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**
Engineering	Develop the working platform for drivers and riders.	Monitoring App and Traffic data
Marketing	As a startup, the Company needs to attract new users.	Targeted advertising
Finance	Finance team needs to record and predict Flyber's Profit and Loss.	Monitoring current P&L

Customer Care	Address customers' needs and collect feedback for Company's development	Provide personalized Surveys to the customer

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?**
Engineering	Monitoring App and Traffic data	Event Data	Monitor the traffic and riders's data, they can respond to riders' needs of the app quickly.
Marketing	Targeted advertising	Entity Data	Identify and acquire potential customers.
Finance	Monitoring current P&L	Entity Data	Collect revenue and cost data, so they can perform analysis of P&L, to find out if the business is profitable or not.
Customer Care	Provide personalized Surveys to the customer	Event Data	To understand customers' need thus improve the performance of the App,

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:

Rides Geographic information

(You may add more columns if necessary.)

Rider ID	*PIckup Location*	*Pickup DateTime*	*Drop-off Location*	*Drop-off Datetime*

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

Engineering Team can leverage these data to match the riders' needs and assign drivers to them. Engineering Team needs data to perform Algorithms Analysis to deliver results.

Table 2:

Customer Demographics

(You may add more columns if necessary.)

Rider ID	*Email*	*Address*	*Phone*	*Gender*	*Age*

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

From these Demographic information, Marketing Team can identify and target potential customers.

Table 3:

Trip

(You may add more columns if necessary.)

Trip ID	*Distance*	*Duration*	*Total Price*

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

Finance Team can calculate total revenue from the analysis, thus perform P&L analysis for the Company.

Table 4:

Customers' Feedback

(You may add more columns if necessary.)

Customer ID	*Trip ID*	*Rating for the Trip*	*Customer's thoughts*

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

Customers Care Team can leverage these data to evaluate satisfaction of customers, and understand where can we improve the App performance,

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. *Data Collection*
    a. *We collect data in Tableau for manipulation, analysis and visualization*
2. *Data Verification*
    a. *We check size of the data, and current data type.*
3. *Assimilate both Records and Source*
    a. *We confirm that the records we have corresponds to what was recorded initially by the source.*
4. *Search the duplications*
    a. *We check the duplicates and delete them for better data analysis.*

Transformation-2

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

1. How many events are being recorded per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Event Count	9891	18056	18202	17963	17600	17694	17595

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

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Choose Car	1498	2843	2953	2769	2727	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	595	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	4373	4380	4482	4500
android	1463	2870	2854	2729	2744	2562	2672
Desktop Web	895	2007	1600	1958	1712	1866	1777
Mobile Web	5149	8842	9531	8903	8764	8784	8646

4. How many events per page type per day?

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

5. How many events for each location per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Manhattan	6869	12591	12807	12180	12270	12371	12201
Brooklyn	2009	3737	3590	4025	3440	3556	1594
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.

From the current data size, the ETL process is efficient. But considering potential growth in the future, the manually ETL process is not Scalability. Also, in some cases, the data we need is not structured, it may cause huge burden for manually Extraction and Transformation. I would suggest less Extraction and Transformation for future process.

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

I would choose :
How many events of each event type per day?

1. *How much is the customer data increasing?*
   The data shows that customers conducted different actions for each day. For example, the number of customers who began ride is 38 on Oct 5, and reached 78 on Oct 11. The number of customers who searched is 1484 on Oct 5, and reached 2821 on Oct 11.

2. *How much is the transactional data increasing?*
   For completed transactions, we can look at the event type of 'Began-Ride', the number of customers who began ride is 38 on Oct 5, and reached 78 on Oct 11.
3. *How much is the event log data increasing?*
   For total events, we can see that the number of total events is 9891 on Oct 5, and reached 17595 on Oct 11.

Which of the following data is ***most*** important to answer this question? Why?
Event-log data is the most important, because we can get information for each event type.

