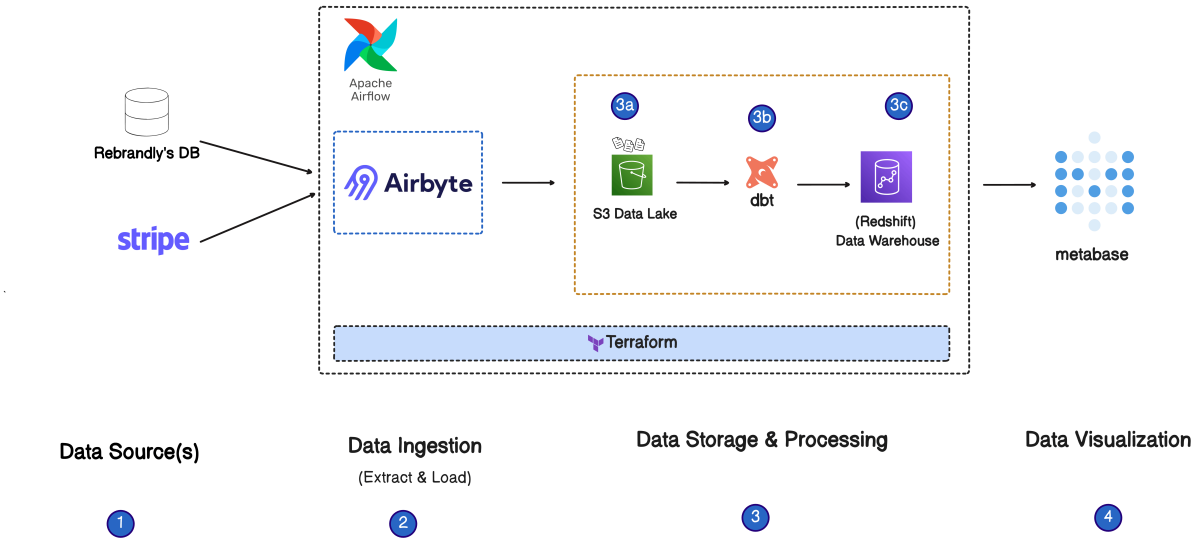
Rebrandly POC Architecture

Overview

Data Architecture

Below is the high-level proposed Data architecture platform for Rebrandly.f



Full Data Workflow: Stripe Data to Redshift

1. **Data Ingestion (Stripe to S3):** Airbyte extracts data from Stripe's APIs and delivers it to an S3 bucket as Parquet files.
2. **Orchestration (Airflow):** The whole data flow is orchestrated using a self-hosted Apache Airflow. This tool schedules and orchestrates the end-to-end data pipeline and manages tasks like triggering Airbyte jobs, S3 file checks, and downstream transformations.
3. **Data Transformation (S3 to Redshift):** dbt processes the raw data from S3, applying transformations and modelling to create optimized tables (e.g., facts and dimensions) in Redshift.
4. **Infrastructure Automation (Terraform):** Terraform provisions and manages resources, including S3 buckets, Airflow instances, Airbyte connectors, Redshift clusters, and Metabase deployments, ensuring consistency across environments.

Layers

1. Data Source(s)

Stripe Stripe is a payment platform that Rebrandly uses to manage subscription payments. It provides access to a wide range of data, including customers, billing, charges, invoices, etc. Refer to the Stripe Data Schema for details on the available data sources.

Rebrandly's Database This online transaction processing (OLTP) database stores the transactional tables essential for Rebrandly's service operations. It contains user details (businesses or individuals) and other relevant tables.

2. Data Ingestion

There are various data ingestion tools; however, each has pros and cons. For a scaleup like Rebrandly, I highly recommend an ingestion platform with many connectors to different sources & destinations, provides open source and cloud offerings, and has a large support or community.

The data ingestion tool of choice I recommend is Airbyte.

Airbyte

Airbyte is a data ingestion tool offering 550+ connectors. This implies that it allows the functionality of ingesting from various data sources, such as different cloud storage providers, SAAS platforms, Databases, and more.

It offers both cloud and open-source options. The latter makes it easy for Rebrandly to **manage, **as a self-hosted option means more ownership. Because it is open-source, it can also be used for POCs, as it can be easily deployed on a virtual machine in AWS.

It also has great support, with a rich community of about 23,000+ users on Slack and GitHub repo stars of around 10,000+, and it is trusted by leading brands such as Calendly, SIEMENS, and many more.

3. Data Storage & Processing

When choosing a Data storage layer, the most important considerations are scalability, ownership, cost efficiency, and avoiding vendor lock-in.

