CERTLABZ CERTIFICATION BOARD

# Certified Big Data Professional (CBDP)

## Ten Sample Questions

These ten questions are written at the difficulty of the live exam. Six carry a figure, a data table, a configuration or a code listing; four are plain scenario questions, as on the live paper. Every one is a scenario a working big data professional would meet, and none of them appears in the live question bank.

The live exam is multistage adaptive: how you score on one stage selects the difficulty of the next, so no two candidates sit the same paper. It is a performance-based exam: each form can carry up to 5 to 12 performance-based questions alongside the multiple-choice questions, which means you build, sequence or complete something rather than only choosing from a list. Answers and full explanations follow each question here; the live exam does not disclose item-level answers.

### Question 1 — Domain 1: Big Data Foundations — DATA

Sarah Miller reviews the core big data characteristics, with one meaning missing.

| Characteristic | Meaning |
|---|---|
| Volume | Scale of data |
| Velocity | (to choose) |
| Variety | Mix of formats |
| Veracity | Trust in data |

Table 1.1 Characteristics and meaning.

What does velocity describe?

A. Number of servers
B. Speed of data
C. Cost of storage
D. Size of each record

**Correct answer: B**
Velocity describes the speed at which data arrives and must be processed, so the missing meaning is speed of data. Volume covers scale, variety covers formats, and veracity covers how far the data can be trusted.

## Question 2 Domain 2: Storage and Data Lakes DATA

James Carter maps each store to what it suits best, with one entry missing.

| Store | Best for |
|---|---|
| Data warehouse | Structured analytics |
| Data lake | (to choose) |
| Key-value store | Fast lookups |
| Object store | Cheap bulk files |

Table 2.1 Stores and best fit.

What is a data lake best for?

A Only clean tables
B Fixed schemas only
C Raw varied data
D Single row lookups

**Correct answer: C**
A data lake holds raw data in many formats and lets schema be applied on read, so it is best for raw varied data. Warehouses expect clean structured tables, key-value stores serve single lookups, and rigid schemas defeat the purpose of a lake.

## Question 3 Domain 3: Batch Processing DATA

Emma Wilson lists the traits of a batch job, with the data scope missing.

| Trait | Batch |
|---|---|
| Latency | Minutes to hours |
| Data scope | (to choose) |
| Trigger | Scheduled runs |
| Use case | Nightly reports |

Table 3.1 Batch job traits.

What is the batch data scope?

A One event at a time
B Live infinite stream
C Only tiny samples
D Large bounded sets

**Correct answer: D**
Batch processing runs over large bounded sets collected before the job starts, so the missing scope is large bounded sets. Single events and infinite live streams describe streaming, and batch is not limited to tiny samples.

## Question 4 Domain 4: Streaming and Real-Time CONFIG

David Brown reviews a streaming job defined by a short pipeline listing.

```
source: sensor topic
window: 5 min tumbling
agg:    average temp
sink:   alert service
```

Listing 4.1 The streaming pipeline.

What does the window clause do?

A Groups events by time
B Removes duplicate events
C Encrypts the stream
D Sorts by arrival order

**Correct answer: A**
A tumbling window groups incoming events into fixed time buckets so the average can be computed per interval. It does not encrypt, deduplicate, or reorder the stream, which are separate concerns handled elsewhere in the pipeline.

## Question 5 Domain 5: Data Pipelines and Orchestration CONFIG

Olivia Davis reads the task dependencies an orchestrator runs each night.

```
extract -> stage
stage   -> clean
clean   -> load
load    -> report
```

Listing 5.1 The pipeline dependencies.

What does the orchestrator guarantee here?

A Tasks run at random
B All steps run at once
C Tasks run in order
D Steps skip on error

**Correct answer: C**
The dependencies form a directed graph, so the orchestrator runs each task only after its upstream task succeeds, meaning tasks run in order. It does not fire steps at random, launch them all at once, or silently skip failures.

## Question 6 Domain 6: Governance, Quality and Security DATA

Ava Thompson maps each governance control to its purpose, with one missing.

| Control | Purpose |
|---|---|
| Data catalog | Find and describe data |
| Access policy | (to choose) |
| Lineage | Track data origins |
| Masking | Hide sensitive fields |

Table 6.1 Controls and purpose.

What does an access policy do?

A Speeds up all queries
B Limits who can read
C Deletes old backups
D Formats the reports

**Correct answer: B**
An access policy defines who may read or change a dataset, so its purpose is to limit who can read. Speeding queries, deleting backups, and formatting reports are unrelated operational tasks, not access controls.

## Question 7 Domain 1: Big Data Foundations

A team stores a few thousand rows but insists on a distributed cluster to process them, adding cost and complexity for no gain. What is the sound guidance?

A Match tool to scale
B Always use a cluster
C Buy more nodes
D Avoid all databases

**Correct answer: A**
Big data tooling earns its cost only at genuine scale, so the sound guidance is to match the tool to the data scale. A small dataset runs fine on a single machine, and adding nodes or clusters here only wastes money and effort.

## Question 8 Domain 2: Storage and Data Lakes

A data lake fills with undocumented files that nobody can find or trust, so analysts stop using it. What practice best prevents this data swamp?

A. Store more raw files
B. Delete the whole lake
C. Lock out all analysts
D. Catalog the metadata

**Correct answer: D**
A lake stays usable when every dataset carries metadata in a catalog, so cataloging the metadata prevents the swamp. Piling in more raw files makes it worse, and deleting the lake or blocking analysts throws away the value rather than fixing it.

## Question 9 Domain 4: Streaming and Real-Time

A fraud team must flag suspicious transactions within seconds of each purchase, but the current job runs only once a night. Which approach fits the need?

A. Nightly batch job
B. Stream processing
C. Manual weekly review
D. Quarterly reports

**Correct answer: B**
Flagging events within seconds calls for stream processing, which acts on each transaction as it arrives. A nightly batch job, weekly review, or quarterly report all add far too much delay to catch fraud in time.

## Question 10 Domain 5: Data Pipelines and Orchestration

A nightly pipeline sometimes runs twice after a restart, and each run appends the same rows, so totals double. What design best avoids the duplicates?

A. Run the job more often
B. Ignore the extra rows
C. Make loads idempotent
D. Turn off all retries

**Correct answer: C**
An idempotent load writes the same result no matter how many times it runs, so making loads idempotent avoids the doubled totals. Running more often worsens it, ignoring rows corrupts the data, and disabling retries leaves failures unhandled.

CBDP-001. https://certlabz.com/certifications/certified-big-data-professional.html