	BMS COLLEGE OF ENGINEERING, BANGALORE-19 (Autonomous Institute, Affiliated to VTU) **Department of Computer Science and Engineering**

1ST Internal Test		
Course Code : 20CS6PEBDA	**Course Title :** Big Data Analytics	
Semester : VI A/B/C/D	**Maximum Marks: 40**	**Date: 17/05/2022**
Faculty Handling the Course:	Dr Pallavi G B, Prof Antara D	
Instructions: No choice in Part A and Part B. *Internal choice is provided in Part C.*		

PART-A

1. Define Big Data. List out the sources from which big data gets generated.

Big Data is high-volume, high- velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.----------2 Marks

-Multitude of sources

- XLS, DOC, Youtube videos, Chat conversations, customer feedback CCTV coverage

1.Typical Internal Data Sources- Data present within an organization firewall

- Data Storage- File systems, SQL, NoSQL..
- Archives- Achieves of scanned documents , paper archives, customer correspondence records, patient's health records, Student admission records and so on.

2. **External Data Sources**- Data residing an organization firewall

- Public Web: Wikepedia, weather, regulatory, census

3. **Both internal and external data sources**

- ***Sensor data:*** Car sensors, smart electric meters, office buildings, air conditioning units, refrigerators, and so on.
- ***Machine log data:*** Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
- ***Social media:*** Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
- ***Business apps:*** ERP, CRM, HR, Google Docs, and so on.
- ***Media:*** Audio, Video, Image, Podcast, etc.
- ***Docs:*** Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

---------3 Marks

PART-B

2 a Compare and contrast various properties supported by SQL, NoSQL and New SQL

	SQL	NoSQL	NewSQL
Adherence to ACID properties	Yes	No	Yes
OLTP/OLAP	Yes	No	Yes
Schema rigidity Adherence to data model	Yes Adherence to relational model	No	Maybe
Data Format Flexibility	No	Yes	Maybe
Scalability	Scale up Vertical Scaling	Scale out Horizontal Scaling	Scale out
Distributed Computing	Yes	Yes	Yes
Community Support	Huge	Growing	Slowly growing

----------5 Marks

2.bYou are at the University library. You see a few students browsing through the library catalog on a kiosk. You observe the librarians busy at work issuing and returning books. You see few students fill up the feedback form on the services offered by the library. Few Students are learning using the e-learning content. Analyze the above scenario and list the different types of data that are being generated. Support your answer with logic.
Structured Data- Library Catalog, Issue and returning books
Semistructured- Issue and returning books
Unstructured- Feedback form,e-learning content----------5 Marks

2.c Create a column family BANK_TRANSACTION with the following fields :
Account_Number int PRIMARY KEY,
Account_Type text,
OTP text.
Demonstrate the usage of map which displays a series of account number and transaction_time by altering the table to add appropriate attribute and inserting the required values.

CREATE TABLE BANK_TRANSACTION (Account_Number int PRIMARY KEY,
Account_Type text , OTP text);----1 Mark

Alter table BANK_TRANSACTION add todo map<timestamp, text>---1 Mark

ii) INSERT INTO BANK_TRANSACTION (Account_Number, Account_Type,
OTP,todo)values(121,'Current' ,12abc, 121:12-2-2020);
----------2 Marks
iii) Select * from BANK_TRANSACTION where Account_Type=121;
----------1 Marks

PART-C

3a Using CQL write queries for the following:

i) Create a Keyspace Hospital and Create the Column Family Doctor (ID, Name, Reg_no, Salary, Department, Designation, Specializations, VisitingHospitals) assuming appropriate data type.
ii) Insert required row to the Column Family.
iii) Display Name and Department whose designation is "Senior Surgeon" and salary is greater than 1,00,000 in decreasing order.
iv) Create a table to add patient_name and disease .Insert values which will be valid for 30 days.
v) Import an existing csv file into the current column family.

i) Create Keyspace Hospitak WITH REPLICATION ={‘class’: ‘Simple Strategy’, Replication_Factor=1};

Create table FamilyDoctor(ID text,Name text , Reg_no int, salary int, Department text, Designation text, Specialization List<text>, VisitingHospitals set(text), primary key (Reg_no, Name);

ii) Insert into (ID , Name, Regno Salary, Department, Designation, Specializations, VisitingHospitals)values(111,’Raju’,20ISCD, 85000, ‘Oncology’, ‘Senior Doctor’, [MBBS,MS],{‘KIMS’, ‘RKR’});

iii) Select name, department from FamilyDoctor where Designation=’ Senior Doctor’’ and salary>1,00,000 order by Name Desc;
Note: Create index on Designation and Salary

iv) Create table Patient(Patient_name text, Disease text);

v) Insert into Patient(Patient_name, Disease)values(‘Rahul’,’Fever’)USING TTL 2592000;

vi) Copy FamDoc(ID, Name, Regno Salary, Department, Designation, Specializations, VisitingHospitals) from d:/Doc.csv;----------5*2= 10 Marks

(OR)

3.b Using CQL write queries for the following:

i) Create a Keyspace GovtService and Create a Column Family Police (ID, Name, Salary, Department, Designation, types-of-cases-handled, Places-posted) assuming appropriate data type.

ii) Insert required row to the Column Family.

iii) Display Name and Department whose designation is “Deputy Commissioner” and salary is greater than 1,00,000 .

iv) Create a table to add Police_Name, Police_Station and Counter. Display police men names who have worked in a police station twice. Insert appropriate values required in the table to display the result.

v) Export this file on to the desktop.
Similar queries as above----------5*2=10 Marks

4.a Demonstrate various features of Cassandra in detail.

