**Shalem Raju**

**Senior Data Engineer**

**Email:** shalemraju0707@gmail.com | **Ph:** 913-283-4989 **| LinkedIn:**  https://www.linkedin.com/in/shalemraju-/



**PROFESSIONAL SUMMARY**

* Over 10+ Years of strong experience as a Senior Data Engineer in Big Data, Data Warehousing and Business Intelligence in Financial, Retail and Telecom domains including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Expertise in designing and developing scalable Big Data solutions, and data warehouse models on large-scale distributed data, performing a wide range of analytics to measure service performance.
* Strong experience in using major components of Hadoop ecosystem components like HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Hue, Spark Streaming and Oozie, with hands-on expertise in Cloudera Hadoop distribution.
* Experience in working with high-performance NoSQL databases like Aerospike, implementing low-latency data solutions for real-time analytics and transactional processing.
* Experience in Microsoft Azure/Cloud Services like SQL Data Warehouse, Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory and worked on Azure development and Azure Databricks, Power BI.
* Strong Experience with Amazon Web Services (AWS) Cloud Platform which includes services like EC2, S3, VPC, ELB, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Elastic Beanstalk (EBS), Auto Scaling, Security Groups.
* Extensive experience in ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
* Experience in Data collection, Data Extraction, Data Cleaning, Data Aggregation, Data Mining, Data validation, Data analysis, Reporting, and data warehousing environments.
* Expertise in troubleshooting, debugging, performance tuning, and optimization of slow-running ETL/ELT jobs using push-down optimization and partitioning techniques to manage large volumes of data.
* Utilized Snowflake data warehouse for building and optimizing data pipelines and ETL processes, ensuring efficient data storage and retrieval.
* Extensive experience with Azure DevOps for CI/CD pipelines and infrastructure automation, as well as Azure Analysis Services for building and optimizing analytical models to support business intelligence
* Good working experience on Spark (spark streaming, spark SQL) with Scala and Kafka. Worked on reading multiple data formats on HDFS using Scala.
* Strong expertise in healthcare data integration, including handling HL7 and FHIR data formats, ensuring regulatory compliance and data integrity.
* Good understanding and knowledge of NoSQL databases like MongoDB, PostgreSQL, HBase and Cassandra.
* Extensive hands-on experience in writing Hadoop jobs for data analysis as per the business requirements using Hive and worked on HiveQL queries for required data extraction, join operations, writing custom UDF's as required and having good experience in optimizing Hive Queries.
* Experience in importing and exporting data using Sqoop from S3 to Relational Database systems and vice-versa and load into Hive tables, which are partitioned.
* Implemented data security measures such as encryption and access controls in AWS and Azure environments to protect sensitive data.
* Hands on experience with data ingestion tools Kafka, Flume, and workflow management tools Oozie.
* Experience in Data Modelling using Dimensional Data Modelling techniques like Star Schema and Snowflake Modeling
* Strong hands-on experience using Teradata utilities - BTEQ, Fast Load, Multiload, Fast Export, Tpump, and Unix Shell scripting.
* Designed and implemented a graph schema for a social networking platform, accommodating dynamic relationships and user interactions.
* Possess in-depth knowledge of Database Concepts, Design of algorithms, SDLC, OLAP, OLTP, Data marts and Data Lake.
* Experience in all stages of the Software Development Lifecycle (SDLC) – Agile, Scrum, and Waterfall methodologies, right from Requirement analysis to development, testing, and deployment.
* Strong interpersonal skills with a proven ability to collaborate effectively across cross-functional teams, ensuring seamless communication and coordination in high-pressure environments.
* Demonstrated problem-solving skills by resolving complex technical challenges, optimizing workflows, and implementing innovative data solutions to meet business objectives.
* Designed APIs for serving AI models in real-time using Fast API and Pydantic, ensuring robust input validation and seamless integration with existing systems.
* Built and deployed ML models for classification, regression, and clustering tasks, leveraging tools like Scikit-learn and PyTorch for real-world applications.

