ARAVINDH SAI GIKURU

Mckinney, Texas • +1 475-287-5733 • aravindhsaigikuru@gmail.com

Experienced Data Engineer with 5+ Years in Building Scalable Data Solutions, Cloud Migrations, and Big Data Ecosystems Using AWS, Azure, GCP, Apache, and Hadoop Technologies

Professional Summary:

- Practical experience with **Python** and **Apache Airflow** to create, schedule, and monitor workflows. Experience in migrating on-premises to **Azure** using **Azure** Site Recovery and **Azure** backups.
- Experience using various **Hadoop Distributions** (**Cloudera**, **MapR**, **Hortonworks**, **Azure**) to fully implement and leverage new **Hadoop features**.
- Working knowledge of **HDFS, Kafka, MapReduce, Spark, Pig, Hive, Sqoop, HBase, Flume,** and **Apache ZooKeeper** as tools for designing and deploying end-to-end big data ecosystems. Developed batch processing solutions using **Data Factory** and **Azure Databricks**.
- Hands-on experience with **Amazon EC2**, **Amazon S3**, **Amazon RDS**, **VPC**, **IAM**, **Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda**, **EMR** and other services of the **AWS** family.
- Practical knowledge in setting up and designing large-scale data lakes, pipelines, and effective **ETL** (Extract/Transform/Load) procedures to collect, organize, and standardize data that can be used to convert a current on-premises application to use **Azure** cloud databases and storage. Worked with **CSV, Avro, Parquet** data formats to load into Data frames and do the analysis.
- Experience in automating day-to-day activities by using **Windows PowerShell.**
- Experienced with Dimensional modelling, Data migration, Data cleansing, Data profiling, and **ETL** Processes features for data warehouses.
- Experience in data integration and building data pipelines in **Airflow** using **Python scripting**. Strong experience in **Agile** (**SCRUM**) and Waterfall **SDLC**.
- Proficient with container systems like **Docker** and container orchestration like **EC2** Container Service, **Kubernetes**, worked with **Terraform**.
- Hands-on experience interacting with **REST APIs** developed using the microservices architecture for retrieving data from different sources.
- Experienced in fact dimensional modeling (Star schema, **Snowflake schema**), transactional modeling and **SCD** (**slowly changing dimension**)
- Hands-on experience in Implementing, Building, and Deployment of **CI/CD** pipelines, managing projects often including tracking multiple deployments across multiple pipeline stages (**Dev**, Test/QA staging, and Production).
- Experience in **Cisco Cloud Center** to more securely deploy and manage applications in multiple data centers, private cloud, and public cloud environments.
- Designed and implemented a scalable data architecture on **AWS** using **Kubernetes**, **Terraform**, and **Snowflake**, enabling seamless data integration and processing across multiple data sources.

WORK EXPERIENCE

Capital One • Plano, Texas, USA • 12/2023 - Present

Capital One is a financial services company offering banking, credit cards, and investment solutions with a focus on technology-driven innovation. Developed and optimized data pipelines and infrastructure solutions, significantly improving data processing efficiency and automation across various systems.

Azure Data Engineer

- Responsible for loading the data from BDW Oracle database, Teradata into HDFS using Sqoop. Implemented AJAX, JSON, and JavaScript to create interactive web screens.
- Used Pig as ETL tool to do Transformations with joins and pre-aggregations before storing the data onto HDFS and assisted Manager by providing automation strategies, Selenium/Cucumber Automation and Jira reports.
- Created and maintained technical documentation for launching Hadoop Clusters and for executing Hive queries and Pig Scripts.
- Effectively scheduled and managed jobs on Azure virtual machines using Control-M, optimizing resource allocation and ensuring reliable execution.
- Imported real time weblogs using Kafka as a messaging system and ingested the data to Spark Streaming and did data quality checks using Spark Streaming and arranged bad and passable flags on the data. Developed business logic using Kafka & Spark Streaming and implemented business transformations. Supported Continuous storage in ADLS and configured Snapshots and wrote entities in Spark along with named queries to interact with database.
- Created several Databricks (Spark) jobs with PySpark to perform several tables to table operations.
- Developed custom aggregate functions using Spark SQL and performed interactive querying.
- Created Data tables utilizing PyQt to display customer and policy information, enabling add, delete, and update operations on customer records.
- Architected Python scripts for automated data extraction and loading from web server output files, reducing manual data entry, and processing time by 75%.
- Worked with AWS Terraform templates in maintaining the infrastructure as code.
- Involved in various phases of Software Development Lifecycle (SDLC) of the application, like gathering requirements, design, development, deployment, and analysis of the application.
- Performed data analysis and data profiling using complex SQL queries on various source systems including Oracle 10g/11g and SQL Server 2012.
- Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. Also Worked with Cosmos DB (SQL API and Mongo API) .
- Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
- Controlling and granting database access and migrating on-premise databases to Azure Data Lake Store using Azure Data Factory.
- Designed and deployed a Kubernetes-based containerized infrastructure for data processing and analytics, leading to a 20% increase in data processing capacity.
- Skilled in monitoring servers using Nagios, Cloud watch and using ELK Stack - Elasticsearch and Kibana.
- Designed and implemented data warehousing solutions using Azure Synapse Analytics, including building data models, ETL processes, and analytical queries.
- Developed and maintained data models and schemas within Snowflake, including the creation of tables, views, and materialized views to support business reporting and analytics requirements.
- Implemented Continuous Integration Continuous Delivery CI/CD for end to end automation of release pipeline using DevOps tools like Jenkins.

