NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | NTU Academy for Professional and Continuing Education

(SCTP) Advanced Professional Certificate

Data Science and AI

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

C2.2 Coaching session

16 Aug 2025

Agenda

- Welcome Instructor - Thomas
- 2.4 Web Scraping (Continued)
- Coaching questions
- Advance Technical Setup for Lesson 2.5-2.7 (DBT, Meltano, Dagster)
- DBT tool overview
- Breakout : Project Data Selection (If time permits)

Coaching

Question 1

Besides the 7 pointers of project brief and Evaluation Criteria 'Focus', and if you are a hiring manager, what other hard skillsets (technical attributes) you would be looking out (your expectations) for, from a job applicant in his/her M2 project posted in GitHub.

Junior Data Engineer (0-2 years)

Core Technical Skills

- Programming: Proficient in Python and SQL with strong fundamentals
- Data Processing: Basic understanding of ETL/ELT processes and data ingestion patterns
- Database Knowledge: Familiarity with relational databases (PostgreSQL, MySQL) and basic SQL optimization
- Data Warehousing: Understanding of dimensional modeling concepts (star schema, fact/dimension tables)
- Cloud Fundamentals: Basic knowledge of at least one cloud platform (AWS, GCP, or Azure)
- Version Control: Proficient with Git and collaborative development workflows

Tools & Technologies

- Languages: Python, SQL, Shell scripting
- Pipeline Tools: Basic experience with dbt or similar transformation tools
- Cloud Platforms: Introductory knowledge of cloud data warehouses (BigQuery, Redshift, Snowflake)
- Development: Git, basic IDE/text editor proficiency
- Data Formats: Understanding of JSON, CSV, and basic binary formats

Educational Requirements

- Bachelor's degree in Computer Science, Engineering, Mathematics, or related field
- Preferred Certifications: Cloud fundamentals (AWS Cloud Practitioner, Google Cloud Digital Leader)

Key Responsibilities

- Assist in data pipeline development and maintenance
- Perform data quality checks and basic validation
- Support data ingestion from various sources
- Collaborate with senior engineers on data transformation tasks
- Document processes and troubleshoot basic pipeline issues

Mid-Level Data Engineer (4-5 years)

Core Technical Skills

- Programming: Advanced Python, SQL, and familiarity with Scala or Java
- Pipeline Architecture: Design and implement scalable ETL/ELT pipelines
- Data Modeling: Proficient in dimensional modeling, slowly changing dimensions (SCD), and data warehouse design
- Big Data Technologies: Hands-on experience with Apache Spark and distributed computing concepts
- Streaming: Understanding of real-time data processing with Kafka or similar platforms
- Cloud Expertise: Deep knowledge of cloud data services and infrastructure
- Performance Optimization: Query tuning, pipeline optimization, and resource management

Tools & Technologies

- Orchestration: Apache Airflow, Dagster, or cloud-native orchestration tools
- Transformation: dbt, Apache Spark, custom Python frameworks
- Streaming: Apache Kafka, Apache Flink, cloud streaming services
- Databases: Both SQL (PostgreSQL, MySQL) and NoSQL (MongoDB, Cassandra, Redis)
- Cloud Platforms: Advanced knowledge of AWS, GCP, or Azure data services
- Data Quality: Great Expectations, custom testing frameworks
- DevOps: CI/CD pipelines, Docker basics, Infrastructure as Code concepts

Educational Requirements

- Bachelor's degree required, Master's preferred in relevant field
- Required Certifications: At least one cloud platform certification (AWS Solutions Architect, Google Cloud Professional Data Engineer)

Key Responsibilities

- Lead end-to-end data pipeline development projects
- Implement complex data transformations and business logic
- Collaborate with data scientists and analysts on data requirements
- Ensure data quality through automated testing and monitoring
- Optimize pipeline performance and cost efficiency
- Mentor junior engineers and conduct code reviews

Principal Data Engineer (7+ years)

Core Technical Skills

- Architecture Design: Expertise in designing large-scale, distributed data architectures
- Multi-Cloud Strategy: Deep knowledge across multiple cloud platforms and hybrid architectures
- Advanced Programming: Expert-level proficiency in Python, SQL, Scala/Java, and emerging languages
- Real-time Systems: Advanced streaming architectures with Kafka, Flink, and event-driven design
- Performance Engineering: Advanced optimization techniques, cost management, and scalability planning
- Data Governance: Comprehensive understanding of data privacy, security, and compliance frameworks
- Infrastructure: Kubernetes, advanced containerization, and Infrastructure as Code

