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 Team	Without an experienced and diverse engineering team it would be difficult to scale the app.	They need event and transactional data to monitor and improve the app. These optimizations can include load balancing, data management for when the user base has reached a certain threshold
Product Team	Product team is responsible to access the usability of the app and drive the retention and acquisition metrics forward.	By getting access to the event, transactional and usage data for the app, product team can get to know the pain points of the customers and then can take step to resolve them.
Marketing Team	Marketing team will be responsible to drive the targeted campaigns to expand the business by introducing newer or underutilized demographics.	They will need access to the event and the transactional data to derive their own campaigns. This data will include customer demographic for identifying the underserved market as we scale our

		services.
Customer Service Team	Customer service team will ensure that our services are satisfactory at all times and that we are dealing with customer queries on priority.	Customer support team will need to have the transactional and profile information of the riders and customers to be able to assist them with the query.
Finance Team	This team will ensure that we are in fact making revenue as without it our product will not survive. This team will also ensure that we are working within our budget so as not to go in debt.	This team will need access to transactional data and will also need to it in an aggregated format to make sense of the same. This will be one of the prime consumers of our processed data.

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 primary use-case?**
Engineering Team	They need event and transactional data to monitor and improve the app. These optimizations can include load balancing, data management for when the user base has reached a certain threshold	Event	Assess how much data is being produced?
Product Team	By getting access to the event, transactional and usage data for the app, product team can get to know the pain points of the customers and then can take step to resolve them.	Event and Entity data	Product Usability improvement
Marketing Team	They will need access to the event and the transactional data to derive their own campaigns. This data will	Entity	User Acquisition and Retention

	include customer demographic for identifying the underserved market as we scale our services.		
Customer Care	Customer team will need to have the transactional and profile information for the riders and customers to be able to assist them with the query	Entity	Manage customer complaints and improve the NPS
Finance	This team will need access to transactional data and will also need to it in an aggregated format to make sense of the same. This will be one of the prime consumers of our processed data.	Entity	Manage the PnL for the business

The tables we need are:

Table 1:

[Customer Table]

Customer ID	First Name	Last Name	Age	Email ID	Phone Number	Address

Each customer needs to be recognized uniquely; hence **Customer ID** is chosen as the **Primary key**. Since this table is specific the customer, hence this will not need any foreign key.

Table 2:

[Rider table]

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

Rider needs to be recognized uniquely; hence **Rider Id** is the **primary key** for this table.

Table 3:

[Vehicle Table]

Vehicle ID	Type of Vehicle	Cost	Status

Vehicles needs to be registered in the system and have their unique identification so that. Hence **Vehicle Id** is the **primary key** for vehicle table. In future we might also have to look into the vehicle maintenance part hence having a separate table to hold current condition (status) will help us out.

Table 4:

[Booking Table]

Booking ID	*Customer Id*	*Rider Id*	*Vehicle ID*	Booking Time	Pick up Location	Drop Off location	Distance Travelled	Payed Amount

Booking Id is the primary Key for the booking table. Since booking will also include the details of customer, rider and vehicle; hence we have customer id, rider id, and vehicle id as the foreign key in this table.

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. Connecting to the source of the data; here the engineering team can load up the csv file on the FTP server.
2. Extract the data from the file in rows to see what is being shared and to find if there is any unnecessary data that is being processed. This is needed because not all the data we have will be required for analysis and also if we need some more information we will be able to gauge that too.

3. Copying the extracted data on a staging environment where the data can be standardized. Since we might have data from multiple sources types like XML, JSON or flat files hence this will help us convert it into single format so that the analysis is efficient.
4. Connecting the data to a data warehouse where it can be accessed by different stakeholders as per their need.

