SOWJANYA Y

**SENIOR CLOUD DATA ENGINEER**

**Phone**: 937-400-9094

**Email**: [y.sowjanya@outlook.com](mailto:y.sowjanya@outlook.com)

**LinkedIn**: <https://www.linkedin.com/in/y-sowjanya/>

# PROFESSIONAL SUMMARY

Seasoned and insightful Data Engineer with 9 years of accomplished experience, offering comprehensive expertise across a spectrum of various data engineering disciplines focusing on Data Acquisition, Modelling, Ingestion, Integration, Processing, Aggregation, and Visualization. Certified **Google Cloud Professional Data Engineer** with expertise in designing and implementing scalable data solutions using Google Cloud Platform (GCP). Proficient in developing, and optimizing end-to-end data pipelines, leveraging a diverse skill set that encompasses Apache Kafka, Python, and Google Cloud Platform (GCP), Amazon Web Services (AWS), Snowflake, MySQL, NoSQL. Proficient in

leveraging AWS services such as S3, **Redshift**, **Glue**, and **EMR** to extract, transform, and load (ETL) data from diverse sources. Adept in designing event-driven architectures powered by **Kafka**, **Amazon Kinesis**, **Google Cloud Pub/Sub** enabling real-time data streaming and fault-tolerant data processing systems. Mastery in **Python**, **PySpark**, **SparkSQL** scripts that empowers intricate data transformations and manipulations, utilizing libraries such as pandas, NumPy, and

scikit-learn. Fluent in harnessing GCP services like **Cloud Data Flow, Big Query, Cloud Data Prep, Cloud Data Fusion, Cloud SQL, Cloud Data Proc, Composer, Cloud Functions, Big Query, and Compute Engine** crafting scalable and efficient data solutions. Expertise in designing reports, performing analysis, interactive dashboards, and visualizations using **Tableau**, **PowerBI**, and **Quick Sight**. Recognized for orchestrating ETL optimizations, driving data processing efficiency enhancements by up to 50%, and ensuring meticulous data quality and governance. Leveraged **CI/CD** automation to streamline code integration, testing, and deployment processes, resulting in accelerated development cycles and enhanced data pipeline reliability using Jenkins and **Kubernetes** to orchestrate automated testing, version control, and deployment of data processing code, facilitating rapid and error-free data pipeline updates. Engaged in the whole **software development lifecycle** with requirements, solution design, development, **QA** implementation, and product support using **Scrum**, and **Jira** to maintain stories and other **Agile** methodologies. Proven collaborator, seamlessly interface with multidisciplinary teams, consistently delivering tailored data solutions aligned with strategic business goals. Comprehensive experience, leadership acumen, and dedication to driving impactful results make me an invaluable asset to data-focused endeavors.

# SKILLS

|  |  |
| --- | --- |
| **AWS Technologies** | EBS, EMR, EC2, S3, EKS, Airflow, Lambda, Elasticsearch, RDS, Kinesis, SQS,  DynamoDB, Redshift, ECS, Aurora PostgreSQL |
| **GCP Technologies** | Cloud Data flow, Cloud data store, Cloud Data Fusion, Cloud pub/sub, Big Query, Cloud Data Prep, cloud functions, Cloud Storage, Cloud Data Proc |
| **Hadoop/Big Data Technologies** | Hadoop, MapReduce, Apache Storm, Apache Sqoop, Hive, Apache Spark, Kafka, Cloudera Manager, YARN, Zookeeper, Apache Ambari, Apache NiFi,  HBase, Pig |
| **SQL Database** | Teradata, Oracle, MYSQL, Oracle 11g |
| **NO SQL Database** | DynamoDB, HBase, Cassandra, MongoDB |
| **Programming & Scripting** | Python, Scala, SQL, HiveQL, Shell Scripting, PySpark |
| **Monitoring and Reporting** | Tableau, Quick Sight, Looker, PowerBI |
| **Version Control** | GIT, GitHub, GitLab |
| **Others** | Jenkins, Snowflake |

EXPERIENCE

## Senior GCP Data Engineer KeyBank | Cleveland, Ohio

**April 2022 – till date**

KeyBank is banking company; team works for Auto loans department. And objective is to compute operations on customer data if customer is eligible for any loans depending on their activity and store aggregated data that is available for analysts.

