ABDUL BASEER SHAIK

AWS DATA ENGINEER

+1 240-245-0114 | abdulbaseershaiks@gmail.com | linkedin.com/in/abdulbaseershaik

PROFESSIONAL SUMMARY

- AWS Data Engineer with 12+ years of experience across software development, cloud platforms, and enterprise data systems.
- Background spans Amazon Web Services, Big Data technologies, and enterprise data warehouse implementations.
- Core AWS expertise: Glue, Amazon S3, EMR, Redshift, Lambda, Step Functions, Snowflake, and CI/CD services.
- Proven ability to build ETL data pipelines leveraging PySpark, Spark SQL, and Scala.
- Developed Spark pipelines with Scala and PySpark, including data loading and integration through AWS Glue.
- Hands-on with AWS Glue, Amazon EMR, AWS Step Functions, AWS Lambda, Snowflake, AWS CodePipeline, and CodeBuild.
- Experience moving data from Data Lake to Amazon S3 across the AWS stack.
- Implemented AWS Lambda, Amazon S3, SQS, and SNS for large enterprise-level ERP integration systems.
- Strong knowledge of in-house UNIX shell scripting for Hadoop Big Data development.
- Built and managed AWS CodePipeline and CodeBuild for CI/CD pipelines.
- Background in data pipeline development and data modeling.
- Developed ingestion workflows reading data from multiple sources into Avro, Parquet, Sequence, JSON, and ORC formats.
- Working knowledge of Hadoop, Java, HDFS, MapReduce, Hive, Tez, Python, and PySpark.
- Hands-on developing large-scale data pipelines using Spark and Hive.
- Used Apache Sqoop to import/export data between HDFS, Hive, and relational systems.
- Set up Apache Oozie workflows for scheduling and managing Hadoop jobs.
- Optimized Hive query performance using bucketing and partitioning techniques.
- Tuned ETL workflows for ingestion, modeling, encryption, and performance; extensive Spark job tuning.
- Developed Spark scripts using Scala shell commands per business requirements.
- Worked with Kafka and Spark Streaming to support real-time data pipelines.
- Experience with Apache Kafka and Amazon Kinesis for messaging/streaming applications using Scala.
- Well-versed in ETL methodology for corporate-wide solutions using Informatica.
- Worked with Parquet, ORC, AVRO, JSON, and CSV; tuned Spark configurations, partitioning, and memory allocation.
- Extensive experience developing EDW, Data Marts, ODS, and Data Warehouses with Star and Snowflake schema.
- Hands-on experience with GitHub for version control.
- Comprehensive knowledge of SDLC and Agile methodology.

CORE COMPETENCIES

Data Engineering • ETL/ELT Pipeline Development • Cloud Data Architecture • Big Data Processing • Data Warehousing • Data Modeling • Data Lake Implementation • Batch & Real-Time Streaming • Data Migration • Data Integration • Data Quality & Validation • Performance Tuning & Optimization • Workflow Orchestration • CI/CD Automation • Data Governance • Master Data Management • Dimensional Modeling • Source-to-Target Mapping • Production Support • Agile/Scrum Delivery

TECHNICAL SKILLS

Cloud Technologies	AWS Glue, Amazon S3 Data Lake, Amazon Redshift, Amazon RDS, Amazon Athena, Amazon DynamoDB, AWS Secrets Manager, Amazon EMR, AWS Lambda, AWS Step Functions, Amazon Kinesis, Amazon CloudWatch, AWS IAM, Databricks
Databases	Oracle, MySQL, SQL Server, MongoDB, DynamoDB, Cassandra, Snowflake, PostgreSQL
Programming Languages	Python, PySpark, Scala, Shell Script, Perl Script, SQL, Java, PL/SQL
Big Data Technologies	Cloudera, Hortonworks, Hadoop, HDFS, MapReduce, YARN
Big Data Tools	Spark, Spark SQL, Spark Streaming, Sqoop, Hive, Oozie, Zookeeper, Kafka, Flume, Druid, Impala, Storm, Solr, Pig
ETL & Data Warehousing	Informatica PowerCenter, SSIS, SSAS, SSRS, ETL/ELT Development, Data Modeling, Data Warehousing, Star Schema, Snowflake Schema, Slowly Changing Dimensions, OLTP, OLAP, Data Marts, ODS
CI/CD Tools	Terraform, Apache Airflow, Amazon MWAA, AWS CodePipeline, CodeBuild, Jenkins
Version Control	SVN, Git, GitHub
Methodologies	Agile, Scrum, SDLC, CI/CD, Data Pipeline Development, Data Integration, Data Governance, Data Quality Assurance
Reporting & BI	Amazon QuickSight, Tableau, Power BI
Operating Systems	Windows, UNIX, Linux, OS