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	7979

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

Number of events grouped by Date and Event Type

	Event Time							
Event Type	Null	5	6	7	8	9	10	11	12
begin_ride		38	49	62	86	57	57	78	18
choose_car		1,498	2,843	2,953	2,769	2,725	2,801	2,804	1,301
open	1	6,594	11,733	11,767	11,662	11,531	11,325	11,370	5,133
request_car		277	540	596	547	538	607	521	220
search		1,484	2,891	2,824	2,899	2,749	2,904	2,821	1,307

3. **How many events per device type per day?**

Number of events grouped by device type and date

	Event Time							
Device Type	Null	5	6	7	8	9	10	11	12
android		1,463	2,870	2,854	2,729	2,744	2,562	2,672	1,231
desktop_web		895	2,007	1,600	1,958	1,712	1,866	1,777	682
ios		2,384	4,337	4,217	4,373	4,380	4,482	4,500	2,026
mobile_web	1	5,149	8,842	9,531	8,903	8,764	8,784	8,645	4,040

4. **How many events per page type per day?**

Number of events grouped by page-type and event time

	Event Time							
Event Page	Null	5	6	7	8	9	10	11	12
book_page		1,977	3,548	3,576	3,572	3,586	3,424	3,506	1,639
driver_page		965	1,823	1,871	1,794	1,755	1,689	1,768	801
search_page	1	3,995	7,219	7,307	7,221	6,979	7,201	7,136	3,174
splash_page		2,954	5,466	5,448	5,376	5,280	5,380	5,184	2,365

5. **How many events for each location per day?**

Events grouped by location and event time

		Event Time							
User Neighbor	Null	5	6	7	8	9	10	11	12
Bronx		250	533	507	469	510	394	558	231
Brooklyn		2,009	3,737	3,590	4,025	3,440	3,400	3,556	1,594
Manhattan	1	6,869	12,591	12,807	12,180	12,270	12,371	12,200	5,580
Queens		595	842	905	893	1,026	1,069	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.

With the growing customer based our data will also increase, hence in order to scale we need to segregate the data into the following type:

- Event data: anything that happens in the app, like, user searching for ride, user registering and user booking a ride will be included.
- Customer data: How many customers are there on the app?
- Transactional data: a transaction is considered when a ride is booked and customer pays at the end of the ride.

In order to analyze this growing data we can introduce Data Lake wherein we can gather all of our data in one place and from our data lakes information is used by different consumers. We might want to look into ELT i.e. Extraction Load and Transformation because since we know that not all of the data needs to transform to reach a result.

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.

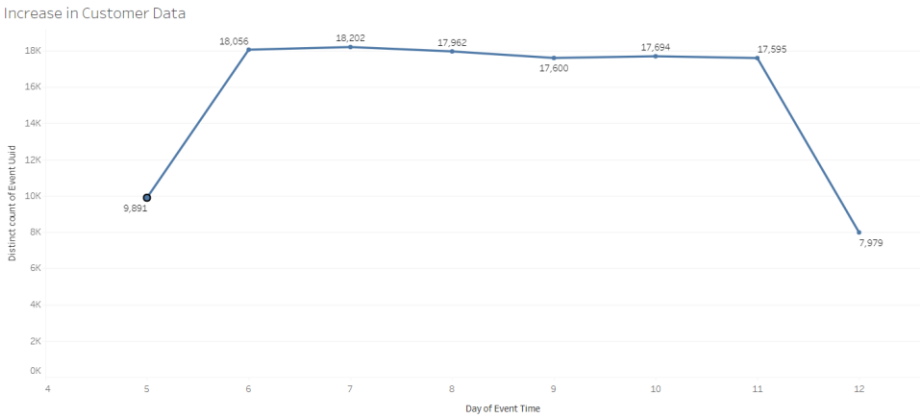
1. **How many events are being recorded per day?**
2. How many events of each event type per day?
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

For analyzing we need to know the growth of the event, **hence event data** is the most useful. Keeping this in mind, let's look at the data and answer the following questions:

