys for the Table 2:

***Partner_ID** acts as **a unique identifier and a Primary Key**. It should be the unique identifier between all partners' info and will be used to connect with other different tables.*

Table 3:

Vehicle

(You may add more columns if necessary.)

Vehicle_ID	*Vehicle_Type*	*Vehicle_Model*	*Vehicle_Colour*	*Serial_Number*	*Max_Passenger*

Rationale for Choosing Primary and Foreign Keys for the Table 3:

***Vehicle_ID** acts as **a unique identifier and a Primary Key**. It should be the unique identifier between all vehicle's info and will be used to connect with other different tables.*

Table 4:

Ride

(You may add more columns if necessary.)

Ride_ID	*Customer_ID(FK)*	*Partner_ID(FK)*	*Vehicle_ID(FK)*	*Pickup_Timetamp*	*Dropoff_Timestamp*	*Pickup_Location*	*Dropoff_Location*

Rationale for Choosing Primary and Foreign Keys for the Table 4:

***Ride_ID** acts as **a unique identifier and a Primary Key**. It should be the unique identifier between all rides' info and will be used to connect with other different tables. **Customer_ID, Partner_ID and Vehicle_ID** should be used as Foreign keys from other tables.*

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 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:
(You may add more steps if necessary.)

1. *Data Gathering*
    a. *Raw data collection and organized.*
2. *Data Wrangling*
    a. *Assessing and cleaning data to better understand it and deal with outliers and duplicates.*
3. *Data Exploring*
    a. *Analyze and model data to find out answers to questions and try to draw conclusions.*
4. *Data Visualizations*
    a. *Communicate results through visual representation.*

Transformation-2
Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
Event Count	9891	18056	18202	17963	17600	17694	17595	7979

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

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
Choose Car	1498	2843	2953	2769	2725	2801	2804	1301
Search	1484	2891	2824	2899	2749	2904	2821	1307
Open	6594	11733	11767	11662	11531	11325	11371	5133
Begin Ride	38	49	62	86	57	57	78	18
Request Car	277	540	596	547	538	607	521	220

3. How many events per device type per day?

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
ios	2384	4337	4217	4373	4380	4482	4500	2026
android	1463	2870	2854	2729	2744	2562	2672	1231
Desktop Web	895	2007	1600	1958	1712	1866	1777	682
Mobile Web	5149	8842	9531	8903	8764	8784	8646	4040

4. How many events per page type per day?

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
Search Page	3995	7219	7307	7221	6979	7201	7137	3174
Book Page	1977	3548	3576	3572	3586	3424	3506	1639
Driver Page	965	1823	1871	1794	1755	1689	1768	801
Splash Page	2954	5466	5448	5376	5280	5380	5184	2365

5. How many events for each location per day?

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
Manhattan	6869	12591	12807	12180	12270	12371	12201	5580
Brooklyn	2009	3737	3590	4025	3440	3400	3556	1594
Bronx	250	533	507	469	510	394	558	231
Queens	595	842	905	893	1026	1069	936	386
Staten Island	168	353	393	396	354	460	344	188

ETL Automation and Scalability:

Provide an analysis about this ETL process. Address and provide rationale for manually extracting, loading and transforming the data from the raw logs. Also address potential preliminary recommendations on improving this process.

I believe connecting over the databse directly would save time and effort for reading only data in real time. There is also an option of having a job where these reports are sent on regular basis. I believe the first option is the best while setting a good environment to identify outliers, and transform it for further analysis. Therefore, I would recommend a better idea that manual extraction as it is not scalable and would talk a lot of time.

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

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

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

- Event Log Data
- Transactional Data
- Customer Data

I would choose: "**How many events of each event type per day?"**

Date	05/10/2019	06/10/2019	07/10/2019	08/10/2019	09/10/2019	10/10/2019	11/10/2019	12/10/2019
Choose Car	1498	2843	2953	2769	2725	2801	2804	1301
Search	1484	2891	2824	2899	2749	2904	2821	1307
Open	6594	11733	11767	11662	11531	11325	11371	5133
Begin Ride	38	49	62	86	57	57	78	18
Request Car	277	540	596	547	538	607	521	220

** I would exclude October 12th as after having a look over the raw data, the data was not fetched from the whole day yet.

How much is the data increasing?

