Shashank Gorle

[**shashank1172019@gmail.com**](mailto:shashank1172019@gmail.com)

**Contact Details: 425-587-1019**

[www.linkedin.com/in/shashank-g-a96915a9](http://www.linkedin.com/in/shashank-g-a96915a9)

**Professional Summary:**

* Data Engineer with around 9 years of experience in Cloud (GCP, AWS, Azure) Data warehousing, Data engineering, Feature engineering, Hadoop big data, ETL/ELT, and Business Intelligence.
* Specialize in **GCP, AWS and Azure frameworks**, Cloudera, Hadoop Ecosystem, Spark/Py Spark/Scala, Data bricks, Hive, Redshift, Snowflake, relational databases, tools like Tableau, Airflow, DBT, Presto/Athena, Data DevOps Frameworks/Pipelines Programming skills in Python.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to **Google Cloud Platform (GCP) using BIG query, Cloud Data Proc, Google Cloud Storage, Composer. ﻿﻿**
* ﻿﻿Hands on experience in **GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/Sub cloud shell, GSUTIL, Data Proc, Stack driver.**
* Hands-on experience with **Google Kubernetes Engine (GKE**) including cluster deployment, management, autoscaling, logging, and monitoring. Experience with Kubernetes concepts like pods, services, deployments, etc.
* Expertise in deploying and managing production workloads on GKE. Experience with using **GKE ingress controllers, node pools, network policies, and integration with other GCP services.**
* Deep knowledge of Google BigQuery including schema design, partitioning, clustering, performance optimization, and querying large datasets. Experience with BigQuery ML for predictive analytics.
* Proficiency in Google Cloud SQL including deployment, configuration, replication, backups, and integration with GCP tools. Experience in right-sizing, scaling, and optimizing Cloud SQL performance.
* Hands on experience on Google Cloud Platform (GCP in all the big data products **Big Query, Cloud Data Proc, Google Cloud Storage, Composer (Airflow as a service)**
* Led **multiple enterprise-scale data migration and ETL transformation initiatives on Google Cloud Platform**; migrated on-premise and legacy data warehouses to BigQuery, architected data pipelines using Cloud Dataflow, Dataproc, and Composer for seamless data integration.
* Experience with an in - depth level of understanding in the strategy and practical implementation of AWS Cloud-Specific technologies including **EC2, EBS, S3, VPC, RDS, SES, ELB, EMR, ECS, Cloud Front, Cloud Formation, Elastic Cache, Cloud Watch, Red Shift, Lambda, SNS, Dynamo DB, Kinesis**.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Hands on experience on **AWS cloud services** (VPC, EC2, S3, RDS, Redshift, Data Pipeline, EMR, DynamoDB, Workspaces, Lambda, Kinesis, RDS, SNS, SQS).
* Experience in Hadoop Ecosystem components like Hive, HDFS, Sqoop, Spark, Kafka, Pig.
* Good understanding of Hadoop architecture and Hands on experience with Hadoop components such as Resource Manager, Node Manager, Name Node, Data Node and Map Reduce concepts and HDFS.
* Expertise in Data Migration, Data Profiling, Data Ingestion, Data Cleansing, Transformation, Data Import, and Data Export using multiple ETL tools such as Informatica Power Centre.
* Hands on Spark RDD, Data frame API, Data set API, Data Source API, Spark SQL, and Spark Streaming.
* Experienced in managing **Azure Data Lakes** (ADLS) and **Data Lake Analytics** and an understanding of how to integrate with other **Azure Services**.
* Proficient in SQL databases **MSSQL Server, MySQL (RDBMS), Oracle DB, Post gres, and MongoDB**.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop Components / Big Data** | HDFS, Hue, Map Reduce, PIG, Hive, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, pyspark Airflow, Snowflake Spark Component. |
| **Cloud Platforms** | GCP, AWS and Microsoft Azure |
| **GCP Cloud Platform** | Big Query, Cloud Data Proc, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Shell, GSUTIL, BQ Command Line, Cloud Data Flow |
| **AWS Cloud Platform** | EC2, EMR, Redshift, RDS, Lambda, Kinesis, SNS, SQS, AMI, IAM, Cloud formation |
| **Azure Cloud Platform** | Azure Data Lake, Data Bricks, Azure Data Factory, Azure Monitoring, Active Directory, Synapse, Key Vault, SQL Azure |
| **Databases:** | Oracle, Microsoft SQL Server, MySQL, DB2, postgres, MongoDB, Teradata. |
| **Programming Languages:** | Java, Scala, Impala, Python. |
| **Web Servers** | Apache Tomcat, WebLogic, GitHub. |
| **Methodologies** | Agile (Scrum), Waterfall, UML, Design Patterns, SDLC. |
| **Cloud Services:** | AWS, Azure Data Factory / ETL / ELT / SSIS Azure Data Lake, Storage Azure Data bricks |
| **Methodologies:** | Agile (Scrum), Waterfall, UML, Design Patterns, SDLC. |
| **Reporting and ETL Tools** | Tableau, Power BI, AWS GLUE, SSIS, SSRS, Informatica, Data Stage |