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

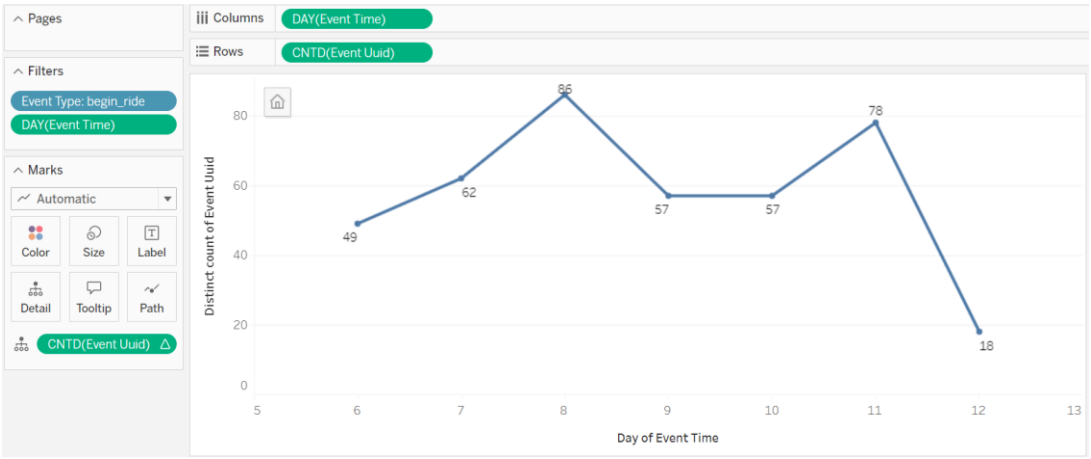
1. **How much is the customer data increasing?**

   From the data below we can see there is a spike on 6th October. We need to analyze this further to know the cause.



2. **How much is the transactional data increasing?**

- Since we established earlier that 'begin ride' will be our transactional data. Hence we filter our results accordingly.
- From the observation we can see the data is steadily rising with steeper in middle of the week then rise again in the weekends. This is in line with our observations in the proposal that most of the ride is happening during the end of the week.



3. **How much is the event log data increasing?**

The event data seems to follow a similar trend to the customer data. There is a sudden spike after **October 6.**

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

- **Event Log Data**
- Transactional Data
- Customer Data

Based on the above analysis we only have a week worth of data however we can clearly observe that we need to look into Event Data to understand and provide some business insights. We can now ask the engineers to provide us with a summarized view of data for a month with the following values: Event Type, Date and number of events.

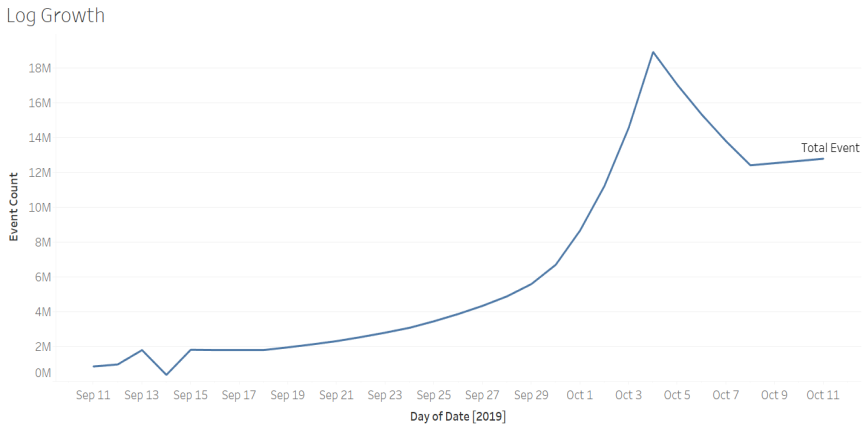
Section 6: Business Insights

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:

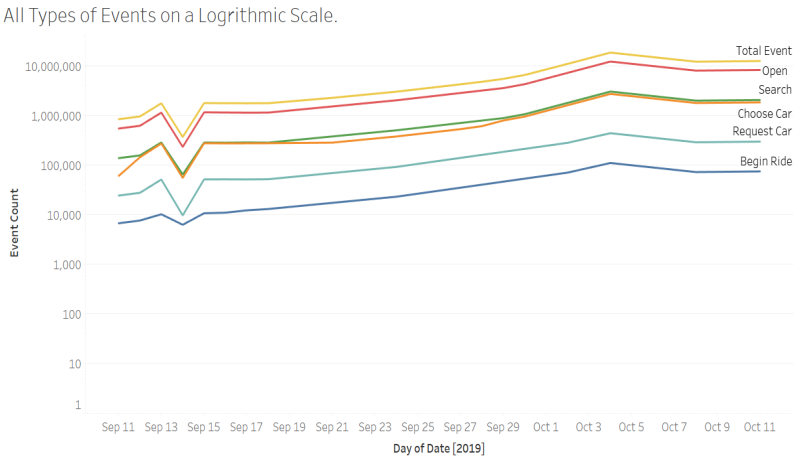


- There seems to be a massive growth in the October as compared to September. Might be a result of marketing campaign.
- From the data we can also see that there is a compounded growth which means **more users are interacting with our app and we have an increased engagement.**

What is the fastest growing data and why?

- According to the below graph the fastest growing data is **event data.** This means that we have a traction and engagement. Our marketing efforts are on point.
- **Search and choose** seems to have flatter curve, this indicates that we might want to optimize our Funnel.

- **Begin ride event is less than the Search and choose,** this tells us that we also need to look at user flow optimization as it might be an issue. This analysis will help us convert leads to customer thus improving our adoption metrics.



All Event Type Data

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

- There is a significant spike in **October** month when compared with September. This shows that our campaign has a positive impact in reaching out to our target customers. However, as a Data PM we know that just based on this we cannot credit the marketing campaigns for a significant spike. We might need to analyze this further to ensure our results are unbiased.
- We can also see that since we might have a huge spike in the data, we were able to have a stable app even with the increased load. This data is also essential to know for the engineers so that in future they are able to optimize this for load balancing.
- Marketing, Product and Engineering team needs to collaborate in order to avoid any negative impact. If a marketing campaign is successful we can see a huge traffic on the app, hence if at this points the product and engineers are not in the loop, it can result in a negative user experience and we might lose some potential customers. Also we also need to ensure that the increasing traffic on the site does not hinder the current functioning of the product otherwise we might lose existing customers with a negative impression of our product.

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.

Based on my understanding of the benefits of On Premise and Cloud DWH, I would propose to have Hybrid approach.
A self-service tool that allows users to explore the data freely while having centralized storage of data.

- Since Flyber is a new business hence we might not want to have very expensive DWH. We can balance it by choosing the most cost effective option to storing and analyzing our data.
- We are looking for scalability in a sense that once we have a large amount of data, we should be able to see a trend in real – time.
- Since we do not have a lot of in-house experts hence having some support from cloud based solution will help
- We need to ensure the latency and connectivity of various tools to our data warehouse here we can use an on premise solution which will have this ability. The also ensures the reliability of the data.

Suggested DWH

Provide an evidence based solution as to which DWH product is best for Flyber. Remember to address the factors above.
From the list provided, **Hadoop** could be a great on premise option as it is well known in the industry, cheap (relatively to its competitors) and is open-source. In terms of scalability is flexible and reliable to have our data on it; with a redundancy mechanism of multiple nodes, it ensure uptime in case there is a node failure

We need a data source that fulfills following needs:
- Support aggregated and transformed data to server variety of users
- Could be easily connected with a visualzation tool for the user to make quick business decision as this is a startup and we need it to scale rapidly.
- Support multiple file formats
- Is secure and possiblity of having customised security mechanism

Keeping in mind the above, **Snowflake** is a great contender for user facing tool for analyst.

