**CHARITHA**

**Email:** **krishna.mkc0208@gmail.com PH: 469-554-9201**

**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 in working with Azure cloud platforms (HDInsight, DataLake, Databricks, Blob Storage, Data Factory, Azure Functions, Azure SQL data warehouse, and Synapse).
* 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 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 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.
* Extensively used Spark Data Frames API over the Cloudera platform to perform analytics on Hive data and used Spark Data Frame Operations to perform required Validations in the data.
* 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.
* Developed Python scripts to do file validations in Databricks and automated the process using Azure Data Factory.
* 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.
* Practical understanding of Data modeling (Dimensional & Relational) concepts like Star-Schema Modelling, Fact, and Dimension tables.
* 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.
* Good exposure to usage of NoSQL databases column-oriented Cassandra and MongoDB (Document Based DB).
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems(RDBMS) and vice-versa.
* Experience in analyzing 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 Data Engineer**

**Fifth Third Bank, Evansville, IN April 2022 to Present**

**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.
* Developed Spark applications in azure Databricks using Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into customer usage patterns.
* 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.

**Data Engineer**

**Ascena Retail Group, Patskala, Ohio November 2019 to March 2022**

**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.
* 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.
* 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.

**Big Data Engineer**

**Early Warning's, Scottsdale, AZ May 2017 to October 2019**

**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

**GrapeSoft Solutions Hyderabad, India September 2015 to February 2017**

**Data Engineer**

**Responsibilities:**

* Good working experience on Hadoop architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node & Map Reduce programming paradigm.
* Configured Spark streaming to receive real-time data from Kafka and store the stream data to HDFS using Python.
* Experienced in designing, built, and deploying multitude of the application utilizing the AWS stack, focusing on high-availability, fault tolerance, and auto-scaling.
* Experienced in developing spark application using Spark RDD API
* profound understanding of Partitions and Bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Worked extensively on fine tuning spark applications to improve performance and troubleshooting failures in spark applications.
* Knowledge of CI/CD pipeline on GCP
* Experience in developing Advanced SQL queries and stored procedures. Ensure appropriate data filters are applied based on business requirements
* Responsible for SQL query performance tuning.
* Strong Experience in working with Databases like Oracle 10g, SQL Server 2016, and MySQL and proficiency in writing complex SQL queries.
* Worked on fine-tuning spark applications to improve the overall processing time for the pipelines.
* Developed Spark data pipeline that reads data from Google Drive, perform transformations and ingest data to Snowflake.

**Environment:** Hadoop, MapReduce, HDFS Hive, Java Hadoop distribution of Horton Works, Cloudera, Pig, HBase, Linux, XML, MySQL, MySQL Workbench Java 6, Eclipse, Oracle 10g PL/SQL, SQL PLUS Sub Version Cassandra.

**Ceequence Technologies Hyderabad, India June 2014 to August 2015**

**Hadoop Developer**

**Responsibilities:**

* Involved in the evaluation of functional and non-functional requirements.
* Installed and configured Hadoop MapReduce HDFS Developed multiple MapReduce jobs in java for data cleaning and pre-processing.
* Installed and configured Pig and also written Pig Latin scripts.
* Wrote MapReduce job using Pig Latin.
* Involved in managing and reviewing Hadoop log files.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis.
* Developing Scripts and Batch Job to schedule various Hadoop Program.
* Written Hive queries for data analysis to meet the business requirements and created hive tables and worked on them using Hive QL.
* Importing and exporting data into HDFS and Hive using Sqoop.
* Experienced in defining job flows.
* Got good experience with NOSQL database SOLR HBase.
* Involved with putting data into Hive tables and constructing hive queries that will execute internally in a map reduce fashion.
* Created a custom Filesystems plug-in for Hadoop that allows it to access files on the Data Platform.
* This plugin enables Hadoop MapReduce applications HBase Pig and Hive to function normally and directly access files.
* Designed and implemented MapReduce-based large-scale parallel relation-learning system.
* Extracted feeds form social media sites such as Facebook Twitter using Python scripts.
* For internal purposes, I have set up benchmarked Hadoop clusters.

**Environment:** Hadoop, MapReduce, HDFS Hive, Java Hadoop distribution of Horton Works, Cloudera, Pig, HBase, Linux, XML, MySQL, MySQL Workbench Java 6, Eclipse, Oracle 10g PL/SQL, SQL PLUS Sub Version Cassandra.