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**
Marketing	Getting new Customers and Leads	Targeted Advertising
Product Management	Improving the Product	Identify Customer Pain points to improve the product
Engineering	Build and maintain working application for reserving rides. Maintenance and repair of mini copters	Monitor App and site performance. Monitor Operations of copters .
Finance	Estimating profit and loss	Monitor current Revenue

Inventory Management	Purchase of Copters and hiring pilots	Monitor demand for Copters and pilots.
Compliance and Regulatory Department	Deal with Air traffic Control Board and make sure that the company and its employees are in compliance.	Monitor copters to ensure they follow air traffic rules. Pilots are trained regularly.
Customer Care	Addressing Customer grievances	Provide personalized response to customers.

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?**
Marketing	Targeted Advertising	Customer details, Ride details. **Entity Data**	Identify new Leads for targeted advertising.
Product Management	Improving the Product	Customer Experience. **Entity+Event Data**	Improve the product based on Customer Pain points.
Engineering	Monitoring Application Copter maintenance and repair.	Website/App events for reserving rides. **Event Data** Data from Copter sensors and Flight recorder.**Event Data**	Monitoring App/site status to ensure that website is working as expected, Monitor Copter status through sensor readings and data recorders on copters to ensure copters are in good condition to fly.
Finance	Estimate P&L	Aggregated Transactional Data, Number of rides, Cost of operating Copters,	Estimate Revenue earned to project if the product is going to be viable financially in near future.

		Charge to customers. **Entity Data**	
Inventory Management	Purchase of Copters and hiring pilots	Number of Reservations made based on pickup zone and drop off zone. **Entity Data**	Determine demands based on popular pick up /drop off zones.
Compliance and Regulatory Department	Safety and Air traffic compliance	Data from flight data recorder. **Event Data**	Ensure that copters are following air traffic rules and ensuring safety of passengers
Customer Care	Addressing customer's grievances	Customer feedback and rating from app/site. **Event Data + Entity Data**	Provide personalized response to customers.

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:

Ride Service

Ride Service ID (Primary Key)	*Pilot ID(FK)*	*Customer ID(FK)*	*Transaction ID(FK)*	*Pick Up Zone*	*Drop off zone*	*Wait time until pick-up*	*Amount paid*

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

Ride Service ID is the primary key here corresponds to unique ride hailed by a customer. Each ride has Foreign key as pilot ID, passenger/customer ID and transaction ID that would map to Pilot/driver table Customer table, Transaction table.

Since we are focusing on rider side of the business, we are not going to describe Pilot Table.

Customer details are present in customer table and transaction details are stored in Transaction table that will be used by Finance and Inventory Department.

Table 2:

Customer

Primary Key: Customer ID

Customer ID	First Name	Last Name	Address	Email

Table 3:

Customer Demographics

Primary Key: Customer ID

Customer ID	Age	Gender	Marital Status	Parental Status	Race

There are no foreign keys, here the two tables save customer details that is one row for each customer.

Table 4:

Transaction

Primary Key: Transaction ID

Foreign Key : CustomerID , PilotID, *Ride Service ID*

Transaction ID	CustomerID	PilotID	*Ride Service ID*	Transaction Date	Total Amount Paid by Customer	Tip to the Pilot	Taxes

Each transaction for a ride is saved in this table, so primary key is transaction ID representing a transaction between Pilot and Driver for one ride service. The fields involve transaction details basically date, payment made by customer and tax information which would help us estimate amount earned by the company.

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. **Identifying Data Source:**

    a. Entity Data /Structured data from Relational database systems like the Tables for Customer ,Ride service and Transaction tables.
    b. Semi structured data/Event Data from Web application and mobile application. This mainly Event Data which constitutes activities like opening the page, searching for ride service, booking a reservation etc.
    c. Unstructured Data/Event Data from Flight recorder which consists of sensor readings and events during ride duration.

2. **Data Collection**: Collect Event Data from web, mobile application and Flight recorder and organize them in tables. Collect Entity data from database by querying (Online Extraction) or using data files(Offline Extraction) needed columns and rows.
    a. Choose event_id, user_id, event time, event date, event type from Event Data.
    b. For Entity data, Choose Ride ID, and get information of amount payed, pick up and dropoff zone and extract Customer information from Customer Table using CustomerID.

3. **Data Verification:**
    a. Check the current Data Type
    b. Get an idea of number of records present in the table
    c. Verify that entity data is retrieved via incremental Extraction because for our current MVP we need latest changes in data.

4. **Data Cleaning:**
    a. Delete duplicate rows.

