A hexagon with text and a logo

Description automatically generated

**Sandeep Rao**

**Professional Summary:**

* **Passionate Data Engineer with around 9+ years of experience on Data Engineering and Data Modeler with excellent knowledge on BI, Data warehouse, ETL, Cloud and Big - Data technologies and hands on experience in IT data analytics projects.**
* Utilized **PySpark** DataFrame and SQL API to perform complex data processing, transformations, aggregations, and filtering on large-scale datasets.
* Worked on developing ETL pipeline using **PySpark** to extract, transform, and load data from various source systems to target data warehouses and analytical platforms.
* Proficient in performance tuning techniques and strategies in **Snowflake**, a cloud-based data warehousing platform.
* Hands-on experience with Snowflake utilities like **SnowSQL, SnowPipe, Big Data** model techniques using Python.
* Experience working with Snowflake cloud data warehouse Snowflake Database, Schema and Table structures.
* Involved in loading the data from AWS to Snowflake using Python Snowflake Connector.
* Experience in Integrating Amazon Connect with other AWS services, such as Amazon S3, Amazon DynamoDB, or Amazon Redshift, as well as external systems, such as CRM or ticketing systems. This involves designing and implementing data flows, APIs, and connectors to ensure seamless data exchange and process automation.
* Strong experience in working on **AWS ECS, Fargate, S3, Glue, EMR, Redshift, Neptune (Graph Database),** **RDS.**
* In depth Knowledge of **AWS**cloud service like Compute, Network, Storage, and Identity & access management and hands-on experience on **AWS services like Athena, Lambda, Step Function and SQL.**
* Proficient in **Azure Cloud technologies**, specializing in **Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure SQL Database, and Azure Cosmos** **NoSQL DB**.
* Hands-on experience in migrating SQL databases to various Azure services including Azure Data Lake, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse, while managing database access and facilitating seamless migration from on-prem databases.
* Hands-on experience on Unified Data Analytics with gcpicks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delts Lake with Python, Delta Lake with Spark SQL.
* **In-depth knowledge of Hadoop architecture and its components like YARN, HDFS, Name Node, Data Node, Job Tracker, Application Master, Resource Manager, Task Tracker and Map Reduce programming paradigm.**
* Have extensively worked in developing ETL program for supporting Data Extraction, transformations and loading using Informatica.
* Worked on designing and implementing conversational flows and dialogues, mapping user intents and entities to enable accurate understanding and response generation.
* Have more than 3+ years of hands-on experience in migrating on-prem ETLs to Google Cloud Platform (**GCP**) using cloud service tools such as **BigQuery, Cloud Data Proc, Google Cloud Storage, Google Cloud Composer, Data Flow**, Pub/Sub and **Cloud SQL**.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow.
* Involved in the entire lifecycle of the project including Design, Development, and Deployment, Testing and Implementation and support.
* Involved in code reviews using GitHub pull requests, reducing bugs, improving code quality, and increasing knowledge sharing.
* Responsible for debugging the project monitored on JIRA (Agile).
* Built development environment with JIRA, Rally & GitHub.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Programming** | Python, PySpark, Scala, C, C++, Shell script, Perl script, SQL |
| **Databases** | Snowflake, Oracle, MySQL, SQL Server, MongoDB, Cassandra, DynamoDB, PostgreSQL, Teradata, Cosmos |
| **Big Data Technologies** | Hadoop, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Flume, Kafka, Yarn, Apache Spark, Sparklib |
| **Cloud Technologies** | AWS, Azure, Google Cloud Platform(GCP) |
| **Frameworks** | Django REST framework, Flask, web2py |
| **Healthcare Tools** | HL7, FHIR, ELR |
| **Tools** | PyCharm, Eclipse, Visual Studio, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Postman, Insomnia  **ETL Tool:** Informatica Power Centre, DBT  **Container: Docker** |
| **DevOps** | Git, GitHub, Jenkins, Docker, SVN |
| **Operating Systems** | Windows, Ubuntu Linux, MacOS |
| **Database Modelling** | Dimension Modelling, ER Modelling, Star Schema Modelling, Snowflake Modelling |
| **Orchestration Tool** | Apache Airflow, Astronomer, Oozie |
| **Visualization/ Reporting** | Tableau, Power BI & matplotlib |

