



สถาบันไอทีจีเนียส
IT GENIUS INSTITUTE

COURSE ID
DBC-53

COURSE OUTLINE · รายละเอียดหลักสูตรอบรม

DATA ENGINEERING / BOOTCAMP

Data Engineering Bootcamp

สร้าง Data Pipeline ครบวงจร จาก Ingestion ถึง Lakehouse และ BI

● 30 ชั่วโมง

● 5 วัน

● พื้นฐาน Python, SQL และ Docker เบื้องต้น

● ระดับ Intermediate



ระยะเวลา: 5 วัน

30 ชั่วโมง (5 วัน วันละ 6 ชม.) · 09:00-16:00 น.



รูปแบบ: Classroom / Hybrid

Onsite หรือ Online ผ่าน Microsoft Teams · Public / In-house



รายละเอียดหลักสูตร

www.itgenius.co.th

องค์กรส่วนใหญ่มีข้อมูลกระจายอยู่หลายระบบ ทั้งฐานข้อมูลของระบบงาน ไฟล์ส่งออกรายวัน และข้อมูลเหตุการณ์แบบ Real-time ขณะที่ฝ่ายธุรกิจและทีม AI ต้องการข้อมูลที่สะอาด เชื่อถือได้ และพร้อมใช้อย่างรวดเร็ว ปัญหาที่พบบ่อยคือ Pipeline ถูกสร้างแบบเฉพาะกิจ ไม่มีการทดสอบคุณภาพข้อมูล ไม่มีการจัดลำดับงานที่ชัดเจน และขยายต่อได้ยากเมื่อปริมาณข้อมูลเพิ่มขึ้น

หลักสูตรนี้พาผู้เรียนสร้าง Data Platform ครบวงจรให้ร้านค้าปลีกจำลอง MetroMart ด้วยเครื่องมือโอเพนซอร์สที่ใช้กันแพร่หลาย ได้แก่ PostgreSQL, Python กับ Polars และ DuckDB, Apache Kafka, Apache Spark (PySpark), Apache Iceberg บน S3-compatible Object Storage, dbt Core และ Apache Airflow 3 ชิ้นทั้งหมดด้วย Docker Compose บนเครื่องผู้เรียน จบหลักสูตรจะได้ Pipeline ที่รับข้อมูลทั้งแบบ Batch และ Streaming จัดเก็บแบบ Lakehouse แปลงเป็น Star Schema มี Data Quality Gate อัตโนมัติ และส่งต่อให้เครื่องมือ BI ใช้งานได้จริง (อบรม 5 วัน วันละ 6 ชั่วโมง รวม 30 ชั่วโมง ระดับ Intermediate)

เครื่องมือ

Kf

Apache Kafka

Sp

Apache Spark

Ib

Apache Iceberg

dbt

dbt

Af

Airflow

ราคาค่าอบรม · PUBLIC / IN-HOUSE

15,900 บาท / ท่าน

ราคายังไม่รวมภาษีมูลค่าเพิ่ม (VAT 7%) · รวมเอกสารประกอบการอบรม Lab Guide รายวัน ไฟล์ Docker Compose ของ Data Stack ซอร์สโค้ด Pipeline ทั้งหมด และ Certificate · รับผิดชอบ In-house ขอใบเสนอราคาได้

