DATA REPLY LEVERAGES AWS POWER TO OFFER CLIENTS A USER-FRIENDLY PLATFORM THROUGH A SERVERLESS END-TO-END WORKFLOW.

Data analytics and data science are constantly associated with a black box where only technical guys understand what is happening. AWS proposes a set of services that make it easier for any user to ‘get their hands dirty’.

AWS applies different kinds of serverless, user-friendly and no-code services to perform data cleaning, data analysis, SQL queries, and dashboarding.

Similarly, Data Reply empowers you and your organization with advanced analytics to achieve outstanding outcomes through the right use of data.

INTRODUCTION

Let's analyze a case study next.

Based on the data provided by NJ Weather we will proceed to identify the main cause or causes for train cancellations in New Jersey, back in March 2018.

Data is provided in the CSV format, separated by commas, with no structure. It describes the traffic of 28 lines for 2 companies (NJ Transit and Amtrak) - with the status of each train stop sequence (departed or cancelled). Figure 1 below displays the rail traffic of these two companies, with the main lines highlighted in different colours.

Figure 1 - Map of rail traffic - New Jersey

Usually, the first step would be to retrieve the appropriate schema, whether in Excel or even Python, in order to be able to analyse the information. Then we'd delve deep into the data to understand its meaning, and do some data exploration, data cleaning and analysis to get descriptive statistics. Let us see how the 5 mentioned AWS services can help us with this challenge. The architecture below describes the end-to-end workflow to get excellent value out of raw data.

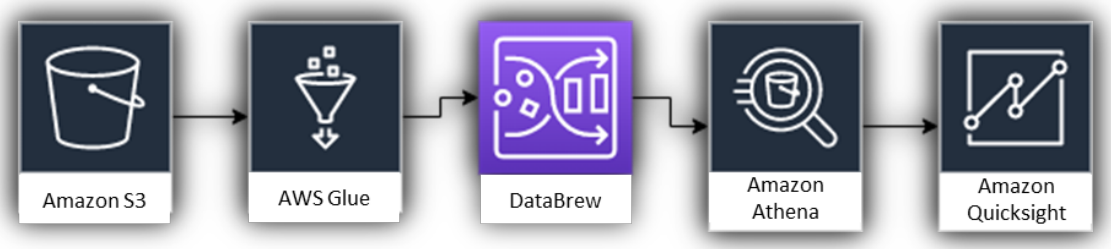

Figure 2 - AWS defined Architecture

1) The very first step is to upload the data in 'a simple storage service' (S3) - a bucket in common language. S3 enables cost reduction, data protection and management. To do so, we first need to create a new bucket to act as a data container: it requires a unique name and its security policies to be defined (such as read-only access, data encryption, etc.). S3 offers a large set of classes that can be chosen depending on the use case; from simple storage with quick access to data, to long-term storage with no need to instantly retrieve data.

   Once the bucket is set we can proceed to upload our data with the click of a button, inheriting the defined bucket policies (Figure 3).

