ANKIT MALLYA

Data Engineer

Bangalore, India
+91 9137210394
ankit.mallya@gmail.com

Driven and results-oriented Data Engineer with nearly 4 years of extensive expertise in data engineering, ETL processes, and cloud technologies. Proven track record in implementing robust data pipelines using **AWS (Glue, EC2, Kinesis, Lambda, Redshift)**, **Apache Airflow** for orchestrating workflows and scheme inference with **Data Build Tool (DBT)**, real-time data streaming with **Apache Spark Streaming** and leveraging **SQL** databases for data modeling and warehousing whilst ensuring compliance with data governance standards (GDPR, HIPAA). Intermediate in machine learning techniques, statistical analysis and automation testing to deliver innovative solutions. Seeking to leverage advanced skills in a dynamic and **Data Engineer**, **Technology/IT Consultant**, or **Junior Product Manager**.

CORE COMPETENCIES

Data Engineering & Data Ingestion

ETL Processes & AWS Data Streams

Apache Airflow & Data Transformation

Data Governance & Compliance (GDPR, HIPAA)

SQL Server Data, Integration and Analytics Tools

Automation Testing & SDLC Management

Data Integration & Real-time Streaming

Data Modeling & Warehousing

Containerization & CI/CD Pipelines

Agile Methodologies & Version Control (Git, SVN)

DevOps Practices & Cloud Storage (AWS, Beginner GCP)

Big Data Technologies (Hadoop, Spark)

Database Management (SQL, NoSQL)

Scripting Languages (Python, Bash)

Business Intelligence Tools (Tableau, Power BI)

Collaborative Tools (Jira, Confluence)

Data Visualization & Predictive Analytics

Continuous Improvement & Optimization Strategies

Problem Solving & Critical Thinking Skills

PROFILE SUMMARY

- **Data Engineering Expertise**: Experienced Data Engineer with 3 years of expertise in designing, implementing, and optimizing data ingestion pipelines and ETL processes. Proficient in using AWS services (S3, Kinesis, Glue, Redshift) for scalable data solutions. Skilled in Spark Structured Streaming and Apache Airflow for large-scale data processing and workflow orchestration. Adept at real-time data streaming with JSON/Parquet files, ensuring data quality, and utilizing data lakes and data warehouses for real-time analytics.

- **Data Governance and Compliance**: Comprehensive knowledge of data governance and compliance standards (GDPR, HIPAA), ensuring data integrity and security throughout all stages of data processing.

- **Statistical and Machine Learning Proficiency**: Proficient in statistical analysis and machine learning techniques, including hypothesis testing, regression, classification, A/B testing, and advanced machine learning methodologies such as ensemble learning (e.g., Random Forests, Gradient Boosting), deep learning (e.g., Convolutional Neural Networks, Recurrent Neural Networks), and unsupervised learning (e.g., clustering, dimensionality reduction).

- **Software Engineering Practices**: Experienced in software engineering practices, including agile methodologies (e.g., Scrum, Kanban), CI/CD pipelines using tools like Jenkins and GitLab CI, and automation testing frameworks such as Selenium and PyTest. Proficient in implementing version control with Git, practicing code reviews, and utilizing containerization technologies like Docker and Kubernetes to contribute to a robust software development lifecycle management.

- **Project Management Skills**: Demonstrated ability to manage complex projects from conception to deployment, ensuring adherence to timelines, budgetary constraints, and quality standards.

- **Communication and Leadership**: Effective communicator and team player with a passion for leveraging data-driven insights to drive business decisions and optimize organizational performance.

- **Collaborative Team Player**: Proven track record of collaborating across teams to enhance data availability and streamline workflows, resulting in measurable improvements in operational efficiency.

PROFESSIONAL EXPERIENCE

Aug 2024 - Present | Coforge

KEY RESULT AREAS:

- Designed Apache Airflow DAGs to sequentially trigger AWS Glue jobs using *TriggerDagRunOperator* and *ExternalTaskSensor*, integrating with *S3KeySensor* for monitoring data availability in S3. Automated dbt model execution to refresh Snowflake target tables post-job completion.
- Implemented end-to-end ETL workflows leveraging Airflow's AWS provider packages, *boto3* for Glue job management, and *snowflake.connector* for data updates, ensuring seamless and efficient data pipeline orchestration.

