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**
Customer Experience Teams	With the collected data if there is a problem or a process that needs to be optimized, CX teams can determine them and take actions.	Checking customer complaints and taking actions.
Product and Project Management Teams	This is the team that plans next steps for the MVP, so they will need data to make analyses about the product's usage.	They will create new futures and work on optimization projects according to data.

Finance Team	This is the team that manages the company's accounting, so they will need data to take actions on all decisions.	They will calculate the general profit, and determine the parts of the product that have the most cost. Take actions and precautions on expenses.

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?**
Customer Experience Teams	This team needs compliances data to determine the user problems and solve them.	Data should contain user id, ticket creation date, ride id and compliance. Also we need user info so we can reach them if necessary.	CX teams focus on user issues. To be able to determine them, they need compliance data. So that they can determine the pain points and increase user satisfaction.
Product and Project Management Teams	This team needs data to decide some optimization work. Are the drop off and pick up nodes right choices? Do we need to expand operations?	Data should contain ride id, pickup and drop off locations, ride durations.	These teams need to decide if the current process is working correctly and if we need to expand our operations, or change our existing one. For example, if one of our pick up nodes is not being used as much as we predicted, we might change its location.
Finance Teams	This team needs data to calculate our expenses and income.	Ride id, fuel cost, driver cost, ride price.	So we can determine if we make profit, do we need to increase our prices.

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:

flyber_customer_tickets

ticket_id *(primary key)*	*ride_id* *(foreign key)*	*User_id* *(foreign key)*	*compliance_message*	*rating*

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

- *Ticket id should be our unique key, so each ticket can be questioned separately. This is a ticket based table.*
- *Ride_id and user_id are our foreign keys, so with these id's we can question ride and user information if necessary.*

Table 2:

flyber_user_info

user_id *(primary_key)*	*name_and _surname*	*email*	*phone_number*

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

- *User_id is our primary key for this table, because every user has one communication info, and we update this row if any changes happen.*
- *I didn't determine a foreign key for this table.*

Table 3:

flyber_rides

ride_id *(primary_key)*	*User_id* *(foreign_key)*	*Driver_id* *(foreign_key)*	*pickup_location*	*dropoff_location*	*duration*

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

- *Every roe determines a ride so ride_id is the primary key*
- *user and driver info can be reached via ids if necessary they are the foreign keys.*

Table 4:

flyber_ride_cost

cost_id *(primary_key)*	*ride_id* *(foreign_key)*	*driver_cost*	*fuel_cost*	*taxes*	*ride_price*

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

- *Every row determines cost data so cost id is the primary key.*
- *Each row is related to a ride so ride_id is the foreign key.*

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. *Determining data sources and products*
    a. *We first need to find the source of the data we need.*
    b. *We need to talk to the product teams and developer to find the proper data source.*
2. *Data verification*
    a. *Checking if the collected data type is usable.*
    b. *If the collected data meets the need of stakeholders.*
3. *Clearing the data*
    a. *Cleaning data for healthy analyses*
    b. *Removing duplicated and meaningless data*
4. *Creating reports and data visualizations*
    a. *Creating the reports and visualizations stakeholders needs.*

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	2834	4337	4217	4373	4380	4482	4500
android	1463	2870	2854	2729	2744	2562	2672
Desktop Web	895	2007	1600	1958	1712	1866	1777
Mobile Web	5149	8842	9531	8903	8764	8784	8646

4. How many events per page type per day?

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

5. How many events for each location per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Manhattan	6869	12591	12807	12180	12270	12371	12201
Brooklyn	2009	3737	4025	4025	3440	3400	3556
Bronx	250	533	469	469	510	394	558
Queens	595	842	893	893	1026	1069	936
Staten Island	168	353	396	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.

We can extract this data and store it in a data warehouse or a data lake. We need an automated pipeline, so we can read large amounts of data automatically if we need to. Data contains columns and types that can have a data type like strings and integers. We can use id's and date columns to read data incrementally, so we don't need the full data everytime a new event is published.

