

Ollie Lopes

SUMMARY

Senior Data Engineer with 10+ years of experience architecting and implementing enterprise-scale data solutions across federal government, healthcare fintech, and telecommunications sectors. Expert in AWS cloud services, Azure cloud platform (ADF, ADLS Gen2, Synapse Analytics), Databricks Delta Lake architecture, Snowflake data warehousing, and Apache Spark ecosystem. Proven track record of delivering large-scale data migrations with zero data loss, implementing medallion architecture patterns, and building real-time and batch analytics pipelines. Strong background in federal security compliance, CI/CD automation, and cross-functional collaboration.

TECHNICAL SKILLS

Cloud & Services AWS (Glue, Athena, Lambda, S3, EMR, EC2, Kinesis, Redshift, DynamoDB, SQS) | Azure (ADF, ADLS Gen2, Synapse Analytics, Azure Databricks, Event Hubs, Key Vault)

Big Data & ETL Apache Spark (PySpark, Spark SQL), Delta Lake, Kafka, Hadoop (HDFS, YARN), Hive

Data Platforms Databricks (Delta Lake, Unity Catalog, Workflows), Snowflake (SnowSQL, SnowPy), Airflow

Programming Python (Pandas, NumPy, Boto3), PySpark, SQL, Java Spark, Scala, Shell Scripting, REST APIs

DevOps & Tools Git, GitHub, Jenkins, Terraform, Docker, JIRA, MLflow, AWS Bedrock, SageMaker, Azure DevOps

Databases DB2, PostgreSQL, MySQL, Oracle, MongoDB, DynamoDB, Redshift, Azure SQL

PROFESSIONAL EXPERIENCE

Sr. Data Engineer | IRS - IBM

July 2024 – Present

- Led architecture design for enterprise Databricks Lakehouse migrating legacy DB2/mainframe data to AWS. Worked closely with data architects and legacy system experts to understand DB2 data structures, relationships, and business logic before designing the migration strategy.
- Implemented medallion architecture (bronze/silver/gold layers) with S3 as underlying storage. Bronze layer holds raw ingested data, silver contains cleaned and enriched data, gold provides curated datasets ready for consumption by analytics and reporting teams.
- Built incremental ingestion patterns to pull data from DB2 sources, creating reusable PySpark notebooks for complex transformations including cleaning, deduplication, joins, and enrichment as data moves through the medallion layers.
- Developed comprehensive data validation frameworks comparing source and target counts, checksums, and business rule checks to ensure data integrity during 1M+ record migrations with zero data loss.
- Leveraged Delta Lake features extensively: ACID transactions for concurrent write safety, time travel for querying historical data states and rollbacks, Z-ordering on frequently filtered columns for query optimization, and MERGE operations for efficient upserts.
- Utilized Delta Lake time travel to debug data corruption issues by querying previous table versions, identifying the exact batch where corruption occurred, and restoring silver tables to last consistent version—preventing downstream propagation and avoiding project timeline delays.
- Optimized Spark performance for large-scale processing: handled data skew using salting keys and broadcast joins, implemented strategic caching for small lookup tables reused across job stages, tuned partitions, and performed file compaction to maintain optimal file sizes (128MB–512MB).
- Optimized slow Delta table MERGE operations (5TB tables) by analyzing Spark UI for shuffle bottlenecks, verifying data skipping statistics, running OPTIMIZE with Z-order clustering on merge keys, and reviewing file compaction to reduce I/O overhead.
- Implemented end-to-end CI/CD workflows with Git and Jenkins, automating testing and validation frameworks. Created separate AWS accounts for development, test, and production environments with proper IAM isolation.
- Ensured federal security compliance implementing IAM roles, S3 encryption at rest and in transit, PII handling through data masking and tokenization, role-based access controls, and comprehensive audit logging aligned with IRS clearance protocols.
- Orchestrated workflows using Databricks jobs and Airflow, managing dependencies, scheduling, retries, and alerting. Exposed gold layer datasets through APIs and data warehouse tables for analytical teams, data science, and operational reporting.

