Flyber Data Strategy MVP

Data Product Manager: Otávio Bastos

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	They take care of the platform clients are using to request Flyber's Volocopters.	Monitoring app performance.
Product Management	They make sure the product is improving through time	Identifying customer pain-points
Operations	They make sure the operation is up and running.	Anticipate, identify and mitigate operational risks (increasing lead time, accidents, equipment maintenance)
Growth	They make sure we grow at an exponential rate	Monitor Flyber's usage
Marketing	They ensure we build an excellent brand reputation, attract new clients and make existing clients love our services.	Targeted advertising and Performance Marketing
Legal	They take care of legal disputes we might have.	Legal Risk Mitigation
Finance	They make sure we are financially healthy	Monitoring current P&L
Customer Satisfaction	They address any customer grievances	Provide personalized responses to the customer
Investors	They have funded our business and will connect us with people for future VC rounds	Monitor main KPIs

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 performance.	Event ID, Timestamp, Event type	In order to make sure Flyber's digital platform is up and running, the Engineering Team needs to monitor basic event data in **streaming view.**
Product Management	Identifying customer pain-points	Event ID, Time stamp, Event Type, product_journey User Research Data: Customer name, user_ID, email, phone, address, question, answer, score	To improve Flyber over time, PMs need basic platform usage information and essential user research data.
Operations	Anticipate, identify and mitigate operational risks (increasing lead time, accidents, equipment maintenance)	Event Data: Event ID, Time Stamp, Event Type, product_journey, location_latitude, location_longitude, user_ID, pilot_ID, flight_key, vehicle_serial_number Entity Data: flight_key, takeoff_timeStamp, landing_timeStamp	In order to make sure Flyber's operation is up and running, Operations needs basic event data in **streaming view** and basic entity data in **batches** every 30 minutes.
Growth	Monitor Flyber's usage	Event ID, Time Stamp, Event Type, location_latitude, location_longitude, user_ID	To grow at an exponential pace, Growth teams need to know basic usage data: what, when and where is happening.

Marketing	Targeted advertising and Performance Marketing	Entity Data: Customer name, user_ID, email, phone, address, customer_rides_history Entity Data: user_ID, age, address, city, state, country, profile_interests	The Marketing Squad must know essential information about Flyber's usage and demographics about its clients and potential clients.
Legal	Legal Risk Mitigation	suit_ID, user_ID, opening_timeStamp, totalCost, status	The Legal Office has to follow every suit that Flyber might face.
Finance	Monitoring current P&L	Aggregated Transactional Data Flight_key, pilot_ID, user_ID, distance_km, duration, fuel_consumed, total_cost, total_charges, ride_price	Monitoring essential financial data will make sure the business is up and running, as well as it is financially healthy.
Customer Satisfaction	Provide personalized responses to the customer	request_ID, user_ID, request_subject, ticket_opening_timeStamp, status, total_wait_time, priority, customer_channel	Basic customer care will be needed if we want to focus on User Engagement.
Investors	Monitor main KPIs	Aggregated Usage Data Total Rides per Day Total Unique Users per Day Customer Retention Total Weekly Cost Total Weekly Revenue	Flyber will probably need to raise money from further VC funding rounds, so reporting to investors should also be one of our priorities.

		Average Customer Life Cycle Time Average Lifetime Value	

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:

Event Data

event_ID	*event_timeStamp*	*event_type*	*product_journey*	*location_latitude*	*location_longitude*	*user_ID*

Primary Key: event_ID

Foreign Key: user_ID

event_ID** a primary key as it is a unique ID that describes and differentiates every row on the table, while **user_ID** is a foreign key as there could be multiple **user_ID** for different **event_ID

Table 2:

Flight Data

flight_key	*user_ID*	*pilot_ID*	*vehicle_serial_number*	*takeoff_timeStamp*	*landing_timeStamp*

Primary Key: flight_key

Foreign Key: user_ID

Foreign Key: pilot_ID

flight_key** a primary key as it is a unique ID that describes and differentiates every row on the table, while **user_ID** and **pilot_ID** are foreign keys as there could be multiple **user_ID** and **pilot_ID** for different **flight_key

---

Table 3:

Customer Care

request_ID	*user_ID*	*request_subject*	*ticket_opening_timeStamp*	*status*	*total_wait_time*	*priority*	*customer_channel*

Primary Key: request_ID

Foreign Key: user_ID

request_ID** a primary key as it is a unique ID that describes and differentiates every row on the table, while **user_ID** is a foreign key as there could be multiple **user_ID** for different **request_ID

Table 4:

CSAT (Customer Satisfaction Score)

survey_ID	user_ID	product_journey	usage_timeStamp	score

Primary Key: survey_ID

Foreign Key: user_ID

survey_ID is the only unique variable here and it will be what differentiates one row from another, that is why it should be our primary key. user_ID would be a foreign key, as it does not describe an unique row and because for a given survey_ID, we could possibly have multiple user_ID, i.e. clients, answering these CSAT surveys.

