SQL Interview Questions
By
www.questpond.com

Dedication

This book is dedicated to my kid Sanjana, whose dad’s playtime has been stolen and given to this book. I am thankful to my wife for constantly encouraging me and to BPB Publication to give newcomer a platform to perform. Finally, at the top of all thanks to two old eyes my mom and

dad for always blessing me. I am blessed to have Raju as my brother who always keeps my momentum moving on.

I am grateful to Bhavnesh Asar who initially conceptualized the idea I believe concept thinking is more important than execution. Tons of thanks to my reviewers whose feedback provided an essential tool to improve my writing capabilities.

Just wanted to point out Miss Kadambari. S. Kadam took all the pain to review for the left outs with out which this book would have never seen the quality light.

About the author

Author works in a big multinational company and has over 8 years of experience in software industry. He is working presently as project lead and in past has led projects in Banking, travel and financial sectors.

However, on the top of all, I am a simple developer like you all guys there doing an 8 hour job. Writing is something I do extra and I love doing it. No one is perfect and same holds true for me .So anything you want to comment, suggest, point typo / grammar mistakes or technical mistakes regarding the book you can mail me at shiv_koirala@yahoo.com. Believe me guys your harsh words would be received with love and treated to the top most priority. Without all you guys I am not an author.

Writing an interview question book is really a great deal of responsibility. I have tried to cover maximum questions for the topic because I always think probably leaving one silly question will cost someone's job there. However, huge natural variations in an interview are something difficult to cover in this small book. Therefore, if you have come across such questions during interview, which is not addressed in this book do, mail at **shiv_koirala@yahoo.com** .Who knows probably that question can save some other guys job.

Features of the book

- This book goes in best combination with my previous book ".NET Interview questions". One takes care of your front-end aspect and this one the back end, which will make you really stand out during .NET interviews.
- Around 400 plus SQL Server Interview questions sampled from real SQL Server Interviews conducted across IT companies.
- Other than core level interview question, DBA topics like database optimization and locking are also addressed.
- Replication section where most of the developer stumble, full chapter is dedicated to replication so that during interview you really look a champ.
- SQLCLR that is .NET integration, which is one of the favorites of every interviewer, is addressed with great care .This makes developer more comfortable during interview.
- XML is one of the must to be answered questions during interview. All new XML features are covered with great elegance.
- Areas like data warehousing and data mining are handled in complete depth.

- Reporting and Analysis services, which can really surprise developers during interviews, are also dealt with great care.
- A complete chapter on ADO.NET makes it stronger from a programmer aspect. In addition, new ADO.NET features are also highlighted which can be pain points for the new features released with SQL Server.
- Must for developers who are looking to crack SQL Server interview for DBA position or programmer position.
- Must for fresher's who want to avoid some unnecessary pitfall during interview.
- Every answer is precise and to the point rather than hitting around the bush. Some questions are answered to greater detail with practical implementation in mind.
- Every question is classified in DB and NON-DB level. DB level question are mostly for guys who are looking for high profile DBA level jobs. All questions other than DB level are NON-DB level, which is must for every programmer to know.
- Tips and tricks for interview, resume making and salary negotiation section takes this book to a greater height.

Introduction

When my previous book ".NET Interview Questions" reached the readers, the only voice heared was more "SQL Server". Ok guys we have heard it louder and clearer, so here is my complete book on SQL Server: - "SQL Server Interview Questions". However, there is a second stronger reason for writing this book, which stands taller than the readers demand and that is SQL Server itself. Almost 90 % projects in software industry need databases or persistent data in some or other form. When it comes to .NET persisting data SQL Server is the most preferred database to do it. There are projects, which use ORACLE, DB2 and other database product, but SQL Server still has the major market chunk when language is .NET and especially operating system is windows. I treat this great relationship between .NET, SQL Server and Windows OS as a family relationship.

In my previous book, we had only one chapter, which was dedicated to SQL Server, which is complete injustice to this beautiful product.

So why an interview question book on SQL Server? If you look at any .NET interview conducted in your premises both parties (Employer and Candidate) pay no attention to SQL Server even though when it is such an important part of development project. They will go talking about stars (OOP, AOP, Design patterns, MVC patterns, Microsoft Application blocks, Project Management etc.) but on database side, there would be rare questions. I am not saying these things are not important but if you see in development or maintenance majority time, you will be either in your IDE or in SQL Server.

Secondly many candidates go really as heroes when answering questions of OOP , AOP , Design patterns , architecture , remoting etc etc but when it comes to simple basic question on SQL Server like SQL , indexes (forget DBA level questions) they are completely out of track.

Third very important thing IT is changing people expect more out of less. That means they expect a programmer should be architect, coder, tester and yes and yes a DBA also. For mission critical data there will always be a separate position for a DBA. Now many interviewers expect programmers to also do a job of DBA, Data warehousing etc. This is the major place where developers lack during facing these kinds of interview.

Therefore, this book will make you walk through those surprising questions, which can sprang from SQL Server aspect. I have tried to not go too deep, as that will defeat the complete purpose of an Interview Question book. I think that an interview book should make you run through those surprising question and make you prepare in a small duration (probably with a night or so). I hope this book really points those pitfalls that can come during SQL Server Interview's.

I hope this book takes you to a better height and gives you extra confidence boost during interviews. Best of Luck and Happy Job-Hunting.....

How to read this book

If you can read English, you can read this book...kidding. In this book, there are some legends, which will make your reading more effective. Every question has simple tags, which mark the rating of the questions.

```
These rating are given by Author and can vary according to companies
and individuals.
```

Compared to my previous book ".NET Interview Questions" which had three levels (Basic, Intermediate and Advanced) this book has only two levels (DBA and NON-DBA) because of the subject. While reading you can come across section marked as "Note" , which highlight special points of that section. You will also come across tags like "TWIST", which is nothing , but another way of asking the same question, for instance "What is replication?" and "How do I move data between two SQL Server database?" , point to the same answer.

All questions with DBA level are marked with (DB) tag. Questions, which do not have tags, are NON-DBA levels. Every developer should have a know how of all NON-DBA levels question. But for DBA guys every question is important. For instance if you are going for a developer position and you flunk in simple ADO.NET question you know the result. Vice versa if you are going for a DBA position and you cannot answer basic query optimization questions probably, you will never reach the HR round.

So the best way to read this book is read the question and judge yourself do you think you will be asked these types of questions? For instance, many times you know you will be only asked about data warehousing and rather than hitting the bush around you would like to target that section more. In addition, many a times, you know your weakest area and you would only like to brush up those sections. You can say this book is not a book that has to be read from start to end you can start from a chapter or question and when you think you are ok close it.

Software Company hierarchy

It is very important during interview to be clear about what position you are targeting. Depending on what positions you are targeting the interviewer shoots you questions. Example if you are

looking for a DBA position you will be asked around 20% ADO.NET questions and 80% questions on query optimization, profiler, replication, data warehousing, data mining and others.

Note: - In small scale software house and mid scale software companies there are chances where they expect a developer to a job of programming , DBA job , data mining and everything. But in big companies you can easily see the difference where DBA job are specifically done by specialist of SQL Server rather than developers. But now a days some big companies believe in a developer doing multitask jobs to remove dependencies on a resource.

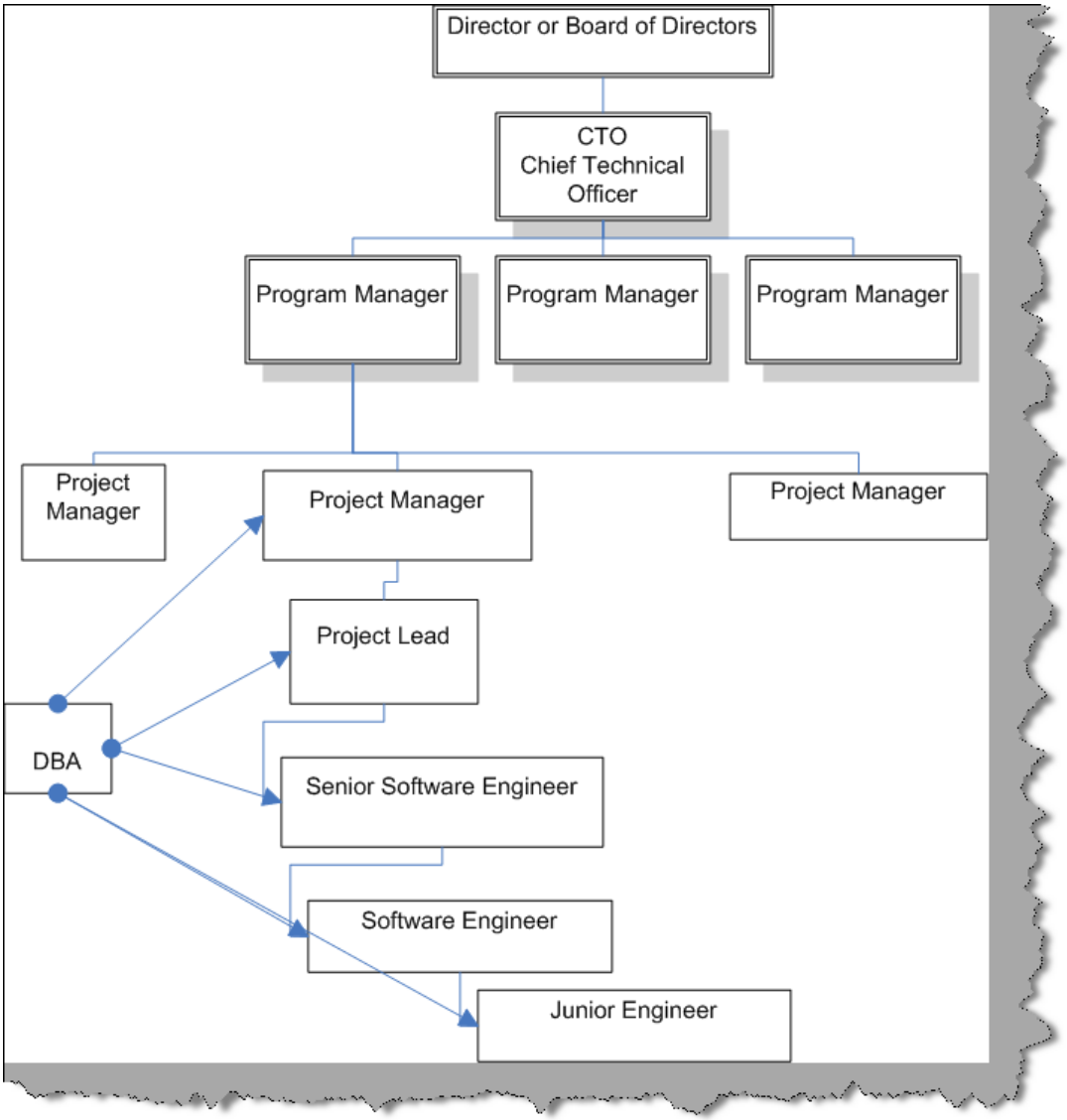


Figure: - 0.1 IT Company hierarchy

Above is a figure of a general hierarchy across most IT companies (Well not always but I hope most of the time). Because of inconsistent HR, way of working you will see difference between companies.

Note: - There are many small and medium software companies which do not follow this hierarchy and they have there own ADHOC way of defining positions in the company.

So why there is a need of hierarchy in an interview?

"Interview is a contract between the employer and candidate to achieve specific goals."

Therefore, employer is looking for a suitable candidate and candidate for a better career. Normally in interviews, the employer is very clear about what type of candidate he is looking for. However, 90% times the candidate is not clear about the positions he is looking for.

How many times has it happened with you that you have given a whole interview and when you mentioned the position you are looking for...pat comes the answer we do not have any requirements for this position. So be clarified about the position right when you start the interview.

Following are the number of years of experience according to position.

- Junior engineers are especially fresher and work under software engineers.
- Software engineers have around 1 to 2 years of experience. Interviewer expects software engineers to have know how of how to code ADO.NET with SQL Server.
- Senior Software Engineers have around 2 to 4 years of experience. Interviewer expects them to be technically very strong.
- Project leads should handle majority technical aspect of project and should have around 4 to 8 years of experience. They are also actively involved in to defining architect of the project. Interviewer expects them to be technically strong plus should have managerial skills.
- Project Managers are expected to be around 40% technically strong and should have experience above 10 years plus. But they are more interviewed from aspect of project management, client interaction, people management, proposal preparation etc.
- Pure DBA's do not come in hierarchy as such in pure development projects. They do report to the project managers or project leads but they are mainly across the hierarchy helping every one in a project. In small companies, software developers can also act as DBA's depending on company's policy. Pure DBA's have normally around 6 and above years of experience in that particular database product.
- When it comes to maintenance, projects where you have special DBA positions lot of things are ADHOC. That means one or two guys work fulfilling maintenance tickets.

So now judge where you stand where you want to go.....

Resume Preparation Guidelines

First impression the last impression

Before even the interviewer meets you, he will first meet your resume. Interviewer looking at your resume is almost a 20% interview happening with out you knowing it. I was always a bad guy when it comes to resume preparation. But when I looked at my friends resume they where gorgeous. Now that I am writing series of book on interviews, I thought this would be a good point to put in. You can happily skip it if you are confident about your resume. There is no hard and fast rule that you have to follow the same pattern but just see if these all checklist are attended.

- Use plain text when you are sending resumes through email. For instance, you sent your resume using Microsoft word and what if the interviewer is using Linux he will never be able to read your resume. You cannot be sure both wise, you sent your resume in Word 2000 and the guy has Word 97…uuhhh.
- Attach a covering letter it really impresses and makes you look traditionally formal. Yes even if you are sending your CV through e-mail, send a covering letter.

Checklist of content you should have in your resume:-

- Start with an objective or summary, for instance, "Working as a Senior Database administrator for more than 4 years. Implemented quality web based application. Followed the industry's best practices and adhered and implemented processes, which enhanced the quality of technical delivery. Pledged to deliver the best technical solutions to the industry."
- Specify your Core strengths at the start of the resume by which the interviewer can make a quick decision are you eligible for the position. For example :-
    - Looked after data mining and data warehousing department independently. Played a major role in query optimization.
    - Worked extensively in database design and ER diagram implementation.
    - Well versed with CMMI process and followed it extensively in projects.
- Looking forward to work on project manager or senior manager position.

This is also a good position to specify your objective or position that makes it clear to the interviewer that should he call you for an interview. For instance if you are looking for senior position specify it explicitly looking for this job profile. Any kind of certification like MCP, MCSD etc you can make it visible in this section.

- Once you have specified briefly your goals and what you have done its time to specify what type of technology you have worked with. For instance RDBMS, TOOLS, Languages, Web servers, process (Six sigma, CMMI).
- After that, you can make a run through of your experience company wise that is what company you have worked with, year / month joining and year / month left. This will

give an overview to the interviewer what type of companies you have associated your self.

Now its time to mention all your projects you have worked till now. Best is to start in descending order that is from your current project and go backwards. For every project try to put these things:-

- Project Name / Client name (It is sometimes unethical to mention clients name; I leave it to the readers).
- Number of team members.
- Time span of the project.
- Tools, language, RDBMS and technology used to complete the project.
- Brief summary of the project.

Senior people who have huge experience will tend to increase there CV with putting in summary for all project. Best for them is to just put down description of the first three projects in descending manner and rest they can say verbally during interview. I have seen CV above 15 pages… I doubt who can read it.

- Finally comes your education and personal details.
- Trying for onsite, do not forget to mention your passport number.
- Some guys tend to make there CV large and huge. I think an optimal size should be not more than 4 to 5 pages.
- Do not mention your salary in CV. You can talk about it during interview with HR or the interviewer.
- When you are writing your summary for project make it effective by using verbs like managed a team of 5 members, architected the project from start to finish etc. It brings huge weight.
- This is essential very essential take 4 to 5 Xerox copies of your resume you will need it now and then.
- Just in case, take at least 2 passport photos with you. You can escape it but many times you will need it.
- Carry you are all current office documents specially your salary slips and joining letter.

Salary Negotiation

Ok that is what we all do it for money… not every one right. This is probably the weakest area for techno savvy guys. They are not good negotiators. I have seen so many guys at the first instance they will smile say “NEGOTIABLE SIR”. So here are some points:-

- Do a study of what‘s the salary trend? For instance, have some kind of baseline. For example, what is the salary trend on number of year of experience? Discuss this with your friends out.

- Do not mention your expected salary on the resume?
- Let the employer first make the salary offer. Try to delay the salary discussion until the end.
- If they say, what do you expect? , come with a figure with a little higher end and say negotiable. Remember never say negotiable on something which you have aimed, HR guys will always bring it down. So negotiate on AIMED SALARY + some thing extra.
- The normal trend is that they look at your current salary and add a little it so that they can pull you in. Do your home work my salary is this much and I expect this much so whatever it is now I will not come below this.
- Do not be harsh during salary negotiations.
- It is good to aim high. For instance, I want 1 billion dollars / month but at the same time be realistic.
- Some companies have those hidden cost attached in salary clarify that rather to be surprised at the first salary package.
- Many of the companies add extra performance compensation in your basic that can be surprising at times. So have a detail break down. Best is to discuss on hand salary rather than NET.
- Talk with the employer in what frequency does the hike happen.
- Take everything in writing , go back to your house and have a look once with a cool head is the offer worth it of what your current employer is giving.
- Do not forget once you have job in hand you can come back to your current employer for negotiation so keep that thing in mind.
- Remember the worst part is cribbing after joining the company that your colleague is getting this much. So be careful while interview negotiations or be sportive to be a good negotiator in the next interview.
- One very important thing the best negotiation ground is not the new company where you are going but the old company, which you are leaving. So once you have offer on hand get back to your old employee, show them the offer, and then make your next move. It is my experience that negotiating with the old employer is easy than with the new one….Frankly if approached properly rarely any one will say no. Just do not be aggressive or egoistic that you have an offer on hand.

Top of all some time some things are worth above money: - JOB SATISFACTION. So whatever you negotiate if you think you can get JOB SATISFACTION aspect on higher grounds go for it. I think its worth more than money.

Points to remember

- One of the first questions asked during interview is “Can you say something about yourself”.

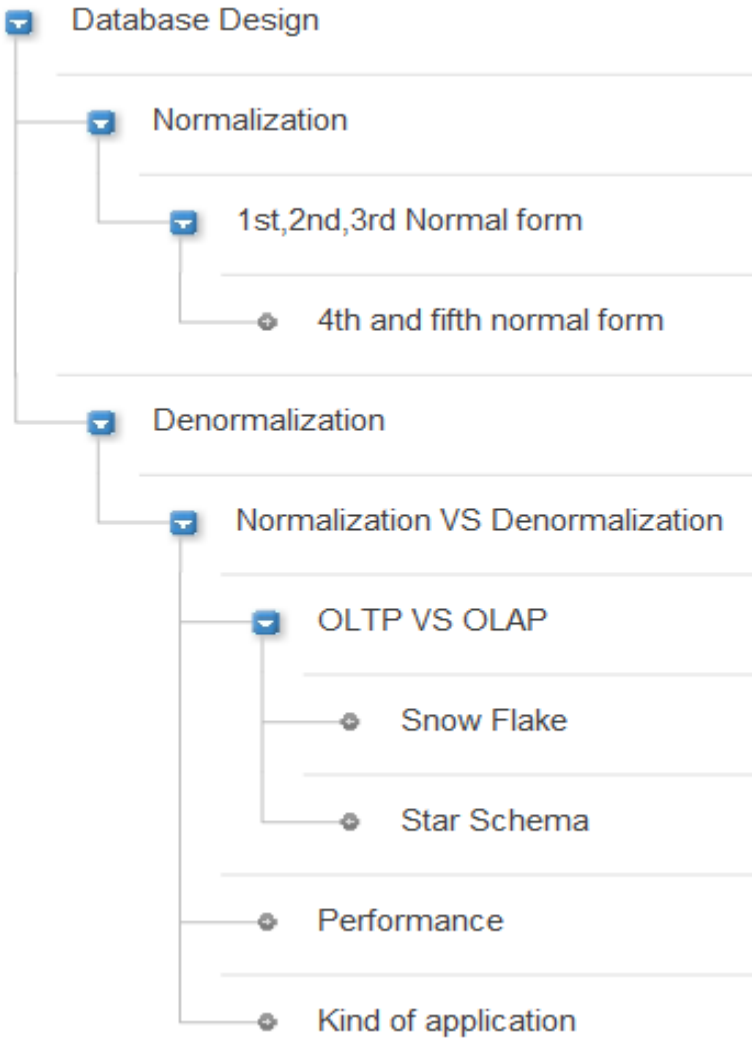
- Can you describe about your self and what you have achieved till now?
- Why you want to leave the current company?
- Where do you see yourself after three years?
- What are your positive and negative points?
- How much do you rate yourself in .NET and SQL Server in one out of ten?
- Are you looking for onsite opportunities? (Be careful do not show your desperation of abroad journeys)
- Why have you changed so many jobs? (Prepare a decent answer do not blame companies and individuals for your frequent change).
- Never talk for more than 1 minute straight during interview.
- Have you worked with previous version of SQL Server?
- Would you be interested in a full time Database administrator job?
- Do not mention client name's in resume. If asked say that it's confidential which brings ahead qualities like honesty
- When you make your resume keep your recent projects at the top.
- Find out what the employer is looking for by asking him questions at the start of interview and best is before going to interview. Example if a company has projects on server products employer will be looking for BizTalk, CS CMS experts.
- Can you give brief about your family background?
- As you are fresher, do you think you can really do this job?
- Have you heard about our company? Say five points about our company? Just read at least once what company you are going for?
- Can you describe your best project you have worked with?
- Do you work on Saturday and Sunday?
- Which is the biggest team size you have worked with?
- Can you describe your current project you have worked with?
- How much time will you need to join our organization? What is notice period for your current company?
- What certifications have you cleared?
- Do you have passport size photos, last year mark sheet, previous companies employment letter, last months salary slip, passport and other necessary documents.
- What is the most important thing that motivates you?

- Why you want to leave the previous organization?
- Which type of job gives you greatest satisfaction?
- What is the type of environment you are looking for?
- Do you have experience in project management?
- Do you like to work as a team or as individual?
- Describe your best project manager you have worked with?
- Why should I hire you?
- Have you been ever fired or forced to resign?
- Can you explain some important points that you have learnt from your past project experiences?
- Have you gone through some unsuccessful projects, if yes can you explain why did the project fail?
- Will you be comfortable with location shift? If you have personal problems say no right at the first stage.... or else within two months you have to read my book again.
- Do you work during late nights? Best answer if there is project deadline yes. Do not show that it is your culture to work during nights.
- Any special achievements in your life till now...tell your best project which you have done best in your career.
- Any plans of opening your own software company...Beware do not start pouring your bill gate's dream to him...can create a wrong impression.

Chapter 1: Database Design

Overview

This is an interesting topic and yes it's the most discussed one when it comes to SQL Server interviews.



In SQL Server interviews database design conversation goes in to two wide discussions one is Normalization and the other is de-normalization.

So in normalization section interviewer can ask you questions around the 3 normal forms i.e. 1st normal form, second normal form and 3rd normal form. This looks to be a very simple question but you would be surprised to know even veteran database designers forget the definition thus giving an impression to the interviewer that they do not know database designing.

Irrespective you are senior or a junior everyone expects you to answer all the 3 normal forms. There are exceptions as well where interviewer has asked about 4th and 5th normal form as well but you can excuse those if you wish. I personally think that's too much to ask for.

When it comes to database designing technique interviewer can query the other side of the coin i.e. de-normalization. One of the important questions which interviewers can query is around Difference between de-normalization and normalization. The expectation from most of the interviewers when answering the differences is from the perspective of performance and type of application.

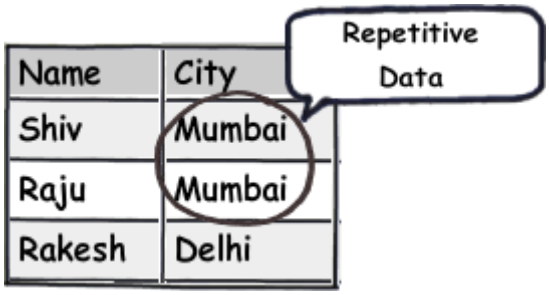
As people discuss ahead there is high possibility of getting in to OLTP and OLAP discussions which can further trigger discussions around database designing techniques Star and Snow flake schema.

What is normalization and what are the benefits of the same?

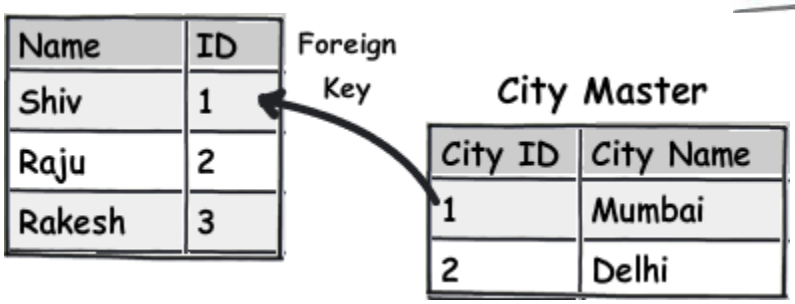
It's a *database design technique* to *avoid repetitive data and maintain integrity* of the data.

Note: - After that sweet one line, I can bet the interviewer will ask you to clarify more on those two words repetitive and integrity word.

Let's first start with repetitive. Let's say you have a simple table of user as shown below. You can see how the city is repeated again and again. So you would like to improve on this.

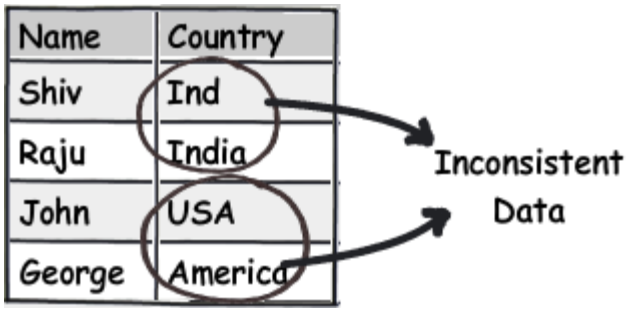


So to solve the problem, very simple you apply normalization. You split that repetitive data in to separate table (city master) and put a reference foreign key as shown in the below figure.



Now the second word “Data integrity”. “Data integrity” means how much accurate and consistent your data is.

For instance in the below figure you can see how the name of the country is inconsistent. “Ind” and “India” means the same thing, “USA” and “United States” means the same the thing. This kind of inconsistency leads to more complication and problems in maintenance.



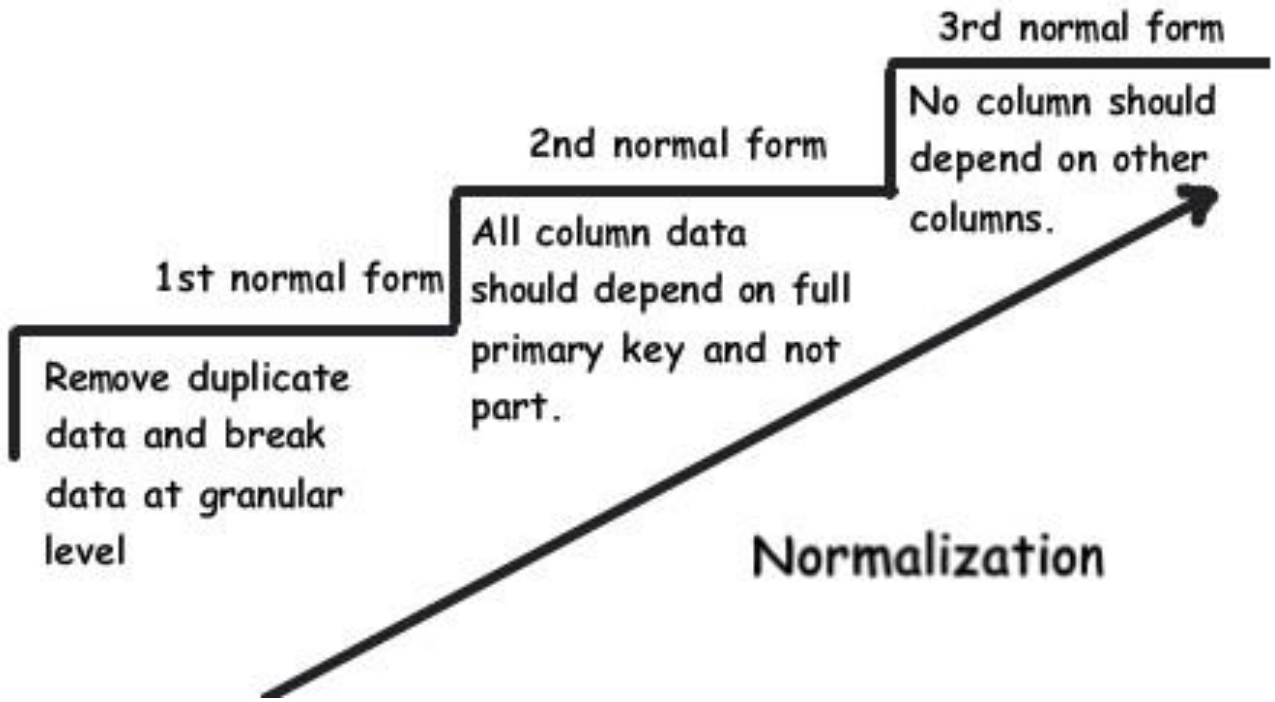
One of the most important thing in technical interviews like SQL , .NET , Java etc is that you need to use proper technical vocabulary. For example in the above answer the word “Data integrity” attracts the interviewer than the word “Inaccurate”. Right technical vocabulary will make you shine as compared to people who use plain English.

What is 1st normal form, second normal form and 3rd normal form?

Its surprising that many experienced professionals cannot answer this question. So below is simplified one liner’s for each of these normal forms.

- First normal form is all about breaking data in to smaller logical pieces.
- In Second normal form all column data should depend fully on the key and not partially.

- In Third normal form no column should depend on other columns.



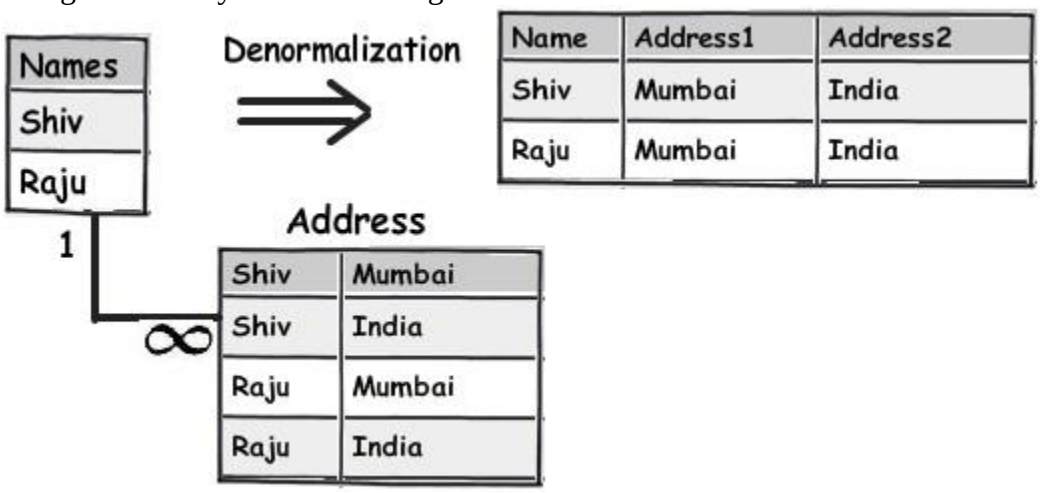
For in-depth explanation you can also see the video :- Can you explain First, Second & Third Normal forms in database? provided in the DVD.

What is denormalization?

Denormalization is exact opposite of normalization. Normalization is good when we want to ensure that we do not have duplicate and inconsistent data. In other words when we want to do operational activities like insert, update, delete and simple reports normalization design is perfectly suited. Putting in other words when data comes in to the system a normalized design would be the best suited.

But what if the we want to analyze historical data, do forecasting, do heavy calculations etc. For these kinds of requirements normalization design is not suited. Forecasting and analyzing are heavy operations and with historical data it becomes heavier. If your design is following normalization then your SQL needs to pull from different tables making the select process slow.

Normalization is all about reducing redundancy while denormalization is all about increasing redundancy and minimizing number of tables.



In the above figure you can see at the left hand side we have a normalized design while in the right hand side we have denormalized design. A query on the right side denormalized table will be faster as compared to the left hand side because there are more tables involved.

You will use denormalized design when your application is more meant to do reporting, forecasting and analyzing historical data where read performance is more important.

What is the difference between OLTP and OLAP system?

Both OLTP and OLAP are types of IT systems.OLTP (Online transaction processing system) deals with transactions (insert, update , delete and simple search) while OLAP (Online analytical processing) deals with analyzing historical data, forecasting etc.

Below is a simple table which chalks out the differences.

	OLTP	**OLAP**
Design	Normalized. (1st normal form, second normal form and third normal form).	Denormalized (Dimension and Fact design).
Source	Daily transactions.	OLTP.
Motive	Faster insert, updates, deletes and improve data quality by reducing redundancy.	Faster analysis and search by combining tables.
SQL complexity	Simple and Medium.	Highly complex due to analysis and forecasting.

For what kind of systems is normalization better as compared to denormalization?

Normalization is best suited for OLTP systems (faster transactions) while denormalizations are best suited for OLAP systems (faster queries and analysis).

What are Facts, Dimension and Measures tables?

The most important goal of OLAP application is analysis on data. The most important thing in any analysis are “NUMBERS”. So with OLAP application we would like to get those numbers , forecast them, analyze them for better business growth. These numbers can be total sales, number of customers etc.

These numbers are termed as “Measures” and measures are mostly stored in “Fact” tables.
“Dimension” describes what these measures actually mean. For example in the below table you can see we have two measures 3000 units and 1500 $. One dimension is “ProductWiseSales” and the other dimension is “AgewiseSalary’.

Dimensions are stored in dimension table.

Dimension	Measures
ProductWiseSales	3000 units
AgeWiseSalary	1500 $

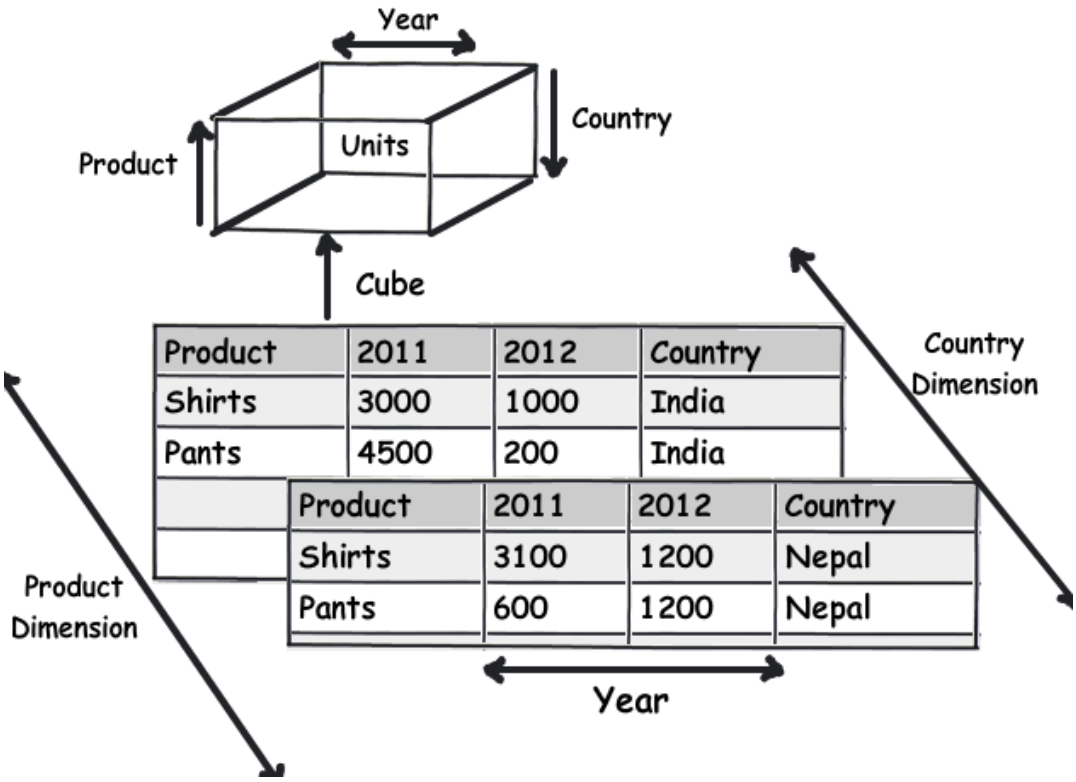
What are cubes?

A cube helps us to get a multi-dimensional view of a data by using dimension and measures. For instance in the below table we have “Units” as measure and we have 3 dimensions Product wise, Year wise and Country wise.

ProductName	Units	Year	Country
Shirts	3000	2011	India
Shirts	1000	2012	India
Pants	4500	2011	India
Pants	200	2012	India
Shirts	3100	2011	Nepal
Shirts	1200	2012	Nepal

Pants	600	2011	Nepal
Pants	1200	2012	Nepal

So if we change the dimensions and measures in a cube format we would get a better picture. In other words cube is intersection of multiple measures and their dimensions. If you visualize in a graphical model below image is how it looks like.



What is the difference between star schema and snow flake design?

Star schema consists of fact and dimension tables. The fact tables have the measures and dimension tables give more context to the fact tables.

In the below figure “Star design” you can see we have four dimension tables and each one of them are referencing the fact tables for measure values. The references between dimension and fact tables are done using simple foreign key relationships.

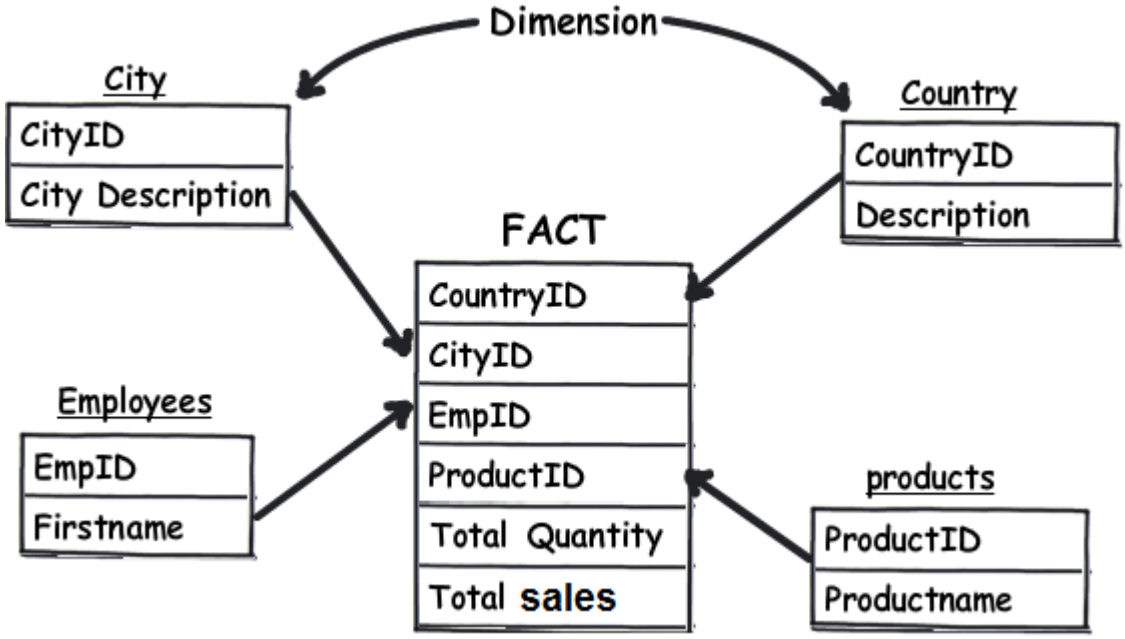


Figure: - Star design

Snow flake design is very much similar to star design. The exception is the dimension table. In snow flake dimension tables are normalized as shown in the below figure "Snow flake design". The below design is very much similar to the star design shown previously but the products table and vendor tables are separate tables.

The relationship is more of a normalized format. So summing in other words Star design is pure denormalized design while snow flake can have normalized dimension tables.

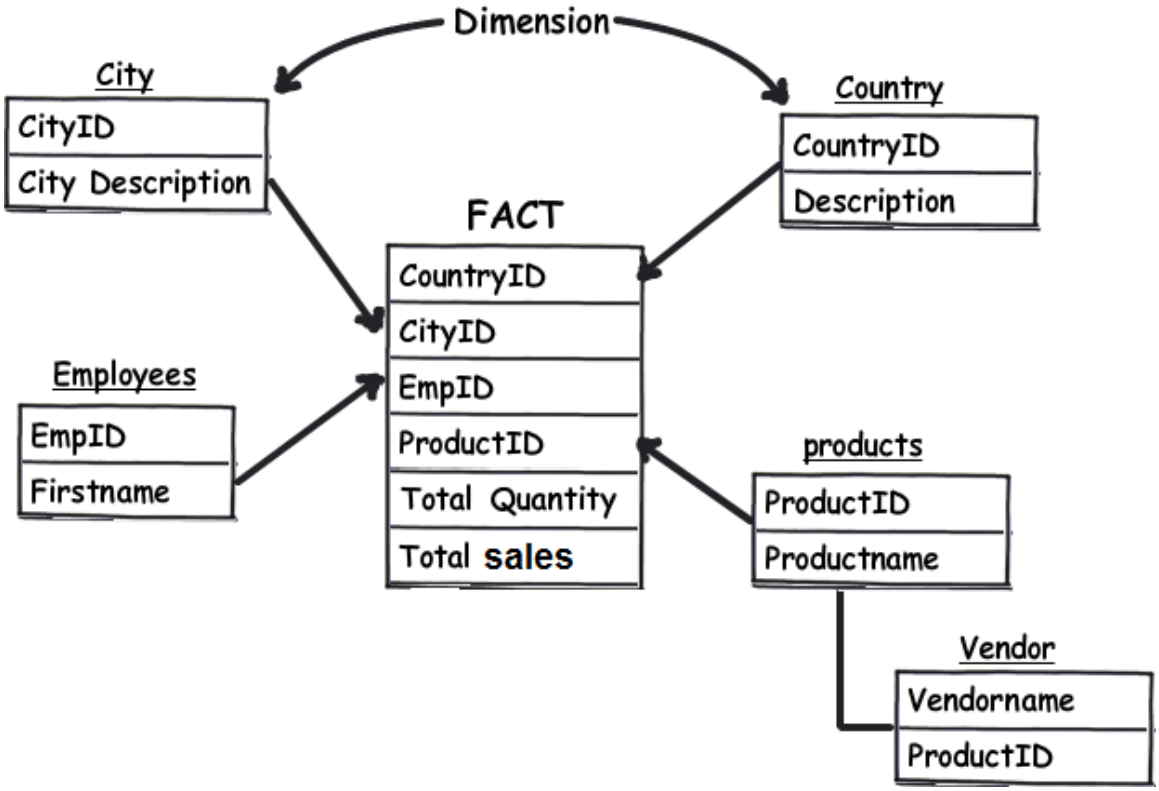


Figure: - Snow flake design

	Snowflake Schema	**Star Schema**
Normalization	Can have normalized dimension tables.	Pure denormalized dimension tables.
Maintenance	Less redundancy so less maintenance.	More redundancy due to denormalized format so more maintenance.
Query	Complex Queries due to normalized dimension tables.	Simple queries due to pure denormalized design.
Joins	More joins due to normalization.	Less joins.
Usage guidelines	If you are concerned about integrity and duplication.	More than data integrity speed and performance is concern here.

Chapter 2 :- Data types

How many bytes does “char” consume as compared to “nchar”?

“char” data types consumes 1 byte while “nchar” consumes 2 bytes. The reason is because “char” stores only ASCII characters while “nchar” can store UNICODE contents.

In case you are new to ASCII and UNICODE. ASCII accommodates 256 characters i.e.english letters ,punctuations , numbers etc.But if we want to store a Chinese character or some other language characters then it was difficult to the same with ASCII , that’s where UNICODE comes in to picture. An ASCII character needs only 1 byte to represent a character while UNICODE needs to 2 bytes.

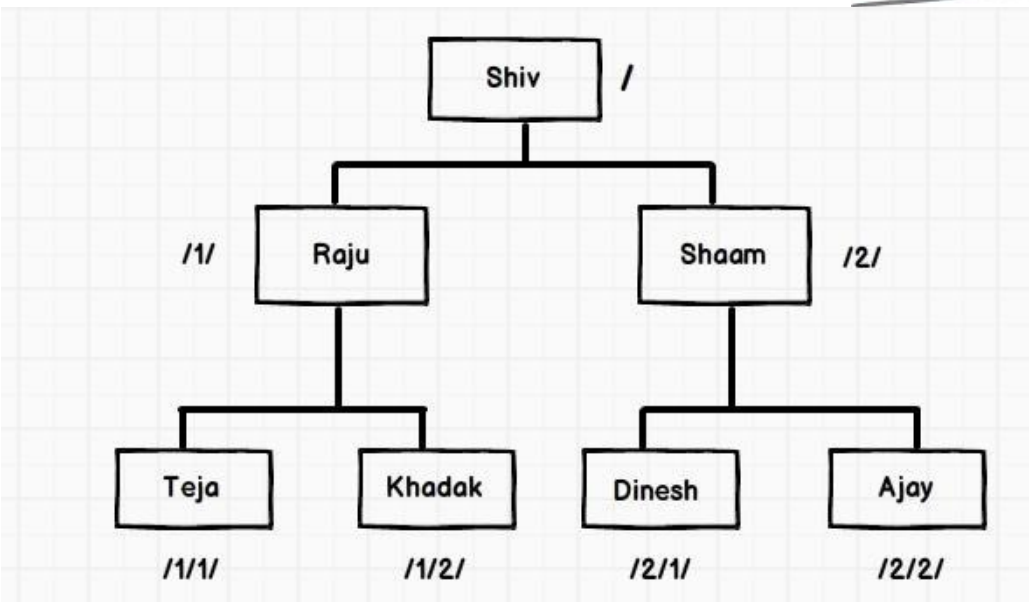
What is the difference between “char” and “varchar” data types?

“Char” is a fixed length data type. That means if you create a char of 10 length, it always consumes 10 bytes, irrespective you store 1 character or 10 character. While “varchar” is a variable length data type. That means if you create a “varchar” of 10 length, it will consume length equivalent to the number of characters. So if we store 3 characters, it will only consume 3 bytes.

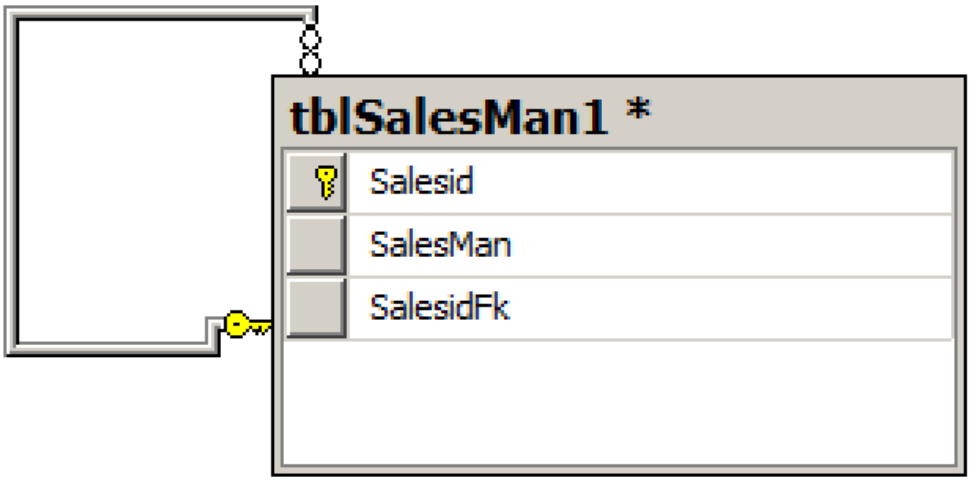
So if you have data like country code which always consumes 3 characters (USA, IND, NEP etc) “char” data type is a good choice. If you are not sure about the number of characters like “Name” of a person, “varchar” is a better fit.

What is the use of ‘hierarchyid’ data type in SQL Server?

Many times we need to store data which has a deep tree structure hierarchy. Below is a simple example which represents a sales team organization hierarchy. Now if you want to get all people who work under “Shaam” you will really need to work hard on both database design and logic.



In the past many developers used to create database design using self-reference primary and foreign key relationship. You can see in the below figure we have a simple table which stores sales person. Every sales person is marked with a primary key called as "SalesId". We have one more column called as "Salesid_fk" which references the primary key "SalesId".



Now to establish the tree structure hierarchy we need to enter data in a linked list format. For instance you can see in the below data entry snapshot. The first row "Shiv" indicates the top sales person in the hierarchy. Now because "Shiv" is the top sales person in the hierarchy the "SalesIdFk" value is null. "Raju" and "Shaam" report to "Shiv", so they

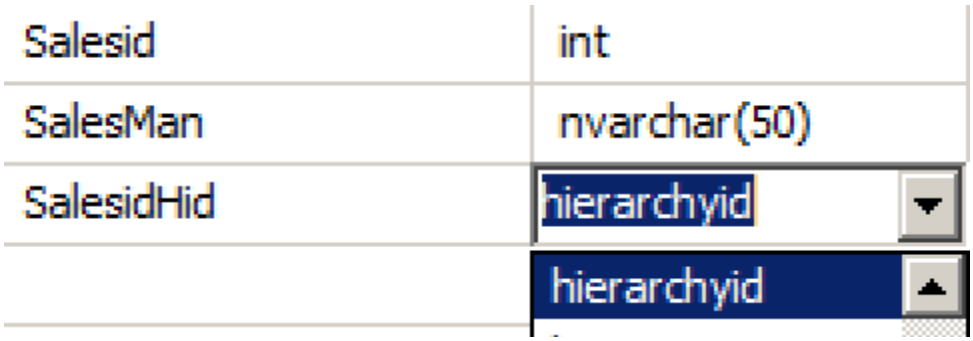
have “Salesidfk’ value as “1” which is nothing but primary key value of “Shiv”. Using this approach we can represent any deep hierarchy.

This approach is great but it needs cryptic DB design and some complicated logic to process data. For instance if I want to get how many people work under “Shaam” , I really need to write some complicated recursive logic at the backend.

So here’s a good news, SQL Server has support of hierarchy data type which can accommodate such complex tree structure.

Salesid	SalesMan	SalesidFk
1	Shiv	*NULL*
2	Raju	1
3	Shaam	1
4	Teja	2
5	Khadak	2
6	Dinesh	3
7	Ajay	3

So first step is to get rid of “SalesIdfk” column and add “SalesHid” column which is of datatype “Hierarchyid”.



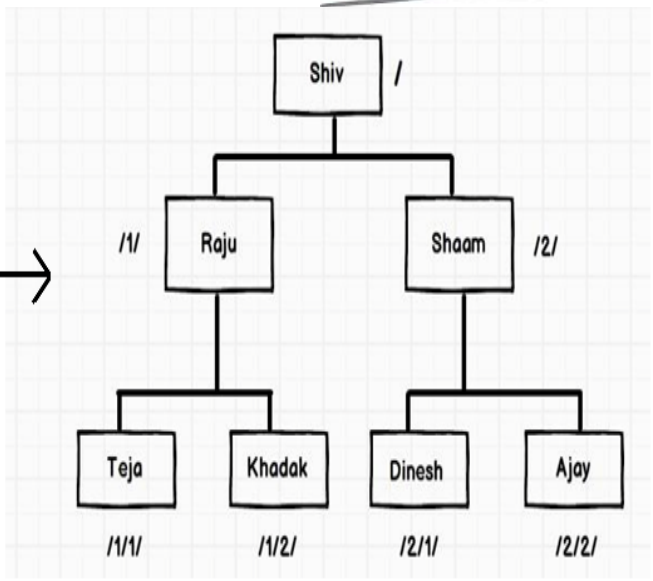
Now the HID (Hierarchy id) data type column uses the below format to represent tree structure data:-

- All data in the HID column should start with “/” and end with “/”.
- The top level root is represented by “/”.
- The next level below the root is represented by “/1/”. If you have one more person on the same level you will enter it has “/2/”.
- If you want to enter one more child below “/1/”, you need to enter “/1/1/”.

Salesid	SalesMan	(No column name)
1	Shiv	/
2	Raju	/1/
3	Shaam	/2/
4	Teja	/1/1/
5	Khadak	/1/2/
6	Dinesh	/2/1/
7	Ajay	/2/2/

Below is a pictorial representation of how HID values map with the tree structure levels.

Salesid	SalesMan	(No column name)
1	Shiv	/
2	Raju	/1/
3	Shaam	/2/
4	Teja	/1/1/
5	Khadak	/1/2/
6	Dinesh	/2/1/
7	Ajay	/2/2/

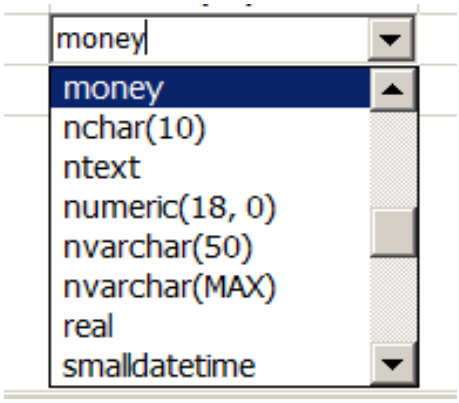


Now if you want to search who are below “Shaam”. You can fire the below query. The “IsDescendantOf” function evaluates to true if the records are child’s of that level. This is easy as compared to creating a custom design and writing logic yourself.

```
SELECT TOP 1000 [Salesid]
      ,[SalesMan]
      ,[SalesidHid].ToString()
  FROM [PracticeSQl2012].[dbo].[tblSalesMan]
where SalesidHid.IsDescendantOf('/2/') = 1
```

If you wish to store financial values which SQL Server data type is more suitable ?

Money data type is the most suited data type for storing financial values as they are accurate ten-thousandth of the unit they represent. In financial figures accuracy matters. Small cents add up to millions later.



Further money data type has two variations “money” and “smallmoney”. Money has 8 bytes of storage while small money has 4 bytes of storage.

Chapter 3 :- SQL Queries

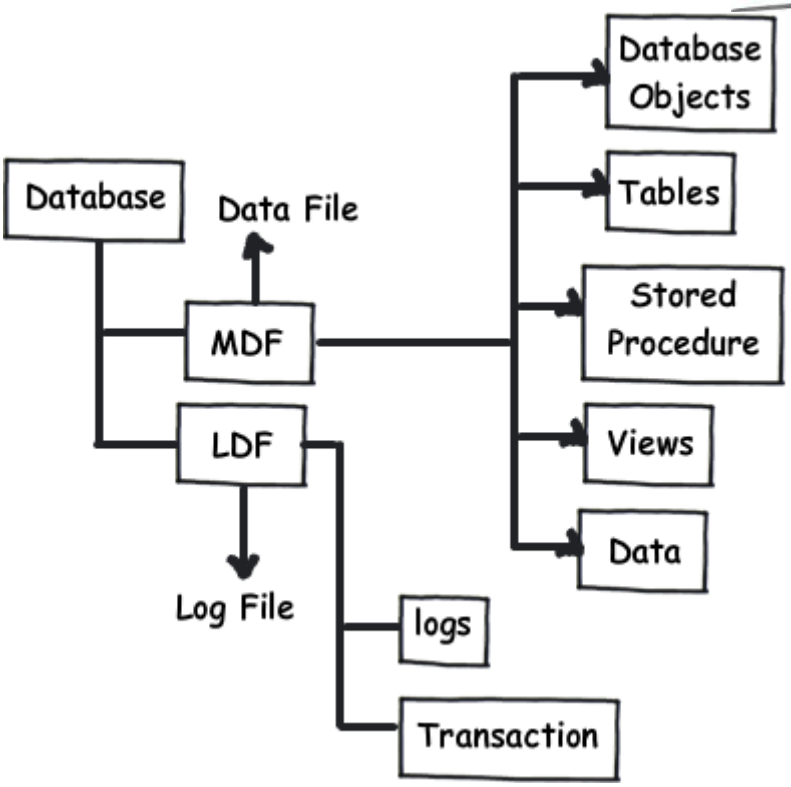
Chapter 4:- Page, Extent and Splits

There are 2 physical files MDF and LDF, what are they?

Both these files are found in “C:\Program Files\Microsoft SQL Server\MSSQL**X**.MSSQLSERVER\MSSQL\DATA”. Please note to replace X with 10 for 2008 and 11 for 2012.

MDF file is the primary database file where the actual user data and schema gets stored. In other words table data, table structure, stored procedures, views and all SQL Server objects gets stored in to this file.

LDF file is a transaction log file to store transaction logs. They do not have the actual user data or schema definition. Every MDF will have associated LDF. While recovering the database it’s always advisable to keep the LDF so that you do not lose any transaction information.