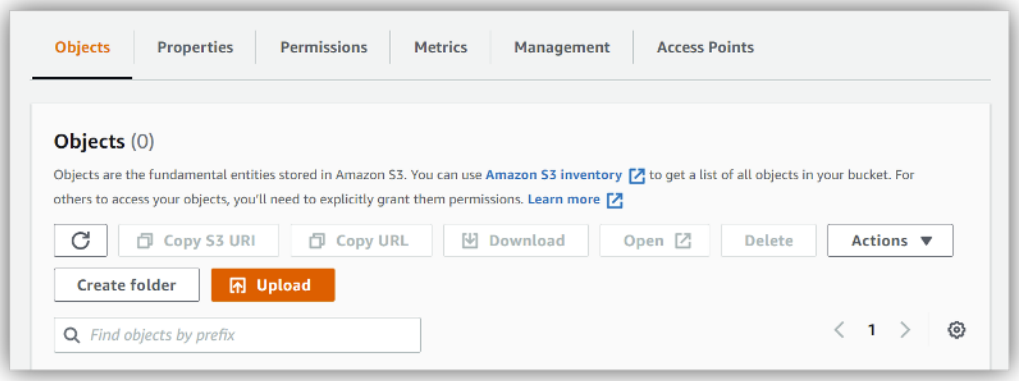

Figure 3 – How to upload Data into S3

2) One can then continue by using Glue, a serverless service that makes it easy to prepare data for analytics, machine learning or app development. Glue provides all the required capabilities to schedule ETL (Extract, Transform, Load) jobs, to transform and automatically retrieve our data schema.

   Prior to that, we need to define our ETL job through a crawler (Figure 4) that will run on our data.

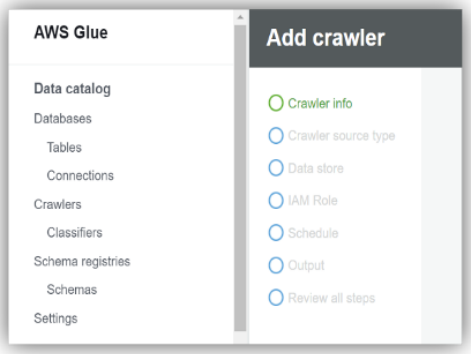

Figure 4 – How to define a glue crawler

Figure 2 - Automatically Retrieved schema

As we can see in Figure 5, the crawler automatically detected that our data is composed of 250,243 records and 13 columns, giving them a data type according to the initially detected format. We can see that the date column was categorized as a string due to its format - that is not a problem as AWS offers other services that enable easy data manipulation.

These tasks may be scheduled as jobs (a job is defined as a set of tasks) to be triggered whenever new data is uploaded to the S3 bucket.

3) At Data Reply, we understand the difficulty for any company to provide access to their data (mostly sensitive data). As such, we want to make sure that the end-user, the client, gets outstanding customer experience.

   How do we manage this? Users can use Glue DataBrew, a data preparation tool with a well-designed interface, that allows them to clean data from outver 250 pre-built transformations without having to write any new code. One of the most interesting fact, once again, is that those preparations are saved as recipes and can be scheduled and automated.

   As a Glue service, DataBrew automatically gets the data in the right schema with the glue crawler and offers:(Figure 6)

- Profiling to evaluate data quality, find patterns or detect anomalies.
- Cleaning and normalization (filtering, converting, correcting values, etc.).