Table 5:

Legal Ops

suit_ID	user_ID	opening_timeStamp	totalCost	*status*

Primary Key: suit_ID

Foreign Key: user_ID

suit_ID** a primary key as it is a unique ID that describes and differentiates every row on the table, while **user_ID** is a foreign key as there could be multiple **user_ID** for different **suit_ID

Table 6:

Financial Control

Flight_key	pilot_ID	user_ID	distance_km	duration	fuel_consumed	total_cost	total_charges	ride_price

Primary Key: flight_key

Foreign Key: user_ID

Foreign Key: pilot_ID

flight_key** a primary key as it is a unique ID that describes and differentiates every row on the table, while **user_ID** and **pilot_ID** are foreign keys as there could be multiple **user_ID** and **pilot_ID** for different **flight_key

Table 7:

KYC Table

customer_name	user_ID	email	phone	address	customer_rides_history

Primary Key: user_ID

***user_ID** a primary key as it is a unique ID that describes and differentiates every row on the table*

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 and Transformation:

1. **Events Per Day**
    a. Create event_date column from timestamp event_time
    b. Create event_day column from event_date
    c. Count distinct event_uuid for every event_day
2. **Number of Events for each Event Type per Day**
    a. Create event_date column from timestamp event_time
    b. Create event_day column from event_date
    c. Count distinct event_uuid for every event_day and event_type
3. **Events per Device Type per Day**
    a. Create event_date column from timestamp event_time
    b. Create event_day column from event_date
    c. Count distinct event_uuid for every event_day and device_type
4. **Events per Page Type per Day**
    a. Create event_date column from timestamp event_time
    b. Create event_day column from event_date
    c. Count distinct event_uuid for every event_day and event_page
5. **Events per Location per Day**
    a. Create event_date column from timestamp event_time
    b. Create event_day column from event_date
    c. Count distinct event_uuid for every event_day and user_neighborhood

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
android	1463	2870	2854	2729	2744	2562	2672
desktop_web	895	2007	1600	1958	1712	1866	1777
ios	2384	4337	4217	4373	4380	4482	4500
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
book_page	1977	3548	3576	3572	3586	3424	3506
driver_page	965	1823	1871	1794	1755	1689	1768
search_page	3995	7219	7307	7221	6979	7201	7137
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
Bronx	250	533	507	469	510	394	558
Brooklyn	2009	3737	3590	4025	3440	3400	3556
Manhattan	6869	12591	12807	12180	12270	12371	12201
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.

Doing a manual ETL is definitely not a good choice for Flyber's expected exponential growth. We have pointed out in previous sections a huge amount of stakeholders and data use cases, we do need to find a more efficient way to feed all those departments with data.

It is critical for the success of our data infrastructure that we apply appropriate processing strategy for this piece. I propose that we reconsider the manual ETL process we have just performed and replace it with a automatic ETL pipeline flowing to a centralized Data Lake.

A centralized data lake will provide us the following advantages:

1. *Easy integration of new data sources as data lakes do not have strict data format requirements*
2. *Less time to integrate new data sources*
3. *Data in data lakes will be in its natural form that will help us use and process data in multiple ways (based on how applications need it).*
4. *It will help us scale, as our data is growing day by day, we need a scalable solution.*

Our ETL processes will run every 1 hour for departments in need of bath data and we should use a ELT process for areas needing a streaming view of these data, such as the Engineering and Operations Teams. In order to feed these areas with a streaming data pipeline, we should create a Extract, Load and Transform these data transforming event data with a matching mask to a more readable structure.

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.

