Anoop Kumar Singh

8958228829| imaks2597@gmail.com | linkedin.com/in/anoop-singh-de |

Education

IIMT Engineering College — Meerut, U.P.
Bachelor of Technology in Civil Engineering — *Aug. 2015 – May 2019*

Aditya Birla Public School — Grasim, C.G.
Intermediate (PCM) — *April 2012 – May 2013*

Aditya Birla Public School — Hirmi, C.G.
High School — *April 2012 – May 2013*

Experience

Junior Technical Consultant (Data Engineer) — Oct 2023 – Present
Digivate Labs — *Gurugram, HR*

- Developed scalable batch and real-time data pipelines using Databricks, Autoloader, and Apache Spark.
- Implemented Unity Catalog for robust data governance and secure access controls.
- Built and deployed data jobs orchestrated via Databricks Jobs, enhancing pipeline automation.
- Partnered with analysts and data scientists to ensure data quality and lineage across projects.

Database Administrator — Feb 2023 – Oct 2023
RR Finance — *Delhi*

- Managed SQL Server environments including user roles, backups, and query optimization.
- Automated stored procedures and improved query performance to support business operations.
- Ensured high availability, data consistency, and security in transactional systems

Technical Executive — Nov 2021 May 2022
Ultratech Cement Pvt. Ltd. — *Deoghar*

Site Engineer — Jun 2021 Nov 2021
Goel Construction Pvt. Ltd. — *Hirmi*

Site Engineer — Jul 2019 Jan 2020
Rudra Constructions

Projects

E-commerce Data Platform Migration — Feb 2025 – Present

- Automated migration 35TB+ of historical and real-time data from vertica to s3 to databricks, ensuring data integrity and minimal downtime.
- Re-engineered and optimized 150+ ETL pipelines and batch workflows using Databricks notebooks and Delta Lake
- Refactored 1100+ SQL scripts to align with updated naming standards
- Implemented robust data governance and access control with Unity Catalog.
- Applied naming rules: tables → lowercase, columns → camelCase
- Automated data quality validation and monitoring to ensure accuracy and consistency post-migration.
- Achieved a 40% reduction in infrastructure costs and improved query performance by 3x.
- Collaborated with cross-functional teams to align migration with business requirements and SLAs.
- Utilized best practices in schema evolution, data transformation, and cost optimization.
- Built metadata-driven batch processing for scalable table migration.
- **Key Skills:** Databricks, AWS, ETL, Data Migration, Delta Lake, Python, SQL, Data Quality, Data Governance, Big Data, Real-time Streaming and Data Processing, CI/CD, Stakeholder Management, data lake

Project: Data Pipeline Modernization — Dec 20w4 – Feb 2025

- Refactored and optimized 10+ legacy Python scripts into scalable PySpark jobs, significantly improving data processing performance for large-scale IoT datasets.
- Migrated complex ETL pipelines to Databricks, utilizing advanced Spark features such as partitioning, caching, and optimized joins to enable distributed, real-time batch analytics.

- Integrated geospatial data processing using GeoPandas and DBSCAN, and implemented robust data enrichment, aggregation, and alerting mechanisms with Slack integration.
- Enhanced data quality and reliability by establishing comprehensive error handling, logging, and monitoring systems.
- Collaborated cross-functionally to ensure seamless migration, validation, and deployment of new data workflows, driving improved operational efficiency and analytics capabilities.
- Refactored inefficient Spark jobs, reducing runtime by over 40% through effective caching, partitioning, and code modularization.
- Applied best practices in Delta Lake storage layout and metadata management to optimize data workflows.
- Key Skills: PySpark, Databricks, Distributed Data Processing, ETL, Kafka, Delta Lake, Geospatial Analytics, Python, Real-Time Data Pipelines, Data Engineering, Automation, Monitoring & Alerting, Problem Solving, Scalability

Real-Time Streaming Analytics Platform Oct 2024 – Dec 2025

- Designed and implemented a real-time streaming data pipeline for an e-commerce client using Databricks Delta Live Tables (DLT) and PySpark, processing live data from Azure Data Lake Storage.
- Developed and deployed a scalable Medallion architecture (Bronze, Silver, Gold layers) to optimize data ingestion, transformation, and analytics workflows, implemented separately in both Python (PySpark) and SQL.
- Built automated data quality monitoring with expectation-based validations, quarantining invalid records and achieving 99.9% data accuracy for downstream reporting.
- Implemented SCD Type 1 and Type 2 transformations for dimension and fact tables, enabling historical change tracking and enhancing data warehouse integrity.
- Integrated a real-time analytics dashboard that automatically refreshes as new data arrives in ADLS, providing instant insights into total revenue, customer retention, discount impact, and product performance KPIs, accelerating business decision-making and market responsiveness.

AWS SQS to Delta Lake Ingestion Pipeline may 2024 – May 2024

- Built a fault-tolerant data ingestion pipeline using Boto3 to stream SQS messages into Databricks Delta Tables.
- Implemented exactly-once delivery and deduplication logic for streaming data to ensure reliability and data integrity.

Technical Skills

Programming & Querying: Python (Pandas, NumPy, Boto3), SQL, PySpark, T-SQL (SQL Server)
Big Data & Processing: Apache Spark (RDD, SQL, DataFrame API), Databricks (Jobs, DLT/workflows, Autoloader, Unity Catalog), Kafka
Data Architecture: Delta Lake, Medallion Architecture, Structured Streaming, ETL/ELT , Data Warehousing
Cloud Platforms: Azure (Azure Data Factory (ADF), Synapse, ADLS / Blob), AWS (S3, Lambda, Glue)
Databases: SQL Server, MySQL, Databricks SQL **Tools & DevOps:** Git, GitHub Actions, Power BI, Tableau, VSCode, draw.IO ,CI/CD **Soft Skills:** Collaboration, Problem-Solving, Quick Learning, Time Management
Languages: English, Hindi