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	Engineering is responsible for building the product hence would always be a primary stakeholder	1. Monitoring real time use of the app, to prevent downtimes 2. Monitor event data generated by the app
Customer Service	Customer service necessary to gauge feedback from customers	1. Access customer data to service them

	during launch stage to iterate over the product	2. Create reports and visualisations
Operations Team	Operations team will work on executing the launch of the product on field, hence are a primary stakeholder	1. Monitoring real time use of the service to identify breakdowns 2. Identify bottlenecks from tracking data, to improve quality of service
Marketing Team	Marketing team will be responsible for customer acquisition during launch	1. Monitor Metrics of Marketing initiatives 2. Creating Customised content targeting customers using data available about them

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 Real time use of the App to prevent downtime	Real Time App usage, Traffic Data	Engineering needs to be able to fix any potential downtime due to high traffic on the app.
Customer Service	Access to customer data stored	Customer Data, Driver Data, Ride tracking	Customer Service needs to be able to identify the issue using data to be able to resolve customer issues
Operations	Monitoring real time use of the service to monitor breakdowns	Customer Data, Driver Data, Trip Data, Real time	Operations team need to be able to track rides and access data generated by them to service flying taxi

	and prevent them in the future	tracking of rides, Data generated by flying taxis	breakdowns during the MVP launch phase
Marketing	Track success of marketing initiatives to lower CAC	Ad performance data, App usage data	Marketing team need access to Ad performance metrics to tweak them to target the right customer segment and lower the CAC. App usage data will tell them what a customer visiting the app after clicking an ad ended up doing

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 Data

This table fits the use cases for Operations and Customer Service

(You may add more columns if necessary.)

Customer ID	*Trip ID*	*Customer Name*	*Customer Device*	*Customer Phone Number*	*Email*	*Payment Information*

Rationale for Choosing Primary and Foreign Keys for the Table 1: Customer ID is a unique identifier in this table for each customer hence the Primary Key, Trip ID is the foreign key as each customer's trips can be referenced from the Trips Table using this ID.

Table 2:

Trip Data

This table fits the use cases for Operations and Customer Service

(You may add more columns if necessary.)

Trip ID	*Customer ID* *Driver ID* *Transaction ID*	*Trip Date*	*Start Time*	*End Time*	*Start Location*	*End Location*	*Total Distance*	*Price Charged*

Rationale for Choosing Primary and Foreign Keys for the Table 2: Trip ID is unique for each trip and can be referenced in other tables as a foreign key. Customer Id, Driver ID and Transaction ID keys are foreign keys that refer to corresponding tables and makes it easier to keep data normalised without redundancy.

Table 3:

Transaction Data

This table fits the use cases for Operations and Customer Service

(You may add more columns if necessary.)

Transaction ID	*Customer ID* *Driver ID*	*Time*	*Date*	*Amount*	*Payment Mode*	*Transaction Complete*

Rationale for Choosing Primary and Foreign Keys for the Table 3: Each entry in this table is an individual financial transaction which can be referenced from the field Transaction ID. Foreign IDs Customer ID and Driver ID allow us to add Customer and Driver data without introducing any redunduncy.

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. *Save the file as .xlsx*
    a. *We must change the format since excel cannot save advanced analysis into csv*
2. *Create a pivot table*
    a. In order to calculate the aggregate data, we can use pivot tables to quickly get an aggregate count of event data
3. *Put the dates on the columns in the pivot table*
    a. *This will help us see the change as time progresses*
4. *Put Event Types/Device/Location/Event Page in the Rows in the pivot table*
    a. Do this for one of the fields

    b. Add 'Count of' the same field in the values section in the pivot table

    c. Copy the table and paste it in the ETL sheet

    d. Repeat these steps for other fields mentioned above

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

1. How many events are being recorded per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Event Count	**9891**	**18056**	**18202**	**17963**	**17600**	**17694**	**17595**

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

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Choose Car	1498	2843	2953	2769	2725	2801	2804
Search	1484	2891	2824	2899	2749	2904	2821
Open	6594	11733	11767	11662	11531	11325	11371
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?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
ios	2384	4337	4217	4373	4380	4482	4500
android	1463	2870	2854	2729	2744	2562	2672
Desktop Web	895	2007	1600	1958	1712	1866	1777
Mobile Web	5149	8842	9531	8903	8764	8784	8646