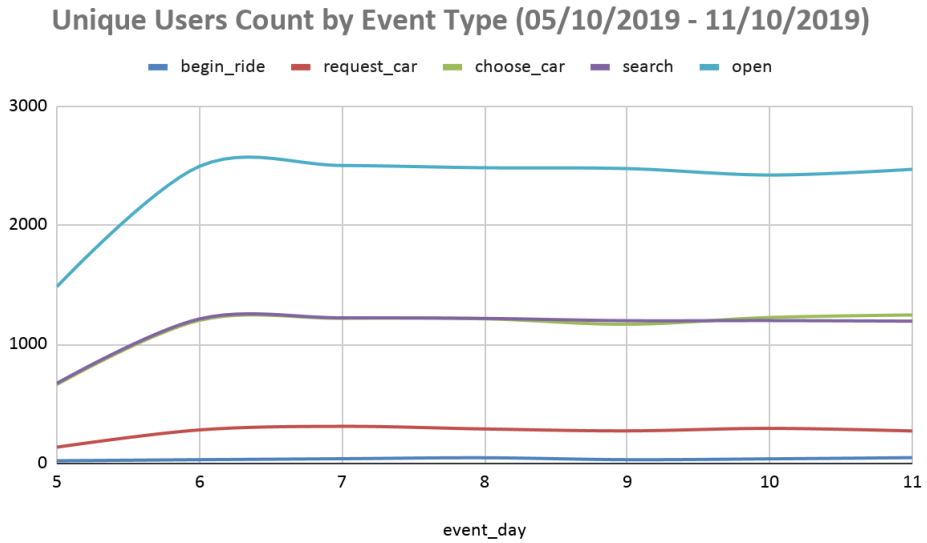
1. How many events are being recorded per day?
2. **How many events of each event type per day?**
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

Choosen prompt

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?

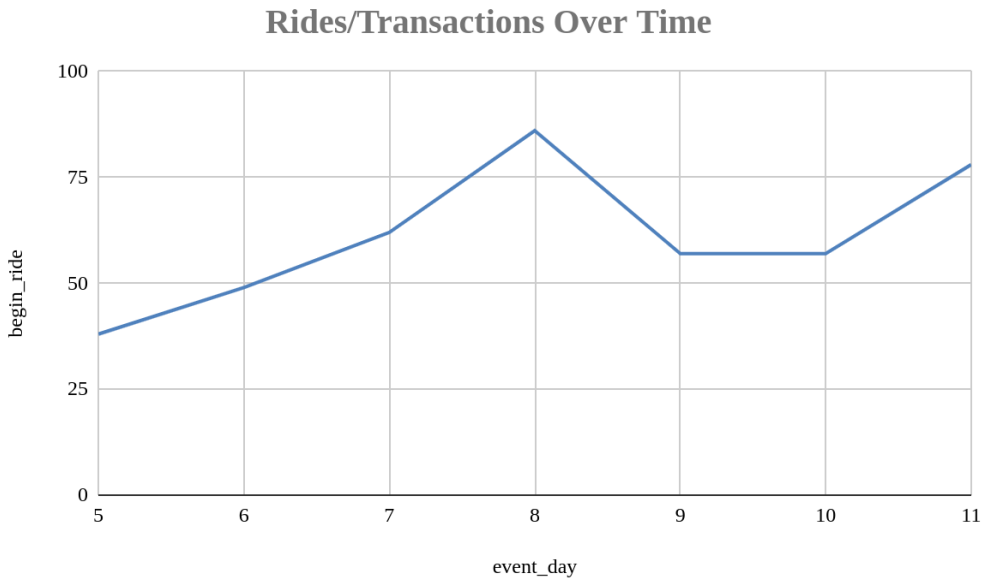
We can clearly see unique users reaching a peak after the 6th October 2019 and remaining stable over time for any event type.



It is also possible to check for overall conversion using the total amount of unique users going through each step on the user's journey. An overall conversion of 3.37% was found considering opening the app as a first step.