3a) S3 Data Lake

One of the most critical reasons for choosing a **data lake** for the raw layer is that it allows us to replay data easily without retrieving it from the source systems. This reduces the load on source systems, which is particularly important in high-traffic production environments or systems with performance constraints.

In the context of audits, having the ability to replay data from the raw layer ensures that we can reproduce historical states and validate transformations or business processes without impacting the source. This capability is essential for:

- **Regulatory Compliance:** Providing full data traceability and reproducing datasets for external or internal audits.
- **Data Quality Assurance:** Comparing the original raw data with processed outputs to ensure transformations are accurate.
- **Disaster Recovery:** Reprocessing data in the event of pipeline failures or downstream errors. The data lake is a reliable, low-cost repository that supports scalability and minimizes operational risks by decoupling the raw data storage from the source systems.

I recommend using S3 as a Data lake for the first storage layer. All data copied from the ingestion platform (Airbyte) is stored as files in an S3 bucket.

I chose S3 as the landing / raw zone storage layer because of the following:

- As Rebrandly's data volume grows, the data architecture will require a highly scalable storage system, and S3 can store up to exabytes of data.
- Data can be structured and unstructured; a service like an S3 bucket can handle this.
- S3 provides different storage classes for different use cases (data lakes, backups, data archives, etc) and still maintains high performance at an affordable cost.
- It allows seamless integration with other AWS services.
- It allows data to be stored in open, standardized formats (Parquet, ORC, AVRO, etc.) best suited for big data applications.
- Since it allows data storage in different formats, it allows easy integration with other platforms and tools.
- It provides various control, encryption, and security levels to help securely store data.

3b) dbt (Data Build Tool)

It was chosen because it aligns well with modern data engineering principles and the project's specific needs. Its ability to transform data in the warehouse using SQL and version-controlled workflows makes it ideal for building scalable, reliable data pipelines. Here is a detailed analysis of why I chose dbt:

- dbt uses SQL, a widely known and accessible language for data analysts and engineers.
- It reduces the learning curve compared to tools requiring custom coding languages, implying that new teammates can learn quickly.
- dbt promotes modular coding with reusable models and macros, making building, maintaining, and scaling transformations easier.
- Dependencies between transformations are automatically managed.
- It integrates seamlessly with Git, enabling version control, code reviews, and team collaboration.
- It supports built-in data testing for schema validation, uniqueness constraints, and data integrity. This ensures high-quality and trustworthy data outputs.
- It auto-generates documentation with data lineage, allowing stakeholders to understand how data flows through the pipeline. This increases transparency and simplifies debugging.
- It supports a variety of data warehouses, provides flexibility, and avoids vendor lock-in.
- It has a strong open-source community that provides plugins, best practices, and resources for support.

3c) Redshift (Data warehouse)

Redshift was selected for its seamless integration into Rebrandly's existing ecosystem, leveraging its compatibility with AWS services. It offers excellent scalability, allowing Rebrandly to efficiently handle growing volumes of data, from terabytes to petabytes. With its columnar storage and massively parallel processing (MPP), Redshift performs complex analytical queries well. Redshift's ability to connect with various ETL tools and visualization platforms ensures smooth data workflows and accessibility. Its cost-effectiveness and elasticity also allow for dynamic scaling, aligning with Rebrandly's data processing needs.

4) Metabase (Data visualization)

Metabase stands out for its simplicity and user-friendly interface, making data visualization accessible to all team members, including non-technical users. Its open-source nature reduces costs while offering flexibility in deployment, whether on-premises or in the cloud. Metabase integrates seamlessly with Redshift, enabling real-time insights from Rebrandly's data warehouse. The platform's no-code query builder and SQL support cater to diverse user needs, ensuring quick and easy data exploration. By empowering teams to create and share dashboards independently, Metabase fosters a data-driven culture at Rebrandly.

Evaluation

Below is a comparison of why I chose specific tools.

Why Airbyte?

