**SAI **

**Email:** [**saik242023@gmail.com**](mailto:saik242023@gmail.com) **Data Engineer**

**954-554-6286 https://www.linkedin.com/in/sai-arun-696b7593/**

* Around 10 years of professional experience. Well versed in installation, configuration, supporting & managing of **Bigdata**, **Data Warehouse** and CRM applications.
* Extensive experience in cloud services, including S3, Lambda, Airflow (MWAA), Secrets Manager, and CloudWatch.
* Proficient in Snowflake and SQL for proficient data ingestion, transformation, and loading, coupled with a robust comprehension of relational database principles.
* Proficient in Python and Unix, leveraging hands-on experience for coding and scripting purposes.
* Implemented and maintained Apache Druid clusters for real-time analytics and data exploration.
* Optimized ingestion processes to handle large volumes of streaming data efficiently. Configured and tuned Apache Druid segments for optimal query performance.
* Developed custom ingestion tasks and data pipelines using Apache Druid APIs.
* Skilled in automating build, test, and deployment processes using Bash scripting and employing build automation tools like Maven or sbt for Scala projects, as well as pip or virtual env for Python projects.
* Proficient in integrating automated testing frameworks like JUnit, ScalaTest, pytest, or unit test into CI/CD pipelines to ensure optimal code quality and reliability.
* Stay current with the latest cloud technologies, patterns, and methodologies; share knowledge by clearly articulating results and ideas to key stakeholders.
* Worked closely with cross-functional teams, including data engineers, analysts, and business stakeholders, to develop and deploy integration solutions that meet business objectives.
* Designed and implemented data integration workflows using SnapLogic’s drag-and-drop interface. Utilized pre-built connectors (Snaps) to integrate applications, databases, SaaS platforms, and APIs.
* Performed complex data transformations and cleansing operations within SnapLogic pipelines.
* Collaborating closely with data scientists, **data analysts**, **software engineers** to build and maintain the infrastructure required for optimal **extraction**, **transformation**, **loading** of data from diverse sources.
* Collaborate with product owners and business subject matter experts to analyze customer requirements and deliver supportable, sustainable engineered solutions.
* Proven track record of partnering with cross-functional teams, including product owners and business SMEs, to analyze customer requirements and design tailored, sustainable solutions that meet business needs.
* **Designing**, **developing**, and owning data pipelines and models that will power the internal analytics used for product and business teams.
* Implemented disaster recovery and high availability strategies for CI/CD pipelines on GCP, ensuring our systems are resilient to failures and capable of recovering quickly in case of downtime.
* OnPrem to Cloud Migrations, Application deployment, troubleshooting and support.
* Hands-on experience with **Amazon EC2, Amazon S3, AWS Glue, AWS Lambda, AWS Redshift, Amazon RDS, VPC, IAM**, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SQS, **Lambda, EMR** and other AWS services
* Experience in developing scalable solutions using **NoSQL** databases including **HBASE**, CASSANDRA and MongoDB.