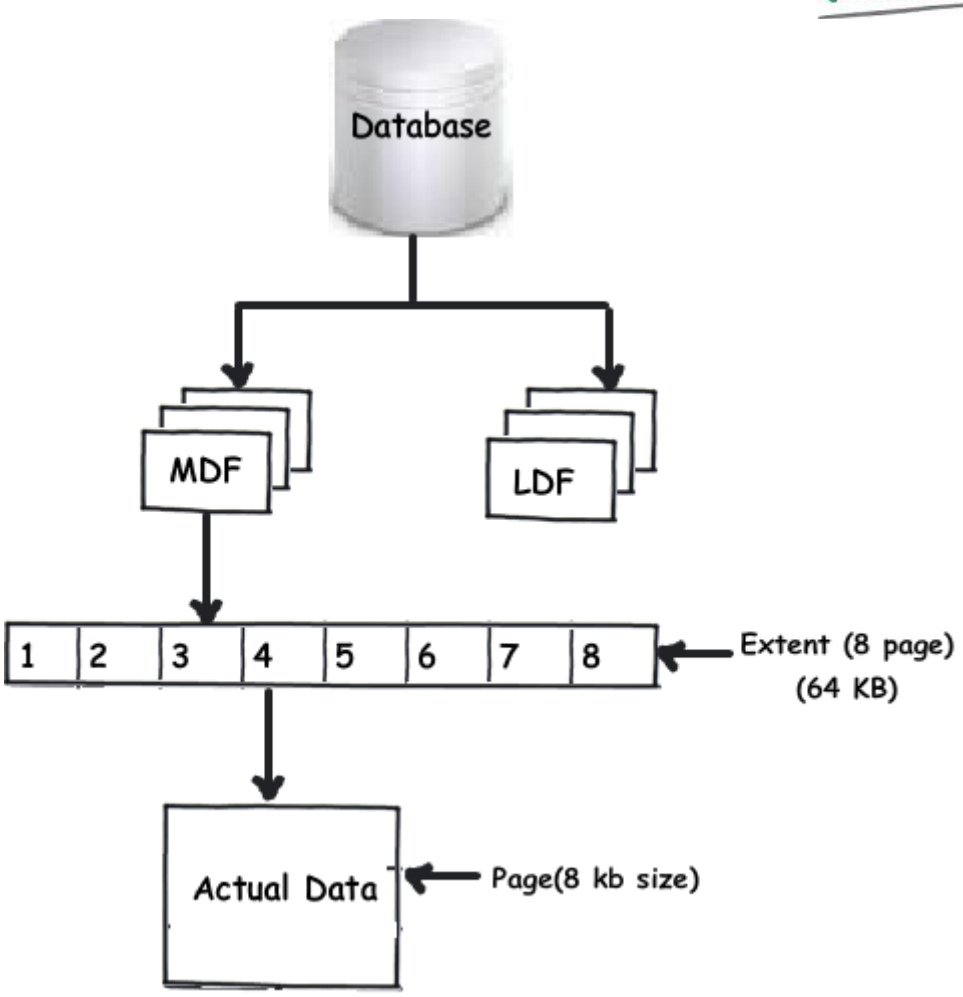


What are page and extents in SQL Server?

Extent and pages define how data is stored in SQL Server. The actual data is stored in pages. Each one of these pages are of 8 KB size. You can also visualize pages as the fundamental unit to store data. In other words your table row data gets stored actually in pages.

Pages are further grouped in to extent. One extent is collection of eight pages. So the size of each extent is 64 KB (i.e. 8 pages x 8 KB size).

Finally these extents sum up in to the MDF physical file which you see on your hard disk.

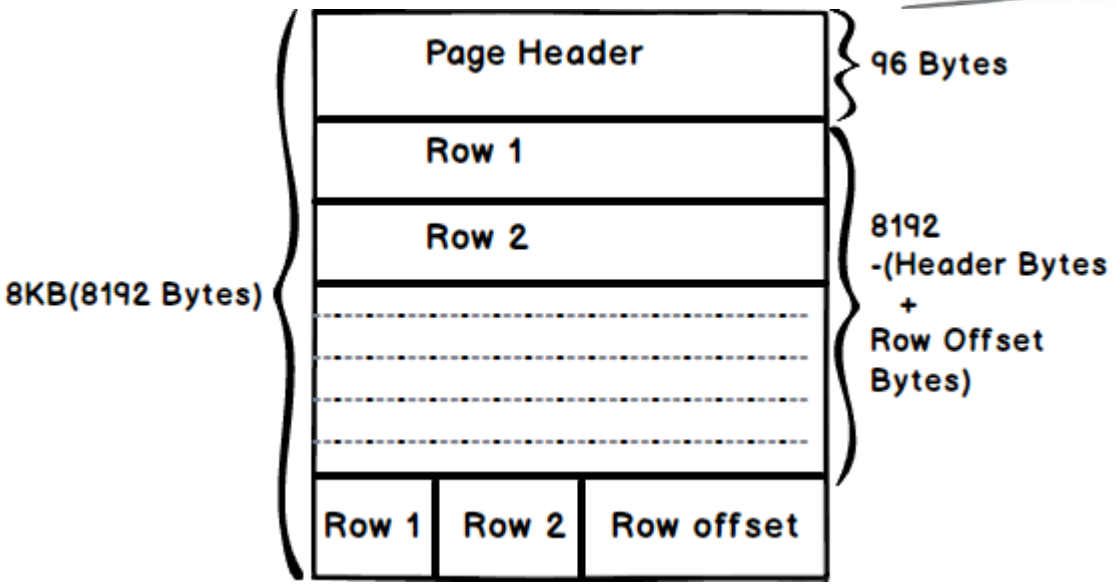


What are heap tables ?

Heap tables are those tables which do not have clustered indexes.

How does SQL Server actually store data internally?

SQL Server stores data in 8 KB pages. This 8 KB page is further divided in to 3 sections Page header, Data row and row offset, see the below figure for visual's.



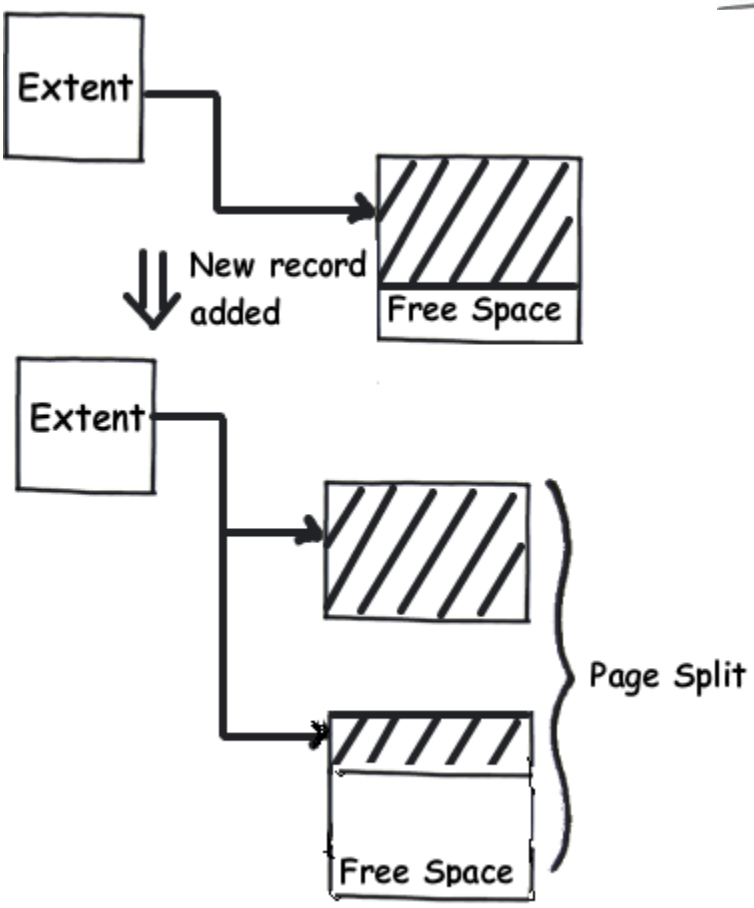
1. Page header stores information about the page like page type, next and previous page if it's an index page, free space in the page etc.
2. After the page header data row section follows. This is where you data is actually stored.
3. Row offset information is stored at the end of the page i.e. after the data row section. Every data row has a row offset and the size of row offset is 2 bytes per row. Row offset stores information about how far the row is from the start of the page.

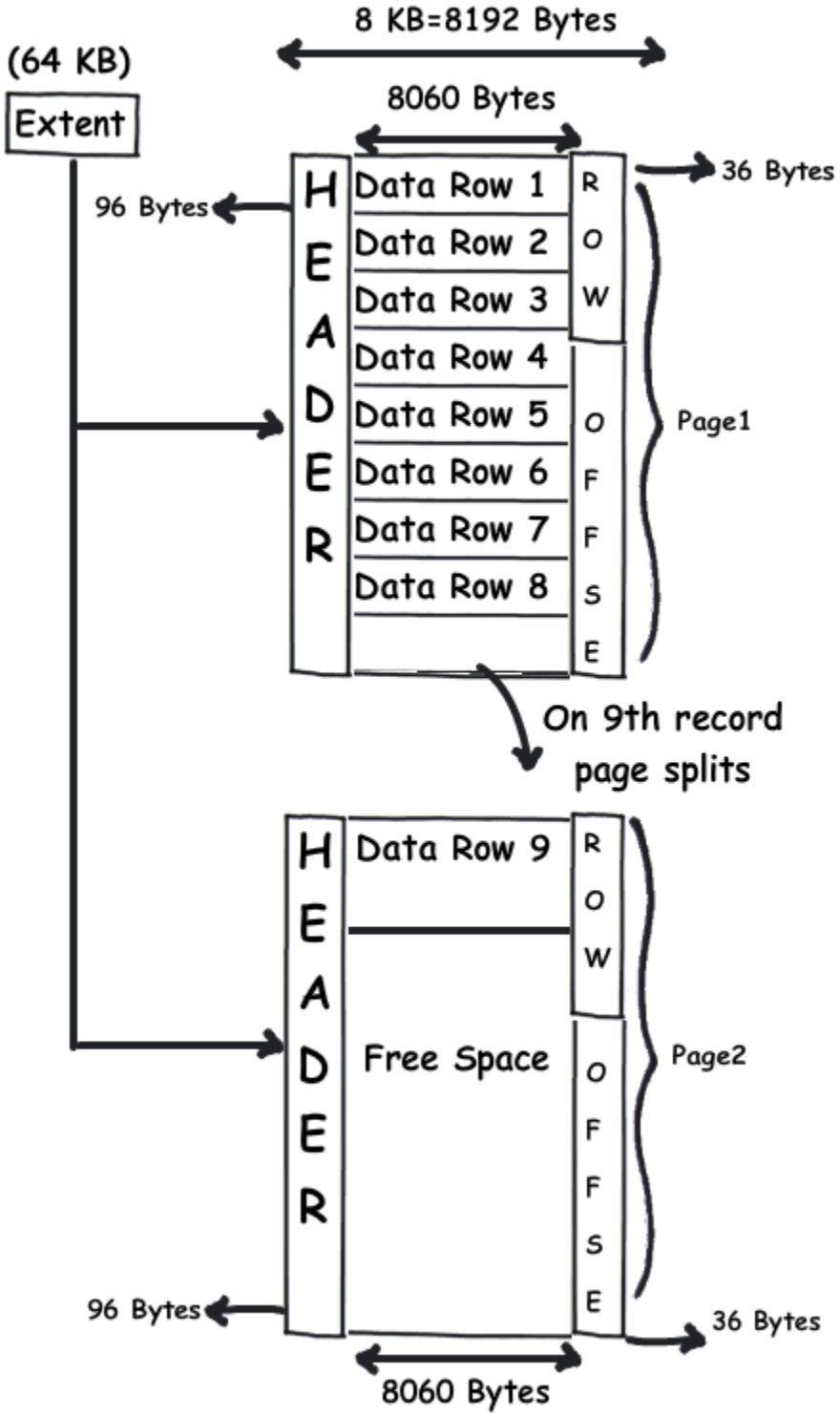
Putting in simple words the complete page equation comes as shown below.

Page (8 KB/8192 bytes) = Page header (96 bytes) + Actual data (Whatever bytes) + Row offset (2 bytes per row).

What is page split?

The actual data (table row data) is stored in pages. Pages are of size 8 KB. So on a page when data exceeds 8 KB SQL Server creates a new page to accommodate this data. This phenomenon is termed as Page split.


Extent
New record
added
Free Space
Extent
Page Split
Free Space




```
SELECT in_row_data_page_count FROM
sys.dm_db_partition_stats
```

```
WHERE object_id = OBJECT_ID('dbo.table1');
```

Chapter 3:- Indexes (Clustered and Non-Clustered)

Why do we need Indexes?

Indexes makes your search faster.

How does index makes your search faster?

When you create an index on a column it creates a "B-Tree" (Balanced-Tree) structure for the column which results in faster searches.

Its very important to answer to the point , short and sweet. For example in the above question many people start explaining the b-tree search which becomes lengthy and boring for the interviewer. See the interviewer's interest as well.

How does Balance tree make your search faster?

When you search data which are not indexed it searches sequentially. For instance in the below figure you can see if you want to search "50" the search engine has to browse through all the records sequentially until he gets "50".

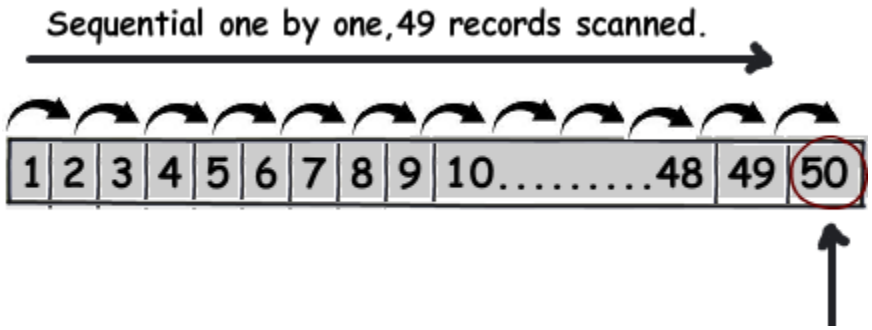


Figure: - Sequential

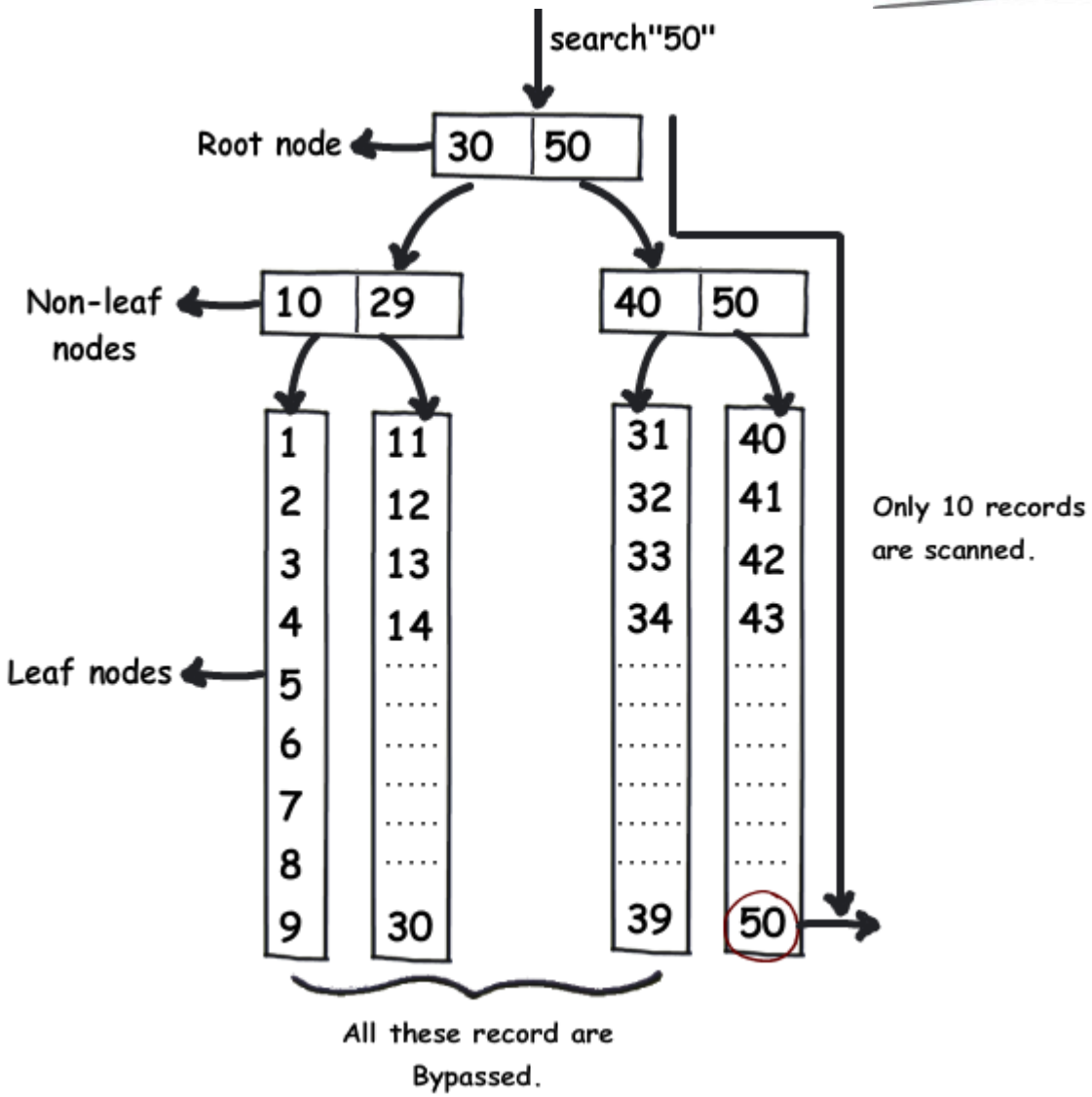


Figure: - B-tree structure

In B-tree data is divided in to root node, non-leaf node and leaf node. With this structure in place if you want to search "50" the following steps happen:-

- It first searches the root node and compares with the first node i.e. 30. It first checks if 50 is equal or less than 30, it's not. So it by passed the complete 30 root node and proceeds to 50 root node.
- It then compares the second root node i.e. 50. Now 50 is equal to 50 so it moves ahead to non-leaf nodes of 50 i.e. 40 and 50.
- In the non-leaf node again the comparison is done, it bypasses the complete 40 node and follow's the non-leaf node 50.

- Finally it sequentially travels from 40 to 50 values to find the exact match.

If you see the steps the closely it has only scanned 10 records as comparison to the sequential scan where it scans 49 records.

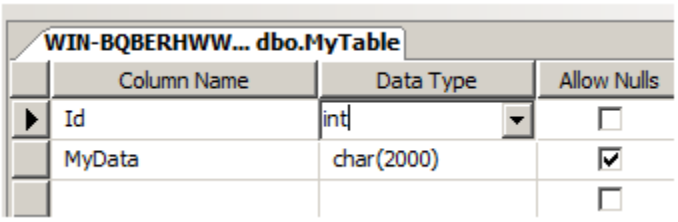
Note :- If you are serious about the job do not shy away from drawing diagrams to communicate better.

What are page splits in indexes ?

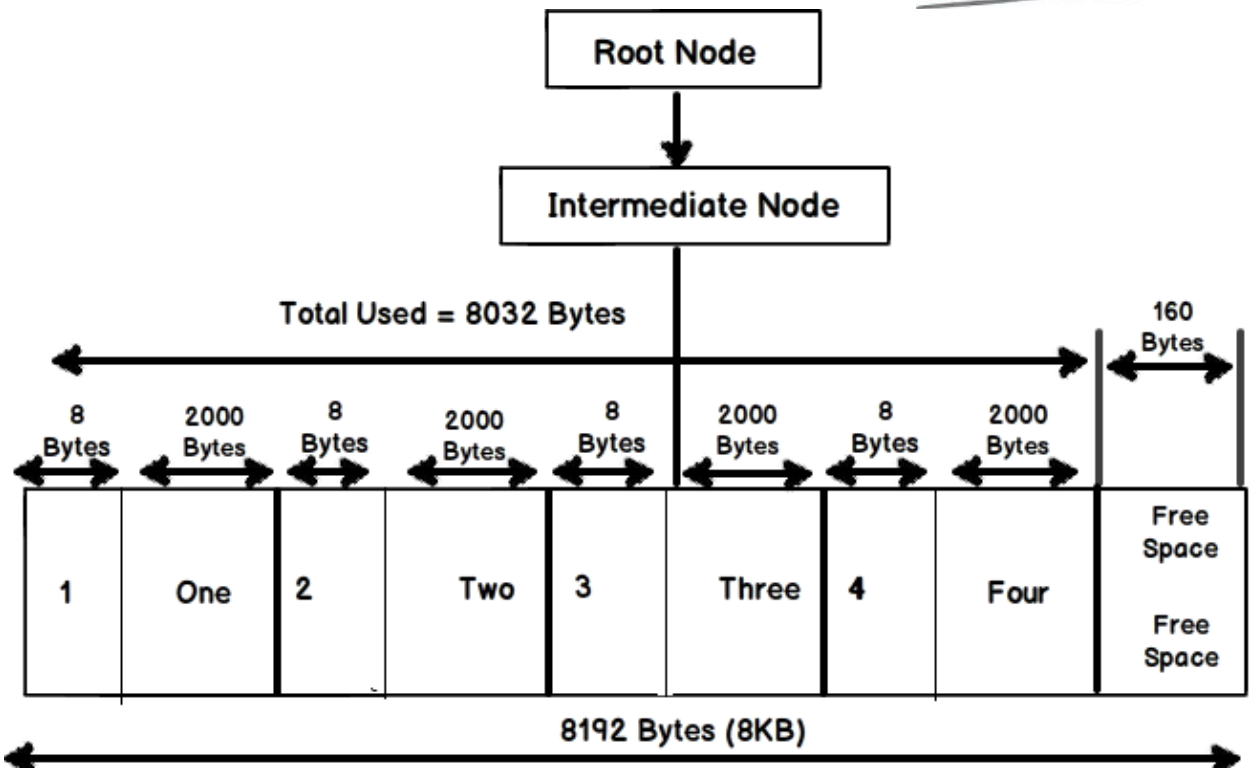
Indexes are organized in B-Tree structure divided in to root nodes, intermediate nodes and leaf nodes. The leaf node of the B-tree actually contains data. The leaf index node is of 8 KB size i.e. 8192 bytes. So if data exceed over 8 KB size it has to create new 8 KB pages to fit in data. This creation of new page for accommodating new data is termed as page split.

Let me explain you page split in more depth. Let's consider you have a simple table with two fields "Id" and "MyData" with data type as "int" and "char(2000)" respectively as shown in the below figure. "Id" column is clustered indexed.

That means each row is of size 2008 bytes (2000 bytes for "MyData" and 8 bytes for "Id").

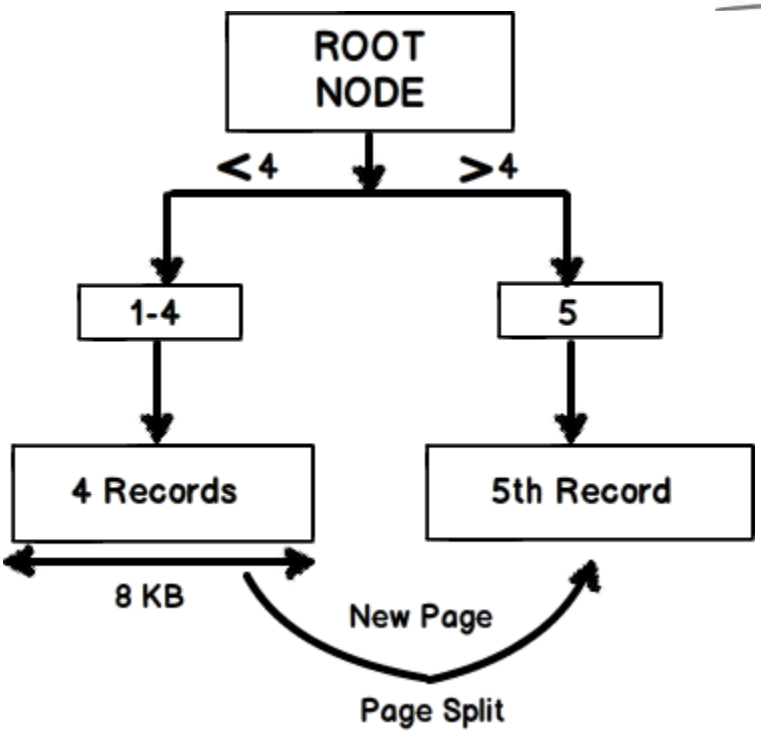


Id	MyData
1	One ...
2	two ...
3	three ...
4	four ...



So if we have four records the total size will be 8032 bytes (2008 * 4) that leaves 160 bytes free. Do look at the above image for visual representation.

So if one more new record is added there is no place left to accommodate the new record and the index page is forced to go for an index page split.



So does page split affect performance?

If your database is highly insert intensive it can lead to large number of page splits. In order to perform those page splits your CPU has to drain some power which can lead to performance issues.

So how do we overcome the page split performance issue?

Page split performance problem can be overcome by assigning a proper "**FILL FACTOR**".

What exactly is fill factor?

What are "Table Scan's" and "Index Scan's"?

These are ways by which SQL Server searches a record or data in table. In "Table Scan" SQL Server loops through all the records to get to the destination. For instance if you have 1, 2, 5, 23,

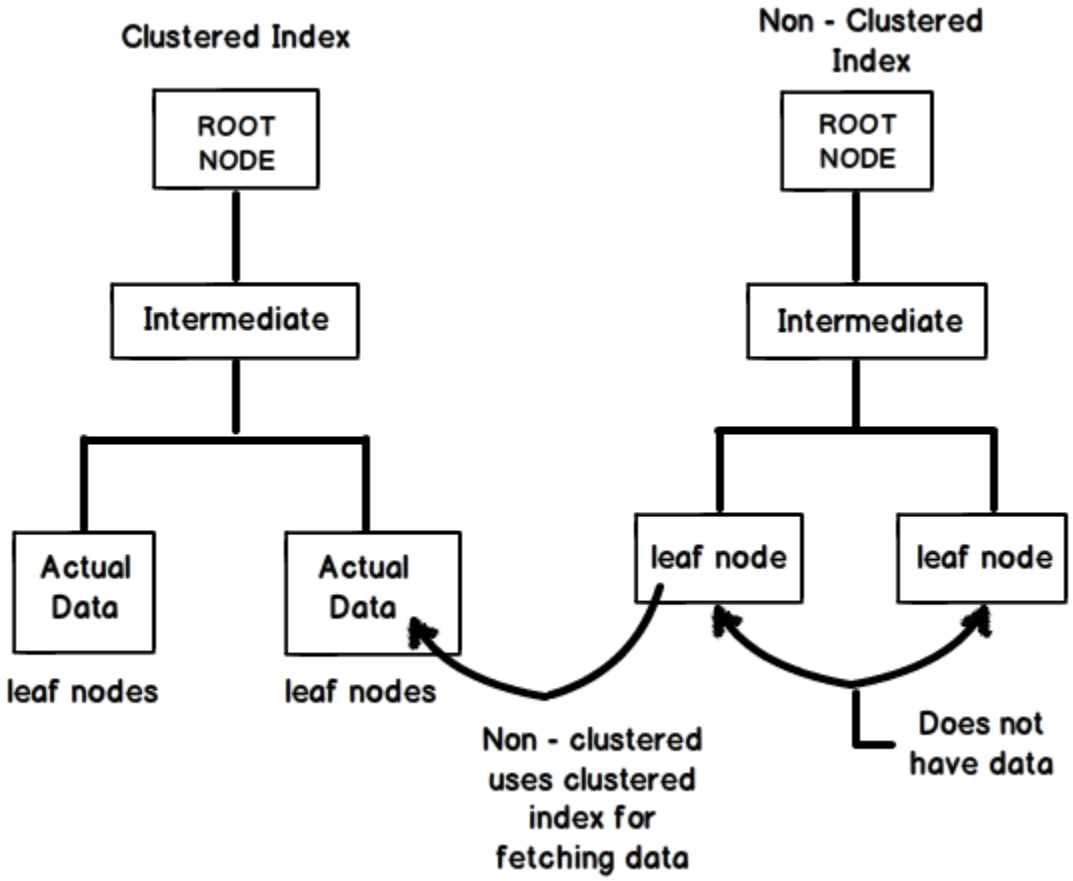
63 and 95. If you want to search for 23 it will go through 1, 2 and 5 to reach it. Worst if it wants to search 95 it will loop through all the records.

While for "Index Scan's" it uses the "B-TREE" fundamental to get to a record. For "B-TREE", refer previous questions.

Note: - Which way to search is chosen by SQL Server engine. Example if it finds that the table records are very less it will go for table scan. If it finds the table is huge it will go for index scan.

(Q) What are the two types of indexes and explain them in detail?

Twist: - What is the difference between clustered and non-clustered indexes?



There are basically two types of indexes:-

- Clustered Indexes.
- Non-Clustered Indexes.

Ok every thing is same for both the indexes i.e. it uses "B-TREE" for searching data. However, the main difference is the way it stores physical data. If you remember the previous figure (give

figure number here), there where leaf level and non-leaf level. Leaf level holds the key, which is used to identify the record. Moreover, non-leaf level actually point to the leaf level.

In clustered index, the non-leaf level actually points to the actual data.

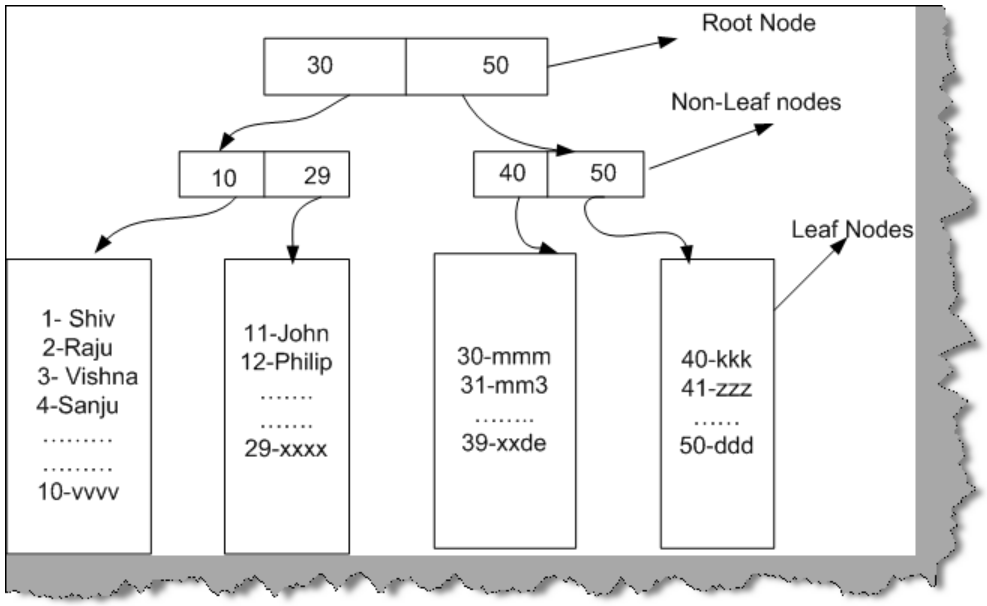


Figure 12.3: - Clustered Index Architecture

In Non-Clustered index the leaf nodes point to pointers (they are rowid's) which then point to actual data.

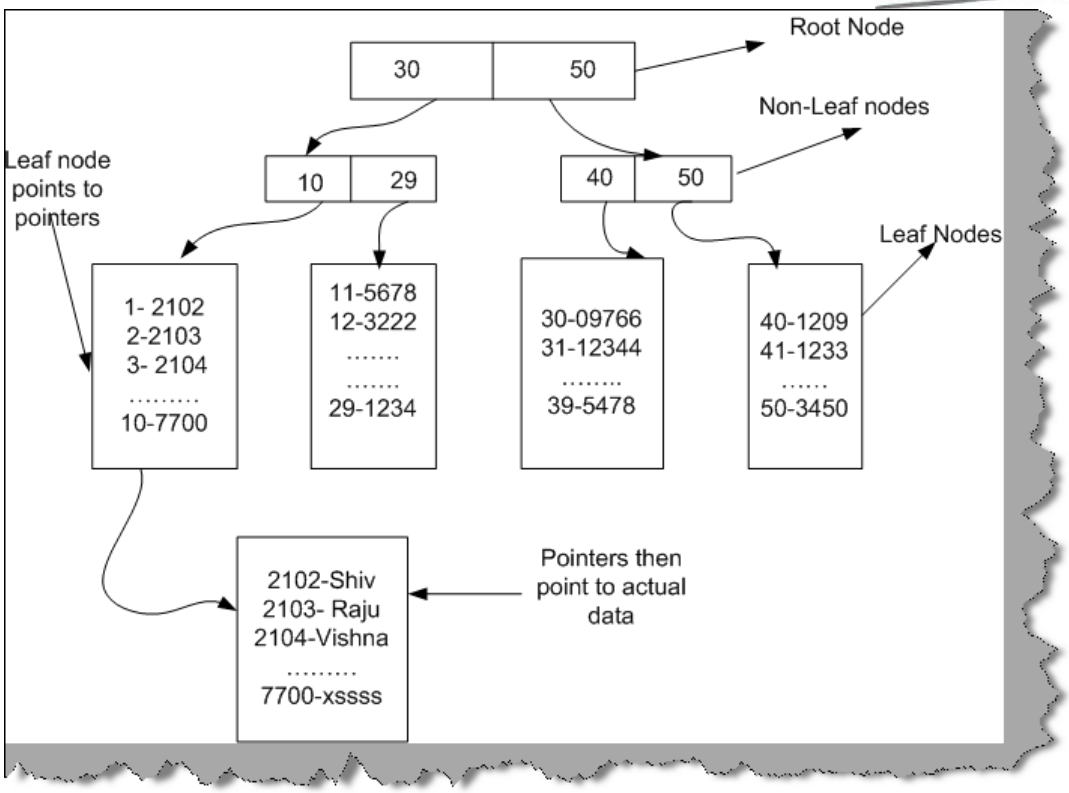


Figure 12.4: - Non-Clustered Index has pointers.

So here is what the main difference is in clustered and non-clustered, in clustered when we reach the leaf nodes we are on the actual data. In non-clustered indexes, we get a pointer, which then points to the actual data.

Therefore, after the above fundamentals following are the basic differences between them:-

- Also, note in clustered index actual data as to be sorted in same way as the clustered indexes are. While in non-clustered indexes as we have pointers, which is logical arrangement we do need this compulsion.
- So we can have only one clustered index on a table as we can have only one physical order while we can have more than one non-clustered indexes.

If we make non-clustered index on a table, which has clustered indexes, how does the architecture change?

The only change is that the leaf node point to clustered index key. Using this clustered index key can then be used to finally locate the actual data. So the difference is that leaf node has pointers while in the next half it has clustered keys. So if we create non-clustered index on a table which has clustered index it tries to use the clustered index.

(DB) What is “FillFactor” concept in indexes?

When SQL Server creates new indexes, the pages are by default full. “FillFactor” is a percentage value (from 1 – 100) which says how much full your pages will be. By default “FillFactor” value is zero.

(DB) What is the best value for “FillFactor”?

“FillFactor” depends on how transactional your database is. Example if your database is highly transactional (i.e. heavy insert’s are happening on the table), then keep the fill factor less around 70. If it is only a read-only database probably used only for reports, you specify 100%.

Remember there is a page split when the page is full. Therefore, fill factor will play an important role.

(Q) What are “Index statistics”?

Statistics are something the query optimizer will use to decide what type of index (table scan or index scan) to be used to search data. Statistics change according to inserts and updates on the table, nature of data on the table etc...In short, “Index statistics” are not same in all situations. So DBA has to run statistics again and again after certain interval to ensure that the statistics are up-to-date with the current data.

Note: - If you want to create index you can use either the “Create Index” statement or you can use the GUI.

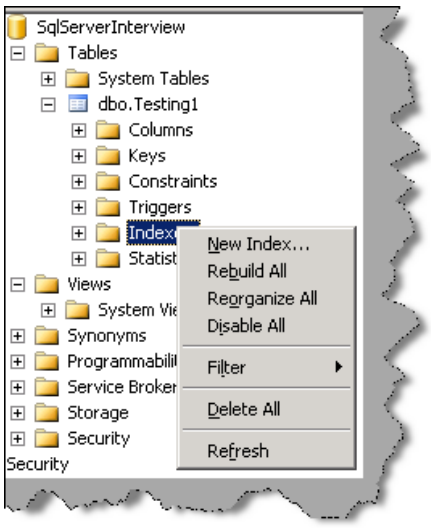


Figure 12.5: - Index creation in Action.

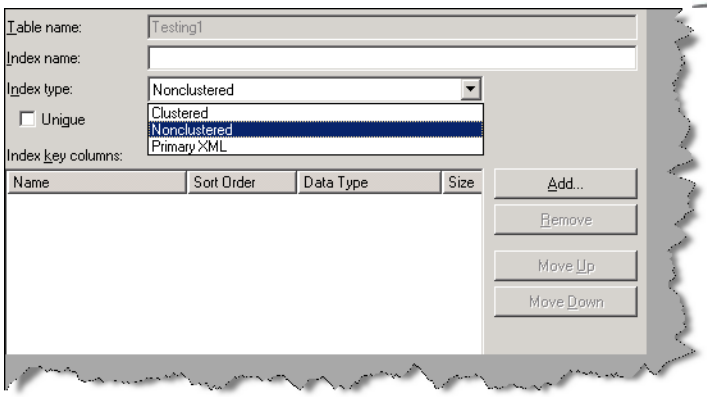


Figure 12.6: - Index Details

(DB) How can we see statistics of an index?

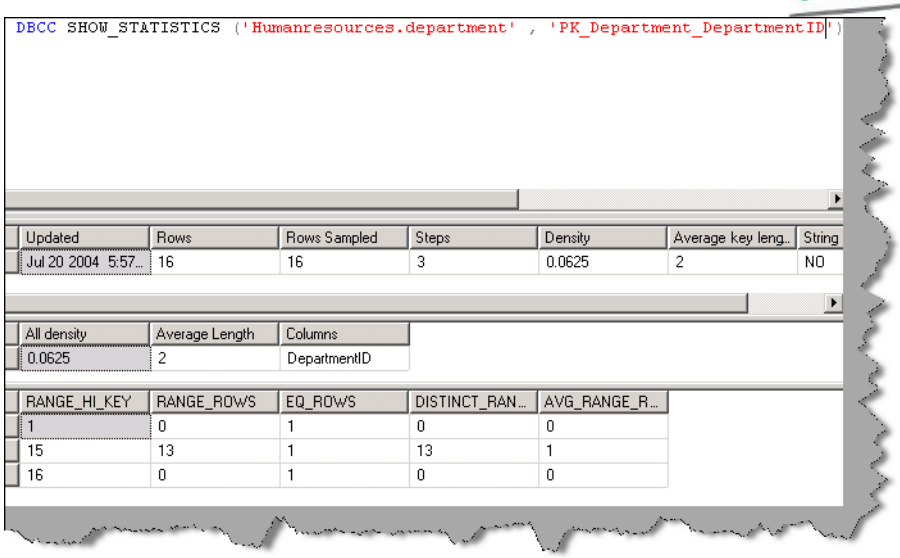
Twist: - How can we measure health of index?

In order to see statistics of any index following the T-SQL command you will need to run.

Note: - Before reading this you should have all the answers of the pervious section clear. Especially about extent, pages and indexes.

```
DECLARE
@ID int,
@IndexID int,
@IndexName varchar(128)
-- input your table and index name
SELECT @IndexName = 'AK_Department_Name'
SET @ID = OBJECT_ID('HumanResources.Department')
SELECT @IndexID = IndID
FROM sysindexes
WHERE id = @ID AND name = @IndexName
--run the DBCC command
DBCC SHOWCONTIG (@id, @IndexID)
```

Just a short note here "DBCC" i.e. "Database consistency checker" is used for checking heath of lot of entities in SQL Server. Now here we will be using it to see index health. After the command is run you will see the following output. You can also run "DBCC SHOWSTATISTICS" to see when was the last time the indexes rebuild.



```
DBCC SHOW_STATISTICS ('Humanresources.department' , 'PK_Department_DepartmentID')
```

Updated	Rows	Rows Sampled	Steps	Density	Average key leng...	String
Jul 20 2004 5:57...	16	16	3	0.0625	2	NO

All density	Average Length	Columns
0.0625	2	DepartmentID

RANGE_HI_KEY	RANGE_ROWS	EQ_ROWS	DISTINCT_RAN...	AVG_RANGE_R...
1	0	1	0	0
15	13	1	13	1
16	0	1	0	0

Figure 12.7 DBCC SHOWSTATISTICS

```
DBCC SHOWCONTIG scanning 'Department' table...
Table: 'Department' (613577224); index ID: 3, database ID: 5
LEAF level scan performed.
- Pages Scanned.................................: 1
- Extents Scanned...............................: 0
- Extent Switches...............................: 0
- Avg. Pages per Extent.........................: 0.0
- Scan Density [Best Count:Actual Count].......: 0.00% [0:0]
- Logical Scan Fragmentation ...................: 0.00%
- Extent Scan Fragmentation ....................: 0.00%
- Avg. Bytes Free per Page......................: 0.0
- Avg. Page Density (full)......................: 7.44%
DBCC execution completed. If DBCC printed error messages, contact your system administrator.
```

Figure 12.8: - DBCC SHOWCONTIG.

Pages Scanned

The number of pages in the table (for a clustered index) or index.

Extents Scanned

The number of extents in the table or index. If you remember we had said in first instance, that extent has pages. More extents for the same number of pages the higher will be the fragmentation.

Extent Switches

The number of times SQL Server moves from one extent to another. More the switches it has to make for the same amount of pages, the more fragmented it is.

Avg. Pages per Extent

The average number of pages per extent. There are eight pages / extent so if you have an extent full with the eight you are in a better position.

Scan Density [Best Count: Actual Count]

This is the percentage ratio of Best count / Actual count. Best count is number of extent changes when everything is perfect. It is like a baseline. Actual count is the actual number of extent changes on that type of scenario.

Logical Scan Fragmentation

Percentage of out-of-order pages returned from scanning the leaf pages of an index. An out of order page is one for which the next page indicated is different page than the page pointed to by the next page pointer in the leaf page. .

Extent Scan Fragmentation

This one is telling us whether an extent is not physically located next to the extent that it is logically located next to. This just means that the leaf pages of your index are not physically in order (though they still can be logically), and just what percentage of the extents this problem pertains to.

Avg. Bytes free per page

This figure tells how many bytes are free per page. If it's a table with heavy inserts or highly transactional then more free space per page is desirable, so that it will have less page splits.

If it's just a reporting system then having this closer to zero is good as SQL Server can then read data with less number of pages.

Avg. Page density (full)

Average page density (as a percentage). It's is nothing but:-

1 - (Avg. Bytes free per page / 8096)

8096 = one page is equal to 8096 bytes

```
Note: - Read every of the above sections carefully, incase you are
looking for DBA job you will need the above fundamentals to be very
clear. Normally interviewer will try to shoot questions like "If you
see the fill factor is this much, what will you conclude?  , If you see
the scan density this much what will you conclude?
```

(DB) How do you reorganize your index, once you find the problem?

You can reorganize your index using "DBCC DBREINDEX". You can either request a particular index to be re-organized or just re-index the all indexes of the table.

This will re-index your all indexes belonging to "HumanResources.Department".

```
DBCC DBREINDEX ([HumanResources.Department])
```

This will re-index only "AK_Department_Name".

```
DBCC DBREINDEX ([HumanResources.Department],[AK_Department_Name])
```

This will re-index with a "fill factor".

```
DBCC DBREINDEX ([HumanResources.Department],[AK_Department_Name],70)
```

You can then again run DBCC SHOWCONTIG to see the results.

(Q) What is Fragmentation?

Speed issues occur because of two major things

- Fragmentation.
- Splits.

Splits have been covered in the first questions. But one other big issue is fragmentation. When database grows it will lead to splits, but what happens when you delete something from the database…HeHeHe life has lot of turns right. Ok let's say you have two extents and each have two pages with some data. Below is a graphical representation. Well actually that's now how things are inside but for sake of clarity lot of things have been removed.

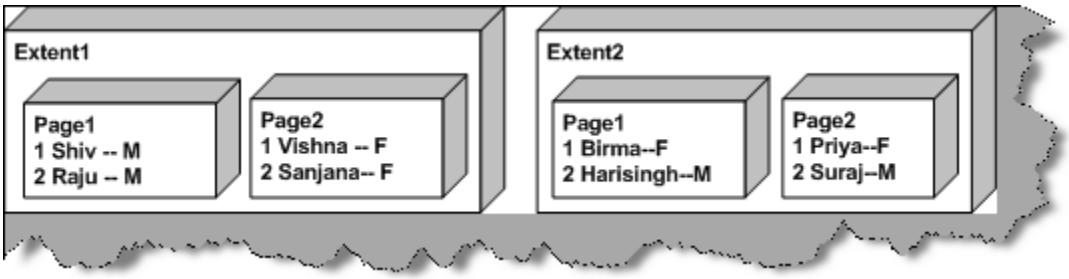


Figure 12.9: - Data Distribution in Initial Stages

Now over a period of time some Extent and Pages data undergo some delete. Here is the modified database scenario. Now one observation you can see is that some page's are not removed even when they do not have data. Second If SQL server wants to fetch all "Females" it has to span across to two extent and multiple pages within them. This is called as "Fragmentation" i.e. to fetch data you span across lot of pages and extents. This is also termed as "Scattered Data".

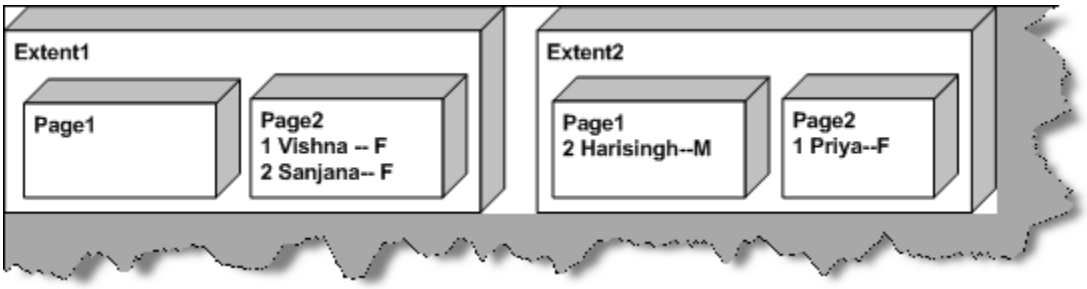


Figure 12.10: - Data Distribution after Deletes

What if the fragmentation is removed, you only have to search in two extent and two pages. Definitely, this will be faster as we are spanning across less entities

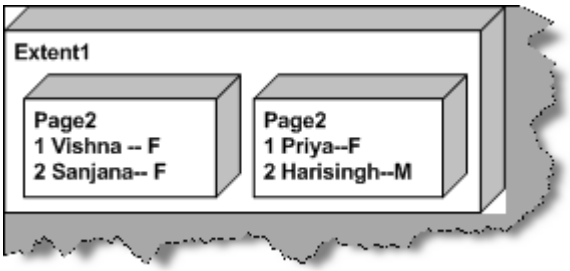


Figure 12.11: - Fragmentation removed

(DB) How can we measure Fragmentation?

Using "DBCC SHOWCONTIG ".

(DB) How can we remove the Fragmented spaces?

- **Update Statistics :-** The most used way by DBA's
- **Sp_updatestats.** :- It's same as update statistics , but update statistics applies only for specified object and indexes , while "sp_updatestats" loops through all tables and applies statistics updates to each and every table. Below is a sample, which is run on "AdventureWorks" database.

Note: - "AdventureWorks" is a sample database which is shipped with SQL Server 2005..

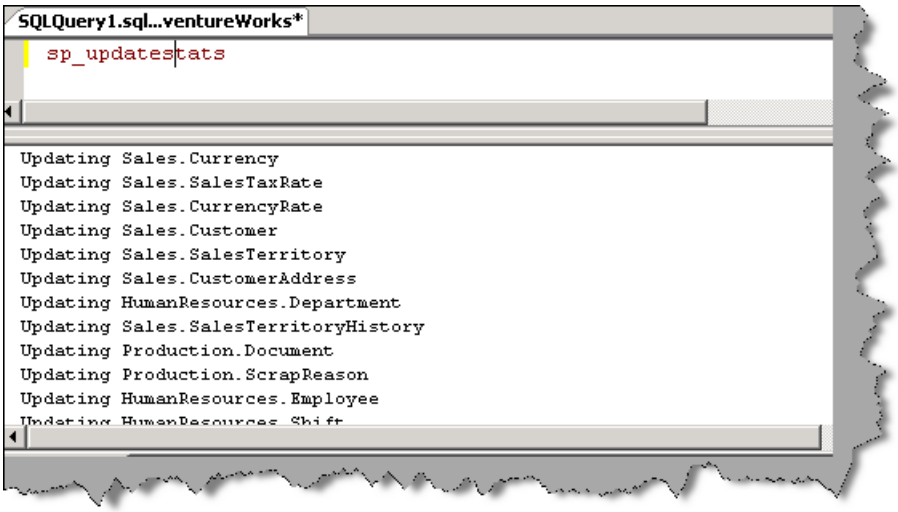


Figure 12.12: - sp_updatestats in action

- **DBCC INDEXFRAG**: - This is not the effective way of doing fragmentation it only does fragmenting on the leaf nodes.

(Q) What are the criteria you will look in to while selecting an index?

Note: - Some answers what I have got for this question.

- I will create index wherever possible.
- I will create clustered index on every table.

That is why DBA's are always needed.

- How often the field is used for selection criteria. For example in a "Customer" table, you have "CustomerCode" and "PinCode". Most of the searches are going to be performed on "CustomerCode" so it's a good candidate for indexing rather than using "PinCode". In short you can look in to the "WHERE" clauses of SQL to figure out if it's a right choice for indexing.
- If the column has higher level of unique values and is used in selection criteria again is a valid member for creating indexes.
- If "Foreign" key of table is used extensively in joins (Inner, Outer, and Cross) again a good member for creating indexes.
- If you find the table to be highly transactional (huge insert, update and deletes) probably not a good entity for creating indexes. Remember the split problems with Indexes.
- You can use the "Index tuning wizard" for index suggestions.

(DB) What is "Index Tuning Wizard"?

Twist: - What is "Work Load File"?

In the previous question, the last point was using the "Index Tuning wizard". You can get the "Index Tuning Wizard" from "Microsoft SQL Server Management Studio" – "Tools" – "SQL Profiler".

Note: - This book refers to SQL Server 2005, so probably if you have SQL Server 2000 installed you will get the SQL Profiler in Start – Programs – Microsoft SQL Server -- Profiler. But in this whole book we will refer only SQL Server 2005.We will going step by step for this answer explaining how exactly "Index Tuning Wizard" can be used.

Ok before we define any indexes let's try to understand what is "Work Load File". "Work Load File" is the complete activity that has happened on the server for a specified period of time. All the activity is entered in to a ".trc" file, which is called as "Trace File". Later "Index Tuning Wizard" runs on the "Trace File" and on every query fired it tries to find which columns are valid candidates for indexes depending on the Indexes.

Following are the step has to use “Index Tuning Wizard”:-

- Create the Trace File using “SQL Profiler”.
- Then use “Database Tuning Advisor” and the “Trace File” for what columns to be indexed.

Create Trace File

Once you have opened the “SQL Profiler” click on “New Trace”.

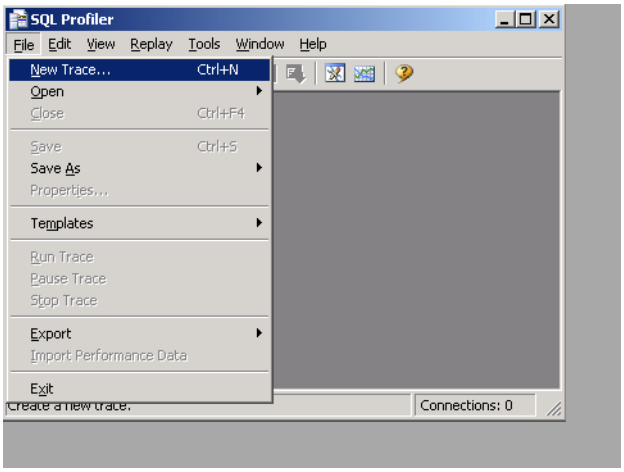


Figure 12.13: - Create New Trace File.

It will alert for giving you all trace file details for instance the “Trace Name”, “File where to save”. After providing, the details click on “Run” button provided below. I have provided the file name of the trace file as “Testing.trc” file.

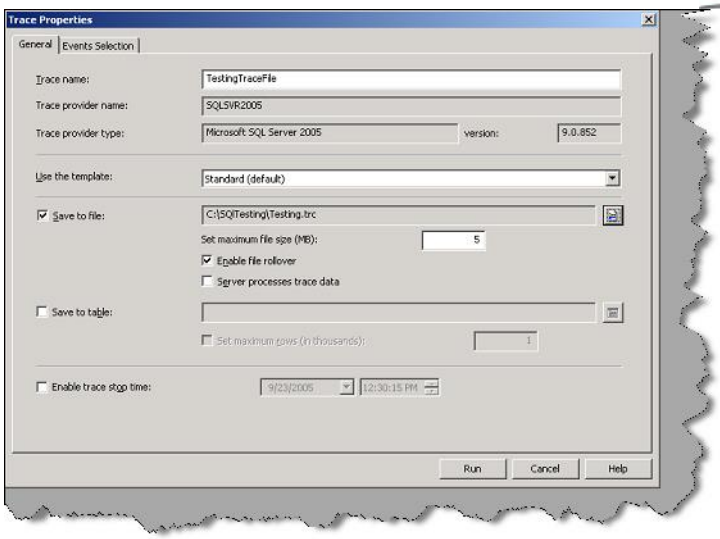


Figure 12.14: - Trace File Details

HUH and the action starts. You will notice that profiler has started tracing queries which are hitting "SQL Server" and logging all those activities in to the "Testing.trc" file. You also see the actual SQL and the time when the SQL was fired.

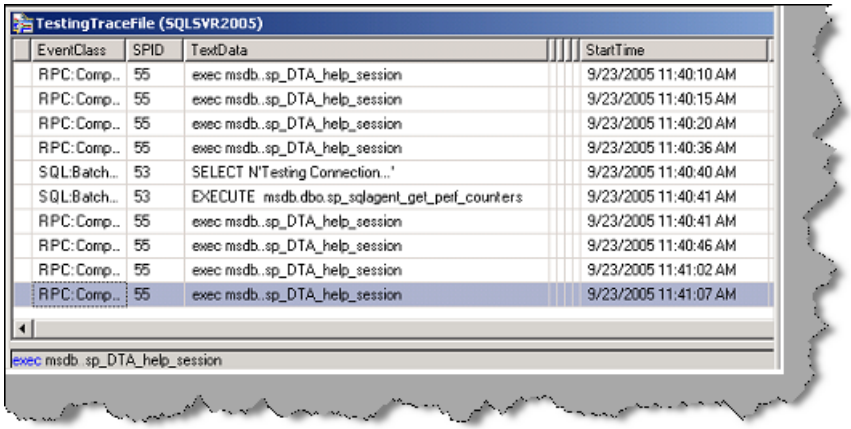


Figure 12.15: - Tracing in Actions

Let the trace run for some but of time. In actually practical environment, I run the trace for almost two hours in peak to capture the actual load on server. You can stop the trace by clicking on the red icon given above.

Figure 12.16: - Stop Trace File.

You can go the folder and see your ".trc" file created. If you try to open it in notepad, you will see binary data. It can only be opened using the profiler. So now that we have the load file we have to just say to the advisor hey advisor here's my problem (trace file) can you suggest me some good indexes to improve my database performance.

Using Database Tuning Advisor

In order to go to "Database Tuning Advisor" you can go from "Tools" – "Database Tuning Advisor".

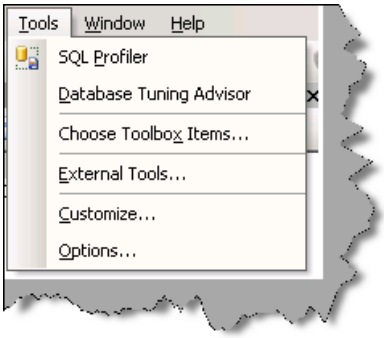


Figure 12.17: - Menu for SQL Profiler and Database advisor

In order to supply the workload file you have to start a new session in "Database tuning advisor".

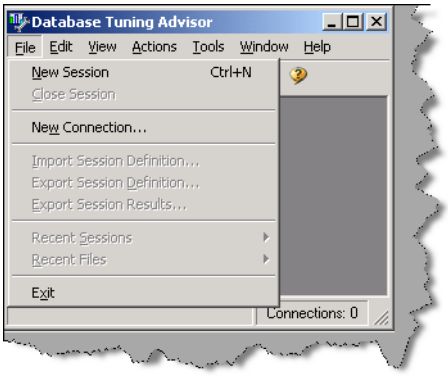


Figure 12.18: - Creating New Session in Advisor

After you have said "New Session", you have to supply all details for the session. There are two primary requirements you need to provide to the Session:-

- Session Name
- "Work Load File" or "Table"

```
(Note either you can create a trace file or you can put it in SQL
Server table while running the profiler).
```

I have provided my "Testing.trc" file, which was created when I ran the SQL profiler. You can also filter for which database you need index suggestions. At this moment, I have checked all the databases. After all the details are filled in you have to click on "Green" icon with the arrow. You can see the tool tip as "Start analysis" in the image below.

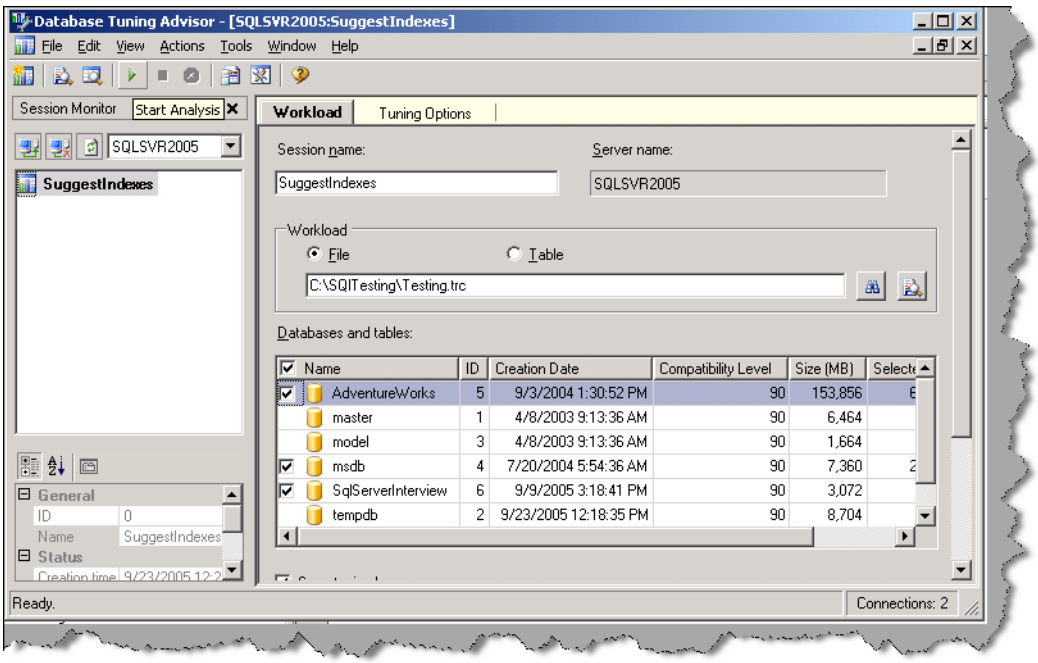


Figure 12.19: - Session Details for Advisor

While analyzing the trace file it performs basic four major steps:-

- Submits the configuration information.
- Consumes the Work load data (that can be in format of a file or a database table).
- Start performing analysis on all the SQL executed in the trace file.
- Generates reports based on analysis.
- Finally give the index recommendations.

