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**
Risk and security team	Focused on safety and security of customers and partners	- Dashboard - Alerts - Historic lead time data - ML engine
Accounting team	Ensuring they are financially compliant	- Expense reports - BI and visualization tool - ML engine for future managing

		prediction
Payments Team	Focused on smooth payment mechanisms for customers and partners	- Monitoring money transactions - BI tool for insights - Visualization tool

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?**
Risk and security team	- Dashboard - Alerts - Historic lead time data - ML engine	Customer: Name, contact information, license details Partner: Name, contact information, car registration details	To ensure safety and security, the risk and security team will be interested in customer and partner screening.
Accounting team	- Expense reports - BI and visualization tool - ML engine for future managing prediction	All types of coast – operations, payments to partners Revenue earned from customers	The financial compliance accounting team will maintain financial books and create accurate balance sheets.
Payments Team	- Monitoring money transactions	Account details of partners, terms of payment, amount	Managing seamless payments to partners is a use case for the Payments

	- BI tool for insights - Visualization tool	to be paid	team.

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:

Customer

Customer ID (PK)	Frist Name	*Last Name*	*Address*	*Email*	License Number

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

I just used a primary key here because we don't need information for partners, payment or car inside. It is information for a customer.

Table 2:

Partner

Partner ID (PK)	Partner Payment ID (FK)	Car ID (FK)	First Name	*Last Name*	*Address*	*Email*

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

Here we need partner payment and the car like foreign keys because we should have payment transactions and the cars used in this entity.

<u>Table 3:</u>

Partner Payment

Partner Payment ID (PK)	*Account Holder Name*	*Account Number*	*Terms of payments*

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

Just the partner payment ID is necessary for the payment table. Not need to foreign keys for this table.

<u>Table 4:</u>

Car

Car ID (PK)	Make	*Year*	*Registration Number*	*Daily Rental price (DRP)*

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

The Car ID is the primary key and we don't need any other information here because we have the car ID in the Partner table.

<u>Table 5:</u>

Booking Details

Booking ID (PK)	Customer ID (FK)	Car ID (FK)	Booking Date	Pick-up Date	Drop-off Date	*Booking Total*	*DRP*	*Taxes and fees*

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

We have Booking ID like primary key and two foreign keys (Customer ID and Car ID) to have historics for customers and car took in our system.

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 and organise them in tables for a visualization.
2. *Data Verification*
    a. Now, we check our current data type, we check for file extensions and probably the size of our dataset.
3. *Assimilate both Records and Source*
    a. Confirm that the records we have corresponds to what was recorded initially by the source.
4. *Search the duplications*
    a. Check that our records to look for duplicates and delete them, since our dataset is large we could use tools as tableau Public.

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	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

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	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.

- Working with the automated ETL process helps in time saving. Also, you will be able to initiate many process and engage automatic extraction which is normally done otherwise.
- Working with Automated ETL helps to adapt rapidly changing data. Also, we can easily handle a huge data as compared to the manual extraction that requires us to work with it bit by bit hence tedious.
- Note that manual ETL will not be suitable with a large amount of data, this is because it will require us to work on small bits of the data sequentially hence time consuming.

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

Event Log Data
The Event log data is used for the questions above. This data contains the different event types that corresponds to the activity of the different users. Furthermore, it contains data for the different months which is specifically what we need to compare the different activities between months. Obtaining the answer to the question asked can be easily done due to the use of Automated ETL which can be done more fluently on the Event Log Data. Also, with an event we can have transactions and the customes assist in this event. The event can give more data for the analysis.