* Extensive expertise in Hadoop Architecture, encompassing key components such as HDFS, MapReduce, Hadoop GEN2 Federation, High Availability, and YARN architecture. Profound understanding of workload management, scalability, and distributed platform architectures.
* Experience in setting up monitoring and logging solutions within CI/CD pipelines, using tools like Stackdriver Monitoring and Logging in Google Cloud Platform.
* Successfully implemented and managed **Control-M**, a workload automation tool, to optimize job scheduling and streamline batch processing workflows.
* Designed and developed Grafana dashboards to visualize metrics and key performance indicators (KPIs) for monitoring purposes.
* Integrated Grafana with various data sources including Prometheus, InfluxDB, and Elasticsearch.
* Customized Grafana dashboards using templating, annotations, and plugins to meet specific business requirements.
* Automated Grafana dashboard deployment and configuration using scripting tools like Ansible or Terraform.
* Provided training and support to teams for effectively utilizing Grafana for monitoring and analytics.
* Hands on experience working with **AWS using Elastic Map Reduce (EMR)** and **EC2** for data processing.
* Extensive experience in working on **IDEs** like **Eclipse, Net Beans, PyCharm** and **Edit** Plus.
* Experience building **Spark applications** and Import data from different sources like **HDFS/HBase** into **Spark RDD**.
* Experience in using **Apache Sqoop** to import and export data to and from (different sources) **HDFS** and **Hive**.
* Utilized Amazon Athena for serverless data analysis, running complex SQL queries on structured and semi-structured data in Amazon S3. Implemented data partitioning and efficient formats like Parquet to optimize query performance.
* Extensive background in Data warehouse - analysis, design, development, implementation and troubleshooting in ETL Tools - SSIS, **Snowflake**, Data ware housing Methodologies and Data Modelling
* Identifying, analyzing, and resolving relational / distributed database operations and application deployment issues
* Experience in writing **UDF’s** (User Defined Functions) to enhance the functionalities of **Spark**, **Hive** and **Pig**.
* Very Good understanding and Working Knowledge of Object-Oriented Programming (**OOPS**), Python and Scala.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Bigdata Technologies** | Hadoop, MapReduce, Pig, Hive, YARN, Kafka, Sqoop, Impala, Oozie, Zookeeper, Spark, Drill, Ambari, Cassandra, Avro, and Parquet. |
| **Programming Languages** | Python, Scala, Java, SQL, PL/SQL. |
| **Cloud Services** | Amazon EC2, AWS, AWS S3, AWS Lambda, AWS Glue, AWS EMR, IAM, Cloud watch, Redshift, Snowflake |
| **Databases/RDBMS** | Oracle 11g/10g, DB2, MS-SQL Server, MySQL |
| **Scripting/Web Languages** | JavaScript, HTML5, CSS3, XML, J query, Angular, Terraform |
| **Operating Systems** | Windows, UNIX, Linux, Mac OS. |
| **Software Life Cycles** | SDLC, Waterfall and Agile models |
| **Webservices** | SOAP, REST Web Services |
| **Utilities/Tools** | Eclipse, Tomcat, ANT, Maven, Automation, PyCharm |
| **Orchestration** | Cron, Oozie, Apache Airflow |
| **DevOps Tools** | git, Azure DevOps, CI/CD, TFS, K8 |
| **Reporting Tools** | Tableau, PowerBI |
| **App/Web servers** | WebLogic, Tomcat |

