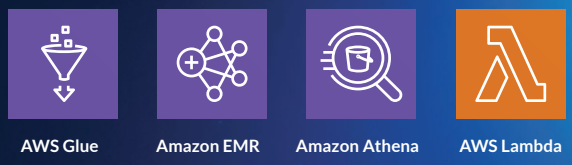
aws partner network

creativecapsule™

Case Study

AWS Glue drives faster data ingestion pipeline



In this case study, we tell the story of:

- An automotive startup whose infrastructure was struggling with large ETL workloads
- How it faced ever-increasing processing times and an escalating infrastructure cost
- Why Creative Capsule recommended a modern ETL architecture based on distributed AWS cloud computing
- How this architecture cut processing times by 70%, and reduced the DevOps time and OpEx by 50%

About the client

OneRequest is a marketplace startup built on the AWS cloud and designed to disrupt the traditional process of buying and selling cars. It aims to personalize the experience of buying, selling, and trading cars by eliminating the typical stress and frustration of the process while protecting the privacy and interests of the buyer. It enhances the user experience with a range of services, such as car-buying concierge, comparison tools, anonymous messaging with car dealerships, automotive reviews, and shopping advice.

A key differentiator of OneRequest is its massive, seller-agnostic car inventory, allowing the buyer to see all the cars on the market that meet their requirements. However, attempting to interact with this huge inventory resulted in issues with performance and efficiency. This case study delves into each of the challenges the startup faced and the measures taken to address them.

onerequest

Creative Capsule played a crucial role in our ETL growth management. As our data grew, it became increasingly challenging to swiftly ingest, normalize, and effectively use the data. Creative Capsule expertly identified our needs, presented viable options to fulfill them, and then successfully implemented the plan within the agreed budget and time, enabling us to achieve the necessary data throughput that is essential for our operation.

Chris Nickols
VP of Technology, OneRequest

Unraveling the problem

During OneRequest's initial years of operation, the startup's vehicle inventory was updated every day with over 2.5 million cars, which involved processing 30 to 150 GB of data from third-party vendors. This process was extremely time-consuming as the raw data from vendors was not optimally organized and needed several pre-processing steps. The entire process would take anywhere between 6 to 18 hours daily.
Moreover, the Amazon Aurora technology used for storage and data processing was not optimal for processing large volumes of data and led to latency issues.

Occasionally, the technical team had to manually synchronize the inventory data with the latest market data to resolve any inconsistent data sent by the third party. Besides this, the AWS infrastructure cost was growing at the rate of 15% per month for AWS RDS alone. Continuing with the existing process would have resulted in the monthly AWS cost increasing by a further 20%.

Another factor that limited scalability was the use of a centralized, monolithic relational database. With the client's growing data processing demands, the existing infrastructure — developed four years earlier during the company's MVP phase — was not easily scalable. A fresh approach using newly available and technically mature technology was called for.

To scale up the product, the client needed to

1. Reduce latency in ingesting the data from third-party vendors
2. Improve processing speeds
3. Optimize the AWS infrastructure cost
4. Rethink the database solution

New architecture for speed and savings

Creative Capsule's team of cloud developers, ETL experts, database architects, and quality assurance engineers undertook a major initiative to increase the speed, efficiency, and cost-effectiveness of the data engine.

By introducing a new distributed processing architecture, Creative Capsule aimed at processing 3 TB of data per day at 3x the speed of the existing data pipeline. This new setup resulted in a 50 percent reduction in costs.

At the core of the new architecture is AWS Glue, a fully managed ETL service. AWS Glue jobs (ETL jobs) are implemented using the Apache Spark framework, an open-source distributed computing system designed for big data processing and analytics.

The storage framework for the data is implemented using Delta Lake, an open-source storage layer that brings greater reliability to data lakes. It not only functions as a library on top of Spark — an open-source unified analytics engine for large-scale data processing — but provides a new output format that will read and write so-called Delta tables.

