Flyber Data Strategy MVP

Introduction

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

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

Section 1: Data Customers & Needs

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

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

Stakeholder	**Why are they primary stakeholders?**	**Use-Case**
Engineering	Creates, designs and maintains the product	Monitor Flyber site and app performance
Accounting	Recording the expenses and earnings of a business	Monitoring current profit and loss, develop future finance predictions

Customer Care	Handles customer grievances and issues	Provide personalized responses and record feedback from customers for Flyber development
Marketing	Acquiring new customers and retaining old	Targeted advertising, customer profiling and identifying opportunities in customer preferences

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	Monitor Flyber site and app performance	Data generated from app and website. Entity- Customer ID(new and existing) Event- Pick-up and drop-off location, duration of travel	Monitoring how customers are using the Flyber app and website is vital to maintain and improve performance levels of the product, as well as fixing any bugs.
Accounting	Monitoring current profit and loss, develop future finance predictions	Entity- Ride ID Event- cost per ride, duration of rides, timing of rides	Without tracking of profit and loss trends, early detection and fixing of issues may not occur and company can incur significant losses before it can fix the problem.
Customer Care	Provide personalized responses and record feedback from customers for Flyber development	Entity- Customer ID, Complaint ID Event- Pick-up/drop-off locations, Flyber usage timings,	If Flyber does not address customer grievances, they can end up losing customers and hence lose profitability.

		Complain or Question submitted	
Marketing	Targeted advertising, customer profiling and identifying opportunities in customer preferences	Entity- Customer IDs Event- Travel details such as pick-up/drop-off, time of using Flyber, Income bracket and demographics	Maintaining and acquiring new customers is critical for sustaining and growth of 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:

[Customer Details]

Customer ID	*First Name*	*Last Name*	*Address*	*Email*	*Phone*

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

Table 1 provides the details of all customers who have used Flyber at least once. This table is has the Primary Key as "Customer ID" and it is used to connect all event data related to individual customers. This table does not have any Foreign Key.

Table 2:

[Ride Details]

Ride ID	*Customer ID*	*Driver ID*	*Ride Cost*	*Pick-up Location*	*Drop-off Location*	*Duration*	*Pick-up Time*

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

The Primary Key here is "Ride ID", and this table contains all the details for each individual ride that has been book and executed through Flyber. This ID helps in uniquely identifying each ride. The Foreign ID here is "Customer ID", which is used to connect individual customers to unique ride events.

Table 3:

[Customer Complain Details]

Complaint ID	*Customer ID*	*Complain*	*Date-Time of Submission*	*Status*	*Solution*

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

This table contains all the details related to customer grievances, where the Primary Key is Complaint ID which uniquely identifies all complaints, and the Foreign ID is Customer ID, which helps link unique customers to each complaint.

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 Gathering and Analysis:*
    a. *Raw data is gathering from data producers and later organized into tables*
    b. *Data is grouped based on Customers, Event, Neighborhood, Event type*

2. *Data Verification*
    a. *Data needs to be checked if it is relevant or not, and check if other data extensions are needed.*

3. *Conversion of 'event time' object column to 'datetime' file type*
    a. *To separate date by date and/or time easily, this event timing needs to be in 'datetime' datatype*

4. *Check for data consistency, where records are checked for consistency of aligning with specific event IDs.*
    a. *This step makes sure there are no garbage data generated and any data incoming error is addressed asap.*

5. *Check for Duplication and Missing Values*
    a. *If missing values or duplicates are present, this may lead to inaccurate summary statistics and inaccurate projections*

6. *Data Organization for Visualization*
    a. *Based on different stakeholder requirements, the data can be grouped based on customers, events or neighborhood and so on. This can lead to easy generation of visualizations as well as quickly accessing the data if deeper dives are needed.*

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	1498	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.

