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** |
| Developers | Necessary to maintain a working product | As a Developer I need access to all data in platform to ensure stability and improve upon user experience. |
| Marketing | Needed to identify new riders, ways to retain existing riders | As a marketer, I need to view our ride usage data to identify the customer demographics most likely to use our service |
| Product Management | Must understand how riders are interacting with platform to identify and prioritize new features. | As a Product Manager, I need to view transaction and customer data to identify and prioritize improvements in our platform. |
| Customer Assistance | Must have access to customer account history in case of an issue using Flyber service | As a Customer Care specialist, I need to view customer event history to identify the appropriate solution to their assistance request. |
| Finance | Must have ability to view revenue / costs to monitor and forecast | As a financial analyst, I need to track our revenues and operating costs to forecast profit / loss |

## **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?** |
| Developers | As a Developer I need access to all data in platform to ensure stability and improve upon user experience. | Data Generated by the website and app with a dashboard and monitoring alerts | Without a 99.999% uptime on platform, we will lose our customers and revenue stream.  Comparing Prediction to Action trip duration will be needed to optimize trip prediction times and targetting where to send Flyber drivers. |
| Marketing | Advertising | Business Intelligence, with Customer Profile and preferences | Marketing needs to understand who to target with ad campaigns to drive growth while minimizing ad spend. |
| Product Management | As a Product Manager, I need to view transaction and customer data to identify and prioritize improvements in our platform. | Event Data, NPS Data, Customer Feedback Data,  BI / Visualization Tool | In order to prioritize Rider needs, PM needs to identify what are their pain points |
| Customer Assistance | As a Customer Care specialist, I need to view customer event history to identify the appropriate solution to their assistance request. | Customer Profile,  Customer Trip History  Visualization Tool | Needed to provide personalized responses to resolve issues encountered with Flyber app/service. |
| Finance | As a financial analyst, I need to track our revenues and operating costs to forecast profit / loss | Monitoring Profit and Loss, Expense Reports, BI / Visualization Tool |  |

**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 \_Table*

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Rider\_ID*  *(PK)* | *Phone* | *Name* | *Email* | *Address* | *Birthday* |

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

*[Your response here.]*

Rider\_ID is needed as primary key to enable this table to joined to other tables, using this table’s primary as Foreign Key. No foreign key is required for this table.

**Table 2:**

*Trip\_Table*

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Trip\_ID*  *(PK)* | *Rider\_ID*  *(FK)* | *Pickup\_Location* | *Dropoff\_location* | *Predicted\_Duration* | *Actual\_Duration* | *Cost* | *Driver\_ID*  *(FK)* | *Trip\_Rating* |

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

*[Your response here.]*

*Trip\_ID is the primary key to differentiate between trips. Rider\_ID is Foreign Key to join to Rider table. Driver\_ID is Foreign Key to join to Driver Table*

**Table 3:**

*Driver\_Table*

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Driver\_ID*  *(PK)* | *Name* | *Phone* | *Vehicle\_ID*  *(FK)* | *Pay\_per\_mile* |

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

*Driver\_ID is primary key to enable one entry per driver and link as foreign key to other tables. Vehicle\_ID enables join to vehicle table*

**Table 4:**

*Vehicle\_Table*

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Vehicle\_ID*  *(PK)* | *Make* | *Model* | *Mileage* | *Service\_Start\_Date* | *Last\_Maintenance\_Date* |

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

*Driver\_ID is primary key to enable one entry per driver and link as foreign key to other tables. Vehicle\_ID enables join to vehicle table*

**Table 5:**

*Event\_Table*

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Event\_ID*  *(PK)* | *Event\_Name* | *Duration* | *Event\_Timestamp* | *Device\_type* | *User\_id (FK)* | *Session\_id* |

Event\_ID is the PK as one entry per event is required. User\_ID is added as a Foreign Key to enable joining this to retrieve additional user or trip details.

## **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. Confirm Event\_Time is set to Date & Time    1. Needed to ensure any timestamp analysis is not considered a number or string |
| 1. Transform Neighborhood to [value] + ", " + "NY"    1. Needed in case of geographic mapping desired |
| 1. Create Pivot Table    1. Needed to aggregate and summarize data |
|  |

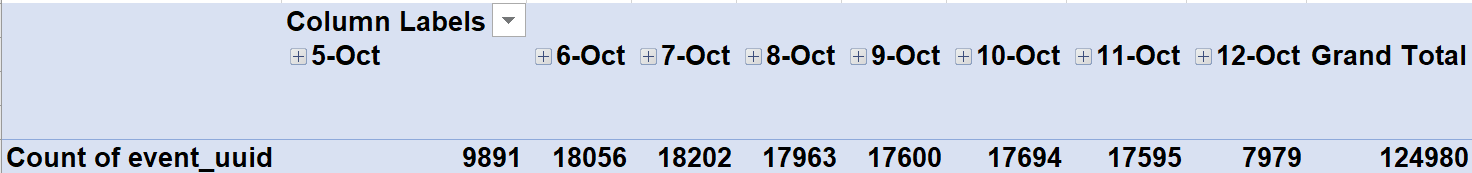
**Transformation-2**

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 |  |  |  |  |  |  |  |

See pivot table below



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

See pivot table below

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row Labels** | **5-Oct** | **6-Oct** | **7-Oct** | **8-Oct** | **9-Oct** | **10-Oct** | **11-Oct** | **12-Oct** | **Grand Total** |
| begin\_ride | 38 | 49 | 62 | 86 | 57 | 57 | 78 | 18 | 445 |
| choose\_car | 1498 | 2843 | 2953 | 2769 | 2725 | 2801 | 2804 | 1301 | 19694 |
| open | 6594 | 11733 | 11767 | 11662 | 11531 | 11325 | 11371 | 5133 | 81116 |
| request\_car | 277 | 540 | 596 | 547 | 538 | 607 | 521 | 220 | 3846 |
| search | 1484 | 2891 | 2824 | 2899 | 2749 | 2904 | 2821 | 1307 | 19879 |
| **Grand Total** | **9891** | **18056** | **18202** | **17963** | **17600** | **17694** | **17595** | **7979** | **124980** |

1. How many events per device type per day?

See pivot table below

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row Labels** | **5-Oct** | **6-Oct** | **7-Oct** | **8-Oct** | **9-Oct** | **10-Oct** | **11-Oct** | **12-Oct** | **Grand Total** |
| android | 1463 | 2870 | 2854 | 2729 | 2744 | 2562 | 2672 | 1231 | 19125 |
| desktop\_web | 895 | 2007 | 1600 | 1958 | 1712 | 1866 | 1777 | 682 | 12497 |
| ios | 2384 | 4337 | 4217 | 4373 | 4380 | 4482 | 4500 | 2026 | 30699 |
| mobile\_web | 5149 | 8842 | 9531 | 8903 | 8764 | 8784 | 8646 | 4040 | 62659 |
| **Grand Total** | **9891** | **18056** | **18202** | **17963** | **17600** | **17694** | **17595** | **7979** | **124980** |

1. How many events per page type per day?

See Pivot Table below

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row Labels** | **5-Oct** | **6-Oct** | **7-Oct** | **8-Oct** | **9-Oct** | **10-Oct** | **11-Oct** | **12-Oct** | **Grand Total** |
| book\_page | 1977 | 3548 | 3576 | 3572 | 3586 | 3424 | 3506 | 1639 | 24828 |
| driver\_page | 965 | 1823 | 1871 | 1794 | 1755 | 1689 | 1768 | 801 | 12466 |
| search\_page | 3995 | 7219 | 7307 | 7221 | 6979 | 7201 | 7137 | 3174 | 50233 |
| splash\_page | 2954 | 5466 | 5448 | 5376 | 5280 | 5380 | 5184 | 2365 | 37453 |
| **Grand Total** | **9891** | **18056** | **18202** | **17963** | **17600** | **17694** | **17595** | **7979** | **124980** |

1. How many events for each location per day?

See Pivot Table below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row Labels** | **5-Oct** | **6-Oct** | **7-Oct** | **8-Oct** | **9-Oct** | **10-Oct** | **11-Oct** | **12-Oct** | **Grand Total** |
| Bronx | 250 | 533 | 507 | 469 | 510 | 394 | 558 | 231 | 3452 |
| Brooklyn | 2009 | 3737 | 3590 | 4025 | 3440 | 3400 | 3556 | 1594 | 25351 |
| Manhattan | 6869 | 12591 | 12807 | 12180 | 12270 | 12371 | 12201 | 5580 | 86869 |
| Queens | 595 | 842 | 905 | 893 | 1026 | 1069 | 936 | 386 | 6652 |
| Staten Island | 168 | 353 | 393 | 396 | 354 | 460 | 344 | 188 | 2656 |
| **Grand Total** | **9891** | **18056** | **18202** | **17963** | **17600** | **17694** | **17595** | **7979** | **124980** |

**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 above steps were completed using Pivot Tables in Tableau. However, this is not scalable and requires manual work each time we want to review the data. A scalable solution would require moving from CSV to a database connected to Tableau. The same data transformation steps would be required, and then a custom dashboard would be generated creating the above views, leveraging the additional visualization functionality available in Tableau.*

**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? ~2700 Customers/Day
2. How much is the transactional data increasing? 60-80 rides/day
3. How much is the event log data increasing? Growth of ~18,000 events/day

Chart

Description automatically generated

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

* Event Log Data
* Transactional Data
* Customer Data

**Event Log Data** because this provides the most complete and largest data set to understand how users are interacting with our app.

