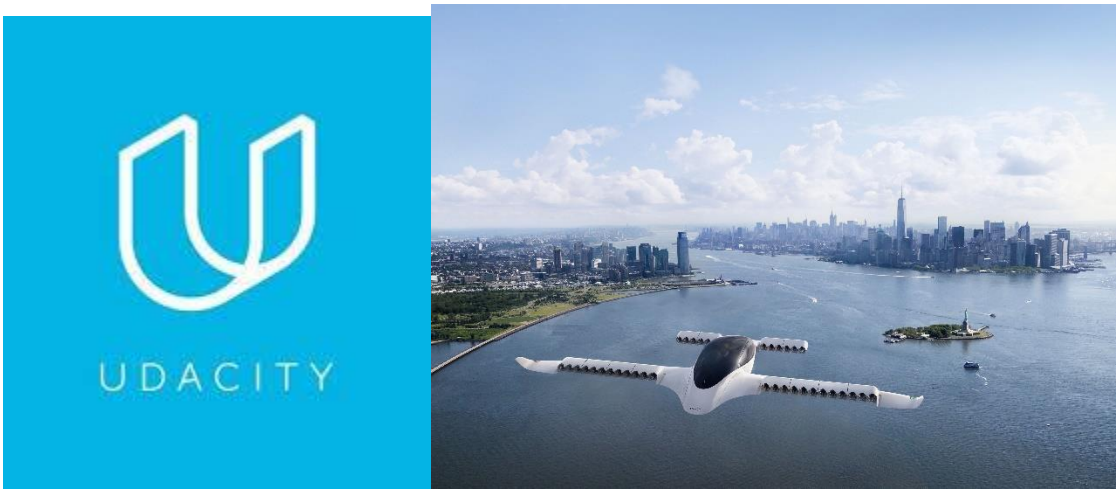
An MVP’s Data Strategy

FLYBER’s Data Infrastructure Strategies



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**
Specific Product Managers & Project Coordinators	Normally, such departments operate 'interdependently'. Such teams usually run the operational procedures. Flyber's success would highly depend on how well these teams run either 'Agile' or 'Waterfall' methodologies.	Over-all monitoring of the business' functionalities, Insights into Flyber's key Metrics, utilizes several forms of 'Information Radiators' (IR) and 'Big Visual Information Radiator' (BVIR)
Flying Taxi Dispatch Unit	A very key/ vital aspect of the operation. Ride/flight routes are managed in certain timelines. Uses a VERY SPECIFIC methodology of	Flight routes and dispatching schedules are managed in real time. Access to real-time analytics dashboard.

	both flight routing and dispatching schedules. This team has very specialized training that has similarities with flight controllers, regular taxi/ride dispatchers and drone operators.	
Human Computer Interaction (HCI) and User Experience (UX) Team	Designs all the User Interface (UI) that customers and stakeholders use. A business' success relies heavily on the over-all customer experience. This starts by a client booking a ride either via mobile app or website.	Monitor's the ongoing User Experience (UX) and Customer Experience (CX) ratings
Cybersecurity Team (includes Computer Forensics & Incident Response Team)	Flyber's survival also relies on the security and integrity of its systems (just like ALL businesses and organizations). The over-all infrastructure needs to be	Monitor's all security related incidents (using Splunk, Grafana & ElasticSearch)
Artificial Intelligence & Machine Learning Group (embedded within the Data Science & Innovation Teams)	There needs to be an ongoing research to find innovative ways of improving both product (app , software capabilities) and the service. Also needed to develop algorithms that may help in flight paths and scheduling , analyzing wait times, 'pickup and drop-off analytics'. **Both the Dispatch and HCI/UX Teams would work hand in hand with this team on certain scenarios	Creates algorithms for further product & service innovations.
Finance & Accounting	Monitoring of all finance related issues. Strong focus on the Profits & Losses (P & L). Helps ensure that the MVP with achieve profitability	Tracks the entire financial records of Flyber. Focuses on P & L.

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?**
Specific Product Managers & Project Coordinators	Over-all monitoring of Flyber's operations. A general 'bird's eye view' of user/customer needs.	ENTITY: Ride_Scheduling (e.g.pick ups, drop offs, gelocations) EVENT: Ride_ID, Time_Stamp, PickUp_TimeStamp, DropOff_TimeStamp, Price)	Flyber's success would highly depend on how well these teams run either 'Agile' or 'Waterfall' methodologies. It is also very critical to have a dedicated, specialized team that monitors if Flyber meets the user needs.
Flying Taxi Dispatch Unit	Flight routes and dispatching schedules are managed in real time. Access to real-time analytics dashboard.	ENTITY: Choose Car , Search, Begin ride, Request Car, Begin Ride Event: ID, timestamp, type, geolocation	A very key/ vital aspect of the operation. Ride/flight routes are managed in certain timelines. Uses a VERY SPECIFIC methodology of both flight routing and dispatching schedules. This team has very specialized training that has similarities with flight controllers, regular taxi/ride dispatchers and drone operators.
Human Computer Interaction (HCI) and User Experience (UX) Team	Monitor's the ongoing User Experience (UX) and Customer Experience (CX) ratings	EVENT: Search, Choose Car	Designs all the User Interface (UI) that customers and stakeholders use. A business' success relies heavily on the over-all customer experience. This starts by a client booking a ride either via mobile app or website.
Cybersecurity Team (includes Computer Forensics & Incident Response Team)	Keeps track of the over-all resiliency & security of the entire platform.	EVENT: Choose Car, Open, Begin Ride, Request Car, Total Event	Flyber's survival also relies on the security and integrity of its systems (just like ALL businesses and organizations). The over-all infrastructure needs to be

Artificial Intelligence & Machine Learning Group (embedded within the Data Science & Innovation Teams)	Responsible for technology innovation. AI & ML would be at the forefront of cutting edge R&D	EVENT: customer feedbacks, incidents, electrical and mechanical issues, analysis of KPIs	All upcoming businesses these days rely heavily on technology & data. Analyzing a huge repository of data needs high level computing power & skill sets.
Finance & Accounting	Tracks the entire financial records of Flyber. Focuses on P & L sheet.	ENTITY: Revenue, Profit, Costs	Any business needs to have their entire financial activities to be monitored properly. This is especially true for a technology startup that is pioneering a service.

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.)

