Database Architecture

For the database exam

**This is kinda cool to see how its after changing**

**Discuss the link between memory management and concurrency (multiple users accessing shared data) -- no concurrency in exam?**

**This was in the architecture paper**

1. An in-memory database (IMDB; also main memory database system or MMDB or memory resident database) is a database management system that primarily relies on main memory for computer data storage. It is contrasted with database management systems that employ a disk storage mechanism.
2. Concurrency is the ability of a database to allow multiple users to affect multiple transactions. This is one of the main properties that separates a database from other forms of data storage like spreadsheets.

**Explain what is meant by load balancing**

Load balancing is dividing the amount of work that a computer has to do between two or more computers so that more work gets done in the same amount of time and, in general, all users get served faster.

“Load balancing is a core networking solution responsible for distributing incoming traffic among servers hosting the same application content.”

-so its just giving even distribution of queries between multiple servers

**Describe the term hit rate and comment on it as a factor in performance.**

The chief measurement of a cache, which is the percentage of all accesses that are satisfied by the data in the cache. Also known as "hit ratio.”

“The term **hit rate** refers to the proportion of successful accesses in the total number of accesses on a particular item of storage. It is a critical indicator of performance. Successful access means that the required data was already available in a cache when it was requested. If you store anything in a cache/buffer, to justify its position there it should have a high hit rate. Good hit rates = good performance.”

**Storing Query data (or QEP) may be useful for improved throughput, explain?**

Storing query data can help with optimising as if there is a very commonly executed query it'll have the data stored so the access time will be quicker in cache

**Describe how CPU scheduling and memory allocation are linked.**

When a process/thread is executing, it uses RAM for the data that it is processing. Therefore, we should note, that **threads and memory allocation are linked**. A thread based system is efficient as it conserves global memory for the overall process, as threads are swapped. In other words, the parent process can manage shared memory between its threads. So in general, memory management is better and the context switching for threads are lower cost while maintaining the same parent address space and control. Because of this threads are called ‘light weight’,

scheduling is the method by which threads, processes or data flows are given access to system resources

Threads under one parent can share common data

**A DBA (**[**Database administrator**](http://en.wikipedia.org/wiki/Database_administrator)**) faced with a choice of optimizer type faces a dilemma, explain?**

When an SQL statement is run, the DBMS must decide on a query execution plan (QEP) to implement that SQL. The DBMS may run an optimizer program to implement the SQL statement in an efficient way i.e. it develops the QEP.

**Explain how the same query may have different executions (or costs of executing) .**

If the execution plan was the same and one was slow, it would be database load, hardware, locking/blocking, etc.

However, if the execution plans are different something is different between the two databases.

**Explain, using an example, the basis of query optimisation.**

Query optimization is a function of many relational database management systems. The query optimizer attempts to determine the most efficient way to execute a given query by considering the possible query plans.

**example:**

**Select \* FROM**

College DB where department =”business” will result in a different QEP than department=”computing” because the number of records for each is different. So we would use an index on computing but NOT business

**What is meta data, and why is accessing meta data in the data dictionary critical for system performance?**

Meta data is “data about data”. An item of metadata describes the specific characteristics about an individual data item. In databases, metadata describes the structural components of tables and their elements. For example, metadata about an element could include data types, name of data, size and many more characteristics about that element. It would also give information about the tables the database is storing, information, such as length of fields, number of columns, where the tables are located and other pertinent information. One of the main uses for met data is to provide a link between the information creator and the information users. Meta data allows the users to speed up the search for individual data. This is done by being able to set parameter for searches.

**What is an SQL dialect? Why it is important for issues such as portability and data migration?**

The dialect of the database is simply a term that defines the specific features of the SQL language that are available when accessing that database.

**Explain using examples why a DBA must know about software developer (programmer) requirements during install/config.**

DBMS may have a setting that relates to the version of SQL supported. Many older programs have code that does not adhere to significant SQL standard releases. The administrator must liaison/agree with software developers that all programs that interact with the database adhere to a SQL standard before configuring/installing the DBMS.

**Describe an N-tier architecture for a database system? IS this recommended in all cases?**

**A three tier architecture is effectively spreading the overall processing load** over 3 machines. Any tier can be expanded but in data intensive applications the data tier is commonly more complex. This an N-tier architecture. Within the data server element of the overall architecture you may for instance be able to run multiple instances of the database on the same machine (or on different machines). Each instance can be configured specifically for the particular data it stores. This is moving into the area of distributed databases (note the terms federated databases, clouds or clustered databases relating to a co-ordinated set of data servers).

**What is thread and what is its role in making a database robust ( or fault tolerant) (or load balanced)**

Threads also enable a more robust system essential where demands of high availability are becoming a standard feature of many business systems. Usually this demand is prompted by web interfaces for transaction systems such as sales, ordering etc. Reliability means that the system is as close to constantly available as is possible.

**In a database there is far more data that can fit in RAM (processing memory); Describe how ‘good’ memory management is essential for efficient DBMS.**

http://en.wikibooks.org/wiki/Design\_of\_Main\_Memory\_Database\_System/Overview\_of\_DBMS

**Explain why multiple instances of a database might be used?**

Multiple instances against single (same) database is required in such environment where that database is distributed among different systems (either on LAN or WAN) forming cluster so users can access the data transparently.Each instance can be configured specifically for the particular data it stores.

