Company 2 Case Study

Ben Winby

Retention Analysis

Executive Summary

Whilst Company 2 has been growing rapidly over the last few years this growth is primarily fuelled by strong customer acquisition and steady retention. Retention is relatively high with over 50% making a 2nd purchase - however, retention has remained steady/slightly declined.

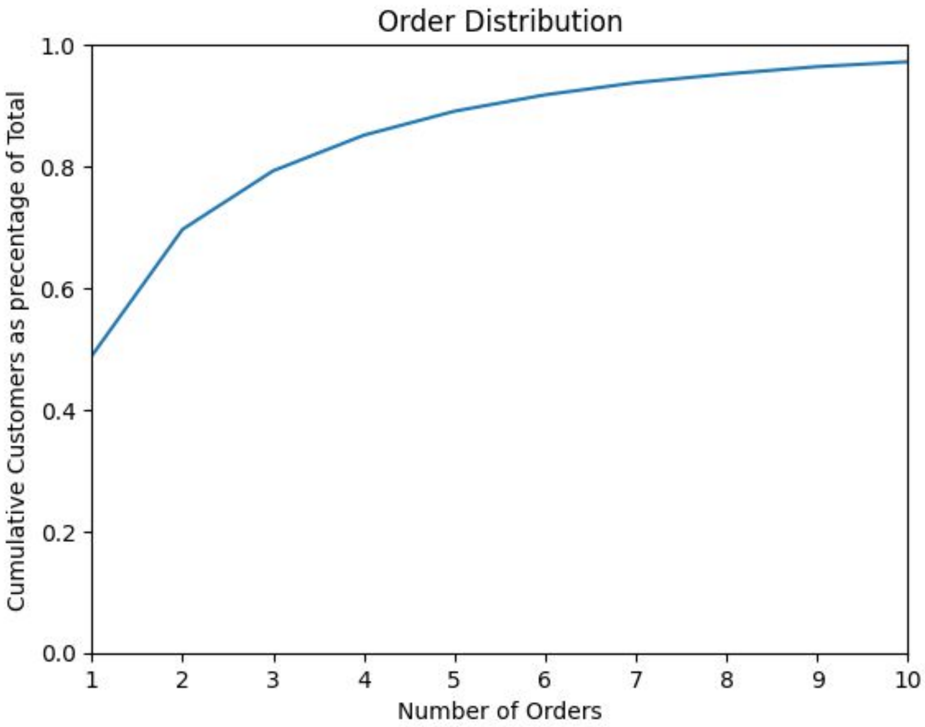
Retention varies significantly across Clinics - with some seeing 4x as many repeat orders as first time orders, whilst others struggle to get 5% to come back for a second order. Some of the largest clinics have the worst retention rates meaning there is lots of potential to be unlocked. It is also clear that the older Clinics have a higher retention than newer clinics. We should seek to better understand this behaviour so that we can accelerate the performance of new clinics.

In Feb 2021 and Feb 2022 we saw a significant improvement in retention across all cohorts. If we can tie this to specific marketing activity we will likely be able to replicate that success.

We also see an uplift in second purchases 2 months after the first purchase. This could be the result of CRM activity or another phenomena. By understanding what is driving this we might be able to extend its success.

We have a higher chance of driving a second order if we target users within the first month after making a purchase. Higher discounts are also correlated heavily with more repeat orders and so we should evaluate how discounts are being used.