Tools & Technologies

- Advanced Orchestration: Custom framework development, advanced Airflow/Dagster implementations
- Big Data Ecosystem: Hadoop ecosystem, advanced Spark optimization, custom big data solutions
- Streaming Platforms: Apache Kafka, Apache Flink, custom streaming architectures
- Infrastructure: Kubernetes, Terraform, CloudFormation, advanced cloud services
- Data Governance: Apache Atlas, Collibra, custom metadata management solutions
- Monitoring: Advanced observability tools, custom monitoring solutions
- Emerging Technologies: MLOps integration, data mesh architectures, modern data stack components

Educational Requirements

- Bachelor's degree required, Master's/PhD preferred
- Required Certifications: Multiple advanced cloud certifications, specialized big data certifications

Leadership & Strategic Skills

- Technical Leadership: Ability to lead and mentor large engineering teams
- Strategic Planning: Drive data strategy and technical roadmap development
- Cross-functional Collaboration: Work with C-level executives, product teams, and external vendors
- Innovation: Evaluate and implement cutting-edge technologies and methodologies

Key Responsibilities

- Architect enterprise-scale data platforms and ecosystems
- Define technical standards and best practices across the organization
- Lead major technical initiatives and transformation projects
- Ensure compliance with data governance, security, and regulatory requirements
- Build and mentor high-performing data engineering teams
- Drive innovation and adoption of emerging data technologies
- Collaborate with business leaders on data strategy and ROI optimization

Desirable Attributes **Technical Excellence & Analytical Thinking**

Technical Excellence

Portfolio Sophistication

- Candidates who demonstrate exceptional project quality over quantity, showcasing end-to-end data solutions that solve real business problems with measurable outcomes and professional-grade documentation.

Solution Architecture Mindset

- The ability to design technical solutions from business requirements, translating stakeholder needs into scalable data architectures while considering cost, performance, and maintenance implications.

Code Craftsmanship

- Commitment to engineering best practices including clean code principles, comprehensive testing, version control proficiency, and maintainable system design that reflects production-ready thinking.

Analytical & Problem-Solving Capabilities

Data Intuition

- Natural sense for data quality, patterns, and anomalies - the ability to quickly identify data issues, understand statistical distributions, and recognize when results seem implausible or require deeper investigation.

Systems Thinking

- Capacity for holistic problem-solving that considers upstream and downstream impacts, data lineage, system dependencies, and the broader ecosystem implications of technical decisions.

Optimization Mindset

- Instinctive drive to improve performance and efficiency, whether through query optimization, architectural refinements, cost reduction, or process automation initiatives.

Desirable Attributes **Communication & Business Acumen**

Technical Storytelling

- Exceptional ability to translate complex technical concepts into compelling narratives for diverse audiences, using data visualization and clear explanations to drive decision-making.

Stakeholder Fluency

- Skill in bridging technical and business domains, facilitating requirements gathering, managing expectations, and ensuring data solutions align with organizational objectives.

Documentation Excellence

- Commitment to comprehensive knowledge transfer through clear technical documentation, runbooks, and architectural decision records that enable team scalability.

Continuous Learning Orientation

- Demonstrated passion for emerging technologies and methodologies, with evidence of self-directed exploration of new tools, frameworks, and industry best practices.

Research Mindset

- Natural inclination to investigate root causes, experiment with alternative approaches, and validate assumptions through systematic testing and analysis.

Innovation Drive

- Proactive identification of improvement opportunities and willingness to propose novel solutions or challenge existing processes for better outcomes.

Next Steps

Thought Leadership & Knowledge Sharing

- Technical blogs
- LinkedIn articles translating previous industry experience to data contexts
- Meetup presentations on applying data solutions to your former industry
- Medium posts explaining complex concepts to fellow career switchers

Industry Recognition

- Hackathon participation showcasing rapid skill acquisition
- Online course completion certificates from Coursera, Udacity, edX
- Kaggle competition rankings demonstrating practical skills
- Portfolio projects from curriculum

Business-Focused Projects

- Industry-specific solutions leveraging your previous domain expertise
- End-to-end pipelines solving real problems from your former field

