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** |
| Ops teams | They manage and optimize the details that keep its organization running. | Improve operational efficiency   * Identify congested areas * Reports around every Flyber stop and how many rides are being taken at every moment * BI tools for insights, and visualization tools for trends. |
| Marketing/commercial teams | The focus on getting new customers and retaining old | Understand the customers more effectively in order to get new customers and retaining all.  - Create targeted advertising.  - Identify customer profiles along with preferences.  - Help customers with virtual assistants.  - Create customized marketing content |
| Finance teams | They focus the company economy, predicting P&L growth | Public reporting of operating and financial data  - Monitor current P&L.  - Accounting reports generated with details such as daily, weekly, monthly expenses and earnings.  - Integrate these reports with their finance tool, that provides them insights and visualizations.  - ML engine to help them make future financial predictions. |
| Product management | They focus on improving the product | - Identifying customer pain points  - Analyze customer interaction data on site and app.  - BI and visualization tools to tell stories about how customers use the app and their needs |
| Engineering | They build and maintain the Flyber’s app | * Monitoring app availability |

## **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?** |
| Ops teams | Improve operational efficiency   * Identify congested areas * Reports around every Flyber stop and how many rides are being taken at every moment * BI tools for insights, and visualization tools for trends. | * Drivers information: Available Drivers, waiting times, average rating * Rides information: Distance, duration, pick up point, drop off point, time between user ordered the ride and driver arrival, reported incidents, time of pickup, time of drop off * Vehicles data: incidents, last checkup, details | Because we need to identify where are the best areas in the city to locate Flyber’s vehicles in order to maximize its use. |
| Marketing/ commercial teams | Understand the customers more effectively in order to get new customers and retaining all.  - Create targeted advertising.  - Identify customer profiles along with preferences.  - Help customers with virtual assistants.  - Create customized marketing content | * Customer information: age, location, payment method, device type * User log data: event type, event count, device type | Flyber is a new company and must focus on acquiring new customers and keep them using our service afterwards. |
| Finance teams | Public reporting of operating and financial data  - Monitor current P&L.  - Accounting reports generated with details such as daily, weekly, monthly expenses and earnings.  - Integrate these reports with their finance tool, that provides them insights and visualizations.  - ML engine to help them make future financial predictions. | * Rides information: date of ride, Payment method, amount charged, amount paid, discounts applied | As every company, Flyber needs to attain to accounting norms and also monitor its revenue and expenses to make better decisions that will maintain the company afloat. |
| Product management | - Identifying customer pain points  - Analyze customer interaction data on site and app.  - BI and visualization tools to tell stories about how customers use the app and their needs | * User log data: event type, event count, device type * Rides information: Distance, duration, pick up point, drop off point, time between user ordered the ride and driver arrival, reported incidents | In order to provide a smooth user experience, Flyber needs to be in constant improvement of their user app. |
| Engineering | * Monitoring app availability | * User log data: event type, event count, device type | Flyber’s main product is the app, so it’s crucial that it’s mantainted and available at all times. |

**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:**

*customers*

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *User\_uuid (Primary key)* | *first\_name* | *last\_name* | *email* | *phone* | *address* | *city* | *birthday* | *gender* |

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

*I chose User\_uuid as the primary key because is the unique identifier from every customer, every customer will appear only once in this table (table granularity = customer) and not 2 customers can have the same user\_uuid. There are no foreign keys because this is an entity that will be referred by the other tables, not the other way around. In a data warehouse, this will be a dimension table.*

**Table 2:**

*rides*

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

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Ride\_id (Primary key)* | *User\_uuid (Foreign key)* | *Vehicle\_uuid (Foreign key)* | *Driver\_uuid (Foreign key)* | *pickup\_location* | *pickup\_timestamp* | *dropoff\_location* | *dropoff\_timestamp* | Payment method | amount charged | amount paid | discounts applied |

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

*Although we could uniquely identify each ride with its data (eg. combination of user and timestamp), I decided to set one unique identifier Ride\_id to make it more efficient.*

*The foreign keys user\_uuid and vehicle\_uuid, are needed to link them to the ride. This table, in a data warehouse, would be a fact table.*

**Table 3:**

*events*

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Event\_uuid (Primary key)* | *User\_uuid (Foreign key)* | *event\_type* | *timestamp* | *app\_version* | *event\_page* | *device\_type* |

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

*The evet\_uuid is the unique identifier of each row, and that’s why it has been chosen as the primary key.*

*Since each event is triggered by a user, we need the user\_uuid as foreign key to link it.*

## **Section 3:** Extraction and Transformation

Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are currently collecting in the pipelines and they provide you with section\_3\_event\_logs template (which you can download from the classroom) generated by rider’s activities on the Flyber App. Also provided in the Project Resources.

