clearscale

aws partner network

Premier Consulting Partner

Decisiv Enhances PaaS Data Infrastructure with Machine Learning

Decisiv

Executive Summary

Virginia-based Decisiv is a leading provider of the largest asset service management ecosystem for the commercial vehicle industry. The Decisiv SRM platform is the foundation for more than 4,700 service locations across North America that manage more than 3.5 million service and repair events for commercial vehicles annually. Through Decisiv's Service Relationship Management (SRM) platform, dealers, service providers, manufacturers and fleet and asset managers can all connect and collaborate on every service event. The SRM solution streamlines the entire service process bringing all the necessary diagnostics, telematics and asset information together for all participants—at the point of service. This level of connectivity and collaboration drives an unrivalled level of service performance and asset optimization. Trucks get back on the road faster. Fleets see better revenues per asset and a lower TCO. Service providers establish the communication, better controls, and reliability in service operations that enables them to become trusted partners to their fleets.

Decisiv's Platform-as-a-Service (PaaS) model depends heavily on the organization's ability to collect, process, and manage data at scale. As Decisiv has grown, the velocity at which the company generates data has increased significantly, making it harder for in-house data engineers to manage critical workflows by hand. Consequently, Decisiv decided to bring in ClearScale, a cloud solutions provider with data management expertise, to help upgrade the company's IT infrastructure.

> "ClearScale demonstrated both its technical cloud expertise and creativity through our recent project. It was clear from the get-go that ClearScale had done this type of data management revamp time and time again. Their expertise in machine learning was a valuable complement to our own application experts. We're now well-positioned to create even greater value for everyone in the SRM Ecosystem from manufacturers, to our dealers, fleets and partners."
>
> **Satish Joshi,** Chief Technology Officer, Decisiv

The Challenge

Decisiv's SRM platform generates a tremendous amount of data at a high velocity. The company relied on manual workflows to manage all of this data. However, that approach was error-prone, inefficient, and unsuitable for running advanced analytics. Decisiv couldn't use its data to drive automated processes because there was no standardized or trustworthy single source of truth. Many data fields accept freeform input in order to accommodate a wide range of possible use cases.

This process left room for errors, redundancies, and inconsistencies. The company needed help figuring out how to overhaul its data infrastructure in a way that would make life easier for users and enhance Decisiv's ultimate value proposition of optimizing asset performance. Enter ClearScale.

As an AWS Premier Consulting Partner with ten AWS competencies, including Data & Analytics, ClearScale knew what Decisiv needed to revamp its data operations.

