Andrew Mackay

ID:3441988

COMP 378

Intro to database assignment 3

1. A) The integrity controls performed in both data integrity and referential integrity are first and foremost that data integrity is referring to the underlying structure of the fields and controls. This means that it is somewhat of the overarching concept which then allows referential integrity to exist somewhat, although they are two totally different things simultaneously.

So things like the default value, null value control, and range control all have to do with data integrity where default value is a general rule that means to reduce errors a default value must be assigned to fields, then in null value control in order to enforce integrity,

C)

Vertical partitioning means some columns are moved to different tables during a partition where each tables columns decrease in the move but rows remain the same.

Horizontal partitioning divides tables into smaller tables , which contain the same number of columns but fewer rows. (bytebytego, 2022)

Benefits of this approach include security, as data can be moved if ways that improves chances of 3rd party access. Recovery time and load balancing as the weight on the database will be reduced in areas of partitioning therefor increasing recovery.

Some disadvantages of partitioning are that partitions may have different access speeds. Data duplication and excessive storage use. Paritioning violates the 3rd normal form and therefor creates a unnecessary complexity and reduces ease of access.

1. A)

When choosing file organization one should consider how much storage space is needed, access of data and manipulation of records should be simple while also not needing to reorganize. In adverse scenarios data corruption must be accounted for and handled appropriately also. The views of which stakeholder might need to be created and the amount of views needed may need to be considered when designing a database at large, and its organization to make said views. To display as many things in as logically structured a way as possible and to define primary and foreign keys when creating said tables so that inserting values from then on is much easier.

B)

Clustering provides improved data management for larger systems, and especially systems that are projected to be growing a sizeable amount. It is an efficient method for handling data and it is achieved by leveraging multiple physical storage servers which mount the file system which makes it accessible and then makes it one singular system. Data storage and data access can be organized across all clusters, which enables them to work together managing the total workload and enables users to access the same files and data at the same time. This improvement of resource availability comes down to if multiple nodes fail it will not have a detrimental affect to the totality of the system like in other storage methods, so the users of this system will always have access to the data.(Weka, 2022)

C)

Hashed file and index file organization are in a way 2 sides of the same coin. They have different methods of organization which provide unique benefits over one another in certain situations.

Hashed file organization, otherwise known as direct or random file organization is a method for storing records a hash function has calculated, which then gives an address to save the record at. Records are stored randomly, after the application of the hash function, which is applied to attributes and columns where if there is a key associated with the column applied to then it is called as a hash key otherwise being a hash column for non key columns.

3.

A)

SELECT Employee.emp\_no, Employee.name, ProjAssigned.proj\_no, ProjAssigned.worked\_hours

FROM Employee

Inner JOIN ProjAssigned ON Employee.emp\_no = projAssigned.emp\_no,

WHERE IN Employee( salary > 66000), ORDER by Employee.name;

4.A)

A powerful firewall, antivirus system with on hand tech support constantly updating and monitoring the entirety of the system and backing it up frequently, along with encryption of user data.

B)

Advanced multifactor authentication could initiallyy provide accessibility to security cleared authorized employees.

C)

A firewall in combination with 2 factor authentication given to golf club with enforced security and extra granted permissions to those in administrative roles, clearly defining the paramaters of each role within the club more clearly.

D)

Again an advanced firewall with encrypted data, and multifactor authentication would likely be enough to keep intruders away in this situation but for more additional security having an IT team constantly monitor and update these security measures.

5.

a) This being type of Database destruction. Switching to a duplicate database in this situation where likely a mass amount of incorrect data entries were made, would be preferable. However this could be also solved by roll forward technique or reprocessing the transactions.

b) Incorrect data entry again may have been made of a large scale in which case reprocessing transactions may be the least appealing technique for the administrator. If payment has already been processed for students, a transaction may be offset and re payed with the payee notified by their email on file or phone number. If the error is found before this however the backward recovery technique would be more useful.

c) Restarting the backup from a earlier checkpoint with activation of the journalizing facility then resuming the data backup.

6.

1. Marts and Warehouses yield similarities like that they are both read only and structured repositories of transnational data. The kind of data that is stored is where they are distinct. Data Marts have are considered better suited for analyzing data sets under 100GB that are structured and of a specific subject, while Data Warehouses are used by data scientists and engineers for storing over 100GB data sets of a greater complexity which may encompass data mining, machine learning and AI for business insights. A data Mart also has fewer source systems that are operationally focused, whereas data warehouses have a wider variety from all across the enterprise. (qlik, 2022)

B.