Data can be misleading and incorrect if not monitored throughout the ETL process. Especially, during the initial development of ETL process, manually extracting data plays a key role in running ad-hoc tests and visualizations in order to catch errors at the source.
After initial development, this process can be automated with the help of standardized third party software or out-sourcing to more specialized companies can be undertaken based on data volumes, standards and analytics requirements.

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

In my opinion, the most relevant information can be obtained from:
'How many events per device type per day?'

Date	ios	android	Desktop Web	Mobile Web	**Grand Total**
2019-10-05	2384	1463	895	5149	9891
2019-10-06	4337	2870	2007	8842	18056
2019-10-07	4217	2854	1600	9531	18202
2019-10-08	4373	2729	1958	8903	17963
2019-10-09	4380	2744	1712	8764	17600
2019-10-10	4482	2562	1866	8784	17694
2019-10-11	4500	2672	1777	8646	17595
2019-10-12	2026	1231	682	4040	7979
Grand Total	30699	19125	12497	62659	124980

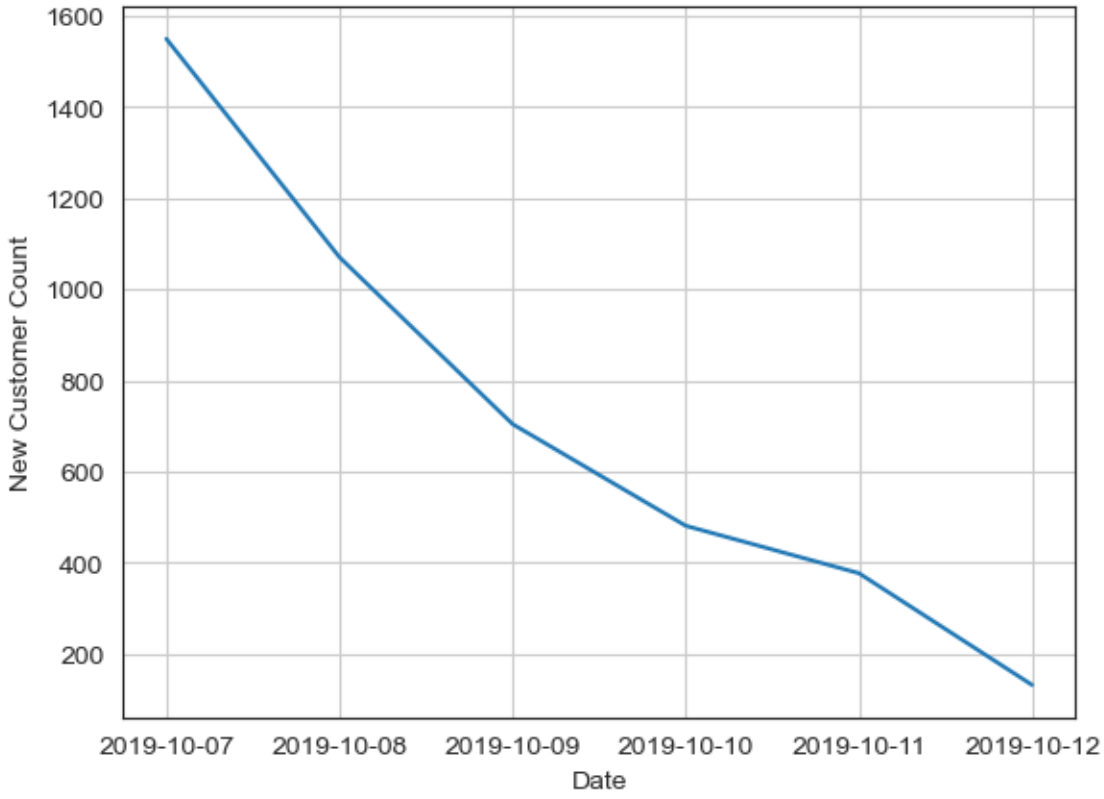
1. How much is the customer data increasing?