## Responsibilities:

* Involved in **Data Ingestion** process and cluster handling in real-time streaming using **Cloud Data flow**.
* Involved in development of pipeline to convert unstructured customer data in **Cloud data store** to structured data by writing **PySpark** scripts in **Cloud Data Fusion** and storing data in **Cloud SQL**.
* Implemented pipeline between **Cloud pub/sub** to **Big Query** via **Cloud Data Flow** for streaming messages.
* Created Cloud Scheduler job that invokes Pub/Sub trigger for triggering **Cloud functions**.
* Implemented pipeline for storing aggregated customer data in **Cloud Storage**, using combination of **Big Query** and

## Cloud Data Prep.

* Worked on job scheduling related to pipeline in **Airflow** using **Python** with **CRON** syntax to schedule triggers for pipeline.
* Built data pipelines in **GCP composer** environment by utilizing Apache airflow operators like **Hadoop**, **Kubernetes**, **branching**, and **Python callable**.
* Monitored and visualized **Kafka** System performance by configuring Kafka Consumer and Producer metrics.
* Used **GIT** for version control, project management, and tracking issues and interaction with in-team or other team repositories.
* Created pipelines in **Snowflake** using **Snow Pipe** and visualized data in **Snow Sight** about changing trends in which service customers showing interest.
* Performed complex **aggregations** and created **Tableau** dashboards to find trends in analyzed data.
* Leveraged advanced **DAX** calculations and visualizations to provide deeper insights.
* Queried data stored in **Cloud SQL** using SQL Analytics and visualized it in **Google cloud data bricks**.
* Designed and optimized data models in **Google Big Query**, leveraging partitioning, clustering, and table **partitioning**

to enhance query performance.

* Implemented data processing pipelines that were executed on different execution engines such as **Apache Spark**, and **Google Cloud Dataflow** with use of **Apache Beam**.
* Implemented **RBAC** models in **GCP** to enforce consistent access controls across distributed and dynamic infrastructures.
* Utilized **Snowflake's** automatic query optimization capabilities, such as query acceleration and caching for better optimization.

## Tech Stack:

Cloud Data Flow, Cloud Storage, Big Query, Cloud Data Prep, Cloud Data Fusion, Cloud SQL, Cloud Pub/Sub, Airflow, Cloud Functions, Apache Beam, Apache Spark, Google Cloud Data Bricks, Snowflake, Snow pipe, Snow Sight, Tableau, GIT

## Senior AWS Data Engineer DaVita Inc | Santa Clara, CA

**Sep 2021 - Mar 2022**

Provides kidney dialysis services. Here, team works on Insurance plans and objective is to transform and mask sensitive information of patient data to protect their privacy and store it based on requirement of API team and analysts according to insurance type and associating that data with existing health records of patient.

## Responsibilities:

* Created pipeline to extract transform and load patient data from **AWS** data source **S3** to **Redshift** using combination of **MYSQL** database and **Glue** service.
* Worked on **Elastic search** engine and **Logstash** for ingesting logs using **Amazon Kinesis Data Firehouse** and **Kibana**

dashboards for searching logs to find out failures of jobs and gather other key metrics.

* Performed tasks in creation and computations on datasets of **S3** buckets using **Lambda Functions** and AWS **Glue**

using **PySpark**.

* Executed **PySpark** scripts in **Glue** to convert unstructured patient data from **Dynamo DB** to **MYSQL** database.
* Implemented server less architecture using **Lambda**, and scheduled glue jobs to run as optimization technique to improve performance.
* Developed data normalization scripts using **PySpark** for data ingested from **PostgreSQL** to **Redshift**.
* Created deployment pipeline by integrating services like Bitbucket **AWS Code Pipeline** and **AWS Elastic Beanstalk**.
* Notified alarms, and logs of glue jobs using **AWS CloudWatch** service.
* Developed **CDC** (Change Data Capture) for identifying and capturing data changes that are done in **Redshift**.
* Used version Control technologies like **GIT** for code commits and deployed project pipelines using **Jenkins** as **CI/CD**

tool.

* Utilized **columnar storage** and appropriate compression techniques to optimize storage efficiency in Redshift, thus resulting in reducing storage costs.
* Interactive data visualizations of patient data are visualized/created in **Quick Sight**.
* Modified **Cloud Formation Templates** to simultaneously create multiple AWS services (S3, Aurora PostgreSQL, Glue pipelines) to retrieve and save raw patient data.
* Extracted data from diverse sources including APIs, databases, and **flat files** to **S3**, cleaned, and transformed using

**Python** and **SQL** for accurate analysis.

* Implemented monitoring and alerting using AWS **CloudWatch**, ensuring proactive identification and resolution of performance issues.
* Leveraged AWS Glue for automatic schema discovery, job generation, and execution, reducing development time.
* Upgraded database from **Oracle 12c** to **Oracle 19c** by enabling features like auto-indexing for query optimization which enhanced performance of solution.
* Implemented data migration and replication strategies to move data between AWS and on-premises environments, and between different AWS regions and accounts.

