Mounika Vakkanti
Sr Data Engineer
Mobile:(806)-401-7299
mounikavakkanti27@gmail.com https://www.linkedin.com/in/mouni-vakkanti-520658237/

SUMMARY:

Around 5 years of experience in data engineering and Python development, and I specialize in orchestrating, optimizing, and maintaining data pipelines using Apache Spark, Kafka, and AWS services. I am proficient in data warehousing, ETL processes, and leveraging tools like Informatica, Teradata, and Apache Airflow. My expertise includes architecting large-scale data solutions on cloud platforms such as AWS and Snowflake, with extensive experience in Big Data technologies including HDFS, MapReduce, Hive, Sqoop, Flume, Kafka, Oozie, Pig, and HBase. I possess strong programming skills in Python and Java, with solid expertise in NoSQL databases like HBase, Cassandra, and MongoDB. I am adept at creating insightful data visualizations using PowerBI and Tableau, generating visualizations with R and Python, and leveraging Databricks for data engineering and analytics. Additionally, I have a proven track record in IT, Big Data technologies, data analysis, and data modeling, committed to high-quality solutions in Agile environments.

EDUCATION:

- Master's in computer science
  Texas Tech University, Lubbock, TX in 2024.
- Bachelor's in Electronics and Communication Engineering.
  Chebrolu Engineering Collage, Affiliated by JNTU Kakinada, Guntur 2018

TECHNICAL SKILLS:

Big Data Systems :	HDFS, mapReduce, YARN, Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP), Cloudera Hadoop, Hortonworks Hadoop, Apache Spark, Spark Streaming, Apache Kafka, Pig Hive, Amazon S3, AWS Kinesis
Databases :	Cassandra, HBase, DynamoDB, MongoDB, BigQuery, SQL, Hive, MySQL, Oracle, PL/SQL, RDBMS, AWS Redshift, Amazon RDS, Teradata, Snowflake
Programming & Scripting:	Python, R, Scala, PySpark, SQL, Java, Bash
Web Programming:	HTML, CSS, JavaScript, XML , Angular
ETL Data Pipelines:	Apache Airflow, Sqoop, Flume, Apache Kafka, DBT, Pentaho, SSIS
Hadoop ecosystem:	Power BI, Quick Sight, Looker, Hive, Sqoop, Flume, ZooKeeper, Oozie, Spark, Knox, Kerberos, Tableau, Cloudra, Hortonworks, Apache Hadoop, AWS, Amezon EC2, S3.
Cloud Platforms:	AWS, GCP, Azure
Scheduler Tools:	Apache Airflow, Azure Data Factory, AWS Glue, Step functions
Spark Framework:	Spark API, Spark Streaming, Spark Structured Streaming, Spark SQL
CI/CD Tools:	Jenkins, GitHub, GitLab
Operating Systems :	Windows, Linux /Unix, Mac OS X
Process:	Agile/Scrum
Data Modeling:	ER Studio, power designer
Hadoop Cluster:	Cloudera, Hortonworks, Apache Hadoop, AWS, Amazon EC2, S3.
Data Science:	Classification, Recommendation, Clustering, Topic Modeling, NLP, R, Spark ML, H2O

PROFESSIONAL EXPERIENCE: .

Client: Fannie Mae, Plano, Texas **March 2024 -Current**
Role: Data Engineer
Responsibilities:

- Designed and Developed **ETL** pipelines using **AWS.**
- Ingested data through **AWS Kinesis Data Stream** and **Firehose** like **Parquet,** csv files from various sources to S3.
- Experience with **AWS cloud** (EMR, EC2, RDS, EBS, S3, Lambda, Glue, Elasticsearch, Kinesis, SQS, DynamoDB, Redshift, ECS)**.**
- Hands on Experience with ETL tools such as AWS Glue, Using Data pipeline to move data to AWS RedShift.

