# **Final Project**

Team: Fantastic Four plus Ethan

A drawing of a face

Description automatically generated

|  |  |
| --- | --- |
| Team Members:  Ethan Hodys  Nitheesha Daka  Bhuwaneshwari Iyer  Siddharth Menon  Sreejith Sreekumaran Nair | Software:  [Snowflake Inc.](http://www.snowflake.com) |

# Scope and Importance

# Use Case

# Software details

# Key alternatives

# Software recommendation

# Future developments

# References

**SCOPE & IMPORTANCE**

A major problem for BI and Data Analytics has been that improvements in data storage has far outpaced the ability to access and process the data. Snowflake is one of a few technologies with unique, fully modern architectures that drastically reduces both the latency and throughput of data analysis for data engineers, data scientists and developers. From the Snowflake website, it is described as “a near-zero maintenance platform delivered as-a-service which features compute, storage, and cloud services layers that are logically integrated but scale independent from one another, making it an ideal platform for many workloads.”

While impressive in itself, its true value to organizations is the significant cost reduction in maintaining a single system which requires almost no maintenance and can be accessed from numerous clients via a cloud-based service. It for this reason we would like to present Snowflake to the class via our final project. We will be using data simulated from live Micro-Market kiosks from 365 Retail Markets (<https://365retailmarkets.com/>)

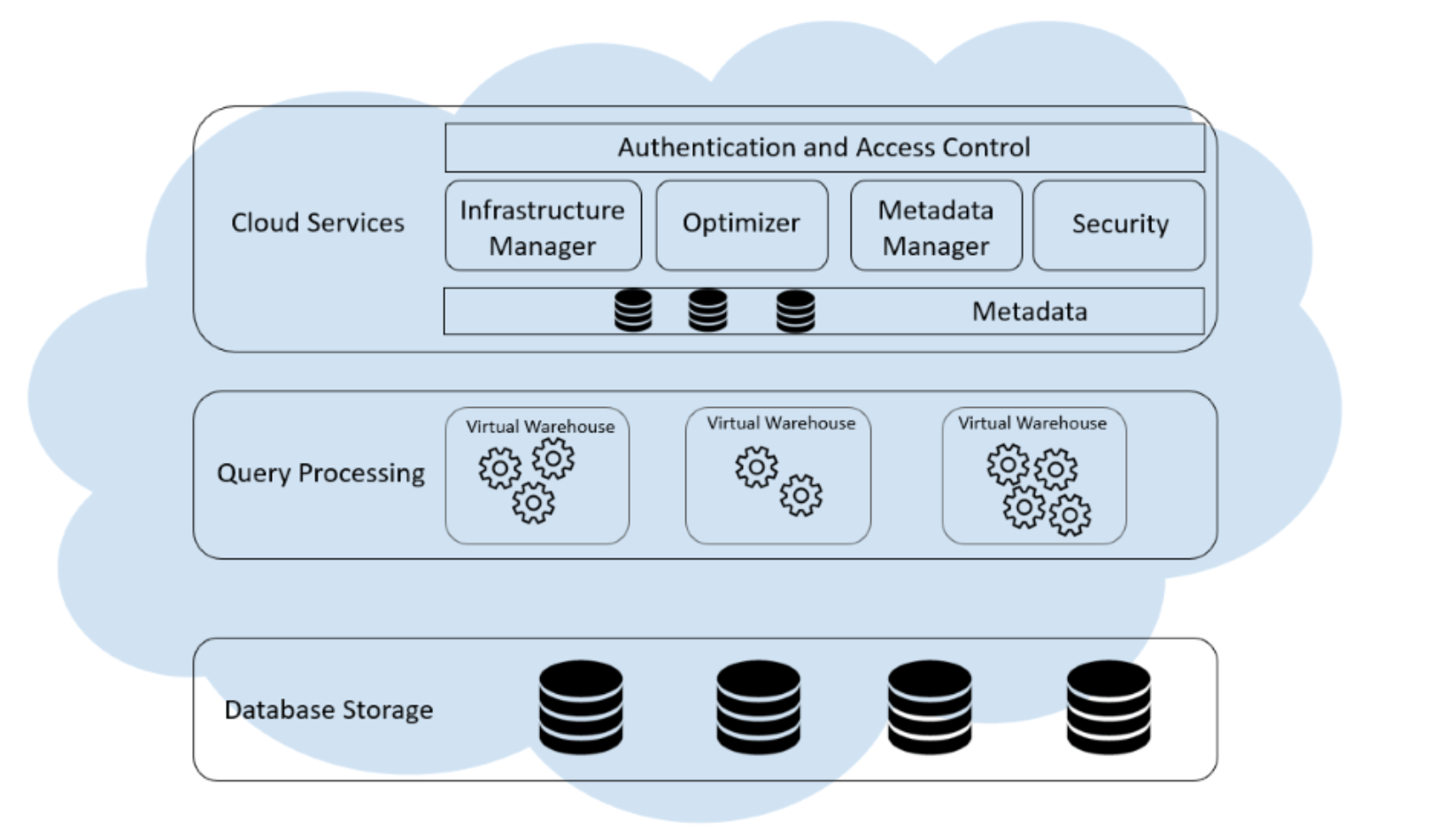
**USE CASE**

With the growing technological advancements, the need to collect, structure and analyze data in order to provide analytics to optimize the performance of the business has increased. A data warehouse is used to store and maintain the bulk of data. Although there are multiple warehouses available, most of them have drawbacks such as lacking some features. For Example, there are some warehouses available which are excellent at performing data manipulation operations but cannot handle big data. Another problem is the infrastructure. Few data warehouses expect the clients to buy storage and servers irrespective of what they need. Traditional warehouses are facing problem in simple, economical and competent service that the swiftly adapting businesses need.