สมัครผ่านเว็บไซต์ ลด 10%



เพิ่มเพื่อน
LINE

☎ 02-570-8449

☎ 088-807-9770

LINE @itgenius

🌐 www.itgenius.co.th

01 วัตถุประสงค์ Objectives

- เข้าใจบทบาท Data Engineer และสถาปัตยกรรม Data Platform สมัยใหม่ ทั้ง Warehouse, Data Lake และ Lakehouse
- ออกแบบ Star Schema พร้อม Slowly Changing Dimension และเขียน SQL เชิงวิเคราะห์ได้อย่างมีประสิทธิภาพ
- สร้าง Batch Ingestion แบบ Incremental และ Idempotent ด้วย Python, Polars และ DuckDB
- เข้าใจ Event Streaming และ Change Data Capture และใช้ Apache Kafka รับส่งข้อมูลแบบ Real-time ได้
- จัดเก็บข้อมูลแบบ Lakehouse ด้วย Apache Iceberg และประมวลผลด้วย PySpark
- ใช้ dbt Core แปลงข้อมูล ทดสอบข้อมูล และสร้างเอกสาร Lineage ได้
- จัดลำดับและตั้งเวลางานด้วย Apache Airflow 3 พร้อม Retry การแจ้งเตือน และ Backfill
- วาง Data Quality Gate และ Data Contract แล้วส่งต่อข้อมูลให้ทีม BI และทีม AI

02 กลุ่มเป้าหมาย Target Audience

- นักพัฒนาซอฟต์แวร์ที่ต้องการก้าวสู่สายงาน Data Engineer
- Data Analyst และ BI Developer ที่ต้องการสร้างและดูแล Pipeline ด้วยตนเอง
- ทีม IT ที่รับผิดชอบการเชื่อมต่อข้อมูลระหว่างระบบในองค์กร
- ทีมที่กำลังวางรากฐานข้อมูลสำหรับโครงการ Analytics และ AI

03 จุดเด่นของหลักสูตร Highlights

- **ครบวงจรในหลักสูตรเดียว** Batch, Streaming, Lakehouse ถึง Dashboard
- **รันได้บนเครื่องผู้เรียน** ทุก Service อยู่ใน Docker Compose
- **โอเพนซอร์สทั้งหมด** ไม่ผูกกับผู้ให้บริการ Cloud รายใด
- **Streaming และ CDC** Kafka, Kafka Connect และ Debezium
- **โปรเจกต์เดียวตลอด 5 วัน** MetroMart จากระบบต้นทางถึง BI

04 ความรู้พื้นฐาน Prerequisites

- เขียน Python ได้ในระดับพื้นฐาน เช่น ฟังก์ชัน, List, Dictionary และการติดตั้ง Package
- เขียน SQL พื้นฐานได้ เช่น SELECT, JOIN และ GROUP BY
- ใช้ Command Line และ Git เบื้องต้นได้ และเข้าใจแนวคิด Container เบื้องต้น
- ไม่จำเป็นต้องเคยใช้ Kafka, Spark, Airflow หรือ dbt มาก่อน วิทยาการปูแนวคิดก่อนลงมือทุกเครื่องมือ

05 สิ่งที่ต้องเตรียมและเครื่องมือ Setup & Tools

รายการ	รายละเอียด
Source และ Batch	PostgreSQL 18, Python 3.12 ขึ้นไป, Polars และ DuckDB
Streaming และ CDC	Apache Kafka 4 แบบ KRaft, Kafka Connect และ Debezium (สาธิต)
Lakehouse	Apache Iceberg (REST Catalog) บน S3-compatible Object Storage แบบโอเพนซอร์ส และ Apache Spark 4 (PySpark)
Transform, Orchestrate และ BI	dbt Core, Apache Airflow 3, dbt Tests และ Metabase หรือ Apache Superset (แนวทางเชื่อม Power BI)
อุปกรณ์ผู้เรียน	Notebook RAM 16 GB ขึ้นไป (แนะนำ 32 GB) พื้นที่ว่าง 40 GB พร้อม Docker Desktop (WSL2) หรือ Docker Engine, Git และ VS Code

หมายเหตุ: ต่อยอดจากหลักสูตร Modern Data Engineering and Lakehouse Architecture (3 วัน) โดยขยายเป็น Bootcamp 5 วันที่เพิ่ม Streaming และ CDC ด้วย Kafka การประมวลผลด้วย Spark และ Orchestration ครบวงจร · เครื่องมือทั้งหมดเป็นโอเพนซอร์สบน Notebook ของผู้เรียน ไม่ต้องใช้บัญชี Cloud · MinIO รุ่น Community หยุดพัฒนาแล้ว Lab จึงใช้ S3-compatible Object Storage แบบโอเพนซอร์สที่ยังได้รับการดูแล เวอร์ชันของเครื่องมือจะอิงรุ่นเสถียรล่าสุด ณ วันอบรม

