

Cloudera Platform Generalist: Concepts, Practice and Exam Preparation

A broad platform programme with version-aware data engineering and operations labs

FA-0721 · 20 days

Data & Analytics · DevOps, Cloud & Infrastructure · Cybersecurity

Introduction

Develop a broad understanding of the Cloudera platform through concepts, selected operations and analytical exercises. The extended programme reviews deployment models, data services, distributed storage/processing, messaging, security and governance, then consolidates knowledge against the current Generalist exam guide.

The source includes extensive older Hadoop-era material. Legacy tools are identified as context or migration examples, while current labs use an explicitly supported Cloudera runtime and compatible services. This course is not the vendor's official preparation product and does not include a credential, exam purchase or guaranteed pass.

Learning outcomes

- Compare Cloudera on cloud, Base on premises and Data Services on premises and describe major platform roles.
- Explain distributed storage, YARN, Spark, SQL engines and operational-database choices.
- Build selected ingestion, query and event-client workflows and inspect their outputs and failures.
- Describe SDX, authentication, authorisation, metadata/governance and relevant encryption controls.
- Review workload isolation, replication, monitoring, costs and recovery considerations.
- Distinguish current supported components from historical tools and identify runtime-specific compatibility needs.

- Consolidate broad platform knowledge using practice assessments and the current Generalist scope.

Prerequisites

- Working familiarity with programming in at least one language, SQL, Linux commands, filesystems and server architecture; this is not an absolute-computing-beginner course.
- Ability to use an approved terminal/SSH client and a supplied virtual or remote lab; LDAP exposure is helpful for security topics.
- Access to an explicitly supported Cloudera training environment with compatible runtime, services and permissions. Paid vendor/cloud access is arranged separately where needed; accounts on every cloud provider are not required.
- Reliable connectivity and conferencing equipment for remote delivery. A second screen is helpful; fixed attendance, administrator privileges and exam-proctor settings are not universal training prerequisites.

Training outline

1. Days 1–2 — Platform concepts, roles and deployment models

- Enterprise data lifecycle and an end-to-end data use case; administrator, engineer, analyst, scientist and steward roles.
- Compare Cloudera on cloud, Base on premises and Data Services on premises; relate historical CDP Public/Private Cloud names to the selected environment.
- Overview of workload/data management, SDX, isolation, migration, replication, resources and costs.
- Introduce data flow, data warehouse, operational database and Cloudera AI capabilities; older CDSW/CML and research references are context, not interchangeable current products.
- Assess conceptual understanding and review the broad Generalist exam domains.

2. Days 3–4 — Architecture, environment and programming orientation

- Compare traditional clusters and cloud-native service architecture; introduce Kubernetes and deployment dependencies.
- Review installation requirements and a supplied installation demonstration; supported infrastructure and provider matrices govern the actual deployment.
- Use the supplied lab rather than rely on the historical Cloudera QuickStart VM.
- Revisit Python, Scala and Java integration paths at an orientation level; learners use their existing language skills for selected examples.

- Inspect remote access, service configuration and basic resource/cost controls; no all-provider installation promise.

3. Days 5–6 — Kafka concepts and clients

- Brokers, topics, record structure, partitions, offsets, producers and consumers.
- KRaft controller roles in a compatible current runtime; ZooKeeper-based Kafka is legacy/deprecated context and remains release-dependent.
- Start or inspect the supplied environment, produce/consume sample events and compare multiple clients and consumer groups.
- Partition assignment, idle groups, lag and selected producer/consumer configuration; review logs and failures.
- Distinguish inter-cluster replication from a producer automatically spanning independent clusters.

4. Day 7 — Kafka resilience and measured behaviour

- Review multiple brokers, replication, storage and failure recovery in a controlled training environment.
- Inspect selected failure simulations, partition behaviour and logs; avoid implying that a short lab proves production resilience.
- Run a small comparable performance experiment and explain the influence of configuration, data and resources.
- Assess the messaging example and document its limitations.

5. Days 8–9 — Hadoop storage and resource management

- Hadoop architecture and relational-versus-distributed processing trade-offs; hardware and installation considerations.
- HDFS reads/writes, permissions, replication, heartbeats, rack awareness and interfaces.
- Use selected hdfs dfs commands, inspect filesystem sizes and metadata, and review distcp, fsck, dfsadmin and balancing.
- YARN ResourceManager, NodeManager, ApplicationMaster and containers; submit or inspect a supplied job.
- Introduce MapReduce data flow, counters, combiner/cache concepts, debugging and selected joins or sorting; review job resource properties.