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

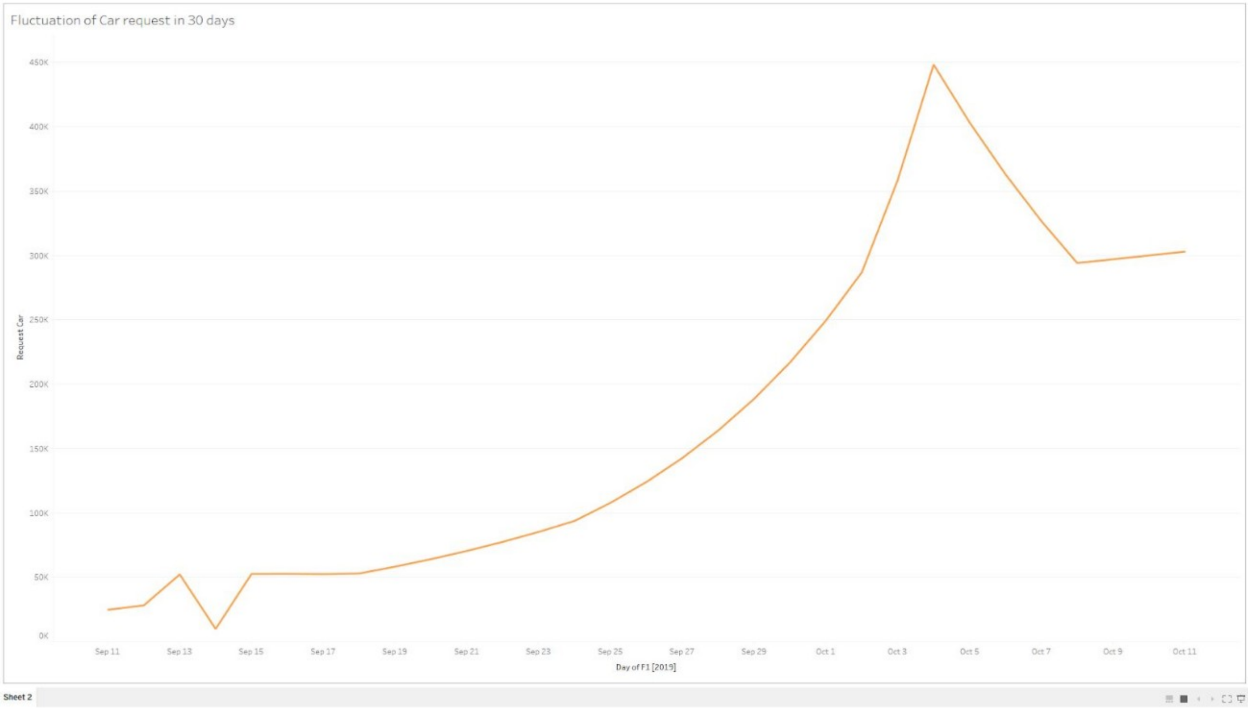
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

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.

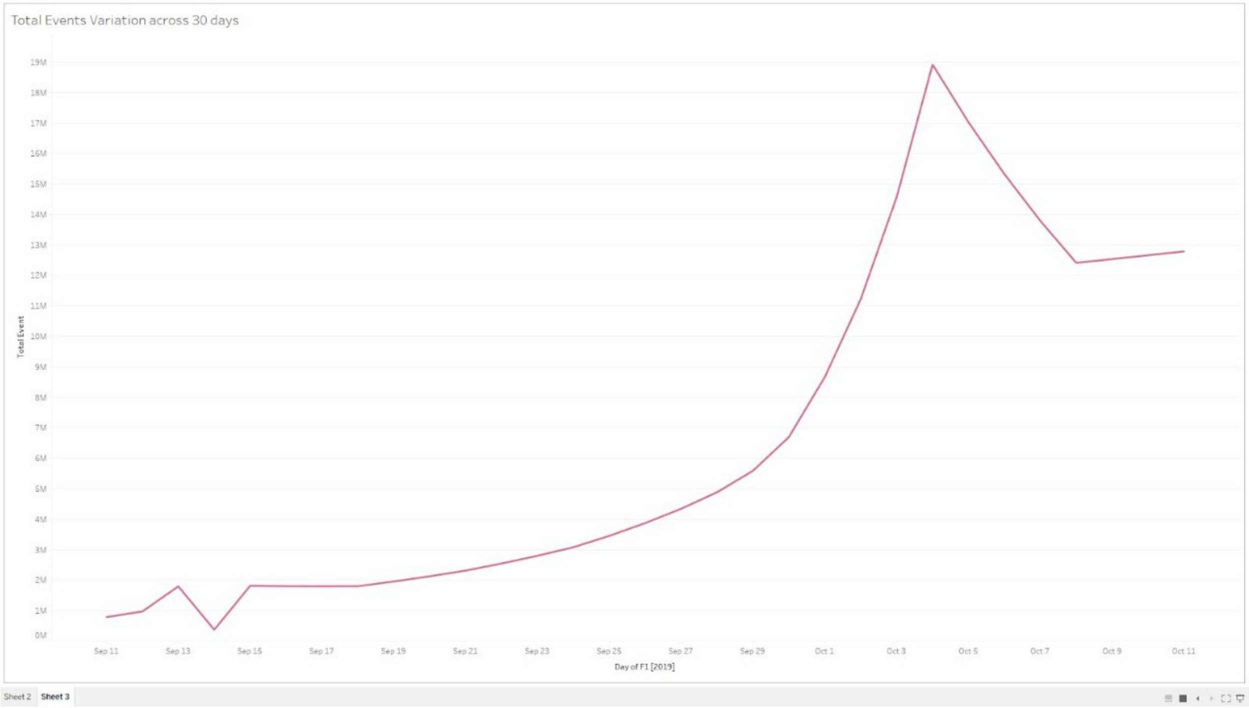
In this section, you can load and visualize the data into whatever platform you would like. A Python Notebook, Tableau or any other visualization tool you are familiar with. Create two visualizations that might help you to better understand your data trends and place either a screenshot or exported image 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:



This graph was created using the following steps:

1. *The Event Log Type data is loaded directly into tableau.*
2. *Create a new sheet this will show where our different fields lie either as a dimension or measurement.*
3. *Choose one measurement and one dimension.*
4. *Choose the line measurement visualization.*



This graph was created using the following steps:

5. *The Event Log Type data is loaded directly into tableau.*
6. *Create a new sheet this will show where our different fields lie either as a dimension or measurement.*
7. *Choose one measurement and one dimension.*
8. *Choose the line measurement visualization.*

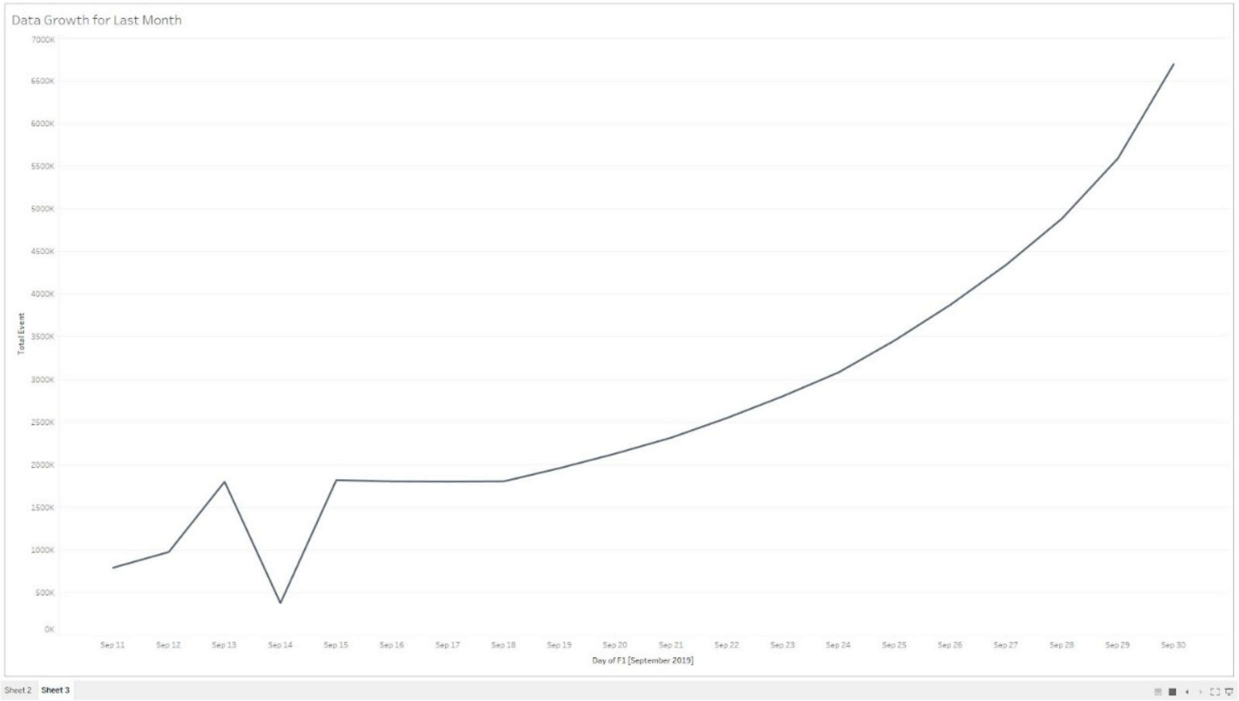
Section 6: Business Insights

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:



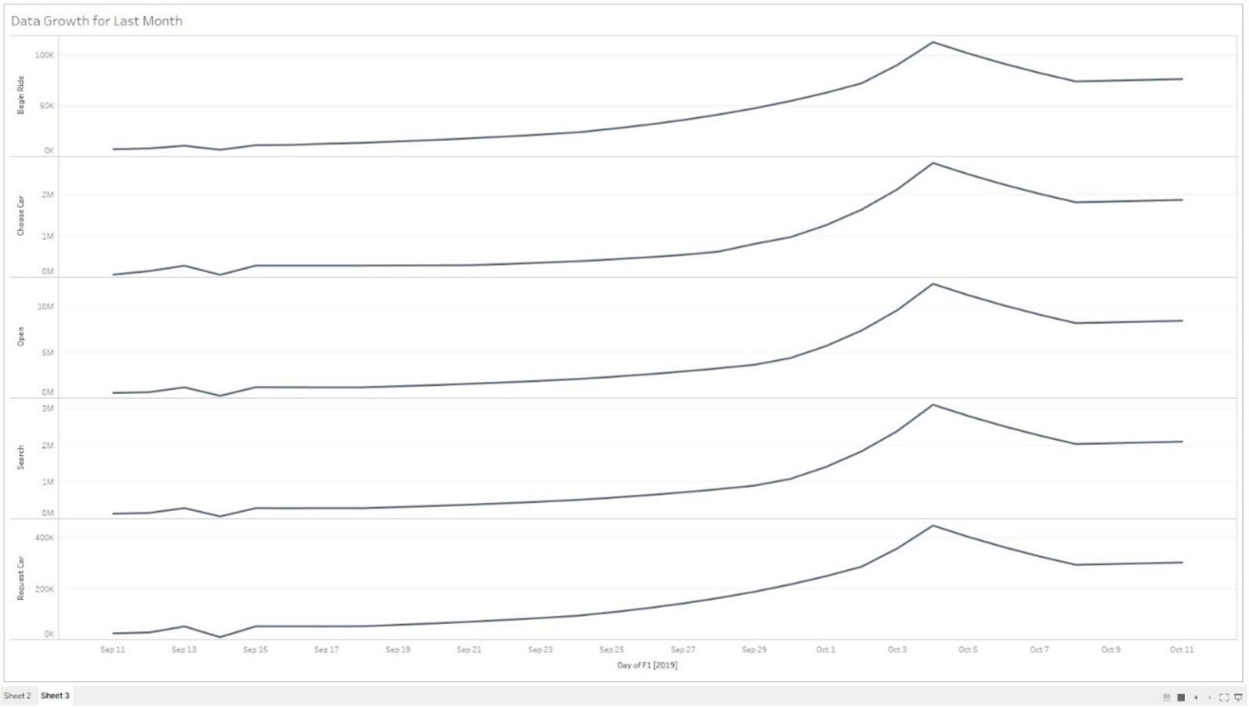
Growth (g) = percentage change in data between the two months
Growth(g) = ((Final value in Oct - Initial value in Sep) / Initial value in Sep) * 100
Initial value in Sep = 790,329
Final value in Oct = 12,788,264

Growth = ((12,788,264 - 790,329) / 790,329) * 100 = 1518%

What is the fastest growing data and why?
The fastest growing data is the total events because it regroups all of the other data sets.

All Event Type Data

Visualization:



- Graph Pattern
- Good or Bad
- October Marketing Campaign
- Marketing Campaign Impact
- Importance of Relationship Between Marketing Campaigns and Data Generation

1. Graph Pattern

The graph pattern demonstrates that the activities are very important towards the end of the month and continues towards the beginning of the month. This may be influenced by factors such as the time salaries are paid this may justify such an increase and also why there is a drop after this.

2. Good or Bad

We think this is a good pattern but it would have been best to experiment a high activity throughout the year. In the above graphs we notices a bad patterns because everything is low we are in the middle of the month but towards the end of the month things become better because we have an increase in activity.

3. October Marketing Campaign

The october marketing was quite successful because we noticed a net increase in the activity in the month of october this is seen through the graph visualization.

4. Marketing Campaign Impact

The impact is clear and obvious from the patterns we observe how there is a net increase in the activities. So the impact was good enough. Initially in the month of September the activity was relatively low but due to the campaign it increased during the month of October.

