CS614 – Final Upto 2015 - MC130401237 Muhammad Kamran Khan

Q-1	Nested loop join variants?

1. Typically used in OLTP environment.
2. Limited application for DSS and VLDB
3. In DSS environment we deal with VLDB and large sets of data.

Nested-Loop Join: Variants

1. Naive nested-loop join
2. Index nested-loop join
3. Temporary index nested-loop join

Q-2	Define click stream?

Clickstream is every page event recorded by each of the company's Web servers

- Web-intensive businesses
- Although most exciting, at the same time it can be the most difficult and most frustrating.
- Not JUST another data source.

Q-3	What SSL stands for?

Secure Socket’s Layer

Q-4	Kimball preferred which approach interview or session?

Kimball prefers using a hybrid approach with interviews to gather the gory details and then facilitation to bring the group to consensus. However, the forum choice depends on the team's skills, the organization’s culture, and what you've already subjected your users to. This is a case in which one size definitely does not fit all.

Q-5	Reason of web searching? P-350

Web is large, actually very large.
To make it useful must be able to find the page(s) of interest/relevance.

Q-6	Limitation of using http secure socket layer briefly explain?

- To track the session, the entire information exchange needs to be in high overhead SSL
- Each host server must have its own unique security certificate.
- Visitors are put-off by pop-up certificate boxes.

Q-7	Write one implication of data ware housing if data documentation is not correctly done?

Lack of documentation may stall system operations as management people cannot manage what they don't know. Also, it may ultimately be used for educating the end users, prior to roll out.

Improper documentation

Usually by this time most, if not all, of the developers will have left the project, so it is essential that proper documentation is left for those who are handling production maintenance. There is nothing more frustrating than staring at something another person did, yet unable to figure it out due to the lack of proper documentation.

Q-8	Inverted index consist of bit vectors correct or not if correct, then justify?

No, Bitmap Index consists of a number of bit vectors or bitmaps.

Q-9	Calculate gender using classification rule of the given table?

The model in our case is a rule that if the per item minutes for any customer is greater or equal than 6 than the customer is female else a male i.e.

IF

Items/Time >= 6

Then

Gender= 'F'

Else

Gender = 'M'

The above rule is based on the common notion that females spend more time.

Q-10	Write 5 activities to measure success criteria of mail campaign? Measure customer's interest for campaigns, ads, e-mails?

Common measurements include:

1. Number of visitors
2. Number of sessions
3. Most requested page
4. Most requested download
5. Most accessed directories
6. Leading referrer sites
7. Leading browser and operating system
8. Visits by geographic breakdown, and many others

This information can be used to modify the design of the Web site, change ad campaigns or develop partnering relationships with other sites.

Q-11

Nested loop join, calculate the cost?

Player

Player_ID	Player_Name	Team
PK-01	Wasim	Pakistan
PK-02	Misbah	Pakistan
SA-03	AB Devillier	South Africa

Award

Award_ID	Match_ID	Player_ID
01	01	PK-01
01	02	PK-01
02	03	PK-02
01	04	SA-03

Total Blocks in Player = 3

Total Blocks in Award = 4

Qualifying Blocks in Player = 2

Qualifying Blocks in Award = 3

First, we consider Player as outer table:

Cost Player & Award = Total Blocks in Player + (Qualifying blocks in Player x Total Blocks in Award)

= 3 + (2 x 4)

= 11

Now, we consider Award as outer table:

Cost Award& Player = Total Blocks in Award + (Qualifying blocks in Award x Total Blocks in Player)

= 4 + (3 x 3)

= 13

Q-12	Write the activities of planning and design phase of Shaku Atre approach?

1. Determine Users' Needs
2. Determine DBMS Server Platform
3. Determine Hardware Platform
4. Information & Data Modeling
5. Construct Metadata Repository

Q-13	Two types of unsupervised clustering? P-271

- One-way Clustering
- Two-way Clustering

Q-14	What is pest scouting? P-333

1. Pest scouting is a systematic field sampling process.
2. It provides field specific information on pest pressure and crop injury.

Q-15	Name authority that controlling pest scouting in Punjab? P-333

The pest scouting data is being constantly recorded by the Directorate of Pest Warning and Quality Control of Pesticides (DPWQCP), Punjab since 1984.

Q-16	"Be a diplomat not a technologist"? P-320

The biggest problem you will face during a warehouse implementation will be people, not the technology or the development.

1. Management: You're going to have senior management complaining about completion dates and unclear objectives.
2. Development Team: You're going to have development people protesting that everything takes too long and why can't they do it the old way?
3. Users: You're going to have users with outrageously unrealistic expectations, who are used to systems that require mouse-clicking but not much intellectual investment on their part.
4. And you're going to grow exhausted, separating out Needs from Wants at all levels. Commit from the outset to work very hard at communicating the realities, encouraging investment, and cultivating the development of new skills in your team and your users (and even your bosses).

Most of all, keep smiling. When all is said and done, you'll have a resource in place that will do magic, and your grief will be long past. Eventually, your smile will be effortless and real.

Q-17	Three types of search? P-350

1. Keyword-based search
2. Querying deep Web sources
3. Random surfing

Q-18	Output of bit streams? P-234
	(a) 111100001111 (b) 0001001100

a. 14#04#14
b. 03#11#02#12#02

Q-19	Identify statement as correct or incorrect justify?
X	“One-way clustering gives local view and Two-way clustering gives global view”
✓	“One-way clustering gives global view and Two-way clustering gives local view”.

Q-20 Calculate the bit map index of given table? P-234

1) Select * from **Player_Country** where Player_type =”Bowler” OR Country_name=”Australia”

2) Select * from **Player_Country** where Player_type =”Batsman” AND Country_name=”Pakistan”

Player_ID	**Player_type**	**Country_name**
PK-01	Batsman	Pakistan
PK-05	Bowler	Pakistan
SA-07	Batsman	South Africa
SA-09	Wicket Keeper	South Africa
AU-01	All Rounder	Australia
AU-04	Bowler	Australia
WI-01	Batsman	West Indies
WI-06	All Rounder	West Indies

First we consider the first query:

Select * from **Player_Country** where Player_type=” Bowler” OR Country_name=”Australia”

We create bitmap index table for “Player_type”.

Player_ID	All rounder	Bowler	Batsman	Wicket Keeper
PK-01			1	
PK-05		1		
SA-07			1	
SA-09				1
AU-01	1			

AU-04		1		
WI-01			1	
WI-06	1			

Bitmap index table for Country_name

Player_ID	Pakistan	South Africa	Australia	West Indies
PK-01	1			
PK-05	1			
SA-07		1		
SA-09		1		
AU-01			1	
AU-04			1	
WI-01				1
WI-06				1

Now we create bitmap vector for Player_type = “Bowler” : 01000100

Now we create bitmap vector for Country_name = “Australia” : 00001100

As there is “OR” operator, so we perform OR operation : 01001100

So, the resultant rows will be 2nd, 5th and 6th i.e. Players with IDs=PK-05, AU-01 and AU-04 will be selected.

Now, we consider the second query:

Select * from **Player_Country** where Player_type =”Batsman” AND Country_name=”Pakistan”

We create bitmap index table for “Player_type”.

Player_ID	All rounder	Bowler	Batsman	Wicket Keeper
PK-01			1	
PK-05		1		
SA-07			1	
SA-09				1

AU-01	1			
AU-04		1		
WI-01			1	
WI-06	1			

Bitmap index table for Country_name

Player_ID	Pakistan	South Africa	Australia	West Indies
PK-01	1			
PK-05	1			
SA-07		1		
SA-09		1		
AU-01			1	
AU-04			1	
WI-01				1
WI-06				1

Now we create bitmap vector for Player_type = “Batsman” : 10100010

Now we create bitmap vector for Country_name= “Paksitan” : 11000000

As there is “AND” operator, so we perform AND operation : 10000000

So only one record will be selected which is the player whose ID is =”PK-01”

Q-21	Why should companies entertain students to visit their company's place? P-328

1. You are students, and whom you meet were also once students.
2. You can do an assessment of the company for DWH potential at no cost.
3. Since you are only interested in your project, so your analysis will be neutral.
4. Your report can form a basis for a professional detailed assessment at a later stage.
5. If a DWH already exists, you can do an independent audit.

