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	Have a working product	Monitor web and app performance
Product Management	Improve the product	Identify user pain points
Marketing	Get new customers and retain old customers	Targeted advertising

Finance	Predict P&L	Monitor current P&L

Section 2: Data Collection and Data Modelling

To support our primary stakeholders' 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	Monitor web and app performance	Event: Event ID, Timestamp, Event Type	To maintain growth, Flyber must continue to deliver a delightful in-app customer experience. To do so, it must monitor and resolve latency, crash, and screen load issues as and when they occur.
Product Management	Identify user pain points	Event: Event ID, Timestamp, Event Type, Event Screen	Product must monitor user behavior to identify and resolve new and existing user pain points so that Flyber can remain competitive in the marketplace, especially against alternatives like taxis and rideshare.
Marketing	Targeted advertising	Entity: Customer Name, Email, Customer Trip History	Targeted advertising is a means to fuel growth, improve engagement, and increase retention.
Finance	Monitor current P&L	Entity: Aggregated Transactional Data, Number of Trips, Cost of Trips, Price of Trips, Taxes, Other Fees	Finance needs to closely monitor P&L as Flyber gets its service off the ground. To create a profitable business, Flyber will need to keep close tabs on revenue growth, costs, and customers, recurring and new.

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:

Rider

Rider ID	Account ID	First Name	Last Name	Email Address	Phone Number

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

The Rider ID primary key should serve as the primary marker for row entries of the Rider table. An Account ID foreign key is necessary to connect the rider with his/her financial account details, which must be stored and protected in accordance with PII and PCI DSS regulatory compliance requirements.

Table 2:

Trip

Trip ID	Rider ID	Route ID	Trip Duration	Trip Fare	Trip Fees

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

The Trip ID primary key should serve as the primary marker for row entries of the Trip table. Rider ID and Route ID are just two of likely several other foreign keys necessary to connect the trip details with rider information and route information.

Table 3:

Route

Route ID	Rider ID	Pickup Time	Dropoff Time	Pickup Address	Dropoff Address

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

The Route ID primary key should serve as the primary marker for row entries of the Route table. A Rider ID foreign key is necessary to connect the route with the corresponding customer.

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:

1. Connect to the source systems
    a. The first step of extractions is connecting to our source data systems. We will need to use an API to connect data sources to our ETL pipeline.
2. Select and collect necessary data

a. We must select and collect the necessary data, which may be located in disparate places. This includes organizing our data in tables for visualization. We must be thoughtful about which data we select and collect so as to not overwhelm ourselves with more data than we can process.

3. Verify data is formatted for transformation processing
    a. We must check our current data type, including file extensions and data set size. If necessary, we should convert our data into a suitable format.
4. Validate data
    a. We validate our data to ensure our records correspond to what was recorded initially by the source. We should also remove null records and duplicates. A tool such as Tableau could be helpful to validate our data.

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	3596	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.

The ETL was initially performed using Excel. The value in performing our ETL in a manual fashion, from extracting, loading, and transforming the data, stems from the value in being able to see each and every one of our steps being done in a systematic way. However, while Excel gives us full transparency into this framework, the process of manual manipulation can become quickly inefficient when dealing with a large volume data set, as with this example. The use of BI tool such as Tableau Public, for instance, can enable scalability of this process. Additionally, we could employ a Python script in order to automate ETL 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. Provide your business question and 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?

