Flyber Data Strategy MVP

Introduction

Flyber has been massively successful. Results have beaten expectations and projections! This is good news for Flyber, but now it's time to plan for what's next. With success came some challenges. While we were able to grow, the original data pipelines to receive and process data are unable to keep up with the current and future growth.

As a Data Product Manager, working with multiple teams and stakeholders is imperative to success. To understand what our needs are, what scale we are growing at, and how we can build for the future, we need to consider all relevant stakeholders. In this proposal, present your findings along with the analysis and reasoning behind the choices made in order to help Flyber continue its success.

Section 1: Data Customers & Needs

Flyber is a two-sided platform. You have customers who are riders, and you have partners who are drivers/pilots (think Uber: riders and drivers). For the Minimum Viable Product, you will be focusing on the **Riders side of the business.** To build an end to end data pipeline the very first step is to understand who needs data and why they need that data. Within Flyber, identify who your primary data customers/stakeholders are, why they are your primary data stakeholders and how they want to use the data (primary use-cases).

Identify your primary internal stakeholders and their use-cases:
(You may add more rows if necessary.)

Stakeholder	Why are they primary stakeholders?	Use-Case
Marketing	Customer retention and acquisition is important	Targeted Advertising, Customized profile of consumers
Product Management	Needs to maintain tradeoff between engineering, business. Also to identify user pain points.	Customer analysis, Big Data

Engineering	Need to build product, features and need to maintain the product.	Monitoring Site, product
Customer Service	Understanding Customer grievances.	Provide personalized responses

Section 2: Data Collection and Data Modelling

To support our primary stakeholders use-cases we need following data:
(You may add more rows if necessary.)

Stakeholder	**Use-Case**	**Data**	**Why is this the primary use-case?**
Marketing	Targeted Advertising, Customized profile of consumers	Trip location data, Referral link, Ad click, new clients	We need to have strong user acquisition and retention which is not feasible completely without marketing. We need data for targeted advertising through promo codes etc.
Product Management	Needs to maintain tradeoff between engineering and business. Also to identify user pain points	Research data, Event data such as booking location	Customers are everything. We need to know what difficulties they faced, what features can be improved to make customer happy. “Happy customers is a happy business.”
Engineering	Need to build product, features and need to maintain the product.	Event data such as User funnel data	The product, app, website should be bug free and hassle free. The app should have Impressive UI/UX.

Customer Service	Understanding Customer grievances.	Customer review data. Customer rating data	From day one, the business should listen to customer reviews. This will help a business to iterate over the product.

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:

User Information

user_id	*user_name*	*user_password*	*name*	*email*	*Address*	*User_payment*

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

Primary key is the user_id. It is unique to all users. Foreign key in this table is the user_id.

Table 2:

Ride information

ride_id	*User_id*	*Pickup_location*	*Dropoff_location*	*timestamp*

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

ride_id + user_id will be the primary key. It is going to be unique identifier. We would have one unique key for every row. Foreign key is the user_id to run queries of the user information.

Table 3:

Ride completed Successfully

user_id	*ride_id*	*Ride_completed_check*	*Timestamp_completed_ride*	*Payment_done*

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

User_id and ride_id is the primary key and foreign key.

Section 3: Extraction and Transformation

Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are currently collecting in the pipelines and they provide you with section_3_event_logs template (which you can download from the classroom) generated by rider's activities on the Flyber App. Also provided in the Project Resources.

Extraction and Transformation-1

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

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

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

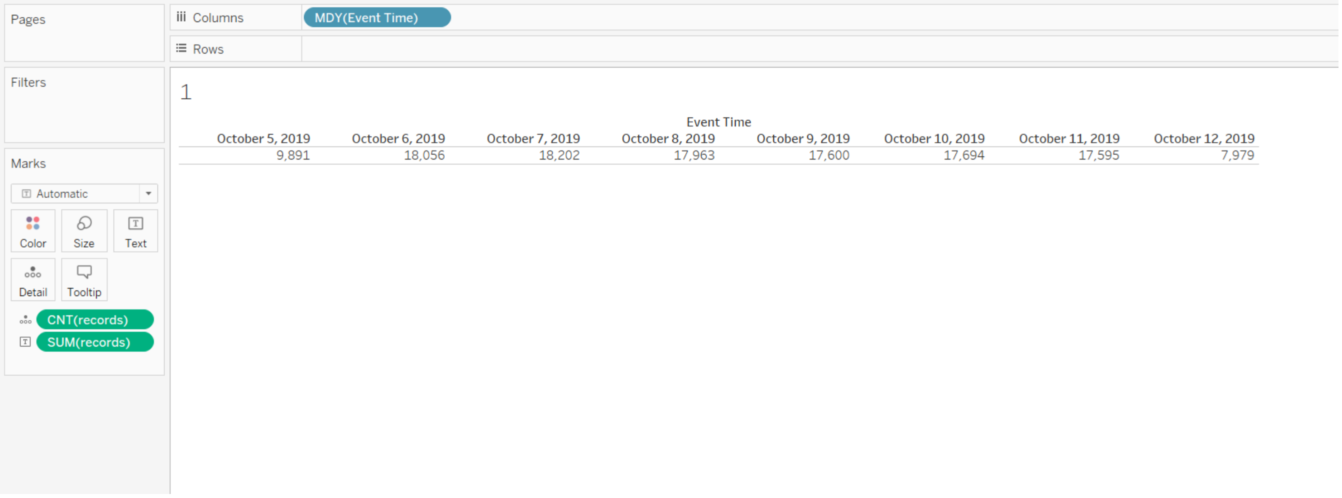
Steps for Extraction:

1. *Data cleaning*
    a. *We need to clean data to remove null values. So that we do not have extra rows in row calculations*
2. *Data Validation*
    a. Removal of *Data duplication has to be performed so that data analysis is proper.*
3. *Tableau*
    a. *Data is loaded into Tableau in Excel format. After this, create a calculated field "records" and input 1 as the entry in it.*
    b. *In first sheet, drag and drop Event time in Columns and select MDY format. Also ,drop records and measure the count.*
    c. *In second sheet, drag and drop event type with records field in marks section.*
    d. *In third sheet, do the above procedure with only device type in column section.*
    e. *In fourth and fifth sheet, drop page and location respectively in the column section.*

Transformation-2

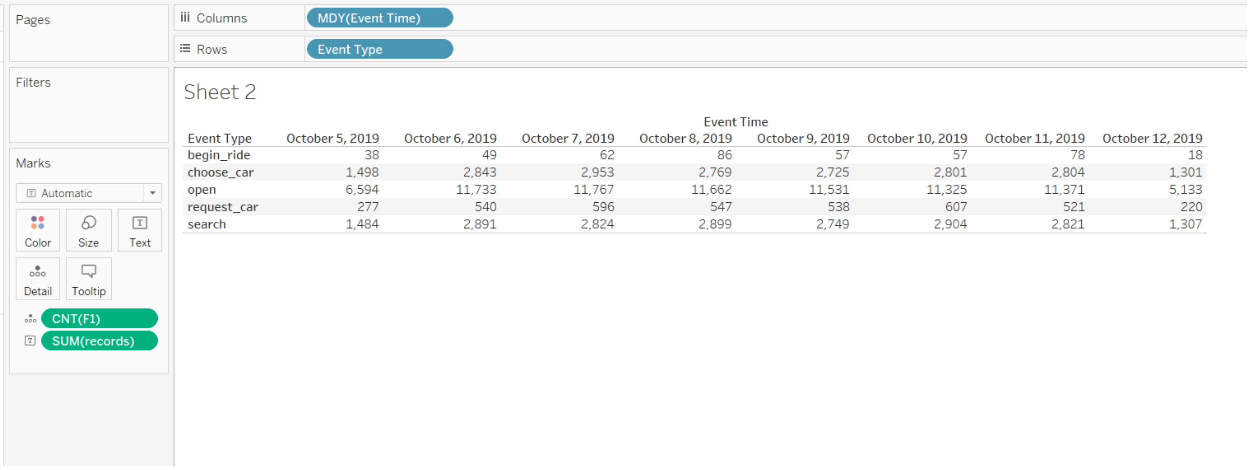
Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?



Event Time							
October 5, 2019	October 6, 2019	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019
9,891	18,056	18,202	17,963	17,600	17,694	17,595	7,979