4. How many events per page type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Search Page	3995	7219	7307	7221	6979	7201	7137
Book Page	1977	3548	3576	3572	3586	3424	3506
Driver Page	965	1823	1871	1794	1755	1689	1768
Splash Page	2954	5466	5448	5376	5280	5380	5184

5. How many events for each location per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019

Manhattan	6869	12591	12807	12180	12270	12371	12201
Brooklyn	2009	3737	3590	4025	3440	3400	3556
Bronx	250	533	507	469	510	394	558
Queens	595	842	905	893	1026	1069	936
Staten Island	168	353	393	396	354	460	344

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.

ETL Works best with Entity Data that is structured. The data we have here is event data that is semi-structured.
Here the data was manually extracted since as a consumer we know what we are looking for from the data and can extract it from raw logs. Since the data was limited to only a week, it was relatively straightforward to transform it.
However with time the data generated by events would be huge. Even if this process was automated, transforming each it everytime before it is loaded would add a processing cost and would not be scalable as the amount of data generated increases. It may also be unnecessary since the needs of the data consumers might differ. Hence it is perhaps more prudent to follow an ELT process for event data, and leave the transformation to be done later depending on the use case of the particular consumer.

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

*The question '*How many events of each event type per day?' Is the prompt which will give us the most relevant information. From the table we can answer the 3 questions as:

1. How much is the customer data increasing?
   Customer data increases whenever the customer interacts with the app, we can estimate this by looking at the event 'open':

Customer Data Increase	05/10/19	06/10/19	07/10/19	08/10/19	09/10/19	10/10/19	11/10/19	12/10/19	Total
open	6594	11733	11767	11662	11531	11325	11371	5133	81116
Cumulative	6594	18327	30094	41756	53287	64612	75983	81116	
%age Increase	0%	178%	64%	39%	28%	21%	18%	7%	

2. How much is the transactional data increasing?
   A transaction occurs whenever a ride is being booked and a ride begins, thus we can measure this by looking at the 'begin_ride' event:

Transaction Data Increase	05/10/19	06/10/19	07/10/19	08/10/19	09/10/19	10/10/19	11/10/19	12/10/19	Total
begin_ride	38	49	62	86	57	57	78	18	445
Cumulative	38	87	149	235	292	349	427	445	
%age Increase	0%	129%	71%	58%	24%	20%	22%	4%	

3. How much is the event log data increasing?

Event Data Increase	05/10/19	06/10/19	07/10/19	08/10/19	09/10/19	10/10/19	11/10/19	12/10/19	Total
Grand Total	9891	18056	18202	17963	17600	17694	17595	7979	124980
Cumulative	9891	27947	46149	64112	81712	99406	117001	124980	
%age increase		182.55	65.13	38.92	27.45	21.65	17.70	6.82	

Since we can answer the three questions using event log data, it is the most important to answer all three questions.

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

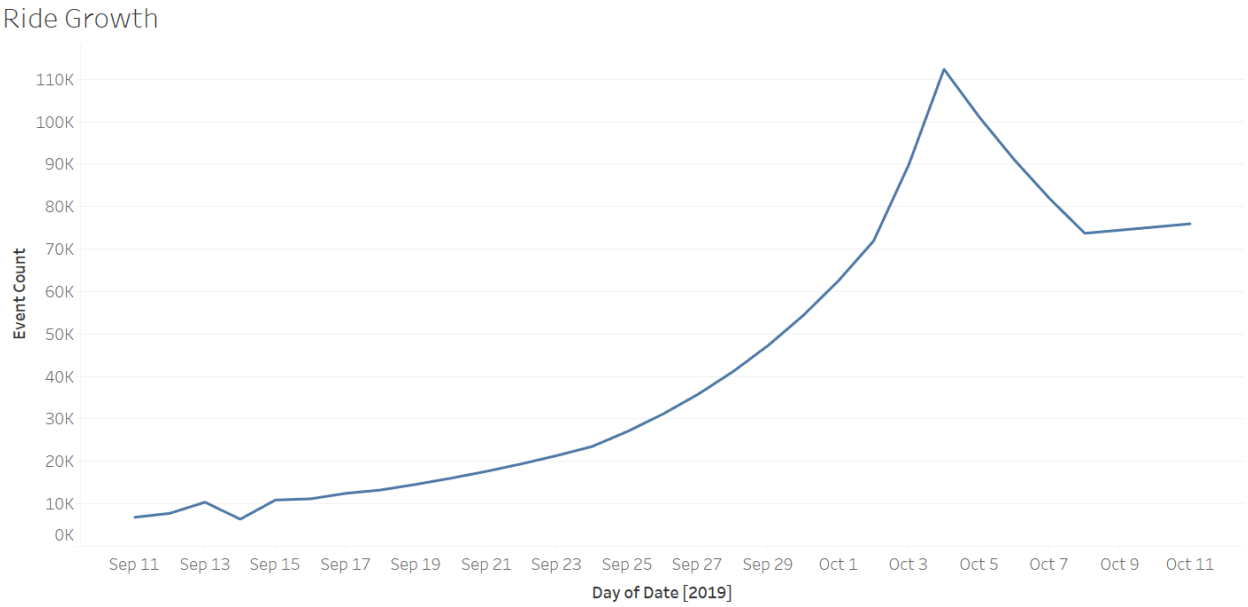
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

