Democratize data access through a self-service Data Platform – Part 1

11 February 2025 - 8 min. read

Data Ingestion | Data Platform | Medallion architecture

In this 3-article series, we will describe how to properly create and structure a self-service Data Platform for data democratization analytics on AWS. We will start with data ingestion and storage and then move through processing tools to create valuable data for analytics, visualizations, and reporting. Moreover, we will focus on data governance, discoverability, and collaboration, with an eye on security and access control.

Follow this article to learn how to build your Data Platform and keep an eye on the website waiting for part 2!

TL;DR

The de facto standard for modern data platforms is the medallion architecture. Ingest all your data sources, streaming and batch, into the bronze layer. Data ingestion can be performed with tools such as DMS and Glue. Raw data needs to be cleaned, normalized, and structured to be stored in the silver area. Further transformations structure data for business use cases (AI, reporting, etc.…) in the golden layer. Golden data is ready to be used by business users.

Why a Data Platform?

In the modern era, there is this phrase that keeps echoing **“Data is the new gold”**, but is it true?

Spoiler alert: YES!

Every click, purchase, and interaction leave behind valuable information that, when properly analyzed, can transform the way businesses operate. By harnessing the power of data, businesses can better understand their market, identify customer patterns, and increase their operational efficiency. Integrating data-driven decisions into businesses' decision-making processes sharpens the strategies, making informed choices, based on real-world data, improving productivity and performance.

Having the data is definitely important, but if you cannot use it, it's barely pointless having it at all!

How to do so?

Here's another concept that is becoming increasingly popular nowadays: a **Data Platform**!

A Data Platform is a solution, composed of several different infrastructural pieces, that collects all organization data in its rawest form and, through multiple steps, refines it to its prime form. Like cutting and polishing for gems, the data platform extracts the highest value from the unstructured dirty data.

For businesses, a Data Platform is a strategic asset that enables data-driven decision-making, powers innovation and drives competitive advantage in today's digital economy.

Medallion architecture

The most common architecture for a Data Platform is the **medallion architecture**, which separates the data refinement process into 3 layers: bronze, silver, and gold.

The core idea here is that the **bronze layer** is the landing area for data in its crudest form. All data collection processes transport data from external sources, like databases, and APIs into the bronze layer, usually adding fields for data lineage and auditability. Common reading strategies involve CDC (Change Data Capture) read, batch-delta loads, and, in some cases, full loading of entire sources. Along with data collection processes, the bronze layer can also be the landing area for systems that directly push data, like CSV/Excel files, or manually provided data, although a “quarantine” area may be beneficial for data from “untrusted systems”.

Once we collect all data from all sources, it’s time to process it. Transformation processes that transport data from the bronze layer deal with data cleansing and validation to ensure some level of data quality. Here, we have processing like data deduplication strategies, null-value imputations, and error data handling. Moreover, data is structured with standard-defined schemas before getting into the next layer: the silver layer.

The **silver layer** of the medallion architecture holds cleaned and structured data. This layer contains a primordial form of data that is queryable and ready for consumption. Business users can explore this data and create new use cases. Properly structuring data inside the data platform makes it easily accessible, thus promoting data democratization and self-service analytics.

Before getting to the last layer, refined data from the silver layer is further transformed and aggregated to answer business needs from business user use cases. The key to these transformation processes is to structure data to optimize query performance to quickly answer business use cases.
The **gold layer** contains data refined at its highest value, its peak form. Data in the gold layer is optimized to serve as a source for business users like Business Analysts, Data Engineers, and Data Scientists for their business use cases, such as business intelligence, financial reports, AI, or machine learning.

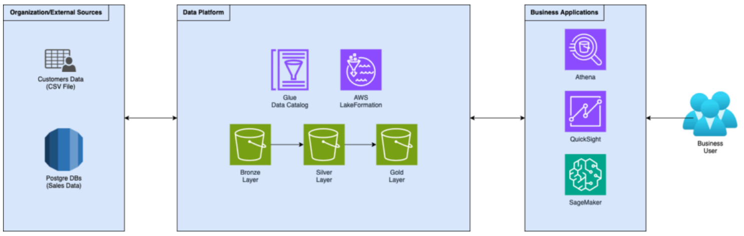
Data Acquisition: the setup

Let’s briefly introduce our sample use case for this article to give you a high-level overview of what you will see, picking up on the concepts discussed earlier, and better understand the solution.

Imagine that we are a fictional company that helps various companies and freelancers, in a variety of industries, increase their sales performance, and thus their revenue. The company wants to start using customers' data to extract valuable information to improve proposed strategies and better help them increase their revenue streams. In particular, the business thinks that gathering data and performing some Business Intelligence should do the job, as a first step!

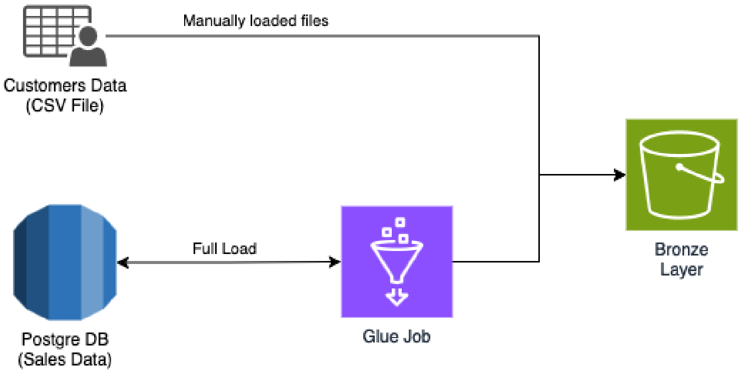
As part of the Data Team, you are tasked with creating the infrastructure and code logic that support these use cases.