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



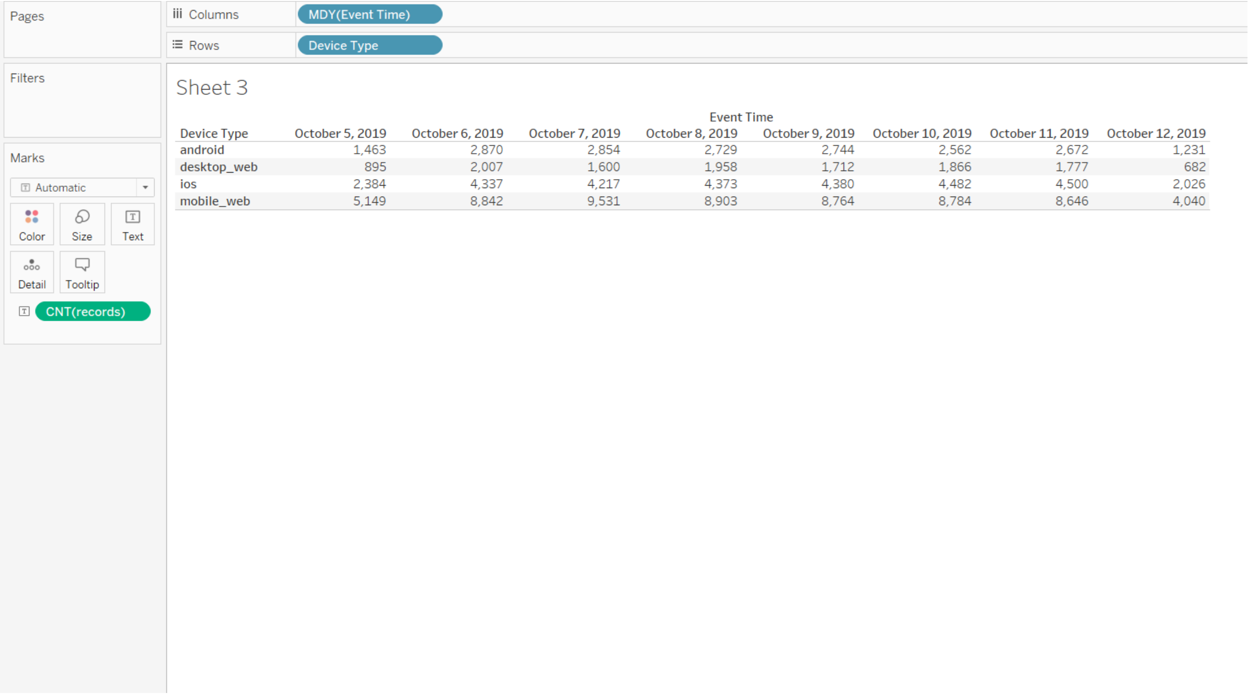
Columns: MDY(Event Time)

Rows: Event Type

Sheet 2

	Event Time							
Event Type	October 5, 2019	October 6, 2019	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019
begin_ride	38	49	62	86	57	57	78	18
choose_car	1,498	2,843	2,953	2,769	2,725	2,801	2,804	1,301
open	6,594	11,733	11,767	11,662	11,531	11,325	11,371	5,133
request_car	277	540	596	547	538	607	521	220
search	1,484	2,891	2,824	2,899	2,749	2,904	2,821	1,307

3. How many events per device type per day?



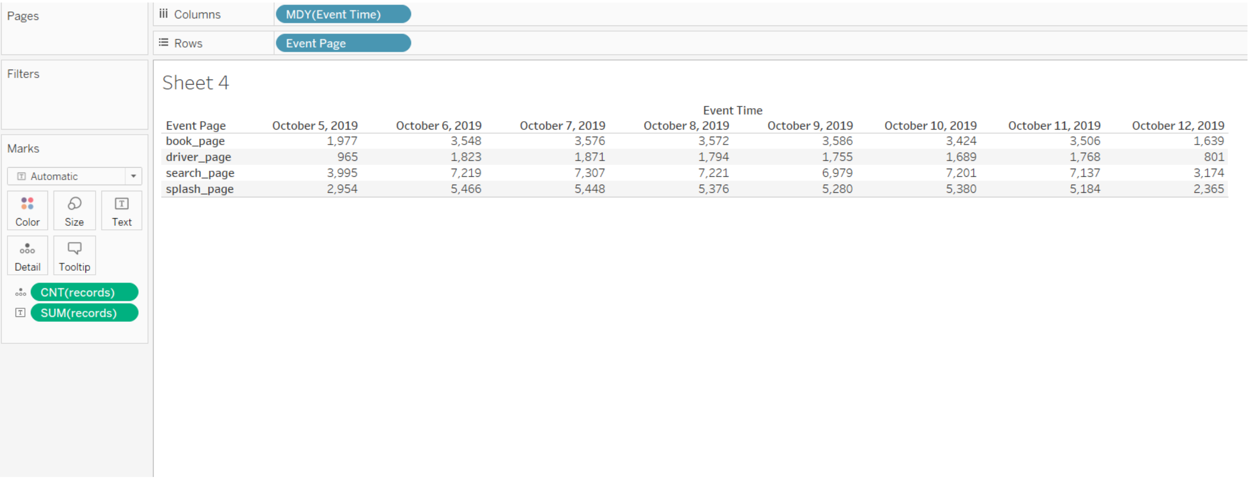
Columns: MDY(Event Time)

Rows: Device Type

Sheet 3

	Event Time							
Device Type	October 5, 2019	October 6, 2019	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019
android	1,463	2,870	2,854	2,729	2,744	2,562	2,672	1,231
desktop_web	895	2,007	1,600	1,958	1,712	1,866	1,777	682
ios	2,384	4,337	4,217	4,373	4,380	4,482	4,500	2,026
mobile_web	5,149	8,842	9,531	8,903	8,764	8,784	8,646	4,040