Feature	Airbyte	Fivetran	Stitch	Estuary
Deployment Options	• Self-hosted: Open-source (free). • Cloud: Paid managed offering.	• Fully cloud-based only.	• Fully cloud-based only.	• Self-hosted: Open-source (Flow) • Cloud: Paid managed offering.
Cost	• Free for self-hosted. • Cloud: $250 for 10 GB of data with 10 rows ingested.*	• Subscription-based pricing depends on (Monthly active rows) MAR and depends greatly on the source.	• Tiered pricing based on MAR (Monthly active rows) starts at ~$100/month for lower usage tiers.	• Open-source version is free. • Cloud pricing is usage-based and varies based on the number of sources and data size

Ease of Setup	Moderate: Requires some technical expertise for self-hosted deployment.	Easy: Plug-and-play with minimal setup.	Easy: Simplified interface for quick integration.	Moderate: Open-source setup requires technical effort; the cloud version is more straightforward.
Connector Availability (integration)	• 550+ connectors, with the possibility of building custom connectors.	• 150+ high-quality connectors with prebuilt transformations.	• 130+ connectors mainly focused on common SaaS platforms.	• Limited connectors compared to Airbyte and Fivetran, but still growing.
Custom Connector Support	• Excellent: Easy to build custom connectors in Java or Python.	• Limited: Requires support from Fivetran's team for custom needs.	• Limited: Custom connectors are not natively supported.	• Moderate: Custom connectors can be added with Flow's scripting capabilities.
Data Governance	• Moderate: The open-source version lacks advanced governance tools; the cloud has basic governance.	• Strong: Offers schema drift handling, metadata, and logging features.	• Moderate: Basic metadata and schema handling.	• Strong: Built-in governance features for managing real-time data pipelines.
Scalability	• Highly scalable with self-hosted deployments, depending on infrastructure.	• Highly scalable with managed cloud infrastructure.	• Scales well but is limited by Stitch's data volume tiers.	• Real-time pipelines are designed to scale with data size and complexity.

Vendor Lock-In	• Low: Open-source nature allows migration or modification.	• High: Proprietary platform with limited customization.	• High: Proprietary platform with limited self-hosting or migration options.	• Low: Open-source version allows flexibility; the cloud has some dependencies.

Takeaways: Airbyte allows more flexibiliy, hence I chose it.

Why DBT?

dbt (Data build tool) pricing

For data transformation, dbt has two offerings:

- Cloud
- Core (Self-hosted) The significant advantage of dbt cloud over dbt core is that it allows for hosting generated dbt documentation and scheduling dbt jobs. However, for Rebrandly, I will not recommend dbt Cloud because it costs around $100 per seat (user). Documentation and scheduling jobs can be self-hosted for less than $100 using Amazon ECS and documentation served easily by self-service.

Feature	dbt Cloud	dbt Core
Deployment	Fully Managed Service: Hosted and maintained by dbt Labs.	Self-hosted: Open-source and can be deployed on your AWS cloud services.
Setup & Maintenance	Easy Setup: Minimal configuration is needed, as dbt Cloud is pre-configured for ease of use.	Manual Setup: This requires the setup of your infrastructure, such as servers and storage.
Pricing	Subscription-based: starts at **$100/mo/seat**	Free: dbt Core is open-source and free to use. However, infrastructure costs are on you.
User Interface	Web UI: Provides a user-friendly web interface for managing models, jobs, and documentation.	CLI (Command Line Interface): Users interact with dbt Core through the command line, requiring familiarity with terminal commands.
Job Scheduling	Built-in Scheduler: Allows scheduling of dbt runs directly from the dbt Cloud platform.	Manual Scheduling: Needs third-party scheduling tools (e.g., ECS, Airflow, cron jobs) to automate runs.
Documentation	Built-in Documentation: Includes automatic model documentation and visualizations.	Requires Setup: Documentation must be manually generated through dbt commands or third-party integrations.

Why Metabase?

1. **Cost-Effectiveness** Metabase's open-source version gives small teams and businesses much control. Although paid plans, like Metabase Cloud, are available, Metabase is more affordable than Tableau or Power BI. A Tableau license costs up to, while Power BI's full-feature access can get expensive as teams scale. Metabase’s pricing aligns with lean and growing organizations.
2. **Ease of Use** Metabase allows non-technical users to explore data quickly. Its query builder is code-free, and SQL support is available for advanced users. Tableau has a steeper learning curve, and Power BI requires familiarity with DAX, which can be complex.
3. **Speed and Simplicity** Setting up Metabase is quick, taking just minutes, regardless of whether you're using Docker, local hosting, or the cloud. It seamlessly connects to Redshift and other Data warehouses.
4. **Self-Serve Analytics** Metabase enables users to quickly generate and share visualizations independently without needing much support from a Data Engineer. It provides straightforward data models and versatile exploration tools. In contrast, Tableau and Power BI require excellent technical expertise to prepare datasets and create dashboards. It also facilitates broader data access, empowering anyone to engage with data seamlessly.
5. **Open Source and Flexibility** As an open-source tool, it provides transparency and customizability and avoids vendor lock-in. This means that Rebrandly can host and adapt it to their specific needs, which is impossible with proprietary options such as Tableau and Power BI.