Vinay Kumar N

Senior Data Engineer at Finoa
Software Engineering, Distributed Systems
AWS Certified Solutions Architect - Professional

narayanavinaykumar@gmail.com
+49-162-1518203 / +91-987-237-5950
/vinay26k /Vinay26k
/vinay.dev *C*redly/vinay26k

Summary

Seasoned Data & Software Engineer with **7+** years of **experience** in **architecting, scaling & implementing data platforms**, First data hire reporting directly to the CTO, spearheading the end-to-end build of a modern data platform - from strategy and architecture to hands-on implementation, while driving data literacy and adoption across the organization.

Core Competencies

- **Data Platforms & Architecture**: Building lakehouses and end-to-end data platforms from scratch; define roadmaps aligning to business strategy.
- **Data Engineering Expertise**: Designing and optimizing data pipelines that scale, implementing governance frameworks that matter, and developing in-house tools that improves efficiency and reliability.
- **Scalable & Cost-Efficient Systems**: Delivering resilient platforms in resource-constrained environments, minimizing vendor lock-in and achieving significant cost savings.

Skills & Tools

- *Programming & Data* - Python, SQL, Postgres, Apache Spark, Apache Kafka, Amazon Redshift, Snowflake, FastAPI, Data Modeling
- *Lakehouse Technologies* - Apache Iceberg, Delta Lake
- *Cloud Platforms* - AWS, Azure, GCP (multi-cloud expertise)
- *Data Engineering Practices* - ETL/ ELT, Streaming pipelines (near real-time, real-time), Lakehouse architectures, Data Governance
- *Platform Engineering & DevOps* - Docker, Terraform (IaC), git
- Contributions to the open-source; actively exploring and building with *'Rust'*

Experience

- **Finoa** Berlin/Germany, DE
  Senior Data Engineer; Data Platform *Oct 2024 - Present*
  - ***As a first Data hire***: Establishing the vision, aligning the strategy and devising road map for analytics & data driven decisions adoption across the business departments.
  - **Data Platform**: Designing and implementing Data Platform architecture from scratch on modern data stack that is handling ingestion, storage, transformation and governance. Data Platform is built-on
    - Kubernetes, Apache Spark, Apache Iceberg, AWS MSK/ Kafka, S3, AWS Aurora Postgres Database and Snowflake
  - **Agentic AI Workflows**: Building agentic workflows & in-house tools, solving the context issues to enable teams use tools with confidence and less hassle
    - Python, FastMCP, AWS Kiro, Amazon Q Developer, RestAPI
  - **Daily**: Work activity involves
    - Cross team collaboration with Platform, Developers, SecOps, Compliance and other business departments
    - On-boarding & analyzing assets to the Data Platform, based on end-users request
    - Building ingestion and transformation pipelines to automate the business processes
    - Working on observability metrics for Data Platform
    - Upgrading & Deploying data infrastructure of the platform
    - Contributing to the Organisation's key initiatives
    - Tracking Data Platform adoption with metrics and making it usable with user's feedback
  - **Improvements/ Performance Gain**:
    - *Data Platform Costs* - reduced costs of the Data Platform for a single source by **$33k**, that is **43**% by building the in-house components and custom pipelines on k8s

- **N26** Berlin/Germany, DE
  Data Engineer; Data Governance *Apr 2023 - Aug 2024*
  - **Data Engineering**: Built scalable data pipelines that adapt to growing datasets on Apache Airflow & in-house orchestrator pipelines within a Kubernetes environment. Developed custom connectors and containerized them for seamless execution. A key focus was also on ensuring data quality by building observability metrics such as accuracy, completeness, and consistency.
  - **Data Discovery Platform**: Designed and implemented a data discovery platform that significantly improved access to diverse data assets across our organization. This platform empowered data consumers to rapidly locate and utilize these assets. By streamlining this process, I've enabled Business Units to reduce the time spent in finding data assets to meet their needs - unlocking new insights and driving business success
  - **Data Governance Frameworks**: Contributed to the design and implementation of data governance frameworks, including automated access management controls and the generation of audit reports to support Security, Audit & Compliance requirements.
  - **Day-to-day**: Work activity involved
    * Developing Airflow data pipelines for ingesting metadata of the data assets
    * Python packaging and development for data governance frameworks
    * Development and integration of ingestion and transformation pipelines to the in-house orchestrator tool
    * Development of Data Observability frameworks
  - **Improvements/ Performance Gain**:
    * Data discovery platform implementation resulted in a 40-50% reduction in the time required for data access
    * Data Access Management tool required manual effort to assess and validate requests. I've worked on automating this process to eliminate 60-80 minutes of manual work per request.

- **Walmart Global Tech India** Remote/ Bangalore, IN
  Software Engineer III; Data Engineering *Nov 2021 - Mar 2023*
  - **Data Processing Pipelines at Scale**: Built data processing pipelines with GCP data stack. Dataproc for compute engine and BigQuery & GCS buckets for storing data for different formats. Areas/ stack that involves in building processing pipelines
    * Python packaging and development for common utilities
    * Pyspark/ Python data processing pipelines includes pulling raw data and publishing refined data
    * Integration of data science related processing scripts
    * Effective logging of data processing pipelines
    * Implementation of recovery from point-of-failure based on requirements
  - **Self Serve Data Platform**: Developed multiple iterations of backend service for the platform with majority of connectors to different data sources in Walmart allowing data consumers specially data analysts and scientists to perform analysis quickly without need for the data engineer resource. Stack - FastAPI, SQLAlchemy
  - **Improvements/ Performance Gain**:
    * Data Processing Pipelines at Scale: Improved upon existing pipelines and built generic data processing pipelines using templates for the new ones. Helped in leveraging already existing code and avoiding redundancy in development
    * Self Serve Data Platform: This project also aligns on the usage of existing code base but this is more on self serve & UI based. Avoided waiting time for data engineer resource allocation which was a sprint timeframe on average

