

AKSHAY MASTER
Senior Data Engineer

Email: masterakshay04@gmail.com

Ph. No.: +1 (361)-945-1255

PROFESSIONAL SUMMARY

- Over **6+ years of professional experience** as a Data Engineer specializing in designing, developing, and optimizing scalable data pipelines across cloud and big data platforms.
- Strong expertise in **AWS cloud technologies**, including **S3, Glue, EMR, Athena, Redshift, RDS, DynamoDB, Lambda, EC2, IAM**, and building cloud-native data solutions.
- Hands-on experience in **Databricks**, working on large-scale distributed data processing, workflow orchestration, performance tuning, and migration of legacy data pipelines into modern cloud environments.
- Experienced in **migrating data pipelines across platforms** by refactoring legacy code, replacing platform-specific components, modernizing workflows, and improving scalability, reliability, and maintainability.
- Expertise in building end-to-end **ETL/ELT pipelines** using **PySpark, Spark SQL, Python, and AWS Glue**, handling large-scale structured and semi-structured data.
- Strong experience in designing and implementing **data lake architectures** using cloud storage solutions, Delta Lake, and partition-based data processing techniques.
- Proficient in optimizing **Spark jobs** by tuning memory allocation, partitions, execution plans, and parallelism, significantly improving performance and reducing processing time.
- Experienced in developing **configuration-driven data pipelines** using **YAML/JSON** to support seamless deployment across development, staging, and production environments.
- Skilled in workflow orchestration using **Apache Airflow and Databricks Workflows**, automating complex dependencies and scheduling enterprise-grade pipelines.
- Hands-on experience integrating **REST APIs, Kafka, Kinesis, and external systems** for batch and near real-time data ingestion.
- Expertise in **Snowflake** for data warehousing, data sharing, performance tuning, and large-scale analytics.
- Strong experience working with big data technologies including **Apache Spark, PySpark, Kafka, Flink, Hadoop, Hive, and EMR** for large-scale batch and streaming data processing.
- Proficient in **Python (Pandas, NumPy)** for data transformation, automation, API integrations, and reusable data engineering utilities.
- Extensive experience working with relational and NoSQL databases including **PostgreSQL, Oracle, SQL Server, MySQL, Redshift, DynamoDB, MongoDB, and Cassandra**.
- Strong expertise in implementing **data quality validation, reconciliation, and monitoring frameworks** to ensure data consistency and integrity.
- Hands-on experience in building **automated migration validation frameworks** for comparing source and target datasets to ensure data accuracy during migration.
- Developed reusable automation scripts for validating **JSON, Parquet, and Delta files** by performing schema comparisons, record count validations, partition checks, and data reconciliation.
- Experienced in implementing **end-to-end testing strategies for migrated data pipelines**, including historical backfill validation, incremental load testing, and performance benchmarking.
- Built automated testing solutions for identifying **schema drift, null inconsistencies, duplicate records, datatype mismatches, and data completeness issues** across large-scale datasets.
- Experience with secure credential management using **IAM, Secrets Management, and environment-driven configurations** for enterprise data pipelines.
- Skilled in **CI/CD pipelines, Git, Terraform, and CloudFormation** for automating deployments and infrastructure provisioning.
- Developed interactive dashboards and analytical reports using **Power BI and Tableau** to deliver business insights.
- Strong understanding of **Agile methodologies, sprint planning, JIRA, and cross-functional collaboration** to deliver scalable data solutions.
- Proven ability to design, develop, test, migrate, and optimize data engineering solutions with a focus on **performance, scalability, data integrity, and cloud modernization**.

PROFESSIONAL EXPERIENCE

Client: Zillow Group, Inc., Seattle, WA

Jan 2026 – Present

Role: Senior Data Engineer

Responsibilities:

- Led the migration of legacy **Apache Airflow DAGs to Databricks Workflows**, converting existing PySpark pipelines into scalable, multi-task workflows with automated scheduling, dependency management, and improved reliability.
- Modernized **AWS EMR-based PySpark processing tasks for Databricks**, migrating data access from Amazon S3 to Unity Catalog Volumes and optimizing ETL pipelines for performance, scalability, and secure data access. Migrated data storage from **Amazon S3 to Databricks Unity Catalog Volumes**, enabling governed data access, improved security, and simplified data management.
- Designed and implemented scalable **PySpark-based ETL pipelines** to process large-scale structured and semi-structured datasets for downstream analytics and reporting.
- Built and configured **Databricks Workflows** by converting legacy Airflow DAGs into multi-task job orchestration pipelines with dependency management and scheduling.
- Developed and maintained **configuration-driven data pipelines** using YAML/JSON to support seamless execution across development, staging, and production environments.
- Refactored platform-specific logic from legacy systems and replaced it with **Databricks-native implementations**, improving portability and operational efficiency.
- Implemented **partition-based data processing strategies** to support efficient incremental data loads and historical backfills.
- Built ingestion pipelines integrating with **external REST APIs** to fetch and process metadata and operational datasets into Delta/Parquet formats.
- Developed reusable **Python utility modules** for parameter management, config parsing, secret handling, and execution control.
- Optimized PySpark jobs by tuning **shuffle partitions, executor memory, caching strategies, and execution plans**, improving performance and reducing runtime.
- Implemented robust **data validation and reconciliation frameworks** to compare source and target datasets during migration.
- Developed automation scripts for comparing **JSON, Parquet, and Delta files**, performing schema validation, row count checks, and partition consistency verification.
- Built end-to-end **migration testing frameworks** to validate data integrity, transformation accuracy, and completeness across migrated pipelines.
- Performed **historical and incremental data validation testing**, ensuring consistency between legacy and migrated datasets.
- Implemented automated checks for **schema drift, duplicate records, null values, and data mismatches**, improving data quality and reliability.
- Created and maintained **Hive and Databricks SQL-based partition checks** to validate source availability before pipeline execution.
- Integrated **Databricks Secrets** for secure management of credentials, API keys, and environment-specific configurations.
- Developed logging and monitoring mechanisms to improve visibility into pipeline execution and simplify troubleshooting.
- Worked closely with cross-functional teams to analyze existing workflows, gather migration requirements, and deliver scalable data engineering solutions.
- Participated in **Agile ceremonies**, sprint planning, code reviews, and production support activities to ensure smooth delivery of migration and modernization initiatives.

Environments: Databricks, PySpark, Spark SQL, AWS EMR, AWS S3, AWS Glue, Amazon Athena, Delta Lake, Unity Catalog, Databricks Volumes, Python, Pandas, NumPy, Apache Airflow, Databricks Workflows, REST APIs, JSON, YAML, Hive, Hadoop, Kafka, PostgreSQL, Snowflake, Data Migration, Data Validation Frameworks, Parquet, Delta Files, Git, CI/CD, JIRA, Agile.

Client: Allstate, Northbrook, IL

Jul 2024 – Dec 2025

Role: Senior Data Engineer

Responsibilities:

- **Designed and built data pipelines using AWS Glue, Amazon Athena, and Spark SQL** to process large volumes of data and support business reports and analytics.
- **Built reliable data pipelines to load data into AWS services** like S3, RDS, Redshift, and DynamoDB, using AWS Glue and Amazon EMR to process and transform the data.

- Took ownership of ETL pipelines using **AWS Glue**, extracting data from diverse sources, transforming it into meaningful formats, and loading it into centralized data warehouses for enterprise use.
- Automated critical data integration processes through **AWS Glue** to reduce manual overhead and ensure consistent, timely delivery of clean datasets for reporting and analysis.
- Enhanced ETL standardization and reliability by developing reusable components in **Talend**, while also monitoring and maintaining Talend jobs in dynamic production environments.
- Handled the performance optimization of **Snowflake's virtual warehouses**, minimizing costs and maximizing query performance across business units.
- Used **Python (Pandas and NumPy)** extensively for complex data transformation, feature engineering, and module creation—ensuring scalable, reusable ETL logic.
- Developed **PySpark and Spark DSL scripts** for efficient big data aggregation and querying, writing the output back into relational databases using tools like **Sqoop**.
- Delivered near real-time analytics by processing **Spark Streaming** mini-batches within **Amazon EMR**, enabling faster decision-making pipelines.
- **Built real-time data ingestion pipelines using Kafka** and scheduled daily batch jobs using **Apache Oozie** to keep data updated and in sync across systems.
- **Built real-time streaming applications using PySpark, Apache Flink and Kafka**, enabling fast insights and processing of continuously changing data.
- Migrated and transformed large datasets into **PostgreSQL**, optimizing the performance for both OLTP and OLAP environments through tailored ETL pipelines.
- **Developed Python and Scala-based data preprocessing workflows** using **AWS Glue** and **Amazon EMR**, handling large-scale raw data transformation and enrichment tasks.
- **Designed and implemented data integration pipelines** to move data between **AWS services (S3, RDS, DynamoDB)** and external **RDBMS/NoSQL systems** using tools like **Apache Sqoop** and **custom Python scripts**.
- **Built NoSQL data processing solutions using AWS-hosted MongoDB and Cassandra**, enabling efficient handling of high-volume, unstructured data in real-time applications.
- Designed scalable data models for **MongoDB** and **Cassandra** to support real-time access and processing of large-scale transactional data.
- Developed and managed **PostgreSQL** environments supporting both real-time operations and analytical queries across business domains.
- Designed and maintained **Apache Airflow DAGs** to orchestrate, schedule, and monitor complex batch processing workflows on a recurring and scalable schedule.
- Delivered automated ETL pipelines in **Apache Airflow**, integrated with AWS components, to support diverse data processing use cases.
- Configured **Snowflake's data sharing and replication** features to facilitate secure, cross-team access to real-time datasets.
- Created and embedded **Power BI dashboards** into **SharePoint**, setting up granular access controls using role-based permissions.
- Developed dashboards and heat maps using **Power BI**, drawing data from sources such as **Teradata**, and delivering visual insights through charts and graphs.
- Used **DevOps practices** by integrating **Git** with **CI/CD pipelines**, enabling rapid and consistent deployment of data workflows across environments.
- Operated within **Agile frameworks**, using **JIRA** for tracking tasks, managing service requests, and contributing to sprint planning cycles for data projects.

