Koushik N S **Mobile: +91-7676885789**
Associate Software Engineer **Email:** koushikns68@gmail.com

Associate Software Engineer

Associate Software Engineer with a proven track record in designing, implementing, and optimizing large-scale data structures for analytical and reporting purposes. Possesses a comprehensive understanding of the Hadoop Ecosystem and its major components, coupled with expertise in Creating and implementing the data pipeline for ETL and data Warehousing methodology.

Technical Skills

Language:	SQL, Python, PySpark
Big Data Tool:	Hive, HBase, HDFS, Spark, Sqoop
Databases:	RDBMS, NoSQL
Cloud Technologies:	AWS (S3, lambda, Glue, Redshift)
Reporting Tool:	Tableau
Version Control System:	GitHub
Operating System:	Windows, Unix

Professional Experience

Associate Software Engineer in Maveric Systems Pvt Ltd in Bangalore from June 2022 to till date.

- Worked on the design and implementation of large-scale data structures, resulting in enhanced data processing efficiency for analytical and reporting needs.
- Developed the data pipeline for ETL processes to extraction, transformation, and loading of data into the data warehouse.
- Implemented and managed Hadoop based solutions, leveraging technologies like Hive and Spark for efficient data processing.
- Designed and implemented Apache Spark-based data processing pipelines to handle large-scale data ingestion, transformation, and analytics tasks.
- Developed HiveQL queries and scripts to extract, transform, and load data from various sources into Hive tables, optimizing performance and efficiency also Implemented Hive partitioning and bucketing strategies to improve query performance and manage large datasets effectively.
- Conducted performance tuning on databases and queries, improving overall system performance.
- Utilized cloud platforms, such as AWS, for scalable and efficient data storage and processing.
- Experience in working with the Different file formats like TEXTFILE, ORC, Parquet and JSON and store it in HDFS/Hive tables.
- Implemented a data governance framework that improved data quality and integrity, resulting in increased stakeholder trust and confidence in data-driven decision-making.
- Collaborated with cross-functional teams to understand data requirements and design solutions to support business analytics and reporting initiatives.

Education

Canara Engineering College Mangalore, VTU Karnataka, India
Bachelor of Engineering Aug 2018 – Jul 2022

Project Summary

Project 1: Client Review Center

Description: CRC is an efficient, User friendly, reporting engine that gives you access to robust library of Merrill Lynch approved exhibits – including information on performance, holdings, and asset allocation. The result is a colorful, easy to read professional report package that will help strengthen client relationship. The report generated using CRC provides the details about Cash, Equity, Fixed Income, Holdings, Mutual Funds, Alternative Investments etc. with Market value and percentage of accounts. The data will be persisted into the Hadoop file system, Hive are injected using Sqoop and Processing using Hive and Spark. Spark SQL load the aggregated data to RDBMS, Hive and Elastic Search for Analytics team and visualization in CRC UI

Responsibly:

- Data Ingestion and acquisition is done through Sqoop from RDBMS into hive table.
- Handling data sets and read data different source files like CSV, Parquet, Avro, Json, Sequence files and store it in HDFS/Hive tables.
- Importing historical transactions and user information into HDFS.
- Creating external and managed Hive tables with appropriate buckets and partitions.
- Data Enrichment, cleansing and common data aggregations through RDD transformations using Spark.
- Interactive analysis of Hive tables through various data frame operations using Spark SQL.
- Data parsing and processing by performing In-memory computation, data persistence, cache, repartition, performance tuning using Spark.
- Providing enriched data availability to frontend team.
- Storing the final processed data into RDBMS, Hive table, HDFS and in Elastic Search for making the final data available for visualization in CRC UI which the Analytics/Reporting teams can make use of it.
- Taking Maven build with the latest changes and provide to QA and Deployment team.
- Involved in story-driven agile development methodology and actively participated in daily scrum meetings.

Technologies: Sqoop, HDFS, Hive, Spark, Pyspark, SQL

Project 2: Health Care Data Analysis (Capstone Project)

Description: Creating the R&D Reports to plan and strategize the newly launched health product in the US market based on the customer transaction data.

Responsibly:

- Understanding the business requirement.
- Data ingestion: Loading Data from local system to HDFS pipeline directory.
- Reading the source file from web shell PySpark.
- Using PySpark data frame Pre-processing/cleansing the data as per the business requirement.
- Applying the transformations on the pre-processed data as per the business rules.
- Preparing and executing the SQL queries to verify target data.
- Loading the final target data to Hive database.
- Preparing the status report.

Technologies: HDFS, Hive, Spark, PySpark, SQL

Certification

- Big Data Certification
- Tableau Certification