## **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:

Chart, line chart

Description automatically generated

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

*This tells us how often a Flyber session results in a ride initiated.*

This graph was created using the following steps:

1. *Create Calculated Field named ‘Ride Conversion’ calculated as (Begin Ride / Open).*
2. *Add Ride Conversion(Sum) to row*
3. *Add Date (F1) as a measure (Day of Month) as Column.*
4. *Convert to Line*

*Chart, line chart

Description automatically generated*

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

*The count of events by event type over time. It enables us to see we were scaling exponentially until an event occurred on October 4th, which reduced our growth which only began to recover a few days later.*

This graph was created using the following steps:

1. *Add F1/Day as Measure to Columns*
2. *Add each of the Event(SUM) to Row*
3. *Drag each line chart up to the next to display on a single axis.*
4. *Drag Measure Names to Color*
5. *Drag Measure Names to Label*

Chart, line chart

Description automatically generated

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

*The sum of total events per day in the 1 month of data we have available.*

This graph was created using the following steps:

1. *Add F1/Day as Measure to Columns*
2. *Add Total Event(SUM) to Row*
3. *Convert to Line*

## **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:

Chart, line chart

Description automatically generated

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

*Text, timeline

Description automatically generated*

Our forecast based on the month of data shows Flyber’s event data growing by 3,275,846 events for the following 7 days. Note that due to the high variation in growth rates over the preceding 30 days, there is high variability in the forecast. While growth is expected, it is possible we see a further reduction in events as seen during the 10/4-10/9 period. See the forecast below.

Chart, line chart

Description automatically generated

What is the fastest growing data and why?

*The fastest growing data is in App Open events. We can see an increase in nearly 250,000 events between 10/8 and 10/11.*

Chart, line chart

Description automatically generated

**All Event Type Data**

Visualization:

Chart, line chart

Description automatically generated

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 data shows we were experiencing strong growth until October 4th, when we see sharply reduced App Opens and subsequent events before resuming on a lower upward trajectory the remaining few days in our data set.*

*We can infer from this that our October Marketing campaign had a negative impact, based on the reduction in sessions following the start of the marketing campaign. We would expect to see an increase in the growth rate of our data generation following marketing spend, so we must re-evaluate the campaign we initiated against the user data we have, to ensure the advertising we commit to both grows new users while maintaining our existing users.*

## **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.

*Given we expect to eventually reach a similar level of data as Uber, we can leverage a similar solution as they have, with an On-Premise solution. Hadoop is affordable, given it is open-source software, offers significant scalability, with its supported tools like Cassandra, Spark, Hive and Uber-generated open-source tool Hudi to support multiple databases, data warehouses and efficient ETL processes.*

*Diagram

Description automatically generated*

[*Per Uber’s architecture*](https://eng.uber.com/uber-big-data-platform/)*, this enables 30 minute latency for raw data, 1 hr for modeled data, and this toolset supports a wide variety of common programming languages (java, scala, python, r, and sql) to enable our existing developers the opportunity to set-up with minimal ramp-up time.*

*By leveraging on-premise, we avoid the chance of spiraling* [*cloud compute costs*](https://www.wsj.com/articles/coronavirus-pandemic-forces-businesses-to-confront-cloud-computing-costs-11595849519)*, while addressing the key requirements needed by our data warehouse strategy.*

**Suggested DWH**

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

*As referenced above, we will leverage a combination of Hadoop DFS with related Apache projects (Hive, Spark, Hudi, Parquet, and Kafka).*

*Uber has already demonstrated this solution meets the need for pedabytes of data updated at sub – 1hr latencies. Leveraging this enables us to start with infrastructure in place to handle exponential growth in data processing without requiring rearchitecting in another 2 years.*

