Maliha Sameen

Computer Scientist - Data Engineer

☎ 0333-4576325 | ✉ malihasameen58@gmail.com | malihasameen | maliha-sameen

BS Computer Science from SEECS, NUST | CGPA 3.52/4.00

Education

Bachelors of Computer Science
2016 - 2020

National University of Science & Technology

Experience

Bentley Systems
Feb 2021 - Present

Data Engineer

- Performed data cleaning, manipulation and anonymization tasks using SQL and Python based scripts.
- Designed multiple end to end data pipelines with Azure Synapse Analytics and Azure Data Factory to perform ELT operations on data from Azure storage services and integrated with AML pipelines, serverless Azure Functions and internal data platforms.
- Defined Spark UDFs for specific data manipulation requirements.
- Designed data storage solutions using Azure Data Lake storage with lifecycle management policies.
- Designed scalable data models and developed PowerBI dashboards using DAX based calculated tables and measures with data cleaning and manipulation using Power Query M.

Rapidev DMCC
Nov 2020 - Feb 2021

Python Developer

- Wrote scraping scripts to scrape data from social media platforms e.g. instagram, etc. and search engines e.g. google, baidu, etc.
- Maintained Rabbit MQ server to get scraping requests from product website backend and entertain those requests accordingly.
- Developed and maintained the overall data pipeline to get scraping requests, scrape data and store data into HDFS storage with Hadoop server hosted on-premises.

VisionX Technologies
June 2019 - Aug 2019

Machine Learning Intern

- Worked on ’Sender Receiver Address Area Localization’ system for PackageX receipts
- Built supervised pipeline for sender-receiver region localization system using deeplearning, with data labeling & preprocessing

TUKL Research & Development Lab
June 2019 - Aug 2019

Research Intern (AI & ML)

- Worked on a DAAD funded project titled, ’Water Body Segmentation through remote sensing.
- Built supervised pipeline for accurate water body extraction around boundary and achieved dice score of 86% on Pakistani dataset.

Projects

Product Portfolio Analysis

- A PowerBI dashboard that uses Product Basket Analysis to simplify product portfolio.
- Visualized processed data from Snowflake and CSV files to highlight cross-sell and bundling, dependence and risks associated with deprecation or replacement of products.

Real Time Sales Streaming

- An end to end sales streaming data pipeline utilizing Docker containerization for seamless environment setup.
- The pipeline comprises of multiple layers, including data ingestion, message brokering, stream processing, serving databases, and visualization.
- Leveraging FastAPI & Kafka, the pipeline enables robust data ingestion layer with JSON validation. Spark streaming process, efficiently transform and load data into MySQL & Cassandra persisting aggregated & raw data. Superset connects to MySQL, providing dynamic visualizations of sales data.

Labelling Platform

- Configured Azure Synapse, ADLS and Azure Key Vault and designed data storage strategy in ADLS to allow monitoring.
- Created ELT pipeline on Azure Synapse using REST endpoints to extract data, load into Azure Data Lake, tranform and call Azure ML pipeline to get predictions and load pre-labelled dataset in ML Platform (REST layer on ADLS allows for RBAC).
- Integrated data pipeline with Azure Function and defined UDFs inside Azure Synapse Spark notebooks for data transformation.

Media Indexing Search Trends

- A PowerBI dashboard that uses Past 90 days of custom search event logs in Application Insights, to show search trends for Media Indexing Search Engine.

Technical Skills

Languages: SQL, Python, C/C++, Java

Big Data & Cloud Tools: Spark, Snowflake, Kafka, Azure AD, Azure Data Lake, Azure Blob Storage, Azure Synapse, Azure Data Factory, Azure Key Vault, Application Insights, Azure Functions

Visualization: PowerBI, Superset, Streamlit, Matplotlib, Plotly

Databases: MySQL, MongoDB, Cassandra

DevOps: Git/Github, Docker, Azure DevOps

AI/ML: Tensorflow, Keras, Pytorch, OpenCV, NLTK, Scikit-Learn

Others: REST API, Flask, Linux, Jupyter Notebook, Postman, Android Studio, Web Scraping with Selenium, Fast API

Courses & Certifications

- Introduction to Data Engineering, **Data Camp**
- Hands on Essentials, **Snowflake**
- Microsoft Office Specialist, 2013, **Microsoft**
- Databricks Lakehouse Fundamentals, **Databricks**
- Introduction to Computer Science CS50, **Harvard**
- Deep Learning Speciaization by Andrew NG, **Coursera**

Achievements

- STAR Award Q1 2022 Advance Analytics Champion, **Bentley**
- Bronze Medal Code Sprint 3, **Hackerrank**
- Rank 10 Identify the apparels, **Analytics Vidhy Datahack**