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:



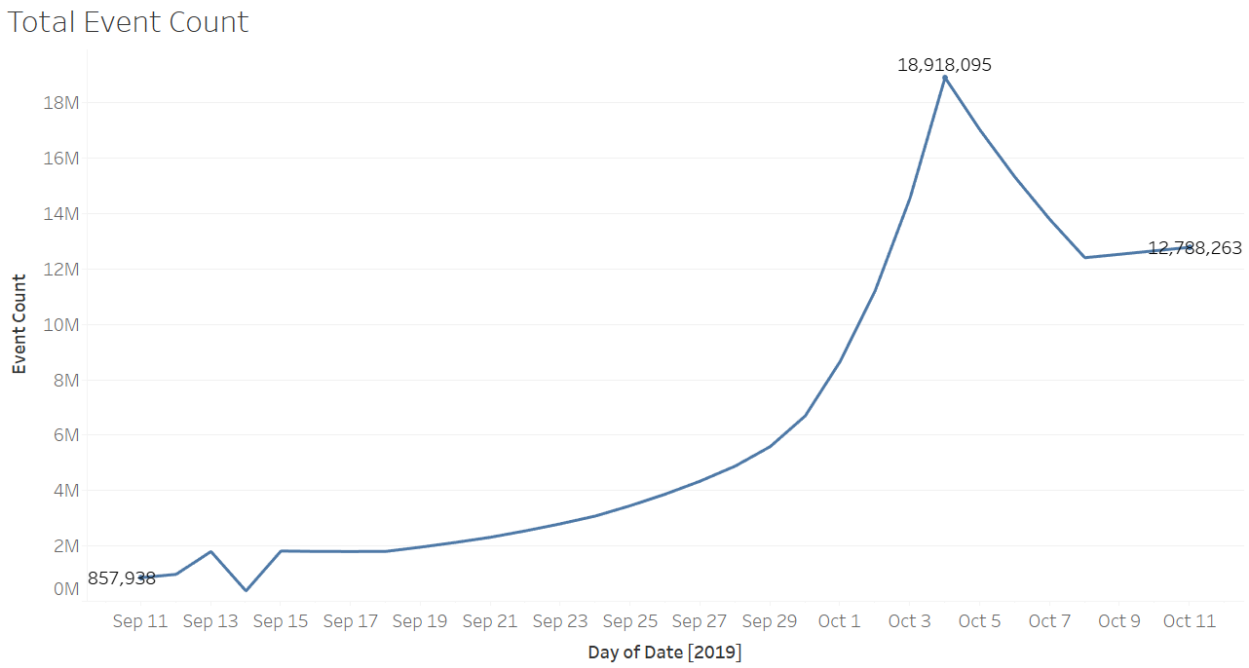
Data Story: This graph tells us:

Growth of rides taken by customers over time

This graph was created using the following steps:

1. *Open Tableau and load data*
2. *Add date to columns and select it as date in the right click menu*
3. *Add 'Begin Ride' to Rows and select it as a dimension*
4. *Select the line graph option from 'Show Me' Menu*

Visualization 2:



Data Story: This graph tells us:

Total events generated by customers over time

This graph was created using the following steps:

1. *Open Tableau and load data*
2. *Add date to columns and select it as date in the right click menu*
3. *Add 'Total Event' to Rows and select it as a dimension*
4. *Select the line graph option from 'Show Me' Menu*

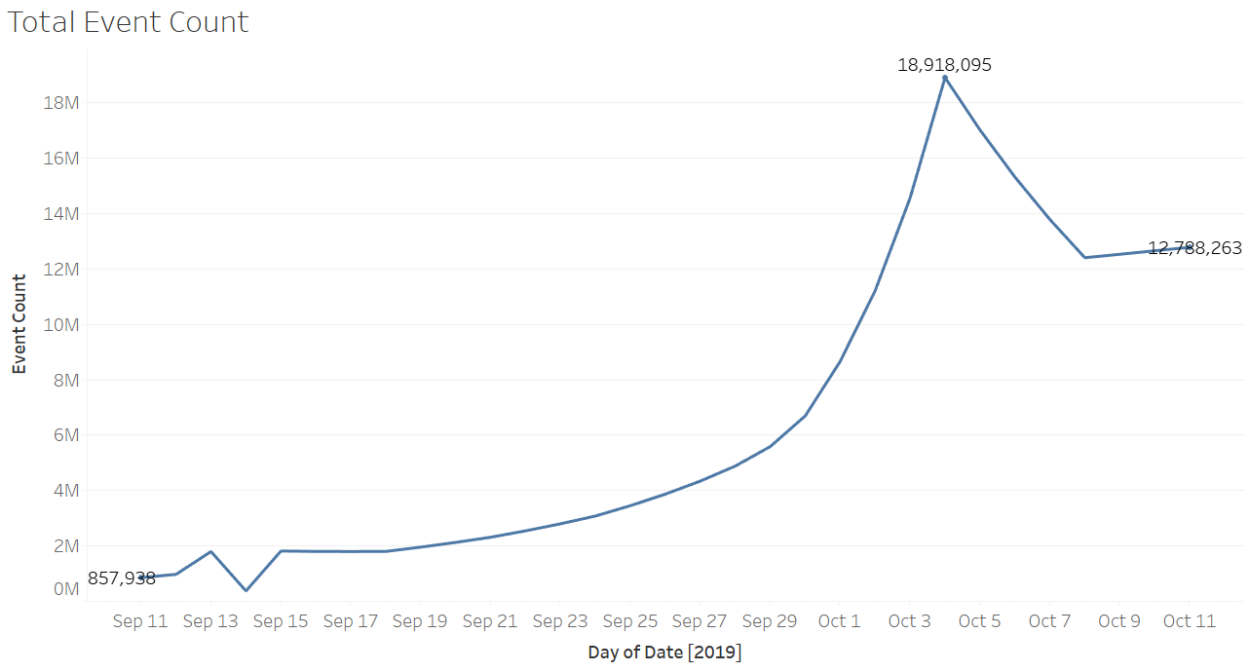
Section 6: Business Insights

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:

Weekly %age Growth

	F1			
	Week 37	Week 38	Week 39	Week 40	Week 41
% Difference in Begin Ride from the Previous along F1		206.41%	107.29%	170.69%	-12.46%
% Difference in Choose Car from the Previous along F1		265.56%	58.15%	285.85%	-3.56%
% Difference in Open from the Previous along F1		244.42%	87.16%	225.44%	-3.56%
% Difference in Request Car from the Previous along F1		250.82%	97.11%	170.69%	-12.46%
% Difference in Search from the Previous along F1		236.42%	87.33%	225.44%	-3.56%
% Difference in Total Event from the Previous along F1		245.87%	83.41%	230.84%	-3.85%

%age Difference

	F1	
	September	October
% Difference in Begin Ride from the Previous along Table (Across)		112.40%
% Difference in Choose Car from the Previous along Table (Across)		195.39%
% Difference in Open from the Previous along Table (Across)		172.67%
% Difference in Request Car from the Previous along Table (Across)		111.21%
% Difference in Search from the Previous along Table (Across)		172.29%
% Difference in Total Event from the Previous along Table (Across)		173.30%

Thus from this we can see that Total event data has increased by 173.3%

What is the fastest growing data and why?

