

🎓 教育背景

工學碩士

國立交通大學

工業工程與管理

2016-09 - 2018-06

<> 技能專長

Spark

PySpark

Spark SQL

Structured Streaming

Delta Lake

Spark UI

Python 生態

Python

Pandas

Pydantic

pytest

Poetry

平台

Databricks

AWS EMR

Airflow

S3

Kubernetes

🏆 證書獎項

Databricks Certified Associate Developer for Apache Spark

Databricks

2023-02

Python Institute PCAP

Python Institute

2020-05

個人總結

專注 PySpark 批次與串流處理，熟悉資料傾斜、partition、checkpoint 與效能剖析。曾把製造、物流資料整理成模型可用的特徵，習慣與資料科學家與領域專家一起定義可驗證的輸出。

🏢 工作經歷

PySpark 數據工程師

智造視界股份有限公司

新北

2021-08 - 至今

用 PySpark 清理 1.8TB 日感測資料，和製程專家定義 94 個訊號特徵，模型訓練資料可用率由 87% 提升至 98.9%。

定位資料傾斜與 shuffle spill，重寫 16 個 join 與 partition 策略，配合 ML 團隊將批次窗口由 2 小時縮至 19 分鐘。

以 Structured Streaming 與 checkpoint 處理 3,400 台設備事件，協作現場維運 8 人，異常告警平均提前 14 分鐘。

導入 pytest、樣本快照與 Spark UI 效能基準，帶領 5 位工程師審查 42 個作業，回歸缺陷減少 49%。

Spark 數據工程師

穀雨供應鏈科技

新竹

2018-09 - 2021-07

以 PySpark 合併倉儲掃碼、車隊 GPS 與訂單資料，和物流規劃師校準時間窗，日運力報表提早 75 分鐘完成。

將 320 個 CSV 來源轉成 Delta Lake，協作 6 位資料分析師完成 schema mapping，重複貨件記錄下降 83%。

使用 Spark UI 找出 8 個高成本 stage 並套用 broadcast join，服務 14 個倉庫的查詢延遲降低 64%。

用 Airflow 與 S3 lifecycle 自動化歷史資料重算與保存，和資安確認保留期限，月度雲端儲存支出減少 21%。

專案經驗

設備預知維護特徵湖

以 PySpark 生成振動、溫度與停機特徵，供維護模型每日訓練。

PySpark

Delta Lake

Databricks

MLflow

物流 ETA 串流處理

將車隊定位與訂單事件轉成分鐘級 ETA 指標。

Structured Streaming

Kafka

S3

Airflow

語言能力

中文

母語

英文

技術流利