Sanjeet Kumar Shukla

Senior Data Engineer | ☎ +91-7040190097
sanjeets1900@gmail.com | Linkedin | Github

Professional Summary

A Data Professional with 12+ years of experience in **architecting**, developing, and optimizing data solutions. Expertise in building modern data lakes, pipelines, and lake-houses using Big Data, Cloud, and DevOps technologies. Proven ability to lead, mentor, and drive innovation across complex projects. I am looking for a Staff/Lead or **Architect** role to apply my leadership, strategy, and technical skills in delivering data-driven solutions.

Core Competencies

- Big Data Technologies: Apache Spark, Snowflake, Kafka, Lakehouse, Iceberg, Hive, Elasticsearch
- Cloud: GCP (Big Query, Dataproc, GKE, Dataflow), AWS (EC2, Glue, EMR, ECS)
- Programming: Python, SQL, Scala, Java
- DevOps & CI/CD: Docker, Kubernetes, GitHub Actions
- Data tools: Looker, Tableau, Collibra, Atlan
- Leadership: Team mentoring, cross-functional collaboration, strategic planning
- Data Modelling: Data Modelling (Star/Snowflake), Database Design, Data Governance

Key Achievements

- Led multiple data lake implementation and **cloud migration** projects, saving operational costs and increasing system reliability. Worked on disaster recovery (DR) plan and setup for business critical operations.
- Well versed with data lake, data-warehouse, lakehouse and ETL/ELT processes, like ingestion, data modelling (fact, dimension), Data Quality Check, Medallion Architecture, SCD, CDC and Database normalization.
- Actively discussed and addressed complex data scenarios, and contributed to development of most effective solutions.
- Created TRDs and NFR documents capturing requirements, complex scenarios, edge cases, & potential challenges ensuring scalable and fail-safe design.
- Have worked in Product based companies for the last 5 Years alongside colleagues from tier-1 colleges and consistently exceed performance expectations. Have presented my Ideas, optimizations and cost saving initiatives to the C suite (CTO and CEO, CFO).

Professional Experience

Yieldmo | Senior Data Engineer – L5 | Remote | June 2022 - Present

- Led the design and development of high-throughput data platform processing 20 billion+ records daily, using **Kafka**, Spark, Snowflake, Airflow, and AWS services (S3, EC2, EventBridge, Lambda, Glue, EMR, Redshift).
- Mentored junior engineers on best practices, code quality, performance optimization, and cloud infrastructure.
- Worked on architecting snowflake pipelines using snowpipe, stages, transient/temp tables, clones, integrations, masking, tags, monitors, datashares. Have used streams, clones, dynamic tables and various other snowflake features.
- Wrote complex transformation, data masking job and cleanup job in Pyspark to enforce GDPR and data governance compliance. Optimized existing spark jobs to reduce run time and execution cost.
- Led cost optimization initiatives that resulted in a 20% reduction in processing and storage costs. This was achieved through strategic implementation of parallelization, sampling, and auto-scaling techniques at the pipeline, platform, and data source layers.
- Continuously working on optimising job performance to meet report SLAs despite increasing data volume.
- Developed and maintained automation tools (snowflake-schema-cloner, SQL-runner, aggregate table generator) that significantly enhanced teams productivity, and streamlined operational tasks.
- Collaborated with cross-functional teams to define data requirements, translate business needs into technical solutions, and ensure successful project delivery.
- Conducted POCs for emerging technologies like; Iceberg with Snowflake, DBT, Blazing SQL, Keebo, and Elasticsearch. Developed solutions based on these evaluations to reduce cost and optimize data pipeline.

- Built a RAG chatbot using historical slack messages with PGVector, sentence transformers, and LangChain during a hackathon.
- Created lookml models, looker reports, datagroups and looker sdk applications. Responsible for managing and scheduling looker reports to internal and external users. Worked on pdt, merged results, admin part of looker.
- Performed end-to-end **upgrade of the data platform**, including OS, Airflow, Python, Java, and Scala across 20+ repositories. Identified the need for modernization, planned the migration, and executed it seamlessly with zero downtime, showcasing strong project management and technical expertise.