**Professional Experience**

**Centene Corporation St Louis, Missouri August 2022 to Present**

**Role: Senior GCP Data Engineer**

**Responsibilities:**

* Involved in full Software Development Life Cycle (SDLC) - Business Requirements Analysis, preparation of Technical Design documents, Data Analysis, Logical and Physical database design, Coding, Testing, Implementing, and deploying to business users.
* Worked on Google Cloud Platform (GCP) in all the big data products big query, Cloud Data Proc, Google Cloud Storage, Composer (Air Flow as a service).
* Involved in full Software Development Life Cycle (SDLC) - Business Requirements Analysis, preparation of Technical Design documents, Data Analysis, Logical and Physical database design, Coding, Testing, Implementing, and deploying to business users.
* Transformed and analyzed the data using Pyspark, HIVE, based on ETL mappings.
* Developed spark programs and created the data frames and worked on transformations.
* Experience in **migrating cron jobs to composer in GCP**.
* Migrated legacy on-prem data warehouses to **Google BigQuery on GCP, increasing scalability and enabling real-time analytics.**
* **Architected data pipelines on GCP using Dataflow, Dataproc, and Composer** to ingest and process data from diverse sources.
* Led offshore resources in developing ETL code and scripts to transform and load data into BigQuery and Cloud SQL.
* Analyzed source data coming different sources like SQL Server tables, XML files and Flat files then transformed according to business rules using Informatica and loaded the data in to target tables.
* Analyzed large and critical datasets using HDFS, HBase, Hive, Scala, HQL, PIG, Sqoop, Kubernetes and Zookeeper.
* Designed and developed several complex mappings using various transformations like Source Qualifier, Aggregator, Router, Joiner, Union, Expression, Lookup, Filter, Git Hub Update Strategy, Stored Procedure, Sequence Generator, etc.
* Skilled in optimizing costs on GCP including usage of committed-use discounts, interrupted VMs, autoscaling, right-sizing, RI utilization, and budget alerts. Experience with GCP Pricing Calculator and Billing console.
* Developed data marts in big data world in Big Query or on – premises Hadoop clusters.
* Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in the UNIX operating system as well.
* Used Apache airflow in GCP Composer environment to build data pipelines and explored various airflow operations like bash operator, Hadoop operators and branching operators.
* Used apache airflow in GCP composer environment to build data pipelines and used various airflow operators like bash operators, Hadoop operators and python callable and branching operators.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big query.
* Extensive use of cloud shell SDK in GCP to configure/deploy the services using GCP Big Query.
* Developed ELT processes from the files from abinitio, google sheets in GCP with compute being dataproc (pyspark) and big query.
* Hands-on experience with other GCP services like Cloud Storage, Cloud Functions, VPC, IAM, Cloud Monitoring, and Stackdriver Logging.
* Worked on Building data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Changing the existing Data Models using Erwin for Enhancements to the existing Datawarehouse projects.
* Worked on Google Cloud Platform (GCP) services like compute engine, Git Hub, cloud load balancing, cloud storage and cloud SQL.
* Used Talend connectors integrated to Redshift – BI Development for multiple technical projects running in parallel.
* Develop and deploy the outcome using spark and scala code in Hadoop cluster running on GCP.
* Built data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in Google Cloud components, Google container builders and GCP client libraries.
* Experienced with reporting tools, software, and other applications, including SQL Database, Looker, Tableau and Salesforce dashboards (reports designing tool) and Data warehouse.
* Supported various business teams with Data Mining and Reporting by writing complex SQLs using Basic and Advanced SQL including OLAP functions like Ranking, partitioning and windowing functions, Etc.
* Using g-cloud function with Python to load Data in to big query for on arrival csv files in GCS bucket.
* Extensively used spark SQL and Data frames API in building spark applications.

