Q1

Saty: answers for a and b are the following instead, please dont use the above :)

a. The expressions can be named, and indexes created on those named columns (aka computed columns) - eg. we can create a column called PossiblePriceIncrease, CurvedScore etc, and index them, then we can specify such indexes, everytime we use those expression columns, ie PossiblePriceIncrease > 50
b. For a bitmap index, **the number of rows in the index will be identical to the number of rows** in the original/raw/unindexed table! In a 'regular' indexed column (eg GPA), this won't be the case.
a. Emphasis on the use of hash indexes. Since the column contains long unique strings, strings can be hashed to a unique key. The hash value can be computed using a hashing algorithm. For example, a user query (eg. LNAME="Johnson") is converted to a hash 'key' which is then used to search through the pre-computed list of hash values and retrieve an exact match or a small set of values that are all stored with the same key (as a result of 'hash collision').
    i. 1 point for emphasising on the use of hash indexes

Q2 (5 points):

XML:

```
<Name of the node = A>
        <Connected Node = B>
        <Connected Node = C>
</Node>
<Name of the node = B>
        <Connected Node = A>
</Node>
<Name of the node = C>
        <Connected Node = A>
        <Connected Node = D>
        <Connected Node = E>
</Node>
<Name of the node = D>
        <Connected Node = C>
</Node>
<Name of the node = E>
        <Connected Node = C>
</Node>
```

Saty: nice. There might be other ways to specify the graph, too.

JSON:

```
{
```

```
"graph":{
  "A":{
    "Connected Nodes":{
      "B":{

      },
      "C":{

      }
    }
  },
  "B":{
    "Connected Nodes":{
      "A":{

      }
    }
  },
  "C":{
    "Connected Nodes":{
      "A":{

      },
      "D":{

      },
      "”E”":{

      }
    }
  },
  "D":{
    "Connected Nodes":{
      "C":{

      }
    },
    "E":{
      "Connected Nodes":{
        "C":{

        }
      }
    }
```

```
    }
  }
}
```

Saty: Nice! Again, there might be other ways..

Plaintext:
Graph G
Node A Edges B,C
Node B Edges A
Node C Edges A,D,E
Node D Edges C
Node E Edges C

Saty: excellent! Once again, there are many other ways :) [so please accept any valid answer]

Rubrics:

- 2 Points for Each XML and JSON Format
- 1 Point for PlainText Format
- Students Can name the tags differently.

Q3

Pick one of these four domains: an airport, an airline, a farm, a retail store (eg. 7-11). For your choice, list 5 pieces of data analysis (mining) you'd perform on each. Eg. if 'USC' was a choice, one of the five answers would be 'outgoing GPAs vs incoming (high school) GPAs, fitted using linear regression'. For each, be sure to clearly list the data (eg GPA) as well as the analysis (eg linear regression fitting).

Student should choose any one of these four domains, for each one, list 5 data analysis, each has 1 point.
E.g.
If choose **airport**, a sample analysis could be:
Data (0.5 points): “ticket **prices** vs flight **attendant rate**”, Analysis (0.5 points): “Nonlinear Regression”
If the data is not relevant to airport, the first 0.5 points will be deducted.
If the analysis doesn’t make sense, the second 0.5 points will be deducted.

Saty: yes, good rubric. For each answer, make sure that the data, and the analysis (eg mining algorithm name) are mentioned.

Question 4

Part A (1 point): Can mention any powerful algorithms. A few examples are:

- AdaBoost
- XGBoost
- Expectation Maximization
- Neural Networks

Saty: to me, the answer would be NN, but, others above might be ok, too.

Part B (2 points): Should give at least one reason for the algorithm picked in part a.

Example: AdaBoost Algorithm

AdaBoost is an iterative ensemble method. AdaBoost builds a strong classifier by combining multiple poorly performing classifiers. The combined classifier increases the accuracy significantly making it a powerful algorithm.

Saty: for NN, the reason is that we can always create enough (deep) layers and neurons in each layer such that any dataset can be learned (the architecture is flexible, in that sense).

Part C (1 point): Can mention any simple algorithms. A few examples are:

- Linear Regression [no]
- Decision Trees [no]
- k-nearest neighbors
- Naive Bayes [no]
- k-means clustering
- hierarchical clustering

Saty: I'd take 'clustering' to be the simplest (to understand, implement…)., or possibly, kNN. For the 'no' ones, you could give partial credit.

Part D (1 point): Should give at least one reason.

A few examples are:

- It reduces cost.
- It could accelerate calculations.
- Also, Give points for any correct or reasonable explanation.
- And, many other similar points can be accepted.

