# 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:**

| **Stakeholder** | **Why are they primary stakeholders?** | **Use-Case** |
|---|---|---|
| Finance & Accounting | Monitoring of all finance related issues. Strong focus on Margin and Profit, ensuring that the MVP is profitable. | Tracks the entire financial records of Flyber (Revenue, Margin Profit, Prices, Costs, ...) |
| Marketing | In the MPV phase it is important to acquire as many customers as possible. Bringing in new customers is another key factor for the success of our business. | There will have to be a great ability to keep up with all the communication and advertising media, as well as all the events generated in customer referrals. It is very important to evaluate and follow the evolution of the use of the platforms. |
| Product & Business Development | Based on data and customer feedback, the product team aims to develop new features that allow the evolution of our service, both in terms of technology and experience. | Through the data collected in the apps or sites, namely user feedback data, crossing this information with business data, i.e., operational data of the trips made by customers, we seek to |

| | | |
|---|---|---|
| | | understand what can be improved and added to the product. |
| Operations & Logistics | Keeping track of aircraft routes in real time is vital to the success of Flyber business. The safety of Flyber customers is one of the key factors in our business. | Flight routes and dispatch rules are certainly managed in real time, so there is a need to have access to real-time data, from routes, generated events, ... |
| Engineering | With the main objective of having a reliable, safe and functional product, the engineering teams constantly evaluate the behavior of the platforms, the vehicles, and the entire technological platform. | It is important to ensure the tracking of data and events collected on the web and mobile platforms, as well as, and not least, to track and monitor actual data from flyber's service. |
| Data Analytics & Data Science (Business Intelligence & ML/AI Teams) | The Data Analytics and Data Science teams, along with the Product team, seek to understand how the product is positioning itself in the market, what the trends are, and future forecasts, with the goal of taking the service and product to the next level. To do this, they need data that allows them to evaluate and predict trends. | The recorded travel data, customer behavior in the apps, frequency of use, structured and unstructured feedback, etc. allow these teams to assess market segmentation, flight forecasting, feedbacck classification into categories, topic creation, among others. The main objective is to have data to create algorithms in order to innovate. |
| IT Business | The integrity, consistency, viability and security of Flyber's systems is | Data related to all incidents of the platforms and services provided. |

| | | |
|---|---|---|
| | critical to its business success. The IT teams (SRE, Infrastructure, Security, ...) need data and tools to monitor these services. | For this there should also be tools that allow better monitoring of this information. |

# **Section 2:** Data Collection and Data Modelling

**To support our primary stakeholders's use-cases we need following data:**

| **Stakeholder** | **Use-Case** | **Data** | **Why is this the primary use-case?** |
|---|---|---|---|
| Finance & Accounting | Tracks the entire financial records of Flyber (Revenue, Margin Profit, Prices, Costs, …) | Entity Data: Price, Costs, Revenue, Profit, Margins, Tax, Region, Driver, Rides | For the business it is fundamental to have its financial activities properly monitored, especially when it is a startup that is just beginning on the market.<br>➔ Revenue<br>➔ ARPU<br>➔ … |
| Marketing | There will have to be a great ability to keep up with all the communication and advertising media, as well as all the events generated in customer referrals. It is very important to evaluate and follow the evolution of the use of the platforms. | Entity Data: Costumer Code, Costumer Name, Email, Phone, Registration, Address, User Rides History, Rating History, Last Login, Last Ride, Price History, Packages<br><br>Event Data: Platform, Channel, Refer History, Download Date | In the MPV phase it is important to acquire as many customers as possible. Bringing in new customers is another key factor for the success of our business.<br>➔ New costumers<br>➔ Conversion Rate<br>➔ DAU<br>➔ MAU<br>➔ ADAU<br>➔ WAU<br>➔ … |
| Product & Business Development | Through the data collected in the apps or sites, namely user | Entity Data: Costumer Code, Costumer Name, | Based on data and customer feedback, the product team aims to develop new features that allow the evolution of our |

| | | | |
|---|---|---|---|
| | feedback data, crossing this information with business data, i.e., operational data of the trips made by customers, we seek to understand what can be improved and added to the product. | Email, Phone, Registration, Address, User Rides History, Rating History, Last Login, Last Ride, Price History, Packages<br><br>Event Data: ID, Costumer, Feedback, NPS, CSAT, Surveys Answers | service, both in terms of technology and experience.<br><br>➔ NPS<br>➔ CSAT<br>➔ CES<br>➔ ... |
| Operations & Logistics | Flight routes and dispatch rules are certainly managed in real time, so there is a need to have access to real-time data, from routes, generated events, ... | Entity Data: Search, Star Ride, End Ride, Choose Motor, Request Motor, Location Pick-up, Location Drop-off, Price<br><br>Event Data: ID, Costumer, Date, Hour, Event Type, Package, Geolocation, | Keeping track of aircraft routes in real time is vital to the success of Flyber business. The safety of our customers is one of the key factors in our business.<br><br>➔ SLA's<br>➔ GIS<br>➔ |
| Engineering | It is important to ensure the tracking of data and events collected on the web and mobile platforms, | Entity Data: Search, Star Ride, End Ride, Choose Motor, Request Motor, Location Pick-up, | It is essential to track and monitor the status of applications and services provided, to understand the problems, |

| | | | |
|---|---|---|---|
| | as well as, and not least, to track and monitor actual data from flyber's service. | Location Drop-off, Price, Ride<br><br>Event Data: ID, Costumer, Date, Hour, Event Type, Package, Geolocation | and what can be improved in terms of software and hardware.<br><br>➔ SLA's<br>➔ ... |
| Data Analytics & Data Science (Business Intelligence & ML/AI Teams) | The recorded travel data, customer behavior in the apps, frequency of use, structured and unstructured feedback, etc. allow these teams to assess market segmentation, flight forecasting, feedbacck classification into categories, topic creation, among others. The main objective is to have data to create algorithms in order to innovate. | Entity Data: Search, Star Ride, End Ride, Choose Motor, Request Motor, Location Pick-up, Location Drop-off, Price, Ride, KPI,<br><br>Event Data: ID, Costumer, Date, Hour, Event Type, Package, Geolocation, Feedback, Incidents | Todos os próximos dias de negócios dependem fortemente da tecnologia &dados. A análise de um enorme repositório de dados necessita de um alto nível de poder computacional e de conjuntos de competências.<br><br>➔ NLP<br>➔ Predictions<br>➔ Forecasting<br>➔ ... |
| IT Business | Data related to all incidents of the platforms and services provided. For this | Event Data: ID, Costumer, Date, Hour, Event Type, Package, | The integrity, consistency, viability and security of Flyber's systems is critical to its business success. The IT teams (SRE, |

| | | | |
|---|---|---|---|
| | there should also be tools that allow better monitoring of this information. | Geolocation, Incidents | Infrastructure, Security, ...) need data and tools to monitor these services.<br>➔ SLA's<br>➔ ... |

**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.

**Table 1 - Customers**

| *Costumer_id (PK)* | *First Name* | *Last Name* | *Email* | *Mobile* | *App Version* | *Last Login Date* | *Last Ride Date* | *Adress* | *Resgistration_id (FK)* |
|---|---|---|---|---|---|---|---|---|---|

This table stores all customers registered with our service who use the app or the website. The primary key is the Costumer_id (PK) which is the unique identifier of the customer. The registration_id (Fk) is the foreign key to the Registrations table which stores the user's registration history.

**Table 2 - Registrations**

| *Registration_id (PK)* | *Timestamp* | *Geo_Long* | *Geo_Lat* | *Version* | *...* |
|---|---|---|---|---|---|

This table stores all applications downloads and web site registrations. The primary key is the registration_id (PK) is the primary key to the Registrations table which stores the user's registration history. There is no foreign key in this table

**Table 3 - Motors (Flyber Cabs)**

| *Motor_id (PK)* | *Registered_Date* | *Description* | *Manufacterd_Date* | *Last_Revision_Date* | *Software_Version* | *Status* |
|---|---|---|---|---|---|---|

This table stores all the information related to the vehicle that Flyber has in its fleet. It uses the motor_id (PK) as its primary key, which allows it to identify a unique vehicle. This table has no foreign keys.

**Table 4 - Drivers**

| *Driver_id (PK)* | *Registered_Date* | *First Name* | *Star_Date* | *Last Name* | *Medical_Status* | *Licence* | *Is_Active* |
|---|---|---|---|---|---|---|---|

This table stores all the information related to the drivers that Flyber has in contract. It uses the driver_id (PK) as its primary key, which allows it to identify a unique vehicle. This table has no foreign keys.