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

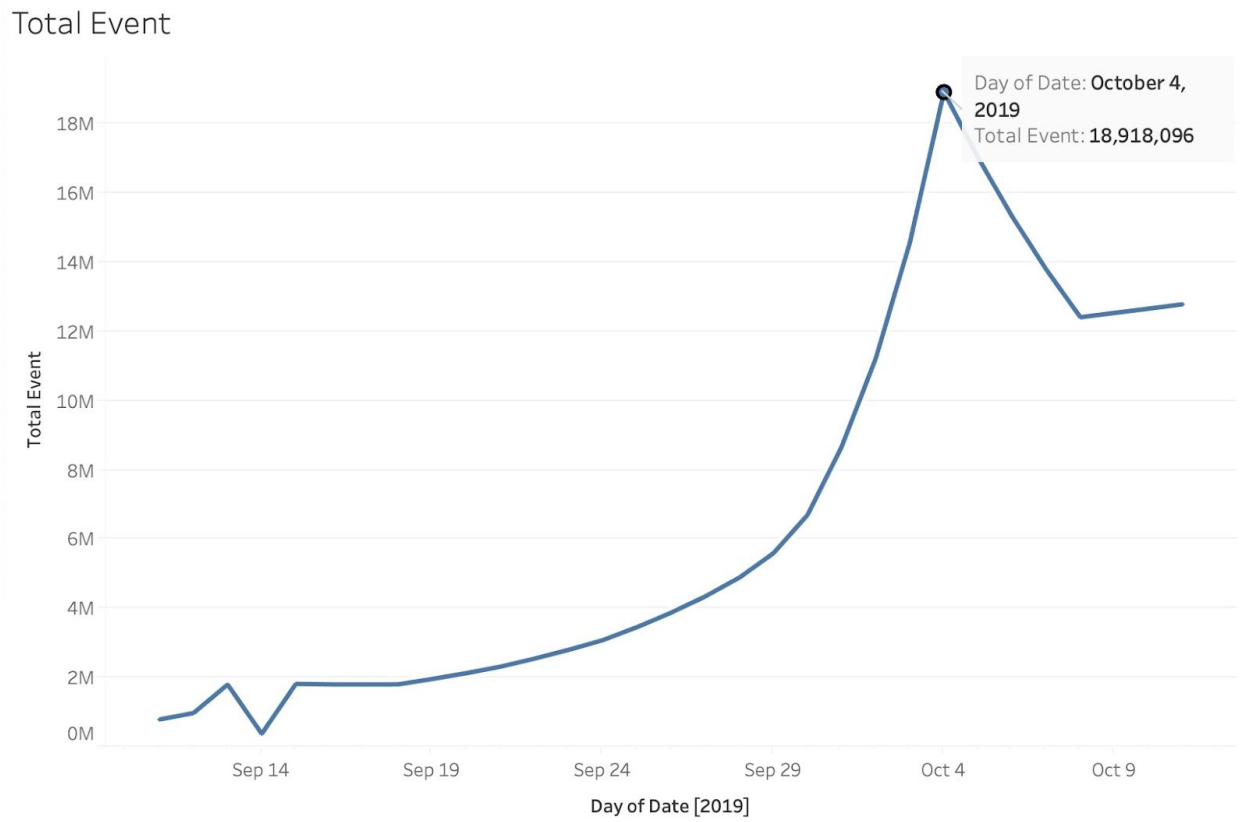
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

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:



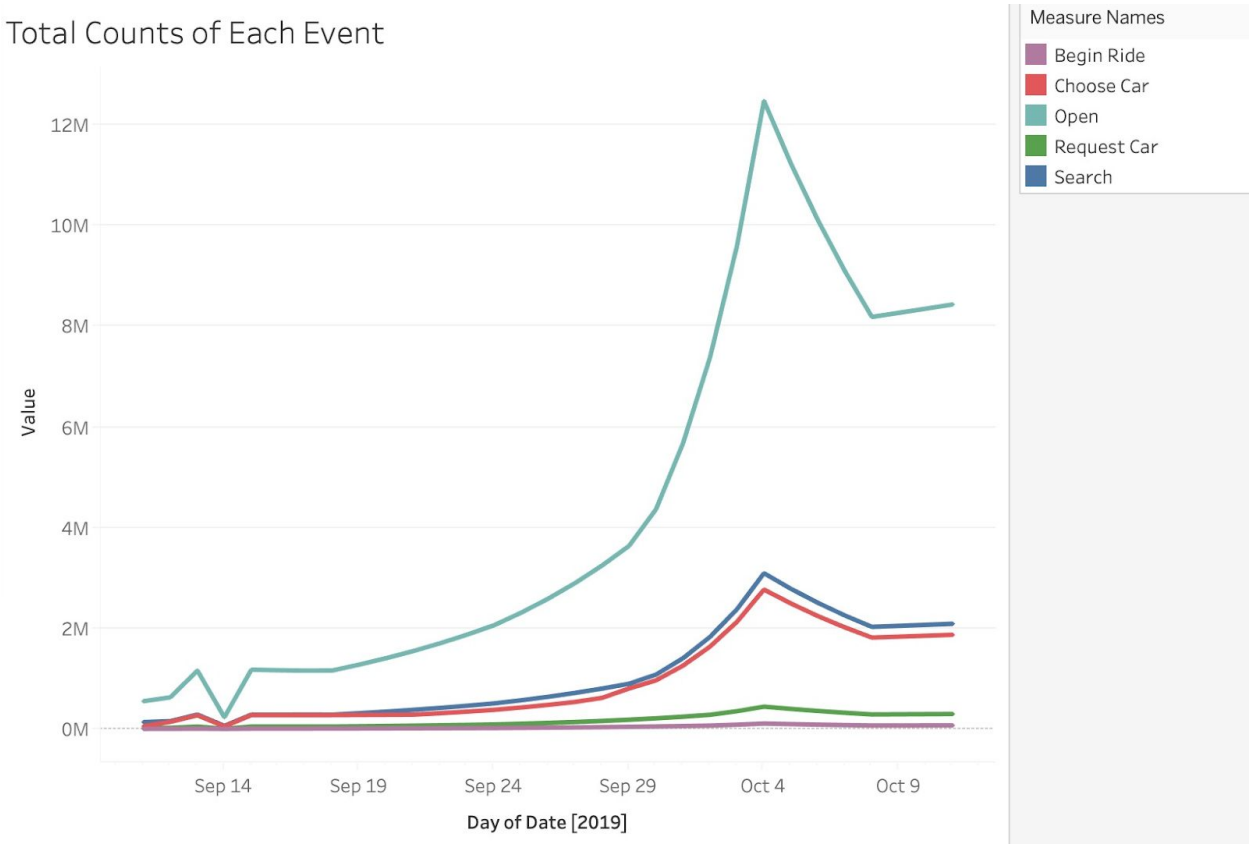
Data Story: This graph tells us:

The total events experienced low growth during Sep 14 to Sep 29, and high growth through Sep 29 to Oct 4, reached peak at 18,918,096 on Oct 4. And after that, decrease slowly.

This graph was created using the following steps:

1. *Load data in Tableau*
2. *Choose related field: Date and Total events count*
3. *Generated line chart in Tableau*

Visualization 2:



Data Story: This graph tells us:

Most users conducted 'Open', but few of them 'Begin Ride'.

This graph was created using the following steps:

1. *Load data in Tableau*
2. *Choose related field: Date and all Event Types*
3. *Generated multiple lines chart in Tableau*