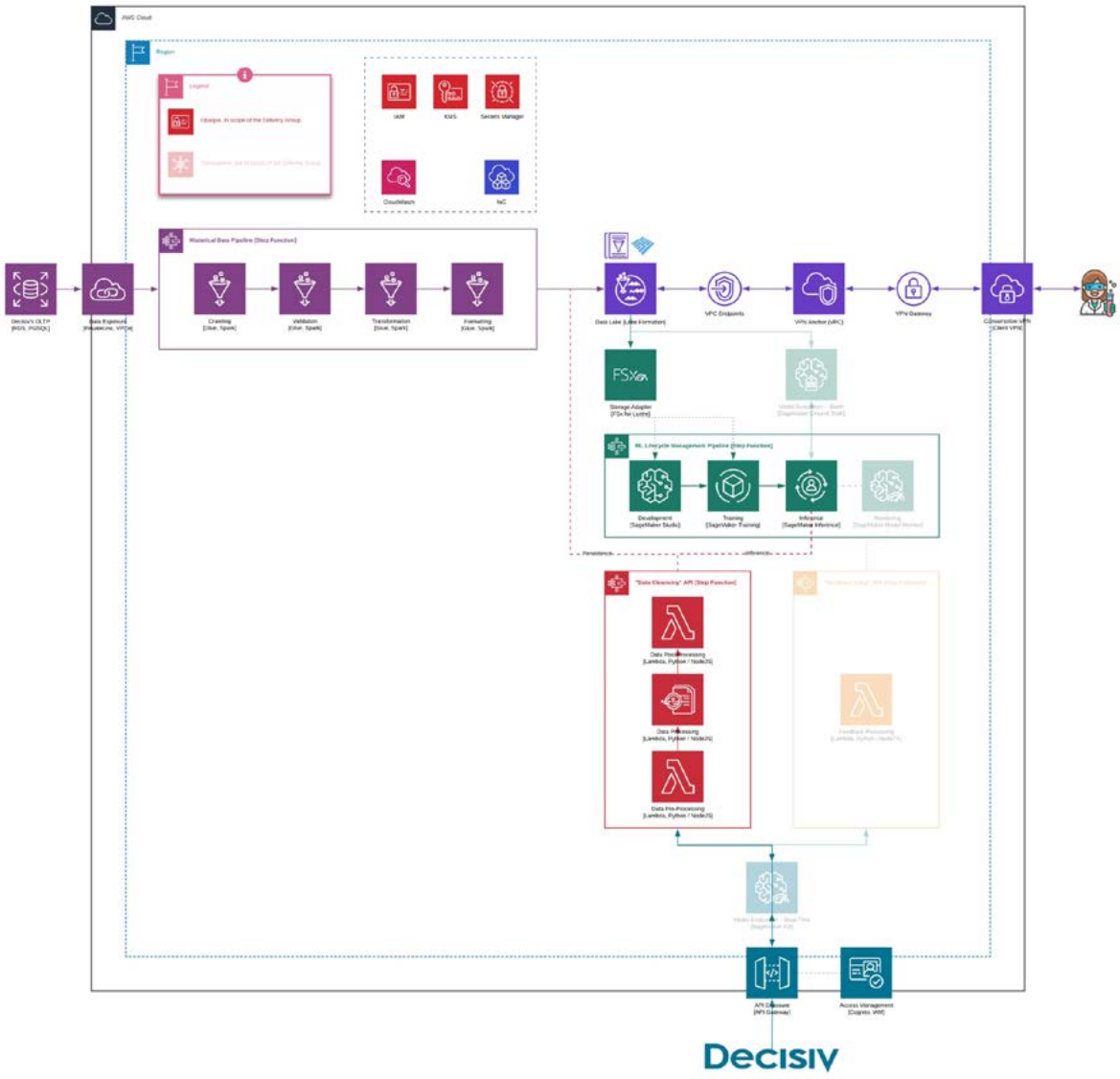
The ClearScale Solution

ClearScale proposed a three-step solution to meet Decisiv's needs:

- Unify and normalize Decisiv's historical data
- Correct and cleanse incoming data
- Deduplicate and reduce data

ClearScale aimed to implement as many automated workflows as possible. Doing so would require the use of advanced machine learning models and data management technologies. The overall architecture is shown in Figure 1. We'll discuss its components in depth.

Architecture Diagram



Data Unification and Normalization

Before creating any new machine learning models, ClearScale has to gather and standardize historical data in one place. Decisiv currently leverages Amazon Relational Database Service (RDS) for PostgreSQL & AWS Aurora Postgres to store operational data as well as AWS Redshift to conduct data analysis, which ClearScale made available to its engineers via AWS PrivateLink.

ClearScale uses AWS Glue, a serverless data integration service, to ingest, validate, and transform the data from RDS as it moves through the pipeline. To streamline this process, ClearScale implements AWS Step Functions, the only serverless stateflow orchestrator. After processing, Decisiv's data is stored in a data lake that utilizes Amazon S3, Glue Data Catalog, AWS Lake Formation, and AWS-managed OpenVPN tunnels.

To normalize ingested data, ClearScale converts existing data points in every column to the most efficient format available. For example, the team implements Google Geocoding and libphonenumber to format addresses and phone numbers in a uniform manner, respectively. To stay efficient under constantly changing subject field rules, the model needs to continuously receive all the new operational data and, hence, the process above is not performed once but automatically repeated in a near-real-time fashion.

Data Correction and Cleaning

The historical data ClearScale has is sufficient for training new machine learning models. However, preparing models to deal with all possible real-world scenarios requires extra effort.

First, SageMaker Notebook instances and Jupyter stack are used to supply data scientists with the modern cloud IDE. It enables them to create, update, and assess models via familiar tools while working from all around the globe. Amazon SageMaker is not a single tool but an umbrella for many features to build an end-to-end machine learning lifecycle management.

Second, ClearScale encapsulates all of the training logic needed in Docker containers. The team uses Amazon FSx for Lustre as a storage adapter in combination with SageMaker Training instances. It is a powerful approach for training machine learning models because it comes with numerous tools not available anywhere else. For example, developers can leverage SageMaker Automatic Model Tuning, SageMaker Experiments, and SageMaker Debugger to generate high-quality models with less development effort overall.

Finally, the model is deployed into the endpoint to be inferenced or, in other words, generate predictions. Again, SageMaker rules this domain with Inference instances.

Data Deduplication and Reduction

On the deduplication front, ClearScale uses AWS Glue FindMatches on top of the Fuzzy Matching algorithm to identify duplicates and inconsistencies between different data entries. This tool is efficient when it comes to comparing records based on subtle relationships between distinct fields. On the downside, it falls short when it comes to finding issues in a single column.

For that reason, in the parallel branch, ClearScale applies three custom machine learning models: FastTEXT, XGBoost, and DistilBERT (as noted in Figure 2 below). Each of these are designated for a specific combination of input parameters such as desired predictions accuracy, velocity, and throughput. Because the three models work differently depending on rules complexity, ClearScale ran them all in parallel. Together, these models are highly accurate at finding duplicate entries across Decisiv's data, ensuring the company has a single source of truth capable of supporting comprehensive analytics.

Custom Machine Learning Models

Machine Learning Models Comparison				
Criteria	**FindMatches**	**FastText**	**XGBoost**	**DistilBERT**
Dimensionality	1,000s	0	10,000s	100,000s
Performance	Best	Better	Good	Good
Coverage	Low	Low	Medium	High
Accuracy	Better	Good	Better	Best
Hardware	CPU	CPU	CPU	GPU
AWS-native	Yes	No	Yes	No

Lastly, end-users can access that pipeline through REST API exposed via Amazon API Gateway in both batch and real-time fashion with access management enforced via Amazon Cognito.

The Benefits

ClearScale's efforts benefitted Decisiv in many ways. The company now has a sophisticated machine learning-driven pipeline that can clean, deduplicate, and improve data quality in both real-time and in bulk. On top of that, Decisiv's engineers can now focus their energy on extending their new data platforms and operations. Collaboration with ClearScale is enabling Decisiv to proactively improve the quality of incoming customer and service data to ensure that it is trustworthy, consistent, and complete, enabling the company to serve its customers more effectively.

Going forward, Decisiv is well positioned continue to lead in providing unrivalled asset service management to the commercial vehicle industry, as it incorporates new and expanded analytics, modeling and data management technologies. Decisiv's engineers can tweak the organization's data infrastructure as needed and build new data and analytic services on their robust foundation, thanks to ClearScale's innovative design and approach. In an industry where Decisiv has revolutionized asset service management, ClearScale will be a part of their very bright future.