There has been increasing day after day specially after the first day with "Open" having the top number of events while "Request Car" is the least. This is a normal distribution however there might be something wrong with the user experience so I would recommend revisiting it. Moreover, I would recommend visiting the number of newly acquired customers to match it with the increase in events to understand the type of increase whether from newly or retained customers. Last but not least, I would recommend creating funnels to better understand the cycle, understand where is the least adoption rates and what is expected from the customer exactly in terms of behavior since they open the app till they reach the goal of making a ride request.

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

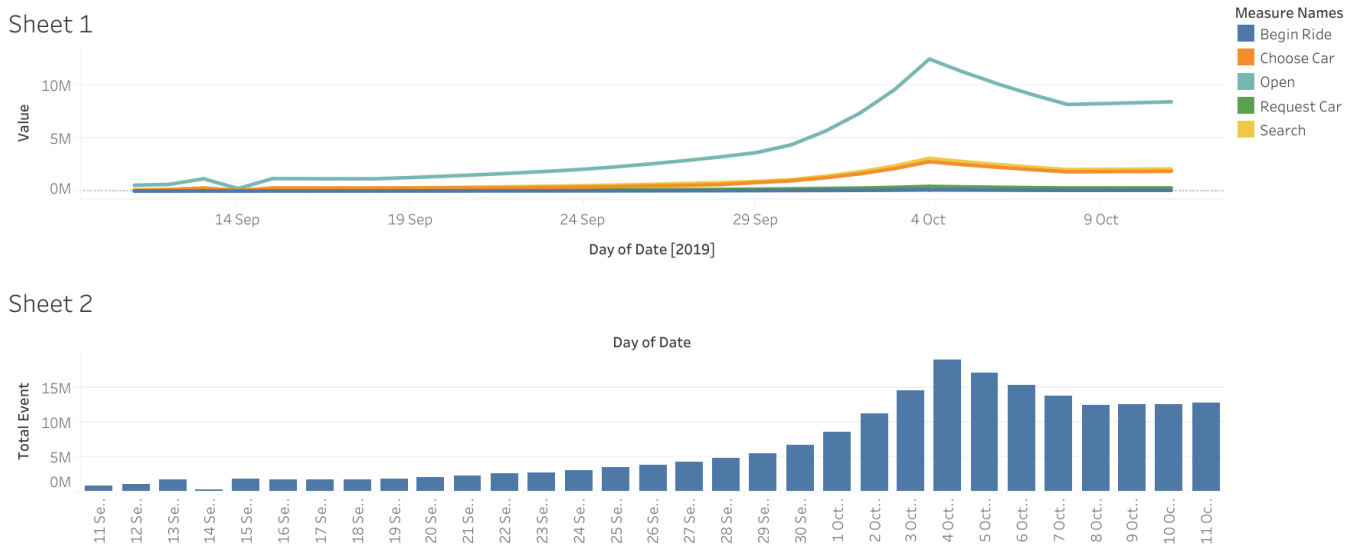
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

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:



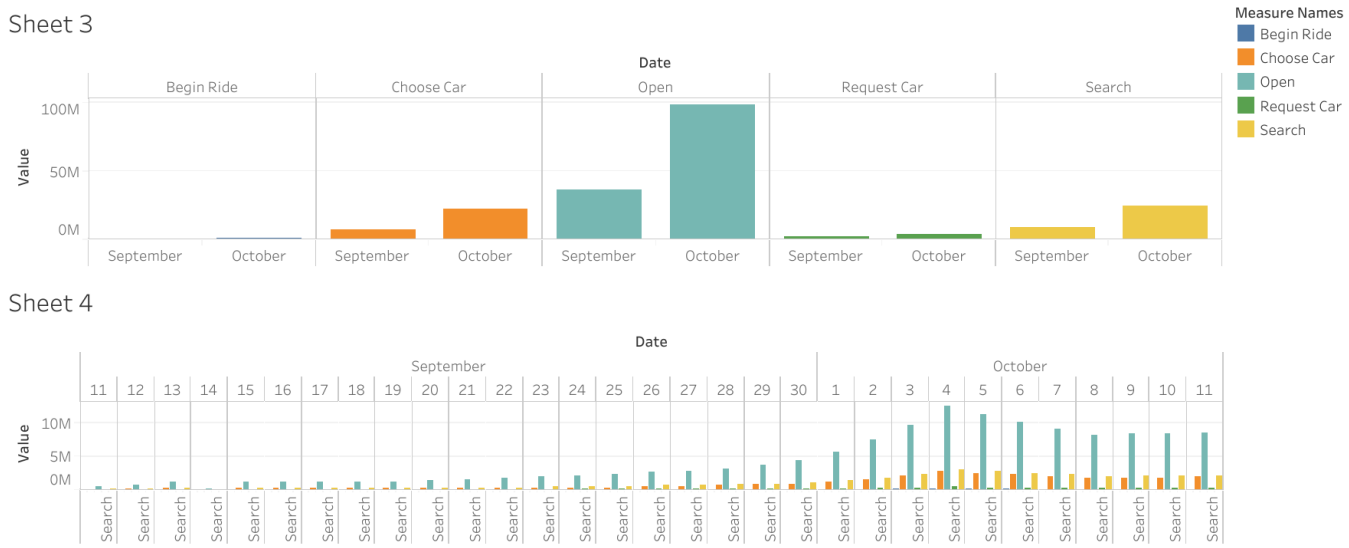
Data Story: This graph tells us:

This dashboard shows that the highest peak was n October 4th and it is increasing and decreasing gradually around it. This might be a good indicator of a buildup that happened before it or the success of a certain campaign. Moreover, there has been an increase in October 1st to 11th compared to the same period in September which shows an increase in the interest and the requests and rides and therefore the whole funnel.

This graph was created using the following steps:

1. *Raw data was connected on Tableau public.*
2. *Sheets were created to see the distribution of both each event and the total using line chart and bar charts respectively.*
3. *A dashboard was created to combine both sheets together.*

Visualization 2:



Data Story: This graph tells us:

Digging deeper to compare events performance over the given period (month). It shows that we were only 11 days into October but there is a huge significant increase in the events compared to almost 20 days in September. A further analysis of what made the huge growth is needed to understand customer targeting.

This graph was created using the following steps:

1. *Raw data was connected on Tableau public.*
2. *Sheets were created to see the distribution of monthly and daily events using bar charts for both.*
3. *A dashboard was created to combine both sheets together.*

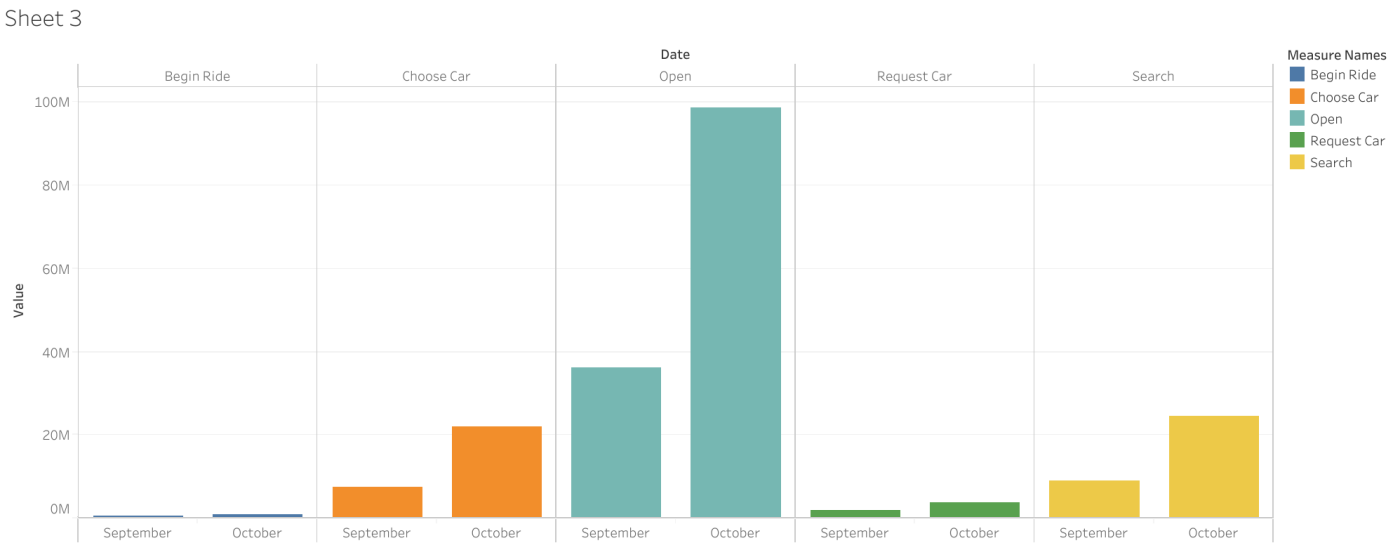
Section 6: Business Insights

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

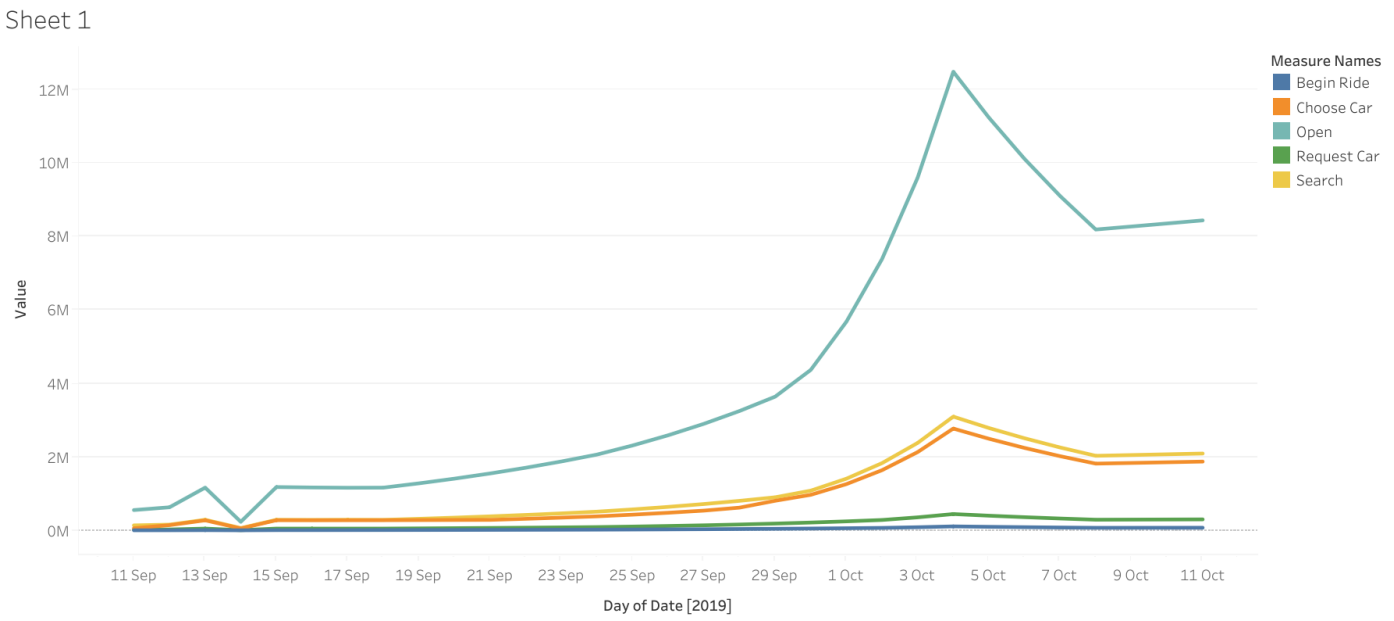
Visualization:



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

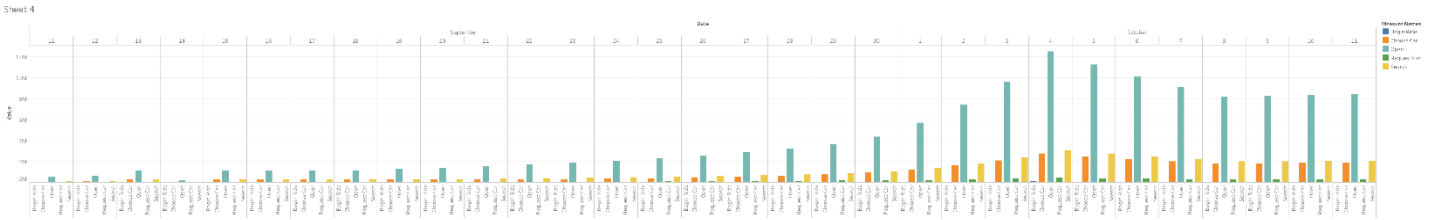
This definitely had shown an increase in all events and specifically ***Open.*** *I find this makes sense as it is the start of the funnel as well.*

What is the fastest growing data and why?