Data federation provides a singular view of the entirety of integrated data in a database, without bringing data into a centralized physical database. The federation engine fetches specified data from sources, then send passes the result along to a requesting application. The data in which is accessible within the data federation type of integration is current data with no delay time. A downside is with bigger hauls of data that they are not as effective and also for more often used integration.

Data Propagation is the replication of databases, with near real time delay. Data is pushed in event driven propagation, and updates are enabled to be either synchronous or asynchronous.This technique provides real time cascade of data changes throughout an organization, although higher specielized technologies are needed to execute these updates.

C.

Hadoop plays in important role in analyzing and managing big data, and scheduling tasks that operate on mass records. Hadoop is an open source framework of MapReduce, which is an algorithm in which enables mass parallel automated processing of various tasks. Hadoop enables MapReduce to apply this paralellization to real world problems, which consists of various components to help carry out this task.

The Hadoop distributed file system(HDFS) manages a mass amount of large files within a distributed environment.

Pig helps integrate a scripting language and execution environment which simplifies the tasks of MapReduce.

Hive derives from Apache, and supports querying and management using something called HiveQl which is like SQL but specifically for managing stored data in Hadoop.

HBase is the final component to Hadoop which is a database modeled after Googles BigTable in which runs on top of HDFS and serves as a data source in MapReduce tasks.

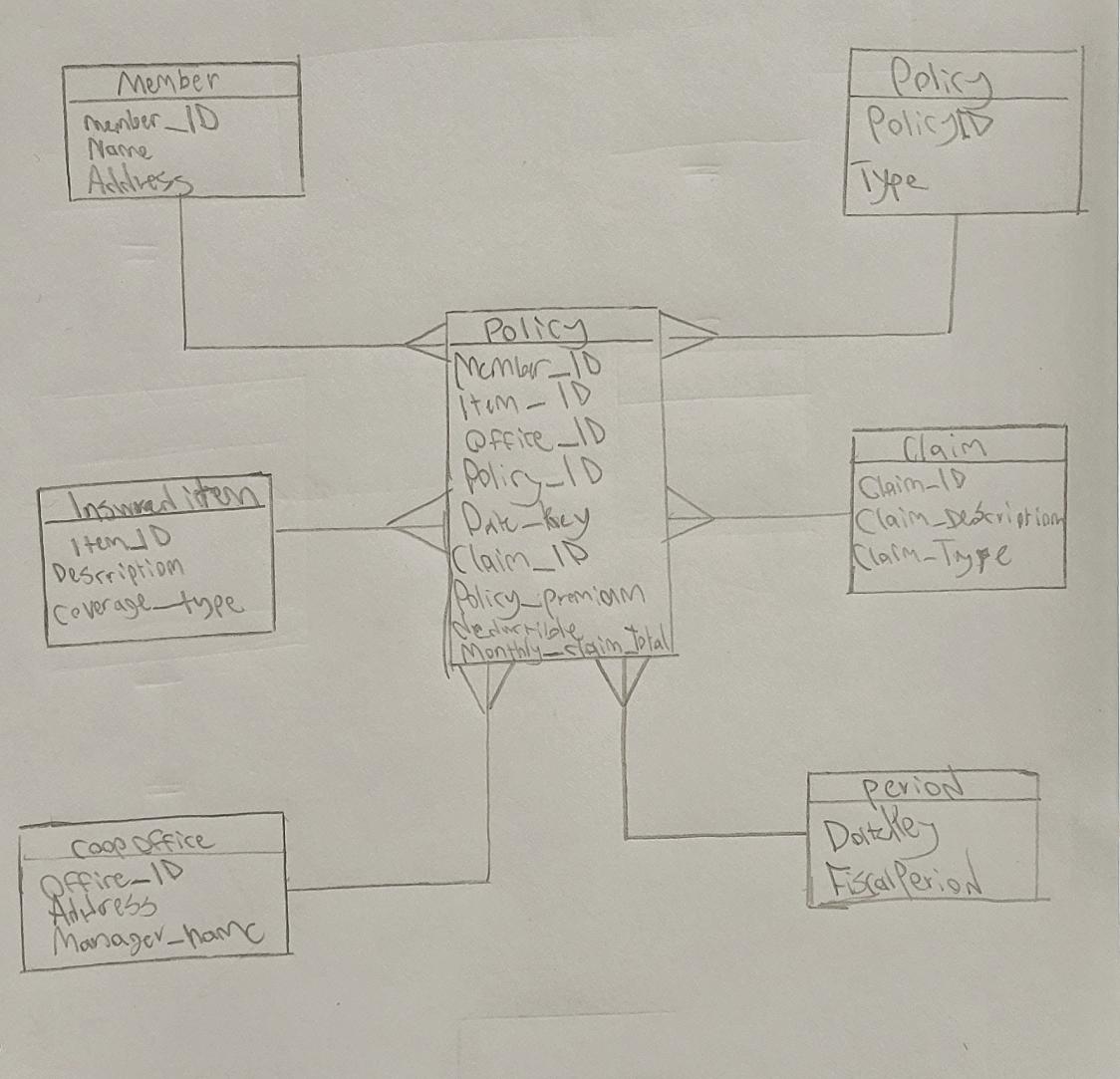
D.

