Yash Raj Singh

Leetcode Global Rank 4K • IEEE / Springer Researcher

Experienced AI Developer with 8+ years of expertise in data engineering, machine learning, and cloud-based data processing. Strong background in data profiling, feature engineering, and model development to extract insights from large datasets. Proven ability to build scalable data pipelines, optimize ML workflows, and leverage AI techniques for business intelligence.

- Analyzed large datasets in Snowflake, identifying key fields (e.g., location, employee count, revenue, industry keywords) for company profiling.
- Used Snowpark, AWS (S3, Redshift, EMR) to process structured and semi-structured data efficiently.
- Designed data pipelines and ETL workflows to ensure clean, high-quality data for ML models.
- Built classification and clustering models using Hugging Face, PyTorch, and Scikit-learn to group companies based on business attributes.
- Implemented nearest neighbour and similarity score algorithms for inter-dataset transferring of insights.
- Engineered ML powered company categorizations and predictive analysis to better business outcomes
- Designed scalable pipelines using Apache Spark, Airflow, and Python for ingestion transformation, and feature engineering.
- Fully automated workflows at real-time, batch levels on data with very minimal manual inputs, enhancing overall system efficiency;
- Used AWS Lambda and Redshift for scalable and real-time-based ML inference reports.
- Extensive experience with AWS services (S3, Redshift, Lambda, EMR) for handling large-scale data processing.
- Optimized data storage and retrieval using Snowflake and AWS Redshift, improving query performance and reducing costs.
- Managed and processed unstructured and structured data from multiple sources, ensuring smooth data operations.
- Developed AI-driven solutions, including text classification, NLP-based keyword extraction, and similarity detection.
- Built APIs integrating ML models to automate business insights and decision-making.
- Experimented with text-to-image and text-to-audio models to explore AI applications.

Skills

- **Cloud Platforms**: AWS (S3, Redshift, EMR, Glue, RDS), GCP (IAM, MySQL, SecretManager)
- **Principles**: TDD, Design Pattern, SOLID
- **Tools**: Apache Spark, Hadoop, Kafka, MLlib, SnowFlake, Pytorch, sklearn.
- **ETL Frameworks**: Apache Airflow, Apache Spark
- **Programming Languages**: Python, SQL, Java, C/C++, Nodejs.
- **Database Systems**: MySQL, PostgreSQL, MongoDB.
- **Monitoring and Deployment**: Kubernetes, Helm Charts, Sleuth Micrometer, Jenkins, Docker.

Experience

APR 2023 – PRESENT
AI Engineer | Nextzen Minds | Singapore

DEC 2016 – FEB 2023
Senior Software Engineer | CBNITS| Portsmouth, US

Key Projects

AI Engage

- Designed a multi-stage objective flow architecture with reusable components for workflows like Research Agent, Ad Generator, and Persona Creation.
- Used TDD & A/B testing.
- Utilized TypeScript and Node.js for API development, ensuring robust and scalable endpoints.
- Created an optimized relational schema supporting objectives, stages, user data, and tags for efficient querying and extendibility.
- Leveraged MySQL for structured storage and implemented tagging mechanisms for improved discoverability.
- Integrated OpenAI APIs for tasks like keyword generation, summarization, and persona creation.
- Implemented Apache Tika for document parsing and Apache Spark for distributed processing of large datasets.
- Deployed the platform on **Google Cloud Platform (GCP)** using managed services for databases, storage, and authentication.
- Designed and implemented a tagging feature enabling users to categorize and search objectives based on relevance.
- Enhanced UX with intuitive tag-based filtering for workflows and personas.

NZM Hospital Management System

- Designed and developed **AI-powered healthcare data processing and analytics workflows** using **Python, Flask, and Apache Spark**. Focused on **scalable data pipelines, machine learning-driven insights, and seamless system integrations** to enhance hospital operations and patient care.
- Built **ETL pipelines** using **PySpark and Apache Spark** to process large-scale healthcare data, including patient records, lab results, and operational metrics.
- Optimized **data ingestion and transformation workflows** for structured and unstructured medical data using **Pandas, NumPy, and Spark DataFrames**.
- Developed **predictive models** with **Scikit-learn** to analyze patient admission trends, disease progression, and treatment effectiveness.
- Implemented **clustering and classification models** to segment patient data and optimize resource allocation.
- Built a **similarity-based recommendation system** for personalized treatment plans using **KNN and cosine similarity**.
- Designed and developed **RESTful APIs** using **Flask**, enabling secure and efficient access to healthcare data.
- Integrated **FastAPI for high-performance API endpoints**, ensuring seamless communication with hospital systems.
- Implemented **JWT-based authentication and role-based access control (RBAC)** for secure data access.
- Utilized **AWS S3** for secure storage of patient records, lab reports, and machine learning models.
- Designed **PostgreSQL and NoSQL (MongoDB) schemas** for storing structured and unstructured healthcare data.
- Implemented **Redis caching** to optimize API response times and reduce database load.
- Built **NLP models** using **Hugging Face Transformers** and **SpaCy** to extract key medical insights from doctor notes and patient reports.
- Developed **automated symptom analysis and disease risk prediction tools** using **Logistic Regression and Random Forest**.
- Implemented **TF-IDF and Named Entity Recognition (NER)** for analyzing unstructured medical text.
- Integrated **FHIR and HL7 standards** to enable seamless data exchange with **Electronic Health Records (EHRs) and Lab Information Systems (LIS)**.
- Designed **Kafka-based real-time streaming pipelines** for processing live patient data and monitoring hospital operations.
- Deployed **Celery workers** to automate background tasks like appointment scheduling, report generation, and ML inference.
- Optimized **Spark jobs** for efficient data processing, reducing execution time by **40%** through **broadcast joins and partitioning strategies**.
- Built **dashboards in Streamlit** for real-time monitoring of hospital KPIs and patient trends.