**Extraction and Transformation-1**

ETL is performed on the provided Event Logs Template and results will be transferred to the proposal template. The project's ETL should be created inside of your copy of the Event Logs template in the tab titled, ETL. Clicking on the link above will create a copy of the Event Logs for you

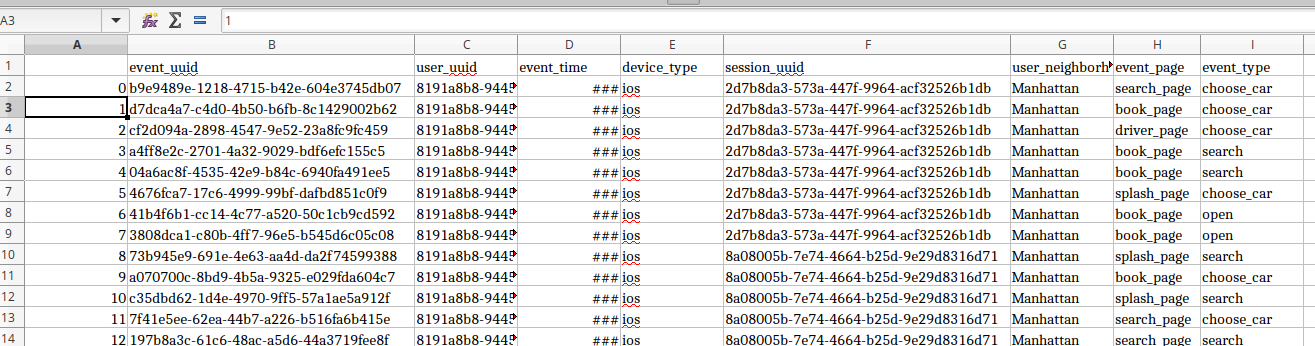
After being provided with a CSV log file, use extraction techniques to be able to get the data into a usable form. Because this needs to be a repeatable process we need to document it in order to assess its feasibility. Below,

1. Write the steps you took to extract the data and provide reasoning for why you used this method *Note: Don't forget to include any file type changes*:
2. Perform cleaning and transformation of the data in the ETL tab and document.
3. Document and provide rationale for all of your steps below as well.

Steps for Extraction:

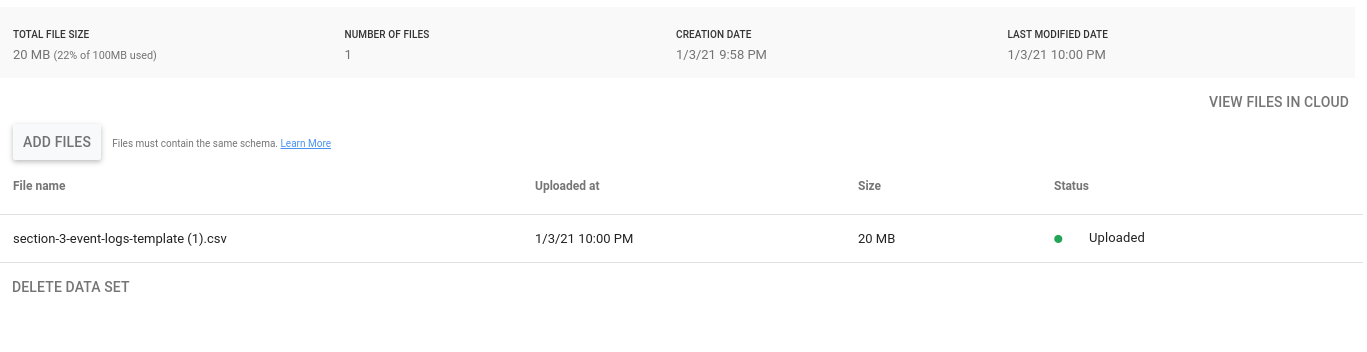
*(You may add more steps if necessary.)*

1. *Drop the first column of the document and convert file to csv*

**

* 1. First row doesn’t provide any extra information, is just the number of row, so there’s no need for it. We convert the data to csv, otherwise it can't be imported in Google Data studio.

1. *Upload file to google data studio*

**

* 1. I chose Google data studio because, similar to Tableu, provides easy data management and visualizations.

