Data Engineer

Adinarayana Sangala +91-9573530034

aadisang400@gmail.com

Professional Summary:

- Overall, I have 4.8 years of professional experience in Clinical Trail industry, with strong debugging and troubleshooting skills in the technologies listed below.
- I am trained in Data Science and Data Engineering, with experience in big data analytics and batch processing using Snowflake, SQL Server, Azure Cloud, AWS Cloud, PySpark, and Spark SQL, with a focus on ETL operations across Snowflake, on-premises systems, and Azure cloud environments.
- I have hands-on experience working with Delta Lake using a medallion architecture. I have processed data using Azure Databricks (ADB) and Amazon Databricks, leveraging Amazon S3 and Azure Data Lake Storage Gen2 (ADLS Gen2) for storage. Additionally, I have practical experience with Azure Machine Learning Studio, AWS SageMaker, EC2, and Microsoft Virtual Machine services.
- To process data in Delta Lake, various commands and optimization techniques were employed to create and manage data frames and partitions. These included repartitioning, performing joins, applying window functions, and cleaning data stored in Parquet format. The structured data was then properly formatted and saved using PySpark code in notebooks, along with Spark SQL queries.
- I am highly proficient in data transformations using PySpark and Spark SQL queries within notebooks. I have extensive experience with data flow components in Azure Data Factory, including Transformation activities such as Lookup, Conditional Split, Join, Sort, and Aggregate.
- My expertise extends to various Azure Data Factory activities, including Copy Data, Stored Procedure, Get Metadata, Lookup, scripting, conditional logic, filtering, looping, pipeline execution, and the use of variables and variable appending. Additionally, I have a strong background in parameterizing datasets, linked services, Integration Runtime services, and triggers notebooks, and pipelines within the Azure Portal.
- My responsibilities included developing PySpark code, writing Spark SQL queries, and handling complex data transformations with exception handling. I also managed data in Azure Synapse Analytics and utilized Azure Pipelines for orchestration, with monitoring and notifications via Azure Data Factory.
- I have good knowledge in utilizing Python, Pyspark, Spark-SQL in Azure Portal. The Prepared Dash boards by Power BI Reporting Tool. I have worked on Power Query, DAX, View, Services.
- My strengths include strong analytical, diagnostic, and troubleshooting skills, along with the ability to excel in a dynamic, team-oriented environment.
- I have experience working within an Agile sprint methodology framework, participating in Scrum Meetings as part of a cross-functional team.
- I utilized the JIRA tool for reviewing client requirements and tracking issues and bugs.

Academic Profile:

Master of Pharmacy in **Acharya Nagarjuna University** /Guntur, A.P/ Pass-2015-2018 / CGPA-6.3.

Technical skills:

Languages **:** Python, PySpark, SQL Server, My SQL.

Cloud Technologies **:** Azure Portal, Microsoft Fabric & AWS.

Azure Services **:** Azure Data Bricks, Azure Data Lake, Azure Data Factory, Azure ML Studio, Data Science and Artificial Intelligence, utilizing Python, Machine Learning (ML), Natural Language Processing (NLP), Deep Learning via Google Colab Notebook, Azure Data Bricks, Azure Blob Storage, Azure Data Lake, Snowflake.

AWS Services **:** IAM, EC2, Aws S3, Aws Glue, Aws Lambda, Aws CloudWatch, Aws SNS, Data Bricks, Aws Athena, RDS, Aws RedShift, Aws SageMaker, QuickSight.

Frameworks&Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit-Learn, SciPy, TensorFlow, Flask Deployment, NLTK, Spacy, Keras, OpenCV.

Tools : SQL Server, T-SQL, Azure Portal, Azure Data Bricks, Snowflake, Airflow, Power BI Reporting, Jira App.

Experience Summary:

Qinecsa Solution(Bioclinica)India.Pvt.Ltd/Drug Safety Associate-II/From Sept2021 - Mar2023

Cohance (CR Bio) CRO Pvt.Ltd/Data Analyst / From Jan 2021 to Aug 2021.

