Assignment 6 Solutions

Wiam Skakri

November 28, 2025

1 Question 1: Graph DBs and NoSQL

1.1 Question 1(a)

Given scenarios and query tasks below, select which type of NoSQL database best fits the application's needs.

Scenario / Application	**Query / Task**	**Best-Fit DB Type?**
Real-time personalization and caching for an online retail website, storing session tokens and user carts for millions of active users.	Fast lookup by session ID or user ID; extremely low latency reads/writes for ephemeral data.	**A**
Global-scale time-series analytics platform monitoring IoT devices, energy grids, or financial tick data.	Range scans and aggregations over time intervals for billions of sensor readings.	**B**
Content management and flexible JSON-like data for an e-commerce platform storing product catalogs, reviews, and metadata with different fields per item.	Search by product category, flexible schema, and nested document queries.	**C**
Social network and recommendation graph connecting users, influencers, and shared interests.	Multi-hop traversal queries like "friends-of-friends," "who influences whom," or "shortest path between two users."	**D**
AI-powered semantic search system for images, research papers, or user queries.	Retrieve top-k items with the most similar vector embeddings (cosine similarity or Euclidean distance).	**E**

1.2 Question 1(b)

Property Graph Model for Tech Reviewer Scenario

Nodes

Person Nodes:

```
(:Person {name: "Taylor Chen", role: "Tech Reviewer"})
(:Person {name: "Lin Zhang", role: "Software Engineer"})
(:Person {name: "Sam Rivera", role: "Tech Creator"})
```

Company Nodes:

```
(:Company {company_name: "Apple"})
(:Company {company_name: "Meta"})
```

Device Nodes:

```
(:Device {device_name: "VisionPro 2", type: "AR/VR Headset"})
(:Device {device_name: "Meta Quest 4", type: "VR Headset"})
```

Feature Nodes:

```
(:Feature {feature_name: "Eye-Track Pro"})
(:Feature {feature_name: "SpatialCast"})
```

Video Nodes:

```
(:Video {video_title: "Why the VisionPro 2 Changes Everything",
         date: "Feb 12, 2025", views: 2300000})
(:Video {video_title: "VisionPro 2 vs Meta Quest 4 Comparison",
         date: "Feb 2025", type: "collaboration"})
```

Comment Node:

```
(:Comment {text: "Super interesting breakdown! I'm curious how
            this compares to Meta Quest 4.", likes: "several"})
```

Relationships

Content Creation Relationships:

```
(Taylor)-[:POSTED]->(Video1: "Why the VisionPro 2 Changes Everything")
(Taylor)-[:POSTED]->(Video2: "VisionPro 2 vs Meta Quest 4 Comparison")
(Sam Rivera)-[:POSTED]->(Video2: "VisionPro 2 vs Meta Quest 4 Comparison")
(Taylor)-[:REVIEWED]->(VisionPro 2)
```

Collaboration Relationships:

```
(Taylor)-[:COLLABORATED_WITH]->(Sam Rivera)
```

Social Interaction Relationships:

```
(Lin Zhang)-[:COMMENTED {on_video: "Video1"}]->(Comment)
(Viewers)-[:LIKED]->(Comment)
```

Employment Relationships:

```
(Lin Zhang)-[:WORKS_AT]->(Meta)
```

Manufacturing Relationships:

```
(VisionPro 2)-[:MANUFACTURED_BY]->(Apple)
(Meta Quest 4)-[:MANUFACTURED_BY]->(Meta)
```

Feature Relationships:

```
(VisionPro 2)-[:HAS_FEATURE]->(Eye-Track Pro)
(VisionPro 2)-[:HAS_FEATURE]->(SpatialCast)
```

Comparison Relationships:

```
(Video2)-[:COMPARES]->(VisionPro 2)
(Video2)-[:COMPARES]->(Meta Quest 4)
```

1.3 Question 1(c)

Describe in natural language what each Cypher query does.