**Environment:** GCP, Cloud SQL, Big Query, Cloud DataProc, GCS, Cloud Composer, Informatica Power Center 10.1, Talend 6.4 for Big Data, Hadoop, Hive, Teradata, SAS, Teradata, Spark, Python, Java, SQL Server.

**Cummins Columbus, Indiana April 2019 to July 2022**

**Role: GCP Data Engineer**

**Responsibilities:**

* Developed ETL pipelines on GCP using Apache Beam and Data flow to process large-scale data in real-time, resulting in a 20% improvement in data processing time.
* Built and deployed data pipelines using Cloud Composer and Cloud Functions, enabling seamless integration with other GCP services such as BigQuery, Pub/Sub, and Cloud Storage.
* Implemented monitoring and alerting mechanisms using Stackdriver, enabling proactive issue identification and resolution in GCP data pipelines.
* Designed and executed end-to-end testing strategies for GCP data pipelines, ensuring the accuracy and completeness of data from ingestion to analysis.
* Utilized DevOps practices and tools such as Jenkins, Terraform, and Ansible to automate GCP infrastructure deployment and configuration, resulting in a 50% reduction in deployment time.
* Worked with Python, SQL, and Bash scripts to develop custom data transformations and data quality rules, resulting in a 25% reduction in data processing errors.
* Migrated batch ETL processes from on-prem Hadoop to GCP Dataproc, reducing costs by 40%.
* Developed data validation scripts on GCP using BigQuery and Dataproc.
* Experience in monitoring resource utilization, planning capacity, and optimizing performance of workloads on GCP.
* Knowledge of infrastructure-as-code tools like Terraform and Ansible to automate deployment and configuration of GCP resources.
* Developed and maintained CI/CD pipelines on GCP using Cloud Build and Cloud Run, enabling seamless code deployment and testing in a controlled environment.
* Implemented data versioning and lineage tracking using tools such as Data Catalog and Data Studio, enabling auditability and traceability of healthcare data in GCP.
* Conducted capacity planning and scaling of GCP data pipelines using Kubernetes and Cloud Autoscaling, ensuring optimal performance and cost-efficiency.
* Developed multi-cloud strategies in better using GCP (for its PAAS).
* Familiarity with database management using Python and SQL, including connecting to various database systems like MySQL, PostgreSQL, or SQLite.
* Applied geospatial data fusion techniques to combine information from GIS, remote sensing, and satellite imagery, enhancing spatial analysis and mapping capabilities.
* Utilized geospatial data fusion for applications such as environmental monitoring, urban planning, and disaster response.
* Designed and optimized real-time data fusion systems, meeting strict latency requirements for time-sensitive applications.
* Designed and developed Spark jobs with Scala to implement end-to-end data pipelines for batch processing.
* Developed data pipeline using Flume, Kafka, and Spark Stream to ingest data from their weblog server and apply the transformation.
* Setting up and configuring of New Relic, improved monitoring and performance of application in  New Relic.
* Developed data validation scripts in Hive and Spark and perform validation using Jupiter Notebook by spinning up the query cluster in EMR.
* Executed Hadoop and Spark jobs on EMR using data stored in Amazon S3.
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
* Developed PySpark script to encrypt the raw data by using hashing algorithms concepts on client-specified columns.
* Developed Stored Procedures, Views, and Triggers, and was responsible for the design, development, and testing of the database.
* Developed Python-based API (RESTful Web Service) to track revenue and perform revenue analysis.

