**Name: Annam Ribka Rani**

**Email:** [**ribkaraniannam6@gmail.com**](mailto:ribkaraniannam6@gmail.com)

**PH: 913-335-0638**

**Linkedin:** [**https://www.linkedin.com/in/ribkaraniannam/**](https://www.linkedin.com/in/ribkaraniannam/)

**Sr Data Engineer**

**Professional Summary**

* 9+ years of overall experience as a Data Engineer, Data Analyst, and ETL developer, with expertise in comprising designing, developing, and implementing data models for enterprise-level applications using Big Data tools and technologies such as Hadoop, Sqoop, Hive, Spark, Flume.
* Experienced with the use of AWS services including S3, EC2, SQS, RDS, Neptune, EMR, Kinesis, Lambda, Step Functions, Terraform, Glue, Redshift, Athena, DynamoDB, Elasticsearch, Service Catalog, CloudWatch, IAM and administering AWS resources using Console and CLI.
* Hands-on experience in building the infrastructure necessary for the best data extraction, transformation, and loading from a range of data sources using NoSQL and SQL from AWS & Big Data technologies (Dynamo, Kinesis, S3, HIVE/Spark)
* Developed and deployed a variety of Lambda functions using the built-in AWS Lambda Libraries and Lambda functions written in Scala and using custom libraries.
* Capable of using AWS utilities such as EMR, S3, and cloud watch to run and monitor Hadoop and spark jobs on Amazon Web Services (AWS).
* Strong knowledge in working with Amazon EC2 to provide a complete solution for computing, query processing, and storage across a wide range of applications.
* Experienced in working with Azure cloud platforms (HDInsight, DataLake, Databricks, Blob Storage, Data Factory, Azure Functions, Azure SQL data warehouse, and Synapse).
* Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks. Use the RUP and agile methodology to conduct new development and maintaining software.
* Proficient in migrating on-premises data sources to Azure data lake, Azure SQL Database, Databricks, and Azure SQL Data warehouse using Azure Data factory and granting access to the users.
* Experienced in Developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into customer usage patterns.
* Experienced with Azure Data Factory (ADF), Integration Run Time (IR), File System Data Ingestion, and Relational Data Ingestion.
* Experienced in configuring Spark Streaming to receive real-time data from Apache Kafka and store the stream data to HDFS and expertise in using spark-SQL with various data sources like JSON, Parquet, and Hive.
* Expertise in developing production-ready Spark applications utilizing Spark-Core, Data Frames, Spark-SQL, Spark-ML, and Spark-Streaming API.
* Expert in using Azure Databricks to compute large volumes of data to uncover insights into business goals.
* Experienced in integrating data from diverse sources, including loading nested JSON-formatted data into Snowflake tables, using the AWS S3 bucket and the Snowflake cloud data warehouse.
* Configured Snow pipe to pull the data from S3 buckets into the Snowflake table and stored incoming data in the Snowflake staging area.
* Expertise in developing Python scripts to build ETL pipelines and Directed Acyclic Graph (DAG) workflows using Airflow.
* Orchestration experience in scheduling Apache Airflow DAGs to run multiple Hive and spark jobs, which independently run with time and data availability.
* Strong experience working with Hadoop ecosystem components like HDFS, Map Reduce, Spark, HBase, Oozie, Hive, Sqoop, Pig, Flume, and Kafka.
* Hands-on experience with Hadoop architecture and various components such as Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and Hadoop MapReduce programming.
* Strong experience troubleshooting failures in spark applications and fine-tuning spark applications and hive queries for better performance.
* Strong experience writing complex map-reduce jobs including the development of custom Input Formats and custom Record Readers.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems(RDBMS) and vice-versa.
* Experience in analysing data using Python, R, SQL, Microsoft Excel, Hive, PySpark, and Spark SQL for Data Mining, Data Cleansing, Data Munging, and Machine Learning.
* Expertise in the development of various reports, and dashboards using various Power BI, and Tableau.
* Excellent Communication skills, Interpersonal skills, problem-solving skills, and a team player. Ability to quickly adapt to new environments and technologies.