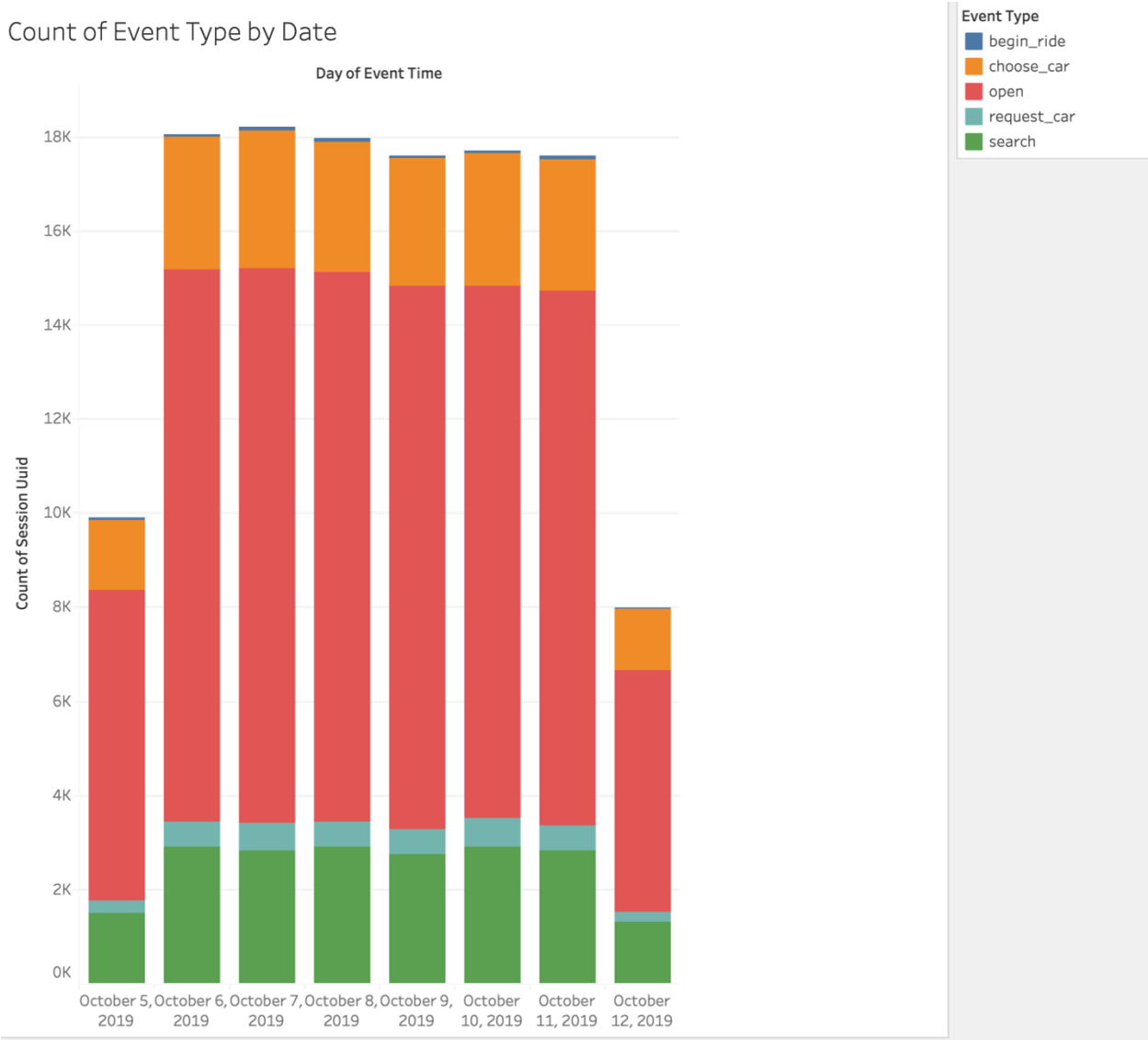
For your chosen question also answer the following using the data from section 3 to support your answer:

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

Given that our resources and time are limited, I would prioritize answering the questions, **"How many events of each event type per day?"** This would allow us to better understand how riders are spending their time on our app, what functionality they find most useful/valuable, and in turn, enable the Flyber product team to develop and prioritize its product roadmap. To answer this question, we will want to collect **event log data** so that we can get a clear picture of user behavior.



How much is the customer data increasing? Customer user data jumped from ~10K on 10-05-2019 to ~18K on 10-06-19, and remained between ~17.75K and ~18.25K from the period beginning 10-06-19 and 10-11-19 before falling to ~8K on 10-12-19.

How much is the transactional data increasing? The behavior of transactional data tracks closely with the volumetric changes of customer data.

How much is the event log data increasing? The behavior of transactional data tracks closely with the volumetric changes of customer data.