SOFT SKILLS

Collaboration Skills

Time Management Skills

Interpersonal Proficiency

Leadership Competence

Decision Maker

Problem-Solving Acumen

EDUCATION

- **2022 | Master of Science - Analytics** | Harrisburg University of Science and Technology, Harrisburg PA | GPA: 3.9/4.0
- **2019 | Bachelor of Arts - Economics** | Miami University, Oxford OH | Lambda Chi Alpha Fraternity: Treasurer, Phi Chi Theta Professional Business Fraternity, Commercial Banking Club

TECHNICAL SKILLS

- **Programming Languages & Databases**: Python, Java, R, SQL (Advanced), Snowflake, Redshift, BigQuery
- **ETL Tools**: Apache Airflow, DBT
- **Cloud Platforms**: AWS
- **Data Versioning & Management**: GitHub, GitLab
- **Data Governance & Compliance**: GDPR, HIPAA
- **Statistics & Machine Learning**: Hypothesis Testing, A/B Testing, Bayesian Statistics, Multi-Armed Bandit, Regression, Classification, Dimensionality Reduction

PROJECT DETAILS

- **Honda Capstone**: Researched Honda's 'Fuel and Brake Pipe' supplier ecosystem, developed supplier scorecards, conducted SWOT analysis, and performed forecasting analysis. Recommendations for new partnerships were adopted by Honda to enhance their lean supply chain management.
- **Optimizing SEC Data for Stock Market Prediction**: Analyzed variables influencing stock sell or pass indices and company revenue. Utilized big data techniques and data visualization in R (GGPlot, PCA, Regression Analysis) to demonstrate findings.

June 2021 - March 2024 | Sisense

KEY RESULT AREAS:

October 2023 - March 2024 | Data Engineer

- Spearheaded the implementation of a robust AWS-based data ingestion pipeline, optimized JSON payload transformation, and orchestrated seamless data flow from S3 to RDBMS using AWS Data streams, Firehose, Lambda functions, dbt, and Airflow.
- Facilitated the decommissioning of Gainsight data using Upsolver, employed Boto3 scripts and S3 operators in Airflow for daily data transfers, and updated final tables in the relational database.
- Developed a real-time data pipeline using Spark Streaming to process nested JSON and Parquet files, enabling dynamic schema handling and efficient transformation of incoming data**.** Applied advanced transformations like flattening deeply nested structures, aggregating time-series data, and performing join operations with reference datasets for enriched analytics. Data was then written back to S3 in a partitioned format, supporting downstream analytics and querying.
- Optimized the Spark Streaming pipeline by tuning batch intervals and utilizing checkpointing to ensure fault tolerance and seamless recovery. Applied key transformations such as filtering, aggregation, and deduplication to process incoming data efficiently while maintaining low processing latency.
- Streamlined storage integration, Snowpipe, tables, and models through YAML-driven macros in dbt, automated processes, and reduced manual efforts.
- Implemented AWS event notifications for real-time data streaming from the data lake to the database, ensured updates, and maintained data integrity.

January 2021 - October 2023 | Software Engineer

- Developed cloud-native automation solutions using AWS, Terraform as IaC and Databricks, enabling a 15% increase in data pipeline efficiency across multiple reporting dashboards. These solutions streamlined data processing and increased visibility into system performance, leading to optimized resource utilization.
- Designed and implemented a comprehensive automation testing suite using JAVA, Selenium, Cucumber, and Python, achieving 95% test coverage and reducing referential integrity errors by 85%.
- Developed monitoring dashboards within Databricks to track PySpark job performance and data lineage, enabling faster identification of data bottlenecks and ensuring consistent data quality across workflows.
- Assisted in developing and enhancing the automation suite for the Notebooks feature using JAVA, Selenium, and Python, which had reduced translation time by 75% and mitigated referential integrity errors by over 80%.
- Engineered Image/Document classification, SVM's, Time-Series, Sentiment Analysis, and Fourier Transform samples for the Playgrounds initiative using Python and R.
- Conducted code quality assurance activities, including identifying and resolving code conflicts using JAVA, TestRail, and Selenium.
- Collaborated with the Product team to create a community blog series titled "From BI to AI using Sisense."

October 2019 - June 2020 | 3M HIS (Contract) | Data Systems Analyst

KEY RESULT AREAS:

- Modeled and analyzed healthcare data, encompassing inpatient, outpatient, professional, and pharmacy claims.
- Collaborated with internal client teams to structure, analyze, and interpret data requests.
- Teamed up with the research team to assess analytic processes and integrate them into new and existing products.
- Developed data visualizations using graphs, charts, and tables to effectively communicate findings to a lay audience.
- Executed SQL queries utilizing joins, subqueries, and indexing to extract and manipulate data.
- Offered analytic support for clients and attended on-site client meetings as needed.
- Conducted data quality assurance activities, identifying and resolving data errors to ensure accuracy and reliability.