SUDHEER CHANDRA KUNAPAREDDY
DATA ENGINEER
660-528-3128 | sudheer.c@mycvhire.com| Fremont, CA

SUMMARY

- Around 5 years of experience as a Data Engineer, specialized in designing and building cloud-based data-intensive applications, data warehouses, and big data solutions. Skilled in the Hadoop ecosystem, big data analytics, data visualization, and data quality.
- Expertise in Spark and Databricks to extract, transform, and aggregate customer data from diverse file formats (CSV, JSON, logs) to improve customer behavior insights.
- Proven ability to leverage Hadoop for big data storage, processing, and analysis and skilled in utilizing tools like MapReduce, Hive, HDFS, Sqoop, Apache Airflow, Apache Kafka, Apache Spark and Apache Flink.
- Experienced in leveraging AWS services like Athena, EC2, S3 Bucket, Amazon Redshift, Glue, Lambda, and AWS Pipeline for scalable and cost-effective big data analytics on Teradata datasets.

SKILLS

Programming Language:	Scala, Python, R, SQL
IDE's:	PyCharm, Jupyter Notebook
Big Data Ecosystem:	Hadoop, Hive, HDFS, Sqoop, Apache Airflow, Apache Kafka, Apache Spark, Apache Flink
ETL and Cloud Technologies:	SSIS, AWS (EC2, S3 Bucket, Amazon Redshift, Glue, Lambda, Athena, AWS Pipeline)
Visualizations:	Tableau, Power BI, Excel
Packages & Data Processing:	NumPy, Pandas, Matplotlib, Seaborn, TensorFlow, PySpark, Data Pipelines
Version Control & Database:	GitHub, Gitlab, SQL Server, PostgreSQL, MongoDB, MySQL, Snowflake
Operating Systems:	Windows, MacOS

EDUCATION

Master of Science in Information Studies | Trine University, AZ — **Dec 2023**
Bachelor in Computer Science | Andhra University, Andhra Pradesh, India — **May 2015**

EXPERIENCE

Northern Trust, CA | Data Engineer   **Aug 2023 – Current**

- Designed and orchestrated complex data pipelines using Airflow DAGs (Directed Acyclic Graphs) to automate data ingestion, transformation, and loading (ETL) processes.
- Utilized Spark SQL for data querying and analysis on distributed datasets stored in HDFS or data warehouses.
- Developed Spark streaming pipeline to batch real-time data, detect anomalies by applying business logic, and write the anomalies to the HBase table.
- Establish Apache Flink streaming jobs to continuously ingest and analyze application logs, enabling faster detection of potential system issues and enhancing operational efficiency.
- Execute data pipelines using Scala and PySpark (including libraries like NumPy and Pandas) to efficiently ingest, clean, and transform large-scale datasets (structured, semi-structured, and unstructured) from various sources.
- Streamlined data acquisition and integration from multiple sources, improving accuracy and completeness by 35%.

Blue light IT Solutions, India | Data Engineer   **Aug 2017 - July 2021**

- Developed complex HiveQL queries (including joins, aggregations, and UDFs) to transform and analyze large datasets for data analysts and scientists.
- Integrate Lambda with other AWS services (API Gateway, SNS, SQS) to build event-driven data architectures, enabling real-time notifications and actions.
- Established complex Hive queries and scripts to perform data extraction, transformation, and aggregation tasks.
- Implement a robust data streaming architecture using Kafka, efficiently handling the ingestion and distribution of real-time data across diverse applications.
- Employed Snowflake materialized views, data masking, and optimization to deliver efficient and secure data solutions for seamless data management and advanced big data analytics.
- Spearheaded a benchmark using Spark and Scala APIs to compare Spark against Hive and SQL, achieving a 20% performance improvement for data processing tasks.

Oval SoftTech, India | Data Scientist   **Jun 2015 - July 2017**

- Employed a range of supervised learning algorithms (Linear Regression, Logistic Regression, Decision Trees, SVM, Naive Bayes) to tackle data classification challenges and enhance predictive modeling accuracy.
- Collaborated with cross-functional teams to implement automated dashboards and reporting using Tableau, empowering stakeholders with data-driven insights, and a remarkable improvement in workforce performance.
- Performed text pre-processing tasks like tokenization and stemming using NLTK to prepare text data for machine learning models in a sentiment analysis project.
- Applied Natural Language Processing (NLP) techniques for deploying a spam detection model and performing sentiment analysis on customer product reviews.
- Operated different types of Python packages in Data Science such as Tensorflow, PyTorch, Scipy, NumPy, Matplotlib, Seaborn, and Scikit-Learn.