Customer_Id	***Registration_id***	***Last_Name***	**First_Name**	**email**	***mobile***	***Home_address***	***D.O.B.***

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

Customer_ID is a **unique identifier**. This has been a 'de facto' standard among majority of businesses. Also, it becomes a LOT easier to use it than to formulate another name. Customer ID also can act to easily differentiate more than 2 customers who have exactly the same names and almost similar addresses. **Registration_ID** becomes a **foreign key**

Table 2:

Vehicle

Vehicle_id	*Vehicle_type*	*Serial_number*	*Max_load*	*Max_passengers*	*Charging_time*	*Car_load*

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

Vehicle_Id is unique*. It is synonymous to a license plate per state or province. Therefore it is a unique identifier. There is no foreign key in this table.*

Table 3:

Ride Reservations

Reservations_id	***Fk_membserhip_id***	*Pick_up_datetime*	*Drop_off_datetime*	*Drop_off_location*	***Fk_vehicle_id***

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

*The **reservations_id is a Unique Identifier.** Both the fk_membership_id and the fk_vehicle_id are referencing the tables that are highly related to reservations. Ride reservations are made via the memberships table. Therefore the PK of the memberships table (**membership_id**) becomes **FK_membership_id** for this table.*

Table 4:

Memberships

membership_id	*Fk_customer_id*	*membership_rates*	*Date_start*	*Date_End*	*Membership_type*

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

Membership_Id is unique*. It is synonymous to either a license plate (as previously mentioned) or social insurance number. But it is also good to note that a unique customer may have more than 1 subscription. Hence this is the reason why I've decided to add customer_id as a **FK_customer_id** .*

Table 5:

Method of Payment

Payment_method_id	***Fk_customer_id***	***Fk_membership_id***	*Payment-type*	*Automated_blockain_ID*

Rationale for Choosing Primary and Foreign Keys for the Table 5:

Payment_Id is unique*. It is synonymous to either a license plate (as previously mentioned) or social insurance number. But it is also good to note that a unique customer may have more than 1 subscription. Hence this is the reason why I've decided to add customer_id as a* ***FK_customer_id*** *. It is also good to note that* ***customer_id*** *is the PK of the* ***customers*** *table. A single customer may have more than 1 specific type or method of payment, just like a customer can have more than 1 type of membership.*

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 Gathering & Analysis*
    a. *Raw data was being collected and later organized for visualization.*
    b. *For simplicity and ease of use, I have decided to use Pivot Tables. Simple & direct to the point.*
2. *Data Verification*
    a. *Checking the current data type is vital in this part. Checking file for possible extensions is necessary*
3. *Source & Records are being assimilated*
    a. *Confirm that records are properly corresponding to the specific records*
4. *Search for Data Duplications*

a. *Verify that there are no duplications among the dataset. This violates 'referential integrity'*

5. *Data Visualization*
   a. *A visual representation of data is a necessity. It is being presented in order for easy 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	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

Count of event_uuid	Column Labels					
Row Labels	begin_ride	choose_car	open	request_car	search	Grand Total
05-Oct	38	1498	6594	277	1484	9891
06-Oct	49	2843	11733	540	2891	18056
07-Oct	62	2953	11767	596	2824	18202
08-Oct	86	2769	11662	547	2899	17963
09-Oct	57	2725	11531	538	2749	17600
10-Oct	57	2801	11325	607	2904	17694
11-Oct	78	2804	11371	521	2821	17595
12-Oct	18	1301	5133	220	1307	7979
Grand Total	445	19694	81116	3846	19879	124980

** The values for both tables 1 and 2 were derived from the 'pivoted table' above.

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	89-3	8764	8764	8646

device_type	ios	
Row Labels	**Count of event_uuid**	**Count of event_type**
⊞05-Oct	**2384**	**2384**
⊞06-Oct	**4337**	**4337**
⊞07-Oct	**4217**	**4217**
⊞08-Oct	**4373**	**4373**
⊞09-Oct	**4380**	**4380**
⊞10-Oct	**4482**	**4482**
⊞11-Oct	**4500**	**4500**
⊞12-Oct	**2026**	**2026**
Grand Total	**30699**	**30699**

device_type	android	
Row Labels	**Count of event_uuid**	**Count of event_type**
⊞05-Oct	**1463**	**1463**
⊞06-Oct	**2870**	**2870**
⊞07-Oct	**2854**	**2854**
⊞08-Oct	**2729**	**2729**
⊞09-Oct	**2744**	**2744**
⊞10-Oct	**2562**	**2562**
⊞11-Oct	**2672**	**2672**
⊞12-Oct	**1231**	**1231**
Grand Total	**19125**	**19125**

device_type	mobile_web	
Row Labels	**Count of event_uuid**	**Count of event_type**
05-Oct	**5149**	**5149**
06-Oct	**8842**	**8842**
07-Oct	**9531**	**9531**
08-Oct	**8903**	**8903**
09-Oct	**8764**	**8764**
10-Oct	**8784**	**8784**
11-Oct	**8646**	**8646**
12-Oct	**4040**	**4040**
Grand Total	**62659**	**62659**

device_type	desktop_web	
Row Labels	**Count of event_uuid**	**Count of event_type**
05-Oct	**895**	**895**
06-Oct	**2007**	**2007**
07-Oct	**1600**	**1600**
08-Oct	**1958**	**1958**
09-Oct	**1712**	**1712**
10-Oct	**1866**	**1866**
11-Oct	**1777**	**1777**
12-Oct	**682**	**682**
Grand Total	**12497**	**12497**

**The above 4 ‘pivoted tables’ were utilized in answering the entire table 3.

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

Count of event_uuid	Column Labels				
Row Labels	book_page	driver_page	search_page	splash_page	Grand Total
05-Oct	1977	965	3995	2954	9891
06-Oct	3548	1823	7219	5466	18056
07-Oct	3576	1871	7307	5448	18202
08-Oct	3572	1794	7221	5376	17963
09-Oct	3586	1755	6979	5280	17600
10-Oct	3424	1689	7201	5380	17694
11-Oct	3506	1768	7137	5184	17595
12-Oct	1639	801	3174	2365	7979
Grand Total	**24828**	**12466**	**50233**	**37453**	**124980**

** The answers from Table 4 were derived from the 'pivoted table' above

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

Count of event_uuid	Column Labels					
Row Labels	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Grand Total
05-Oct	250	2009	6869	595	168	9891
06-Oct	533	3737	12591	842	353	18056
07-Oct	507	3590	12807	905	393	18202
08-Oct	469	4025	12180	893	396	17963
09-Oct	510	3440	12270	1026	354	17600
10-Oct	394	3400	12371	1069	460	17694
11-Oct	558	3556	12201	936	344	17595
12-Oct	231	1594	5580	386	188	7979
Grand Total	**3452**	**25351**	**86869**	**6652**	**2656**	**124980**

** The answers for table 5 were derived from the 'pivoted table' pasted above.

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.]

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

QUESTION Being Chosen:

How many events per device type per day?