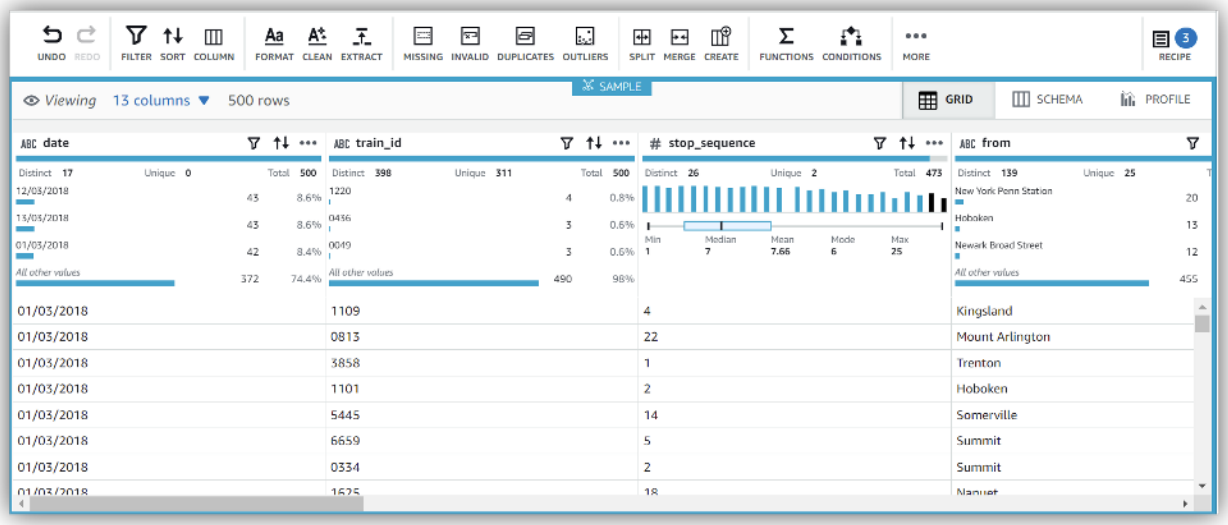

Figure 6 – Glue DataBrew user-friendly interface

As mentioned earlier, the date format was considered a string. Let's see how to transform it back to a date format in a user-friendly and flexible way. The user just has to click on the 'three dots' button over the date column and select the suitable option. (figure 7)

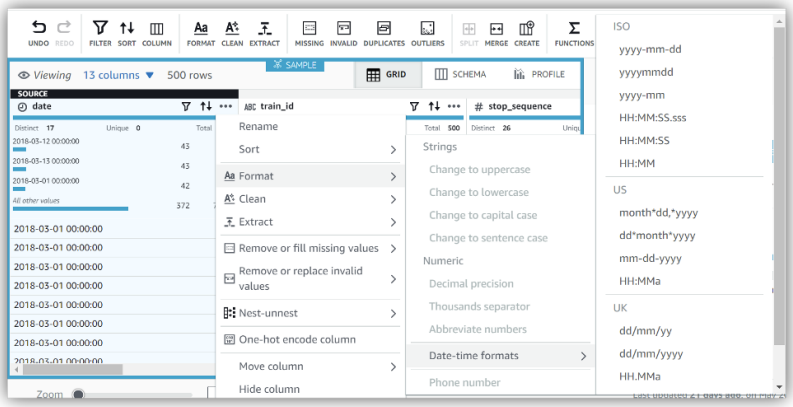

Figure 7 – How to easily apply data transformation

Assuming that we also want to fill the missing values with 0 from the 'delay_minutes' column and change the format to decimal precision, please refer to 'Figure 8' to see what the recipe will look like:

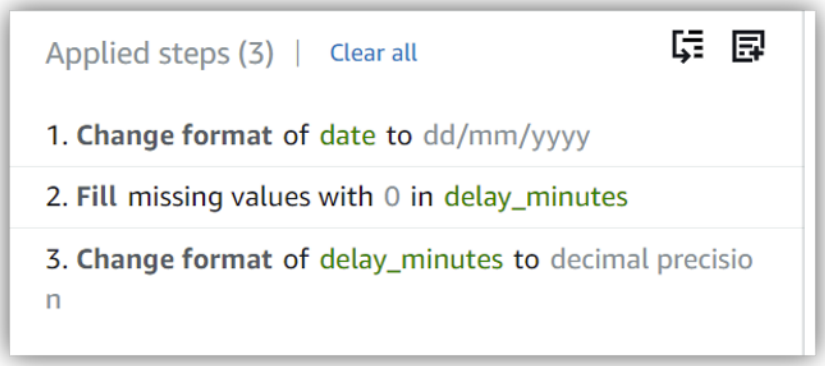

Figure 8 – Saved recipe steps

So DataBrew acts like a super Excel, allowing any user, with or without technical skills, to apply data transformation. It is user-friendly and it guides the user from the first to the last step.

Finally, we get the mapping data lineage to understand all the steps the data has gone through (Figure 9). The recipes can be versioned so that it is possible to get a previous version or build more steps or modify existing ones.