## Tech Stack:

S3, Redshift, Aurora PostgreSQL, Oracle 12c, Oracle 19c, Lambda, Quick Sight, DynamoDB, Glue, AWS Cloud Shell, AWS Cloud Watch, EMR, EKS, Aurora PostgreSQL, UNIX, Apache Flume, Sqoop, Hive, Pig, HBase, Map Reduce, Pig, Mongo DB, Apache Spark, YARN, T-SQL, SparkSQL, Oozie.

## Data Engineer

**The Home Depot | TX**

**Responsibilities:**

**Jan 2019 - July 2021**

* Utilized **pandas** to represent data as **time series** and tabular format for transforming and retrieval of data.
* Involved in splitting, validating, and processing files using **Python scripts**.
* Involved in designing data schemas and improved consistency by reducing redundancy.
* Developed code in Python to store data in **MongoDB** for analyzing metrics. Complex MongoDB queries are written to validate and review data stored through API.
* Used String Interning for betterment of string operations processing time and space utilization by caching them.
* Utilized **NumPy** and Pandas in Python Scientific Programming for Data Manipulation.
* Executed and modified SQL queries and Functions, Triggers, and Cursors as per requirements.
* Performed tasks utilizing scientific computing stack (Matplotlib, pandas, NumPy, SciPy).
* Created pipeline to store data in AWS **MySQL** database from **S3** buckets where data is consumed by technical team or business analysts.
* Maintained data integrity and verified pipeline stability of **ETL** (Extract, Transform, Load) tasks by collaborating with different functional teams.
* Performed optimization and tuning using **Teradata** Views and by writing SQL Queries to improve performance of batch pipelines.
* Created tables in **PL/SQL** and completed setup in various environments DEV, SIT, UAT, and PROD
* Optimizations are performed on PL/SQL queries by utilizing **CURSOR FOR** loops for better control and performance and cursor attributes like Bulk Collect to fetch data in batches.
* Created **PL/SQL** packages to encapsulate and reuse business logic across multiple applications, improving code maintainability and reducing development time.
* Collaborated with database administrators to design and implement database schemas and structures that support

**PL/SQL** code and optimize database performance.

* Developed **Stored Procedures** to **test ETL Load** per batch and provided performance-optimized solutions to eliminate duplicate records.
* Executed **SparkSQL** Scripts to handle data from MySQL and streaming sources named **Kafka**.
* Implemented Python pandas for predictive analytics by performing data pre-processing and feature engineering.
* Worked on documentation including Source-to-Target Data Mapping Document, Set of Unit Test Cases, and Set of

**Data Migration** Documents.

* Automated pipeline workflow by using **Oozie** Scheduler system to orchestrate map-reduce jobs that extract data.
* Executed Shell scripting and Teradata utilities such as **Mload**, **Fast Load**, and **Query Scheduler** to develop process.
* Worked with new features in **Oracle Database 19c**, such as Real-Time Statistics, SQL Quarantine, and Automatic Indexing.
* Optimized **T-SQL** queries and procedures to enhance database performance, reducing query execution and improving overall system responsiveness.
* Implemented data validation checks and constraints using **T-SQL** to ensure data accuracy and integrity, contributing to high-quality and reliable data.

## Tech Stack:

Python, Pandas, MongoDB, Oracle, MySQL, Pl/SQL, NumPy, Jupyter Notebook, AWS RDS, S3, Teradata, Oozie, SparkSQL, Kafka, Query Schedular, ETL.

## Hadoop Engineer

**ZenQ | Hyderabad, India, IN**

**Responsibilities:**

**Feb 2016 - Dec 2018**

* Executed **pig** and **hive** scripts by utilizing **Oozie** Workflow engine and **Kafka** for real-time streaming/processing to navigate to datasets in **HDFS** file storage by loading Log file data directly into HDFS.
* Used **Apache Flume** for collecting and aggregating large amounts of unstructured data from **MongoDB** to **HBase**.
* Developed data pipelines using **Sqoop**, Hive, and Pig to ingest customer data into HDFS and perform Analytics.
* Used **YARN** to get control of volume of resource thus increasing efficiency.
* Designed cluster for handling huge amounts of structured, semi-structured, and unstructured data loaded in Hive and **Map Reduce** jobs by writing **Hive Base** Scripts.
* Continuously Monitored **Hadoop** Cluster jobs performance for maintaining data integrity using **Hive** Scripts.
* Implemented data ingestion pipelines using **Apache Sqoop** and **Apache NiFi** to extract and load data from various sources like **RDBMS**, and log files into Hadoop.
* Utilized execution engine Spark to procure optimized **HiveQL/ pig** scripts.
* Converted few existing hive scripts to Spark applications using RDDs for transforming data and loading into HDFS.
* Optimized Spark jobs to enhance performance and reduce processing time through techniques like data

