Rishav Anand

IBM certified in Big Data Technologies with AWS

anand.rishav7856@gmail.com
+91-7856978876
LinkedIn | Git-Profile | Portfolio

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

- Implemented **data quality checks** using **Pandas** to validate nulls and field dependencies before pushing data to the pipeline.
- Designed a lightweight pre-processing step on **EC2** to offload basic validation from the cluster, reducing resource usage on **EMR**.
- Enabled **partition pruning** by writing partitioned Parquet files, improving downstream query performance by up **to 40%** by reading only relevant partitions.
- Built an **incremental data loading** mechanism using Spark filters and partitioned writes, avoiding full data reloads and minimizing compute cost.
- **Used broadcast join with the daily exchange rate reference data to eliminate shuffle overhead, contributing to a ~25% improvement in Spark job execution time.**
- Leveraged **predicate pushdown** and column pruning via Parquet, which minimized unnecessary I/O and memory usage during reads.
- Used **Apache Airflow to orchestrate** the entire data pipeline — automating EC2 and EMR provisioning, script transfers, dependency setup, and transformation job execution.
- **Owned the end-to-end automation** of daily client report generation and delivery using **Apache Airflow** and **Python**, optimizing scheduling, data transformation, and distribution.
- **Tech-stack – Boto3, EC2, EMR, Glue Crawler, PySpark, Python scripting, Pandas, 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.**

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.

Technical Skills:

Language

- C
- Python
- Java & JDBC (Basic)
- DSA
- SQL/NoSQL
- PySpark

Big Data Tools

- Hadoop
- Hive
- Spark
- Kaka (Streaming)
- Datawarehouse/Data Modeling
- Airflow (Orchestration)

Tools

- Git
- Jira
- Jenkins

Education Qualification:

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