You can see all the above steps have run successfully which is indicated by "0 Error and 0 Warning".

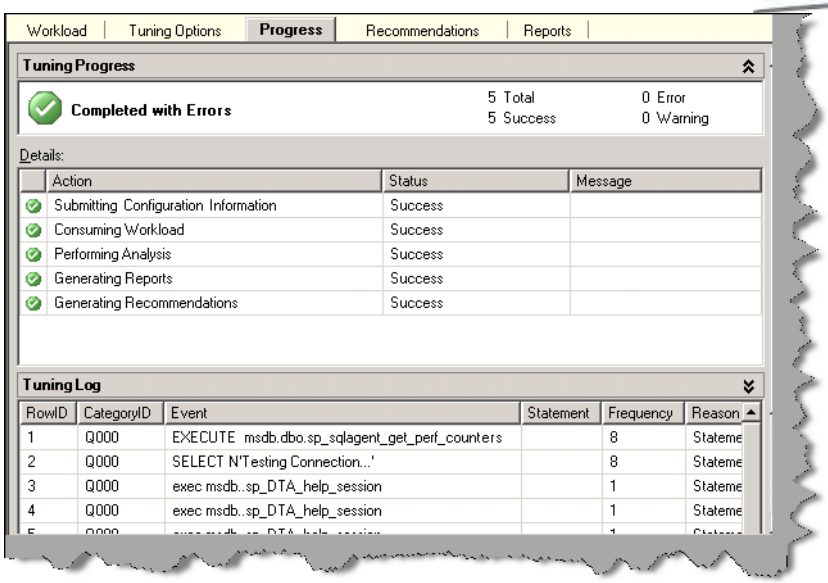


Figure 12.20: - Session completed with out Errors

Now its time to see what index recommendations SQL Server has provided us. Also, note it has included two new tabs after the analysis was done "Recommendations" and "Reports".

You can see on "AdventureWorks" SQL Server has given me huge recommendations. Example on "HumanResources.Department" he has told me to create index on "PK_Department_DepartmentId".

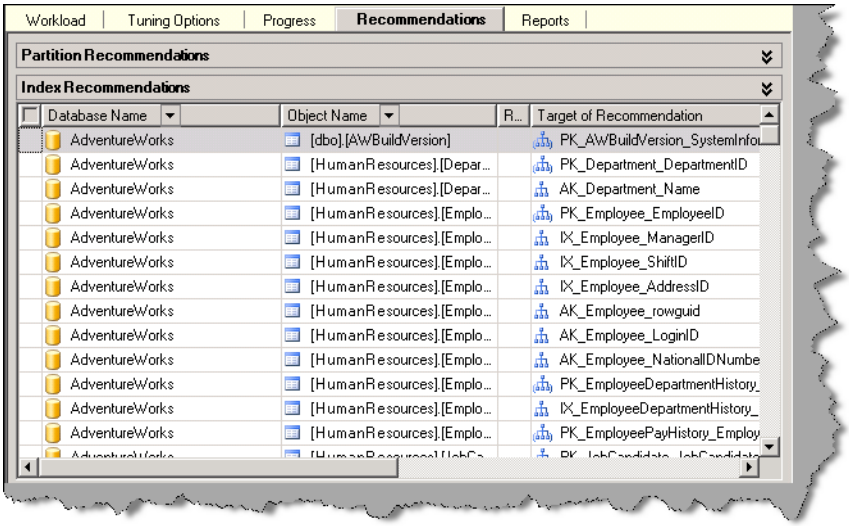


Figure 12.21: - Recommendations by SQL Server

In case you want to see detail reports you can click on the "Reports" tab and there are wide range of reports which you can use to analyze how you database is performing on that "Work Load" file.

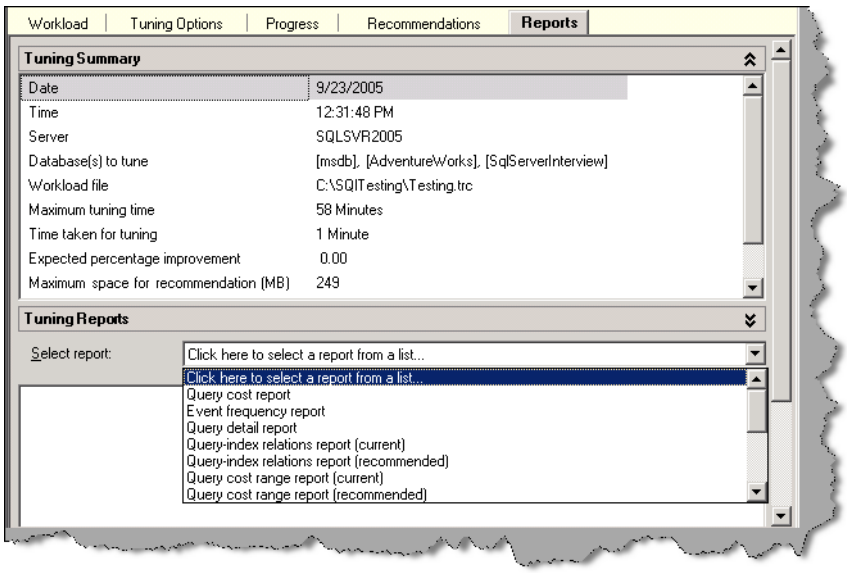


Figure 12.22: - Reports by Advisor

Note: - The whole point of putting this all step by step was that you have complete understanding of how to do "automatic index decision" using SQL Server. During interview one of the question's that is very sure "How do you increase speed performance of SQL Server? " and talking about the "Index Tuning Wizard" can fetch you some decent points.

(DB)What is an Execution plan?

Execution plan gives how the query optimizer will execute a give SQL query. Basically it shows the complete plan of how SQL will be executed. The whole query plan is prepared depending on lot of data for instance:-

- What type of indexes do the tables in the SQL have?
- Amount of data.
- Type of joins in SQL (Inner join , Left join , Cross join , Right join etc)

Click on the ICON in SQL Server management studio as shown in figure below.

Figure 12.23: - Click here to see execution plan

In bottom windowpane, you will see the complete break up of how your SQL Query will execute. Following is the way to read it:-

- Data flows from left to right.
- Any execution plan sums to total 100 %. For instance in the below figure it is 18 + 28 + 1 + 1 + 52. So the highest is taken by Index scan 52 percent. Probably we can look in to that logic and optimize this query.
- Right most nodes are actually data retrieval nodes. I have shown them with arrows the two nodes.
- In below figure you can see some arrows are thick and some are thin. More the thickness more the data is transferred.
- There are three types of join logic nested join, hash join and merge join.

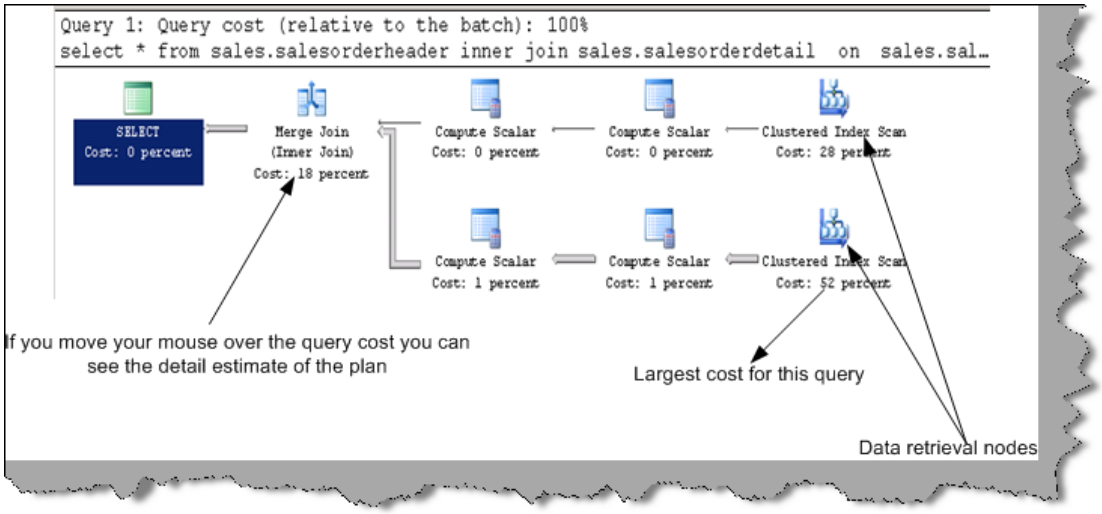


Figure 12.24: - Largest Cost Query

If you move your mouse gently over any execution, strategy you will see a detail breakup of how that node is distributed.

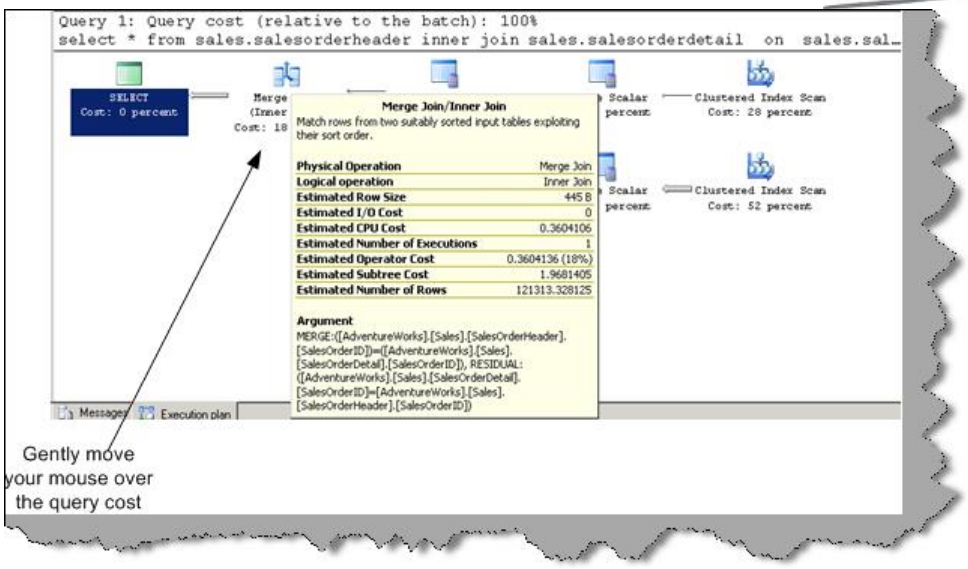


Figure 12.25: - Complete Break up estimate

(Q) How do you see the SQL plan in textual format?

Execute the following "set showplan_text on" and after that execute your SQL, you will see a textual plan of the query. In the first question, what I discussed was a graphical view of the query plan. Below is a view of how a textual query plan looks like. In older versions of SQL Server where there was no way of seeing the query plan graphically "SHOWPLAN" was the most used. Today if any one is using it that I think he is doing a show business or a new come learner.

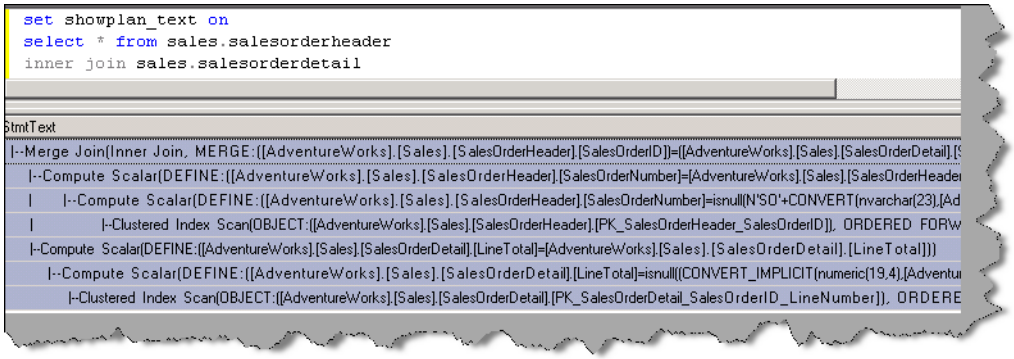


Figure 12.26: - Textual Query Plan View

(DB) What is Nested join, Hash join and Merge join in SQL Query plan?

A join is whenever two inputs are compared to determine and output.

There are three basic types of strategies for this and they are:

- Nested loops join,
- Merge join and

- Hash join.

When a join happens the optimizer determines which of these three algorithms is best to use for the given problem, however any of the three could be used for any join. All of the costs related to the join are analyzed the most cost efficient algorithm is picked for use. These are in-memory loops used by SQL Server.

Nested Join

If you have less data this is the best logic. It has two loops one is the outer and the other is the inner loop. For every outer loop, its loops through all records in the inner loop. You can see the two loop inputs given to the logic. The top index scan is the outer loop and bottom index seek is the inner loop for every outer record.

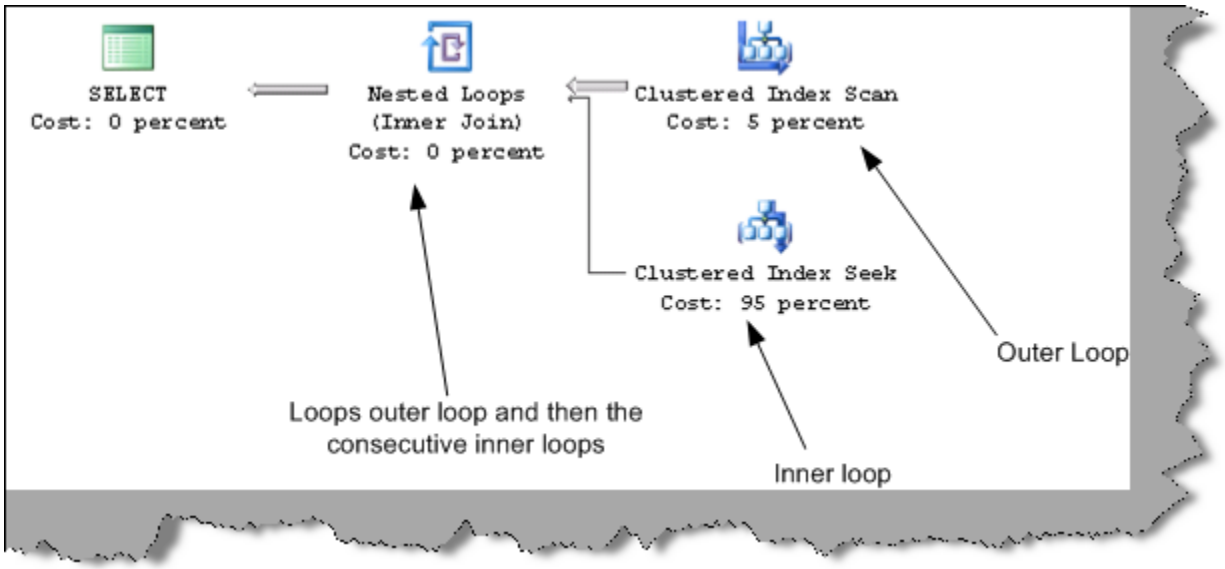


Figure 12.27: - Nested joins

It is like executing the below logic:-

```
For each outer records
For each inner records
Next
Next
```

So you visualize that if there fewer inner records this is a good solution.

Hash Join

Hash join has two input “Probe” and “Build” input. First, the “Build” input is processed and then the “Probe” input. Whichever input is smaller is the “Build” input. SQL Server first builds a hash table using the build table input. After that, he loops through the probe input and finds the

matches using the hash table created previously using the build table and does the processing and gives the output.

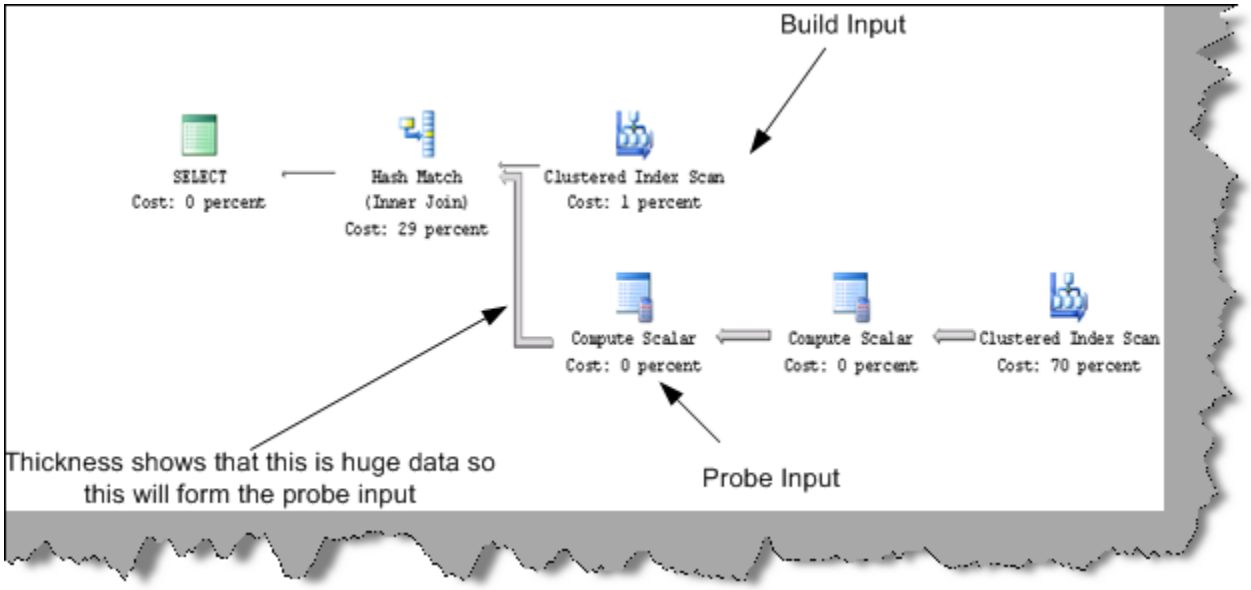


Figure 12.28: - Hash Join

Merge Join

In merge joins both the inputs are sorted on the merge columns. Merge columns are determined depending on the inner join defined in SQL. Since each input join is sorted merge join takes input and compares for equality. If there is equality then matching row is produced. This is processed till the end of rows.

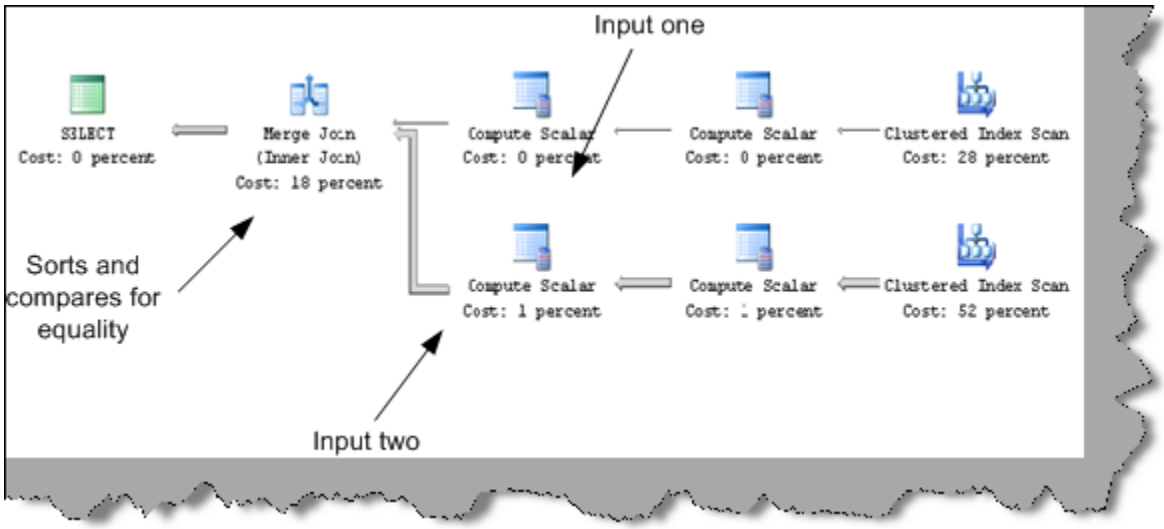


Figure 12.29: - Merge Join

(Q) What joins are good in what situations?

Nested joins best suited if the table is small and it is a must the inner table should have an index.

Merge joins best of large tables and both tables participating in the joins should have indexes.

Hash joins best for small outer tables and large inner tables. Not necessary that tables should have indexes, but would be better if outer table has indexes.

Note: - Previously we have discussed about table scan and index scan do revise it which is also important from the aspect of reading query plan.

(DB) What is RAID and how does it work?

Redundant Array of Independent Disks (RAID) is a term used to describe the technique of improving data availability using arrays of disks and various data-striping methodologies. Disk arrays are groups of disk drives that work together to achieve higher data-transfer and I/O rates than those provided by single large drives. An array is a set of multiple disk drives plus a specialized controller (an array controller) that keeps track of how data is distributed across the drives. Data for a particular file is written in segments to the different drives in the array rather than being written to a single drive.

For speed and reliability, it is better to have more disks. When these disks are arranged in certain patterns and use a specific controller, they are called a Redundant Array of Inexpensive Disks (RAID) set. There are several numbers associated with RAID, but the most common are 1, 5 and 10.

RAID 1 works by duplicating the same writes on two hard drives. Let us assume you have two 20-Gigabyte drives. In RAID 1, data is written at the same time to both drives. RAID1 is optimized for fast writes.

RAID 5 works by writing parts of data across all drives in the set (it requires at least three drives). If a drive failed, the entire set would be worthless. To combat this problem, one of the drives stores a "parity" bit. Think of a math problem, such as 3 + 7 = 10. You can think of the drives as storing one of the numbers, and the 10 is the parity part. By removing any one of the numbers, you can get it back by referring to the other two, like this: 3 + X = 10. Of course, losing more than one could be evil. RAID 5 is optimized for reads.

RAID 10 is a bit of a combination of both types. It does not store a parity bit, so it is fast, but it duplicates the data on two drives to be safe. You need at least four drives for RAID 10. This type of RAID is probably the best compromise for a database server.

```
Note :- It's difficult to cover complete aspect of RAID in this
book.It's better to take some decent SQL SERVER book for in detail
knowledge , but yes from interview aspect you can probably escape with
this answer.
```

Chapter 4:- Stored procedures , Views Cursors , Functions and triggers

What are triggers and what are the different kinds of triggers ?

Triggers are special kind of stored procedure. They can be executed after or before data modification happens on a table. There are two types of triggers “Instead of triggers” and “After triggers”.

"Instead of triggers" executes prior to data modification while "after trigger" executes after data modification.

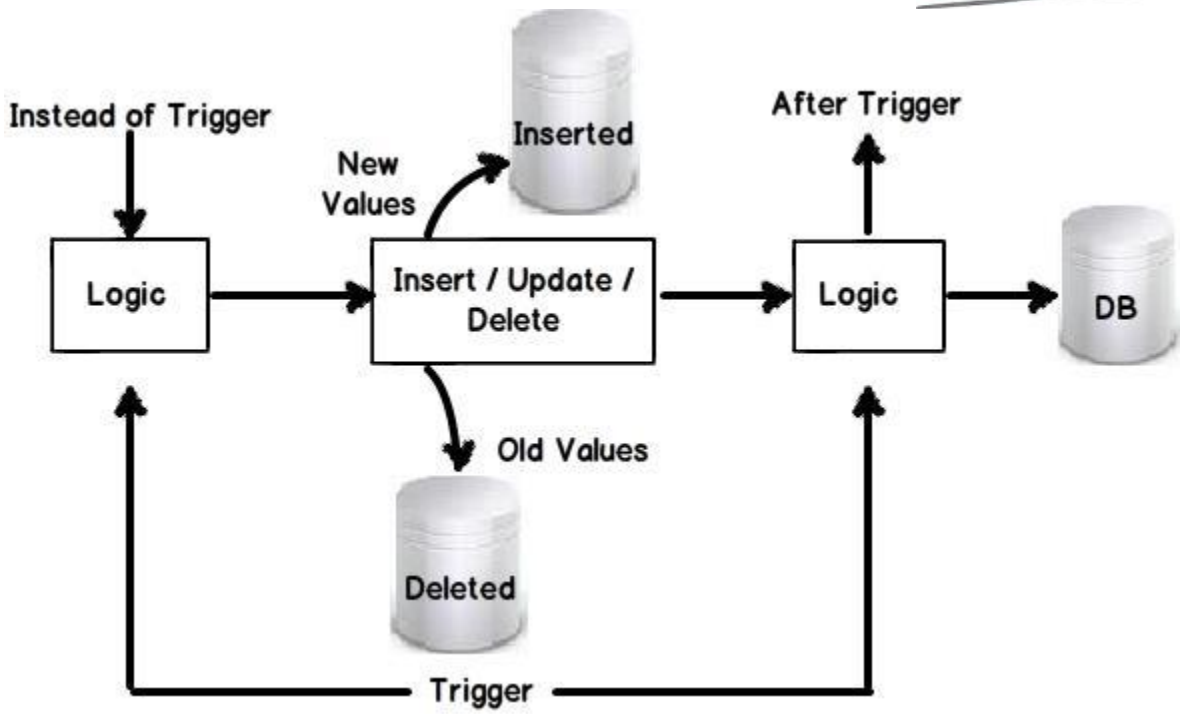


Figure: - Triggers

In what scenarios will you use instead of trigger and after trigger?

You will use "INSTEAD trigger" to take alternative actions before the update happens.

Some of the uses of instead of trigger's are:-

- Reject updates which are not valid.
- Take some alternative action if any error occurs.
- To implement cascading deletes. For instance you want to delete a customer record. But in order to delete the customer record you also have to delete address records. So you can create a instead of trigger which will first delete the address table before executing delete on customer table.

While "AFTER trigger" is useful when you want to execute trigger logic after the data has been updated.

Some uses of after triggers are:-

- For recording Audit trail where you want new and old values to be inserted in to audit table.

- Updating values after the updation has happened. For instance in a sales table if number of products and per product is inserted, you can create an after trigger which will calculate the total sales amount using these two values.

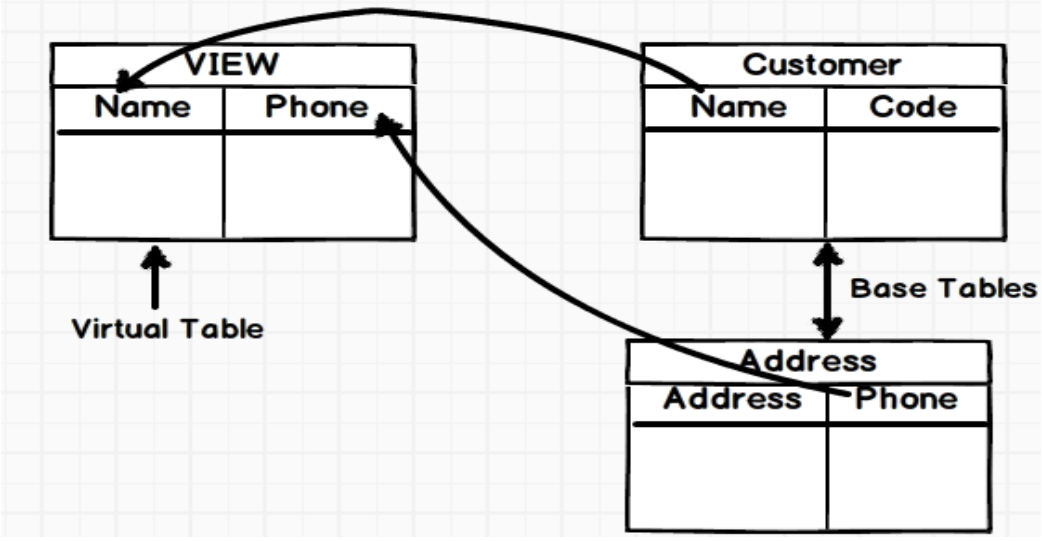
What are inserted and deleted tables?

During triggers we sometimes need old and new values. Insert and deleted tables are temporary tables created by SQL server itself which have new and old values. Inserted tables have one record for newly added data and deleted table has one record of the old version of the data.

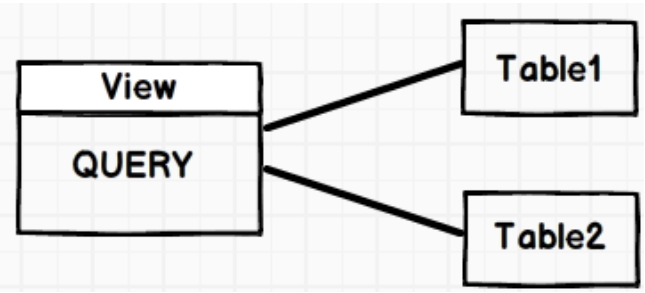
So for instance let's say we add a new record called as "Shiv". The inserted table will have "Shiv" and deleted table will have nulls because the record did not exist. Now let's say some user updates "Shiv" to "Raju". Then inserted table will have "Raju" and deleted tables will have "Shiv".

What is a SQL Server view?

A view is a virtual table which contains data from different base tables. A base table means actual physical tables. For instance you can see in the below figure we have a view which takes data from two base tables "Customer" and "Address". From customer the view fetches "Name" field and from "Address" it fetches the phone number field.

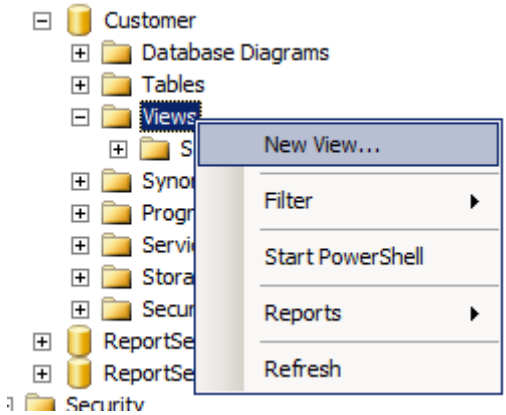


In real essence view does not contain data it's an encapsulated query which fetches data from multiple tables.



How do you create a view?

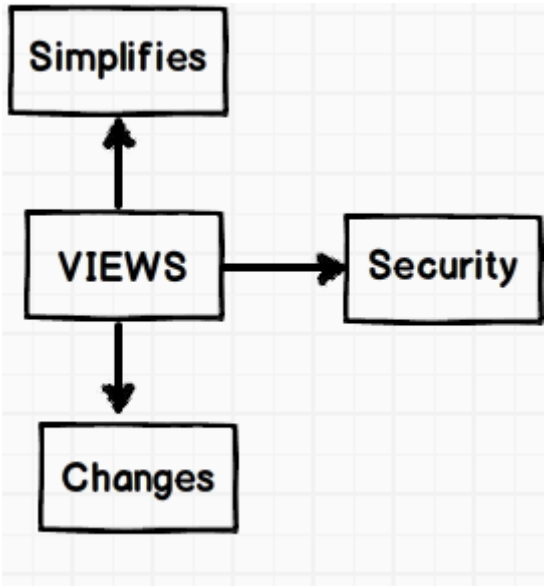
To create a view right click on the “Views” folder and then the write the query for the view.



Or you can also use the “Create View” statement as shown in the below code snippet.

```
CREATE VIEW  [View Name]
AS
[SELECT Statement]
```

What are the benefits of using a view?



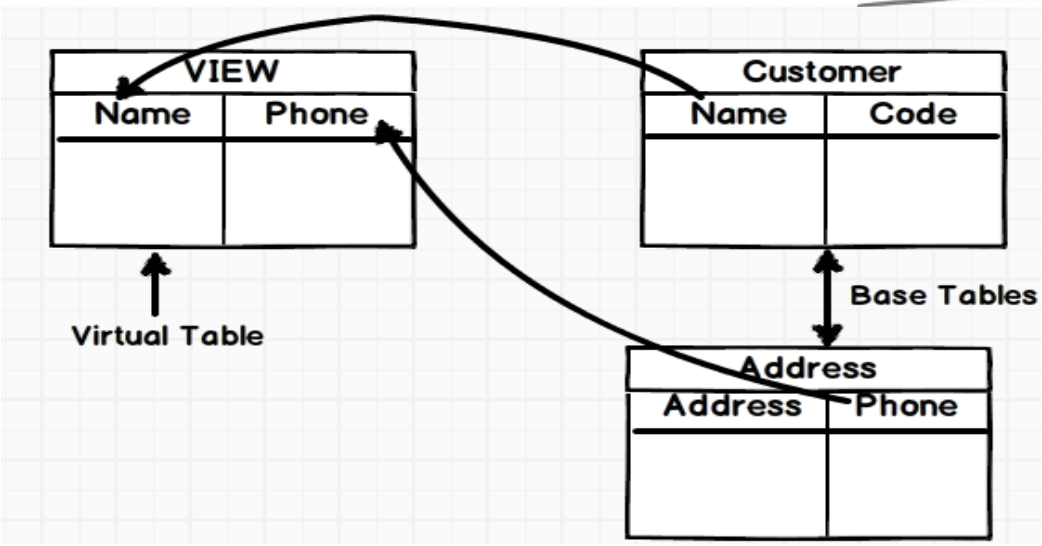
There are three big benefits of using a view:-

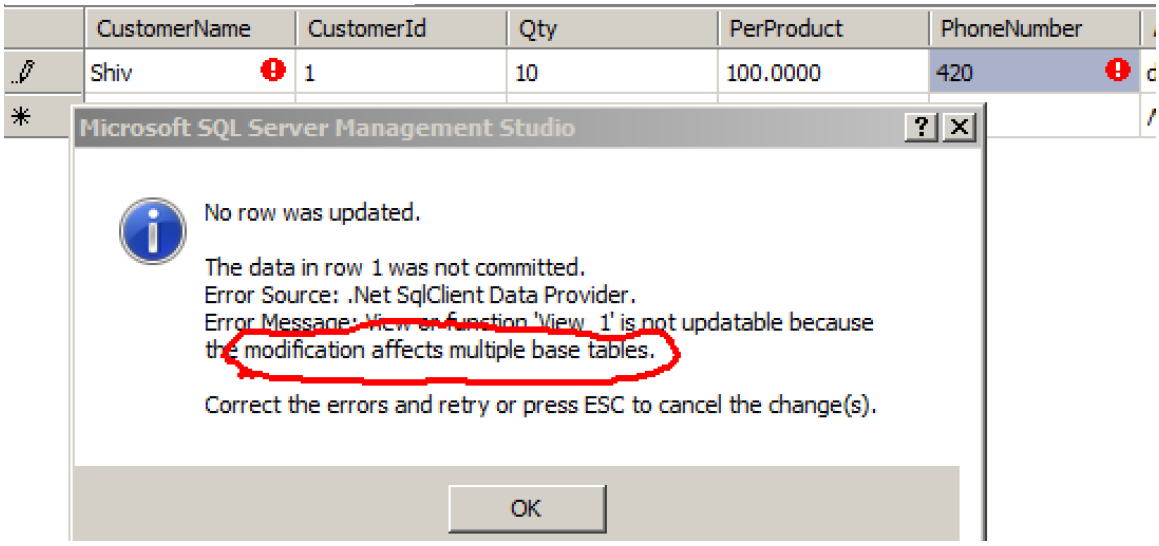
- **Simplifying things i.e. hide complexity**: - So if you have a complex SQL rather than writing them again and again create a view and keep calling the view.
- **Security**: - Sometimes you would like to give a controlled access to a table. So you can create a view and provide specific security accordingly. For instance you want a particular user to access only certain fields, so you can create a view of those fields and give access on that view. On the actual physical table the user is not allowed to access.
- **Changes**: - Due to unavoidable situation we sometimes change field names. Now if those tables are used in queries, the queries will crash. So what we can do is create views and use those views in the queries. Now if any field name changes we just need to change in the view and not across all queries.

Are SQL Server views updatable?

You can only update columns of view if they come from the same base table. If your view is using multiple base tables, you cannot update multiple columns from multiple base tables.

For instance if you have a view which uses “Customer” and “Address” base table , you can either update “Customer” base tables columns or “Address” base table columns at time.But not both of them in one go.





In case you trying to update multiple base tables you will get an error as shown in the above figure.

Chapter 2:- SQL server Data types

Chapter 2:- Constraints (Primary keys, unique keys)

Is it possible to insert NULL value in to unique keys ?

The whole point about unique keys is to have unique value. So it treats NULL has one the values and allows you to insert one NULL value (stressing it one NULL value) in to the table.

Chapter 4:- MSBI (SSIS , SSAS and SSRS)

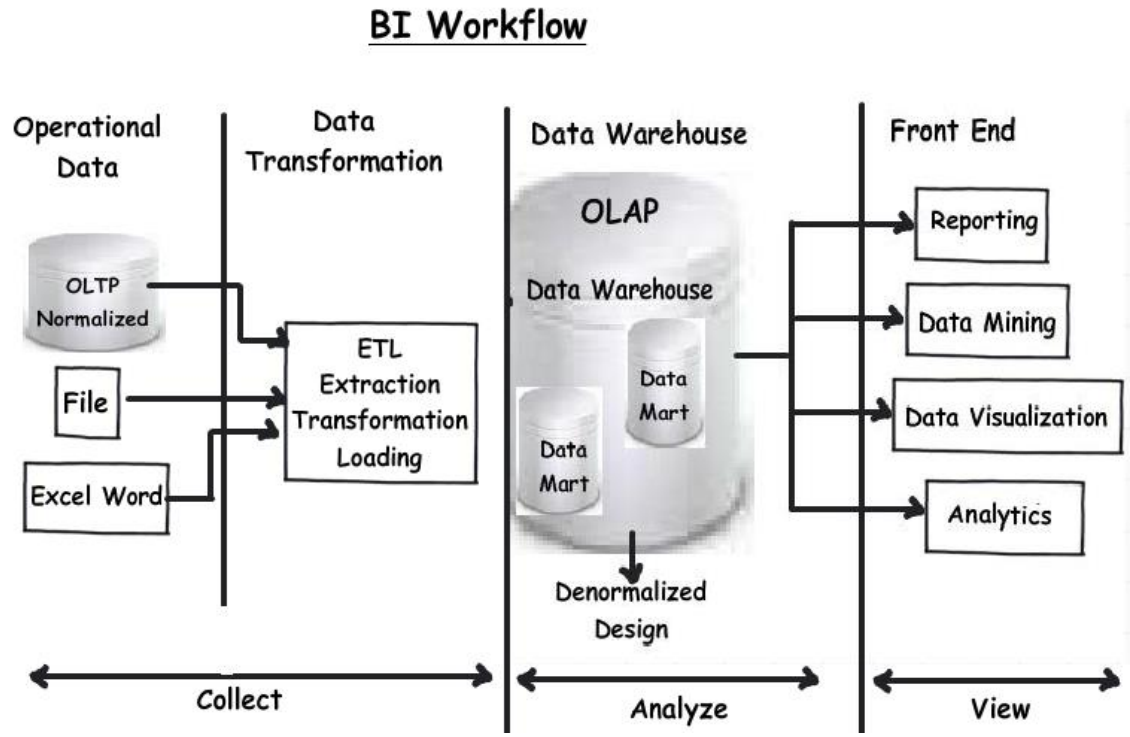
Explain Business intelligence and ETL?

Business intelligence is a collection of skills, technologies and practices, by which we can gather data, analyze data and provide data for forecasting and better decision making.

BI system has 3 important phases, Collect, Analyze and View:-

- **Collect data**: - Data in an enterprise can be stored in various formats. Now these formats can vary from normalized structured RDBMS to excel sheets or probably unstructured file formats. So the first step in BI is to collect all these unstructured and scattered data and bring them in to one uniform format.

  This can be achieved in 3 steps:-
  - **Extract**: - In this step we read unorganized data from these sources by understanding their data structure.
  - **Transform**: - In this step we transform the data in to a standard format.
  - **Load**: - Finally we load the standard format in to a data ware house.
- **Analyze data**: - Once the data is loaded in to Data ware house, you run tools, algorithms so that you can analyze and forecast information.
- **View data**: - Once you have analyzed the data you would like view it. Now again how people want to view data can vary from simple tabular format to complex graphical chart. So in this section we would need good reporting tools to achieve the same.



What is the difference between data warehouse and data mart?

Data warehouse is a database which is used to store data for reporting and data analysis. Data in data ware house come from disparate sources like structured RDBMS, Files or any other source. ETL fetches data from these sources and loads it in to data warehouse.

Data warehouse can be further divided in to data marts. Data warehouse focuses on wide data while data mart focuses on single process.

Data warehouse are also termed as OLAP systems. The database designs for these systems do not follow conventional normalization (1st, 2nd or 3rd normal form) design. Most them use denormalized design like star schema and snow flake design. They normally store data in fact and dimension tables.

What is the difference between OLTP and OLAP system?

Please refer chapter database design

What is the difference between star schema and snow flake design?

Please refer chapter database design

What are Facts, Dimension and Measures tables?

Please refer chapter database design

What are Cubes?

Please refer chapter database design

Can you explain ROLAP, MOLAP and HOLAP?

Let first quickly run through full forms.

ROLAP stands for Relational Online Analytical Processing, MOLAP stands for Multidimensional Online Analytical Processing and HOLAP stands for Hybrid Online Analytical Processing.

When we talk about BI it has two kinds of data, one is the actual data and other is the aggregated data. For instance you can see in the below table we have customer name, quantity, per price and total amount. In this case the total amount column is a derived or aggregated data while all the other fields are actual data (detail data).

Customer name	Quantity	Per price	**Total amount**
Shiv	10	100	**1000**
Raju	5	10	**50**

ROLAP, MOLAP and HOLAP define storage structure for Business intelligence. It defines how the actual (detail) data and aggregated data are stored.

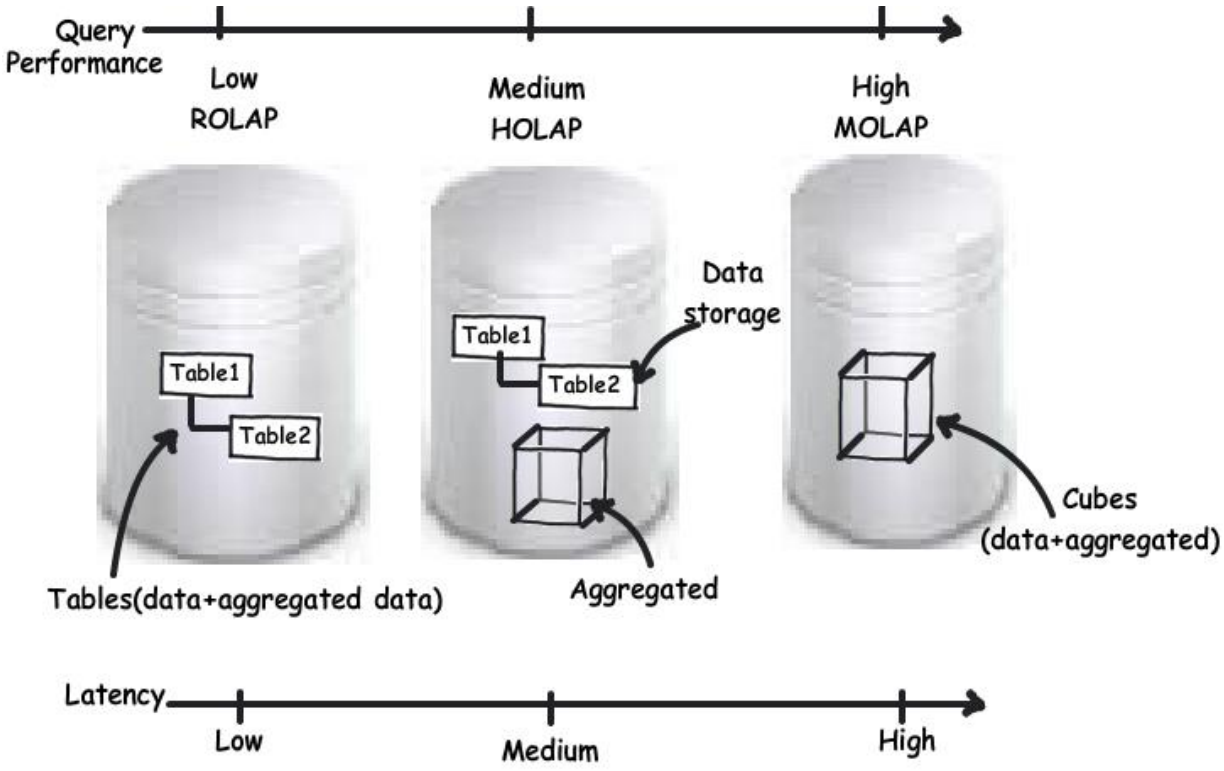
ROLAP stores data and aggregated data in relational format. So in this case query performance is low and also latency is low. Low latency means you get aggregated/calculated data instantly.

MOLAP stores data in cubes i.e. multi-dimensional format. It uses facts, measures and dimension table structure to form a cube (read the previous question for cubes). Multi-dimensional database design format is optimized for better query performance. So the performance is much better as compared to ROLAP but the latency is high.

The latency is high in MOLAP because it needs to fetch data from relational database (operational data), do calculations / aggregations / and convert the relational structure data to cube structure data.

HOLAP is a hybrid approach. It's a combination of MOLAP and ROLAP. HOLAP stores detail data in to relational database i.e. (ROLAP) and the aggregated data is stored in to MOLAP (cubes).

Due to this approach query is faster than ROLAP but not as fast as MOLAP. Latency is less than MOLAP but higher than ROLAP.



Below is comparison summary table.

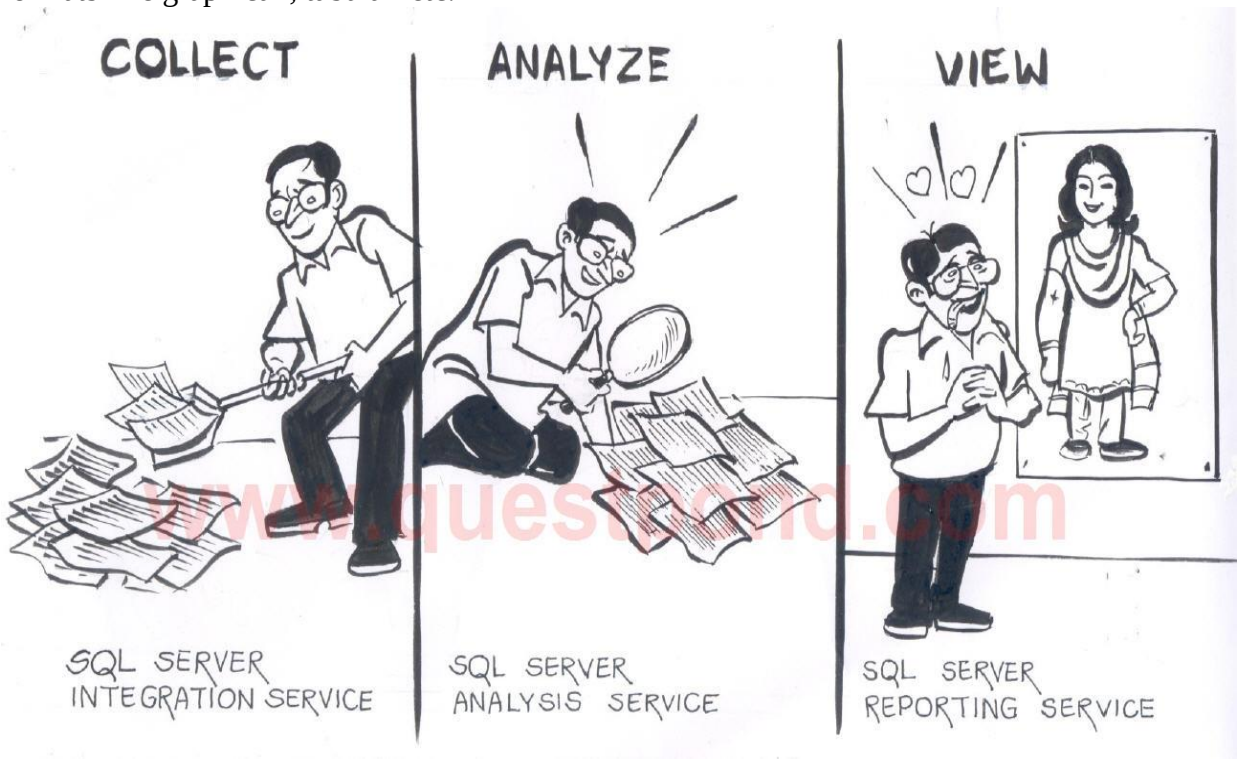
	Storage structure	Query performance	Latency
ROLAP	Relational	Low	Low
MOLAP	Cubes	High	High
HOLAP	Relational + Cubes	Medium	Low

Where does SSIS, SSAS and SSRS fits in?

SSIS (SQL Server integration services) helps to **collect data**. In other words it does ETL (Extract transformation and loading) as explained in the previous question.

SSAS (SQL Server analysis services) helps us to **analyze data** by creating cubes, facts , dimensions and measures.

SSRS (SQL Server reporting services) helps us to **view** this analyzed data in different formats like graphical , tabular etc.



Chapter 4:- Business intelligence (SSIS)

What role does SSIS play in BI?

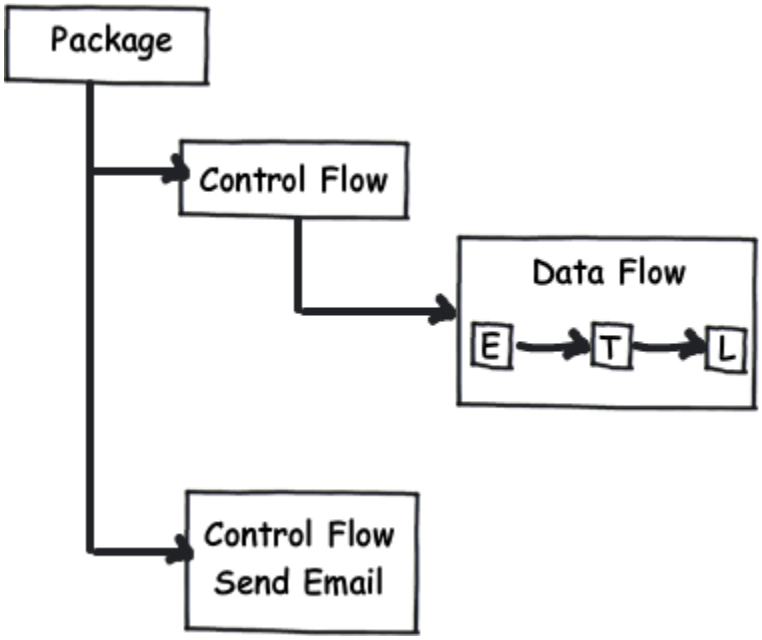
SSIS does three important things extraction, transformation and loading.

What is a package, control flow and data flow?

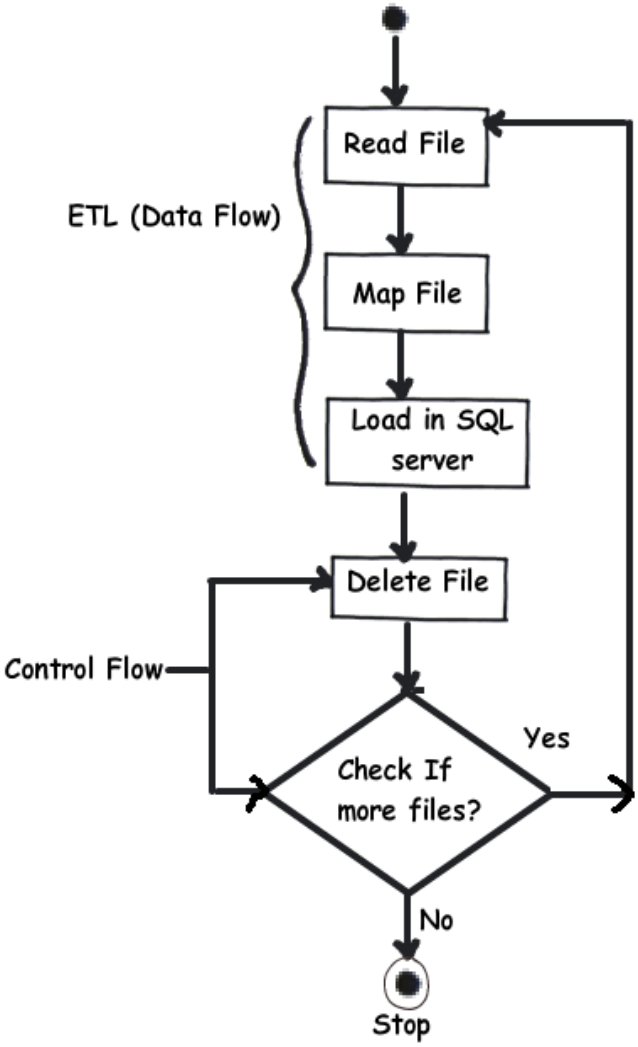
Data flow defines flow of data from a source to a destination. In other words it defines individual ETL.

Control flow defines work flow (iterations, conditional checks etc.) and component execution (example: - send email, copy files, invoke web services, FTP etc.). Control flow has nothing to do with data. The data flow only deal with data i.e. movement, transformation and loading.

Package is a collection of control flows.



For instance below is a simple example, lets understand where data flow and control flow will fit in.In the below example we need to read files from a folder and load the same in SQL Server. Once this file is loaded, delete the file and proceed with the next file in the folder. This continues until there are no more files present in the folder.



In the above process data flow will take care of reading the file, mapping and loading it to SQL Server. While control flow will delete the file once data flow has loaded the file in to database. It will check if there more files in the folder and the accordingly invoke the data flow.

Can you explain architecture of SSIS (SQL Server integration services)?

The complete SSIS architecture comprises of 8 components.

Data flow task: - It defines source, mapping and transformation. It defines the core ETL process.

Control flow task: - This section helps to define logic and invoke task like data flow tasks, send email task, web service task etc.

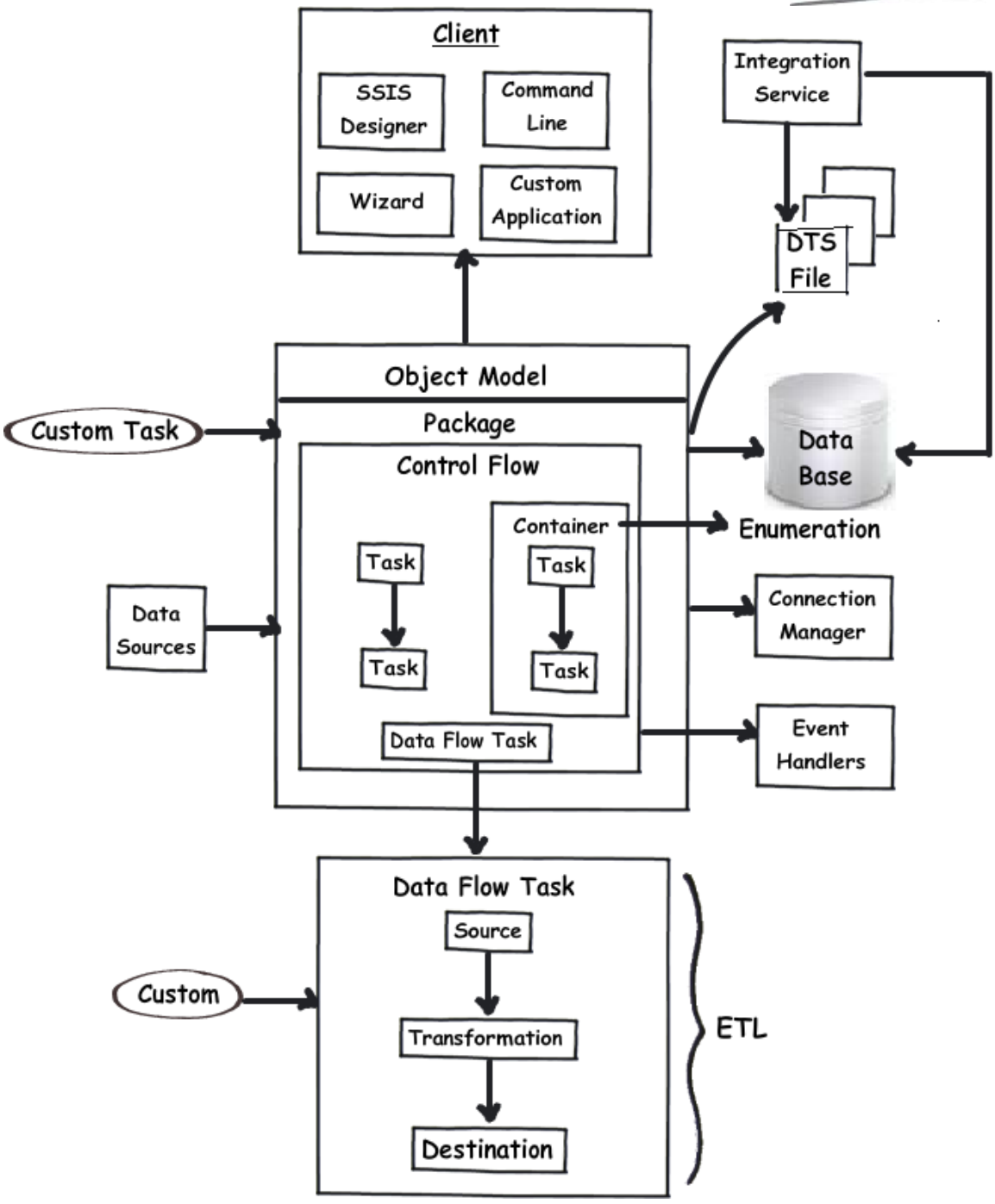
Package: - It's a collection of control flow tasks. It's the unit of work which is retrieved, executed and saved.

Client: - SSIS system can be connected via various clients like SSIS designer which comes with BI IDE, custom applications, SSIS wizard etc. These clients use the object model to communicate with the SSIS system.

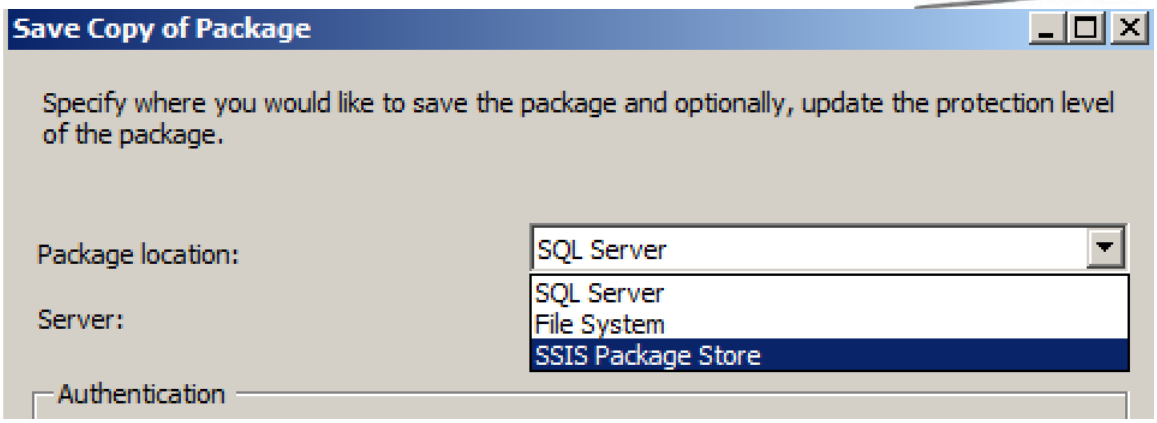
Connection manager and data sources: - They help us to define connection objects with data sources which can be reused in data flow task.

DTSX files, MSDB and Integration service: - The complete package can be stored in DTSX files or database. These packages can later be connected and invoked using integration service which runs at the background as windows services.

Event handlers: - When you run any SSIS package you would like to trap various events like OnError, OnPostExecute, OnPreExecute etc. to run certain logic like logging in to database, send emails etc.



What are the different locations of storing SSIS packages?



You can store SSIS packages in three locations Physical files (File system), SSIS package store or SQL Server.

- File system stores SSIS package in simple folders.
- "SSIS package store" saves package in "C:\Program Files\Microsoft SQL Server\..\DTS\Packages".
- SQL Server stores package in SQL Server database called as MSDB.

What is the different between project deployment and package deployment ?

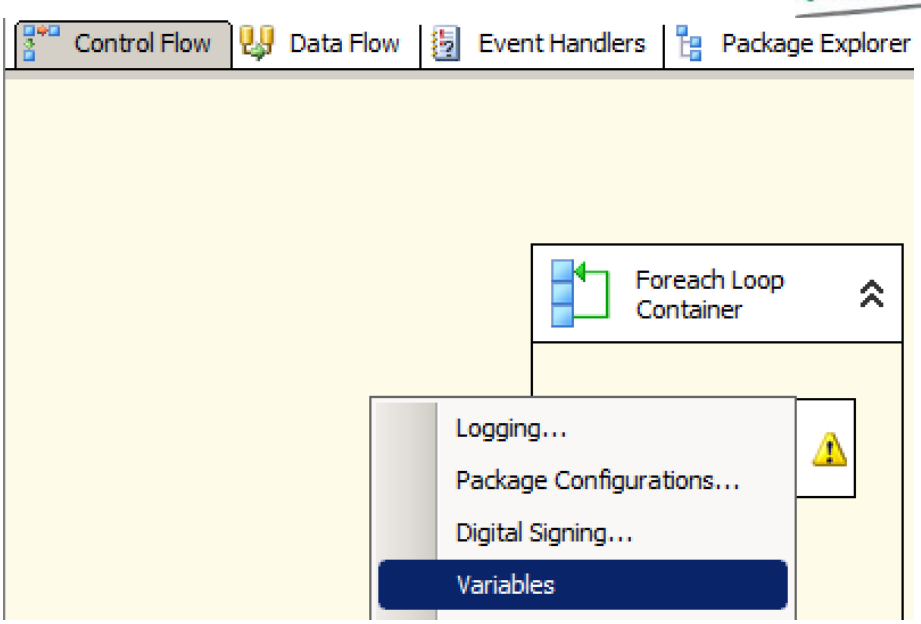
How can we execute SSIS packages?

We can execute SSIS packages by using DTEXEC or DTEXECUI. DTEXEC is a command line utility while DTEXECUI gives a nice user interfaces to execute the package.

What are the different types of variables in SSIS?

There are two types of variables user defined and system defined variables. System defined variables are those variables which are ready made and given by the system. Some of the examples of system defined variables are package name, last modified etc.

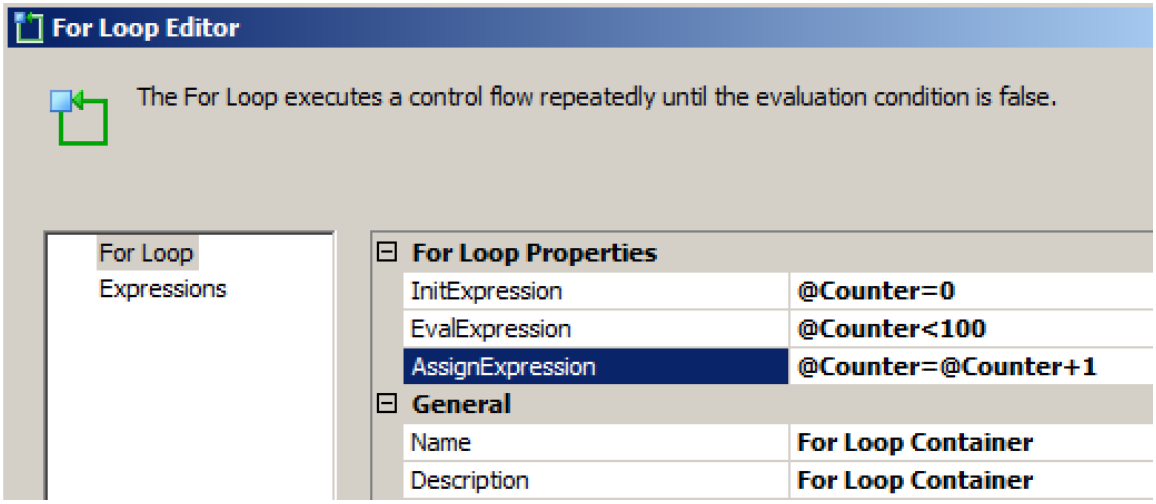
User defined variables are created by SSIS developers. User defined variables can be created by right clicking on package / data flow – variables – add variables.



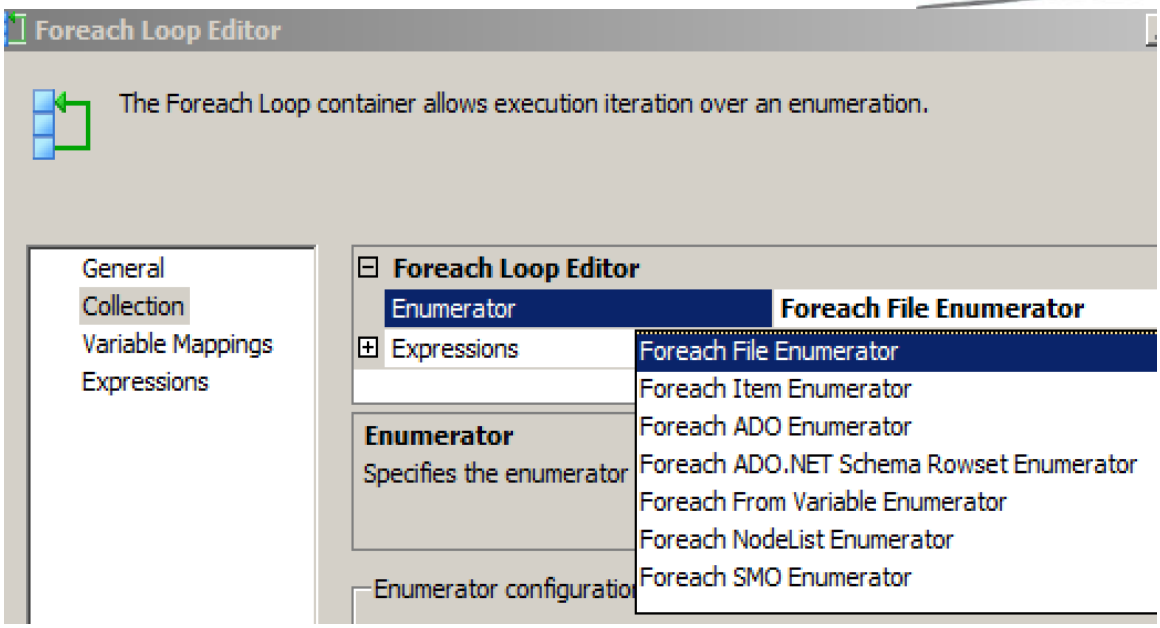
User defined variables can have a package scope, data flow scope, container scope etc.

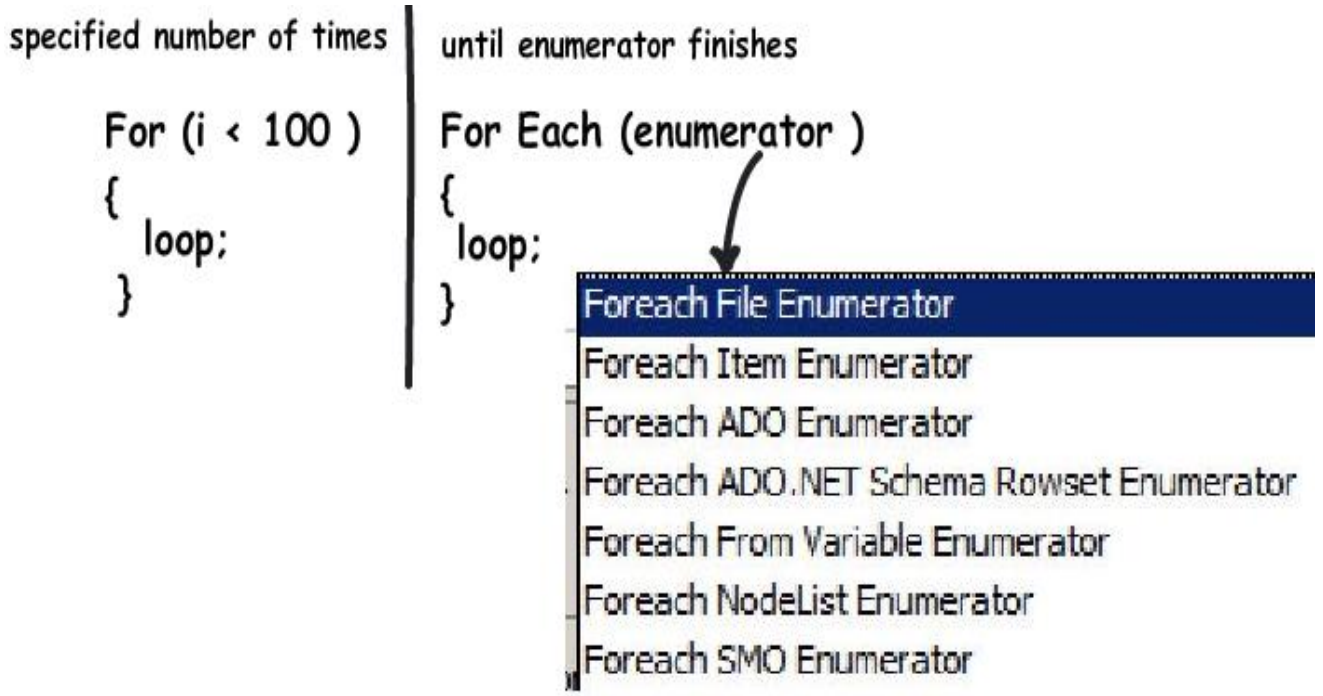
Explain difference between “For loop container” and “Foreach loop container”?

The “For Loop Container” executes specified number of times like 10 times, 20 times until the specified condition is met.



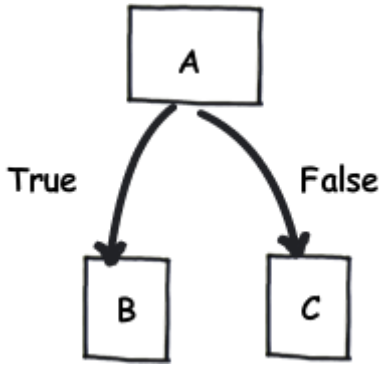
The “Foreach Loop Container” runs over an iterator. This iterator can be files from a folder, records from ADO, data from a variable etc.



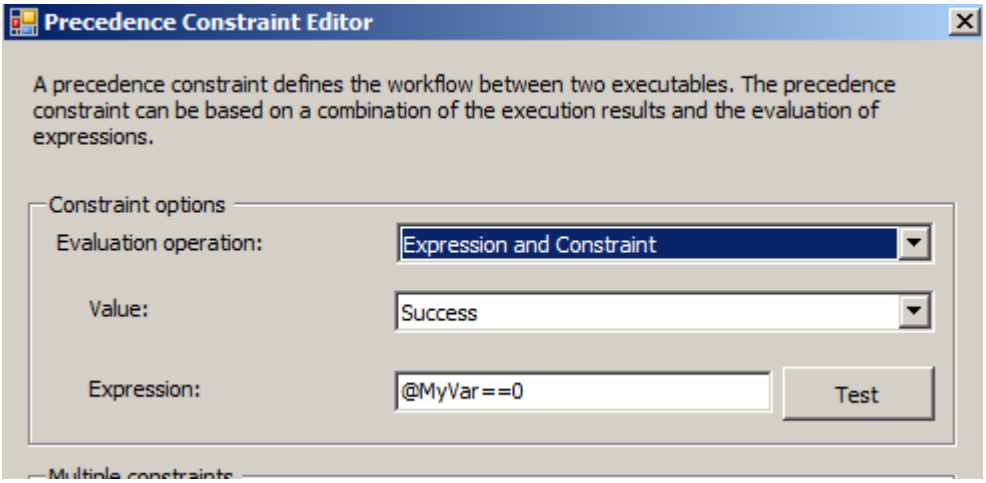


What are precedence constraints in SSIS?

Precedence constraints links tasks/ containers and also specifies conditions on which those tasks/containers should execute.

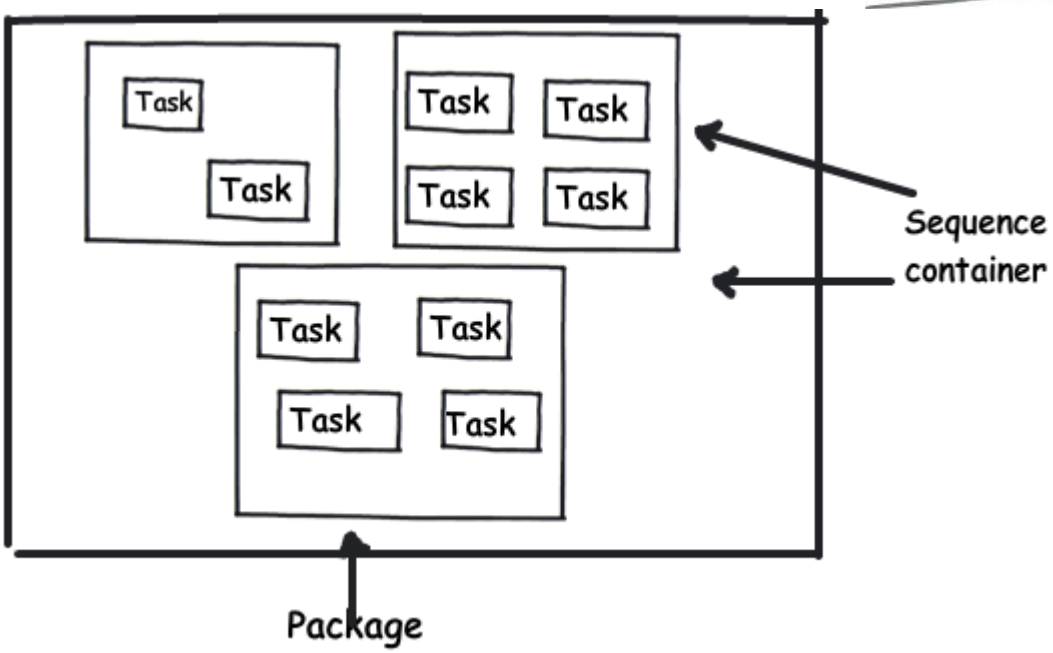


To add a precedence constraint, right click on any task and click “Add precedence constraint”. You can then specify precedence constraint using the editor as shown in the below figure.



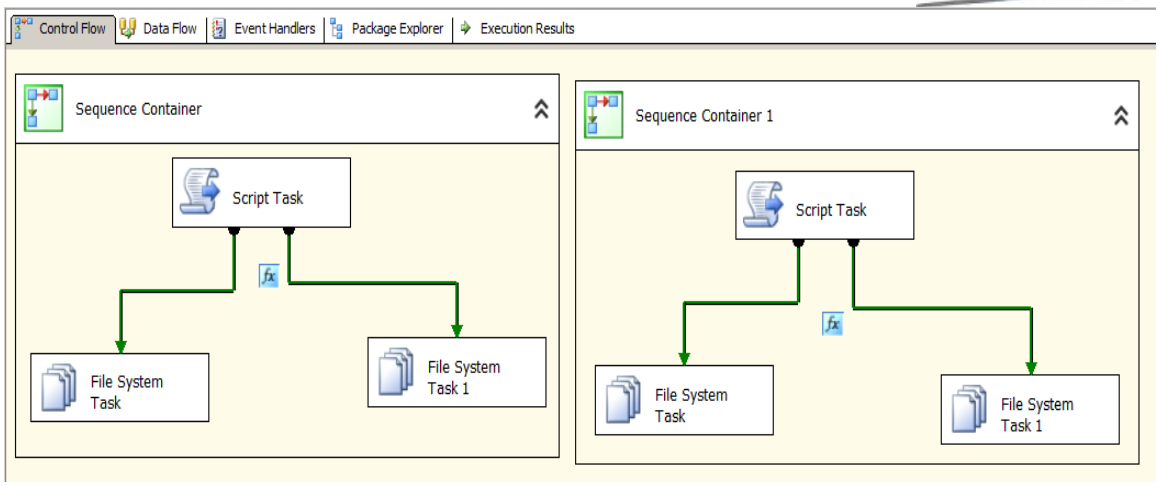
What are sequence containers in SSIS and how do they benefit?

Sequence containers group’s set of tasks logically. They help to define multiple control flows inside a package.



Following are the benefits of using sequence containers:-

- Makes your package more readable and easy to maintain. You can expand and collapse the container for ease of reading during design mode.
- You can define transactions for a group of task by specifying transaction attribute at the sequence container level.
- Define variables which are scoped only for the container.
- You can do focus debugging on a particular container and disable other containers.
- Manage properties at container level rather than on individual tasks. For instance you can enable and disable a sequence container rather than enabling and disabling each task.



How can we consume web services in SSIS?

We can consume web services in SSIS by using the "Webservice" task. The data which is received from the web service can be directed to a file or a SSIS variable.

How to check quality of data using SSIS?

Many times you get raw data (as the one shown below) and you would like to understand what kind of quality does this data have ?. For example for the below data you would probably like to know:-

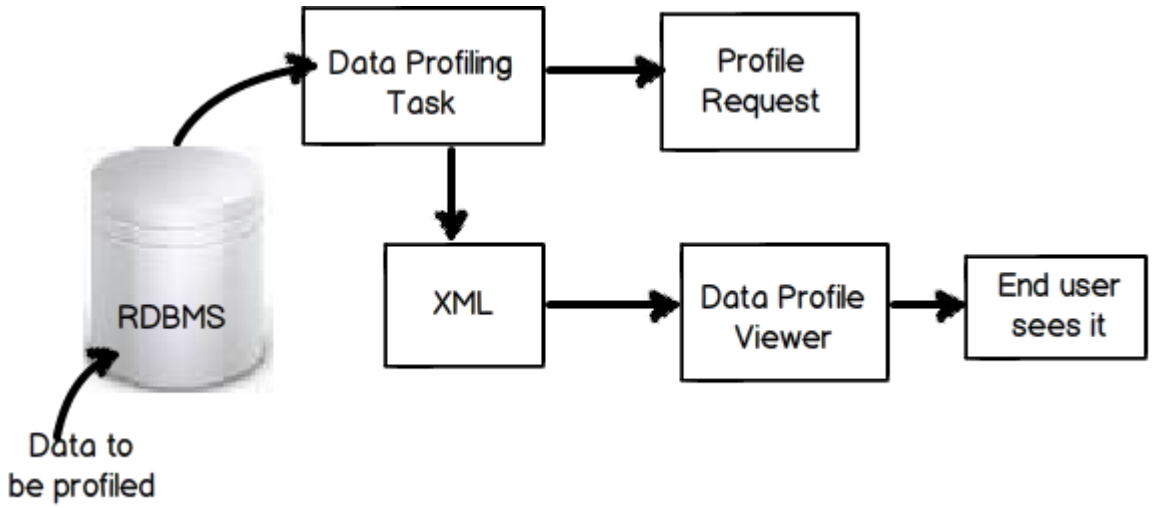
- How many null values exist in the name field?
- What are the types of contact information, email, phone, address etc.
- What kind of salary range exists?

Name	Contact	D O B	Salary	Country	EMP Code	Pan card	CountryTaxcode	Tax%
Shiv	shiv_koirala@yahoo.com	3/12/1980	1000	IND	E001	D001	IND	5
Raju	91-022-2130928933	11/2/1975	1500	IND	E002	D002	IND	5
	shaam@yahoo.com	3/16/1988	1000	NEP	E003	D003	NEP	2
Ajay	ajay@yahoo.in	5/22/1986	1000	IND	E004	D004	IND	5
Kumar	kumar@gmail.in	9/24/1977	6000	USA	E005	D005	USA	6
Neeraj	neeraj@yahoo.com	4/16/1971	4000	USA	E006	D006	USA	6
	suraj@yahoo.com	2/19/1973	8000	IND	E006	D007	IND	5
Vishal	vishal@gmail.in	6/24/1978	3000	USA	AMDK	D008	USA	6
sharma	sharma@yahoo.com	3/26/1976	2000	IND	E005	D009	IND	5

Yadav	91-022-2130928933	8/13/1983	1000	IND	AQPR	D010	IND	5
Dinesh	dinesh@yahoo.com	1/17/1966	5000	IND	E007	D011	IND	5

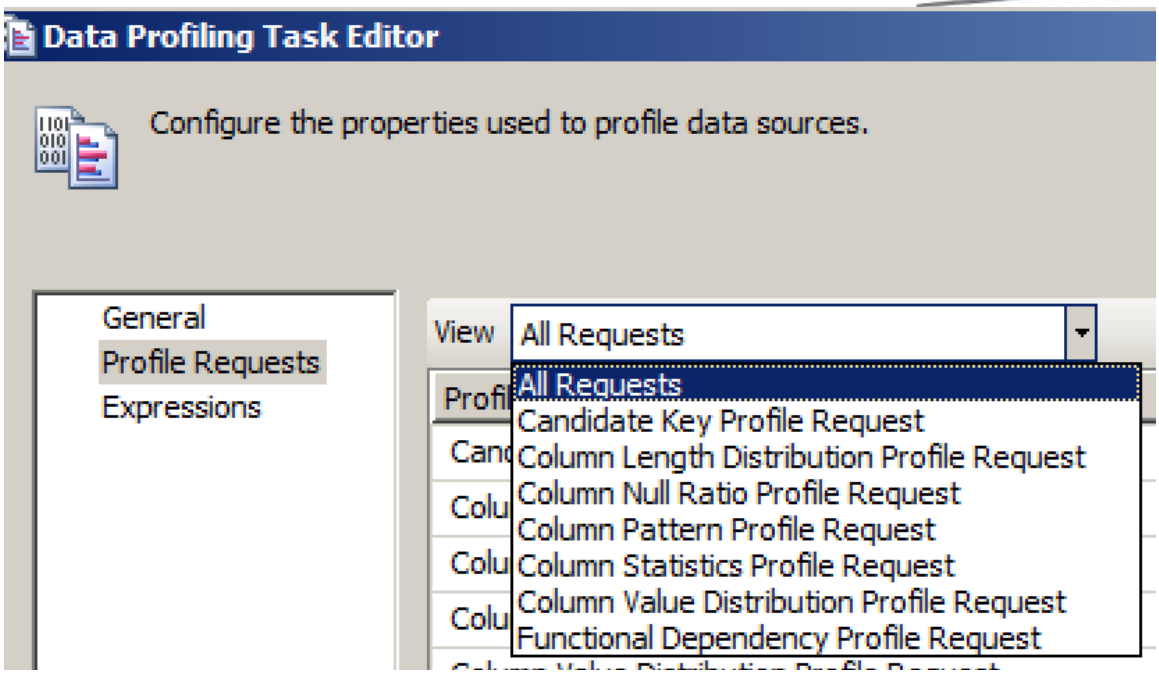
This can be achieved by using data profiling task. Data profiling task is available in the control flow toolbox.
Following steps needs to be followed:-

- Create profile request in data profiling task.
- Once you run the data profiling task it creates a XML output.
- You can then view the XML output using data profile viewer. Data profile viewer exists in "C:\Program Files (x86)\Microsoft SQL Server\100\DTS\Binn" directory.



What kind of profile requests exists in SSIS?

There are 8 ways by which you can profile requests.



Below are more details of what kind of data analysis is performed by these 8 profile requests.

Type of data analysis	**Profile request**
How many NULL values exist?	Column null ratio profile request
Detects what kind of pattern does the data have email address , website URL etc.	Column pattern profile request.
What are the minimum, maximum, average values in column?	Column Statistics profile request.
What are the distinct lengths of string values?. For instance you have a country code column you would like to ensure that the length should be equal to 3 (IND , USA). In case there are some other lengths you would like to take necessary actions ahead.	Column Length Distribution Profile
Finds out how many distinct values exists for a column.	Column Value Distribution Profile
How much does one columns depend on other column?. It helps you to find out at how many places the dependency has been violated.	Functional Dependency Profile
Which columns are good candidates for primary keys ?.	Candidate Key Profile
Checks if there is overlap of values between two	Value Inclusion Profile

columns ?. Helps to detect a likely foreign key column ?.	

What is the difference between Merge and Merge join transformation?

In merge the data from two inputs are merged as one. In merge join two inputs are merged on a basis of a common key. In merge join you can specify left join, right join or inner join depending on the join key.

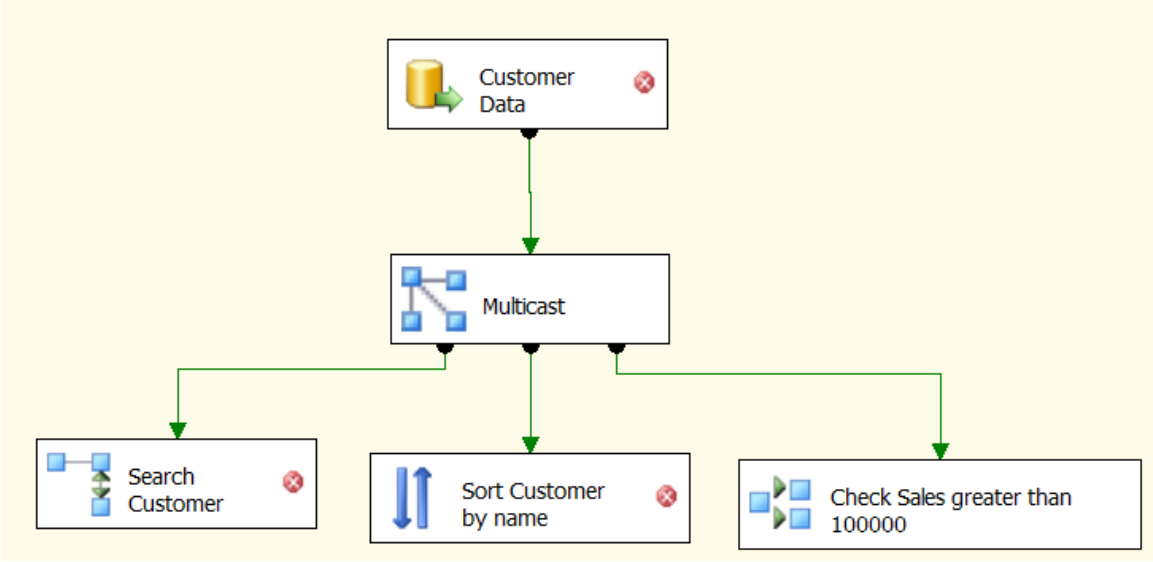
If we have data unsorted will merge and merge join work ?

A big “NO”.

How can you send a single data source output to multiple SSI controls?

This can be achieved by using Multicast control. For instance you can see in the below figure we have a single ADO.NET source. Data coming from this single data source needs to be used in lookup control (search customer), sort control (for sorting by customer name) and conditional split control (check sales > 10000).

You can see how SSIS multicast control has three outputs which is broadcasted to all these three controls.



You have millions of records in production, you want to sample some data to test a SSIS package ?

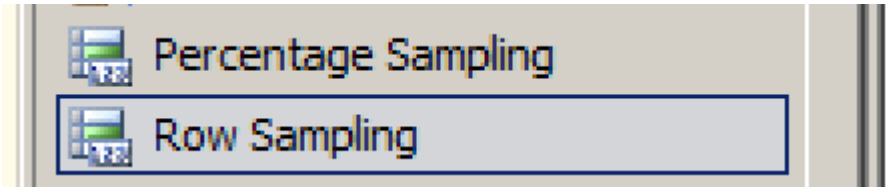
```
Note :- Sometimes  interviewer can also twist this question by asking
"What is percentage sample and row sampling in SSIS ?".
```

Many times we would like to test SSIS project against production test data. Normally Production data is huge in quantity. If you run the project against the full production data it can take ages for running and testing your SSIS project.

So for those kinds of scenarios we have two great SSIS components "Percentage sampling" and "Row Sampling". "Row sampling" extracts exact number of records. So for example you have 150 records and you want to sample only 50 "Row sampling" is the way to go.

But in some scenarios we want to sample percentage of data rather than absolute data. For instance from 150 records we want to sample 22% of 150 which comes to 33 records. So for those kinds of scenarios we can use "Percentage sampling".

Both these components are found in the SSIS toolbox as shown in the below figure.



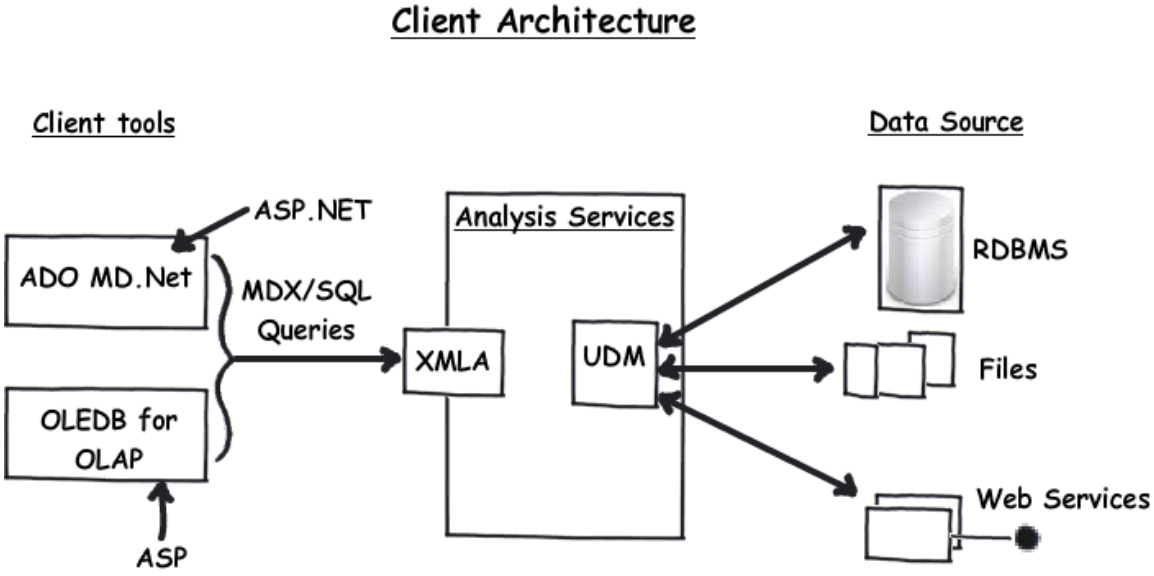
What is the use of SCD ?

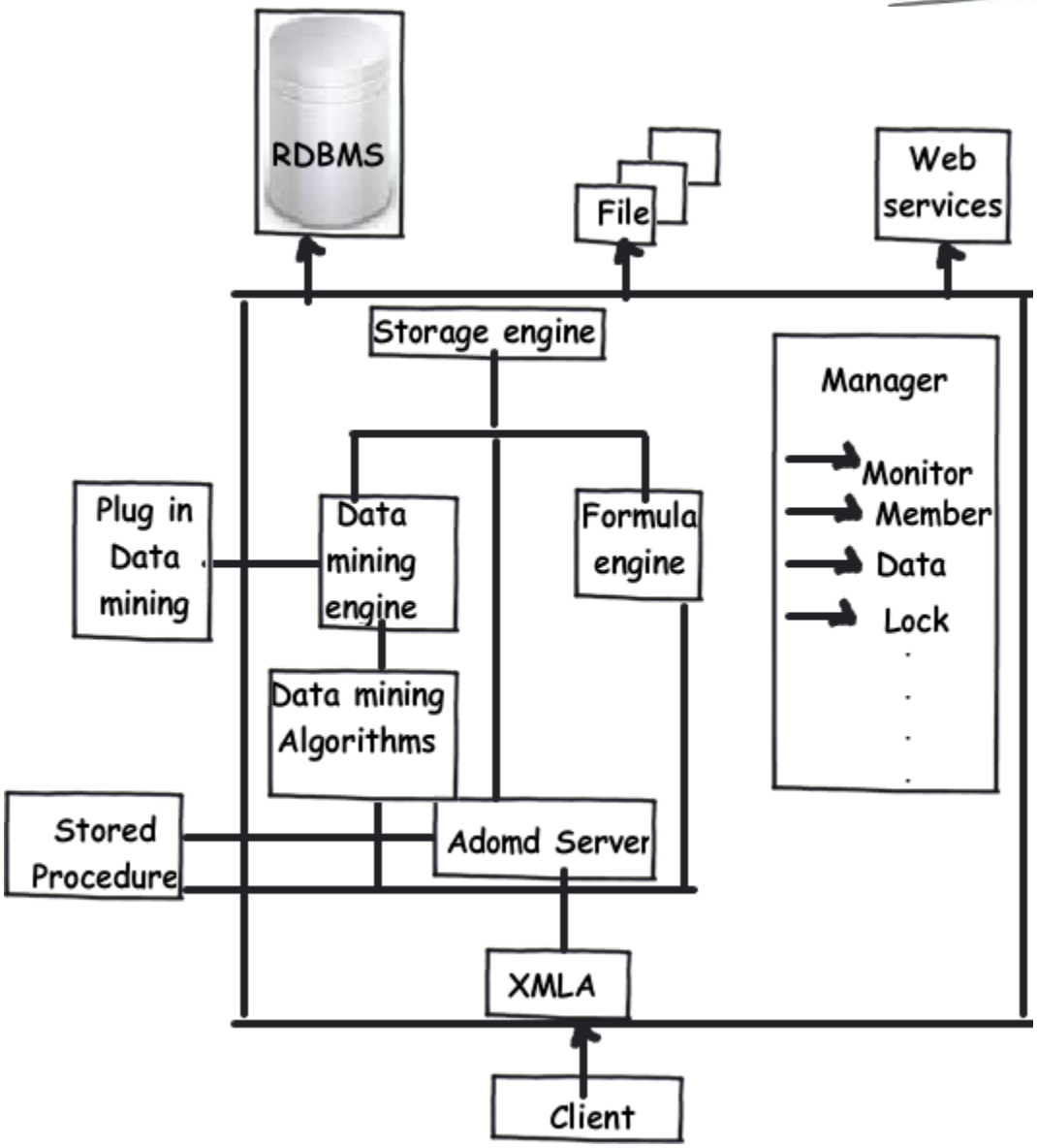
Using SSIS how can we standardize "Indian","India" and "Ind" to "Ind"?

How can we convert "string" to "int" data type in SSIS ?

What is the use of “Audit” component ?

Chapter 5:- Business intelligence (SSAS)





How can we apply scale-out architecture for SQL Server Analysis Services?

Before we answer this question lets first define scalability, scale-up and scale-out.

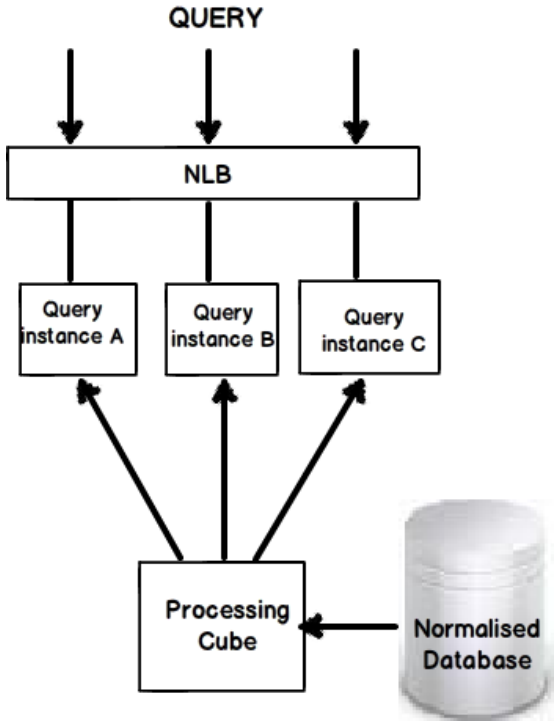
Scalability is the ability of a system to handle growing amount of load without degrading performance. Consider you have SSAS system which runs with 100 users efficiently. Let's say over a period of time your load increases to 500 users, your system still has the ability to handle the load and provide the same efficiency.

You can achieve scalability by two ways either you can Scale-up or Scale-out. In scale-up we have only one machine and we increase the processing power of the machine by adding more processor, more ram, more hard disk etc. So if your load increases you add more ram , more processor etc , but you do not add extra physical machines.

In Scale-out, as your load increases you add more computers and you have load balancers in front of those machines to distribute load appropriately.

Now when we talk about sql server analysis services we have two big tasks one is querying the analysis service cube and other is processing of the cube. So we can create scale-out architecture by having dedicated machines for processing the SQL Server analysis services queries and dedicated machine for processing the cube.

Below is how the physical architecture shapes up. The boxes represent physical computers. So you can see how we have separate physical machines to handle sql server analysis services queries and separate physical machines which call pull data from data source, process the cube and replicate the cube data to query physical servers.



How do you create cubes SSAS SQL Server Analysis Services?

You want your cube to support localization ?

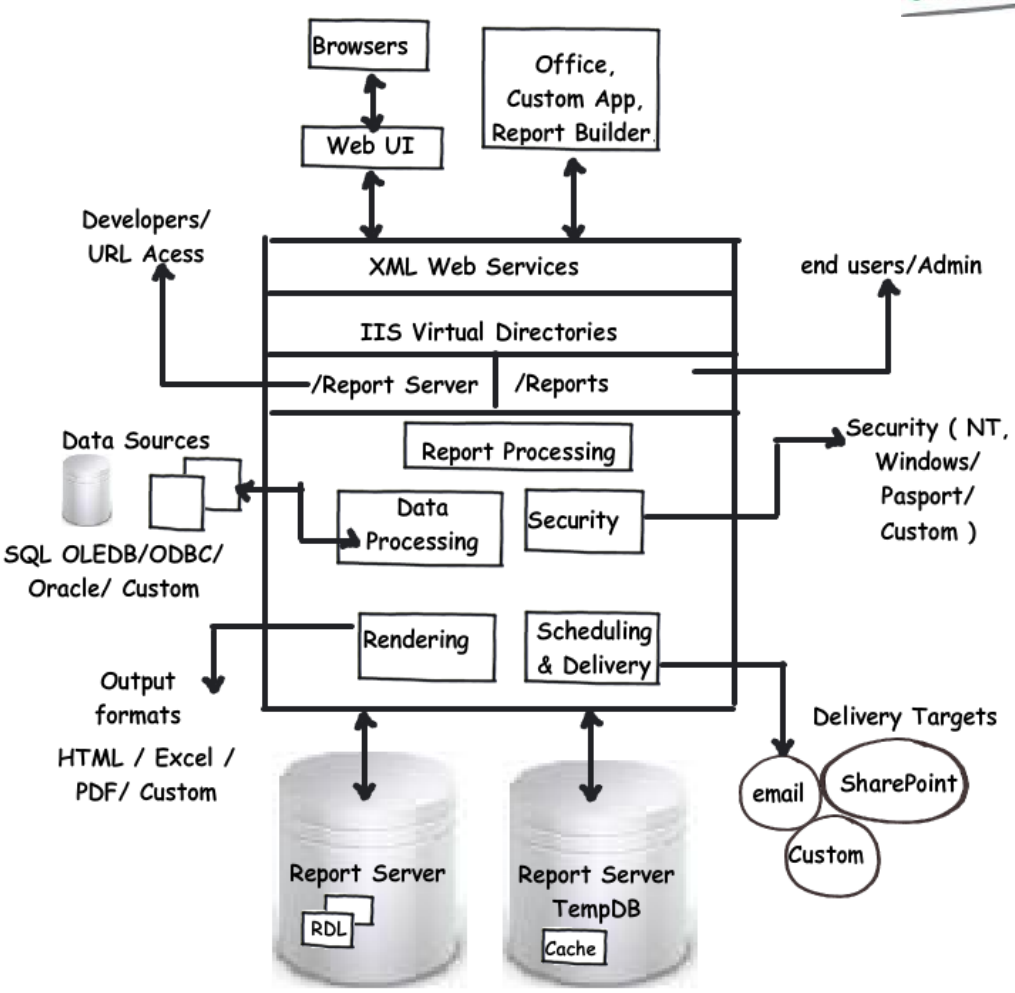
What kind of tables will go in fact and dimension tables ?

In what kind of scenario will you use a KPI ?

How can you create a pre-calculated measure in SSAS ?

Chapter 6:- Business intelligence (SSRS)

Can you explain SSRS architecture?



SSRS Architecture

Chapter 2:- SQL Server 2012

What are the new features which are added in SQL Server 2012?

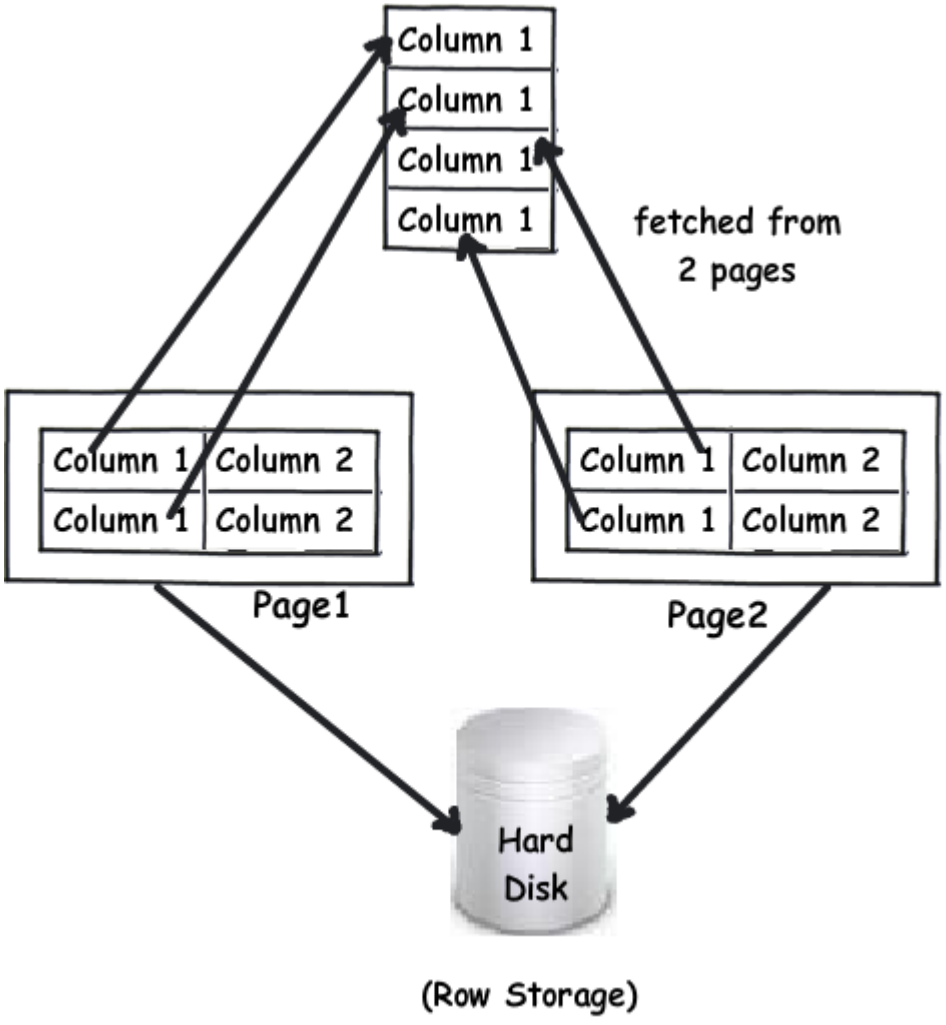
- Column store indexes.

What are column store indexes?

Relational database store data “row wise”. These rows are further stored in 8 KB page size.
For instance you can see in the below figure we have table with two columns “Column1” and “Column2”. You can see how the data is stored in two pages i.e. “page1” and “page2”. “Page1” has two rows and “page2” also has two rows.

Now if you want to fetch only “column1”, you have to pull records from two pages i.e. “Page1” and “Page2”, see below for the visuals.

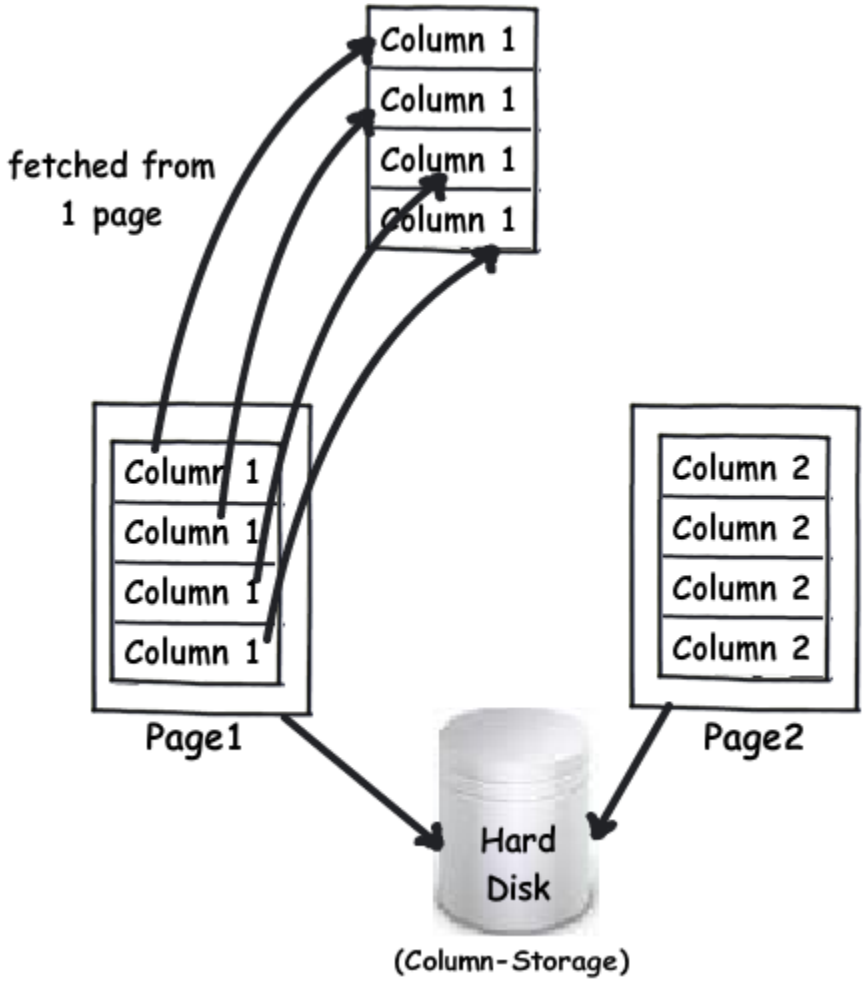
As we have to fetch data from two pages its bit performance intensive.



(Row Storage)

If somehow we can store data column wise we can avoid fetching data from multiple pages. That’s what column store indexes do. When you create a column store index it stores same column data in the same page.

You can see from the below visuals, we now need to fetch "column1" data only from one page rather than querying multiple pages.



What are the other benefits of column stored indexes?

There are two benefits of column store indexes:-

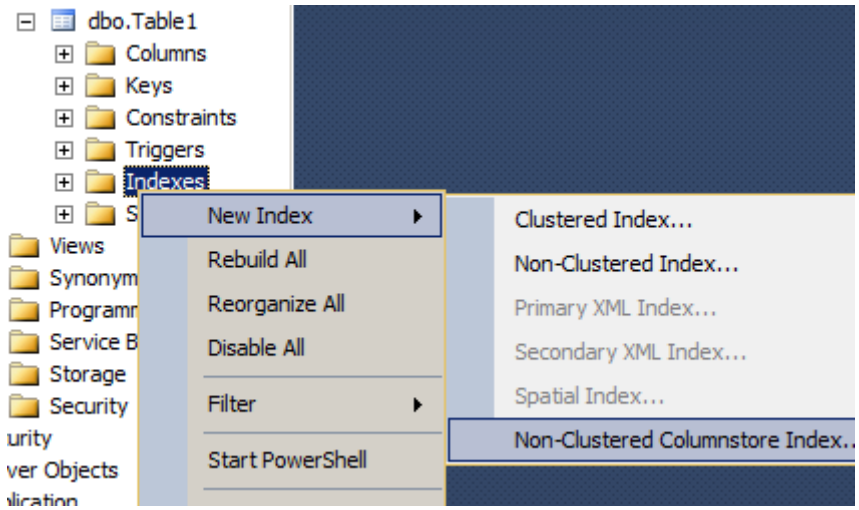
- Increases select performance (already discussed in the previous question).
- Efficient compression due repeating and similar data.

Let's discuss the second point. When data is stored row wise it's not ideal for compressing data as you get disparate data from various columns. When you create

column store indexes compression algorithm exploits repeating/ similar data and makes compression more efficient. You can expect 50% or more benefit in space saving when you use column store indexes.

How can we create column store indexes?

To create column store indexes, expand the table, right click on the indexes folder and click on "Non-Clustered column stored index". Once you click on this menu you can specify the columns on which column store indexes can be applied. See the below figure for visuals.



Are there any limitations of Column store indexes?

Column stored indexes was mainly created for OLAP applications to increase the select performance. Due to this focused implementation below are some the important limitations:-

- Insert, update and delete SQL commands do not work on table which have column store indexes.
- Column store indexes do not support the following data types :-
    - decimal greater than 18 digits
    - binary and varbinary
    - BLOB
    - CLR
    - (n)varchar(max)

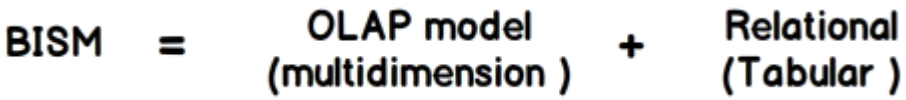
- datetimeoffset with precision greater than 2

- Replication cannot be implemented.
- Indexed views cannot be applied.

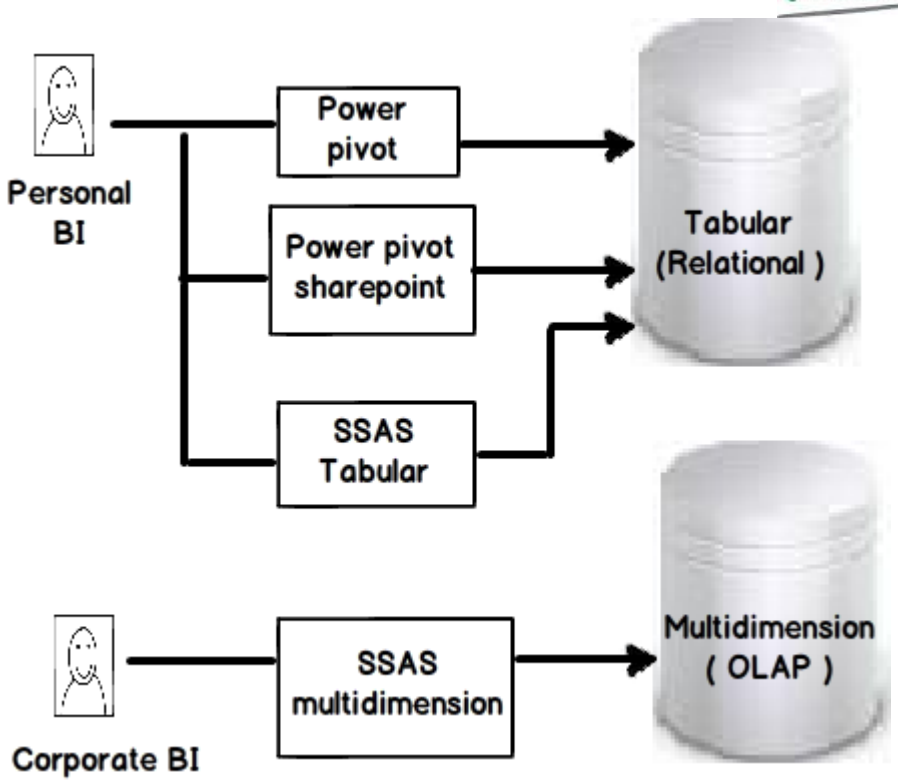
What is BISM in SQL Server 2012?

BISM stands for Business intelligence semantic model. It's actually unifies Multi-dimensional model plus tabular model. In the previous SSAS (SQL Server analysis) version if we want to do analysis, your database design structure needs to comply with OLAP model. In other words we need to define the structure in Fact and measures (star schema or snow flake).

But then there are certain classes (developers, people working with excel) of people who are very comfortable with tabular formats. Where people think in terms of tables and relationships. That's where tabular model is introduced.



So if you are pure corporate BI guy, use your Multi-dimensional model and if you are happy going personal BI person (developer, simple end user who uses excel) you can use Tabular model



What are Sequence objects ?

A sequence object generates sequence of unique numeric values as per specifications. Many developers would have now got a thought we have something similar like this called "Identity" columns. But the big difference is sequence object is independent of a table while identity columns are attached to a table.
Below is a simple code to create a sequence object. You can see we have created a sequence object called as "MySeq" with the following specification:-

- Starts with value 1.
- Increments with value 1
- Minimum value it should start is with zero.
- Maximum it will go to 100.
- No cycle defines that once it reaches 100 it will throw an error. If you want to restart it from 0 you should provide "cycle".
- "cache 50" specifies that till 50 the values are already incremented in to cache to reduce IO. If you specify "no cache" it will make input output on the disk.

```
create sequence MySeq as int
      start with 1  -- Start with value 1
      increment by 1-- Increment with value 1
```

```
      minvalue 0 -- Minimum value to start is zero
      maxvalue 100 -- Maximum it can go to 100
      no cycle -- Do not go above 100
      cache 50 -- Increment 50 values in memory rather than
incrementing from IO
```

To increment the value we need to call the below select statement. This is one more big difference as compared to identity.In identity the values increment when rows are added here we need to make an explicit call.

```
SELECT NEXT VALUE FOR dbo.MySequence AS seq_no;
```

What is the use of "OFFSET" and "FETCH" commands?

These are new commands in SQL Server 2012. They help to do pagination. See the next question which answers the same in more detail.

How to do pagination in SQL Server ?

There are instances when you want to display large result sets to the end user. The best way to display large result set is to split them i.e. apply pagination. So developers had their own hacky ways of achieving pagination using "top", "row_number" command etc. But from SQL Server 2012 onwards we can do pagination by using "OFFSET" and "FETCH' commands.
For instance let's says we have the following customer table which has 12 records. We would like to split the records in to 6 and 6.

CustomerCode	CustomerName
1001	shiv
1002	Raju
1003	Anil
1004	Jaideep
1005	Anil
1006	Raju
1007	Rony
1008	Rodney
1009	Vinod
1010	Ajay
1011	Rakesh
1012	Amit

So doing pagination is a two-step process:-

- First mark the start of the row by using "OFFSET" command.

- Second specify how many rows you want to fetch by using “FETCH” command.

You can see in the below code snippet we have used “OFFSET” to mark the start of row from “0”position. A very important note order by clause is compulsory for “OFFSET” command.

```
select * from
tblcustomer order by customercode
offset 0 rows
```

In the below code snippet we have specified we want to fetch “6” rows.

```
fetch next 6 rows only
```

Now if you run the above SQL you should see 6 rows.

CustomerCode	CustomerName
1001	shiv
1002	Raju
1003	Anil
1004	Jaideep
1005	Anil
1006	Raju

To fetch the next 6 rows just change your “OFFSET” position. You can see in the below code snippet I have modified the offset to 6. That means the row start position will from “6”.

```
select * from
tblcustomer order by customercode
offset 6 rows

fetch next 6 rows only
```

The above code snippet displays the next “6” records , below is how the output looks.