- **Environment** : Azure, Azure Synapse Analytics, CI/CD, Cluster, Cosmos DB, Data Factory, ETL, Data Factory, HDFS, HDInsight, Hive, Java, Jenkins, Jira, JavaScript, Kafka, Kubernetes, Oracle, Pig, PySpark, Python, Selenium, Snowflake, Spark, Spark SQL, Spark Streaming, SQL, Sqoop, Teradata

Celanese • Irving, Texas, USA • 12/2022 - 11/2023

Celanese is a global leader in chemistry, producing specialty material solutions used across most major industries and consumer applications. Played a key role in enhancing data processing systems, developing security frameworks, and automating infrastructure management to improve overall efficiency and scalability.

AWS Data Engineer

- Looked into existing Java/Scala spark processing and maintained, enhanced the jobs.
- Developed ETL pipelines using PySpark. Used both Data frame API and Spark SQL API.
- Designed and developed Security Framework to provide fine grained access to objects in AWS S3 using AWS Lambda, DynamoDB.
- Wrote oozie scripts and setting up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs.
- Used Kafka functionalities like distribution, partition, replicated commit log service for messaging systems by maintaining feeds.
- Responsible for estimating cluster size, monitoring, and troubleshooting the Spark Databricks cluster.
- Provisioned high availability of AWS EC2 instances, migrated legacy systems to AWS, and developed Terraform plugins, modules, and templates for automating AWS infrastructure.
- Involved in the entire lifecycle of the projects including Design, Development, and Deployment, Testing and Implementation, and support.
- Worked on SQL and PL/SQL for backend data transactions and validations
- Built and configured Jenkins slaves for parallel job execution. Installed and configured Jenkins for continuous integration and performed continuous deployments.
- Working on migrating Data to the cloud (Snowflake and AWS) from the legacy data warehouses and developing the infrastructure.
- Integrated Kubernetes with cloud-native services, such as AWS EKS and GCP GKE, to leverage additional scalability and managed services.
- Developed Kibana Dashboards based on the Logstash data and Integrated different source and target systems into Elasticsearch for near real time log analysis of monitoring End to End transactions.
- Developed metrics based on SAS scripts on legacy system, migrated metrics to Snowflake (AWS S3) .
- Executed full CI/CD pipeline by coordinating SCM (Git) with computerized testing instrument Gradle and Deployed utilizing Jenkins (Declarative Pipeline) and Dockerizing containers with various DevOps tools such as AWS CloudFormation, AWS CodePipeline, Terraform, and Kubernetes.
- Designed and deployed multi-tier applications using all AWS services (EC2 , Route53 , S3 , RDS, DynamoDB, SNS, SQS, IAM) with an emphasis on high-availability, fault tolerance, and auto-scaling in AWS CloudFormation.
- **Environment** : AWS, CI/CD, Docker, DynamoDB, EC2, Elasticsearch, ETL, GCP, Git, Java, Jenkins, Kafka, Kubernetes, lake, Lambda, Oozie, PL/SQL, PySpark, RDS, S3, SAS, Scala, Snowflake, Spark, Spark SQL, SQL

Capgemini (Intel) • Chennai, Tamil Nadu, India • 02/2022 - 10/2022

Intel Corporation is an American multinational corporation and technology company. Developed ETL processes using Google Cloud Dataflow or Apache Beam to move data from various sources into a centralized platform.

GCP Data Engineer

- Creating Data Studio report to review billing and usage of services to optimize the queries and contribute in cost saving measures.
- Loaded and transformed large sets of structured, semi structured, and unstructured data using Hadoop/Big Data concepts.

