databricks

Welcome to Advanced Data Engineering with Databricks

Learning Objectives

Advanced Data Engineering with Databricks

1. Design databases and pipelines optimized for the Databricks Lakehouse Platform.
2. Implement efficient incremental data processing to validate and enrich data driving business decisions and applications.
3. Leverage Databricks-native features for managing access to sensitive data and fulfilling right-to-be-forgotten requests.
4. Manage code promotion, task orchestration, and production job monitoring using Databricks tools.

Course Prerequisites

Advanced Data Engineering with Databricks

1. Design databases and pipelines optimized for the Databricks Lakehouse Platform.
2. Implement efficient incremental data processing to validate and enrich data driving business decisions and applications.
3. Leverage Databricks-native features for managing access to sensitive data and fulfilling right-to-be-forgotten requests.
4. Manage code promotion, task orchestration, and production job monitoring using Databricks tools.

Course Overview

Advanced Data Engineering with Databricks

Module 1: Incremental Processing with Spark Structured Streaming and Delta Lake

Module 2: Streaming Architecture Patterns with DLT

Module 3: Data Privacy and Governance

Module 4: Performance Optimization with Spark and Delta Lake

Module 5: CI/CD Workflows with DLT Pipelines

Module 6: Automate Production Jobs

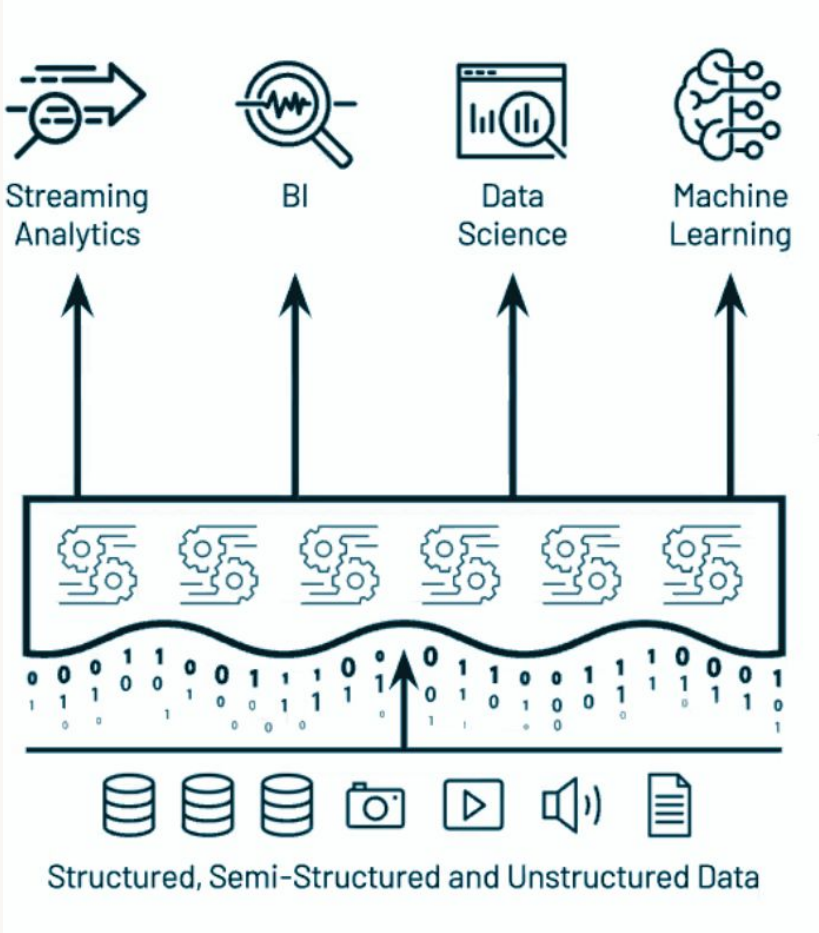
databricks

Data Overview

Case Study

Health Tracker Device Company

moovio



Data Sources

Heart rate (bpm)

- BPM measurements collected by user devices
- Largest volume of data

Workouts

- When users start and complete series of exercises within our application
- Much lower volume of data
- Users complete workouts a few times per month to a few times per day

User information

- Mostly static
- New users processed after device activation and registration
- Includes confidential PII

Gym visitors

daily-stream (Kafka)

streams records for 3 topics: bpm, workout, user_info

field	type	description
key	BINARY	
value	BINARY	
topic	STRING	bpm, workout, user_info
partition	LONG	
offset	LONG	
timestamp	LONG	

bronze (source: Kafka)

key, value, topic, partition, offset, timestamp, date, week_part

field	type	description
key	BINARY	
value	BINARY	
topic	STRING	bpm, workout, user_info
partition	LONG	
offset	LONG	
timestamp	LONG	
date	DATE	
week_part	STRING	

gym_mac_logs (source: JSON)

first_timestamp, gym, last_timestamp, mac

field	type	description
first_timestamp	double	
gym	long	
last_timestamp	double	
mac	string	

registered_users (source: JSON)

device_id, mac_address, registration_timestamp, user_id

field	type	description
device_id	long	
mac_address	string	
registration_timestamp	double	
user_id	long	

user_lookup

Pseudonymization, hashing -- alt_id, device_id, mac_address, user_id

field	type	description
alt_id		sha2(concat(user_id,'BEANS'), 256)
device_id	LONG	Table Content
mac_address	STRING	Table Content
user_id	LONG	

databricks