Query 1:

```
MATCH (d:Device {name: "VisionPro 2"})-[:HAS_FEATURE]->(f:Feature)
RETURN f.name;
```

Natural Language Description: "Find all features of the VisionPro 2 device, and return the name of each feature."

Query 2:

```
MATCH (p:Person)-[:COMMENTED_ON]->(v:Video {title: "Why the VisionPro 2
                                                Changes Everything"})
RETURN p.name;
```

Natural Language Description: "Find all people who commented on the video titled 'Why the VisionPro 2 Changes Everything', and return each person's name."

Query 3:

```
MATCH (t:Person {name: "Taylor Chen"})-[:COLLABORATED_WITH]->(s:Person)
RETURN s.name;
```

Natural Language Description: "Find all people that Taylor Chen has collaborated with, and return their names."

2 Question 2: Vectors Similarity

Given movie embedding vectors:

- Movie A: "Laser Quest" (Sci-Fi Action): $\mathbf{a} = (1, 2, 3)$
- Movie B: "Sunny Days" (Family Comedy): $\mathbf{b} = (2, 1, 0)$
- Movie C: "Nebula Dreams" (Sci-Fi Drama): $\mathbf{c} = (1, 2, 2)$

2.1 Question 2(a)

Compute the similarities or distances for the following queries.

(i) Cosine Similarity: "Laser Quest" to "Nebula Dreams"

Formula: $\cos\_sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$

Step 1: Compute the dot product $\mathbf{a} \cdot \mathbf{c}$

$$\begin{aligned} \mathbf{a} \cdot \mathbf{c} &= (1)(1) + (2)(2) + (3)(2) \\ &= 1 + 4 + 6 \\ &= 11 \end{aligned}$$

Step 2: Compute the norms

$$\begin{aligned} \|\mathbf{a}\| &= \sqrt{1^2 + 2^2 + 3^2} = \sqrt{1 + 4 + 9} = \sqrt{14} \\ \|\mathbf{c}\| &= \sqrt{1^2 + 2^2 + 2^2} = \sqrt{1 + 4 + 4} = \sqrt{9} = 3 \end{aligned}$$

Step 3: Compute cosine similarity

$$\cos\_sim(\mathbf{a}, \mathbf{c}) = \frac{11}{\sqrt{14} \times 3} = \frac{11}{3\sqrt{14}} = \frac{11}{11.225} \approx \boxed{0.980}$$

Interpretation: A cosine similarity of 0.980 (very close to 1) indicates that "Laser Quest" and "Nebula Dreams" are highly similar in their embedding space, which makes sense as both are Sci-Fi movies.

(ii) Euclidean Distance: "Laser Quest" to "Sunny Days"

Formula: $d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$

$$\begin{aligned} d_E(\mathbf{a}, \mathbf{b}) &= \sqrt{(1-2)^2 + (2-1)^2 + (3-0)^2} \\ &= \sqrt{(-1)^2 + (1)^2 + (3)^2} \\ &= \sqrt{1 + 1 + 9} \\ &= \sqrt{11} \\ &\approx \boxed{3.317} \end{aligned}$$

Interpretation: The Euclidean distance of $\sqrt{11} \approx 3.317$ indicates a moderate distance between "Laser Quest" (Sci-Fi Action) and "Sunny Days" (Family Comedy), reflecting their different genres.

(iii) Manhattan Distance: "Laser Quest" to "Nebula Dreams"

Formula: $d_{L1}(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$

$$\begin{aligned} d_{L1}(\mathbf{a}, \mathbf{c}) &= |1-1| + |2-2| + |3-2| \\ &= 0 + 0 + 1 \\ &= \boxed{1} \end{aligned}$$

Interpretation: A Manhattan distance of only 1 confirms that "Laser Quest" and "Nebula Dreams" are very close in the embedding space, differing only in the third dimension (action level: 3 vs 2).