CustomerCode	CustomerName
1007	Rony
1008	Rodney
1009	Vinod
1010	Ajay
1011	Rakesh
1012	Amit

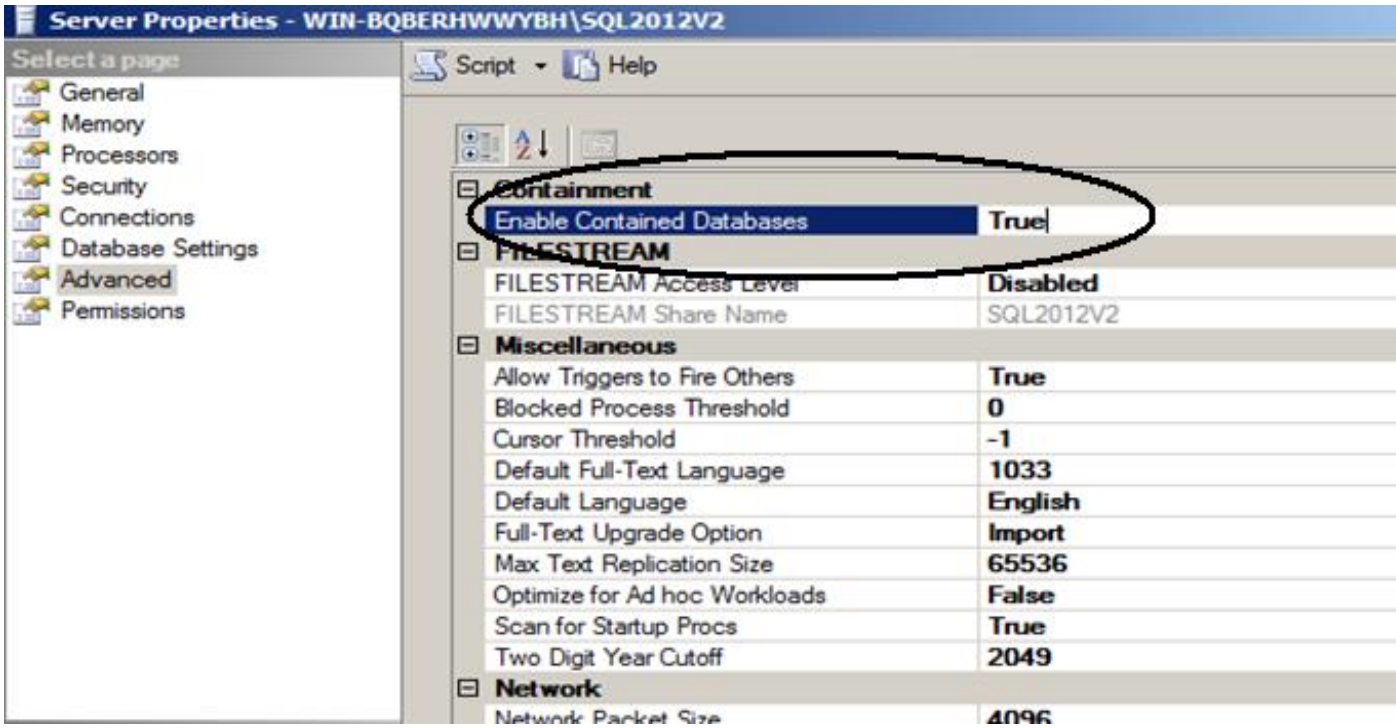
What are contained databases in SQL Server?

This is a great feature for people who have to go through pain of SQL Server database migration again and again. One of the biggest pains in migrating databases is user accounts. SQL Server user resides either in windows ADS or at SQL Server level as SQL Server users. So when we migrate SQL Server database from one server to other server these users have to be recreated again. If you have lot's of users you would need one dedicated person sitting creating one's for you.

So one of the requirements from easy migration perspective is to create databases which are self-contained. In other words, can we have a database with meta-data information, security information etc with in the database itself. So that when we migrate the database, we migrate everything with it. There's where "Contained" database where introduced in SQL Server 2012.

Creating contained database is a 3 step process:-

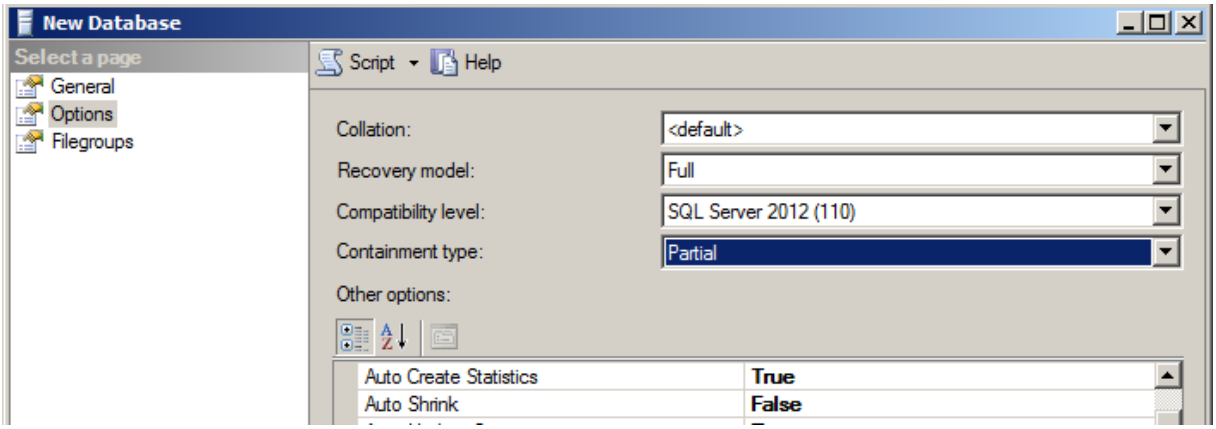
Step 1:- First thing is to enable contained database at SQL Server instance level. You can do the same by right clicking on the SQL Server instance and setting "Enabled Contained Database" to "true".



You can achieve the same by using the below SQL statements as well.

```
sp_configure 'show advanced options',1
GO
RECONFIGURE WITH OVERRIDE
GO
sp_configure 'contained database authentication', 1
GO
RECONFIGURE WITH OVERRIDE
GO
```

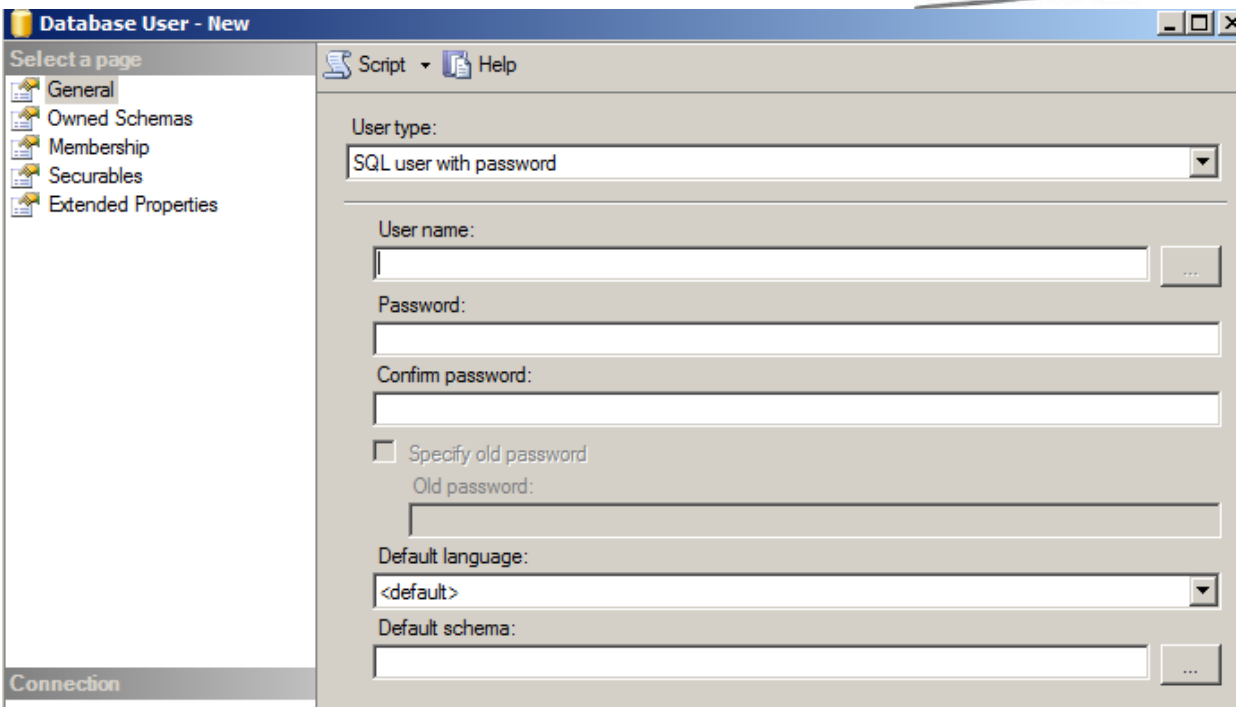
Step 2:- The next step is to enable contained database at database level. So when create a new database set “Containment type” to partial as shown in the below figure.



You can also create database with “containment” set to “partial” using the below SQL code.

```
CREATE DATABASE [MyDb]
CONTAINMENT = PARTIAL
ON PRIMARY
( NAME = N'My', FILENAME = N'C:\My.mdf')
LOG ON
( NAME = N'My_log', FILENAME =N'C:\My_log.ldf')
```

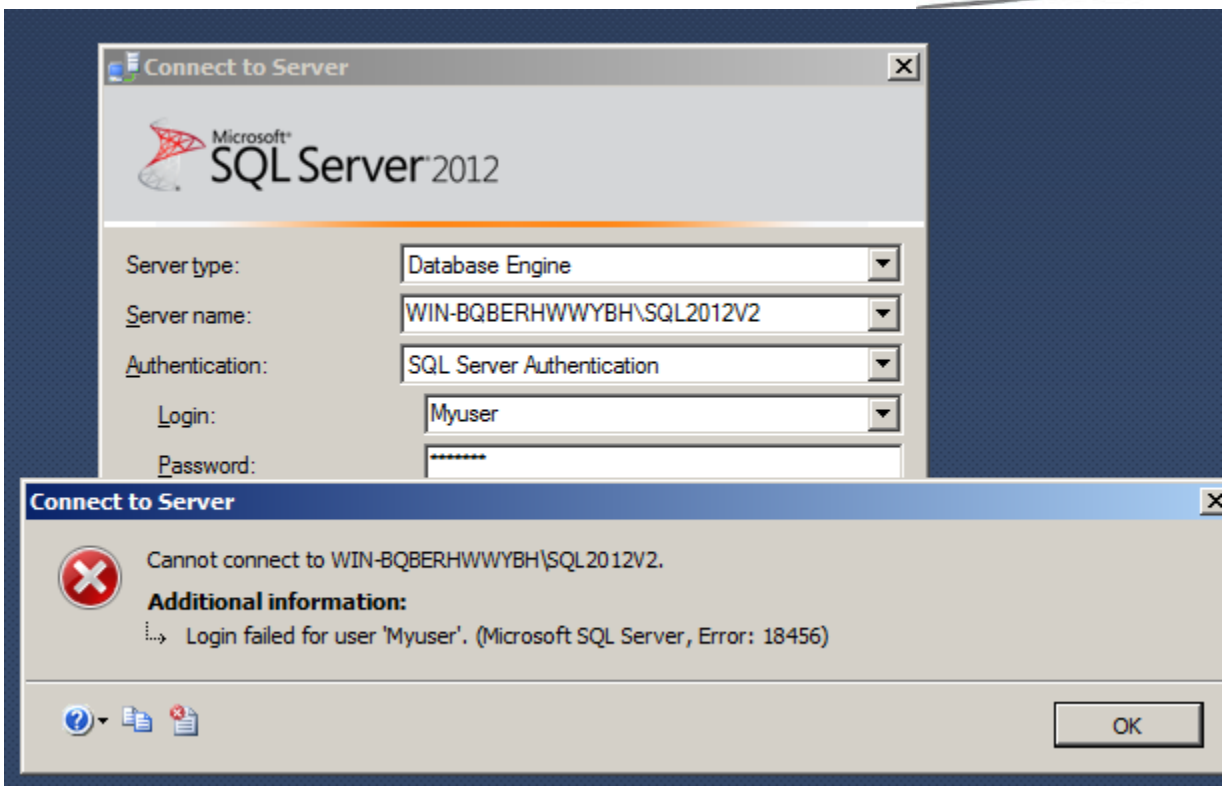
Step 3:- The final thing now is to test if “contained” database fundamental is working or not. Now we want the user credentials to be part of the database , so we need to create user as “SQL User with password”.



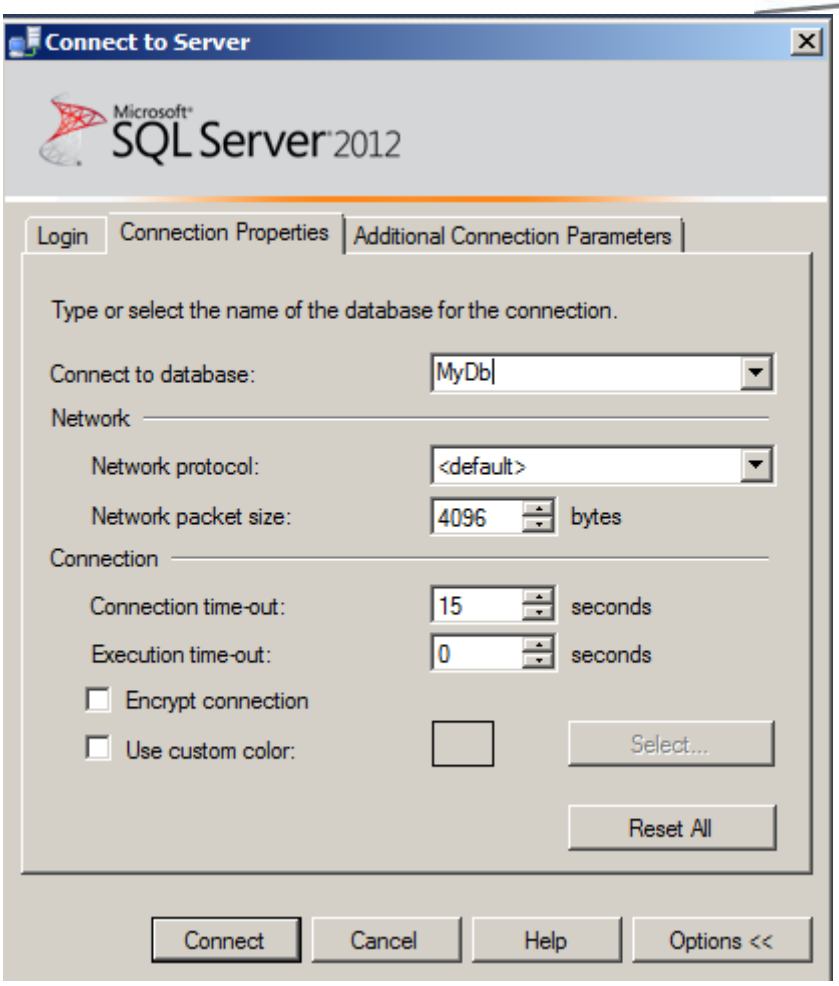
You can achieve the same by using the below script.

```
CREATE USER MyUser
WITH PASSWORD = 'pass@123';
GO
```

Now if you try to login with the user created, you get an error as shown in the below figure. This proves that the user is not available at SQL Server level.



Now click on options and specify the database name in “connect to database” , you should be able to login , which proves that user is part of database and not SQL Server instance.



(Q) What is Collation in SQL Server?

Collation refers to a set of rules that determine how data is sorted and compared. Character data is sorted using rules that define the correct character sequence, with options for specifying case-sensitivity, accent marks, kana character types, and character width.

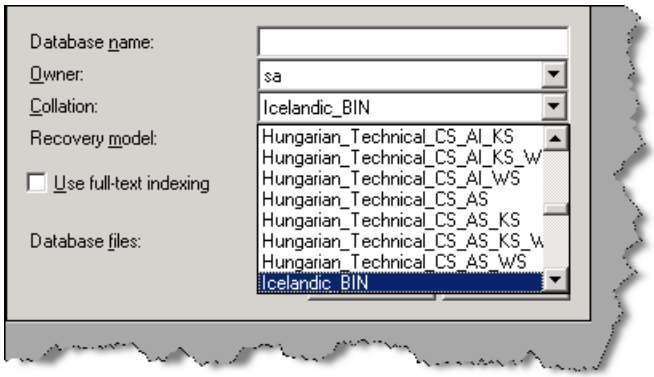


Figure 1.19: - Collation according to language

Note:- Different language will have different sort orders.

Case sensitivity

If A and a, B and b, etc. are treated in the same way then it is case-insensitive. A computer treats A and a differently because it uses ASCII code to differentiate the input. The ASCII value of A is 65, while A is 97. The ASCII value of B is 66 and b is 98.

Accent sensitivity

If “a” and “A”, o and “O” are treated in the same way, then it is accent-insensitive. A computer treats “a” and “A” differently because it uses ASCII code for differentiating the input. The ASCII value of “a” is 97 and “A” 225. The ASCII value of “o” is 111 and “O” is 243.

Kana Sensitivity

When Japanese kana characters Hiragana and Katakana are treated differently, it is called Kana sensitive.

Width sensitivity

When a single-byte character (half-width) and the same character when represented as a double-byte character (full-width) are treated differently then it is width sensitive.

(DB) Can we have a different collation for database and table?

Yes, you can specify different collation sequence for both the entity differently.

Chapter 2: SQL

Note: - This is one of the crazy things which I did not want to put in my book. But when I did sampling of some real interviews conducted across companies I was stunned to find some interviewer judging developers on syntaxes. I know many people will conclude this is childish but it’s the interviewer’s decision. If you think that this chapter is not useful you can happily skip it. But I think on fresher’s level they should not

Note: - I will be heavily using the "AdventureWorks" database which is a sample database shipped (in previous version we had the famous'NorthWind' database sample) with SQL Server 2005. Below is a view expanded from "SQL Server Management Studio".

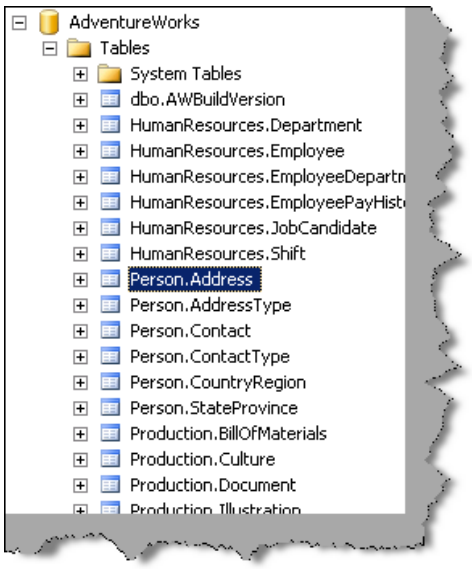


Figure 2.1: - AdventureWorks

(Q) Revisiting basic syntax of SQL?

```
CREATE TABLE ColorTable
(code VARCHAR(2),
 ColorValue VARCHAR(16)
)
INSERT INTO ColorTable (code, colorvalue) VALUES ('b1', 'Brown')
DELETE FROM ColorTable WHERE code = 'b1'
UPDATE ColorTable SET colorvalue ='Black' where code='bl'
DROP TABLE table-name {CASCADE|RESTRICT}
GRANT SELECT ON ColorTable TO SHIVKOIRALA WITH GRANT OPTION
REVOKE SELECT, INSERT, UPDATE (ColorCode) ON ColorTable FROM
Shivkoirala
COMMIT [WORK]
ROLLBACK [WORK]
```

```
Select * from Person.Address
Select AddressLine1, City from Person.Address
Select AddressLine1, City from Person.Address where city ='Sammamish'
```

(Q) What are "GRANT" and "REVOKE' statements?

GRANT statement grants rights to the objects (table). While revoke does the vice-versa of it, it removes rights from the object.

(Q) What is Cascade and Restrict in DROP table SQL?

Twist: - What is "ON DELETE CASCADE" and "ON DELETE RESTRICT"?

RESTRICT specifies that table should not be dropped if any dependencies (i.e. triggers, stored procedure, primary key, foreign key etc) exist. Therefore, if there are dependencies then error is generated and the object is not dropped.

CASCADE specifies that even if there dependencies go ahead with the drop. That means drop the dependencies first and then the main object. So if the table has stored procedures and keys (primary and secondary keys) they are dropped first and then the table is finally dropped.

(Q) How to import table using "INSERT" statement?

I have made a new temporary color table which is flourished using the below SQL. Structures of both the table should be same in order that this SQL executes properly.

```
INSERT INTO TempColorTable
SELECT code,ColorValue
FROM ColorTable
```

(Q) What is a DDL, DML and DCL concept in RDBMS world?

DDL (Data definition language) defines your database structure. CREATE and ALTER are DDL statements as they affect the way your database structure is organized.

DML (Data Manipulation Language) lets you do basic functionalities like INSERT, UPDATE, DELETE and MODIFY data in database.

DCL (Data Control Language) controls you DML and DDL statements so that your data is protected and has consistency. COMITT and ROLLBACK are DCL control statements. DCL guarantees ACID fundamentals of a transaction.

```
Note: - Refer to "Transaction and Locks" chapter.
```

(Q) What are different types of joins in SQL?

INNER JOIN

Inner join shows matches only when they exist in both tables. Example in the below SQL there are two tables Customers and Orders and the inner join in made on Customers.Customerid and

Orders.Customerid. So this SQL will only give you result with customers who have orders. If the customer does not have order, it will not display that record.

```
SELECT Customers.*, Orders.* FROM Customers INNER JOIN Orders ON
Customers.CustomerID =Orders.CustomerID
```

LEFT OUTER JOIN

Left join will display all records in left table of the SQL statement. In SQL below customers with or without orders will be displayed. Order data for customers without orders appears as NULL values. For example, you want to determine the amount ordered by each customer and you need to see who has not ordered anything as well. You can also see the LEFT OUTER JOIN as a mirror image of the RIGHT OUTER JOIN (Is covered in the next section) if you switch the side of each table.

```
SELECT Customers.*, Orders.* FROM Customers LEFT OUTER JOIN Orders ON
Customers.CustomerID =Orders.CustomerID
```

RIGHT OUTER JOIN

Right join will display all records in right table of the SQL statement. In SQL below all orders with or without matching customer records will be displayed. Customer data for orders without customers appears as NULL values. For example, you want to determine if there are any orders in the data with undefined CustomerID values (say, after a conversion or something like it). You can also see the RIGHT OUTER JOIN as a mirror image of the LEFT OUTER JOIN if you switch the side of each table.

```
SELECT Customers.*, Orders.* FROM Customers RIGHT OUTER JOIN Orders ON
Customers.CustomerID =Orders.CustomerID
```

(Q) What is "CROSS JOIN"?

Twist: - What is Cartesian product?

"CROSS JOIN" or "CARTESIAN PRODUCT" combines all rows from both tables. Number of rows will be product of the number of rows in each table. In real life scenario, I cannot imagine where we will want to use a Cartesian product. However, there are scenarios where we would like permutation and combination probably Cartesian would be the easiest way to achieve it.

(Q) You want to select the first record in a given set of rows?

```
Select top 1 * from sales. salesperson
```

(Q) How do you sort in SQL?

Using the "ORDER BY" clause, you either sort the data in ascending manner or descending manner.

```
select  * from sales.salesperson order by salespersonid asc
```

```
select  * from sales.salesperson order by salespersonid desc
```

(Q) How do you select unique rows using SQL?

Using the "DISTINCT" clause. For example if you fire the below give SQL in "AdventureWorks" , first SQL will give you distinct values for cities , while the other will give you distinct rows.

```
select distinct city  from person.address

select distinct *  from person.address
```

(Q) Can you name some aggregate function is SQL Server?

Some of them which every interviewer will expect:-

- **AVG: -** Computes the average of a specific set of values, which can be an expression list or a set of data records in a table.
- **SUM**: - Returns the sum of a specific set of values, which can be an expression list or a set of data records in a table.
- **COUNT: -** Computes the number of data records in a table.
- **MAX: -** Returns the maximum value from a specific set of values, which can be an expression list or a set of data records in a table.
- **MIN: -** Returns the minimum value from a specific set of values, which can be an expression list or a set of data records in a table.

(Q) What is the default "SORT" order for a SQL?

ASCENDING

(Q) What is a self-join?

If you want to join, two instances of the same table you can use self-join.

What is the difference between DELETE and TRUNCATE?

	Delete	**Truncate**
Filters	In delete we can specify a where clause.	Truncate deletes all records, we can not specify filters.
Data removal procedure	Delete removes data one row at a time.	Truncate removes data by removing pages. Pages are 8 KB units which stores row data.
Performance	Slower, as delete happens row wise.	Truncate is faster as compared to delete as data is removed in pages.
Triggers	Triggers are executed in delete as the data is removed	In truncate triggers are disabled.

	row wise.	
Identity	In delete the identity value is retained.	In truncate the identity is reset back to zero.

(Q) Select addresses which are between '1/1/2004' and '1/4/2004'?

Select * from Person.Address where modifieddate between '1/1/2004' and '1/4/2004'

(Q) What are Wildcard operators in SQL Server?

Twist: - What is like clause in SQL?

Note: - For showing how the wildcards work I will be using the "person.address" table in adventureworks.

Basically, there are two types of operator:-

- **"%" operator (During Interview you can spell it as "Percentage Operator").**

"%" operator searches for one or many occurrences. So when you fire a query using "%" SQL Server searches for one or many occurrences. In the below SQL I have applied "%" operator to "S" character.

Select AddressLine1 from person.address where AddressLine1 like 'S%'

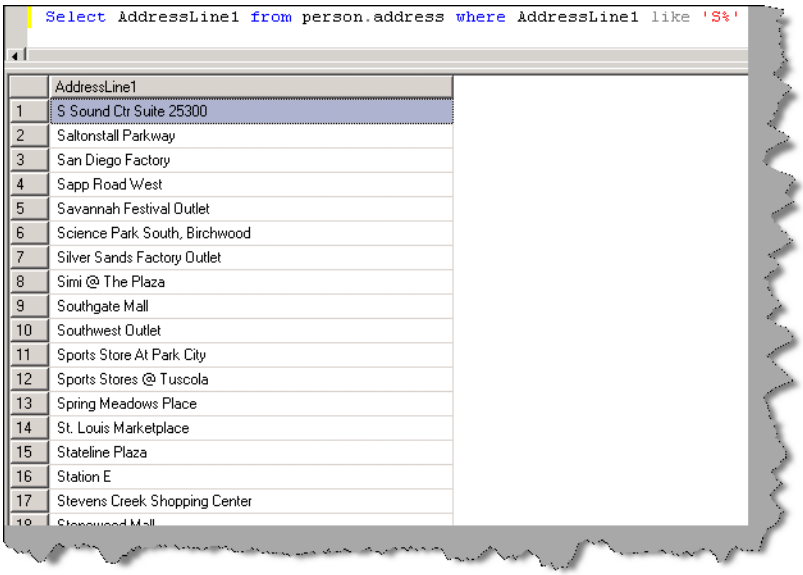


Figure 2.2: - "%" operator in action.

- **"_" operator (During Interview you spell it as "Underscore Operator").**

"_" operator is the character defined at that point. In the below sample I have fired a query

Select AddressLine1 from person.address where AddressLine1 like '_h%'

So all data where second letter is “h” is returned.

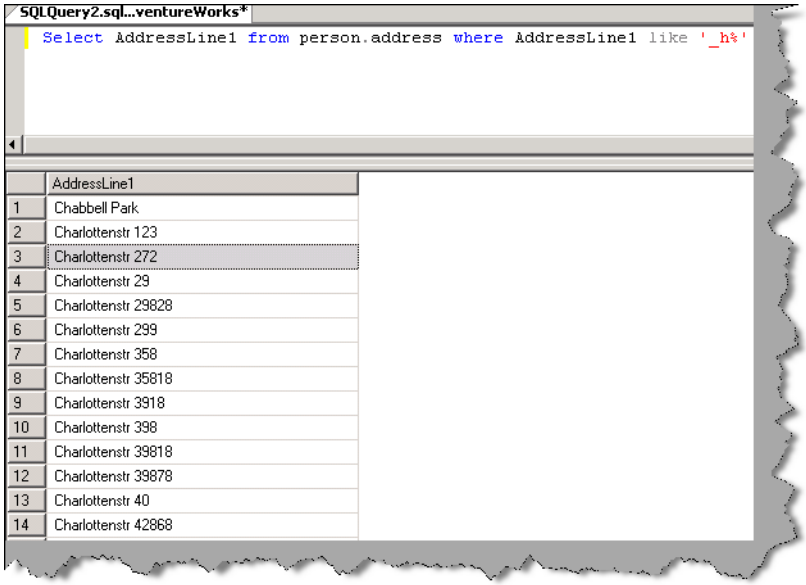


Figure 2.3: - “_” operator in action

(Q) What is the difference between “UNION” and “UNION ALL”?

UNION SQL syntax is used to select information from two tables. But it selects only distinct records from both the table. , while UNION ALL selects all records from both the tables.

To explain it practically below are two images one fires “UNION” and one “UNION ALL” on the “person.address” table of the “AdventureWorks” database.

```
Select * from person.address
Union
Select * from person.address
This returns 19614 rows (that’s mean it removes duplicates)
Select * from person.address
union all
Select * from person.address
```

This returns 39228 rows (“unionall” does not check for duplicates so returns double the record show up)

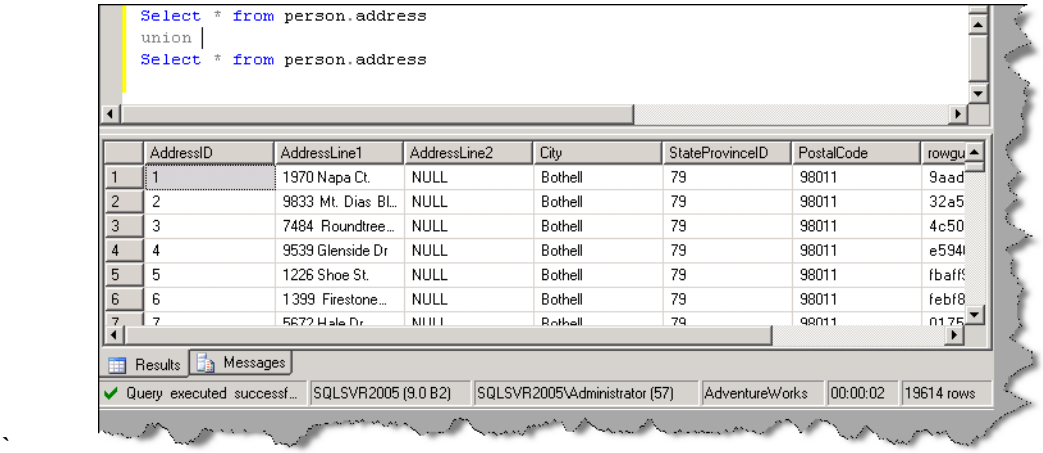


Figure 2.4: - Union keyword in action (19614 rows)

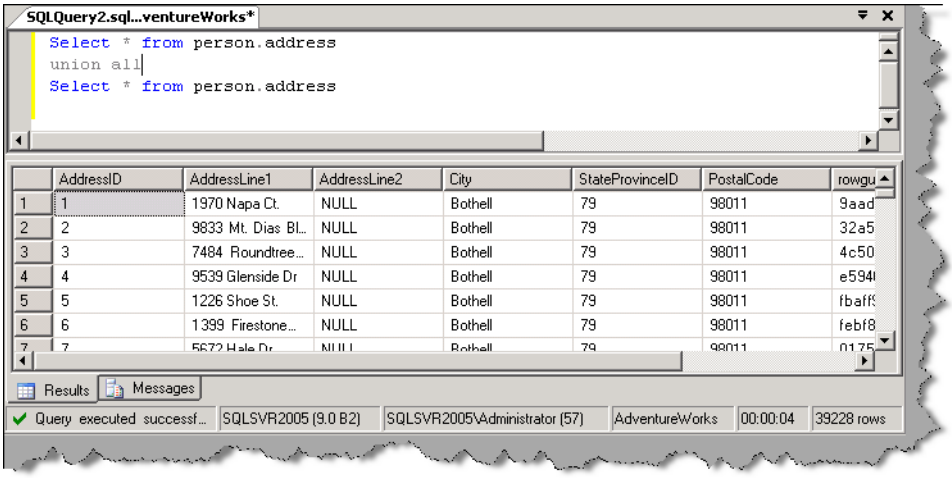


Figure 2.5: - Union All in action (39228 rows)

Note: - Selected records should have same data type or else the syntax will not work.

Note: - In the coming questions you will see some 5 to 6 questions on cursors. Though not a much discussed topic but still from my survey 5% of interviews have asked questions on cursors. So let’s leave no stone for the interviewer to reject us.

(Q) What are cursors and what are the situations you will use them?

SQL statements are good for set at a time operation. So it is good at handling set of data. But there are scenarios where you want to update row depending on certain criteria. You will loop through all rows and update data accordingly. There is where cursors come in to picture.

(Q) What are the steps to create a cursor?

Below are the basic steps to execute a cursor.

- Declare
- Open
- Fetch
- Operation
- Close and Deallocate

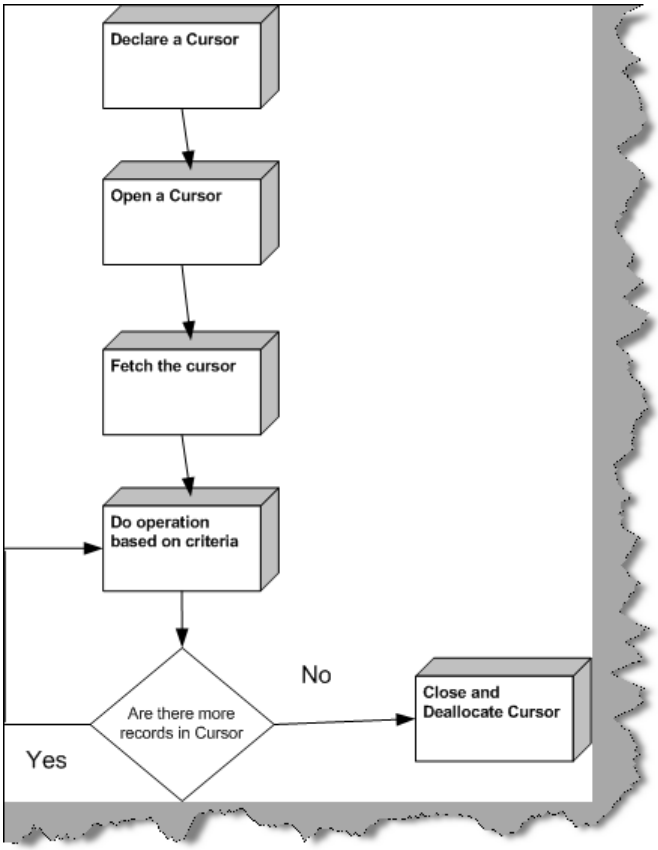


Figure 2.6: - Steps to process a cursor

This is a small sample, which uses the "person.address" class. This T-SQL program will only display records, which have "@Provinceid" equal to "7".

```
DECLARE @provinceid int
-- Declare Cursor
DECLARE provincecursor CURSOR FOR
SELECT stateprovinceid
FROM Person.Address
-- Open cursor
OPEN provincecursor
-- Fetch data from cursor in to variable
FETCH NEXT FROM provincecursor
INTO @provinceid
WHILE @@FETCH_STATUS = 0
BEGIN
-- Do operation according to row value
if @Provinceid=7
begin
PRINT @Provinceid
end
-- Fetch the next cursor
FETCH NEXT FROM provincecursor
INTO @provinceid
END
-- Finally do not forget to close and deallocate the cursor
CLOSE provincecursor
DEALLOCATE provincecursor
```

(Q) What are the different Cursor Types?

Cursor types are assigned when we declare a cursor.

```
DECLARE cursor_name CURSOR
[LOCAL | GLOBAL]
[FORWARD_ONLY | SCROLL]
[STATIC | KEYSET | DYNAMIC | FAST_FORWARD]
[READ_ONLY | SCROLL_LOCKS | OPTIMISTIC]
[TYPE_WARNING]
FOR select_statement
```

```
[FOR UPDATE [OF column_list]]
```

STATIC

STATIC cursor is a fixed snapshot of a set of rows. This fixed snapshot is stored in a temporary database. As the cursor is using private snapshot any changes to the set of rows external will not be visible in the cursor while browsing through it. You can define a static cursor using "STATIC" keyword.

```
DECLARE cusorname CURSOR STATIC
FOR SELECT * from tablename
WHERE column1 = 2
```

KEYSET

In KEYSET the key values of the rows are saved in tempdb. For instance, let us say the cursor has fetched the following below data. So only the "supplierid" will be stored in the database. Any new inserts happening is not reflected in the cursor. But any updates in the key-set values are reflected in the cursor. Because the cursor is identified by key values you can also absolutely fetch them using "FETCH ABSOLUTE 12 FROM mycursor"

SupplierId	Supplier Name
17	Evan and Evan limited
18	Hari brothers
19	European Suppliers
20	Stockers
21	New Suppliers
22	Sasan Enterprises

Figure 2.7: - Key Set Data

DYNAMIC

In DYNAMIC cursor you can see any kind of changes happening i.e. either inserting new records or changes in the existing and even deletes. That's why DYNAMIC cursors are slow and have least performance.

FORWARD_ONLY

As the name suggest they only move forward and only a one time fetch is done. In every fetch the cursor is evaluated. That means any changes to the data are known, until you have specified "STATIC" or "KEYSET".

FAST_FORWARD

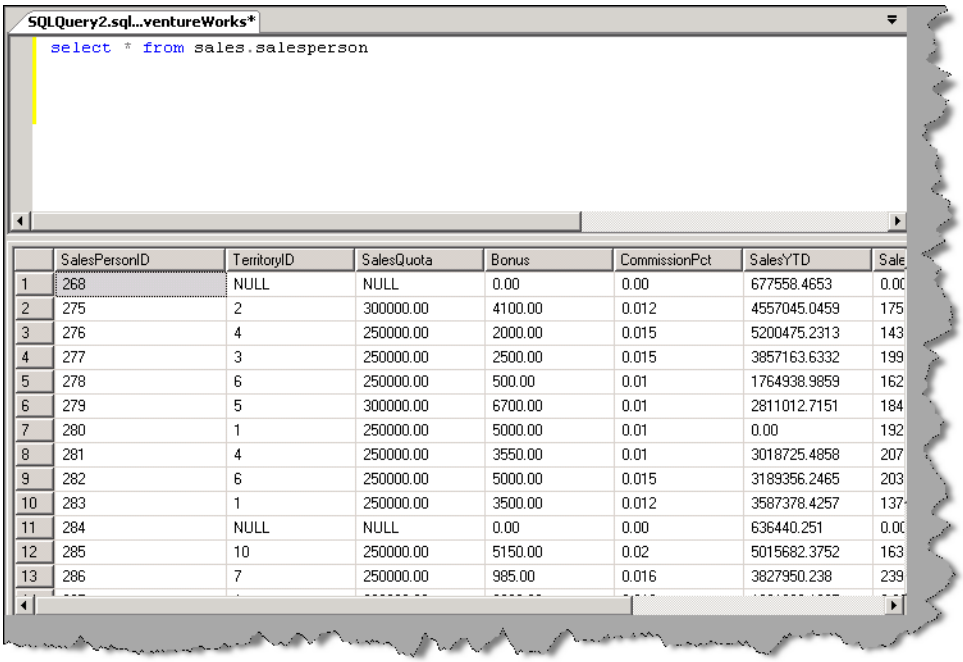
These types of cursor are forward only and read-only and in every fetch they are not re-evaluated again. This makes them a good choice to increase performance.

(Q) What are “Global” and “Local” cursors?

Cursors are global for a connection. By default cursors are global. That means you can declare a cursor in one stored procedure and access it outside also. Local cursors are accessible only inside the object (which can be a stored procedure, trigger or a function). You can declare a cursor as “Local” or “Global” in the “DECLARE” cursor syntax. Refer the “DECLARE” statement of the cursor in the previous sections.

(Q) What is “Group by” clause?

“Group by” clause group similar data so that aggregate values can be derived. In “AdventureWorks” there are two tables “Salesperson” and “Salesterritory”. In below figure “Actual data” is the complete view of “Salesperson”. But now we want a report that per territory wise how many sales people are there. So in the second figure I made a group by on territory id and used the “count” aggregate function to see some meaningful data. “Northwest” has the highest number of sales personnel.



	SalesPersonID	TerritoryID	SalesQuota	Bonus	CommissionPct	SalesYTD	Sale
1	268	NULL	NULL	0.00	0.00	677558.4653	0.00
2	275	2	300000.00	4100.00	0.012	4557045.0459	175
3	276	4	250000.00	2000.00	0.015	5200475.2313	143
4	277	3	250000.00	2500.00	0.015	3857163.6332	199
5	278	6	250000.00	500.00	0.01	1764938.9859	162
6	279	5	300000.00	6700.00	0.01	2811012.7151	184
7	280	1	250000.00	5000.00	0.01	0.00	192
8	281	4	250000.00	3550.00	0.01	3018725.4858	207
9	282	6	250000.00	5000.00	0.015	3189356.2465	203
10	283	1	250000.00	3500.00	0.012	3587378.4257	137
11	284	NULL	NULL	0.00	0.00	636440.251	0.00
12	285	10	250000.00	5150.00	0.02	5015682.3752	163
13	286	7	250000.00	985.00	0.016	3827950.238	239

Figure 2.8: - Actual Data

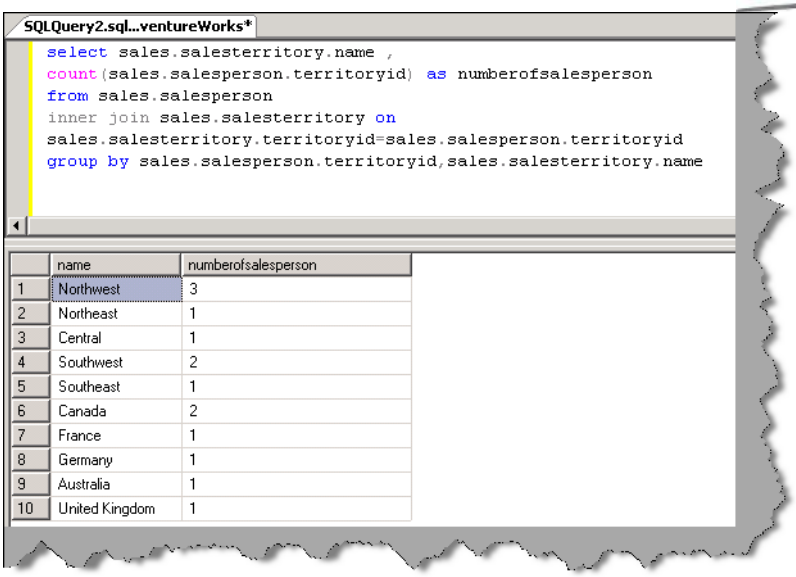
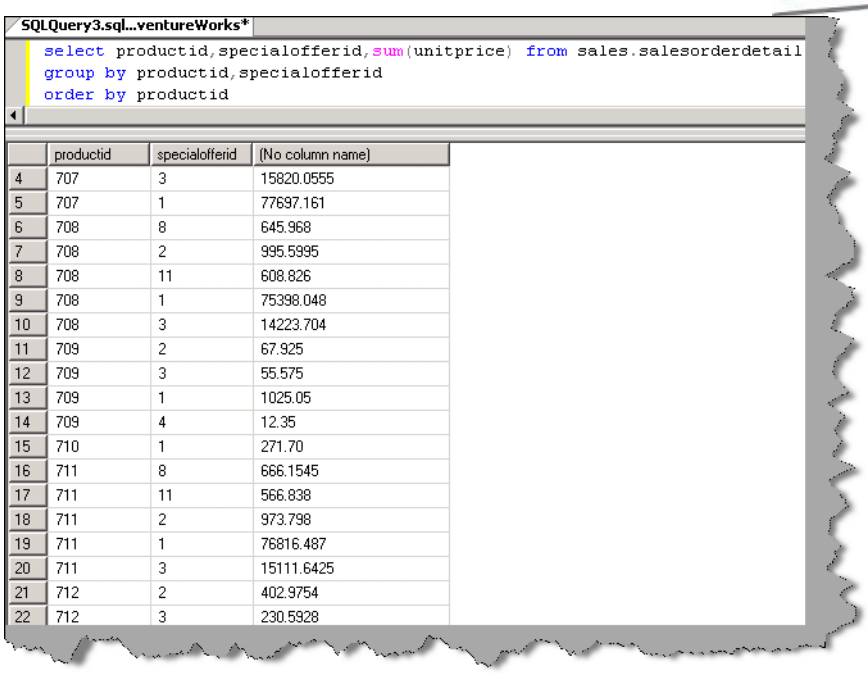


Figure 2.9: - Group by applied

(Q) What is ROLLUP?

ROLLUP enhances the total capabilities of "GROUP BY" clause.

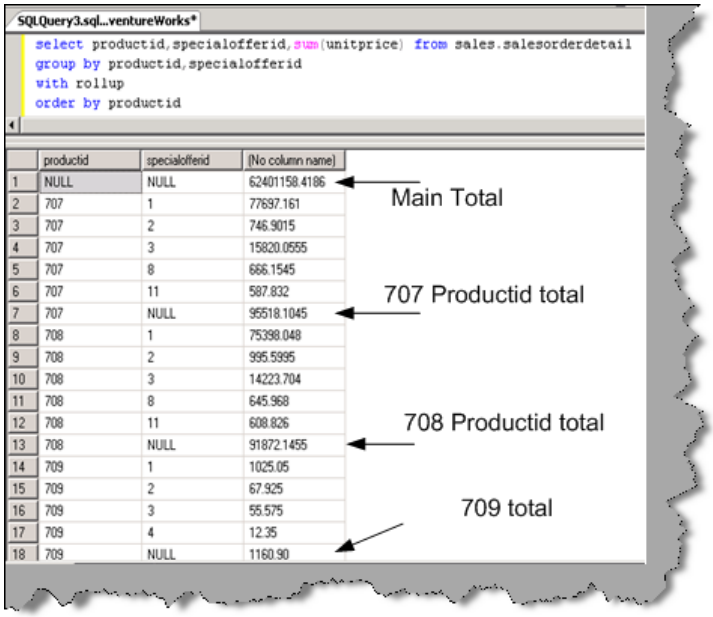
Below is a GROUP BY SQL, which is applied on "SalesorderDetail" on "Productid" and "Specialofferid". You can see 707,708,709 etc products grouped according to "Specialofferid" and the third column represents total according to each pair of "Productid" and "Specialofferid". Now you want to see sub-totals for each group of "Productid" and "Specialofferid"



	productid	specialofferid	(No column name)
4	707	3	15820.0555
5	707	1	77697.161
6	708	8	645.968
7	708	2	995.5995
8	708	11	608.826
9	708	1	75398.048
10	708	3	14223.704
11	709	2	67.925
12	709	3	55.575
13	709	1	1025.05
14	709	4	12.35
15	710	1	271.70
16	711	8	666.1545
17	711	11	566.838
18	711	2	973.798
19	711	1	76816.487
20	711	3	15111.6425
21	712	2	402.9754
22	712	3	230.5928

Figure 2.10: - Salesorder displayed with out ROLLUP

So after using ROLLUP you can see the sub-total. The first row is the grand total or the main total, followed by sub-totals according to each combination of "Productid" and "Specialofferid". ROLLUP retrieves a result set that contains aggregates for a hierarchy of values in selected columns.



	productid	specialofferid	(No column name)
1	NULL	NULL	62401158.4186
2	707	1	77697.161
3	707	2	746.9015
4	707	3	15820.0555
5	707	8	666.1545
6	707	11	587.832
7	707	NULL	95518.1045
8	708	1	75398.048
9	708	2	995.5995
10	708	3	14223.704
11	708	8	645.968
12	708	11	608.826
13	708	NULL	91872.1455
14	709	1	1025.05
15	709	2	67.925
16	709	3	55.575
17	709	4	12.35
18	709	NULL	1160.90

.**Figure 2.11: - Subtotal according to product using ROLLUP**

(Q) What is CUBE?

CUBE retrieves a result set that contains aggregates for all combinations of values in the selected columns. ROLLUP retrieves a result set that contains aggregates for a hierarchy of values in selected columns.

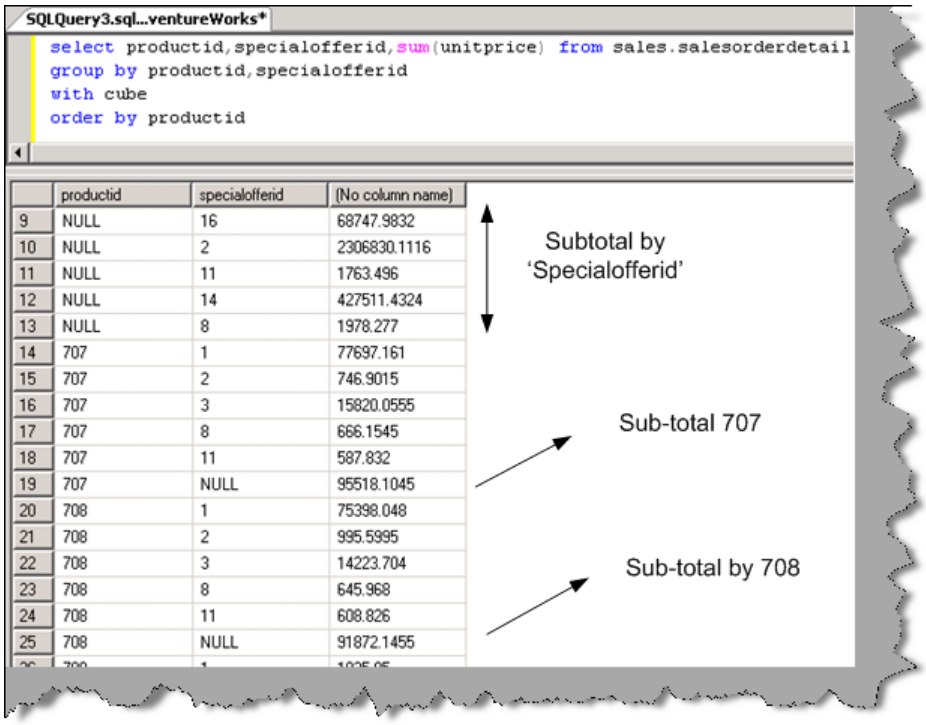


Figure 2.12: - CUBE in action

(Q) What is the difference between "HAVING" and "WHERE" clause?

"HAVING" clause is used to specify filtering criteria for "GROUP BY", while "WHERE" clause applies on normal SQL.

In the above example we if we want to filter on territory which has sales personnel count above 2.

```
select sales.salesterritory.name ,
count(sales.salesperson.territoryid) as numberofsalesperson
from sales.salesperson
inner join sales.salesterritory on
sales.salesterritory.territoryid=sales.salesperson.territoryid
group by sales.salesperson.territoryid,sales.salesterritory.name
```

```
having count(sales.salesperson.territoryid) >= 2
```

Note:- You can see the having clause applied. In this case you can not specify it with "WHERE" clause it will throw an error. In short "HAVING" clause applies filter on a group while "WHERE" clause on a simple SQL.

(Q) What is "COMPUTE" clause in SQL?

"COMPUTE "clause is used in SQL to produce subtotals for each group.

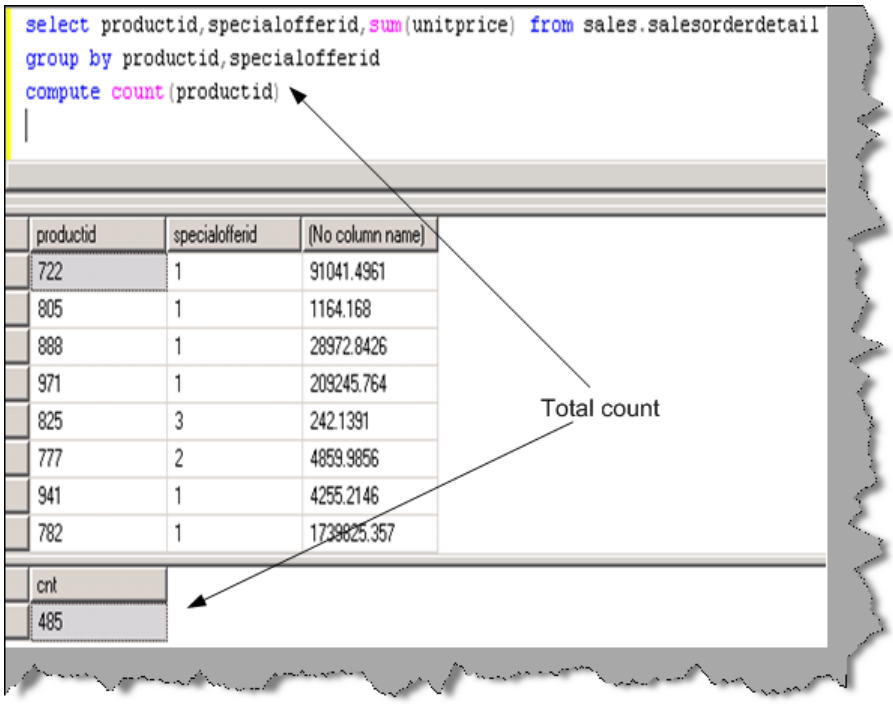


Figure 2.13: - "Compute" in action

(Q) What is "WITH TIES" clause in SQL?

"WITH TIES" clause specifies that additional rows be returned from the base result set with the same value in the ORDER BY columns appearing as the last of the TOP n (PERCENT) rows. So what does that sentence mean? See the below figure there are four products p1,p2,p3 and p4. "UnitCost" of p3 and p4 are same.

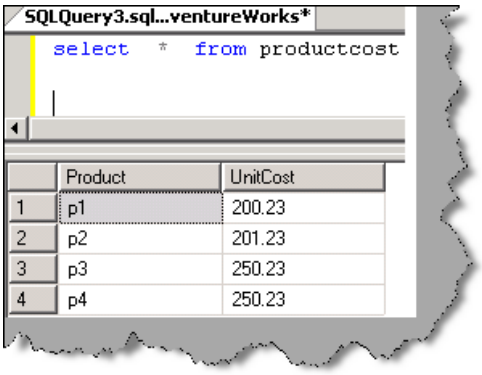


Figure 2.14: - Actual Data

So when we do a TOP 3 on the "ProductCost" table we will see three rows as show below. But even p3 has the same value as p4. SQL just took the TOP 1. So if you want to display tie up data like this you can use "WITH TIES".

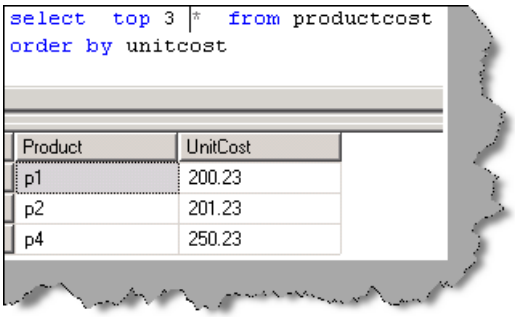


Figure 2.15: - TOP 3 from the "productcost" table

You can see after firing SQL with "WITH TIES" we are able to see all the products properly.

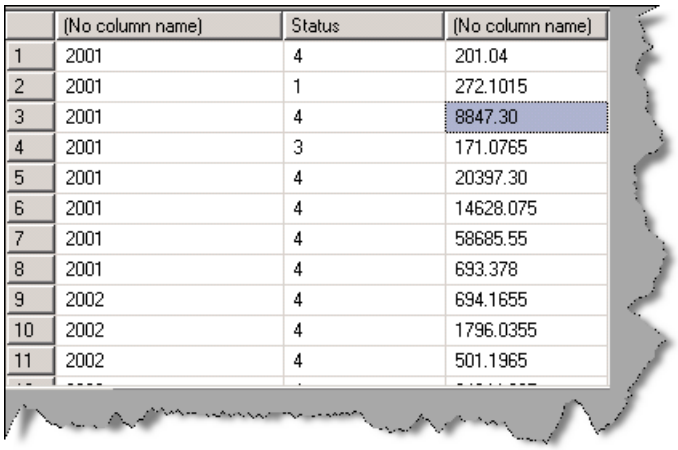


Figure 2.16: - WITH TIES in action

Note: - You should have an "ORDER CLAUSE" and "TOP" keyword specified or else "WITH TIES" is not of much use.

(Q) What does “SET ROWCOUNT” syntax achieves?

Twist: - What is the difference between “SET ROWCOUNT” and “TOP” clause in SQL?

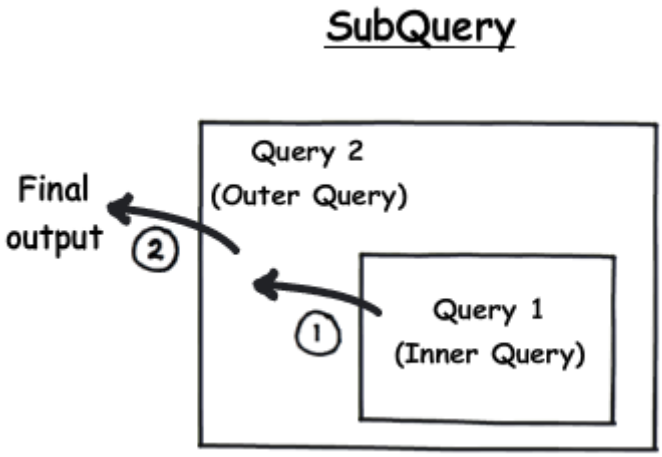
“SET ROWCOUNT” limits the number of rows returned. Its looks very similar to “TOP” clause, but there is a major difference the way SQL is executed. The major difference between “SET ROWCOUNT” and “TOP” SQL clause is following:-

“SET ROWCOUNT is applied before the order by clause is applied. So if "ORDER BY" clause is specified it will be terminated after the specified number of rows is selected. ORDER BY clause is not executed”

What are Sub-Queries ?

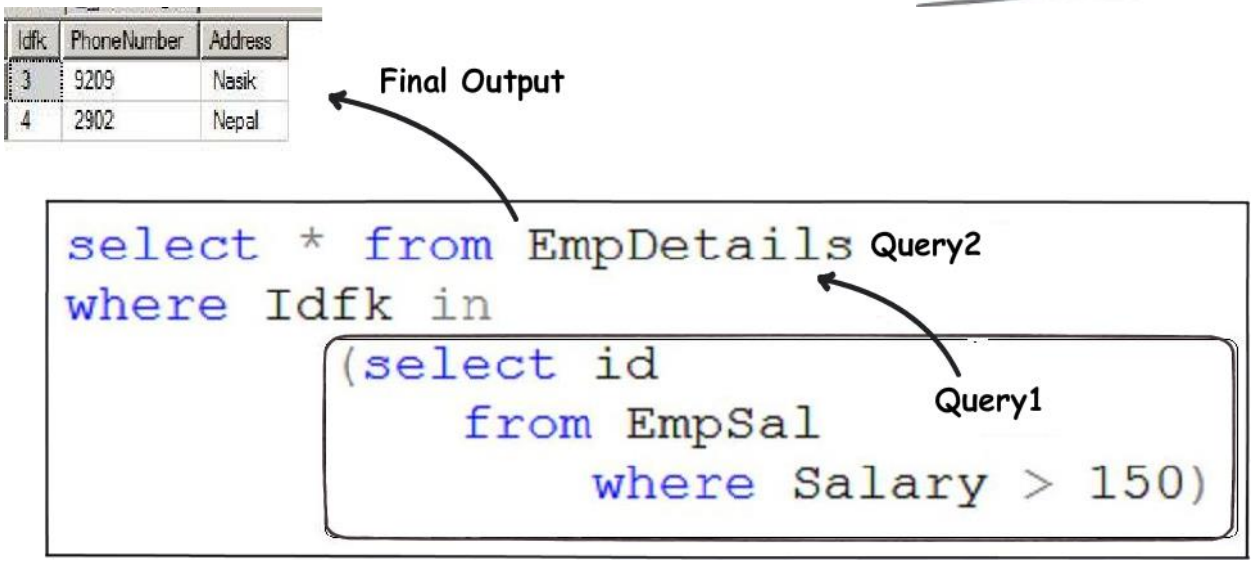
```
Note :- Subqueries are also termed as nested queries.
```

Subquery is a Query inside query. Many times we would like to have chain of SQL statements, where output of one SQL statement serves as an input to the other SQL statement.