Q-22	Describe problems that can occur due to the use single Server for Development and production?

1. Sometimes it is possible that the server needs to be rebooted for the development environment. Having single development environment will prevent the production environment from working.
2. There may be interference while having different database environments on a single server. For example, having multiple long queries running on the development server could affect the performance on the production server, as both are same.

Q-23	Attributes of Web page dimension?

1. Describes the page context for a Web page event
2. Definition of page must be flexible
3. Assume a well-defined function that
    a. Characterizes the page
    b. Describe the page
4. The page dimension is small

Q-24	Pitfalls of session information using Ping-pong?

1. Requires a great deal of control over the Web site's page-generation methods
2. Approach breaks down if multiple vendors are supplying content in a single session

Q-25	Hash Index?

In contrast to B-tree indexing, hash based indexes do not (typically) keep index values in sorted order.

1. Index entry is found by hashing on index value requiring exact match.

    SELECT * FROM Customers WHERE AccttNo= 110240
2. Index entries kept in hash organized tables rather than B-tree structures.
3. Index entry contains ROWID values for each row corresponding to the index value.
4. Remember few numbers in real-life to be useful for hashing.

A hash scan requires an exact match of a key value. It's extremely fast and efficient as long as the exact value (typically a number) is known. It's best for things like account number, NID number, or part vehicle registration number.

For example, if a customer wanted to buy a car on an installment plan from a bank, the bank would typically need his/her exact account number, for instance 110240. The SQL syntax to look it up would be similar to the following:

SELECT * FROM Customers WHERE AccttNo= 110240

B-tree vs. Hash Indexes

1. Indexing (using B-trees) good for range searches, e.g.:

   SELECT * FROM R WHERE A > 5

2. Hashing good for match based searches, e.g.:

   SELECT * FROM R WHERE A = 5

Q-26	How many clusters, type of cluster?

1) There are two cluster made from the available data. The clusters are identified based on Pool_ID because there are total two pools in the table and there are further two teams in each Pool_ID. The two clusters are i.e. A and B.

Question				**Answer**
Player_ID	**Player_Name**	**Team_ID**	**Pool_ID**	**Cluster**
WI-01	Ambrose	WI	A	one
WI-06	Richordson	WI	A	
PK-01	Hafiz	PK	A	
PK-05	Misbah	PK	A	
AU-04	Maxwell	AU	B	two
AU-01	Steve Waugh	AU	B	
SA-07	AB Devillier	SA	B	
SA-09	Dal Styn	SA	B	

2)

The clustering is one way clustering because we have selected one column for clustering.

Q-27	How many types of partitioning? Write at least three name of Shared nothing RDBMS architecture?

- Hash partitioning
- Key range partitioning.
- List partitioning.
- Round-Robin
- Combinations (Range-Hash & Range-List)

Q-28	Web warehouse 5 attributes?

Timestamp	Log Tag	URL	Request Method
Elapsed Time	HTTP Code	User Identity	Size
Client Address	Hierarchy Data	Content Type	

Q-29	Click stream issues?

Clickstream data has many issues.

1. Identifying the Visitor Origin
2. Identifying the Session
3. Identifying the Visitor
4. Proxy Servers
5. Browser Caches

Q-30	Web dimensions?

- Some are standard dimensions.
- Other are non-standard and different.

Some of dimensions for a Web retailer could include:

Date, Time of day, Part, Vendor, Transaction, Status, Service policy, Internal organization, Employee. These are same as for DWH.

Page, Event, Session, Referral, these are created additionally to cater the web requirements.

Q-31	Precedence's constraints?

Unconditional: If you want Task 2 to wait until Task 1 completes, regardless of the outcome, link Task 1 to Task 2 with an *unconditional precedence* constraint.