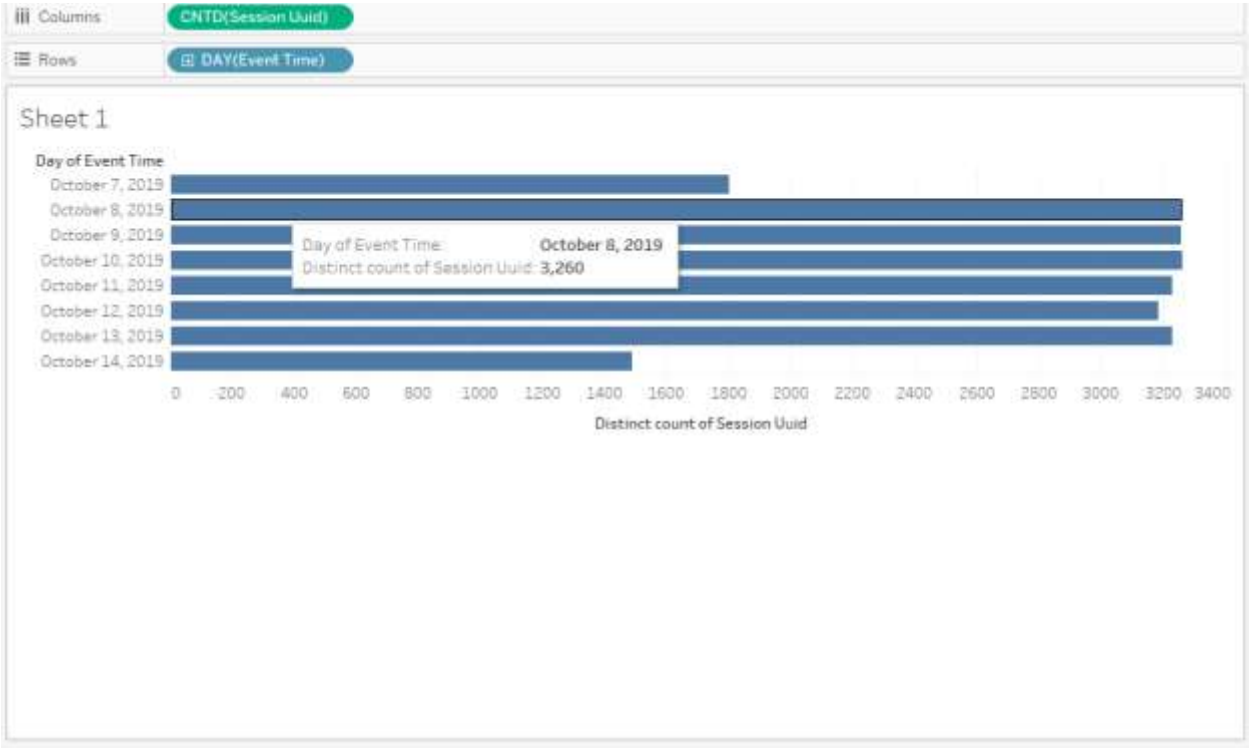
    b. Confirm that records we have corresponded to what was recorded inititally.

Transformation-2

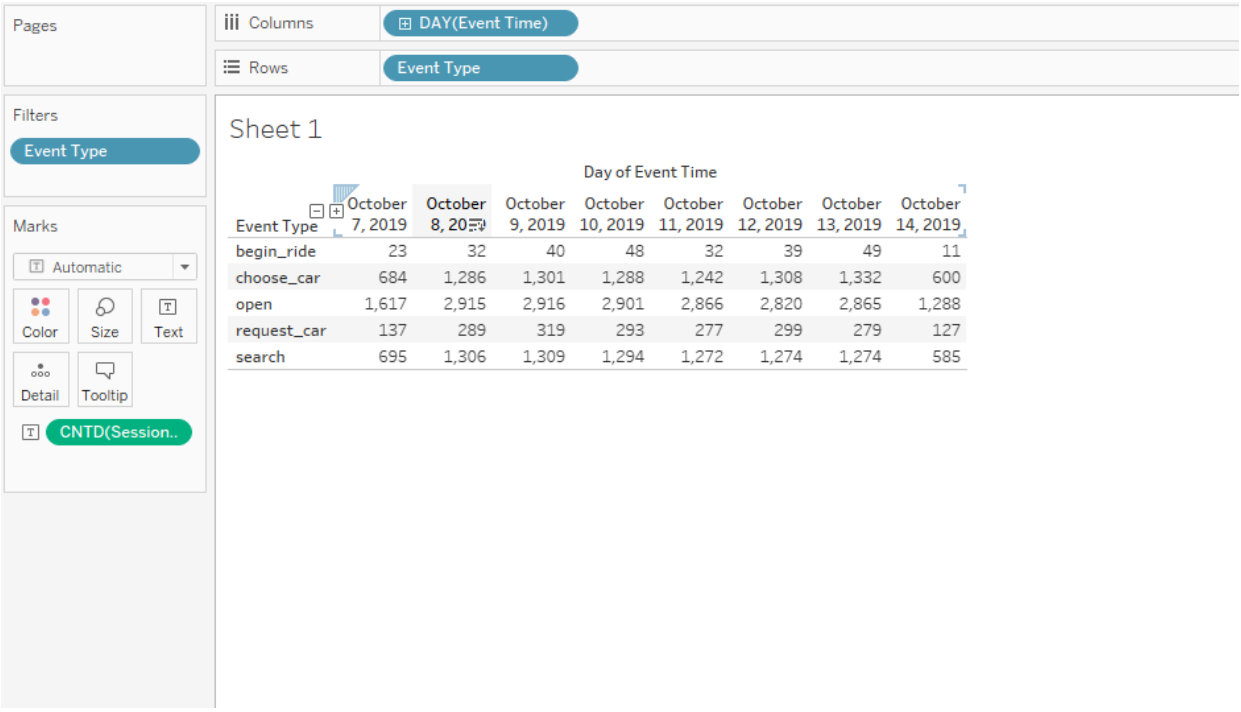
Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?