Count of event_uuid	Column Labels				
Row Labels	android	desktop_web	ios	mobile_web	Grand Total
05-Oct	1463	895	2384	5149	9891
06-Oct	2870	2007	4337	8842	18056
07-Oct	2854	1600	4217	9531	18202
08-Oct	2729	1958	4373	8903	17963
09-Oct	2744	1712	4380	8764	17600
10-Oct	2562	1866	4482	8784	17694
11-Oct	2672	1777	4500	8646	17595
12-Oct	1231	682	2026	4040	7979
Grand Total	19125	12497	30699	62659	124980

For Android devices, we have 19,125. If excluding Oct 12th , we have 17,894
For desktop devices, we have 12,497. If excluding Oct 12th, we have 11,815
For ios devices, we have 30,699. If excluding Oct 12th, we have 28,673
For mobile devices, we have 62,659. If excluding oct 12th, we have

**Oct 12 appears to have been taken out in the initial questions hence I gave the option of excluding it.

Follow Up Questions:

1. How much is the customer data increasing?
   Answer: There has been a steady increase with most of the data within the devices. By far, IOS has the most steady increase.

2. How much is the transactional data increasing?
   Answer :
3. How much is the event log data increasing?

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

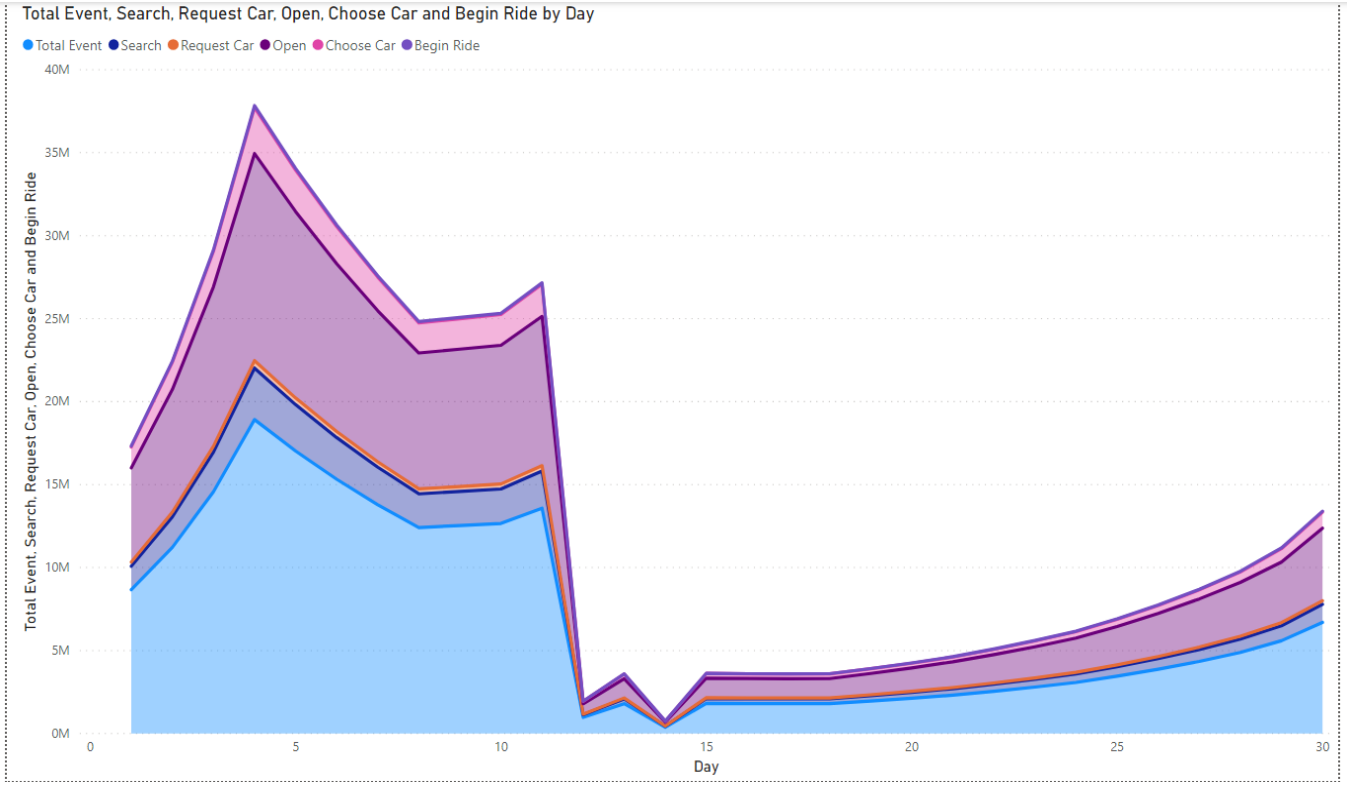
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

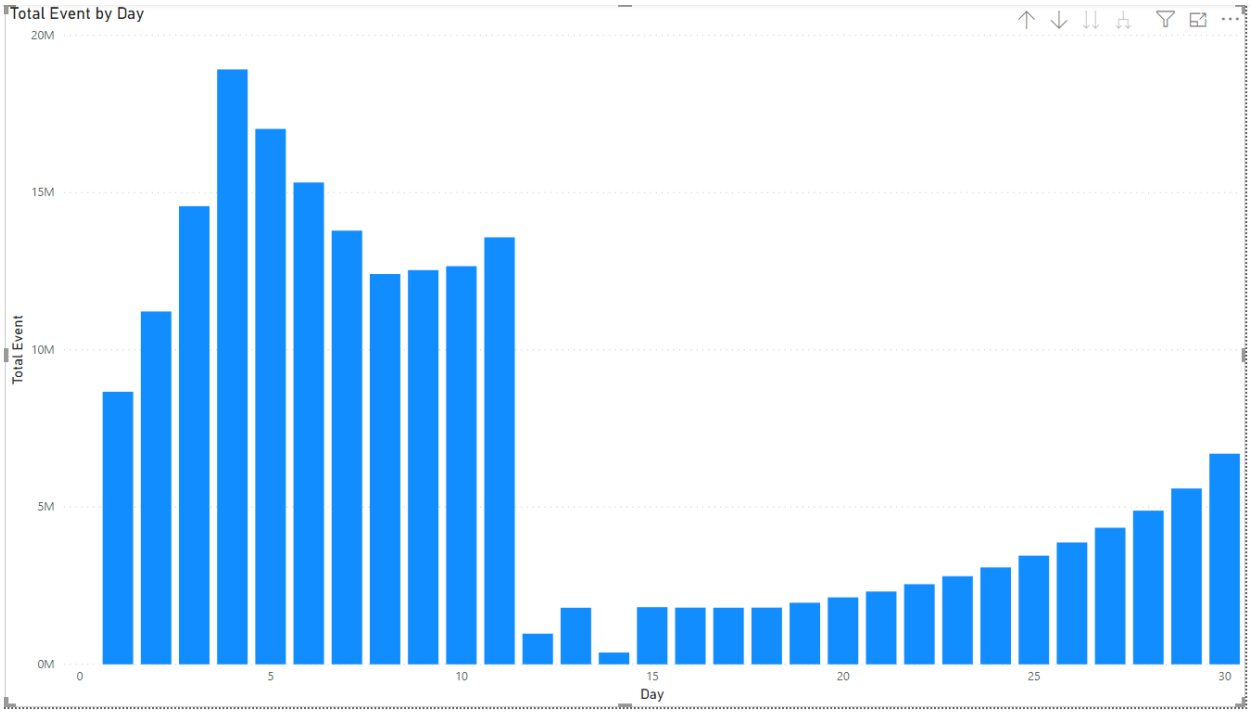
After sharing your criterion with engineering, they give you a new set of data: Section 5 Event Type Log also available in the classroom resources. Also provided in the project resources section.