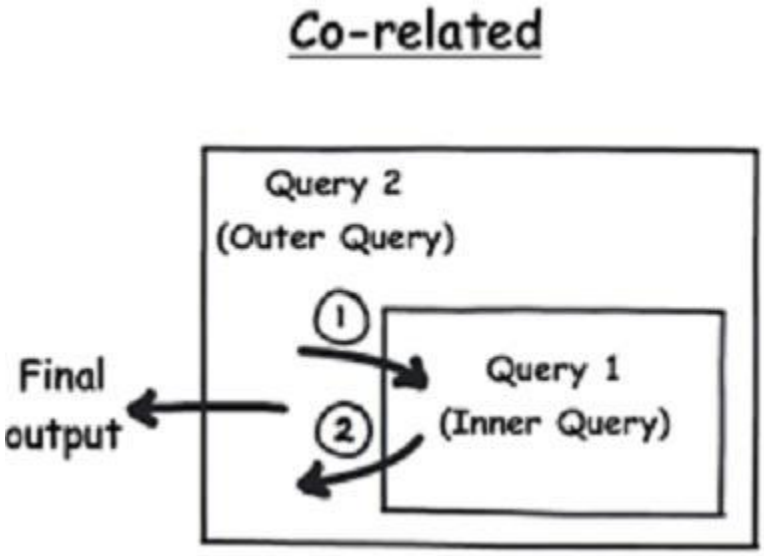
For example in the below figure you can see we have two queries (Query1 and Query 2). Query 1 (Inner query) fetches record whose salaries are greater than 150 and that is fed to Query 2 (outer query). The outer query takes data given by inner query and displays address and phone number.

This type of query is called as Sub query.



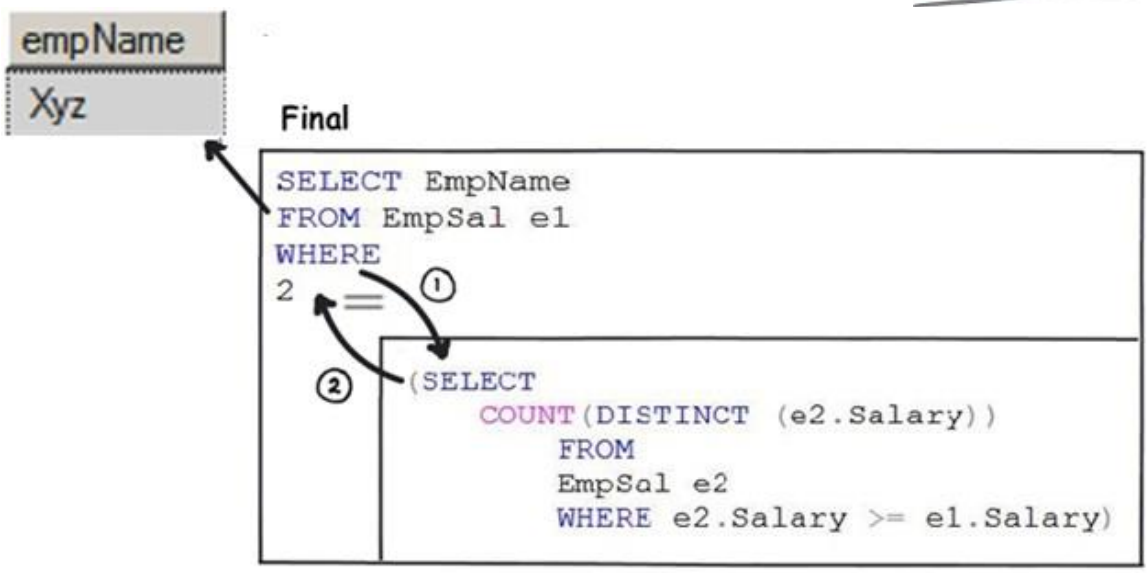
What are co-related queries?

Co-related queries are same like subquery i.e. its query inside query. But in correlated queries the data passes to and fro between inside and outside query.



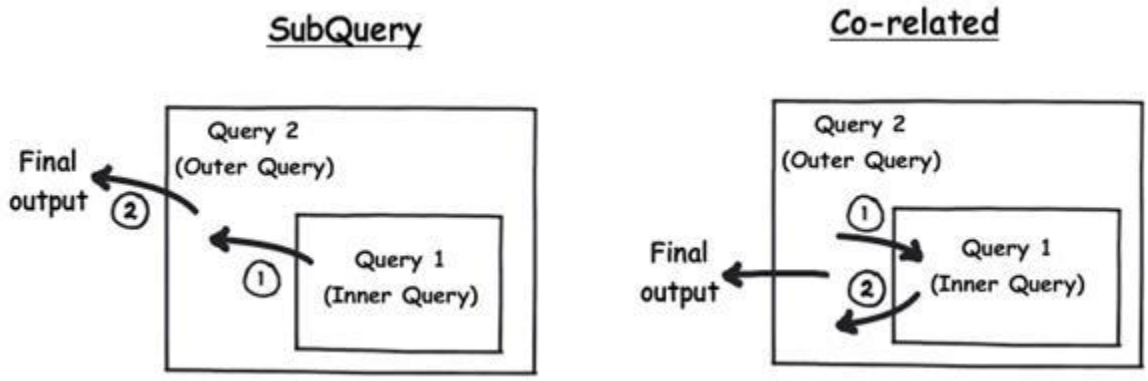
One of the scenarios where co-related queries comes handy is to find second highest, third lowest etc from a table. Below is a simple co-related query which finds the second highest from a table.

So first the record from the outer query is passed to the inner query, inner query then evaluates and if the evaluation is true then output is displayed or this process continues until all records are finished.



What is the difference between co-related query and sub query?

Sub-query	**Co-related query**
Inner query is self-contained and independent.	Inner query is dependent on outer query and references data from outer query.
In sub query data passes unidirectional i.e from inner query to outer query.	In co-related query first data passes from outer query to inner query, its evaluated and then the final output is displayed.



Can you explain Coalesce in SQL Server ?

Coalesce returns the first non-null column from more than one columns. For instance you can see in the below table either “FirstName” has null values or “SurName” as null value. Now you would like to pull values using coalesce function and it will return the column which will have null values.

FirstName	SurName
Prasad	NULL
Raju	NULL
NULL	Koirala
NULL	Shinde

So if you fire the below SQL statement with coalesce.

```
SELECT   coalesce(FirstName, SurName) as Name
FROM         tblPerson
```

So when:-

- If “Surname” is NULL and “FirstName” is not NULL , then “FirstName” is returned.
- If “FirstName”is NULL and “SurName” is not NULL , then “SurName” will be retuned.

Below is the output/

Name
Prasad
Raju
Koirala
Shinde

What is CTE (Common table expression)?

CTE is a temporary result set which can be used within a execution of a SINGLE insert,update,delete or select query.

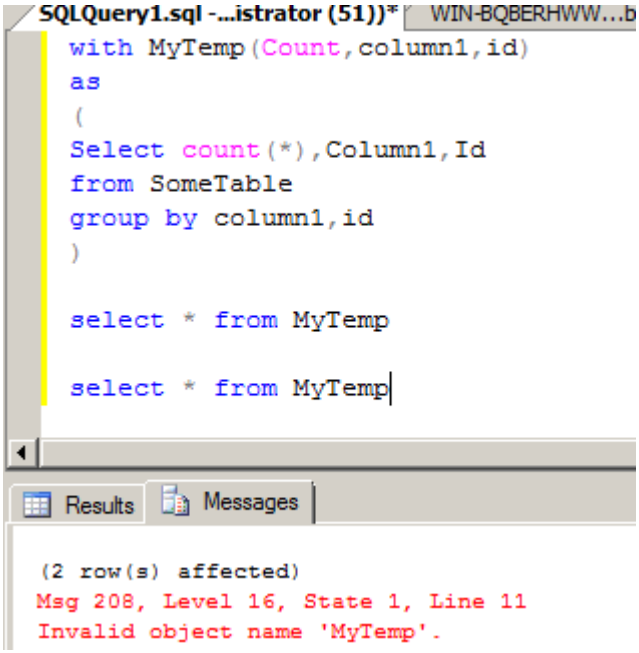
Using CTE is a 4 step process:-

- All CTE starts with “with” clause.
- After with you need to define CTE name and the field names. For instance in the below code snippet I have 3 fields Count,Column and Id. The name of CTE is “MyTemp”.
- Once you have defined CTE we need to specify the SQL which will give the result for the CTE.
- Finally you can use the CTE in your SQL query.

```
with – Step 1 :-  Start with a with
MyTemp(Count,column1,id) – Step 2:- column names with CTE name
as
(
Select count(*),Column1,Id  -- Step 3 :-define the SQL Query
from SomeTable
group by column1,id
)
Select * from MyTemp – Step 4 :-Use the CTE
```

Can we use CTE multiple times in a single execution ?

CTE can be used only once in the same execution. You can see the below code snippet where we have created a simple CTE called as “MyTemp”. In the same execution I have tried to use it 2 times and you can see how did not identify the “MyTemp” in the second select execution.



Can you give some real time examples where CTE is useful ?

CTE can be used in the following scenarios:-

- Complex SQL queries can be broken down using CTE which will make your code more readable. Note it does have side effect of performance.
- Recursive query.
- Replacement for views if you do not want to store the metadata.
- You want use aggregate functions in WHERE clause.
- You can group by scalar values which are derived from a result set.

How to delete duplicate records which does not have primary key ?

Let me first explain what this question is all about. Let's say you have table which has names as shown in the below figure. Now in the names table you have duplicate records (ex. Shiv) and this table does not have a primary key.

So the question is how can we delete duplicate records and keep one of the records from the duplicates. For example from the below table how can we delete one "Shiv" and keep the other one.

Name's
Shiv
Shiv
Raju
XYZ

```
Note: - One more constraints many SQL Server interviewer puts here is
you cannot add an identity column to the table.
```

Now there are lots of ways of doing and the best I found personally was by using “Row_Number” and “CTE” (common table expression). In case you are not aware of CTE and row_number please refer previous questions.

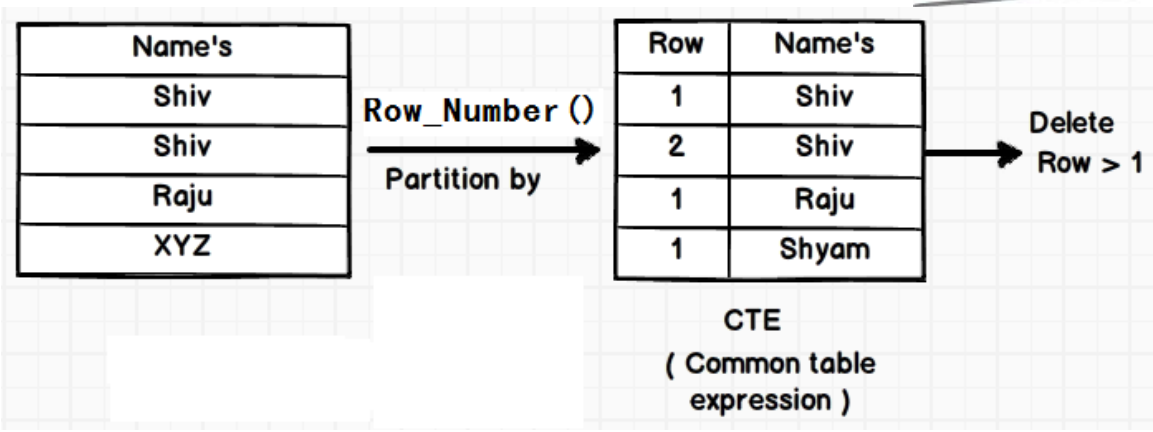
It’s a 2 step process:-

- Create a temp result set using CTE which has a new column which uses “row_number” using partition.
  Now the “row_number” with partition will create unique numbers for unique records with same. In case the records are same it will increment the number. For instance you can see for duplicate Shiv record, he has numbered them 1 and 2.But for “Raju” and “Shyam” he has created fresh number sequence.

```
with TempNames as -- Step 1 create the CTE
(
select row_number() over (partition by Name order by name) as
RowNo,Name from Names
)
```

- Once the CTE is created delete the records whose row sequence number is greater than one.

```
Delete from Tempnames where rowno>1 -- Delete  from CTE
```



Below is the complete code snippet for the same.

```
with TempNames as
(
select row_number() over (partition by Name order by name) as
RowNo,Name from Names
)
Delete from Tempnames where rowno>1
```

Temp variables VS Temp tables

	Temp tables	**Temp variables**
Big difference	Temp tables are real temporary SQL Server tables , you can create indexes , they can participate in transactions , it will use SQL Server optimization techniques etc. So if you are operating on large number of records use Temp tables.	As the name says these are variables. So they do not participate in transactions, you can not create indexes directly, they do not use SQL server optimization techniques etc. Good for small number of records.
Should be used when?	Large number of records.	Less than 100 records.
Scope	Outside procedure	Only Inside the procedure.
Transaction	Yes	No
Indexes	Yes	No (Note: - Indexes get indirectly created if you great a unique primary

		key.)
Truncate	Yes	No
Alter Table	Yes	No it's just variable.
Affected by SQL Server optimization	Yes	No
Parallelism	Yes	No.

(Q) What is "ALL" and "ANY" operator?

(Q) What is a "CASE" statement in SQL?

(Q) What does COLLATE Keyword in SQL signify?

(Q) What is TRY/CATCH block in T-SQL?

No I am not referring to .NET TRY/CATCH block this is the new way of handling error in SQL Server. For instance in the below T-SQL code any error during delete statement is caught and the necessary error information is displayed.

```
BEGIN TRY
DELETE table1 WHERE id=122
END TRY
BEGIN CATCH
SELECT
ERROR_NUMBER() AS ErrNum,
ERROR_SEVERITY() AS ErrSev,
ERROR_STATE() as ErrSt,
ERROR_MESSAGE() as ErrMsg;
END CATCH
```

(Q) What is PIVOT feature in SQL Server?

PIVOT feature converts row data to column for better analytical view. Below is a simple PIVOT fired using CTE. Ok the first section is the CTE which is the input and later PIVOT is applied over it.

```
WITH PURCHASEORDERHEADERCTE(Orderdate,Status,Subtotal) as
(
Select year(orderdate),Status,isnull(Subtotal,0) from
purchasing.PURCHASEORDERHEADER
)
Select Status as OrderStatus,isnull([2001],0) as 'Yr 2001'
,isnull([2002],0) as 'Yr 2002' from PURCHASEORDERHEADERCTE
pivot (sum(Subtotal) for  Orderdate in ([2001],[2002])) as pivoted
```

You can see from the above SQL the top WITH statement is the CTE supplied to the PIVOT. After that PIVOT is applied on subtotal and orderdate. You have to specify in what you want the pivot (here it is 2001 and 2002). So below is the output of CTE table.

```
select  top 3 with ties *  from productcost
order by unitcost
```

Product	UnitCost
p1	200.23
p2	201.23
p3	250.23
p4	250.23

Figure 2.17: - CTE output

After the PIVOT is applied, you can see the rows are now grouped column wise with the subtotal assigned to each. You can summarize that PIVOT summarizes your data in cross tab format.

	OrderStatus	Yr 2001	Yr 2002
1	3	171.0765	383552.904
2	1	272.1015	0.00
3	4	103452.643	3842580.126

Figure 2.18: - Pivoted table

(Q) What is UNPIVOT?

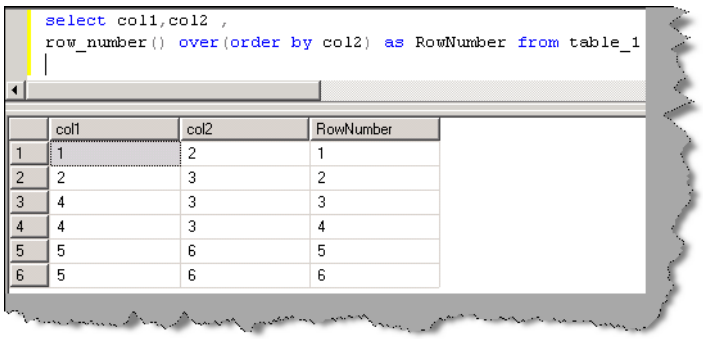
It's exactly the vice versa of PIVOT. That means you have a PIVOTED data and you want to UNPIVOT it.

(Q) What are RANKING functions?

They add columns that are calculated based on a ranking algorithm. These functions include ROW_NUMBER(), RANK(), DENSE_RANK(), and NTILE().

(Q) What is ROW_NUMBER()?

The ROW_NUMBER() function adds a column that displays a number corresponding the row's position in the query result . If the column that you specify in the OVER clause is not unique, it still produces an incrementing column based on the column specified in the OVER clause. You can see in the figure below I have applied ROW_NUMBER function over column col2 and you can notice the incrementing numbers generated.



```
select col1,col2 ,
row_number() over(order by col2) as RowNumber from table_1
```

	col1	col2	RowNumber
1	1	2	1
2	2	3	2
3	4	3	3
4	4	3	4
5	5	6	5
6	5	6	6

Figure 2.19:- ROW_NUMBER in action

(Q) What is RANK()?

The RANK() function works much like the ROW_NUMBER() function in that it numbers records in order. When the column specified by the ORDER BY clause contains unique values, then ROW_NUMBER() and RANK() produce identical results. They differ in the way they work when duplicate values are contained in the ORDER BY expression. ROW_NUMBER will increment the numbers by one on every record, regardless of duplicates. RANK() produces a single number for each value in the result set. You can see for duplicate value it does not increment the row number.

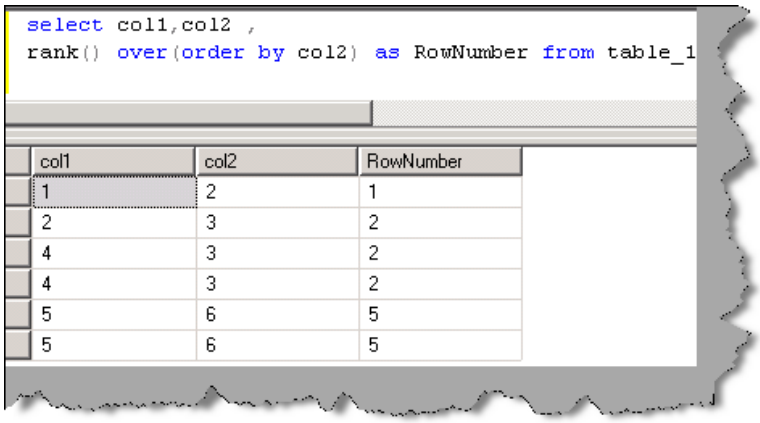


Figure 2.20: - RANK

(Q) What is DENSE_RANK()?

DENSE_RANK() works the same way as RANK() does but eliminates the gaps in the numbering. When I say GAPS, you can see in previous results it has eliminated 4 and 5 from the count because of the gap in between COL2. But for dense_rank it overlooks the gap.

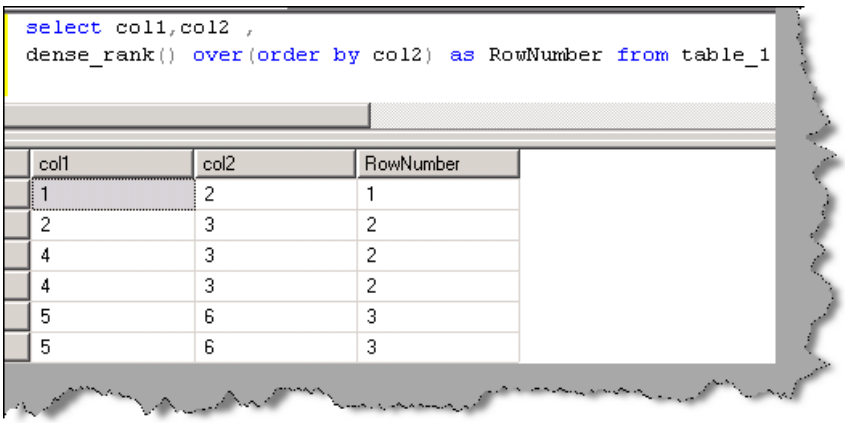


Figure 2.21 :- DENSE_RANK() in action

(Q) What is NTILE()?

NTILE() breaks the result set into a specified number of groups and assigns the same number to each record in a group. Ok NTILE just groups depending on the number given or you can say divides the data. For instance, I have said to NTILE it to 3. It has 6 total rows so it grouped in number of 2.

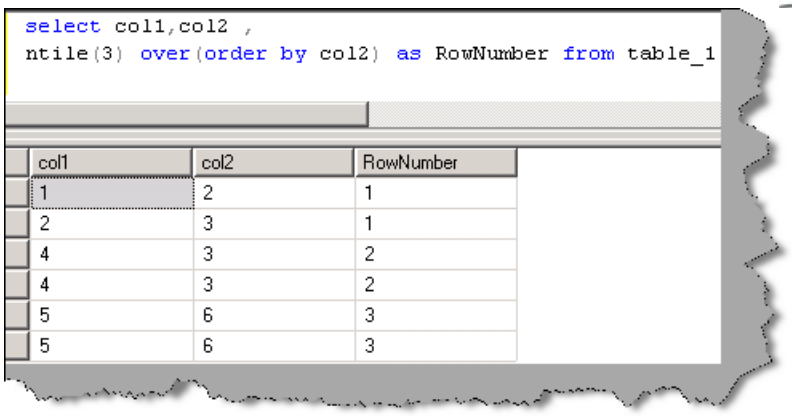


Figure 2.22: - NTILE in Action

(DB) What is SQl injection?

It is a Form of attack on a database-driven Web site in which the attacker executes unauthorized SQL commands by taking advantage of insecure code on a system connected to the Internet, bypassing the firewall. SQL injection attacks are used to steal information from a database from which the data would normally not be available and/or to gain access to an organization's host computers through the computer that is hosting the database.

SQL injection attacks typically are easy to avoid by ensuring that a system has strong input validation.

As name suggest we inject SQL which can be relatively dangerous for the database. Example this is a simple SQL

```
SELECT email, passwd, login_id, full_name
FROM members
WHERE email = 'x'
```

Now somebody does not put "x" as the input but puts "x ; DROP TABLE members;". So the actual SQL which will execute is :-

```
SELECT email, passwd, login_id, full_name
FROM members
WHERE email = 'x'  ; DROP TABLE members;
```

Think what will happen to your database.

Chapter 3: .NET Integration

(Q) What are steps to load a .NET code in SQL SERVER 2005?

Following are the steps to load a managed code in SQL SERVER 2005:-

- Write the managed code and compile it to a DLL / Assembly.
- After the DLL is compiled using the "CREATE ASSEMBLY" command you can load assembly in to SQL SERVER. Below is the create command which is loading "mycode.dll" in to SQL SERVER using the "CREATE ASSEMBLY" command.

```
CREATE ASSEMBLY mycode FROM 'c:/mycode.dll'
```

(Q) How can we drop an assembly from SQL SERVER?

```
DROP ASSEMBLY mycode
```

(Q) Are changes made to assembly updated automatically in database?

No, it will not synchronize the code automatically. For that you have to drop the assembly (using the DROP ASSEMBLY) and create (using the CREATE ASSEMBLY) it again.

(Q) Why do we need to drop assembly for updating changes?

When we load the assembly in to SQL SERVER, it persist it in sys.assemblies. So any changes after that to the external DLL / ASSEMBLY will not reflect in SQL SERVER. So you have to DROP and then CREATE assembly again in SQL SERVER.

(Q) How to see assemblies loaded in SQL Server?

```
Select * from sys.assemblies.
```

(Q) I want to see which files are linked with which assemblies?

Assembly files system tables have the track about which files are associated with what assemblies.

```
SELECT * FROM sys.assembly_files
```

Note: - You can create SQL SERVER projects using VS 2005 which provides ready made templates to make development life easy.

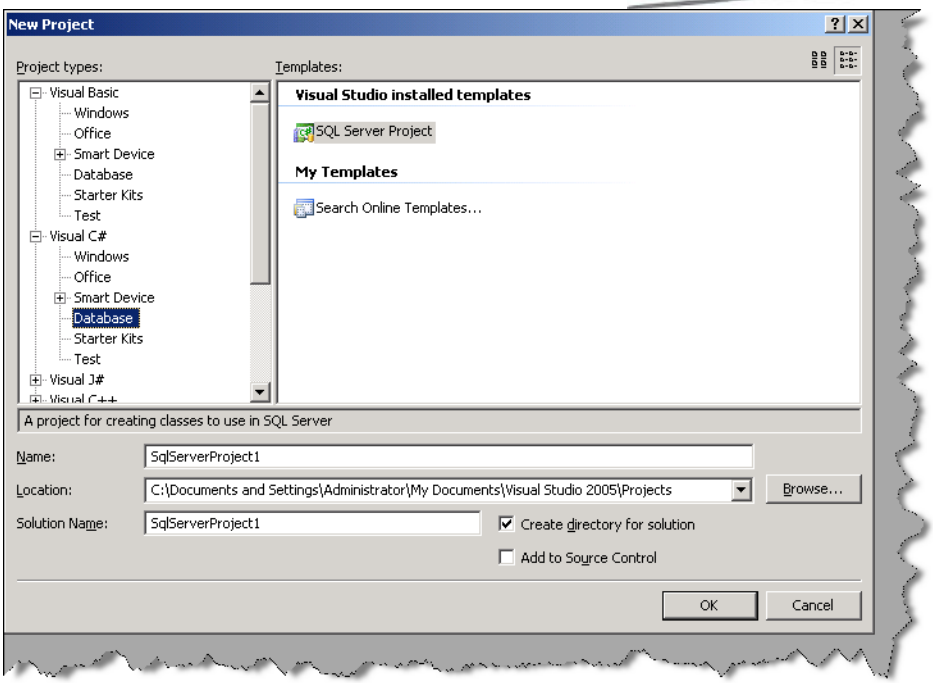


Figure 3.1 : - Creating SQL SERVER Project using VS2005

(Q) Does .NET CLR and SQL SERVER run in different process?

.NET CLR engine (hence all the .NET applications) and SQL SERVER run in the same process or address space. This "Same address space architecture" is implemented so that there no speed issues. If the architecture were implemented the other way (i.e. SQL SERVER and .NET CLR engine running in different memory process areas), there would have been reasonable speed issue.

(Q) Does .NET controls SQL SERVER or is it vice-versa?

SQL Server controls the way .NET application will run. Normally .NET framework controls the way application should run. But in order that we have high stability and good security SQL Server will control the way .NET framework works in SQL Server environment. So lot of things will be controlled through SQL Server example threads, memory allocations, security etc.
SQL Server can control .NET framework by "Host Control" mechanism provided by .NET Framework 2.0. Using the "Host Control" framework external application's can control the way memory management is done, thread allocation's are done and lot more. SQL Server uses this "Host Control" mechanism exposed by .NET 2.0 and controls the framework.

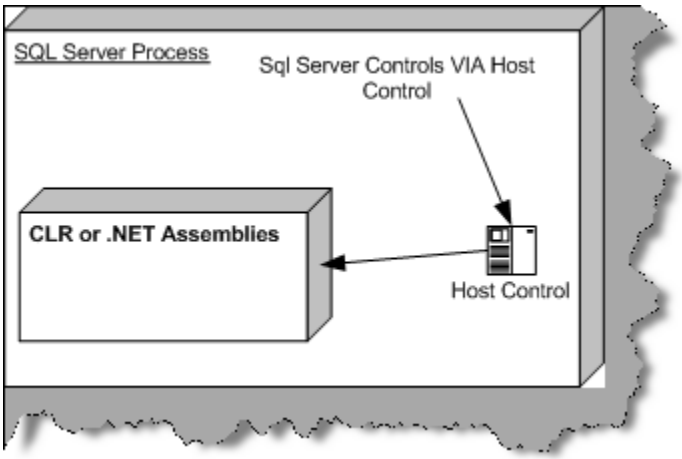


Figure 3.2:- CLR Controlled by Host Control

(Q) Is SQLCLR configured by default?

SQLCLR is not configured by default. If developers want to use the CLR integration feature of SQL SERVER, it has to be enabled by DBA.

(Q) How to configure CLR for SQL SERVER?

It is an advanced option you will need to run the following code through query analyzer.

```
EXEC sp_configure 'show advanced options', '1';
go
reconfigure
go
EXEC sp_configure 'clr enabled' , '1'
go
reconfigure;
go
```

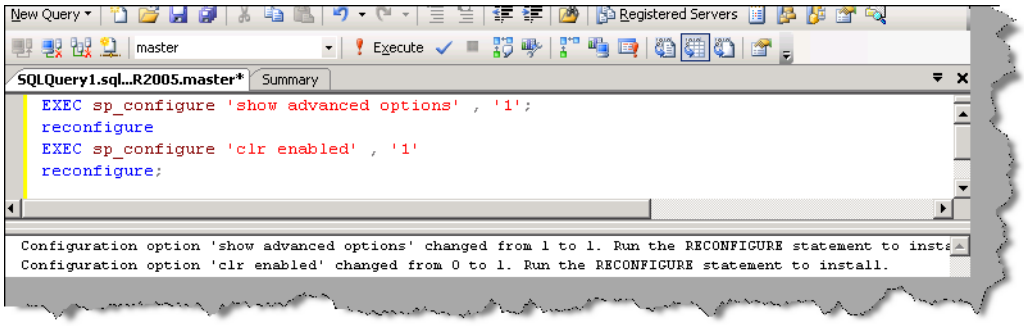


Figure 3.3:- sp_configure in action

Note: - You can see after running the SQL "clr enabled" property is changed from 0 to 1 , which indicates that the CLR was successfully configured for SQL SERVER.

(Q) Is .NET feature loaded by default in SQL Server?

No it will not be loaded, CLR is lazy loaded that means it's only loaded when needed. It goes one-step ahead where the database administrator has to turn the feature on using "sp_configure".

Note: - Loading .NET runtime consumes some memory resources around 20 to 30 MB (it may vary depending on lot of situations). So if you really need .NET Integration then only go for this option.

(Q) How does SQL Server control .NET run-time?

.NET CLR exposes interfaces by which an external host can control the way .NET run time runs.

(Q) In previous versions of .NET it was done via COM interface "ICorRuntimeHost".

In previous version, you can only do the following with the COM interface.

- Specify that whether its server or workstation DLL.
- Specify version of the CLR (e.g. version 1.1 or 2.0)
- Specify garbage collection behavior.
- Specify whether jitted code may be shared across AppDomains.

In .NET 2.0 it is done by "ICLRRuntimeHost".

However, in .NET 2.0, you can do much above what was provided by the previous COM interface.

- Exceptional conditions
- Code loading
- Class loading
- Security particulars
- Resource allocation

SQL Server uses the "ICLRRuntimeHost" to control .NET run-time as the flexibility provided by this interface is far beyond what is given by the previous .NET version, and that's what exactly SQL Server needs, a full control of the .NET run time.

(Q) What is a "SAND BOX" in SQL Server 2005?

Twist: - How many types of permission levels are there and explain in short there characteristics?

Ok, here is a general definition of sand box:-

"Sandbox is a safe place for running semi-trusted programs or scripts, often originating from a third party."

Now for SQL Server it is .NET the external third party that is running and SQL Server has to ensure that .NET runtime crashes does not affect his working. So in order that SQL Server runs properly there are three sandboxes that user code can run:-

Safe Access sandbox
This will be the favorite setting of DBA's if they are every compelled to run CLR - Safe access. Safe means you have only access to in-proc data access functionalities. So you can create stored procedures, triggers, functions, data types, triggers etc. But you can not access memory, disk, create files etc. In short, you cannot hang the SQL Server.

External access sandbox
In external access you can use some real cool features of .NET like accessing file systems outside the box, you can leverage you classes etc. But here you are not allowed to play around with threading, memory allocation etc.

Unsafe access sand box
In Unsafe access, you have access to memory management, threading etc. So here developers can write unreliable and unsafe code which destabilizes SQL Server. In the first two access levels of sand box its difficult to write unreliable and unsafe code.

(Q) What is an application domain?

Previously "PROCESS" where used as security boundaries. One process has its own virtual memory and does not over lap the other process virtual memory; due to this, one process cannot crash the other process. Therefore, any problem or error in one process does not affect the other process. In .NET, they went one-step ahead introducing application domains. In application domains, multiple applications can run in same process with out influencing each other. If one of the application domains throws error it does not affect the other application domains. To invoke method in a object running in different application domain .NET remoting is used.

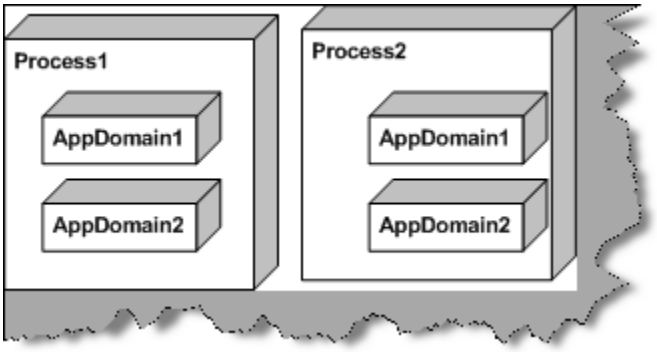


Figure 3.4:- Application Domain architecture

(Q) How is .NET Appdomain allocated in SQL SERVER 2005?

In one-line, it's "One Appdomain per Owner Identity per Database".

That means if owner "A" owns "Assembly1" and "Assembly2" which belong to one database. They will be created in one Appdomain. But if they belong to different database, two Appdomains will be created.
Again if there are different owners for every the same assembly then every owner will have its own Appdomain.

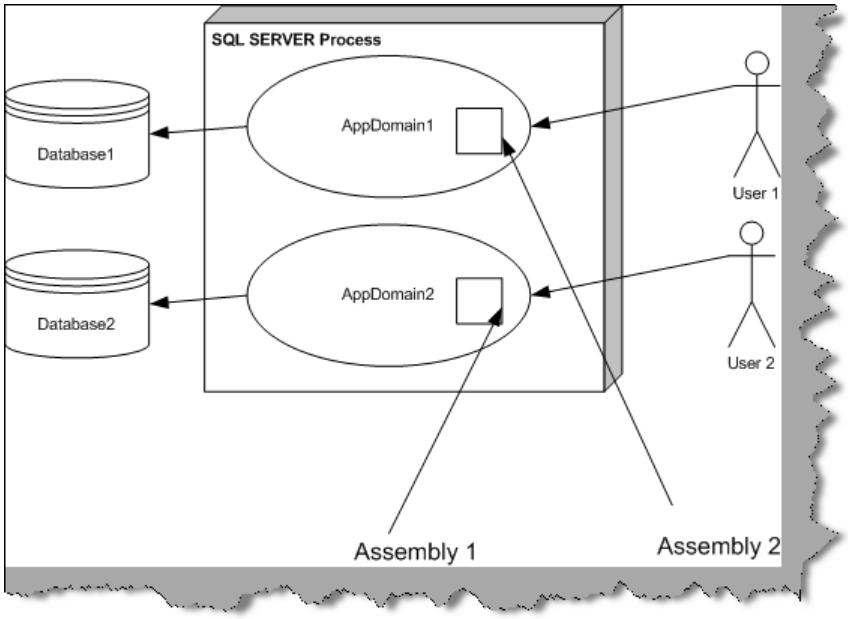


Figure 3 .5: - One Appdomain / Owner / Database

Note: - This can be pretty confusing during interviews so just make one note "One Appdomain per Owner Identity per Database".

(Q) What is Syntax for creating a new assembly in SQL Server 2005?

```
CREATE ASSEMBLY customer FROM 'c:\customers\customer.dll'
```

(Q) Do Assemblies loaded in database need actual .NET DLL?

No, once the assembly is loaded you do not need the source. SQL Server will load the DLL from the catalog.

(Q) You have an assembly, which is dependent on other assemblies; will SQL Server load the dependent assemblies?

Ok. Let me make the question clearer. If you gave "Assembly1.dll" who is using "Assembly2.dll" and you try cataloging "Assembly1.dll" in SQL Server will it catalog "Assembly2.dll" also? Yes it will catalog it. SQL Server will look in to the manifest for the dependencies associated with the DLL and load them accordingly.

Note: - All Dependent assemblies have to be in the same directory, do not expect SQL Server to go to some other directory or GAC to see the dependencies.

(Q) Does SQL Server handle unmanaged resources?

SQL Server does not handle the unmanaged resource of a framework. It has to be guaranteed by the framework DLL that it will clean up the unmanaged resource. SQL Server will not allow you to load .NET framework DLL that do not have clean up code for unmanaged resources.

(Q) What is Multi-tasking?

It is a feature of modern operating systems with which we can run multiple programs at same time example Word, Excel etc.

(Q) What is Multi-threading?

Multi-threading forms subset of Multi-tasking instead of having to switch between programs this feature switches between different parts of the same program. Example you are writing in word and at the same time word is doing a spell check in background.

(Q) What is a Thread?

A thread is the basic unit to which the operating system allocates processor time.

(Q) Can we have multiple threads in one App domain?

One or more threads run in an AppDomain. An AppDomain is a runtime representation of a logical process within a physical process. Each AppDomain is started with a single thread, but can create additional threads from any of its threads.

Note: - All threading classes are defined in System.Threading namespace.

(Q) What is Non-preemptive threading?

In Non-preemptive threading every thread gives control to other threads to execute. So for example we have "Thread1" and "Thread2", let us say for that instance "Thread1" is running. After some time it will give the control to "Thread2" for execution.

(Q) What is pre-emptive threading?

In pre-emptive threading operating system schedules, which thread should run, rather threads making there own decisions.

(Q) Can you explain threading model in SQL Server?

SQL Server uses the "Non-preemptive" threading model while .NET uses "pre-emptive" threading model.

(Q) How does .NET and SQL Server thread work?

Note: - From hence onwards I will refer .NET assemblies running on SQL SERVER as SQLCLR that's more of industry acronym.

As said in the previous section-threading model of .NET and SQL Server is completely different. Therefore, SQL Server has to handle threads in a different way for SQLCLR. So a little different threading architecture is implemented termed as "Tasking of Threads". In tasking thread architecture there is switching between SQLCLR threads and SQL Server threads, that .NET threads do not consume full resource and go out of control. SQL Server introduced blocking points, which allows this transition to happen between SQLCLR and SQL Server threads.

(Q) How is exception in SQLCLR code handled?

If you remember in the previous section's we had mentioned that there is One Appdomain / User / Database. So if there is an error in any of the Appdomain SQL Server will shut down the Appdomain, release all locks if the SQLCLR was holding and rollback the transaction in case if there are any. So the Appdomain shut down policy ensures that all other Appdomain including SQL Server process is not affected.

(Q) Are all .NET libraries allowed in SQL Server?

No, it does not allow all .NET assemblies to execute example: - System.Windows.Forms, System. Drawing, System. Web etc are not allowed to run's Server maintains a list of .NET namespaces which can be executed, any namespaces other than that will be restricted by SQL Server policy. This policy checks are made on two instances:-

- When you are cataloging the assembly.
- When you are executing the assembly.

Readers must be wondering why a two time check. There are many things in .NET which are building on runtime and can not be really made out from IL code. So SQL Server makes check at two point while the assembly is cataloged and while it's running which ensures 100 % that no runaway code is going to execute.

Note : - Read Hostprotectionattribute in next questions.

(Q) What is "Hostprotectionattribute" in SQL Server 2005?

As said previously .NET 2.0 provides capability to host itself and that is how SQL Server interacts with the framework. But there should be some mechanism by which the host who is hosting .NET assemblies should be alerted if there is any out of serious code running like threading, synchronization etc. This is what exactly the use of "Hostprotectionattribute". It acts like a signal to the outer host saying what type of code it has. When .NET Framework 2.0 was in development Microsoft tagged this attribute on many assemblies , so that SQL Server can be alerted to load those namespaces or not. Example if you look at System. Windows you will see this attribute.

So during runtime SQL Server uses reflection mechanism to check if the assembly has valid protection or not.

Note :- HostProtection is checked only when you are executing the assembly in SQL Server 2005.

(Q) How many types of permission level are there for an assembly?

There are three types of permission levels for an assembly:-

Safe permission level

Safe assemblies can only use pre-defined framework classes, can call any COM based components or COM wrapper components, can not access network resources and, can not use PInvoke or platform invoke

External Access

It is like safe but you can access network resources like files from network, file system, DNS system, event viewer's etc.

Unsafe

In unsafe code you can run anything, you want. You can use PInvoke; call some external resources like COM etc. Every DBA will like to avoid this and every developer should avoid writing unsafe code unless very much essential. When we create an assembly we can give the permission set at that time.

Note: - We had talked about sand boxes in the previous question. Just small note sandboxes are expressed by using the permission level concepts.

(Q) In order that an assembly gets loaded in SQL Server what type of checks are done?

SQL Server uses the reflection API to determine if the assembly is safe to load in SQL Server. Following are the checks done while the assembly is loaded in SQL Server:-

- It does the META data and IL verification, to see that syntaxes are appropriate of the IL.
- If the assembly is marked as safe and external then following checks are
- Check for static variables, it will only allow read-only static variables

- Some attributes are not allowed for SQL Server and those attributes are checked.
- Assembly has to be type safe that means no unmanaged code or pointers are allowed.
- No finalizer's are allowed.

```
Note: - SQL Server checks the assembly using the reflection API, so the code should be IL compliant.
```

You can do this small exercise to check if SQL Server validates your code or not. Compile the simple below code, which has static variable defined in it. Now because the static variable is not read-only it should throw an error.

```
using System; namespace StaticDll { public class Class1 { static int i; } }
```

After you have compiled the DLL, use the Create Assembly syntax to load the DLL in SQL Server. While cataloging the DLL you will get the following error:-

```
Msg 6211, Level 16, State 1, Line 1 CREATE ASSEMBLY failed because type 'StaticDll.Class1' in safe assembly 'StaticDll' has a static field 'i'. Attributes of static fields in safe assemblies must be marked readonly in Visual C#, ReadOnly in Visual Basic, or initonly in Visual C++ and intermediate language.
```

(Q) Can you name system tables for .NET assemblies?

There are three mail files, which are important:-

- **sys.assemblies:-** They store information about the assembly.
- **sys.assembly_files:-** For every assembly you will can one or more files and his is where actual file raw data, file name and path is stored.
- **sys.assembly_references:–** All references to assemblies are stored in this table.

We can do a small practical hand on to see how the assembly table looks like. Let us try to create a simple class class1. Code is as shown below.

```
using System;
using System.Collections.Generic;
using System.Text;
namespace Class1
{
      public class Class1
      {
      }
}
```

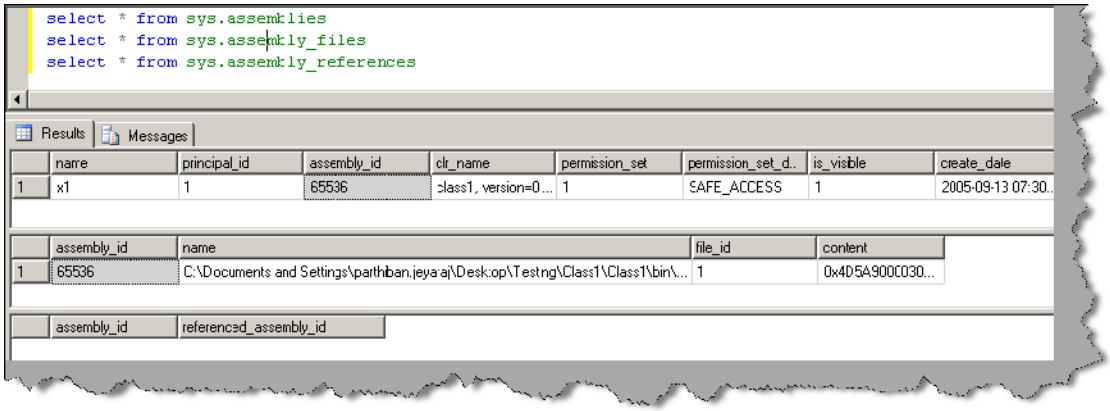


Figure 3.6: - Assembly related system files.

Then we create the assembly by name "X1" using the create assembly syntax. In above image is the query output of all three main tables in this sequence sys.assemblies, sys.assembly_files, and sys.assembly_references.

Note :- In the second select statement we have a content field in which the actual binary data stored. So even if we do not have the actual assembly it will load from this content field.

(Q) Are two version of same assembly allowed in SQL Server?

You can give different assembly name in the create statement pointing to the different file name version.

(Q) How are changes made in assembly replicated?

ALTER ASSEMBLY Customers ADD FILE FROM 'c:\mydir\CustomerAsm.pdb'

Note:- You can drop and recreate it but will not be a good practice to do that way. Also note I have set the file reference to a "PDB" file which will enable my debugging just in case if I want it.

(Q) Is it a good practice to drop a assembly for changes?

- Dropping an assembly will lead to loose following information:-
- You will loose all permissions defined to the assembly.
- All stored procedure, triggers and UDF (User defined functions) or any SQL Server object defined from them.

Note: - If you are doing bug fixes and modifications its good practice to Alter rather than Drop? Create assembly.

(Q) In one of the projects following steps where done, will it work?

Twist: - Are public signature changes allowed in "Alter assembly" syntax?

Following are the steps:-

- Created the following class and method inside it.

```
public class clscustomer
{
Public void add()
{
```

Compiled the project success fully.

- Using the create assembly cataloged it in SQL Server.
- Later made the following changes to the class

```
} public class clscustomer
{
Public void add(string code)
{
}
```

```
Note: - The add method signature is now changed.
```

- After that using "Alter", we tried to implement the change.

Using alter syntax you cannot change public method signatures, in that case you will have to drop the assembly and re-create it again.

(Q) What does Alter assembly with unchecked data signify?

"ALTER ASSEMBLY Customer Assembly FROM 'c:\cust.dll' WITH UNCHECKED DATA"

This communicates with SQL Server saying that you have not made any changes to serialization stuff, no data types are changed etc. Only you have changed is some piece of code for fixing bugs etc.

(Q) How do I drop an assembly?

DROP ASSEMBLY cuts

```
Note: - If the assembly is referencing any other objects like triggers,
stored procedures, UDT, other assemblies then the dependents should be
dropped first, or else the drop assembly will fail.
```

(Q) Can we create SQLCLR using .NET framework 1.0?

No at this moment only .NET 2.0 versions and above is supported with SQL Server.

(Q) While creating .NET UDF what checks should be done

Following are some checks are essential for UDF (User Defined Function):-

- The class in which the function is enclosed should be public.
- .NET Function must be static.

```
Note: - When we want to catalog the assembly and function. Then first
we catalog the class and the function. In short we use "Create
Assembly" and then "Create Function".
```

(Q) How do you define a function from the .NET assembly?

Below is a sample "create function "statement and following are the legends defined in it:-

Sort Customer: - Name of the stored procedure function and can be different from the .NET function.In this case the .NET function is by name sort which is defined in the external name.

Customer Assembly: - The name of the assembly.

Customer Namespace: - The Namespace in which the class is lying.

Sort: - The .NET function name.

Other things are self-explanatory.

```
Create Function Sort Customer (@Strcustcode int) returns int As
EXTERNAL NAME Customer Assembly.[CustomerNameSpace.CustomerClass].Sort
```

```
Note: - One important thing about the function parametersis allinput
parameter will go by the order mapping. So what does that mean if my
.NET function name func1 has the following definitions:-
```

- `func1 (int i, double x,string x)`
- Then my stored procedure should be defined accordingly and in the same order. That means in the stored procedure you should define it in the same order.

(Q) Can you compare between T-SQL and SQLCLR?

```
Note: - This will be one the favorite questions during interview.
Interviewer will want to know if you know when is the right decision to
take T-SQL or .NET for writing SQL Server objects. When I say SQL
Server objects I am referring to stored procedure, functions or
triggers.
```

- Pure Data access code should always be written using T-SQL as that is what they where meant for. T-SQL does not have to load any runtime, which make the access much faster. Also to note T-SQL will have access directly to internal buffers for SQL Server and they

where written in probably assembly and “C” which makes them much faster for data access.

- Pure Non-data access code like computation, string-parsing logic etc should be written in .NET. If you want to access web services or want to exploit OOP is programming for better reusability and read external files its good to go for .NET.

We can categorize our architect decision on there types of logic:-

- Pure Data access functionality – Go for T-SQL.
- Pure NON-Data access functionality – Go for .NET.
- Mixture of data access and NON-Data access – Needs architecture decision.

If you can see, the first two decisions are straightforward. But the third one is where you will have do a code review and see what will go the best. Probably also run it practically, benchmark and see what will be the best choice.

(Q) With respect to .NET is SQL SERVER case sensitive?

Following are some points to remember regarding case sensitiveness:-

- Assembly names are not case sensitive.
- Class and Function names are case sensitive.

So what does that mean? Well if you define a .NET DLL and catalog it in SQL Server. All the methods and class name are case sensitive and assembly is not case sensitive. For instance, I have cataloged the following DLL, which has the following details:-

- Assembly Name is “CustomerAssembly”.
- Class Name in the “CustomerAssembly” is “ClsCustomer”.
- Function “GetCustomerCount ()” in class “ClsCustomer”.

When we catalog the above assembly in SQL Server. We cannot address the “ClsCustomer” with “CLSCUSTOMER” or function “GetCustomerCount ()” with “getcustomercount ()” in SQL Server T-SQL language. But assembly “CustomerAssembly” can be addressed by “customer assembly” or “CUSTOMERASSEMBLY”, in short the assemblies are not case sensitive.

(Q) Does case sensitive rule apply for VB.NET?

The above case sensitive rules apply irrespective of whether the .NET language is case sensitive or not. So even if VB.NET is not case sensitive the rule will apply.

(Q) Can nested classes be accessed in T-SQL?

No, you cannot access nested class in T-SQL its one of the limitation of SQLCLR.

(Q) Can we have SQLCLR procedure input as array?

SQL Server has no data types like arrays so it will not be able to map array data type of .NET. This will throw an error.

(Q) Can object data type be used in SQLCLR?

You can pass object data type to SQLCLR but then that object should be defined as UDF in SQL Server.

(Q) How is precision handled for decimal data types in .NET?

```
Note: - Precision is actually the number of digits after point which
determines how accurate you want to see the result. Example in "9.29"
we have precision of two decimal places (.29).
```

In .NET, we declare decimal data types with out precision. However, in SQL Server you can define the precision part also.
decimal i; --> .NET Definition
decimal (9,2) --> SQL Server Definition
This creates a conflict when we want the .NET function to be used in T-SQL as SQLCLR and we want the precision facility.
Here is the answer you define the precision in SQL Server when you use Create syntax. So even if .NET does not support the precision facility we can define the precision in SQL Server.

```
.NET definition
func1(decimal x1)
{
}
SQL Server definition
create function func1(@x1 decimal(9,2))
returns decimal
as external name CustomerAssembly.[CustomerNameSpace.ClsCustomer].func1
```

If you see in the above code, sample func1 is defined as simple decimal but later when we are creating the function definition in SQL Server, we are defining the precision.

(Q) How do we define INPUT and OUTPUT parameters in SQLCLR?

.NET has following type of variable directions by value, byre and out (i.e. for C# only).Following is how the mapping goes:-

- Byval definition for function maps as input parameters for SQL Server.
- Byref definition maps to input and output parameters for SQL Server.

But for "out" types of parameters there are no mappings defined. Its logical "out" types of parameter types does not have any equivalents in SQL Server.

```
Note: - When we define byref in .NET that means if variable value is
changed it will be reflected outside the subroutine, so it maps to SQL
Server input/output (OUT) parameters.
```

(Q) Is it good to use .NET data types in SQLCLR?

No, it is always recommended to use SQL Server data types for SQLCLR code as it implements better integration. Example for "int" data type in .NET we cannot assign NULL value it will crash, but using SQL data type SqlInt32 NULLS will be handled. All SQL data types are available in "system.data.SQLtypes", so you have to refer this namespace in order to get advantage of SQL data types.

```
Note: - NULL is a valid data in SQL Server which represents no data,
but .NET datatype does not accept it.
```

(Q) How to move values from SQL to .NET data types?

You have to use the value property of SQL data types.

```
SqlInt32 x = 3;
int y = x.Value;
```

```
Note :- Direct assigning the values will crash your program.
```

(Q) What is System.Data.SqlServer?

When you have functions, stored procedures etc written in .NET you will use this provider rather than the traditional System.Data.SQLClient. If you are accessing objects created using T-SQL language then you will need a connection to connect them. Because you need to specify which server you will connect, what is the password and other credentials? But if you are accessing objects made using .NET itself you are already residing in SQL Server so you will not need a connection but rather a context.

(Q) What is SQLContext?

As said previously when use ADO.NET to execute a T-SQL created stored procedure we are out of the SQL Server boundary. So we need to provide SQLConnection object to connect to the SQLServer. But when we need to execute objects, which are created using .NET language, we only need the context in which the objects are running.

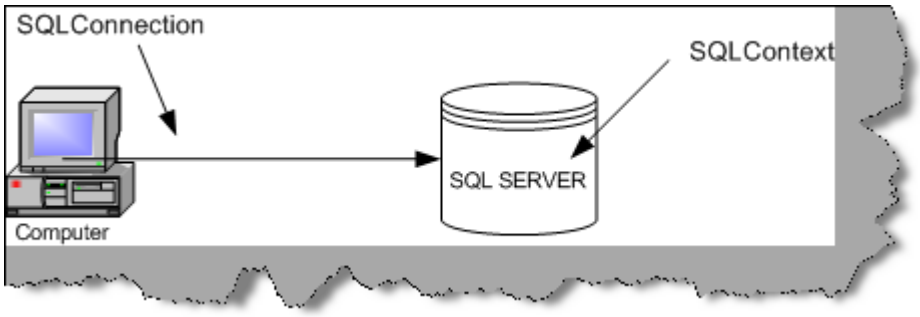


Figure 3.7 : - SQLConnection and SQLContext

So you can see in the above figure SQLConnection is used because you are completely outside SQL Server database. While SQLContext is used when you are inside SQL Server database. That

means that there is already a connection existing so that you can access the SQLContext. And any connections created to access SQLContext are a waste as there is already a connection opened to SQL Server.

These all things are handled by SQLContext.

Which are the four static methods of SQLContext?

Below are the four static methods in SQLContext:-

Get Connection ():- This will return the current connection

Get Command ():- Get reference to the current batch

Get Transaction ():- If you have used transactions, this will get the current transaction

Get Pipe ():- This helps us to send results to client. The output is in Tabular Data stream format. Using this method you can fill in data reader or data set, which can later be used by client to display data.

Note: - In top question I had shown how we can manually register the DLL's in SQL Server but in real projects no body would do that rather we will be using the VS.NET studio to accomplish the same. So we will run through a sample of how to deploy DLL's using VS.NET and paralelly we will also run through how to use SQLContext.

(Q) Can you explain essential steps to deploy SQLCLR?

This example will make a simple walk through of how to create a stored procedure using visual studio.net editor. During interview, you can make the steps short and explain to the interviewer. Therefore, we will create a stored procedure, which will retrieve all products from adventure works database. All products are stored in "Production. Product" table.

Let start step1 go to visual studio --> new project --> expand the Visual C# (+)--> select database, you will see SQL Server project. Select SQL Server project template and give a name to it, then click ok.

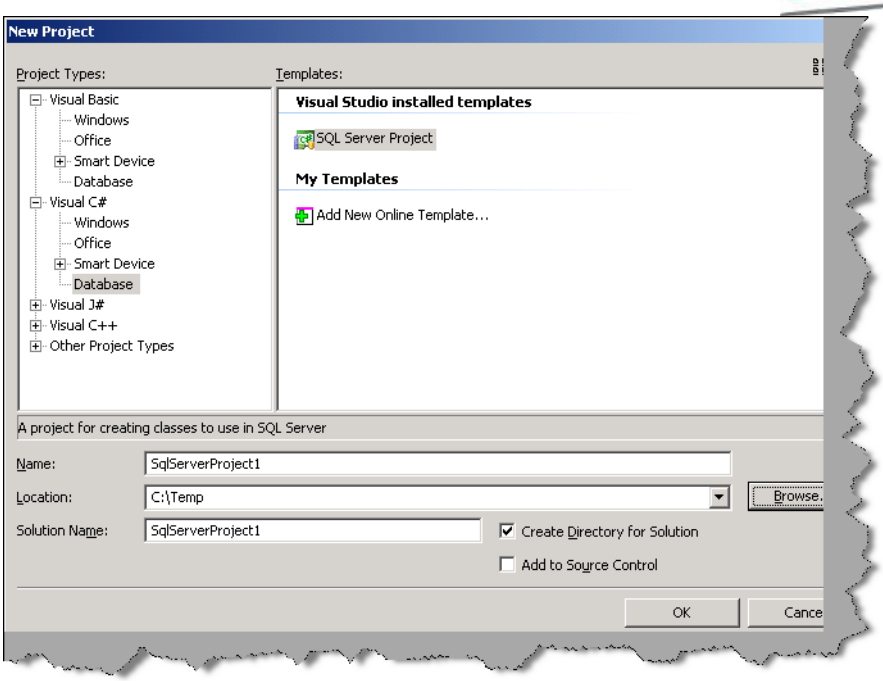


Figure 3.8: - Template dialog box

As these DLL's need to be deployed on the server, you will need to specify the server details also. So for the same you will be prompted to specify database on which you will deploy the .NET stored procedure. Select the database and click ok. In case you do not see the database, you can click on "Add reference" to add the database to the list.