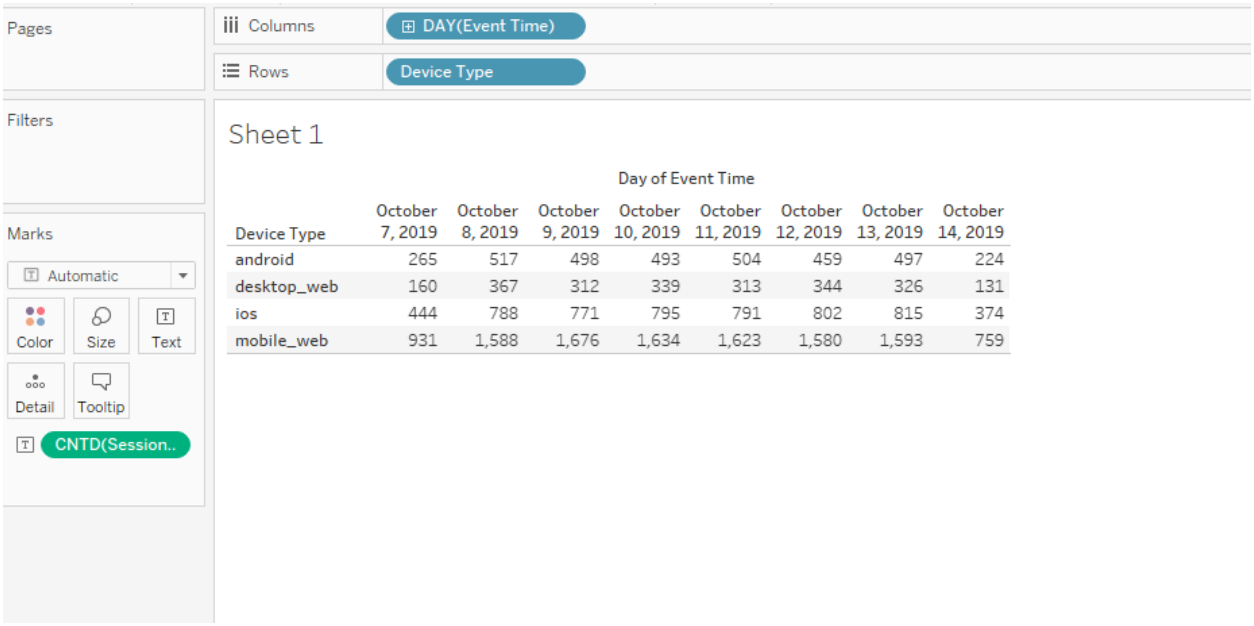
Date	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019	10/12/2019	10/13/2019	10/14/2019
Event Count	1800	3260	3257	3261	3231	3185	3231	1488



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

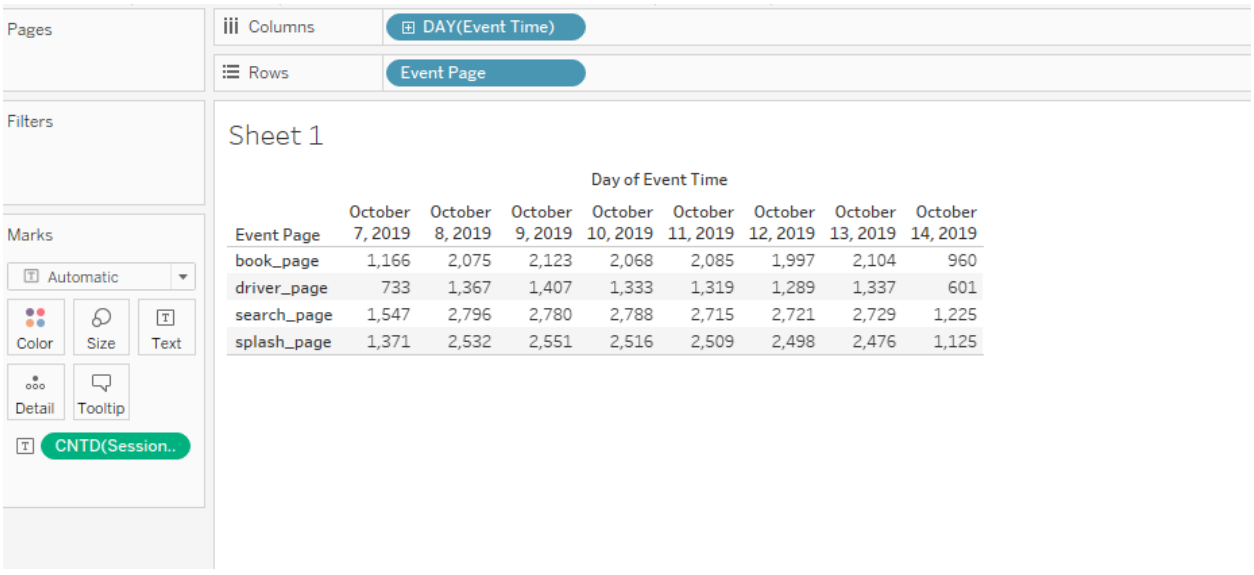


3. How many events per device type per day?



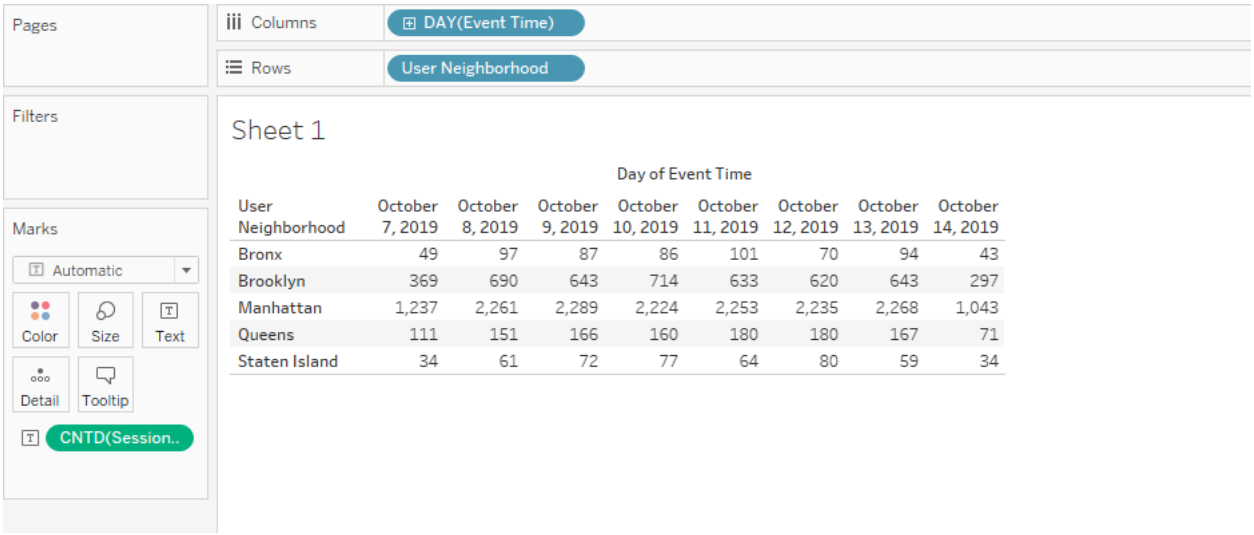
Device Type	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019	October 13, 2019	October 14, 2019
android	265	517	498	493	504	459	497	224
desktop_web	160	367	312	339	313	344	326	131
ios	444	788	771	795	791	802	815	374
mobile_web	931	1,588	1,676	1,634	1,623	1,580	1,593	759

