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 | They form the backbone for this app-based Flyber as they maintain the app and the backend infrastructure in Ship-Shape condition. | They make use of the data to monitor the app health, maintain status quo and to improve the app. |
| Payments team | They track the payments made by the customers to the Partners and resolve any issue faced in the process. | They make use of the data to validate customers are genuine, the partners get paid and the transactions are ACID compliant. |
| Product Management Team | They enhance the product and User Experience paving the way for user retention and enabling Flyber to take the game to the next level. | They make use of the event data to enable the Product iterations. |
| Customer Care Team | For Flyber to remain in the growth Phase the customer grievances should be handled properly for a better Customer experience. | They make use of the data to personalize the Customer experience. |
| Sales & Marketing Team | They capture new customers and help in existing customer retention. | They make use of the Data to do Targeted advertising. |

## **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 | They make use of the data to monitor the app health, maintain status quo and to improve the app. | Event Id, Event Timestamp, Event Type & Event Page | The app’s functionality and stability is paramount as Flyber is app-based. |
| Payments team | They make use of the data to validate customers are genuine, the partners get paid and the transactions are ACID compliant. | Customer Id, Payment Id, Payment amt, Partner Id, Customer account details, Partner account details | The business depends on these transactions. |
| Product Management Team | They make use of the event data to enable the Product iterations. | Event Id, Event Timestamp, Event Type & Event Page | Product improvement is key to Flyber sustaining it’s growth and increasing the Customer Base. |
| Customer Care Team | They make use of the data to personalize the Customer experience. | Customer Id, Customer name, Address, email id, Phone No., Order Id, Retail amount, Product Id, Tax rate | Personalized customer experience results in customer satisfaction, which is very important for the business. |
| Sales & Marketing Team | They make use of the Data to do Targeted advertising. | Customer name, Address, email id, Phone No, Age, Marital Status, Region, Parental Status, Order Id, Quantity, Product Id, Customer Id, Retail amount, Tax rate | This helps in reaching the Target segment thereby increasing the Product/Market fit and generating sales. |

**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:**

*[Payments]*

*(You may add more columns if necessary.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Payment\_Id | Customer\_Id | Payment\_Amt | Partner\_Id | Partner\_acct\_details | Customer\_acct\_details |

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

*[Payment\_id is the unique id for Payments and customer \_id is the foreign key from Customer table]*

**Table 2:**

*[Customer Details]*

*(You may add more columns if necessary.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Customer\_Id | Order\_Id | Customer\_Name | Address | Email\_Id | Phone\_No |

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

*[Customer\_Id is the unique identifier for Customers and Order\_id is the Foreign key from the Orders table]*

**Table 3:**

*[Customer Demographics]*

*(You may add more columns if necessary.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Customer\_Fullname | Age | Gender | Region | Marital\_Status | Parental\_Status |

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

*[Composite Primary Key of Customer\_Fullname and Age, no foreign key here.]*

**Table 4:**

*[Orders]*

*(You may add more columns if necessary.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Order\_Id | Product\_id | Customer\_Id | Quantity | Retail\_Amt | Tax\_Rate |

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

*[Order\_Id is the unique identifier here hence it’s the primary key and Product\_Id and Customer\_Id are the foreign keys from the Product and Customer tables respectively]*

**Table 5:**

*[Product]*

*(You may add more columns if necessary.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Product\_Id | Customer\_Id | Size | Colour | Cost | SKU No. |

Rationale for Choosing Primary and Foreign Keys for the Table 5:

*[Product\_Id is the unique identifier hence the Primary Key and Customer\_Id is the foreign key from Customer 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:

*(You may add more steps if necessary.)*

1. *[For less Event Data We can use Countif for aggregation]*
   1. *[In less time we can do Event data aggregation and get the required information also]*
2. *[For large Event Data set we should start with Pivot Table]*
   1. *[When many columns and rows are there Pivot Table helps in Analysing the data faster]*
3. *[First we should be clear about the fields that should go into the columns and values section respectively]*
   1. *[It makes for faster and simple analysing]*
4. *[Then Place the field to be analysed in the rows section and change the fields as per different requirements]*
   1. *[This helps in getting a tabular data]*

*As Flyber data is growing exponentially manual ETL won’t help and it isn’t scalable. Hence automated ETL pipelines will be required for faster ingestion and processing.*

**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** |

1. 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** |

1. 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 |

1. 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 |

1. 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.