06 ภาพรวมการเรียนรู้ 5 วัน Big Picture

วัน	หัวข้อหลัก
Day 1 Section 1-5	บทบาท Data Engineer, Warehouse กับ Lakehouse, Medallion Architecture, ติดตั้ง Data Stack ด้วย Docker Compose, SQL เชิงวิเคราะห์ และออกแบบ Star Schema พร้อม SCD
Day 2 Section 6-10	Batch Ingestion แบบ Incremental และ Idempotent, Polars และ DuckDB, Apache Kafka แบบ KRaft, Producer และ Consumer ด้วย Python และ CDC ด้วย Debezium (สาธิต)
Day 3 Section 11-15	S3-compatible Object Storage, Apache Iceberg เชิงลึก (Snapshot, Time Travel, Schema Evolution), PySpark บน Spark 4 และสร้าง Bronze กับ Silver Layer ด้วย MERGE INTO
Day 4 Section 16-20	dbt Core: Model, Materialization, Macro และ Snapshot, Apache Airflow 3 ด้วย TaskFlow API และ Assets, Retry, Deadline Alerts, Backfill และ DAG ที่รับ Spark กับ dbt
Day 5 Section 21-25	Data Quality และ Data Contract, Quality Gate, Observability ของ Pipeline, Serving Layer สำหรับ BI และ AI, Governance และ Master Workshop รวม Pipeline ครบวงจร



DAY
1

Data Engineering Foundations และ Dimensional Modeling

เนื้อหาการอบรม · Section 1-5

รวม
6 ชั่วโมง

1 บทบาท Data Engineer และสถาปัตยกรรม Data Platform

- ความต่างของ Data Engineer, Data Analyst และ Data Scientist
- เปรียบเทียบ Data Warehouse, Data Lake และ Lakehouse
- ETL กับ ELT และ Medallion Architecture (Bronze, Silver, Gold)
- Batch กับ Streaming ควรเลือกใช้เมื่อใด

2 Lab: ติดตั้ง Data Stack ด้วย Docker Compose

- โครงสร้าง docker-compose.yml ของ Data Stack หลาย Service
- Volume, Network และ Environment Variable
- ตรวจสอบ Service และอ่าน Log เมื่อเกิดปัญหา
- จัดสรร RAM และ CPU ให้เหมาะกับเครื่องผู้เรียน
- Lab: เปิด Data Stack ของ MetroMart และตรวจสอบว่าทุก Service พร้อมใช้งาน

3 Lab: SQL สำหรับงานวิเคราะห์ข้อมูล

- ทบทวน JOIN, GROUP BY และ Subquery
- Common Table Expression (CTE) และ Window Function เช่น ROW_NUMBER, LAG และ SUM OVER
- อ่าน Execution Plan และ Index เบื้องต้น
- Lab: วิเคราะห์ยอดขายสะสมและอันดับสินค้าของ MetroMart

4 Dimensional Modeling (Star Schema)

- Fact Table กับ Dimension Table และการกำหนด Grain
- Surrogate Key และ Natural Key
- Slowly Changing Dimension (SCD Type 1 และ Type 2)
- ออกแบบ Date Dimension

5 Workshop Day 1: Star Schema ของ MetroMart

- โหลดข้อมูลจำลองของ MetroMart เข้า PostgreSQL
- ออกแบบ fact_sales, dim_product, dim_customer, dim_store และ dim_date
- กำหนด Grain และ Key ของแต่ละตารางพร้อมเหตุผล
- Workshop: เขียน SQL วิเคราะห์ยอดขายตามสาขา หมวดสินค้า และช่วงเวลา