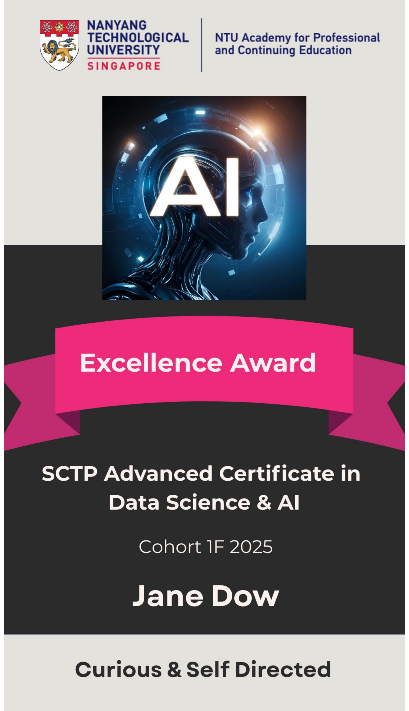
Transferable Skill Documentation

- Cross-industry case studies applying data engineering to familiar problems
- ROI calculations for data initiatives in your previous sector
- Stakeholder communication examples bridging technical and business needs

Build New Connections

- Informational interviews with data engineers in your target companies
- Network of senior practitioners
- Peer networks with other career switchers
- Conference attendance at data-focused events

Reminder Excellence Awards



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NTU Academy for Professional and Continuing Education

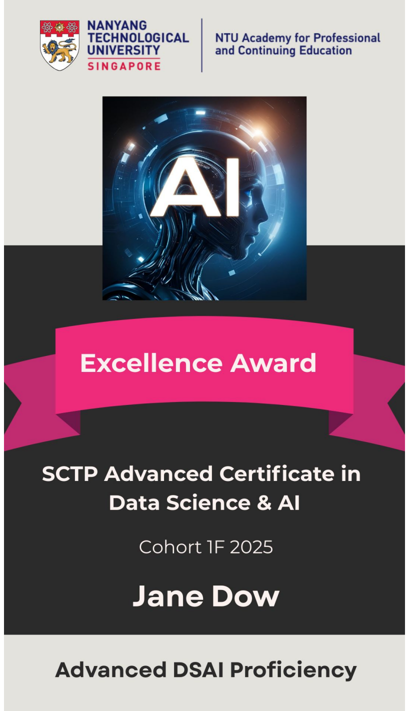
Excellence Award

SCTP Advanced Certificate in Data Science & AI

Cohort 1F 2025

Jane Dow

Curious & Self Directed



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NTU Academy for Professional and Continuing Education

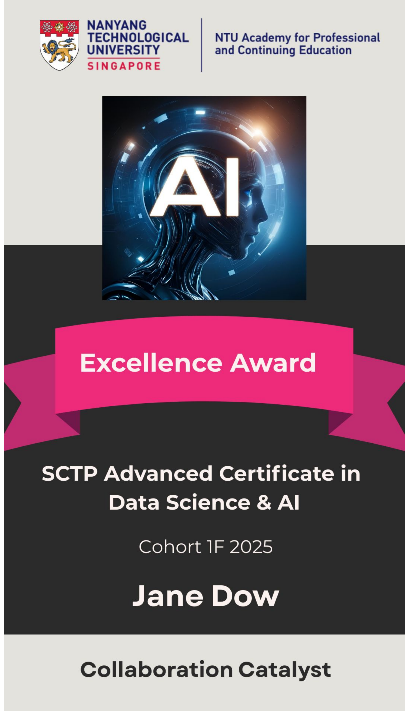
Excellence Award

SCTP Advanced Certificate in Data Science & AI

Cohort 1F 2025

Jane Dow

Advanced DSAI Proficiency



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NTU Academy for Professional and Continuing Education

Excellence Award

SCTP Advanced Certificate in Data Science & AI

Cohort 1F 2025

Jane Dow

Collaboration Catalyst

Curious & Self Directed | Advanced DSAI Proficiency | Collaboration Catalyst

Coaching

Question 2

For M2 Project, please also share with us on the ppt deck and presentation duration

M2 Project Presentation & Deck

Presentation

- 10 minutes presentation, 5 minutes Q&A
- Focus on both technical design and business impact / value

Deck

- 10-15 pages
- Presentation Flow Sample:
  - Setup (Introduction/Problem)
  - Methodology (How you approached it)
  - Analysis (What you found)
  - Implementation/Recommendations (What to do about it)
  - Technical Details (How it works)

