**Project Title:** USA – UK Pharma Data Pipeline

**Domain:** Pharma

**Client :** Tibsovo ( services pharmaceutical )  
**Tech Stack:** PySpark, Python, SQL, Pandas, AWS (Lambda, S3, Glue, SNS/SQS, Step Functions, Redshift), ETL

**📄 Summary:**  
Developed a robust AWS-based data pipeline for processing pharmaceutical sales data from multiple external sources. Data ingestion was handled via S3 with event notifications triggering AWS Lambda for validation and error handling. AWS EMR executed PySpark ETL jobs that transformed and enriched data across the Bronze → Silver → Gold layers following Medallion Architecture principles. AWS Step Functions orchestrated the workflow, managing validation, enrichment, and data delivery stages. Curated datasets were stored in Amazon Redshift for analytics and reporting. This architecture improved data quality, reduced manual intervention, and enabled faster decision-making for business teams.

**🔍 Problem Statement:**  
The client’s pharmaceutical sales data came from multiple external sources (APIs, monitoring tools, broker-specific formats) with no centralized, scalable pipeline. The existing system lacked automation, validation, and real-time reporting. Manual processes caused performance bottlenecks and inconsistent data delivery.

**🎯 Goal:**  
Modernize the sales data processing system using cloud-native, big data technologies. Create a scalable and automated data lakehouse with Bronze, Silver, Gold layers for enriched and validated data delivery to business tools and CRM platforms.

**🏗️ Solution & Architecture:**  
**🟠 Bronze Layer – Raw Data Ingestion**

* External API → Lambda trigger → Raw data stored in S3
* S3 triggers → SNS/SQS → Lambda for validation; errors logged in S3 error bucket

**⚪ Silver Layer – Validation & Enrichment**

* AWS Glue Jobs: transformation, validation, error logging, schema mapping

**🟠 Gold Layer – Aggregation & Delivery**

* Partitioned and enriched data stored in S3, grouped by broker/region
* Delivered clean datasets to Amazon Redshift for analytics

**👨‍💻 Role & Responsibilities:**

* Designed and implemented ETL workflows using AWS Glue and PySpark
* Developed PySpark transformation logic and managed Delta tables in S3
* Set up external partitions, optimized queries using caching and partitioning
* Managed schema evolution, error handling, and dynamic data ingestion
* Followed best practices in Medallion Architecture for data layering

**📈 Business Impact:**

* 80% reduction in manual data handling
* Improved processing time by up to 40%
* Enabled real-time visibility into broker-wise sales data
* Handled 1 TB of data per month with scalable, reusable pipelines

**🧗‍♂️ Challenges Faced:**

* Schema drift across multiple broker data formats
* Handling large file sizes and optimizing performance
* Integrating legacy systems with modern AWS tools
* Maintaining cost-efficiency at scale

**✅ Outcome:**

* Built a scalable, cloud-native AWS pipeline
* Ensured data reliability, consistency, and availability for analytics
* Delivered clean, structured datasets to CRM and BI systems
* Aligned pipeline with modern Data Lakehouse architecture