Shriyansh Singh

+1 930 333 5141 | shriyansh.singh24@gmail.com | linkedin.com/in/shriyansh-bir-singh

SUMMARY

Aspiring Data Engineer with hands-on experience in building scalable data pipelines, optimizing ETL processes, and deploying real-time data architectures using Apache Spark, Kafka, and cloud platforms such as AWS and Google Cloud. Proficient in Python, SQL, and Scala with a strong foundation in big data technologies and cloud-native solutions.

PROFESSIONAL EXPERIENCE

Machine Learning Intern *Apr 2024 – Present*

Hyphenova AI *Los Angeles, California*

- Designed and deployed a scalable data pipeline using Apache Kafka and Spark Streaming, processing over 500,000 transactions per second with sub-second latency, leveraging partitioning and fault-tolerance mechanisms to enhance detection speed and accuracy by 30%.
- Built automated data pipelines using SQL and Python to extract and transform data from various sources into a centralized data warehouse, supporting Power BI dashboards that tracked key performance metrics and increased reporting efficiency by 25%.
- Enhanced data pipeline robustness by implementing automated data validation and quality checks using PySpark and Delta Lake, implementing schema validation and anomaly detection on datasets exceeding 10TB, improving data reliability by 20% and ensuring accurate downstream analytics.
- Architected and optimized a scalable data pipeline using Apache Spark on AWS (EC2, S3, and Lambda), processing 10+ million records daily, reducing processing time by 40%, and cutting operational costs by 15%, thereby improving analytics capabilities and decision speed.
- Developed a real-time data quality framework using Apache Airflow and Python, incorporating checks for data completeness, consistency, and conformity, which enhanced data reliability by 30% and minimized analytics errors by 20%, thus maintaining data integrity across pipelines.

Data Engineering Intern *May 2022 – Oct 2022*

Enterprise Business Technologies Pvt. Ltd *Mumbai, India*

- Built automated data pipelines using SQL and Python to extract and transform data from various sources into a centralized data warehouse, supporting Power BI dashboards that tracked key performance metrics and increased reporting efficiency by 25%.
- Engineered ETL processes with Python and Apache Airflow to automate data integration from SQL databases and APIs, improving data accuracy by 15% and reducing refresh time by 40%, enabling scalable and reliable data ingestion into the enterprise data warehouse.
- Developed scalable data pipelines using Apache Airflow and Python, handling over 2 million records daily for predictive analytics, which improved data processing speed by 30% and enhanced forecasting accuracy, contributing to a 10% increase in revenue.
- Implemented data validation and anomaly detection within ETL pipelines using Python and SQL scripts, which reduced data errors by 20% and ensured the integrity of data for critical business operations, enhancing the reliability of downstream analytics.

EDUCATION

Indiana University Bloomington Indiana, United States

Master's of Science in Data Science *Aug 2023 – May 2025*

- Relevant Courses: Information Visualization, Data Mining, Applied Machine Learning, Statistics, Big Data Applications, Cloud Computing, Graph Analytics, Applied Database Technologies, Intelligent Systems

University of Mumbai Maharashtra, India

Bachelor's of Engineering in Electronics *Aug 2019 – May 2023*

PROJECTS

Fraud Detection in Financial Transactions | *Python, XGBoost, Apache Spark* *Jan 2024 – Apr 2024*

- Engineered a real-time data pipeline using Apache Kafka and Spark Streaming, processing over 500,000 transactions per second with sub-second latency, leveraging partitioning and fault-tolerance mechanisms to enhance detection speed and accuracy by 30%.
- Developed scalable data ingestion pipelines using Kafka and AWS Glue to preprocess transactional data for anomaly detection, increasing data throughput by 50% and ensuring system scalability, which improved reliability and customer satisfaction by 40%.
- Optimized data processing pipelines with Apache Spark by implementing data partitioning and in-memory computations, which improved processing efficiency by 40% and scaled the system to handle a 3x increase in transaction volume, reducing performance bottlenecks.

