Sirogiddin D.
Senior Senior Data Engineer, DataOps with ML & Data Science skills

SUMMARY

- Experienced Data Engineer and BI Developer with 6+ years of expertise in Database Design and Business Intelligence Development.
- Proficient in cloud technologies such as Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure.
- Skilled in building high-performance data integration and workflow solutions, including ETL operations for data warehousing and supporting OLAP, OLTP, and Data warehouse systems. Experience in optimizing DWH performance and automating data pipelines;
- Modern data engineer skills such as data modeling, data warehousing, data lake, data governance, and data quality.
- Experience with big data technologies such as Hadoop, Spark, and Kafka, and experience with data streaming and real-time data processing.
- Proficiency in SQL and NoSQL databases, Snowflake, and ClickHouse
- Data visualization tools such as Tableau or Power BI.
- Programming languages such as Python, Java, or Scala, and understanding of machine learning concepts, with experience building and deploying machine learning models.
- Experience with CI/CD, data governance, and security best practices.

TECHNICAL SKILLS

Main Technical Skills	Python (6 yr.), SQL (6 yr.), Apache Airflow, Apache Spark, AWS
Programming Languages	Python (6 yr.)
Java Frameworks	Apache Spark
Scala Frameworks	Apache Spark
AI & Machine Learning	AWS SageMaker, AWS SageMaker (Amazon SageMaker), TensorFlow
Python Frameworks	FastAPI
Python Libraries and Tools	Pandas, PySpark, TensorFlow
Data Analysis and Visualization Technologies	Airbyte, Apache Airflow, Apache Hive, Apache Spark, Azure Data Factory (2 yr.), Azure Data Lake Storage, Data Analysis Expressions (DAX), Databricks (2 yr.), ETL, Jupyter Notebook, Looker Studio, Pandas, Power BI, Sigma Compute, Superset, Tableau
	Apache Hadoop, Apache Hive, Apache Spark, Aurora, AWS Redshift, Clickhouse, dbt, DWH, Firebase Realtime

Databases & Management Systems / ORM	Database, HDFS, Microsoft Azure SQL Server, Microsoft SQL Server, MySQL, Oracle Database, PL/SQL, PostgreSQL, Snowflake, SQL (6 yr.)
Cloud Platforms, Services & Computing	AWS, GCP
Amazon Web Services	Amazon RDS, AWS Aurora, AWS Cloud9, AWS CloudTrail, AWS CloudWatch, AWS EMR, AWS Lambda, AWS Quicksight, AWS R53, AWS Redshift, AWS S3, AWS SageMaker, AWS SageMaker (Amazon SageMaker)
Azure Cloud Services	Azure Databricks, Azure Data Factory (2 yr.), Azure Data Lake Storage, Azure MSSQL, Databricks (2 yr.), Microsoft Azure SQL Server
Google Cloud Platform	Firebase Realtime Database, Google BigQuery, Google Cloud Storage
Deployment, CI/CD & Administration	CI/CD
Virtualization, Containers and Orchestration	Docker, Kubernetes
SDK / API and Integrations	FastAPI
Version Control	Github Actions
Logging and Monitoring	Grafana, Prometheus
Message/Queue/Task Brokers	Kafka
Other Technical Skills	Apache Kafka, database, DAX Studio, Google Cloud SQL, OpenMetadata, Relational, Spark EMR, Trino, Unix\Linux

WORK EXPERIENCE

Data Engineer, CPLUS DM Platform Migration

Duration: Jul 2024 – Present

Summary: Played a key role in migrating CPLUS DM-related tables and dashboard metrics, including WBR (Weekly Business Review) metrics, from another department into the central Data Warehouse (DWH), ensuring smooth data integration and system alignment.

Technologies: Hive, Spark EMR, Trino, Airflow, Superset, DWH, OpenMetadata, Qdrant, OpenAI API, Python, SARIMA.