On Success: If you want Task 2 to wait until Task 1 has successfully completed, link Task 1 to Task 2 with an *On Success precedence* constraint.

On Failure: If you want Task 2 to begin execution only if Task 1 fails to execute successfully, link Task 1 to Task 2 with an *On Failure precedence* constraint. If you want to run an alternative branch of the workflow when an error is encountered, use this constraint.

Q-32	Business process diagram from Kimball?

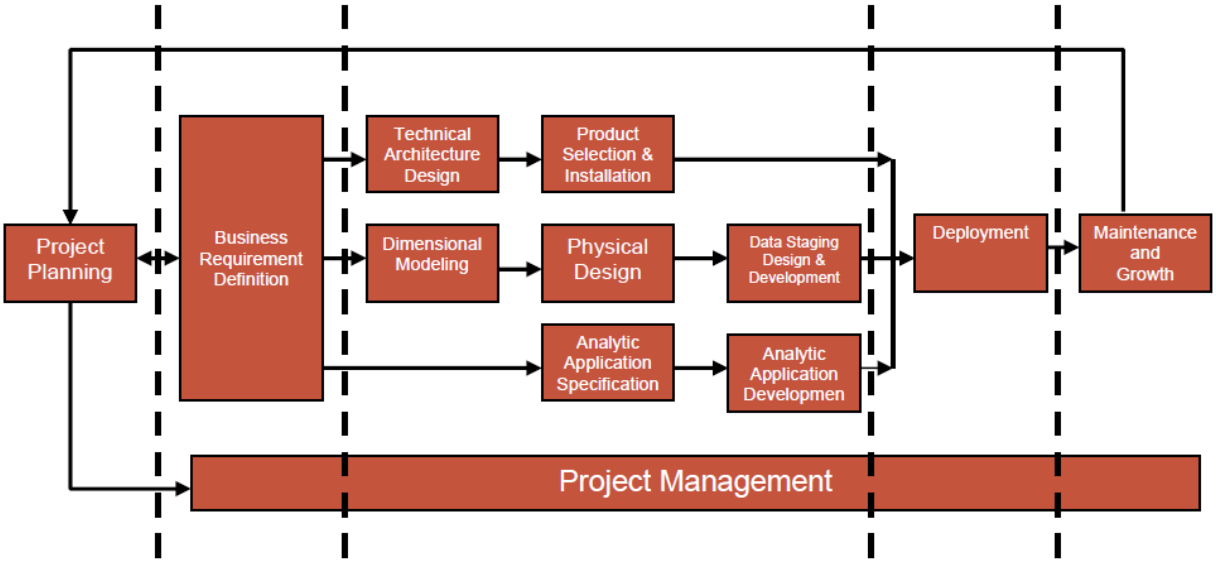


Figure -33.1: Business Dimensional Lifecycle (Kimball's Approach)

Q-33	What is web page, attributes of web page?

A hypertext document connected to the World Wide Web.

Timestamp	Log Tag	URL	Request Method
Elapsed Time	HTTP Code	User Identity	Size
Client Address	Hierarchy Data	Content Type	

Q-34	Do you agree a single technology/tool is sufficient to fulfill all needs of users?

No, different users require different technologies, e.g. MS Office.

Q-35	How would you determine outer / inner table in Nested-Loop join?

The outer table is usually the one that has:

- The smallest number of qualifying rows, and/or
- The largest numbers of I/Os required to locate the rows.

The inner table usually has:

- The largest number of qualifying rows, and/or
- The smallest number of reads required to locate rows.

Q-36	Write names of steps of Kimball DWH lifecycle?

DWH Lifecycle: Key steps

1. Project Planning
2. Business Requirements Definition
3. Parallel Tracks

3.1 Lifecycle Technology Track

3.1.1 Technical Architecture

3.1.2 Product Selection

3.2 Lifecycle Data Track

3.2.1 Dimensional Modeling

3.2.2 Physical Design

3.2.3 Data Staging design and development

3.3 Lifecycle Analytic Applications Track

3.3.1 Analytic application specification

3.3.2 Analytic application development

4. Deployment
5. Maintenance

Q-37	Write Drawbacks of traditional web searches?