Leading companies in finance, media, healthcare and manufacturing are all using Snowflake to transform their organization into a data-driven enterprise. What sets snowflake apart is that it uses a tiered system approach i.e. storage of the data and processing, querying of it is done distinctly. Operations like data manipulation and processing can be grouped depending on the needs of the organization. Its system supported both structured as well as unstructured data. Also, snowflake Is a cloud-based database. Users don’t need to pay or install any kind of software and can pay as and when the need arises. Snowflake provides a data warehouse that is faster, easier to use, and far more flexible than traditional data warehouse offerings.

Our project simulated a real-world project at one of the team’s current company: 365 Retail Markets in Troy Michigan. 365 Retail Markets is a producer of self-service kiosks for purchasing various snacks, drinks, and other items typically found in office and retail break rooms. While strong growth over the past 2 -3 years has been great for revenue it has also highlighted weaknesses in the data service layer as customer load increases as well as the number data and analytics savvy customers are integrated. The specific business case challenging 365 was maintaining a legacy system using Sql Server with Sql Server Analysis Service (SSAS) in conjunction with the current generation kiosks running against MariaDB (MySql clone) with only integrated web application reports. Consequentially a new project was initiated to product a single Data Warehousing and BI solution for all internal and external customers. The decision was made to use Snowflake for the Data Warehouse implementation but the larger implementation challenge was the migration of the data from two different platforms on two different sub-nets with near real-time replication int the Data Warehouse.

**SOFTWARE DETAILS**



Snowflake is a parallel processing (MPP) database that is fully relational, ACID compliant, and processes standard SQL natively without translation or simulation. It was designed as a software service that can take full advantage of cloud infrastructure, while retaining the positive attributes of existing solutions. Snowflake architecture has the hybrid of both shared-disk and shared-nothing architecture so that it provides users with the best from both. The 3 main components of the Snowflake architecture are:

**Database storage:** When data is loaded into Snowflake, Snowflake reorganizes that data into its internal optimized, compressed, columnar format. Snowflake stores this optimized data in cloud storage. Snowflake manages all aspects of how this data is stored in the organization such as file size, structure, compression, metadata, statistics, and other aspects of data storage are handled by Snowflake.

**Query Processing:** The processing of data is done in this layer which is done by the Virtual Warehouses where each one is an independent compute cluster that doesn’t share resources with other virtual warehouses. As a result, each warehouse doesn’t impact the performance of other warehouses.

**Cloud Services:**  The cloud services layer is a collection of services that coordinate activities across Snowflake like sessions, authentication, SQL compilation, encryption, etc.

**KEY ALTERNATIVES**

In the Data Warehouse category, Snowflake has a market share of about 1.7%. Other major and competing products in this category include: Microsoft Azure, IBM DB2, SAP Business Warehouse, Amazon Redshift, Apache Hive, Google Big Query, Oracle Autonomous Data Warehouse Cloud. Snowflake has a wide range of customers which include industries like Computer Software (41%), Information Technology and Services (9%) and Financial Services (6%) are the largest segments. The size of the industries includes 39% are small (<$50M), 21% are medium-sized and 35% are large (>$1000M).

Snowflake should be chosen over RedShift since it provides more robust support for JSON-based functions, includes wider variety of security and compliance options and has better automated database maintenance features.

The main differences between Snowflake and Google Big Query are that snowflake is better in terms of pricing, optimal performance, better scalability, more interactive and provides better security.

**SOFTWARE RECOMMENDATION**

Snowflake is highly preferred as although it is built specifically for the cloud, it’s designed to address many of the challenges found in older hardware-based data warehouses, such as limited scalability, data transformation issues, and delays or failures due to high query volumes. Here are some of Its strengths:

**Performance and speed:** Snowflake exceeds the traditional methods for executing data workloads. Efficient query optimization delivers answers in a fraction of the time and performance challenges can be addressed in seconds.

**Storage and support for structured and semi structured data:** Both structured and semi structured data can be combined for analysis and loaded into the cloud database without the need for conversion or transformation into a fixed relational schema. Thus, Snowflake automatically optimizes how the data is stored and queried.

**Seamless data sharing:** Snowflake’s architecture enables data sharing among Snowflake users. It also allows organizations to seamlessly share data with any data consumer. This functionality allows the provider to create and manage a Snowflake account for a consumer.

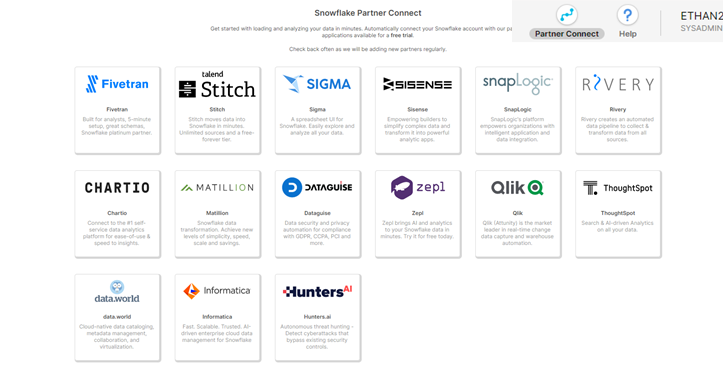
**Availability and security:** The system has high tolerance to failures with minimal impact. In case of hardware failure, Snowflake can continue operations and withstand the loss of availability. It is also extremely secure as data is encrypted at rest and in transit.

**PROJECT IMPLEMENTATION**

Below is the equivalent of a high-level block diagram showing the data flow. Legacy and Current applications are constantly streaming sales and other data to the databases which need to replicate those changes to the Data Warehouse against an SLA of a lag no greater than 30 minutes.