Responsibilities:

- Data Migration: Led the migration of key CPLUS DM-related tables from a different department into the central DWH, ensuring data consistency and integrity during the transition;
- Dashboard Metrics Migration: Migrated critical dashboard metrics, including WBR metrics, ensuring accurate reporting for business stakeholders;

- Data Processing: Utilized Hive and Spark EMR for processing and partitioning large datasets to ensure optimized query performance post-migration;
- Orchestration & ETL Pipelines: Automated the migration and integration processes using Airflow, creating efficient ETL pipelines to manage the transfer of data and dashboard metrics;
- Data Governance: Ensured high data quality and governance standards using OpenMetadata to track data lineage and validate migrated data;
- Visualization: Updated Superset dashboards to reflect the migrated data and ensure that business-critical metrics for the CPLUS business were accurately reported.
- Integrated Qdrant vector database for storing embeddings and implemented a Retrieval-Augmented Generation (RAG) model using OpenAI API to enable semantic search and knowledge retrieval across internal documentation and data assets;
- Developed and deployed a SARIMA model to forecast time series data for the sales department, improving demand planning and resource allocation.

Achievements:

- Successfully migrated complex data structures and metrics to the central DWH with zero downtime;
- Improved dashboard reporting accuracy and efficiency, leading to better insights for business stakeholders;
- Streamlined ETL processes for ongoing data synchronization, reducing manual intervention post-migration.
- Enabled intelligent document retrieval and assisted insights generation through RAG-based NLP integration;
- Enhanced business forecasting accuracy through advanced statistical modeling.

Data Analytics Engineer, Big Data Processing and Analytics Platform

Duration: Dec 2023 - May 2024

Summary: At a leading investment company, we focus on harnessing the power of data to drive strategic decisions across CRM, fraud detection, transaction processing, and marketing. Our dedication to advanced data analytics helps us maintain a competitive edge in the dynamic financial sector. The project aimed to modernize our data warehouse capabilities by transitioning from a legacy DWH system to a state-of-the-art analytics platform. This upgrade was essential for handling large volumes of investment and transactional data and for supporting real-time analytics that inform critical business operations.

Technologies: ClickHouse, Kafka, Kubernetes, AWS S3, Snowflake, Airflow, DBT, Tableau, Python, SQL.

Responsibilities:

- Worked during the migration of a legacy DWH system to a modern analytics platform using ClickHouse, enhancing data processing capabilities.
- Designed and implemented a Kafka cluster on Kubernetes for real-time data ingestion to AWS S3, streamlining data flow and accessibility.
- Automated the importation of data into the Snowflake SRC layer from S3 using Snowpipes, facilitating efficient data management.
- Developed and orchestrated staging layer (STG) processes with Airflow DAGs, incorporating incremental DBT jobs for optimized data transformation in Snowflake.
- Scheduled hourly updates for data mart and model layers, supporting dynamic reporting needs.
- Utilized Tableau for insightful reporting, enabling data-driven decision-making across departments.

- Employed a variety of technologies, including Python, SQL, Snowflake, Kafka, Kubernetes, Airflow, and Tableau to enhance data analytics infrastructure.
- Integrated Amazon Bedrock to embed LLM-powered analytics and generate dynamic insights, summaries, and anomaly explanations directly within the AWS-based reporting environment.

Data Engineer, MultiSim

Duration: Jun 2022 - Nov 2023

Summary: Multisim is industry-standard SPICE simulation and circuit design software for analog, digital, and power electronics in education and research.

Technologies: MS SQL, PL/SQL, PySpark, Power BI, Kafka, Airflow, HDFS, Oracle.

