Biraj Parikh

Google Cloud Certified Professional Data Engineer

Bloomington, Indiana | 812-650-2716 | birajparikh16@gmail.com | GitHub | LinkedIn | Portfolio

EDUCATION

Indiana University, Bloomington, USA | Master of Science in Data Science | **GPA: 3.9** **August 2019 - December 2020**

University of Mumbai, India | Bachelor of Engineering in Mechanical Engineering | **GPA: 3.7** **August 2013 - May 2017**

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Intern | Apothecary, Inc | Massachusetts**,** USA 6 months | **May 2020 - present**

- Successfully increased reviews database by 70 percent by scrapping over 200000+ product reviews and ratings.
- Implemented a niche **Collaborative Filtering** model to provide personalized product recommendations based on the user skin input which are stored on Amazon RDS database.
- Orchestrated and containerized an ETL pipeline using Apache Airflow and Docker for scheduling, monitoring, and troubleshooting issues when needed.

Advanced tech skills & tools: Python, PostgreSQL, Apache Airflow, Docker, AWS RDS, FASTAPI, Tableau.

Data Engineer | Reliance Jio Inc. | Mumbai, India 1 yr 2 mos | **May 2018 - June 2019**

- Streamlined reliable and robust data pipelines to ingest, transform, and process petabytes of customer-centric streaming data using **Spark Structured Streaming** and **Apache Kafka** resulting in a 20 percent redundancy reduction.
- Leveraged Asia's biggest On-Prem Hadoop **Data Lake** for efficiently storing and accessing transformed data.
- Reduced code execution time from 20 mins to 5 mins by optimizing and tuning Spark Jobs using Scala which resulted in significant improvement in the overall performance.
- Reported the analyses by developing a real-time & detailed dashboard on ZoomData for making data-driven decisions.

Advanced tech skills & tools: Apache Spark, Spark Streaming, Scala, Hadoop, Apache Kafka, Hive, Nifi, Apache Airflow, ZoomData.

Data Science Intern | Piramal Corporate Service Ltd | Mumbai, India 7 mos | **October 2017 - April 2018**

- Implemented a predictive and prescriptive model for a **Fraud Detection** use case to predict the feasibility of debtor loan repayment, utilizing past loan history and customer behavior metrics to make a tangible business impact.
- Innovated web-scraping framework as another measure to validate user information accounting for 5% of the business decision.

Advanced tech skills & tools: Python, R, OOPs, Selenium, Microsoft Excel, PowerPoint, Word.

SKILLS

- **Programming Languages:** Python, R, PySpark, Scala, PyTorch, SQL, Tensorflow, Keras, shell-scripting
- **Databases:** MySQL, PostgreSQL, MongoDB (NoSQL), Cassandra, Amazon Redshift
- **Libraries:** Pandas, Numpy, Scikit-learn, NLTK, Requests, Matplotlib, tidyverse, ggplot2, dplyr
- **Machine Learning:** Classification, Regression, Clustering, Neural Networks, Anomaly Detection, Forecasting, CNN, Dimension Reduction, Natural Language Processing, Recommender Systems
- **Framework/Tools:** Apache Spark, Spark Streaming, Hadoop, Hive, Apache Kafka, Apache Airflow, Tableau, Git, GitHub, Flask, Docker, Kubernetes, Amazon Web Service (AWS), Google Cloud Platform (GCP)

PROJECTS

Real-Time Virtual Store Data Analysis (GCP, Pub/Sub, Apache Beam, Dataflow, Google Data Studio, Dash) **October 2020**

- Utilized GCP Pub/Sub and Apache Beam deployed on Dataflow for ingesting streaming data from virtual online store designed using Dash Plotly and saved aggregated results on Cloud SQL for downstream applications.
- Developed a dashboard on Google Data Studio to communicate insights and perform analytics in real-time.

Real-Time Server Status monitoring (Spark, Hadoop, Kafka, PostgreSQL, Tableau, Docker, PySpark) **September 2020**

- Engineered an ETL data pipeline using Apache Spark and Kafka for processing and monitoring the data center's event status in real-time and reporting the resolution in case of issues occurring, for better server stability.
- Built a Tableau dashboard that shows the data center's event and resolution time status in real-time across the world.

Music Data Analysis on AWS (Spark, Apache Airflow, AWS Redshift, AWS EMR, S3, Star Schema model) **August 2020**

- Developed an ETL pipeline which extracts data from the data lake hosted on S3, stages them in Redshift, and transforms into a set of dimensional tables using Spark application deployed on AWS EMR cluster.
- Orchestrated the ETL pipeline using Apache Airflow to schedule and routinely monitor the workflow.

Human Protein Multi-Label Image Classification (PyTorch, Convolutional Neural Networks, Transfer Learning) **June 2020**

- Implemented ResNet34 model architecture, to identify and classify (multilabel classification) mixed patterns of proteins in microscopic images to accelerate biomedical image analysis.
- Optimized the model performance using regularization and state-of-the-art techniques like Transfer Learning, learning rate finder, augmenting, batch normalization, gradient clipping which improved the accuracy up to 87 percent.