The descriptive analytics seek to answer what has happened? This is the groundwork before other analytical approaches are considered. It is to look at past data related events and identify patterns of occurrences within the data.

Predictive uses descriptive and diagnostic techniques to first establish what exactly occurred and why. As the name suggests predictive analytics seek to predict what will happen next. This involves techniques such as forecasting, regression analysis, multivariate statistics, predictive modeling and pattern matching. These techniques require high quality data in large amounts and understanding these techniques requires an in depth knowledge of programming language Python and R.

Prescriptive analysis seeks to answer what is the remedy? How can this be fixed? This is the final technique within the ladder of analytics approaches. Types of techniques which are used within this scope are simulation, graph analysis, complex event processing, recommendation engines, neural networks, machine learning and heuristics. In order to reach this step all other steps need to succesfully be completed to move on in the hierarchy, therefor automatically making it the most difficult to achieve, although it is difficult for other reasons as well. The quality of data needed for this step is of the highest and how well the other steps in the hierarchy are performed determines this.(Lupoka, 2022)

1. A)



Star schema is assembled in star shape with fact over event although fact is policy event is to the side as to resemble star shape. There is a fairly fine grain to the model, dimensional data is split from fact data as displayed.

B)

(Percent of policies experience = 0.05 \* number of policy = 1,000,000

\* number of coverage item = 10 \* number of insured party = 2 \* number of years of historical data = 5 \* number of month = 12)

Total number of rows = 60,00,000 rows

C)

If 5 bytes for field and 60,000,000 rows each containing 9 fields

Formula is (Number of rows\*Number of fields per row\*Number of bytes per field)

60,000,000\*9\*5 = 2,700,000,000 bytes estimate

8.

1. Column store: Database supports a complex and varied structure for each record
2. Graph Base: Best suited for organizing as a network of association between the data items

C)Document store: Is able to provide fast access to individual records using a key value, while also being accessible.

9.



Descriptive analytics/predictive:

In order to scale the hierarchy of analytics we must first begin using descriptitve analytics which then technically would need to build to the top to use predictive but descriptive being especially helpful for viewing customers history, which is a large part of what is needed for a prediction of future best practises.

B)

Diagnostic analytics:

Viewing the root of a problemand determining what exactly is wrong and what caused it is known as a diagnoses, therefor diagnostic analysis.

C)

Predictive analytics:

Yet again the prediction aspect of the analytics hierarchy is at the top so all other steps are important to refining data but the most suiting is predictive as it is focusing on the future in this case.

D)

Descriptive analysis:

In Ubisofts case descrpitive analysis would likley be enough to show the metrics related to their revenue with numbers of downloads falling into this category.

References

Hoffer, J., Venkataraman, R., & Topi, H. (2018). Modern Database Management (13th ed.). Pearson Education (US). https://online.vitalsource.com/books/9780134792293

What is Clustered File System | Benefits and more | WekaIO. (2022). Retrieved 20th November 2022, from https://www.weka.io/learn/clustered-file-system#:~:text=In%20fact%2C%20clustering%20provides%20better,of%20servicing%20multiple%20application%20clients.

Xu, A. (2022). Vertical partitioning vs horizontal partitioning. Retrieved 20th November 2022, from https://blog.bytebytego.com/p/vertical-partitioning-vs-horizontal

What is a data warehouse? definition, concepts, types - qlik (no date) qlik.com. Available at: https://www.qlik.com/us/data-warehouse (Accessed: November 30, 2022).

LUPOKA, R. (2022). Where to Start With the 4 Types of Analytics. Retrieved 5th December 2022, from https://iterationinsights.com/article/where-to-start-with-the-4-types-of-analytics/