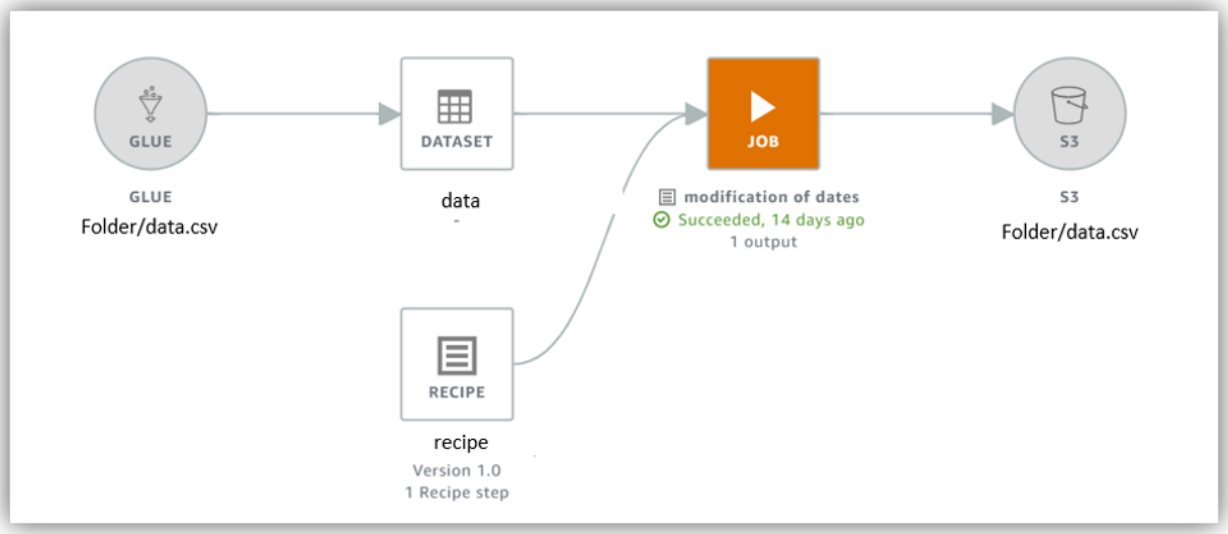

Figure 9 – Mapping Data Lineage

4) Once the data is prepared, we can use Athena, an interactive service using Presto that makes it easy to run simple and standard Structured Query Language (SQL) commands and export the required data. The resulting data becomes the input for Quicksight (Figure 10). Athena is an easy-to-use service, for anyone with SQL knowledge.

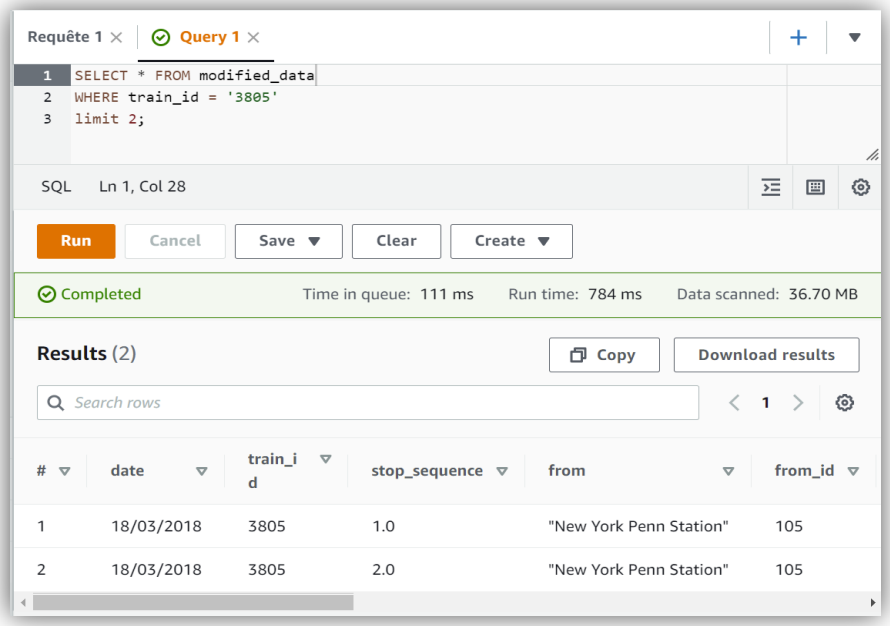

Figure 10 – Athena interface with a query example

5) How about some magic now? (if this isn't already the case!)