Features:

1. Peer to Peer Network
   -Designed to distribute and manage large data loads across multiple nodes in a clustering constituted of commodity hardware
   -Does NOT have a master slave architecture-NOT have a single point of failure
   -Graceful degradation where everything does not come crashing at any instant owing a node failure
   Ensures data is distributed across all nodes in the cluster
   Each node exchange information across the cluster every second
2. Gossip and Failure Detection
   -Gossip protocol is used for intra ring communication
   -Peer to peer communication protocol which eases the discovery and sharing of location and state information with other nodes in the cluster
   -A node only has to send out the communication to a subset of other nodes
3. Partitioner
   -A partioner takes a call on how to distribute data on the various nodes in the cluster
   -It also determines the node on which to place the very first copy of the data

-A partitioner is a hash function is used to compute the token of the partion key

4. Replication factor

- As hardware problem can occur or link can be down at any time during data process, a solution is required to provide a backup when the problem has occurred.
-So data is replicated for assuring no single point of failure.
-Cassandra places replicas of data on different nodes based on these two factors.

1. Where to place next replica is determined by the Replication Strategy.
2.While the total number of replicas placed on different nodes is determined by the Replication Factor

There are two kinds of replication strategies in Cassandra

1. SimpleStrategy
   - SimpleStrategy is used when you have just one data center.
2. NetworkTopologyStrategy
   - NetworkTopologyStrategy is used when you have more than two data centers.

5. Anti Entropy and Read Repair

A client can connect to any node in the cluster to red data
How many nodes will be read before responding to the client is based on the Consistency level specified by the client
If few of the nodes respond with an out of date value Cassandra will initiate a read repair to bring the replicas with stale values up to date
This is done using Anti Entropy gossip protocol.

6. Write operation in Cassandra

When write request comes to the node, first of all, it logs in the commit log. Commit log is used for crash recovery.
After data written in Commit log, data is written in Mem-table
Data written in the mem-table on each write request also writes in commit log separately.
Mem-table is a temporarily stored data in the memory while Commit

7. Hinted Handoffs

Three nodes A,B, C. C is down.Replication factor is 2.Write operation on node A which is the coordinator and serves as a proxy.When row k is wriiten by the client to Node A, it will write row K to node B and stores a hint for Node C.
Hint has the following information
Location of the node on which replica is to be placed
Version Mtada
Actual data

8. Tunable Consistency

→ Strong Consistency- Each update propagates to all location where that piece of data resides
→ Eventual Consistency- Client is acknowledged with success as soon as part of the cluster acknowledges the write----------10 Marks

(OR)

4.b Distributed data bases relaxes ACID properties. Analyze the properties supported by distributed systems with a scenario. Reflect databases that follow one of the three possible combinations with a neat diagram

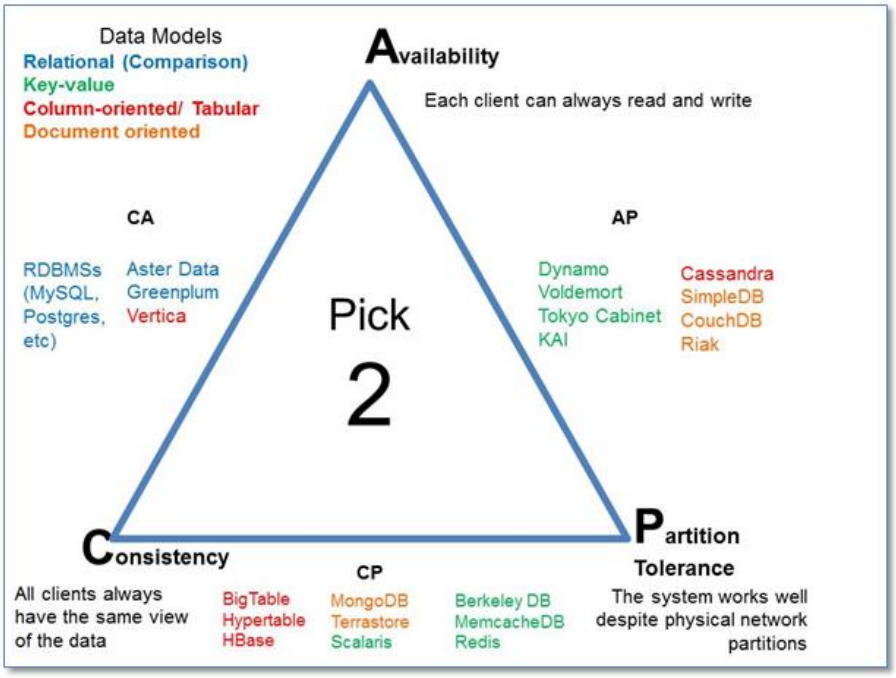
The CAP Theorem(Brewer's CAP) states that, in a **distributed computing environment** (a collection of interconnected nodes that share data.), **it is possible to provide the following guarentees**– At best you can have two of the following three. : Consistency, Availability, and Partition Tolerance

Consistency - A read is guaranteed to return the most recent write for a given client.
Availability - A non-failing node will return a reasonable response within a reasonable amount of

time (no error or timeout).

Partition Tolerance - The system will continue to function when network partitions occur.

----------7 Marks



----------3 Marks