*[Insert Response Here.]*

**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 2nd Question:* **How many events of each event type per day?** is specific in nature and helps us to get to the relevant data faster

1. Here the Customer Data’s increase can be measured by looking at the “Request Car” event type data. There has been a net 1124% increase in the data.

2. Here the Transactional Data can be measured by looking at the “Begin Ride” event type data. There has been a net 1011% increase in the same.

3. Here the event log data increase can be observed by looking at the Total Event Data, which has seen an increase of net 1518%.

Event Log data is most suitable for answering this question.

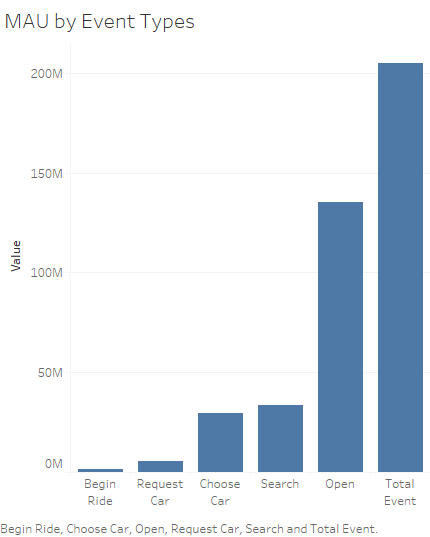
## **Section 5: [Optional]** Loading and Visualization On Your Own This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

## 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:



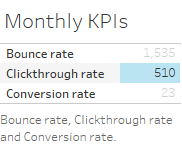
**Data Story:** This graph tells us:

*Monthly Active Users by Different Event Types.*

This graph was created using the following steps:

1. *Click on the Measure values from Measures section.*
2. *Automatically horizontal bars come up.*
3. *Then sort ascending.*

Visualization 2:



**Data Story:** This graph tells us:

*KPI metrics for event data. MAU, shown in the previous graph, is very high but the conversion rate is comparatively less at 23%. For further explosive growth the conversion rate needs to be increased. The Clickthrough rate is great at 510%. The bounce rate is very high so we need to improve our landing page as well as Splash page. Eventhough Flyber is doing pretty well, over a period of time we can increase the metrics further to take the game to the next level.*

This graph was created using the following steps:

1. *First create the calculated fields Bounce rate, Clickthrough rate and Conversion rate.*
2. *Then click on each of them one by one.*
3. *The text tables come automatically.*