**Technical Skills**

|  |  |
| --- | --- |
| **Big Data Technologies** | Kafka, Cassandra, Apache Spark, Spark Streaming, Apache Flink, HBase, Flume, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper |
| **Hadoop Distribution** | Cloudera CDH, Apache, AWS, Hortonworks HDP |
| **Programming Languages** | SQL, PL/SQL, T-SQL, Python (Pandas, NumPy, Scikit-learn, Pydantic), R, PySpark, Pig, Hive QL,  Scala, Shell Scripting, Regular Expressions |
| **Spark components** | RDD, Spark SQL (Data Frames and Dataset), and Spark Streaming |
| **Cloud Infrastructure** | AWS, Azure, GCP, Azure Analysis Services |
| **Databases** | Oracle, Teradata, My SQL, SQL Server, NoSQL Database (HBase, MongoDB, Aerospike), PostgreSQL |
| **Scripting & Query Languages** | Shell scripting, JSON, SQL |
| **Version Control** | CVS, SVN and Clear Case, GIT, ADO (Azure DevOps) |
| **Build Tools** | Maven, SBT |
| **Containerization Tools** | Kubernetes, Docker, Docker Swarm |
| **Reporting Tools** | Junit, Eclipse, Visual Studio, Net Beans, Azure Databricks, UNIX Eclipse, Linux, Google Shell, Unix, Power BI, SAS, and Tableau |
| **API Frameworks** | Flask, Fast API, Django REST Framework |
| **Artificial Intelligence & Machine Learning** | Scikit-learn, PyTorch, Pydantic, Natural Language Processing (NLP), and AI model deployment. |

**Project Experience**

# Senior Azure Data Engineer

**Client:** *Fifth Third Bank, Evansville, IN* **[June 2022 –Present]**

**Responsibilities:**

* Worked on Python, shell scripting, and bash, with expertise in connecting to on-premises Data Lake and analyzing application programs.
* Proficient in writing PySpark scripts for data transformations, aggregations, and machine learning tasks, and having experience in using PySpark and DataFrames for efficient data processing.
* Exposed transformed data in Azure Databricks platform to parquet formats for efficient data storage.
* Created Azure Data Factory pipelines for ETL processes, bulk copying multiple tables from relational databases to Azure Data Lake Gen2.
* Created SSIS Packages using different types of tasks and with error handling.
* Ingested data into Azure Blob Storage and processed it using Databricks, writing Spark, Scala scripts and UDFs for ETL transformations on large datasets.
* Designed and implemented real-time data streaming solutions using Apache Kafka, enabling continuous data ingestion and processing.
* Experience in complete SSIS life cycle in creating SSIS packages, building, deploying, and executing the packages in both the development and production environments.
* Integrated Azure Functions with Kafka for event-driven data processing and integration workflows, automating data processing tasks based on triggers and events.
* Implemented event-driven ETL workflows in Azure Data Factory using Azure Functions and Logic Apps, enabling real-time data processing and automation of business processes within data pipelines.
* Developed and optimized Spark ETL applications on Azure Databricks, leveraging Python and Scala programming languages.
* Leveraged Azure DevOps for version control, CI/CD pipelines, and infrastructure automation, streamlining deployment processes and ensuring consistency across environments.
* Integrated Azure DevOps pipelines for managing CI/CD workflows and implemented Azure Analysis Services for advanced data analysis and reporting.
* Integrated Snowflake with Azure Data Factory to orchestrate data pipelines and automate ETL workflows, enabling efficient data movement from Azure data lakes to Snowflake.
* Designed scalable and fault-tolerant Kafka architectures on Azure Cloud, utilizing Azure VMs and managed Kafka services such as Azure HDInsight.
* Wrote optimized T-SQL scripts to handle large datasets efficiently, supporting critical ETL pipelines and improving query performance.
* Developed advanced T-SQL queries for complex data analysis and reporting.
* Conducted unit testing and performance tuning of Informatica mappings and workflows to ensure optimal data processing performance.
* Experience in creating complex SSIS Packages with error handling using control and dataflow elements.
* Developed and maintained Power BI reports and dashboards to visualize and analyze data trends, patterns, and insights, enabling stakeholders to make informed decisions.
* Utilized Power BI's advanced features such as DAX (Data Analysis Expressions) for creating complex calculations and measures to derive actionable insights from data.
* Integrated Python scripts into Power BI for advanced data preprocessing, cleansing, and analysis, leveraging Python's extensive libraries for statistical analysis and machine learning.
* Integrated Informatica PowerCenter with Azure services such as Azure Data Lake Storage and Azure SQL Database to streamline data movement and processing workflows.
* Implemented and optimized Snowflake data warehouse on Azure Cloud platform for efficient data storage and processing.
* Designed and developed Snowflake schemas, tables, and views to support complex analytical queries and reporting requirements.
* Day-to-day responsibility includes developing ETL Pipelines in and out of the data warehouse, developing major regulatory and financial reports using advanced SQL queries in Snowflake.
* Hands-on experience with Informatica Cloud Integration services, including Cloud Data Integration, Cloud Application Integration, and Cloud B2B Gateway.
* Developed and deployed SSIS packages, configuration files, and scheduled jobs to run the packages to generate data in CSV files. Maintained TFS Source Control server for versioning the database objects and production releases.
* Integrated Azure Key Vault for secure key management, Azure Disk Encryption to encrypt virtual machine disks at rest, and Azure Functions for automating data encryption and decryption processes within Snowflake workflows, ensuring robust data security in the Azure environment.
* Integrated Azure Active Directory (AAD) authentication and authorization for securing access to Azure data services and resources.