Different types of applications require different configurations of the database to run efficiently. Rather than tuning the single database to give acceptable performance to all applications, one solution is to install multiple instances of the DB server, configuring each specifically to suit a limited subset of applications. Multiple instances are used for load balancing and targeted configuration for optimum performance; Instances can also be used in development and test environments.

**Explain what is the role of the administrator in Data/File Management on 2nd storage devices.**

In MySql these are called storage engines, but in general they are called file organisations. A database administrator might find it useful to change the file organisation and process the table in a different way i.e re-organise the table and therefore process it using a different internal subprogram e.g. change from a Btree to a Hash file

**Why not create indexes on all attributes? Should a QEP always use an index if available?**

Reason you don't use indexes for every table is your index is suppose to be for quick access

If you have every table indexed its the same size as the amount of data you got

Data Migration/Cleansing/Portability

**Describe and explain the following terms:**

Data migration is the process of transferring data between storage types, formats, or computer systems. It is a key consideration for any system implementation, upgrade, or consolidation.

**Data import** , Allowing you to import tables and database to your dbma system.

**Data export** allowing you to export your database and tables in exel,csv and sql files. It is vendor locked.

**Data cleansing,** data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

**data loading/unloading,**

|  |  |
| --- | --- |
| **Unloading** | **Copying database data to external text files for consumption by another database or another application (such as a spreadsheet application). The text files are in an industry-standard format such as tab-delimited or comma-delimited (CSV).** |
| **Loading** | **Copying data into the database from external text files that are in either a standard delimited format or in any of the formats that are supported by the SQL\*Loader utility.** |

**Database portability** Portability and data migration are closely related. SQL code that is portable means that it’ll work on any DBMS, so you can port or migrate your programs to a new system without encountering SQL dialect errors.

**legacy system** , In the general context, it can refer to any of the older database technologies. In a more specific context, it can refer to a database system that was inherited by a team from previous project owners.

**Data extraction** is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration).

**ETL**, In computing, Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that: Extracts data from homogeneous or heterogeneous data sources Transforms the data for storing it in proper format or structure for querying and analysis purpose .Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)

**Explain the principle components of a data import utility**

Command default values.

Character set used in the data file (source) and the database (destination). A Given DBMS e.g. MySql may not support all character sets. Windows ASCII, Unicode, Vendor specific.

Source/Destination specification e.g. path names, any restrictions on where data can be loaded from (source) etc.

Effects of performance; impact on concurrent transactions (reader, writers). This implies that the destination database / table already exists and contains data.

Option to delay execution of the import to a quiesced state of the database.

Need flexibility in handling various text formats/structures. Fields separated by commas, spaces, tabs. Lines ended by carriage return, carriage return & line feed, special control characters e.g. control Z.

Ability to skip file headers e.g. the column names in a spreadsheet, if so we may need to be able to skip the header. Note the header would confuse any format/structure analysis tool.

Ability to process data as it is imported e.g. calculate, skip fields, assign to specific table column names

Error handling: Errors may occur during the import process e.g. end of line encountered (i.e. not enough text in the input line); too much data on a line for the destination table definition. Blanks and Null values.

Security implications: privileges required to execute the import command (run the utility). Irrespective of the assigned privileges to users, the entire DBMS may be set up to allow/disallow data import (this would over ride the user privilege grants)

**Give examples of the application of data migration**

1. **Data migration** is the process of transferring **data** between storage types, formats, or computer systems. It is a key consideration for any system implementation, upgrade, or consolidation.

**Perform a data migration to include a data cleanse & data import**

**Explain, using an example, what concatenated fields in a database means, and comment on why concatenated fields are an issue for the data migration.**

Concatenation, in the context of databases, refers to the joining together two or more things into a large one. In database parlance, the things being joined are generally two table fields which may be from the same or different tables.

In relational database system design, we do not need concatenated values. W**e can break up the composite value and store the elements in separate columns. We then can create a composite key using the attributes themselves.** The database designer must decide whether to keep the composite codes in one column (design A below) or break up the elements into separate attributes using string functions, and import the set of attributes into the database (Design B below). This is an example of a transformation that could be programmed for regular migration.

**Explain the term multi-threading.**

1. a technique by which a single set of code can be used by several processors at different stages of execution.

**Explain the fundamental techniques used to implement SQL joins efficiently**

### Join Indexes - did this in labs

### Join algorithms

Three fundamental algorithms for performing a join operation exist: nested loop join, sort-merge join and hash join.

**Stored Procedures**

Stored Procedures are a batch of SQL statements that can be executed in a couple of ways. Most major DBMs support stored procedures; however, not all do. You will need to verify with your particular DBMS help documentation for specifics. As I am most famillar with SQL Server I will use that as my samples

In a database management system ([DBMS](http://searchsqlserver.techtarget.com/definition/database-management-system)), a stored procedure is a set of Structured Query Language ([SQL](http://searchsqlserver.techtarget.com/definition/SQL)) statements with an assigned name that's stored in the database in compiled form so that it can be shared by a number of programs. The use of stored procedures can be helpful in controlling access to data ([end-users](http://whatis.techtarget.com/definition/end-user) may enter or change data but do not write procedures), preserving [data integrity](http://searchdatacenter.techtarget.com/definition/integrity) (information is entered in a consistent manner), and improving productivity (statements in a stored procedure only need to be written one time).

^^ benefits = security (SQL injection attacks) + stored procedures can be fully optimized for fast execution + compiled