Karthik Soma

Data Engineer

940-290-2319 karthik.s@mycvtalent.com

Professional Summary

- 4+ years of hands-on experience as a Data Engineer, specializing in architecting data-intensive applications leveraging Cloud Data engineering, Data Warehouse, the Hadoop Ecosystem, Big Data Analytical tools, Data Visualization, Reporting platforms, and Data Quality solutions.
- Expertise in designing and implementing enterprise-grade batch processing solutions with Databricks and streaming frameworks such as Spark Streaming, Apache Kafka, and Apache Flink.
- Leveraged advanced Big Data tools including Hadoop, Spark, HDFS, HBase, MapReduce, Hive, PIG, OOZIE, and SQOOP to develop scalable data solutions and perform comprehensive analytics.
- Excelled in deploying a broad spectrum of cloud services, mastering Amazon Web Services (AWS) technologies such as EC2, S3, RDS, and more. Expertly engineered robust data pipelines on Google Cloud Platform (GCP) using Apache Airflow and Google Cloud Dataflow; proficient in dynamic data management with technologies like Dataproc and BigQuery. Streamlined and enhanced Azure data solutions including Data Lake, Data Factory, and Databricks, significantly boosting analytics capabilities and operational efficiency.
- Strong knowledge of RDBMS concepts, Data Modeling (Facts and Dimensions, Star/Snowflake schemes), Data Migration, Data Cleansing and ETL Processes.

Education

University of North Texas **Jan 2023 – May 2024**

Advance Data Analytics – Master's Degree *Tx, USA*

Work Experience

CitiGroup **January 2024 – Current**

Data Engineer *Texas, USA*

- Demonstrated expertise in leveraging Spark and Scala APIs for performance comparison with Hive and SQL, and adept at manipulating Data Frames with Spark SQL, enhancing data analysis and processing capabilities.
- Designed and developed Tableau dashboards to monitor performance metrics, leading to a 10% improvement in team productivity by providing actionable insights.
- Optimized PySpark scripts for distributed data processing, improving processing speed by 30% and enhancing data quality for analytics.
- Led a successful AWS migration project, utilizing Lambda, S3, and Redshift, achieving a 20% cost reduction while showcasing deep AWS expertise. Additionally, established a data normalization and consolidation process with AWS Glue, reducing redundancy by 40% and improving data quality scores by 50%, which increased the accuracy of machine learning models by 15%.
- Optimized data processing efficiency by enhancing Hive schema design and implementing advanced performance techniques, leading to a 30% improvement in processing time and a 45% increase in query efficiency. Also, worked on improving Spark Streaming and Kafka for real-time data processing, including stateless/stateful transformations.
- Integrated Apache Airflow with AWS to monitor multi-stage ML workflows, with tasks running on Amazon SageMaker, and contributed to CI/CD solutions using Git and Jenkins for setting up and configuring the big data architecture on the AWS cloud platform.
- Designed and implemented robust data pipelines on GCP using Apache Airflow.
- Developed and deployed streaming and batch data processing workflows using Google Cloud Dataflow and Python SDK.
- Responsible for real-time data ingestion and processing across GCP and Hadoop ecosystems, utilizing technologies such as Dataproc, Cloud Functions, and BigQuery to execute comprehensive data engineering tasks.

R-Financial **Jan 2019 – Dec 2022**

Data Engineer *India*

- Preprocessed and integrated over 100 GB of data daily using Spark SQL, ensuring optimal data quality for analysis.
- Enhanced data pipelines for efficient storage of raw data in Hive, resulting in a 20% improvement in data retrieval speed.
- Developed and maintained Azure Data Lake Storage and Azure Data Factory pipelines for handling both structured and unstructured data, ensuring high data quality and availability. Led the design and implementation of data pipelines in Azure, boosting data collection, storage, and analytics capabilities.
- Developed a Python-driven live data conduit leveraging Apache Kafka and AWS Lambda, resulting in a 25% enhancement in data flow and facilitating immediate actionable insights.
- Utilized Airflow's scheduling capabilities to enhance pipeline execution efficiency, considering factors such as data availability and resource limitations.
- Optimized batch processing efficiency by 25% with Spark, resulting in accelerated analysis of data volumes up to 50GB.
- Implemented data cleansing workflows using Python and SQL, reducing data inaccuracies by 35% and enabling more reliable reporting and decision-making.
- Created a Power BI dashboard to effectively visualize critical business Key Performance Indicators (KPIs), resulting in a weekly saving of 10 hours previously spent on manual reporting duties.
- Designed and Implemented ETL Processes using AWS Glue to seamlessly transfer data from external repositories such as S3 and Parquet into AWS Redshift.

Skills

Programming Languages: Scala, Python, R, SQL.
IDEs: Eclipse, IntelliJ IDEA, PyCharm, Jupyter Notebook, Visual Studio Code.
Big Data Ecosystem: Hadoop, MapReduce, Hive, HDFS, Spark, Kafka, PySpark, Apache Airflow, Sqoop, Flume, Nifi, Oozie, Zookeeper, Apache Flink.
Machine Learning: Linear Regression, Logistic Regression, Decision Trees, SVM, K-Means, Random Forest.
Cloud Platforms and Services: AWS (S3, EMR, EC2, RDS, Glue, Lambda, SDK, DynamoDB, Elasticsearch, QuickSight, Kinesis, Athena, VPC, Redshift, CloudWatch, CloudFormation, Route53), Docker, Google Cloud (BigQuery, Cloud SQL, Cloud Composer/Airflow, Cloud Storage, Dataflow/Data Fusion, Dataproc, Pub/Sub), Azure (Data Lake, Data Factory, Databricks, Logic Apps, HDInsight, Synapse Analytics, Azure DevOps).
Packages: NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow.
CI-CD/Other Tools: Jenkins, Tableau, Power BI, SSIS, SSRS, SSMS.
Databases: SQL Server, PostgreSQL, MongoDB, HBase.
Operating Systems: Windows, MacOS, Linux.