1. *Drop duplicates*
   1. It can sometimes happen that some events are inserted duplicated in the database, those should be removed to avoid miscalculations.
2. *Validate data types*
   1. It’s important to validate and check that dates are in the correct format to avoid errors in calculations

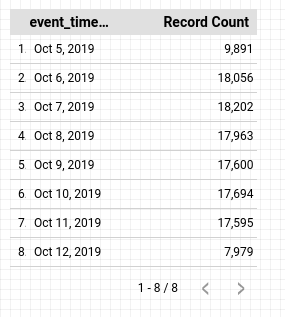
**Transformation-2**

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 |

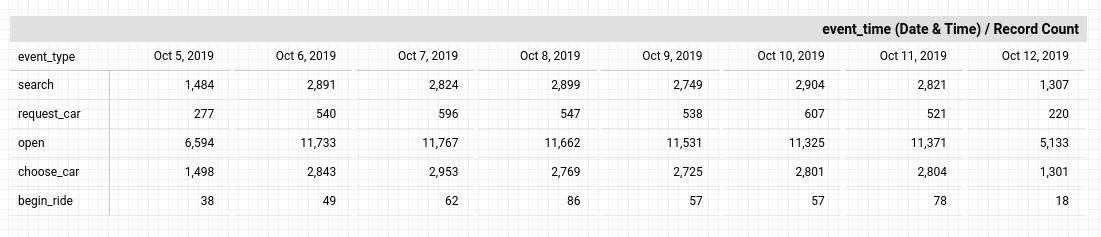
Source calculation:



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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 10/5/2019 | 10/6/2019 | 10/7/2019 | 10/8/2019 | 10/9/2019 | 10/10/2019 | 10/11/2019 |
| Choose Car | 1498 | 2843 | 2953 | 2769 | 2725 | 2801 | 2804 |
| Search | 1484 | 2891 | 2824 | 2899 | 2749 | 2904 | 2821 |
| Open | 6594 | 11733 | 11767 | 11662 | 11531 | 11325 | 11371 |
| Begin Ride | 38 | 49 | 62 | 86 | 57 | 57 | 78 |
| Request Car | 277 | 540 | 596 | 547 | 538 | 607 | 521 |

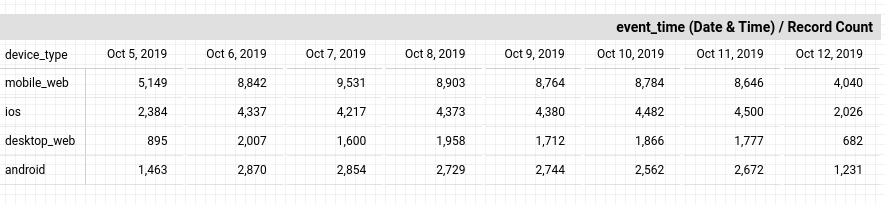
Source calculation:



1. How many events per device type per day?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 10/5/2019 | 10/6/2019 | 10/7/2019 | 10/8/2019 | 10/9/2019 | 10/10/2019 | 10/11/2019 |
| ios | 2384 | 4337 | 4217 | 4373 | 4380 | 4482 | 4500 |
| android | 1463 | 2870 | 2854 | 2729 | 2744 | 2562 | 2672 |
| Desktop Web | 895 | 2007 | 1600 | 1958 | 1712 | 1866 | 1777 |
| Mobile Web | 5149 | 8842 | 9531 | 8903 | 8764 | 8784 | 8646 |

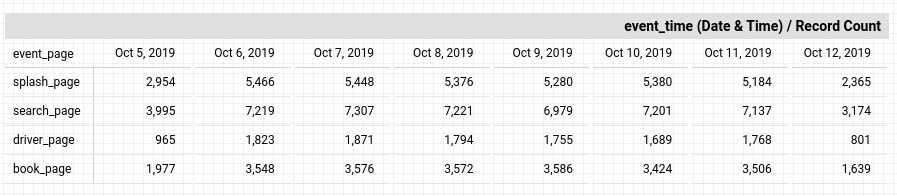
Source calculation:



1. How many events per page type per day?

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

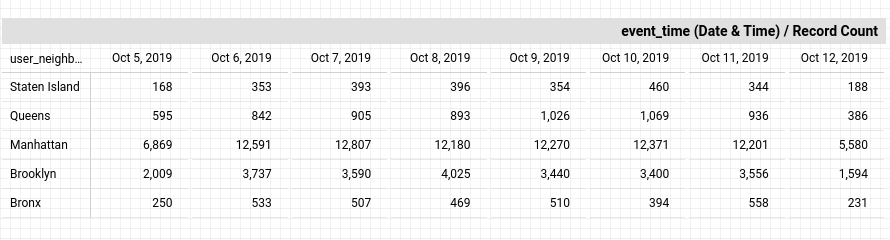
Source calculation:



1. How many events for each location per day?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 10/5/2019 | 10/6/2019 | 10/7/2019 | 10/8/2019 | 10/9/2019 | 10/10/2019 | 10/11/2019 |
| Manhattan | 6869 | 12591 | 12807 | 12180 | 12270 | 12371 | 12201 |
| Brooklyn | 2009 | 3737 | 3590 | 4025 | 3440 | 3400 | 3556 |
| Bronx | 250 | 533 | 507 | 469 | 510 | 394 | 558 |
| Queens | 595 | 842 | 905 | 893 | 1026 | 1069 | 936 |
| Staten Island | 168 | 353 | 393 | 396 | 354 | 460 | 344 |

Source calculation:



**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.

*This ETL process is simple, fast and straightforward, but it’s not production ready. Although it accomplishes what it’s needed, extracting, transforming and loading the data, it can’t be easily automated and won’t be scalable following this manual way.*

*The best way to implement this ETL pipeline in production would be using tools developed for such purpose. I would store the files with the event data in s3 buckets, pick them up with airflow, execute datatype validation and drop of duplicates with this tool and then insert into redshift.*

## 

## **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.

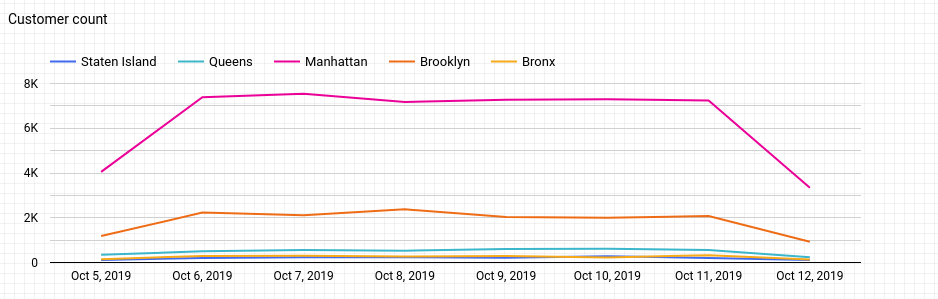
Choose one of the following prompts that you think can get you the most relevant information to proceed further.

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:

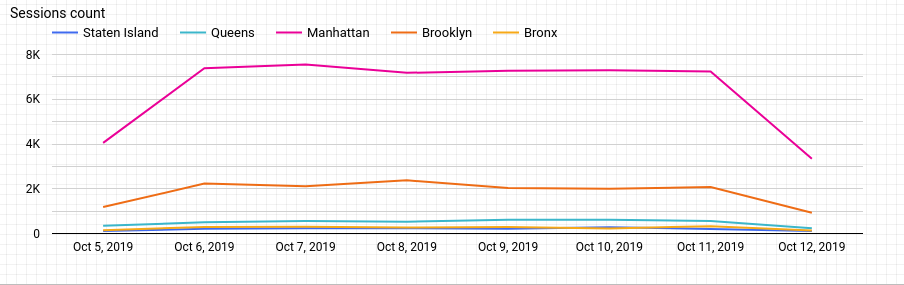
1. How much is the customer data increasing?

Customer data remains at a stable amount every day, there’s not much increase from one day to the other. The only exception is the first and last day of the dataset, that has around half of the amount of events the rest of the days have.



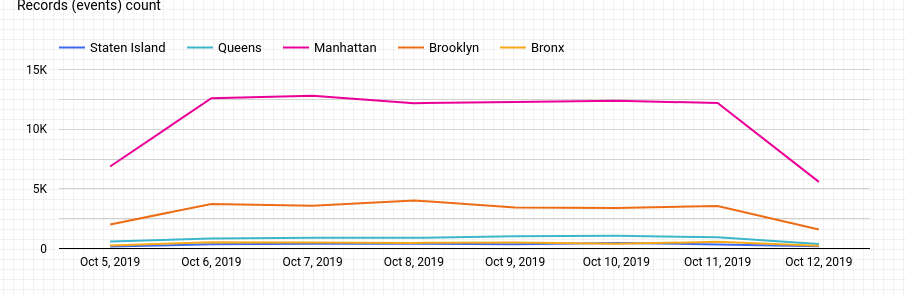
1. How much is the transactional data increasing?

Similar to customer data, transactional data is stable across the sample.



1. How much is the event log data increasing?

As seen with customer and transactional data, event log data remains pretty much the same on daily basis.



