Naga Sai Kalyan Mikkilineni
+1 945-217-0276 | kalyan.mikkilineni@gmail.com | LinkedIn | Dallas, Texas

PROFESSIONAL SUMMARY:

- Around **6 years** of experience in the IT industry, specializing in data mining with large structured and unstructured datasets, including data acquisition, validation, predictive modeling, and visualization.
- Implemented various solutions for big data analytics, cloud data engineering, data warehousing, data mart, data virtualization, and data quality across multiple projects.
- Proven experience as a Data Engineer working on **Amazon cloud services, Big Data/Hadoop applications,** and **product development.**
- Expertise in data ingestion, transformation, and integration using tools such as **BigQuery, Cloud Run, and Qlik**
- **(Attunity)**
- Extensive hands-on expertise with AWS services such as **EC2, S3, Glue, Athena, DynamoDB,** and **Snowflake**.
- Designed and deployed **monitoring, metrics, and logging systems** on AWS, ensuring optimal performance.
- Experience in **MongoDB** cluster deployment, management, and development for scalable NoSQL solutions.
- Proficient in **Docker container orchestration** using **ECS, Glue,** and **Lambda** for efficient deployment and scaling.
- Skilled in **SQL Azure, Amazon Redshift, RDS**, and other cloud databases and data warehouses.
- Strong Experience in working with Databases **like Oracle 10g, DB2, SQL Server 2008** and **MySQL** and proficiency in writing complex SQL queries.
- Migrated **Cassandra** and **Hadoop clusters** on AWS, optimizing data processing strategies.
- Strong SQL development skills, including writing **Stored Procedures, Triggers,** and **Views**.
- Proficient in **Databricks, Spark-Scala, PySpark**, and **MapReduce** for unified data analytics.
- Developed **ETL** applications using **SSIS/DTS** Packages for large data volumes.
- Configure Jenkins to build **CI/CD pipeline** which includes triggering auto builds, auto promote builds from one environment to another, code analysis, auto version etc. for various projects.
- Proficient in **Tableau, Power BI,** and other data visualization tools for reporting and dashboards.

EDUCATION:

The University of Texas at Dallas, Dallas **Dec 2023**
Master of Science, Business Analytics, 3.5
• Coursework: Advanced Statistics, Predictive Analytics, Prescriptive Analytics, Time Series and Econometrics, Data Visualization, Big Data.

Sathyabama Institute of Science and Technology, Chennai **May 2018**
***Bachelor of Engineering, Mechanical Engineering,* 3.4**

CERTIFICATIONS:

Python Essentials, AWS Cloud Practitioner, and SQL.

TECHNICAL SKILL:

DBW Tools	Business Objects XIR2 and XI 3.0, XI 3.1, BI 4.0 / 4.1
Programming Languages	Python, Java, C++, SQL, Scala, Oracle, GraphQL, CloudFormation
Dashboard Tools	Tableau 8.x, 9.x, Power BI, Jupyter Notebooks, Qlik Sense, SAS, MATLAB
Frameworks/Platforms/Cloud	AWS, GCP, Docker, Flask, Django, Angular, Kubernetes, Helm, API Gateway, Microservices
Environment	Linux, Windows 2008, windows 2003, XP, Jenkins, CI/CD environment
Database	Oracle Exadata, MongoDB, NoSQL, Cassandra
Machine Learning	Scikit-learn, KNN, Random Forest, K-Means Clustering, Logistic Regression, Regression Analysis, Statistical techniques

Other Tools	HADOOP (HDFS, SQOOP, PIG, HIVE, HBASE), HPQC, WINSCP, DataDog, Git, Apache Kafka, AWS Kinesis, Apache Spark, Jupyter Notebook, Terraform

PROFESSIONAL EXPERIENCE:

Client: Meta Bank **May 2023 – Till date**
Data Cloud Engineer with AWS

- Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
- Hands-on experience with CI/CD practices and tools such as Jenkins, Terraform, and Gitlab, ensuring seamless integration and continuous delivery of software, leading to a 15% reduction in deployment time.
- Developed and maintained robust data pipelines using DBT, Airflow, and Kafka, ensuring timely and error-free data flow across various financial systems, improving operational efficiency by 30%.
- Automated 80% of data ingestion pipelines by developing and integrating RESTful APIs, reducing manual intervention and enhancing data flow consistency across systems.
- Configured and managed replica/snapshot settings in the pipeline to handle CDC files in GCP storage buckets.
- Leveraged AWS (Lambda, Glue, EC2, SNS, RDS) to build and scale secure, high-performing data infrastructure, achieving a 40% improvement in data processing times and enhancing system reliability.
- Set up security views in BigQuery to control and restrict access to sensitive data, ensuring compliance with data governance policies.
- Implemented analytics applications using multiple database technologies, such as relational, multidimensional (OLAP), key-value, document, or graph.
- Engineered high-performance data pipelines using DataFlow and DataProc, leading to a 40% reduction in batch processing times and enabling real-time analytics for business teams.
- Built scalable, API-driven data ingestion pipelines, leveraging Google Cloud Storage (GCS) to process and store over 500GB of data daily.
- Implemented automated unit and integration tests using PySpark, reducing testing time by 40% and ensuring data accuracy in ETL workflows.
- Utilized Git for version control, promoting collaboration across the data engineering team of 15+ and ensuring seamless integration of new features and updates into the data architecture.
- Developed user-friendly UI screens using ReactJS and NodeJS, improving data visualization efficiency by 25% for business intelligence teams.
- Conducted data validation and testing to ensure accuracy and reliability of data pipelines and ETL processes.
- Integrated Jenkins with Kubernetes clusters (AKS/EKS) to facilitate continuous integration and delivery, enhancing deployment efficiency by 35%.
- Developed Python-based data pipelines integrating with Airflow and Cloud Composer, reducing processing time by 25% for large-scale ETL processes.
- Designed and implemented distributed data processing workflows using PySpark and HDFS, improving data processing speeds by 50% for datasets exceeding 1TB.

Mordor Intelligence Private Limited, India **January 2020 - December 2021**
Data Engineer

- Leveraged cloud-based ETL pipelines using BigQuery and Cloud SQL, improving query performance by 40% and reducing costs by 15% through optimized resource management.
- Streamlined API calls to automate the validation of data flows between systems, reducing errors by 30% in real-time data processing.

- Migrated legacy on-premises data stores to BigQuery, leading to a 20% increase in query performance and scalability for analytics workloads.
- Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
- Designed and developed scalable data pipelines to ingest data from various sources into Amazon Redshift using AWS Glue, Kinesis, and Lambda functions.
- Implemented ETL processes to transform and load data into Amazon Redshift, ensuring data quality and integrity.
- Collaborated with front-end teams to integrate API-driven data streams into React-based dashboards, enhancing real-time data analytics by 15%.
- Configured and managed real-time data streams using Amazon Kinesis for efficient data processing and analytics.
- Set up and maintained security configurations, including IAM roles, policies, and data encryption, to ensure compliance with data governance and security standards.
- Developed 20+ ad-hoc reports and visualizations using AWS Athena and QuickSight, facilitating effective decision-making for key stakeholders.
- Implemented data replication strategies using AWS Database Migration Service (DMS) to ensure data consistency and availability.
- Led the design of end-to-end automated testing pipelines for APIs, increasing test coverage by 35% and improving fault detection rates by 50%.
- Automated infrastructure deployments using Cloud Shell scripts, improving deployment speeds by 30% and ensuring consistency across cloud environments.

Invesco, India **June 2018 – December 2019**
Data Engineer

- Managed Unix-based environments for deploying, configuring, and maintaining applications in production.
- Defined data contracts, and specifications including REST APIs.
- Extensive experience in data transformation and analytics using Alteryx, optimizing data workflows, and automating processes to increase efficiency and productivity.
- Responsible to build and run resilient data pipelines in production and implemented ETL/ELT to load a multi-terabyte enterprise data warehouse.
- Worked closely with Data science team and understand the requirement clearly and create hive table on HDFS.
- Communicated with cross-functional teams to streamline data pipeline monitoring and troubleshooting using Splunk and Grafana, while documenting run books in confluence, saving 10 hours per week.
- Solved performance issues in Spark with understanding of groups, joins and aggregation.
- Designed a Live Analysis dashboard in Power BI to track real-time finance data and predict customer behavior metrics and trends, revising and narrowing analytical operations by 80%.
- Optimized PySpark jobs for Hive and HDFS, reducing processing time by 30% and ensuring efficient resource utilization in a cluster environment.
- Responsible for loading data from UNIX file systems to HDFS. Installed and configured Hive and written Hive UDFs.
- Was responsible for creating on-demand tables on S3 files using Lambda Functions using Python and PySpark.
- Created End-to-end ETL pipeline for data processing for created dashboards to business using PySpark.
- Developed spark programs using python APIs to compare the performance of spark with HIVE and SQL and generated reports monthly and daily basis.
- Developed dataflows and processes for the Data processing using SQL (SparkSQL & Data frames).
- Understood business requirements and prepared design documents, coding, testing and go on live production environment.
- Involved in planning process of iterations under the Agile Scrum methodology, performed gap analysis, collected requirements, and performed data discovery and data profiling for the application.