

## PRACTICE QUESTIONS

Microsoft

# AI-200

Azure AI Cloud Developer Associate

SHARED BY [u/happybyte1981](#) on aiexamcerts

---

|           |                                                      |
|-----------|------------------------------------------------------|
| EDITION   | Free Set 1                                           |
| QUESTIONS | 10                                                   |
| SHARED    | October 18, 2026                                     |
| EXAM      | Microsoft AI-200, Azure AI Cloud Developer Associate |

---

Shared by [u/happybyte1981](#) on aiexamcerts

10 practice questions for the Microsoft AI-200 exam, Azure AI Cloud Developer Associate, each followed by its answer and a short explanation. More free sets and the full practice exam: [aiexamcerts.com/microsoft/ai-200](https://aiexamcerts.com/microsoft/ai-200). Spotted a mistake? Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam code AI-200 and the question number.

1. In pgvector, an HNSW index on an embedding column speeds up approximate nearest-neighbor queries on large tables, at the cost of extra memory for the index.

- A True
- B False

**Answer: A** (True)

HNSW builds a layered graph over the vectors, which lets queries find near neighbors without scanning every row, trading some recall for speed. The graph consumes memory and must be maintained on writes. The statement is true.

2. An application needs geo-replicated images and private endpoints for Azure Container Registry. Which registry tier is required?

- A Basic
- B Standard
- C Developer
- D Premium

**Answer: D**

Geo-replication and private endpoints are features of the Premium tier of Azure Container Registry, which is the tier that adds the networking and replication capabilities. The Basic and Standard tiers do not provide geo-replication. The Developer option is not an Azure Container Registry tier.

3. An AI service opens a new PostgreSQL connection for every embedding lookup, and throughput collapses under load. Which change most directly improves throughput?

- A Routing connections through a connection pooler such as the built-in PgBouncer
- B Raising max\_connections to 10000 on the server, so every client keeps its own connection open
- C Disabling SSL on every client connection
- D Dropping all indexes on the embedding table

**Answer: A**

A connection pooler keeps a set of server connections open and reuses them, so each lookup avoids the cost of establishing a new connection. Raising max\_connections only allows more connections to pile up and consume memory. Disabling SSL and dropping indexes do not address connection setup cost and the second would slow queries.

4. The Redis command that sets a time to live, in seconds, on an existing key is \_\_\_\_\_.

**Answer: EXPIRE**

EXPIRE sets a timeout on a key, after which Redis deletes it automatically, which is how cached values get bounded lifetimes in Azure Managed Redis. The command name is written as EXPIRE. Azure Managed Redis is Redis compatible, so the standard command set applies.

5. A RAG application caches generated answers in Azure Managed Redis and must never serve an answer older than 15 minutes. Which approach meets the requirement?

- A Set a time to live of 15 years and rely on memory eviction
- B Set a time to live of 900 seconds on each cached key
- C Store the answer in a Redis list with no expiration
- D Run KEYS \* on every request to find stale entries

**Answer: B**

A time to live of 900 seconds expires each key 15 minutes after it is written, so Redis removes stale answers without application code. Relying on memory eviction gives no time bound, and a list without expiration keeps answers indefinitely. Running KEYS on every request is slow and does not enforce any age limit.

6. A multi-stage Dockerfile reduces the final image size by copying only the build outputs into the last stage, leaving compilers and build tools out of the shipped image.

- A True
- B False

**Answer: A (True)**

In a multi-stage build, the earlier stages can install compilers and build dependencies, and the final stage copies just the artifacts it needs, so the shipped image is smaller and has less attack surface. The statement is true. This is the standard way to keep runtime images lean.

7. A Kubernetes Deployment manifest for an AKS cluster must run three replicas of a pod. Which field in the manifest sets that number?

- A `metadata.labels`
- B `spec.template.spec.env`
- C `spec.replicas`
- D `status.phase`

**Answer: C**

The `spec.replicas` field of a Deployment declares the desired number of pod replicas, and the controller creates or removes pods to match it. The `labels` field is metadata used for selection, not for counting. The status phase is reported by the cluster and is not set by the author.

8. A chat service stores messages in a multi-tenant Cosmos DB container. Which property is the best partition key for reading a conversation's messages together?
- A A constant value such as messages for every item
  - B The message timestamp
  - C A random GUID generated for every write, which spreads writes but scatters each conversation
  - D The conversation identifier, which keeps one conversation in one logical partition

**Answer: D**

A conversation identifier groups all messages of one conversation in one logical partition, so reading that conversation is an efficient single-partition query. A constant value puts every write in one partition, and a timestamp concentrates recent writes on one partition. A random GUID spreads writes well but forces a cross-partition query to read a conversation.

9. Which two statements about the Azure Cosmos DB change feed are correct? (Choose two.)
- A Changes are ordered within each partition key value
  - B Deletes are not reported by default, so applications often model them with a soft-delete flag
  - C Only Azure Functions can read the change feed
  - D Changes are returned in one global order across all partition keys
  - E The change feed requires a separate Service Bus namespace

**Answer: A, B**

The change feed preserves the order of changes for each logical partition key value, and it does not emit delete events by default, so teams commonly write a soft-delete marker. Any client using the SDK can read the change feed, not only Azure Functions. Global ordering across partition keys and a required Service Bus namespace are not properties of the change feed.

10. A PostgreSQL table stores document chunks, their text, source metadata and 1536-dimension embeddings. Which design is most appropriate?
- A Use a vector(1536) column and keep chunk text and metadata in typed columns
  - B Store the embeddings as a comma-separated text column
  - C Store the embeddings in a JSON column with no declared dimension
  - D Store each embedding as one row per float value

**Answer: A**

A vector(1536) column enforces the dimension and supports pgvector distance operators and HNSW indexes, while typed columns keep the chunk text and metadata easy to filter. A text or JSON column cannot use the pgvector distance operators or vector indexes directly. Splitting each float into its own row makes similarity search slow and complicated.