Walmart Labs | Senior Data Engineer – IN4 | Remote | May 2020 – June 2022

- Worked on design and development of a data lake and data pipelines for Walmart's supply chain domain, utilizing Apache Spark, Kafka, Airflow, Druid, Elasticsearch, and GCP services (Dataproc, GKE, Big Query).
- Worked on Spark Scala to develop data pipelines for omnichannel business and replace existing SQL pipelines.
- Worked on various migration tasks like on-premise to GCP; Hive to Spark code migration, and do performance testing and Integration testing after migration in Hortonworks and GCP environments.
- Successfully migrated critical workloads from on-premise to GCP, created a python utility to automate the migration task, reducing manual efforts in migration and testing by 60%.
- Created Lookml models and looker reports using liquid variables, complex calculations and joins.
- Create POCs to test various use case; like Blazing SQL POC to run complex hive queries efficiently,
- Improved run time from 7 hours to 20 minutes. Created Spark on GPU POC with the help of Nvidia Team.
- Responsible for maintaining clean and maintainable codes with highest code coverage across departments.
- Conducted POCs on “Spark on GPU” to enhance performance, achieved a 90% reduction in query runtime.
- Enhanced the reliability and trust in new data pipelines by implementing robust data quality checks, establishing data lineage, and maintaining a comprehensive data catalog.
- Created end to end data pipelines involving getting data from NFS mounts. Loading data into data lakes, processing using Spark and loading into Big query, Druid, ElasticSearch and SQL Server.

Cognizant | Pune-India/USA **| Senior Associate**-Data Engineering | August 2014 - May 2020

- Worked closely with the product engineering team to design a **multi-layered** on-premise data-lake for network traffic analysis using open source big data stack (Sqoop, HDFS, Hive, PySpark).
- Created ELT pipeline and data-lake, reading structured and semi-structured data from 25 different source systems using Sqoop, Spark and python and stored that data in data lake (HDFS, Hive). Source systems included comm-vault, mcafee, remedy, sql-server and other sources.
- Used SparkML to implement anomaly detection model on network traffic data, using Isolation forest and KNN.
- Wrote Sqoop Jobs for incremental imports of data from RDBMS, used inhouse scheduling tool to orchestrate jobs, create bash scripts and python scripts for file transfer, file validation and DQ checks.
- Designed Hive external tables with partitioning, dynamic partitioning & buckets for better performance.
- Used sql server spark connector to write a huge volume of data from spark to SQL server in parallel.
- **Worked on on-premise to AWS migration** to move multi tiered data lake into aws cloud. Used S3, EC2, Glue, Athena, EMR, lambda and Redshift to migrate existing data lakes into AWS.
- Transformed complex JSON and XML data into structured Hive tables via JSON and XML SerDe libraries, improving data accessibility for analysts and accelerating report generation.
- Implemented data-warehouse features like SCD, CDC, Star Schema, snowflake schema, Unique Key (PK) etc using spark and hive. Implemented data quality checks at various stages of the pipeline.
- Worked on various Machine learning POCs like recommendation engine, CNN based Image classifier, patient classification etc. Created Spark Streaming application to get live feed from Yammer and do sentiment analysis on it. Used sparkML, Spark pipeline, sklearn and Keras for these POCs

TCS | Mumbai, India | Business Intelligence Developer | October 2011 - August 2014

- Worked as a BI developer for World’s leading consulting firm(McKinsey) in their market research project, and for India’s largest banking client (SBI). Developed mobile friendly dashboards in Tableau and IBM Cognos.
- Created reports and dashboards of varied complexity like tables, crosstab, custom charts, geographical maps, reports with drill down and drill through features.
- Published paper in the company's knowledge portal for customising BI tools to enable additional features that are not available by default. Eg. column limits, sorting, text box filters, tabular-sql.

- Worked on customising cognos reports using javascript and css. Created a custom template as per the specification given by the client and using cognos global stylesheet.
- Worked on IBM Cognos administration, **Dimensional and relational data modelling** using Cognos Framework Manager tool. Worked on models, deployment, packaging, scheduling, triggers etc.
- Worked on IBM Cognos Data Manager **ETL tool** to create ETL pipeline, reading data from various source systems and writing it to data marts and data warehouse in fact and dimension tables.
- Wrote complex SQL queries (pivot, views, and analytical queries) to fetch data from RDBMS in tableau.

Education

- **Degree:** **Bachelor of Technology** In Applied Electronics and Instrumentation
- **Year:** 2011
- **Institute:** ITER, Bhubaneswar, India

Github: https://github.com/SanjeetShukla01
Medium: https://medium.com/@sanjeets1900