Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2023-2024

Mid-Semester Test
(EC-2 Regular – ANSWER KEY)

Course No.	: DSECLZG529/AIMLCZG529
Course Title	: Data Management for Machine Learning
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: ~~06/03/2021 or 19/03/2021 (FN/AN)~~

No. of Pages = 6
No. of Questions = 3

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. Assume that you are working on a Real Estate app that allows the users to Search New Projects, Homes, Apartments, Offices, Shops, and Showrooms for Buy, Sell, and Rent. This property app is required for fulfilling all the real estate needs for searching, shortlisting and finalizing property of their choice. Be it a ready-to-move flat in a new project, investment in shops, offices or showrooms, or girls PG or boys PG in city, app should have listings from qualified property agents & top builders as well as no brokerage properties. This house search app should make property search on mobile simpler, faster and smoother. Search for residential and commercial property listings from owners, brokers and top builders on this user-friendly app should help millions of users to start their real estate journey. To make it more interesting and meaningful, when the users are online searching for a property on the app, the recommendations about the relevant properties should be shown to her. Also relevant advertisements has to be shown on the app screen to increase the chances of converting these paid advertisements into increasing the foot falling and result into actual selling. **[3 + 2 + 1 + 1.5 + 2 + 2.5 + 1 + 2 = 15]**

a) Who are the primary users of this systems and what type of user interfaces are suitable for them to interact with this application?

Answer:

- Property Owners – Web / Mobile interfaces
- Property Searchers – Web / Mobile interfaces
- Real-Estate Agents – Web / Mobile interfaces / APIs
- Portal Executives – Web / Mobile interfaces / APIs

b) Discuss the important data modalities of data that are suitable for such applications?

Answer:

- Structured data – user profiles
- Semi-structured data – property details

crackbitswilp.in

Assume a builder who is promoting his real-estate project on this application needs a periodic report on the searches happened for the project, type of project information accessed, brochure downloads, queries raised against property etc.

c) What type of data abstraction will be required to generate such a report? Why?

Answer:

- Data Warehouse / Data Mart – 0.5M
- as most of the data is structured data which can be stored in RDBMS backed DW / DM – 0.5M

d) What type of data processing is suitable for the above requirement? Why?

Answer:

- Batch processing - 0.5M
- as it needs to be generated periodically – 1M

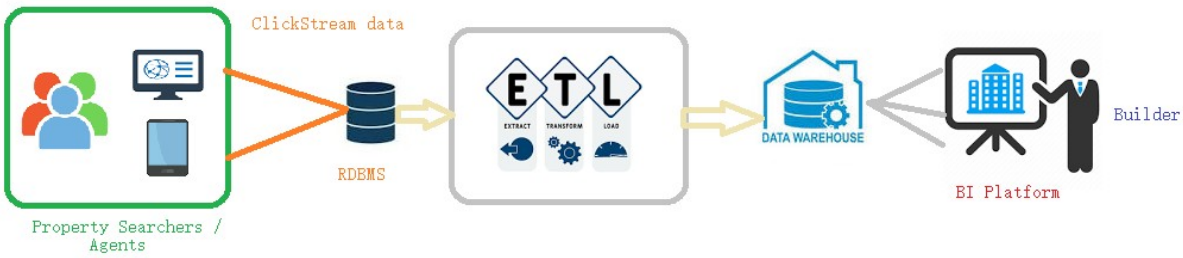
e) Identify any two approaches through which this prepared report can be served to the builder or his representative?

Answer:

- Through web browser / BI application interface
- Report that can be downloaded in CSV, Spreadsheet or pdf format
- Report link that can be shared over email or other communication channels

f) Discuss briefly the data flow (and the stages involved) that will be required to meet this requirement along with a pictorial representation.

Answer:



Components

- Users
- OLTP System – RDBMS
- ETL pipeline
- Data Warehouse
- BI Platform

crackbitswilp.in

Each component – 0.5M

Imagine that now you have to implement a feature where immediate property recommendations needs to be shown to the users when they are looking for a property.

g) What is the significant change will be required in dataflow mentioned in (f) to match the new requirement?

Answer:

- Streaming storage and streaming processors needs to be implemented in the dataflow.

h) Which data format will be suitable to exchange the data between users of the application and recommendation engine? Provide a snippet of the data item exchanged.

Answer:

- JSON will be a suitable data format for data exchange. 1M
- Input – JSON structure having user applied filters and some user profile data - 0.5M
- Output – JSON list of properties matching to the property user is looking upon - 0.5M

Q2. Predictive Automotive Components Services (PACS) Company renders customer services for maintenance and servicing of (Internet) connected cars and its components. Assume that number of centers are 8192 ($=2^{13}$), number of car serviced by each center per day equals 32 ($=2^5$). Each car has 256 ($=2^8$) components, which requires maintenance or servicing in the Company's car. The service center also collects feedback after every service and send responses to customer requests. The feedback and responses text takes on average 128 B ($=2^7$ B) and each service or responses records in a report of average 512 B ($= 2^9$) text. Company saves the centers data for maximum 10 years and follows last-in first-out data replacement policy. **[1 + 1.5 + 1.5 + 1 + 2 = 7]**

a) Justify why PACS is example of big data use case.
b) What are the big data system characteristics that you think will be important in this service?
c) How will the files of PACS be saved using big data file system like GFS or HDFS?
d) If the data block size is 64MB, how many records will be stored per block?
e) What shall be the minimum memory requirement (in MB) for 10 years?

