Sachendra Chandra

6394633084, 9410411981 07sachendra@gmail.com LinkedIn GitHub

Work Experience

Data Engineeer 2, Rakuten India **August 2021 – Present**

Mart Migration

- Part of data platform migration project from Teradata data to Bigquery. Converted the Systemwalker Jobnet into Airflow Python Scripts. These jobnets helped in creating the data marts in Teradata. The corresponding Airflow scripts orchestrated the same data mart creation in BigQuery.
- Analyzed the translated Teradata SQL to BigQuery SQLs and the data marts created in BigQuery to ensure exact match between the data of Teradata that with BigQuery and automated the validation process for translated queries on Google BigQuery and Hive and involed in data ingestion from Teradata to BigQuery via Jenkins.
- Part of migration task from C5000 to Minio and Bigquery. Translated the Airflow scripts for daily ingestion from Hive to Minio and BigQuery.
- Skills Used: Python, SQL, Apache Airflow, BigQuery, Teradata, HiveQL

POC of Data Mgmt. Tool like Spline and DataHub

- Integrated Spline Data Lineage tool with existing code base to track Lineage of Spark job used for mart creation in BigQuery and Minio.
- Worked on creating a Custom dispatcher using the spark spline agent code base so as to extract relevant lineage information for Spark mart creation jobs.
- Explored DataHub for capturing, tracking, and visualizing data lineage, allowing organizations to gain valuable insights into their data ecosystems.
- Skills Used: Python, Spline, DataHub, Apache Spark

Internal Metadata Management

- Worked on building framework to capture the dataset and datasource lineage and dataset metadata for spark job used for mart building.
- Developed REST Client SDK that leverages Hive Metastore service for capturing matadata information of spark job and dispatching Metadata and lineage information to API layer.
- Skills Used: Apache Spark, Hive, ArangoDB, Docker

Operational Artificial General Intelligence (OAGI)

- Thousands of jobs running every day and they consume cores, memories, storages and network bandwidth to full-fill their functional goals; those cores, memories, storages and network bandwidth are having the cost based on usage; if those processes are unnecessarily retried and rerun in a circumstance where errors are non recoverable, business has to pay unnecessary cost for those resources consumed by processes.
- Before retrying, the process asks for the opinion from the OAGI service which is powered by LLM(Large Language Model) which has given the enough knowledge of what kind of exceptions/errors are re-triable and non re-triable. This saves many resources and directly impacts business revenue.
- Skills Used: Langchain, Llama model, Huggingface

Education

Indian Institute of Science, Bangalore **Aug. 2019 – July 2021**

Master of Technology in Computer Science

Institute of Engineering and Technology, Lucknow **Aug. 2015 – July 2019**

Bachelor of Technology in Information Technology

M.Tech. Thesis (GitHub Link)

Transformer Models for Assertion Generation in JAVA Unit Tests **Feb. 2020 – June 2021**

Guide: Prof. Aditya Kanade (SEAL Lab) *IISc. Bangalore, Karnataka*

- Automatic Assert Statement Generation for Java Unit Test Cases using multiple Transformers Models Fused with Pretrained CuBERT (**C**ode **U**nderstanding **BERT**) Encoder - BERT Large Transformer Encoder by Google.
- Fed the JUnit Test methods as input to the CuBERT encoder and used it's output as sentence embeddings for the Junit Test method. Fed this output as input to the modified Tranformer models.
- Used drop-net technique to outperform the baseline models.

Projects

AI for Code

* Developed a Langchain application that scrapes any code repository and uses it's source files as a Knowledge document for creating a general purpose application for developers.
* Made use of quantized CodeLama model to leverage the existing code base along with Retreival Augmented generation mechanism and Chroma database to generate required results.
* The developers could ask general questions in Natural language and even write down unit tests for the existing source code

ML Pipeline for Wine quality application

* Build an Flask web app to detect wine quality using ElasticNet Regression along with Data Version Control(DVC) framework for tracking versions of models, data and ML pipelines.
* TOX framework was used for creating testing modules for the application code. The app is deployed on cloud and can be accessed via this **link**

Data Pipeline for Kaggle Datasets

* Written spark programs to read the kaggle dataset and saved the data to partitioned parquet format and automated some transformation and analytics spark job using airflow and written unit test using Funsuite for analytic logic.
* Created a new SBT project with all dependencies and builded the jar using SBT assembly.

Data Collection and Modeling NMT model for Source code to Natual Language comments

* Cloned most forked Python Repositories from GitHub and recursively extracted Python files from each repository and writing Python Script to scrape Functions and corresponding Docstrings form each Python File.
* 2,22,513 (Function, Docstrings) pairs were extracted for training Machine Translation Model to generate Docstrings for corresponding Python Functions using Transformer Encoder-Decoder Architecture with Self-Attention mechanism.

Movie Recommender System

* Involves predicting the unknown ratings(ranging between 0.5-5) given a user-movie pair (Netflix Prize Dataset) using concepts of Collaborative Filtering, Non Negative Matrix Factorization(NNMF), Soft Impute etc.

Image Captioning

* The project involves giving textual caption to an image with multi-model and Transfer learning technique using both RNN and CNN with pretrained VGG16 model.
* The frontend of the web app is built using Flask and backend is created using LSTM python code and restructuring VGG16 pretrained model.The Docker image of the app is deployed on cloud platforms like Heroku or Azure for complete deployment.

Other Key projects

* XGBoost Regression, ARIMA Forecasting on COVID19 data, Quora Questions Auto-complete suggester, Question Answer retrieval system, Extractive Text Summarization

Internship

* Cyber Security internship at BSNL ALTTC Ghaziabad, U.P. (Nov. 2017- Dec 2017)
* Performed Case study on Network Security and Networking by implementing security protocols like TELNET,SSH along with routing configurations and implementing cyber attacks such as SQL injection, Cross-site scripting, Denial of Service and Buffer Overflow attack at RPI Consultants Pvt. Ltd. (Jan 2019- Feb 2019)

Technical Skills

Languages: C, Python, Java, SQL
Libraries: Tensorflow, Pytorch, Keras,Data Version control(DVC)
Technologies/Frameworks: Hadoop, Apache Airflow, BigQuery, Kafka, PySpark, Docker, LangChain
ML Algorithms: Decision Tree based models, Attention Networks, Transformer models, BERT