Faculty of Exact Sciences

Department of Computer science

For first-year Master: Internet of Things & Cybersecurity

Big Data: Exam 2025

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv

EX 01 (07 P): Convert the unstructured data in the following two paragraphs to semi-structured data and then to structured data.

Paragraph 01 "Global Finance Corp is a leading financial company established in 1995 and headquartered in HASSI MASSOUED, ALGERIA. The company is led by Chief Executive Officer (CEO) DRISS ELHARETH and employs over 12,000 professionals worldwide. It provides a wide range of services, including corporate banking, retail banking, investment solutions, and insurance. In 2024, the company reported an annual revenue of $4.2 billion. For inquiries, you can contact the company via email at info@globalfinance.com or phone at +213-800-555-0199".

Paragraph 02 "Connectify Solutions is a leading tech company specializing in social networking applications. Founded in 2010 and headquartered in ELOUED, ALGERIA, the company is under the leadership of Chief Executive Officer (CEO) YACIN LABIDI. It employs over 5,000 talented professionals worldwide. Connectify Solutions offers services such as instant messaging, video conferencing, team collaboration tools, and social networking platforms. In 2024, the company reported an annual revenue of $1.5 billion. For inquiries, you can reach the team via email at support@connectify.com or call their hotline at ++213-415-555-1234.

EX 02 (07 P) : Given the nature of the data, processing, and storage that each company (Global Finance Corp and Connectify Solutions) handles, **explain for each:**

- ✓ **The nature of the data and processing needs;**
- ✓ **The characteristics (ACID; BASE; CAP theorem Compliance) of appropriate databases;**
- ✓ **The recommended databases**.

EX 03 (06 P) : For HDFS Disk Rebalancer , suppose we have a machine with four disks – Disk1, Disk2, Disk3, Disk4.

	Disk1	Disk2	Disk3	Disk4
Capacity	400 GB	600 GB	700 GB	1000 GB
DfsUsed	**200 GB**	**152 GB**	**600 GB**	**950 GB**

- ✓ **Calculate the Volume Data Density of the disks. What do these results indicate?**

Faculty of Exact Sciences

Department of Computer science

For first-year Master: Internet of Things & Cybersecurity

Big Data: Exam Solution   2025

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv

EX 01 (07 P): Convert the unstructured data in the following two paragraphs to semi-structured data and then to structured data.

Paragraph 01 "Global Finance Corp is a leading financial company established in 1995 and headquartered in HASSI MASSOUED, ALGERIA. The company is led by Chief Executive Officer (CEO) DRISS ELHARETH and employs over 12,000 professionals worldwide. It provides a wide range of services, including corporate banking, retail banking, investment solutions, and insurance. In 2024, the company reported an annual revenue of $4.2 billion. For inquiries, you can contact the company via email at info@globalfinance.com or phone at +213-800-555-0199**"**.

Paragraph 02 "Connectify Solutions is a leading tech company specializing in social networking applications. Founded in 2010 and headquartered in ELOUED, ALGERIA, the company is under the leadership of Chief Executive Officer (CEO) YACIN LABIDI. It employs over 5,000 talented professionals worldwide. Connectify Solutions offers services such as instant messaging, video conferencing, team collaboration tools, and social networking platforms. In 2024, the company reported an annual revenue of $1.5 billion. For inquiries, you can reach the team via email at support@connectify.com or call their hotline at ++213-415-555-1234.

Solution

- ✓ **To : Semi-structured data** 2*2

```
]
}
company_name": "Global Finance Corp","
established": 1995,"
headquarters": " HASSI MASSOUED, ALGERIA. ","
CEO": " DRISS ELHARETH ","
employees": 12000,"
services": ["
corporate banking","
retail banking","
investment solutions","
```

```
insurance""
‘[
annual_revenue": 4.2,"
currency": "billion USD","
contact": {"
email": "info@globalfinance.com","
phone": "+1-800-555-0199""
{
‘{
}
company_name": "Connectify Solutions","
established": 2010,"
headquarters": " ELOUED, ALGERIA ","
CEO": " YACIN LABIDI ","
employees": 5000,"
services": ["
instant messaging","
video conferencing","
team collaboration tools","
social networking platforms""
‘[
annual_revenue": 1.5,"
currency": "billion USD","
contact": {"
email": "support@connectify.com","
phone": "+1-415-555-1234""
{
{
]
```