- Elasticsearch was used to filter data from S3 buckets, and data was then loaded into external Hive tables.
- Performed analysis on terabytes of data using **Python, SparkSQL, S3** and **Redshift** In order to obtain customer insights.
- Collaborated with the infrastructure team to ensure seamless external integrations, implementing scalable and maintainable IaC solutions using Terraform and AWS CloudFormation.
- Involved in building a custom Rest API to help data scientists and applications with real-time customer analytics.
- Built ETLs to load the data from Presto, PostgreSQL, Hive, SQL Server to Snowflake using Apache Airflow, Python and Spark.
- Independently migrated scheduled tasks and pipelines to Apache Airflow DAGs, optimizing scheduling and dependencies for efficient and reliable data workflows.
- Maintained and optimized DBT data models for efficient transformations and loading, while managing and securing the Snowflake Data Warehouse.

Environment: Python, PySpark, SQL, Scala, Git, Data Lake, Hive, Flume, HBase, SQL Server, RDBMS, Tableau, Teradata, Cassandra, Talend, AWS, Apache Airflow DAG, snowflake, NumPY, Pandas, Apache Parquet.

Client: Texas Tech University Health Sciences Centre, Texas **March 2023 – February 2024**

Role: Software Developer

Responsibilities:

- Involved in the analysis, design, development, and testing phases of Software Development Lifecycle (SDLC).
- Engaged in Agile development processes to gather and analyze application requirements.
- Involved in analyzing microservices architecture challenges, design analysis, development of the user stories, assigning tasks, and testing the application.
- Developed and enhanced Python GUI applications to diagnose speech-impaired children remotely, leveraging advanced technology solutions.
- Worked in the HPCC department to conduct an in-depth analysis of an open-source stress-ng program for stress testing computer systems, gaining insights into system performance.
- Enhancing the stress-ng functionality with the development of an open API, expanding stress-testing functionalities.
- Demonstrated expertise in utilizing stress testing tools and methodologies, ensuring robustness and reliability of computer systems.
- Collaborated within the team to implement technological advancements, fostering innovation and continuous improvement in software development.
- Worked with Collections, Exceptions, and Interfaces to develop Python applications.
- Cleaned and preprocessed large datasets using PySpark to ensure data quality and consistency. Identify and handle missing or incorrect data values, outliers, and duplicates.
- Cleaned and preprocessed raw data using Python and R to ensure data quality and usability for analysis.
- Web Development using Asia Specific Seach language and hearing, conference in Viathnam using Angual, Java Script, HTML, CSS

Environment: Java, Python, WebLogic, Spring Framework, Spring MVC, json, Eclipse, R, HTML5, CSS3, GIT, R studio, pySpark, Angualr, Typescript, restAPIs, Panda, numpy

Client: FIS GLOBAL, Bangalore, Karnataka, India **April 2021 – August 2022**

Role: Sr. Data Engineer

Projects: Big Data Fabric, Customer Segmentation and Personalization, Data Quality Monitoring System

Responsibilities: -

- Design robust, scalable data-driven solutions and pipeline frameworks to automate ingestion, processing, delivery of structured/unstructured batch and real-time streaming data using Python.
- Built data warehouse structures, dimensional modeling, Star/Snowflake schemas.
- Applied transformations on Spark Data Frames for in-memory computation.
- Troubleshot, tuned Spark applications, Hive scripts for optimal performance. Utilized Spark Data Frames API for analytics on Hive data, validation.
- Developed UDFs, SQL queries in Spark SQL.
- Modified data ingestion pipelines using Kafka, Sqoop for database tables, streaming data into HDFS.
- Encoded, decoded json objects using PySpark.