**Environment:** GCP, Gcs Bucket, G-Cloud Function, Apache Beam,  New Relic, Cloud Data flow, Cloud Shell, Gsutil, Cloud SQL, Big Query, Cloud Data Proc, GCS, Cloud SQL, Cloud Composer, Data Fusion, Talend for Big Data, Airflow, Hadoop, Hive, Teradata, SAS, Teradata, Spark, Python, SQL Server, Kubernetes, Docker.

**Travelport Englewood, CO May 2017 to March 2019**

**Role: SR Data Engineer**

**Responsibilities:**

* Hands-on major components in Hadoop Echo Systems like Spark, HDFS, HIVE, HBase, Zookeeper, Sqoop, Oozie.
* Conducting Exploratory data analysis in Jupyter notebooks using Python libraries and sharing the data analysis
* Developing Sqoop jobs to ingest data from various system of records into Enterprise Data Lake.
* Development of Spark jobs in PySpark and Spark SQL to run on top of hive tables and create transformed data sets for downstream consumption.
* Working with business analysts to convert functional requirements into technical requirements and build appropriate data pipelines.
* Ingesting data from various source systems like Oracle, SQL Server, Flat files, JSONs.
* Performance tuning spark and hive jobs by reading execution plans, DAGs and Yarn logs.
* Creating generic shell scripts to submit Hadoop and spark jobs on EMR and on-prem edge node.
* Writing Complex SparkSQL codes to clean, join, transform and aggregate the datasets and publish them for Power BI team to produce operational scorecards.
* Writing custom python modules for reusable python code.
* Working on designing the MapReduce and Yarn flow and writing MapReduce scripts, performance tuning and debugging.
* Worked on migrating on-prem Hadoop cluster data and data pipelines to AWS Cloud.
* Created multiple proof-of-concepts using PySpark and deployed them on the Yarn cluster, comparing Spark's performance to that of Hive and SQL/Teradata.
* Integrated Teradata Warehouse into EMR cluster. Developed BTEQ scripts to load data from Teradata Staging area to Teradata data mart. Handled Error & tuned performance in Teradata queries and utilities.
* Solving appropriate partition, bucketing schemes and making sure correct load policies are employed so data can be stored as per requirements.
* Generates ETL scripts to transform, flatten, and enrich the data from source to target using AWS Glue and created event driven ETL pipelines with AWS Glue.
* Documenting data flow diagrams and technical logic in confluence
* Worked on setting up AWS DMS and SNS for data transfer and replication and used SQL on the new AWS Databases like RedShift and Relation Data Services.
* Created external tables with partitions using Hive, AWS Athena, and Redshift.
* Created and deployed Lambda functions in AWS using pre-built AWS Lambda Libraries, as well as Lambda functions in Scala using custom libraries.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
* Creating Oozie workflows, Coordinators and scheduling handshake jobs in Control-M.
* Created Session Beans and controller Servlets for handling HTTP requests from Talend
* Using Control-M for orchestration, workflow will be executed through Airflow, AWS EMR and Glue to run the jobs and glue catalog to capture the metadata, Lambda to process the source data freshness check and capturing the data quality metrics
* Working with governance teams to ensure metadata management, data lineage and technical metadata are correctly updated for each data asset.
* Working with production support teams and administration teams to ensure correct access controls are setup on each hive database.
* Working with master-feature branch model and commit the code with appropriate comments.
* Attending sprint planning, agile ceremonies and demo the work products on bi-weekly basis.

**Environment:** Cloudera Hadoop distribution, AWS EMR, Glue, S3, Athena, Hive, Impala, Lambda, Scala, PySpark, Spark SQL, Oracle 11g/12c, Teradata, Jira, BitBucket, Power BI, Control-M, Talend, CI/CD pipelines, GKE, Redshift, SAS.

