Vishal B

dataengineervishal@gmail.com 9994405350 in/dataengineervishal github.com/vishalmighty

SUMMARY

Data Engineer with 2 years of professional experience in designing and optimizing ETL pipelines, real-time streaming data solutions, and cost-effective data workflows. Skilled in Python, SQL, GCP, Apache Flink, DBT, and Airflow, with a proven track record of reducing ETL costs by up to 70% and enabling real-time analytics. Adept at building scalable data pipelines and creating analytical solutions to drive operational efficiency and business insights, with experience working in sprint/agile frameworks.

SKILLS

Languages: SQL, Python, PySpark, C, Java
Database & Datawarehouse: Mysql, Postgres, Hive, Snowflake, Nosql
Bigdata Frameworks: Flink, Hadoop, Spark, Databricks, ETL, ELT
AWS Services: S3, Lambda, Glue, Athena, Redshift, EMR, Kinesis, DynamoDB, Step Function, SNS, SQS, Ec2
GCP Services: Cloud Storage, Datastream, Biguquery
Azure Services: Azure Data Factory(ADF), Azure Synapse Analytics, Azure Functions
Dashboard Tools: Metabase
Scheduler & Orchestrator : AirFlow
Development Methodologies & Tools : Agile, Scrum, Jira, Github

EXPERIENCE

Data engineer

Shopup **June 2023 - Present, Bangalore**

- Improved query performance and reduced costs by 70% by Transforming log tables to incremental structures , Implementing efficient partitioning and clustering strategies.
- Leveraged Docker containers for experimenting various data warehouse deployments.
- Evaluated and experimented with various data warehouse solutions.

Data Analyst/Engineer Intern

July 2022 - May 2023

- Developed analytical reports using metabase for operational teams.
- Liberated the operational team by 50% through building insightful reports.
- Created ETL pipelines using Hevo & airbyte.
- Developed Python scripts for cross-database querying (Metabase, BigQuery APIs), enabling improved data access and contributing to a 2 point increase in product team satisfaction.

PROJECTS

Cost-Optimized Data Stream Pipeline

- Designed and implemented a **cost-optimized ETL and CDC pipeline** that reduced BigQuery costs by **over 70%**.
- Previously, data was ingested directly into BigQuery from Datastream, incurring high costs due to inefficient table scans without partitioning.
- Developed a new mechanism by **loading data from Datastream into GCS** (zero loading cost) and using **Data Transfer Service** to ingest data into BigQuery raw tables (also zero cost).
- Configured **scheduled queries** to load data into a **partitioned Silver layer**, significantly minimizing scan costs.
- Shared this cost-optimization model with **Google's Data Engineering team**, who appreciated and drew inspiration from the approach during a collaborative discussion.

Building a Scalable CDC Data Lake with DynamoDB, Kinesis, and Athena

Shopup

- Architected and implemented a scalable Change Data Capture (CDC) data lake infrastructure using DynamoDB Streams, Kinesis Data Streams, and Firehose to ingest real-time data into Amazon S3. Optimized query performance by 40% with Parquet storage and Amazon Athena for analytics.
- Used AWS Lambda to transform event data before storing it in Parquet format on S3, reducing storage and query costs.
- Automated schema detection with AWS Glue Crawler and enabled ad-hoc analytics using Amazon Athena.

Developed Gold-Layer Data Transformation Using DBT

- Designed and implemented data transformation logic in **DBT on top of BigQuery** to create a Gold layer for analytics.
- Added extensive test cases in DBT to ensure data sanity, including validations for manually uploaded data, improving data reliability and accuracy. Streamlined **data quality checks**, enabling efficient monitoring and validation of critical business datasets.

High-Throughput Data Pipelines with Apache Flink

- Developed and deployed streaming data pipelines using Flink SQL to ingest data from MySQL/PostgreSQL into StarRocks, reducing ETL costs by 60% through the adoption of open-source technology.
- Explored and set up Apache Flink in the environment, leveraging its capabilities for efficient, real-time data processing.
- Optimized pipeline performance by **fine-tuning Flink parameters**, ensuring scalability and stability while maintaining cost efficiency.
- Delivered a robust and cost-effective solution for streaming data integration, enabling near **real-time analytics** for the first time in my organization.
- Developed custom Python scripts utilizing StarRocks migration tools to **automate the generation of Flink SQL statements**.

Automated Data Ingestion and Processing Pipeline using AWS

- Engineered a scalable pipeline that ensured seamless data flow from **AWS S3 to Amazon Redshift**, employing AWS Glue for data cataloging and transformation, which improved data consistency and reduced processing time by 40%.
- Configured **AWS EventBridge** to trigger an **AWS Step Function** when new data arrives from a third-party source.
- Developed **AWS Glue Crawler** to catalog data and **AWS Glue Job** to process incremental data.
- Optimized **incremental data loading** into Redshift for efficient storage and query performance.
- Implemented error handling and failure notifications using **Amazon SNS**, ensuring proactive monitoring.