Responsibilities:

- Building and maintaining data pipelines to support near real-time data processing for network and device performance monitoring and analysis using Apache Kafka and Apache Airflow;
- Designing and implementing data warehousing solutions to support network performance analysis, capacity planning, and traffic management using Oracle Database;
- Developing and deploying machine learning models to predict network outages, service degradation, and capacity requirements with Python and Airflow, Grafana;
- Ensuring data quality and integrity through data cleansing, data validation, and data lineage tracking using Apache NiFi;
- Implementing data security and privacy measures to ensure the confidentiality and integrity of sensitive customer information;
- Collaborating with cross-functional teams, including network operations, IT, and data science, to ensure data solutions meet business needs and technical requirements(Atlassian, Jira);
- Conducted a variety of data analysis and modeling tasks, utilizing statistical and econometric techniques to gain insights into business trends and customer behavior;
- Utilized PySpark to analyze and interpret customer insights, leading to a significant increase in profits for services;
- Developed advanced BI reports and dashboards using modern practices and migrated existing reports from Oracle and Excel to MS SQL Server and Power BI;
- Consulted with users on self-service reporting utilizing Power BI and made improvements to report performance using Oracle PL/SQL and MS SQL Server;
- Analyzed and interpreted large datasets using PySpark and data warehousing technologies, providing valuable insights for business decision-making;
- Conducted ETL of data using PL/SQL and SQL to support data analysis and reporting needs.

Data Analytics Engineer, End-to-End Data Platform for Investment Fund

Duration: Oct 2021 - May 2022

Summary: As a U.S.-based venture capital firm with a proactive investment strategy in early-stage and growth companies globally, we've backed industry leaders like Airbnb, Stripe, and Snap. Our focus on leveraging cutting-edge technology to make informed investment decisions sets us apart in the competitive venture capital landscape. The project aimed to develop and launch a comprehensive data platform for a startup investment fund within an enterprise environment. This initiative was crucial for integrating and analyzing data across various sources to support dynamic investment decisions and operational efficiencies.

Technologies: GCP, BigQuery, GCS, Airbyte, DBT, GitHub Actions, Airflow, Power BI, Sigma Compute.

Responsibilities:

- Spearheaded the development of a comprehensive data platform for an investment fund startup within an enterprise environment;
- Implemented a cloud-based solution using Google Cloud Platform (GCP), leveraging BigQuery as the data warehousing solution for optimized data analysis and storage;
- Integrated Google Cloud Storage (GCS) for secure and scalable data storage, enhancing data management efficiency;
- Utilized Airbyte for seamless data ingestion, ensuring near real-time data availability for analytics and reporting;
- Automated data transformation processes in BigQuery using DBT, coupled with GitHub Actions for continuous integration and deployment, improve data reliability and accuracy;
- Orchestrated and scheduled data workflows with Airflow (GCP supported), streamlining data operations and ensuring timely data updates;
- Employed Power BI and Sigma Compute as BI tools for cost-effective data visualization and analysis, offering versatile reporting solutions.

Data Analytics Engineer, TC Data Platform

Duration: Jul 2020 – Sep 2021

Summary: In the competitive telecom sector, our company not only provides communication services but also retails devices, dealing with complex data interactions across customer and dealer domains. Our focus on robust data analytics supports strategic decisions, enhances customer experience, and optimizes operations. The TC Data Platform project was centered on developing a scalable and secure data platform tailored to the needs of the telecom industry. This initiative aimed to support advanced business intelligence and machine learning capabilities while transitioning from traditional OLAP data warehousing to a more dynamic Snowflake-based data lake architecture.

Technologies: AWS, AZURE, Matillion, Airflow, DWH, MySQL, MSSQL, DMS, Snowflake, S3.

Responsibilities:

- Designed and implemented a scalable and secure data platform, utilizing AWS and Azure, to support business intelligence and machine learning initiatives;
- Developed and maintained several data pipelines to transfer data from various sources, including transactional databases, log files, and third-party APIs, to the data lake;
- Built and optimized data warehousing solutions, utilizing Amazon Redshift and Snowflake, to support reporting and analytics needs;
- Supporting legacy data platform, which is built on MS SQL as DWH in an Azure environment;
- Collaborated with data analysts and data scientists to develop data models, algorithms, and data visualizations, utilizing Tableau and Power BI;
- Built and optimized data warehousing solutions, utilizing Amazon Redshift and Snowflake, to support machine learning needs;
- Led a team of data engineers in the design and implementation of data-driven solutions, utilizing Agile methodologies and project management tools, such as JIRA;
- Conducted regular code reviews, performance tuning, and quality assurance to ensure the data platform meets business needs and technical standards;
- Mentored and trained junior data engineers on best practices for data engineering, data modeling, and data security;

