DHIRAJ C MISHRA

Phone: (+91) 9082318513
Email: dhiraj.m.jobs@gmail.com

Summary

AWS Data Engineer with 7+ years of proven track record in designing and implementing scalable real-time and batch data pipelines on AWS. Adept at optimizing big data workflows and delivering actionable insights through robust analytics solutions. Architected and deployed data models supporting both OLAP and OLTP use cases for a major sports ticketing platform. Designed and implemented a data governance framework for PII encryption, and used AWS KMS (server-side and client-side) and secure data migration across environments. Expertise includes Spark performance tuning, large-scale data migrations, and leading cross-functional teams to enable data-driven decision-making.

Projects

Elevate Sports & Ventures – Quantiphi Mumbai | July 2023 – Present

Associate Technical Architect - Data

- Designed and implemented a scalable near real-time streaming platform on AWS using Postgres RDS, DMS, Kinesis Data Streams, Lambda, and Redshift for a sporting client. Identified and mitigated various system bottlenecks, such as multiple upsert operations, late-arriving data, and duplicate records. Setup prerequisite to start the DMS data movement from RDS to Kinesis using WAL.
- Created a staging layer to process data from multiple sources, resolving serializable isolation errors. Implemented a checkpointing process to ensure only new or updated data was processed.
- Successfully migrated over 200 GB of raw data using AWS Glue, Python, Pyspark, and API integration, loading it into the Redshift data warehouse. The process was orchestrated using Step Functions and optimized for high-performance analytics. Used DynamoDB to store metadata for handling retries, failures, and preventing data duplication.
- Developed an automated incremental data loading script using checkpointing and watermarking techniques, ensuring seamless updates after historical data migration. Orchestrated the entire workflow using Step Functions.
- Created over 30 comprehensive reports and interactive charts in real time, empowering business stakeholders with actionable insights and supporting data-driven decision-making across operational and strategic functions.
- Designed an analytical platform to help a client answer critical business questions about the U.S. population. Processed terabytes of data through preprocessing, transformation, sampling, and extrapolation to generate the final dataset.
- Improved Spark job performance by optimizing the dataset scope, reducing execution time from 1.5 hours to under 20 minutes. Experienced in identifying and resolving Spark performance issues, such as data skew and out-of-memory errors.
- Designed and implemented a governance framework for performing encryption of PII data and facilitating secure migration to lower environments for consumption. Developed a validation framework to verify data integrity post-loading. Additionally, developed a validation framework to verify data integrity post-loading.
- Established a landing account to securely hold all incoming data, applying encryption and masking protocols before transferring to lower environments.
- Currently migrating vanilla Parquet-based data lakes to Apache Iceberg, improving query performance through optimized data scanning and reduced job runtimes.
- Leading a team of over five data engineers with varying levels of experience, providing mentorship, technical guidance, and delivery oversight.
- Collaborated closely with the platform team to identify cost-saving opportunities and implement continuous AWS cost optimization strategies, reducing unnecessary compute/storage usage across Databricks and other AWS services.

IntentPro – Blenheim Chalcot Mumbai | May 2023 - July 2023

Senior Data Engineer

- Worked on building business using Generative AI like ChatGPT v4, Dall-e.

Emirates NBD – Synechron Mumbai | Dec 2022 - May 2023

Senior Software Developer - Data

- Setting up an AWS Lambda function to retrieve files uploaded to S3 and perform a data validation and completeness check on the data, and creating Lambda functions to move data from raw to lake layer as per the load strategy.
- Developing Pyspark code for performing business transformations and moving the data to the hub layer in order to facilitate this process.

The AA – Deloitte USI Mumbai | July 2021 - Dec 2022

Consultant - Data

- Worked for a Finance project to build AWS pipelines to load data from multiple source systems using Kinesis data firehouse, S3, EMR on EKS, lambda, DynamoDB.
- Implemented customer segmentation based on the logic delivered by client.
- Airflow is used to execute the data pipeline, which is running on an EMR EKS Fargate cluster, while CloudWatch is used to monitor job statistics. In order to ensure that downstream extracts are performed efficiently and accurately, DynamoDB is loaded for downstream extracts.

Takeda Pharmaceuticals – Accenture Mumbai | Dec 2017 - July 2021

ETL Developer

- This project was designed to migrate datasets from Legacy system to AWS.
- Designed and developed scalable ETL pipelines using Pyspark SQL on Databricks, processing diverse datasets such as Customer, Address, and Patient records by re-engineering legacy workflows originally built in Informatica PowerCenter 10.1.
- Extracted data from structured sources like Oracle and Amazon Redshift, applied advanced Pyspark transformations (pivot, custom UDFs in Python), and loaded the results into Amazon S3 in columnar formats (e.g., Parquet) to optimize performance for analytics and reporting.
- Setting Involved in preparation of high, low-level design documents, run books and automated multiple manual processes using Unix and Python Scripting.

Skills

- Programming Languages: Python, Shell
- Cloud: AWS
- Processing & Streaming: Spark, MapReduce, Kafka
- Artificial Intelligence & LLMs: Amazon Bedrock, OpenAI (ChatGPT, DALL·E)
- Storage / DWH / Databases: Redshift, Oracle, Postgres, DynamoDB, Hadoop, Hive
- Data Orchestration: Step functions, Airflow
- AWS services: EMR, Glue, Glue catalogue, Athena, Lambda, S3, Kinesis Data Streams, DMS, IAM, EC2, SNS, ECS
- Data Integration/Management: Informatica (PowerCenter, BDM, MDM)
- Version control: Git, Jira, AWS CI/CD, Docker, Linux, Jira, SVN, Confluence

Education

P.G.Diploma in Machine learning & AI – I.I.I.T Bangalore, Online | 2019 – 2020

B.E. IN ELECTRONICS AND TELECOMMUNICATION – K.J. Somaiya College of Engineering, Mumbai | 2013 – 2017

Achievements & Certifications

- AWS Certified Developer - Associate certificate