50% of customers have purchased more than once

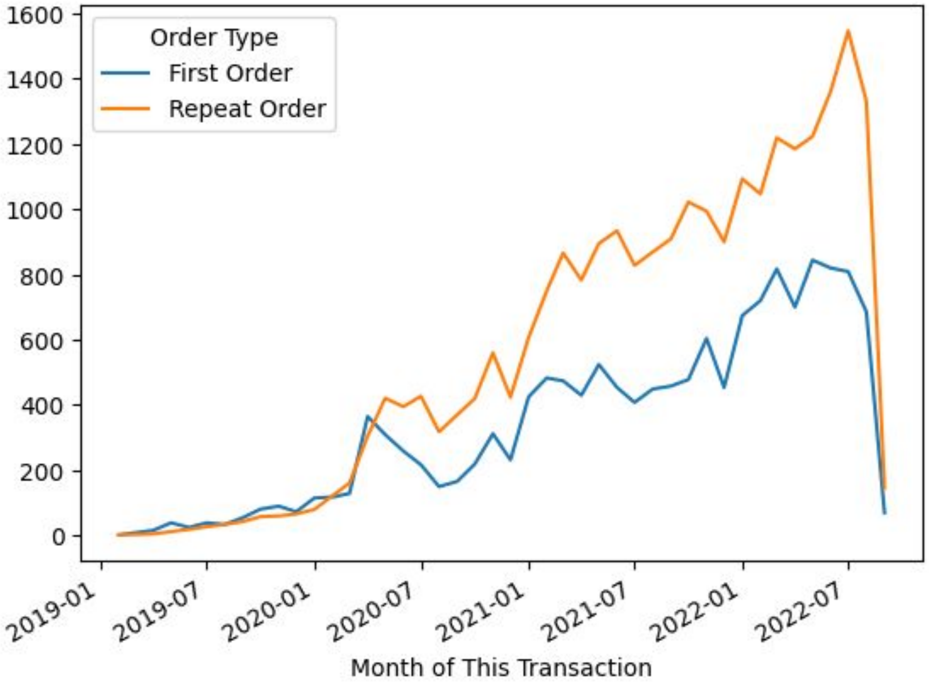


50% of customers have purchased just one time.

This is a large group of customers who we might be able to target.

In comparison to other ecommerce businesses this is relatively good

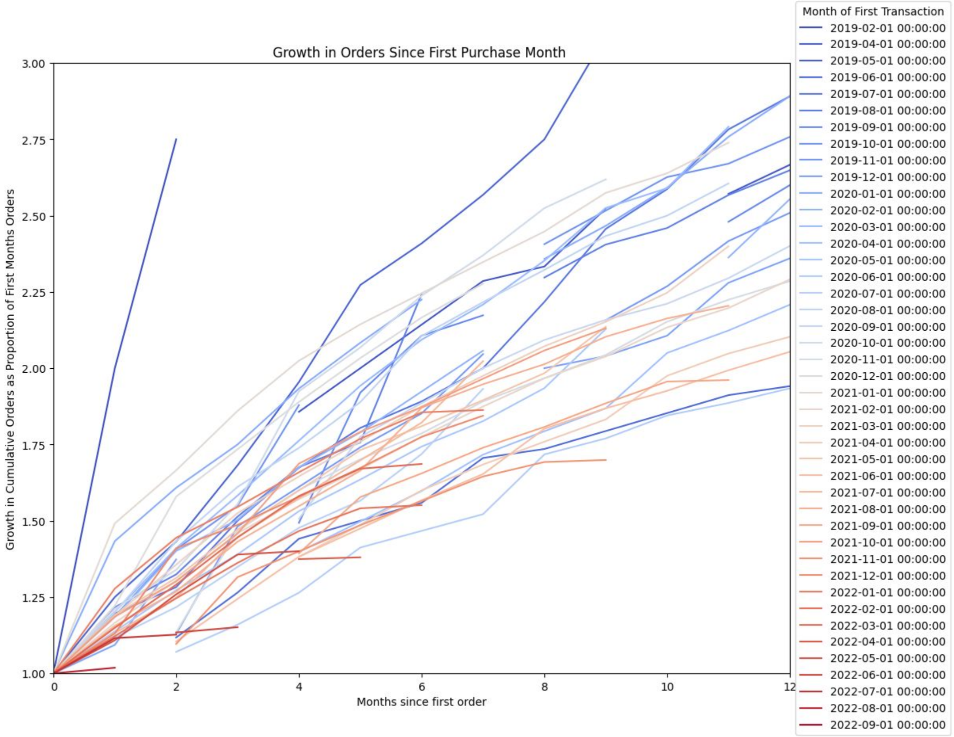
Repeat orders are growing very steadily



Repeat orders have grown significantly over the last few years

Repeat orders now account for over 65% of orders per month

Growth in repeat orders is being driven by acquisition since retention rates have weakened slightly