5. Importance of Relationship Between Marketing Campaigns and Data Generation
The relationship is very important because we are capable to determine whether our product is successful or not. Furthermore, this permits us to calculate our growth rate and anticipate for the upcoming months (years).

Section 7: Data Infrastructure Strategy

Thus far we have:

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

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

Data Warehouse Options:

Cloud:

- Amazon Redshift
- Google BigQuery
- Snowflake
- Microsoft Azure

On-Premise:

- Oracle Exadata
- Teradata, Vertica
- Apache
- Hadoop

You will address the following factors with a rationale as to why the DWH chosen is the best for Flyber:

- Cost
- Scalability
- In-house Expertise
- Latency/Connectivity
- Reliability

Cloud vs On-Premise

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

Flyber would be best on the Cloud than On-Premise

1. Cost

One of the main reasons being that there is reduced staff to operate also we will use less electricity because we will not be using the heavy machines to perform our computation. Moreover, we will need lesser capital cost because for On-premise we need heavy capital to install the machines.

2. Scalability

In the cloud, we will enjoy more processing power because we have several machines working. When compared to the On-premise we are limited to the local machines we have and this can be extended to other factors such as storage space. This is a huge advantage because we are no longer limited to the local constraints.

3. In-house Expertise

With the cloud services we do not need to have experts to operate the machines, an average computer guy can handle the system with ease. But compared with the on-premise where you need many experts that understand the heavy machinery and are capable of operating them.

4. Reliability

You can always get instantly updated about the changes.Most cloud providers are extremely reliable in providing their services, with many maintaining 99.99% uptime. The connection is always on and as long as workers have an Internet connection, they can get to the applications they need from practically anywhere. Some applications even work off-line.

5. Connectivity

Cloud computing allows you to deploy your service quickly in fewer clicks. This faster deployment allows you to get the resources required for your system within fewer minutes.

Suggested DWH

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

Amazon Redshift will be the most suitable cloud service provider for Flyber.

Some reasons:

1. Cost

Pay only for what you use and know how much you'll spend with predictable monthly costs. Amazon Redshift is at least 50% less expensive than all other cloud data warehouses. Scale and pay for storage and compute separately and get the optimal amount of storage and compute for diverse workloads.

2. Scalability

The flexibility and in particular, scalability, of Amazon Redshift makes it appealing to businesses of all shapes and sizes. With a few simple clicks, you can easily scale the number or type of nodes in your Redshift data warehouse to suit your capacity requirements.

3. In-house Expertise

There is little or no technical expertise required to operate the amazon redshift , everything is managed by amazon.

4. Reliability

Amazon Redshif is also one of the most reliable cloud proving services out there. This is of course due to their excellent servers and also they are available practically 99.9% of the time.

5. Connectivity

Amazon have clusters that are very efficient and reliable, they are considered to be the fastest available on the market presently.

