Pallavi Yellisetty

☎ +1 8177933220 📧 pallaviyellisetty@gmail.com

🔗 Pallavi Yellisetty | LinkedIn 🌐 My portfolio: https://lucent-dolphin-4da8a4.netlify.app/

Professional Summary

Data Engineer with 3 years of experience in building and optimizing data pipelines, ETL processes, and scalable architectures for big data platforms. Proficient in Apache Spark, Python, Scala, and AWS, with a strong focus on improving data quality, system efficiency, and delivering actionable insights. Adept at collaborating with cross-functional teams to implement solutions that meet business goals and technical requirements.

Technical Skills

Programming Languages	Python, Scala, Java, SQL
Big Data Tools	Apache Spark, Hadoop, Hive, Kafka, HDFS
ETL & Workflow Management	Apache Nifi, Airflow, AWS Glue
Databases	PostgreSQL, Snowflake, MySQL, MongoDB, Redshift
Cloud Platforms	AWS (S3, EMR, Redshift, Lambda, Glue), Azure
Data Analytics & Visualization	Tableau, Power BI
DevOps Tools	Docker, Jenkins, Git
Frameworks & Libraries	PySpark, NumPy, Pandas

Experience

Data Engineer

Virtusa-Hyderabad, India **Jul 2022 – Jul 2023**

- Developed and optimized distributed data pipelines using Apache Spark and Scala, processing over 5TB of data daily.
- Designed ETL workflows in AWS Glue, reducing data processing times by 40% and improving scalability.
- Built and managed data models in Snowflake and PostgreSQL, enhancing query performance by 30%.
- Implemented real-time data ingestion pipelines with Apache Kafka and Spark Streaming to support time-sensitive analytics.
- Collaborated with the data science team to preprocess datasets, boosting ML model accuracy by 25%.
- Optimized Spark jobs by fine-tuning configurations, reducing execution time by 35% and lowering cloud compute costs.
- Implemented data validation frameworks using Apache Airflow and AWS Lambda, ensuring 99.9% data accuracy across pipelines.

- Developed reusable ETL frameworks, reducing code redundancy by 60% and accelerating new pipeline deployments.
- Automated data quality checks using Great Expectations and Apache Iceberg, leading to a 50% decrease in data anomalies.
- Designed scalable microservices to support data transformation, leveraging AWS Lambda and API Gateway for seamless integration.
- Led the migration of on-premise ETL pipelines to AWS, improving resilience, security, and reducing operational costs by 45%.
- Enhanced query performance in Snowflake by implementing clustering and materialized views, achieving 2x faster report generation.
- Developed CI/CD pipelines for data engineering workflows using GitHub Actions and Terraform, ensuring zero downtime deployments.
- Optimized Kafka consumer performance, handling peak loads of 1M+ events per second with minimal latency.
- Integrated logging and monitoring with Prometheus and Grafana, improving issue detection and reducing debugging time by 40%.

Data Engineer Intern
Altimetrik – Hyderabad, India May 2021 – May 2022

- Designed and implemented ETL pipelines for IoT sensor data integration into PostgreSQL, improving data accessibility for real-time analytics.
- Optimized PostgreSQL indexing and partitioning, enhancing query performance by 40% for time-series IoT sensor data.
- Developed scalable data ingestion workflows using AWS Lambda, AWS Glue, and S3, reducing pipeline costs by 20%.
- Engineered real-time streaming solutions with AWS Kinesis, Apache Kafka, and Spark Streaming, reducing ingestion latency by 50%.
- Automated data preprocessing and cleansing with Python (Pandas, NumPy) and PySpark, reducing missing data by 35% and improving dataset quality.
- Designed a robust data lake architecture leveraging AWS S3, AWS Glue, and Amazon Athena, reducing storage costs by 30% while improving accessibility.
- Implemented anomaly detection algorithms for IoT sensor data using Apache Spark and Python, enhancing operational insights and reducing error rates.
- Developed Airflow DAGs to schedule, monitor, and optimize ETL workflows, improving data pipeline efficiency by 35%.
- Integrated data validation and quality checks using Great Expectations and AWS Step Functions, ensuring 99.9% data accuracy.
- Enhanced monitoring and observability using AWS CloudWatch, Grafana, and Prometheus, maintaining 99.99% uptime for mission-critical data pipelines.
- Implemented schema evolution strategies for evolving IoT device data, ensuring seamless adaptability and reducing schema-related failures.

- Optimized S3 storage with lifecycle policies, compression, and partitioning, cutting storage overhead by 20%.
- Developed CI/CD pipelines for ETL workflows using GitHub Actions and Terraform, ensuring zero-downtime deployments.

Software Engineer - Intern
Genpact -Hyderabad, India Jan 2020-May-2020

- Built a scalable data ingestion pipeline for processing 2M+ cybersecurity log events daily using Apache Kafka.
- Utilized Apache Hive and Spark SQL to query and analyze large datasets, improving query execution by 40%.
- Conducted data quality validation and implemented automation scripts, ensuring consistent and accurate data processing.
- Collaborated with cross-functional teams to deliver insights, supporting threat detection and prevention systems.

Education

Master's in Computer Science
University of Texas at Arlington – Arlington, TX
Graduation: May 2025

Certifications

AWS Certified Solutions Architect – Associate
Databricks Certified Associate Developer for Apache Spark
Google Professional Data Engineer Certification

Projects

Fraud Detection with Neural Networks
Developed an algorithm in Python to detect fake profiles across social networks, increasing fraud detection accuracy by 50%.
Trained neural network models with advanced preprocessing techniques, improving dataset reliability and scalability.
Biometric-Based Secure Access
Designed a biometric-based authentication system for cloud services, reducing unauthorized access by 50%.
Implemented backend systems using Python and integrated encryption protocols for enhanced data security.

Key Achievements

Reduced ETL processing costs by 20% through optimization techniques in AWS Glue.
Boosted query performance by 30% in Snowflake and PostgreSQL by redesigning data models.
Delivered real-time analytics with Kafka and Spark, reducing latency for critical business insights.