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: In this optional section, I opted to use Microsoft's Power BI Platform.





Data Story: This graph tells us:

The 2 graphs simply shows that on day 4, there was a very heavy presence of activity. There was a total of 18,918,096 events that occurred that day. This was immediately followed by Day 5 with 17,026,286 events. Day 6 was a strong 3rd with a total of 15,323,656 events.
Interestingly, Day 3 was the 4th highest with 14,569,632 events.

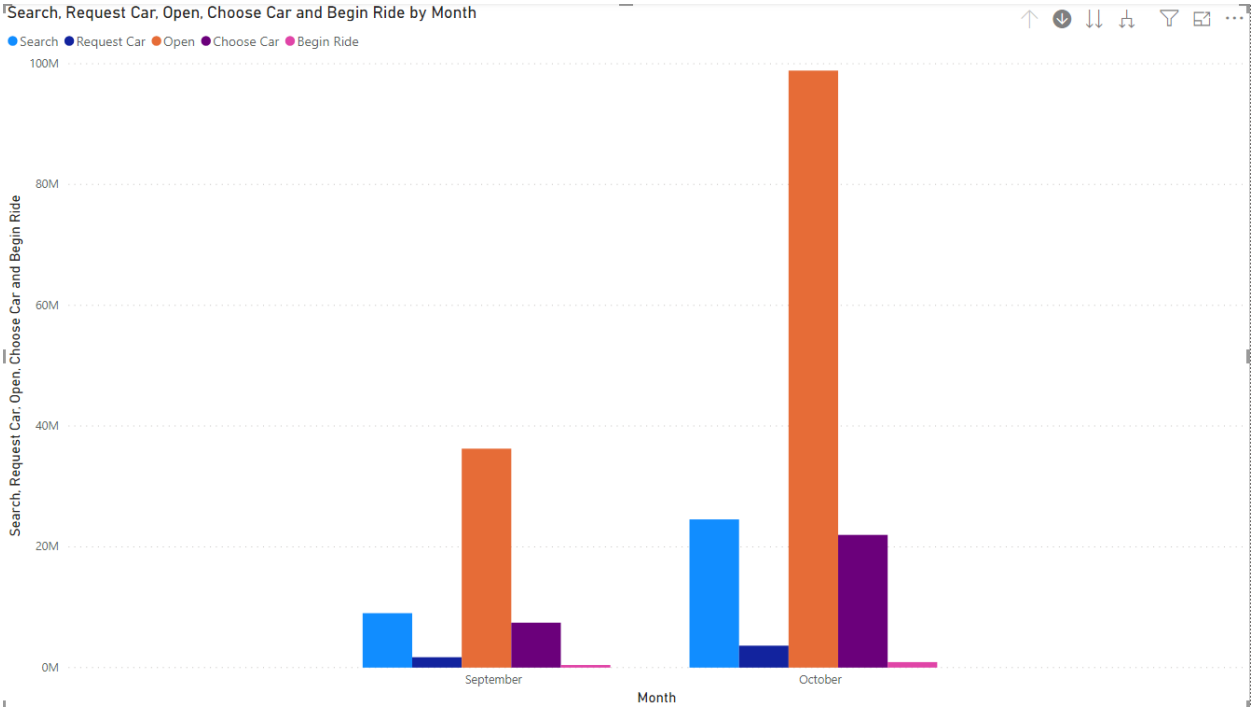
From this, we could infer that there was an 'Initial Burst' of interest in Flyber's services.

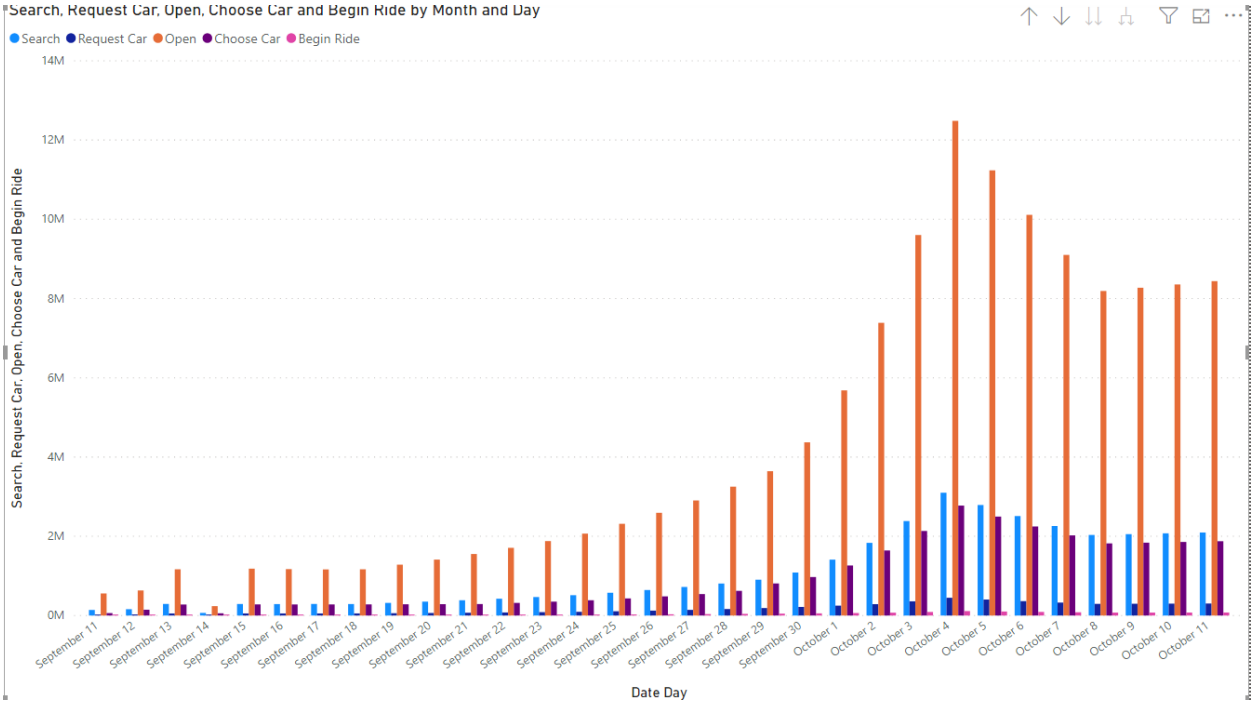
The graph also clearly shows that after Day 11, it drastically plummeted and barely recovered in over-all events until the 30th day.