**Table 5 - Rides**

| *Ride_id (PK)* | *Costumer_id (FK)* | *Driver_id (FK)* | *Motor_id (FK)* | *Price* | *Pick_up_datetime* | *Drop_off_datetime* | *Pickup_Location* | *Drop_off_location* | *Status* |
|---|---|---|---|---|---|---|---|---|---|

This table stores all the information about the consumer's rides and all the information derived from this ride. As primary key it has the unique identifier of a ride (Ride_id) and as foreign key, it has Costumer_id (FK), Driver_id (FK) and Motor_id (FK) that allows to associate the information of customers, drivers and vehicles used in this trip.

**Other tables related:**

***Customer Type*** - allows Flyber to see if the consumer has bought a travel package, staying in plafond. Making him the VIP consumer.

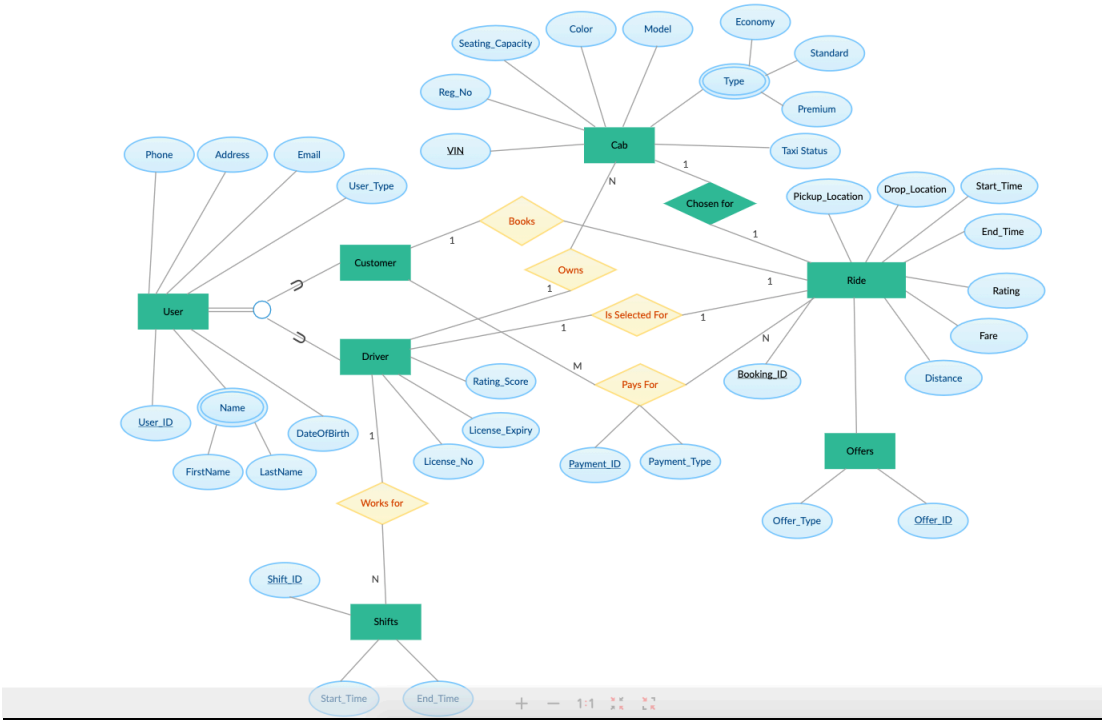
***Packages*** - relates to the consumer type and to the customers, perceiving the customer type and the customer history.

***Payment Method*** - A ride may have a direct payment method or via the package that the customer has purchased.

**Entity Relationship (ER) Diagram:**

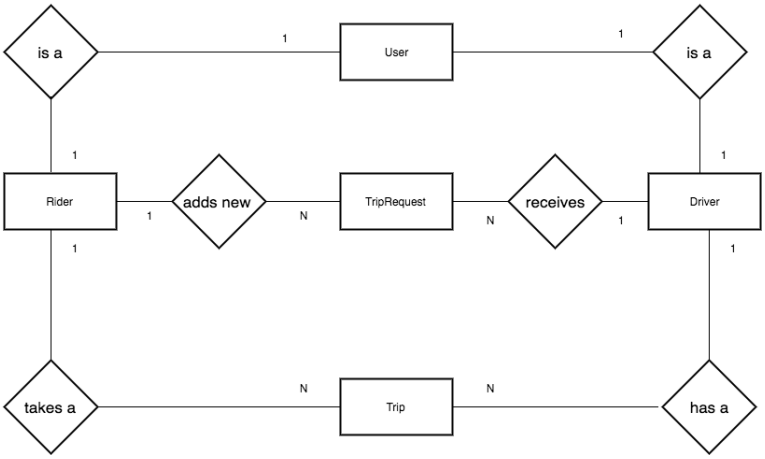
ER is a visual representation of different data using conventions that describe how these data are related to each other.

Example 1:




Taxi Service System ( Entity Relationship Diagram) *[7]*

Example 2:




ER diagram used in Taxi App Backend Architecture *[8]*

# **Section 3:** Extraction and Transformation

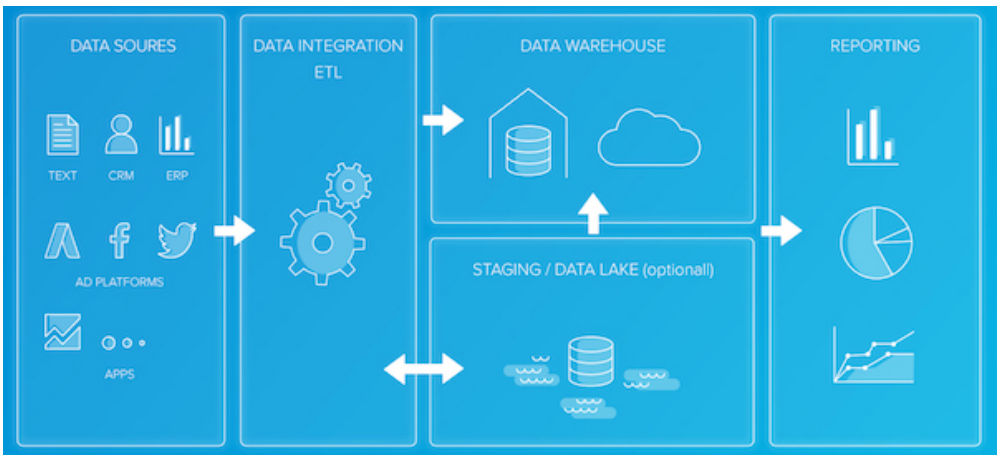
Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are currently collecting in the pipelines and 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.




ETL pipeline [11]

**Steps for extraction:**

*1. Data Collection:*

*1.1. Raw data is available locally in csv files.*

*1.2 In a first step the data is analyzed in its original format in order to know what information is available.*

*2. Data Verification:*

*2.1 Realize which data need transformations and corrections for further analysis.*

*2.2 Since the original data has problems, it will be necessary to map columns, filter valid columns, convert units of measure such as numerical and temporal.*

*3. Data Transformation:*

*3.1. Activity that allows to transform the data according to the above checks, from transforming the data type, filtering, converting and removing duplicates.*

*4. Data Loading:*

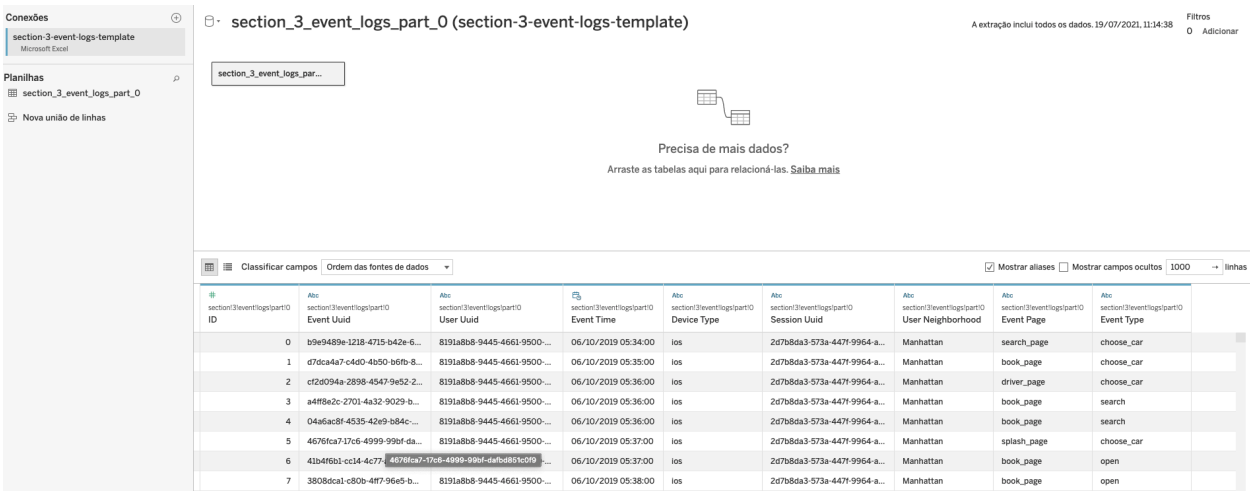
*4.1 In this process I chose to load the data into the final database.*