PROFESSIONAL EXPERIENCE

High Radius Technology | Houston, TX **Nov 2024 – Present**

AWS Data Engineer

- Completed requirements analysis, application design, development, testing, and production operations for financial SaaS data solutions.
- Supported end-to-end AWS Glue pipelines for invoices, payments, remittances, deductions, collections, and customer-account data.
- Improved finance-data retrieval by tuning SQL queries, indexes, partitions, and access paths used by operational dashboards and reconciliation workflows.
- Developed DDL, DML, views, stored procedures, functions, triggers, and packages for financial staging, transformation, and reporting layers.
- Integrated ERP, CRM, bank, MySQL, Cassandra, API, and file data via AWS Glue and loaded curated datasets into Amazon Redshift and Snowflake.
- Designed custom AWS Lambda, Amazon EMR, and Databricks components for cleansing, standardization, validation, and exception processing.

- Orchestrated multi-stage finance pipelines with Amazon MWAA and Apache Airflow, including dependencies, retries, alerts, and recovery procedures.
- Created reusable Airflow operators and parameterized workflows for daily, intraday, month-end, and backfill processing.
- Automated data-quality checks for invoice totals, payment matching, customer balances, duplicate records, and source-to-target completeness.
- Managed Oozie workflows for legacy Hadoop jobs while modernizing workloads to cloud-native orchestration.
- Implemented Spark SQL and PySpark transformations in Amazon EMR and Databricks for receivables analytics, cash application, and financial reporting.
- Monitored scheduled workflows via AWS Step Functions and Amazon CloudWatch; investigated failures and restored processing within delivery SLAs.
- Delivered AWS Lambda code to ingest finance data from databases, APIs, secure files, and event-driven sources.
- Oversaw integration programs spanning Hadoop, relational databases, Snowflake, and cloud data platforms.
- Established CI/CD for data pipelines using AWS CodePipeline, CodeBuild, Jenkins, Git, automated tests, and controlled release workflows.
- Worked closely with DevOps and QA teams to promote repeatable builds, test-driven delivery, configuration management, and production validation.
- Developed PL/SQL utilities to automate reconciliation, cleansing, transformation, and validation of high-volume financial transactions.
- Applied Informatica Data Validation Option to compare source and target results and confirm financial-data accuracy across platforms.
- Designed Shell and UNIX scripts for secure file handling, job automation, archival, operational checks, and restart processing.
- Supported pipeline stability and data integrity via control totals, audit checks, exception handling, and reconciliation procedures.
- Handled payment, remittance, and transaction events with Amazon Kinesis, Kafka, and Spark Streaming for near-real-time use cases.
- Created Kafka, Spark, and Hive pipelines to ingest, transform, and analyze high-volume financial and customer-account events.
- Leveraged Jira to manage epics, stories, defects, development tasks, QA validation, and partner acceptance activities.
- Took part in Agile ceremonies, release planning, daily stand-ups, retrospectives, and internationally coordinated PI planning.

Environment: *Amazon EMR, AWS Glue, AWS Step Functions, AWS Lambda, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Git, JIRA, Jenkins, Kafka, AWS Glue Pipeline, Amazon QuickSight.*

State of Illinois | Springfield, IL **Jan 2023 – Oct 2024**

AWS Data Engineer

- Implemented public-sector data solutions using Hadoop, Cloudera, Spark, and AWS cloud services for agency operational/reporting workloads.
- Ingested structured/semi-structured agency data into Amazon S3, relational stores, Amazon Redshift, and Snowflake for analytics and public-service reporting.
- Delivered ingestion and orchestration workflows with AWS Glue, Amazon EMR, Databricks, and AWS Step Functions for scheduled agency-data processing.
- Migrated on-premises Oracle ETL workloads to Amazon EMR, Databricks, Amazon Redshift, and Snowflake while preserving source-to-target controls.
- Modernized SQL databases and file-based feeds into Amazon S3, cloud relational services, and analytical warehouse layers.
- Established access controls, service identities, and secrets management with AWS IAM and Secrets Manager for protected government datasets.
- Worked with bulk-load and warehouse-native transfer patterns to move high-volume data efficiently between Amazon S3, Redshift, and Snowflake.

