**Sudhir Saxena**

**: sudhirsxn84@gmail.com** **: +1-214-940-8509**

**LinkedIn:** [**https://www.linkedin.com/in/sudhir-saxena-45b9a524/**](https://www.linkedin.com/in/sudhir-saxena-45b9a524/)

 **TOOLS & TECHNOLOGIES:**

# AWS Certified Data Engineer | AWS Certified Solutions Architect | AWS Certified Machine Learning – Specialty |

# PySpark | AWS | GCP | Spark | Databricks | Glue | Lambda | Python | Snowflake | Hudi

**PROFESSIONAL EXPERIENCE**

Results-driven AWS Data Engineer with over 15 years of experience designing, developing, and optimizing large-scale, cloud-native data platforms. Proven track record of leading end-to-end data architecture initiatives on AWS, including data lake design, ETL pipeline automation, real-time streaming, and cloud migration projects. Skilled in leveraging services like AWS Glue, Redshift, EMR, Lambda, and Athena to drive business insights, reduce infrastructure costs, and enable real-time analytics. Strong leadership in cross-functional teams, mentoring junior engineers, and aligning data strategies with enterprise goals. Passionate about scalable, secure, and high-performance data engineering solutions.

* Designed and implemented end-to-end data pipelines using AWS Glue, Lambda, S3, Redshift, and Athena, enabling real-time analytics for enterprise applications, reducing ETL processing time.
* Led a team of data engineers in developing scalable data lake architectures with AWS Lake Formation, improving data governance and security compliance.
* Migrated on-premises data warehouse (Teradata) to AWS Redshift, handling over TB of data with zero downtime and achieving cost savings on infrastructure.
* Developed CI/CD workflows using AWS CodePipeline and Terraform for infrastructure as code (IaC), reducing manual deployment errors.
* AWS Certified data engineer, Leading the design and implementation of cloud-native architectures using core AWS services such as EC2, EMR, Glue, S3, Glue, Lambda, RDS, ECS, Step Function, Redshift, CloudFormation, and IAM. End-to-end design, development, and implementation of software systems, ensuring high performance, scalability, and security.
* Worked extensively in Hadoop ecosystem, Bigdata, Pyspark, Python, Spark, AWS, Redshift, Glue, Lambda, Warehousing, ETL tool, Business Intelligence, Data Analytics, Data Integration, Implementation and Maintenance and the related Database Platforms.
* Working as technical lead to play a key role in software development team, make high-level design and architecture decisions, review code for quality, defect fixing during QA and prod phase, performance, and maintainability.
* 5 years of AWS experience with data warehouse application design, development, and production go-live deployments.
* 2 years of Google Cloud Platform experience with data warehouse application design, development, and production go-live deployments.
* Hands on experience in writing python as ETL framework and Pyspark to process huge amount of data daily. implementing Spark using python and Spark SQL for faster data processing.
* Created framework which pulls data from AWS S3 bucket, implement all the transformation using pyspark, CDC logic using apache hudi, and load the data into Redshift.
* Leading and mentoring a team of developers and cloud engineers in delivering cloud-based projects, ensuring adherence to AWS well-architected frameworks.
* Excellent understanding of project issues, tracking of issues, solving issues and closing issues.

**TECHNICAL SKILLS:**

| Skills Summary  | PySpark, AWS, AWS EMR, Hadoop, Bigdata, Hive, GCP, Big Query, DataProc, Sqoop, Control-M, Autosys, Teradata, Oracle, DB2, Unix, ETL Glue, Redshift, Apache Hudi  |
| --- | --- |
| Programming Language  | Python, Unix |
| Databases  | AWS Redshift, Snowflake, Oracle, DB2, PL/SQL, SQL, Dynamo DB, AWS RDS |
| Monitoring tools  | Control-m, Autosys, Airflow, Step function  |

**PROJECT DESCRIPTION:**

| Project Name  | FMCSA |
| --- | --- |
| Role  | Senior Data Engineer  |
| Organization  | Reveal Global Consulting (RGC)  |
| Duration  | Apr 2024 to at Present  |

FMCSA Working on Data modernization from Legacy system to AWS cloud project to process huge amount of data which was loading into AWS redshift cloud Datawarehouse.