*5. Data Visualization:*

*5.1 The data becomes available to be analyzed using visualization tools.*

**Transformation-2**

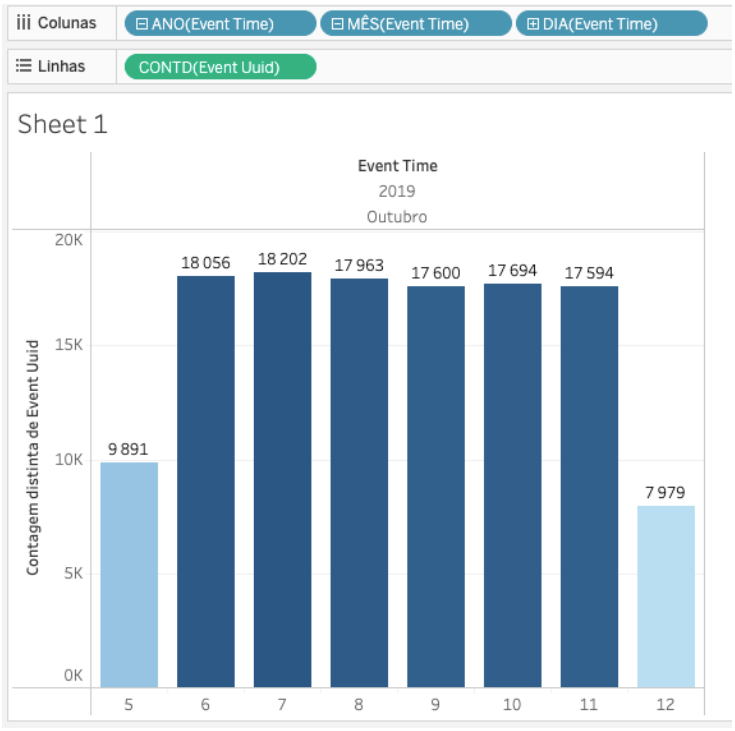
Analyze the data from part 1 to answer the following questions:
The analysis can be obtained from the tableau that has been created for this purpose. Example of a public visualisation here.




Data source tableau

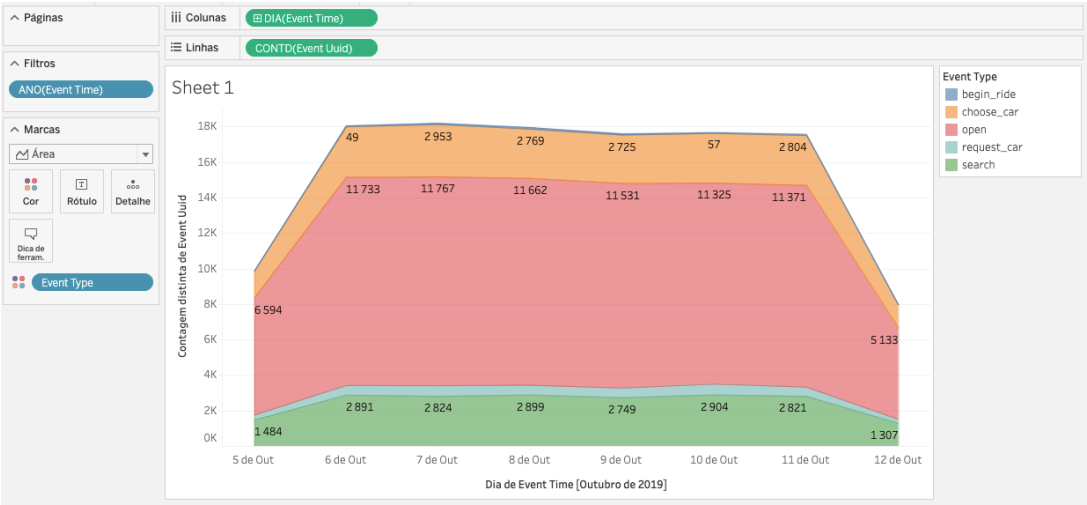
1. How many events are being recorded per day?

| 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 | 17594 |



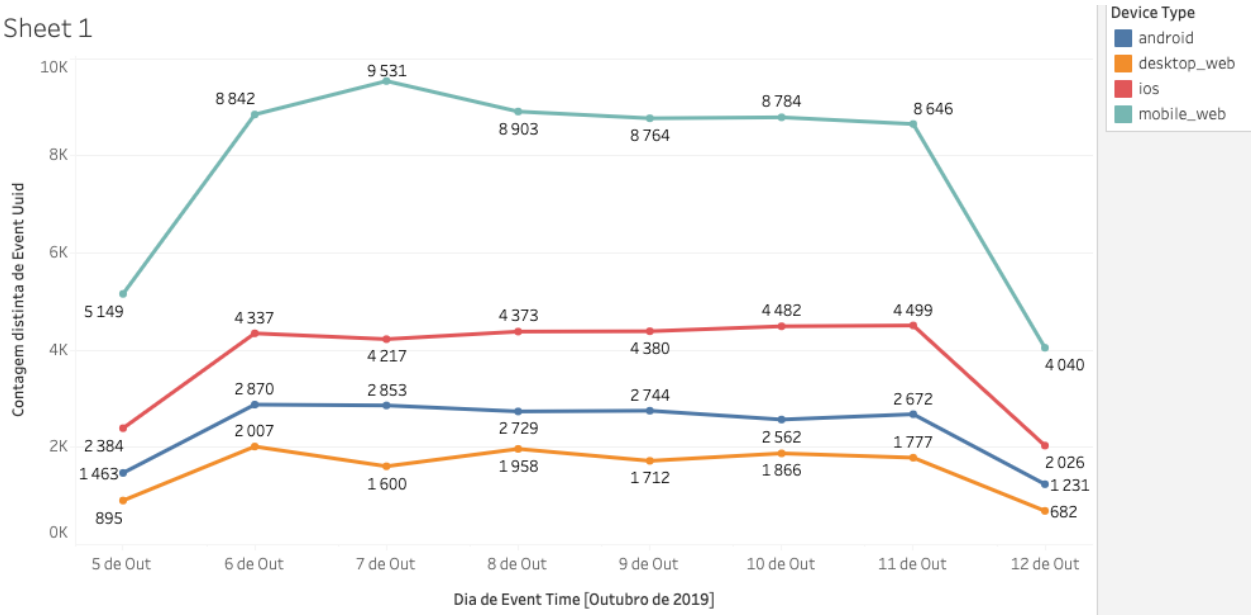
2. How many events of each event type per day?

| 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 | 2892 | 2824 | 2899 | 2749 | 2904 | 2821 |
| Open | 6594 | 11733 | 11767 | 11662 | 11531 | 11525 | 11371 |
| Begin Ride | 38 | 49 | 62 | 86 | 57 | 57 | 78 |
| Request Car | 277 | 540 | 596 | 547 | 538 | 607 | 521 |



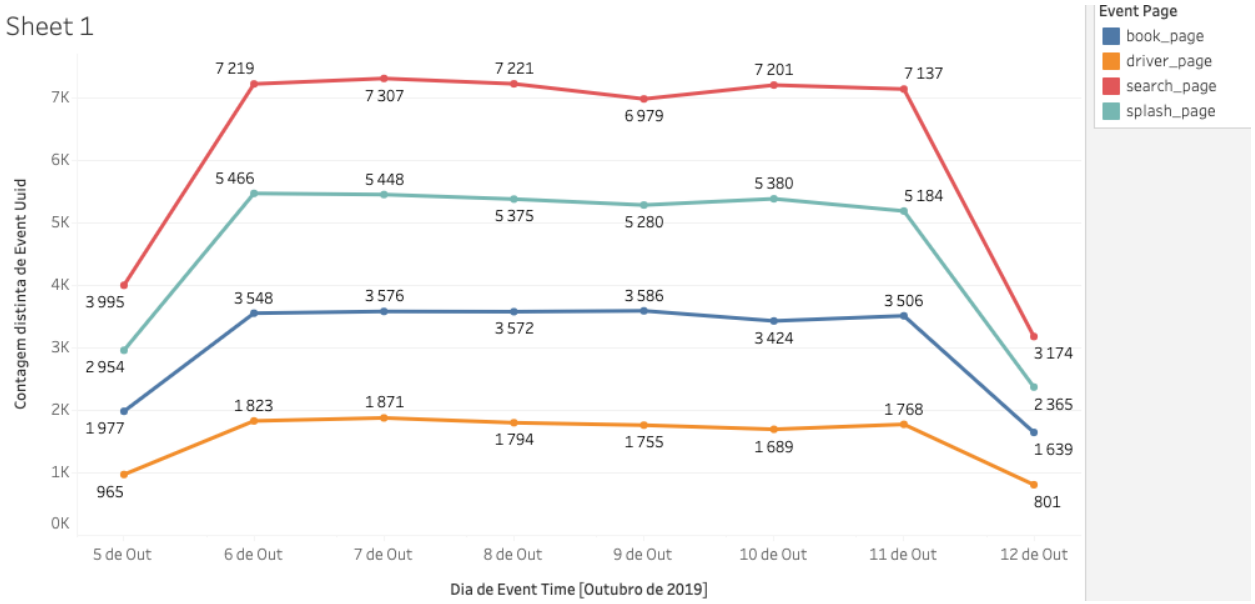

3. How many events per device type per day?

| Date | 10/5/2019 | 10/6/2019 | 10/7/2019 | 10/8/2019 | 10/9/2019 | 10/10/2019 | 10/11/2019 |
|---|---|---|---|---|---|---|---|
| ios | 2384 | 4337 | 4217 | 4380 | 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 |