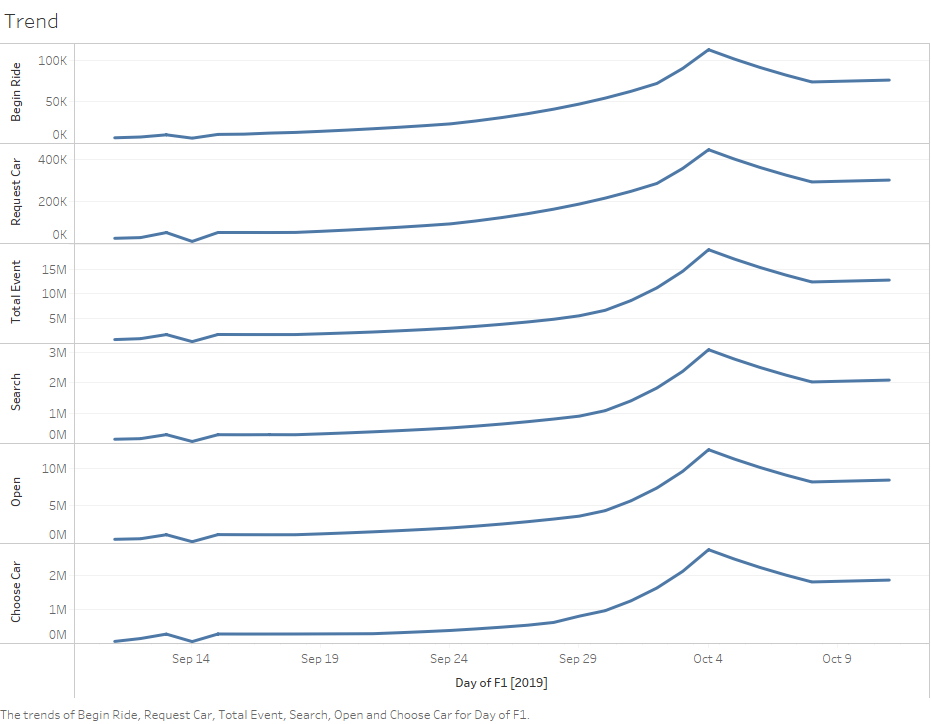
## **Section 6:** Business Insights

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 Time-period dimension F1 was plotted on x-axis and the different event types data on the y-axis for the continous lines visualization.*

What is the fastest growing data and why?

*Total Event data is the fastest growing data as it’s the sum of all the event types, which are growing individually.*

**All Event Type Data**

Visualization:

*[Insert Visualization Here] If you didn't create your own, please use the images in the appendix to guide you through this section.*

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
* *The Graph pattern has shown exponential growth for all the event Data types.*
* *It’s good.*
* *Yes, the October Marketing campaign has done great.*
* *The campaign impact has been very positive as the growth had picked up exponentially during the October month and maintained the level during the same.*
* *There has been a strong positive correlation between the Marketing campaigns and the 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 DWH will best serve Flyber as in long term the cost is less compared to on-prem. We can scale it fast according to our requirements and also Auto-scaling is there. As per our usage more resources can be added and removed too when the usage is less hence optimizing resource usage and cost. As in-house expertise is not too big a factor we can go for cloud. Similarly latency isn’t a concern now as we are serving in the US geographic region and for future cloud serves well if we expand internationally. Regarding connectivity we have different Availability zones for each region hence downtime is very less as per the Cloud service provider SLAs. So Reliability is very good in Cloud as we can fall back on another availability zone when disaster hits a zone.

**Suggested DWH**

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

*Google BigQuery DWH is the best Cloud DWH for Flyber as it charges only for the storage and computing used. Auto scaling is there which makes it easy to scale up and down. It also has a massive Parallel processing architecture which allows multiple workloads to run concurrently at great speed. Also the compute and storage resources are separate making it highly flexible to scale. It supports stream processing enabling real time processing and also it supports multiple formats. Within a region or for multiple regions data is replicated allowing for hassle free uptime. It also has AI/ML support functionality thus allowing for more data analytics use cases. It’s also easier to manage from end user perspective. Also if Flyber is using other google platforms it’s easier to integrate into Google Bigquery. In conclusion, for the varied and multiple needs of Flyber it makes great sense to go with Google Bigquery.*