Quicksight can connect to all of your data in AWS, on-premises or on a third-party cloud. The data is sourced from Athena. This is a powerful tool that enables end-users to ask questions in natural language or business analysts to create serverless dashboards without any software.

All previous steps were aimed at preparing the data to become inputs for Quicksight in order to build interactive dashboards, automatically look for patterns, do some forecasting, etc. - so that every user can understand the data, get the insights and use these to make decisions.

We will not expand further on how to build a dashboard in this article, but instead, we will focus on how to get the best out of one that is already built. The idea is to highlight which days of March 2018 were the ones with most trains of NJ Transit and Amtrak cancelled and which lines were affected.

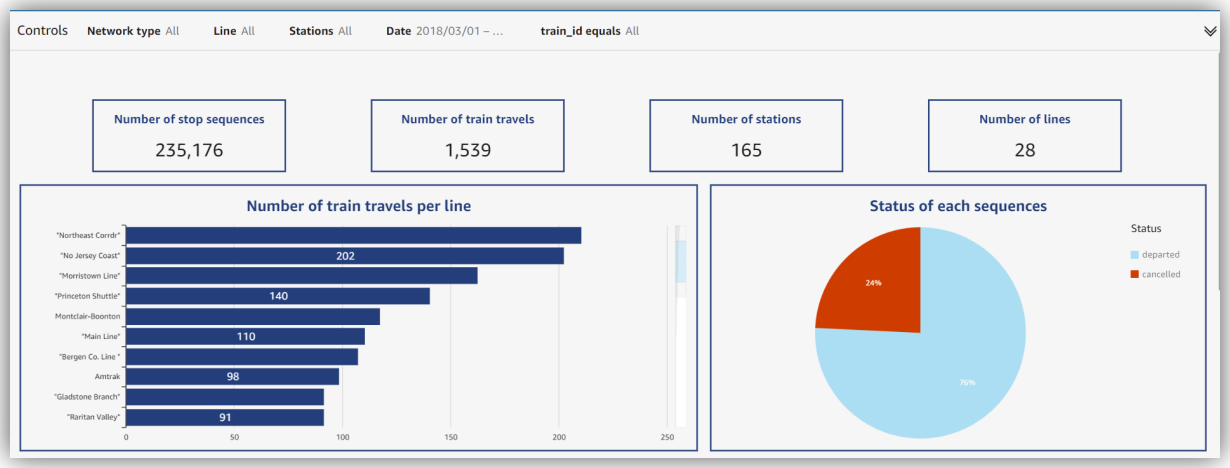

Figure 11 – Quicksight dashboard – a global overview of the data

Let's have an overview of our data; Figure 11 depicts 1,539 train travels (travel is defined as a train getting from 'station a' to 'station b' regardless of the number of stop sequences) for 28 lines in March 2018. Those travels stopped in 165 different stations a total of 235,176 times. Out of the 28 different lines, 'Northeast Corrdr' seems to be one of the most represented in the number of travels directly followed by 'No Jersey Cost'. The pie chart shows us that 76% of the sequences departed on time. Yet, the cancelled departures still represent an important 24% of the train journeys that happened on the day. The key here would be to understand the reason behind these cancellations in order to prevent these in the future.