Answer:

Date	Total Unique Customers	Total New Customers
2019-10-05	1644	1644
2019-10-06	2743	2239
2019-10-07	2746	1549
2019-10-08	2735	1070
2019-10-09	2723	704
2019-10-10	2692	481
2019-10-11	2724	377
2019-10-12	1382	132

None of the measures tell us whether all customers are new or not for Oct 5th, but if we assume them to be all new customers, the new customers added from Oct 7th onwards can be calculated.

Excluding data from 5th and 6th, the trend of new customers is:



The trend shows a steady decline in new customers added. This trend may be further lowered if majority of the customers added on 5th and 6th October are existing customers.

2. How much is the transactional data increasing?

Answer:

In this case, Transactional data is considered as the 'event_type' to be 'begin_ride' as this would indicate the customer has started the ride which would only happen after a transaction is booked.

Range of 'begin_ride' is:

Date	**'begin_ride' count**
2019-10-05	38
2019-10-06	49
2019-10-07	62
2019-10-08	86
2019-10-09	57
2019-10-10	57
2019-10-11	78
2019-10-12	18

Hence, the maximum 'begin_ride' transactions happened on Oct 8th with 86 count and the minimum was on Oct 12th with 18 count.

3. How much is the event log data increasing?

Answer:

Date	**Total of Device Event Count**
2019-10-05	9891
2019-10-06	18056
2019-10-07	18202
2019-10-08	17963
2019-10-09	17600
2019-10-10	17694
2019-10-11	17595
2019-10-12	7979

Based on the table above, the maximum event based data is seen 7th Oct with 18202 count and minimum is on 12th Oct with 7979 count..

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

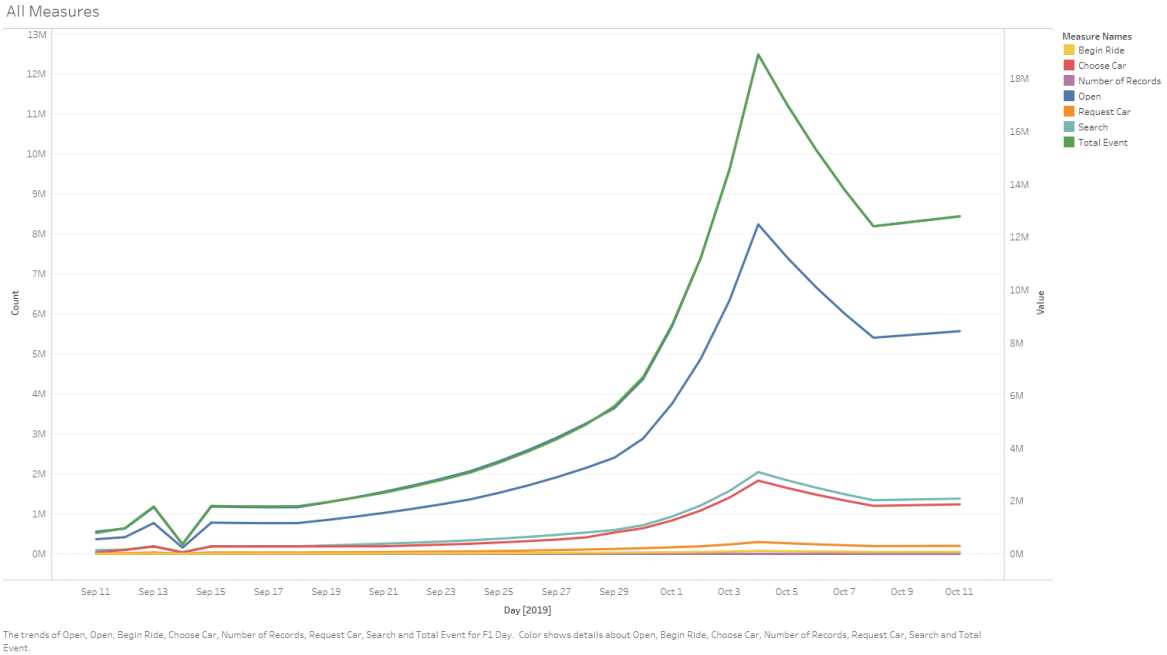
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

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:



The trends of Open, Open, Begin Ride, Choose Car, Number of Records, Request Car, Search and Total Event for F1 Day. Color shows details about Open, Begin Ride, Choose Car, Number of Records, Request Car, Search and Total Event.