(iv) Dot Product: "Sunny Days" and "Nebula Dreams"

Formula: $\text{dot}(\mathbf{x}, \mathbf{y}) = \sum_i x_i y_i$

$$\begin{aligned} \text{dot}(\mathbf{b}, \mathbf{c}) &= (2)(1) + (1)(2) + (0)(2) \\ &= 2 + 2 + 0 \\ &= \boxed{4} \end{aligned}$$

Interpretation: The dot product of 4 measures the alignment of the two vectors. A positive value indicates some thematic alignment between "Sunny Days" and "Nebula Dreams", though not as strong as between the two Sci-Fi movies.

2.2 Question 2(b)

K-Means Clustering Algorithm

Given:

- Data points: $D = \{2, 4, 5, 8, 12, 13\}$
- Number of clusters: $k = 2$
- Initial centroids: $\mu_1(0) = 4$, $\mu_2(0) = 12$

Round 1

Step 1: Assign points to clusters based on distance to centroids

Current centroids: $\mu_1 = 4$, $\mu_2 = 12$

Point	Distance to $\mu_1 = 4$	Distance to $\mu_2 = 12$	Assigned Cluster
2	$\|2-4\| = 2$	$\|2-12\| = 10$	Cluster 1
4	$\|4-4\| = 0$	$\|4-12\| = 8$	Cluster 1
5	$\|5-4\| = 1$	$\|5-12\| = 7$	Cluster 1
8	$\|8-4\| = 4$	$\|8-12\| = 4$	Cluster 1 (tie $\rightarrow$ first cluster)
12	$\|12-4\| = 8$	$\|12-12\| = 0$	Cluster 2
13	$\|13-4\| = 9$	$\|13-12\| = 1$	Cluster 2

Clusters after Round 1 Assignment:

- Cluster 1 (Low-engagement): $\{2, 4, 5, 8\}$
- Cluster 2 (High-engagement): $\{12, 13\}$

Step 2: Update centroids

$$\mu_1(1) = \frac{2+4+5+8}{4} = \frac{19}{4} = \boxed{4.75}$$
$$\mu_2(1) = \frac{12+13}{2} = \frac{25}{2} = \boxed{12.5}$$

Round 2

Step 1: Reassign points based on new centroids

Current centroids: $\mu_1 = 4.75$, $\mu_2 = 12.5$

Point	Distance to $\mu_1 = 4.75$	Distance to $\mu_2 = 12.5$	Assigned Cluster
2	$\|2-4.75\| = 2.75$	$\|2-12.5\| = 10.5$	Cluster 1
4	$\|4-4.75\| = 0.75$	$\|4-12.5\| = 8.5$	Cluster 1
5	$\|5-4.75\| = 0.25$	$\|5-12.5\| = 7.5$	Cluster 1
8	$\|8-4.75\| = 3.25$	$\|8-12.5\| = 4.5$	Cluster 1
12	$\|12-4.75\| = 7.25$	$\|12-12.5\| = 0.5$	Cluster 2
13	$\|13-4.75\| = 8.25$	$\|13-12.5\| = 0.5$	Cluster 2

Step 3: Clusters reassigned?

No change — The cluster assignments remain the same:

- Cluster 1 (Low-engagement): $\{2, 4, 5, 8\}$
- Cluster 2 (High-engagement): $\{12, 13\}$

Step 2: Update centroids

Since cluster memberships did not change:

$$\mu_1(2) = \frac{2+4+5+8}{4} = \frac{19}{4} = \boxed{4.75}$$
$$\mu_2(2) = \frac{12+13}{2} = \frac{25}{2} = \boxed{12.5}$$

2.3 Question 2(c)

K-NN Search for Genre Prediction

Given users and their coordinates:

User	Coordinates (x, y)	Favorite Genre
A	(1, 2)	Sci-Fi
B	(2, 4)	Comedy
C	(3, 3)	Sci-Fi
D	(6, 5)	Drama
E	(7, 3)	Drama