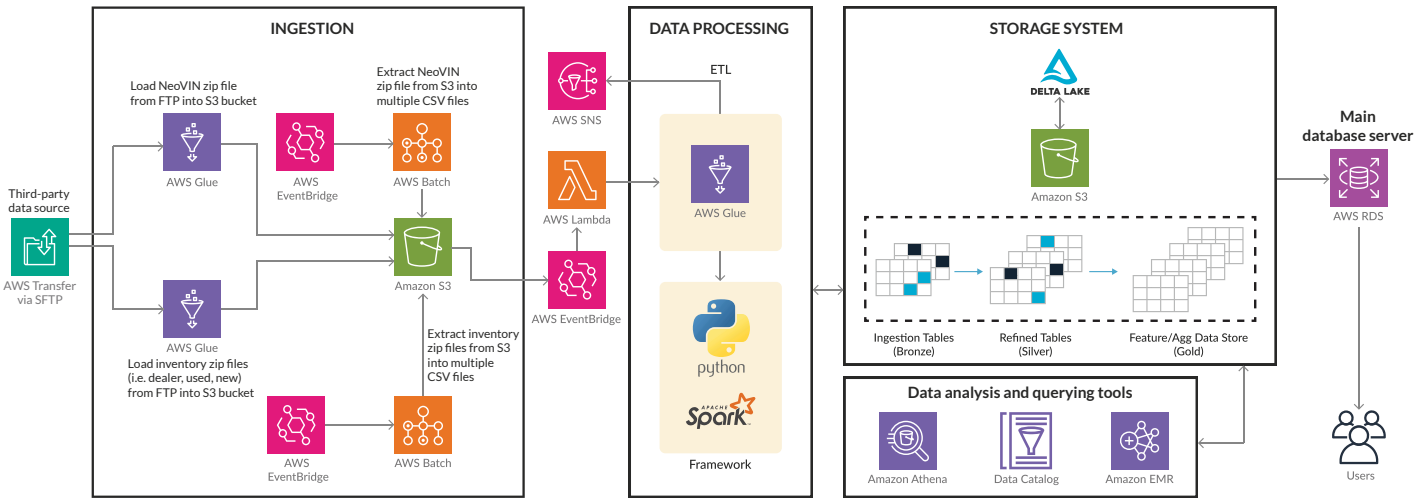
Delta tables store Parquet files, which contain the actual data, as well as metadata such as the schema and a transaction log. The transaction log enables Delta Lake to perform ACID transactions which help protect data validity despite errors, power failures, and other mishaps. The data used by Spark is stored on the Amazon S3 standard storage class.

The pipeline begins by loading the third-party data files via SFTP to Amazon S3. This process is facilitated by AWS Glue jobs, which are triggered as soon as the data files arrive at the SFTP server. The new vehicle data is received in two different files, NeoVIN, which has the pricing information, and Inventory Schema, which has the latest inventory in the market. AWS EventBridge listens for the event of loading data files into S3 and triggers Amazon EC2 Batch jobs. Instead of processing the massive daily load files at once, the files are now split into smaller chunks using the EC2 Batch jobs, enabling distributed processing.

Once the files have been split, an automated workflow step sends them to AWS Glue for processing. The processing is divided into multiple steps or stages, each handled by a specific Glue job. The workflow orchestrates the execution of these Glue jobs, ensuring that each job is executed in the correct order with the proper dependencies.

This precise, automated orchestration of steps ensures the smooth flow of data through the various stages of processing. Finally, to complete the workflow, some Glue jobs are dedicated to merging the processed data with the existing data in Delta Lake.

Fast and efficient data engine architecture



Going beyond: Streamlining the data engine

Once the new architecture was implemented, Creative Capsule's team focused their attention on other areas that were impeding performance and elevating the infrastructure costs:

1 **Optimizing the data storage policy:**
The team devised lifecycle management policies for Amazon S3, which were implemented to remove unwanted feed data older than six days. This reduced the storage space occupied by S3 and, consequently, reduced costs.

2 **Revamping jobs to reduce AWS Glue DPUs (Data Processing Units):**
Once the NeoVIN data was partitioned, the team reconfigured the AWS Glue jobs responsible for processing NeoVIN data to operate with a reduced number of DPUs (Data Processing Units). This optimization led to a remarkable 50% reduction in the cost of processing the NeoVIN data.

3 **Eliminating idle processing using AWS Glue triggers:**
The earlier system used a scheduled AWS Batch job which ran on a dedicated Amazon EC2 instance configured specifically for this task. The job would consistently monitor the FTP server at 30-minute intervals to check if files were available. This added significant costs and proved highly inefficient as the EC2 instance remained idle for 90% of the time. The Creative Capsule team replaced this batch job with an AWS Glue job, activated every 30 minutes via AWS Glue triggers, thus removing the need for an EC2 instance.