**EDUCATION:** Bachelor’s Degree in Engineering (CSE), JNTUH (Ganapathy Engineering College)

**Client: Marsh Digital, Phoenix, AZ July 2023 to Till Now**

**Data Engineer**

**Responsibilities:** -

* Collaborated with cross-functional teams within the data science and analytics team to design, develop, and execute solutions for deriving business insights and solving clients' operational and strategic problems.
* Designed and developed an automated framework to create and automate the development process in the **Data Lake**.
* Developed and managed API integrations to connect internal services with third-party applications, enhancing system interoperability.
* Leveraged SnapLogic’s elastic scaling capabilities to manage varying data volumes and processing loads.
* Led the migration of data from RDS PostgreSQL to Snowflake using AWS services such as Lambda, Glue ETL, and EventBridge.
* Designed and implemented data pipelines using Glue ETL for extracting, transforming, and loading data into Snowflake.
* Orchestrated the data migration process using EventBridge to schedule and trigger Lambda functions and Glue ETL jobs.
* Created on-demand tables on **S3** files using Lambda Functions and AWS Glue with **Python** and **PySpark**.
* Migrated an existing on-premises application to AWS, utilizing services like **EC2** and **S3** for small dataset processing and storage, while also maintaining a **Hadoop cluster** on **AWS EMR**.
* Built scalable ETL pipelines with Google Cloud Dataflow, processing large datasets from
* diverse sources into Google Cloud Storage and BigQuery.
* Designed and implemented data warehousing solutions using BigQuery, reducing query
* processing time and improving data accessibility for business analysts.
* Developed and maintained automated test scripts in Java and Python for web applications using Cypress framework.
* Utilized Java and Python expertise to develop automated tests for Apache Spark-based data processing applications.
* Conducted performance tuning and optimization of Spark jobs, resulting in improved processing efficiency and reduced resource consumption.
* Managed and optimized databases, including Postgres (Aurora RDS), Snowflake, and Oracle.
* Designed and implemented data models to ensure efficient data management and retrieval.
* Developed complex SQL queries for data extraction, transformation, and reporting.
* Managed data transformation workflows with Google Cloud Dataproc, integrating data
* from various systems and providing clean, consistent datasets for analytics.
* Utilized AWS Data Pipeline to configure data loads from **S3** into **Redshift**.
* Worked extensively with JSON and XML data formats for data exchange and transformation processes within SnapLogic.
* Used Amazon Athena to query large datasets stored in Amazon S3, providing interactive data analysis for key projects. Applied SQL to extract insights and generate reports for business decision-making.
* Designed, implemented, and tested major subsystems on the **AWS cloud platform** and **core service** offerings.
* Automated DBT runs using tools like Airflow, AWS Step Functions, or DBT Cloud to ensure consistent data pipelines.
* Scheduled and monitored DBT runs to maintain optimal data processing workflows.
* Developed complex DBT models for transforming raw data into refined, analytics-ready datasets.
* Created and managed DBT projects, including configurations, model directories, and dependency trees.
* Developed and implemented data integration processes using Azure Data Factory (ADF) to efficiently extract, transform, and load (ETL) data from diverse sources into centralized data warehouses or data lakes.
* Successfully integrated multiple data sources into a centralized data lake, enabling advanced analytics and real-time reporting capabilities
* Ensured data quality and integrity by implementing robust data cleaning, conversion, and validation techniques as part of the ETL process.
* Leveraged **AWS Glue** catalog with crawler to extract data from **S3** and perform **SQL query** operations.
* Developed a **Spark-based ingestion** framework for ingesting data into **HDFS**, creating tables in **Hive**, and executing complex computations and parallel data processing.
* Implemented **Spark** using **Python**, utilizing **DataFrames** and **Spark SQL API** for faster data processing.
* Developed **Spark** applications using **PySpark** and **SparkSQL** for data **extraction**, **transformation**, and **aggregation** from multiple file formats, aiming to analyze and transform the data to uncover insights into customer usage patterns.
* Integrated Scala-based data analysis with data visualization tools to create interactive dashboards and visualizations.
* Developed complex transactional queries and stored procedures to support real-time processing in an OLTP environment.
* Ensured ACID compliance through robust data validation and consistency checks.
* Optimized OLTP systems to handle high transaction volumes while maintaining low latency and high availability.
* Utilized Amazon Aurora's read replicas to scale read-heavy workloads, improving query response times.
* Developed a microservices-based backend system in Scala with persistent storage on Amazon Aurora.
* Designed and maintained scalable data storage solutions leveraging Azure services such as Azure SQL Database, Azure Data Lake Storage, Azure Cosmos DB, and Azure Blob Storage.
* Optimized data storage systems to meet organizational requirements for performance, reliability, and cost-effectiveness.
* Developed **Terraform** scripts to automate **AWS** services such as **Lambda**, **Glue**, **EventBridge**, **ELB**, **CloudFront** distribution, **RDS**, **EC2**, **database security groups**, and **S3 buckets**.
* Conducted data migration from **On-Premises** systems into a **Snowflake** Cloud data **warehouse**, automating workloads in the process.
* Implemented Informatica Data Quality (IDQ) tools to profile, cleanse, and standardize data, ensuring high data quality standards were met post-migration. Conducted thorough data validation and reconciliation to verify the accuracy and completeness of migrated data.
* Designed and implemented snowflake schemas with appropriate dimensions and facts to support business intelligence and analytics.
* Profiled **structured**, **unstructured**, and **semi-structured** data across various sources to identify patterns and implemented data quality metrics using necessary queries or Python scripts based on the source.
* Installed and **configured** Apache **Airflow** for **S3 buckets** and **Snowflake** data **warehouse**, creating **DAGs** to run the **Airflow**.
* Implemented complex analytical queries and multi-dimensional analysis in OLAP environments.
* Created data cubes and star/snowflake schemas to facilitate efficient data analysis and reporting.
* Optimized OLAP systems for high-performance analytics and rapid query response times.
* Created **DAGs** to utilize **Email Operator**, **Bash Operator**, and **Spark** Livy Operator for execution in an **EC2 instance**.
* Deployed code to **EMR** via **CI/CD** using Jenkins.
* Experienced with **APIs** and **REST** services in collecting data and publishing to downstream applications.