Carried out the following activities:

* Understand business requirements from Client, created source to target mapping to design Data pipeline through PySpark to fetch the data from S3 bucket where Upstream places the API joson file on daily basis as batch process.
* Designed and implemented end-to-end data pipelines using AWS Glue, Lambda, S3, Redshift, and Athena, enabling real-time analytics for enterprise applications, reducing ETL processing time by 40%.
* Led a team of 5 data engineers in developing scalable data lake architectures with AWS Lake Formation, improving data governance and security compliance.
* Migrated on-premises data warehouse (Teradata) to AWS Redshift, handling over 30TB of data with zero downtime and achieving 60% cost savings on infrastructure.
* Developed CI/CD workflows using AWS CodePipeline and Terraform for infrastructure as code (IaC), reducing manual deployment errors by 90%.
* Implemented enterprise-wide data cataloging using AWS Glue Data Catalog, improving data discoverability and access for analytics teams.
* Integrated third-party data sources via REST APIs and automated ingestion with Python and AWS Lambda.
* Leading the design and implementation of cloud-native architectures using core AWS services such as EC2, EMR, Glue, S3, Glue, Lambda, RDS, ECS, Step Function, Redshift, CloudFormation, and IAM. End-to-end design, development, and implementation of software systems, ensuring high performance, scalability, and security.
* Mentored junior engineers and established data engineering best practices and coding standards.
* Developed monitoring dashboards using CloudWatch, Datadog, and custom Python scripts for real-time pipeline observability.
* Created Data Lake for MCMIS application which pulls the data from Safer DB and create staging file on S3 bucket using Database management services (DMS) and loaded into S3 bucket in parquet format.
* Created pyspark script framework which pulls data from S3 bucket in parquet format and exploded, flattened data and implemented all the logics in spark SQL, Dataframe , created temp view, cached the dataframe to reuse the dataframe and finally prepared final dataframe to load the data into Redshift.
* Created Step functions which is serverless orchestration service that easily coordinated multiple Lambda functions to integrate all the data pipeline, schedule, ran, debug and changes whenever needed.
* Deployed the objects and application code using Aws Code commit.
* Worked on defects for any issues, data mismatches which are created by QA Team.

**Skills:** PySpark, AWS, Spark,S3, Step Functions, AWS Glue, Lambda, Redshift, Python, SQL, EMR, Athena

| Project Name  | BPPSL  |
| --- | --- |
| Role  | Senior Data engineer |
| Organization  | Kommforce Solution, Dallas, TX  |
| Duration  | Aug 2023 to Apr 2024 |

Working on Data modernization from Abinitio to AWS cloud project to process huge amount of data which was loading into Teradata currently that needs to be loaded into AWS redshift cloud Datawarehouse.

Carried out the following activities:

* Understand business requirements from Client, created source to target mapping to design Data pipeline through PySpark to fetch the data from S3 bucket where Upstream places the xml file on daily basis as batch process.
* Created Data Lake for BPPSL application which pulls the data from S3 bucket as raw data in .gz file and loaded into S3 bucket in parquet format using AWS Glue.
* Created pyspark script framework which pulls data from S3 bucket in parquet format and exploded, flattened data and implemented all the logics in spark SQL, Dataframe , created temp view, cached the dataframe to reuse the dataframe and finally prepared final dataframe to use in Apache Hudi.
* Using Apache Hudi, implemented CDC logic between target Redshift DB and data coming from S3 bucket in Hudi table to insert, Update, delete data into target Redshift table through EMR Clusters.
* Created Step functions which is serverless orchestration service that easily coordinated multiple Lambda functions to integrate all the data pipeline, schedule, ran, debug and changes whenever needed.
* Deployed the objects and application code using bitbucket, CI/CD Bamboo and GitHub.
* Worked on defects for any issues, data mismatches which are created by QA Team.

**Skills:** PySpark, AWS, S3, EMR, Step Functions, AWS Glue, Lambda, Redshift, Apache Hudi, Python, Unix

| Project Name  | Cloud Data Warehouse  |
| --- | --- |
| Role  | Technical Lead |
| Organization  | Mphasis Corporation, Dallas, Texas  |
| Duration  | Dec 2021 to July 2023  |

Working on Cloud Data Warehouse Migration project to migrate the huge amount of data from Hadoop Hive On-Prem to GCP Big Query. As part of this project, we fetched the data from hive table and copy to GCS bucket using gsutil cp or Hadoop distcp and created the stage table. We did transformation, added a few business columns and created BQ Target table.

Carried out the following activities:

* Understand business requirements from Business, created source to target mapping to design Data pipeline through PySpark, Python, GCP Dataproc cluster, GCS, BigQuery.
* Worked on Framework to extract the data from Teradata and Hive table and loaded into GCP Big Query through GCS bucket. It's a migration framework to load all the historical data from prem data to Cloud BigQuery.
* Designed spark job to pick GCS files and performs transformation and pushes to BQ Stage table and from BQ stage data are type casted and sent over to BQ Target tables.
* Scheduled the jobs on Airflow cloud composer - Orchestration Framework using Python, Airflow packages, GitLab for data processing pipeline. designed Airflow Cloud composer DAGs with TaskGroup, XComs, Operators (Bash, Branch, Python, Google Clouse Storage,
* Designed the python module, DAG to automate the process for create the historical views for source CDC tables in GCP BigQuery.

**Skills:** PySpark, GCP, BigQuery, Dataproc, Airflow, Hadoop, Bigdata, Hive, Sqoop, Control-M, Python, Teradata, Unix

| Project Name  | Securities Pricing Platform (SPP)  |
| --- | --- |
| Role  | Technical Lead |
| Organization  | Randstad Technologies  |
| Duration  | Dec 2019 to Dec 2021   |

Application is designed to get the pricing details for majority of the firm’s brokerage system (GWIM & GBAM) every end-of-day by running the Merrill’s Lynch nightly batch jobs.