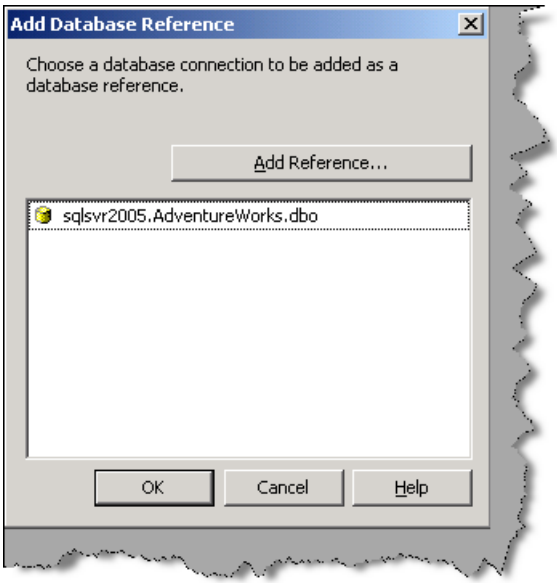


Figure 3.9: - Select database

Once you specify the database, you are inside the visual studio.net editor. At the right hand side, you can see the solution explorer with some basic files created by visual studio in order to deploy the DLL on the SQL Server. Right click on SQL Server project and click on ADD --> New items are displayed as shown in figure below.

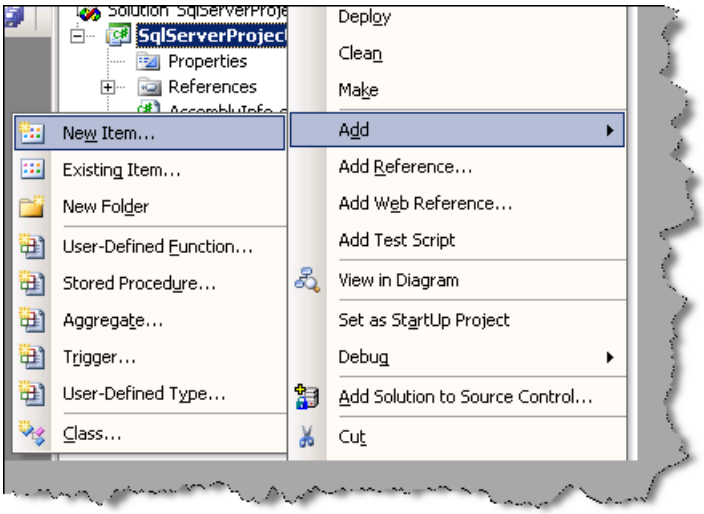


Figure 3.10: - Click add item

You can see in the below figure you can create different objects using VS.NET. For this point of time, we need to only create a stored procedure, which will fetch data from "Product. Product".

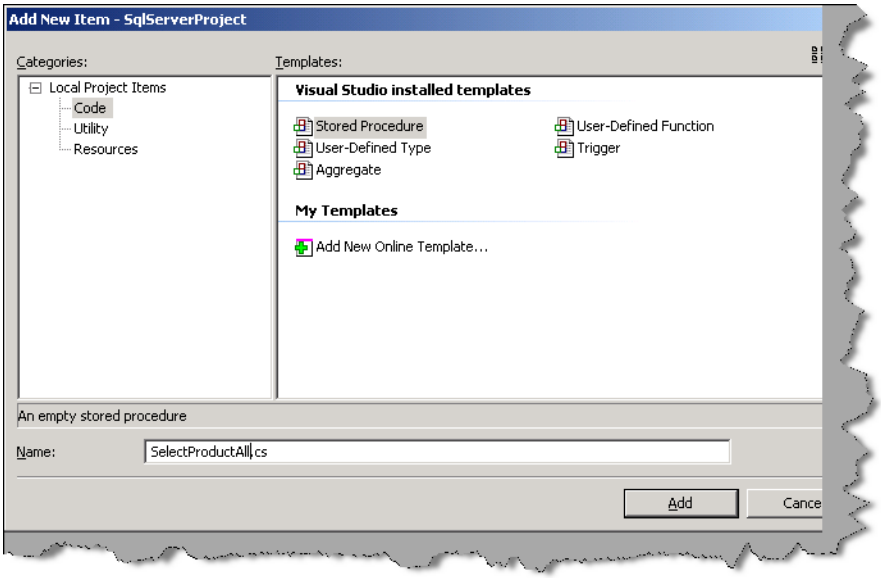


Figure 3.11: - Select stored procedure template

This section is where the real action will happen. As said previously you do not need to open a connection but use the context. So below are the three steps:-

- Get the reference of the context.

- Get the command from the context.
- Set the command text, at this moment we need to select everything from “Production. Product” table.
- Finally get the Pipe and execute the command.

```
using System;
using System.Data;
using System.Data.Sql;
using System.Data.SqlServer;
using System.Data.SqlTypes;
public partial class StoredProcedures
{
    [SqlProcedure]
    public static void SelectProductAll()
    {
        // Put your code here
        SqlCommand sqlCmd = SqlContext.GetCommand();
        sqlCmd.CommandText = "select * from production.product";
        SqlContext.GetPipe().Execute(sqlCmd);
    }
};
```

Figure 3.12: - Simple code to retrieve product table

After that, you need to compile it to a DLL form and then deploy the code in SQL Server. You can compile using “Build Solution” menu to compile and “Deploy Solution” to deploy it on SQL Server.

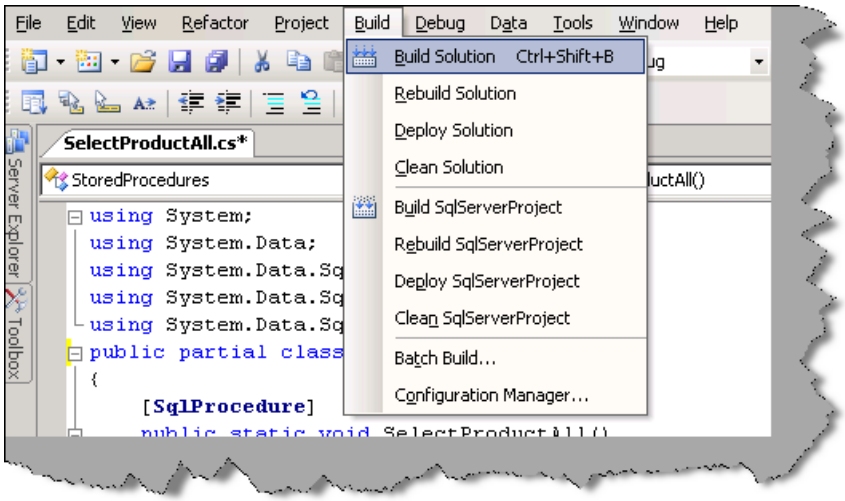


Figure 3.13: - Finally build and deploy solution

After deploying the solution, you can see the stored procedure “SelectProductAll” in the stored procedure section as shown below.

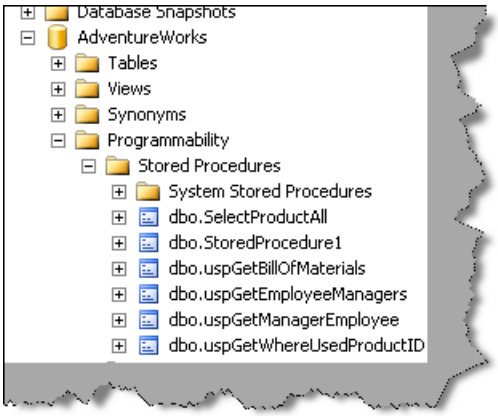
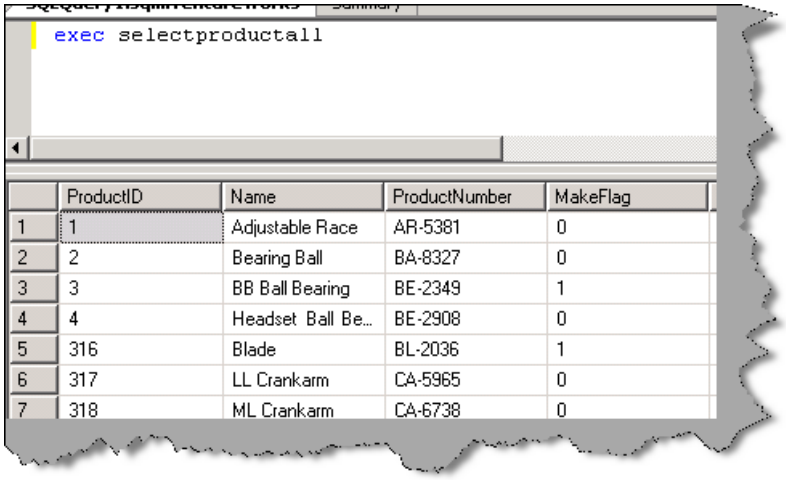


Figure 3.14: - SelectProductall listed in database

Just to test I have executed the stored procedure and everything working fine.



	ProductID	Name	ProductNumber	MakeFlag
1	1	Adjustable Race	AR-5381	0
2	2	Bearing Ball	BA-8327	0
3	3	BB Ball Bearing	BE-2349	1
4	4	Headset Ball Be...	BE-2908	0
5	316	Blade	BL-2036	1
6	317	LL Crankarm	CA-5965	0
7	318	ML Crankarm	CA-6738	0

Figure 3.15: - Execute stored procedure selectproductall

(Q) How do create function in SQL Server using .NET?

In order to create the function you have to select in Visual studio installed templates, User defined function template. Below is the sample code. Then follow again the same procedure of compiling and deploying the solution.

(Q) How do we create trigger using .NET?

For trigger, you have to select trigger template. But you can see some difference in code here. You have to specify on which object and which event will this method fire. The first attribute specifies the name, target (on which the trigger will fire) and event (insert , update or delete).

```
[SqlTrigger (Name="Trigger1", Target="Table1", Event="FOR INSERT")]
      public static void Trigger1()
      {
```

```
            // Put your code here
            SqlTriggerContext objtrigcontext =
SqlContext.GetTriggerContext();
            SqlPipe objsqlpipe = SqlContext.GetPipe();
            SqlCommand objcommand = SqlContext.GetCommand();
            if (objtrigcontext.TriggerAction == TriggerAction.Insert)
            {
                  objcommand.CommandText = "insert into table1
values('Inserted')";
                  objsqlpipe.Execute(objcommand);
            }
      }
```

(Q) How to create User Define Functions using .NET?

The below code is self-explanatory. Compiling and deploying remains, same for all object created using .NET.

```
using System;
using System.Data.Sql;
using System.Data.SqlTypes;

public partial class UserDefinedFunctions
{
    [SqlFunction]
    public static SqlInt16 Function1(SqlInt16 num1 , SqlInt16 num2)
    {
        // Put your code here
        return num1+num2;
    }
};
```

Figure 3.16 :- Function source code

Note: - Some home work for readers down.

(Q) How to create aggregates using .NET?

(Q) What is Asynchronous support in ADO.NET?

One of features which was missing in ADO.NET was asynchronous processing. That means once your SQL command is executed your UI has to wait until it is finished. ADO.NET provides support where you do not have to wait until the SQL is executed in database. You can see from the code below you have to issue "BeginExecuteReader" and then proceed ahead with some other

process. After you finish the process, you can come back and see your results. Long running queries can really be benefited from Asynchronous support.

```
conn.Open();
    IAsyncResult myResult = mycommand.BeginExecuteReader();
    while (!myResult.IsCompleted)
    {
        // execute some other process
    }
    // Finally process the data reader output
```

```
SqlDataReader rdr = mycommand.EndExecuteReader(myResult);
```

Note: - Here's a small project which you can do with Asynchronous processing. Fire a heavy duty SQL and in UI show how much time the SQL Server took to execute that query.

(Q) What is MARS support in ADO.NET?

In previous versions of ADO.NET, you should one connection on every result set. But this new feature allows you to execute multiple commands on the same connection. You can also switch back and forth between the command objects in a connection. There is nothing special to do for MARS (Multiple Active Result Sets) just you can allocate multiple command objects on a single connection.

(Q) What is SQLbulkcopy object in ADO.NET?

With SQlbulkcopy, you can insert bulk records in to table. The command is pretty simple you can see we have provided the data reader object to the SQlbulkcopy object and he will take care of the rest.

```
SqlBulkCopy objbulkData = new SqlBulkCopy(conn);
objbulkData.DestinationTableName = "table1";
objbulkData.WriteToServer(datareader1);
```

(Q) How to select range of rows using ADO.NET?

Twist: - What is paging in ADO.NET?

By paging, you can select a range of rows from a result set. You have to specify starting row in the result set and then how many rows you want after that.

```
command.ExecutePageReader(CommandBehavior.Default, 1, 10);
```

You can see in the above example I have selected 10 rows and starts from one. This functionality will be used mainly when you want to do paging on UI side. For instance, you want to show 10 records at a time to the user this can really ease of lot of pain.

(Q) If we have multiple AFTER Triggers on table how can we define the sequence of the triggers.

If a table has multiple AFTER triggers, then you can specify which trigger should be executed first and which trigger should be executed last using the stored procedure sp_settriggerorder. All the other triggers are in an undefined order, which you cannot control.

(Q) How can you raise custom errors from stored procedure?

The RAISERROR statement is used to produce an ad hoc error message or to retrieve a custom message that is stored in the sysmessages table. You can use this statement with the error handling code presented in the previous section to implement custom error messages in your applications. The syntax of the statement is shown here.

```
RAISERROR ({msg_id |msg_str }{,severity ,state }
   [ ,argument [ ,,...n ] ] ))
[WITH option [ ,,...n ] ]
```

A description of the components of the statement follows.

msg_id:-The ID for an error message, which is stored in the error column in sysmessages.

msg_str:-A custom message that is not contained in sysmessages.

Severity: - The severity level associated with the error. The valid values are 0–25. Severity levels 0–18 can be used by any user, but 19–25 are only available to members of the fixed-server role sysadmin. When levels 19–25 are used, the WITH LOG option is required. State A value that indicates the invocation state of the error. The valid values are 0–127. This value is not used by SQL Server.

Argument…

One or more variables that are used to customize the message. For example, you could pass the current process ID (@@SPID) so it could be displayed in the message.

WITH option . . .

The three values that can be used with this optional argument are described here.

LOG - Forces the error to logged in the SQL Server error log and the NT application log.

NOWAIT - Sends the message immediately to the client.

SETERROR - Sets @@ERROR to the unique ID for the message or 50,000.

The number of options available for the statement make it seem complicated, but it is actually easy to use. The following shows how to create an ad hoc message with a severity of 10 and a state of 1.

RAISERROR ('An error occurred updating the Nonfatal table', 10,1)

--Results--

An error occurred updating the nonfatal table

The statement does not have to be used in conjunction with any other code, but for our purposes, it will be used with the error handling code presented earlier. The following alters the ps_NonFatal_INSERT procedure to use RAISERROR.

```
USE tempdb
go
ALTER PROCEDURE ps_NonFatal_INSERT
@Column2 int =NULL
AS
DECLARE @ErrorMsgID int
INSERT Nonfatal VALUES (@Column2)
SET @ErrorMsgID =@@ERROR
IF @ErrorMsgID <>0
BEGIN
RAISERROR ('An error occurred updating the Nonfatal table',10,1)
END
```

When an error-producing call is made to the procedure, the custom message is passed to the client. The following shows the output generated by Query Analyzer.

Chapter 6: Service Broker

(Q) What do we need Queues?

There are instances when we expect that the other application with which we are interacting are not available. For example when you chat on messaging system like yahoo, MSN, ICQ etc, you do not expect that the other users will be guaranteed online. So there is where we need queues. So during chatting if the user is not online all the messages are sent to a queue. Later when the user comes online, he can read all messages from the queue.

(Q) What is “Asynchronous” communication?

Once a client has send messages to the other end application, he can continue with some other task without waiting for any notifications from the end client. For instance, take an example of any online email systems. Once you have sent a mail to the end user, you do not have to wait for

notification from the ends user. User just sends the message to queue, which is later picked up by the mailing system and sent to the desired end-user.

Note: - MSMQ does the messaging and queuing, but now the queuing functionality is leveraged to SQL Server 2005, due to its practical needs.

(Q) What is SQL Server Service broker?

SQL Server Service broker provides asynchronous queuing functionality to SQL Server. So now the end client will not have to wait. He can just say add these 1000 records and then come back after one hour or so to see has the work been done or not.

(Q) What are the essential components of SQL Server Service broker?

Following are the essential components of SQL Server:-

- **End-Points**

The endpoints can be two applications running on different servers or instances, or they can be two applications running on the same server.

- **Message**

A message is an entity that is exchanged between Server Brokers. A message must have a name and data type. Optionally, a message can have a validation on that type of data. A message is part of a conversation and it has a unique identifier as well as a unique sequence number to enforce message ordering.

- **Dialog**

Dialog ensure messages to be read in the same order as they where put in to queue between endpoints. In short, it ensures proper ordered sequence of events at both ends for a message.

- **Conversation Group**

Conversation Group is a logical grouping of Dialog. To complete a task you can need one or more dialog. For instance an online payment gateway can have two Dialog's first is the "Address Check" and second is the "Credit Card Number" validation, these both dialog form your complete "Payment process". So you can group both the dialogs in one Conversation Group.

- **Message Transport**

Message transport defines how the messages will be send across networks. Message transport is based on TCP/IP and FTP. There are two basic protocols "Binary Adjacent Broker Protocol" which is like TCP/IP and "Dialog Protocol" which like FTP.

(Q) What is the main purpose of having Conversation Group?

There two main purpose of having conversation group:-

- You can lock a conversation group during reading, so that no other process can read those queue entries.
- The most difficult thing in an asynchronous message system is to maintain states. There is huge delay between arrivals of two messages. So conversation groups maintains state using state table. Its uses instance ID to identify messages in a group.

(Q) How to implement Service Broker?

Below are the steps for practical implementation:-

- Create a Message type, which describes how the message is formed. If the message type is XML, you can also associate a schema with it.
- Further, you have to assign these Message type to Contract. Message type is grouped in Contracts. Contract is an entity, which describes messages for a particular Dialog.So a contract, can have multiple message type's.
- Contracts are further grouped in service. Service has all the dialogs needed to complete one process.
- Service can further be attached to multiple queues. Service is the basic object from SQL Server Service broker point of view.
- So when any client wants to communicate with a queue he opens a dialog with the service.

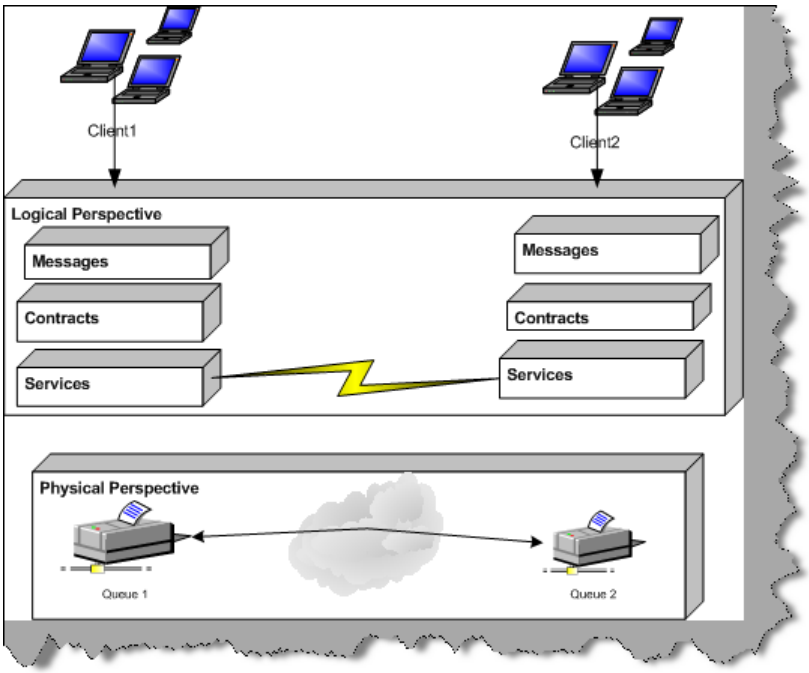


Figure 6.1: - SQL Server Service Broker in Action.

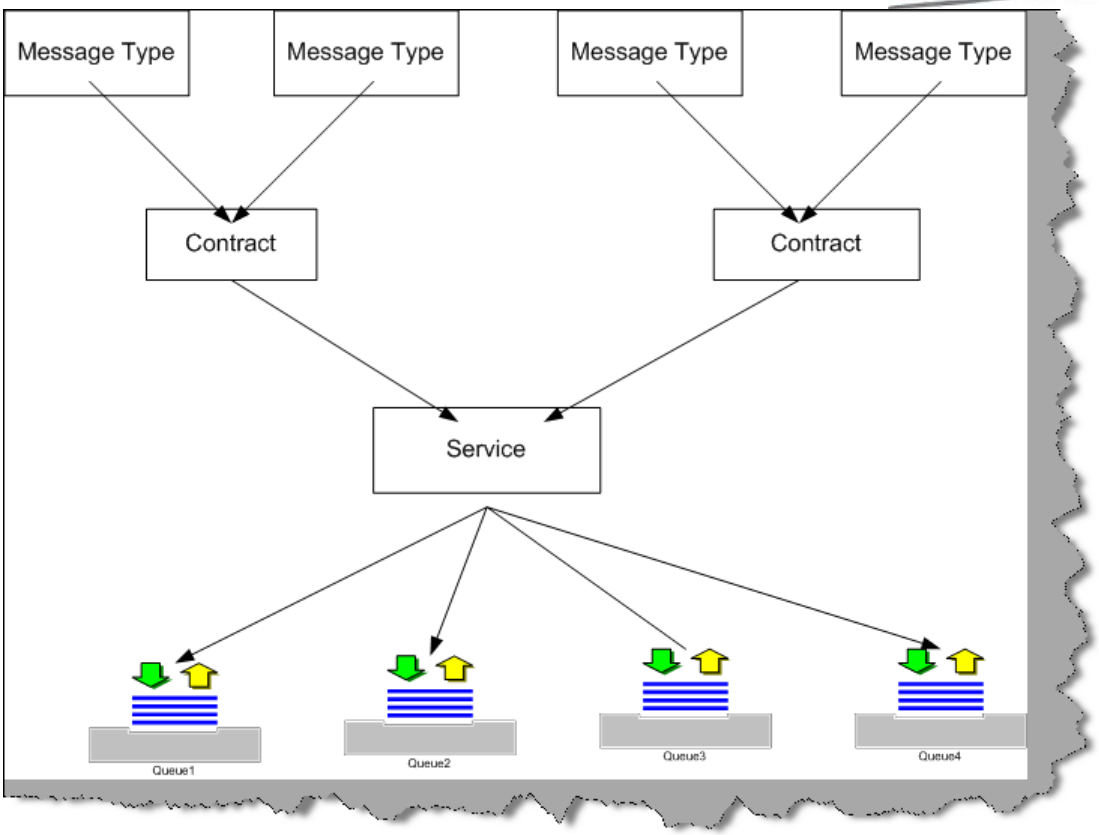


Figure 6.2: - Message, contract and service

Above figure shows how SQL Server Service broker works. Client who want to use the queues do not have to understand the complexity of queues. They only communicate with the logical view of SQL Server Service broker objects (Messages, Contracts, and Services). In turn, these objects interact with the queues below and shield the client from any physical complexities of queues.

Below is a simple practical implementation of how this works. Try running the below statements from a T-SQL and see the output.

```
-- Create a Message type and do not do any data type validation for
this
CREATE MESSAGE TYPE MessageType
VALIDATION = NONE
GO
```

```
-- Create Message contract what type of users can send these messages
at this moment we are defining current as an initiator
```

```
CREATE CONTRACT MessageContract
(MessageType SENT BY INITIATOR)
GO
-- Declare the two end points that's sender and receive queues
CREATE QUEUE SenderQ
CREATE QUEUE ReceiverQ
GO
-- Create service and bind them to the queues
CREATE SERVICE Sender
  ON QUEUE SenderQ
CREATE SERVICE Receiver
  ON QUEUE ReceiverQ (MessageContract)
GO
-- Send message to the queue
DECLARE @conversationHandle UNIQUEIDENTIFIER
DECLARE @message NVARCHAR(100)
BEGIN
  BEGIN TRANSACTION;
  BEGIN DIALOG @conversationHandle
        FROM SERVICE Sender
        TO SERVICE 'Receiver'
        ON CONTRACT MessageContract
```

```
- Sending message
  SET @message = N'SQL Server Interview Questions by Shivprasad
Koirala';
  SEND  ON CONVERSATION @conversationHandle
        MESSAGE TYPE MessageType (@message)
  COMMIT TRANSACTION
END
GO
-- Receive a message from the queue
```

```
RECEIVE CONVERT(NVARCHAR(max), message_body) AS message
 FROM ReceiverQ
-- Just dropping all the object so that this sample can run
successfully
DROP SERVICE Sender
DROP SERVICE Receiver
DROP QUEUE SenderQ
DROP QUEUE ReceiverQ
DROP CONTRACT MessageContract
DROP MESSAGE TYPE MessageType
GO
```

After executing the above T-SQL command you can see the output below.

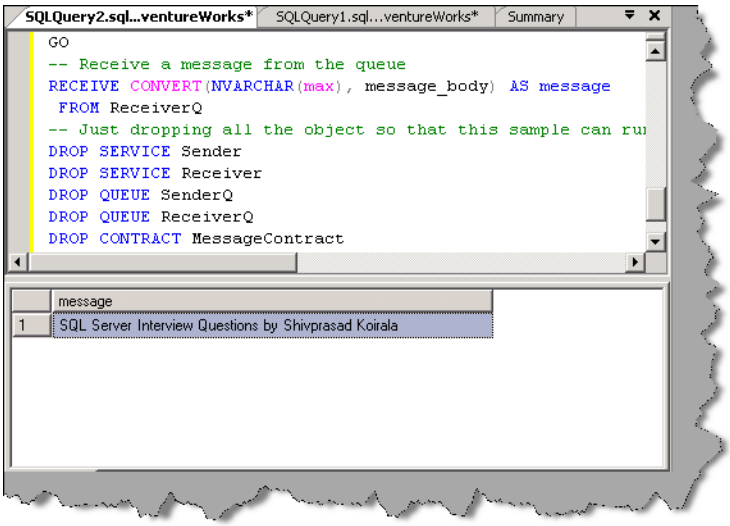


Figure 6.3: - Output of the above sample

Note:- In case your SQL Server service broker is not active you will get the following error as shown below. In order to remove that error you have to enable the service broker by using

```
Alter Database [DatabaseName] set Enable_broker
```

At this moment I have created all these samples in the sample database "AdventureWorks".

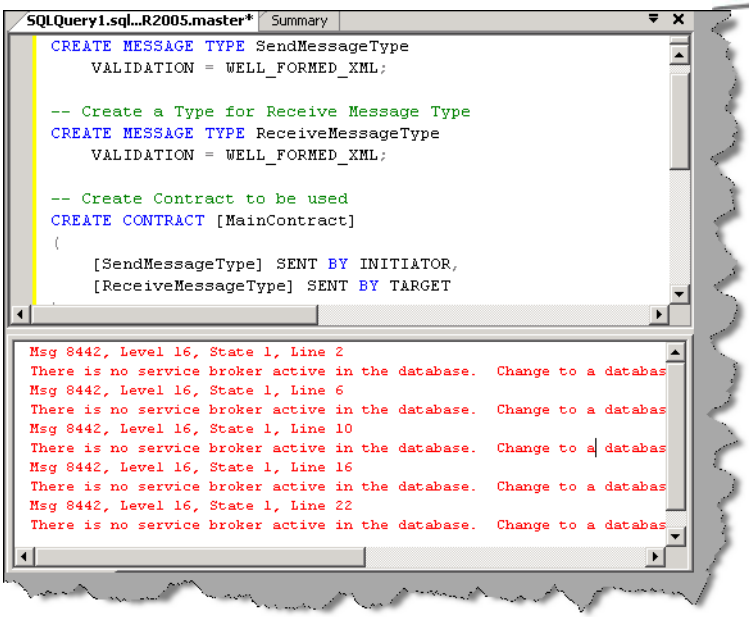


Figure 6.4: - Error Service broker not active

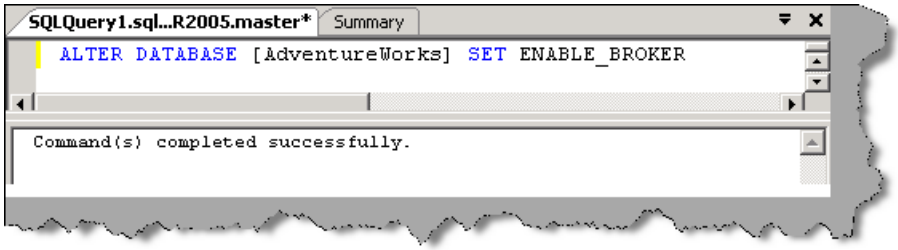


Figure 6.5: - Enabling Service broker

(Q) How do we encrypt data between Dialogs?

If you create a dialog using "WITH ENCRYPTION" clause a session key is created that is used to encrypt the messages sent between dialog.

7. XML Integration

Note: - In this chapter we will first just skim through basic XML interview questions so that you do not get stuck up with simple questions.

(Q) What is XML?

XML (Extensible markup language) is all about describing data. Below is a XML, which describes invoice data.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<invoice>
<productname>Shoes</productname>
<qty>12</qty>
<totalcost>100</totalcost>
<discount>10</discount>
</invoice>
```

An XML tag is not something predefined but it is something you have to define according to your needs. For instance in the above example of invoice all tags are defined according to business needs. The XML document is self-explanatory; any one can easily understand looking at the XML data what exactly it means.

(Q) What is the version information in XML?

"Version" tag shows which version of XML is used.

(Q) What is ROOT element in XML?

In our XML sample given previously <invoice></invoice> tag is the root element. Root element is the top most elements for a XML.

(Q) If XML does not have closing tag will it work?

No, every tag in XML which is opened should have a closing tag. For instance in the top if I remove </discount> tag that XML will not be understood by lot of application.

(Q) Is XML case sensitive?

Yes, they are case sensitive.

(Q) What is the difference between XML and HTML?

XML describes data while HTML describes how the data should be displayed. So HTML is about displaying information while XML is about describing information.

(Q) Is XML meant to replace HTML?

No they both go together one is for describing data while other is for displaying data.

(Q) Can you explain why your project needed XML?

Note: - This is an interview question where the interviewer wants to know why you have chosen XML.

Remember XML was meant to exchange data between two entities as you can define your user-friendly tags with ease. In real world scenarios, XML is meant to exchange data. For instance, you have two applications who want to exchange information. But because they work in two complete opposite technologies, it is difficult to do it technically. For instance, one application is

made in JAVA and the other in .NET. But both languages understand XML so one of the applications will spit XML file which will be consumed and parsed by other applications

You can give a scenario of two applications, which are working separately and how you chose XML as the data transport medium.

(Q) What is DTD (Document Type definition)?

It defines how your XML should structure. For instance in the above XML we want to make it compulsory to provide "qty" and "total cost", also that these two elements can only contain numeric. So you can define the DTD document and use that DTD document with in that XML.

(Q) What is well formed XML?

If a XML document is confirming to XML rules (all tags started are closed, there is a root element etc) then it is a well-formed XML.

(Q) What is a valid XML?

If XML is confirming to DTD rules then it is a valid XML.

(Q) What is CDATA section in XML?

All data is normally parsed in XML but if you want to exclude some elements, you will need to put those elements in CDATA.

(Q) What is CSS?

With CSS, you can format a XML document.

(Q) What is XSL?

XSL (the extensible Stylesheet Language) is used to transform XML document to some other document. So its transformation document which can convert XML to some other document. For instance, you can apply XSL to XML and convert it to HTML document or probably CSV files.

(Q) What is Element and attributes in XML?

In the below example invoice is the element and the in number the attribute.

```
<invoice in number=1002></invoice>
```

(Q) Can we define a column as XML?

Yes, this is a new feature provided by SQL Server. You can define a column data type as XML for a table.

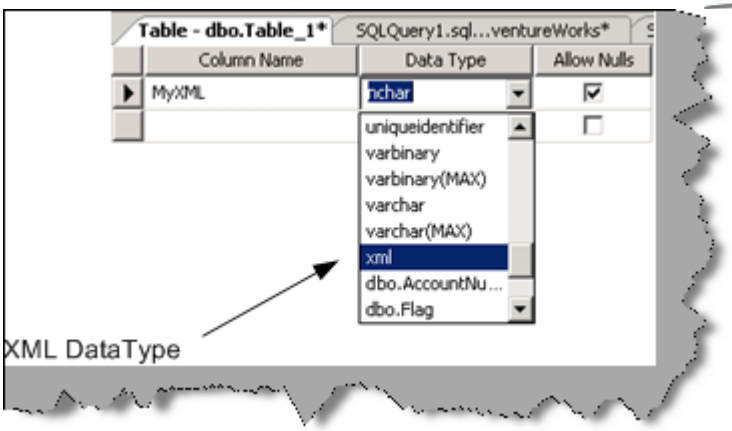


Figure 7.1: - Specify XML data type

(Q) How do we specify the XML data type as typed or untyped?

If there is a XSD schema specified to the data type then it is typed or else it's untyped. If you specify XSD then with every insert SQL Server will try to validate and see that is the data adhering to XSD specification of the data type.

(Q) How can we create the XSD schema?

Below is the DDL statement for creating XML schema.

```
CREATE XML SCHEMA COLLECTION MyXSD AS
N'<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" targetNamespace="http://MyXSD">
<xs:element name="MyXSD">
<xs:complexType>
<xs:sequence>
<xs:element name="Orderid" type="xs:string" />
<xs:element name="CustomerName" type="xs:string" />
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>'
```

After you have created the schema, you see the MYXSD schema in the schema collections folder.

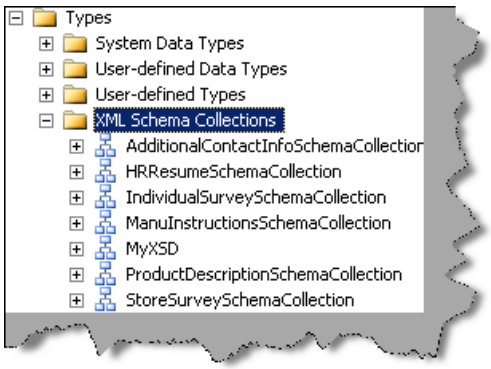


Figure 7.2: - You can view the XSD in explorer of Management Studio

When you create the XML data type, you can assign the MyXsd to the column.

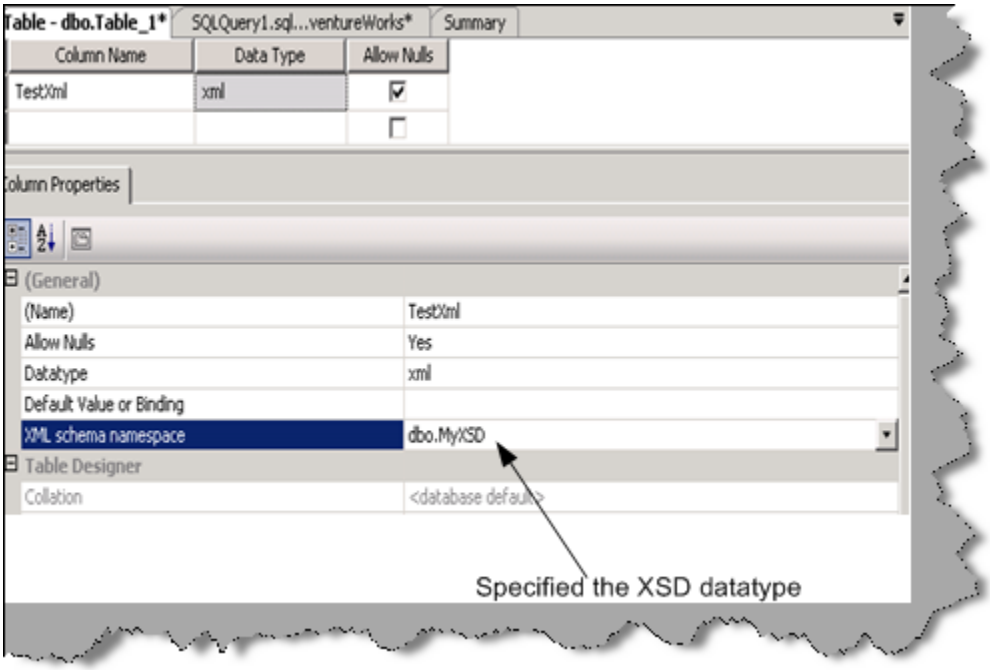


Figure 7.3: - MyXSD assigned to a column

(Q) How do I insert in to a table that has XSD schema attached to it?

I know many developers will just say what the problem with simple insert statement. Well guys it's not easy with attaching the XSD its now a well formed datatype.The above table I have named as xmltable. So we had specified in the schema two nodes one is ordered and the other customer name. So here is the insert.

```
Insert into xmltable values ('<MyXSD
xmlns="http://MyXSD"><Orderid>1</Orderid><CustomerName>Shiv</CustomerNa
me></MyXSD>')
```

(Q) What is maximum size for XML data type?

2 GB and is stored like varbinary.

(Q) What is Xquery?

In a typical XML table below is the type of data, which is seen. Now I want to retrieve orderid "4". I know many will jump up with saying use the "LIKE" keyword. Ok you say that interviewer is very sure that you do not know the real power of XML provided by SQL Server.

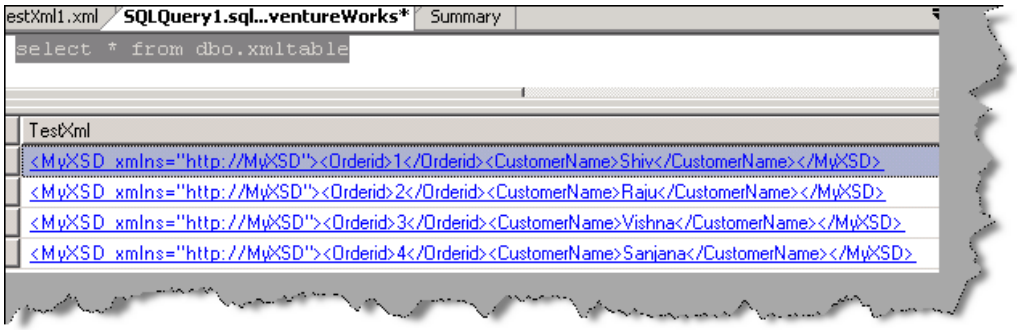


Figure 7.4: - XML data

Well first thing, XQUERY is not that something Microsoft invented, it's a language defined by W3C to query and manipulate data in a XML. For instance in the above scenario we can use XQUERY and drill down to specific element in XML.

So to drill down here's the XQUERY

```
SELECT * FROM xmltable
WHERE TestXml.exist('declare namespace
xd=http://MyXSD/xd:MyXSD[xd:Orderid eq "4"]') = 1
```

```
Note: - It's out of the scope of this book to discuss XQUERY.  I hope
and only hope guys many interviewers will not bang in this section. In
case you have doubt visit www.w3c.org or SQL Server books online they
have a lot of material in to this.
```

(Q) What are XML indexes?

XML data types have huge size 2 GB. But first thing is that you should have a primary key on the XML data type column. Then you can use the below SQL statement to create index on the XML column:-

```
CREATE PRIMARY XML INDEX xmlindex ON xmltable(TestXML)
```

(Q) What are secondary XML indexes?

Secondary indexes are built on document attributes.

(Q) What is FOR XML in SQL Server?

FOR XML clause returns data in XML rather than simple rows and columns. For instance if you fire the below query on any table, you will get XML output:-

```
SELECT * FROM MyTable FOR XML AUTO
```

(Q) Can I use FOR XML to generate SCHEMA of a table and how?

The below SQL syntax will return the SCHEMA of the table.

```
SELECT * FROM MyTable FOR XML AUTO, XMLSCHEMA
```

(Q) What is the OPENXML statement in SQL Server?

We had seen that FOR XML returns a XML format of a table data. And FOR XML does the vice versa of it. If you pass XML document to it will convert it to rows and columns.

(Q) I have huge XML file which we want to load in database?

Twist: - Can I do a BULK load of XML in database?

Below is the SQL statement, which will insert from "MyXml.xml" in to "MyTable".

INSERT into MyTable (MyXMlColumn) SELECT * FROM OPENROWSET

(Bulk 'c:\MyXml.xml', SINGLE_CLOB) as abc

(Q) How to call stored procedure using HTTP SOAP?

Twist: - Can I create web services for SQL Server objects?

Note: - Ok every one reading this answer out of dedication I have switched off my mobile and I am writing this answer.

You can call a stored procedure using HTTP SOAP. This can be done by creating END POINTS using the "CREATE ENDPOINT" DDL statement. I have created a TotalSalesHttpEndPoint, which can be called later through "web services".

```
CREATE ENDPOINT TotalSalesHttpEndPOint
STATE = STARTED
AS HTTP(
PATH = '/sql',
AUTHENTICATION = (INTEGRATED ),
PORTS = ( CLEAR ),
```

```
SITE = 'server'
)
FOR SOAP (
WEBMETHOD 'http://tempUri.org/'.'GetTotalSalesOfProduct'
(name='AdventureWorks.dbo.GetTotalSalesOfProduct',
schema=STANDARD ),
BATCHES = ENABLED,
WSDL = DEFAULT,
DATABASE = 'AdventureWorks',
NAMESPACE = 'http://AdventureWorks/TotalSales'
)
```

(Q) What is XMLA?

XMLA stand for XML for Analysis Services. Analysis service is covered in depth in data mining and data ware housing chapters. Using XMLA, we can expose the Analysis service data to the external world in XML. So that any data source can consume it as XML is universally known.

Chapter 8: Data Warehousing / Data Mining

Note: - "Data mining" and "Data Warehousing" are concepts which are very wide and it's beyond the scope of this book to discuss it in depth. So if you are specially looking for a "Data mining / warehousing" job its better to go through some reference books. But below questions can shield you to some good limit.

(Q) What is "Data Warehousing"?

"Data Warehousing "is a process in which the data is stored and accessed from central location and is meant to support some strategic decisions. "Data Warehousing" is not a requirement for "Data mining". But just makes your Data mining process more efficient.

Data warehouse is a collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time.

(Q) What are Data Marts?

Data Marts are smaller section of Data Warehouses. They help data warehouses collect data. For example, your company has lot of branches that are spanned across the globe. Head-office of the company decides to collect data from all these branches for anticipating market. In order to

achieve this IT department can setup data mart in all branch offices and a central data warehouse where all data will finally reside.

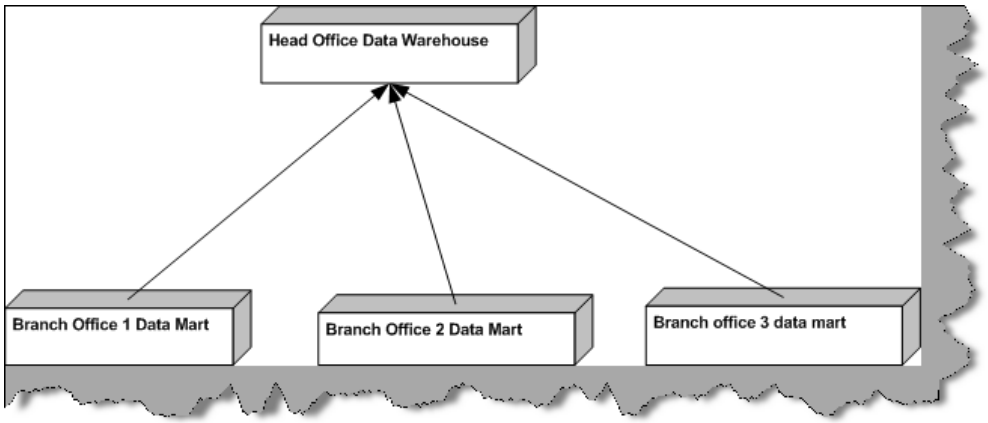


Figure 8.1: - DataMart in action

(Q) What are Fact tables and Dimension Tables?

Twist: - What is Dimensional Modeling?

Twist: - What is Star Schema Design?

When we design transactional database we always think in terms of normalizing design to its least form. However, when it comes to designing for Data warehouse we think more in terms of "denormalizing" the database. Data warehousing databases are designed using "Dimensional Modeling". Dimensional Modeling uses the existing relational database structure and builds on that.

There are two basic tables in dimensional modeling:-

- Fact Tables.
- Dimension Tables.

Fact tables are central tables in data warehousing. Fact tables have the actual aggregate values that will be needed in a business process. While dimension tables revolve around fact tables. They describe the attributes of the fact tables. Let us try to understand these two conceptually.

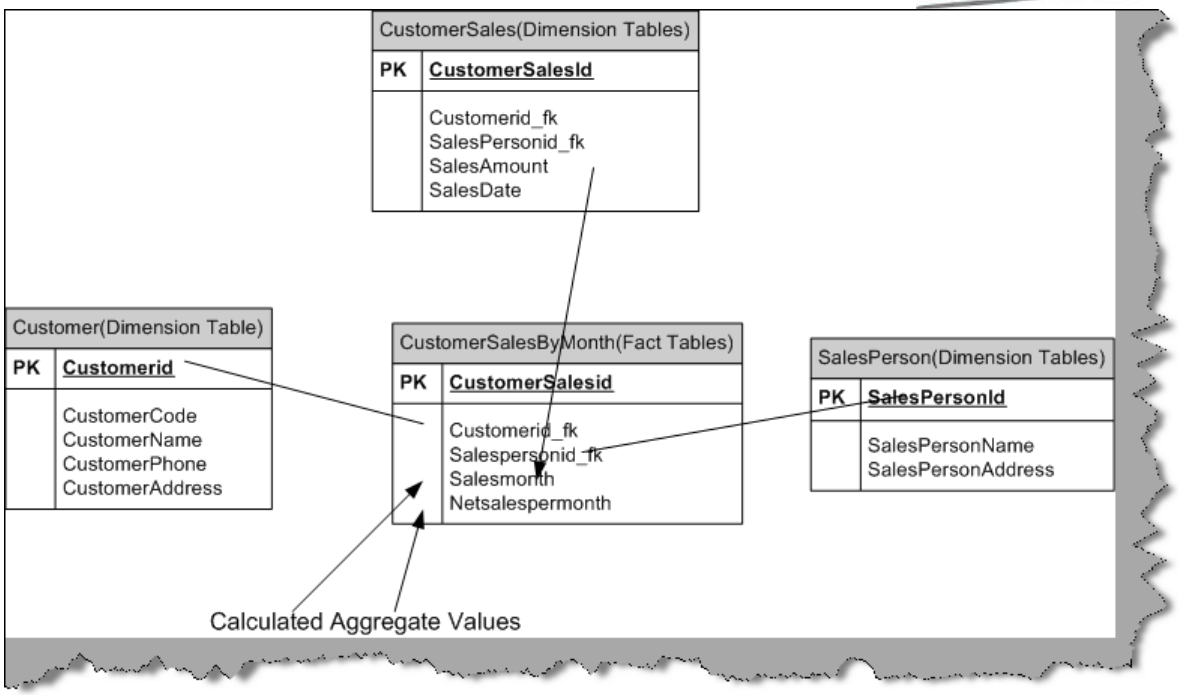


Figure 8.2: - Dimensional Modeling

In the above example, we have three tables, which are transactional tables:-

- **Customer: -** It has the customer information details.
- **Salesperson:** - Sales person who are actually selling products to customer.
- **CustomerSales: -** This table has data of which sales person sold to which customer and what was the sales amount.

Below is the expected report Sales / Customer / Month. You will be wondering if we make a simple join query from all three tables, we can easily get this output. However, imagine if you have huge records in these three tables, it can really slow down your reporting process. So we introduced a third dimension table "CustomerSalesByMonth" which will have foreign key of all tables and the aggregate amount by month. So this table becomes the dimension table and all other tables become fact tables. All major data warehousing design use Fact and Dimension model.

Customer Name	Sales Person Name	Month	Sales Amount Per Month
Man Brothers	Rajesh	Jan	1000
Suman Motela	Shiv	Jan	2000
KL enterprises	Rajesh	feb	500
KL enterprises	Shiv	Jan	1000

Figure 8.3: - Expected Report.

The above designs are also known as Star Schema design.

Note: - For a pure data warehousing job this question is important. So try to understand why we modeled out design in this way rather than using the traditional approach - normalization.

(DB)What is Snow Flake Schema design in database?

Twist: - What is the difference between Star and Snowflake schema?

Star schema is good when you do not have big tables in data warehousing. However, when tables start becoming really huge it is better to denormalize. When you denormalize star schema it is nothing but snow flake design. For instance below "customeraddress" table is been normalized and is a child table of "Customer" table. Same holds true for "Salesperson" table.

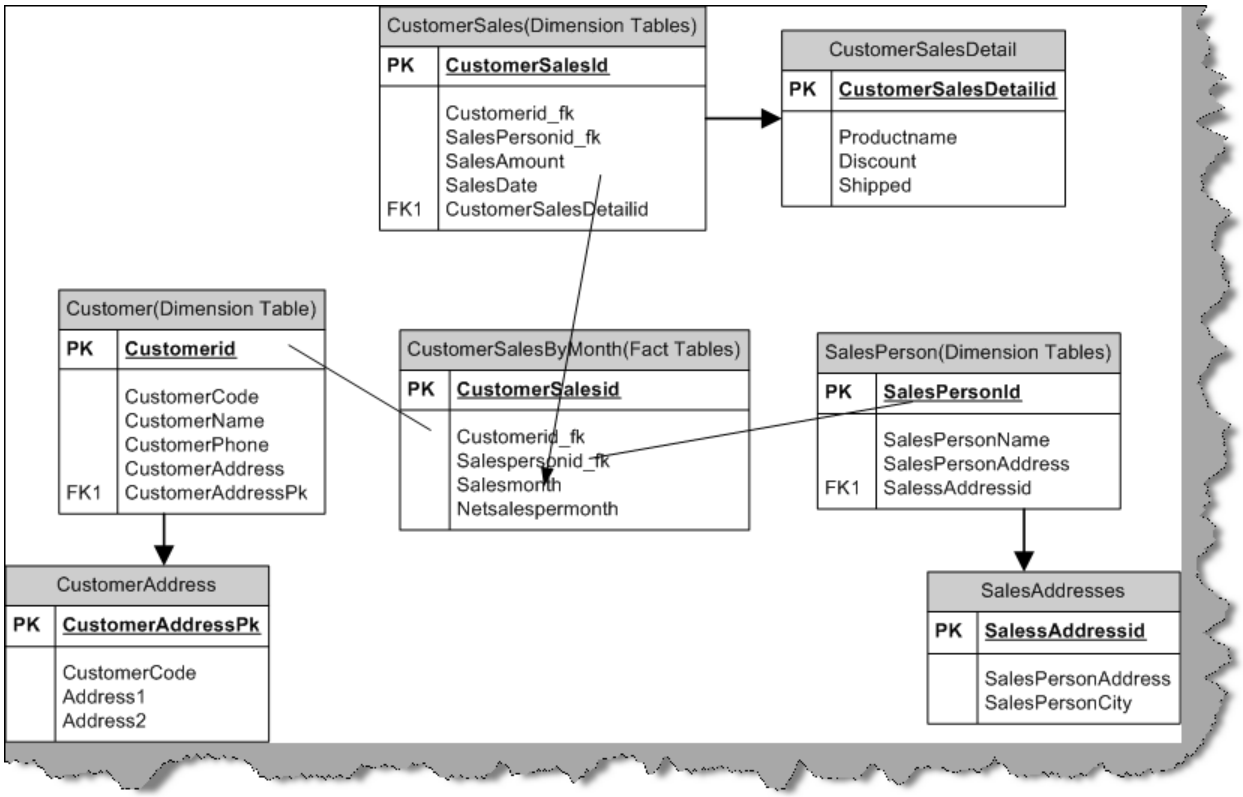


Figure 8.4: - Snow Flake Schema

(DB) What is ETL process in Data warehousing?

Twist: - What are the different stages in "Data warehousing"?

ETL (Extraction, Transformation and Loading) are different stages in Data warehousing. Like when we do software development, we follow different stages like requirement gathering, designing, coding and testing. In the similar fashion, we have for data warehousing.

Extraction:-

In this process, we extract data from the source. In actual scenarios data source can be in many forms EXCEL, ACCESS, Delimited text, CSV (Comma Separated Files) etc. Therefore, extraction process handle is the complexity of understanding the data source and loading it in a structure of data warehouse.

Transformation:-

This process can also be called as cleaning up process. It is not necessary that after the extraction process data is clean and valid. For instance, all the financial figures have NULL values but you want it to be ZERO for better analysis. Therefore, you can have some kind of stored procedure that runs through all extracted records and sets the value to zero.

Loading:-

After transformation, you are ready to load the information in to your final data warehouse database.

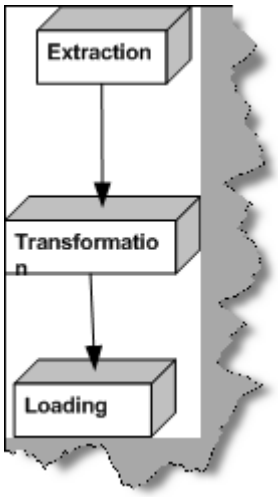


Figure 8.5: - ETL stages

(DB) How can we do ETL process in SQL Server?

I can hear that scream: - Words and words, show us where does this ETL practically fit in SQL Server.

SQL Server has following ways with which we can import or export data in SQL Server:-

- BCP (Bulk Copy Program).

- Bulk Insert
- DTS (Data Transformation Services).DTS is now called as Integration Services.

(Q) What is “Data mining”?

“Data mining” is a concept by which we can analyze the current data from different perspectives and summarize the information in more useful manner. It is mostly used either to derive some valuable information from the existing data or to predict sales to increase customer market.

There are two basic aims of “Data mining”:-

- **Prediction: -** From the given data, we can focus on how the customer or market will perform. For instance, we are having a sale of 40000 $ per month in India, if the same product is to be sold with a discount how much sales can the company expect.
- **Summarization:** - To derive important information to analyze the current business scenario. For example, a weekly sales report will give a picture to the top management how we are performing on a weekly basis?

(Q) Compare “Data mining” and “Data Warehousing”?

“Data Warehousing” is technical process where we are making our data centralized while “Data mining” is more of business activity which will analyze how good your business is doing or predict how it will do in the future coming times using the current data.

As said before “Data Warehousing” is not a need for “Data mining”. It is good if you are doing “Data mining” on a “Data Warehouse” rather than on an actual production database. “Data Warehousing” is essential when we want to consolidate data from different sources, so it’s like a cleaner and matured data which sits in between the various data sources and brings then in to one format. “Data Warehouses” are normally physical entities, which are meant to improve accuracy of “Data mining” process. For example, you have 10 companies sending data in different format, so you create one physical database for consolidating all the data from different company sources, while “Data mining” can be a physical model or logical model. You can create a database in “Data mining” which gives you reports of net sales for this year for all companies. This need not be a physical database as such but a simple query.

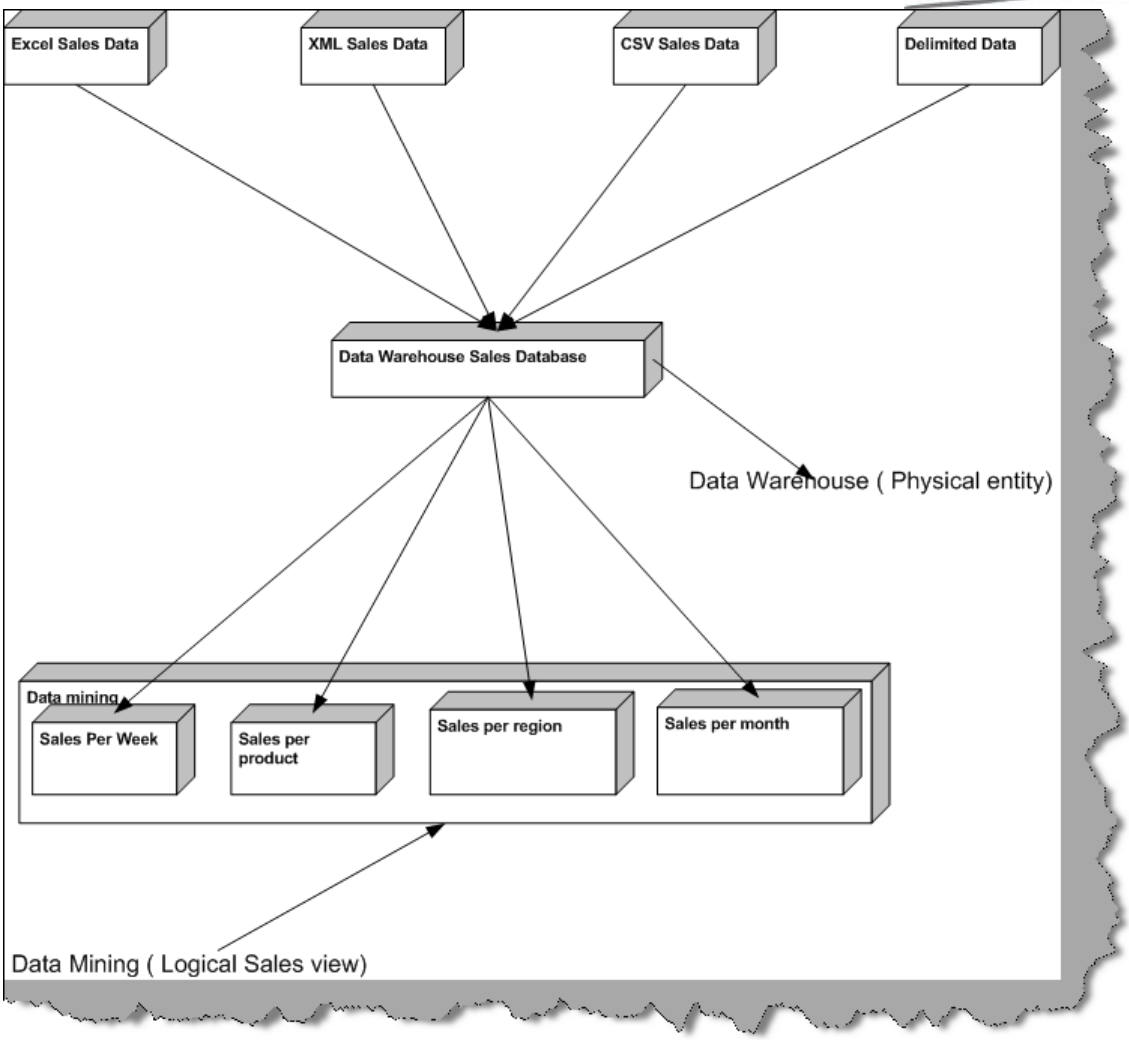


Figure 8.6: - Data Warehouse and Data mining

The above figure gives a picture how these concepts are quiet different. "Data Warehouse" collects cleans and filters data through different sources like "Excel", "XML" etc. But "Data Mining" sits on the top of "Data Warehouse" database and generates intelligent reports. Now either it can export to a different database or just generate report using some reporting tool like "Reporting Services".

(Q) What is BCP?

Note: - It is not necessary that this question will be asked for data mining. But if a interviewer wants to know your DBA capabilities he will love to ask this question. If he is a guy who has worked from the old days of SQL Server, he will expect this to be answered.

There are times when you want to move huge records in and out of SQL Server, there is where this old and cryptic friend will come to use. It is a command line utility. Below is the detail syntax:-

```
bcp {[[<database name>.][<owner>].]{<table name>|<view
name>}|"<query>"}
{in | out | queryout | format} <data file>
[-m <maximum no. of errors>] [-f <format file>] [-e <error file>]
[-F <first row>] [-L <last row>] [-b <batch size>]
[-n] [-c] [-w] [-N] [-V (60 | 65 | 70)] [-6]
[-q] [-C <code page>] [-t <field term>] [-r <row term>]
[-i <input file>] [-o <output file>] [-a <packet size>]
[-S <server name>[\<instance name>]] [-U <login id>] [-P <password>]
[-T] [-v] [-R] [-k] [-E] [-h "<hint> [,...n]"]
```

UUUHH Lot of attributes there. However, during interview, you do not have to remember so much. Just remember that BCP is a utility with which you can do import and export of data.

(Q) How can we import and export using BCP utility?

In the first question, you can see there is huge list of different command. We will try to cover only the basic commands that are used.

-T: - signifies that we are using windows authentification

-t: - by default, every record is tab separated. But if you want to specify comma separated you can use this command.

-r: - This specifies how every row is separated. For instance specifying –r/n specifies that every record will be separated by ENTER.

Bcp adventureworks.sales.salesperson out c:\salesperson.txt -T

Bcp adventureworks.sales.salespersondummy in c:\salesperson.txt –T

When you execute the BCP syntax, you will be prompted to enter the following values (data type, length of the field and the separator) as shown in figure below. You can either fill it or just press enter to escape it. BCP will take in the default values.

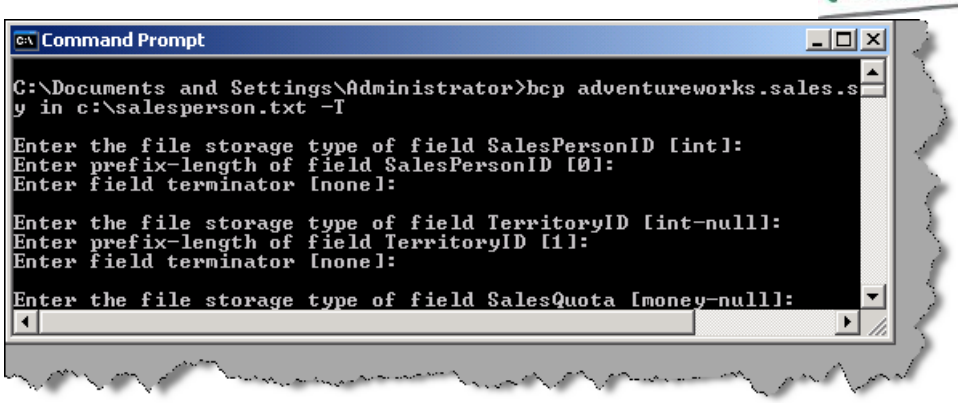


Figure 8.7: - After executing BCP command prompts for some properties

(Q) During BCP we need to change the field position or eliminate some fields how can we achieve this?

For some reason during BCP, you want some fields to be eliminated or you want the positions to be in a different manner. For instance, you have field1, field2 and field3. You want that field2 should not be imported during BCP. Alternatively, you want the sequence to be changed as field1, field2 and then finally field3. This is achieved by using the format file. When we ran the BCP command in the first question, it has generated a file with ".fmt" extension. Below is the FMT file generated in the same directory from where I ran my BCP command.

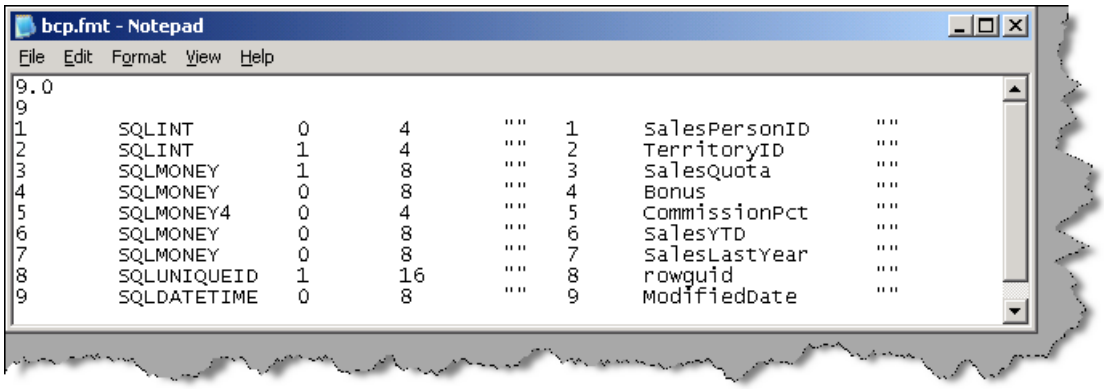


Figure 8.8: - Format file generated due to BCP.

FMT file is basically the format file for BCP to govern how it should map with tables. Let us say, in from our salesperson table we want to eliminate commissionpct, salesytd and saleslastyear. Therefore, you have to modify the FMT file as shown below. We have made the values zero for the fields that has to be eliminated.

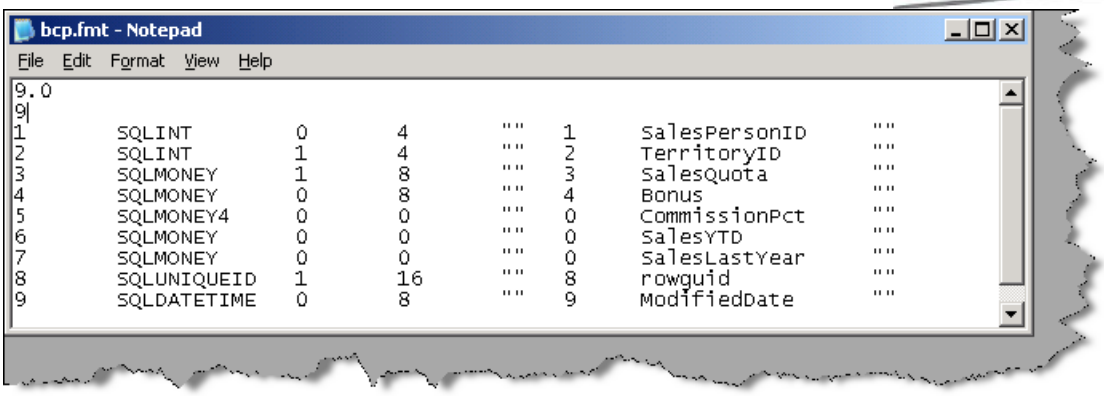


Figure 8.9: - FMT file with fields eliminated

If we want to change the sequence, you have to just change the original sequence number. For instance we have changed the sequence from 9 to 5 --> 5 to 9 , see the figure below.

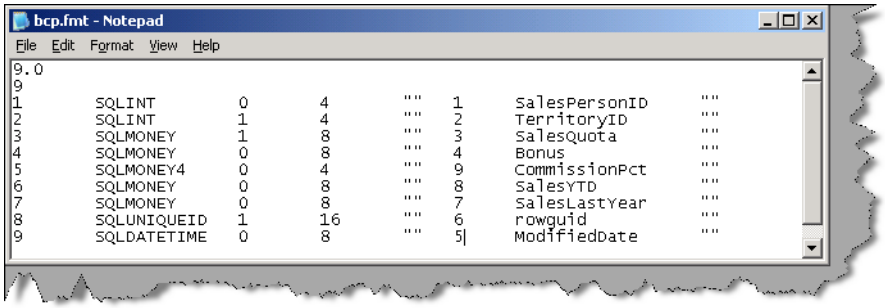


Figure 8.10 : - FMT file with field sequence changed

Once you have changed the FMT file you can specify the .FMT file in the BCP command arguments as shown below.