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:

[Insert Visualization Here.]

Data Story: This graph tells us:

[Insert Response Here.]

This graph was created using the following steps:

1. *[Insert Step Here.]*
2. *[Insert Step Here.]*
3. *[Insert Step Here.]*

Visualization 2:

[Insert Visualization Here.]

Data Story: This graph tells us:

[Insert Response Here.]

This graph was created using the following steps:

1. *[Insert Step Here.]*
2. *[Insert Step Here.]*
3. *[Insert Step Here.]*

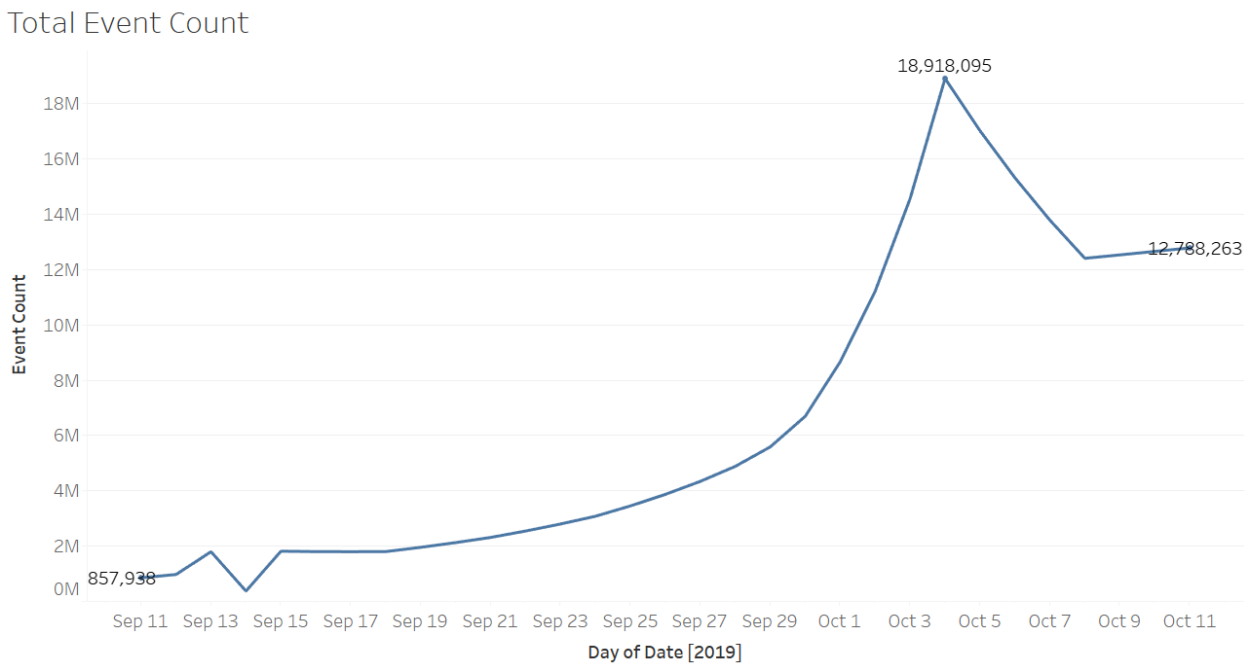
Section 6: Business Insights

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:

The above visualization was taken from the Appendix.

What is the fastest growing data and why?