Environments: AWS Glue, Amazon Athena, Amazon EMR, S3, RDS, Redshift, DynamoDB, AWS Glue, Talend, Snowflake, Python, Pandas, NumPy, Spark, Spark DSL, Sqoop, Streaming, Kafka, Oozie, PySpark, Apache Flink, Hive, Scala, Pig, Flume, NoSQL, MongoDB, Cassandra, PostgreSQL, Airflow, Snowflake, Power BI, Git, CI/CD, JIRA, Agile

Client: Max Healthcare, New Delhi, India

Jun 2018 – Jul 2022

Role: Data Engineer

Responsibilities:

- Designed and maintained **ETL pipelines** for healthcare datasets, integrating structured and semi-structured data from operational systems, APIs, relational databases, and distributed data platforms using **Python, PySpark, SQL, Kafka, and PostgreSQL**.
- Integrated both **framework** and **CloudFormation** to automate **Azure** environment creation along with ability to deploy on **Azure**, using build **scripts (Azure CLI)** and automated solutions using terraform.
- Developed an automated process in **Azure cloud**, which can ingest data daily from web **service** and load into **Azure SQL DB**.
- Managed **Informatica** metadata and ensured proper documentation for all **ETL** processes and transformations.
- Developed end-to-end **ETL** solutions using **Talend Data Integration** for various data transformation projects.
- Optimized **Python scripts** for performance, reducing execution times and improving resource efficiency.
- Designed and developed data pipelines using **Apache Spark** for **large-scale** distributed data processing-maintained **Python-**

based APIs to expose data services.

- Integrated **Kafka** with log aggregation systems to capture and analyze application logs for real-time monitoring and alerting.
- Created on reading and writing multiple data formats such as **JSON, Parquet**, and delta from various sources using **PySpark**.
- Installed/Configured/Maintained **Apache Hadoop clusters** for application development based on the requirements.
- Developed data pipelines that integrated **NoSQL** databases with distributed systems, enhancing the scalability and fault tolerance of the overall architecture.
- Worked with **JSONB** and other data types in **PostgreSQL** to handle semi-structured data within a relational environment, enhancing data flexibility.
- Developed custom **Airflow** operators to extend functionality and integrate with third-party tools such as **Spark, Hadoop**, and **Snowflake**.
- Developed interactive and visually compelling dashboards in **Tableau** to enable business users to explore and understand data insights effectively.
- Connected **Tableau** to diverse data sources, including **SQL** databases, cloud storage, and **APIs**, to create integrated, real-time reporting solutions

Environments: Azure, CloudFormation, CLI, Terraform, SQL DB, Informatica, ETL, Talend Data Integration, Python, Apache Spark, Kafka, PySpark, JSON, Parquet, Delta, Apache Hadoop, NoSQL, PostgreSQL, JSONB, Apache Airflow, Tableau, SQL, APIs.

EDUCATION

- **Master of Science in Computer Science**
Texas A&M University – Corpus Christi, TX
- **Bachelor of Technology in Electronics and Communication Engineering**
Jawaharlal Nehru Technology University, Hyderabad, Telangana, India