**Environment:** Azure Data Factory, Azure Databricks, Data Pipelines, Azure Data lakes, Azure Active Directory, Azure Key Vault, Informatica, Python, Pyspark, Scala, Bash, HDFS, Hadoop, Snowflake, Yarn, MapReduce, Hive, Sqoop, Oozie, Kafka, Spark SQL, Spark Streaming, Eclipse, Oracle, Teradata, PL/SQL, T-SQL, UNIX Shell Scripting, Power BI.

# Senior AWS Data Engineer

**Client:** *Change Healthcare, Nashville, TN***, [Nov 2020 – June 2022]**

**Responsibilities:**

* Gathered business requirements and converted them into SQL stored procedures for database-specific projects.
* Collaborated with cross-functional teams to gather requirements and translate them into SSRS report designs and SSIS package specifications.
* Developed Python-based Spark applications for ETL tasks, utilizing AWS Glue for seamless data ingestion and transformation.
* Developed Python-based Spark applications for performing data cleansing, event enrichment, data aggregation, denormalization, and data preparation needed for machine learning and reporting teams to consume.
* Created ETL workflows in AWS Glue to extract data from various sources like S3, DynamoDB, and Oracle databases, transforming it for analytics purposes.
* Designed and implemented ETL processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift, improving data accessibility and analytics.
* Streamed real-time data by integrating Kafka with Spark for dynamic price surging using machine learning algorithms.
* Designed and implemented end-to-end data pipelines using AWS services such as Lambda, Glue, and Step Functions, orchestrating data movement and processing.
* Utilized Python and AWS SDKs to create automated data ingestion processes from S3 buckets, DynamoDB tables, and Oracle databases into Redshift, ensuring data accuracy and timeliness.
* Configured DynamoDB Streams and Lambda functions to capture real-time data changes and load them into Redshift for near real-time analytics.
* Leveraged Databricks to design, build, and deploy scalable data pipelines, enabling efficient data ingestion and processing for large-scale datasets.
* Constructed data pipelines in Airflow within AWS for ETL-related jobs using different Airflow operators while scheduling all data pipelines by creating DAGs in Airflow, ensuring timely data processing and delivery.
* Worked on SQL Server Integration Services (SSIS) and SQL Server Reporting Services (SSRS), migrated multiple databases, and fetched huge data using SSIS packages and scheduled as jobs.
* Utilized SSIS control flow tasks and data flow components to orchestrate and manage complex ETL workflows, optimizing data loading performance.
* Integrated Oracle databases into data pipelines, leveraging PL/SQL stored procedures, bulk operations, and procedural logic for efficient data extraction, transformation, and loading (ETL)
* Processed healthcare datasets in HL7 v2 and FHIR formats, leveraging Python and custom parsers to normalize and transform clinical records for downstream analytics and ETL ingestion
* Ensured efficient data processing and optimization within Oracle environments, handling large datasets effectively.
* ETL pipelines in and out of the data warehouse using a combination of Python and Snowflake’s Snow SQL. Writing SQL queries against Snowflake.
* Redesigned the Views in Snowflake to increase performance.
* Installed Docker Registry for local upload and download of Docker images and from Docker Hub and created Docker files to automate the process of capturing and using the images.
