**Database Exam Answers**

**Question 1 ( A ) Primary Keys**

|  |  |
| --- | --- |
|  | [**Primary keys should be immutable**](http://programmers.stackexchange.com/questions/8320/should-a-primary-key-be-immutable) (or [as stable as possible](http://stackoverflow.com/questions/3632726/what-are-the-design-criteria-for-primary-keys) since **immutability can not be enforced in the DB**).  While there is nothing that will prevent you to update a primary key (except integrity constraint), it may not be a good idea:  From a performance point of view: **IOUS**   * **Indexed foreign keys -** the **update will lead to the update of the index** (delete+insert in the index structure), generally more expensive than the actual update of the base table. * **ORGANIZATION INDEX tables** - rows **physically sorted by the primary key**. update results in physical D+I. * **Unindexed foreign keys** – require **lock** to ensure integrity. * **Single** **updated key** –potentially lots of tables/rows **updates**.   **Other considerations:**   * If this key **is referenced in any external system** (application cache, another DB, export...), the reference will be broken upon update. * Additionally, **some RDBMS don't support CASCADE UPDATE**, [in particular Oracle](http://asktom.oracle.com/pls/asktom/f?p=100:11:0%3a%3a%3a%3aP11_QUESTION_ID:5773459616034). |

**Question 1 ( B ) Part I – SQL Statement and DBMS commands**

**Data Definition Language** (DDL) statements are used to define the database structure or schema.

* **CREATE** - to create objects in the database

**Data Manipulation Language** (DML) statements are used for managing data within schema objects.

* **SELECT** - retrieve data from the a database
* **INSERT** - insert data into a table
* **UPDATE** - updates existing data within a table
* **DELETE** - deletes all records from a table

**Data Control Language (DCL) statements.**

* **GRANT** - gives user's access privileges to database
* **REVOKE** - withdraw access privileges given with the GRANT command

**Transaction Control** (TCL) statements are used to manage the changes made by DML statements. It allows statements to be grouped together into logical transactions.

* **ROLLBACK** - restore database to original since the last COMMIT
* **SET TRANSACTION** - Change transaction options like isolation level and what rollback segment to use

**Question 1 ( B ) Part II – Optional Commands / Statements**

**Data Control Language** and **Transaction Control**

**Question 1 ( B ) Part III – SQL Statements to import from a file**

* Select From , Insert Into, Join , Where

**Question 1 (B) Part IIII – SQL Vs Application Side**

* **Security** - Application can make use of stored procedures and prepared statements to protect the database bridge
* **Reduces Erroneous Input** - Can reduce on errors made.
* **Direct control and Manipulation** – Generate new Queries
* **User Interface** – Not as simplistic as application

**Question 2 (A) 2012 – Model Types**

**Object Oriented Model -** (**OODB**)

Provides all the facilities associated with object oriented paradigm. It enables us to create classes, organize objects, structure an inheritance hierarchy and call methods of other classes. it also provides the facilities associated with standard database systems

**Advantage**

* Designed to **store, retrieve and manage objects** created by programs written in OOP.
* Models the **real world more closely**.
* **Extensible** – Allows new data types to be built.

**Disadvantages**

* **Software is required to translate objects into tuples** of a relation . *– Middle layer software*
* **Lack of universal data model -** Uses Its **own DML**
* **Lack of experience** – focused more towards programmer
* **Lack of standards** – no standards OO query language

**MOLAP (multidimensional OLAP)**

tools utilize a **pre-calculated** data set, e.g. data cube, that contains all the possible answers to a given range of questions.

MOLAP tools are best used for users who have **"bounded" problem sets** (*they want to ask the same range questions every day/week/month on an updated cube, e.g. finance).*

**Advantages**

* **very fast response - Unnormalised**
* ability to **quickly write back data into the data set** (budgeting and forecasting are common applications).
* **Can perform complex calculations**: All calculations have been pre-generated when the cube is created

**Disadvantages**

* **Limited scalability** (the cubes get very big, very fast when you start to add dimensions and more detailed data),
* **Inability to contain detailed data** (you are forced to use summary data unless your data set is very small)
* **Load time of the cubes**.
* **Requires additional investment:** additional investments in human & capital resources needed.

**ROLAP (relational OLAP)**

tools **do not use pre-calculated data** cubes. Instead, they intercept the query and pose the question to the standard relational database and its tables in order to bring back the data required to answer the question.

ROLAP tools are best used for users who have "**unbounded" problem set** (they don't have any idea what they want to ask from day to day; e.g., marketing).

**Advantages of ROLAP tools are**

* **ability to ask any question** (you are not limited to the contents of a cube)
* ability to **drill down to the lowest level of detail** in the database.

**Primary downsides of ROLAP tools are**

* **Slow response**
* **Some limitations on scalability** (depending on the technology architecture that is utilized).