This graph was created using the following steps:

1. *Raw Data set was loaded into Power BI Platform*
2. *Chose the entire set of activities/events that were present in the dataset*
3. *The top 2 simplest forms of visualizations were chosen. These were bar charts and area chart.*

Visualization 2:





Data Story: This graph tells us:

The first viz was a 'high level' overview of the 2 month comparison: October has was more activity / events than September. The 2nd image shows a deeper level of granularity. This further gives justification to the first 2 visualizations that October 4th was indeed the BUSIEST DAY ever!

This graph was created using the following steps:

1. *Loaded Raw Dataset into Power BI Platform*
2. *Once loaded, I decided to choose bar charts. I utilized all the events in the dataset*
3. *I first focused on a month by month comparison (September & October)*
4. *I then decided to have a deeper look into granularities. It then proved the exact insights in the first 2 visualizations.*

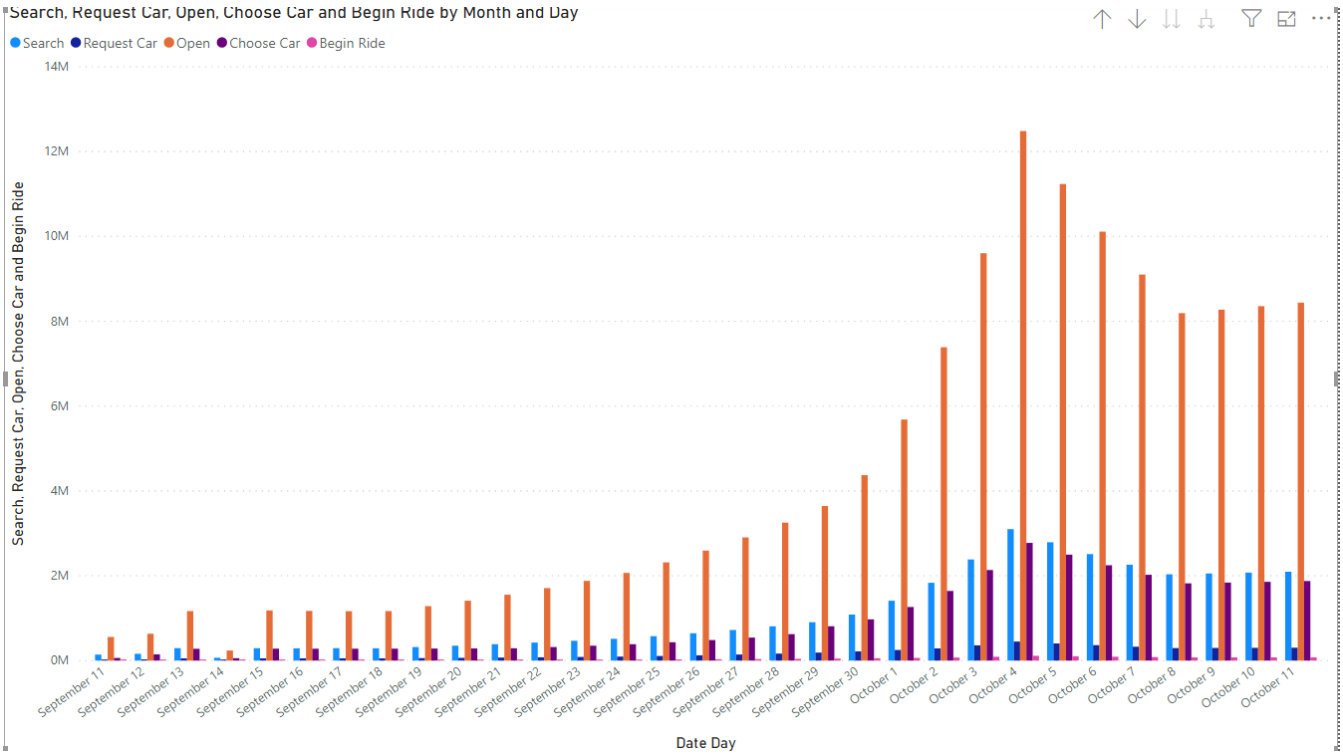
Section 6: Business Insights

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

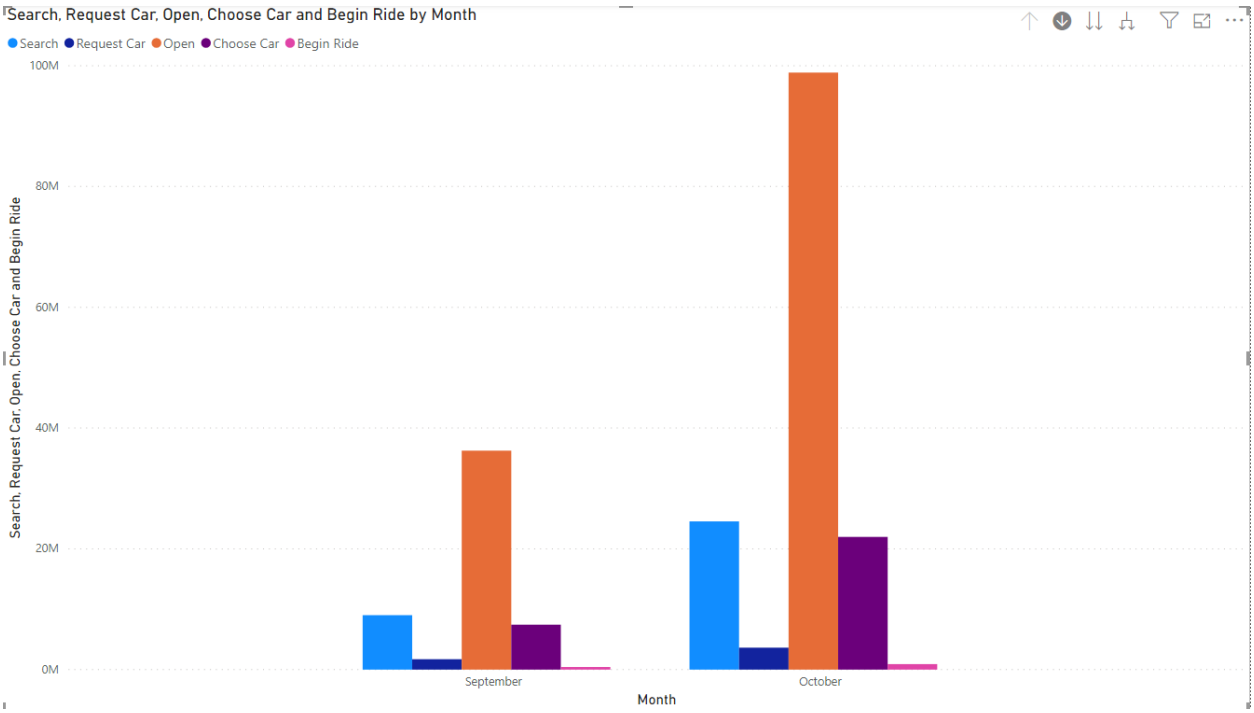
Visualization:



This data viz displays the EXACT insight that

September did not have as much events as October. From October 3rd until October 7th, they seem to have peaked. It would be very interesting to note on what triggered such events within these 4 days. Was it our marketing & advertising? Was it the presence of good reviews? Was there any downtown NYC event that made the public try a flying taxi service?