- Saty: another reason: connectivity to the cloud - if this is lost, edge can permit the calcs to continue (eg for an SDC, medical diagnosis in a remote location etc)

Q5) There are clear/unmistakable/irreversible 'trends' in the way data is stored, analyzed, interpreted etc. Name 5 of them (trends) we looked at, say a line or two about each. In other words, where is all this going, compared to what used to be, and, what is practiced now?

1. More languages are available e.g. Julia, Wolfram
2. Edge processing - Edge processing refers to the execution of aggregation, data manipulation, bandwidth reduction and other logic directly on an IoT sensor or device. The idea is to put basic computation as close as possible to the physical system, making the IoT device as “smart” as possible.
3. Rise of DIY ML - https://dlabs.ai/blog/machine-learning-off-the-shelf-models-or-custom-build-pros-and-cons/
4. GPU/TPU development - deploy thousands of ALUs in a single processor, allowing you to perform thousands of parallel computations simultaneously, making GPUs capable of extensive machine learning training and inferencing.
5. Cloud computing and storage - AWS, GCP, Azure, IBM Cloud etc
6. Rise of DataOps - DataOps is a methodology and practice that borrows from the DevOps framework often deployed in software development. While those in DevOps roles manage ongoing technology processes around service delivery, DataOps is concerned with the end-to-end flow of data through an organization.
7. Data provenance techniques - With the rise of AI and deep fakes in advertising, the quality and reliability of data are now being called into question more than ever. When analyzing data for marketing or financial purposes, one of the biggest initial hurdles is deciding whether the data can be trusted.
8. Saty: Dataflow-based software, eg KNIME, RapidMiner - a ‘no code’ option, useful for business analysts and casual users

+1 for any answer that makes sense

Q6) (5 points)
Atleast 5 valid statements/assumptions (1 point each)
Examples:

- Extensive user interaction in higher dimensions (3D graphs - add, remove data, turn the graph, change the axis)
- Visualize higher dimensions (X-Y-Z axis) which earlier needed to be reduced in order to visualize
- Make graphs, charts, data points easier to understand by everyone, not only data scientists, allowing more collaboration

- The collected data can be used for new forms of human-computer interaction by giving people feedback such as tactile information, which can be different from traditional 2d data visualization.
- Engineers could specify the direction, speed and quality for collected data points, assigning meanings to the data, making the static visualization become vivid for users.

-1 for any point that is incorrect (eg. explaining regular data visualization, not relating it to metaverse)

Saty: look for the mentioning and discussion of AR and VR, for presenting (eg sensor data in 3D), and analyzing, visualizing (via natural gestures (using our hands) and voice), etc. There needs to be 3D CG involved, in the answer. The points above are good, but we don't 5 of them with a point for each :)

7)

Saty: please ignore the above, consider the following instead!!

a. KML, tables in Postgres, shapefiles (.shp), JSON
b. The plaintext format could be sets (blocks) of these:

Lat
Long
Label
Popularity
Lat
Long
…

Q8

A modern alternative is a "data lake".(+3 'data lake')
It offers a more continuous form of analytics, driven by the rise of unstructured data, streaming, cloud storage, ets. In a data lake, data is stored in its raw form . (+2 for reasonable explanation)
Saty: yes - last slide in the BI lecture!

Q9

A:

Saty: please ignore the above, use this instead. Dataflow is powerful because of the facility for **selective** node execution! In other words, when we change a node parameters and re-execute, **the entire graph does not get re-executed, only that node and its 'downstream' (affected) ones do. That is what offers enormous savings in computation.**

B:
1 point for each data flow based tool. Examples: (1) RapidMiner, (2) KMine, (3) TIBCO Data Science (Statistica), (4) Lobe, (5) Synapse ML (6) SmartPredict.AI (7) Bonsai
0 points for tools that are not related to building data flows.
Saty: nice list!

Q10 (5 points)

Any point discussing these is fine:
Data misuse is the use of information in ways it wasn't intended for.
Data abuse happens as a result of a cyberattack or when data is collected without the owner's consent. Saty: also, abuse would be when our face/body images/videos, our voice, are used for deepfakes.

Possible answers:
Undermining users' privacy by selling users' personal information to 3rd parties or sharing/selling personal information to advertisers
Governments can monitor people by collecting large amounts of data
Improper anonymizing might lead to leaking sensitive information when combined with other data
Inaccurate algorithms can result in a company bringing in data it never meant to gather, endangering customers and leaving businesses outside of compliance regulations.
Employees might copy confidential work files or data over to their personal devices, so they make that information accessible outside of its intended, secure environment.
Leaked or stolen credentials can be employed in widespread data breaches
…

Rubrics:
- 1 point for each distinct abuse/misuse
- 0 point for an abuse/misuse that is from the same aspect as another one
- 0 point if the abuse/misuse is inaccurate or wrong

Saty: nice rubrics and possible answers :) Mention of almost ANY 'bad' use of data is ok :)