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

* Event Log Data
* Transactional Data
* Customer Data

The most important data to answer how many events per day per location we have is the Event Log Data. This data provides the events associated to a date and a user\_neighborhood that would allow us to check events per day by each area.

## **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.]*

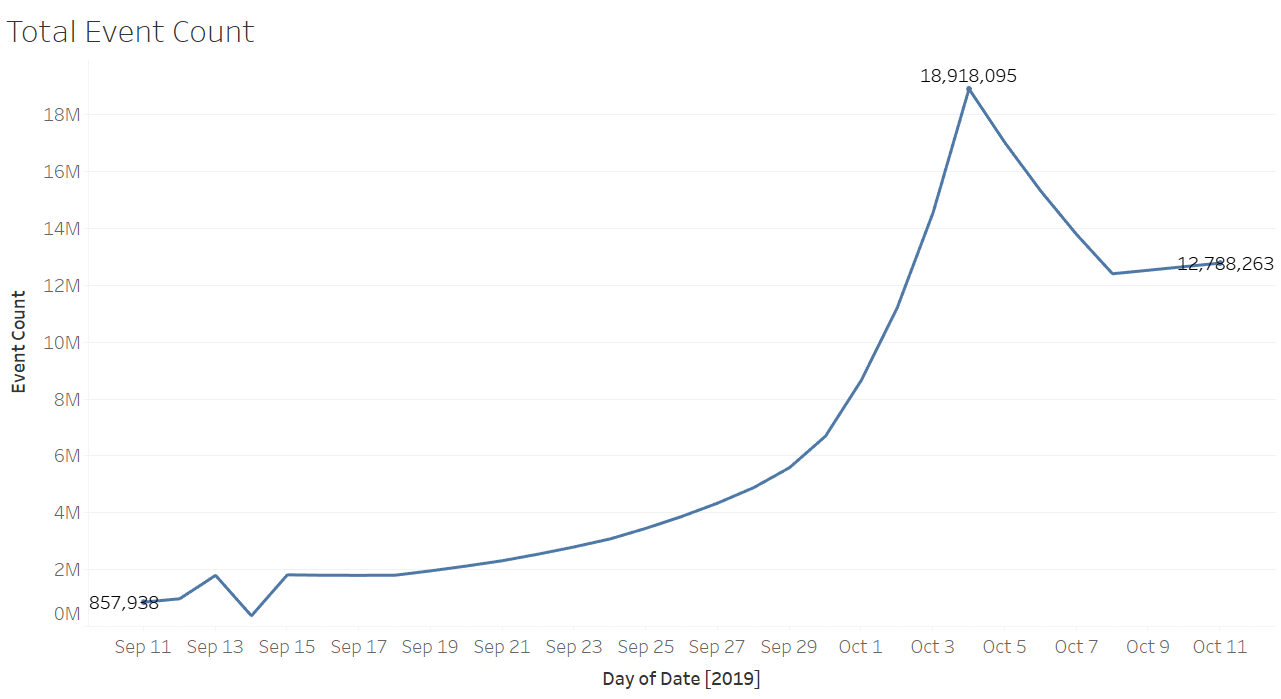
## **Section 6:** Business Insights

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:



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

*The data used for quantifying Flyber’s data growth is the total event count because it is directly related with the log growth and the total amount of data storage that will be needed.*