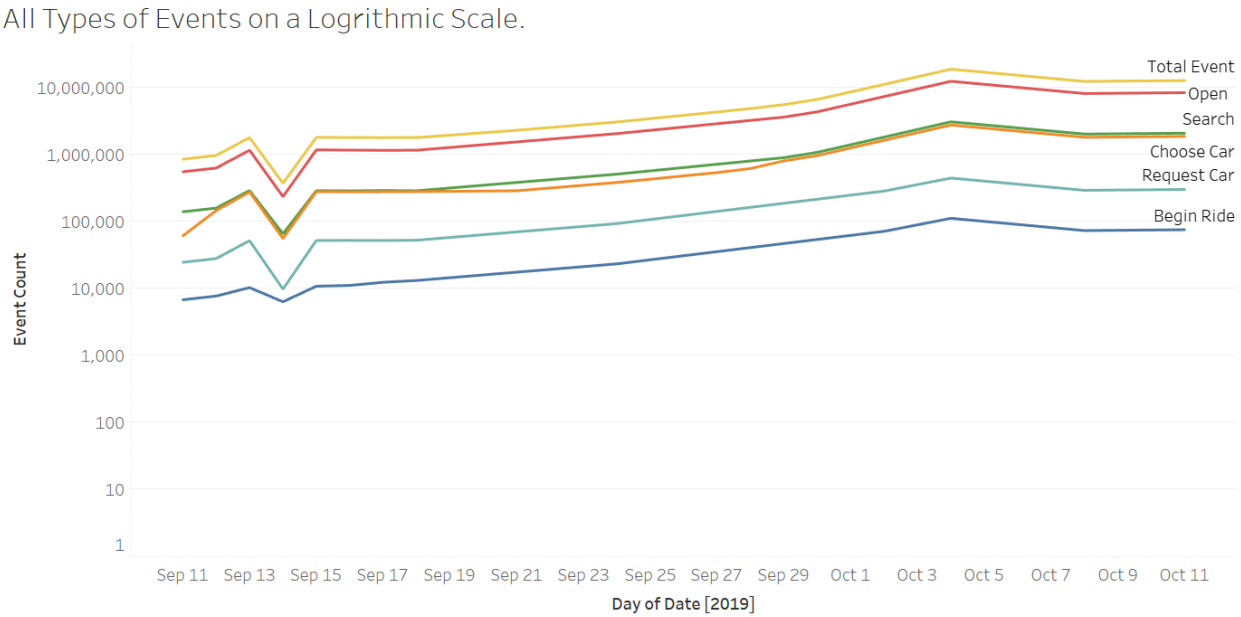
	Choose Car	Search	Open	Begin Ride	Request Car	Total Event
9/11/2019	61,416	140,598	556,714	6,838	24,763	790,329
10/11/2019	1,875,536	2,094,974	8,438,774	75,956	303,024	12,788,264
Times Growth	31	15	15	11	12	16
% Growth	2954%	1390%	1416%	1011%	1124%	1518%

From the plot, we see that total event data grew exponentially over the past month from 857,938 on 09-11-2019 to 12,788,264 on 10-11-2019. We see that events logged peaked 11-04-2019 reaching 18,918,096 that one day. We see a corresponding trend across all data types.

We can compute percentage growth by each event type for the period beginning 09-11-2019 and ending 10-11-2019. Doing so reveals that the Choose Car event type is the fastest growing data. Likely, returning riders will have developed a preference for a particular flying car and will want to book their next trip with this particular vehicle. Coupled with the demand for certain vehicles, riders may be more discerning in how they structure trips as time goes on.

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

Looking at the graphs, we see a sharp spike in the total events generated between 09-29-2019 and 10-05-2019. This is a positive trend, as it indicates a sharp rise in demand for ridership. The fact that a marketing campaign was launched in October is also consistent with this result, as targeted advertising likely played a role in driving rider demand up relative to the other months of Flyber's operation where demand grew organically. The data growth rate is similar across all event types.

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.

Flyber is best served by choosing a cloud data warehouse solution. Cloud offers Flyber (a) the most flexibility in terms of scaling up or scaling down its service, (b) will save Flyber the up-front costs of having to build an on-premise data warehouse for an MVP, (c) will allow Flyber to benefit from always-available support through its third-party data warehouse provider, (d) latency/connectivity will not be a concern since the test program will be limited to New York City, (e) we can optimize our engineering resources for innovation and improvement rather than data infrastructure setup.