**Education:**

**Master of Science in Computer Science, Campbellsville University 2022 - 2023**

**Bachelor of Technology in Information Technology, Kakatiya University 2011 - 2015**

**Professional Experience:**

|  |
| --- |
| **Client: USAA**  **Nov 2023 - Present**  **Role: Sr. AWS Data Engineer**  **Location: San Antonio, Texas** |
|  |

**Responsibilities**:

* Analyzing the data for multiple reports for automation to filter and validate the data before migrating to **Snowflake** from ERP database.
* Worked on retrieving data from different source systems like Netezza DB, SQL Server, DB2 database and also external sources like **Salesforce** to target database.
* Responsible for data warehouse migration from Netezza at on-prem to Redshift DB in **AWS.**
* On a day-to-day basis I involve in collaborating with stakeholders to gather requirements and designing solutions that leverage **Amazon Connect** capabilities to meet business needs. This involves analyzing customer service processes, call flows, and integration requirements to architect effective solutions.
* Worked in implementing data integration workflows using IBM DataStage by setting up the connections, designing ETL processes, and orchestrating workflows to move data between DataStage and Snowflake.
* Deployed the project onto the **Jenkins** server and organized the configurations and versions of the code using **Git** version control system.
* Developed **Python** scripts using **AWS Lambda** to gather data from various sources to the Snowflake data warehouse.
* Working on incremental data load from Snowflake sharable object/secured view to BOD Snowflake tables using DBT ELT tool.
* Worked on **Snowflake's** storage optimization features, such as micro-partitioning and automatic clustering to minimize data scan and improve query performance.
* Used **PySpark** to load data from AWS S3 to Snowflake by setting up the AWS credentials, installing required libraries and used snowflake-connector-python.
* Integrated Amazon Lex with other AWS services such as AWS **Lambda**, DynamoDB, and S3 to enable seamless backend processing, data storage, and content retrieval.
* Involved in building the data pipelines for data ingestion from SQL Server to S3 bucket.
* Implemented a CI/ CD pipeline with **Docker, Jenkins** and **GitHub** by virtualizing the servers using **Docker** for the Dev and Test environments by achieving needs through configuring automation using Containerization.
* Creating tables, views, user defined functions in **Snowflake** Cloud Data Warehouse.
* Worked with **Snowflake** and **PySpark** to create effective data models and to analyze data to provide valuable insights for every Sprint.
* Worked on ETL pipelines in and out of data warehouse using combination of Python and Snowflakes SnowSQL writing SQL queries against Snowflake.

**Environment**: PySpark, AWS, Amazon Connect, Git, Jenkins, Snowflake, Neptune, IBM DataStage tool, DBT tool, SQL, Python, Netezza DB AWS Lambda, S3, EC2, EMR, Redshift, AWS Glue.

|  |
| --- |
| **Client: CVS Health Feb 2023 – Nov 2023**  **Role: Sr. GCP Big Data Engineer**  **Location: Richardson, TX** |