Image Appendix

Image 1: Log Growth

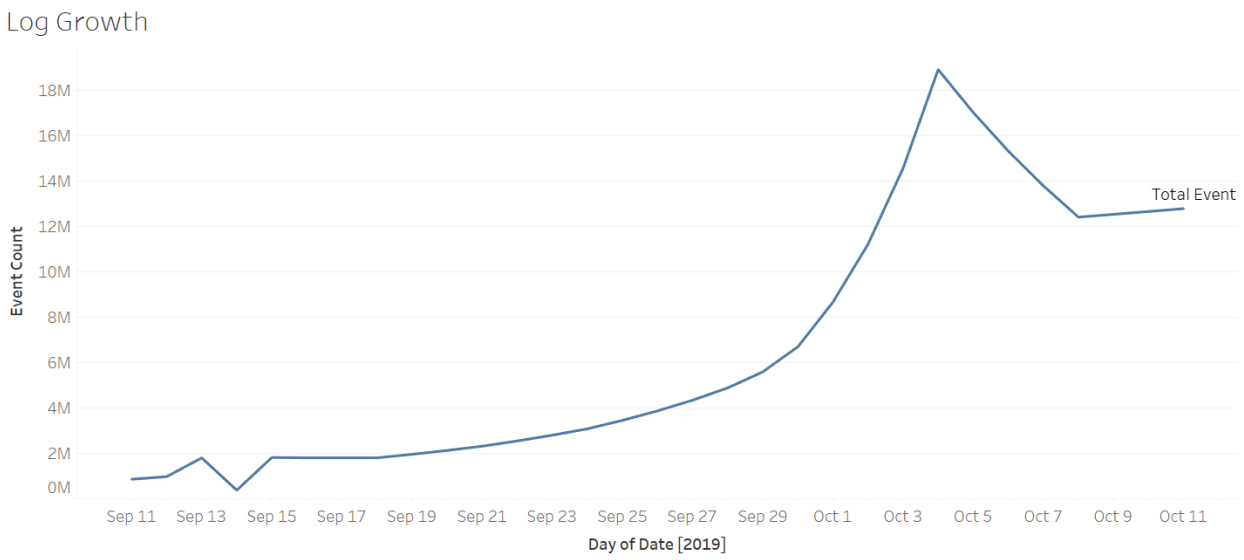


Image 2: Ride Growth