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

We can get the most insight from the question:

- *How many events of each event type per day?*

	Event Time							
Event Type	5	6	7	8	9	10	11	Grand Total
begin_ride	38	49	62	86	57	57	78	**427**
choose_car	1.498	2.843	2.953	2.769	2.725	2.801	2.804	**18.393**
open	6.594	11.733	11.767	11.662	11.531	11.325	11.371	**75.983**
request_car	277	540	596	547	538	607	521	**3.626**
search	1.484	2.891	2.824	2.899	2.749	2.904	2.821	**18.572**
Grand Total	**9.891**	**18.056**	**18.202**	**17.963**	**17.600**	**17.694**	**17.595**	**117.001**

How much is the event log data increasing?

- *Event log data almost doubles after the first date, and then stabilizes around 17 to 18 thousand per day after the first date.*

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

- *Event log data is the most important. Because using these events, we can determine insights like how many rides are being made, how many of the customers don't do the rides even though they*

search for the ride, etc.

- *These questions will help us to understand if we need an improvement on our MVP and make us question the reasons why aren't customers using it even though they use the app.*
- *Begin ride events can only be generated when a ride actually begins. So this will be the correct data to determine the transactions per day. We can see that in the first couple days it increased and then after the 4th day there was some decrease. But overall we managed to increase this data after the first day.*
- *We can't really determine if we have new customers or the same customers making the ride since there is no indicator in data for that. But if we check the distinct user id that begins a ride per day, we can see that in the first couple days it increased and then after the 4th day there was some decrease. But overall we can see some growth.*

Section 6: Business Insights

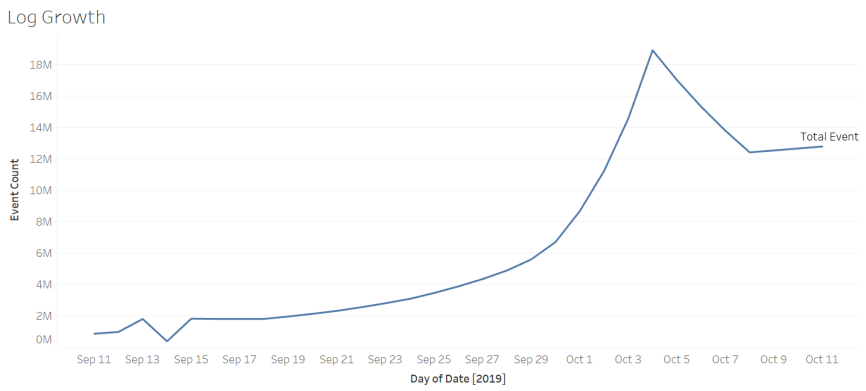
The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received 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

In early September logs were under 2 million and by the first couple of days of October it peaked to 20 million. So it increased approximately 20 times in 1 month.