4. How many events per page type per day?



Event Page	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019	October 13, 2019	October 14, 2019
book_page	1,166	2,075	2,123	2,068	2,085	1,997	2,104	960
driver_page	733	1,367	1,407	1,333	1,319	1,289	1,337	601
search_page	1,547	2,796	2,780	2,788	2,715	2,721	2,729	1,225
splash_page	1,371	2,532	2,551	2,516	2,509	2,498	2,476	1,125

5. How many events for each location per day?



User Neighborhood	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019	October 13, 2019	October 14, 2019
Bronx	49	97	87	86	101	70	94	43
Brooklyn	369	690	643	714	633	620	643	297
Manhattan	1,237	2,261	2,289	2,224	2,253	2,235	2,268	1,043
Queens	111	151	166	160	180	180	167	71
Staten Island	34	61	72	77	64	80	59	34

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.

It is a lot of repetitive work of extracting and transforming the data. If the dataset is small this might be feasible. Even if process is available as scripts, it is still a cumbersome process of execution and involves lot of manual work. What we need is an automated pipeline. Since we are still in the process of building MVP and we cannot afford to build our own infrastructure and have a scalable datacenter, we can use existing cloud platform service providers such as AWS or GCP.

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.

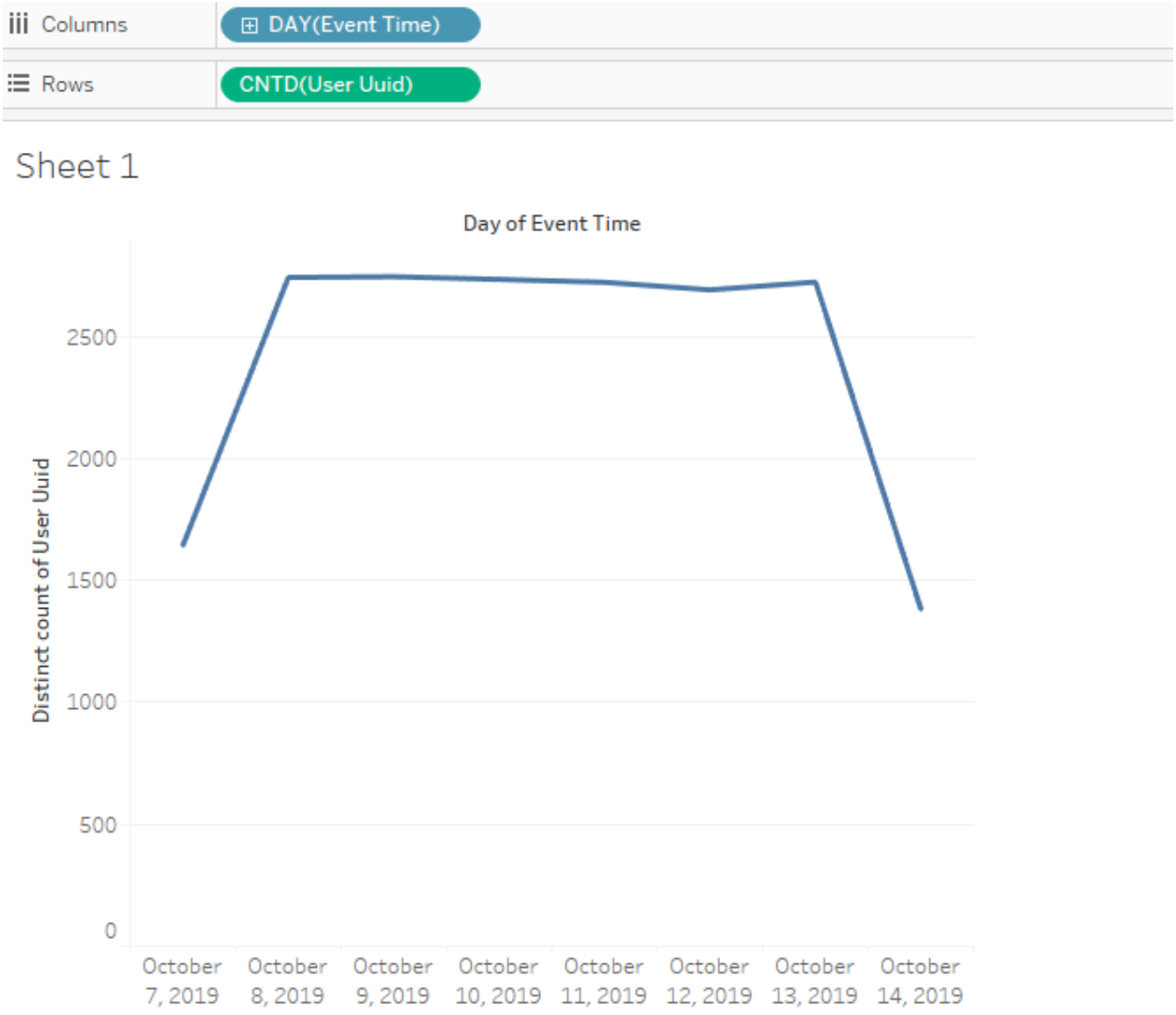
Engineering is willing to provide some data, but they have asked for the criterion that is most important. To First provide your business question and provide a rationale for why this is the most important.

Choose one of the following prompts that you think can get you the most relevant information to proceed further.

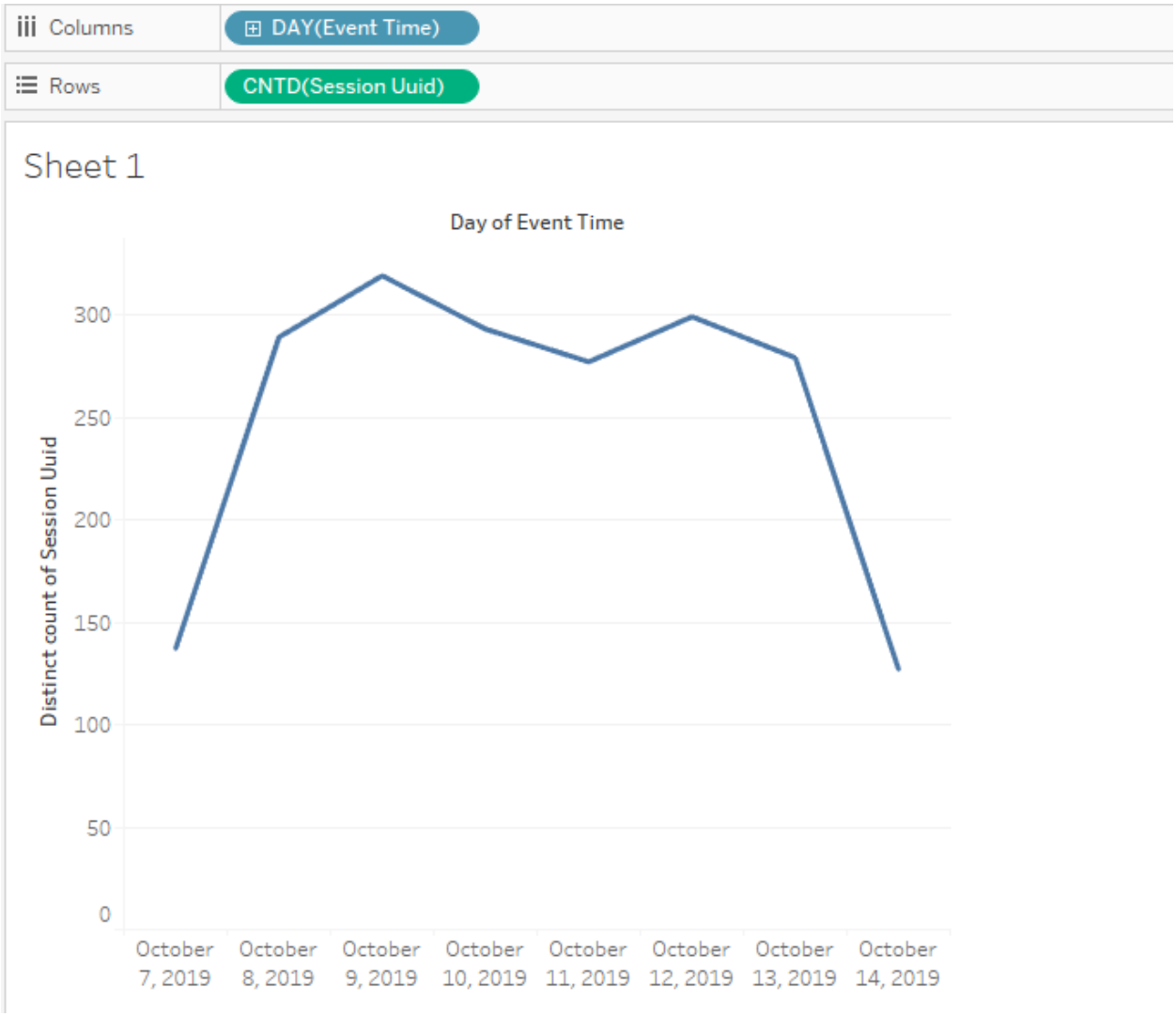
1. How many events are being recorded per day?
2. How many events of each event type per day? This is what I choose
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

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?

   From the data above, Monday has the least event logs when compared to rest of the days of the week. There is an increase on Tuesday and Wednesday, Thursday, Friday, Saturday and Sunday have pretty much same data with no significant increase.

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

- Event Log Data : This gives you events occurred based on User interaction with Web or mobile application. It tells you how many people who opened the app vs how many actually booked the ride using web/mobile application.This is most important as you can pretty much extract all other data from events.

- Transactional Data :
  Transaction data corresponds to request_car event. At this point we know that customers have made a reservation successfully. The data is least on Monday. But significantly higher on Sunday and Wednesday.

- Customer Data: This is basically count of unique User Uuid in the event log data.

Section 5: [Optional] Loading and Visualization On Your Own

This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

After sharing your criterion with engineering, they give you a new set of data: Section 5 Event Type Log also available in the classroom resources. Also provided in the project resources section.

Engineering provided you with the data you want, but you still have yet to achieve your ultimate goal as a Data Product Manager. Now, utilize the data to make business decisions. Your executives do not want you to give them a bunch of data tables; instead, they prefer visualizations to help convey the key insights succinctly. Visualizing this data will help you understand the underlying trends and help you determine the story that needs to be told in your proposal to executives.

In this section, you can load and visualize the data into whatever platform you would like. A Python Notebook, Tableau or any other visualization tool you are familiar with. Create two visualizations that might help you to better understand your data trends and place either a screenshot or exported image of your visualizations and the details of each below. Please provide the steps you took to visualize your data and what the visualization tells you about your data.

Visualization 1:

[Insert Visualization Here.]

Data Story: This graph tells us:

[Insert Response Here.]

This graph was created using the following steps:

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

Visualization 2:

[Insert Visualization Here.]

Data Story: This graph tells us:

[Insert Response Here.]

This graph was created using the following steps:

1. *[Insert Step Here.]*

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

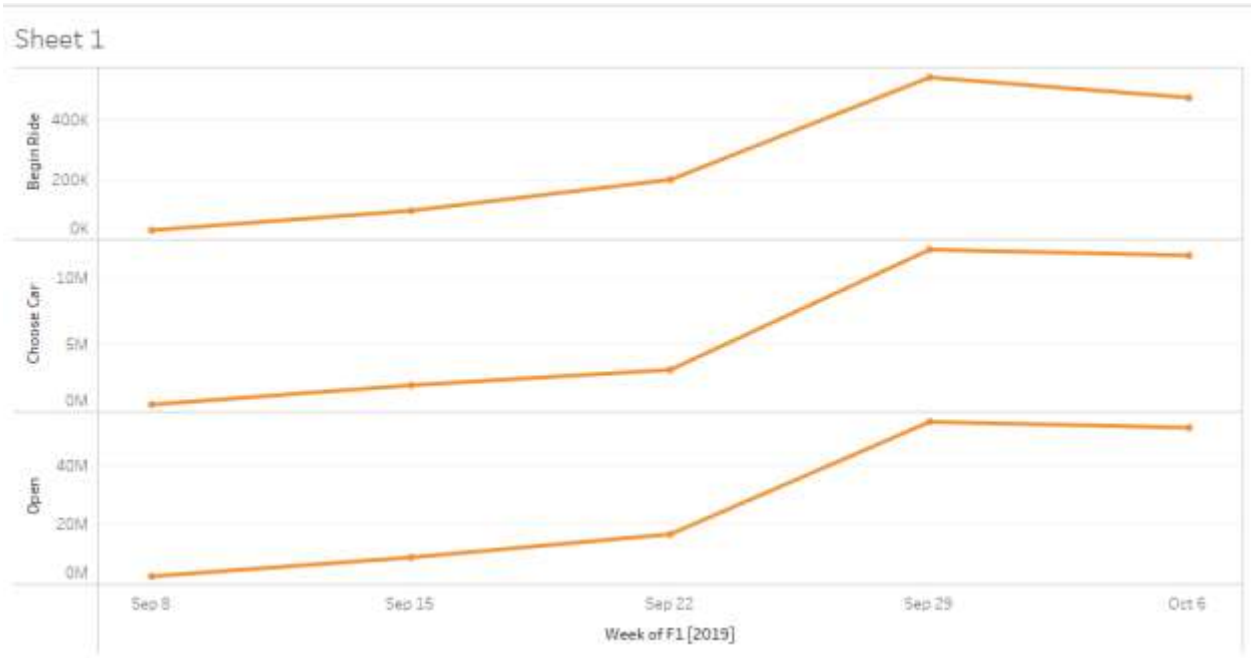
Section 6: Business Insights

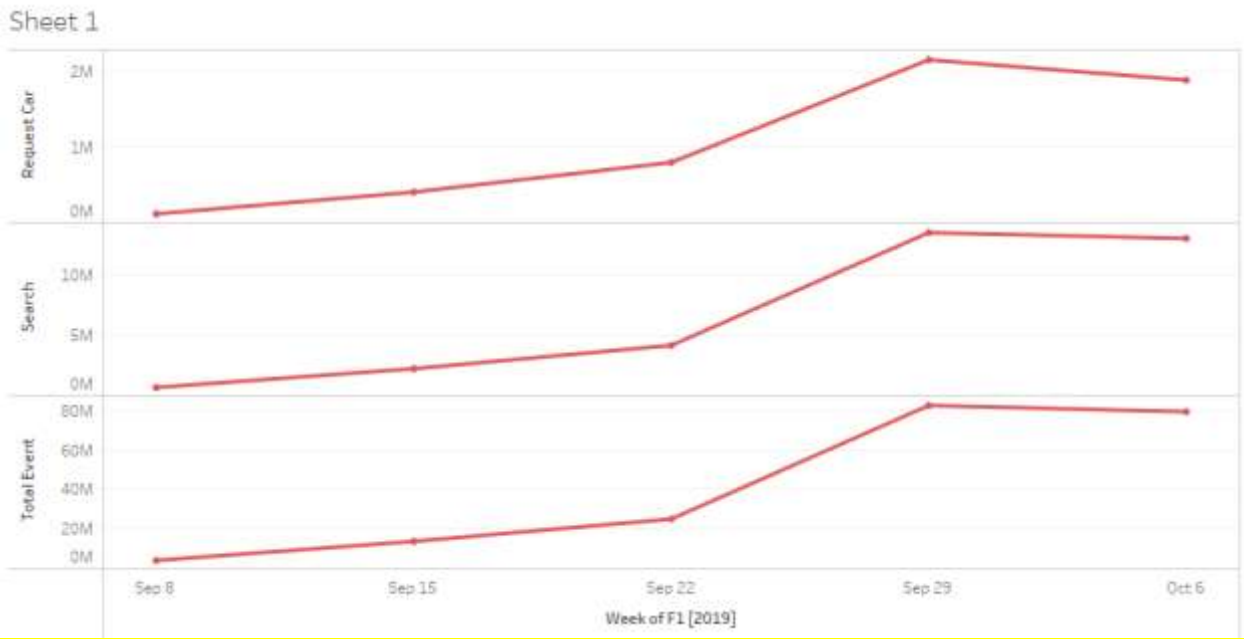
The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received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

What is the story the data is telling you about Flyber's data growth? If you created Visualizations, you can use them as well, but they are not required). Include any data and calculations that were made to help tell that story and quantify the data growth.

Data Growth for Last Month

Visualization:





We are seeing that the data is growing rapidly from Sep 22nd onwards, and after 29th Total event count stabilizes and even drops a little. The same behavior is observed in general for all types of events as shown above. To understand better we need more data spanning across few more months.

What is the fastest growing data and why?
Observing the graphs of all the events above there is no fastest growing data, all event types have the same trend. However It is interesting to observe that out of total "open" events only 10% are actually requesting a car.

All Event Type Data

Visualization:

What is the Data Story our data tells for each of the following:

- Graph Pattern
  Data rapidly increase during last few weeks of September and then stabilizes, the stabilization could be due to reduction in new customers and existing customers not using the application frequently.
- Good or Bad
  This is not enough data to predict if this is good pattern or not we need few more months of data.
- October Marketing Campaign
  Total events seem to have stabilized on October, with no further increase which means we have to rampup marketing effort to gain new leads.

- Marketing Campaign Impact: *Marketing Campaign conducted in September resulted in new customers and increased events*
- Importance of Relationship Between Marketing Campaigns and Data Generation : *Better the marketing campaign higher the number of Event activities.*

Section 7: Data Infrastructure Strategy

Thus far we have:

- identified data stakeholders and their data needs.

- Identified what data is currently being collected and what data needs to be collected.
- Identified data insights and growth trends.

Now, it's time to tie all the loose threads together and bring this process to its logical conclusion by suggesting which Data Warehouse (DWH) Flyber should invest in and why. Using data warehouse options below, suggest whether Flyber should choose an on-premise or Cloud data warehouse system and which specific data warehouse would best serve Flyber's data needs.

Data Warehouse Options:

Cloud:
- Amazon Redshift
- Google BigQuery
- Snowflake
- Microsoft Azure

On-Premise:
- Oracle Exadata
- Teradata, Vertica
- Apache
- Hadoop

You will address the following factors with a rationale as to why the DWH chosen is the best for Flyber:
- Cost
- Scalability
- In-house Expertise
- Latency/Connectivity
- Reliability

Cloud vs On-Premise
Provide an evidence based solution as to why Flyber would be best served by a Cloud or on-premise DWH. In this response, you don't need to specify *which* specific Cloud or on-premise DWH product you will choose, just if it will be Cloud or on-premise. Remember to address the factors above.

Flyber is still in very early stage working on MVP, our primary expenditure need to focus on gaining more customers, increasing number of mini copters to serve a larger customer base.

If we go with On-Prem solution will have a storage cost and computation cost which will grow exponentially as we have more data coming in. Building our own infrastructure might not be scalable and need heavy investments as our data grows.

. With growing customer base we need a more scalable infrastructure which is might turn out to be too expensive with On-prem services .Also hiring in-house expertise is out of question because our investment goals is not to hire more employees but to invest on more customer leads and services.

Cloud is scalable which means as our data grows our data needs will be met. We do not need to hire skilled expertise to run our infrastructure.

Reliability is an issue with cloud but we can install additional tools and applications to ensure secure data computation and result extractions.

Suggested DWH

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

Our best bet is using Cloud,especially go with Snowflake which is compatible with Google Cloud,AWS and Azure. Snowflake uses mapreduce to compute petabytes of data closer to where data is stored and sending results to the requester which minimizes latency.

Snowflake needs comparatively simple SQL queries to run massive computations and send the results to the requester so we do not need to hire an expert to deal with infrastructure.

Snowflake becomes more affordable as our data grows exponentially along with exponential computation needs.

On premise are affordable if we want to just store with limited computation needs. That's not the usecase we need to store exponentially growing data we will need lots of computation.

*Snowflake is **very reliable** and allows for auto-scaling on large queries meaning that you're only paying for the power you actually use. Storage support for both Structured and un structured data available.*

Snowflake also provides features to keep data secure and safe.