**Professional Experience**

**Sr. AWS Data Engineer**

**Apple, Sunnyvale, CA May 2022 to Present**

**Description:** Led an AWS-centric project at Apple, orchestrating fault-tolerant infrastructure, creating data pipelines, optimizing Spark processing, and utilizing Tableau for insightful customer data, refining digital experiences.Transforming AWS services for high-performance data engineering. Led multi-tier app design with AWS tools, optimized Spark processing, and orchestrated seamless data flow for RedShift and Tableau-driven insights.

**Responsibilities:**

* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53,S3, RDS, Dynamo DB, SNS, SQS, and IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Developed the AWS Data pipelines from various data resources in AWS including AWS API Gateway to receive responses from AWS Lambda and retrieve data and converted responses into JSON format and stored them in AWS redshift.
* Developed the scalable AWS Lambda code in Python for nested JSON files, converting, comparing, sorting, etc.
* Developed Spark Applications by using Scala and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Optimized the performance and efficiency of existing spark jobs and converted the Map-reduce script to spark SQL.
* Experienced in collecting data from an AWS S3 bucket in real time using Spark Streaming, doing the appropriate transformations and aggregations, and persisting the data in HDFS.
* Implemented AWS glue catalog with crawler to get the data from S3 and perform SQL query operations.
* Developed robust and scalable data integration pipelines to transfer data from the S3 bucket to the RedShift database using Python and AWS Glue.
* Use the RUP and agile methodology to conduct new development and maintaining software.
* Built and maintained the Hadoop cluster on AWS EMR and has used AWS services like EC2 and S3 for small data sets processing and storage.
* Identified, evaluated, and documented potential data sources in support of project requirements
* within the assigned departments as per agile methodology
* Developed Python code for different tasks, dependencies, and time sensors for each job for workflow management and automation using the Airflow tool.
* Scheduling Spark/Scala jobs using Oozie workflow in Hadoop Cluster and generated detailed design documentation for the source-to-target transformations.
* Designed the reports and dashboards to utilize data for interactive dashboards in Tableau based on business requirements.

**Environment**: AWS EMR, S3, EC2, Lambda, Apache Spark, Spark-Streaming, Spark SQL, Python, Scala, Shell scripting, Snowflake, AWS Glue, Oracle, Git, Tableau.

**Sr. Data Engineer (Azure)**

**Finra, Rockville, MD August 2020 to April 2022**

**Description:** Led pivotal Azure-based market surveillance project at FINRA, optimizing data workflows, compliance, and analytics for robust regulatory adherence. Transforming data landscapes with expertise in Azure cloud services, Databricks, Data Factory, and SQL DB. Driving impactful projects across ingestion, processing, and ETL pipelines, optimizing Azure ecosystems for efficiency and scalability. Spearheaded data engineering initiatives at FINRA, orchestrating Azure cloud based ETL pipelines, enabling seamless data migration, transformation, and streamlining complex Spark applications. Leveraging Power BI, enhanced reporting insights, and collaborated on repository strategies for robust version control.