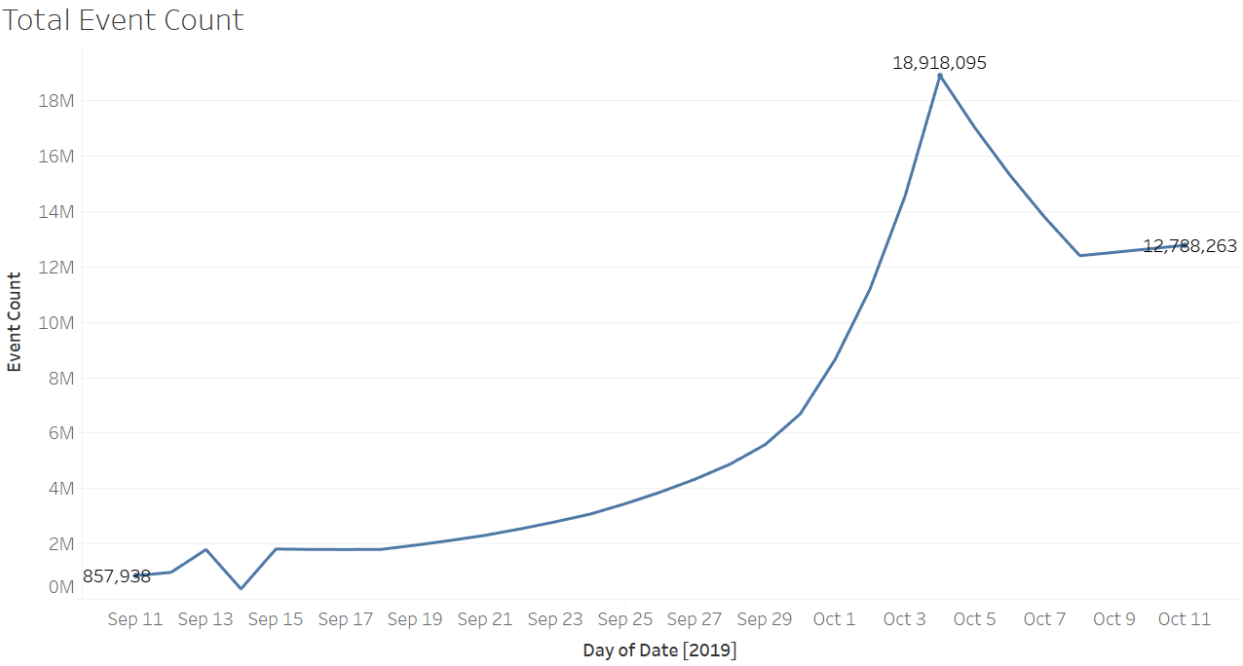
Visualization:



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

We used the event count data to observe the data growth over the days.

What is the fastest growing data and why?

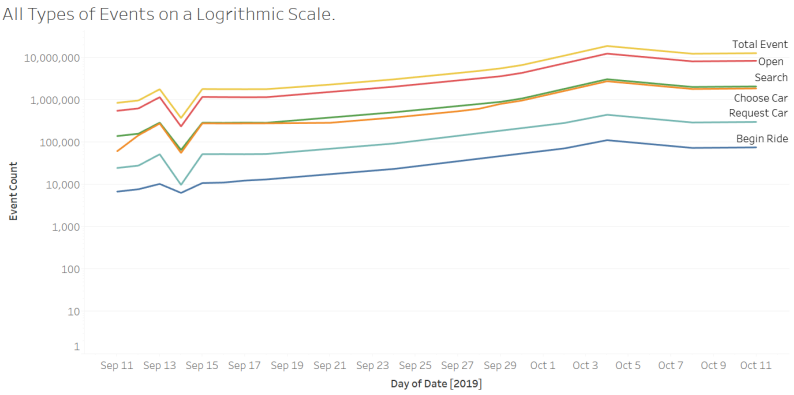


Total event count data is the fastest growing one. As the ride count increases, multiple events are published per ride so they increase the most.

It started with 857.000 logs per day and peaked at nearly 19.000.000. So it increased approximately 22 times over the month.

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

1) *Graph Pattern: We can see that generally over the days, there are more events published from producers. This means our product is being used and it is being used more and more over the days.*
2) *We can say that it is a good growth.*
3) *The October marketing campaign seems to be very successful considering there is a huge jump on the graphics at the beginning of October.*
4) *Even though there were some decreases after the campaign's first days, we still managed to increase our interaction significantly, compared to the month before the campaign.*
5) *Marketing campaigns let users know about the product. So users download the app, check for the rides even if they don't take rides. So these interactions increase our event counts. With the event logs 'begin ride' we can see how many of them are actually using the app, or how many of them are just checking the 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.

Cloud DWH would be the best option for our product.

- *Since we still are a new product, we don't know how much data growth we will have in the future.*
- *Cloud DWH has the capacity to scale up and down.*
- *They mostly have pay as you go futures, so we can arrange the price if our data increases.*
- *Infrastructure doesn't require large, regular investments.*
- *Most of them have automated back-up systems and support.*
- *They are highly encrypted.*

Suggested DWH

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

We can use Google BigQuery as our DWH. (I checked bigquery's website for the reasons below.)

