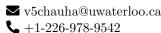
Vivek Chauhan



in linkedin.com/in/v5chauhan github.com/chauhan-vivek

EDUCATION

University of Waterloo
 Master of Engineering in Computer Engineering (Specialization: AI/ML); Percentage: 89 %

 Guru Gobind Singh Indraprastha University
 Bachelor of Technology in Electronics and Communication Engineering; GPA: 8.85/10.0

 EXPERIENCE
 Waterloo, Canada Jan. 2022 – Apr. 2023
 New Delhi, India Aug. 2015 – May. 2019

UbiLab, University of Waterloo

Waterloo, Canada

Data Scientist

Jul 2022 - Present

- Deep Learning: Researched, analysed and modeled novel algorithm to quantify quality of sleep using LSTM and Variational AutoEncoders.
- Machine Learning: Collaborated to model impact of Covid and heatwaves on 14k households sensors data, implemented multi-feature, time-series clustering in distributed architecture.
- **Big Data**: Created ingestion and feature engineering data pipelines using Pyspark and Azure Databricks to analyze behavioral patterns in over 7TB of IoT data.

ZS Associates

Gurugram, India

Data Engineer

Jun 2021 - Dec 2021

- Data Warehousing: Integrated 4 vendors across 10 products in pharmaceutical space in cloud-based data warehouse solution hosted using AWS services, optimized SQL quering time across large datasets by 15%.
- Data Migration: Migrated Oracle database to Redshift, performed historic refresh for claims, sales, and MCM healthcare data to analyze competitive products over last 10 years.
- Transaction Processing: Designed schema to schedule and track ETL jobs and tasks, integrated detailed logging for easier access to operation teams.
- ETL: Developed batch ingestion pipelines applying business rules using SnapLogic and AWS Glue, built a framework to prototype and test pipelines before deployment.
- Master Data Management: Architectured MDM system for conforming data from 12 vendors using Pyspark, python, PostgreSQL and bash on AWS infrastructure, reduced turnaround time for one cycle from 8 hrs to 1 hr.
- QC Framework: Built exhaustive QC framework using PostgreSQL applying data source-specific business rules.

Airtel Data Analyst

Gurugram, India

Jun 2019 - Jun 2021

- Network IAM: Constructed end-to-end network identity solution employing Elasticsearch-Logstagsh-Kibana stack to map resource utilization and traffic at over 40 proxy server points.
- Analytics: Integrated mechanisms for suspicious activities and potential points of failure detection for every session user using Telegraph - InfluxDb - Grafana stack, cutting down malicious attempts into network by 30%.
- Data Analytics: Built real-time, pan-India dashboard in Kibana, enhancing network performance by 5% due to insights generated from root cause analysis of uncorrelated network events using Elasticsearch.
 Dashboards: Designed customized dashboards in IBM Cognos and filters for restricted view access at probe and OSS level
- in MySQL, decreasing traffic between endpoints by 20%, optimizing database performance.

 Notifications: Engineered Python-flask framework for automated impact analysis, sending email and notifications in
- Notifications: Engineered Python-flask framework for automated impact analysis, sending email and notifications in real-time about network outages; saving US \$1M in accrued costs during 1 year.
- Log data: Processed log data using jq utility and regexp to enrich network performance counters with description in MongoDB, analyzed faults using visualization created in Metabase.

PROJECTS

- **Prototype to Deployment**: AWS cloud-native centralized IoT end-to-end platform for healthcare monitoring with analytics, reporting, auditing and failure notifications built-in.
- Fake News Challenge: Stance Detection: Developed a news article stance classifier using RoBERTa, a transformer-based language model, achieving an accuracy of 92% on the test set, improving baseline GradientBoosting Classifier by 12%
- Customer Churn Prediction: Carried out statistical data analysis and built Regression model to accurately predict candidate customers for attrition, utilized seaborn, plotly, and matplotlib to create rich visualizations and make strategic recommendations.
- RecLix: This project delves into the realm of Exploratory Data Analysis, using the Cosine similarity metric to identify and recommend movies that are semantically similar based on their vectorized bag-of-words representation.

CERTIFICATIONS

- Google Cloud Certified Professional Data Engineer, by Google | Issued: 24 Jan'23
- Google Data Analytics, by Google | Issued: 9 May'21

SKILLS SUMMARY

- Languages: Python (scikit-learn, scipy, numpy, pandas, keras/tensorflow), Go, SQL, R, bash
- Cloud: AWS(S3, RS, RDS, Glue, EMR, Lambda), Azure(DataBricks, ADLS, Synapse), GCP(BigQuery, DataFlow)
- ML/AI: Regression, Classification, Clustering, Deep Learning, NLP, Transformers, Statistical Data Analysis
- Other: Data Warehousing, PySpark, Airflow, ETL/ELT, Tableau, ELK stack, Git, Trello