- Established naming standards, Data dictionary for Meta Data Management.
- Developed ETL workflows in Python for processing data in HDFS, HBase using Flume.
- Analyzed large datasets using HDFS, HBase, Hive, HQL, Pig, Sqoop, Zookeeper.
- Conducted POCs using Spark, Scala on Yarn Cluster, compared performance with Hive, SQL.
- Built scalable distributed data solutions in Hortonworks Hadoop Cluster environment.
- Utilized AWS EC2 for computational tasks, S3 for storage.
- Leveraged AWS utilities (EMR, S3, CloudWatch) to run, monitor Hadoop, Spark jobs. Maintained AWS Data pipeline for data movement between S3, EMR, RDS.
- Conducted data cleaning, preprocessing, modeling using Spark, Python.
- Implemented real-time secured REST APIs for data consumption using AWS Lambda, API Gateway, Kinesis, Swagger, Okta, Snowflake. Developed automation scripts for data transfer to Google Cloud Platform.
- Loaded files data from ADLS Server to Google Cloud Platform Buckets, created Hive Tables. Performed performance tuning, optimization of Spark jobs, and queries (Hive/SQL).
- Implemented real-time streaming of AWS CloudWatch Logs to Splunk using Kinesis Firehose.
- Developed AWS CloudWatch Dashboards for monitoring API Performance.
- Providing Solutions & Architecture for business use cases.
- Work with business units to understand Solution scope and suggest possible alternatives.
- Predictive analytics using Classification and clustering principles: LR, RF, NB, SVM, GBT, K-Means.
- Sentiment Analysis, Text mining, Supervised and Un Supervised Machine Learning.
- Real-time in-memory and Batch processing solutions implemented for Billing and Vision data sources.
- EDW migration/ETL to Big Data lakes.
- Real-time streaming with Kafka, and Sqoop ingestion tools for batch processing.
- Transformations & Predictive models in Spark/Scala, Data Warehouse in Hive/Cassandra.
- Indexing on Solr and Visualizing on Tableau & Banana/kibana.
- Extensive Hands-on experience with big data ecosystem tools.
- End-to-end implementation, Scheduling jobs, Deployments, and providing metrics.
- Coordinating/Tracking of implementation teams/tasks, Data Scientists, and business users
- Optimized DBT data models for efficient transformations and loading, while managing and securing the Snowflake Data Warehouse.

Environment: Hadoop, YARN, Spark/Scala, Java, Teradata, Hive, Snowflake, Cassandra, Kafka, Spark MLlib, Solr, Banana, R, Machine Learning, Tableau, Oozie, Knox, Kerberos, Ldap, Linux, AWS, Agile, Sqoop, Hive, HBase, Agile methodology, Teraforms.

ZENMONICS PVT LTD, Hyderabad, India **February 2019 – March 2021**

Role: Data Engineer

Responsibilities:

- Work with business units to understand project scope, suggest possible alternatives, and provide detailed design.
- Data migration: Real-time in-memory and Batch processing solutions implemented.
- Spark/Python, TDCH, and Sqoop ingestion tools for batch.
- Extensive Hands-on experience with big data echo system tools.
- End-to-end implementation, Scheduling jobs, Deployments, and providing metrics.
- Coordinating/Tracking of implementation teams/tasks and business users.
- Implemented agile process, Scrum techniques, and Story reviews.
- Designed and developed complex queries using Hive and Impala for a logistics application.
- Using Informatica, developed sophisticated SQL queries and scripts to extract, aggregate, and validate data from MS SQL, Oracle, and flat files and put it into a single data warehouse repository.
- Configure and monitor resource utilization throughout the cluster using Cloudera Manager, Search, and Navigator.
- Using Apache Flume to collect and aggregate huge volumes of log data, then stage the data in HDFS for later analysis.
- Used relational and non-relational technologies such as SQL and NoSQL to create data warehouse, data lake, and ETL systems for processing transforming a business demand into a technical design document.

Environment: Power Designer, ER Studio Cloudera Hadoop, pySpark, Python, Scala, Java, Teradata, Hive, Oozie, Linux, Kafka, GitLab, PyCharm, Hadoop, AWS S3, Tableau, Impala, Flume, Apache Nifi, Shell-scripting, SQL, Sqoop, Oracle, SQL Server, HBase, PowerBI, Linux, Agile Methodology.