Image Appendix

Image 1: Log Growth

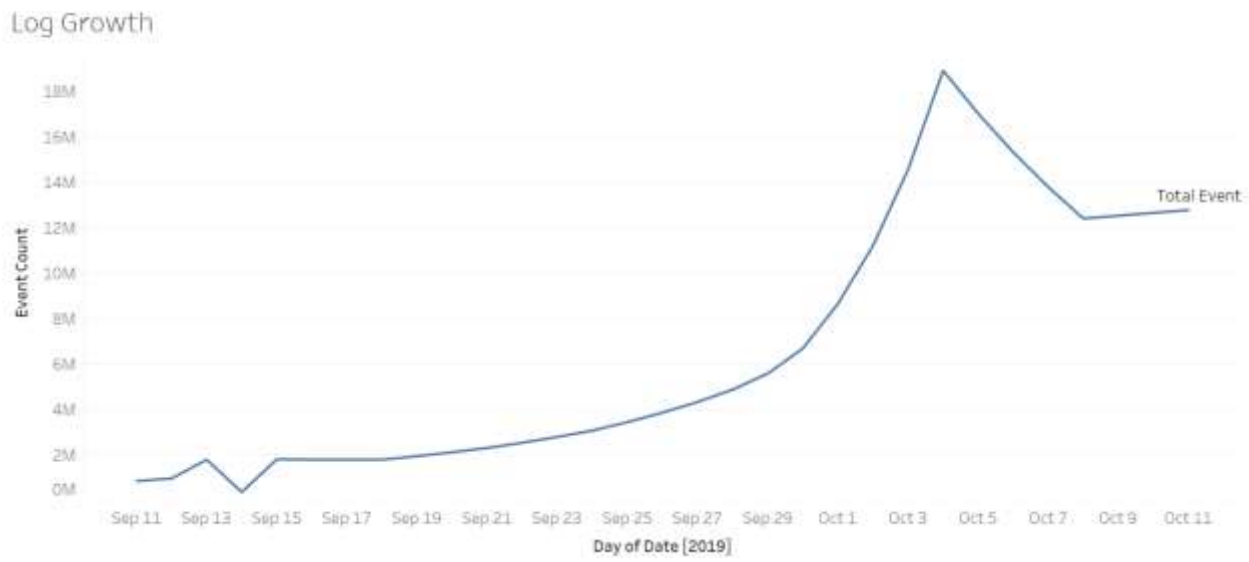


Image 2: Ride Growth

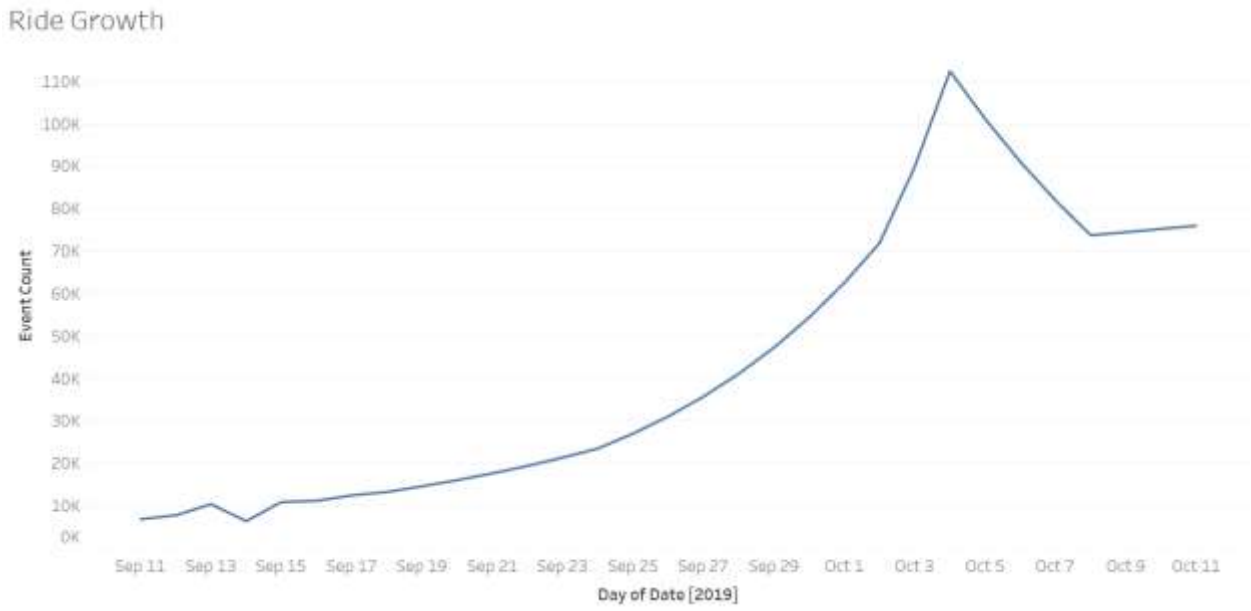


Image 3: Total Event Count

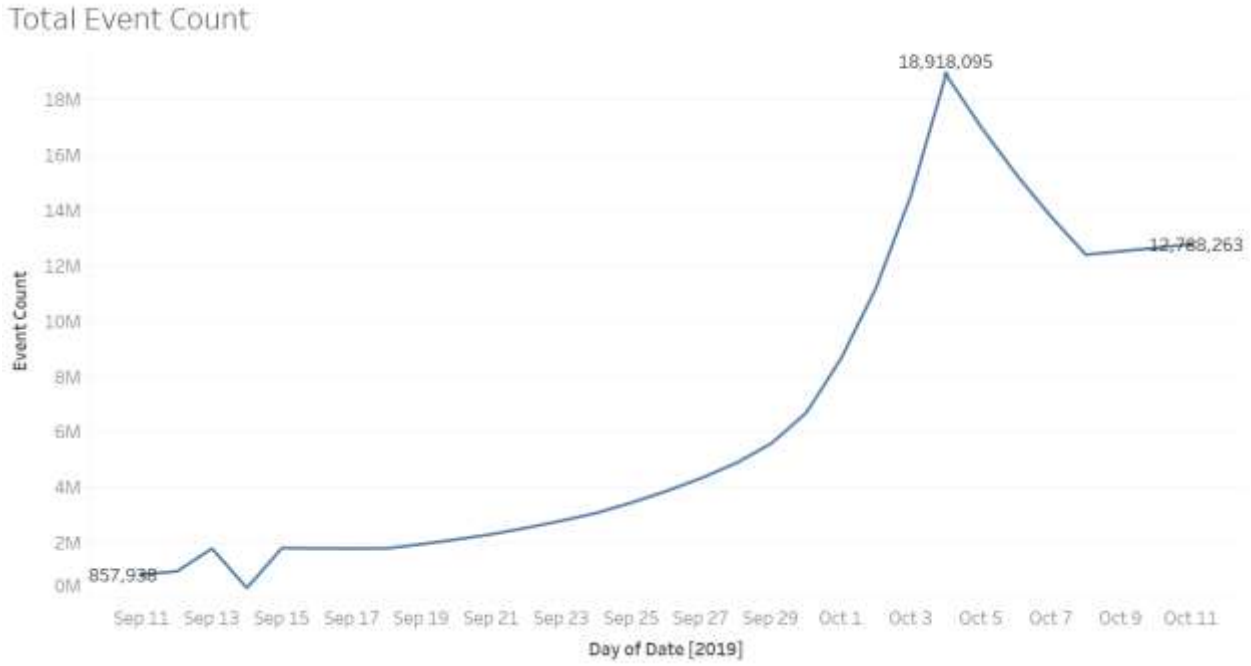


Image 4: All Events Log Scale

All Types of Events on a Logrithmic Scale.