New user X has embedding: $\mathbf{X} = (3, 1)$

Step 1: Calculate Euclidean Distance from X to Each User

Using the formula: $d_E(\mathbf{X}, \mathbf{p}) = \sqrt{(x_1 - p_1)^2 + (x_2 - p_2)^2}$

$$
\begin{aligned}
d(X, A) &= \sqrt{(3-1)^2 + (1-2)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.236 \\
d(X, B) &= \sqrt{(3-2)^2 + (1-4)^2} = \sqrt{1+9} = \sqrt{10} \approx 3.162 \\
d(X, C) &= \sqrt{(3-3)^2 + (1-3)^2} = \sqrt{0+4} = \sqrt{4} = 2 \\
d(X, D) &= \sqrt{(3-6)^2 + (1-5)^2} = \sqrt{9+16} = \sqrt{25} = 5 \\
d(X, E) &= \sqrt{(3-7)^2 + (1-3)^2} = \sqrt{16+4} = \sqrt{20} \approx 4.472
\end{aligned}
$$

Step 2: Sort Users by Distance and Find 3 Nearest Neighbors

Rank	User	Distance	Favorite Genre
1	C	$\sqrt{4} = 2$	Sci-Fi
2	A	$\sqrt{5} \approx 2.236$	Sci-Fi
3	B	$\sqrt{10} \approx 3.162$	Comedy
4	E	$\sqrt{20} \approx 4.472$	Drama
5	D	$\sqrt{25} = 5$	Drama

3 Nearest Neighbors: C, A, B

Step 3: Majority Vote for Genre Prediction

Neighbor	Genre
C	Sci-Fi
A	Sci-Fi
B	Comedy

Vote Count:

- Sci-Fi: 2 votes (C, A)
- Comedy: 1 vote (B)

Prediction

Using majority vote among the 3 nearest neighbors:

User X's predicted favorite genre is **Sci-Fi**

Explanation: The two closest users to X (users C and A) both prefer Sci-Fi, while only one neighbor (B) prefers Comedy. Therefore, by majority vote (2 vs 1), we predict that user X will prefer Sci-Fi content.

3 Question 3: IVF and PQ

3.1 Question 3(a)

IVF Index with Two Coarse Centroids

Given centroids:

- $c_1 = (1.5, 1.5)$ — Low-intensity / lightweight movies
- $c_2 = (6.0, 5.5)$ — High-intensity / epic movies

Assign each movie to its nearest centroid using Euclidean distance.

Distance Calculations

Movie A: Laser Quest (1, 2)

$$d(A, c_1) = \sqrt{(1-1.5)^2 + (2-1.5)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707$$
$$d(A, c_2) = \sqrt{(1-6)^2 + (2-5.5)^2} = \sqrt{25 + 12.25} = \sqrt{37.25} \approx 6.103$$

Nearest centroid: c_1 ✓

Movie B: Sunny Days (2, 1)

$$d(B, c_1) = \sqrt{(2-1.5)^2 + (1-1.5)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707$$
$$d(B, c_2) = \sqrt{(2-6)^2 + (1-5.5)^2} = \sqrt{16 + 20.25} = \sqrt{36.25} \approx 6.021$$

Nearest centroid: c_1 ✓

Movie C: Nebula Dreams (4, 5)

$$d(C, c_1) = \sqrt{(4-1.5)^2 + (5-1.5)^2} = \sqrt{6.25 + 12.25} = \sqrt{18.5} \approx 4.301$$
$$d(C, c_2) = \sqrt{(4-6)^2 + (5-5.5)^2} = \sqrt{4 + 0.25} = \sqrt{4.25} \approx 2.062$$