4. How many events per page type per day?



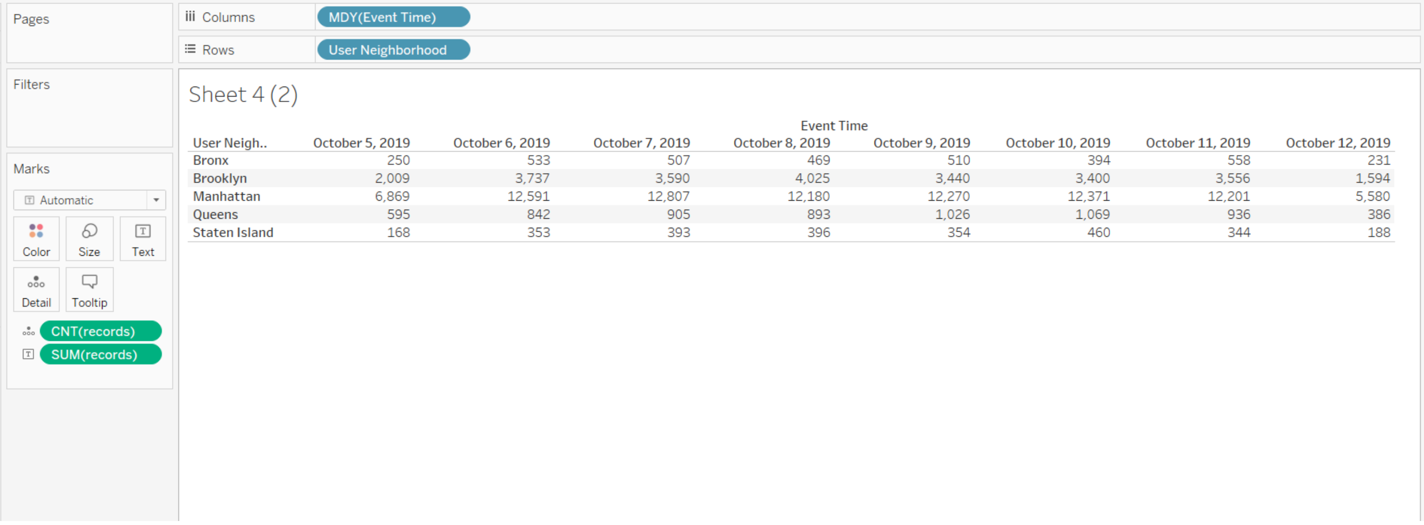
Columns: MDY(Event Time)

Rows: Event Page

Sheet 4

	Event Time							
Event Page	October 5, 2019	October 6, 2019	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019
book_page	1,977	3,548	3,576	3,572	3,586	3,424	3,506	1,639
driver_page	965	1,823	1,871	1,794	1,755	1,689	1,768	801
search_page	3,995	7,219	7,307	7,221	6,979	7,201	7,137	3,174
splash_page	2,954	5,466	5,448	5,376	5,280	5,380	5,184	2,365

5. How many events for each location per day?



Columns: MDY(Event Time)

Rows: User Neighborhood

Sheet 4 (2)

	Event Time							
User Neigh..	October 5, 2019	October 6, 2019	October 7, 2019	October 8, 2019	October 9, 2019	October 10, 2019	October 11, 2019	October 12, 2019
Bronx	250	533	507	469	510	394	558	231
Brooklyn	2,009	3,737	3,590	4,025	3,440	3,400	3,556	1,594
Manhattan	6,869	12,591	12,807	12,180	12,270	12,371	12,201	5,580
Queens	595	842	905	893	1,026	1,069	936	386
Staten Island	168	353	393	396	354	460	344	188

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.

[Insert Response Here.]

Data extraction, Data validation cannot be done in Excel as the data grows larger and larger. We will need warehouse and cloud services.

Section 4: Choosing Relevant Dataset

The previous exercise gave you a sneak peek into the Extraction and Loading aspects of ETLs in data pipelines. For making business decisions, a data consumer would like to have all the data they want. However, for any ecosystem, it is impossible to collect or provide everything that the customers need. In this exercise, you will get a taste of real world scenarios wherein:

- All the resources are not always available to get what you need.
- You have to get creative and get the most insights with a minimal data set.

Oftentimes your stakeholders/customers will “ask for the moon”, but you’ll have to push them to work with the small amount of information you have and get creative.

Note: As you learned in the course, being a Data Project Manager involves an extraordinary amount of collaboration. Complete the next sections based on the following scenario.