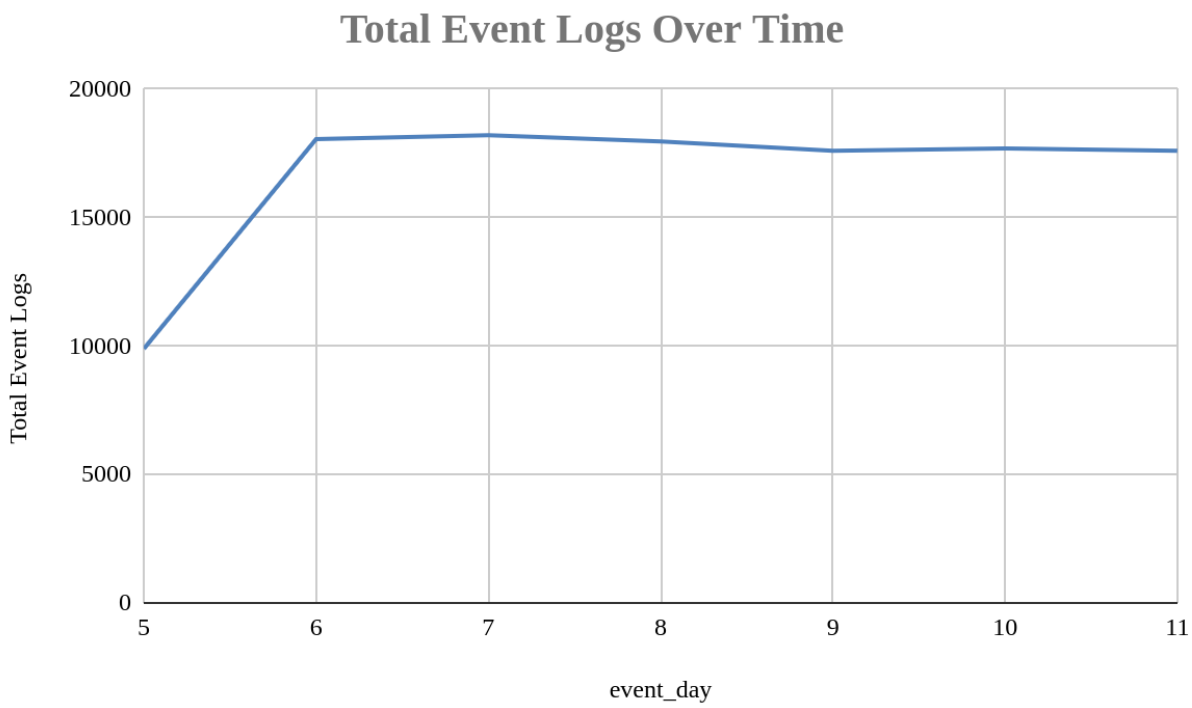
		Step Conversion	% of 1st Step
open	**7975**	100,00%	100,00%
choose_car	**5716**	71,67%	71,67%
search	**5679**	99,35%	71,21%
request_car	**1830**	32,22%	22,95%
begin_ride	**269**	14,70%	**3,37%**

2. How much is the transactional data increasing?



We can see transactions evolving over time with 2 peaks on the 8th October, 86 rides in total, and on the 11th October, 78 in total.

3. How much is the event log data increasing?



Event Log Data reaches a peak on the 6th October with 18202 events collected on a single day and remains below 18000 events a day for the following days.

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

- Event Log Data
- Transactional Data
- Customer Data

As we're currently focusing on **User Engagement**, understanding current ridership behavior and forecasting future pattern changes in behaviour is definitely most important. Event Log Data will also provide as with an optimized understanding about the conversion funnel between every step of the user's journey, allowing us to optimize user experience by tackling main friction steps.

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.

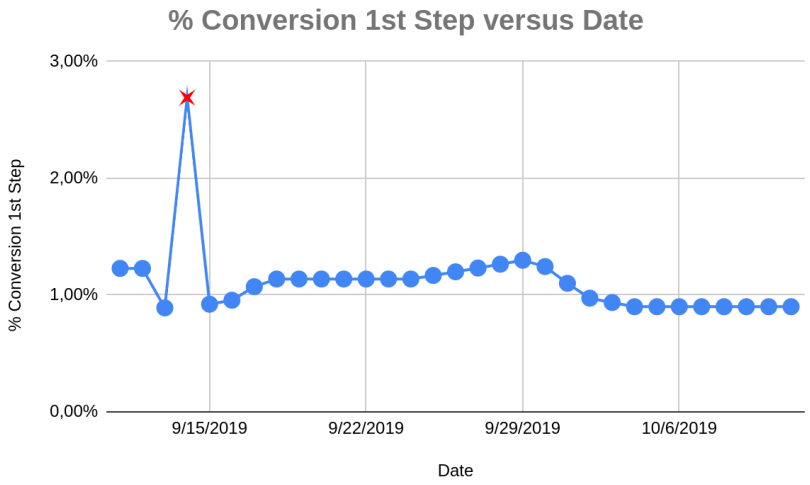
Engineering provided you with the data you want, but you still have yet to achieve your ultimate goal as a Data Product Manager. Now, utilize the data to make business decisions. Your executives do not want you to give them a bunch of data tables; instead, they prefer visualizations to help convey the 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.

