Yifan(Steven) Wen

Toronto | 778-952-9969 | yifan.wen@mail.mcgill.ca

EDUCATION

McGill University, Montreal,QC Sep 2023 – Dec 2024

Master of Computer Science

- Selected Courses: Advance Big Data with Spark and Hadoop/ Deep Learning/ Natural Language Processing

Queen's University, Kingston Sep 2019 – May 2023

Bachelor of Science in Computing

Selected Courses: **Database I&II**/ Data Structures / intro & advance to ML/ Advanced Data Analytics

TECHNICAL SKILLS

- Data Engineering (2 yrs): Data ETL pipeline, pandas, Hadoop, PySpark, MapReduce, Airflow, Github, CI/CD
- Data Science (2 yrs): Natural language processing (NLP), Python, NLTK, TensorFlow, PyTorch, Sklearn, SQL, NoSQL
- Cloud Engineering (2 yrs): AWS Certified Cloud Practitioner, S3, EMR, RDS, Lambda, Glue, Redshift

WORK EXPERIENCE

Data Engineer Intern, Invision Trading Inc.,Toronto Jan 2025 – Present

- **Data Processing and Modeling**: Extract and cleanse data from AWS RDS, ensuring it is free from duplicates, missing values are handled, and data types are appropriate for analysis. Format the data according to the requirements of the competitor metrics model, using transformations such as normalization or encoding, and input the formatted data into the model for prediction.
- **Analysis and Dashboard Interpretation:** Review thresholds and check for data restatements while analyzing key metrics like billing CP and trip CP. Use Tableau or Power BI to create dashboards for a clear visual assessment of company performance against competitors.
- **Reporting and Presentation:** Compile insights into a report with clear methodologies, results, and recommendations, and prepare presentations for stakeholders.

Assist ML Engineer, McGill University Department of Pathology, Montreal April 2024 – Sep2024

- **Scalable Medical Image Processing Platform:** Built AWS cloud pipelines (S3/Glue/EMR) to automatically process 10,000+ daily microscope scans, using Spark for fast image conversion and Airflow for workflow management, reducing storage costs by 58% while maintaining instant data access.
- **AI-Assisted Cancer Detection Tools:** Developed deep learning models (PyTorch/YOLOv8) that automatically identify cancer cells with 95% accuracy, cutting manual analysis time from 3 days to 4 hours per case, with results directly integrated into hospital lab systems via PostgreSQL.
- **Real-Time Diagnostic Collaboration System:** Created serverless analytics (Redshift/Lambda) that standardized test reports, reducing diagnosis turnaround time by 80% and improving agreement between doctors' evaluations by 40% through automated quality checks.
- **Treatment Optimization Analytics**:Transformed 500K+ AI analysis results into Tableau dashboards that identified critical treatment delay patterns, enabling hospitals to prioritize high-risk patients 22% faster and improve resource allocation efficiency by 35% through predictive case management reports.

Data Engineer Intern, Big-Data Analytics and Management Laboratory**, **Kingston Oct 2022 – June 2023

- **High-Volume Streaming Pipeline**: Built a Kafka-based ingestion system processing 20K+ sensor events/sec, with Airflow-orchestrated Glue ETL jobs converting raw CSV to Parquet in S3, achieving 80% lower latency than batch processing and reducing storage costs by 65% via tiered lifecycle policies.
- **Automated ML Feature Engineering**: Designed windowed feature aggregation (5-sec intervals) using Glue Spark, training PySpark ML models on EMR to detect 15+ human activities (98% F1-score). Deployed models via SageMaker endpoints with DynamoDB feature stores, enabling sub-50ms real-time predictions
- **Git-Style Data Governance**: Implemented lakeFS for version-controlled data lakes, integrated with Athena for SQL analytics and Glue Catalog for schema management. Enforced HIPAA compliance via IAM roles and KMS encryption, reducing audit preparation time by 40%.