Rishav Anand

IBM certified in Big Data Technologies with AWS

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

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:

❖ 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 and implemented 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.**

❖ Capgemini Technology Services Limited (Working as BI Specialist/Lead - 0.6 yrs. of Experience)

- Developed data ingestion pipelines using **PySpark** to read and process data from various sources like **Hive**.
- Designed and implemented a solution to publish data to **Kafka** using **Spark's Kafka connector**.
- Used **structured streaming** to produce JSON events, with data serialized as key-value pairs, sending them to Kafka topics for further processing.
- Worked with **complex data transformations**, and creating structured event payloads.
- Implemented business logic in transformations by creating contract and relation identifiers, handling null checks, and applying header details to event data before sending.
- Configured secure Kafka connections with **SASL_SSL and JAAS** authentication, managing Kafka API keys and secrets securely.
- **Optimized Spark jobs by configuring spark executor cores, memory, and shuffle partitions for efficient execution in production environments.**
- Enabled **broadcast joins and filter operations** to handle large datasets.
- **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 – Hive, pyspark, Kafka, Airflow, Jenkins-CICD Pipeline.**

Certifications

- AWS Certified Cloud Practitioner – CF-02
- IBM Certified in Data Engineering

Achievements

- Hacker Rank - SQL Intermediate certified
- Golden Badge achiever in SQL in Hacker Rank Platform

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