DAY
2

Data Ingestion: Batch และ Streaming

เนื้อหาการอบรม · Section 6-10

รวม
6 ชั่วโมง

6 Batch Ingestion ด้วย Python

- รูปแบบไฟล์ CSV, JSON และ Parquet ข้อดีและข้อจำกัด
- Full Load กับ Incremental Load และการใช้ Watermark
- ออกแบบ Pipeline ให้ Idempotent รับซ้ำแล้วผลไม่เพี้ยน
- Lab: ดึงข้อมูลสินค้าและลูกค้าจาก PostgreSQL แบบ Incremental

7 Lab: Polars และ DuckDB

- Polars DataFrame, Lazy API และ Expression
- DuckDB Query ไฟล์ Parquet และ CSV ได้โดยตรงด้วย SQL
- เปรียบเทียบกับ pandas และเลือกเครื่องมือให้เหมาะกับขนาดข้อมูล
- Lab: แปลงไฟล์ CSV ของสาขาเป็น Parquet ด้วย Polars แล้วตรวจด้วย DuckDB

8 Event Streaming ด้วย Apache Kafka

- Topic, Partition, Offset, Producer, Consumer และ Consumer Group
- Kafka 4 ทำงานด้วย KRaft โดยไม่ต้องใช้ ZooKeeper
- Delivery Semantics: At-most-once, At-least-once และ Exactly-once
- Lab: สร้าง Topic และดู Partition กับ Offset ด้วยเครื่องมือ Command Line

9 Change Data Capture (CDC)

- หลักการ CDC จาก Transaction Log ของฐานข้อมูล
- Kafka Connect และ Debezium กับ PostgreSQL (สาธิต)
- จัดการ Schema ของ Event และ Schema Evolution
- เมื่อใดควรใช้ CDC แทนการดึงข้อมูลแบบ Batch

10 Workshop Day 2: Ingestion ของ MetroMart

- Batch Pipeline ดึงข้อมูลจาก PostgreSQL และไฟล์ CSV ของสาขาเป็น Parquet
- Producer ด้วย Python จำลองเหตุการณ์คำสั่งซื้อส่งเข้า Kafka Topic
- Consumer ที่บันทึกข้อมูลลงพื้นที่จัดเก็บข้อมูลดิบ
- Workshop: รัน Pipeline ซ้ำและหยุด Consumer กลางคัน แล้วตรวจว่าข้อมูลไม่ซ้ำและไม่หาย



DAY
3

Lakehouse และ Distributed Processing

เนื้อหาการอบรม · Section 11-15

รวม
6 ชั่วโมง

11 Object Storage และ Lakehouse

- S3-compatible Object Storage การจัดการ Bucket และสิทธิ์
- ปัญหาของ Data Lake แบบไฟล์ล้วน และเหตุผลที่ต้องมี Table Format
- เปรียบเทียบ Apache Iceberg, Delta Lake และ Apache Hudi
- Lab: สร้าง Bucket และสิทธิ์สำหรับ Bronze, Silver และ Gold

12 Lab: Apache Iceberg เชิงลึก

- โครงสร้าง Metadata, Manifest และ Snapshot
- REST Catalog และการเชื่อมต่อจากหลาย Engine
- Schema Evolution, Partition Evolution และ Hidden Partitioning
- Time Travel และการ Rollback ข้อมูล
- Lab: เพิ่มคอลัมน์และย้อนดูข้อมูลก่อนแก้ไขด้วย Time Travel

13 Lab: Apache Spark และ PySpark

- สถาปัตยกรรม Spark: Driver, Executor, Job, Stage และ Task
- DataFrame API และ Spark SQL
- Transformation กับ Action และ Lazy Evaluation
- อ่านเขียนตาราง Iceberg ด้วย Spark 4
- Lab: อ่าน Spark UI เพื่อดู Stage และ Shuffle ของงานจริง