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



What is the fastest growing data and why?

Based on the findings from Section 5, the month of October was way busier than September. The first week of October showed tremendous results. I'd also like to review the data gathered from Section 3:

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

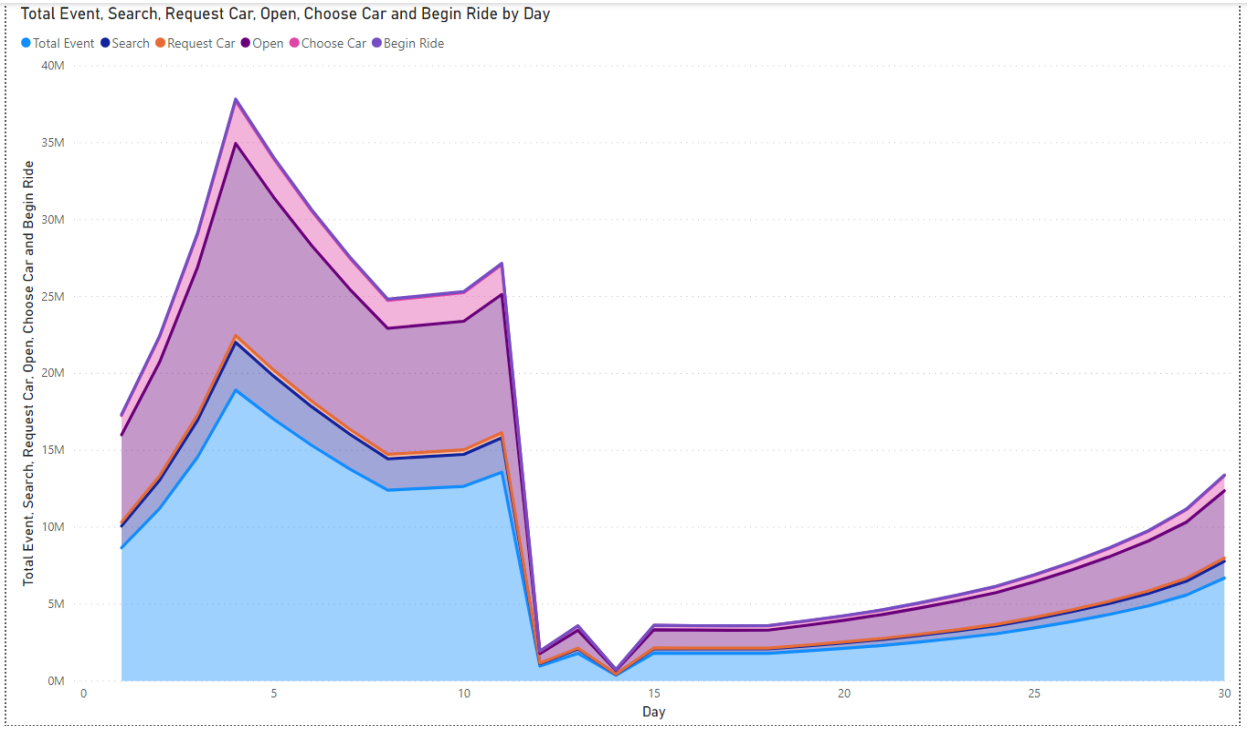
Count of event_uuid	Column Labels					
Row Labels	begin_ride	choose_car	open	request_car	search	Grand Total
05-Oct	38	1498	6594	277	1484	9891
06-Oct	49	2843	11733	540	2891	18056
07-Oct	62	2953	11767	596	2824	18202
08-Oct	86	2769	11662	547	2899	17963
09-Oct	57	2725	11531	538	2749	17600
10-Oct	57	2801	11325	607	2904	17694
11-Oct	78	2804	11371	521	2821	17595
12-Oct	18	1301	5133	220	1307	7979
Grand Total	445	19694	81116	3846	19879	124980

** The values for both tables 1 and 2 were derived from the 'pivoted table' above

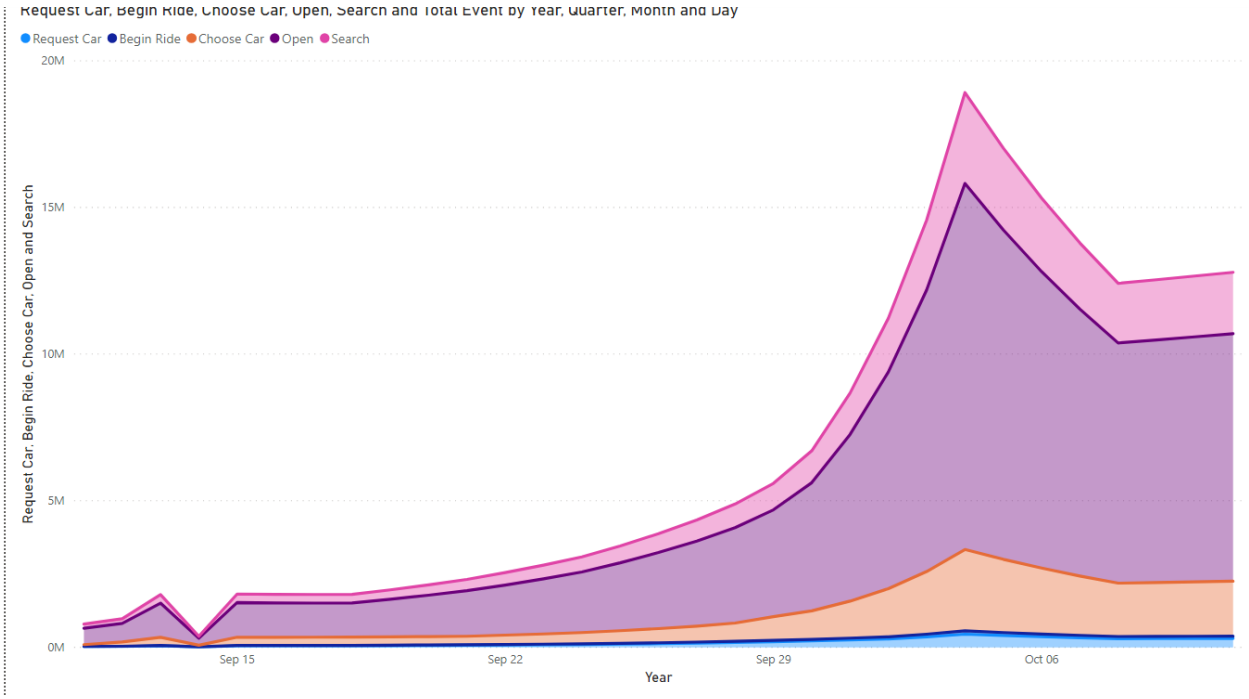
'searching' and 'choosing' a car were having very high totals. 'Search' had 19,879 in a 8 day range. 'Choose_car' had 19,694 in this 8 day range.