The 3rd party MariaDB vendor has implemented an older architecture of using HADOOP to pull the data from MariaDB and migrate it into Snowflake. This is a full snapshot replication and the Data Warehouse is inaccessible for 8 hours of the day as the process is executing. Our team member made the internal recommendation to use “Snowpipe” instead. From the documentation section of the Snowflake website “Snowpipe enables loading data from files as soon as they’re available in a stage. This means you can load data from files in micro-batches, making it available to users within minutes, rather than manually executing COPY statements on a schedule to load larger batches.” (Introduction to Snowpipe 2019) . Since adopting the use of Snowpipe, the migration time has been reduced from 8 hours to just 2 hours: a 75% reduction in execution time and a 25% increase to system availability. While a great improvement this is still a far away from achieving the stated SLA. Our project implementation has solved this problem and served as a proto-type / proof-of-concept and has been approved for adoption for 365 Retail Markets production environment.

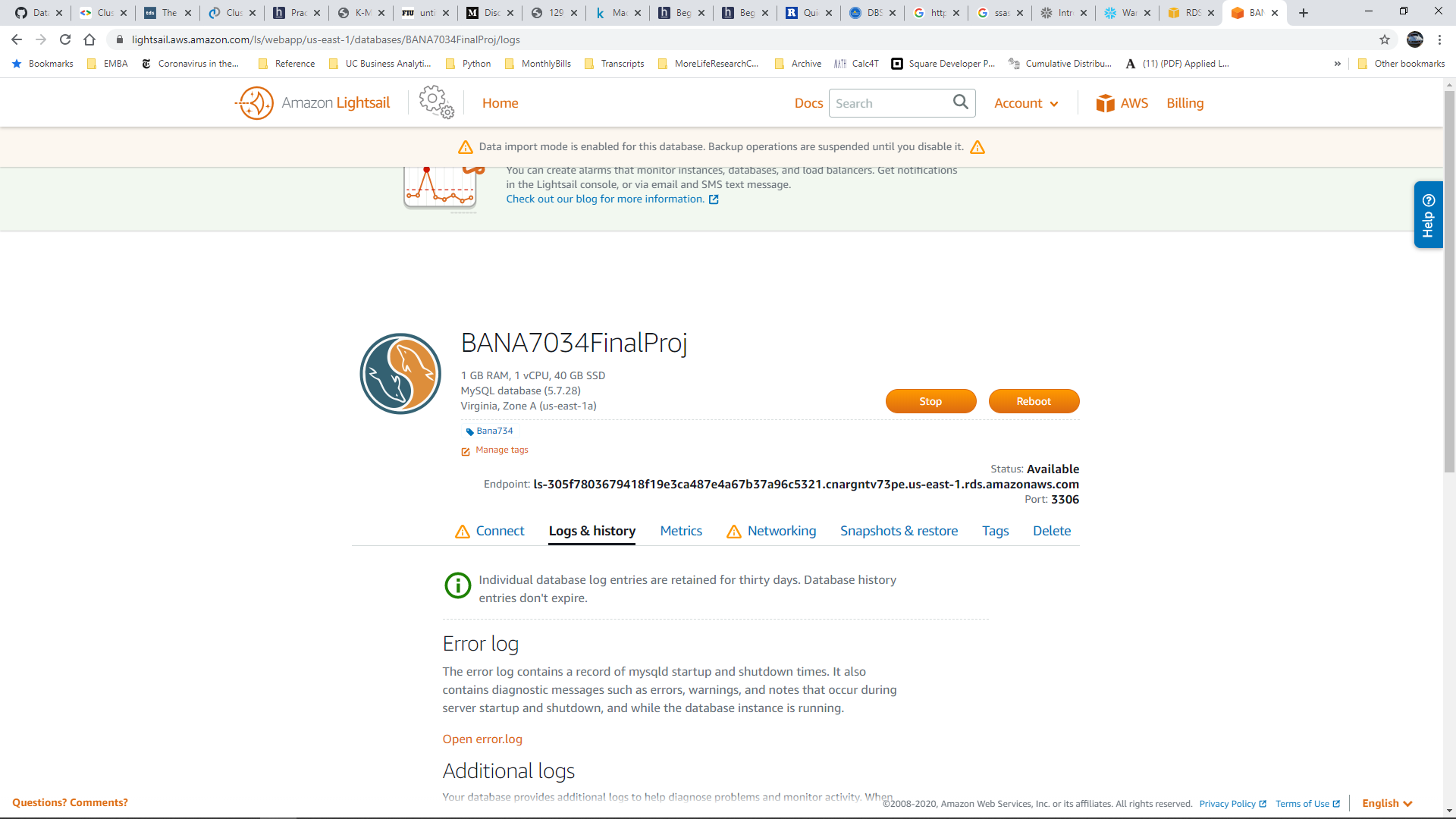
While by no means unique to Snowflake, one the advantages are the integration tools under the Partner Connect section of the web site. These integrations help with loading and analyzing your data in minutes. For our project we opted to create an account with Stitch to handle the data migration from our MariaDb instance to our Snowflake instance.