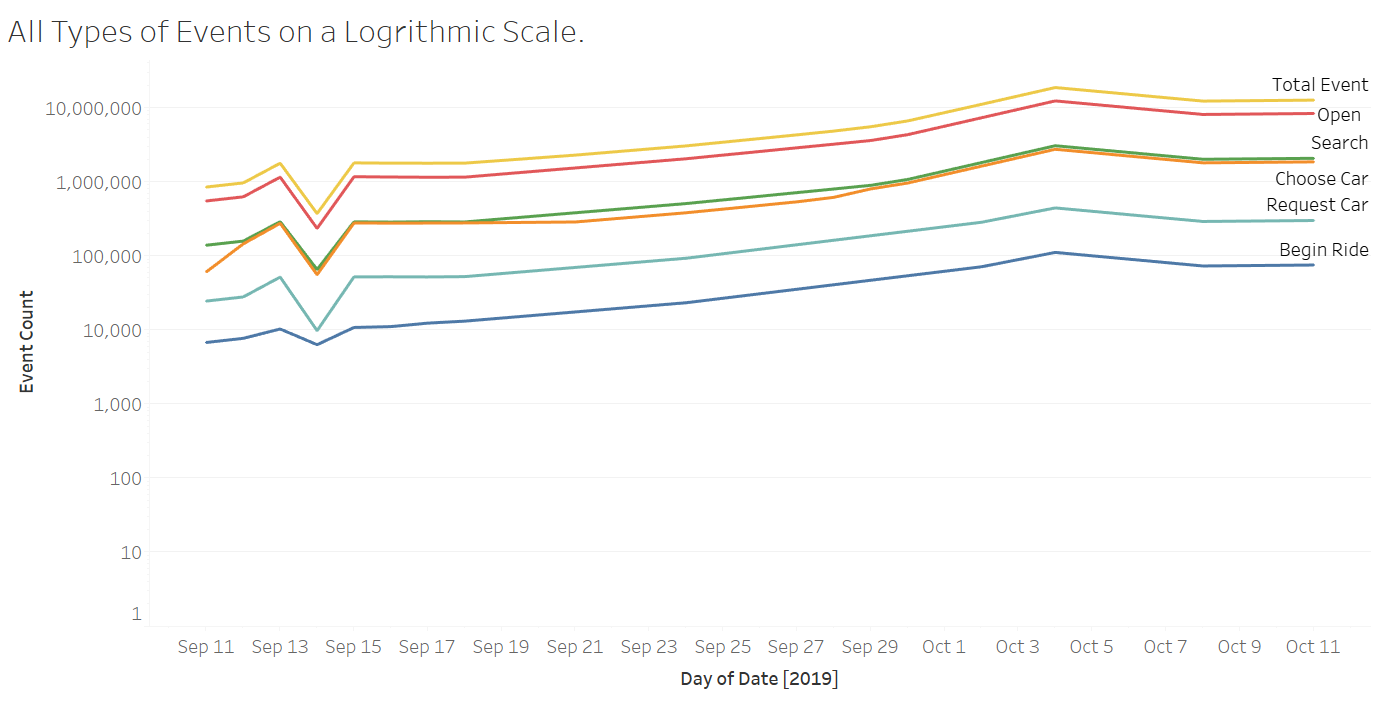
*As seen in the graph above, Total Event count grew exponentially over the past month, from 857,930 events on 11 September to 18,918,095 events on 3 October which was the highest peak in the last month.*

What is the fastest growing data and why?

This data, the events' data, is the fastest growing data because it comprises all the datasets. No other datasets will grow without events data growing.

**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

*By analyzing the data growth in the events' data over the last month an increasing linear pattern can be seen. The amount of events grew at a stable pace except for a peak seen on October 3.*

*This is very good because it notes how our customers are using more our app, generating more and more events on daily basis.*

*Judging by the peak in the data seen on October 3rd, we can assume that the October Marketing Campaing had a huge impact and although not all of those newly acquired customers seem to have been retained, Flyber’s services kept growing.*

*There’s a high correlation between marketing campaigns and data generation, because if a marketing campaign is successful the data generation will increase significantly to show it.*

## **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.

*I think the best option for Flyber is to chose a cloud solution DWH. FLyber is a startup that expects to grow rapidly, but doesn’t have the money, in-house expertise, nor the time to invest in an on-premise solution.*

* Cost: On-premise solutions require a lot of investement at the beggining because big machines have to be installed thay may not be fully needed at the start, and then everytime more capacity is needed, more equipment has to be bought. Cloud based solutions provide lower costs for starting to set-up a datawarehouse, because it will depend only on the computng resources needed, and that will add up with time as needed.
* Scalability: On-premise solutions require to buy more expensive hardware everytime more compute resources are needed, that takes time and a lot of investement. Cloud based solutions provide scalability in a matter of minutes, and only as much is needed.
* In-house Expertise: On-premise solutions require at least one data engineer in house, but cloud based solutions are more manageable and have less failures.
* Latency/Connectivity: Depending on the location where the data is being fetched, if the zone remains the same as the server, then On-premise solutions provide a better connectivity and latency rates that cloud based solutions.
* Reliability: Cloud based solutions always provide SLA’s that ensure a high availability, unlike on-premise solutions where the burden relies all in the in-house expertise.

**Suggested DWH**

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

*I think the best option for Flyber is* Redshift. This DWH is ideal for small startups that are setting up their first DWH, it’s cheaper than Google BigQuery and Snowflake, options that are focused on bigger companies.

Redshift is also easy to use and has connectivity with other AWS tools that will be useful for Flyber’s data strategy like S3 buckets.