For calculating fastest growing data among:

- *Event logs: Total Events*
- *Transactional data: Estimated by 'begin_ride'*
- *Customer data: Estimated by 'Open'*

Weekly %age Growth

	F1				
	Week 37	Week 38	Week 39	Week 40	Week 41
% Difference in Begin Ride from the Previous along F1		206.41%	107.29%	170.69%	-12.46%
% Difference in Open from the Previous along F1		244.42%	87.16%	225.44%	-3.56%
% Difference in Total Event from the Previous along F1		245.87%	83.41%	230.84%	-3.85%

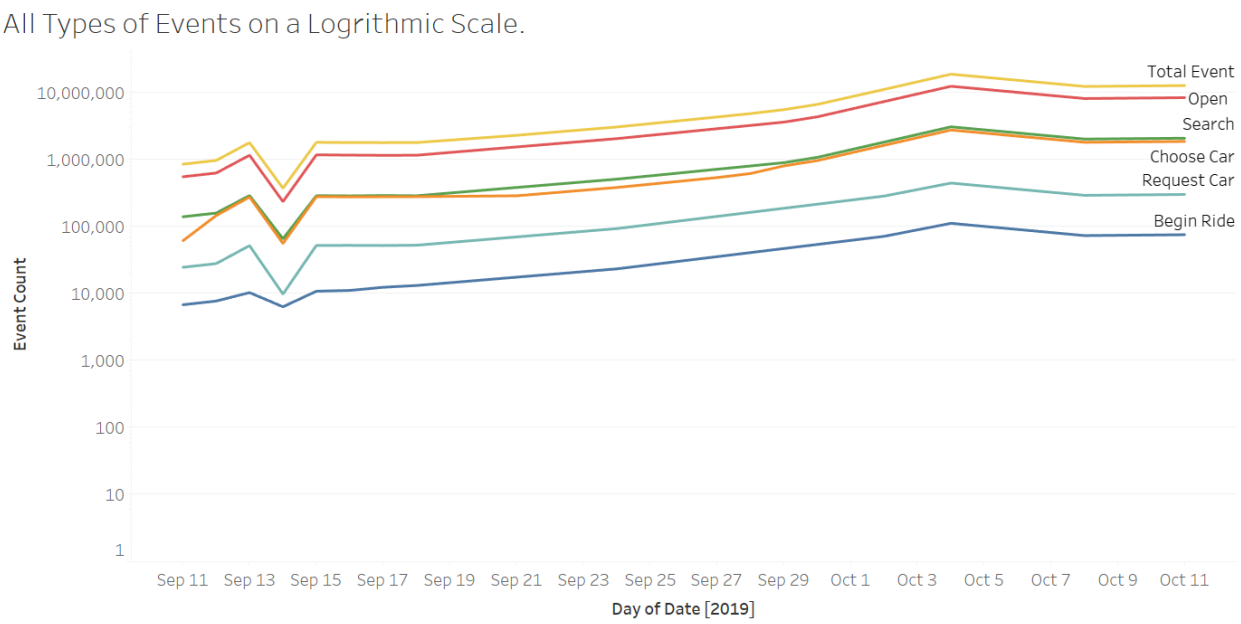
Overall %age Growth

	F1	
	Septem..	October
% Difference in Begin Ride from the Previous along F1		112.40%
% Difference in Open from the Previous along F1		172.67%
% Difference in Total Event from the Previous along F1		173.30%

Fastest growing data type is 'event log'. This is the fastest because this encapsulates and is an addition of all other data types.

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

All events are following the same pattern. This is good for our business since it means that a uniform %age of people are falling through our funnel. If the patterns were different, it would mean one or more of the pages/features that trigger that event may have affected the customer behaviour negatively. We can see that the October marketing campaign resulted in a huge upsurge in data generation from the peak seen around Oct 5. This tells us that our marketing campaign is working and bringing in new customers that are trying our product. This has lead to a dramatic increase in data generated by our product. It is important to note this about marketing campaigns so that we can preemptively be prepared for scalability in our data pipelines around marketing campaigns as it can lead to a spike in data generation.