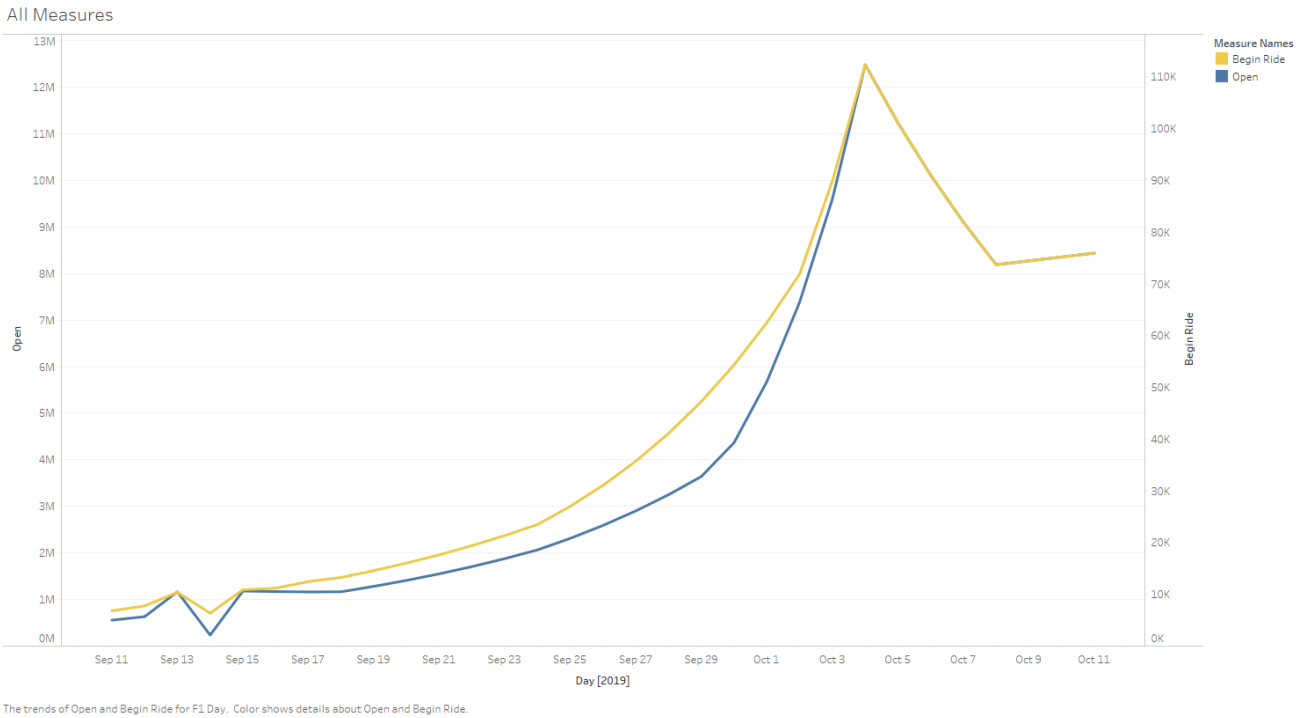
Data Story:

Based on the trends, all measures grow exponentially up to October 4th and then plummet drastically with gradual flattening of the slope. This may be due to initial interest in Flyber.

