UMANG SHAH

Data Engineer – New York
umangshah97@gmail.com | umangshah.me | linkedin.com/in/shah-umang-me | github.com/umangshah-js

Summary

Data Engineer with 4+ years building scalable ETL, real-time streaming, and cloud-native platforms across fintech and healthcare.

Education

New York University – M.S. in Computer Science, GPA: 3.96/4.0   *Sep 2022 – May 2024*
VIT Chennai – B.Tech in Computer Science and Engineering, GPA: 9.16/10   *Jul 2015 – May 2019*

Experience

Analyst (Data / Software Engineering), Nomura Holdings America – New York, NY   *Jul 2024 – Present*

- Delivered 67+ end-to-end features on a 15+ year .NET C# pricing platform by modernizing payoff logic and trade-pricing workflows.
- Added buffer-coupon pricing to provide fast provisional charge estimates on incoming trades, enabling desks to act on quotes before final accurate valuation completed.
- Reduced manual trader workflows by replacing Excel-based pricing with request-driven pipelines using hybrid REST and pub-sub compute.
- Automated the 144A post-trade lifecycle with bulk import, instrument creation, booking-line management, and contract generation.
- Improved pricing accuracy and trade visibility by integrating Euclid APIs, Radar, and Bloomberg into front-to-back workflows.

Software Engineer Intern, MIO Partners (McKinsey Investment Office) – New York, NY   *Jun 2023 – Aug 2023*

- Improved ETL reliability and developer velocity by migrating ActiveBatch/Talend workflows to AWS Step Functions + S3 and building a portal for logs, trigger history, and one-click runs.

Senior Software Engineer (Data Engineering), UnitedHealth Group – Hyderabad, India   *Jul 2019 – Jul 2022*

- Standardized healthcare onboarding by building DXP pipelines ingesting flat files, CCD XML, HL7 messages, and Kafka streams into partitioned Hive Parquet datasets.
- Improved data quality and reliability through ingestion validation, Airflow orchestration, and Python/Angular controls for client scheduling and source formats.
- Delivered MTD/YTD analytics pipelines across 26M+ members on large Hive/Spark clusters; used single-query Hive for straightforward workloads and spearheaded PySpark jobs for complex transformations, tuning memory by queue and node capacity for stable throughput.
- Reduced near-real-time data SLA from 6 hours to 10 minutes by designing and deploying Kafka-based streaming pipelines.
- Reduced infrastructure cost and production downtime by migrating core services to Kubernetes.
- Improved incident detection and recovery by implementing Grafana and Prometheus monitoring/alerting for mission-critical workflows.

Skills

Data Engineering: Spark, Kafka, Hive, HDFS, Snowflake, ETL, PostgreSQL, SQL, MongoDB, HBase, Redis, KDB+
Cloud & DevOps: AWS, Azure, Kubernetes, Docker, CI/CD, Jenkins, Nginx
Web & Backend: Node.js, Flask, .NET Framework, FastAPI
Analytics & ML: Pandas, Tableau, D3.js, Chart.js, NLTK, StanfordNLP
Programming & Version Control: Python, JavaScript, Java, C#, Git

Projects

Media Magic (Live Demo)   *Sep 2023 – Dec 2023*

- Built a developer-first image delivery platform on AWS (S3 → SQS → Lambda → DynamoDB) plus SDK/CLI tooling, enabling path-based integration and sub-200ms image responses.
- Improved visual consistency and processing efficiency by using OpenCV + YOLO smart-focus cropping, then adding hash-based idempotency and lease-based request coalescing to prevent duplicate transforms.
- Improved delivery reliability and observability by orchestrating CloudFront-backed on-demand WebP transforms with readiness-aware retries and analytics for throughput, transform latency, and cache-hit efficiency.

Forest Fire Tracker (GitHub)   *Feb 2023 – Apr 2023*

- Built a Kafka-driven streaming pipeline to process Unity-generated satellite-image grids across ingestion, preprocessing, and analytics stages.
- Generated near-real-time fire-risk progression states by implementing OpenCV-based tree/cluster detection, fire-region extraction, and iterative spread prediction.
- Delivered reliable, live operational visibility by orchestrating Redis state management, Dask aggregation, and a Dash metrics dashboard with flow-control signaling.

Awards

- Won "Open Gov Data Hack" 2018, a national-level competition organized by NIC, Government of India.
- Won best project award in Technical Expo'18, VIT Chennai.