To complete the setup, let’s say that your company has two customers, from two different industries, electronics and food, and you have their sales data along with their respective customers’ information from different sources: relational databases and CSV files. Here is a high-level diagram of the setup:



As you can see, and as we previously described, data flows into the data platform and is iteratively refined through the layers of the medallion architecture. Inside the gold layer data is designed for, and ready to be consumed by, business applications, like AI (Sagemaker), BI (QuickSight), or CI, pun intended here!

Since we are creating a simple data platform with sample data, we decided to go with the fastest method possible: customers' data files were manually loaded into the bronze layer bucket, while we created a Glue Job and a Glue Connection to load sales data.



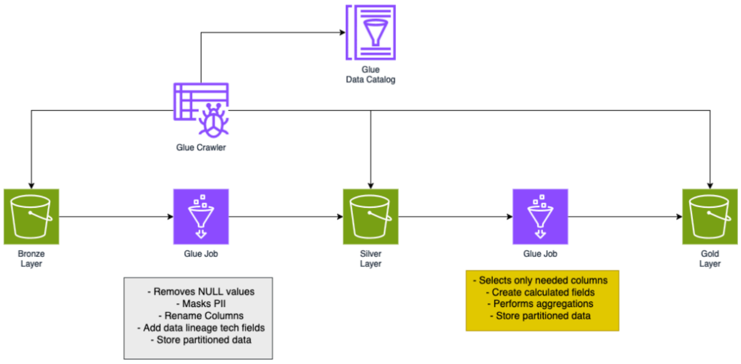
Obviously, while creating a real Data Platform, the context will be very different from this sample example and the choices are multiple. Here is some advice for different kinds of data:

Glue Jobs offer the Bookmark feature that gives you the ability to read a data source from where you ended last time. Still on databases, you can use AWS DMS to read databases with the CDC strategy. For APIs you can use both Lambda functions. For IoT data, AWS offers the IoT Core and IoT Analytics suites, an MQTT broker that gets your IoT data and stores it on S3, with the additional possibility of already transforming the data.

Data Pipeline: from data acquisition to golden data

Now that we have described some data acquisition strategies, let's dive into the realm of the data pipeline: from the bronze layer up to gold data, ready to be consumed.

Here is an image that shows you the structure of the data pipeline:



As you can see, we leverage the power of Glue Jobs, thus a Spark Cluster, to make data flow inside the pipeline. Moreover, we use the Glue Data Catalog, along with Glue Crawlers, to make our data queryable.

From Bronze to Silver

As we described earlier, transformations that take data into the silver layer mainly focus on data cleansing and standardization.

Data cleaning practices can vary widely. Common practices include handling missing values, data deduplication, and managing outliers. In our case, we just needed to handle missing values, and since there were very few rows, we decided to remove them directly. Moreover, we masked sensitive data before proceeding.

It is very common to find data quality checks paired with data cleaning. You may want to validate columns that must have a particular structure, like "email," or a specified range of values.

After we cleaned our data, it was time to standardize the schema. To do so, we renamed some columns and added some technical fields to help us track the data lineage, namely the job ID and the current time.

Finally, we were ready to store our data in the silver area. We stored data using the parquet format, a compressed and efficient format. For sales data, we partitioned it by store and product so that read queries can be optimized.

From Silver to Gold

Transformations from silver to gold layer are very specific to business needs.

In our case, the business just wanted some graphs of the total revenues over a given time period.

To do so, we started by selecting only the fields that we would use to calculate the total revenues, dropping everything that was not useful.

We calculated the revenues by multiplying the price and the quantity sold, thus creating what is called a “calculated field”.

Next, we performed an aggregation over the calculated to calculate the total revenues. We aggregated by date and store and performed a sum operation.

Lastly, data was ready to be stored in the gold area.

We stored data, using the parquet format, partitioned by store and date.

Now the business can understand better how to help their customers grow their sales.

Key Take-Aways and little spoilers

In this article, we described an example of a data platform.

We started with the WHY: why is a Data Platform useful? Then, we covered its most common architecture, the medallion architecture, and the various transformations performed on data from its rawest form up to the highest value.

Now that you have an understanding of a data platform, in the next chapter of this series of articles we will take this to the next level, explaining solutions to upgrade the data platform to a self-service platform, thus promoting data democratization. Moreover, we will use AWS LakeFormation to handle all the user and permissions on data.

Have you ever tried to create a data platform on your own?

Let us know your journey in the comments! See you in part 2!

About Proud2beCloud

Proud2beCloud is a blog by beSharp, an Italian APN Premier Consulting Partner expert in designing, implementing, and managing complex Cloud infrastructures and advanced services on AWS. Before being writers, we are Cloud Experts working daily with AWS services since 2007. We are hungry readers, innovative builders, and gem-seekers. On Proud2beCloud, we regularly share our best AWS pro tips, configuration insights, in-depth news, tips&tricks, how-tos, and many other resources. Take part in the discussion!

Matteo Goretti

DevOps Engineer @ beSharp. Passionate about Artificial Intelligence, in particular, Machine Learning and Deep Learning, and interested in Cloud Computing. I love trekking and nature in general. I relax with my guitar, play video games, and watch TV series.