Carried out the following activities:

* Understood business requirement from Business, created source to target mapping to design Data pipeline through PySpark, Python, AWS, S3, ETL Glue, Athena, EC2 cluster.
* Developed Spark and Hive jobs for data transformation, enrichment, and aggregation to support downstream analytics, risk, and compliance functions.
* Integrated legacy Merrill Lynch nightly batch jobs with modern Big Data architecture, ensuring compatibility and seamless data flow.
* Ensured data integrity through validation, deduplication, and anomaly detection mechanisms
* Created Data Lake for SPP application process which pulls the data from Oracle database using Sqoop import and loads the data into Hive table using PySpark, Hive and Hadoop ecosystem.
* Unloaded the data from Oracle SPP DB and loaded the data into S3 bucket and processed those files along with supporting files on AWS EMR cluster using PySpark code in SPP application.
* Loaded bulk data load from S3 to Snowflake using copy statement through python programming and SnowSql.
* Scheduled Control-m job to invoke shell script to extract the data from Oracle DB and copy into AWS S3 bucket and finally load into Snowflake.
* Created Snowpipe to enable loading latest brokerage data from files as soon as new files arrive in a stage table for one of the modules in SPP application. Created snowflake stage table from AWS S3 bucket and using snowpipe loaded data from stage tables to Snowflake table.

**Skills:** PySpark, AWS, S3, ETL Glue, Snowflake, EC2, AWS Lambda, Hadoop, Bigdata, Hive, Sqoop, Control-M, Python

|  Project Name  | LTB- Real-estate Mortgage  |
| --- | --- |
| Client  | USAA, Texas  |
| Role  | Technical Lead |
| Organization  | HCL America Inc  |
| Duration  | Oct 2014 to Dec 2019   |

LTB- Real-estate MortgageProject is the project where we are extracting data from SQL Server as source for all the Loan Borrowers and Loading the data into data Warehouse and from Data warehouse to Data Mart based on business requirement and Subject area. also, loading the data into Hadoop (Hive DB) through ETL E3 Framework.

Carried out the following activities:

* Understand business requirements from Business, does data profiling, prepare source to target mapping to create data modelling.
* Created Data Lake for LTB application which pulls the data from Oracle database using Sqoop import and loads the data into Snowflake using Python, Hive and Hadoop ecosystem.
* Created Snowpipe to enable loading data from files as soon as new file arrives in a stage table. Created snowflake stage table from AWS S3 bucket and using snowpipe loaded data from stage tables to Snowflake table.
* Extracted the data from source system and loaded it into amazon S3. from S3, loaded into Snowflake internal stage and finally copied into snowflake table.

**Skills:** PySpark, Snowflake, AWS, S3, ETL Glue, AWS EC2, AWS Lambda, Hadoop, Bigdata, Hive, Control-M, Python,

| Project Name  | Insurance Data Warehouse (IDW)  |
| --- | --- |
| Role  | Senior ETL Developer  |
| Organization  | Mindtree Ltd, Chennai  |
| Duration  | Aug 2011 to Oct 2014  |

* Involved as an ETL lead in gathering requirements, analyzing and developing end-to-end applications.
* Created DataStage job to load the data into Dimension and Fact table to have data into warehouse.
* Helped team on-board data, create various knowledge objects, good knowledge on ETL, data warehousing concepts, Unix shell scripting for advance backend integrations.

**Skills:** DataStage 8.7, Unix Shell scripting, DB2, SQL Developer

**EDUCATION DETAILS:**

Master of Computer Applications (MCA) from College of Engineering Guindy, Anna University, Chennai, India in 2011.

**CERTIFICATIONS:**

**AWS Certified Machine Learning – Specialty:** <https://www.credly.com/earner/earned/badge/f9b41b83-dfc2-4d74-aef8-44e7cd7a5b12>

**AWS Certified Solutions Architect:** <https://www.credly.com/earner/earned/badge/87e98ea4-4107-4372-b7bf-49d1399189e7>

**AWS Certified Data Engineer:** <https://www.credly.com/earner/earned/badge/2d3368ef-5142-4afb-8eeb-9f7c268100c3>

**AWS Certified Cloud Practitioner**:  <https://www.credly.com/badges/6528a645-7c8b-4f1e-800c-00d79906a60a>

**AWARDS/ACHIEVEMENT**

* Got Appreciation Certificate and Award for ‘**Training Delivery and Development’** for Oct –Dec 2014 in HCL Technologies Ltd
* Got **‘Certificate of Excellence’** for Certified Datastage trainer in HCL technologies Ltd in 2015-2016.
* **Most Innovative Campus Mind** in March 2012, by Mindtree Ltd.