Suggested DWH

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

When evaluating the options for a cloud data warehouse provider, Flyber should choose Google BigQuery. While RedShift requires periodic management tasks like vacuuming tables, BigQuery has automatic management. In addition, BigQuery is "serverless" meaning compute and storage resources can scale independently and all scaling issues are handled automatically. Finally, compared with Azure, BigQuery turns on encryption by default. Other advantages of BigQuery are its fast speed, reliability, use of SQL-like queries making it easy to use, and finally cost. BigQuery costs $20 per TB per month for the storage line and $5 per TB processed on that storage line, whereas RedShift costs $306 per TB per month for storage and unlimited processing on that

storage. Given that processing needs are likely minimal in the context of the Flyber app, BigQuery will likely result in greater cost savings.

Image Appendix

Image 1: Log Growth

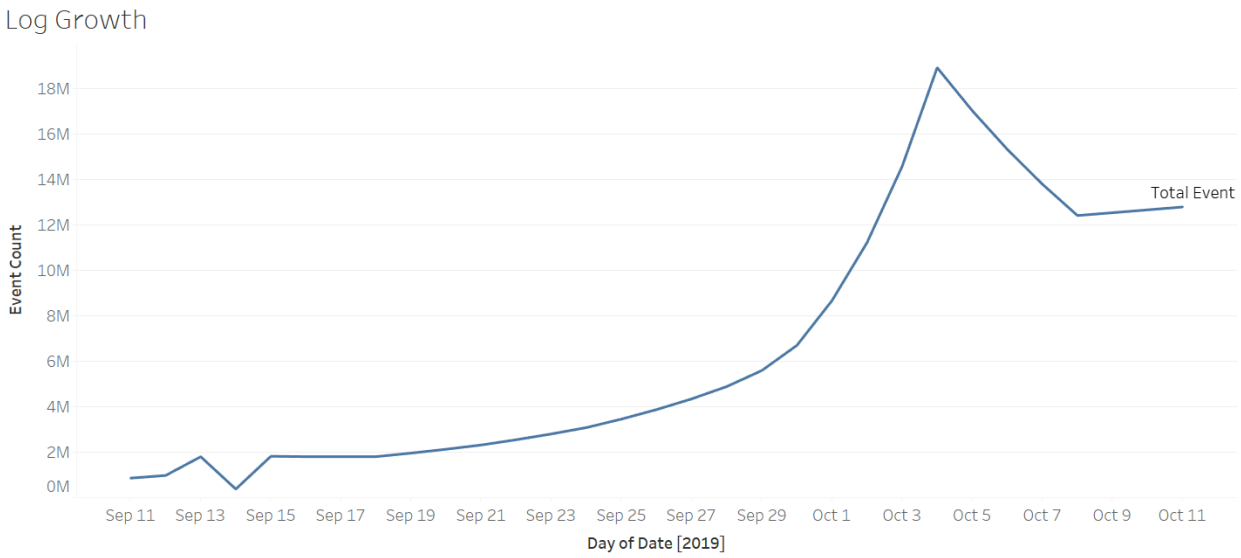


Image 2: Ride Growth

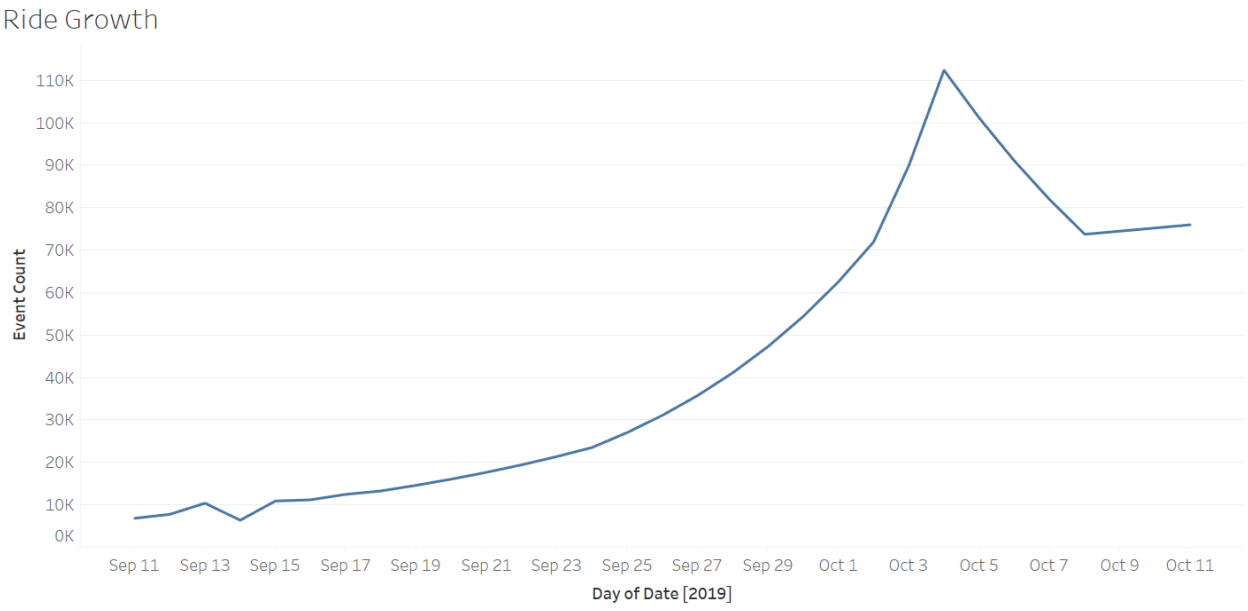


Image 3: Total Event Count

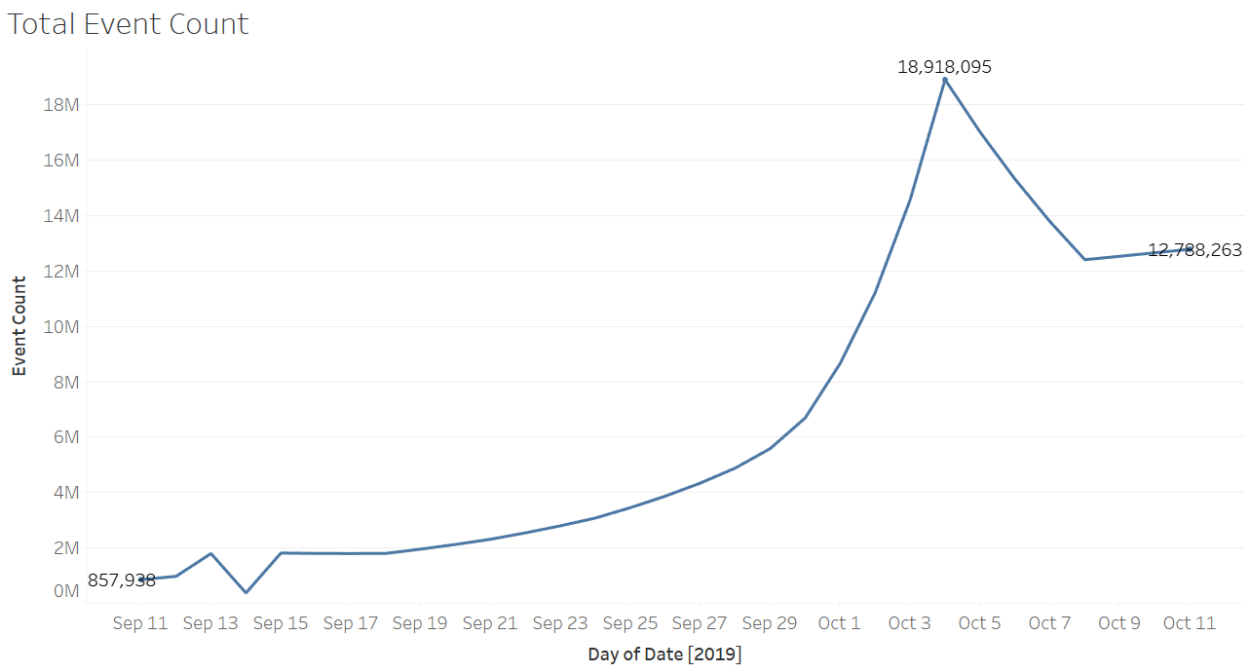


Image 4: All Events Log Scale