- Used PowerBI as a front-end BI tool to design and develop dashboards, workbooks, and complex aggregate calculations.
- Using Dataproc, Big Query to develop and maintain GCP cloud base solutions.
- Experienced in GCP features which include Google Compute Engine, Google Storage, VPC, Cloud Load Balancing, and IAM.
- Performed the migration of large data sets to Databricks (Spark) , created and administered cluster, loaded data, configured data pipelines, loading data from ADLS Gen2 to Databricks using ADF pipelines.
- Good knowledge in using Cloud Shell for various tasks and deploying services.
- Involved in writing Spark applications using Scala/Java.
- Developed a Front-End GUI as stand-alone Python application.
- Experienced in Google Cloud components, Google Container Builders and GCP client libraries and Cloud SDK'S.
- Managed large datasets using Panda DataFrames and SQL.
- Build and deployed the code artefacts into the respective environments in the Confidential Azure cloud.
- Work related to downloading BigQuery data into pandas or Spark DataFrames for advanced ETL capabilities.
- Worked with GCP cloud using in GCP Cloud storage, Dataproc, Dataflow, BigQuery, EMR, S3, Glacier, and EC2 with EMR Cluster
- Processed and loaded bound and unbound Data from Google pub/sub topic to BigQuery using Cloud Dataflow with Python.
- Build data pipelines in Airflow/Composer for orchestrating ETL related jobs using different Airflow operators.
- **Environment**: Airflow, Azure, BigQuery, Cluster, EC2, EMR, ETL, GCP, Java, Python, S3, Scala, SDK, Spark, SQL, VPC

HSBC • **Chennai, Tamil Nadu, India • 05/2019 - 01/2022**

HSBC is a British universal bank and financial services group. Managed the execution of data extraction, transformation, and loading processes, overseeing pipeline creation and performance optimization to meet analytical needs and project goals.

Data Engineer

- Executed Extract, Transform, and Load (ETL) operations, extracting data from source systems and loading it into Azure Data Storage services.
- Involved in creating Hive tables, loading, and analyzing data using Hive scripts.
- Monitored Spark clusters using Log Analytics and Ambari Web Ul. Transitioned log storage from Cassandra to Azure SQL Data Warehouse and improved the query performance.
- Developed Spark Streaming programs to process near real time data from Kafka, and process data with both stateless and state full transformations. Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
- Took personal responsibility for meeting deadlines and delivering high quality work and create POCs to demonstrate new technologies including Jupyter Notebooks, PySpark. Designed and implemented Infrastructure as code using Terraform, enabling automated provisioning and scaling of cloud resources on Azure.
- Involved in various phases of Software Development Lifecycle (SDLC) of the application, like gathering requirements, design, development, deployment, and analysis of the application.
- Storing different configs in No SQL database MongoDB and manipulating the configs using PyMongo.
- Integrated Azure Data Factory with Blob Storage to move data through Databricks for processing and then to Azure Data Lake Storage and Azure SQL Data Warehouse.
- Implemented Azure data lake, Azure Data Factory and Azure data bricks to move and conform the data from on -premises to cloud to serve the analytical needs of the company.
- Implemented Kubernetes namespaces and RBAC Role-Based Access Control policies to enforce security and access controls in the data infrastructure.

- Develop metrics based on SAS scripts on legacy system, migrating metrics to Snowflake (Google Cloud) .
- Implemented Docker containers to create images of the applications and dynamically provision slaves to Jenkins CI/CD pipelines
- **Environment** : Azure, Azure Data Lake, Blob, Cassandra, CI/CD, Data Factory, Docker, EC2, EMR, ETL, Factory, Hive, Java, Jenkins, Kafka, Kubernetes, Lambda, PySpark, S3, SAS, Spark, Spark Streaming, SQL, Sqoop.

EDUCATION

Masters

Indiana Wesleyan University • USA

CERTIFICATIONS

Microsoft Certified Azure Data Fundamentals

Google Certified Associate Cloud Engineer

AWS Certified Associate Data Engineer

SKILLS

ETL: Azure Data Factory, Azure Data-bricks, Informatica Cloud, Kafka, Spark, MS SSIS, AWS Glue, Am-azon EMR

Programming Languages: C, Python, PySpark, Scala, SQL, PL/SQL, T-SQL

Tools: Microsoft Visual Studio, SQL Server management Studio, Py-Charm, PostgreSQL, Azure Data Studio

Reporting Tools: Tableau, PowerBI, MS Excel

Cloud Technologies: Microsoft Azure, AWS, GCP

Version Control Tools: Git and GitHub

IDE’s Utilities: Eclipse, PyCharm

Methodologies: Agile, Waterfall

Operating Systems: UNIX, Linux, Windows, Mac OS

RDBMS Databases: MYSQL, MS SQL Server, Terada-ta, Oracle, PL/SQL

Google Cloud Platform: GCP Composer, BigQuery, GCS, Cloud Dataproc, Cloud SQL, Cloud Functions, Cloud Dataflow, Cloud Datafusion, Cloud Pub/Sub