14 ออกแบบ Bronze และ Silver Layer

- ทำความสะอาดข้อมูล จัดการค่า Null และข้อมูลซ้ำ
- Merge และ Upsert ข้อมูลด้วย MERGE INTO
- จัดการไฟล์ขนาดเล็ก (Compaction) และบำรุงรักษาตาราง
- Lab: Compaction และลบ Snapshot เก่าของตาราง Iceberg

15 Workshop Day 3: Lakehouse ของ MetroMart

- นำข้อมูลดิบจาก Day 2 จัดเก็บเป็นตาราง Iceberg ใน Bronze Layer
- เขียนงาน PySpark ทำความสะอาดและลบข้อมูลซ้ำ
- Upsert เข้าสู่ Silver Layer ด้วย MERGE INTO
- Workshop: ทดลอง Schema Evolution และ Time Travel บนข้อมูลคำสั่งซื้อ



DAY
4

Transformation ด้วย dbt และ Orchestration ด้วย Airflow

เนื้อหาการอบรม · Section 16-20

รวม
6 ชั่วโมง

16 dbt Core พื้นฐาน

- โครงสร้าง dbt Project, Profile และ Adapter
- Model, Source, ref และ Dependency Graph
- Materialization: View, Table, Incremental และ Ephemeral
- Lab: สร้าง Staging Model จาก Silver Layer

17 Lab: dbt สำหรับงานระดับองค์กร

- แบ่งชั้น Model เป็น Staging, Intermediate และ Mart
- Jinja, Macro และ Package
- Snapshot สำหรับ SCD Type 2
- dbt Docs และ Lineage Graph
- Lab: สร้าง fact_sales และ Dimension พร้อม Snapshot ของลูกค้า

18 Apache Airflow 3 พื้นฐาน

- สถาปัตยกรรม Airflow 3: Scheduler, API Server, DAG Processor, Triggerer และ Worker
- เขียน DAG และ Task ด้วย TaskFlow API
- Assets และการตั้งเวลาแบบ Data-aware
- ติดตามผลการรันผ่าน UI ของ Airflow 3

19 Orchestration ระดับใช้งานจริง

- Retry, Timeout, Callback และ Deadline Alerts สำหรับแจ้งเตือนงานที่ล่าช้า
- ส่งค่าระหว่าง Task และจัดการ Connection กับ Variable อย่างปลอดภัย
- เรียกงาน Spark และ dbt จาก Airflow
- Backfill และการรันซ้ำย้อนหลัง

20 Workshop Day 4: Pipeline อัตโนมัติ

- เขียน DAG ที่รับ Ingestion, งาน PySpark และ dbt ตามลำดับ Dependency
- ตั้ง Retry และการแจ้งเตือนเมื่อ Task ล้มเหลว
- ใช้ Assets ให้ Gold Layer รันต่อเมื่อ Silver Layer อัปเดต
- Workshop: ทำให้ Task ล้มเหลวโดยตั้งใจ แก้ไข แล้วรันซ้ำเฉพาะส่วนที่พลาด



DAY
5

Data Quality, Serving และ Master Workshop

เนื้อหาการอบรม · Section 21-25

รวม
6 ชั่วโมง

21 Lab: Data Quality และ Data Contract

- มีคุณภาพข้อมูล: Completeness, Accuracy, Freshness และ Uniqueness
- dbt Tests แบบ Generic และ Singular และ Source Freshness
- Data Contract ระหว่างทีมผู้ผลิตและผู้ใช้ข้อมูล
- Lab: วาง Quality Gate ให้ Pipeline หยุดเมื่อข้อมูลผิดปกติ

22 Observability และการดูแล Pipeline

- ติดตาม Log, Metric และระยะเวลาการรัน
- ออกแบบการแจ้งเตือนให้ทีมรู้ปัญหาได้รวดเร็วโดยไม่แจ้งเตือนเกินไป
- แนวทาง Incident Response สำหรับ Data Pipeline
- Runbook สำหรับเหตุการณ์ที่พบบ่อย เช่น ข้อมูลมาล่าช้าหรือ Schema เปลี่ยน