4 **Addressing NeoVIN data processing memory errors:**
As the data size increased over time, we encountered out-of-memory issues with Glue jobs that processed the NeoVIN data. To address these issues, the team implemented vertical scaling, which involves increasing the number of workers or changing the worker type from G.1X to G.2X. This approach helped resolve the issue in the short term; however, it led to an increase in the monthly AWS cost. So our team further optimized the approach by implementing the following steps:

- Archived inactive NeoVIN data, older than 6 months, to Amazon S3 to reduce data size and improve performance.
- Implemented Delta table partitioning, organizing the data into separate folders for each year.
- Implemented partition pruning logic that resulted in iterative processing of each year's data with reduced file scanning. This optimization significantly alleviated memory utilization during job execution.
- Adjusted the Spark configurations to decrease the size of Parquet files generated after processing.

5 **Selecting a storage solution with CRUD support:**
The development team extensively researched a storage solution that could efficiently support CRUD operations and be rapidly set up. After narrowing down the options, Delta Lake emerged as the ideal choice for this project due to its superior capabilities in handling CRUD operations.

6 **Cleaning up, transforming, and migrating data:**
The data exported from Amazon Aurora in CSV format had escape characters and multiline data. The NeoVIN files were large, with 30-40 columns, including JSON data. During the export, AWS Aurora split the files into smaller chunks, which led to missing headers in some files. This was resolved using Spark read options for reading CSV data. The team optimized the export process to be cost-effective while also making it more reliable than before.

The Result: A sustainable, cost-effective solution

The resulting architecture and solution led to a more dependable, sustainable, cost-effective, and optimized solution that would allow the business to process ever-increasing amounts of data and allow the business to scale without being bogged down by performance and data issues. The new architecture also saves significant DevOps time in provisioning and managing the complex infrastructure.

The team also ensured faster deployments by processing smaller chunks of data and having a multi-step process for deployments to optimize the usage of AWS Glue jobs.

This project took a little over 2.5 months to complete and the team ensured that costs were kept low throughout the development cycle with faster turnarounds.

The Creative Capsule team not only resolved the client's immediate performance and cost concerns but also delved deeper into understanding the pain points and effectively addressed them. As a result, they delivered a robust solution equipped with technology that aligned with the client's present and future requirements.

Benefits of the new data engine

- **50%** reduction in DevOps time and operational expenses
- Streamlined operations using AWS serverless technology
- **3x** improvement in efficiency and performance
- Enhanced data handling scalability from GB to TB
- **70%** reduction of data processing time and faster deployments using a multi-step process
- Automated infrastructure provisioning and management

creativecapsule™

Software Development Services
Startups • Enterprise SaaS • Pharma

About Us

Creative Capsule is a software outsourcing firm established in the USA in 2003 with subsidiaries in Switzerland and India. With our staff of over 300 full-time employees, we provide blended teams of local and offshore tech resources that offer software consulting and development, DevOps, and cyber security services. Our expertise spans diverse industries including tech startups, FSI, logistics, healthcare, FinTech, and smart-<homes/energy/cleaning>.

For companies looking to scale up and secure their software operations, we offer a range of advanced services such as AWS elastic Kubernetes clustering, Azure DevOps, cyber threat prevention, and custom scalability solutions.

In addition to being an AWS Consulting Partner, we are also certified on Azure and Google Cloud. Whether you are a SaaS startup born in the cloud or an SMB interested in adopting and leveraging cloud infrastructure, our approachable team of certified cloud experts are equipped to support you through every step of the cloud-computing lifecycle, from initial roadmap to implementation to ongoing support.

Expertise Areas

- Cloud Computing
- ETL Processing
- Lambda Serverless Computing
- CloudFront CDN
- AWS Batch Jobs
- EC2, S3, RDS
- DevOps
- Cloud Cost Optimization
- QA Automation Testing
- Mobile and Web apps

Our Partnerships

aws PARTNER Select Tier Services

Google Cloud

Veeva Content Partner

Microsoft Azure Certified

www.creativecapsule.com
info@creativecapsule.com

Follow us: