

Jack Toke

Data Engineer

jacktoke.com · hi@jacktoke.com · 0455 149 451 · github.com/jackptoke · linkedin.com/in/jackptoke

SUMMARY

Data and IT professional with a Graduate Certificate in Data Analytics (QUT), Databricks Certified Data Engineer Associate, and hands-on experience across enterprise application support, systems integration, and business reporting. At AME Systems, integrated Siemens Capital with enterprise systems, automated data synchronisation in Python and SQL, and delivered custom reporting for Sales and Engineering. Currently building AI and data engineering solutions for regional businesses as an independent consultant, including a production AI telephone receptionist and lead-qualification system. Four years of multi-site program coordination in Southeast Burma; bilingual Karen and English communicator, comfortable with stakeholders at all technical levels.

TECHNICAL SKILLS

Languages: Python, SQL, T-SQL, Java, C#, Kotlin

Databases: SQL Server, PostgreSQL, MySQL (relational); MongoDB, Cassandra, Firebase (NoSQL)

Warehouses & lakehouses: Snowflake, Databricks (Delta Lake), Amazon Redshift, BigQuery, DuckDB / MotherDuck

Cloud & platforms: Microsoft Azure (Databricks, Data Factory, Functions, ADLS Gen2), AWS (S3, Glue, Athena, MWAA), GCP (BigQuery, Cloud Storage), Docker / Docker Compose

Transformation & orchestration: dbt, Apache Airflow, PySpark / Spark, Databricks Workflows, Auto Loader, Change Data Feed

Ingestion: Pub/Sub, GTFS-Realtime / protobuf

DataOps & governance: Git / GitHub, CI/CD (GitHub Actions, Databricks Asset Bundles), dbt testing, Unity Catalog

Reporting: Tableau, Power BI (advanced DAX), Plotly Dash, Streamlit

SELECTED PROJECTS

Real-time public transport pipeline – *Azure Function, Databricks, Streamlit*

Production lakehouse measuring V/Line train punctuality from Public Transport Victoria's GTFS feeds. A timer-triggered Azure Function (Python) downloads real-time protobuf trip updates to ADLS Gen2. A parallel Databricks job ingests the weekly static GTFS schedule across 11 bronze tables. Protobuf is decoded on the JVM with `from_protobuf()`, then modelled through a bronze-silver-gold medallion architecture with an SCD2 stops dimension. The gold fact flags on-time, severely late and skipped stops per served stop. Deployed serverless on Databricks with Unity Catalog and Bicep infrastructure-as-code, and served through a live Streamlit dashboard.

[GitHub](#) · [Live dashboard](#)

Australian real estate lakehouse & analytics dashboard – *Databricks, dbt, Streamlit*

End-to-end Databricks lakehouse over Australian residential property listings, ingesting buy, rent and sold data from the realty-in-au API. Raw JSON lands via Auto Loader into a bronze-silver-gold medallion architecture on Delta Lake and Azure Data Lake Storage (ADLS Gen2). dbt transforms it into a dimensional star schema that separates physical properties from listings, with conformed dimensions, bridge tables and snapshot facts. Hash-based surrogate keys, referential-integrity checks and dbt data-quality tests (uniqueness, not-null, relationships) keep the data trustworthy. The property-versus-listing split enables repeat-sales analysis across 2007–2026, identifying 640 properties sold twice and 82 sold three times. Orchestrated with decoupled Databricks Jobs for nightly ingestion and file-arrival-triggered transformation, deployed as infrastructure-as-code with Databricks Asset Bundles across separate dev and prod catalogs, and served through an interactive Streamlit dashboard.

[GitHub](#) · [Live dashboard](#)

Citi Bike medallion pipeline – *Databricks*

Event-driven lakehouse over roughly 10 years of Citi Bike trip data (2016–present). Auto Loader ingests raw CSVs and canonicalises their changing historical schemas into one format in testable pure Python. Change Data Feed streams typed, deduplicated trips into silver via MERGE upserts. The gold layer recomputes only the affected days per batch, staying correct under late-arriving data and deletes. Deterministic ride IDs are derived by SHA-256 hashing for records that lack native identifiers. Deployed on serverless compute with three GitHub Actions workflows for CI, gated

deployment and job runs, using service principals across separate dev, test and prod workspaces. Served through a live Streamlit dashboard.