23 Lab: Serving Layer สำหรับ BI และ AI

- ออกแบบ Mart ให้เหมาะกับ Dashboard
- เชื่อมต่อ Metabase หรือ Apache Superset และแนวทางเชื่อม Power BI
- เตรียมข้อมูลให้ทีม AI นำไปใช้ต่อ เช่น Feature Table
- Lab: Dashboard ยอดขายรายวันแยกตามสาขาและหมวดสินค้า

24 แนวทางนำไปใช้ในองค์กร

- Data Governance เบื้องต้น การกำหนดเจ้าของข้อมูลและสิทธิ์การเข้าถึง
- การย้าย Stack ขึ้น Cloud หรือ Kubernetes
- Stream Processing ขั้นสูงอย่าง Apache Flink และ Kafka Streams (ภาพรวม)
- วางแผนพัฒนากับ Data Engineering

25 Master Workshop: MetroMart Data Platform

- รวม Batch และ Streaming Ingestion, Lakehouse บน Iceberg, PySpark, dbt และ Airflow เป็น Pipeline เดียว
- เพิ่ม Data Quality Gate และการแจ้งเตือน
- ทดสอบสถานการณ์ข้อมูลผิดปกติรูปแบบและข้อมูลมาล่าช้า
- สรุปผลบน Dashboard และนำเสนอสถาปัตยกรรมของทีม

07 เรียนจบแล้วได้อะไร Outcomes

- เข้าใจสถาปัตยกรรม **Data Platform** Warehouse, Data Lake และ Lakehouse
- ออกแบบ **Data Model** ได้ Star Schema และ SCD
- Ingestion ทั้ง **Batch** และ **Streaming** Polars, DuckDB และ Kafka
- ใช้ **Lakehouse** ได้จริง Iceberg และ PySpark
- Transform และ **Orchestrate** อัตโนมัติ dbt Core และ Airflow 3
- ดูแลคุณภาพข้อมูล Quality Gate และ Data Contract

08 สิ่งที่คุณเรียนจะได้รับ Deliverables

- เอกสารประกอบการอบรมครบทั้ง 5 วัน และ Lab Guide รายวันพร้อม Checkpoint สำหรับตามให้ทัน
- ไฟล์ Docker Compose ของ Data Stack ทั้งชุด และซอร์สโค้ด Pipeline, งาน PySpark, dbt Project และ Airflow DAG
- Checklist Data Quality, ตัวอย่าง Data Contract และ Runbook ดูแล Pipeline สำหรับนำไปปรับใช้
- ประกาศนียบัตรจบหลักสูตร (Certificate of Completion) และช่องทางสอบถามวิทยากรหลังจบหลักสูตร

09 ข้อมูลผู้สอน Instructor

อ.สาธิต ไทยม วิทยากรผู้พัฒนาซอฟต์แวร์และระบบข้อมูลที่ต้องเชื่อมข้อมูลจากหลายระบบเข้าสู่งานวิเคราะห์จริง สอนผ่านโปรเจกต์เดียวที่ต่อยอดทุกวันตั้งแต่ระบบต้นทางจนถึง Dashboard ปูแนวคิดของแต่ละเครื่องมือก่อนลงมือ และให้ผู้เรียนเจอปัญหาจริงอย่างข้อมูลซ้ำ ข้อมูลมาล่าช้า และ Task ล้มเหลว เพื่อให้กลับไปสร้างและดูแล Pipeline ขององค์กรได้อย่างมั่นใจ

ราคาค่าอบรม 15,900 บาท/ท่าน · สนใจ IN-HOUSE ขอใบเสนอราคา

15,900 บาท / ท่าน (ยังไม่รวม VAT 7%)

สมัครผ่านเว็บไซต์ ลด 10%

☎ 02-570-8449

📞 @itgenius

☎ 088-807-9770

✉ contact@itgenius.co.th



เพิ่มเพื่อน
LINE