- Implemented data security and privacy measures, including encryption, firewalls, and access controls, to ensure the confidentiality, integrity, and availability of sensitive data;
- Performing the role of Data Analyst with Data Engineer capabilities;
- Implementing data quality checks and processes to ensure the accuracy and consistency of data;
- Building and optimizing data pipelines using Python and SQL, as well as big data technologies like Spark, interacting with Data Lakes;
- Creating interactive dashboards and reports in Power BI to present insights and findings to stakeholders, leveraging the data visualization capabilities;
- Applying statistical models and performing data analysis using Python to draw conclusions and support business decisions;
- Collaborating with cross-functional teams, IT, and data science to ensure data solutions meet business needs and technical requirements(MS Teams, Atlassian).

ML/Data Engineer, Under NDA

Duration: Sep 2019 - June 2020

Summary: Implemented a scalable and secure data platform and maintained data pipelines for machine learning initiatives.

Technologies: AWS, GCP, S3, Airflow, Python, SQL, Spark.

Responsibilities:

- Designed and implemented a scalable and secure data platform utilizing AWS, Google Cloud Platform, and other cloud technologies to support machine learning initiatives;
- Developed and maintained data pipelines using Apache Spark and Apache Airflow to transfer data from various sources, including transactional databases and log files, to the data lake in Amazon S3 and Google Cloud Storage;
- Collaborated with data scientists and machine learning engineers to develop machine learning models, utilizing Amazon SageMaker, to solve business problems and improve business outcomes;
- Implemented data security and privacy measures, including encryption, firewalls, and access controls, to ensure the confidentiality, integrity, and availability of sensitive data;
- Conducted regular code reviews, performance tuning, and quality assurance to ensure the data platform meets technical standards and business needs;
- Stayed current with the latest trends and best practices in cloud technologies, machine learning, and big data technologies, including Hadoop, Spark, Kafka, and data streaming.

ML/Data Engineer, FinTech

Duration: Feb 2019 – Aug 2019

Summary: Managed financial data processes, machine learning model development for fraud detection, credit risk analysis, and more.

Technologies: AWS, QuickSight, SageMaker, Airflow, Python, SQL.

Responsibilities:

- Building and maintaining a data platform for financial data processing, storage, and analysis using AWS cloud services, and technologies such as Spark and Kafka;

- Designing and implementing ETL pipelines to extract data from various sources, transform it into the required format, and load it into a data warehouse for analysis using tools such as Apache Spark and Apache Airflow;
- Implementing data security and privacy measures such as encryption, firewalls, and access controls to ensure the confidentiality and integrity of sensitive financial data;
- Developing and deploying machine learning models for fraud detection, credit risk analysis, and investment recommendations using tools such as Python, FastAPI, and machine learning frameworks such as TensorFlow;
- Collaborating with data scientists and business stakeholders to identify and define data requirements for specific use cases, and to develop data-driven solutions that address business challenges;
- Keeping up to date with the latest trends and best practices in financial technologies, data engineering, and data science;
- Utilizing AWS SageMaker for building and deploying ML models and visualizing results through Amazon QuickSight to validate and analyze predictions.

ML/Data Engineer, Retail

Duration: Jul 2018 – Jan 2019

Summary: Handled sales and marketing data platform, creating visualization dashboards for sales metrics.

Technologies: DAX, PowerBI, MSSQL, Pandas, Python, Jupyter notebook.

Responsibilities:

- Building and maintaining a data platform for sales and marketing data processing, storage, and analysis using Google Cloud Platform;
- Creating dashboards and reports to visualize sales and marketing metrics using Looker;
- Keeping up to date with the latest trends and best practices in sales and marketing analytics, data engineering, and data science.

EDUCATION

- Bachelor’s degree in Nuclear Physics and Technology, MEPHI Tashkent Branch, Uzbekistan