Clinsyc CRO Pvt Ltd./Executive/ from Jul 2018 to Nov 2020.

Projects Undertaken:

Azure Cloud Project

Client: PFIZER

Environment: Azure Databricks, Python, PySpark, Snowflake, SQL Server, Azure Data Factory, Azure Data Lake Gen2 Storage, Azure Synapse Analytics, Azure Machine Learning Studio, Power BI, Agile Scrum Methodology, Jira App.

Role:Data Engineer

Description:

Pfizer is an American multinational pharmaceutical and biotechnology corporation that specializes in the development of medicines and medical devices. The company, founded in 1849, is headquartered in New York, with its consumer division also based in New York. Pfizer operates through approximately 49 subsidiary companies, conducting business in over 158 countries, and its products are distributed in more than 181 countries.

Responsibilities:

- Created a staging database for temporary data processing, data consolidation, and database loading using Azure Data Factory.
- Developed various activities in Azure Data Factory, including copy activities, stored procedure activities, and custom activities to process data.
- Established an Azure Data Lake Store for data storage and data processing using Data Lake Analytics.
- Deployed tabular models into Azure Analysis Services.
- Utilized Azure data storage services such as Azure Blob Storage, Azure SQL Databases, Azure Data

Warehouse, and Azure Data Lake Store for data storage.

- Implemented SCD1 (Slowly Changing Dimension Type 1) transformations using Azure Data Factory.
- Implemented incremental/delta loading strategies in Azure Data Factory.
- Designed and implemented a data migration mechanism from the existing on-premises SQL Server system to Azure.
- Designed Azure SQL Database tables to store transformed data.
- Conducted data cleaning operations, including variable transformations and handling missing values, using Azure Databricks (ADB) with PySpark.
- Gained experience with Azure Databricks (ADB) by creating clusters and notebooks for data pre-processing, leveraging PySpark and Python libraries for processing large datasets.

AWS Cloud Project

Client: American International Group (AIG)
Environment: Python, SQL Server, PySpark, IAM, EC2, S3, Aws Glue, Lambda, CloudWatch, Aws SNS, Data Bricks, Aws RDS, RedShift, Aws SageMaker, Power BI, Snowflake, Agile Scrum Methodology, Jira App.
Role: Data Engineer
Description:

American International Group (AIG) is a leading global insurance organization with operations in approximately 80 countries and jurisdictions. They provide a wide range of property casualty insurance, life insurance, retirement solutions, and other financial services to support clients in both business and personal aspects through their General Insurance and Life and retirement business units.

Responsibilities:

- Structured data is securely transferred through SFTP and migrated into AWS RDS using Excel and CSV source files stored in Amazon S3, with access control managed by AWS IAM.
- The files stored in the S3 bucket were analyzed by AWS Lambda to detect triggered activities, with monitoring performed using Amazon CloudWatch.
- The files stored in the S3 bucket were analyzed using the Medallion architecture, transformed and processed through AWS Glue pipelines, and monitored using Amazon CloudWatch.
- The Glue pipelines were utilized to perform multiple data-cleansing activities within the Medallion architecture using Databricks with PySpark and SQL queries.
- In Databricks, using PySpark and SQL scripts, I aggregated and scaled Delta tables, performed joins between them, and managed nodes and data partitions through incremental data ingestion.
- The pipelines failed during execution due to schema mismatches and data corruption, and failure notifications were sent via the SNS service.
- The curated data is stored in Amazon Redshift for reporting and machine learning purposes. It is used for visualization in Power BI and Amazon Quick Sight, as well as for data science workflows in AWS SageMaker.
- The external objects are sourced from an S3 bucket and accessed through Athena and the Snowflake database.

Professional Certifications:

- Certified Data Engineer in **Microsoft Fabric (DP-700)** with strong proficiency in Lakehouse design, Spark processing, Delta tables, and end-to-end data pipelines.
- Full stack Data Science & AI With **Azure Cloud** from Naresh IT.
- Cloud Computing with **AWS** From GDSC KIIT.

Place: Hyderabad (Adinarayana)