Answer:

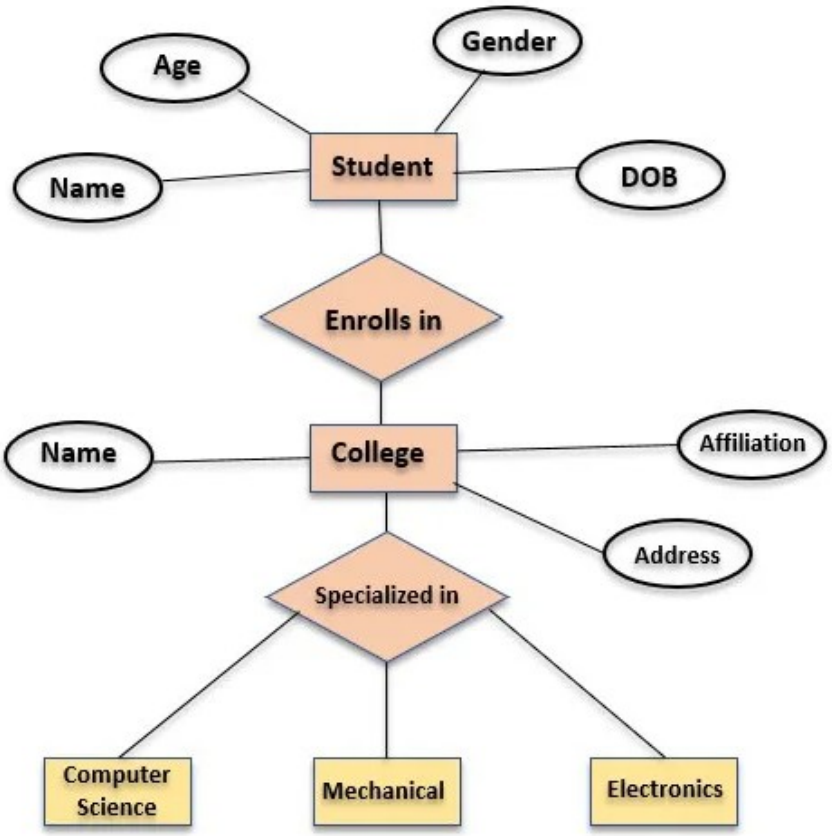
a) Its big data use case as it caters to big data's "Volume" property.
b) Big data system characteristics: scalability, reliability and maintainability
c) The HDFS will use the following mechanism to store the file
   - NamedNode will identify the DataNodes on which data needs to be stored
   - Data is stored in data blocks on the DataNodes
   - HDFS replicates those data blocks and distributes them so they are replicated within multiple nodes across the cluster.
d) Data files size per car = 512B (Service Report) and 128B (feedback) = 640B (512B+128B)
   Therefore, each data block may store 64MB/640B that is ~ 100K records.

crackbitswilp.in

e) The minimum memory requirement = [No. of centers * No. of cars serviced per day per center * No. of days in a year * 10 years * (Size of feedback + size of Service Report)] B
= [$2^{13} * 2^{5} * 365 * 10 * 640$] B
= 584000 MB

Q3. Consider the following Entity-Relationship Diagram representing the relationship between students and college. Represent this ERD in the following data format using sample data. **[3 + 2 + 1 + 2 = 8]**

I. Row-major format
II. Columnar format
III. CSV
IV. JSON



Answer:

I. Row-major format – RDBMS
RDBMS relation – three tables – student, college and Specialization

Student

StudentID	Name	Age	DOB	CollegeID

crackbitswilp.in

1	A	21	1-jan-2001	1
2	B	23	1-march-2001	2

College

CollegeID	Name	Affiliation	Address	SpecializationID
1	SBC	Pune	Pune	1
2	BITS	Pialni	Pilani	2

Specialization

SpecializationID	Name
1	CS
2	Mech

II. Columnar format – NoSQL stores like Cassandra – will use flat structure – store all column values next two each other

Student Name	Age	DOB	CollegeName	Affiliation	Address	Specialization Name
A	21	1-jan-2001	SBC	Pune	Pune	CS
B	23	1-march-2001	BITS	Pialni	Pilani	Mech

{

StudentName: {A, B},
Age:{ 21, 23},
DOB:{ 1-jan-2001, 1-march-2001 },
COllegeName:{ SBC , BITS },
Affiliation:{ Pune , Pialni },
Address:{ Pune , Pilani },

crackbitswilp.in

Specialization Name{ CS , Mech }

}

III. CSV - – can embed all the data in a single file

StudentName, Age, DOB, College Name, Affiliation, Address, Specialization

A, 21, 1-jan-2001, SBC, Pune, Pune, CS

B, 23, 1-march-2001, BITS, Pialni, Pilani, Mech

IV. JSON

```
{
    {
        Name: A,
        Age: 21,
        DOB: 1-Jan-2001,
        College :{
            Name: S B C,
            Affiliation: Pune ,
            Address: Pune ,
            Specialization: CS
        }
    },
    {
        Name: B,
        Age: 23,
        DOB: 1-march-2001,
        College :{
            Name: S B College,
            Affiliation: BITS,
            Address: Pilani,
            Specialization: Mech
        }
    }
}
```

****

crackbitswilp.in