**HOLAP (hybrid OLAP)**

HOLAP addresses the shortcomings of both of these technologies by combining the capabilities of both approaches. HOLAP tools can utilize both **pre-calculated cubes and relational data sources**

**Key Terms**

* **Drill Down** – to the lowest level of details
* **Granularity** - a measure of level of details of your data in a database – e.g. Address

**Measure -** measure is a **performance indicator** that is quantifiable and used to determine how well a business is operating. useful measures - Quantity Sold or Revenue.

**Dimension** - dimension is a broad grouping of related data about a major aspect of your business. For example, you have a dimension called Products.

**Measures** - Revenue, Total Rentals

**Dimensions** – Car, Year, Rental, Accessories

**2(a) Main Alphanumeric data types in RDMS**

|  |  |  |
| --- | --- | --- |
| **Data Type** | **Abbreviation** | **Description** |
| CHARACTER(n) | CHAR | Character string, fixed length n. |
| CHARACTER VARYING(n) | VARCHAR | Variable length character string, maximum length n. |
| CHARACTER LARGE OBJECT | CLOB | Variable length character string measured in characters. |

**2(B) Data types expending database to handle other alphanumeric data.**

* **Blob** – Stores a large object as binary in the database
* **File Reference** – file pointer - record outside the database

**2(C) 3 Different Strategies employed in multimedia databases**

* **Like** – Searching via substrings on CLOBs
* **Content based image retrieval** – CBIR

**3 (a) 3 different strategies for searching**

**Blobs , Clobs, & File Reference**

* **Large Queries** - Takes long time
* **System intensive** - Nothing would be pre-calculated and Would require the RDBMS would have to be locked
* **Changing Data** – no historic data, changing answers

**3(a) Multimedia Data Types**

**Text:** **ASCII based files** / **text is typically stored in processor files**, spreadsheets, databases and annotations on more general multimedia objects.

**Images:** great variance in the quality and size of storage for still images. The space overhead for still images varies on the basis of **resolution**, **size**, **complexity**, and **compression** **scheme** used to store image.

**Audio:** Its quite space intensive. Several techniques are used to compress it in suitable format.

**Video:** space consuming. The digitalized videos are stored as sequence of frames. Depend upon its **resolution** and **size**. Also to have **realistic video playback**, the **transmission**, **compression**, and **decompression** of digitalized require **continuous transfer rate**.

**Graphic Objects:** These consist of special data structures used to define 2D & 3D shapes through which we can define multimedia objects.

**3 (b) Why is MOLAP faster than RDBMS**

Relational database structures are **largely vendor independent** and **especially dimensional structures** are pretty easily ported.

**Performance**

**pre-aggregation** allows it to provide very fast responses to queries that would have required reading, grouping and summarizing millions of rows of relational star-schema data.

The **drilling and slicing and dicing** that an analyst would want to perform to **explore the data** would be immediate using a cube but could have lengthy **pauses when using a relational data source**.

**Drill down functionality**.  Many reporting software tools will automatically allow drilling up and down on dimensions with the data source is an OLAP cube

**Availability of software tools**.  OLAP Reporting Tools

**4(a) Hypermedia database**

* Approach to data management that **organized data as a network of nodes linked in any pattern the user specifies**.
* Hypermedia databases **spread across millions of independent computing systems**
* Any kind of database used for the **purpose of storing multimedia web pages**.
* **Data is stored online** and it can be **accessed by multiple users, simultaneously**. E.g. WWW

**4(B) Data Protection Act**

1. **Processed** fairly and lawfully.

2. **Obtained** for specified and lawful purposes.

3. **Adequate**, relevant and not excessive.

4. **Accurate** and up to date.

5. Not **kept** any longer than necessary.

6. Processed in accordance with the “data subject’s” (the **individual’s) rights.**

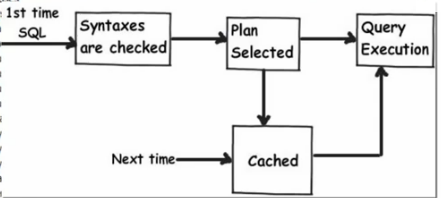
7. **Securely** kept.

8. Not transferred to any other country without adequate protection in situ.

**4 ( C ) Security and Transactions**

**Stored procedures -** Contains one or more SQL statements

SQL 2005, all queries and procedures are compiled / cached



To execute a stored procedure you can use EXEC statement.

*CREATE PROC spGetProductPrice*

*@Product varchar(50)As*

SELECT Product.Price FROM Product WHERE Product.name = @Product

call/execute this procedure from Query Analyzer.