Sr. Data Engineer | VisitIQ

April 2023 – June 2024

- Designed scalable AWS data lake architecture consolidating data from multiple sources: PostgreSQL transactional database, third-party APIs, and clickstream data from web applications into a centralized platform for analysts and data scientists.
- Leveraged Azure Data Factory (ADF) pipelines alongside AWS Glue to orchestrate hybrid data ingestion from on-premises SQL sources into ADLS Gen2, enabling unified reporting across multi-cloud environments.
- Implemented multi-source ingestion strategy: CDC using Debezium for PostgreSQL capturing real-time changes to Kafka topics, Python Lambda functions on schedule for API data pulls landing in S3, and Kinesis for clickstream event streaming.
- Built 30–40 AWS Glue PySpark jobs handling transformations across medallion layers—basic cleaning and deduplication in bronze-to-silver, complex joins and aggregations for silver-to-gold. Tuned Glue jobs by adjusting DPUs, optimizing partitions, and reducing shuffles.
- Resolved critical real-time streaming performance bottleneck where existing architecture struggled with high-throughput data streams causing performance degradation and data loss. Implemented Amazon Kinesis after conducting POC, validating feasibility through performance testing.
- Orchestrated complex ETL workflows using Apache Airflow with 50+ DAGs managing task dependencies, scheduling, and monitoring. Configured Slack notifications for failures with exponential backoff retries across data lakes and Snowflake warehouses.
- Optimized Snowflake query performance using materialized views, clustering keys, and data partitioning, reducing query execution time by up to 60%. Used Snowflake for SQL analytics while leveraging Databricks for complex transformations requiring more control.
- Reduced overall data latency from 24 hours to under 15 minutes, enabling analytics team to transition from ad-hoc queries on production replicas to self-service access to clean, governed data.

Sr. Data Engineer | Fiserv, Frisco TX

October 2021 – March 2023

- Architected enterprise data warehouse solutions using Snowflake with SnowSQL and SnowPy, implementing stored procedures, views, and complex data transformations integrated with Tableau for business intelligence reporting.
- Designed and deployed Azure Synapse Analytics pipelines for financial reporting workloads, integrating ADLS Gen2 as the data lake layer and Azure Key Vault for secret management, ensuring compliance with PCI-DSS standards.
- Orchestrated production ETL workflows with Apache Airflow managing 100+ daily tasks, integrating PySpark, Python, and SQL for automated transformations. Implemented data quality checks including validation for missing values, outliers, and inconsistencies.
- Collaborated with data scientists to operationalize fraud detection model by building scalable data pipelines that preprocess data in real-time and batch. Automated model retraining workflows using MLflow for experiment tracking, model versioning, and deployment.
- Built reusable Databricks notebooks for high-volume dataset processing, enabling data scientists to perform exploratory analysis on billions of financial transactions with pre-configured Spark settings and common transformation templates.
- Led successful cloud migration from legacy on-premises infrastructure to Databricks and AWS, improving system scalability by 10x and reducing infrastructure costs by 35% through strategic use of spot instances and resource optimization.
- Implemented data governance frameworks including data lineage tracking, metadata cataloging, data classification policies, and retention policies aligned with financial compliance regulations.

Big Data Engineer | Avaya Inc, Irving TX

February 2014 – September 2021

- Engineered distributed Spark ETL jobs on AWS EMR clusters utilizing Spark Core, Spark SQL, DataFrames, and Pair RDDs for processing terabytes of telecommunications data daily. Configured mixed on-demand and spot instances for cost optimization.
- Optimized cluster resource management using Spark YARN, tuning memory settings and executor counts to improve job execution performance by 40% and enable efficient handling of concurrent big data workloads.
- Managed wide transformations (joins, shuffles, aggregations) by monitoring shuffle read/write metrics in Spark UI, optimizing partitioning before transformations, and using broadcast joins to reduce shuffle overhead and bottlenecks.
- Implemented CI/CD pipelines using Git and Azure DevOps for automated code deployment, testing, and version control across development, staging, and production environments — complementing existing AWS CodePipeline workflows.
- Designed scalable Hive data models with optimized partitioning and bucketing strategies, processing multi-terabyte datasets with 50% improved storage efficiency and query performance on managed and external tables.

- Built end-to-end data pipelines integrating Hadoop ecosystem (HDFS, MapReduce, Hive) with AWS services for ETL, exporting transformed datasets to Amazon Redshift for business analytics and reporting.

EDUCATION

Bachelor of Science in Information Systems | University of Texas at Dallas

KEY ACHIEVEMENTS

- Successfully migrated 1M+ records with zero data loss during IRS mainframe modernization using Delta Lake ACID transactions, time travel for validation, and comprehensive data integrity frameworks.
- Reduced data latency from 24 hours to under 15 minutes by implementing real-time streaming architecture with Amazon Kinesis and Kafka CDC pipelines.
- Improved query performance by up to 60% through Snowflake optimization (materialized views, clustering) and Delta Lake Z-ordering techniques.
- Reduced infrastructure costs by 35% through strategic cloud migration, spot instance utilization, and resource optimization across AWS and Databricks environments.
- Delivered multi-cloud data solutions spanning AWS and Azure, integrating ADF, ADLS Gen2, Synapse Analytics, and Azure Databricks to support hybrid and cloud-native enterprise workloads.