- Automated build and deployment workflows via AWS CodePipeline, CodeBuild, and Jenkins for Python, SQL, Spark, and data-pipeline components.
- Developed batch and streaming solutions with Spark Streaming, Amazon Kinesis, Kafka, and schema-aware processing frameworks.
- Processed records in multiple data formats using Scala, PySpark, Spark SQL, and reusable transformation libraries.
- Designed Hive partitions and buckets to improve joins and analysis across high-volume agency datasets.
- Created Hive UDFs to apply policy-driven business rules consistently across programs and reporting periods.
- Implemented SQL and shell utilities to extract, standardize, and load data from legacy agency databases and secure file exchanges.
- Applied Sqoop to migrate recurring MySQL and relational extracts into HDFS for legacy downstream processing.
- Supported hybrid data-lake and big-data environments using Hadoop, Spark, Hortonworks, Cloudera, and cloud services.
- Loaded and transformed large, varied data formats for internal analysis and external reporting.
- Developed Hive queries to reconcile agency records, calculate reporting measures, and satisfy documented business requirements.
- Delivered Hive tables and HiveQL transformations that reproduced required MapReduce logic for repeatable analytical processing.
- Leveraged PySpark RDDs, DataFrames, and Spark SQL for data profiling, cleansing, standardization, and aggregation.
- Established Spark Streaming micro-batches for time-sensitive feeds and downstream batch processing.
- Developed CI/CD pipelines for Hadoop, Spark, SQL, and cloud data components with controlled promotion across environments.
- Designed PowerShell scripts for file movement, scheduling, validation, system administration, and operational support.
- Worked with Jira to manage requirements, defects, release tasks, testing evidence, and project workflow.
- Optimized PySpark and Spark SQL jobs via partitioning, caching, join strategies, and configuration tuning.
- Managed Git repositories, pull-request workflows, branching standards, and versioned deployment artifacts.

Environment: *Amazon EMR, AWS Glue, AWS Step Functions, AWS Lambda, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Git, JIRA, Jenkins, Kafka, AWS Glue Pipeline, Amazon QuickSight.*

InnovaCare Health | Athens, GA **Oct 2020 – Dec 2022**

Data Engineer

- Applied AWS Glue to ingest member, provider, eligibility, claims, encounter, and clinical-quality data from disparate healthcare systems.
- Orchestrated upstream-to-downstream healthcare workflows with AWS Glue, dependency controls, retries, and operational alerts.
- Automated scheduled, event-driven, and recurring jobs for daily claims, membership, provider, and quality-measure processing.
- Leveraged Amazon DynamoDB for reference, catalog, and event data for member and provider processing workflows.
- Designed source-to-target healthcare data flows and documented the corresponding AWS cloud architecture.
- Created high-level and application design documents covering ingestion, transformation, security, lineage, recovery, and downstream interfaces.
- Prepared mapping specifications and data-flow rules for claims, eligibility, provider, member, and clinical-quality datasets.
- Implemented build and release definitions with AWS CodePipeline, CodeBuild, and Jenkins for CI, testing, and controlled deployment.
- Documented secure interfaces for exchanging healthcare files and curated datasets via governed cloud-sharing services.
- Delivered pipelines and complex PySpark transformations with AWS Glue, Amazon EMR, and Databricks for healthcare analytics.

- Handled claims and event data in micro-batches with Spark Streaming and applied RDD/DataFrame transformations for timely analytics.
- Provisioned Amazon EMR and Databricks clusters for batch and continuous processing and applied AWS IAM and Secrets Manager controls to protected healthcare data.
- Established reusable AWS Glue pipelines for copy, filter, conditional, transformation, and Spark-based healthcare processing.
- Developed PySpark jobs in Amazon EMR and Databricks for member, provider, claims, encounter, and quality-measure table operations.
- Worked with SQL import/export utilities for controlled movement and reconciliation of healthcare data across database environments.
- Designed complex SQL, stored procedures, triggers, and packages for healthcare staging, audit, transformation, and reporting databases.

Environment: *Hadoop, Hive, Oozie, Spark, Spark SQL, Python, PySpark, AWS Glue, Amazon RDS, Amazon EMR, Amazon Redshift, Amazon S3, Scala, AWS, Linux, Maven, Oracle 11g/10g, Zookeeper, MySQL, Snowflake.*