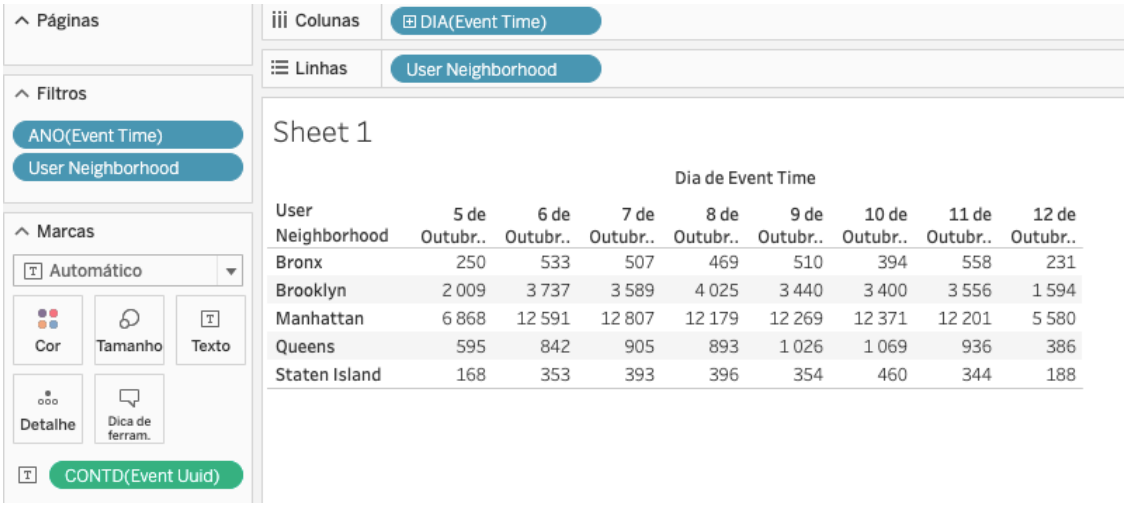

4. How many events per page type per day?

| 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 | 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.

*The ETL process defined for the data processing and exploration was adequate given the data set I had in hand. However, it is clear that if the data set were a more complex data set, larger and broken into several files partitioned by day, for example, the complexity of the ETL would be much greater. So, I admit that the advantage of this manual processing is that it allows non-technical people to explore and work with less complex source data. Self-service tools are able to load the data, transform it and at the same time analyze it in an easy and intuitive way.*

*As I mentioned if it was a complex, dynamic and partitioned dataset per day, then we would have to implement a more robust ETL process to run daily, using data engineering techniques like pipelines, partitioning, data lakes, data warehouse, etc.*

# **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

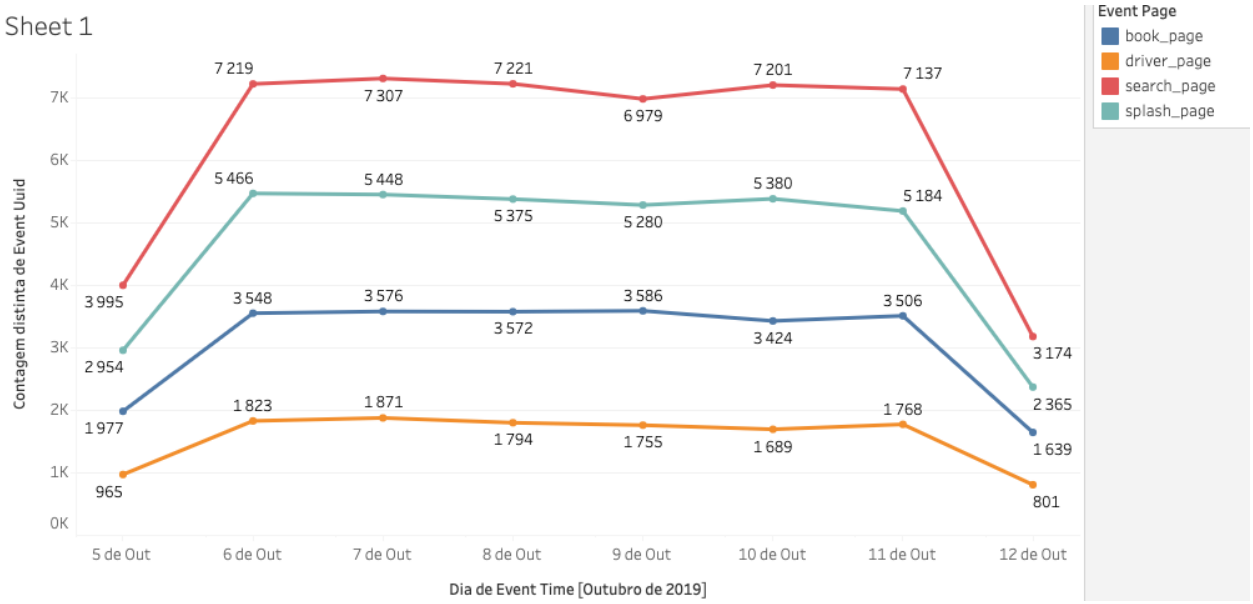
The data provided represents the events collected in a specific time period (between October 05, 2019e October 12, 2019). These events (5 distinct events, spread over 4 channels) concern the activity of the different flyber customers.

It is assumed that there are data for different time periods, namely several months, and therefore it is considered that it is data that allows to compare different events.

In order to automate the generation of valid insighst, we would have to implement an ETL process, with a daily dimension, to extract, transform and load the universe of data derived from the registered events.

Here we have Event Log Data.

Question: How many events per page type per day?



Ignoring the 5th and 12th of October, which appear to be the beginning and end of the collection of events, there is a sharp rise and fall on the 5th and 12th respectively. It appears that the events were recorded from mid-afternoon >10AM on the 5th to mid-afternoon on the 12th (<10AM).

| | Event Time | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2019 | | | | | | Total geral |
| | T4 | | | | | | |
| | Outubro | | | | | | |
| Event .. | 6 | 7 | 8 | 9 | 10 | 11 | |
| search_page | 7 212 | 7 307 | 7 221 | 6 979 | 7 201 | 7 137 | **43 057** |
| splash_page | 5 459 | 5 448 | 5 376 | 5 280 | 5 380 | 5 184 | **32 127** |
| book_page | 3 548 | 3 576 | 3 572 | 3 586 | 3 424 | 3 506 | **21 212** |
| driver_page | 1 820 | 1 871 | 1 794 | 1 755 | 1 689 | 1 768 | **10 697** |
| Total geral | **18 039** | **18 202** | **17 963** | **17 600** | **17 694** | **17 595** | **107 093** |

- ➔ For search_page event type, we have 43.057.
- ➔ For splash_page event type, we have 32.127.
- ➔ For book_page event type, we have 21.212.
- ➔ For driver_page event type, we have 10.697.

1. How much is the customer data increasing?

   A: Based on the analysis above, I can see that the data remains line over the days. However, it is clear that customer many more searches on the platforma compared to bookings (roughly more than 50%).

2. How much is the transactional data increasing?

   A: From the analysis the data shows a steady growth over time.

3. How much is the event log data increasing?

   A: From the analysis the data shows a steady growth over time.