**partitioning** and caching.

* Developed **Kafka** consumers and producers to capture and process real-time data streams, facilitating timely data updates and actionable insights.
* Implemented data partitioning and bucketing strategies in **HDFS** for optimized data storage and retrieval, minimizing query response times.
* Used **Index Keys** to sort documents and evaluated indexing strategies that support queries.
* Handled change data capture for processing incremental records between updated and existing data in RDMS tables by developing Sqoop scripts.
* Implemented data compression techniques such as **Parquet** and **ORC** file formats to reduce storage overhead and improve data access speed.
* Developed automated data validation framework using **SSIS** and **T-SQL** scripts, reducing manual effort and ensuring data integrity across data warehouse.
* Aggregated data to useful insights according to requirement by executing **PySpark** and **SparkSQL** transformation scripts.
* Utilized version control systems to manage **T-SQL** scripts, ensuring traceability and collaboration among team members in a multi-developer environment.
* Employed Map Reduce Programs to import bulk data into **HBase**.
* Implemented simple to complex **MapReduce** streaming jobs using Python.
* Scheduled automated tasks with Oozie for loading data into HDFS through Sqoop and pre-processing data with Pig and Hive.
* Utilized Spark RDD, **Scala**, and Python to convert Hive/SQL queries into Spark transformations.
* Involved in Directed Acyclic Graph (**DAG**) of operations and control flows, defined by Oozie for scheduling tasks to manage **Apache Hadoop** workloads.
* Worked with **Impala** to enable massively parallel Hive query processing.

## Tech Stack:

Apache Flume, Hadoop, Sqoop, Hive, Pig, HBase, Map Reduce, Pig, Mongo DB, Apache Spark, YARN, T-SQL, SparkSQL, Oozie, Impala, Scala.

## Data Analyst

**Syntax Soft-Tech Pvt.Ltd | Bengaluru, India, IN**

**Responsibilities:**

**Dec 2014 - Jan 2016**

* Used Python packages like **NumPy**, and **Matplotlib** to generate different planning reports graphically.
* Integrated **SQL**, **Excel**, and Analytics tools to analyze and assess marketing trends.
* Created reports and dashboards using **Tableau** to explain and communicate data insights, features, and performance to technical teams and stakeholders.
* Worked on creating and sending complete test status reports to business teams.
* Implemented and run **SQL** script to perform **DML** operations like updating and adding indexes.
* Performed reverse and forward engineering to evaluate and upgrade numerous data sets.
* Performed exploratory data analysis (**EDA**) on large datasets using **SQL**, and **Python** to uncover patterns, trends, and insights.
* Joined multiple database tables to aggregate data for reporting purposes, optimizing query performance using indexing techniques like **clustered index, non-clustered index**, and **complex index**.
* Designed and built interactive dashboards and reports to track key performance indicators (**KPIs**) and monitor business metrics in real-time.
* Developed automation scripts using Python and pandas to streamline data cleaning processes, saving time, and improving performance.
* Implemented **Oracle Database 11g** feature i.e., automatic memory management, to dynamically manage allocation of memory of database components to optimize performance.
* Optimized query performance for large-scale **SSAS** solutions by fine-tuning cube design, aggregations, partitions, and cache settings.
* Streamlined data validation procedures using **Python** scripts and automated anomaly detection, reducing data errors.
* Implemented parallel processing by creating well-designed partitions to improve query performance and resource utilization within **SSAS** solution.
* Implemented data validation checks to ensure data accuracy and integrity, reducing errors and mitigating potential risks associated with incorrect data.
* Cleaned and preprocessed messy data using **Python**, addressing missing values and outliers to ensure data integrity and accuracy for reporting purposes.

## Tech Stack:

Python, Tableau, Excel, MySQL, Oracle 11g, NumPy, matplotlib

# EDUCATION

## Bachelor Of Technology

Gayatri Vidya Parishad College of Engineering for Women, India

## Masters - computer science

University of Dayton, Dayton, OH

# CERTIFICATIONS

Google Cloud Certified Professional Data Engineer.