1. Limited to keyword based matching.

2. Cannot distinguish between the contexts in which a link is used.
3. Coupling of files has to be done manually.

Q-38	Why RAD methodology is successful, write at least two reasons?

It is much better suited to the development of a data warehouse because of its iterative nature and fast iterations. There are 5 keys to a successful rapid prototyping methodology:

1. Assemble a small, very bright team of database programmers, hardware technicians, designers, quality assurance technicians, documentation and decision support specialists, and a single manager.

2. Define and involve a small "focus group" consisting of users (both novice and experienced) and managers (both line and upper). These are the people who will provide the feedback necessary to drive the prototyping cycle. Listen to them carefully.

3. Generate a user's manual and user interface first. These will prove to be amazing in terms of user feedback and requirements specification.

4. Use tools specifically designed for rapid prototyping. Stay away from C, C++, COBOL, SQL, etc. Instead use the visual development tools included with the database.

5. Remember a prototype is NOT the final application. It servers a means of making the user more expressive about requirements and developing in them a clear understanding and vision of the system. Prototypes are meant to be copied into production models. Once the prototypes are successful, begin the development processing using development tools, such as C, C++, Java, SQL, etc.

Q-39	What is reverse proxy?

Another type of proxy server, called a reverse proxy, can be placed in front of our enterprise's Web servers to help them offload requests for frequently accessed content. This kind of proxy is entirely within our control and usually presents no impediment to Web warehouse data collection. It should be able to supply the same kind of log information as that produced by a Web server.

Q-40	List down three steps which are performed in requirement definition phase of Kimball's approach in data warehouse development?

Requirements preplanning: This phase consists of activities like choosing the forum, identifying and preparing the requirements team and finally selecting, scheduling and preparing the business representatives.

Business requirements collection: The requirements collection process flows from an introduction through structured questioning to a final wrap-up. The major activities involved are, launching, determining interview flow, wrapping up and conducting data centric interviews.

Post collection: The phase consists of steps like debriefing, documentation, prioritization and consensus**.**

Q-41	Being a part of training team specify three guidelines that you consider as part of effective user education program?

1. Continuing education program.
2. Formal refresher, as well as advanced courses and repeat introductory course.
3. Informal education for developers and power users for exchange of ideas.

Q-42	Brief Intro to Parallel Processing? List down any 2 parallel hardware architecture?

Parallel Hardware Architectures

- Symmetric Multi-Processing (SMP)
- Distributed Memory or Massively Parallel Processing (MPP)
- Non-uniform Memory Access (NUMA)

Parallel Software Architectures

- Shared Memory
- Shard Disk
- Shared Nothing

Types of parallelism

- Data Parallelism
- Spatial Parallelism

Q-43	A data warehouse project is more like scientific research than anything in traditional informational system do you agree or not justify in either case?

The normal Information System (IS) approach emphasizes on knowing what the expected results are before committing to action. In scientific research, the results are unknown up front, and emphasis is placed on developing a rigorous, step-by-step process to uncover the truth. The activities involve regular interactions between the scientist and the subject and also among the project participants. It is advised to adopt an exploratory, hands-on process involving cross-disciplinary participation.

Q-44	In a distributed memory machine a processor can write a value into a shared memory and all process can read this value (Correct or incorrect statement)?

In a shared memory machine, a processor can simply write a value into a particular memory location, and all other processors can read this value. In a distributed-memory machine, exchanging values of variables involves explicit communication over the network, thus need for a high speed interconnection network.

Q-45	In context of nested loop join mention 2 guideline for selecting a table as inner table?

A nested loop join involves the following steps:

1. The optimizer determines the major table (i.e. Table-A) and designates it as the outer table. Table-A is accessed once. If the outer table has no useful indexes, a full table scan is performed. If an index can reduce I/O costs, the index is used to locate the rows.

