Digit88

Digit88 Technologies
www.digit88.com
info@digit88.com
+91 99011 59096

CASE STUDY 2025

Scalable Data Engineering in the Energy & Utilities Sector

Highlights

- Consolidated high-frequency, multi-format data into a **centralized data lake.**
- Implemented **hybrid processing** using Spark for batch and Akka for real-time workloads.
- Ensured scalability, low-latency, and resilience through optimized workflows and orchestration.
- Enabled downstream **BI and ML systems** with clean, high-quality data feeds.

Client

The client is a cloud-based platform provider offering enterprise rate solutions that modern, customer-centric utilities, fortune-500 energy suppliers, and technology providers rely on to usher in clean energy.

Challenge

In the energy and utilities sector, vast amounts of data are generated from smart meters, billing systems, and customer interactions. This includes detailed **consumption metrics, billing records, and rate plan attributes,** captured every 5 to 15 minutes across **millions of consumers.**

Key technical challenges in this space include:

- **Data Variety**: Information scattered across different formats & across **multiple databases and storage systems**.
- **Volume and Velocity**: High-frequency data streams require scalable infrastructure and optimized processing to avoid performance bottlenecks.
- **Data Consolidation**: Aggregating and normalizing data across sources to support downstream analytics and machine learning.
- **Latency and Access**: Need for both real-time and batch processing capabilities to support operational decision-making and predictive modeling.

Strategy & Solution

Data Ingestion & Consolidation

- Data enters the system via **SFTP** and is stored in **Amazon S3**, forming the foundation of a centralized **data lake**.
- Ingestion pipelines standardize diverse file formats and schemas, consolidating structured and semi-structured data from multiple databases.
- Normalized data is persisted into **Apache Cassandra** to support scalable storage of time-series data with high write throughput.

Workflow Automation with Apache Airflow

- **Apache Airflow** is used to manage complex ETL workflows through Directed Acyclic Graphs (DAGs), allowing fine-grained scheduling, monitoring, and dependency management.
- Supports modular pipeline design and provides resiliency through retry logic, alerting, and audit trails.

Scalable Processing: Batch and Real-Time

- **Apache Spark** is used for large-scale **batch processing**, enabling distributed data transformation and enrichment. Jobs are optimized across **Scala, Python, and Java** to ensure efficient resource usage and fast execution.
- **Akka** powers **real-time streaming**, enabling event-driven pipelines with low-latency message processing and backpressure handling — critical for handling high-throughput ingestion scenarios.

Data Delivery

- Processed data is made available through **streaming outputs, REST APIs**, and **query interfaces** to serve operational and analytics applications.
- Ensures secure and low-latency access for business-critical workloads.

Analytics & Machine Learning

- Structured data supports **BI platforms like Tableau** for dashboards and reporting.
- Data science teams use this foundation for **rate prediction models,** customer segmentation, and **personalized rate recommendation engines.**

Technologies Deployed

- **Big Data Engineering:** Scala, Spark, SparkSQL, Akka, Athena, Airflow, Azkaban, Postgres, Cassandra, AWS Keyspace, AWS DynamoDB, Apollo Configuration Service
- **Data Analysis:** Tableau, AWS Quicksight (moving to Tableau now)