MICHAEL WINSOR

SKILLS

Programming

Python · SQL · Bash · Java · PowerShell

Data

Spark · Data Modelling · Data Quality · Pandas · Polars · Matplotlib · Seaborn

Analytics

Time Series Forecasting · KNN · SVM · Kalman Filters · Ridge/Lasso Regression · Logistic Regression · PCA

Cloud & DevOps

Cloudera · AWS · Azure · Linux · Docker · Kubernetes · GitLab CI/CD

CERTIFICATIONS

AWS Solution Architect Associate

Azure Data Engineer Associate

ACHIEVEMENTS

- **Played a lead role in major data engineering tasks for a $19M software project**
  Took on critical project tasks such as production deployments, legacy pipeline maintenance, and pipeline development within a few months of joining the project
- **40% improved Hive query efficiency**
  Optimized file storage to align with the HDFS storage block size across several terabytes of historical and newly ingested data. This decreased the time taken for Hive to read and search through the data.
- **$100,000 cost savings from ETL job optimization**
  Refactored PySpark code to be more memory efficient, which drastically improved performance and avoided investment in expanding the server infrastructure

Data Engineer

mgwinsor@gmail.com github.com/mgwinsor Singapore

SUMMARY

Highly motivated Data Engineer with a passion for learning and a strong background in statistical modelling. Adaptable problem-solver with a keen eye for troubleshooting and the ability to streamline data pipelines. Collaborative team player, dedicated to delivering robust and efficient data solutions through effective cross-functional collaboration. Possess broad technical knowledge, including experience with Python, PySpark, AWS, Azure, and Kubernetes to make valuable contributions to a fast-paced, dynamic environment.

PROFESSIONAL EXPERIENCE

Data Engineer — 11/2021 - Present

ST Engineering — Singapore

- Developed batch and streaming ETL pipelines with PySpark on Hadoop and AWS
- Created a GitLab CI/CD pipeline to test, build, and deploy docker containers to Azure
- Maintained data engineering code repository and oversaw production deployments
- Implemented strong data quality pipeline checks based on lessons learned from previous projects

Software Engineer — 09/2020 - 11/2021

TUMCREATE — Singapore

- Implemented a driving simulation model of a road network to facilitate public transport research
- Lead efforts for a proof-of-concept project to test the real-world application of new public transport concept
- Used Python to collect and analyze vehicle trajectory data via AI to provide insights on the effectiveness of the proof-of-concept project

EDUCATION

Master of Science, Transportation Systems — 10/2017 - 06/2020

Technical University of Munich — Munich, Germany

Bachelor of Science, Mechanical Engineering — 09/2012 - 06/2017

Drexel University — Philadelphia, PA, USA

PROJECTS

MRT Station Crowd Density Model — 12/2023 - Present

ST Engineering

Implemented a model with Python to process raw data streamed from all MRT stations in Singapore and assign density value of high, medium, or low. Next phase is to implement time series forecasting to predict future station crowding.

- Output from this model will be available to commuters through MyTransport.SG app
- Implemented a function to read raw data from a binary file
- Model ran with HA on three servers before results were consolidated and sent to the client's servers as JSON lines via SFTP

SAP Financial Data Pipeline — 01/2022 - 11/2022

ST Engineering

Built an ETL pipeline with PySpark and NiFi to map SAP financial data with other data sources. This allowed downstream models to unlock cost analysis and prediction capabilities. Delivery of this pipeline to production marked a major milestone for the project.

- Implemented AES encryption in motion and at rest to protect sensitive financial information
- Successfully delivered the pipeline under tight deadlines and maintained it throughout its life cycle
- Concurrently managed the data engineering team to ensure that less critical tasks were still accomplished on time