Sai Keerthana Gudimetla
Data Engineer

Location: TX | **Email ID:** saikeerthana.gk@gmail.com | **Phone:** 669-295-9595 | LinkedIn

Summary

- Over 4 years of professional Data Engineer experience with a strong focus on Big Data, Hadoop Ecosystem, and Cloud Engineering.
- Advanced proficiency in AWS, including EC2, S3, EMR, RDS, and more, for effective resource management.
- Skilled in constructing data pipelines using Azure Data Factory and Databricks, managing Azure SQL Data Lake, Database, and Data Warehouse access.
- In-depth experience with Azure's HDInsight, Stream Analytics, Active Directory, Blob Storage, and Cosmos DB.
- Competent in Hadoop platforms like Cloudera, Amazon EMR, Azure HDInsight, and Hortonworks.
- Adept at managing large-scale data using Kafka, Storm, and automation via Oozie and Airflow.
- Profound understanding of PySpark components for application development, including PySpark SQL, Data Frames, and PySpark Streaming.
- Specialized in Kafka streaming data processing and storage in HDFS via PySpark.
- Extensive knowledge of SQL and NoSQL databases (Cosmos DB, MongoDB, HBase, Cassandra) for data modeling, pipeline creation, and disaster recovery.
- Experienced in Python (PySpark), and PySpark-SQL scripting for various data formats.
- Proficient in data analysis with HiveQL, Pig Latin, and MapReduce for performance enhancement.
- Comprehensive skills in data management, covering acquisition, warehousing, processing, and transformation.
- Skilled in MongoDB cluster management, including document growth monitoring and storage estimation.
- Hands-on with Kubernetes cluster creation using cloud formation templates and PowerShell for cloud automation.
- Proficient in developing scalable APIs, ETL solutions, and integration using Informatica.
- Experienced in Jira, Remedy, Git, SVN, and engaging with Agile Scrum and Waterfall methodologies.
- Skilled in legacy application cloud migration using DevOps tools like GitHub, Jenkins, and Docker.
- Collaborative in working with stakeholders, support teams, and engineering units to bolster analytics platforms and data-driven strategies.

Technical Skills

Programming Languages:	Python, R, J2EE, SQL, Pig Latin, HiveQL, Unix Shell Scripting, DotNet
Databases:	MS-SQL SERVER, Oracle, MS-Access, MySQL, Teradata, PostgreSQL, DB2, Snowflake.
Big Data Technologies:	Yarn, MapReduce, Pig, Hive, HBase, Cassandra, Oozie, Apache PySpark, Kafka.
Hadoop Distributions:	Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP, Amazon EMR (EMR, EC2, EBS, RDS, S3, Glue, Elasticsearch, Lambda, Kinesis, SQS, DynamoDB, Redshift, ECS) Azure HDInsight (Databricks, Data Lake, Blob Storage, Data Factory, SQL DB, SQL DWH, Cosmos DB, Azure DevOps, Active Directory).
NoSQL Database:	Cassandra, MongoDB.
Reporting Tools/ETL Tools:	Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI.
Methodologies:	Agile/Scrum, Waterfall.
Operating Systems:	Windows, Macintosh, Linux, Ubuntu, Unix.
Others:	Machine learning, NLP, Stream Sets, Spring Boot, Jupyter Notebook, Docker, Kubernetes, Jenkins, Jira.

Work Experience

Data Engineer **Aug 2023 - Current**
JPMorgan Chase & Co, TX

- Designed and maintained a robust ETL framework in PySpark for daily data processing, including error handling and logging, resulting in increased efficiency.
- Pioneered PySpark applications utilizing PySpark and PySpark-SQL to effectively extract, transform, and integrate data into SQL databases.
- Built data pipelines in PySpark leveraging Azure Data Factory and PySpark SQL within Azure Databricks for proficient handling of diverse data formats.
- Utilized SQLpy queries to generate insightful reports, enabling data-driven decision-making.
- Leveraged PySpark SQL to manipulate data within SQL databases, streamlining data pipeline operations.
- Developed Python scripts for ingesting data into Azure platforms (Data Lake, Storage, SQL) to optimize processing in Azure Databricks.
- Enhanced organizational security by implementing Brinqa for vulnerability risk management and improved risk assessments.
- Managed Azure Data Lake and Azure SQL by creating tables and crafting queries on Azure SQL servers, fostering improved data management.
- Extracted data efficiently from neo4j databases using neo4j in conjunction with PySpark, ensuring a vital component of the ETL pipeline.
- Maintained pipeline reliability through close monitoring of Splunk dashboards for failures and log analysis for resolution.
- Created impactful reports and dashboards using Tableau visualizations to facilitate clear data interpretation and presentation.
- Developed detailed summary reports and dashboards using comprehensive Tableau bar graphs and scatter plots, leading to better data insights.
- Optimized development workflows by implementing Looper and Concord for continuous integration and deployment.
- Automated data processing workflows by demonstrating expertise in Linux commands and scheduling them using crontab.