- ✓ **To : Structured data : 1.5*2**

Company Name	Established	Headquarters	CEO	Employees	Services	Annual Revenue (Billion USD)	Email	Phone
Global Finance Corp	1995	HASSI MASSOUED, ALGERIA.	DRISS ELHARETH	12,000	Corporate banking, Retail banking, Investment solutions, Insurance	4.2	info@globalfinance.com	+1-800-555-0199
Connectify Solutions	2010	ELOUED, ALGERIA	YACIN LABIDI	5,000	Instant messaging, Video conferencing, Team collaboration tools, Social networking	1.5	support@connectify.com	+1-415-555-1234

EX 02 (07 P) : Given the nature of the data, processing, and storage that each company (Global Finance Corp and Connectify Solutions) handles, **explain for each :**

- ✓ **The nature of the data and processing needs;**
- ✓ **The characteristics (ACID; BASE ; CAP theorem Compliance) of appropriate databases;**
- ✓ **The recommended databases**.

Solution (07 P = 3.5*2) :

To determine the characteristics of the appropriate databases for **Global Finance Corp** and **Connectify Solutions**, we can analyze the nature of their data, processing, and storage needs based on two primary database paradigms: **ACID (Relational databases based on SQL)** and **BASE (Distributed systems such as NoSQL, Hadoop)**. Additionally, the **CAP theorem** helps us evaluate the trade-offs involved in distributed database design.

1. Global Finance Corp Company

Nature of Data and Processing Needs:

- **Data Characteristics:** Transactional data (e.g., banking transactions, customer information, financial records) that require strict consistency, durability, and reliability.
- **Storage Requirements:** Medium to high data volumes with structured formats.
- **Critical Needs:**
  - High reliability and data consistency (e.g., for auditing and compliance).
  - Support for complex queries and transactions.

Appropriate Database Characteristics:

- **ACID Compliance:** Ensures **atomicity**, **consistency**, **isolation**, and **durability**, crucial for financial transactions.
- **CAP Theorem Trade-offs:** Prioritize **Consistency** and **Partition Tolerance**. Availability may be slightly compromised during partitioning to maintain strict consistency.

Recommended Database:

- **Traditional Relational Databases (SQL):** Examples include:
  - **Oracle Database**: Robust for large-scale financial systems.
  - **Microsoft SQL Server** or **PostgreSQL**: For handling transactions with high consistency and performance.
  - **MySQL**: A cost-effective option for smaller or medium-scale setups.

2. Connectify Solutions Company

Nature of Data and Processing Needs:

- **Data Characteristics:** Social networking data such as user interactions, messaging, real-time collaboration, and media (e.g., videos and images).
- **Storage Requirements:** Massive, unstructured, and semi-structured data volumes, often distributed globally.
- **Critical Needs:**
  - Scalability to handle spikes in user activity.
  - High availability to ensure uptime.
  - Eventual consistency is acceptable for certain operations.

Appropriate Database Characteristics:

- **BASE Compliance:** Focuses on **availability** and **partition tolerance**, with eventual consistency being acceptable for social networking use cases.
- **CAP Theorem Trade-offs:** Prioritize **Availability** and **Partition Tolerance** over immediate consistency.

Recommended Database:

- **NoSQL and Distributed Systems:** Examples include:
  - **Apache Cassandra**: Ideal for high availability and scalability.
  - **MongoDB**: Supports flexible schemas for semi-structured data.
  - **Hadoop:** For batch processing and analytics on massive datasets.

2- **For HDFS Disk Rebalancer , suppose we have a machine with four disks – Disk1, Disk2, Disk3, Disk4. (06 P = 1*4 + 2)**

	Disk1	Disk2	Disk3	Disk4
Capacity	400 GB	600 GB	700 GB	1000 GB
DfsUsed	**200 GB**	**152 GB**	**600 GB**	**950 GB**
DfsUsedRatio	**0.5**	**0.25**	**0.85**	**0.95**
VolumeDataDensity	**0.20**	**0.45**	**-0.15**	**-0.24**

Calculate the Volume Data Density of the disks. What do these results indicate?

In this example,
Total capacity= 400 + 600 + 700 + 1000 = 2700GB and
Total Used= 200 + 152 + 600 + 950 = 1902 GB
Therefore, the ideal storage on each volume/disk is:
Ideal storage = total Used ÷ total capacity = 1902÷2700 = 0.70 or 70% of capacity of each disk.
Also, volume data density is equal to the difference between ideal-Storage and current dfsUsedRatio.
Therefore, volume data density for disk1 is: VolumeDataDensity = idealStorage – dfs Used Ratio
= 0.70-0.50 = 0.20

A Positive value for volumeDataDensity indicates that disk is under-utilized and, a Negative value indicates that disk is over-utilized in relation to the current ideal storage target.