Liberty Coin – A Payment Gateway Solution

- Designed and developed a scalable, data-driven payment platform supporting multi-currency transactions, real-time fraud detection, and predictive analytics using Python, Apache Spark, Snowflake, and traditional ML algorithms.
- Migrated transactional and user data from on-premise databases to Snowflake and AWS S3, optimizing data partitioning, clustering, and compression for improved query performance and cost efficiency.

- Designed ELT pipelines in Snowflake using Snowpark and SQL transformations to streamline data ingestion and processing.
- Integrated Apache Spark and Snowflake connectors to handle large-scale payment data efficiently.
- Developed fraud detection models using Scikit-learn, XGBoost, and Random Forest, identifying suspicious transactions with 95% accuracy.
- Implemented unsupervised learning algorithms (K-Means, DBSCAN) for anomaly detection, reducing false positives in fraud alerts.
- Used Logistic Regression and Decision Trees for risk scoring of transactions, optimizing fraud detection thresholds.
- Designed a high-performance data warehouse in Snowflake, enabling real-time analytics and fraud detection at scale.
- Created materialized views and optimized queries in Snowflake for faster insights on payment trends, customer behavior, and chargeback risks.
- Implemented time-series forecasting (ARIMA, Prophet) to predict transaction volumes and detect unusual spending patterns.
- Developed real-time payment processing pipelines using Apache Kafka and Spark Streaming, ensuring sub-second latency for fraud alerts and transaction monitoring.
- Integrated Snowflake Streams and Tasks to process incremental payment data efficiently.
- Designed event-driven architectures using Flask APIs and WebSockets for instant fraud notifications and user alerts.
- Optimized Snowflake warehouse scaling policies, reducing compute costs by 40% while maintaining low query latencies.
- Implemented Spark job optimizations (broadcast joins, caching, and partition pruning), improving data processing speeds by 30%.
- Utilized AWS Lambda and Step Functions to orchestrate event-driven data workflows, minimizing infrastructure costs.

Netskope Security

- Designed and developed a scalable file processing system with advanced data security and performance optimization techniques using Azure services.
- **Spark ETL Optimization**: Enhanced Spark DAG by resolving stack/heap issues, integrating lazy loading, and using **JNI with C/C++**, improving performance by 40%.
- **Data Pipeline Orchestration**: Built robust workflows using **Spark** and **Apache Airflow** for file ingestion, validation, and sensitive data detection, reducing manual intervention by 60%.
- **Multi-Source Integration**: Developed a connector framework to integrate data from S**3**, SharePoint Online, OneDrive, and Gmail into a unified pipeline, supporting downstream processing and analytics.
- **Advanced PII Detection**: Implemented **OpenAI Service** and fine-tuned BERT models to identify and classify sensitive data across large datasets with confidence scoring and advanced relationship mapping.
- **Monitoring & Scalability**: Deployed the solution on **Kubernetes**, enabling high availability and automated scaling to handle increased file processing loads.

Connect Buds – E-Learning Platform

Engineered a scalable data pipeline for an e-learning platform connecting teachers, students, and administrators with real-time analytics capabilities.

- **Data Migration**: Migrated student and teacher data from legacy systems (Excel and MySQL) to **BigQuery** using **Apache Airflow**, enabling faster query performance and better data accessibility.
- **ETL Pipelines**: Built ETL workflows using **Apache Spark** and **Python** to process and cleanse student activity data, ensuring high data quality for analytics.
- **Recommendation System**: Designed and deployed a content recommendation engine leveraging collaborative filtering and ML models on **SageMaker**, improving user engagement by 40%.
- **Deployment**: Created helm chart and deployment script for kubernetes and docker container.

Publications

Graded Classification of Liver Cirrhosis using Machine Learning Algorithms on a Highly Unbalanced Dataset, received best paper award of the conference

Achieved an accuracy of 84.24% in the graded classification of liver cirrhosis using machine learning algorithms, including hyperparameter tuning and a stacked model approach. Employed a dataset with 18 features and 6,800 data instances, effectively handling the challenge of an unbalanced dataset. Demonstrated expertise in implementing and optimizing machine learning models, with a focus on healthcare applications. Results contribute to the advancement of accurate disease classification and underscore the potential of machine learning in improving medical diagnostics.

Performance Analysis of Machine Learning Algorithms for Prediction of Cerebral Attack (Stroke)

Conducted a study on stroke prediction using twelve ML algorithms. Random Forest (RF) achieved 97.36% accuracy, while Gradient Boost Classifier (GBC) reached 97.73% after hyper-parameter tuning. Stacking GBC with other models resulted in an impressive accuracy of 98.85%. Our findings highlight GBC as the top-performing algorithm for stroke prediction.

Education

Bachelor of Technology in Computer Science & Engineering | Techno India University | Kolkata, India