Nishant Arora

Lead Data Engineer | BITS Pilani

Chandigarh, India   nishantarora@protonmail.com   nishaantt

Motivated Developer with over 7 years of industry experience. Skilled at designing and building scalable data systems and pipelines. Passionate about solving challenging industry problems using data and code. Good at picking up new work, technologies quickly and getting things done.

EDUCATION

Master of Science - M.Sc. (Hons.) : Mathematics, Birla Institute Of Technology And Science, Pilani Campus

Bachelor of Engineering - B.E. (Hons.) : Mechanical, Birla Institute Of Technology And Science , Pilani Campus

COURSES AND CERTIFICATIONS

I have **25+ professional and course certifications** including three Databricks certified professional certifications which are **Databricks Certified Data Engineer, Databricks Certified Spark Developer and Databricks Lakehouse professional**. The complete list can be found here ↗ .

WORK EXPERIENCE

Velotio Technologies (acquired by RSystems International), Sep 2024 – Present
Lead Data Engineer ↗
Leading data engineering efforts across multiple projects, also driving data engineering R&D for the organisation. Managed and Mentoring a group of data engineers. I am working on Snowflake, Airflow, dbt, Census, Salesforce, SAP, MS Access, BigQuery, Redis, ArangoDB, Flyte, Fivetran, dlthub, datahub, great expectations and with a number of popular data sources.

- *Articul8*: Architected and developed the whole data connector service for their GenAI platform. Supported 15+ data connectors constituting of structured, semi-structured as well as unstructured sources. Worked on the core GenAI pipeline - my focus was on ingesting the core data at scale into arangoDB, milvus and redis.

- *Accel*: Planned and implemented migration of legacy data syncing pipelines which were using Access, VBA and ODBC drivers to more scalable, robust, cloud focused pipelines utilising Snowflake, dbt and Census. It resulted in frictionless, reliable data sync to Salesforce. Improved performance of access -> snowflake data replication pipelines by ~70%.
- *Spinmaster*: Led the efforts for self-serve warehouse ingestion workflows to be used by the whole org. Implemented a number of business critical SAP reports in BigQuery. I also worked on various high-impact data quality, monitoring and observability initiatives which directly resulted in efficient, effective error detection and resolution.

Gemini, Senior Data Engineer ↗ Oct 2023 – Aug 2024
Successfully led multiple data engineering projects and mentored a group of four data engineers. I worked with Databricks, Snowflake, Airflow, Fivetran, AWS, Python, Scala, Soda and Looker.

- Architected and developed the ledger difference reporting system which is the backbone for the treasury and accounting teams. It manages 8 fiat currencies and 200+ cryptocurrencies, and is essential for cash reconciliation and ledger inconsistency detection. It's built to be highly scalable, reliable and maintainable.
- Designed the automated data quality and validation framework for the ledger difference reporting system.
- Spearheaded the successful migration of blockchain reconciliation pipelines from Snowflake to Databricks.
- Led the databricks monitoring and observability project which involved productionising overwatch and other alerting pipelines, and building key dashboards for jobs, clusters, notebooks and serverless SQL. Helped bring down the Databricks cost by ~20%.
- Designed and implemented ETL processes that empower leadership and product teams to monitor crucial north star metrics effectively. Optimized the performance of critical Looker dashboards, accelerating SQL query execution by 2.5x.
- Played a key role in developing and scaling processes to generate and send daily currency transaction reports to the New York State Department of Financial Services, ensuring compliance with regulatory requirements.
- Led the optimization of various inefficient and costly jobs, reducing the runtimes by 45x in some instances, saving the organisation over $50k annually.

Velotio Technologies

Senior Data Engineer Jan 2022 – Oct 2023

Led the data engineering efforts for multiple customers. Managed a team of 5 data engineers.

