RISHAV ANAND

IBM certified in Big Data Technologies with AWS

+917856978876 | anand.rishav7856@gmail.com | LinkedIn | Git-Profile | Portfolio | India

Technical summary:

- Skilled in creating systems to organize, process, and analyze big datasets.
- Experienced in designing and improving complex data pipelines and structures.
- Collaborates well with different teams, always learning, and dedicated to improving data-driven processes for better.

Professional Experience:

Capgemini Technology Services Limited (Working as BI Specialist/Lead) – Present

- Built **Pandas-based data quality checks** for nulls and field dependencies before pipeline ingestion.
- Designed EC2 pre-processing to offload validations, reducing EMR resource usage.
- Optimized queries by writing **partitioned Parquet files,** enabling **partition pruning** (↑40% performance).
- Developed **Spark incremental loading**, reducing overhead and improving reporting.
- Applied **broadcast join** with exchange rate reference data, cutting shuffle and improving Spark runtime (~25%).
- Used **predicate pushdown** and **column pruning** in Parquet to minimize I/O and memory.
- Orchestrated full pipeline with **Airflow**, automating EC2/EMR provisioning, transfers, and transformations.
- Captured **CDC from MySQL** and ingested via **S3 staging to Redshift**.
- Implemented **SCD logic in PySpark + SQL** to manage updates and maintain history.
- Automated **CDC-to-SCD workflows in Airflow**, optimizing Redshift COPY and Spark transformations.
- Delivered **end-to-end report automation** with **Airflow + Python** for scheduling, transformation, and delivery.
- **Tech stack – Boto3, MySQL, EC2, EMR, Glue Crawler, Redshift, PySpark, Python, Pandas, Data-Warehousing, Airflow**.

Infosys Pvt. Ltd. (Worked as System Engineer) – (21 July 21 to 10 August 24)

- Developed an **OLAP system** focused on improving product recommendations, service quality, and customer satisfaction.
- Emphasized **real-time data processing** from user interfaces (UIs).
- Designed data pipelines using Apache **Spark and Kafka** to handle large-scale data efficiently.
- Extracted data from external **APIs** and streamed it into HBase via Kafka.
- **Developed PySpark/Scala scripts for migrating data** from HBase to Hive tables.
- Created efficient Spark jobs to accelerate data processing and reduce job execution times.
- **Implemented data partitioning and caching strategies** in Spark to optimize performance and resource utilization.
- Ensured data quality checks for real-time data coming from UIs.
- Improved **Spark job execution plans and memory management**, enhancing overall system efficiency.
- **Tech Stack - Shell Scripting, Hive, Kafka, PySpark, NoSQL.**

Internships

Sonora Engineering OPC (P) Ltd, New Delhi

- Contributed to backend database design, data modeling, and warehouse structuring for operational analytics.
- Gained hands-on experience in data extraction, transformation, and production data validation.

Certifications

- AWS Certified Cloud Practitioner-(CF-02)
- IBM Certified in Data Engineering
- Data Management with Databricks: Big Data with Delta Lakes
- Fundamentals of Project Planning and Management
- Hacker Rank - SQL Advanced certified

Achievements

- Golden Badge achiever in SQL in Hacker Rank Platform
- Spot-on Award for performance, delivering result and ownership (Infosys)

Additional Project (POC):

Live Data Streaming using Kafka

- This architecture facilitates real-time data flow from the MySQL database to a JSON file.
- It uses Kafka Streaming for efficient and scalable communication between the upstream and downstream components.
- Live Data streaming using AWS service (kinesis)
- This project sets up a comprehensive data pipeline, starting from mock data generation (python script pushing data to DynamoDB) in DynamoDB.
- Then streaming through Kinesis, applying transformations with Lambda, storing in S3 through Firehose, and making the data query-ready in Athena using Glue Crawlers.

Skills:

- **Language:** C, Python. Java, DSA, SQL/NOSQL, PySpark(Framework)
- **Big Data Tools:** Hadoop, Hive, Spark, Kafka, Datawarehouse/Data Modeling, Airflow (Orchestration)
- **Tools:** Git, Jira

Education Qualification:

- **DEGREE NAME** - B-Tech
- **College** - Silicon Institute of Technology
- **Branch** – Electronics and Communication
- **CGPA** - 7.74
- **Year** – 2017-2021