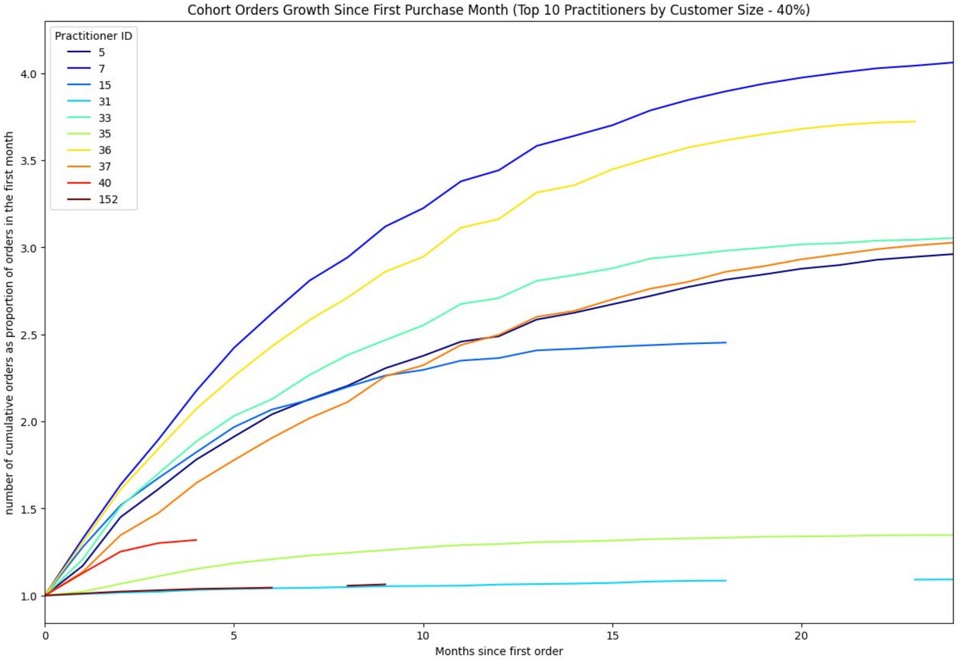
Retention rates have weakened slightly over the last couple of years

Growth in repeat orders is driven primarily by an improvement in user acquisition and a growing customer base

There is an opportunity to further drive growth by improving retention rates

However - continued growth shows that the business has a very strong footing.

Retention is significantly impacted by the clinic



Retention varies significantly depending on which clinic a user visits

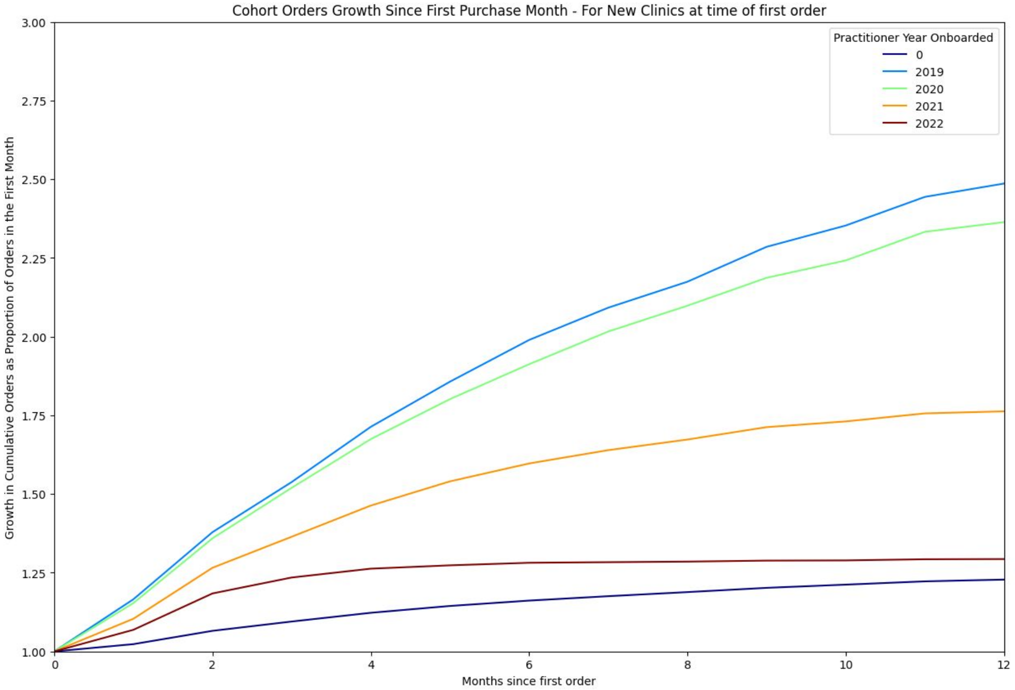
Some clinics are growing by over 4x and are still taking orders 2 years after first seeing someone.