Customer Churn Prediction for Telecom Industry | *TensorFlow, Keras, AWS SageMaker* *Aug 2023 – Oct 2023*

- Designed and implemented a data pipeline using Apache Airflow to automate the ingestion and preprocessing of customer

- Developed a data integration platform with Apache NiFi and Kafka, consolidating over 5 million customer records from CRM, social media, and web analytics into a unified Snowflake data warehouse, enhancing data accessibility and integration reliability.
- Automated data ingestion and transformation workflows using Apache NiFi, integrating ETL processes with Snowflake to streamline data flow, which increased data accessibility by 60% and reduced manual processing errors by 25%.
- Implemented comprehensive error handling and data validation using Apache NiFi, which ensured consistent data quality across integrated sources, reducing data discrepancies by 20% and improving the accuracy of downstream analytics.

Log Data Processing Pipeline for Security Analytics | *ELK Stack, AWS S3, Kibana* *Project*

- Developed a real-time log processing pipeline using the ELK Stack, ingesting over 2 TB of log data daily for security event monitoring, with optimized Logstash configurations to enhance data parsing speed and accuracy.
- Configured Logstash pipelines with custom grok patterns and filtering rules, reducing irrelevant log data by 30% and enhancing focus on critical security alerts, thereby improving incident response times by 20%.
- Developed interactive Kibana dashboards for real-time visualization of security threats, including alert trends and anomaly detection, which enabled faster detection and a 25% reduction in response times to potential breaches.

Scalable ETL Framework for Financial Data | *Apache Beam, Google Cloud Dataflow, BigQuery* *Project*

- Designed a scalable ETL framework using Apache Beam and Google Cloud Dataflow, processing over 100 million financial transactions daily with end-to-end latency under 2 seconds, leveraging dynamic resource scaling and optimized data partitioning.
- Implemented robust data validation and cleansing using custom Apache Beam transforms, which ensured data accuracy and completeness, reducing processing errors by 30% and supporting high-quality inputs for financial analytics.
- Integrated Google BigQuery with ETL pipelines for high-performance querying and analytics, utilizing clustered tables and partitioning strategies to enhance data throughput by 50% and reduce query costs by 20%.

Predictive Maintenance Data Pipeline for Manufacturing | *Apache Kafka, Spark MLlib, AWS Redshift* *Project*

- Engineered a predictive maintenance pipeline with Apache Kafka and Spark MLlib, processing 10 million sensor events daily, leveraging feature extraction and ML model training to reduce maintenance costs by 20% and downtime by 25%.
- Deployed ML models for failure prediction using Spark MLlib and integrated them into streaming pipelines with Kafka, enabling real-time maintenance alerts that reduced downtime by 25% and maintenance costs by 20%.
- Integrated predictive maintenance pipeline with AWS Lambda for event-driven processing and Redshift for scalable data storage, enabling automated insights and reducing response times by 30% through seamless data orchestration.

Automated Data Quality Monitoring System | *Great Expectations, Apache Airflow, AWS S3* *Project*

- Developed an automated data quality monitoring system with Great Expectations and Apache Airflow, running over 100 validation checks per pipeline to ensure completeness, accuracy, and consistency, reducing data errors by 30%.
- Configured Airflow DAGs for automated scheduling of data quality checks, optimizing resource use and reducing manual validation effort by 40%, while maintaining high standards of data integrity across pipelines.
- Implemented real-time alerting with Slack and email integrations within Airflow, enabling immediate notification of data quality issues, which improved response times by 50% and ensured swift corrective actions to maintain data integrity.

SKILLS

Programming Languages: Python, SQL, Scala, Java
Data Engineering Tools: Apache Spark, Apache Kafka, Apache Airflow, Hadoop
Cloud Platforms: AWS (S3, Lambda, Glue, Redshift), Google Cloud (BigQuery, Dataflow), Azure (Data Factory, Synapse Analytics)
Big Data & Databases: NoSQL (MongoDB, Cassandra), Relational (PostgreSQL, MySQL), Data Warehouses (Snowflake, BigQuery, Redshift)
Containerization & DevOps: Docker, Kubernetes, Terraform, Git
Data Visualization: Power BI, Tableau, D3.js
Data Quality: Great Expectations, Deequ