**Responsibilities:**

* Designed and automated the data pipeline to transfer the data to Stakeholders which provides centralized KPIs and reports for CVS.
* Worked on building end to end ETL process by extracting the health care data from various sources, filtering required data and ingesting data into **GCP (BigQuery)**.
* Worked on running data flow jobs using Apache beam integrated in python for performing historical data loads into BigQuery tables.
* Built multiple programs using Python and Apache beam and executed it in cloud Dataflow to run Data validation between raw source file and BigQuery tables.
* Worked on Cleansing the raw data for Immunization, provenance, patient, and practitioner resources to load into **FHIR/HL7** **viewer**.
* Optimized and fine-tuned the scripts to achieve efficient processing in cloud data platform and FHIR/HL7 load.
* Implemented automated data loading workflows in Apache Airflow, orchestrating the extraction of high-volume data from Google Cloud Storage (GCS) into Google BigQuery.
* Utilized Python scripting within **DAG’s** to execute tasks and configured email notifications for DAG updates. Ensured workflow integrity by validating 'SUCCESS' status upon completion of all tasks within the DAG.
* Worked on creating Cloud Composer environment also configured, developed and scheduled multiple DAGs in Dev, UAT and Prod Env’s.
* Hands-on experience building, scheduling and running multiple data pipelines leveraging core Apache Airflow services within Google Cloud Composer, while also proficient in deploying workflows and monitoring them via Airflow UI and Cloud Monitoring.
* Proficient in utilizing Python libraries such **as Pandas, NumPy** and others to effectively analyze and transform diverse raw file formats including JSON, CSV, XML, and RRF.
* Proficient in reading data from Google BigQuery tables, publishing it to a Pub/Sub topic, and seamlessly processing loading both Structured (bound) and Unstructured (unbound) data from Google Pub/Sub to BigQuery using Cloud Dataflow with Python, including the creation of a Pub/Sub topic and its configuration within the codebase.
* Executed multiple MapReduce operations utilizing PySpark and Numpy, while also implementing **Jenkins** for seamless continuous integration.
* Ensuring uninterrupted data flow through our **24X7 data pipeline**, which has necessitated the implementation of a parallel data ingestion approach, utilizing **PySpark, S3 and Python**, across five distinct data sources.
* Utilized Spark Context, Spark SQL, Spark MLlib, DataFrames, Pair RDD, and YARN, while also leveraging Spark Streaming APIs to perform real-time transformations and actions, facilitating the construction of common data pipelines.
* Created Spark Jobs to extract data from **Hive tables** and process the same using Data proc. Also used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables.
* Monitoring **BigQuery**, **Data proc** & **Cloud Dataflow** jobs via Stack driver for all the environments.
* Written python program to extract the zip files from NIH website to update the latest data every month and apply required filters as well as joining using SQL query and finally write the output to Json file.
* Worked on centralizing logs, visualization and monitoring in google-cloud composer and Dataflow console.

**Environment**: GCP Services: Data proc, Dataflow, BigQuery, Pub/Sub, Google Cloud Composer, FHIR, AVRO, Git, Spark. Hive, Python (Pandas, NumPy, TensorFlow, Matplotlib), Airflow.

|  |
| --- |
| **Client: Kaodim**  **Sep 2019 - June 2022**  **Role: Sr. Data Engineer**  **Location: Hyderabad, India** |
|  |

**Responsibilities:**

* Created Azure Storage Planning, Migrated Blob Storage for document and media file, table storage for structured datasets, Queue storage for reliable messaging for workflow processing and file storage to share file data.
* Written Templates for Azure Infrastructure as code using **Terraform** to build staging and production environments.
* Worked with Terraform features such as Infrastructure as code, Execution plans, Resource Graphs, Change Automation.
* Deployed windows **Kubernetes (K8s) cluster** with **Azure Container Service** from Azure CLI, store in repository (ACR) and Utilized Kubernetes and Docker for the runtime environment of the CI/CD system to build, test and Octopus Deploy.
* Used Docker to virtualize deployment containers and built additional Docker Slave that are running in Kubernetes using custom-built Docker images and instances.
* Created Docker images and deployed Docker containers for Tomcat Servers, Prometheus, Grafana, and alert manager.
* Used Docker to virtualize deployment containers and built additional Docker Slave that are running in Kubernetes using custom-built Docker images and instances and created Docker images using a Docker file.
* Leveraged AWS **Lambda** for serverless computing, seamlessly triggering code execution through **S3** and **SNS** events. Utilized Amazon API Gateway to efficiently manage multiple concurrent API requests.
* Setup and build AWS infrastructure various resources like **VPC, EC2, S3, IAM, EBS, Security Group, Auto Scaling** and **RDS** in **Cloud Formation** using **Python Boto3** framework and JSON templates.
* Used Datadog Ops monitoring servers and integrated in **Cloud Watch, Chef** and **Ansible.**
* Implemented Large Scale Cloud Infrastructure using AWS resources - **IAM, Elastic IP, Elastic Storage, Auto Scaling**, **VPC, EC2, EBS, APIs ELB, AMI, Route 53, RDS, SES, SNS, SQS, EMR**, OpenShift, **Cloud Formation**, CDN, **KMS, S3, Cloud Front, Cloud Trail, Elastic Beanstalk**.
* Managed AWS resources in the cloud and maintained CI/CD pipeline.
* Use of Docker, **Kubernetes** to manage micro services for development of Continuous Integration and Continuous Delivery and Architecture/design and deployment of solutions using VMWare, and Amazon Web Services.
* Installed and configured an automated tool Puppet that included the installation and configuration of the Puppet master, agent nodes and an admin control workstation.
* Built and Deployed Java/J2EE to a web application server in an Agile continuous integration environment and automated the process of application from code to deployment.
* Involved in Beta Testing WebSphere 6.0. Experienced in creating and designing test plans for understanding the new & enhanced features of WebSphere Application Server.