This graph was created using the following steps:

1. *All data was loaded as an excel file in Tableau*
2. *The column was selected as Date*
3. *The row was selected as all Measure Values and with the axis being set to 'Dual Axis'*

Visualization 2:



Data Story:
In this graph we see both 'Opening of the App' and 'Begin Ride' to be growing exponentially, but the scales for both plots being very different with roughly 1% of Opens was converted to Ride Begins.

This graph was created using the following steps:

1. *All data was loaded as an excel file in Tableau*
2. *The column was selected as Date*
3. *The row was selected as 'SUM(Open)' and 'SUM(Begin Ride)' with Dual Axis*

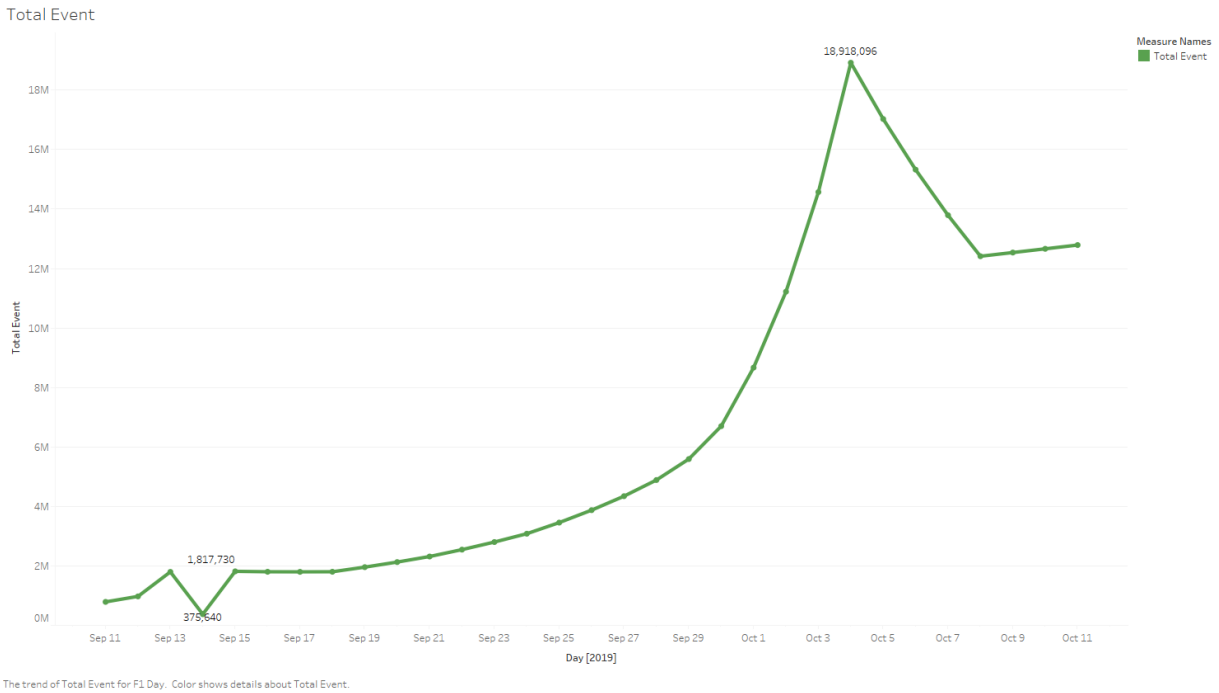
Section 6: Business Insights

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:

Highest Data Point = 18,918,096
Lowest Data Point = 375,640

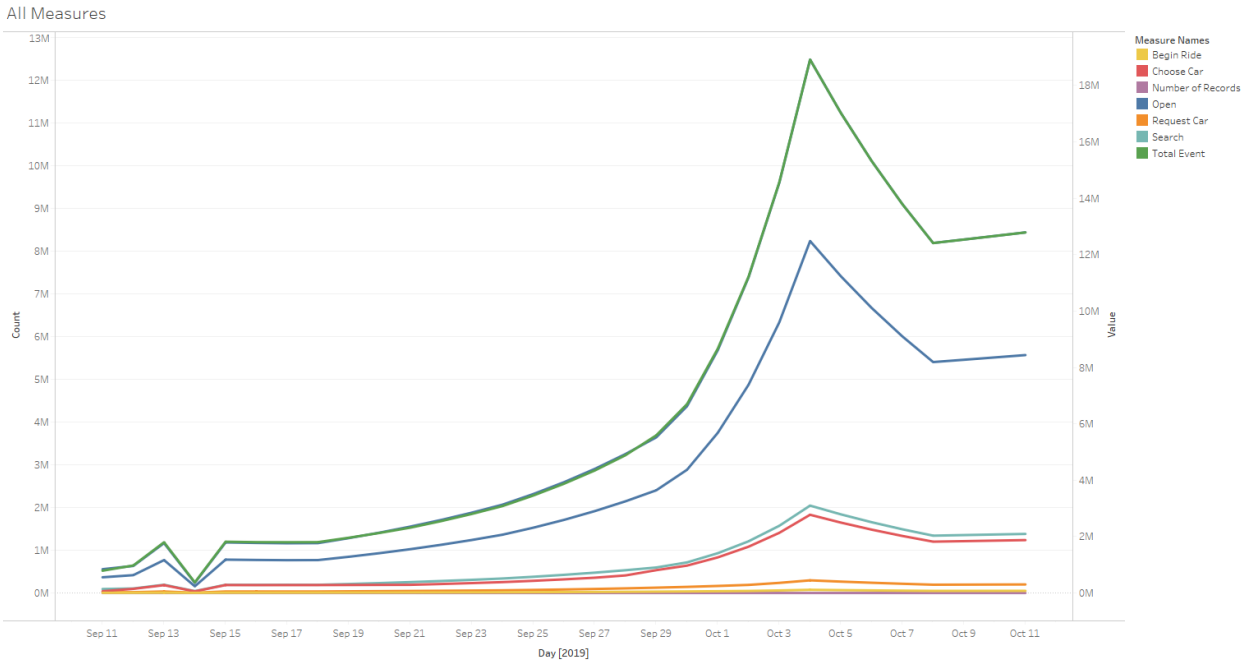
Growth = (18,918,096 – 375,640) / 375,640 = 4936.23%

What is the fastest growing data and why?

The fastest growing data is 'Open' where people are opening the app or website. This can be attributed to the initial interest created due to the launch of the product and good marketing tactics.

All Event Type Data

Visualization:



The trends of Open, Open, Begin Ride, Choose Car, Number of Records, Request Car, Search and Total Event for F1 Day. Color shows details about Open, Begin Ride, Choose Car, Number of Records, Request Car, Search and Total Event.

[Insert Visualization Here] If you didn't create your own, please use the images in the appendix to guide you through this section.

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

- *Graph Pattern- the pattern of the graph is an exponential one up to October 4th and then an exponential drop to a steady state.*
- *Good or Bad- There may be promos or good marketing tactics at pay up to October 4th, where the interest spiked. After that the interest dropped to a steady state. The Steady state indicates that this is a good situation to be in.*
- *October Marketing Campaign- The trends suggest that Marketing campaigns may not be needed to drive demand and interest in Flyber.*
- *Marketing Campaign Impact- The exponential growth in every measure suggests that the Marketing Campaign has been highly successful in driving up interest.*
- *Importance of Relationship Between Marketing Campaigns and Data Generation- Without the Data we would not have been able to correlate the timing and impact of marketing with interest and profits.*

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 being a start-up, cost, and in-house expertise would be limited and in such a situation, developing on-prem and in-house systems may be too cost intensive. In such a situation, using a Cloud DWH such as Snowflake or Amazon Redshift along with their technical support can be a great solution. As the Flyber grows, these solutions would be scale with the company.