After the analysis in section 3, we made sense of the numbers, and realized the total number of events seems to be too small (this was a week's worth of data, but you need at least a month). Further investigation reveals that this was a subset of logs, but the actual data that is being collected is much bigger. Working through this small data set was tedious, and repeating this exercise on a much bigger data set manually won’t be feasible. Considering the time constraints of this project, engineering is willing to help with some automation. They also have limited bandwidth and are busy scaling systems up.

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

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

1. How many events are being recorded per day?
2. How many events of each event type per day?
3. How many events per device type per day?

4. How many events per page type per day?
5. How many events for each location per day?

For your chosen question also answer the following using the data from section 3 to support your answer:

1. How much is the customer data increasing?
2. How much is the transactional data increasing?
3. How much is the event log data increasing?

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

- Event Log Data
- Transactional Data
- Customer Data

*We would choose **event type per day**. **Event log** is this is the most important among all is because event data will help us know the behavior of user from signing up to booking a ride. We will be able to know the user funnel which is important for conversion rate. This data will be also used to marketing team as they can target Advertisement location and can have personalized marketing campaign.*

Customer data increasing: From the data we have, it is quite hard to deduce customer data is increasing or not. Since we do not know whether the data is from customer acquisition or retention. Acquiring customers will show increase in data.

Transactional Data increasing: We can clearly infer that transactional data is increasing till 11 October from looking at the analysis of "book_page" from events per page per day done in section 3. The book_page shows that transaction are increased over the period of time. Basically, on booking page how much users click on booking button would give us the event data.

Event log data increasing: The total of event page is increasing from around 8k to around 17k on 11 October. This data contains the sum of all the various events page.

The most important is the event log data because of the reasons explained above.

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.]*

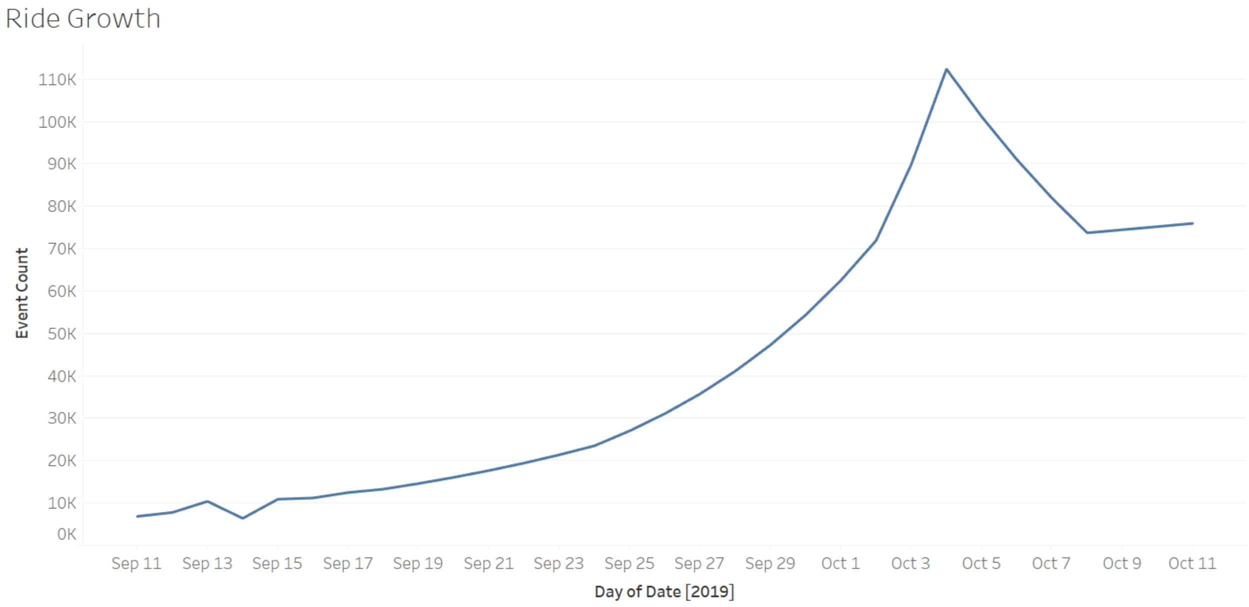
Section 6: Business Insights

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 total event data grew exponentially during the period from Sep 11 to Oct 11 from around 790k events to 12M with highest on 4 oct. Same trend is observed in Ride growth.