## Image Appendix

Image 1: Log Growth

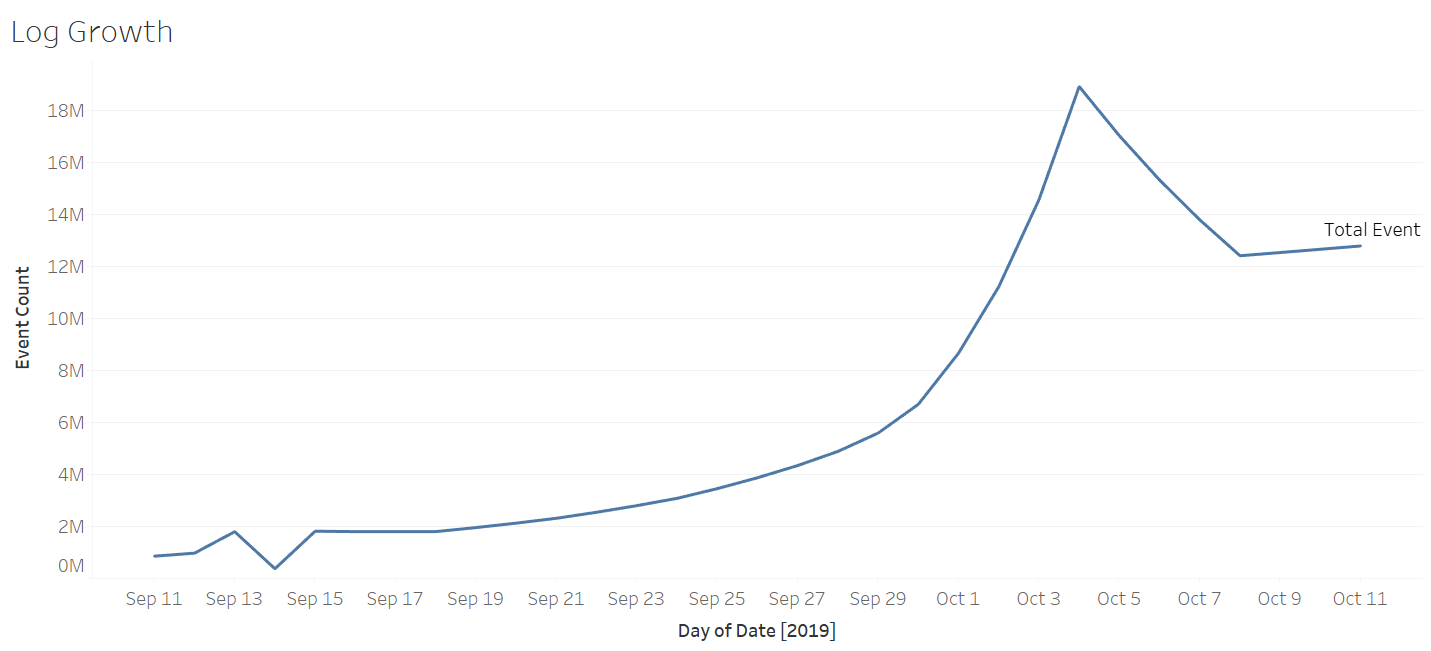


Image 2: Ride Growth