[GitHub](#) · [Live dashboard](#)

End-to-end analytics pipeline – *Airflow, dbt, DuckDB & Streamlit (QUT Graduate Certificate)*

Dockerised ELT pipeline over JSON receipt data from four small businesses, running fully self-contained with no external services. Apache Airflow DAGs orchestrate ingestion and trigger downstream transformation. Modular dbt models transform the data in DuckDB. A Streamlit dashboard answers business questions on seasonal profitability, customer loyalty and employee turnover.

[GitHub](#) · [Live dashboard](#)

AI Telephone Receptionist & Lead Qualifier – *Freelance product*

Production AI voice system that answers inbound calls, qualifies and captures leads, and surfaces activity through live reporting for regional SMBs. Handles call routing, structured lead capture and automated follow-up, giving small businesses a first, practical step into AI adoption.

Private repository

CAREER HISTORY

AI & Data Engineer | Freelance Jan 2026 – Present

- Build an AI telephone receptionist and lead-qualification solution for small and medium-sized businesses.
- Help small and regional businesses take their first step in AI adoption, from scoping to deployment.
- Build AI chatbots, automation workflows and live reporting systems tailored to real operational needs.

ICT Application Officer | AME Systems Nov 2022 – Feb 2025

- Integrated Siemens Capital wiring harness design software with enterprise systems via webhooks and SOAP web services, enabling seamless data flow for the Engineering department.
- Automated component and engineering design data synchronisation using Python and SQL, eliminating manual update processes and reducing costly data inconsistencies.
- Designed and delivered custom reports for Sales, Engineering and cross-functional teams to drive data-informed business decisions.
- Built a label-printing application using .NET MAUI and an ASP.NET Core Minimal API backend, supporting Warehouse and Dispatch operations.
- Developed native Android applications in Kotlin and Jetpack Compose for Zebra barcode scanner hardware, reducing human error in warehouse picking.

Computer Operator | Hazeldene's Chicken Farm Oct 2020 – May 2022

- Ensured smooth operation of computerised packaging machinery and label printers in a high-throughput production environment.
- Maintained packaging machinery and troubleshoot faults to minimise production downtime.

QA Engineer (Intern) | JB Hi-Fi Mar 2020 – Apr 2020

- Verified that online promotional banners, product specifications and EDM links functioned and displayed correctly before go-live, working alongside the marketing and digital teams.

School Coordinator | Education Opportunity Foundation (Thai–Burmese border) May 2015 – Jun 2019

- Coordinated the development and rollout of an English Teacher Toolkit adopted across 20 schools to lift student confidence and English fluency.
- Designed and delivered annual teacher training programs for 76 teachers across remote Southeast Burma, strengthening instructional practice.
- Administered end-to-end payroll for all 20 schools, including in-person distribution and verification against service records.

EDUCATION

Master of Biostatistics | The University of Adelaide In progress

Graduate Certificate of IT Practice (Data Analytics) | Queensland University of Technology Dec 2025

Bachelor of Science (Computer Science) | Victoria University Feb 2010

CERTIFICATIONS & TRAINING

Databricks Certified Data Engineer Associate | Databricks Jul 2026

Lakehouse architecture and Delta Lake; ELT with Spark SQL and PySpark; incremental ingestion with Auto Loader; Lakeflow Declarative Pipelines; job orchestration with Databricks Workflows; data governance with Unity Catalog.

Data Engineering Bootcamp | Le Wagon Jun 2025

ETL orchestration with Apache Airflow; dbt pipelines; PySpark for large-scale transformations; streaming with Pub/Sub; warehousing with BigQuery and Postgres; Docker Compose environments.

Data Engineering with AWS | RMIT / Udacity Apr 2025

Amazon Redshift warehousing with dimensional modelling (star / snowflake schemas); S3, AWS Glue and Athena data lakes and lakehouses with Apache Spark (Iceberg, medallion, CDC); pipeline orchestration with Apache Airflow on Amazon MWAA.

Web Development Bootcamp | Coder Academy Mar 2020

Full-stack web development fundamentals; front-end and back-end application build, version control and collaborative development.