We're using the Data Visualizations on the Appendix for Section 6. **We'll perform on this section only alternative Data Visualizations that could further feed us with quality insights**.

Check for graphs here: **LINK**

Visualization 1:



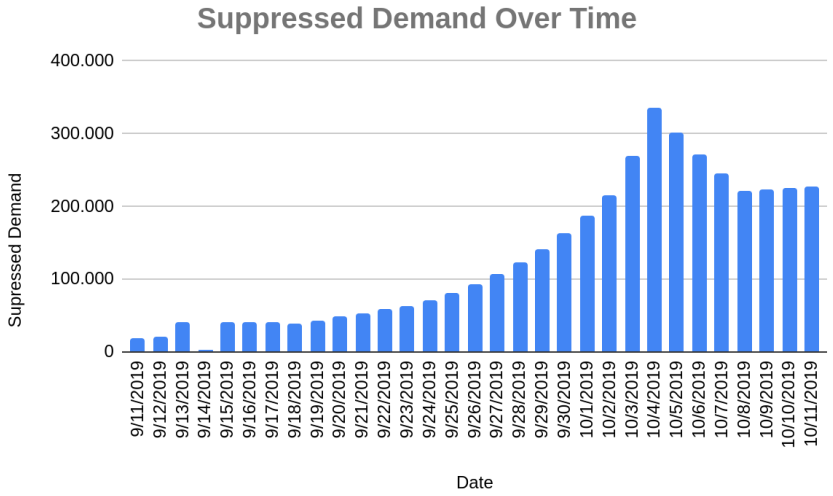
Data Story: This graph tells us:

There was a peak of conversion (% of [#rides]/[#app_opening]) on the 14th September 2019, we should definitely investigate it in deep details in order to understand why this maximum conversion happened and maybe reinforce this same behaviour for future days of operation.

This graph was created using the following steps:

1. *A column called [% Conversion 1st Step] was created dividing the total number of events of the column [Begin Ride] by the total events of the column [Open] day by day;*
2. *A Line Graph was generated with the [% Conversion 1st Step] and [Date] columns.*

Visualization 2:



Data Story: This graph tells us:

There is an increasing Suppressed Demand over time. Not every ride requested becomes an actual ride, day after day. This is a living proof for future VC rounds which proves that an expansion in our operations makes sense.

This graph was created using the following steps:

1. *A column called [Suppressed Demand] was created subtracting the total number of events of the column [Request Card] by the total events of the column [Begin Ride] day by day;*
2. *A Bar Graph was generated with the [Suppressed Demand] and [Date] columns.*