Environment: **AWS Lambda**, AWS **Glue**, **Airflow**, **Spark**, Kinesis, Hive, **Python**, **SQL** **Scripting**, **PySpark**, **Microservices**, CI/CD, Kubernetes, Oracle, **Snowflake**, **MySQL**.

**UK Power Networks Dec 2021 to June 2023**

**Data Engineer**

**Responsibilities:**

* Worked with cross functional teams within the data science and analytics team to design, develop, and execute solutions to derive business insights and solve clients' operational and strategic problems.
* Responsible for designing and developing an **automated framework** which creates and automates the development process in **Data Lake**.
* Led the migration of databases from DB2 Z/OS to AWS, ensuring data integrity and minimal downtime. Utilized Informatica PowerCenter and PowerExchange for efficient data extraction, transformation, and loading (ETL) processes.
* Created on-demand tables on **S3** files using **Lambda Functions** and **AWS Glue** using Python and PySpark.
* Tuned ETL processes and database queries to improve performance and reduce execution time.
* Monitored and optimized the performance of data migration jobs, ensuring timely completion and adherence to project timelines.
* Led ETL processes using SnapLogic, integrating data across various sources and targets.
* Utilized SnapLogic for building data pipelines, ensuring seamless data flow and transformation.
* Worked with Informatica PowerCenter and IICS for additional data integration tasks.
* Managed and integrated APIs using SnapLogic to ensure secure and efficient data exchange.
* Handled both real-time data integration and batch processing tasks to meet diverse business needs.
* Utilized big data technologies including Azure Databricks and Apache Spark to process and analyze large volumes of healthcare data efficiently.
* Developed data processing workflows and pipelines to support advanced analytics, machine learning, and other data-driven applications in healthcare.
* Migrated an existing on-premises application to AWS. Used AWS services like **EC2** and S3 for small data sets processing and storage, experienced in maintaining the Hadoop cluster on AWS EMR.
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift.
* Designing, implementing, and testing major subsystems AWS cloud platform and core service offerings.
* Used AWS glue catalog with crawler to get the data from S3 and perform sql query operations
* Developed spark-based ingestion framework for ingesting data into **HDFS**, creating tables in Hive and executing complex computations and parallel data processing.
* Implemented Spark using Python and utilizing Data frames and Spark SQL API for faster data processing.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage pattern
* Developed Terraform scripts to automate AWS services such as Lambda, **Glue**, Event bridge, ELB, **CloudFront** distribution, **RDS**, EC2, database security groups, and S3 bucket.
* Data migration from On-Premises systems into a Snowflake Cloud data warehouse and automate workloads.
* Upgrade Java application framework to Spring boot, micro services and deploy.
* Java development for different components. Developed applications using **Core Java**, Multi-Threading, **Spring Boot, Beans, Hibernate, Transaction**, Maven to name a few.
* Profile structured, unstructured, and semi-structured data across various sources to identify **patterns in data and Implement data quality metrics**using necessary queries or **python** scripts based on source.
* Install and configure**Apache Airflow** for S3 bucket and Snowflake data warehouse and created **DAGs**to run the Airflow.
* Created DAG to use the**Email Operator, Bash Operator, and spark Livy operator** to execute and in **EC2**instance.
* Deploy the code to**EMR** via**CI/CD using Jenkins.**
* Experienced with API & Rest services in collecting the data and publishing to downstream applications.

**Environment:** Aws Lambda, AWS Glue, Airflow, Spark, Kinesis, Hive, Python, SQL Scripting, Pyspark, Micro services, CI/CD, K8, Oracle, snowflake, MySQL.