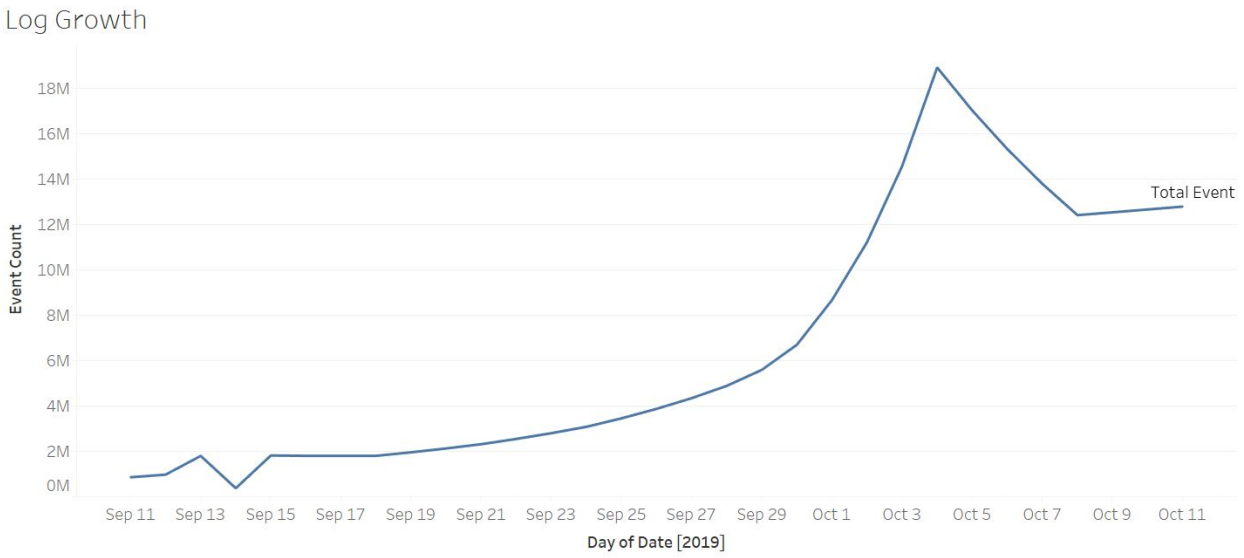
Section 6: Business Insights

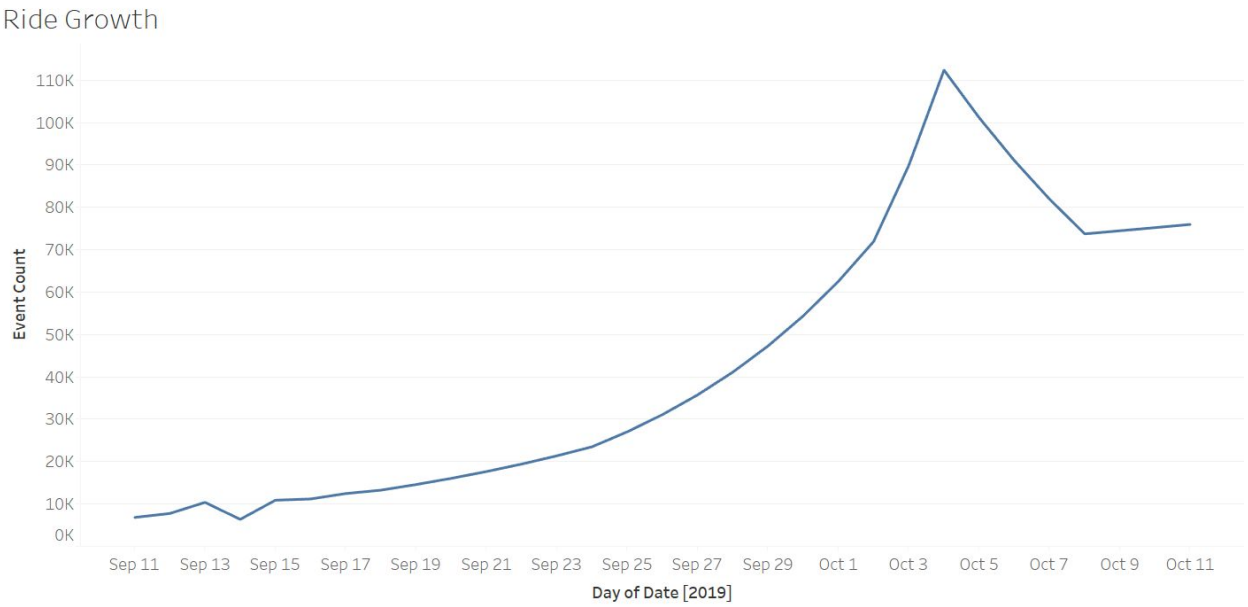
The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received 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 number of total events on Sep 11 is 790,329 and reached a peak at 18,918,096 on Oct 4, which increased 2394% compared to Sep 11.