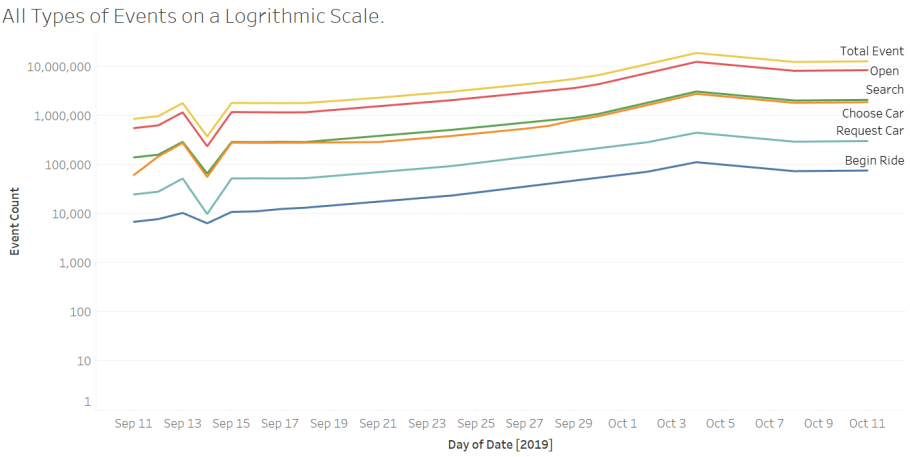
Section 6: Business Insights

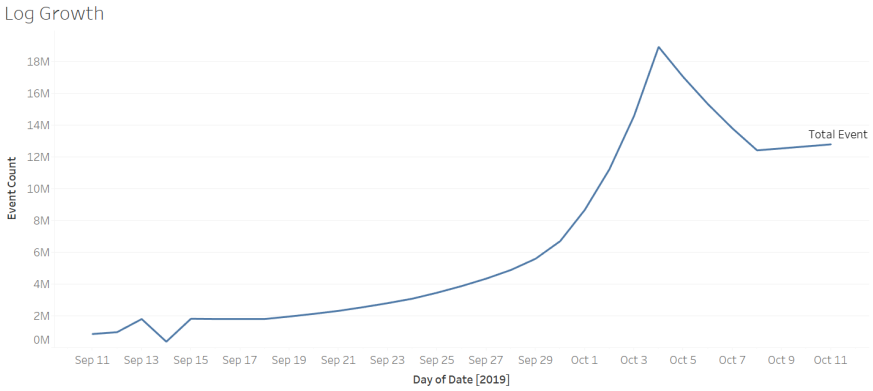
The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received 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:

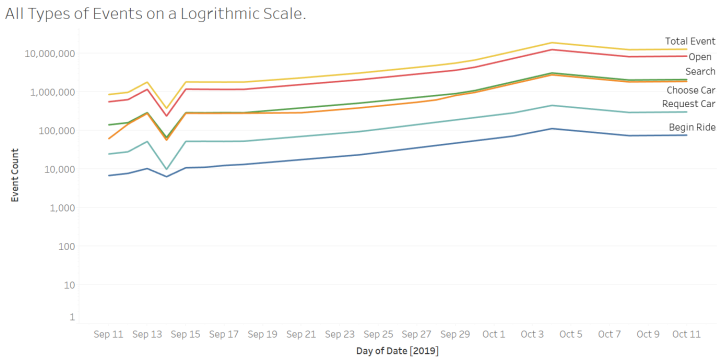
Log Growth *graph was plotted aggregating total event count on event data day by day.*
All Types of Events on Logarithmic Scale *graph was plotted using total number of events by event_type day after day.*

What is the fastest growing data and why?

Being a top-of-funnel event, the "Open" event type is the fastest growth perceived. Indeed, it is quite normal that top-of-funnel events grow at a faster pace than end-of-funnel events.

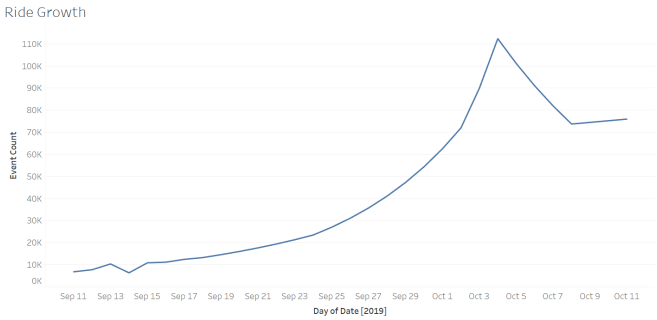
All Event Type Data

Visualization:



What is the Data Story our data tells for each of the following:

- Graph Pattern
  - Valley around the 14th September followed by a consistent growth, intensified from the 1st October.
- Good or Bad
  - Overall behaviour seems good, Flyber's usage is growing at a fast pace, but considering Open->Search->Choose Card->Request Car->Begin Ride as a conversion funnel, overall conversion is still very low.
- October Marketing Campaign



  - Results from October Marketing Campaign were quite positive, as we can notice a surge of number of rides at a faster pace from the 1st October.
- Marketing Campaign Impact
  - If we project same growth seen in September, we would have something around 80K rides on the 4th October, but we had 110K rides as a result of the marketing campaign, that is a 37.5% increase from the no-campaign cenario.
- Importance of Relationship Between Marketing Campaigns and Data Generation
  - Marketing Campaigns are not only a tool for increasing revenue, it is also a way to collect data from new users registering in our platform in order to leverage consistent growth over time.

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.

We were growing 2x rides every 7 days on September and 2x rides every 4 days in October after the Marketing Campaign, we don't have the time and opportunity cost to waste waiting for the hiring process on new professionals in order to have in-house expertise on how to build a complete Data Warehouse and Data Lake Infrastructure. Moreover, an in-house solution wouldn't be fully reliable.