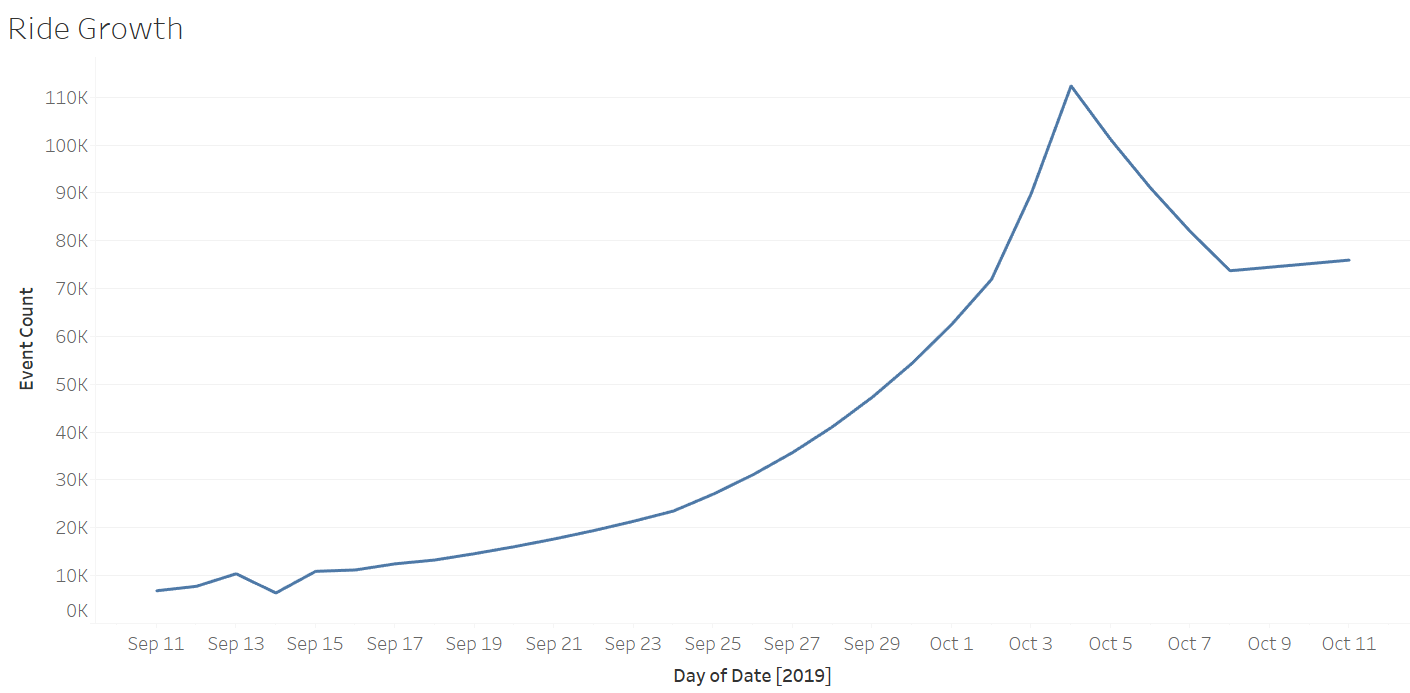


Image 3: Total Event Count

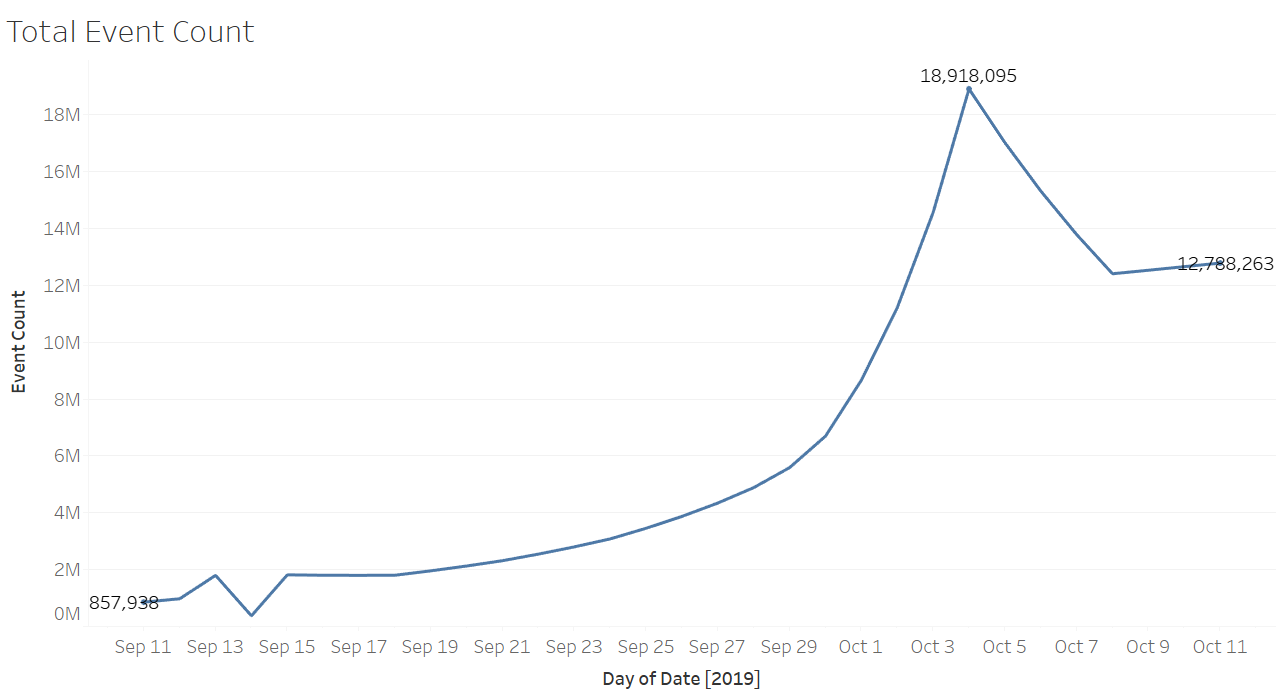


Image 4: All Events Log Scale