- **Neudesic Technologies Pvt Ltd** Remote/ Hyderabad, IN
  Senior Data Engineer; Big Data Engineer *Nov 2020 - Nov 2021*
  - **Azure Migration**: Migration of Workstation telemetry data processing from existing big data stack (Hadoop, Cloudera systems) to Azure Data Stack leveraging processing engine as Azure Databricks, streaming platform as Azure Event Hubs
  - **Spark Structured Streaming**: Developed a total of **5** Streaming jobs some read data from Azure Event hubs and write to multiple sinks along with joins on batch data and other write data to Event hubs
  - **Azure Data Integration Patterns**: Developed multiple integration patterns and have been involved as Architect to suggest and make changes to the patterns. Worked on following data movement patterns namely
    * GCP BigQuery to ADLS Gen2 - Databricks as compute

    - Hive to ADLS Gen2 - Kafka MirrorMaker + Azure Event Hubs as streaming platforms, Databricks as compute
    - Nifi to ADLS Gen2 (& vice-versa) - Azure Event Hubs as streaming platform, Databricks as compute
    - ADLS Gen2 to QlikView - Databricks as compute
  - **Azure Databricks Logging Framework**: Developed a robust end-to-end logging framework on top of Azure Application Insights to log messages, traces and exceptions for Databricks notebooks execution
  - **Azure Databricks QlikView Integration**: Performed PoCs with different authentication mechanisms to Databricks from QlikView, thereby reporting ADLS Gen2 refined data through **external tables**
    - Article - **Azure Databricks QlikView Integration**
  - **Performance Gain/ Time Save**: Developed data pipelines reduced processing time and allowed to accommodate up-to **4 runs/day** which was earlier **2 runs/day** at the same cost

- **MAQ Software** Hyderabad, IN
  Software Engineer II *Sep 2018 - Oct 2020*
  - **Azure Migration**: Migration of multiple projects from On-Premise VM based systems to Cloud-centric Azure Data Stack. Successfully migrated projects to Azure Cloud leveraging majorly using following stack
    - Azure Data Factory - SSIS replacement & Jobs orchestration
    - Azure Databricks - compute for jobs execution
    - Azure Function App - replacement for our custom SSIS module
    - Azure Logic Apps - to trigger processing based on some events
    - Azure Event Hubs - streaming platform
    - Azure SQL Database - replacement for On-Premise SQL Servers
    - Azure Synapse Analytics (formely Azure SQL Datawarehouse) - replacement for On-Premise SQL Data Marts, PDWs
  - **Big Data Fuzzy String Matching**: Developed a Databricks Notebook as Utils for large scale fuzzy string matching that required to perform across **80M x 20K records** dataset. Performed various levels of performance tuning and able to achieve best possible time for the job.
    - Article - **big-data-fuzzy-matching-in-databricks**
  - **Azure Databricks Utils**: Developed Azure Databricks Utils that are used across the project that helps leveraging parallel processing like database interactions, data writes to ADLS Gen2, calculating row counts, verifying schema etc
  - **Stream Analytics - Real time**: Developed a data integration solution based on Cosmos Change Feed system, that has:
    - Azure Function that picks only change feed from Cosmos DB, process it and ingests to event hub
    - Azure Stream Analytics does processing on the Event hub data and ingests to Power BI dataset
    - Power BI dashboard refreshes in real-time thereby providing insights
  - **On-Premise SQL Data Warehousing**: Worked on this field for a limited time of 6 months and there after started working on Azure Migration. During this timeline maintained ETL and ELT jobs as various database stored procedures orchestrated as job and finally insights, facts, dimensions are loaded into Mart Server
  - **Recommendation Engine - Neo4j Graph DB**: Has experience in Neo4j Graph DB, worked on PoC for recommendation engine based on events data. Clients were impressed and chose to have solution developed based on this pitch
  - **Azure Server-less Solution**: Developed a custom Neo4j image and hosted on Azure Server-less solutions i.e., Azure Container Services. Worked actively in this stack and have reported, requirements during it's early stages issues
  - **Time Series Forecasting**: Delivered a project based on Time Series Forecasting and events that are related to it. One of the key topics in Causality of events based prediction system

Certifications

[Recent ↑]

- ☆ AWS Certified Solutions Architect – Professional
- ☆ AWS Certified Data Engineer – Associate

[ordered by earliest ↓]

- ✹ Exam 761: Querying Data with Transact-SQL
- ✹ Exam 778: Analyzing and Visualizing Data with Power BI
- ✹ Exam 767: Implementing a Data Warehouse
- ✹ Exam 768: Developing SQL Data Models
- ✹ Microsoft Certified: Azure Developer Associate
- ✹ MCSA: SQL 2016 Business Intelligence Development - Certified 2019
- ✹ Neo4j Certified Professional
- ✹ Python for Data Science
- ✹ Applied Data Science I: Scientific Computing & Python (with honors)
- ✹ Applied Data Science II: Machine Learning & Statistical Analysis (with honors)

Education

- **International Institute of Information Technology** — Online/ Bangalore, IN
  Executive PGP in ML & AI; GPA: 3.82/4 — *Sept 2021 – Nov 2022*
- **Lovely Professional University** — Punjab, IN
  Bachelor Of Technology Computer Science; GPA: 8.09/10 — *Aug. 2014 – May 2018*
- **Sri Chaitanya Junior College** — Andhra Pradesh, IN
  Higher Secondary (MPC); Percentage: 91% — *June 2014*
- **Ravindra Bharati School** — Andhra Pradesh, IN
  Secondary School (SSC); GPA : 9.8/10 — *May 2012*