

NIKHIL KUMAR

Data Engineer | Secure & Scalable Data Systems | Streaming & Batch | Azure & AWS

+91-9325526453 | nk2242696@gmail.com | linkedin.com/in/nikhil-k-82b003ba | github.com/nk2242696 | nk2242696.github.io

PROFESSIONAL SUMMARY

Senior Data Engineer with 7+ years of experience, designing and operating large-scale, secure data systems across Azure and AWS. Proven track record of building batch and streaming platforms processing 12TB+ data daily across 250+ pipelines serving 700+ users using medallion (lakehouse) architecture and Delta Lake. Led high-impact initiatives in cost optimization (reducing spend by \$80K/month), data reliability (40% reduction in pipeline failures), and platform security under Microsoft's Secure Future Initiative. Deep expertise in Spark, Kafka, Databricks, and distributed data systems with a strong focus on event-driven architecture, data governance, and performance optimization.

EXPERIENCE

Microsoft

Aug 2022 - Present

Data Engineer

- Designed and built a scalable event-driven data platform on Azure (Synapse, ADF, Event Hub, ADLS, SQL Server) using **medallion architecture (Bronze/Silver/Gold)** and **Delta Lake**, orchestrated via **Airflow and ADF**, processing **12TB+ data daily** across **250+ pipelines** serving **700+ users** including forecasters, analysts, and planners.
- Led end-to-end design and org-wide adoption of a config-driven **Data Quality Framework** using Great Expectations with custom alerting, rules engine, and auto-imputation across 250+ pipelines, **reducing failures by 40%** and strengthening data governance practices.
- Led security hardening under Microsoft's **Secure Future Initiative (SFI)** across **16 subscriptions and 116 resources**: implemented VNET isolation, zero outbound connectivity, and network security perimeters.
- Migrated **10 production workloads** from single VMs to VMSS Flex, enabling zero-downtime patching, auto-scaling, and improved resilience.
- Drove **28% reduction in Azure spend (~\$80K/month)** through SQL DW SKU right-sizing, idle resource cleanup, and storage tier optimization; built Power BI dashboards to track cost trends; led a 5-person Kaizen initiative.
- Improved Cosmos DB workload distribution by **12%** through telemetry-driven autoscale provisioning and performance tuning, reducing tail latency for high-throughput queries.
- Built a **production AI-powered data assistant** using Azure OpenAI and Autogen, enabling 25+ analysts to query ADLS via natural language, **reducing time-to-insight by 60%**.
- Developed **12 reusable Synapse pipeline templates** with CI/CD integration for cross-tenant data movement and ingestion, accelerating new project onboarding by **5x** and enforcing engineering standards.
- Mentored junior engineers during onboarding; led design reviews for cost monitoring and data quality initiatives in Agile sprints across the platform.

Epsilon

Aug 2021 - Aug 2022

Software Engineer 2

- Optimized **12-15 Kafka and Spark** (batch + streaming) pipelines processing **300B+ records/day**, improving throughput and stability.
- Reduced HDFS storage footprint by **3x (~9TB/day to ~3.4TB/day)** through ORC compression tuning, Kafka Sticky Partitioner, batch size optimization, and high-cardinality column sorting.
- Extended the open-source CAMUS (LinkedIn) framework with custom source, sink, and config modules for efficient migration from MapReduce to Spark pipelines, improving job efficiency and reducing processing latency.
- Built a **Spark observability dashboard** using Elasticsearch, Logstash, Kibana (ELK), and Prometheus/Grafana, improving job visibility and reducing debugging time.

Hashmap

Jun 2018 - Aug 2021

Cloud Data Engineer

- Migrated enterprise data warehouses from Oracle/Netezza to **Snowflake** for 3 enterprise clients, enabling scalable cloud analytics across multiple data marts.
- Built end-to-end AWS data pipelines using **S3, Lambda, SQS, SNS, and Matillion** for IoT/manufacturing data ingestion; automated infrastructure deployment with CloudFormation.
- Designed **Lambda architecture** (batch + streaming) using Spark and HBase for real-time semiconductor event processing at TB-scale daily volume via Kafka.
- Developed reusable parsers and ingestion frameworks for CSV, JSON, Excel, and PDF data formats, improving flexibility and reducing manual effort.

TECHNICAL SKILLS

Languages: Python, Java, SQL, Bash

Data Engineering:	Apache Spark (Batch & Streaming), Kafka, Databricks, Delta Lake, Airflow, Great Expectations, Hadoop, Hive
Cloud & Platforms:	Azure (ADF, Synapse, ADLS, Event Hub, Cosmos DB), AWS (S3, Lambda, SQS, SNS), Snowflake
Databases:	PostgreSQL, MySQL, Cassandra, HBase, Cosmos DB
Infra & DevOps:	Docker, Kubernetes, Linux, ELK Stack, Prometheus, Grafana, Azure Monitor, AWS CloudWatch, CloudFormation
Security:	VNETs, Network Security Perimeters, Zero Trust Architecture
Data Modeling:	Data Warehousing, Schema Design, Distributed Systems
Familiar With:	Microsoft Fabric
Other:	Power BI, REST APIs, Git

SELECTED PROJECTS

AI-Powered Data Assistant | *Azure OpenAI, Autogen, Python, ADLS*

Built a natural language querying system enabling data access over data lakes without requiring SQL knowledge. Designed prompt orchestration and query translation workflows for structured data retrieval. Used in production by 25+ analysts.

HBase SQL-like Query Engine | *Python, HBase, DSL*

Developed a Python-based DSL to enable SQL-style querying on HBase time-series data, simplifying access for analysts and reducing query complexity.

Excel ETL Automation Framework | *Python, crealytics*

Automated Excel-based ingestion pipelines using Python and crealytics, reducing manual effort by 30% and improving data reliability.

EDUCATION

B.E. in Information Technology

Army Institute of Technology (AIT), Pune | 2014 - 2018

CERTIFICATIONS

- Snowflake SnowPro Core Certification
- Google Cloud Professional Data Engineer
- Applied Data Science - WorldQuant University

AWARDS

- **Mountain Mover Award** - Microsoft