What is the fastest growing data and why?

growth in percentage = Number of events recorded on 11 October / number of events recorded on Sep 11

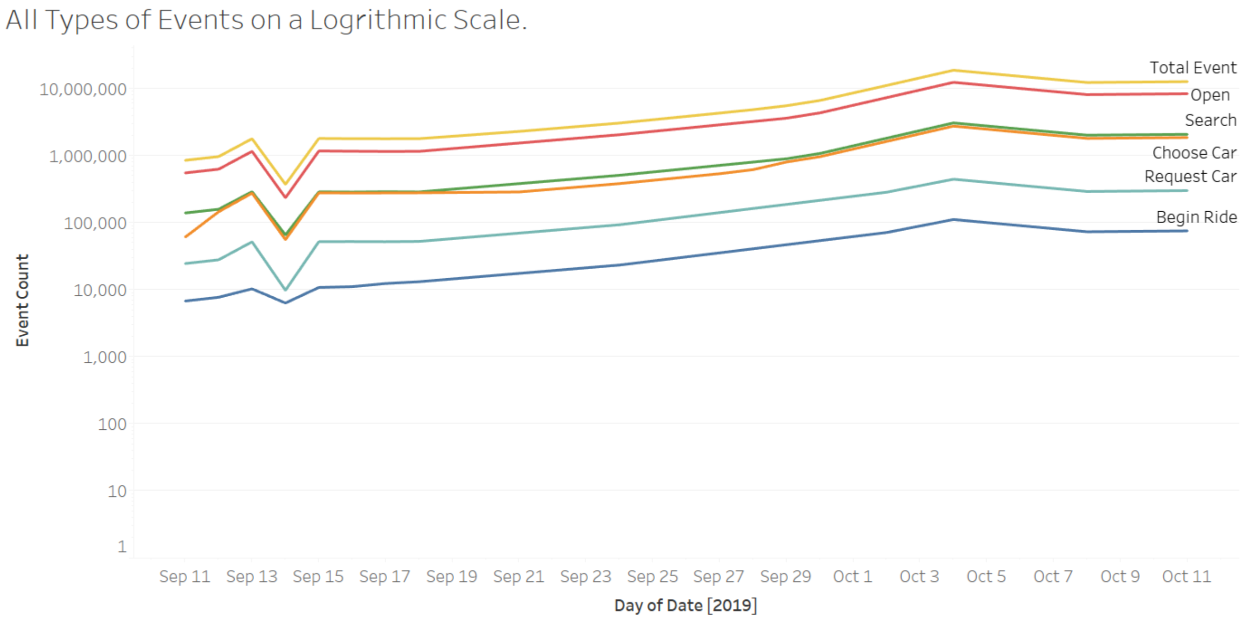
- ➔ *Begin ride = 1000%*
- ➔ *Request car = 1120%*
- ➔ *Choose car = 2960%*
- ➔ *Search = 1390%*
- ➔ *Open = 1415%*

(Figures approximated)

The fastest growing data is the total event type as it is the summation of all the events.

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

The graph pattern indicates we are growing, and the peak was recorded on 4 October. The growing sign is healthy for the company's prospect. It's a good sign that the company is able to acquire new users and retain them too. The marketing campaign done in October has been fruitful to the company has event "begin ride" as shown good growth too from Sept to October. However, the conversion rate is low and this issue has to be solved. The bounce rate is high. The relationship between data engineers/ data scientist and Marketing team is crucial as the data generated would help marketing to have personalized marketing as per user details. Data would help them with demographics, geographics and can identify target users.

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 would highly recommend using Cloud service.
Cost: *We have to pay only for the infrastructure used.*
Scalability: *Can be scaled up or downgraded whenever we need.*
In-house Expertise*: The cloud offering services have experts, we wont need to hire any experts for the same. It would save us lot of expenditure and we will be able to focus on core products, features.*
Latency: *With cloud, they have data centers and local data center region. Combining this with bandwidth and CPU latency is not an issue.*
Reliability: *The providers are reliable as they have layered security compliances. Also, they have in built backup systems.*

Suggested DWH

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

I would recommend Redshift for the following reasons.

It is horizontally scalable meaning, whenever storage has to be increased or speed has to be increased, we just need to add nodes using AWS console and it will be scaled immediately. The loading of data and analytics is super-fast in Redshift. The security comes with VPC and due to the massively parallel processing, It is more reliable. The pricing is based on demand and it is "pay as you go". They charge for each node in cluster. Therefore, as we need nodes our cost would increase.