*EXEC spGetProductPrice @ Product = 'Sandwich' OUTPUT*

**Prepared Statements** - execute the same or similar database

statements repeatedly with high efficiency

**Transactions – ACID**

operations which are executed as a whole to process user

requests for retrieving data or updating the database

* **Atomicity – Completeness – All or nothing**
* **Consistency – Keep it in a consistent state. One or another**
* **Isolation – Another transaction cannot interfere.**
* **Durability – permanence – permanent changes.**

**Dare Warehousing**

* Used to **collect and combine** data from other databases
* **Updates on regular** basis
* **Aggregates Data –** A whole formed by coming several
* **Looses some grainilatiry.**
* **Sub set** of data can be created – **Data Mart**

**ETL – Extract, Transform, Load**

**Extract – extract data from source**

**Transform –** transforms data in transit

**Load –** data into target storage of choice

* **Schedule** - Can be scheduled to happen.
* **Flexible** – e.g. nightly back ups
* **Rolling out** transaction data for business people to work with – data mart / data warehouse
* **Data migrations** – mergers and acquisitions

**Data connectivity**- should connect to any source of data

**Performance** – data requires processing power – *saleability*.

**Transformation flexibility** –tools to make data modifications

**Data quality** – Data is not clean, consistence and clean

**Flexible data acquisitions option** – scheduling.

[**Inmon's definition**](http://www.intranetjournal.com/features/datawarehousing.html)

**Subject-Oriented**:

A data warehouse can be used to analyze a **particular subject area**. *"sales" can be a particular subject.*

**Integrated**:

Integrates data from **multiple data sources**. source A and B may have different ways of identifying a product, but in a data warehouse, only a **single way of identifying a product.**

**Time-Variant**:

**Historical data is kept in a data warehouse**. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a **transaction system may hold the most recent address** of a customer, where a **data warehouse can hold all addresses associated with a customer**.

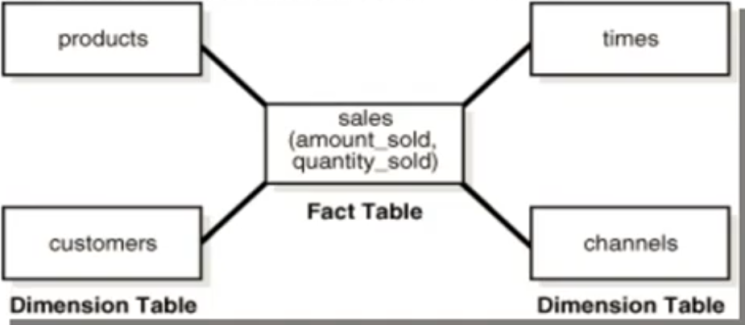
**Non-volatile**: Once data is in the warehouse, it **will not change**. So, historical data in a data **warehouse should never be altered**.

**Data Mining**

* **Slice and Dice – Dividing data into smaller parts** or views that yield more information
* **Drilldown** – **dividing information** up into finer and finer layers in a **hierarchy**, purpose of narrowing it 1small **Predictive analytics** - encompasses a variety of techniques from [statistics](http://en.wikipedia.org/wiki/Statistics), [modeling](http://en.wikipedia.org/wiki/Predictive_modelling), [machine learning](http://en.wikipedia.org/wiki/Machine_learning), and [data mining](http://en.wikipedia.org/wiki/Data_mining) that analyze current and historical facts to make [predictions](http://en.wikipedia.org/wiki/Prediction) about future, or otherwise unknown, events

**Star Schema**

* **Dimensional** simplifies the data model to facilitate access
* **Drill paths, hierarchy, and query profile** are embedded in the data model rather than the data.
* **Easier for inexperienced** users to navigate
* **Selfish model –** subject oriented – already aggregated.

****

**Star schemas** - facts 3NF, dimensions 2NF;

**Snowflake Schema**

**Snowflake schemas** - facts 3NF, dimensions 3NF.

* The snowflake is an extension to a dimension intended to **reduce storage and duplication.**
* Undesirable side **effect of complicating and slowing queries.**

**Dimensional modelling**

**Fact and dimension tables**

All **foreign keys** between fact and dimension tables should be **surrogate keys**, not reused keys from operational data.

Recommended simple integer as key value is meaningless and is only used to be join fields between the fact and dimension tables.

* **Performance** - efficient if a single field surrogate key
* **Handle unknown** or not applicable connections
* **Track changes** in dimension attribute values

**Content based Image Retrieval ( CBIR )**

* Application of [**computer vision**](http://en.wikipedia.org/wiki/Computer_vision) **techniques** to **the** [**image retrieval**](http://en.wikipedia.org/wiki/Image_retrieval) **problem** - *searching for* [*images*](http://en.wikipedia.org/wiki/Digital_image) *in large* [*databases*](http://en.wikipedia.org/wiki/Database)
* Content based image retrieval is opposed to **concept based approaches** also variably named as "description-based" or **"text-based" image indexing/retrieval**
* Generally **make use of lower-level features** like **texture**, **color**, and **shape**, although some systems take advantage of very common higher-level features like faces
* Not good at answering queries such as find all pictures of german shepards. E.g. GS may look like another dog.