Data Engineer (Intern) **Jan 2023 - Jul 2023**
McKesson, TX

- Spearheaded the development of ETL pipelines using SQL for data warehousing and reporting, improving data accuracy and accessibility.
- Orchestrated S3 bucket creation and management with stringent IAM roles policies to reinforce data security and access control.

- Authored PySpark code for AWS Glue jobs and EMR, streamlining data processing and integration within the AWS environment.
- Designed and executed an efficient ETL process in AWS Glue using PySpark to migrate data from various sources (S3, RDBMS) to AWS Redshift, complemented by Athena for streamlined reporting.
- Demonstrated expertise in AWS Redshift by engineering ETL jobs that enhanced data extraction and loading processes.
- Played a key role in constructing ETL pipelines using AWS Data Pipelines to facilitate seamless data transfer and processing.
- Implemented Apache Airflow for superior orchestration of data pipelines, improving scheduling, monitoring, and workflow authoring.
- Established a comprehensive monitoring framework using CloudWatch for Lambda functions, Glue Jobs, and EC2 hosts to ensure optimal system performance and reliability.
- Designed and implemented 30+ data pipelines to transform live data into Azure Data Factory pipelines, performing ETL operations using Snowflake.
- Developed serverless orchestration for data flow between S3, Redshift, and RDS using Lambda and Step Functions, streamlining data processing.
- Employed Step Functions for seamless integration and orchestration of Glue jobs, Lambda functions, data pipelines, and data warehouse.
- Managed MongoDB databases on the cloud, designing efficient database structures and executing MongoDB queries for optimal database operations.

Data Engineer **April 2018 - Jan 2021**
Cipla, India

- Agile Champion: Implemented Agile methodologies in data analysis projects, resulting in a 30% faster adaptation to changing requirements and a 20% boost in project completion rate.
- Enhanced Decision Making: Developed interactive dashboards and reports using Power BI, leading to a 25% increase in stakeholder engagement and a 40% faster decision-making process.
- Statistical Expertise: Utilized R for advanced statistical analysis and graphics, improving data analysis efficiency by 40% and generating more precise data-driven strategies.
- Data Visualization Master: Skilled in crafting clear and informative visualizations using Matplotlib, Seaborn, and Tableau (Desktop & Server). This improved comprehension of key performance metrics, leading to a 30% increase in the accuracy of data-driven decisions and a 25% rise in stakeholder interaction.
- Production Analytics Hero: Effectively supported all production analytics, contributing to a 40% decrease in system downtime and a 35% improvement in overall system reliability.
- Database Optimization Expert: Proficient in managing MySQL schema objects and converting user requirements into technical solutions. Achieved a 30% improvement in database performance and a 25% increase in user satisfaction through efficient data structures.

Projects

Real estate home price prediction using Machine Learning

- Developed a machine learning model to predict real estate prices with an accuracy of 85%
- Analyzed factors like area, location, number of bedrooms, bathrooms, year built, and proximity to amenities to improve prediction power.
- Utilized data cleaning techniques like handling missing values, outliers, and normalization to ensure model effectiveness.
- Trained a model using a Random Forest Regression, XGBoost algorithm and hyperparameter tuning to optimize performance.

Real-time Decision Making with Streaming Data from IoT Sensors

- Built a real-time data processing system for continuous data streams from IoT sensors using AWS Kinesis and S3.
- Enabled timely and accurate insights for critical decision-making by processing streaming data.
- Streamlined data ingestion, analytics, and storage for efficient data utilization..

Automated Weather Data Workflow with Apache Airflow

- Automated weather data pipeline by integrating Apache Airflow with Weather API and S3 storage.
- Facilitated seamless extraction, transformation, and secure storage of real-time weather data.
- Established a scalable and reliable workflow for readily available weather data for analysis and decision-making.

Education

University of North Texas, Denton
Master of Science in Advanced Data Analytics **Aug 2022 – Dec 2023**

NRI Institute of Technology, India
Bachelor of Technology in Computer Science **Jul 2014- Mar 2018**

Certifications

- Microsoft Certified: Azure Data Engineer Associate