I propose we should be completely cloud based. *Costs will be proportional to our event log generation, so we should definetely optimize our conversion funnel in order to make sure conversions and revenue will grow as event log data grows too. We will be fully scalable and our infrasctructure will follow our growth and size. We won't have any infrastructure in-house expertise, but it doesn't matter, we're a digital product platform, not a traditional Computer Software Company. In terms of latency and reliability, there are plenty of plug-and-play solutions on the market that will definetely ensure these, as well as a full support 24-7.*

Suggested DWH

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

Solution/Criteria	***Cost***	***Scalability***	***In-house Expertise***	***Latency/Connectivity***	***Reliability***
Amazon Redshift	*On-demand, managed storage*	*Scales horizontally and vertically*	*PostgreSQL and RDMs*	*Good/AWS Ecosystem, Data Integrations, BI and Reporting Tools*	*Highly Reliable*
Google BigQuery	*Flat rate, on-demand*	*Scales horizontally and vertically*	*SQL and ETL Tools*	*Good/Google Workplace, Data Integrations, BI and Reporting Tools*	*Highly Reliable*
Snowflake	*On-demand, pre-purchased*	*Scales horizontally and vertically*	*SQL and DW Architecture*	*High/Data Integrations, BI and Reporting Tools*	*Highly Reliable*
Microsoft Azure	*Compute charge, storage charge*	*Scales horizontally and vertically*	*SQL and Spark*	*High/Microsoft Software Integrations, Data Integrations, BI and Reporting Tools*	*Highly Reliable*

Legend: winner on criteria

Amazon Redshit *was the winner in 4 out of 5 criteria, we should use* ***Amazon Redshift*** *for our fully cloud based platform.*