2. The other table is designated as the inner table or Table-B. Table-B is accessed once for each qualifying row (or tuple) in Table-A.

3. For every row in the outer table, DBMS accesses all the rows in the inner table. The outer loop is for every row in outer table and the inner loop is for every row in the inner table.

Q-46	What is purpose that data profiling must fulfill?

Data profiling is a process of gathering information about columns, it must fulfil the following purposes

- Identify the type and extent to which the transformation is required
- The number of columns which are required to be transformed and which transformation is required, meaning date format or gender convention.
- It should provide us a detailed view about the quality of data. The number of erroneous values and the number of values out of domain.

Q-47	Differentiate between static and data mining w.r.t. No. of parameters (dimension) and type of date?

Statistics is useful only for data sets with limited parameters (dimensions) and simple relationships (linear). Statistical methods fail when the data dimensionality is greater and the relationships among different parameters are complex. Data mining proves to be viable solution in such situations.

One difference is on the type of data. While statistician traditionally work with smaller and first hand data" that has been collected or produced to check specific hypotheses, data miners work with huge and second hand data" often assembled from different sources. The idea is to find interesting facts and potentially useful knowledge hidden in the data and often unrelated to the primary purpose why the data have been collected.

Q-48	3 activities that u will consider as part of required preplanning phase?

This phase consists of activities like:

1. Choosing the forum

2. Identifying and preparing the requirements team
3. Finally selecting, scheduling and preparing the business representatives.

Q-49	How gender guide is used, when gender is missing?

If for very large number of records gender is missing, it would become impossible for us to manually check each and every individual's name and identify the gender. In such cases we can formulate a mechanism to correct gender.

1. We can either use a standard gender guide or create a new table Gender guide.
2. Gender guide contains only two columns name and gender.
3. Populate Gender guide table by a query for selecting all distinct first names from student table. Then manually placing their gender.
4. This table can serve us as guide by telling what can be the gender against this particular name. For example if we have hundred students in our database with first name equal to 'Muhammed'. Then in our Gender guide table we will have just one entry 'Muhammed' and we will manually set the gender as 'Male' against 'Muhammed'.
5. Now to fill missing genders in exception table we will just do a inner join on Error table and Gender_guide table. We will get the gender against matching names.

Q-50	Do you think it will create the problem of non-standardized attributes, if one source uses 0/1 and second source uses 1/0 to store male/female attribute respectively? Give a reason to support your answer.

Yes, it will create the problem of non-standardized attributes, if one source uses 0/1 and second source uses 1/0 to store male/female attribute respectively. Because one campus consider male other will consider it as female.

Q-51	Two ways in which parallelism can reduce system's performance?

Parallelism when not observed or practices carefully can actually degrade the performance, in case the system is over utilized and the law of diminishing returns sets in or there is insufficient bandwidth and it actually becomes the bottleneck and chokes the system.

Q-52	Why Analytic Track is considered as "fun part"?

We're finally using the investment in technology and data to help users make better decisions. The applications provide a key mechanism for strengthening the relationship between the project team and the business community. They serve to present the data warehouse's face to its business users, and they bring the business needs back into the team of application developers

Q-53	Write any three complete warehouse deliverables?

- Data
- Analytic applications
- Data access tools
- Education tools

Q-54	This query was given SELECT*FROM R WHERE A= 5 We have to tell which indexing technique is better dense, sparse, B-tree or hash?

Hash as this query requires exact match instead of range or anything else.

Q-55

Two tables were given employee and exception table.

Employee Table

EmpID	EmpName	Age
1	Ali	28
2	Faisal	32
3	Waseem	389
4	Arham	398

Exception table

EmpID	IsAgeValid
1	1
2	1
3	1
4	1

We have to write a query to access employee table and set the value of IsAgeValid =0 where age is greater than and equal to 25 and less than and equal to 75?

Update Exception x JOIN Employee e ON
e.EmpID = x.EmpID
Set IsAgeValid = 0
Where e.Age Between 25 AND 75