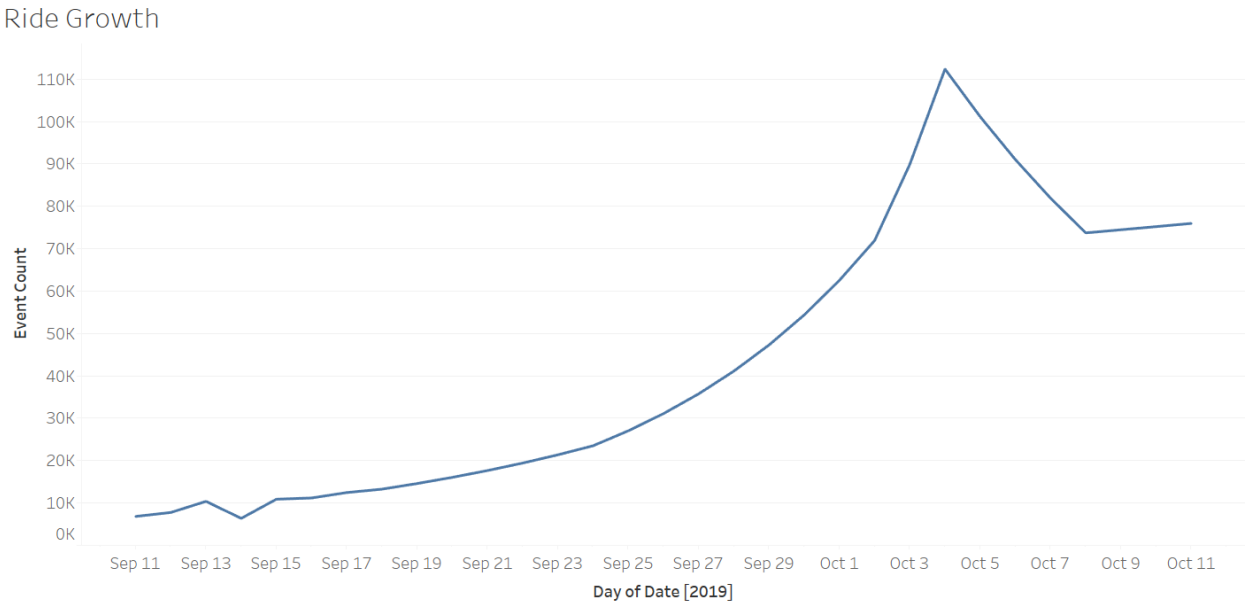


Image 3: Total Event Count

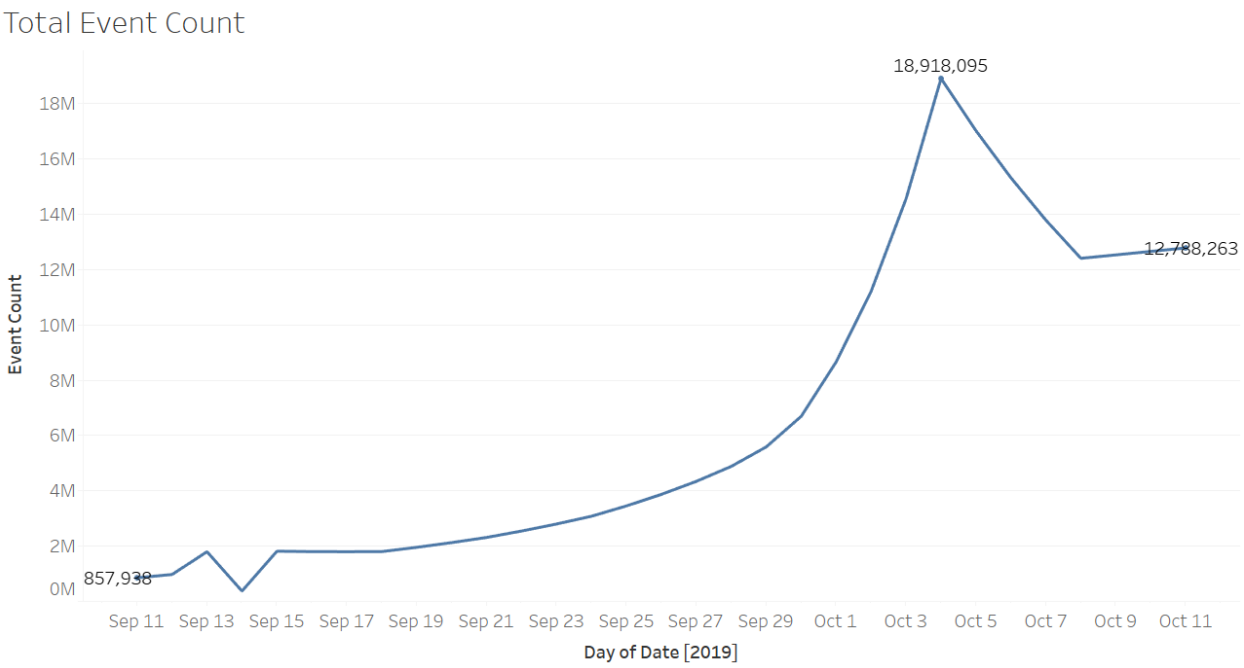


Image 4: All Events Log Scale