Image Appendix

Image 1: Log Growth

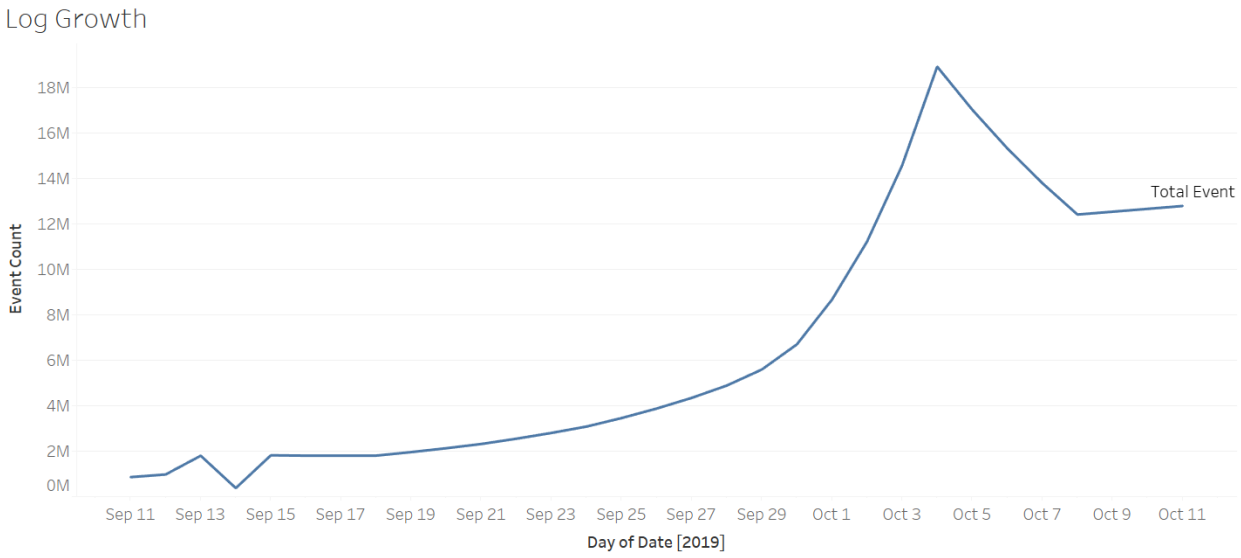


Image 2: Ride Growth

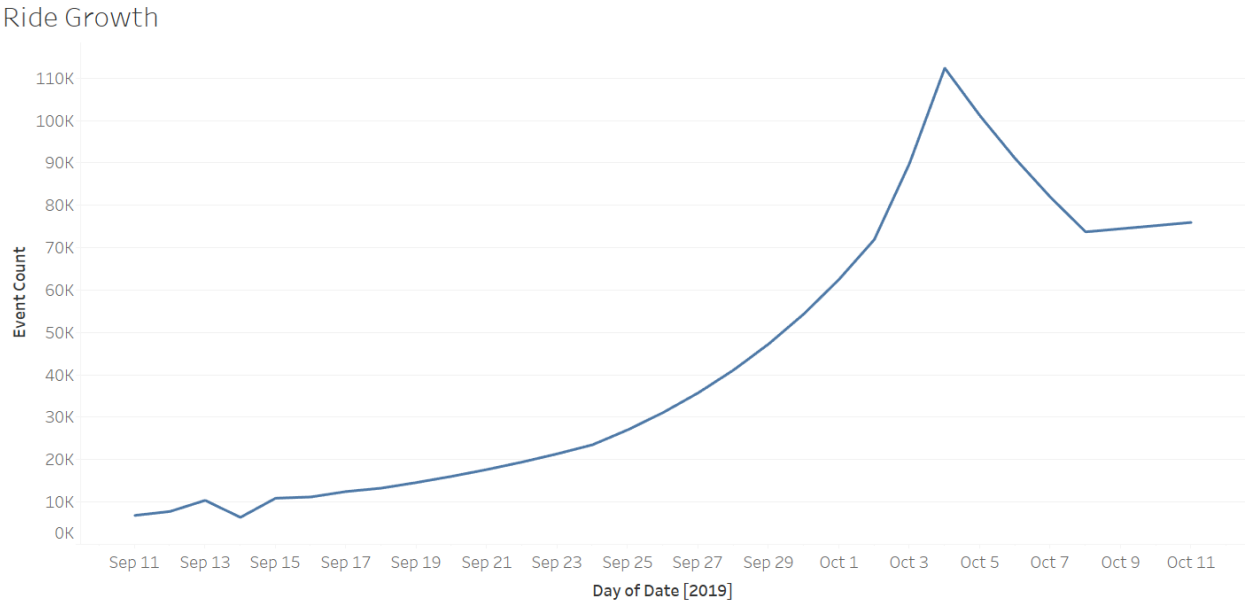


Image 3: Total Event Count

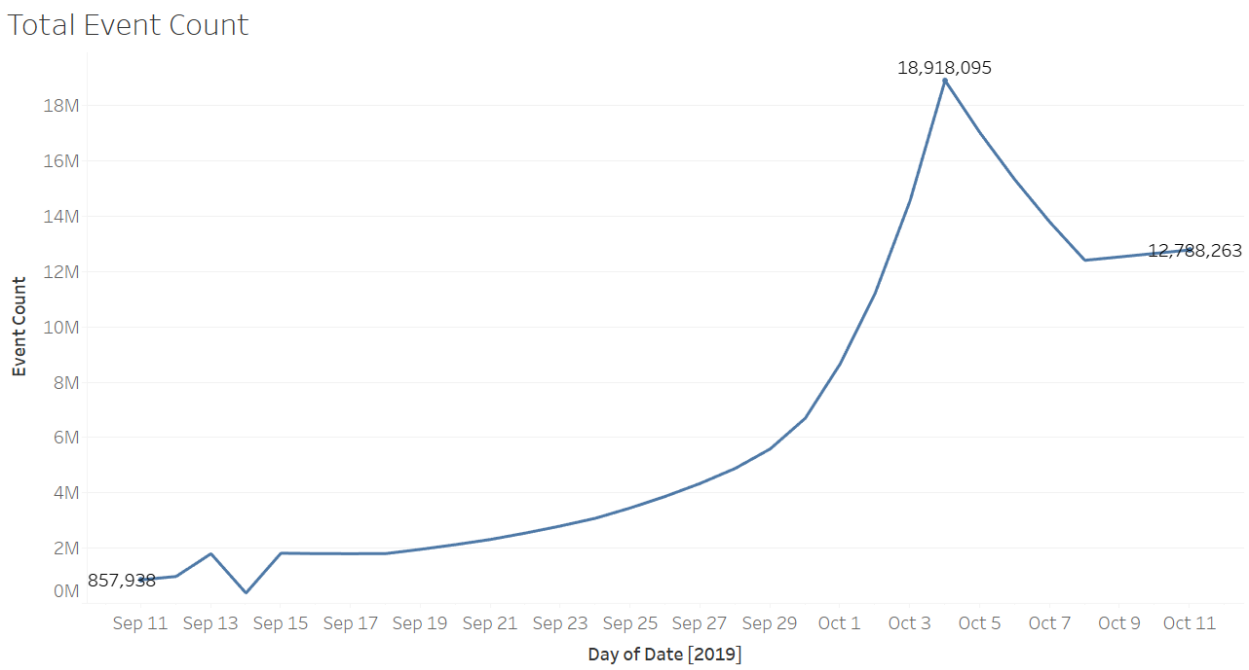


Image 4: All Events Log Scale