**UnitedHealthGroup, Hyd, India     Dec 2018 - Dec 2021**

**Sr AWS Data Engineer**

**Responsibilities:**

* Interacting with **dev team**, **business team**, **data analyst**, and **data architects**.
* Work with stakeholders to assist in the data-related technical issues and support their data infrastructure needs.
* Led a project to migrate legacy ETL processes to SnapLogic, resulting in a X% reduction in processing time and operational costs.
* Designed a complex integration workflow for real-time data synchronization between ERP and CRM systems, enhancing sales and operations alignment.
* Utilized Informatica PowerCenter for designing and managing data extraction, transformation, and loading (ETL) processes.
* Employed Informatica PowerExchange for seamless data integration and real-time data capture from DB2 Z/OS to target systems.
* Developed and maintained web applications using Java, Angular, Node.js, HTML, CSS, and JavaScript.
* Implemented RESTful APIs using Spring Boot, ensuring robust and scalable backend services.
* Automate manual **ingest processes** and **optimize data** delivery subject to service level agreements, work with infrastructure on re-design for greater scalability.
* Developed **data pipeline** using **SQOOP**, **HQL**, **Spark**, **AWS Glue** and other Hadoop technologies to **ingest data** from **RDBMS** to **Hadoop**.
* Automate cloud infrastructure using Terraform scripts and develop python scripts for data pipelines.
* Developed custom **ETL solutions**, **batch processing** and **real-time data** ingestion pipeline to move data in and out of **Hadoop** using **Pyspark** and **shell scripting**.
* Handle huge datasets with **Partitions**, **Spark** in **Memory features**, and **Broadcasts** in **Spark** with **Scala** and **Python**.
* Developed **Python** and **Pyspark scripts** to transfer data from on-premises storage to **AWS S3**.
* Migrated data from **on-premises** to **AWS storage buckets**.
* Develop **oozie workflows** for daily incremental loads to load data from **RDBMS** to **Hive tables**.
* Designed & developed **Spark workflows** to extract data from **AWS S3** bucket and apply transformations using Scala
* Maintaining 99.9% data pipeline uptime while ingesting **streaming** and **transactional** data from main data sources using **Redshift**.
* Used **AWS EMR** to execute **spark/hive** applications to read and write data from/to **S3**.
* Built **dashboards** using **Tableau** to allow internal and external teams to **visualize** and extract insights from big data platforms.

**Environment:** Hadoop, HDFS, Hive, YARN, UNIX shell scripting, PySpark, Python, Data Pipeline, Sqoop, Big data, MS SQL Server, SQL, ETL, AWS Glue.