Example Presentation Outline

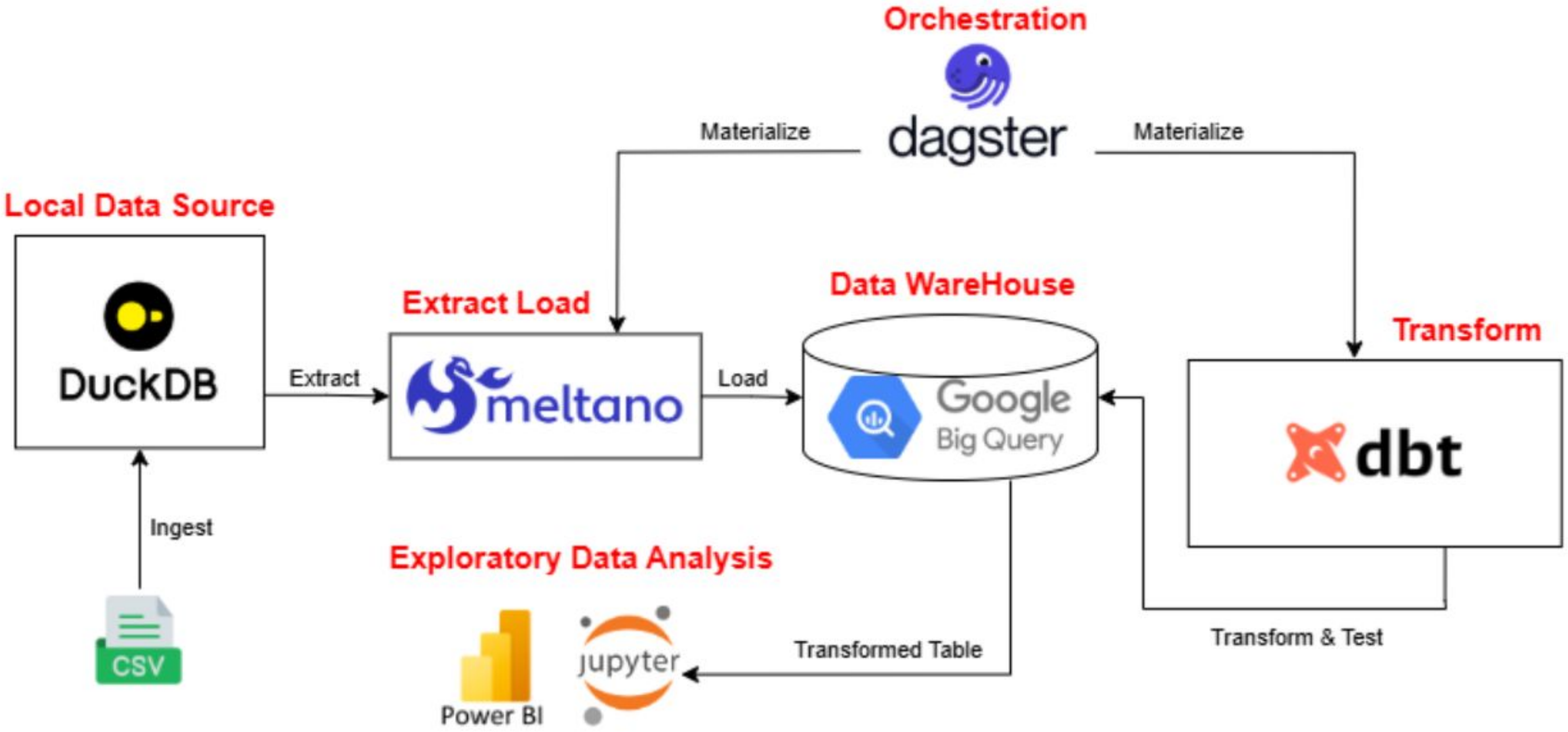
Olist E-commerce Analysis Outline:

1. Executive Summary
   - High-level analysis purpose and focus
2. Introduction & Context
   - Problem statement and scope
   - Performance insights methodology
3. Methodology & Data
   - Dataset description (Olist e-commerce data)
   - Approach and assumptions
   - Data description and exclusions
4. Results & Insights
   - Key metrics dashboard
   - E-commerce sales insights
   - Analysis and findings
5. Strategic Recommendations
   - Actionable recommendations based on findings
6. Conclusion
   - Summary and strategic implications