- *It has storage based pricing so we can pay as we go. If hour data increases, we can buy more storage.*
- *It provides full support for database transaction semantics. BigQuery storage is automatically replicated across multiple locations to provide high availability.*
- *You have full control over who has access to the data stored in BigQuery. BigQuery makes it easy to maintain strong security with Cloud Identity and Access Management, and your data is always encrypted at rest and in transit.*

Image Appendix

Image 1: Log Growth

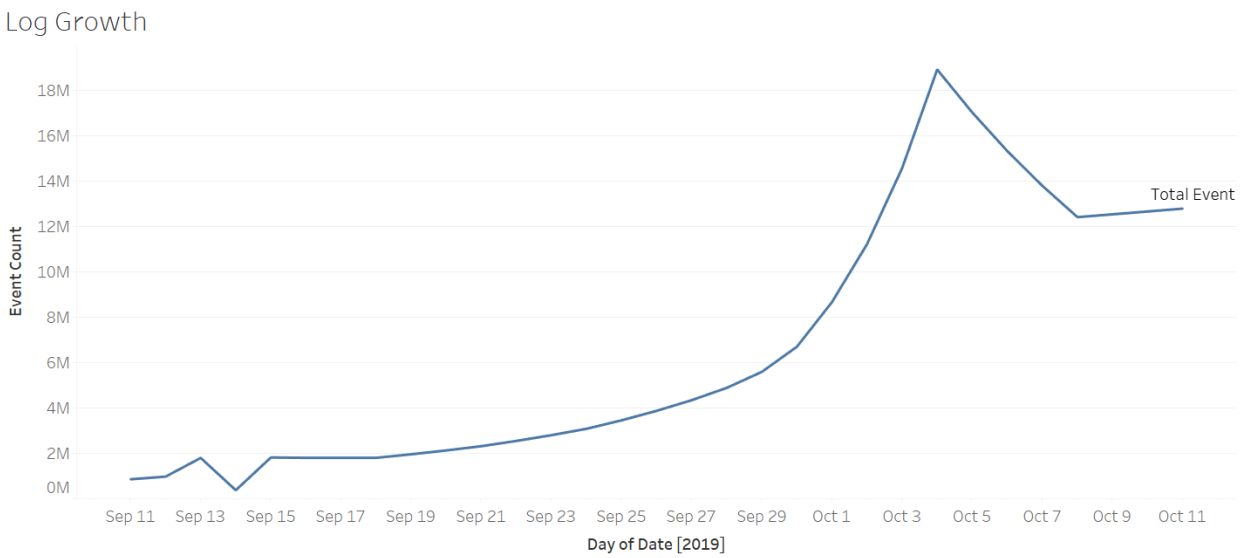


Image 2: Ride Growth

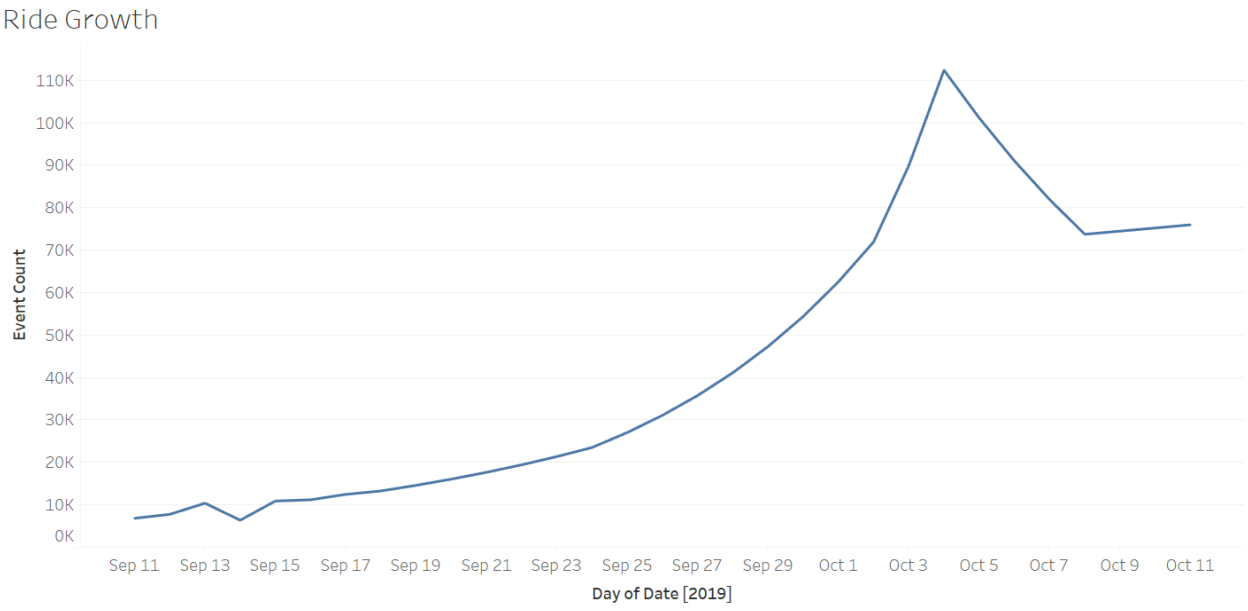


Image 3: Total Event Count

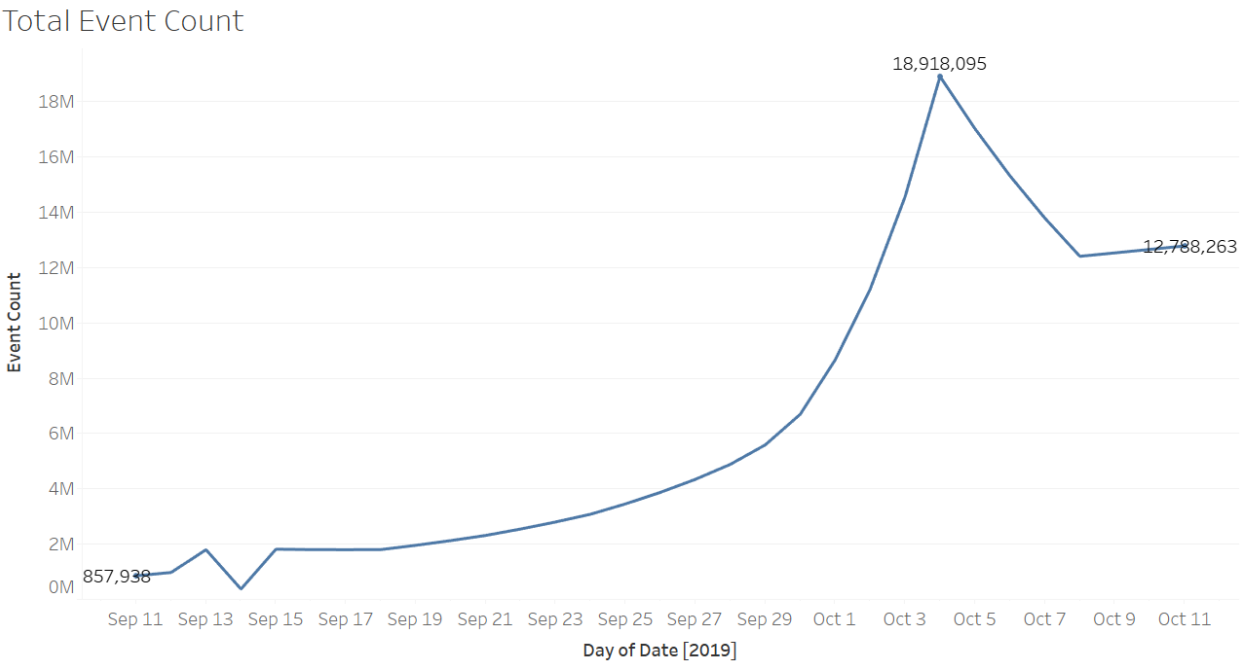


Image 4: All Events Log Scale