SHIKA SHYAM

Boston, MA | (617) 510 - 6560 | shika.shyam@yahoo.com | www.linkedin.com/in/shshyam | https://shikashyam.github.io

EDUCATION

NORTHEASTERN UNIVERSITY **Boston, MA**
Master of Science in Information Systems | GPA : 3.87/4.0 **Expected May 2023**
Relevant Coursework: Data Mgmt & Database Design, Data Science Engg Tools & Methods, Big Data Systems & Intelligence Analytics

VELLORE INSTITUTE OF TECHNOLOGY UNIVERSITY **Vellore, India**
Bachelor of Technology in Computer Science Engineering | GPA : 3.99/4.0 **May 2018**
Relevant Coursework : Data Structures & Algorithms, Data Mining & Data Warehousing, Management Information Systems

KEY SKILLS

Programming Languages : SQL, C, Python, Java, C++, BigQuery, HQL
Databases : Oracle SQL, Oracle 12g, PostGRE SQL, MySQL, Google Big Query, Hive, Hadoop, Firebase
Data Integration and Business Intelligence Tools : Oracle Data Integrator, Looker, PowerBI, Tableau, Data Studio
Machine Learning : Data Exploration, Keras, TensorFlow, scikit-learn, PyTorch, CNNs, Neural Networks, HuggingFace, SHAP
Enablement tools : GCP, Microsoft Azure, AWS ECR, AWS Lambda, Docker, GIT, Airflow, Apache Beam, Azure DevOps, Pub/Sub

PROFESSIONAL EXPERIENCE

WAYFAIR LLC. **Boston, MA**
Data Engineer (Co-op) **July 2022 – (expected to complete by December 2022)**

- Engineered a proof of concept in Python to **automate reporting** for the Finance team by querying the Data Platform in **BigQuery** directly and providing results in Excel workbooks hence **enabling** the **depreciation** of long outdated and expensive OLAP cubes
- Created **ETL Script** to filter **CSV files** based on upload date from **GCS Bucket** and move it to **BigQuery** tables eliminating manual effort
- Facilitated easy access to data for Marketing team stakeholders by designing and developing **Looker explores,** creating Measures and Dimensions and defining relationships between **20+** marketing data tables

DELOITTE CONSULTING **Bengaluru, India**
Consultant – Data Engineer **August 2018 – August 2021**

- As a **functional SME** for Customer Data, **Reverse Engineered 7 Data Models and ETLs** from **23 complex and non-optimized SQL queries** frequently used by the Data Scientist team **reducing the performance overhead** of running heavy queries redundantly
- Confidently **drove conversations** and **built consensus** among client, downstream and upstream data consumers to create an end-to-end data processing pipeline for a new **unstructured data source**. Optimized the ETL using **reusable mappings**, **staging tables** and smaller scripts running in **PySpark** to bring down run time from **9 hours to 12 minutes**.
- **Hot-fixed a wide-spread issue in production** in all timestamp conversion columns occurring across **12 ETL** mappings by **automating an SQL script** to make schema changes and data corrections **in less than 24 hours** compared to manual effort of 4 days
- Created alternate code base of ETLs to run in **PySpark** instead of Hive during specific client operating conditions to **refresh data every hour instead of daily**, orchestrated the pipeline to **automatically switch from daily run to hourly run** based on a flag value.
- Developed a Python Script to automatically **scrape** Integration testing queries from **Azure DevOps** VSTS tickets to create a unified tracker to store all testing queries and automatically run these scripts for any object for regression testing or production backfills validation
- Created a **Chatbot** using **Dialogflow** and **GCP Cloud functions** to read handwritten medical PDF reports using open-source **OCR** libraries and extract fields of interest into a **POSTGRES SQL database**
- Spearheaded **4+ weeks** of intensive training, evaluations and technical support to get **30+ colleagues** certified in **Azure and Oracle Cloud** technologies and presented **Knowledge Transfer** sessions on **Azure Cloud**, **Power BI**, **Oracle Data Integrator** and **GCP**

RELEVANT PROJECTS

END-TO-END DATA PIPELINE FOR STORM FORECASTING USING SATELLITE IMAGERY **Jan 2022 - Feb 2022**

- Using SEVIR dataset provided by the NOAA, pakcaged the nowcasting model as a **model-as-a-service** with a robust front-end built in **Streamlit** by exposing endpoints on **Fast API with JWT keys authentication** hosted on **GCP App Engine**
- User inputs a location or date and time on the **Streamlit WebApp** and gets storm forecast for the next hour. Frequently queried locations would return results faster due to **caching** implemented with **hourly orchestrated Airflow** jobs running on **GCP Cloud Composer**
- **Dockerized NLP models** (**Summarization and Named Entity Recognition**) deployed on **Amazon ECR** using Lambda functions and **Serverless framework** return the summarization of the Storm Event description and named entities.

TWITTER MARKETING CAMPAIGN ANALYSIS USING REAL-TIME DATA PIPELINE **Mar 2022 – May 2022**

- Based on the hashtag input and date range given by the user, the WebApp can analyze the relevant tweets, utilizing **Huggingface**'s **Sentiment Analysis** and **Named Entity Recognition** to **model-as-a-service** deployed on **AWS ECR** as **Docker containers** on the **Serverless framework** and **AWS Lambda** to perform Sentiment analysis on tweets and derive entity names
- Locations from the named entities were used to find and display relevant news using an **open-source News API** along with real-time metrics on a Google Data Studio dashboard embedded in the WebApp.
- **Ingested Twitter Stream** using Twitter API 2.0 using Python Tweepy library and **Google Pub/Sub** and associated **Cloud functions** to Store Raw data in **Google BigQuery** all packaged within **Airflow** running on **Google Cloud Composer** (**Extract part of ETL**)
- For **Transform and Load in ETL**, leveraged **Apache Beam** to set up an **ad-hoc pipeline** and an hourly pipeline ingesting raw data, cleaning and loading into a processed **BigQuery dataset** using **Google Cloud Dataflow** as the orchestration tool for Beam.