All Event Type Data

Visualization:



The above graph shows ALL EVENTS that occurred on a ‘day’ type granularity.



The above graph shows ALL EVENTS on a ‘Quarterly’ level of granularity.

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 2 above visualizations shows that our graph pattern states there is a huge spike in interest in Flyber’s services. Although it slowed down after the 8thOctober, there still appeared to be steady interest.

This tells us that the Marketing efforts for October were really effective. There is a huge possibility that our marketing team ‘PLAYED ALL CARDS’ during the first 12 days of October hence the interest went downwards after the 1st week.

All the visual analytics displayed greatly proved that there is a very vital relationship between creating marketing campaigns and data. The data that was being generated clearly shows how visual data could guide our over-all marketing strategies. Also, the data visualizations gives us an ‘unbiased’ view of what the results actually were.

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.

Even if Flyber is a cutting edge, new technology, it would be great to 'emulate and simulate' those companies have used. Based on several studies, both Uber and Lyft have been utilizing Both On Premise and Cloud Computing datawarehouse resources. These are mainly Apache Spark and Hadoop for Uber and AWS's DynamoDB, Kinesis, Redshift and EC2.

Since Uber has been having tremendous success and has majority of the market, I would highly suggest to use the exact technologies that they use: ON PREM – Apache Spark and Hadoop.

Suggested DWH

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

Apache Spark and Hadoop Are The DWH of choice

Despite having Horizontal Scalability being put in place, there would still appear to be some issues. A business like Flyber operates in real time, just like it's 'ancestors' like Uber and Lyft, and also any drone or UAV companies. Our company relies on data that is as fresh as possible.

A sample of Uber's data platform could be seen below:

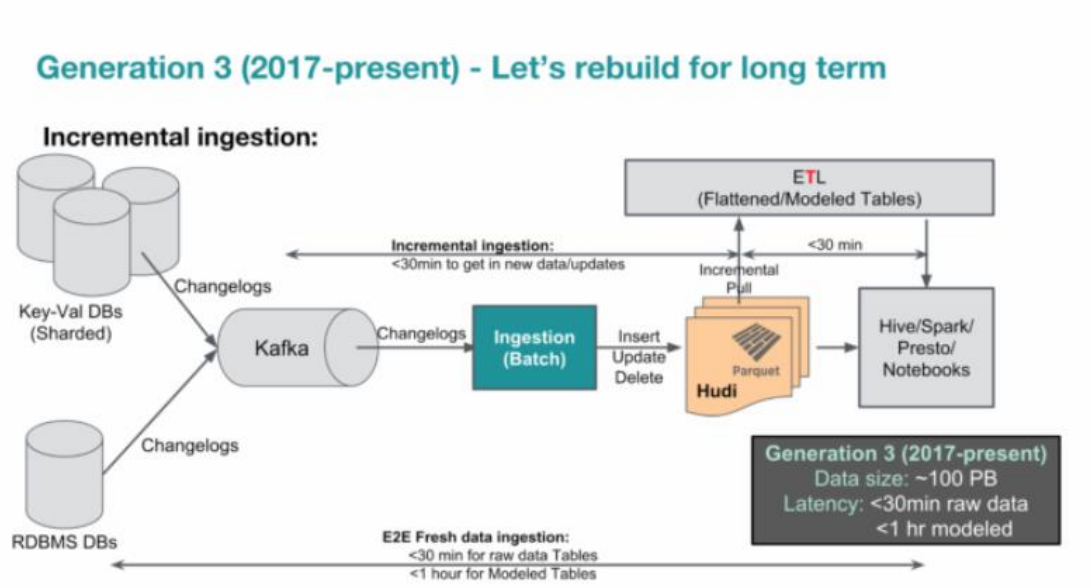


Image courtesy - https://eng.uber.com/uber-big-data-platform/

Flyber has various teams of specialists. Majority of which would use big data in order to forecast rider demands, fraud detection, geospatial analysis, geospatial computation and addressing specific bottlenecks.

Some of the key aspects that Flyber wants to address in using such technologies are:

1. *Incremental Processing via "small" batches*
2. *Increased Efficiency*
3. *Having Real Time Data Analytics*
4. *Hyper Cluster Utilization*
5. *Having Datawarehouse Efficiency*

Taken into key considerations were:
1.Calculating Costs and Utility
2.Operational Cost Reductions
3.Measuring Operational Efficiency

Appendix:
Sources for Cloud Computing & Data Warehouse resources:

Data Center Dynamics : Uber uses AWS and GCP
https://www.datacenterdynamics.com/en/opinions/buying-cloud-scale-lessons-lyft-and-uber/#:~:text=Uber%20has%20argued%20it%20can,m%20a%20year%20for%20Google%20Maps.

Less is More: Uber
https://eng.uber.com/data-warehouse-efficiency/

Uber VS Lyft : How Rivals Approach Cloud Computing, AI & Machine Learning
https://www.zdnet.com/article/uber-vs-lyft-how-the-rivals-approach-cloud-ai-machine-learning/

Datawarehousing By LYFT
https://www.skillsire.com/read-blog/241_data-warehousing-in-lyft.html

Why Serverless is the Uber Infrastrtucture
https://thenewstack.io/why-serverless-is-the-uber-of-infrastructure/

Uber's Data Framework
https://events.static.linuxfound.org/sites/events/files/slides/Apache-Data-Uber-Mayank-Bansal.pdf

Uber Uses Vertica as a DWI
https://eng.uber.com/uber-big-data-platform/

How Uber Uses Spark & Hadoop to Optimize Customer Experience
https://www.datanami.com/2015/10/05/how-uber-uses-spark-and-hadoop-to-optimize-customer-experience/

How LYFT Built its Business Using AWS
http://www.enterpriseappstoday.com/business-intelligence/how-lyft-built-app-business-on-aws.html

Uber's Evolution of Data
https://www.infoq.com/news/2018/11/uber-big-data-evolution/

Uber's Case on Incremental Processing in Hadoop
https://www.oreilly.com/content/ubers-case-for-incremental-processing-on-hadoop/