Nearest centroid: c_2 ✓

Movie D: Shadow Empire (7, 6)

$$d(D, c_1) = \sqrt{(7-1.5)^2 + (6-1.5)^2} = \sqrt{30.25 + 20.25} = \sqrt{50.5} \approx 7.106$$
$$d(D, c_2) = \sqrt{(7-6)^2 + (6-5.5)^2} = \sqrt{1 + 0.25} = \sqrt{1.25} \approx 1.118$$

Nearest centroid: c_2 ✓

Resulting Clusters

Cluster	Centroid	Movies
Cluster 1 (Low-intensity)	$c_1 = (1.5, 1.5)$	Laser Quest (A), Sunny Days (B)
Cluster 2 (High-intensity)	$c_2 = (6.0, 5.5)$	Nebula Dreams (C), Shadow Empire (D)

3.2 Question 3(b)

IVF Search for Nearest Neighbor of Q

Query: New movie $\mathbf{Q} = (3, 3.5)$

Step 1: Find Nearest Centroid to Q

$$d(Q, c_1) = \sqrt{(3-1.5)^2 + (3.5-1.5)^2} = \sqrt{2.25+4} = \sqrt{6.25} = 2.5$$
$$d(Q, c_2) = \sqrt{(3-6)^2 + (3.5-5.5)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.606$$

Nearest centroid: c_1 (distance = 2.5)

Step 2: Search Only in Cluster 1 (One Probe)

Cluster 1 contains: Laser Quest (A) at (1, 2) and Sunny Days (B) at (2, 1)

Calculate distances from Q to movies in Cluster 1:

$$d(Q, A) = \sqrt{(3-1)^2 + (3.5-2)^2} = \sqrt{4+2.25} = \sqrt{6.25} = 2.5$$
$$d(Q, B) = \sqrt{(3-2)^2 + (3.5-1)^2} = \sqrt{1+6.25} = \sqrt{7.25} \approx 2.693$$

IVF Search Result

Nearest Neighbor (IVF): **Laser Quest (A)** with distance 2.5

3.3 Question 3(c)

Brute Force Verification: Is the IVF Answer the True Nearest Neighbor?

Compute Distance from Q to ALL Movies

$$d(Q, A) = \sqrt{(3-1)^2 + (3.5-2)^2} = \sqrt{4+2.25} = \sqrt{6.25} = 2.5$$
$$d(Q, B) = \sqrt{(3-2)^2 + (3.5-1)^2} = \sqrt{1+6.25} = \sqrt{7.25} \approx 2.693$$
$$d(Q, C) = \sqrt{(3-4)^2 + (3.5-5)^2} = \sqrt{1+2.25} = \sqrt{3.25} \approx 1.803$$
$$d(Q, D) = \sqrt{(3-7)^2 + (3.5-6)^2} = \sqrt{16+6.25} = \sqrt{22.25} \approx 4.717$$

Ranked List (Brute Force)

Rank	Movie	Distance to Q	Note
1	Nebula Dreams (C)	$\sqrt{3.25} \approx 1.803$	**TRUE Nearest Neighbor**
2	Laser Quest (A)	$\sqrt{6.25} = 2.5$	IVF Answer
3	Sunny Days (B)	$\sqrt{7.25} \approx 2.693$	
4	Shadow Empire (D)	$\sqrt{22.25} \approx 4.717$	

Conclusion

The IVF answer is **NOT** the true nearest neighbor.

Analysis:

- **True NN:** Nebula Dreams (C) with distance ≈ 1.803
- **IVF Answer:** Laser Quest (A) with distance $= 2.5$
- **IVF Rank:** The IVF answer is the **2nd closest** movie (not the 1st)

Why did IVF miss the true NN?

The true nearest neighbor (Nebula Dreams) is in **Cluster 2**, but the IVF search only probed **Cluster 1** because Q was closer to centroid c_1. This demonstrates the fundamental trade-off of IVF indexing:

- **Advantage:** Faster search (examined only 2 movies instead of 4)
- **Disadvantage:** May return approximate results, missing the true NN if it lies in a different cluster