Suggested DWH

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

Amazon Redshift and Snowflake would be the Cloud based DWH solutions I would select. The reason being both Snowflake and Amazon have a flexible architecture to accommodate elastic storage and data retrieval demands as well as high operational efficiency. This combined with Flyber being a relatively new team, the technical support would also help them build their business without much overhead.

Image Appendix

Image 1: Log Growth

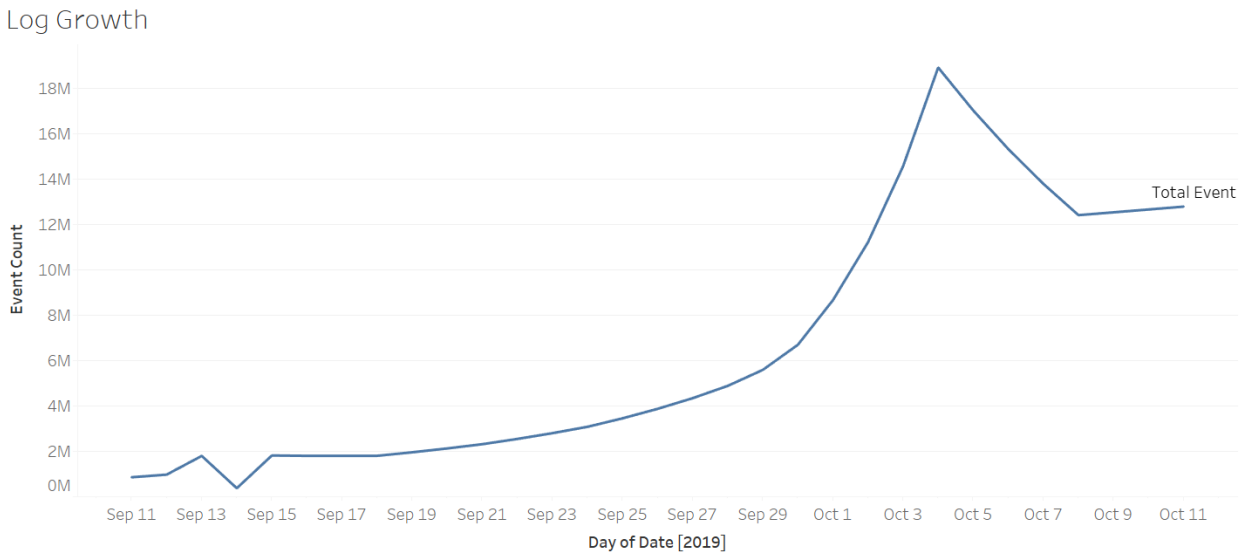


Image 2: Ride Growth

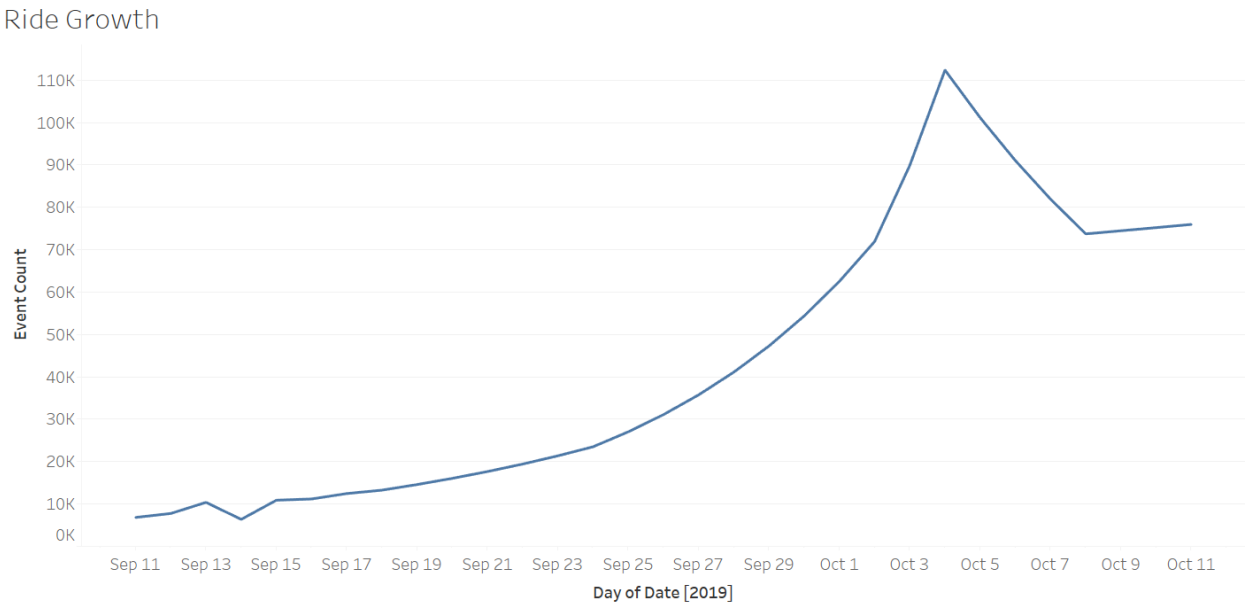


Image 3: Total Event Count

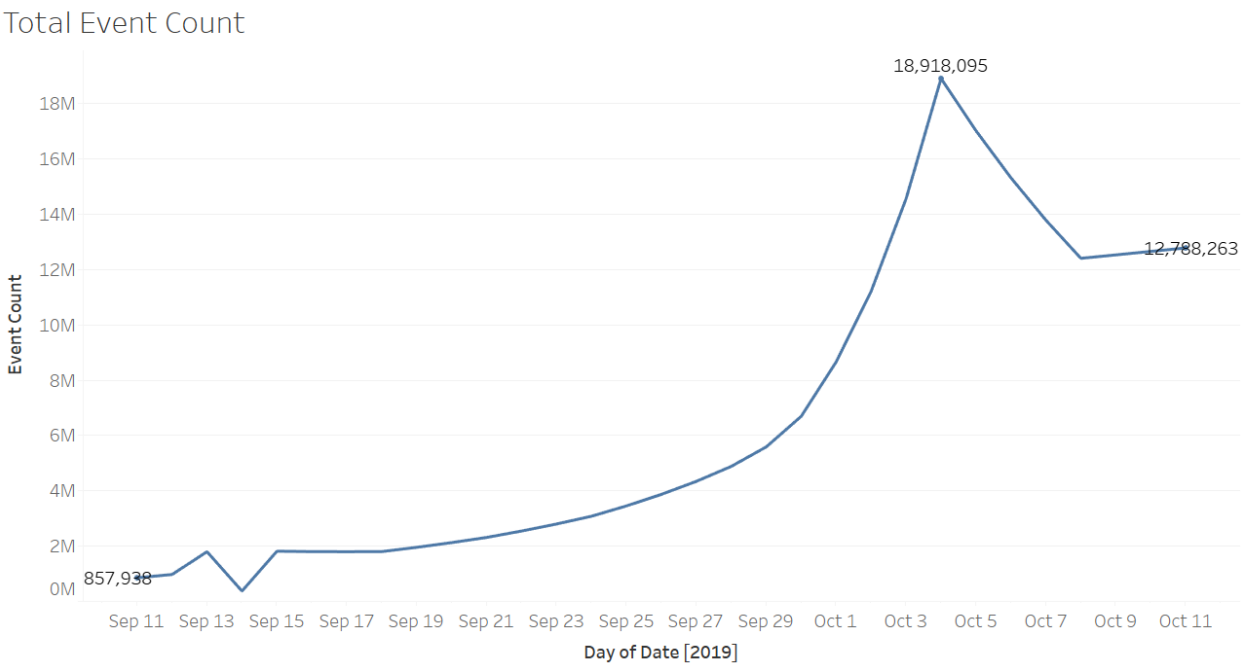


Image 4: All Events Log Scale