**Environment:** Azure Cloud, Azure Data factory, Azure Data Lake, Azure DeVops portal, Agile Methodologies, Kubernetes, Python, SQL.

|  |
| --- |
| **Client: NerdWallet**  **Jan 2017 - Sep 2019**  **Role: Data Engineer**  **Location: Hyderabad, India** |
|  |

**Responsibilities:**

* Created DDL's for tables and executed them to create tables in the warehouse for **ETL data loads**.
* Designed, Implemented and maintained Database Schema, Entity Relationship diagrams, Data modelling, Tables, Stored procedures, **Functions** and **Triggers, Constraints, clustered** and **non-clustered indexes**, partitioning tables, Schemas, Functions, Views, Rules, Defaults, and complex SQL statement for business requirements and enhancing performance.
* Designed and Implemented Big Data Analytics architecture, transferring data from Oracle.
* Implemented logical and physical relational database and maintained Database Objects in the data model using Erwin.
* Worked on AWS cloud services i.e., **EC2, S3, EMR**. Defined and deployed monitoring, metrics, and logging systems on AWS.
* Managed security groups on AWS, focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along CI/CD with AWS Lambda and AWS code pipeline.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloud watch.
* Used AWS Glue for the data transformation, validate and data cleansing.
* Used python Boto 3 to configure the services AWS glue, EC2 & S3.
* Developed and implemented data normalization jobs for newly ingested data into Redshift, ensuring optimal data organization and efficiency. Additionally, led the migration of on-prem database structures to a confidential Redshift data warehouse, streamlining data storage and accessibility.
* Created and managed a Docker deployment pipeline for custom application images in the AWS cloud using Jenkins.
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Designed packages to bring data from existing OLTP databases to new data warehouse using various transformations.
* Developed ETL framework using **Spark and Hive** (including daily runs, error handling, and logging) to useful data.
* Optimizing existing algorithms in Hadoop using Spark Context, **Spark-SQL, Data Frames and Pair RDD's.**

**Environment:** Erwin 9.6, Oracle, SQL, Docker, PL/SQL, Redshift, SQL Server, Hadoop, Spark, Python, GitHub, AWS.

|  |
| --- |
| **Client: Citizens Bank**  **Oct 2015 - Dec 2016**  **Role: Data Analyst**  **Location: Chennai, India** |
|  |

**Responsibilities:**

* Worked with users to identify the most appropriate source of record required to define the asset data for financing.
* Created Teradata SQL scripts using OLAP functions like RANK () to improve the query performance while pulling the data from large tables.
* Imported the customer data into Python using Panda’s libraries and performed various data analysis - found patterns in data which helped in key decisions for the company.
* Extracted data from the database using SAS/Access, SAS SQL procedures and created SAS data sets.
* Design and deploy rich Graphic visualizations with Drill Down and Drop-down menu option and Parameterized using Power BI
* Utilized Python and RESTful Web Services to retrieve, manipulate, and update data from various sources, including XML content from suppliers. Worked on data manipulation and analysis with Python.
* Developed **Python scripts** to **automate** data sampling process. Ensured the data integrity by checking for completeness, duplication, accuracy, and consistency.
* Designed synchronized data models that maintained consistency between OLAP and OLTP systems, guaranteeing accurate analysis while preserving transactional integrity.
* Designed and developed weekly, monthly reports by using **MS Excel Techniques** (Charts, Graphs, Pivot tables) and PowerPoint presentations.
* Strong Excel skills, including pivots, VLOOKUP, conditional formatting, large record sets. Including data manipulation and cleaning.
* Good experience in **Agile Methodologies**, Scrum stories, and sprints experience in a Python-based environment, along with data analytics and Excel data extracts.
* Experience with continuous integration and automation using **Jenkins**.
* Familiarity with **GitHub** for project management and versioning.

**Environment:** REST API, SAS/Access, SAS SQL, MS Excel, Python Pandas, Power BI, GitHub, Tableau, RDBMS, Python, SQL.