Others are growing by less 5%

Recommendations:

- Investigate high performing vs low performing Clinics to understand what is driving the different behaviour.
- Look at commission structure/bonuses to incentivise clinics to support retention efforts
- Provide regular reporting to enable clinics to understand trends and purchasing behaviour of their clients

The clinics most recently onboarded have much worse retention



The clinics onboarded in 2019 and 2020 have a much higher retention that those onboarded in 2021 and 2022.

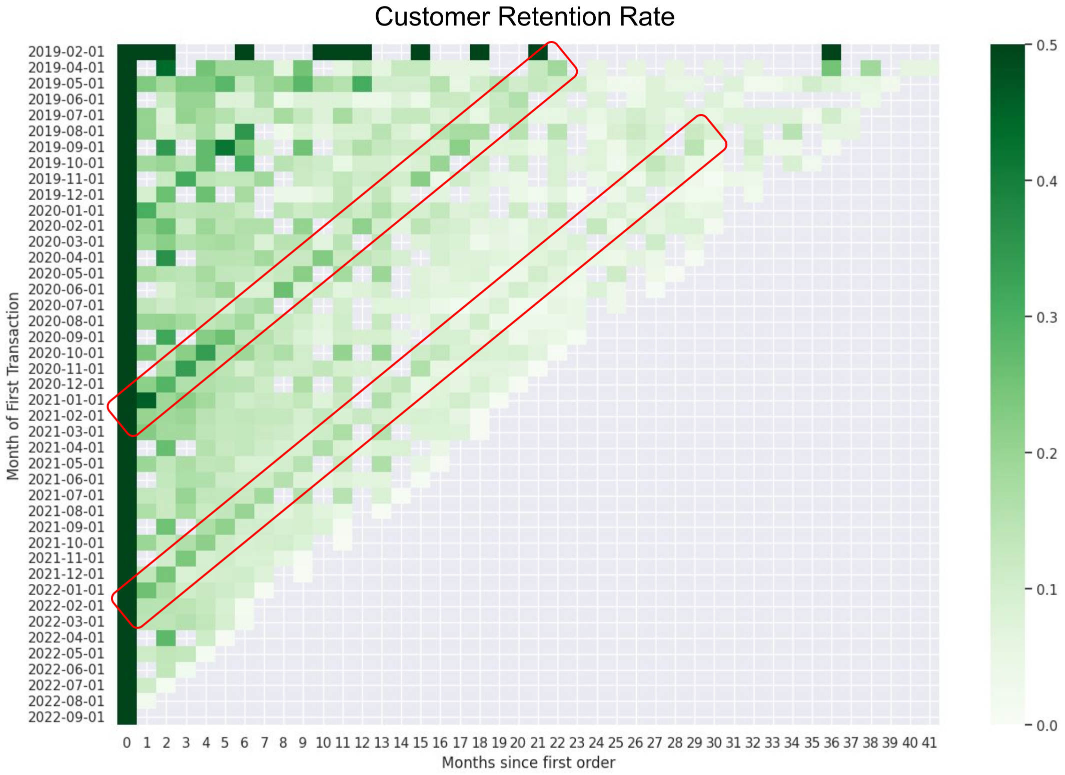
This suggests we are signing up lower quality clinics that we did to start with.

Retention does not improve the longer they have been onboarded - the trend remains the same.

Recommendations:

- Investigate why the oldest clinics are performing better than the newer ones. Use this when targeting clinics for onboarding onto the platform.

Specific events in the past appear to be driving retention across all cohorts



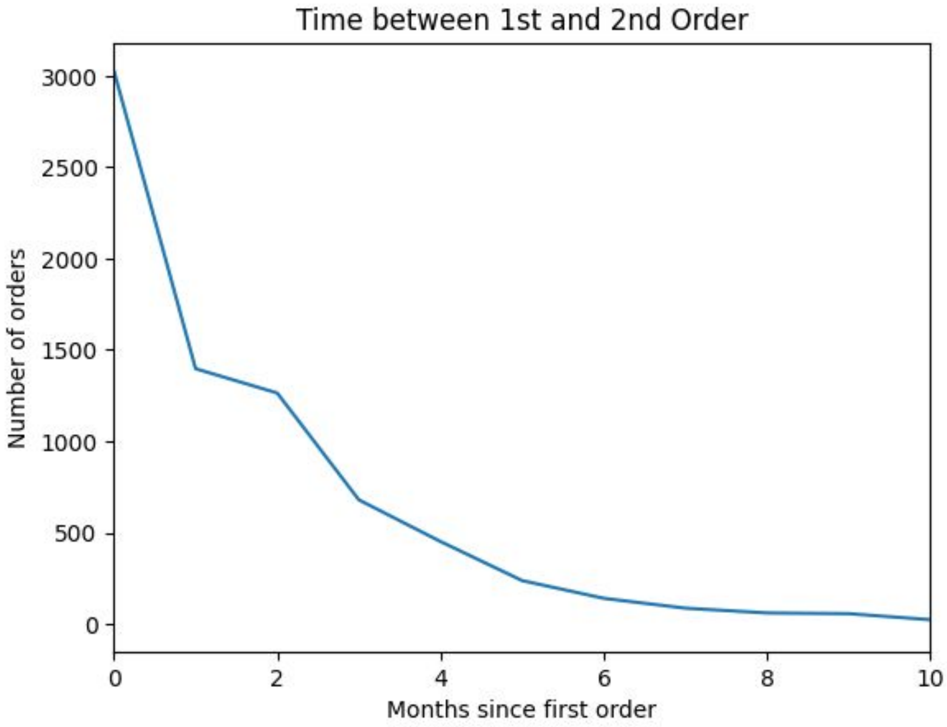
Feb-2021 and Feb-2022 saw strong uptick in retention across all cohorts

The assumption is that this is being driven by marketing activity. It could also be a seasonal trend within skincare but this seems unlikely

Recommendation:

- Understand what drove this specific behavior in those 2 months and attempt to replicate it more often

Users are most likely to purchase again straight after first purchase



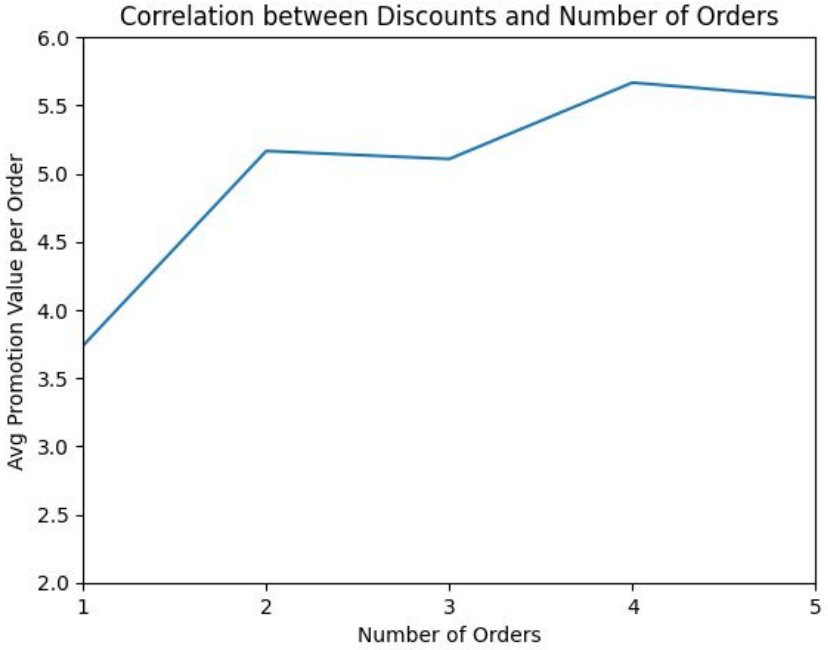
Most 2nd purchases happen in the same month as the first purchase.

Slight bump seen in natural decay of at 2 months. This might be related to length of treatment/replenishment emails.

Recommendations:

- Investigate what is driving users to purchase again in the same month. Is it the same product or a different one?
- Target users whilst they are still “warm” straight after they have made their first purchase
- Investigate the 2 month bump to see if there is anything we can learn.

Discounts are correlated to more repeat orders



Customers with more orders tend to have a higher discount. There is a particularly large jump between those who have purchased once vs twice.

This could suggest that by increasing the discounts we will get more orders and improve retention.

Recommendation:

- Test increasing discounts - particularly for one-time purchasers who might be at risk of not making a second ie they haven’t purchased for a while

Summary of Recommendations

Actionable recommendations:

1. Target users whilst they are still “warm” straight after they have made their first purchase
2. Test increasing discounts - particularly for one-time purchasers who might be at risk of not making a second ie they haven’t purchased for a while

Investigations for wider team:

1. Investigate high performing vs low performing Clinics to understand what is driving the different behaviour
2. Investigate why the older clinics are performing better than the newer clinics
3. Understand what drove the improvement in retention across all cohorts in Feb 2021 and Feb 2022 and attempt to replicate it more often

Further analysis (requires more data)**:**

1. Investigate what is driving users to purchase again in the same month (is it the same product or a different one?)
2. Investigate the bump in second orders 2 months after the first order (is there a CRM campaign?)
3. Investigate customers who had over 100 orders

Appendix 1: Assumptions and Data Quality Issues

- **Customers with over 100 orders**
  - There were a number of customers who had over 100 orders - which appeared to spike over 1-2 months
  - They were excluded from the analysis but we should further investigate to understand whether:
    - These are legitimate orders and the service is being used in a way we didn't expect
    - Or there is a data quality issue - which is potentially affecting the whole dataset? Possibly fraudulent orders?

- **Calendly set as Practitioner Onboarding Year**
  - These were ignored

- **Duplicated rows**
  - Removed from data

- **Unique_id_clean**
  - Assumed that this was a unique customer ID

- **Refunds**
  - Assumed that records are updated when order is refunded - ie a new row is not created just for the refund

Improving Cadence and Accuracy of Cohort Reporting

Executive Summary

To update the monthly cohorts dashboard currently takes x hours, is a manual process and as such is only updated once per month. Additionally, there are known data quality issues and the current process opens

We can solve this by:

- implementing a data stack to automatically extract, transform and visualise the data
- pushing the Customer ID and Order ID into Stripe

This should cost somewhere in the region of £200 per month - and should take 2 months to set up.

This will enable us to:

- Reduce manual effort to near 0 hours
- Schedule the dashboard to be updated daily
- Ensure accurate data

Additionally, this architecture will scale to other use cases enabling fast and efficient analysis and reporting going forwards.