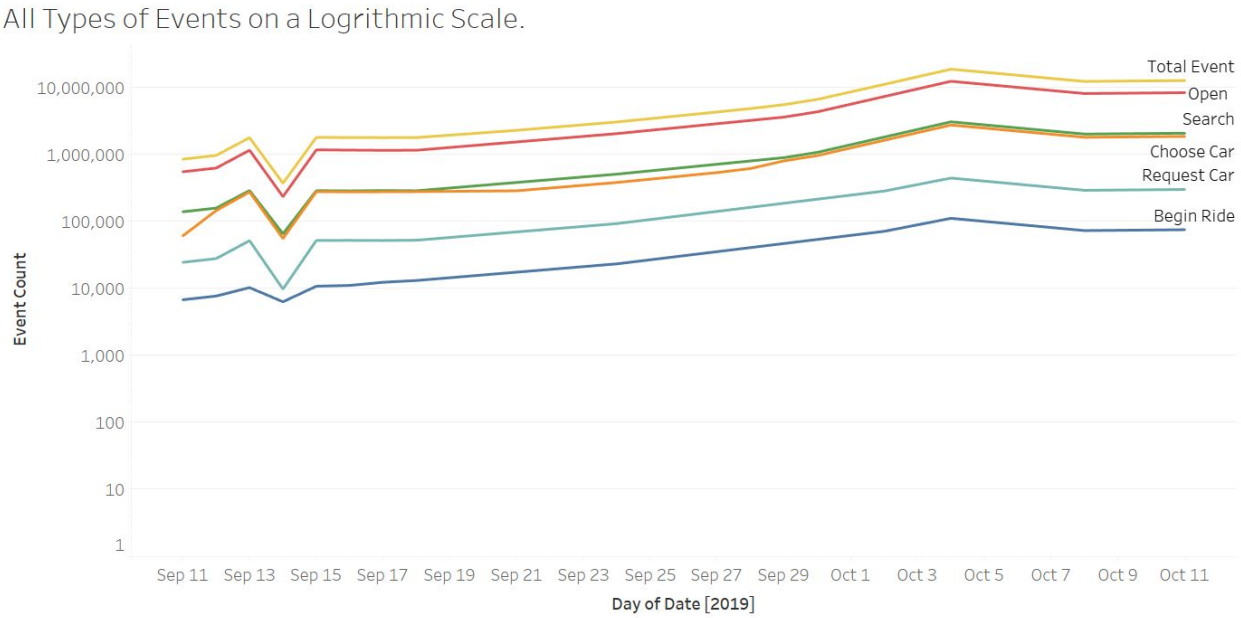
What is the fastest growing data and why?

We can see that since Sep 22, the growth rate for the number of total events exceeds 10% everyday, and exceeds 20% from Sept 30 to Oct 4. The data growth in for Ride experienced a similar pattern to growth of total events.

Day of Date	
September 11, 2019	
September 12, 2019	123.4%
September 13, 2019	184.6%
September 14, 2019	20.9%
September 15, 2019	483.9%
September 16, 2019	99.2%
September 17, 2019	99.9%
September 18, 2019	100.2%
September 19, 2019	108.6%
September 20, 2019	108.7%
September 21, 2019	108.8%
September 22, 2019	110.0%
September 23, 2019	110.0%
September 24, 2019	110.0%
September 25, 2019	112.1%
September 26, 2019	112.1%
September 27, 2019	112.1%
September 28, 2019	112.5%
September 29, 2019	114.4%
September 30, 2019	119.8%
October 1, 2019	129.4%
October 2, 2019	129.5%
October 3, 2019	129.8%
October 4, 2019	129.8%

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

- *All events types experienced a similar growth pattern, steady growth from Sep 14 to Oct 4.*

- *Good. Because it suggested consistency in the users behavior. Thus we do not need to worry about discrepancy in different event types.*
- *We can see during the first week of October, the events experienced high growth (>20%), showing the Marketing Campaign impact on Data generation.*
- *If the Marketing Campaign can target users correctly, it will significant increase the events and thus increase the generated Data.*
- *It is very important to know the relationship Between Marketing Campaigns and Data Generation, because efficient Marketing Campaigns will significantly impact the user behaviors and thus increase total events for Data Generation, which are useful for further business analysis.*

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 propose we should move to the cloud.

