Creating a serverless Big Data Analyser with Amazon Athena

CASE STUDY

We used AWS services and the Cloud to visually interpret and analyse Big Data with a cost-effective, fast, and scalable solution.

BUSINESS NEED

Big Data is becoming increasingly relevant for today's businesses. **Companies have access to larger and larger amounts of information**, but often lack the means to effectively make use of this.

This is a problem relating to both hardware and software. To collect, store, and process large amounts of data typically requires large server space and efficient processing power. However, few companies have the resources to set up their own hardware, especially for something that might not be running all of the time; but will nonetheless add expenses in terms of ongoing maintenance and support. We had an idea that would solve these problems, so we set out to create a working prototype to prove it.

Any data solution needs to be able to accurately convey results to users as quickly as possible. That's why it was crucial to come up with a solution that would easily and effectively express results derived from Big Data. One of the best ways to do this is with graphs and other visualisations, as images that can readily convey important trends and results are vital for more accurate decision making.

Another important factor was ensuring a **cost-effective solution at scale**. As the volume of data increases, we need a solution that doesn't rapidly scale up in expenses. For traditional means, such as using our own servers to store and process all of this information, this can be nearly impossible, due to hardware, maintenance, and other related costs, (all of which still apply, even when the product isn't in use).

SOLUTION

Very early on, we concluded that **the Cloud was the best option** for containing both the data and all the processes required to analyse it. This was achieved using various AWS services.

Specifically, we utilised an **Amazon S3 bucket** to both store and retrieve data. We stored data in two folders - one with compressed data and one with uncompressed data - for the purposes of testing the most efficient solution. These files contained yearly data and were stored in a CSV format.

To extract this, we used **AWS Glue**, which is an ETL (extract, transform, and load) service that works exceptionally well with other AWS services, such as our bucket. This takes raw data and transforms it into various tables and schema to meet our needs. AWS Glue is smart enough to take a query and, if the CSV files were properly formatted, it can even provide the correct column headers.

The key benefit of AWS Glue, however, was the ability to create a scheduler. This enables regular crawling of the database, allowing for the ongoing, automated generation of results.

When it came to conducting a more detailed analysis, we used **Amazon Athena** for the purposes of answering our queries. Amazon Athena can be used to produce interactive queries and even analyse data stored directly in our Amazon S3 bucket; however, it is also entirely compatible with AWS Glue, which we are already using to better transform our data. This means that, when combined, the database and table is automatically configured to use the names and headers already provided by AWS Glue.

All of this gives us the data we need, but lacks a method of easy interpretation from human users. For this last, vital step, we used **Amazon QuickSight.** This business intelligence tool provides data visualisation, integrating directly with Amazon Athena to collect this information, or even the raw files from our Amazon S3 bucket.

By combining these services, we created an **entirely serverless process that provides us with various graphs and visuals** needed to quickly identify trends and changes as required. When testing Amazon Athena, we found that queries could be resolved between 1.75 and 35.5 seconds, depending on both the scale of the data set (we tested as much as approximately 30,000,000 records at once) and whether it was compressed or not.

BUSINESS BENEFITS

By using AWS services and the Cloud, we created a means to visually interpret and analyse Big Data that is **cost-effective**, **fast**, **and easy to scale**.

One of our biggest challenges was providing cost-effective means to store and process large amounts of data. Fortunately, the Cloud solves this. **By storing data on remote servers, we remove all of the hardware costs from our end**.

→ **Using an AWS server costs us just $1.40 per month, with storage costing just $0.023 per gigabyte (GB) per month.**

Similarly, the combination of Amazon S3 buckets, AWS Glue, Amazon Athena, and Amazon QuickSight allowed us to create an **entirely serverless solution for data manipulation** and analysis. Because it's serverless, we only pay for processes that are actually used, rather than the ongoing storage of such data tools - keeping costs at an absolute minimum. With Amazon Athena, costs were kept to just $5 per terabyte (TB) when scanning queries, giving us plenty of opportunities for Big Data without worrying about expenses.

All of this provided a **solution that's easy to expand** and **significantly cheaper** than on-premise solutions. It's also useful across a wide range of data sets - in fact, we've used variants for multiple clients to meet their unique data needs and challenges. We can even take this further, combining it with **Machine Learning** systems and **Artificial Intelligence** capabilities to offer even more advanced functions and relevant insights.

PROJECT DETAILS

Solutions — Amazon Athena, Data Lake, Serverless, Infrastructure as a Code, Big Data, ETL

Technologies —SQL, Python

Tools — Amazon S3 Bucket, AWS Glue, Amazon QuickSight, AWS Lambda, Terraform

Team — 1 Developer, 1 DevOps

ABOUT PGS SOFTWARE

PGS Software is one of the largest public listed custom software & services providers in Poland. As an AWS Advanced Consulting Partner, we specialise in Cloud projects – consulting, cloud-native development, application modernisation, & migration. Working according to agile methodologies (Scrum, DevOps, & Continuous Delivery), we create mobile & web applications as well as provide Business Analysis, Visual Design, UX, UI, & QA services to clients worldwide. We have development & business entities in Poland, UK, Germany, & Spain.

For more information about our services:
— please call us at: +44 (0) 770 353 6786
— visit our website www.pgs-soft.com

PGS SOFTWARE