All Event Type Data

Visualization:



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

- Graph Pattern
- Good or Bad
- October Marketing Campaign
- Marketing Campaign Impact
- Importance of Relationship Between Marketing Campaigns and Data Generation

A growth took place gradually during September and massively during October. Marketing teams have pushed it through either by marketing campaigns to create awareness, create need, activate referral or even discounts and promocodes. Further investigation is highly recommended to understand whether these are new or old customers. Moreover, it is important to understand the demand to match with the supply. Furthermore, it is highly recommended to further investigate the drop after October 4th.

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

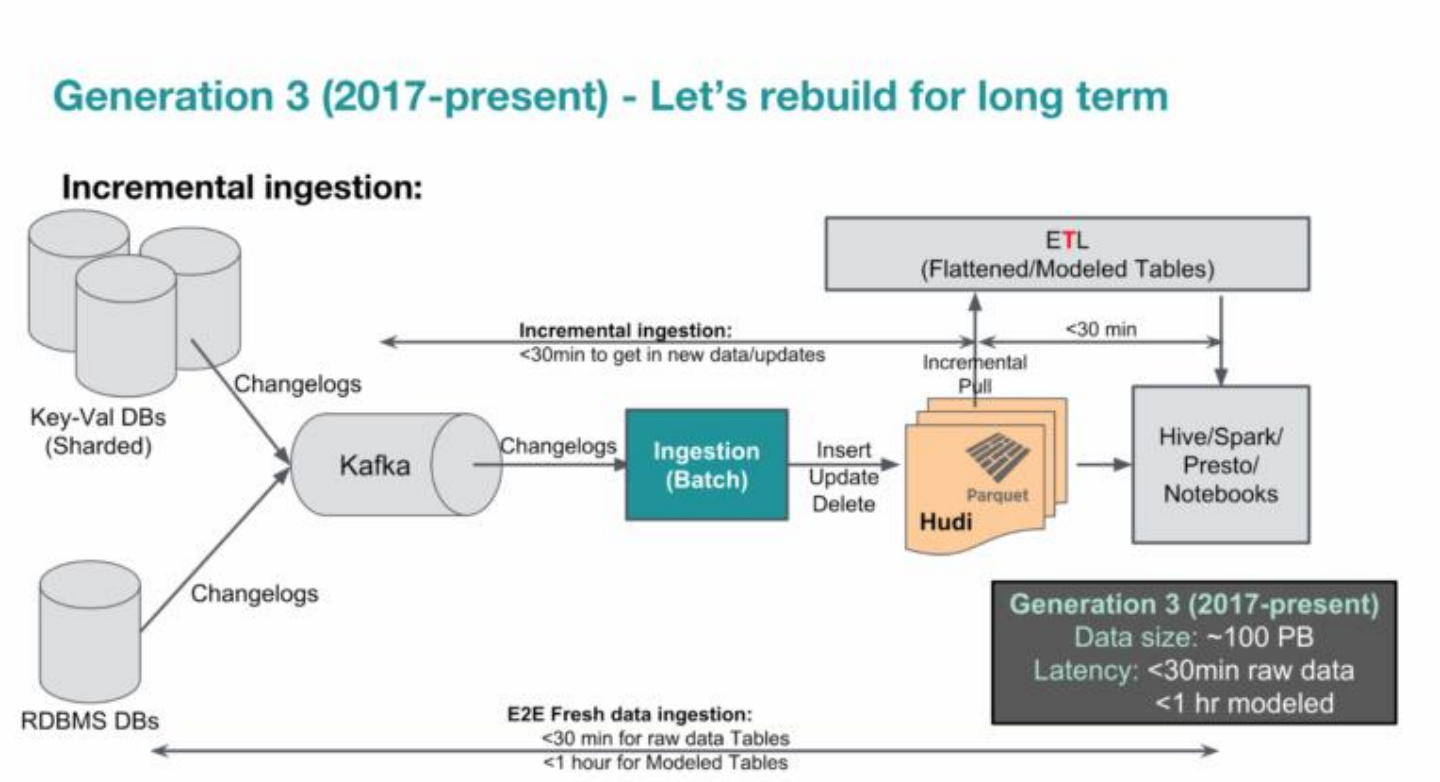
Provide an evidence-based solution as to why Flyber would be best served by a Cloud or on-premise DWH. In this response, you don't need to specify *which* specific Cloud or on-premise DWH product you will choose, just if it will be Cloud or on-premise. Remember to address the factors above.