## Image Appendix

Image 1: Log Growth

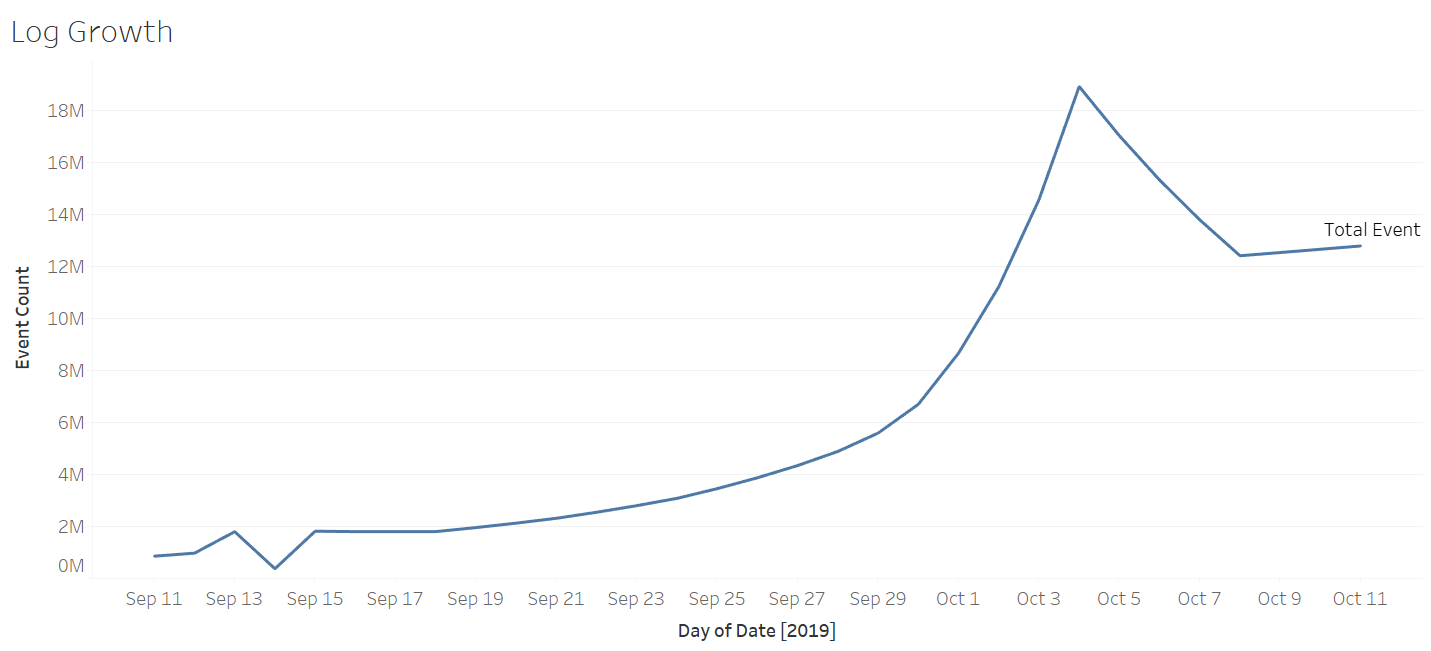


Image 2: Ride Growth