Recommendations

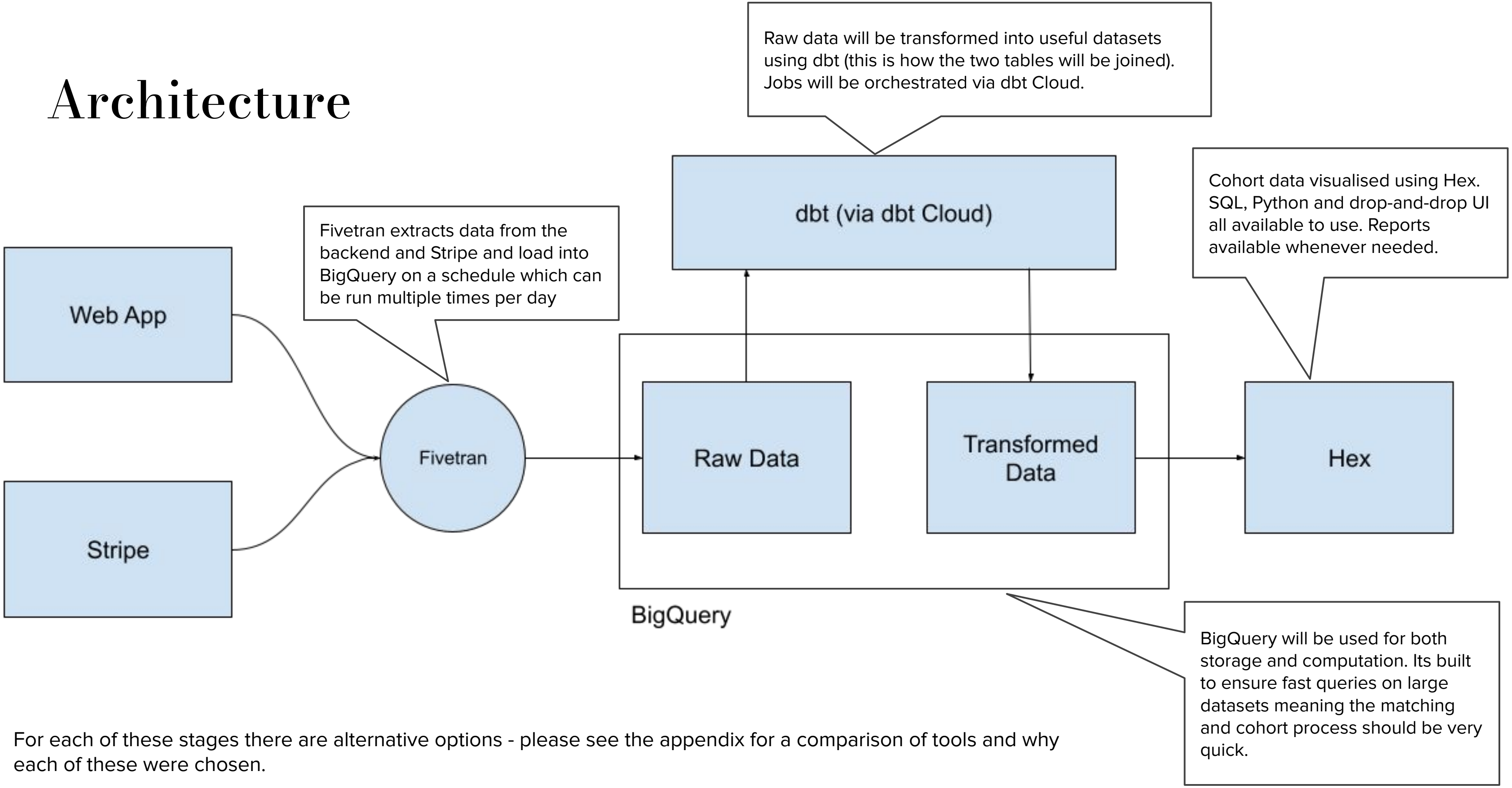
1. **Define Customer ID and Order ID centrally and push into 3rd party systems ie Stripe**
    - By using consistently across all tools we have an easy method for matching records
    - This will increase accuracy by removing processes based on top of flaky matching processes

1. **Extract data using an automated tool**
    - Automated process will be much faster - and can happen in the background
    - Data can be extracted much more often meaning reports can be updated at a much higher cadence

1. **Load data into a cloud database**
    - This will enable quick and easy transformation of data
    - It also removes security/privacy/legal risk of having sensitive customer data on local machines

1. **Implement a visualisation tool**
    - Enables anyone in the business to view the reports whenever they want
    - Removes manual steps of having to run

Architecture



Fivetran extracts data from the backend and Stripe and load into BigQuery on a schedule which can be run multiple times per day

Raw data will be transformed into useful datasets using dbt (this is how the two tables will be joined). Jobs will be orchestrated via dbt Cloud.

Cohort data visualised using Hex. SQL, Python and drop-and-drop UI all available to use. Reports available whenever needed.

BigQuery will be used for both storage and computation. Its built to ensure fast queries on large datasets meaning the matching and cohort process should be very quick.

For each of these stages there are alternative options - please see the appendix for a comparison of tools and why each of these were chosen.

This decision should also be taken in conjunction with the engineering team as there could be cost savings down depending on the technologies they use. For example - if the rest of the business uses AWS it might make more sense from a cost perspective to use Snowflake on AWS, rather than BigQuery.

Costs

	Cost per Month (£)	Comments
Fivetran	0	First 500k rows per month free
BigQuery	<10	Analysis: first TB free Storage: negligible at these volumes
dbt Cloud	80	15,000 successful model builds per month
Hex	90	
TOTAL	**180**	

Costs are based on the specific task and use case outlined

However, each of the tools will scale beyond this use case relatively easily. Some costs will scale whilst others will see a step change as they enter new tiers.

Forecasted volumes are based on 80% growth (as seen over last 12 months)

Not included are internal development costs related to pushing the Customer ID and Order ID into Stripe

Roadmap and Estimated Timeline*

Stage	Tasks	Owners	wk 1	wk 2	wk 3	wk 4	wk 5	wk 6	wk 7	wk 8
Fix metadata in tools	Push IDs into Stripe	Engineering	🟦	🟦	🟦	🟦	🟦	🟦		
Procure Tools	Negotiate Terms	Procurement/Finance	🟦	🟦	🟦	🟦				
	Review contracts	Legal	🟦	🟦	🟦	🟦				
	Check security of tools	Security/Engineering	🟦	🟦	🟦	🟦				
	Set up billing	Procurement/Finance	🟦	🟦	🟦	🟦				
Provision and set up tools	BigQuery	Analytics with support from Engineering					🟦	🟦		
	Fivetran	Analytics with support from Engineering					🟦	🟦		
	dbt Cloud	Analytics					🟦	🟦		
	Hex	Analytics					🟦	🟦		
Build pipeline	Extract and load data from the backend	Analytics							🟦	
	Extract and load data from Stripe	Analytics							🟦	
	Load historical data that has already been through matching process	Analytics							🟦	
	Build models and join two datasets	Analytics								🟦
	QA data	Analytics								🟦
	Schedule dbt job	Analytics								🟦
	Create visualisation in Hex	Analytics								🟦

*Timelines are very rough estimates and would need to be sized by the relevant team

Summary

We can improve the accuracy and cadence of the cohort dashboard by implementing a new architecture comprised of:

- Fivetran
- BigQuery
- dbt Cloud
- Hex

Additionally, we will need to do some engineering work to push some IDs into Stripe.

The new tooling will cost just under £200 per month (for this very simple use case).

We believe this will take around 2 months for us to implement.

Additional benefits include:

- No more manual processes
- Fresh reports can be accessed whenever required
- Reduced security/privacy concerns
- Scalable architecture ready for other use cases

Appendix: Architecture Design Choices

Why don’t we just run a python job in the cloud?

As the volumes of data grow a SQL environment will scale much better than a Python one. A framework such as dbt will also make it much easier to handle the complexity of multiple data sources.

Additionally, SQL is the standard within the Analytics profession making it much easier to hire someone with SQL skills than python.

Why have you chosen to use a tool such as Fivetran when the Engineering team could build the ingestion pipeline themselves?

Whilst it is possible to build an tool for ingestion - it is a lot of work meaning longer lead times and will require ongoing maintenance. Ultimately, the total cost of ownership will be much lower.

Why are we using a visualisation tool - rather than visualising in python?

By having this tool we are able to easily schedule dashboard rebuilds as well as having a simple method for presenting/distributing any visualisations. The tool we have chosen does have the capability to run python so visualisations could still be built in python.

Additionally, a tool such as this could help facilitate a self service environment at some point in the future.

Why don’t we just increase processing power to speed up the python jobs - rather than pushing data into Stripe?

Speed is not the only issue. By not having an explicit match between records we open ourselves up to potential issues in the future.

Appendix: Comparison of Tools

Ingestion	**Fivetran** - Cost: £0 (free tier) - Lots of data sources	**Stitch** - Cost: $100 per month - Smaller selection of data sources			
Database	**BigQuery** - Cost: £0 (free tier) - Minimal admin required - Can handle queries of varying sizes with ease	**Snowflake** - Cost: $40 per month (pricing structure complex) - Pay for compute resources, rather than data processed (more complex to manage costs and scale)	**Redshift** - Cost: unclear (complex pricing) - Requirement to size compute resources adds extra complexity and has potential to lock database if done wrong		
Transformation	**dbt** - Cost: free (open source) - Industry standard	**DataForm** - Cost: free - Integrates well with BQ			
Orchestration	**dbt Cloud** - Cost: $100 per month - Native tool for dbt Core - Comes with IDE for developers - Semantic layer also an option in the future	**Airflow (Cloud Composer)** - Cost: $100 per month (but pricing structure unclear) - Tried and tested tool - lots of support - Can be used to schedule python as well as SQL			
Visualisation	**Hex** - Cost: $108 per month - Great for collaboration and reporting	**Mode** - Cost: unclear (need to talk to sales) - Great for collaboration and reporting	**Tableau Online** - Cost: more than £2k per month	**Looker** - Cost: large - Large overhead to implement	**Looker Studio** - Cost: free - Visualisations are limited

Appendix: Assumptions

1. Sumup data is being recorded in Stripe - and there is a method for us to push metadata into Sumup too LINK

1. The current python workflow requires a local machine to be used meaning:
   a. possible security/privacy/legal issues due to customer data being on a laptop
   b. the process requires manual intervention

1. The business will be interested in other analysis beyond cohorts going forwards