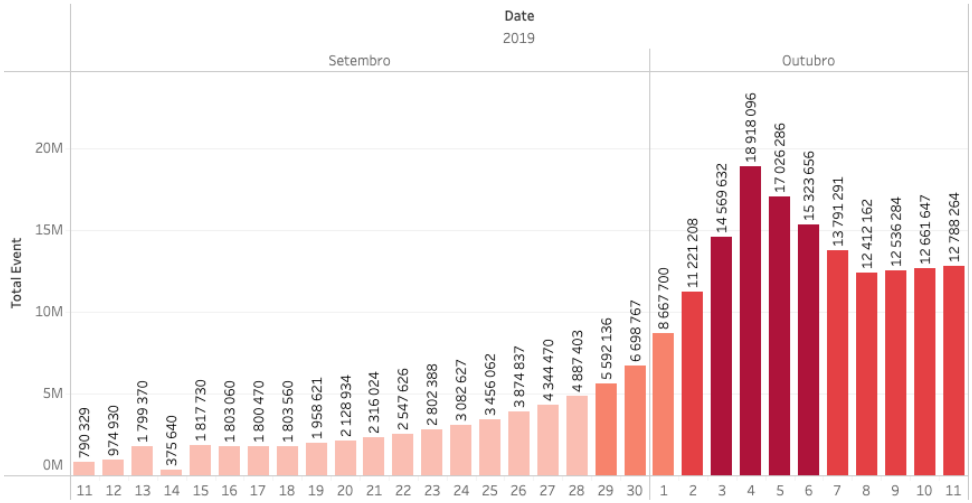
# **Section 5:** Loading and Visualization On Your Own

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:**

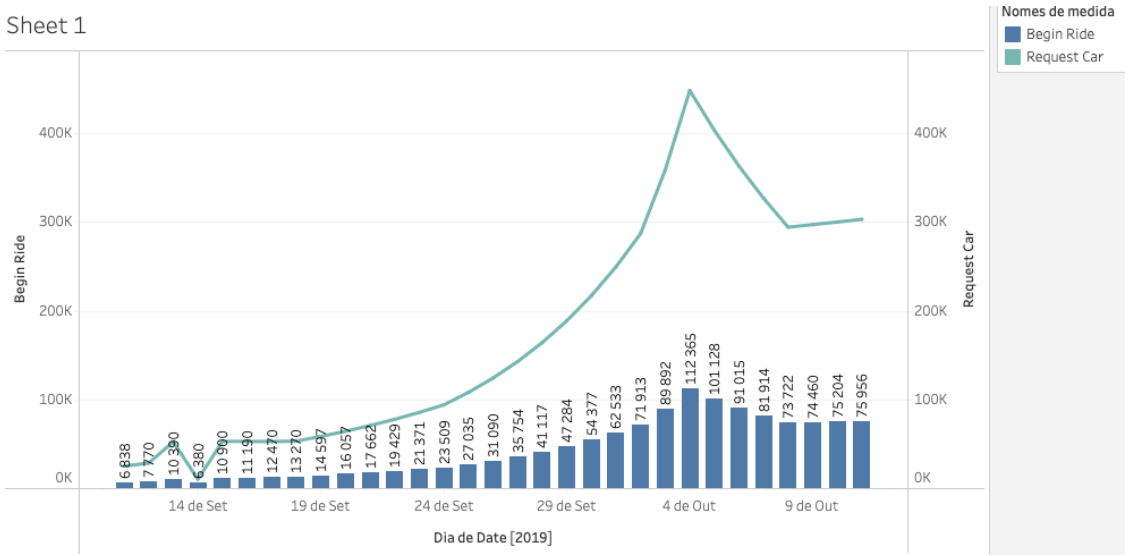


**Data Story**: *That the number of event over time is increasing, but a sharp increase in October. The 4th is the day with the greatest volume of events. For a more detailed analysis it will be necessary to present the events in a differentiated way.*

This graph was created using the following steps:

1. The raw data (excel) set was loaded into Tableau
2. I selected the entire available dataset
3. Considered the metrics relative to the events that were present in the data set
4. A simple form of visualization, based on a bar graph, was chosen.

**Visualization 2:**




**Data Story:** *Shows a comparative analysis between the scheduling and ordering of the car and the start of the trip, and there seems to be a marked gap between the number of events recorded on each day.*

This graph was created using the following steps:

1. The raw data (excel) set was loaded into Tableau
2. I selected the entire available dataset
3. Considered the metrics relative to the events that were present in the data set
4. A simple form of visualization, based on a bar graph, was chosen.

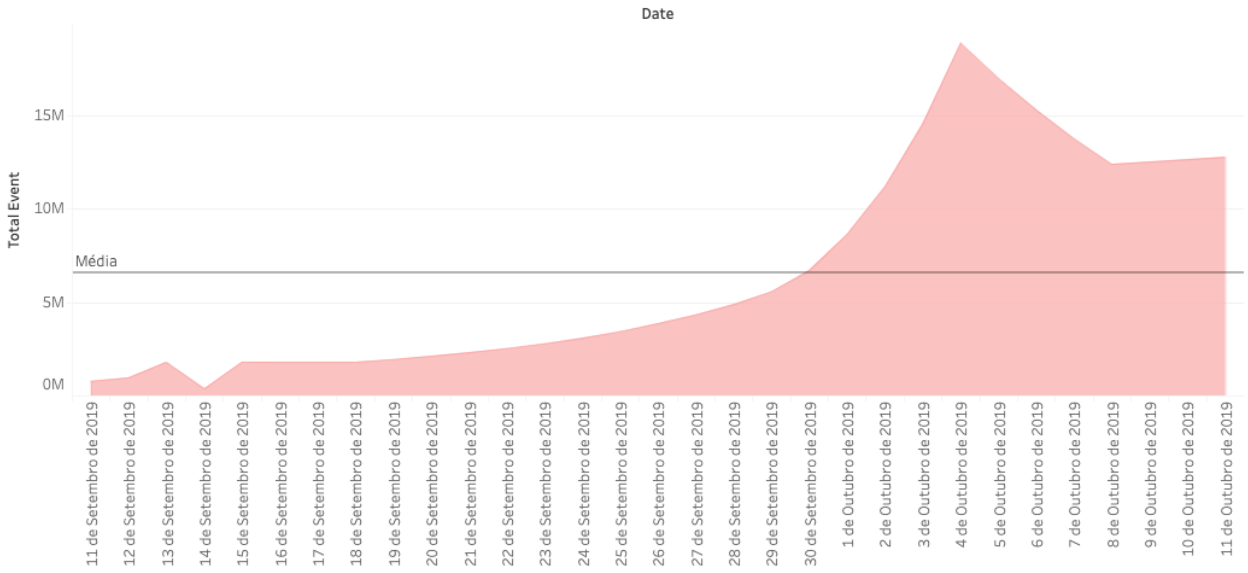
# **Section 6:** Business Insights

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:

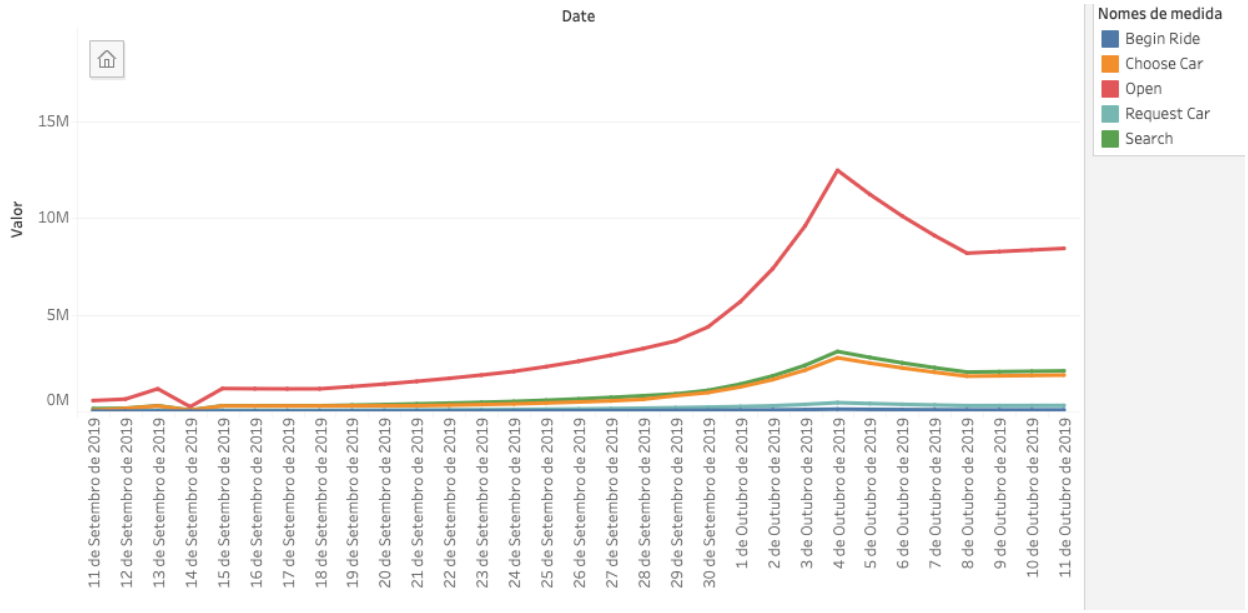
A: Used Log growth (Total Event)

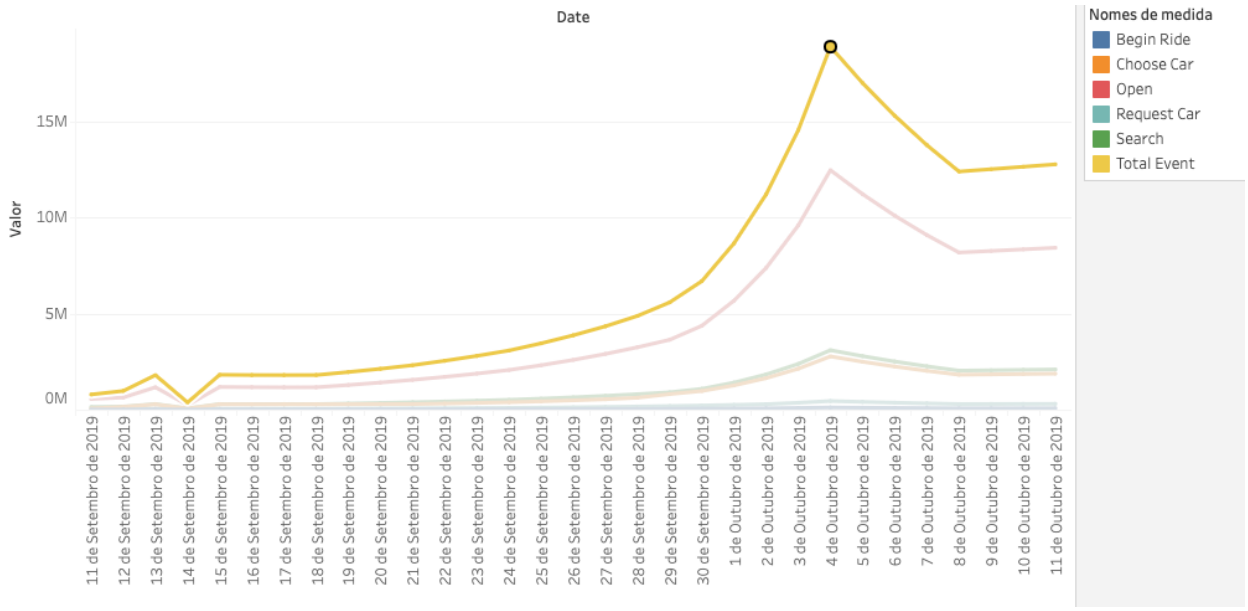
What is the fastest growing data and why?

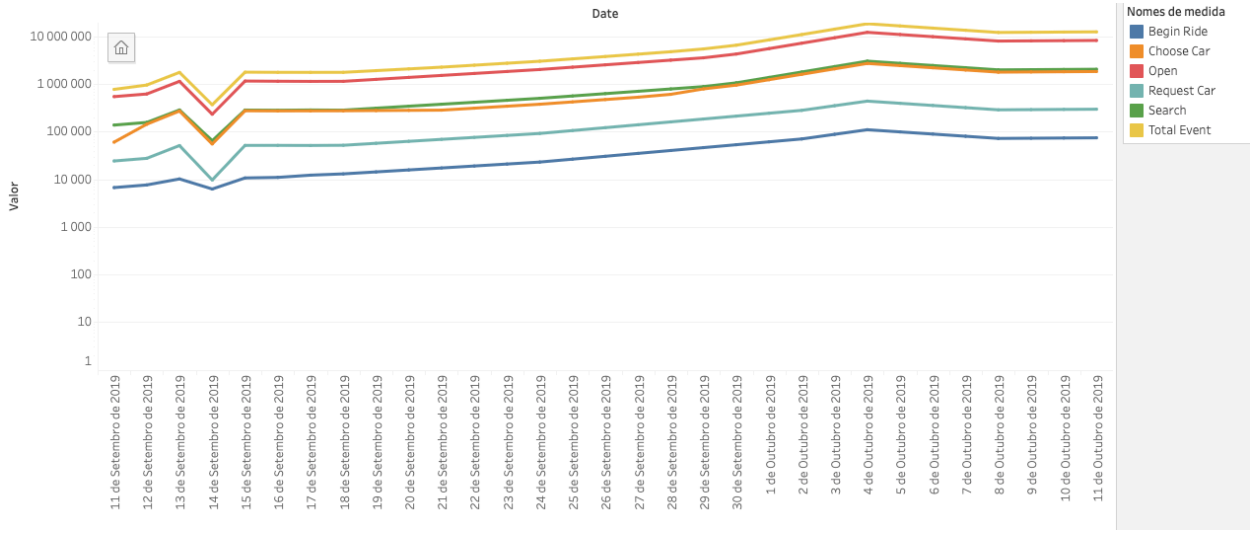
A: Total Events measure summarize all events. With this analysis we can see that the month of October was much busier than the month of September, because the number of events grew exponentially, making the first week of October a very good one.

**All Event Type Data**

Visualization:






In a nutshell, and as mentioned before it is possible to see from the analysis that the pattern of the event graphs shows a steady growth during September and then in October shows an exponential growth, reaching very high levels, although it then slows down a bit from October 8th onwards. However, it is possible to conclude that something happened in October and it seems that this increase is due to the impact of marketing campaigns to promote the brand and attract customers. It seems that in the first days of October, there was a strong focus, leading people to try to use the platform, even if they didn't make a trip.

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

- **Graph Pattern** - the pattern of the event graphs shows a steady growth during September and then in October shows an exponential growth, reaching very high levels, although it then slows down a bit from October 8th onwards. Because it is a short period of time, it is not possible to see a trend in the data, beyond the pattern already explained.

- **Good or Bad -** The impact of marketing campaigns is vital to event growth in October, however there is no growth in the number of conversions from visits to travel.

- **October Marketing Campaign -** it was successful, in the sense that I managed to bring more users to the application.

- **Marketing Campaign Impact -** October shows exponential growth, reaching very high levels of use in the application.

- **Importance of Relationship Between Marketing Campaigns and Data Generation -** Marketing campaigns greatly influence the generation of information, events and data that are usually stored to evaluate the impact of such measures. This means that with more data it will always be possible to have more insight into customers.

# **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.

**First Analysis:**

In order to respond to a solution of this type that should be adopted by Flyber, it is important to carry out a small research of the state of the art in reputable companies, whether in a similar or even different area of activity. We can evaluate the cases of Uber, Cabify, Lyft, among others [13] [14] [15] [16].

Upon analysis it turns out that modern architectures are in the cloud, and trend is increasing. The companies show that despite presenting On-Premise solutions the core is in the Cloud. There are several technologies that companies use to meet the needs of services and products provided.
It is believed that Flyber has the potential to be in the Cloud, because it is a cutting-edge technology. Even though in a first phase it could bet on-premise and then, as the business evolves, migrate to the Cloud. But a migration of this type may bring increased costs in manpower, time and others. It is important to understand the volume of data that will be generated, because storage is also very expensive, which locally can be a cost too high. Ja system maintenance, creation of environments, scalability, among other aspects should be considered in this decision. In the Cloud the management of environments, infrastructure, platforms, services, accesses, are scalable and easily managed.

The focus of DWH should be evaluated by the number of technologies that are needed, as well as the possibility of storage capacity, because as I said, data storage is very expensive. So, making sure that the product is in the Cloud, Flyber choose to include DWH (underlying architecture) also in the Cloud.

In a company like Flyber it will be important to define a set of requirements for the data platform:

- ➔ Flyber is a business that takes place by the second, for this reason the data must be made available as quickly as possible, in real time (or near real time).
- ➔ The scalability of the platform must be ensured, because resources must be expandable, storage must be increased, performance increase must be easy, etc.
- ➔ The correct monitoring of the processes that must run in real time, follow all the issues of data availability and the performance and security of the infrastructure.
- ➔ It should not put in question, at any time, the services provided, that is, it should not affect the 100% operation of the app, sites, etc...
- ➔ Ensure that the specialized teams have the ability to improve and correct problems that are identified in the platform
- ➔ Must be able to collect data from Experimentation, Events and real data from the platforms.
- ➔ Must be able to store and process geospatial data.
- ➔ Must be available 24/7, with efficiency and effectiveness assured.
- ➔ Finance and related teams must be able to monitor platform costs.

- ➔ Others ...

Flyber must assume that structure and functionality are not the only factors that condition the choice of architecture, as the business may or may not benefit from the cloud approach or an on-premises solution. What is important is to understand how much the business will grow while taking into account cost savings and increasing efficiency.