## Image Appendix

Image 1: Log Growth

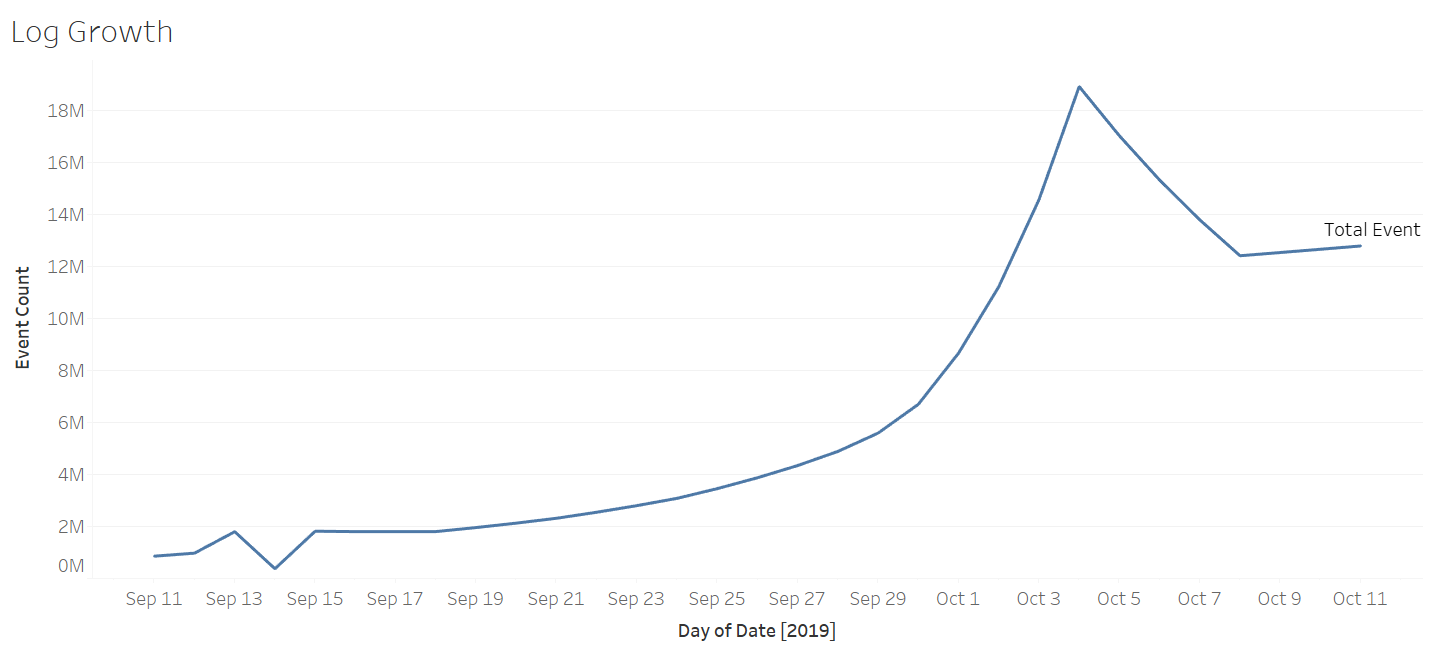


Image 2: Ride Growth