Image Appendix

Image 1: Log Growth

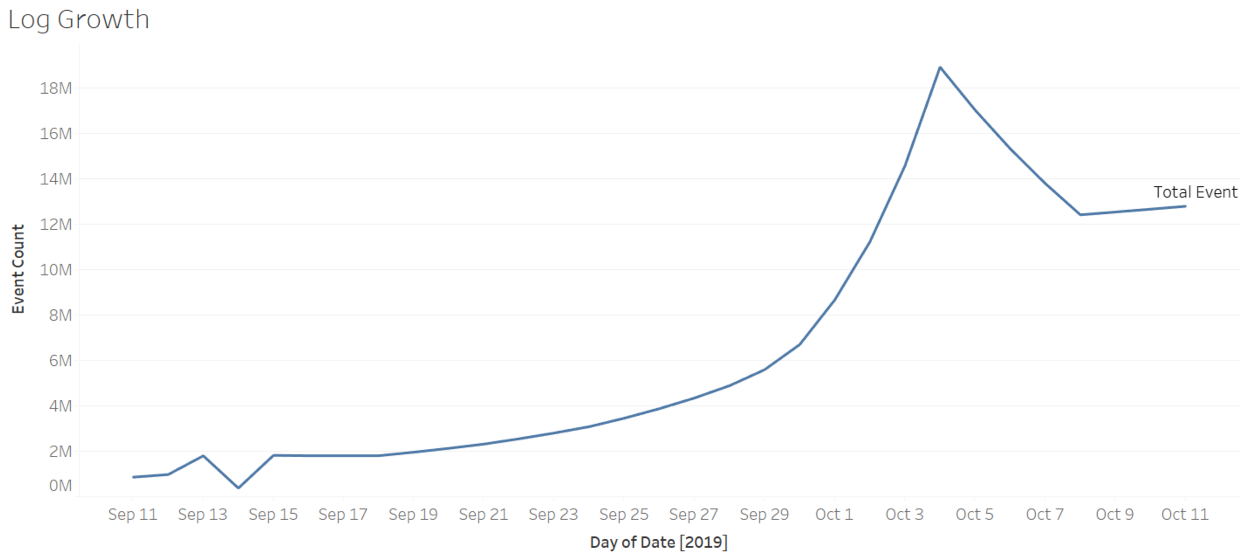


Image 2: Ride Growth

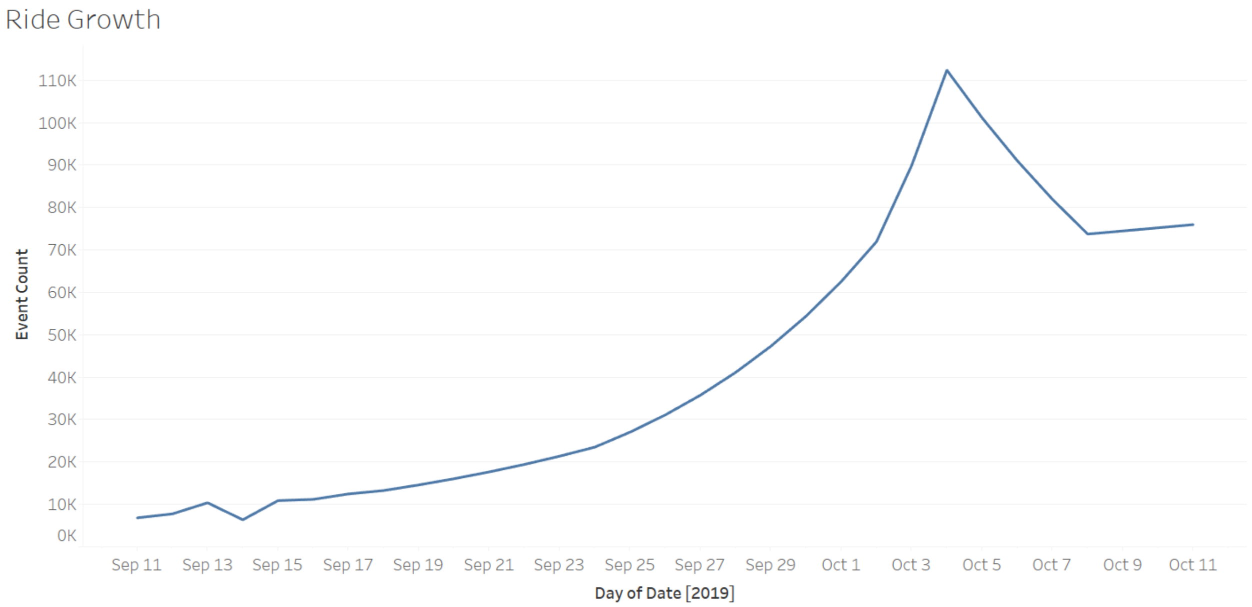


Image 3: Total Event Count

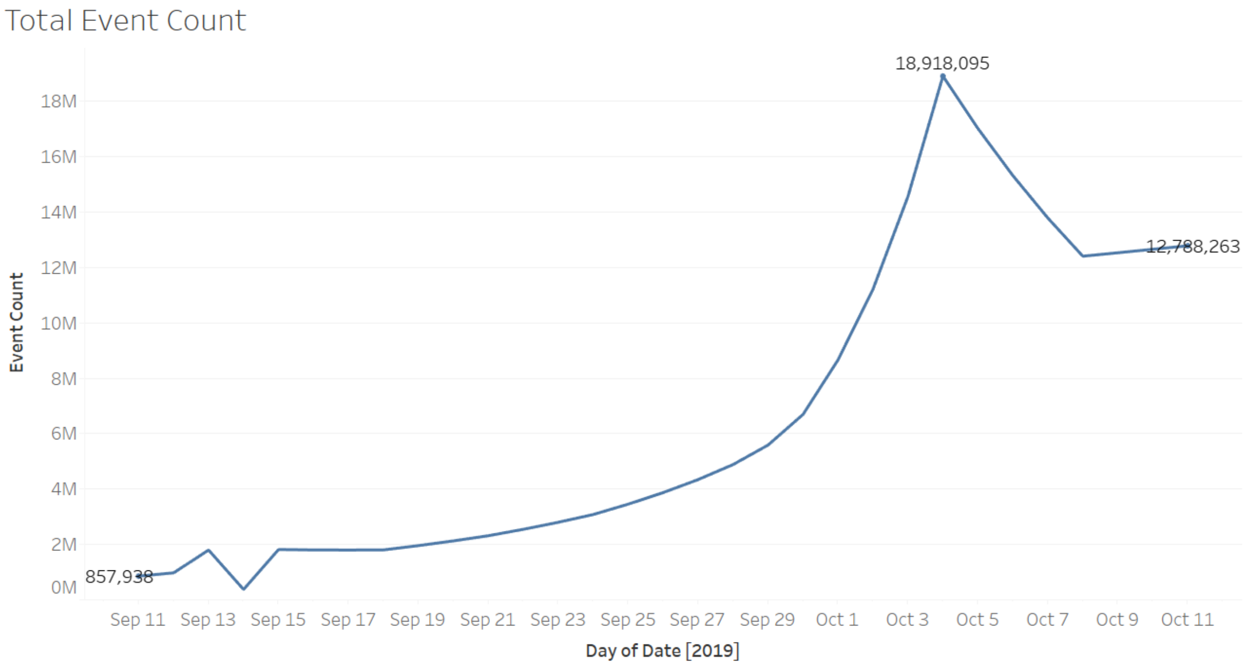