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 a better solution for our use case because:

- *Cost: Its cheaper for us to use cloud since on-prem would add infrastructure costs*
- *Scalability: Flyber's service use depends on traffic. Hence the data generated would vary accordingly. Thus we would need a solution that is scalable. Cloud provides us with that scalability.*
- *In-house Expertise: We need an always available support model. This it is better for us to choose cloud than hire in-house expertise.*
- *Latency/Connectivity: While it is important for our product to have a good latency so that we can track in real time. It is not paramount once a trip is started. Thus cloud is a better option for us.*
- *Reliability: Although we lose some control with cloud, they have become reasonably reliable for our use case, such that we do not need on-premise DWH for uptime.*

Suggested DWH

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

Referring to external sources:
Cloud Data Warehouse Performance Testing – Gigaom
65871.pdf (scitepress.org)
We can see that Azure and Redshift have the best Price-Performance ratio. However we need a solution that is scalable as well. From 65871.pdf (scitepress.org), we can see that Redshift is ahead of Azure in terms of scalability. Redshift also costs lesser. Thus our choice of cloud DWH provider would be Amazon Redshift

Image Appendix

Image 1: Log Growth

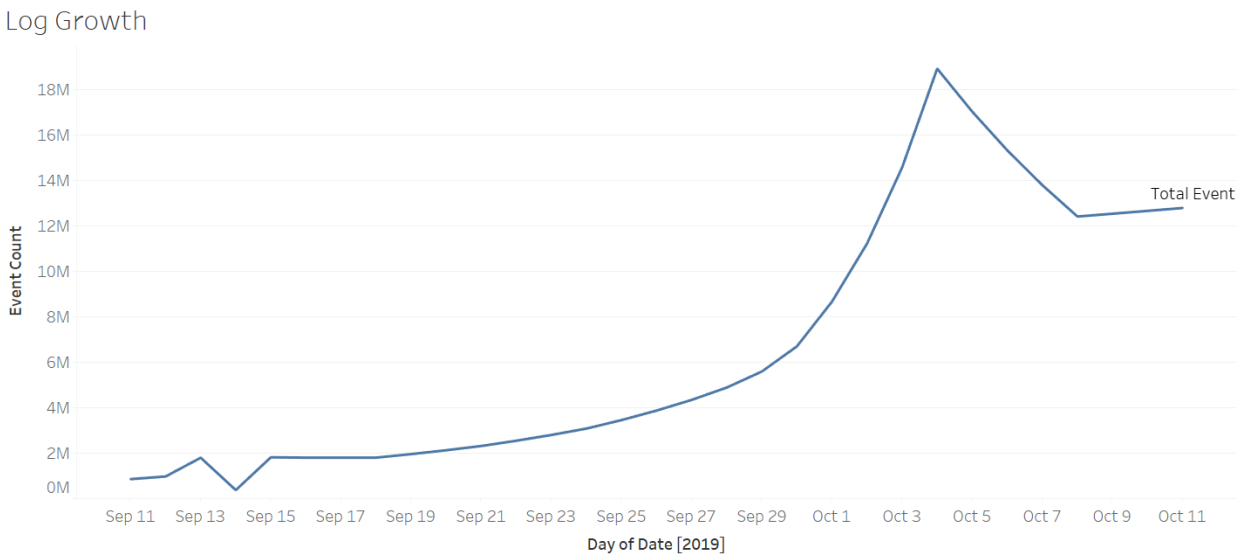


Image 2: Ride Growth

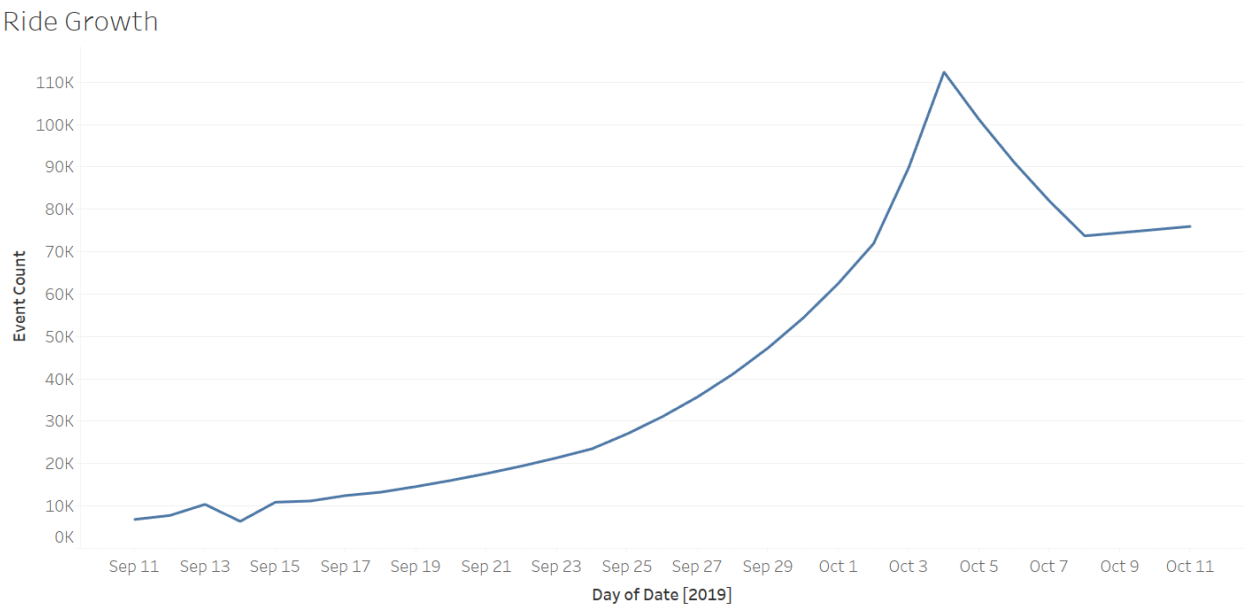


Image 3: Total Event Count

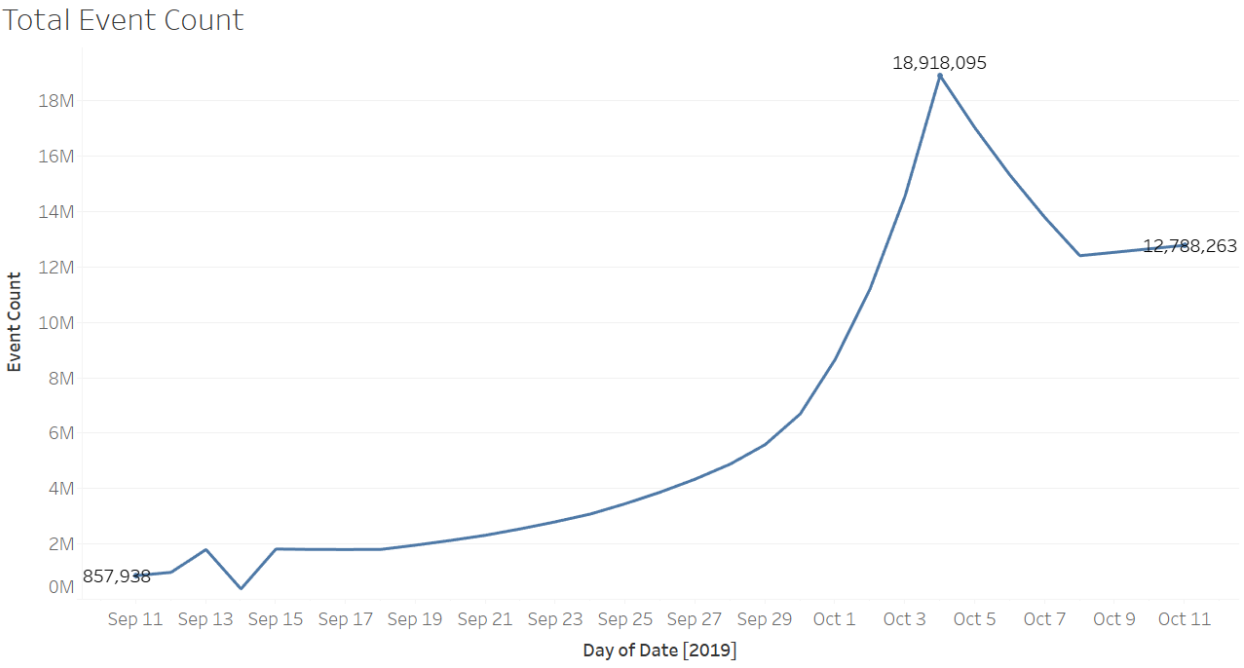


Image 4: All Events Log Scale

All Types of Events on a Logrithmic Scale.