1. Given two sets, S={a,b,c,d,e,f,g,r} and T={a,c,d,f,w,x,y}, calculate the Jaccard distance of these two sets.

$S = \{a, b, c, d, e, f, g, r\}$ $\quad S \cap T = \{a, c, d, f\}$

$T = \{a, c, d, f, w, x, y\}$ $\quad S \cup T = \{a, b, c, d, e, f, g, w, r, x, y\}$

$$d(S,T) = 1 - \frac{|S \cap T|}{|S \cup T|} = 1 - \frac{4}{11} = \frac{7}{11}$$

2. Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and {2, 4, 6}.

(i) $S_1 = \{1, 2, 3, 4\}$ $\quad S_1 \cap S_2 = \{2, 3\}$ $\quad sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{2}{6} = \frac{1}{3}$

$S_2 = \{2, 3, 5, 7\}$ $\quad S_1 \cup S_2 = \{1, 2, 3, 4, 5, 7\}$

(ii) $S_1 = \{1, 2, 3, 4\}$ $\quad S_1 \cap S_3 = \{2, 4\}$ $\quad sim(S_1, S_3) = \frac{|S_1 \cap S_3|}{|S_1 \cup S_3|} = \frac{2}{5}$

$S_3 = \{2, 4, 6\}$ $\quad S_1 \cup S_3 = \{1, 2, 3, 4, 6\}$

(iii) $S_2 = \{2, 3, 5, 7\}$ $\quad S_2 \cap S_3 = \{2\}$ $\quad sim(S_2, S_3) = \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|} = \frac{1}{6}$

$S_3 = \{2, 4, 6\}$ $\quad S_2 \cup S_3 = \{2, 3, 4, 5, 6, 7\}$

3. What are the first eight 3-shingles in the following sentence? A relation can be stored as a file in a distributed file system.

(3) A relation can be stored as a file in a distributed file system.

1 A_r
2 _re
3 rel
4 ela
5 lat
6 ati
7 tio
8 ion

4. The following is a matrix with 6 rows.
(a) Compute the minhash signature for each column if we use the following three hash functions.

Element	S_1	S_2	S_3	S_4	$h_1(x)$ $2x+1 \bmod 6$	$h_2(x)$ $3x+2 \bmod 6$	$h_3(x)$ $5x+2 \bmod 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Minhash Signature Matrix:

	S_1	S_2	S_3	S_4
$h_1(0)$	∞	1	∞	1
$h_2(0)$	∞	2	∞	2
$h_3(0)$	∞	2	∞	2
$h_1(1)$	∞	1	∞	1
$h_2(1)$	∞	2	∞	2
$h_3(1)$	∞	1	∞	2
$h_1(2)$	5	1	∞	1
$h_2(2)$	2	2	∞	2
$h_3(2)$	0	1	∞	0
$h_1(3)$	5	1	1	1
$h_2(3)$	2	2	5	2
$h_3(3)$	0	1	5	0
$h_1(4)$	5	1	1	1
$h_2(4)$	2	2	2	2
$h_3(4)$	0	1	4	0
$h_1(5)$	5	1	1	1
$h_2(5)$	2	2	2	2
$h_3(5)$	0	1	4	0

Final minhash signature matrix:

S_1	S_2	S_3	S_4
5	1	1	1
2	2	2	2
0	1	4	0

↳ Only $h_3(x)$ is true permutation

(b) Which of these hash functions are true permutations?
The third hash function is true permutation with no duplicates.

5. Cloud introduces several security issues including multi-tenancy, lack of

control, and lack of trust. Please explain the above three security issues and discuss possible solutions.

	Issues	Solutions
Multi-tenancy	When multiple tenants share the same cloud infrastructure, there's a risk of data leakage or unauthorized access if proper isolation mechanisms aren't in place, allowing one tenant's data to potentially be accessed by another.	**Data segregation:** Implement strong data isolation techniques to separate tenant data at the storage and application levels. **Access control:** Utilize robust identity and access management (IAM) systems to restrict access to data based on user roles and permissions. **Network segmentation:** Create separate virtual networks for each tenant to further isolate traffic and prevent cross-tenant communication.
Lack of control	When using cloud services, organizations may lose some level of control over their data and infrastructure compared to managing their own on-premises systems.	**Configuration management tools:** Leverage cloud provider tools to consistently configure security settings and manage system updates across the environment. **Auditing and logging:** Implement robust logging mechanisms to monitor cloud activities and identify potential security incidents. **Cloud security posture management (CSPM):** Utilize dedicated tools to assess cloud security posture and identify misconfigurations.
Lack of trust	Concerns about the cloud provider's ability to safeguard sensitive data, especially when data is stored in a third-party data center.	**Compliance certifications:** Choose cloud providers with robust security certifications (SOC 2, HIPAA, PCI DSS) to demonstrate adherence to industry standards. **Data encryption:** Encrypt sensitive data both at rest and in transit for protection even if accessed by the cloud provider. **Third-party audits:** Conduct regular independent audits of the cloud provider's security practices.

6. Explain side-channel attacks in cloud security.

A method where a malicious actor extracts sensitive information from a cloud system by analyzing unintended data leaks from the physical implementation of the system, such as power consumption, timing variations, or electromagnetic radiation, rather than directly attacking the encryption algorithms themselves.