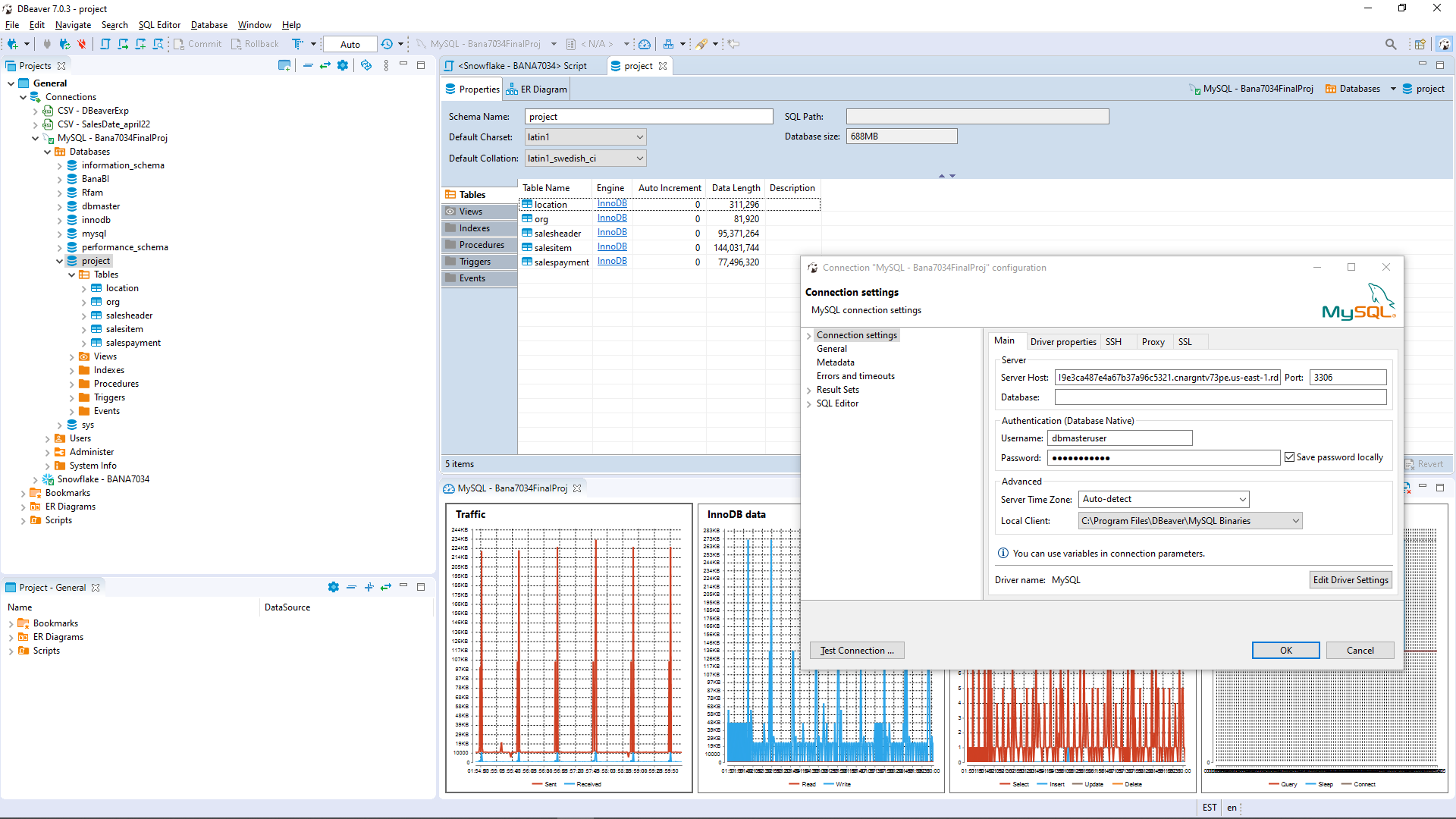
**PROCESS FLOW**

Below is the process flow for our implementation. This is better illustrated during the demo but is present here just for completeness. Following the process flow are an ordered set of screenshots for the respective technology in the process flow.

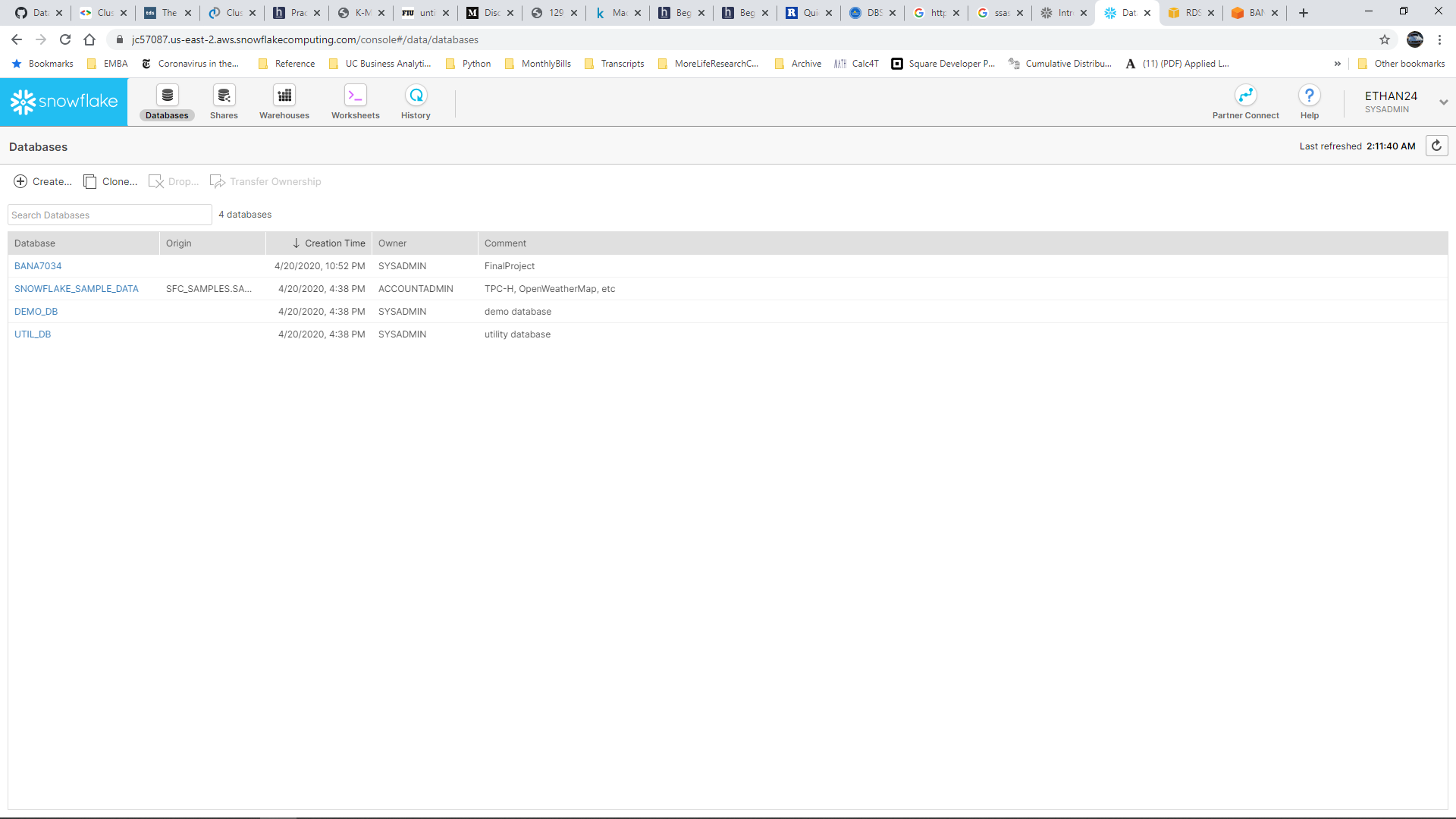
1. Amazon Lightsail implementation for MariaDb hosting