6. Day 10 — Cluster management and storage choices

- Plan, configure, monitor and benchmark a small supplied cluster; inspect logs and node lifecycle.
- Discuss backup, authentication, confidentiality, high availability and federation with stated recovery assumptions.

- Compare HDFS with Ozone and cloud storage concepts; verify service availability in the selected runtime.
- Review HBase architecture, schema, commands, security and integration, with Phoenix and Kudu as contrasting access/storage choices.

7. Days 11–12 — Spark and distributed workflows

- Spark execution, RDDs and structured DataFrame/SQL APIs; distinguish typed Scala/Java Datasets from PySpark DataFrames.
- Read and transform sample data, use libraries, inspect jobs and evaluate selected joins or aggregations.
- Review accumulators, shuffles and tuning; use a small measured comparison rather than assumed speed gains.
- Introduce clustering, MLlib, Structured Streaming and GraphX/graph-search examples at a bounded awareness level.
- Compare additional stream-processing approaches, including Storm, only where relevant to the supplied environment.

8. Days 13–14 — Hive SQL and analytical structure

- Hive architecture, metastore, warehouse layout and a supported client such as Beeline; legacy Hive CLI settings are historical context.
- Create sample databases/tables, inspect types, delimiters and file formats, and load data.
- Compare partitioning, bucketing and ORC; ACID/update/delete support depends on table properties and selected Hive/runtime versions, not a universal bucketing rule.
- Use string, date, numeric and casting functions; handle nulls and validate query results.
- Write projections, filters, aggregations, group sorting, joins, unions and subqueries; inspect logs and query behaviour.

9. Day 15 — Window functions and Impala

- Use a sample employee-style dataset for window aggregates, rank, dense_rank and row_number; understand ordering and derived-column filtering.
- Impala daemons, catalog and state-store roles, shell/scripts and the relationship with Hive metadata.
- Load/query sample data and inspect query logs; apply the required metadata refresh/invalidation for the selected deployment.
- Compare selected analytical queries and assess their correctness.

10. Day 16 — Ingestion, data flow and legacy migration

- Use NiFi/DataFlow concepts for supported ingestion and consider relational, file and API sources.
- Review legacy Sqoop import/export, compression, Avro inspection, incremental jobs and logs as a historical migration case; Apache Sqoop is retired, not the default new ingestion stack.
- Introduce Pig/Pig Latin, Flume and older workflow patterns as environment-dependent legacy context; do not require every historical tool in the current lab.
- Oozie orchestration is a retired Apache project; evaluate supported scheduling alternatives for the actual platform.
- Preserve practical data-movement and verification objectives without claiming legacy tools are current general-purpose recommendations.

11. Days 17–18 — Security, SDX and governance

- Separate authentication, authorisation, auditing and encryption; compare cloud SSO/storage controls with on-premises LDAP/Kerberos integration.
- SDX with Ranger, Atlas, Knox and catalog concepts; review access policies, lineage and stewardship.
- Use TLS terminology, certificates and client trust; demonstrate a small training CA/mTLS example only in the supplied environment.
- Introduce Kafka SASL/GSSAPI, principals, keytabs and ACLs with compatible broker/client configuration.
- Distinguish Kafka KRaft security from ZooKeeper security still relevant to selected other services or legacy deployments; version-specific migration controls apply.
- Review HDFS transparent encryption, storage encryption and historical Navigator Encrypt references without assuming identical availability across environments.

12. Day 19 — Workload management, data services and consolidation

- Review replication policies, workload monitoring, Cloudera Observability and historical Workload XM/Manager contexts; compare baselines and resource contention.
- Relate Data Engineering, Data Warehouse, Operational Database, AI and Data Flow services to the platform use case.
- Inspect an end-to-end small workflow and explain service boundaries, access, costs and recovery considerations.
- Consolidate the cloud/on-premises deployment and governance topics included in the current Generalist guide.

13. Day 20 — Review, practice and next steps

- Complete two bounded practice assessments and review the reasoning behind incorrect answers.
- Summarise the tested lab versions, useful operational evidence and remaining learning gaps.
- Use the official current Generalist guide for exam domains and registration requirements; practice tests are not official questions or a prediction of passing.

Certification context

The original source targeted CDP-0011, now presented by Cloudera as its Generalist exam. Provider scope and proctoring requirements can change. Consult the current official guide before separately arranging an assessment; no exam fee, credential or affiliation is implied by this course.