Image Appendix

Image 1: Log Growth

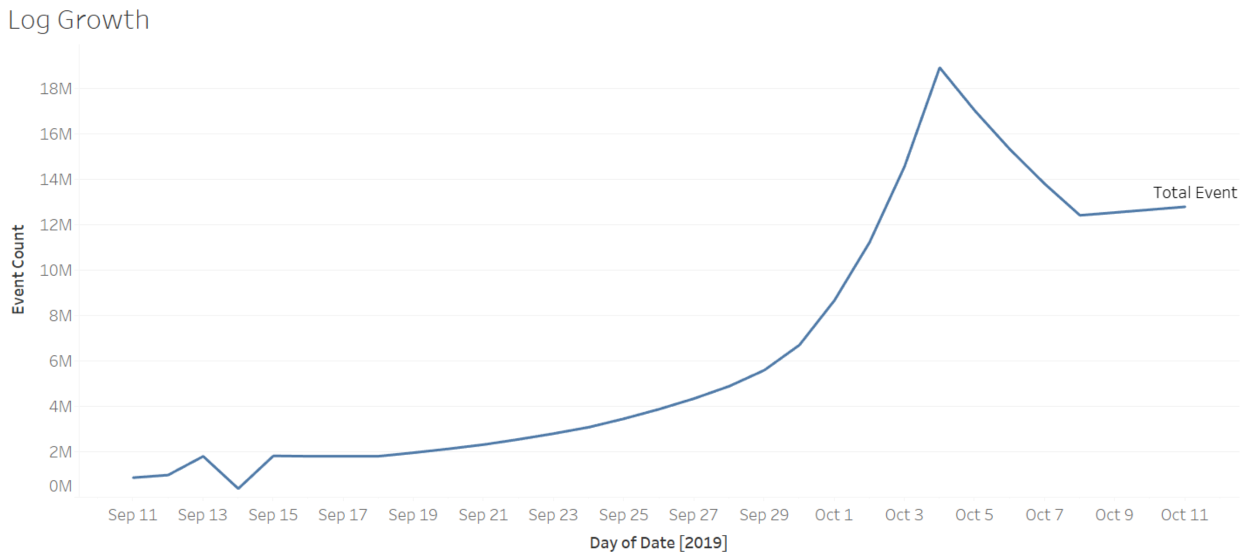


Image 2: Ride Growth

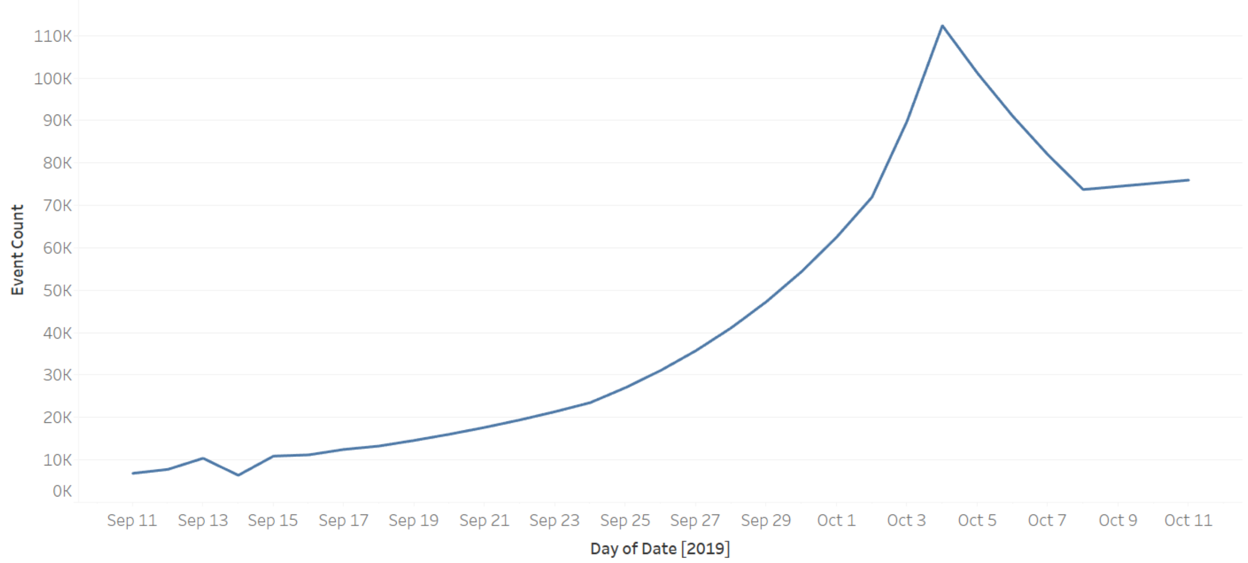

Ride Growth

Image 3: Total Event Count

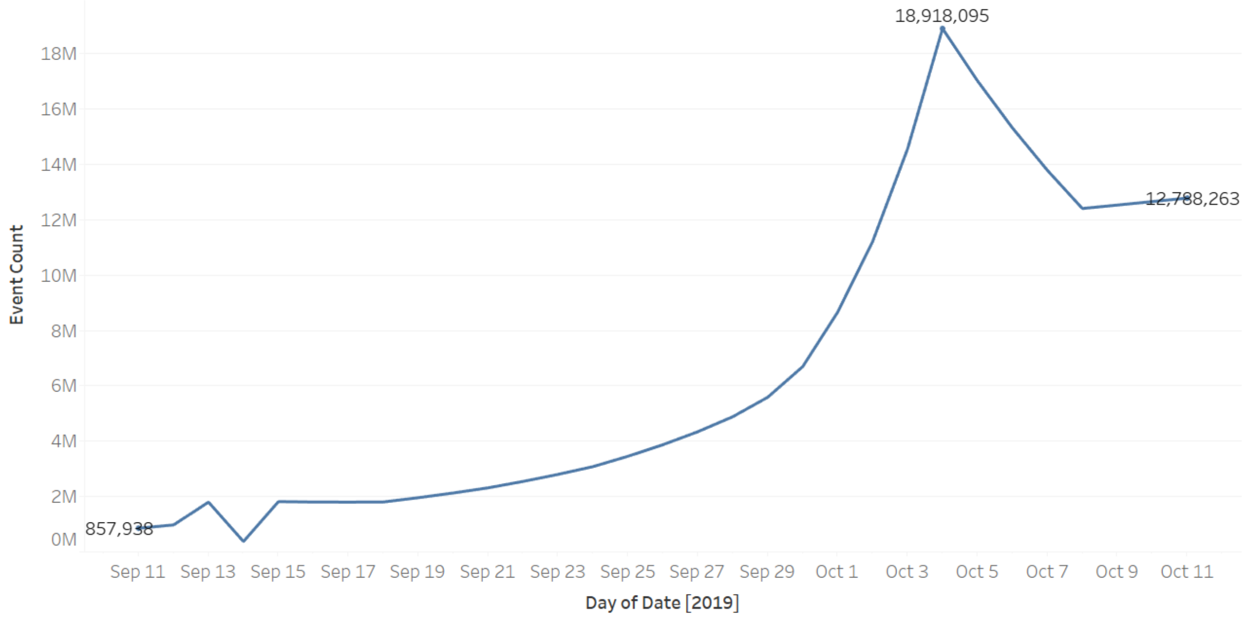
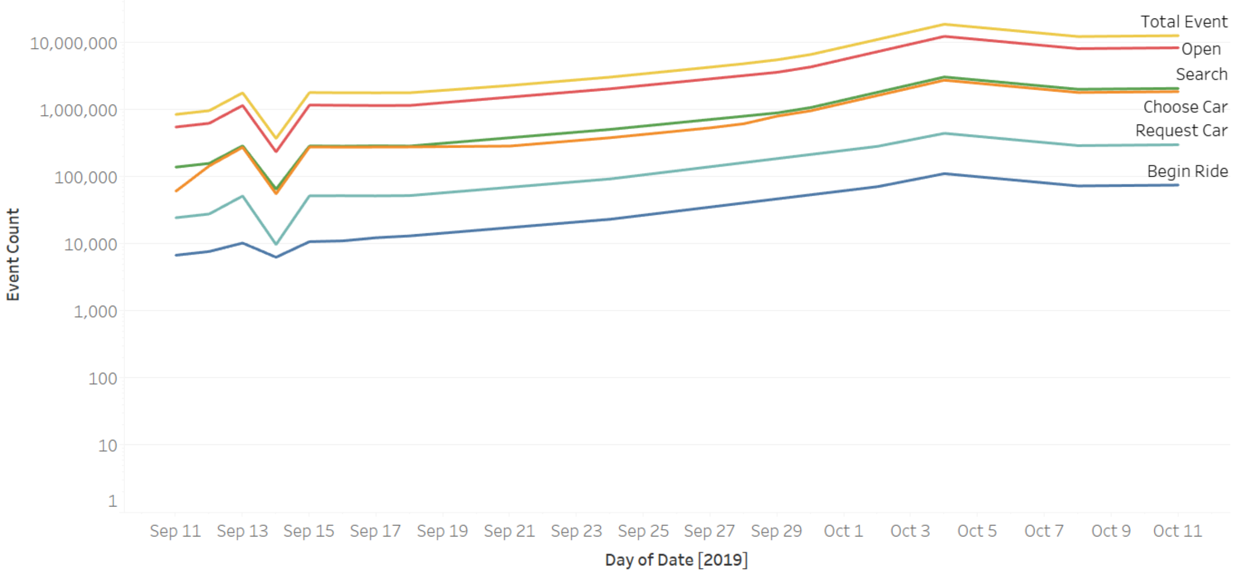

Total Event Count

Image 4: All Events Log Scale

All Types of Events on a Logrithmic Scale.


Total Event
Open
Search
Choose Car
Request Car
Begin Ride
Event Count
Sep 11
Sep 13
Sep 15
Sep 17
Sep 19
Sep 21
Sep 23
Sep 25
Sep 27
Sep 29
Oct 1
Oct 3
Oct 5
Oct 7
Oct 9
Oct 11
Day of Date [2019]