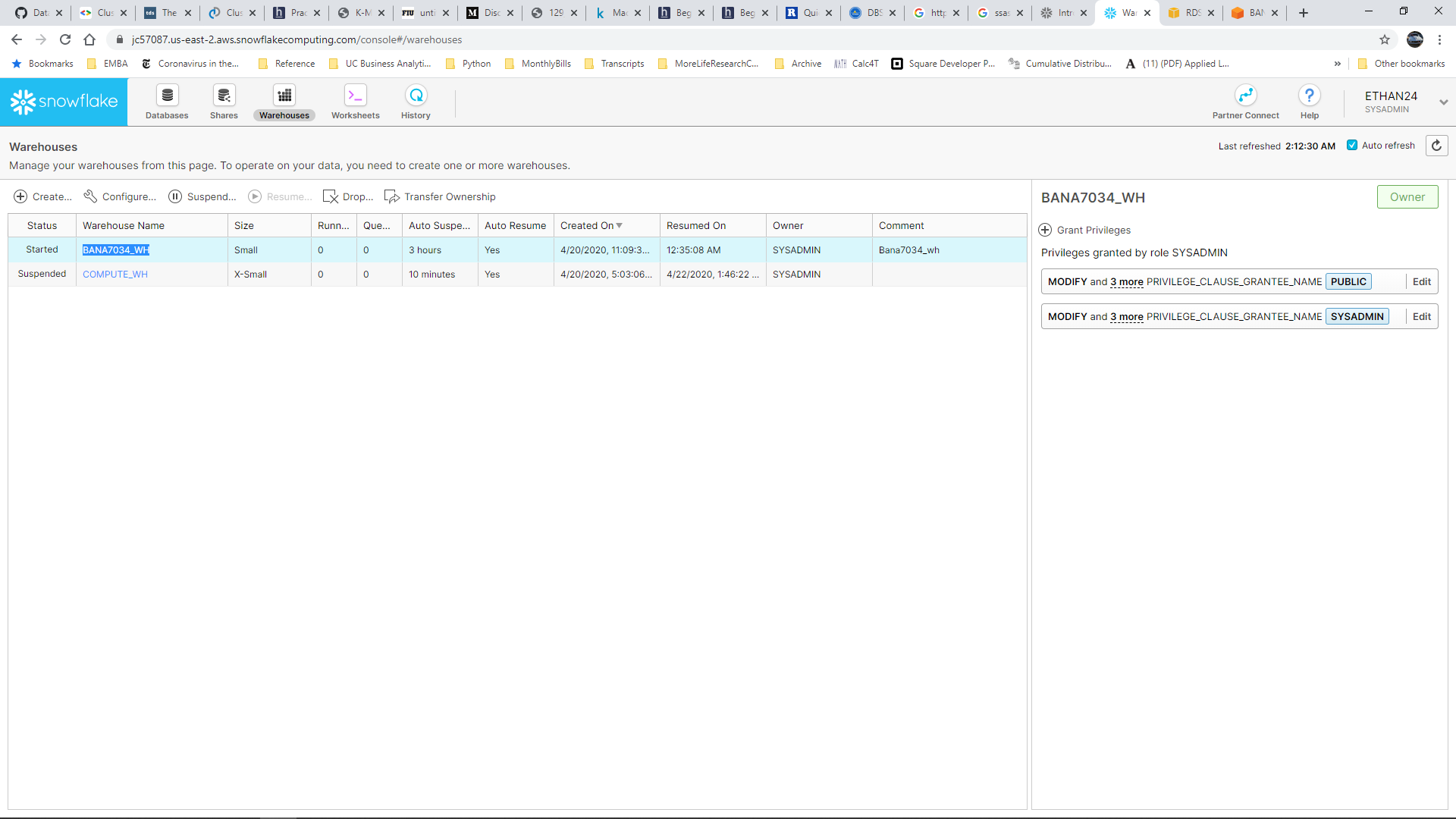
1. MariaDb implementation shown within DBeaver Ide
   1. Database name: Bana7034FinalProj
   2. Schema: project
   3. Server Host: AWS Lightsail URL
   4. Table: [org],[location],[salesheader],[salesitem],[salespayment]
      1. Sizes of tables can be sign in the middle of the screen shot (zoom for readability)
   5. Database size: 688MB. We migrated close to 1gb of actual sales data without any personal identifiers. This was imported via DBeaver. CSV connections can be sign in the upper part of the Projects tab on the left



