Mayank Arora

Noida, Uttar Pradesh | +91-9717240160 | mayankarora.1701@gmail.com | LinkedIn

Hands-on professional with 3 years of experience working extensively with big data tools like Hadoop and Spark. Whether it's optimizing SQL queries, tuning Spark jobs, or improving cloud-based architectures, I focus on making data work smarter and faster. I've built data systems that improved accuracy, enabling better decision-making across teams.

Previously spent 2 years in IT infrastructure management, handling private cloud environments before transitioning into data engineering. This experience gave me a strong grasp of system performance, scalability, and exposure to Azure cloud.

Skills

Programming Language – Python, SQL

Big Data Technologies – Hadoop, PySpark(Spark SQL, DataFrame API), Databricks(Delta Lake, DLT), Hive, Kakfa

Cloud Platform – Azure Data Factory, Azure Synapse, Azure Databricks, Azure Data Lake Gen2

Database – MySQL, Microsoft SQL Server, Azure SQL Database, NoSQL (MongoDB)

Version Control – Git, GitHub

Professional Skills – Agile Practitioner, Cross-functional Collaboration, Technical Writing, Documentation

Professional Experience

Data Engineer — Mar 2020 – Present

Tata Consultancy Services Ltd.

- **Migrated on-prem Hadoop-based workflows** to Azure cloud, reducing infrastructure costs by **30%** while improving data processing efficiency.
- Built scalable batch and streaming ETL pipelines using **PySpark** and **Spark SQL** in **Azure Databricks** with **Delta Lake** & **Auto Loader**, processing **over 5TB daily data** and reducing ingestion time by **40%.**
- Leveraged Azure Data Factory to **ingest large datasets** from databases and APIs into a data lake, improving ingestion efficiency and **reducing data transfer time by up to 30%.**
- Ingested, processed, and transformed **semi-structured data (JSON, Parquet, Avro)** from APIs and event streams into **structured formats** for analytical processing.
- Designed and implemented **DLT pipelines** in Databricks using **Spark Structured Streaming**, ensuring real-time **data quality enforcement** and lineage tracking.
- Optimized data models using **Star and Snowflake schemas** in Synapse Analytics, enabling **10x faster queries** for BI dashboards.
- Enhanced daily job efficiency in **Azure Synapse** by applying **columnar file formats** and **advanced techniques**, leading to significant reductions in data scan sizes.
- Simplified **Spark SQL** queries and **optimized data models** for reporting dashboards, **reducing query execution time by 60%.**
- Implemented **schema evolution** and **CDC (Change Data Capture)** with **Delta Lake**, ensuring seamless data updates with minimal downtime.
- Monitored and troubleshot data pipelines using **Azure Monitor, Log Analytics, and Databricks Job Monitoring**, achieving **99.9% uptime.**
- **Created and maintained documentation** for data engineering processes, systems, and standards, ensuring clarity and consistency across 100+ workflows.

Achievements

- Earned Microsoft Certified: Azure Data Engineer Associate credential (DP-203)
- Earned certification in 'Azure Cloud and Databricks', demonstrating expertise in large-scale data processing.

Education

Vivekananda Institute of Professional Studies — 2016 - 2019

Guru Gobind Singh Indraprastha University, Delhi

Bachelor of Computer Applications