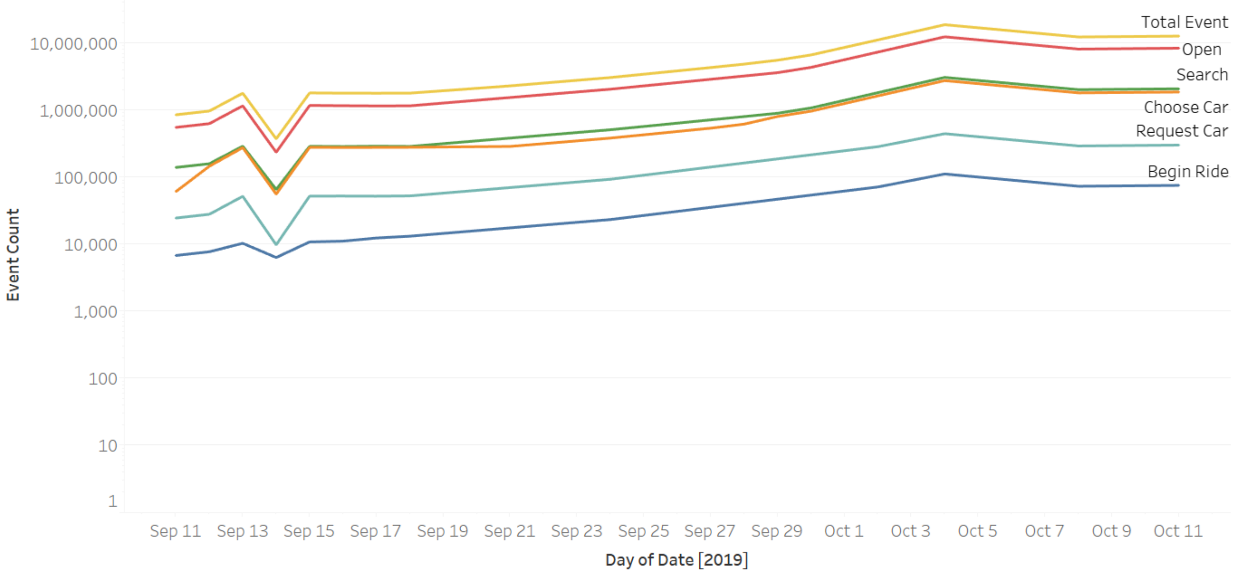


Image 4: All Events Log Scale

All Types of Events on a Logrithmic Scale.


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]
Total Event
Open
Search
Choose Car
Request Car
Begin Ride