On-Prem gives us full control and governance in our hands, but this requires high costs, which is not applicable for a start-up like us.
Also, Cloud has advantages at:

- *We can scale up and down quickly, and pay only for the infrastructure used, which lower infrastructure costs.*
- *Our teams can focus on innovation and core business functionality instead of focusing on making underlying infrastructure work, which saves us lots of resources considering we are a start-up.*
- *Cloud can be accessible from anywhere, and is always support- available.*
- *Cloud technologies have evolved tremendously and give us an opportunity to use the latest and greatest.*

Suggested DWH

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

I propose to use Amazon Redshift.
Cost: Amazon Redshift is very common in the market, the infrastructure cost of Amazon Redshift is acceptable for Flyber.
Scalability: We can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change. We can manage up to eight times (according to Amazon) more riders during peak times and Amazon Redshift to gain customer insights.
In-house Expertise: Amazon Redshift decreases resource wastes in establishing expertise in DWH.
Latency/Connectivity: Amazon Redshift has low latency. We can easily monitor the read/write latency, and database connection in the Amazon Redshift.
Reliability: Cloud service providers have an alacarta of time tested services and tools that can be readily used. Cloud will also help with data privacy and data residency laws. Amazon has lots of experience in the industry market, and also in the international market.

Image Appendix

Image 1: Log Growth

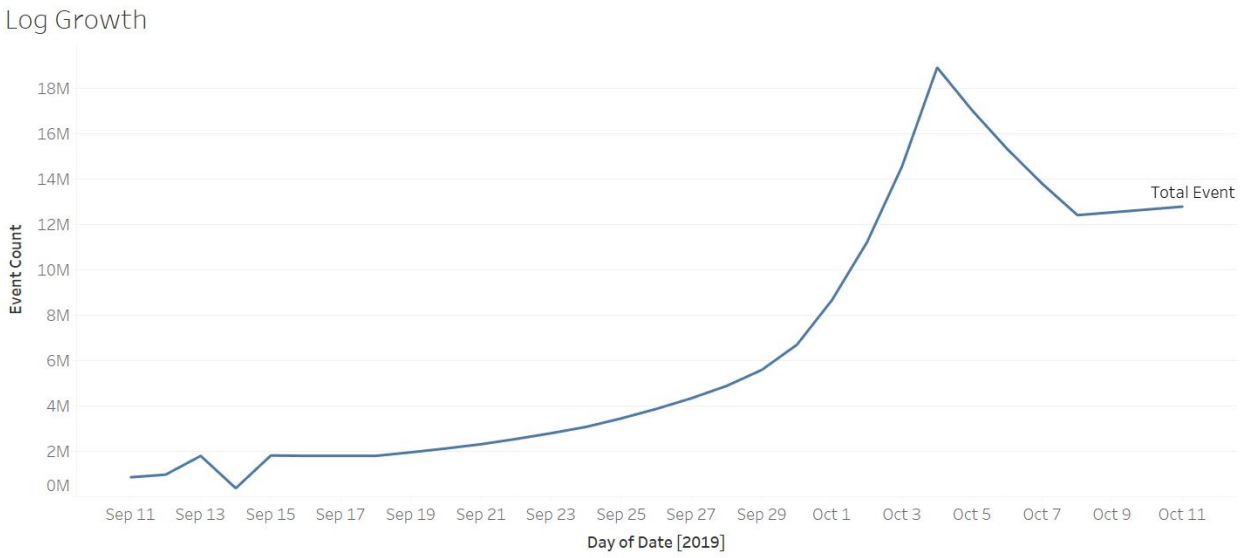


Image 2: Ride Growth

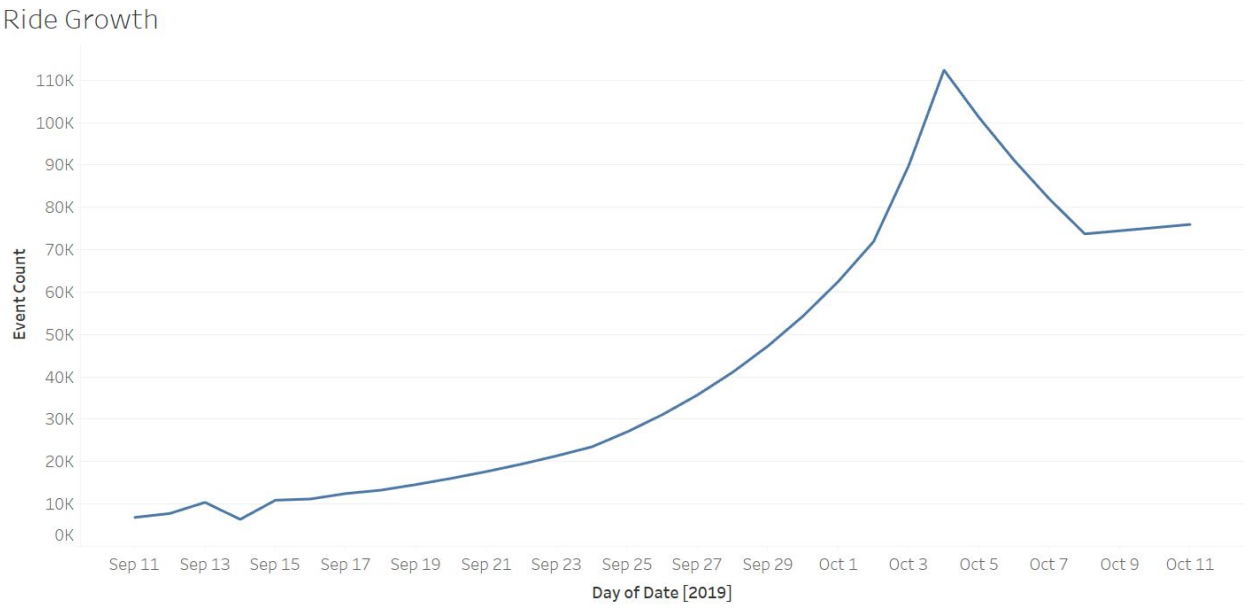


Image 3: Total Event Count

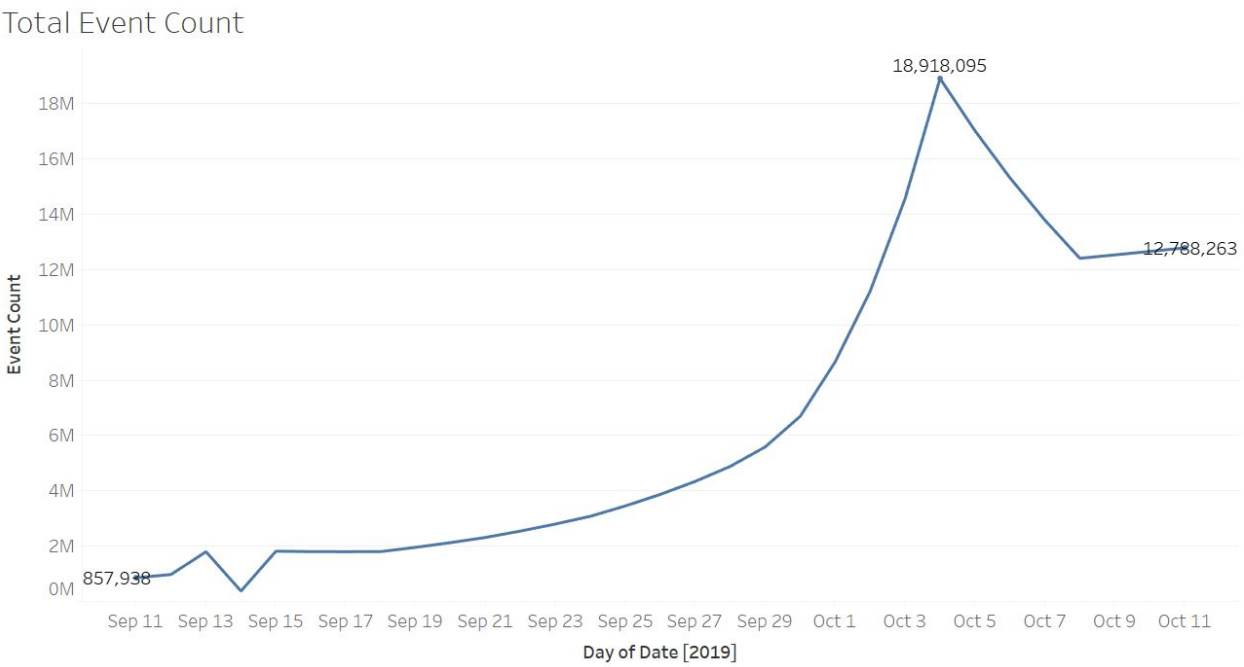


Image 4: All Events Log Scale