**Responsibilities:**

* Extensively worked with Azure cloud platform (HDInsight, Data Lake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH, and Data Storage Explorer).
* Ingested data to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processed the data in In Azure Databricks.
* Designed and configured Azure Cloud relational servers and databases, analyzing current and future business requirements.
* Developed data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL.
* Configured Input & Output bindings of Azure Function with Azure Cosmos DB collection to read and write data from the container whenever the function executes.
* Developed robust ETL pipelines in Azure Data Factory (ADF) using Linked Services from different sources and loaded them into Azure SQL Datawarehouse.
* Developed Elastic pool databases and scheduled Elastic jobs to execute T-SQL procedures.
* Proficient in performing ETL operations in Azure Databricks by connecting to different relational database source systems using JDBC connectors.
* Migrated data from Azure Blob storage data to Azure Data Lake using Azure Data Factory (ADF).
* Developed the robust & scalable ETL Azure Data Lake to Data warehouse applications for Medicaid and Medicare data using the Azure Databricks.
* Built and automated data engineering ETL pipeline over Snowflake DB using Apache Spark and integrated data from disparate sources with Python APIs like PySpark and consolidated them in a data mart (Star schema).
* Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Databricks.
* Worked on Power Shell scripts to automate the Azure cloud system creation of Resource groups, Web Applications, Azure Storage Blobs & Tables, and firewall rules.
* Designed and provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Experience in working with Spark applications like batch interval time, level of parallelism, and memory tuning to improve processing time and efficiency.
* Orchestrate the airflow to migrate the data from Hive external table to azure blob storage and optimized the existing hive jobs using the concepts like partition and bucketing.
* Used Azure DevOps and VSTS (Visual Studio Team Services) for CI/CD, Active Directory for authentication, and Apache Ranger for authorization.
* Experience in working with Spark applications like batch interval time, level of parallelism, and memory tuning to improve processing time and efficiency.
* Used Scala for amazing concurrency support, and Scala plays a key role in parallelizing processing of the large data sets.
* Used Enterprise GitHub and Azure DevOps Repos for version control.
* Created branching strategies while collaborating with peer groups and other teams on shared repositories.
* Developed various interactive reports using Power BI based on Client specifications with row-level security features.

**Environment**: Azure (Data Lake, HDInsight, SQL, Data Factory), Databricks, Cosmos DB, Git, Blob Storage,Power BI, Scala, Hadoop, Spark, PySpark, Airflow.

**Big Data Engineer**

**NantHealth, Providence, RI February 2018 to July 2020**

**Description:** Revamped healthcare analytics at Nant Health by orchestrating diverse data extraction, processing, and team-led pipeline optimization, cutting insights time by 50%. Orchestrated a pioneering healthcare data integration pipeline, optimizing data extraction, processing, and analytics. Streamlined critical healthcare datasets, driving informed decision-making. Spearheaded a multidisciplinary team to design and implement a high-efficiency data pipeline, reducing insights' time by 50%. Employed cutting-edge technologies and methodologies to elevate healthcare data management.

**Responsibilities:**

* Extracted data from HDFS, including customer behaviour, sales and revenue data, supply chain, and logistics data.
* Transferred the data to AWS S3 using Apache Nifi, which is an open-source data integration tool that enables powerful and scalable dataflows.
* Validated and cleaned the data using Python scripts before storing it in S3.
* Used PySpark to process and transform the data, which is a distributed computing framework for big data processing with Python API.
* Loaded the transformed data into AWS RedShift data warehousing to analyze the data.
* Scheduled the pipeline using Apache Oozie, which is a workflow scheduler system to manage Apache Hadoop jobs.
* Developed and maintained a library of custom Airflow DAG templates and operators, which improved consistency and code quality across the team.
* Led a team of three data engineers in designing and implementing a complex data ingestion and processing pipeline for a new data source, which reduced time to insights by 50%.
* Analyzed the data in HDFS using Apache Hive, which is a data warehouse software that facilitates querying and managing large datasets.
* Converted Hive queries into PySpark transformations using PySpark RDDs and Data Frame API.
* Monitored the data pipeline and applications using Grafana.
* Configured Zookeeper to support distributed applications.
* Used functional programming concepts and the collection framework of Scala to store and process complex data.
* Used GitHub as a version control system for managing code changes.
* Developed visualizations and dashboards using Tableau for reporting and business intelligence purposes.

**Environment**: S3 buckets, red shift, Apache flume, PySpark, Oozie, Tableau, Scala, Spark RDDs, Hive, HiveQL, HDFS, HQL, Scala, Zookeeper, Grafana MapReduce, Sqoop, GitHub

**ETL Developer**

**Charter Communications, St Louis, MO November 2016 to January 2018**

**Description:** Revolutionized data integration at Charter Communications, employing Informatica, Kafka, and Sqoop for real-time analytics, fostering informed decisions and operational efficiency. Orchestrated robust data workflows at a financial institution, leveraging Informatica tools for ETL design, implementing Kafka and Spark Streaming for real-time data ingestion, and utilizing Sqoop and Flume for seamless data transfer from diverse sources, enhancing data accuracy and accessibility. Spearheaded end-to-end data management strategies, ensuring adherence to ETL standards through meticulous SQL query evaluations. Drove process optimization by converting business requirements into technical design documents and collaborating with stakeholders to develop comprehensive business requirement documents (BRD), thereby streamlining data processes and enhancing system efficiency.