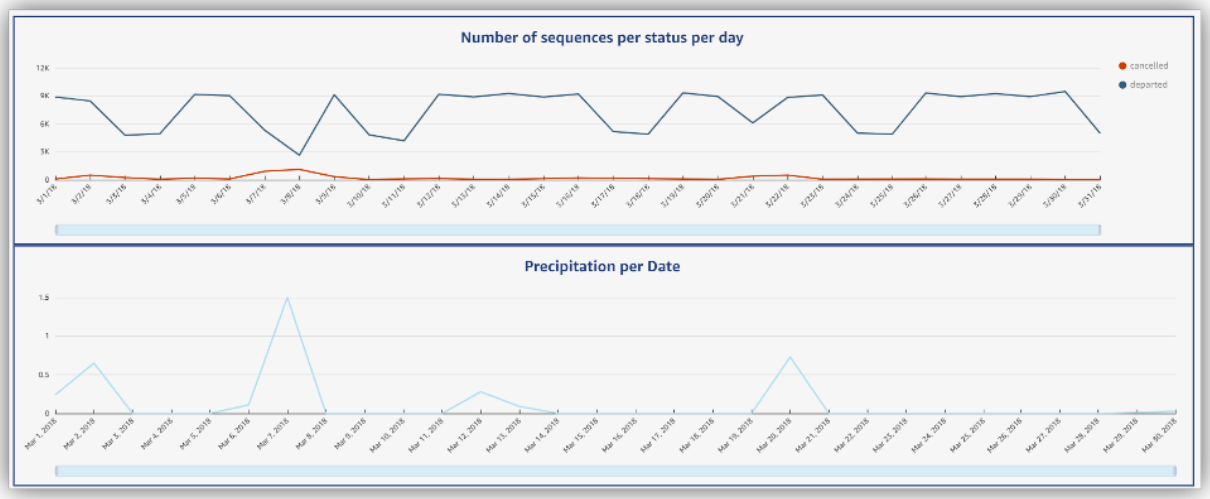

Figure 12 – Quicksight dashboard – Line charts

Now, the first line on the chart (Figure 12) shows the number of sequences departed on time or cancelled per day. First, we identify trends (a prevailing tendency that seems to repeat itself multiple times - 3/10/18 to 3/17/18, then 3/18/18 to 3/14/18 and finally 3/25/18 to 3/31/18). It does not come as a surprise as it represents the traffic from one week to another (Saturday to Saturday). On weekends, fewer departures are listed while on weekdays, traffic increases.

Two obvious anomalies are easily detected with this visualization of the data, from 6th to 8th and 21st of March 2018 - with unusual low departures and high cancellations. The combination of multiple source information which is conveniently displayed enables users to determine the root cause. As an additional information source, the second line chart displays the precipitation per date. We are able to identify that on March 7th and 20th, abnormal precipitations might have caused the anomalies from 6th to 8th and 21st March 2018.

CONCLUSION

In conclusion, AWS services assist with decision making converting raw data into final dashboards, showcasing relevant insights from our data and preventing the users to lose track of the processes applied to their data (black box phenomenon). Within the same platform, we were able to perform data ingestion, data cleaning and transformation, SQL queries, dashboarding and, last but not least, we were able to solve our initial question.

There are multiple benefits for the client:

- Step-by-step view of your data processing, where each step is detailed and versioned.
- Facilitation of the collaborative work. Reply's main objective is to provide its clients with the best customer experience. This platform allows for smooth communication and monitoring as notifications can be triggered whenever a new action is performed and whenever transformations are applied to data.
- Multiple data sourcing.

At Data Reply we put at your disposal our expertise in analytics & BI, machine learning and data engineering to help you add value to businesses and make decisions based on processed data and insights with the aid of AWS services.

For any further information, please feel free to contact us via email at info.data.fr@reply.com.

DATA REPLY

Data Reply is a subsidiary of Reply Group offering a wide range of advanced analytics and AI-powered data services. We operate across a range of industries and business functions, working directly with senior professionals and general managers to enable them to achieve meaningful results through the effective use of data.

We help companies design and implement human-centric data products to turn embryonic ideas into business solutions.

Our goal is to find compromises between business requirements and technical constraints, develop holistic Big Data architectures, and ensure cybersecurity and data protection