**Currency Holdings limited, Hyd, India Sep 2016 - Dec 2018**

**Data Engineer**

**Responsibilities:**

* Developed multiple **MapReduce** jobs in Python for data cleaning and preprocessing and assisted with data capacity planning and node forecasting.
* Involved in design and ongoing operation of several **Hadoop** **clusters** and Configured and deployed **Hive** Meta store using **MySQL** and thrift server
* Implemented and operated on-premises Hadoop clusters from the hardware to the application layer including compute and storage.
* Uploaded and processed more than 30 terabytes of data from various structured and unstructured sources into **HDFS** (**AWS** **cloud**) using **Sqoop** and **Flume**.
* Designed custom deployment and configuration automation systems to allow for hands-off management of clusters via Cobbler, FUNC, and Puppet.
* Prepared complete description documentation as per the Knowledge Transferred about the Phase-II Talend Job Design and goal and prepared documentation about the Support and Maintenance work to be followed in **Talend**.
* Deployed the company's first Hadoop cluster running **Cloudera's** **CDH2** to a 44-node cluster storing 160TB and connecting via 1 GB Ethernet.
* Debug and solve the major issues with **Cloudera** manager by interacting with the **Cloudera** **team**.
* Modified reports and **Talend** **ETL** jobs based on the feedback from QA testers and Users in development and staging environments.
* Upgrade Java application framework to Spring boot, micro services and deploy.
* Java development for different components. Developed applications using **Core Java**, Multi-Threading, **Spring Boot, Beans, Hibernate, Transaction**, Maven to name a few.
* Involved in requirements specification, Analysis and Design of the application.
* Implemented **REST** Microservices using **Spring MVC, Spring IOC**, **and Spring Boot**. Generated Metrics with method level granularity and Persistence using **Spring** AOP and Spring Actuator.
* Implemented distributed **Micro** **Service** Monitoring. Integrated **Swagger** **UI** and wrote integration test along with **REST** document.
* Used Agile Methodology in developing applications, with weekly sprints, daily stand-up SCRUM meetings.
* Worked on **Spring Web Flow on Spring MVC** to build an application from scratch using **Spring XML configurations, annotations.** Wrote controllers to handle incoming requests and send output to the UI.
* Developed **REST** based **Microservices** using **Spring Boot.**
* Designed and developed **REST APIs** using **Spring Boot and Microservices** for exposing the services utilized for risk assessment by various other applications.
* Implemented a proof of concept deploying this product in**AWS S3 bucket**.
* Deployed and managed applications on AWS, leveraging cloud services for scalability and reliability.
* Utilized AWS services such as EC2, S3, and RDS for application hosting and data storage.
* Created and maintained data models, ensuring data integrity and consistency.
* Developed complex SQL scripts for data analysis, reporting, and ETL tasks.
* Developed **pyspark**scripts, **UDF**'s using both **data frames/SQL and RDD** in **Spark**for data aggregation, queries and writing back into S3 bucket.
* Experience in**data cleansing and data mining**.
* Design and Implement ETL for data load from heterogeneous source to SQL server and Oracle as target databases and for Fact and slowly changing dimensions SCD-type1 and SCD-type2.
* Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation and used Spark engine,**Spark SQL** for**data analysis**and provided to the data scientists for further analysis.
* Expertise in Transact **-SQL** (DDL, DML) and in Design and Normalization
* Designed and Developed Spark workflows using Scala for data pull from AWS S3 bucket and Snowflake applying transformations on it.
* Created scripts to read**CSV, JSON, and parquet files**from S3 buckets in **Python**and load into **AWS S3.**

**Environment:** Core Java, Multi-Threading, Spring Boot, Beans, Hibernate, Spring MVC, Spring IOC, and REST APIs, AWS S3, SCD-type1 and SCD-type2, SQL (DDL, DML).

**Aayathi Software Private Limited, Hyd June 2014 - Aug 2016**

**Java Developer**

**Responsibilities:**

* Experience in installing configuring and using Hadoop ecosystem components.
* Experience in administration installing upgrading and managing CDH3 **Pig Hive HBase**
* Importing and exporting data into HDFS and Hive using **Sqoop**.
* Experienced in defining job flows.
* Knowledge in performance troubleshooting and tuning Hadoop clusters.
* Experienced in managing and reviewing Hadoop log files.
* Developed full-stack applications using Java, Angular, Node.js, and Spring Boot, enhancing user experience and backend reliability.
* Implemented RESTful APIs, providing robust and scalable solutions for data access.
* Participated in development/implementation of **Cloudera** Hadoop environment.
* Load and transform large sets of structured semis structured and unstructured data.
* Responsible to manage data coming from different sources.
* Analyzed the SQL scripts and designed it by using **PySpark** SQL for faster performance.
* Worked on reading and writing data formats using **PySpark**.
* Manage configuration of Web App and Deploy to **AWS cloud server.**
* Experienced in automating, configuring and deploying instances on **AWS**, cloud environments and in Data centers
* Designed and developed the UI of the website using Python, HTML, **XHTML, AJAX, and CSS**.
* Developed entire frontend and backend modules using **Python on Django** Web Framework.
* Implement code in **Python** to retrieve and manipulate data.
* Got good experience with **NOSQ**L database.
* Supported Map Reduce Programs those are running on the cluster.
* Involved in loading data from UNIX file system to **HDFS**.
* Involved in creating Hive tables loading with data and writing hive queries which will run internally in map reduce way.
* Created **HBase** tables to store variable data formats of PII data coming from different portfolios.
* Load and transform large sets of structured semis structured and unstructured data
* Exported the analyzed data to the relational databases using **Sqoop** for visualization and to generate reports for the BI team.
* Supported in setting up QA environment and updating configurations for implementing scripts with Pig and **Sqoop**.

**Environment:** Python 3.x/2.7, Java, Spring MVC, Spring Boot, Rest API, Django, Pyspark, AWS, HTML, CSS, Oracle, SQL, Red Hat Linux