Image Appendix

Image 1: Log Growth

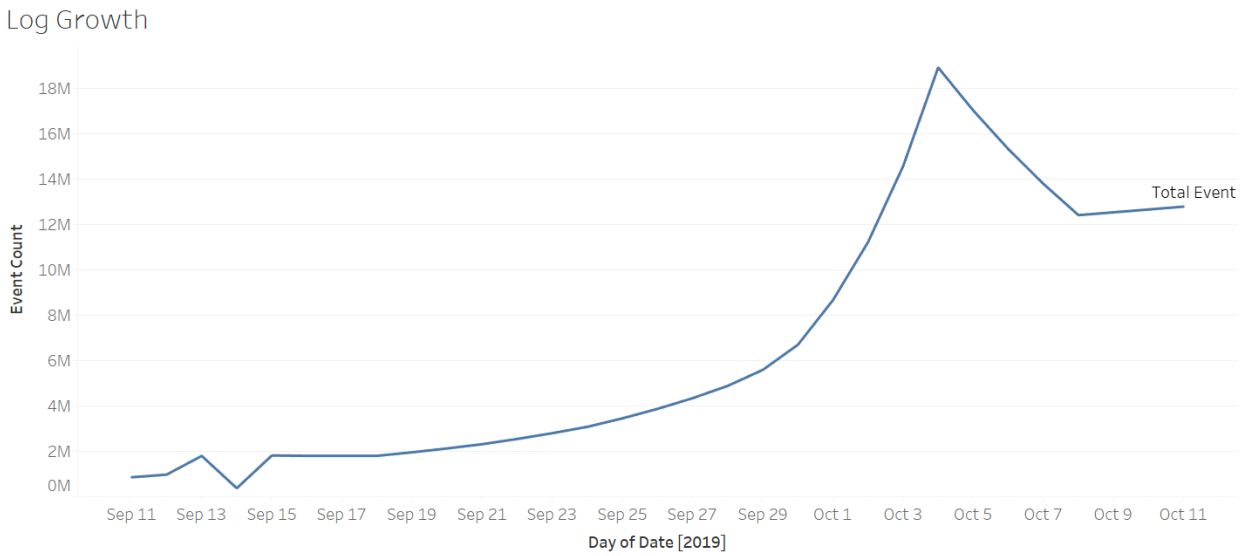


Image 2: Ride Growth

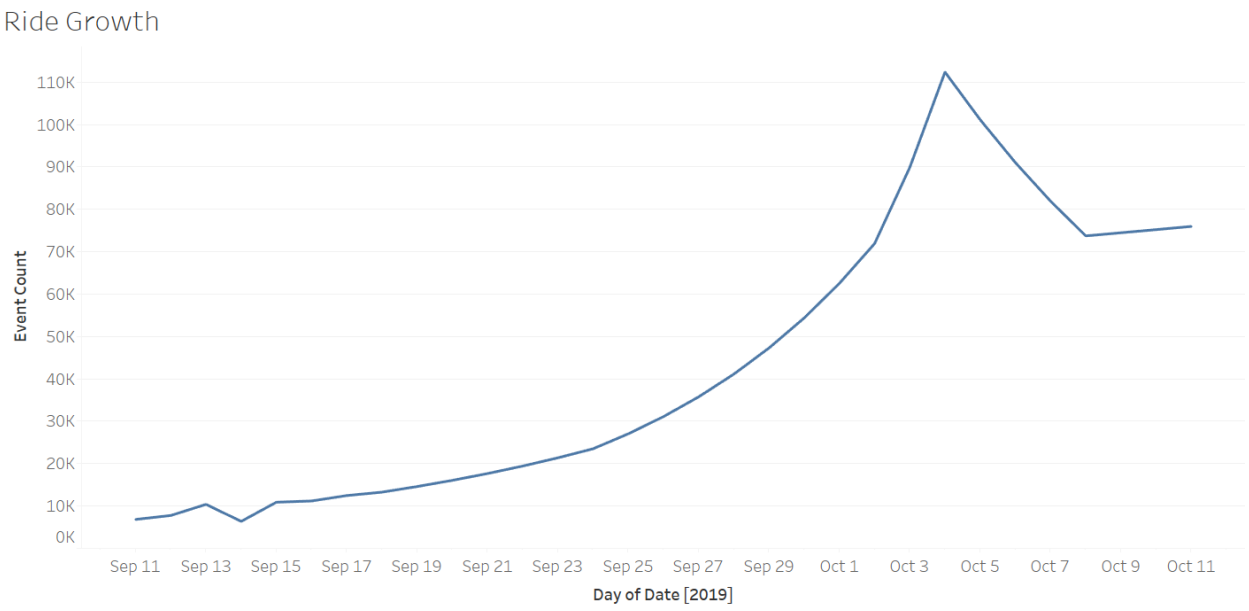


Image 3: Total Event Count

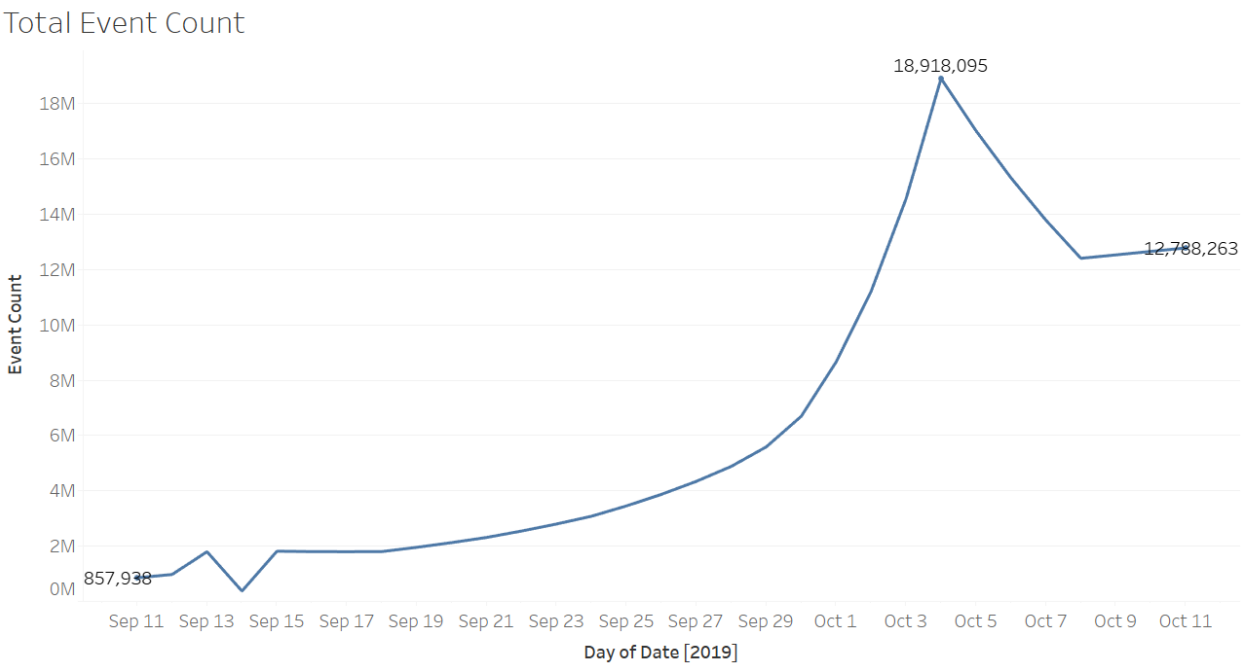


Image 4: All Events Log Scale