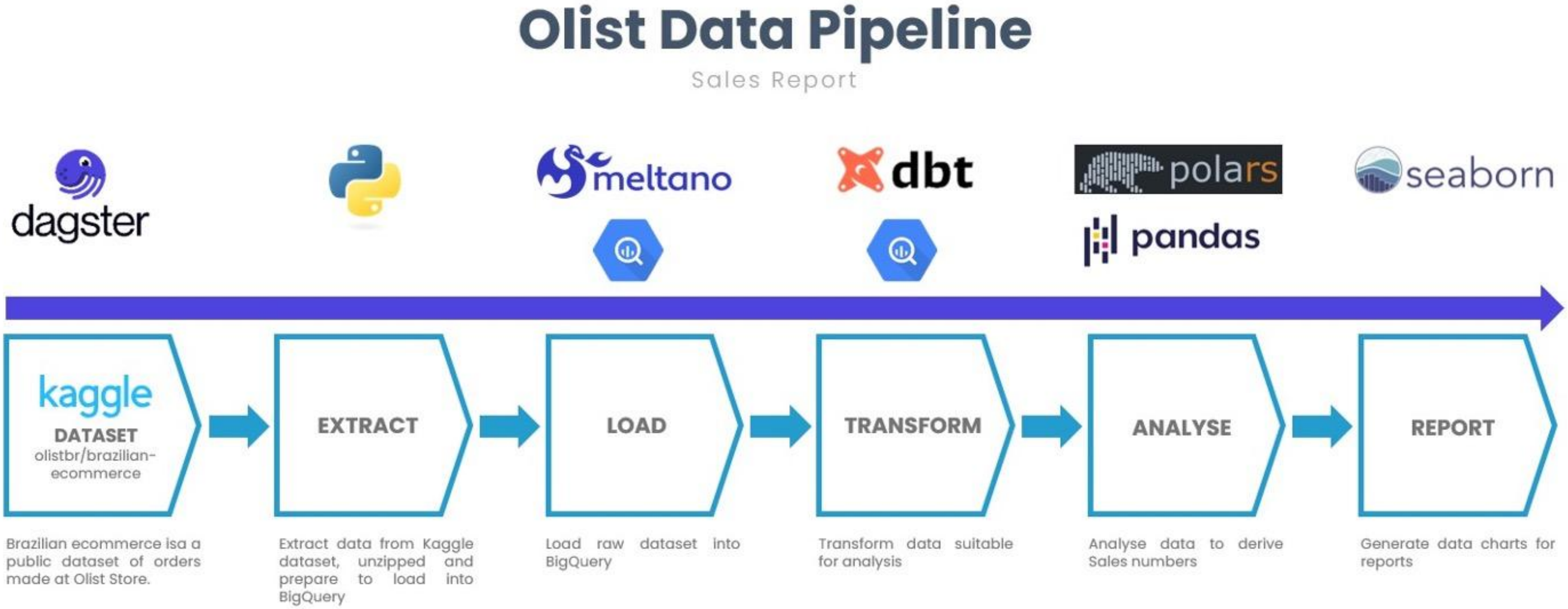
Citi Bike Analysis Outline :

1. Introduction & Setup
   - Dataset description (Citi Bike Trip Histories)
   - Project goals (automated ELT pipeline)
   - Architecture overview
2. Exploratory Data Analysis (EDA)
   - Usage trends analysis
   - Revenue and trip composition
   - Station popularity analysis
   - Anomaly detection
   - Potential failure rate analysis
3. Technical Implementation
   - Database schema design (star schema)
   - Data ingestion process (Meltano orchestration)
   - Key implementation logic for target tables
   - Dagster orchestration and assets

Sample Implementation 1



Sample Implementation 2



Olist Data Pipeline

Sales Report

DATASET
olistbr/brazilian-ecommerce

EXTRACT

LOAD

TRANSFORM

ANALYSE

REPORT

Brazilian ecommerce isa a public dataset of orders made at Olist Store.

Extract data from Kaggle dataset, unzipped and prepare to load into BigQuery

Load raw dataset into BigQuery

Transform data suitable for analysis

Analyse data to derive Sales numbers

Generate data charts for reports

Module 2 Final Project - Excellence Criteria

1. Data Ingestion

Successfully ingested data. Able to justify on decision of the ingestion to the data warehouse and showed all the comparisons between the various methods: listing pros and cons

2. Data Warehouse Design

Designed a star schema. Created most dimension and fact tables. Able to provide good justifications on the design decisions.

3. ELT Pipeline

Implemented ELT pipelines with data cleaning and validation. Also implemented derived columns (business logics) (E.g fact tables). Was able to justify the design process with clear reasons

4. Data Quality Testing

Used testing tools to a large degree. Good coverage in test cases. Went beyond to implement other test cases that was not taught in class

5. Data Analysis with Python

Data analysis was attempted with key metrics and provided reasons why they are important. Was able to link back to some of the design decisions made in the data pipelines based on the analysis or powers such analysis.

6. Pipeline Orchestration (Optional)

Entire pipeline is automated with some orchestration framework. Team was able to provide very clear justifications on why the framework was chosen and list pros vs cons on other frameworks

7. Documentation

Documented code, data lineage, and pipeline architecture using tools like DRAW.IO or EXCALIDRAW. Prepared a comprehensive report summarizing the technical approach, findings, and insights with relevant tables/charts/graphs. Explained tool choices and schema design justification.