**Avon Technologies Pvt Ltd Hyd India October 2015 to February 2017**

**Role: Data Engineer**

**Responsibilities:**

* Hands-on experience with Azure Synapse, Stream Analytics, Azure Event Hubs, and Azure Event Grid.
* Experience in Migrating SQL database to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, and Azure SQL Data Warehouse.
* Data Ingestion to Azure Services, including Azure Data Lake, Azure Storage, Azure SQL, and Azure DW as well as data processing in Azure Data bricks.
* Worked on design and development of Informatica mappings, workflows to load data into staging area, data warehouse and data marts in SQL Server and Oracle.
* Using Server Manager, created batches and sessions to transport data at present intervals and on-demand.
* Extracted, converted, and loaded data from various sources to Azure Data Storage Services utilizing Azure data factory and T-SQL for data lake analytics.
* Performed data transformations for ML OPs, including adding calculated columns, maintaining relationships, establishing various metrics, merging & appending queries, changing values, splitting columns, and grouping by Date & Time Column.
* Involved in Data Migration process using Azure by integrating with GitHub repository and Jenkins.
* Developed Python scripts to do file validations in Databricks and used ADF to automate the process.
* Developed audit, balancing, and control architecture utilizing SQL DB audit tables to manage the Azure ingestion, transformation, and load processes.
* Created tables in Azure SQL DW for business-related data reporting and visualization.
* Utilized Power BI desktop to generate visualization reports, dashboards, and KPI scorecards.
* Conceived, designed, and implemented ETL solutions with SQL Server Integration Services (SSIS).
* Created Pipelines which were built in Azure Data Factory using Linked Services/Datasets/Pipeline/ to extract, transform, & load data from a variety of sources including Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, & reverse.
* Created CI-CD Pipelines using Azure DevOps.
* Worked on various issues on existing Informatica Mappings to Produce correct output.
* Native integration with Azure Active Directory (Azure AD) and other Azure services enables to build modern data warehouse and machine learning and real-time analytics solutions.
* Used a blend of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics, gather, convert, and load the data from source systems to Azure Data Storage services.
* Configured Spark Streaming to receive real time data from Kafka. Used Backpressure to control message queuing in the topic.
* Created and monitored notifications for data integration events (success/failure)
* Collaborated with product managers, scrum masters, and engineers to build Agile processes and documentation projects for retrospectives, the backlog, and meetings.

**Environment:** Azure Data Storage, Azure Data Factory, Azure Services, Azure SQL server, Azure data warehouse, MySQL, ETL, Kafka, PowerBI, SQL Database, T-SQL, U-SQL, GitHub, Azure Data Lake, Azure Databricks, SSIS.

**Amigos Software Solutions, Hyd, India June 2013 to September 2015**

**Role: Big Data Engineer**

**Responsibilities:**

* Work closely with Business Analysts and Product Owner to understand the requirements.
* Used Joins in SPARK for making smaller datasets to large datasets without shuffling of data across nodes.
* Developed Spark Streaming jobs using Python to read messages from Kafka.
* Write Scala program for spark transformation in Dataproc.
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra
* Prototyped analysis and joining of customer data using Spark in Scala and processed it to HDFS
* Implemented Spark in EMR for processing Big Data across our One Lake in AWS System
* Consumed and processed data from DB2.
* Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources.
* Down loaded JSON files from AWS S3 buckets.
* Implemented ETL using AWS RedShift/Glue.
* Developed applications using spark to implement various aggregation and transformation functions of Spark RDD and Spark SQL.
* Used Scala components to implement the credit line policy based on the conditions applied on spark data frames.
* Data sources are extracted, transformed and loaded to generate CSV data files with Python programming and SQL queries.
* Using g-cloud function with Python to load Data into Big query for on arrival csv files in GCS bucket.
* Write a program to download a SQL Dump from there equipment maintenance site and then load it in GCS bucket. On the other side load this SQL dump from GCS bucket to MYSQL (hosted in Google cloud SQL) and load the Data from MYSQL to Big query using Python, Scala, spark and Dataproc.
* Process and load bound and unbound Data from Google pub/sub topic to Big query using cloud Dataflow with Python.
* Create firewall rules to access Google Data proc from other machines.

**Environment:** Spark, Scala, AWS, EMR, Redshift, EC2, Python, SQL, Nifi, Airflow Jupiter, Kafka.

**Education:**

Bachelor of Technology in Computer Science and Engineering from PSG Technology, Coimbatore, India in 2013.