

5-Day Detailed TOC – DataHub Training for Banking

Day 1: Data Governance & DataHub Fundamentals

◆ Session 1: Banking Data Governance Foundations

Importance of data governance in banking

Regulatory requirements (PII, compliance, audit)

Data challenges:

Data silos

Lack of lineage

Poor discoverability

Governance roles:

Data Owner

Data Steward

Data Custodian

◆ Session 2: Introduction to DataHub

What is DataHub?

Core components:

Metadata Store

GMS (Graph Metadata Service)

Frontend UI

Architecture overview (K8s-based deployment)

Supported integrations:

Kafka, Spark, Airflow, DBs

◆ Session 3: DataHub UI Deep Dive

Navigating DataHub UI

Dataset overview

Schema visualization

Lineage visualization

Search & filtering

Lab 1:

Login to DataHub

Explore datasets

Search datasets using keywords

View lineage graph

Explore schema metadata

Outcome:

Participants understand why + what + how of DataHub

Day 2: Metadata Ingestion & Integration

◆ Session 1: Metadata Concepts

What is metadata?

Types:

Technical metadata

Business metadata

Operational metadata

◆ Session 2: Ingestion Framework in DataHub

Ingestion architecture

Connectors overview:

Kafka

PostgreSQL / Oracle

ClickHouse

Power BI

Batch vs real-time ingestion

◆ Session 3: Creating Ingestion Pipelines

YAML-based ingestion config

Source configuration

Sink configuration

Scheduling ingestion

◆ Session 4: Debugging & Monitoring

Logs analysis

Common ingestion errors

Retry strategies

Lab 2:

Create ingestion pipeline for:

PostgreSQL / Oracle

Run ingestion job

Validate metadata in DataHub

Debug ingestion failure

Outcome:

Participants can ingest data into DataHub independently

Day 3: Data Lineage, Profiling & Governance

◆ Session 1: Data Lineage

What is lineage?

Types:

Table-level

Column-level

Importance in banking (audit & compliance)

◆ Session 2: Lineage Integration

Kafka → Spark → DB → BI flow

Automatic lineage capture

Manual lineage creation

◆ Session 3: Data Profiling

Schema profiling

Data statistics

Data quality indicators

◆ Session 4: Governance Framework

Business glossary

Data classification:

Sensitive

Confidential

Public

Data ownership

Lab 3:

Create lineage for sample pipeline

Assign classification tags

Create glossary terms

Assign data owners

Outcome:

Participants can track data flow and apply governance policies

Day 4: Security, RBAC & Administration

- ◆ Session 1: Authentication & Authorization

LDAP / Active Directory integration

SSO concepts

- ◆ Session 2: Role-Based Access Control (RBAC)

Role design:

Data Engineer

Analyst

Security Team

Permissions & policies

- ◆ Session 3: Data Security in Banking

Data masking (metadata level)

Audit logging

Compliance considerations

- ◆ Session 4: System Administration

Monitoring DataHub

Logs & alerts

Backup & restore

High availability (HA basics)

Lab 4:

Create roles and users

Assign permissions

Test restricted access

Monitor ingestion logs

Outcome:

Participants can secure and manage DataHub environment

Day 5: Real Banking Use Case & End-to-End Implementation

◆ Session 1: Banking Data Architecture

Core Banking (T24)

ATM/Card systems

Mobile banking data

Data Lakehouse integration

◆ Session 2: End-to-End Pipeline

Kafka ingestion

Spark transformation

Storage in DB (ClickHouse/Postgres)

Visualization in Power BI

◆ Session 3: Compliance & Audit Use Case

Trace data origin (lineage)

Identify sensitive data usage

Audit reporting

◆ Session 4: Best Practices

Governance model for bank

Data stewardship workflow

Scaling DataHub

Lab 5 (Capstone Project):

Build full pipeline:

Ingest data

Create lineage

Apply classification

Assign ownership

Perform audit trace

Outcome:

Participants can implement real-world banking governance scenario