- *LeoLabs*: Designed and developed scalable data pipelines which process 350 million+ records daily and load them into the warehouse. Built pipelines to efficiently backfill historical data having a multi-billion scale. Integrated Deequ with the data pipelines to enforce the data quality. The primary tech stack was Spark, Redshift, RDS, Lambda, Deequ, AWS.
- *Axio*: Led the development efforts for the data team. Built multiple cdc, incremental and full-load etl/elt pipelines. Worked on managing and optimizing the warehouse and was responsible for a number of business critical reports and integrations. The data pipelines were built to get the data from various sql and nosql data sources, process them and load to Redshift.The primary tech stack used was Redshift, DocumentDB, Lambda, Python, AWS and Holistics.
- Space and Time: Worked in their database team. I wrote the grammar and backend Scala code to support materialized and parameterized views in their data warehouse. Worked on implementing new SparkSQL functions which were supported by Apache Ignite but not Spark. I also led testing and benchmarking for the beta product release. The primary tech stack was Spark, Ignite, Scala, Python, ANTLR.
- LevelAI: Designed and authored a scalable data self-serve pipeline to be used by the AI team to fetch data to train their deep learning models.Built a pipeline to auto-ingest all the data generated by the data annotation team using sensors. Implemented a data archiving pipeline that archives stale data from databases to partitioned data in datalake. Set up AirByte and BigQuery. The primary tech stack used was Spark,Airflow, AirByte, Postgres, BigQuery, GCP.
- *Atlan*: Designed and built workflows to fetch metadata from various data sources, owned the databricks and salesforce connectors. Authored argo packages to help benchmark and improve the metastore performance. The primary tech stack used was Argo, Python, k8s, SQL and majority of data sources in the modern data stack.

Data Engineer Jul 2020 – Dec 2021

My primary responsibilities were designing and building scalable data pipelines, and data warehousing.

- *Seagate*: Worked on an end-to-end ETL pipeline handling petabytes of data daily. Managed hundreds of production tables and airflow dags. Handled huge amounts of structured as well as unstructured data. The primary tech stack was Hive, Spark, Python, Airflow, Presto, Sqoop and AWS.
- Zylotech: Designed an end-to-end scalable customer data platform pipeline handling over 600 million records daily. The pipeline involved many sub-pipelines including ingestion, data preprocessing, deduplication, and unification pipelines following various configurable business rules. Primary tech stack was Spark, Python, Airflow, Snowflake, GCP.

Ecom Express, Data Engineering Intern Jan 2020 – Jun 2020

Designed and built an automated scalable end to end forecasting pipeline from fetching the raw data everyday to delivering the forecasts, using Apache Spark, Apache Airflow, Python, S3, MySQL. Implemented pipelines which ingest raw as well as processed data from MySQL database to the S3 datalake which use various connectors between Apache Spark code, S3, MySQL and Python. Authored a module for Analysing Data by fitting various probability distributions to it using PySpark and Scipy. Accomplished Locality extraction on a Pincode level by clustering on latitude and longitude using Spark,Scala.

couture.ai, Data Science Intern Jul 2019 – Dec 2019

Designed and implemented a NLP Pipeline in Apache Spark and Scala which does various basic text cleanups to spell correction, part-of-speech based filtering, phrase extraction.Authored the implementation of a pipeline which does automated feature selection from data. Introduced sales prediction pipelines using xgBoost and lightGBM models in Apache Spark and Scala. Worked on the implementation of recommendation and similar products pipeline using Topic Modeling in Apache Spark. Researched on AutoML and the implementation of such a platform.

SKILLS

Apache Spark | Python | Snowflake | Databricks | SQL | Airflow | Scala | AWS/GCP/Azure | DeltaLake | ML/AI | Redshift | Trino | Hive | DBT | Distributed Systems | Kafka | Spark Streaming | Fivetran/AirByte | Flink | Argo | NoSQL | Redis | Java | Terraform | k8s | GraphDBs | Salesforce | Business Intelligence

ACCOMPLISHMENTS

- **ACM ICPC Asia Amritapuri Site Regional Contest 2016 Participant**, team name : treeBitsians
- **Rockstar Employee Award**, Velotio Technologies and **Best Intern Award**, nearbuy.com
- **Competitive Programming Online Judges :** Hackerrank Algorithms Domain Rank of 200 with 99% percentile. SPOJ world rank of 2500. Codeforces Expert with 500+ problems solved
- **Programming Contests :** Ranked 3rd in CodeGenesis Contest, 4th in National College Hackathon, 8th in National Inter University Programming Contest, selected for Microsoft codefundo++ 2018.
- **National level Chess player, Gold Medalist at State level**