Thus, Flyber should consider the following aspects [1] [2] [3] [4]:

- ➔ **Speed/Latency/Connectivity -** It is natural that on-premises data stores generally provide more speed than their cloud counterparts because they are not as susceptible to latency issues. Unlike cloud solutions that send queries to servers in other regions and have to wait for responses to come back, local on-premises servers minimise travel time so Flyber can get the answers that need faster. As the business is not spread across multiple geographical locations, then a cloud solution is not going to provide much advantage in this regard.

- ➔ **Scalability** - "Data warehouses grow very quickly over time, which means there will be a need to expand available storage on a regular basis. " If Flyber's business grows/changes, there will likely be a need to buy new software and hardware to accommodate this large-scale growth. This is often if Flyber has a warehouse on site. However if Flyber opts for a cloud warehouse it completely eliminates this need, making it much easier and faster to scale up (increase production or storage) or downsize.

- ➔ **Integration and In-house Expertise** - consequently a cloud data warehouse will make it easier to connect and integrate with other cloud services to help better manipulate Flyber's data. The freer Flyber's business is, the more freely its data can flow through cloud-based integrations. Otherwise, if Fyber's business has associated restrictions and these are a concern, then the on-premises approach is more appropriate as all security remains fully under the control of the Business IT team.

- ➔ **Reliability** - Both on-premises data warehouses and cloud warehouses can offer the highest uptime and reliability, but on-premises has an additional variable: the level of uptime and reliability is solely dependent on Flyber's teams and the equipment Flyber has in their possession. If there are problems with reliability then these will be Flyber's responsibility. If Flyber opts for a cloud warehouse, uptime and reliability are guaranteed through the service provider's SLA which is usually 99.9%. For example, Amazon promises a minimum uptime of 99.99% for its EC2 DWH service. Google promises a 99.9% monthly uptime percentage for Cloud Storage and BigQuery. Google, Amazon, and other cloud DWH providers will replicate their data across multiple clusters to ensure maximum reliability. The AWS

network services provide an abstraction layer to highly scalable and available networking components [12].

- ➔ **Costs** - From the research conducted a cloud data warehouse costs significantly less compared to on-premises options as it does not require hardware, human resources, servers and server rooms. Cloud-based data storage eliminates most of the upfront costs. In addition, Flyber only pay for the resources use, which improves operational efficiency.

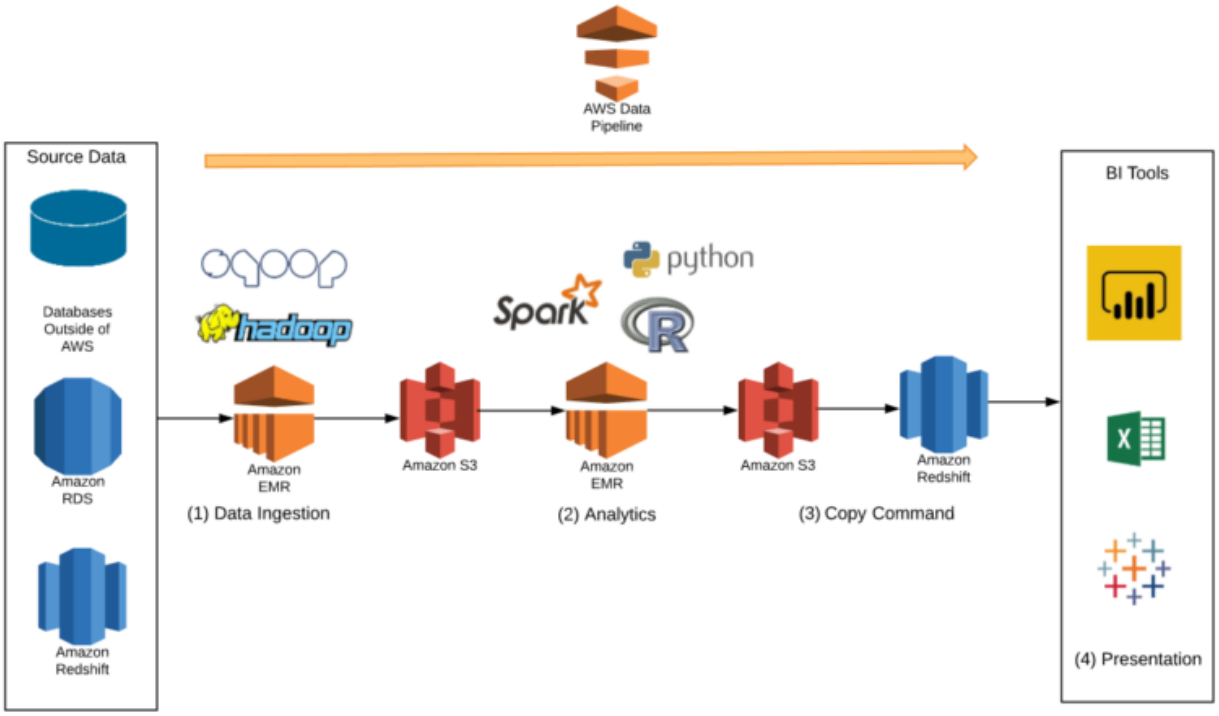
- ➔ **Security** - Cloud solutions are more secure than on-premises solutions in most use cases. On-premises data stores are the most secure option when supported by a strict data access policy (GDPR), whereas cloud storage offers more flexible security that keeps data safe.

It is a challenge, for startups like Flyber that are just starting out and the financial resources are not many to select such an architecture.
But in the case of Flyber, it is considered that despite some infrastructure challenges, the best option is undoubtedly to opt for the Cloud approach.
Assuming that the core business infrastructure is based on the Cloud, and assuming also that there is in house expertise to develop, even over the cloud, highly sophisticated services**, the best option is undoubtedly to assume right away the cloud approach for the data warehouse.**

In terms of stability, costs, integration, built-in ecosystem, security, availability and time to market, these are very significant advantages compared to on-premise architectures.

The primary cost survey [3], as mentioned above, and considering the SLA levels and expertise of the engineering teams it is considered that the best Cloud service provider could be AWS, creating a close cooperation and collaboration partnership. Thus, we share a high level architecture of what could be Flyber's solution in a first in terms of DWH.

**Cloud-based architecture**, with **AWS service** provider resources (AWS S3 for Data lake and AWS Redshift for DWH). *This architecture is scalable and allows Flyber to have data in near real time:*




*AWS Data Pipeline for DWH [10]*

Once the architecture is operational and collecting data is on a daily basis, then with the capability and expertise of the engineering and data team, Flyber can move forward also with implementing an on-premises and in house architecture to leverage the existing architecture. Interaction mechanisms can be created with the sources, such as easy definition of pipelines, creation of dashboards, etc. so that the teams become autonomous in the creation of their data silos according to the informational needs of the team.

In terms of challenges, FLyber may encounter latency issues that can be easily overcome by the solution architecture and the security issues of sensitive information, personal and confidentiality can be taken into consideration.

Considering [2] Flyber may in the short term raise two issues related to security and data latency. So there is therefore an opening to consider a hybrid solution between on-premise and cloud.

In a hybrid strategy, cloud repositories are used for day-to-day storage, with specialised data storage on-premise: (1) Sensitive data that Flyber don't want to travel off-network, (2) Personally Identifiable Information (PII) and (3) Data related to low latency processes.

This approach can help address specific security, compliance or performance issues that may arise when using the Cloud, while offering the flexibility of the Cloud. Flyber also have more scalability options as can expand the on-premises system or change Cloud subscription depending on her needs.
However, this approach may have an upfront cost consideration as Flyber will need to install and configure DWH on-premise. This will require sorting and segmenting the data and ensuring that everything ends up in the right place.