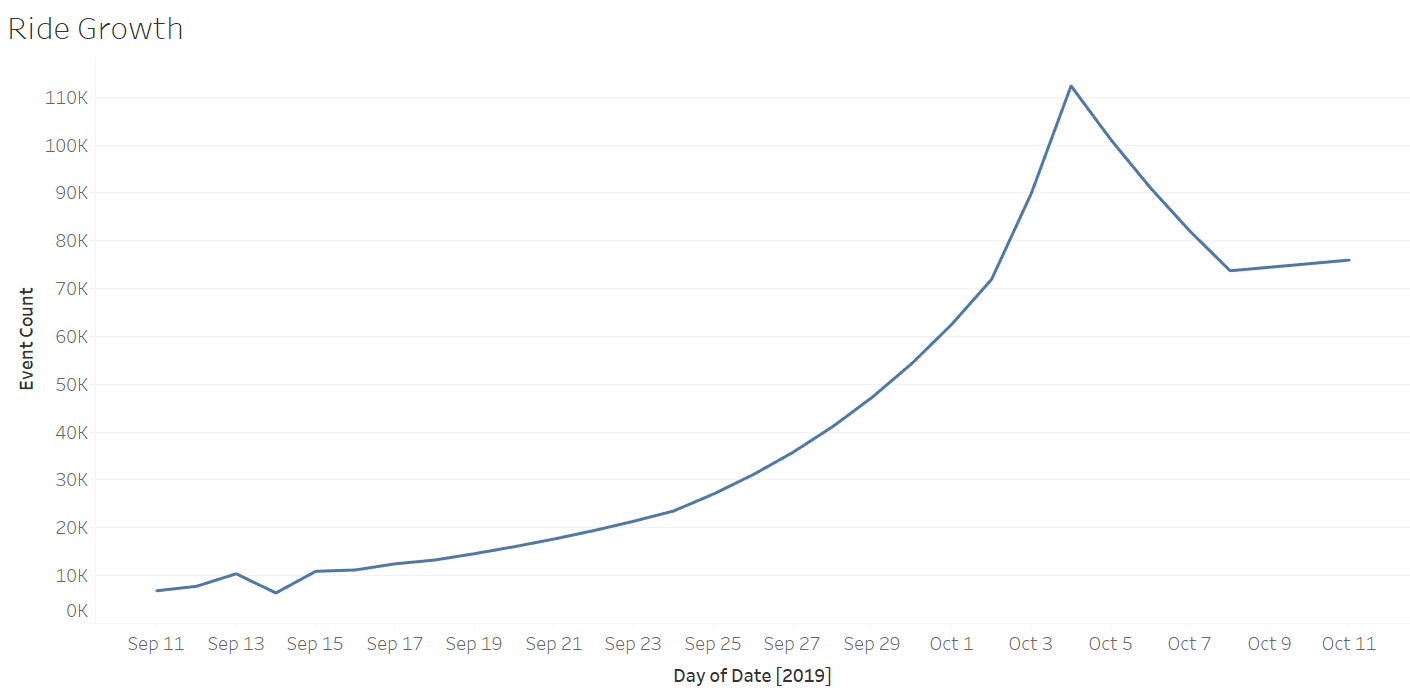


Image 3: Total Event Count

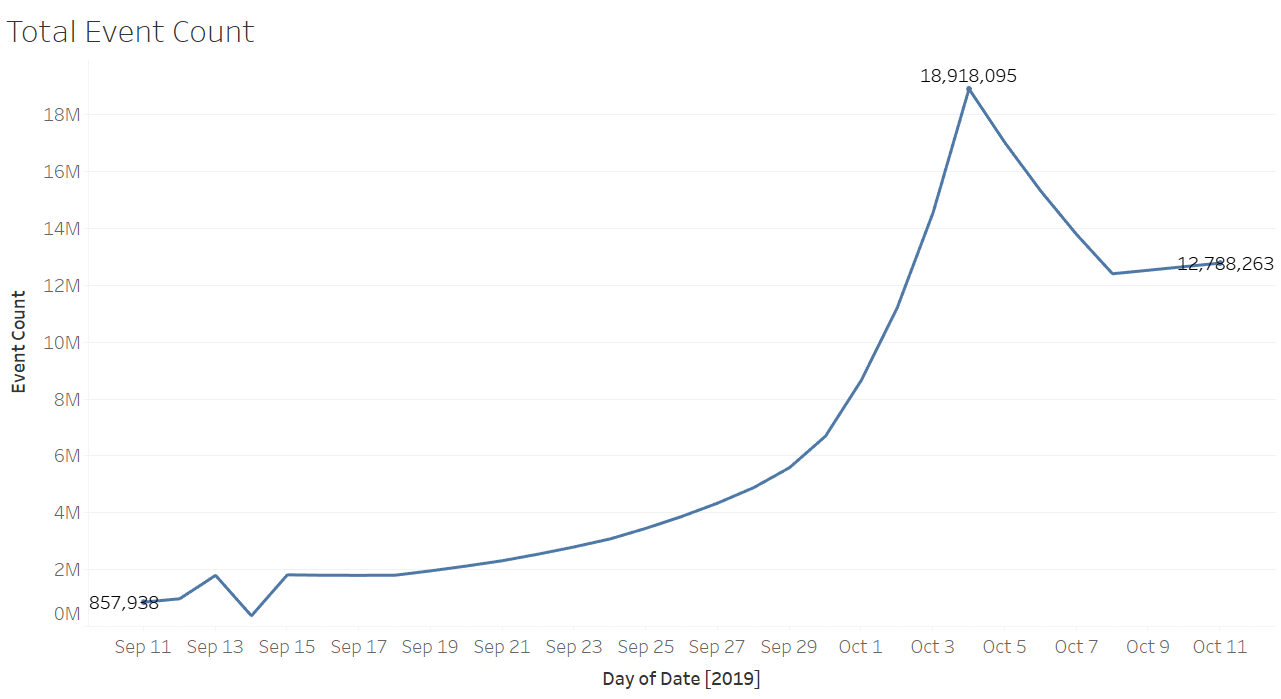


Image 4: All Events Log Scale