* Experience in integrating Jenkins with various tools like Maven (Build tool), Git (Repository), SonarQube (code verification), Nexus (Artifactory), and implementing CI/CD automation for creating Jenkins pipelines programmatically, architecting Jenkins Clusters, and scheduling builds day and overnight to support development needs.
* Designed and developed Tableau reports and dashboards to provide actionable insights for business decision-making.
* Involved in troubleshooting, performance tuning of reports, and resolving issues within Tableau Server and Reports.
* Created SSIS packages and scheduled jobs through SQL agents.
* Implemented AWS IAM roles and policies for fine-grained access control, ensuring data security and compliance.
* Utilized AWS Key Management Service (KMS) for encryption of sensitive data at rest and in transit, maintaining data integrity and confidentiality.
* Documented ETL workflows, data mappings, and transformation rules for future reference and maintenance, maintaining data lineage and transparency.

**Environment:** Python, Pyspark, Scala, Kafka, HBase, Docker, Kubernetes, AWS, EC2, S3, Lambda, Cloud Watch, Auto Scaling, EMR, AWS Key, IAM, GIT, Airflow, Redshift, Jenkins, ETL, Spark, Hive, Athena, Sqoop, Pig, Oozie, Spark Streaming, Data pipelines, Dynamo DB, Databricks, Snowflake Tableau, GIT Micro Services.

# Data Engineer

**Client:** *Global Atlantic financial group***,** *Indianapolis, IN* **[June 2018 – Oct 2020]**

**Responsibilities:**

* Collaborated with team members and stakeholders in the design and development of the data environment.
* Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.
* Designed solutions to process high-volume data stream ingestion, processing, and low-latency data provisioning using Hadoop Ecosystems Hive, Pig, Scoop, Kafka, Python, Spark, Scala, and NoSQL.
* Designed and implemented big data ingestion pipelines to ingest multi-TB data from various data sources using Kafka and Spark Streaming, including data quality checks, transformation, and efficient storage formats.
* Performed data wrangling on multi-terabyte datasets from various data sources for a variety of downstream purposes such as analytics using Spark.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.
* Used Hive SQL, Presto SQL, and Spark SQL to query the data and to load/extract the data.
* Designed and built multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Redshift for large-scale data handling, processing millions of records every day.
* Migrated on-premise database structure to Redshift Data Warehouse.
* Built performant, scalable ETL processes to load, cleanse, and validate data.
* Created various complex SSIS/ETL packages to extract, transform, and load data.
* Developed SSRS reports and SSIS packages to extract, transform, and load data from various source systems.
* Worked on SQL Server Integration Services (SSIS) and SQL Server Reporting Services (SSRS), migrated multiple databases, and fetched huge data using SSIS packages and scheduled as jobs.
* Used ADO.Net data objects such as Connection, Command, Data Adapter, Data Reader, Dataset, Data Table, and XML for consistent access to SQL data sources.
* Managed security groups on AWS, focusing on high availability, fault tolerance, and auto-scaling using Terraform templates.
* Implemented Continuous Integration and Continuous Deployment (CI/CD) with AWS Lambda and AWS Code Pipeline.
* Worked on publishing interactive data visualization dashboards, reports, and workbooks on Tableau and SAS Visual Analytics.
* Analyzed the existing application programs and tuned SQL queries using Execution Plan, Query Analyzer, SQL Profiler, and Database Engine Tuning Advisor to enhance performance.