# References


1. *On-premises vs. cloud data warehouses: a comparison - https://www.stitchdata.com/resources/compare-on-premises-and-cloud-data-warehouse/*

2. *On-Premise vs. Cloud Data Warehouse -https://www.xplenty.com/blog/cloud-vs-onpremise/*

3. *Deciding on a Data Warehouse: Cloud vs. On-Premises - https://www.alooma.com/blog/deciding-on-a-data-warehouse-cloud-vs-on-premise*

4. *Emerging Architectures for Modern Data Infrastructure - https://a16z.com/2020/10/15/the-emerging-architectures-for-modern-data-infrastructure/*
5. *The Economic Advantages of Google BigQuery versus Alternative Cloud-based EDW Solutions -* https://services.google.com/fh/files/blogs/esg_economic_validation_google_bigquery_vs_cloud-based-edws-september_2019.pdf
6. *React Native Taxi App - https://docs-market.nativebase.io/taxi-app-with-backend/*
7. *ER Creatly for taxis - https://creately.com/diagram/example/iw2ugrv5/Taxi%20Service%20System*
8. *Calling-all-cabs-a-database-model-for-a-taxi-service - https://vertabelo.com/blog/calling-all-cabs-a-database-model-for-a-taxi-service/*
9. *ER-diagram-for-cab-services - https://www.conceptdraw.com/examples/er-diagram-for-cab-services*
10. *https://www.mydatahack.com/wp-content/uploads/2018/01/AWS-Data-Science-Pipeline-With-Data-Pipeline-768x455.png*
11. ETL Pipeline - https://cdn.buttercms.com/ZEBc4uz5QT2qFGlXYv9p
12. https://docs.aws.amazon.com/whitepapers/latest/hybrid-connectivity/aws-hybrid-connectivity-services.html
13. https://www.appsrhino.com/lyft-tech-stack-uber
14. https://medium.com/cabify-product/the-cabify-engineering-stack-2020-edition-34edcaff5ad0
15. https://eng.uber.com/tech-stack-part-one-foundation/
16. https://stackshare.io/uber-technologies/uber
17. https://public.tableau.com/app/profile/s.rgio.da.costa/viz/FlyberData/Sheet1#1

# Image Appendix

Image 1: Log Growth

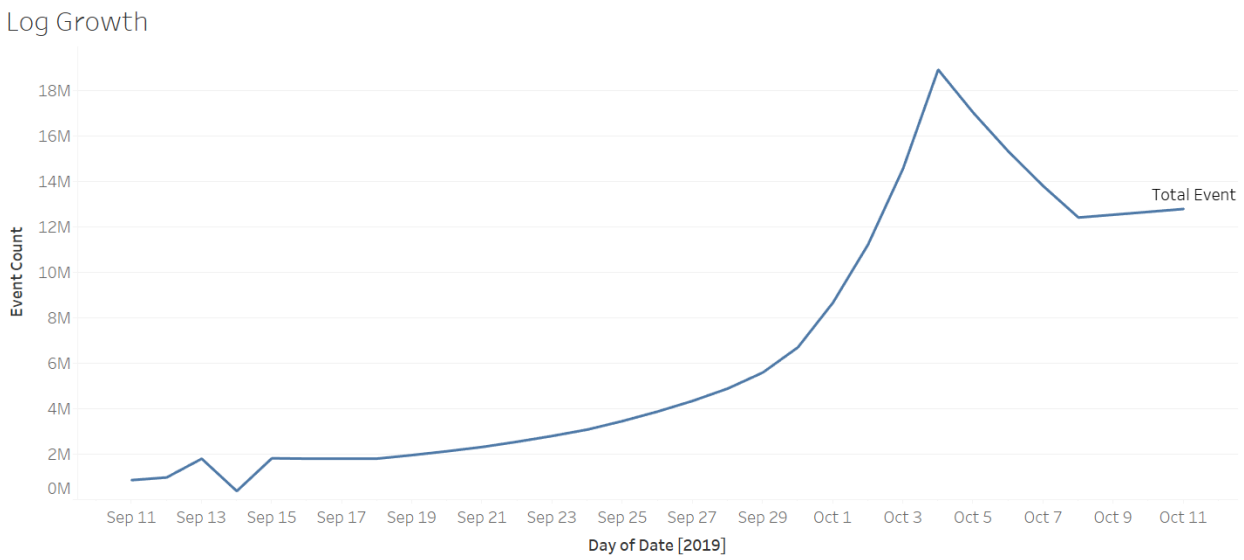



Image 2: Ride Growth

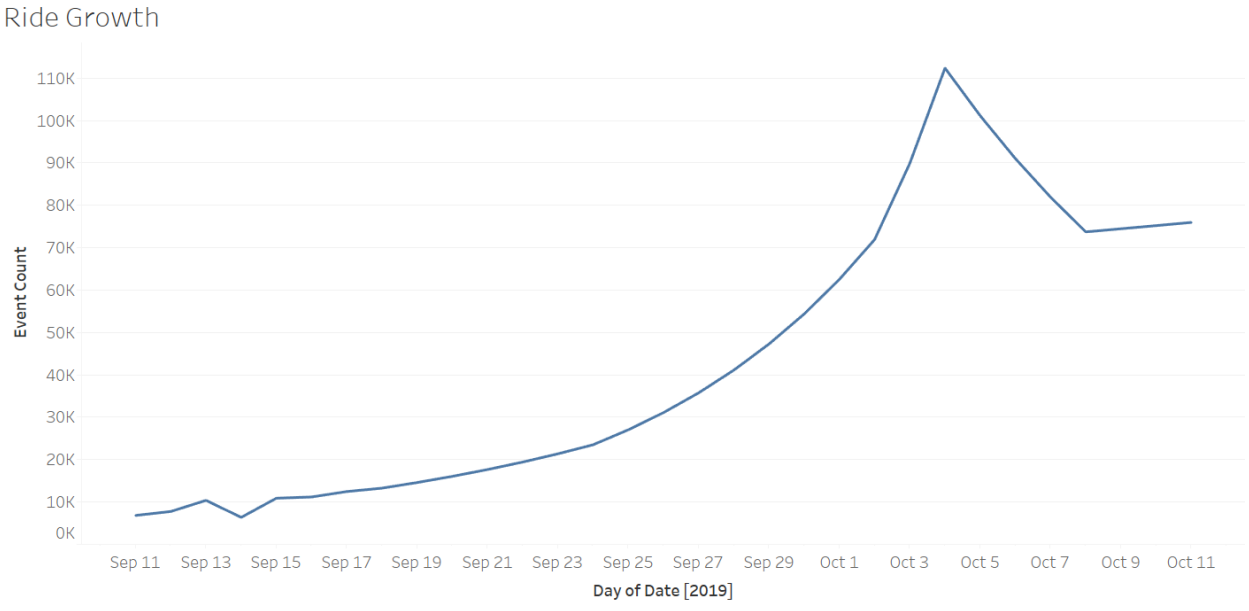



Image 3: Total Event Count

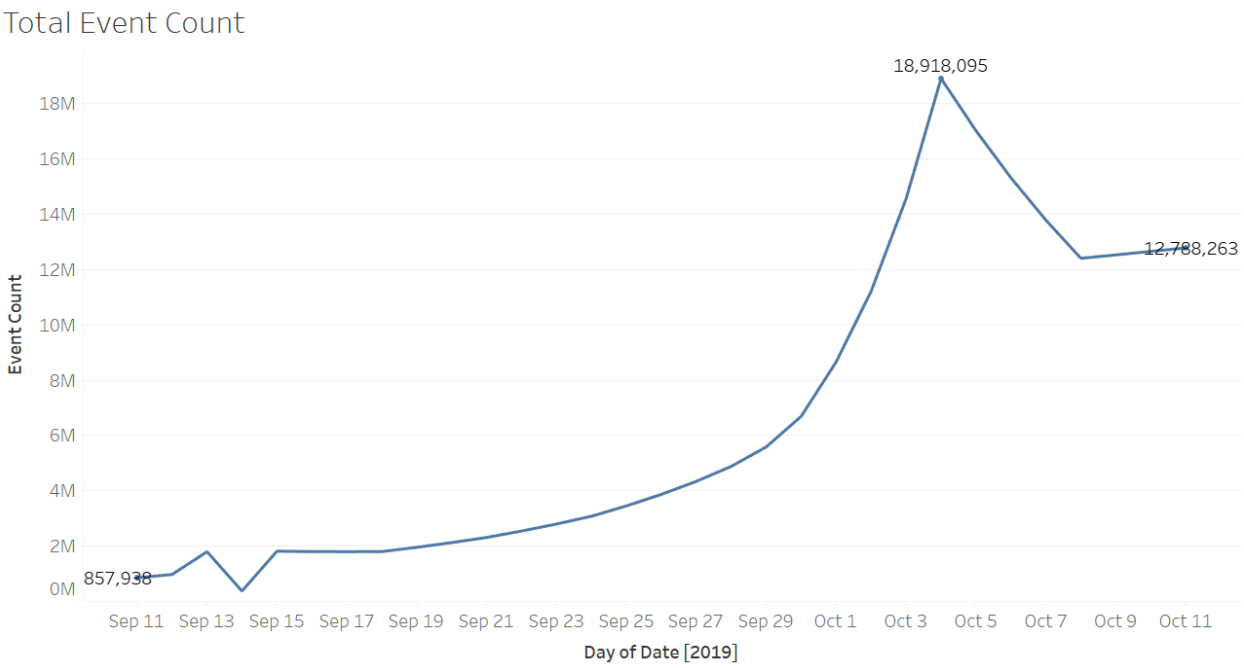



Image 4: All Events Log Scale