Giant Food | McLean, VA **Apr 2018 – Sep 2020**

Big Data Developer

- Created scalable Hadoop data solutions for retail point-of-sale, product, pricing, promotion, inventory, loyalty, pharmacy, and store-operations data.
- Installed and configured Hive, Pig, Sqoop, Flume, and Oozie components for retail batch-processing workloads.
- Configured and benchmarked Hadoop and HBase clusters for high-volume transaction and product data.
- Optimized MapReduce jobs and compression strategies to improve processing of daily store and distribution-center feeds.
- Ingested POS, inventory, product, supplier, and promotion data; transformed it with Hive and MapReduce and stored curated results in HDFS.
- Developed UDFs to apply retail business rules for product hierarchy, store, promotion, and transaction processing.
- Applied Impala to query Hadoop and HBase datasets for faster operational and merchandising analysis.
- Implemented Spark applications for faster processing of retail transactions, inventory movements, and customer activity.
- Delivered Java and Pig UDFs for product, pricing, promotion, and store-level transformation logic.
- Monitored Hadoop clusters with Cloudera Manager and resolved capacity, performance, and job-execution issues.
- Coordinated Hadoop operating-system updates, patches, and version upgrades with infrastructure and application teams.
- Established Oozie workflows to schedule dependent Hive and Pig jobs for daily/weekly retail processing.
- Leveraged Storm for real-time processing of selected transaction and operational data feeds.
- Worked with Solr to search and navigate product, store, and customer-related datasets stored in HDFS.
- Loaded application and transaction logs into HDFS via Flume for operational analytics.
- Exported curated retail datasets to relational databases for BI dashboards, merchandising analysis, and operational reporting.
- Developed Java MapReduce programs to extract, transform, and aggregate XML, JSON, CSV, and compressed retail files.

Environment: *Hadoop, MapReduce, HDFS, Hive, Spark, Pig, Java, SQL, Cloudera Manager, Sqoop, Storm, Solr, Mahout, Flume, Oozie, Eclipse.*

Morningstar, Inc. | Chicago, IL **Jun 2016 – Mar 2018**

ETL Developer

- Analyzed investment-data requirements from data architects and translated them into source-to-target ETL specifications.
- Coordinated delivery estimates and SLAs for market, security, fund, portfolio, pricing, and reference-data requirements.
- Prepared technical design documents describing sources, business rules, transformations, controls, and target datasets.
- Designed Informatica mappings, sessions, workflows, UNIX scripts, and parameter files to process investment and market data.
- Developed test cases and reconciled source data against processed security, fund, pricing, and portfolio datasets.

- Deployed ETL code to integration environments via release and configuration-management procedures.
- Supported UAT and resolved data-quality, transformation, reconciliation, and performance findings.
- Prepared production move checklists covering dependencies, schedules, parameters, rollback, and validation steps.
- Provided support-team walkthroughs for new investment-data feeds, transformations, workflows, and production controls.
- Applied Informatica pushdown optimization to improve performance of high-volume investment-data processing.
- Applied parameter files and parallel workflow execution to meet time-sensitive market-data delivery windows.
- Created Slowly Changing Dimensions Types 1 and 2 and star schemas for investment research and analytical reporting.
- Provided 24x7 on-call support for production ETL workflows and critical market-data deliveries.

Environment: *Informatica, Oracle, ETL, Manual Testing, Windows.*

Transamerica | Baltimore, MD **Jan 2014 – May 2016**

ETL Developer

- Implemented enterprise data-warehouse solutions for insurance, retirement, policy, premium, claims, customer, and financial reporting data.
- Profiled source data and metadata for mapping, validated business assumptions, and identified insurance-data quality issues.
- Delivered logical and physical models and ER diagrams for relational and dimensional insurance databases using Erwin.
- Extracted policy, customer, premium, claims, and financial data from Oracle databases and secure flat files.
- Established Informatica PowerCenter mappings and Mapplets to load cleansed data into the Operational Data Store.
- Developed automated testing utilities for SOA services and data-driven validation of insurance integration workflows.
- Designed complex mappings with joiners, lookups, expressions, filters, routers, aggregators, sorters, stored procedures, and update strategies.
- Reconciled sampled policy and financial records across source, test, and target database environments.
- Designed extraction processes that cleansed, transformed, integrated, and delivered insurance data via controlled flat-file interfaces.
- Created SSIS packages to load SQL Server and SSAS structures for insurance and retirement reporting.
- Implemented and deployed SQL Server, Microsoft Office, VBA, and macro-based utilities for business reporting.
- Worked with architecture and modeling teams on SOA services and enterprise insurance-data integration.
- Completed data alignment, cleansing, mapping debugging, defect correction, and source-to-target validation.
- Delivered sequential and concurrent Informatica sessions for reliable execution of dependent mappings.
- Supported SSRS and SSIS reporting projects requiring policy, customer, financial, and operational data.
- Took part in System Integration Testing and User Acceptance Testing for insurance data releases.

Environment: *Informatica 8.6.1, SQL Server 2005, RDBMS, FTP, SFTP.*

EDUCATION

Lovely Professional University **Aug 2008 – Apr 2012**

Bachelor of Computer Science and Technology

Catholic University of America **Aug 2012 – Dec 2013**

Master's in Information Technology