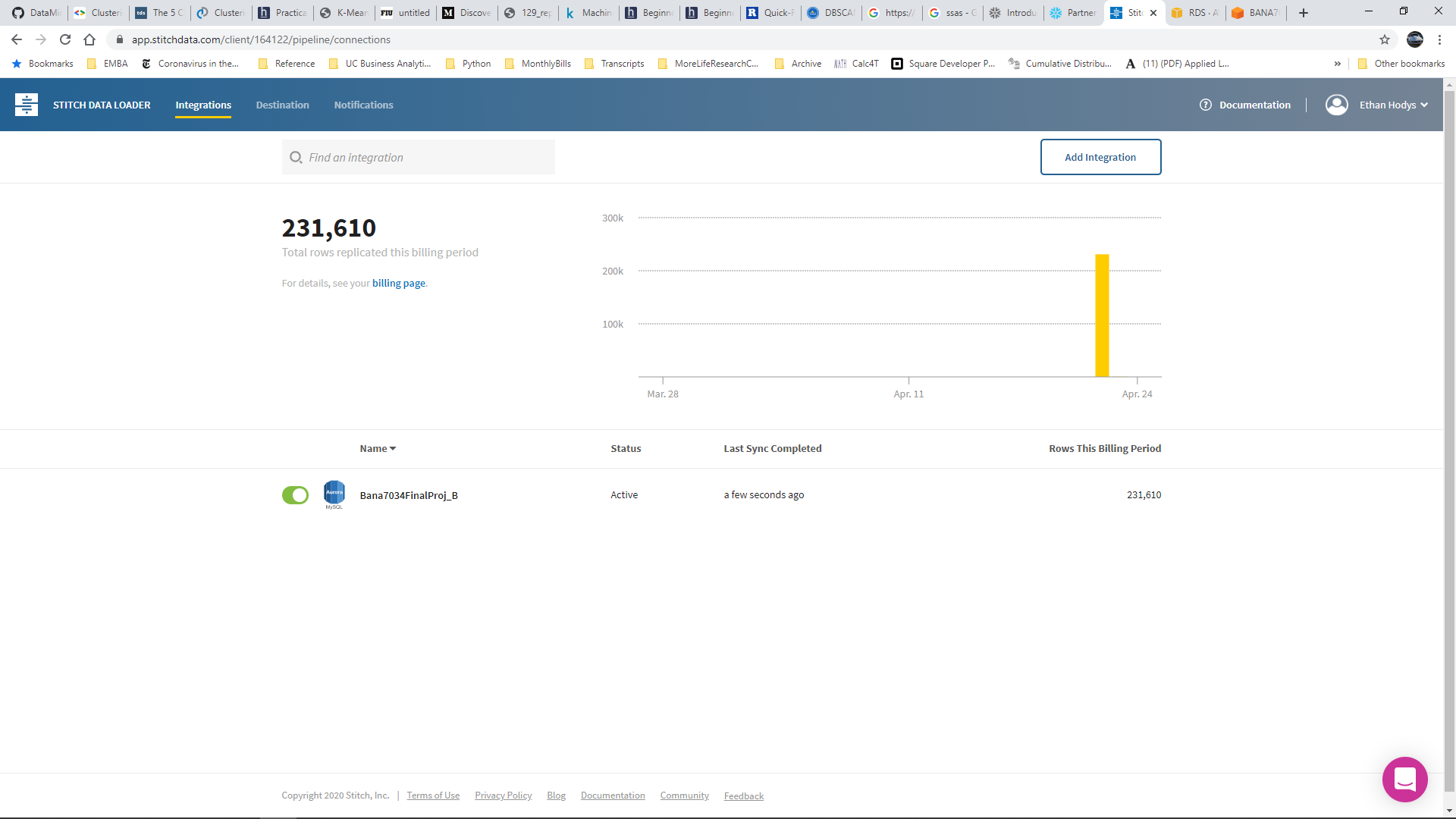
1. Snowflake
   1. Database

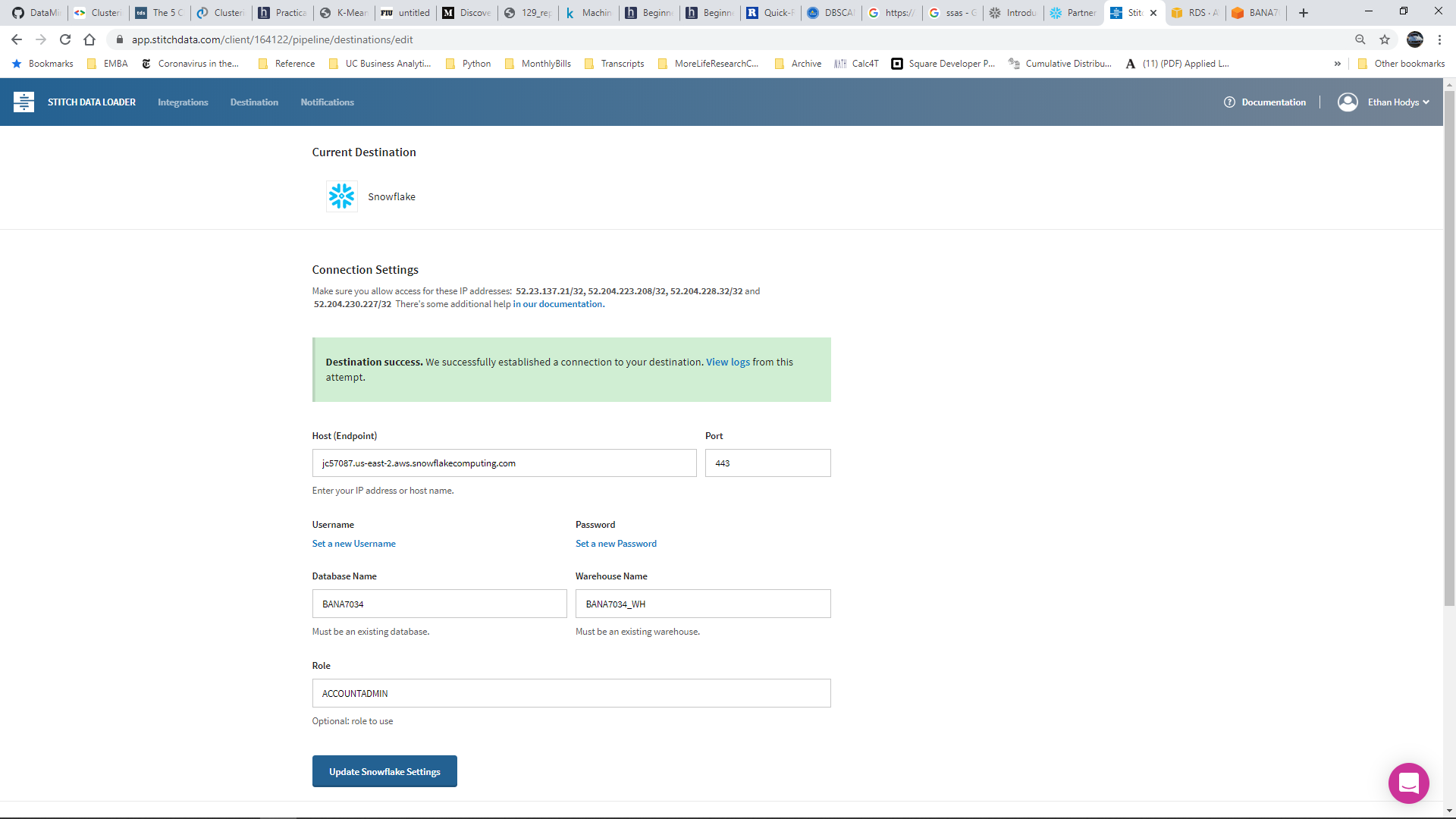


* 1. Warehouse

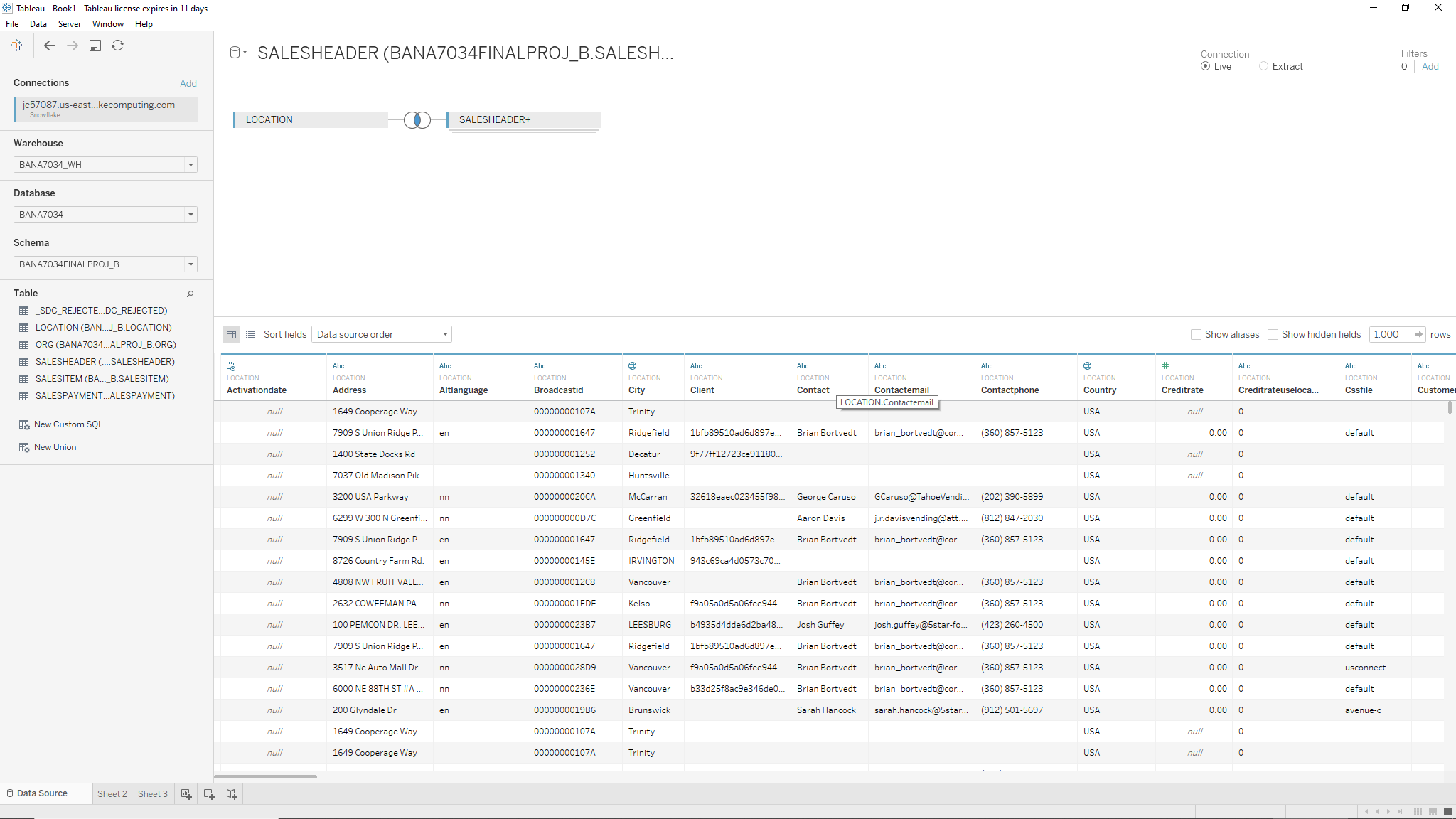


1. Stitch Data