**Environment:** Oracle, Kafka, Python, Redshift, Informatica, ETL, Scala, AWS, EC2, S3, SQL Server, Erwin, RDS, NOSQL, MySQL, Dynamo DB, PostgreSQL, Tableau, Git Hub, SSIS, SSRS, SQL, PL/SQL, T-SQL, Pig, Scoop.

# Jr. Data Engineer

**Client:** *Apple***,** *Cupertino, California* **[April 2017 – June 2018]**

**Responsibilities:**

* Utilized Azure Analysis Services for building and deploying OLAP models to support business intelligence efforts.
* Experienced in the progress of real-time streaming analytics data pipeline. Built connections between Event Hub and Streaming Analytics.
* Data extraction from various sources, transformation, and loading into the target SQL Server Database. Implemented Copy activity, Custom Azure Data Factory Pipeline Activities for On-cloud ETL processing.
* Worked on migrating SQL databases to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Databricks, and Azure SQL Data Warehouse and controlling and granting database access.
* Migrated on-premises databases to Azure Data Lake Store using Azure Data Factory.
* Developed data pipeline using Sqoop to ingest customer behavioral data and purchase histories into HDFS for analysis.
* Involved in importing real-time data to Hadoop using Kafka and implemented the Oozie job for daily imports.
* Involved in complete big data flow of the application starting from data ingestion from upstream to HDFS, processing, and analyzing the data in HDFS.
* Data ingestion to one or more Azure data services like Azure Data Factory and Databricks.
* Responsible for importing data to HDFS using Sqoop from different RDBMS servers and exporting data using Sqoop to the RDBMS servers.
* Exposed transformed data in the Azure Spark Databricks platform to Parquet formats for efficient data storage.
* Enhanced and optimized product Spark code to aggregate, group, and run data mining tasks using the Spark framework.
* Good hands-on Azure Data Factory, worked on creating dependencies of activities in Azure Data Factory.
* Created partitioned and bucketed Hive tables in Parquet file formats with Snappy compression and then loaded data into Parquet Hive tables from Avro Hive tables.

**Environment:** Azure, HDFS, Hadoop, Yarn, MapReduce, Hive, Sqoop, Oozie, Kafka, Spark SQL, Spark Streaming, Eclipse, Oracle, Teradata, PL/SQL, UNIX, Shell Scripting.

# ETL Developer

**Client:** *Wipro Technologies, Hyderabad, India,* **[July 2014 – Dec 2016]**

**Responsibilities:**

* Conducted data comparison across various databases to ensure data integrity and investigate data quality issues during data transformation and loading processes.
* Analyzed and investigated data quality to identify any data loss or corruption during data loads.
* Created and organized HDFS over a staging area for efficient data storage and retrieval.
* Imported legacy data from SQL Server and Teradata into target systems for processing.
* Developed a raw layer of external tables containing copied data from HDFS.
* Created internal tables in Hive for data manipulation and organization in the data service layer.
* Exported data into SQL Server by creating staging tables to load and process data.
* Wrote Python scripts to standardize and manipulate data frames for uniform formatting.
* Developed SQL scripts in Teradata and SQL Server Management Studio for data upload, retrieval, manipulation, and handling of sensitive data.
* Implemented UNIX scripts to define workflow and automate data processing tasks.

**Environment:** HDFS, Python, Hadoop, Hive, HBase, MapReduce, Spark, SQL Server, PostgreSQL, Unix.

**Educational Background:**

SASTRA Deemed University of Engineering Thanjavur, Tamil Nadu

Bachelor of Technology in Electrical and Electronics Engineering **August 2009 – May 2013**