Having a look over Uber as one of the main competitors, they have mentioned that there is no one answer to this question that fits all as there are big players going in and out of the cloud. However, Uber has preferred to stay on prem "Uber's Big Data Platform is mostly on-prem as well." They have said that the main factor of decision was the cost and yet it is a hot topic and hard to be determined.

Suggested DWH

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

*Taking into consideration other ride-hailing companies and highlighting Uber as well, a DWH that needs to operate in real time is needed. Therefore, we will follow in Uber's shoes and choose "**Apache Spark**" and "**Hadoop**" for the DWH. A sample of Uber's data platform is:*



*Section 7 has been referenced from Uber Engineering website: **https://eng.uber.com/***

Image Appendix

Image 1: Log Growth

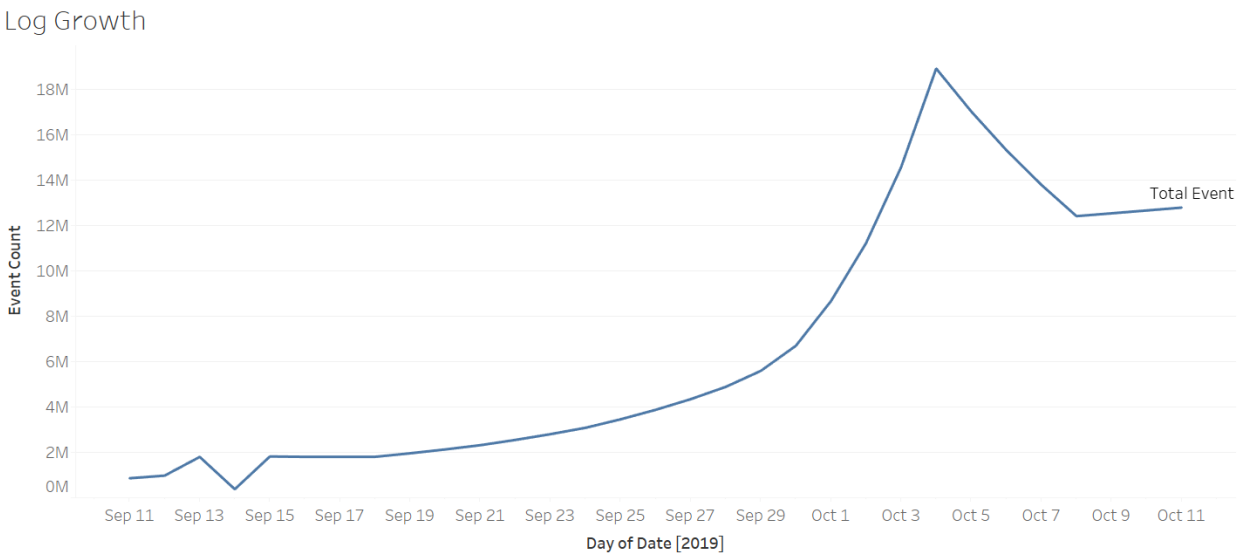


Image 2: Ride Growth

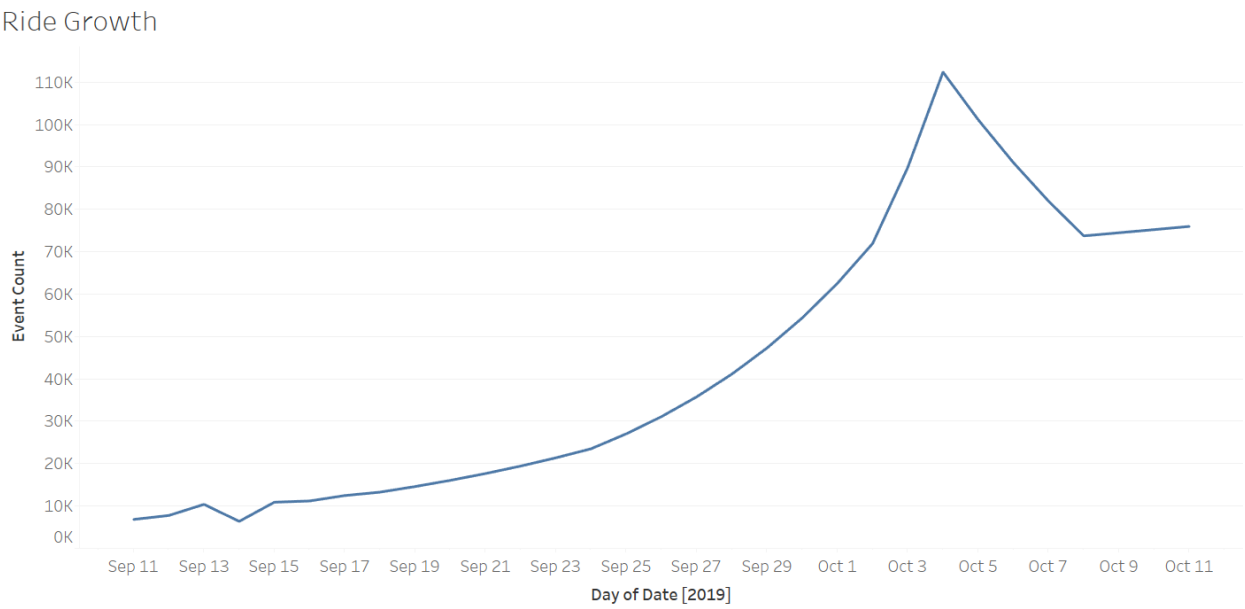


Image 3: Total Event Count

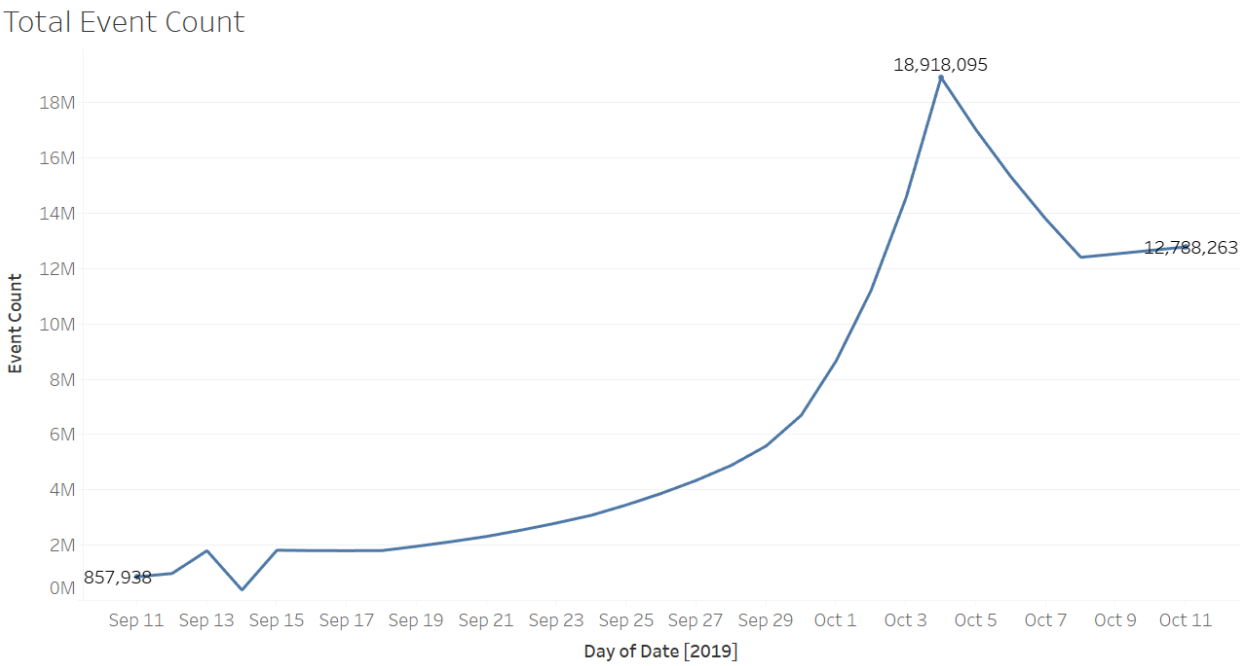


Image 4: All Events Log Scale