   1. Integration : 231, 610 records migrated from MariaDb instance

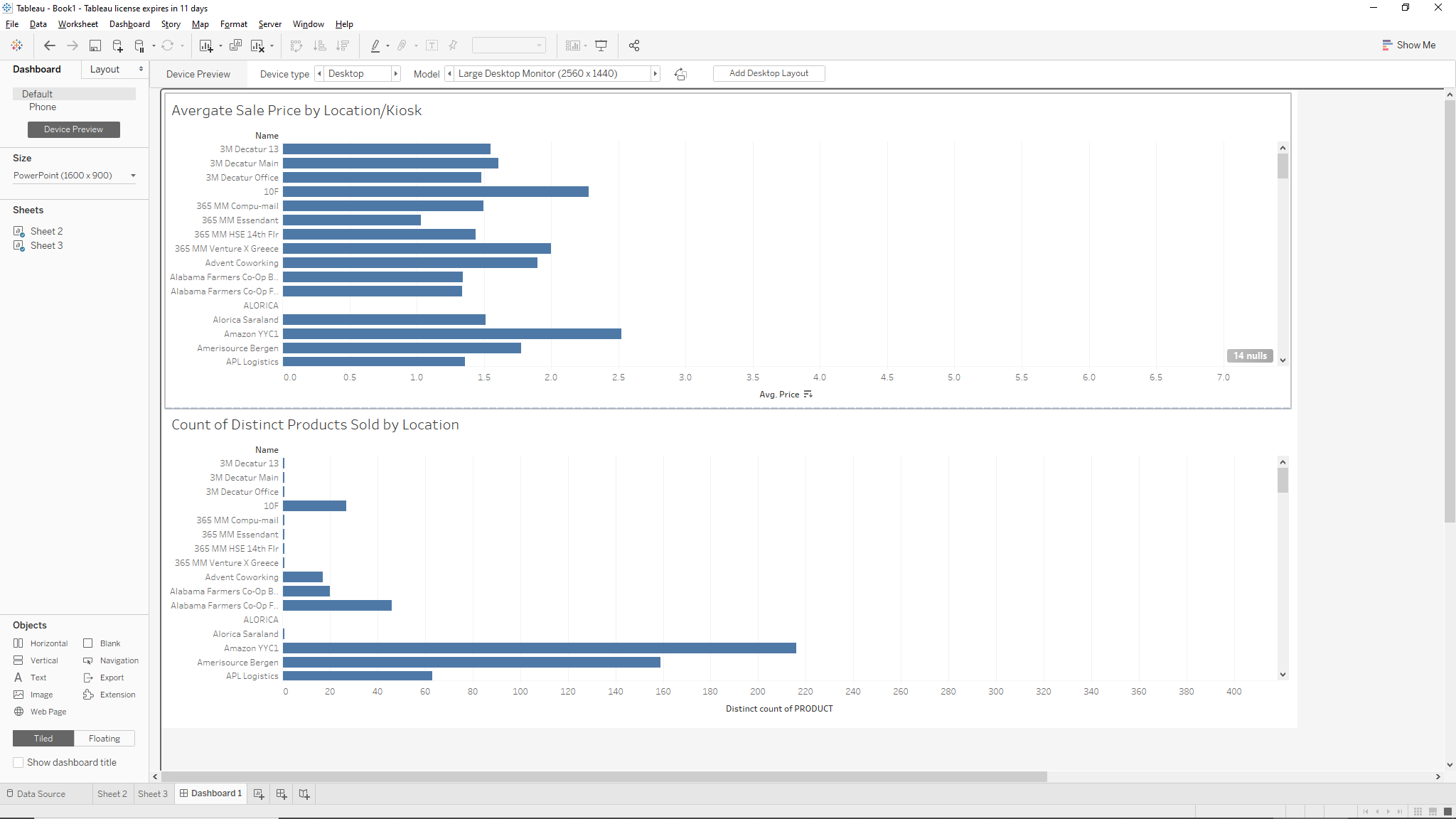


* 1. Destination 

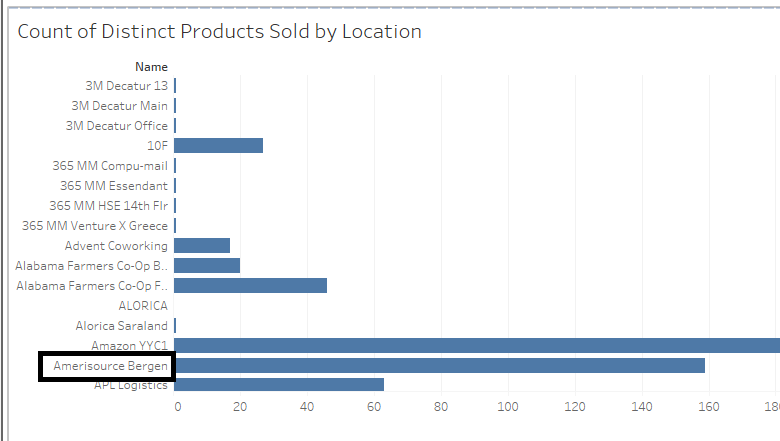
1. Tableau
   1. Snowflake Data Source (zoom for readability)



* 1. Rudimentary Dashboard



* 1. Validation of COUNT of DISTINCT(PRODUCT) for ‘AmerisouceBergen’
     1. Tableau Report



* + 1. Query from Snowflake Worksheet