**Responsibilities:**

* Extensively used Informatica Client tools Power Center Designer, Workflow Manager, Workflow Monitor, and Repository Manager.
* Used Kafka for live streaming data and performed analytics on it. Worked on Sqoop to transfer the data from relational database and Hadoop.
* Configured in building real-time data pipelines using Kafka for streaming data ingestion and Spark Streaming for real-time consumption and processing.
* Loaded data from Web servers and Teradata using Sqoop, Flume, and Spark Streaming API.
* Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS. Implemented a Python-based distributed random forest via Python streaming.
* Written multiple MapReduce programs for data extraction, transformation, and aggregation from multiple file-formats including XML, JSON, CSV & other compressed file formats.
* Extracted data from various heterogeneous sources like Oracle, and Flat Files.
* Developed complex mapping using the Informatica Power Center tool.
* Extracted data from Oracle and Flat files, Excel files, and performed complex joiner, Expression, Aggregate, Lookup, Stored procedure, Filter, Router transformation, and Update strategy transformations to load data into the target systems.
* Worked with Data modeler in developing STAR Schemas.
* Handling a high volume of day-to-day Informatica workflow migrations.
* Review Informatica ETL design documents and work closely with development to ensure correct standards are followed.
* Worked on SQL queries to query the Repository DB to find the deviations from Company's ETL Standards for the objects created by users such as Sources, Targets, Transformations, Log Files, Mappings, Sessions, and Workflows.
* Leveraging the existing PL/SQL scripts for the daily ETL operation.
* Experience in ensuring that all support requests are properly approved, documented, and communicated using the QMC tool. Documenting common issues and resolution procedures
* Extensively involved in enhancing and managing Unix Shell Scripts.
* In converting the business requirement into a technical design document.
* Documenting the macro logic and working closely with Business Analyst to prepare BRD.
* Involved in setting up SFTP setup with the internal bank management.
* Building UNIX scripts in cleaning up the source files.
* Involved in loading all the sample source data using SQL loader and scripts.

**Data Analyst- Python**

**ICICI Bank, Hyderabad, India June 2014 to October 2016**

**Description:** Enhanced ICICI Bank's analytics: Integrated diverse data sources, performed detailed analysis, and optimized operations, refining customer insights and financial strategies.

**Responsibilities:**

* Experience working on projects with machine learning, big data, data visualization, R and Python development, Unix, and SQL.
* Performed exploratory data analysis using NumPy, matplotlib, and pandas.
* Expertise in quantitative analysis, data mining, and the presentation of data to see beyond the numbers and understand trends and insights.
* Experience analyzing data with the help of Python libraries including Pandas, NumPy, SciPy, and Matplotlib.
* Created complex SQL queries and scripts to extract and aggregate data to validate the accuracy of the data and Business requirements gathering and translating them into clear and concise specifications and queries.
* Prepared high-level analysis reports with Excel and Tableau. Provides feedback on the quality of Data including identification of billing patterns and outliers.
* Worked on sort & filters of tableau like Basic Sorting, basic filters, quick filters, context filters, condition filters, top filters, and filter operations.
* Identify and document limitations in data quality that jeopardize the ability of internal and external data analysts' ability; wrote standard SQL Queries to perform data validation; created excel summary reports (Pivot tables and Charts); and gathered analytical data to develop functional requirements using data modeling and ETL tools.
* Read data from different sources like CSV files, Excel, HTML pages, and SQL and performed data analysis and wrote to any data source like CSV file, Excel, or database.
* Experience in using Lambda functions like filter (), map (), and reduce () with pandas Data Frame and performing various operations.
* Used Pandas API for analyzing time series. Creating regression test framework for new code.
* Developed and handled business logic through backend Python code.

**Environment**: Python, SQL, UNIX, Linux, Oracle, NoSQL, PostgreSQL, and python libraries sch as PySpark, and NumPy.