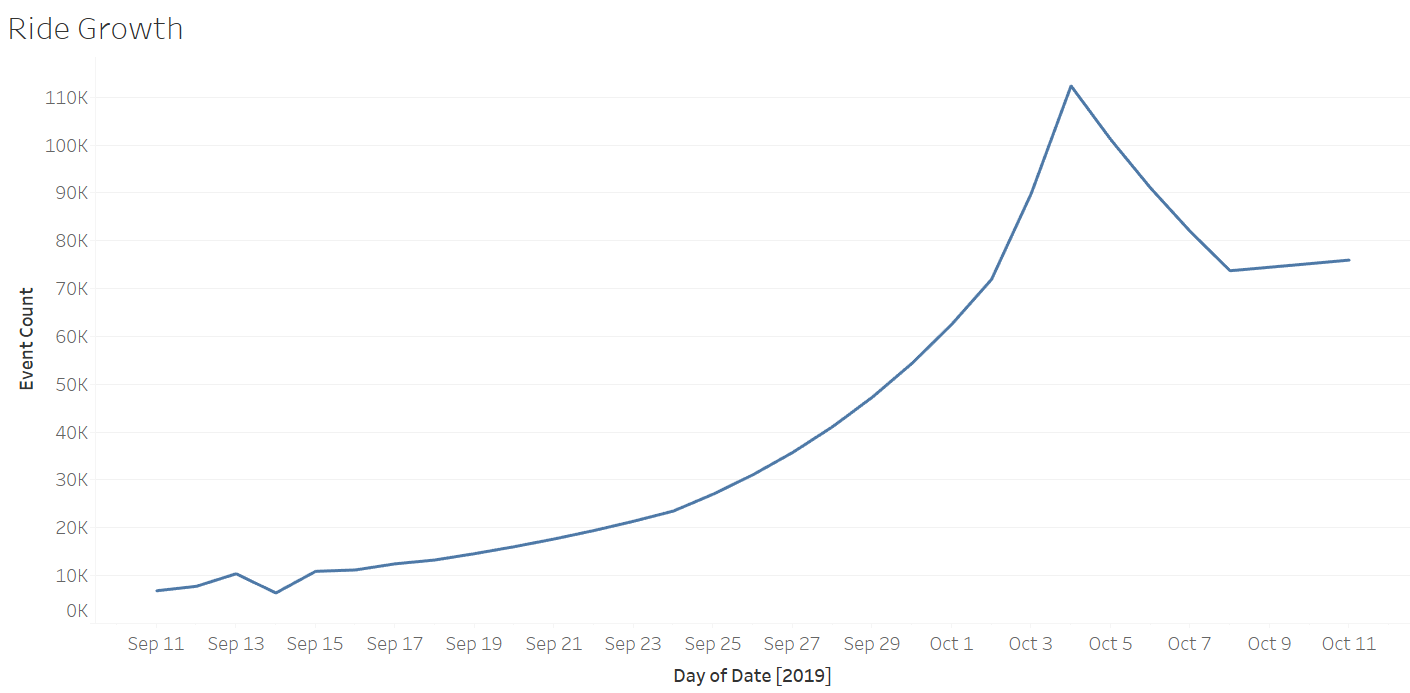


Image 3: Total Event Count

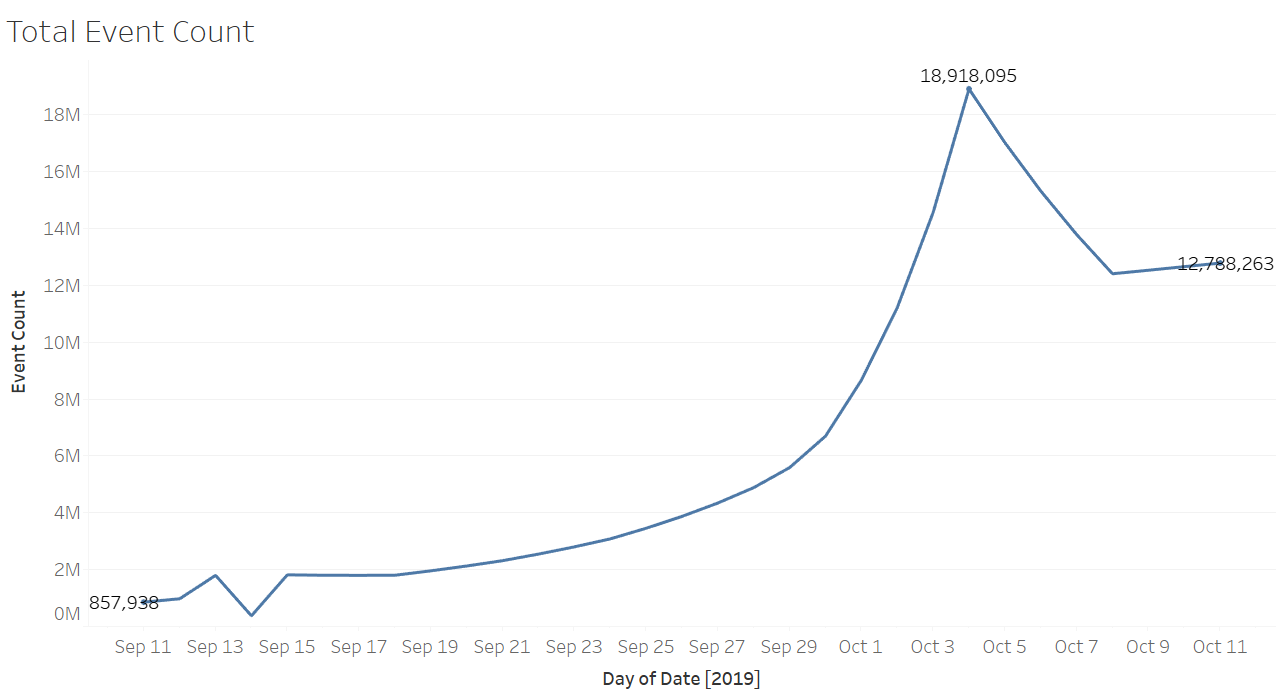


Image 4: All Events Log Scale