```
bcp adventureworks.sales.salesperson in c:\salesperson.txt -
c:\bcp.fmt -T
```

```
Note: - we have given the .FMT file in the BCP command.
```

(Q) What is Bulk Insert?

Bulk insert is very similar to BCP command but we cannot do export with the command. The major difference between BCP and Bulk Insert:-

- Bulk Insert runs in the same process of SQL Server, so it can avail to all performance benefits of SQL Server.
- You can define Bulk insert as part of transaction. That means you can use the Bulk Insert command in BEGIN TRANS and COMMIT TRANS statements.

Below is a detailed syntax of BULK INSERT. You can run this from "SQL Server Management Studio", TSQL or ISQL.

```
BULK INSERT [['database_name'.]['owner'].]
{'table_name' | 'view_name' FROM 'data_file' }
[WITH (
[BATCHSIZE [ = batch_size ]]
[[,] CHECK_CONSTRAINTS ]
[[,] CODEPAGE [ = 'ACP' | 'OEM' | 'RAW' | 'code_page' ]]
[[,] DATAFILETYPE [ = {'char'|'native'|
'widechar'|'widenative' }]]
[[,] FIELDTERMINATOR [ = 'field_terminator' ]]
[[,] FIRSTROW [ = first_row ]]
[[,] FIRETRIGGERS [ = fire_triggers ]]
[[,] FORMATFILE [ = 'format_file_path' ]]
[[,] KEEPIDENTITY ]
[[,] KEEPNULLS ]
[[,] KILOBYTES_PER_BATCH [ = kilobytes_per_batch ]]
[[,] LASTROW [ = last_row ]]
[[,] MAXERRORS [ = max_errors ]]
[[,] ORDER ( { column [ ASC | DESC ]}[ ,…n ])]
[[,] ROWS_PER_BATCH [ = rows_per_batch ]]
[[,] ROWTERMINATOR [ = 'row_terminator' ]]
[[,] TABLOCK ]
)]
```

Below is a simplified version of bulk insert, which we have used to import a comma separated file in to "Salesperson Dummy". The first row is the column name so we specified start importing from the second row. The other two attributes define how the fields and rows are separated.

```
bulk insert adventureworks.sales.salespersondummy from
'c:\salesperson.txt' with
(
FIRSTROW=2,
FIELDTERMINATOR = ',',
ROWTERMINATOR = '\n'
)
```

(Q) What is DTS?

Note : - It is now a part of integration service in SQL Server 2005.

DTS provides similar functionality as we had with BCP and Bulk Import. There are two major problems with BCP and Bulk Import:-

- BCP and Bulk import do not have user friendly User Interface. Well some DBA does still enjoy using those DOS prompt commands, which makes them feel doing something worthy.
- Using BCP and Bulk imports, we can import only from files, what if we wanted to import from other database like FoxPro, access, and oracle. That is where DTS is the king.
- One of the important things that BCP and Bulk insert misses is transformation, which is one of the important parts of ETL process. BCP and Bulk insert allows you to extract and load data, but does not provide any means by which you can do transformation. So for example you are getting sex as "1" and "2", you would like to transform this data to "M" and "F" respectively when loading in to data warehouse
- It also allows you do direct programming and write scripts by which you can have huge control over loading and transformation process.
- It allows lot of parallel operation to happen. For instance while you are reading data you also want the transformation to happen in parallel , then DTS is the right choice.

You can see DTS Import / Export wizard in the SQL Server 2005 menu.

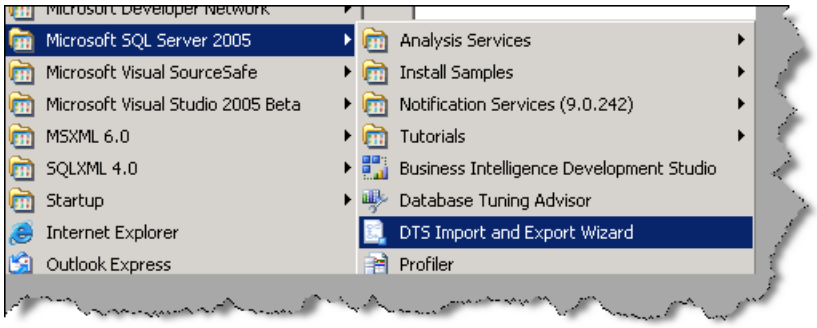


Figure 8.11: - DTS Import Export

Note: - DTS is the most used technology when you are during Data warehousing using SQL Server. In order to implement the ETL fundamental properly Microsoft has rewritten the whole DTS from scratch using .NET and named it as "Integration Services". There is a complete chapter which is dedicated to "Integration Services" which will cover DTS indirectly in huge details. Any interviewer who is looking for data warehousing professional in SQL Server 2005 will expect that candidates should know DTS properly.

(DB) Can you brief about the Data warehouse project you worked on?

Note: - This question is the trickiest and shoot to have insight, from where the interviewer would like to spawn question threads. If you have worked with a data warehouse project you can be very sure of this. If not then you really have to prepare a project to talk about…. I know it's unethical to even talk in books but?

I leave this to readers, as everyone would like to think of a project of his own. But just try to include the ETL process which every interviewer thinks should be followed for a data warehouse project.

(Q) What is an OLTP (Online Transaction Processing) System?

Following are the characteristics of an OLTP system:-

- They describe the actual current data of the system
- Transactions are short. For example user fills in data and closes the transaction.
- Insert/Update/Delete operation is completely online.
- System design expected to be in the maximum Normalized form.
- Huge volume of transactions. Example lots of online users are entering data in to an online tax application.
- Backup of transaction is necessary and needs to be recovered in case of problems

Note: - OLTP systems are good at putting data in to database system but serve no good when it comes to analyzing data.

(Q) What is an OLAP (On-line Analytical processing) system?

Following are characteristic of an OLAP system:-

- It has historical as well as current data.
- Transactions are long. They are normally batch transaction that is being executed during night hours.
- OLAP systems are mainly used for reporting or batch processing, so "Denormalization" designs are encouraged.
- Transactions are mainly batch transactions that are running so there are no huge volumes of transaction.
- Do not need to have recovery process as such until the project specifies specifically.

(Q) What is Conceptual, Logical and Physical model?

Depending on clients requirement first you define the conceptual model followed by logical and physical model.

Conceptual model involves with only identifying entities and relationship between. Fields / Attributes are not planned at this stage. It is just an identifying stage but not in detail.

Logical model involves in actually identifying the attributes, primary keys, many-to-many relationships etc of the entity. In short, it is the complete detail planning of what actually has to be implemented.

Physical model is where you develop your actual structure tables, fields, primary keys, foreign leys etc. You can say it is the actual implementation of the project.

Note: - To Design conceptual and logical model mostly VISIO is used and some company combine this both model in one time. So you will not be able to distinguish between both models.

(DB) What is Data purging?

You can also call this as data cleaning. After you have designed your data warehouse and started importing data, there is always a possibility you can get in lot of junk data. For example, you have some rows which have NULL and spaces, so you can run a routine, which can delete these kinds of records. Therefore, this cleaning process are known as “Data Purging”.

(Q) What is Analysis Services?

Analysis Services (previously known as OLAP Services) was designed to draw reports from data contained in a "Data Warehouses”.” Data Warehouses" do not have typical relational data structure (fully normalized way), but rather have snowflake or star schema (refer star schema in the previous sections).

The data in a data warehouse is processed using online analytical processing (OLAP) technology. Unlike relational technology, which derives results by reading and joining data when the query is issued, OLAP is optimized to navigate the summary data too quickly return results. As we are not going through any joins (because data is in denormalized form), SQL queries are executed faster and in more optimized way.

(DB) What are CUBES?

As said in previous question analysis services do not work on relation tables, but rather use “CUBES”. Cubes have two important attributes dimensions and measures. Dimensions are data like Customer type, country name and product type. While measures are quantitative data like dollars, meters and weight. Aggregates derived from original data are stored in cubes.

(DB) What are the primary ways to store data in OLAP?

We store information in OLAP in primarily in three ways:-

MOLAP

Multidimensional OLAP (MOLAP) stores dimension and fact data in a persistent data store using compressed indexes. Aggregates are stored to facilitate fast data access. MOLAP query engines are usually proprietary and optimized for the storage format used by the MOLAP data store. MOLAP offers faster query processing than ROLAP and usually requires less storage. However, it does not scale as well and requires a separate database for storage.

ROLAP

Relational OLAP (ROLAP) stores aggregates in relational database tables. ROLAP use of the relational databases allows it to take advantage of existing database resources, plus it allows ROLAP applications to scale well. However, ROLAP's use of tables to store aggregates usually requires more disk storage than MOLAP, and it is generally not as fast.

HOLAP

As its name suggests, hybrid OLAP (HOLAP) is a cross between MOLAP and ROLAP. Like ROLAP, HOLAP leaves the primary data stored in the source database. Like MOLAP, HOLAP stores aggregates in a persistent data store that is separate from the primary relational database. This mix allows HOLAP to offer the advantages of both MOLAP and ROLAP. However, unlike MOLAP and ROLAP, which follow well-defined standards, HOLAP has no uniform implementation.

(DB) What is META DATA information in Data warehousing projects?

META DATA is data about data. Well that is not an enough definition for interviews we need something more than that to tell the interviewer. It is the complete documentation of a data warehouse project. From perspective of SQL Server all Meta data is stored in Microsoft repository. It is all about way the structure is of data warehouse, OLAP, DTS packages.

Just to summarize some elements of data warehouse Meta data are as follows:-

- Source specifications — such as repositories, source schemas etc.
- Source descriptive information — such as ownership descriptions, updates frequencies, legal limitations, access methods etc.
- Process information — such as job schedules, extraction code.
- Data acquisition information — such as data transmission scheduling, results and file usage.
- Dimension table management — such as definitions of dimensions, surrogate key.
- Transformation and aggregation — such as data enhancement and mapping, DBMS load scripts, aggregate definitions &c.
- DMBS system table contents,
- descriptions for columns
- network security data

All Meta data is stored in system tables MSDB. META data can be accessed using repository API, DSO (Decision Support Objects).

(DB) What is multi-dimensional analysis?

Multi-dimensional is looking data from different dimensions. For example, we can look at a simple sale of a product month wise.

Month	Product	Amount
January		
	Shoes	500$
	Shirts	100$
	Caps	50$
February		
	Shoes	100$
	Shirts	600$
	Caps	50$
March		
	Shoes	900$
	Shirts	200$
	Caps	70$

Figure 8.12: - Single Dimension view.

However, let us add one more dimension "Location" wise.

	Products	Mumbai	Delhi	Bangalore	Calcutta	Total
January						
	Shoes	100$	100$	100$	200$	500$
	Shirts	-	-	-	100$	100$
	Caps	-	-	-	50$	50$
February						
	Shoes	100$	-	-	-	100$
	Shirts	-	-	-	600$	600$
	Caps	-	-	-	50$	50$
March						
	Shoes	300$	300$	300$	-	900$
	Shirts	-	-	-	200$	200$
	Caps	-	-	-	70$	70$

Figure 8.13: - Multi-Dimension View

The above table gives a three-dimension view; you can have more dimensions according to your depth of analysis. Like from the above multi-dimension view I am able to predict that "Calcutta" is the only place where "Shirts" and "Caps" are selling, other metros do not show any sales for this product.

(DB) What is MDX?

MDX stands for multi-dimensional expressions. When it comes to viewing data from multiple dimensions SQL lacks many functionalities, there is where MDX queries are useful. MDX queries are fired against OLAP databases. SQL is good for transactional databases (OLTP databases), but when it comes to analysis queries MDX stands the top.

Note: - If you are planning for data warehousing position using SQL Server 2005, MDX will be the favorite of the interviewers. MDX itself is such a huge and beautiful beast that we cannot cover in this small

book. I will suggest at least try to grab some basic syntaxes of MDX like select before going to interview.

(DB) How did you plan your Data warehouse project?

Note: - This question will come up if the interviewer wants to test that had you really worked on any data warehouse project. Second if he is looking for a project manager or team lead position.

Below are the different stages in Data warehousing project:-

- **System Requirement Gathering**

This is what every traditional project follows and data warehousing is no different. What exactly is this complete project about? What is the client expecting? Do they have existing data base which they want to data warehouse or do we have to collect from lot of places. If we have to extract from lot of different sources, what are they and how many are they?. For instance you can have customer who will say this is the database now data warehouse it. Or customer can say consolidate data from EXCEL, ORACLE, SQL Server, CSV files etc etc. So if more the disparate systems more are the complications. Requirement gathering clears all these things and gives a good road map for the project ahead.

Note: - Many data warehouse projects take requirement gathering for granted. But I am sure when customer will come up during execution with, I want that (Sales by month) and also that (consolidate data from those 20 excels) and that (prepare those extra two reports) and that (migrate that database)…. and the project goes there (programmer work over time) and then there (project goes over budget) and then (Client looses interest)…. Somewhere (software company goes under loss).

- **Selecting Tool.**

Once you are ok with requirement, its time to select which tools can do good work for you. This book only focuses on SQL Server 2005, but in reality, there are many tools for data warehousing. Probably SQL Server 2005 will sometimes not fit your project requirement and you would like to opt for something else.

- **Data Modeling and design**

This where the actual designing takes place. You do conceptual and logical designing of your database, star schema design.

- **ETL Process**

This forms the major part for any data warehouse project. Refer previous section to see what an ETL process is. ETL is the execution phase for a data warehouse project. This is the place where you will define your mappings, create DTS packages, define workflow, write scripts etc. Major issue when we do ETL process is about performance which should be considered while executing this process.

Note: - Refer “Integration Services” for how to do the ETL process using SQL Server 2005.

- **OLAP Cube Design**

This is the place where you define your CUBES, DIMENSIONS on the data warehouse database, which was loaded by the ETL process. CUBES and DIMENSIONS are done by using the requirement specification. For example, you see that customer wants a report "Sales Per month" so he can define the CUBES and DIMENSIONS which later will be absorbed by the front end for viewing it to the end user.

- **Front End Development**

Once all your CUBES and DIMENSIONS are defined, you need to present it to the user. You can build your front ends for the end user using C#, ASP.NET, VB.NET any language which has the ability to consume the CUBES and DIMENSIONS. Front end stands on top of CUBES and DIMENSION and delivers the report to the end users. With out any front end the data warehouse will be of no use form user's perspective.

- **Performance Tuning**

Many projects tend to overlook this process. However, just imagine a poor user sitting to view "Yearly Sales" for 10 minutes….frustrating no. There are three sections where you can really look why your data warehouse is performing slow:-

  - **While data is loading in database "ETL" process.**

    This is probably the major area where you can optimize your database. The best is to look in to DTS packages and see if you can make it better to optimize speed.

  - **OLAP CUBES and DIMENSIONS.**

    CUBES and DIMENSIONS are something that will be executed against the data warehouse. You can look in to the queries and see if some optimization can be done.

  - **Front-end code.**

    Front end are mostly coded by programmers and this can be a major bottleneck for optimization. So you can probably look for loops and you see if the front end is running too far away from the CUBES.]

- **User Acceptance Test (UAT)**

UAT means saying to the customer "Is this product ok with you?" Either it is a testing phase, which can be done by the customer (and mostly done by the customer), or by your own internal testing department to ensure that its matches with the customer requirement that was gathered during the requirement phase.

- **Rolling out to Production**

Once the customer has approved your UAT, its time to roll out the data warehouse in production so that customer can get the benefit of it.

- **Production Maintenance**

I know the most boring aspect from programmer's point of view, but the most profitable for an IT company point of view. In data warehousing this will mainly involve doing back ups, optimizing

the system and removing any bugs. This can also include any enhancements if the customer wants it.

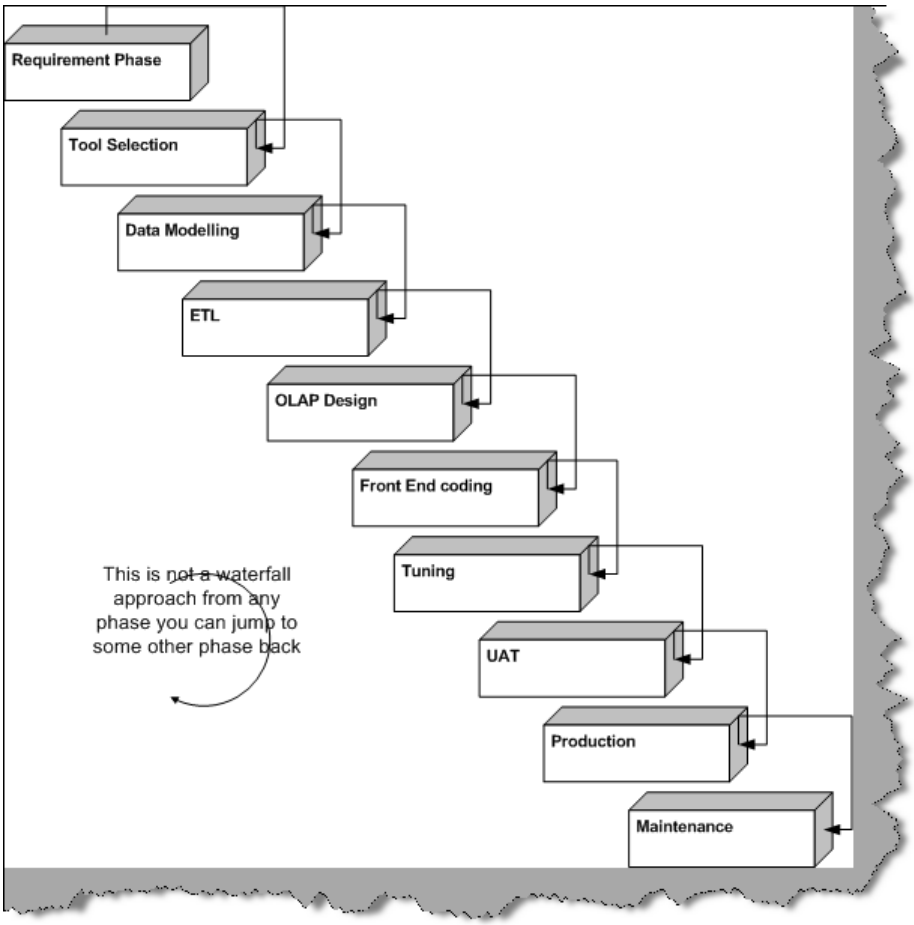


Figure 8.14: - Data ware house project life cycle

(Q) What are different deliverables according to phases?

Note: - Deliverables means what documents you will submit during each phase. For instance Source code is deliverable for execution phase, Use Case Documents or UML documents are a deliverable for requirement phase. In short what will you give to client during each phase?.

Following are the deliverables according to phases:-

- **Requirement phase: -** System Requirement documents, Project management plan, Resource allocation plan, Quality management document, Test plans and Number of reports the customer is looking at. I know many people from IT will start raising there eyeballs hey do not mix the project management with requirement gathering. But that's a debatable issue I leave it to you guys if you want to further split it.

- **Tool Selection: -** POC (proof of concept) documents comparing each tool according to project requirement.

Note: - POC means can we do?. For instance you have a requirement that, 2000 users at a time should be able to use your data warehouse. So you will probably write some sample code or read through documents to ensure that it does it.

- **Data modeling: -** Logical and Physical data model diagram. This can be ER diagrams or probably some format that the client understands.
- **ETL: -** DTS packages, Scripts and Metadata.
- **OLAP Design:-**Documents which show design of CUBES / DIMENSIONS and OLAP CUBE report.
- **Front end coding: -** Actual source code, Source code documentation and deployment documentation.
- **Tuning: -** This will be a performance-tuning document. What performance level we are looking at and how will we achieve it or what steps will be taken to do so. It can also include what areas / reports are we targeting performance improvements.
- **UAT: -** This is normally the test plan and test case document. It can be a document, which has steps how to create the test cases, and expected results.
- **Production: -** In this phase, normally the entire data warehouse project is the deliverable. But you can also have handover documents of the project, hardware, network settings, in short how is the environment setup.
- **Maintenance: -** This is an on going process and mainly has documents like error fixed, issues solved, within what time the issues should be solved and within what time it was solved.

(DB) Can you explain how analysis service works?

Note: - Ok guys this question is small but the answer is going to be massive. You are going to just summarize them but I am going to explain analysis services in detail, step by step with a small project. For this complete explanation I am taking the old sample database of Microsoft "NorthWind".

First and foremost ensure that your service is started so go to control panel, services and start the "Analysis Server "service.

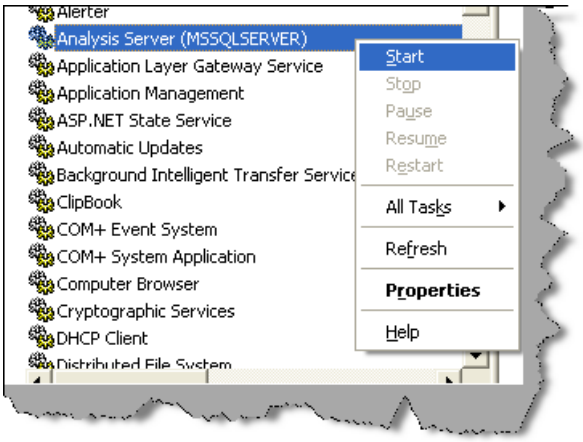


Figure 8.15: - Start Analysis Server

As said before we are going to use "North Wind" database for showing analysis server demo.

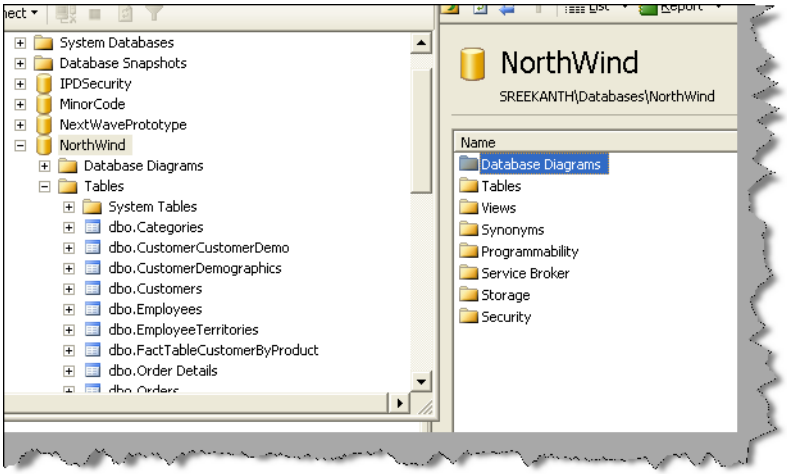


Figure 8.16: - North Wind Snapshot.

We are not going use all tables from "North Wind". Below are the only tables we will be operating using. Leaving the "FactTableCustomerByProduct" all other tables are self-explanatory. Ok I know I have still not told you what we want to derive from this whole exercise. We will try to derive a report how much products are bought by which customer and how much products are sold according to which country. So I have created the fact table with three fields Customerid, Productid and the Total Products sold. All the data in Fact table I have loaded from "Orders" and "Order Details". Means I have taken all customerid and productid with there respective totals and made entries in Fact table.

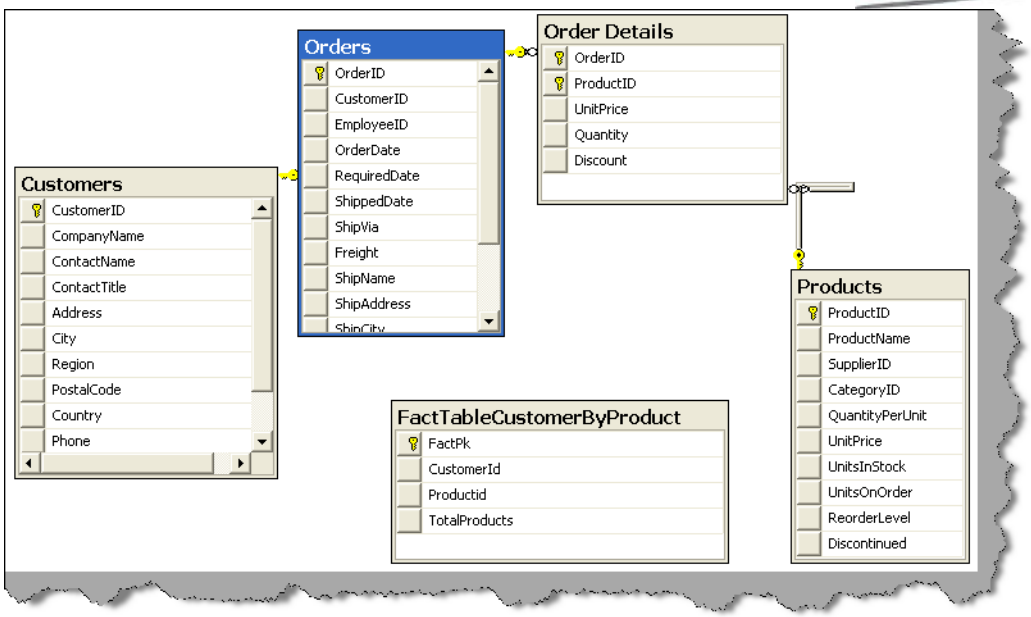


Figure 8.17: - Fact Table

Ok I have created my fact table and populated using our ETL process. Now its time to use this fact table to do analysis.

So let us start our BI studio as shown in figure below.

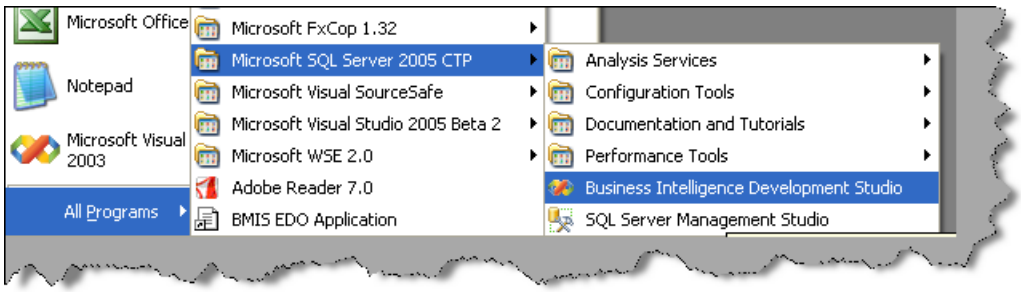


Figure 8.18: - Start the Business Development Studio

Select “Analysis” project from the project types.

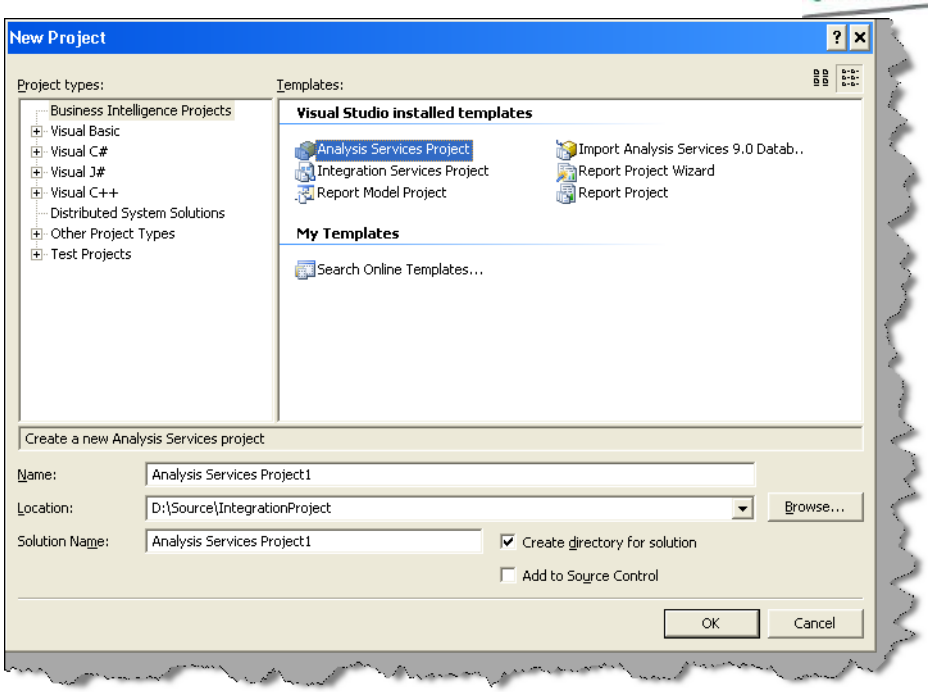


Figure 8.19: - Select Analysis Services Project

I have name the project as "Analysis Project". You can see the view of the solution explorer.

Data Sources: - This is where we will define our database and connection.

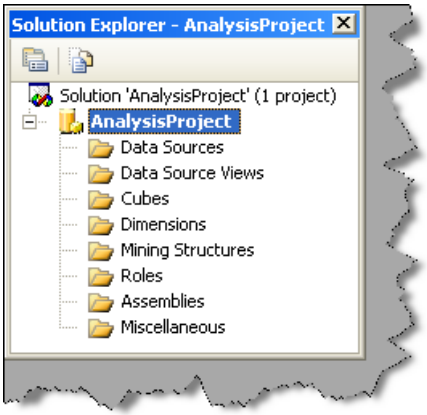


Figure 8.20 : - Solution Explorer

To add a new "data Source" right click and select "new Data Source".

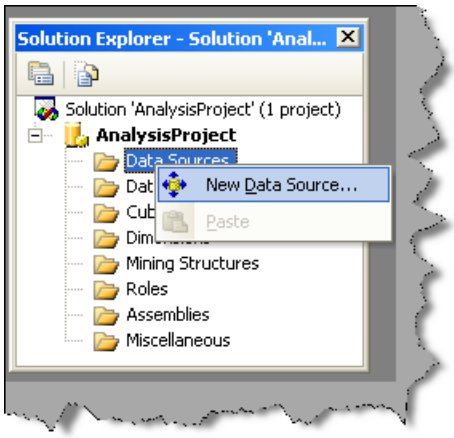


Figure 8.21: - Create new data Source

After that click next and you have to define the connection for the data source, which you can do by clicking on the new button. Click next to complete the data source process.

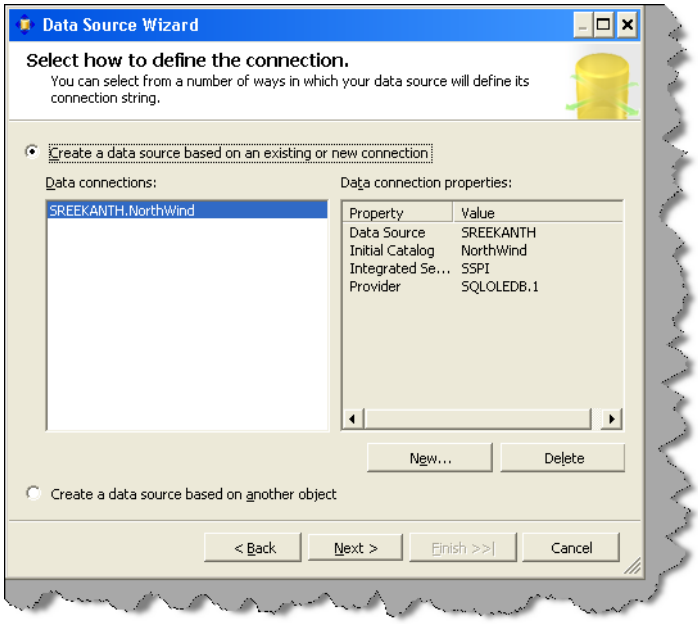


Figure 8.22: - Define Data source connection details

After that its time to define view.

Data Source View: - It is an abstraction view of data source. Data source is the complete database. It is rare that we will need the complete database at any moment of time. Therefore, in "data source view", we can define which tables we want to operate on. Analysis server never operates on data source directly but it only speaks with the "Data Source" view.

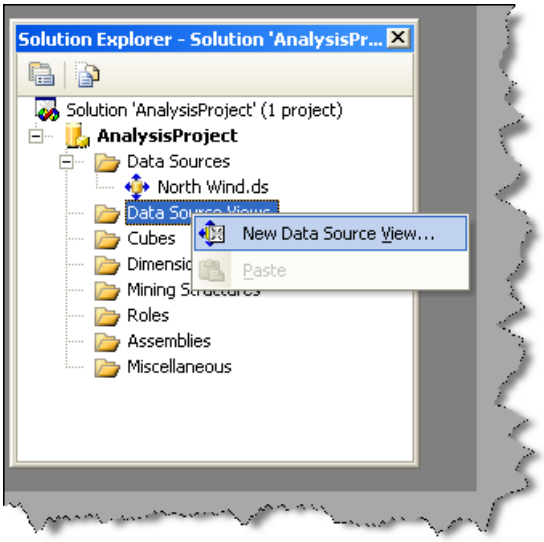


Figure 8.23: - Create new Data source view

So here, we will select only two tables “Customers”, “Products” and the fact table.

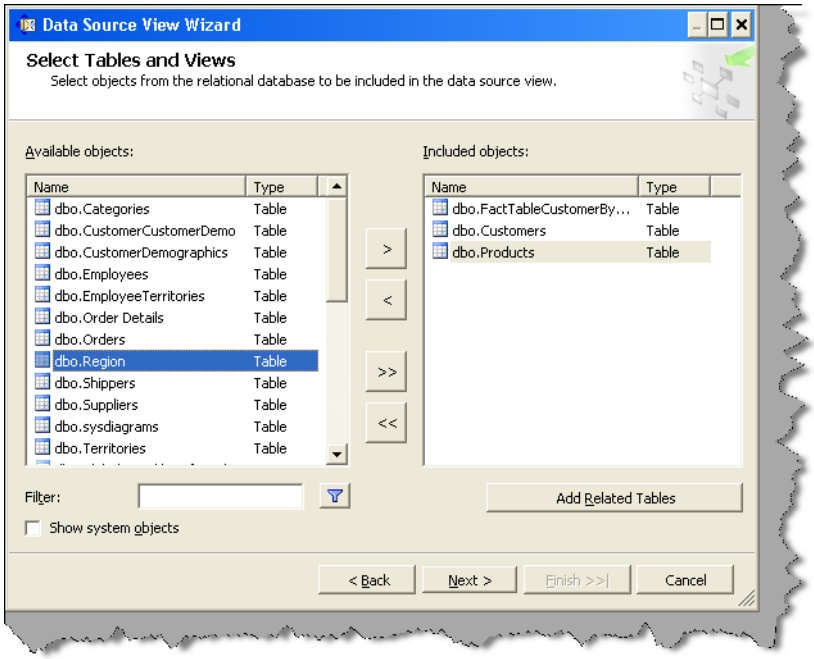


Figure 8.24: - Specify tables for the view

We had said previously fact table is a central table for dimension table. You can see products and customers table form the dimension table and fact table is the central point. Now drag and drop from the “Customerid” of fact table to the “Customerid” field of the customer table. Repeat the same for the “productid” table with the products table.

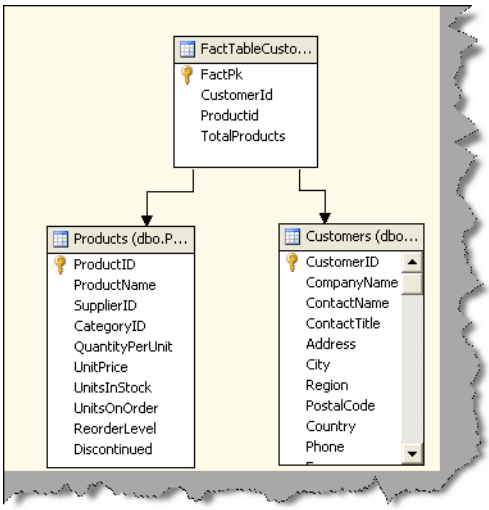


Figure 8.25: - Final Data Source view

Check “Auto build” as we are going to let the analysis service decide which tables he want to decide as “fact” and “Dimension” tables.

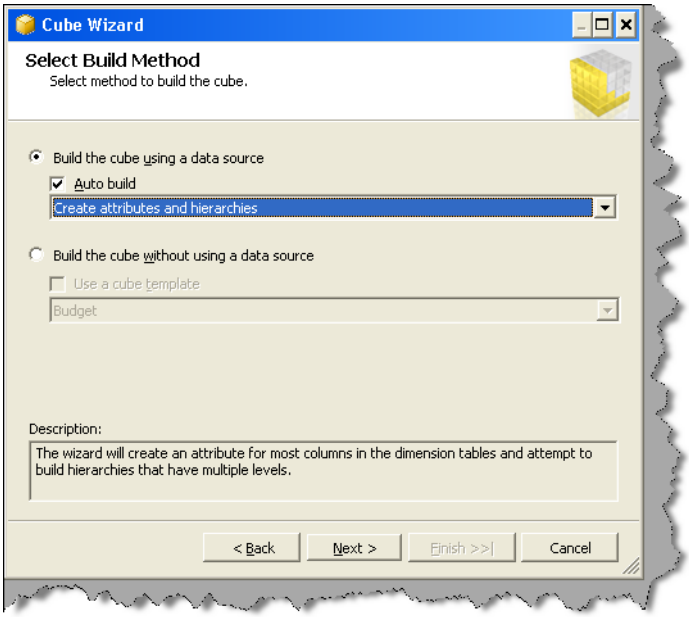


Figure 8.26: - Check Auto build

After that comes the most important step, which are the fact tables and which are dimension tables. A SQL Analysis service decides by itself, but we will change the values as shown in figure below.

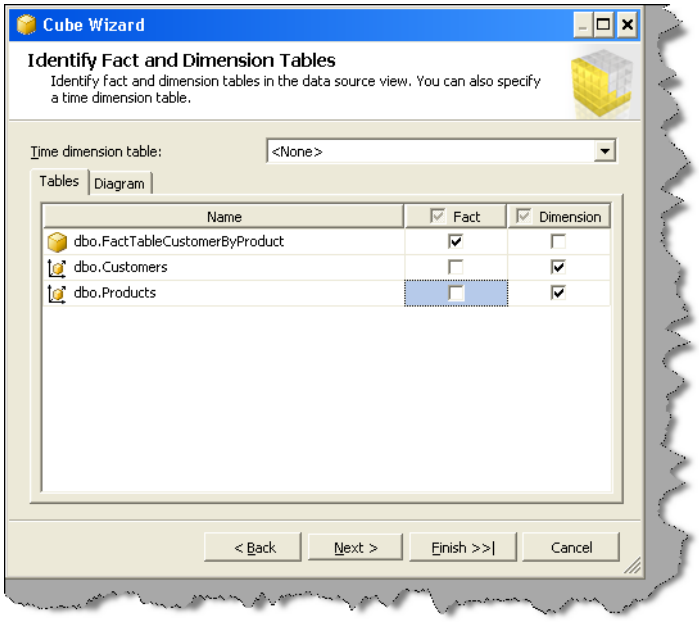


Figure 8.27: - Specify Fact and Dimension Tables

This screen defines measures.

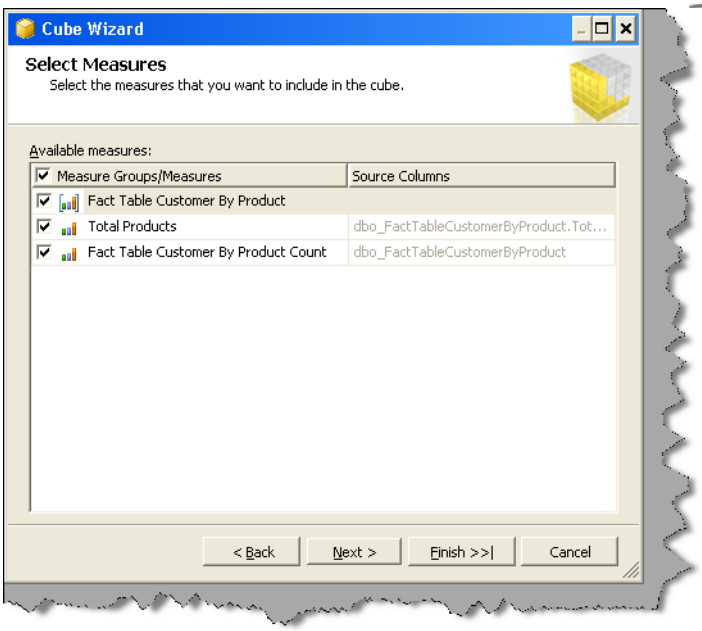


Figure 8.28: - Specify measures

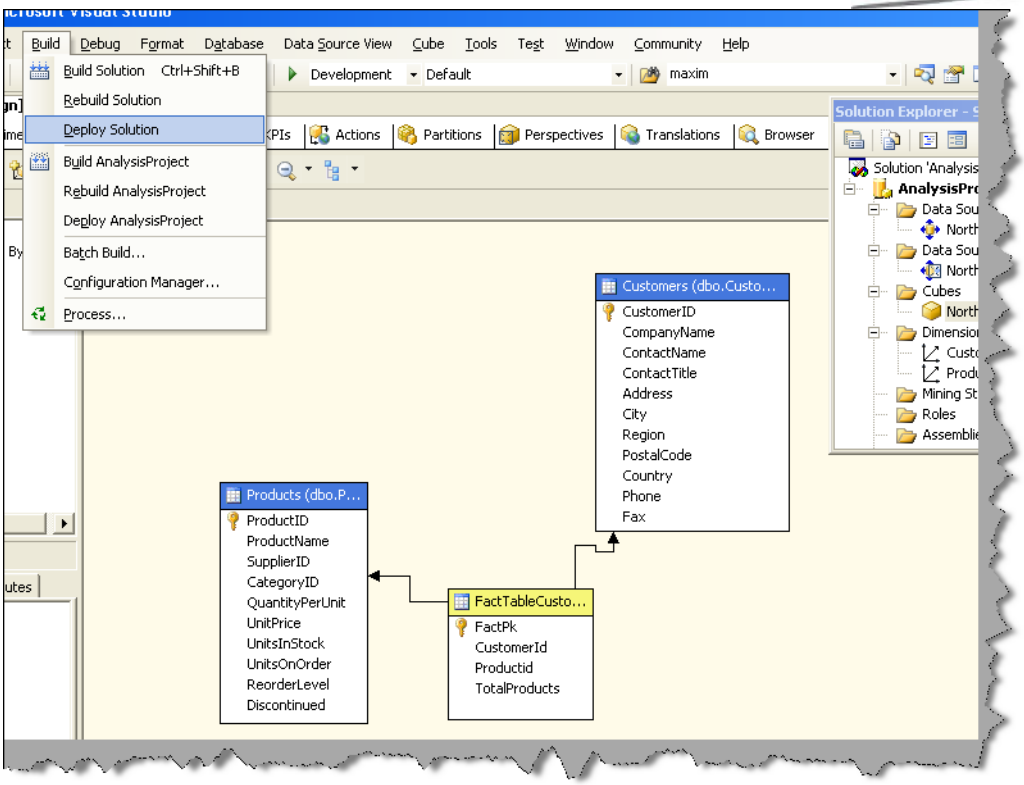


Figure 8.29: - Deploy Solution

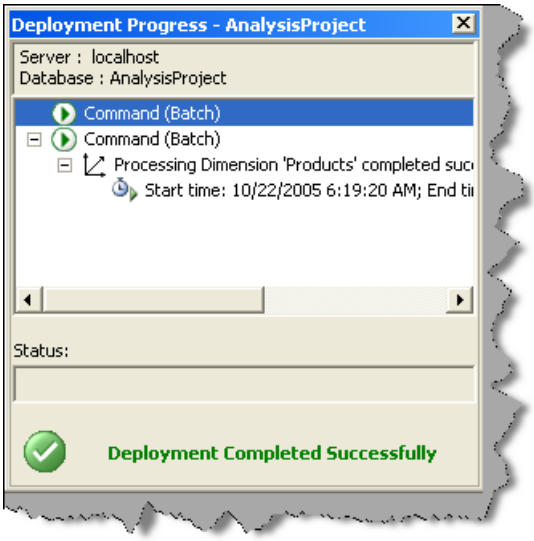


Figure 8.30: - Deployment Successful

- Cube Builder Works with the cube measures

- Dimensions Works with the cube dimensions
- Calculations Works with calculations for the cube
- KPIs Works with Key Performance Indicators for the cube
- Actions Works with cube actions
- Partitions Works with cube partitions
- Perspectives Works with views of the cube
- Translations Defines optional transitions for the cube
- Browser Enables you to browse the deployed cube

Figure 8.31: - View of top TAB

Once you are done with the complete process drag drop the fields as shown by the arrows below.

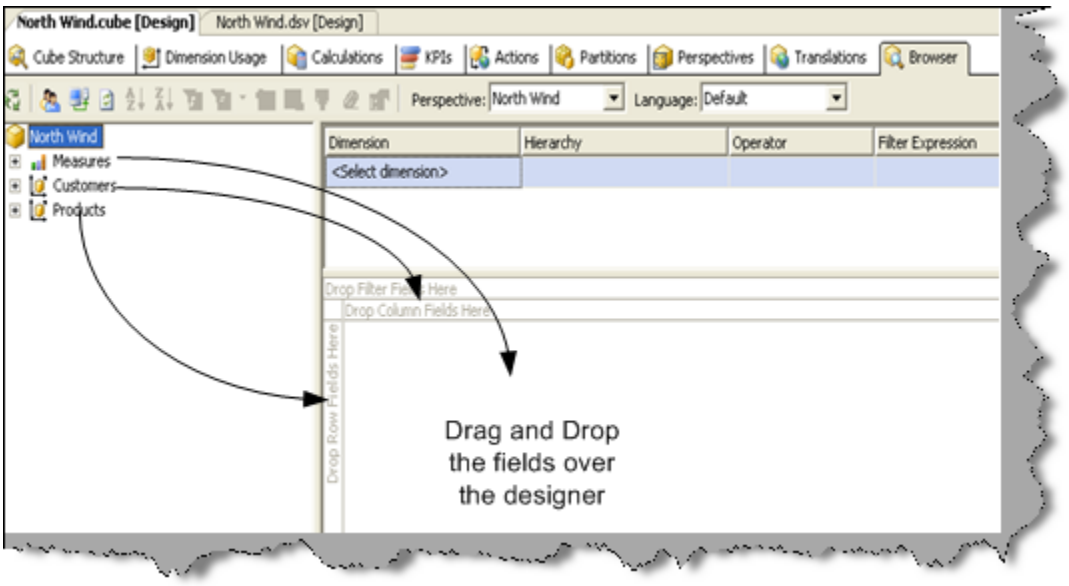


Figure 8.32: - Drag and Drop the fields over the designer

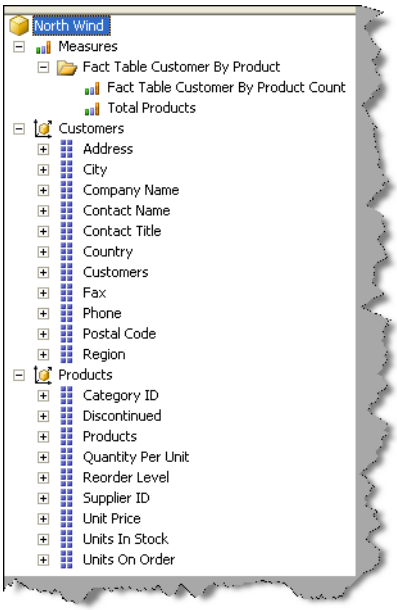


Figure 8.33: - Final look of the CUBE

Once you have dragged dropped the fields you can see the wonderful information unzipped between which customer has bought how many products.

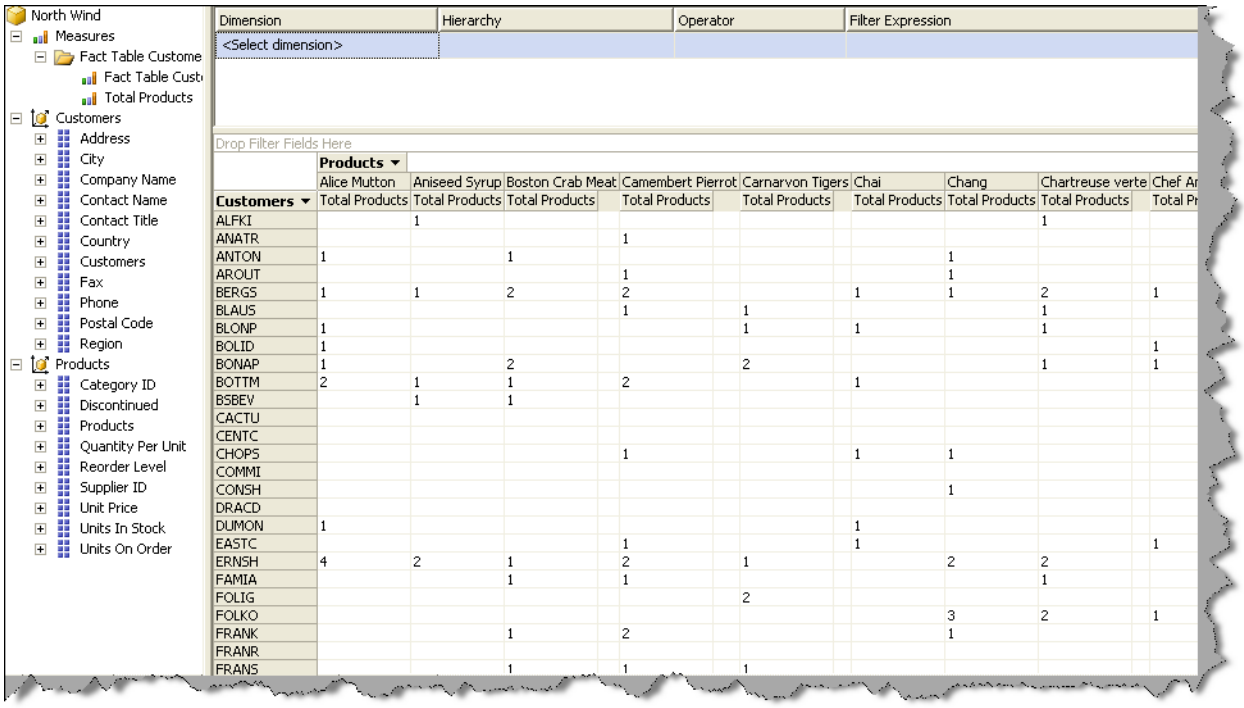


Figure 8.34: - Product and Customer Report

This is the second report that says in which country I have sold how many products.

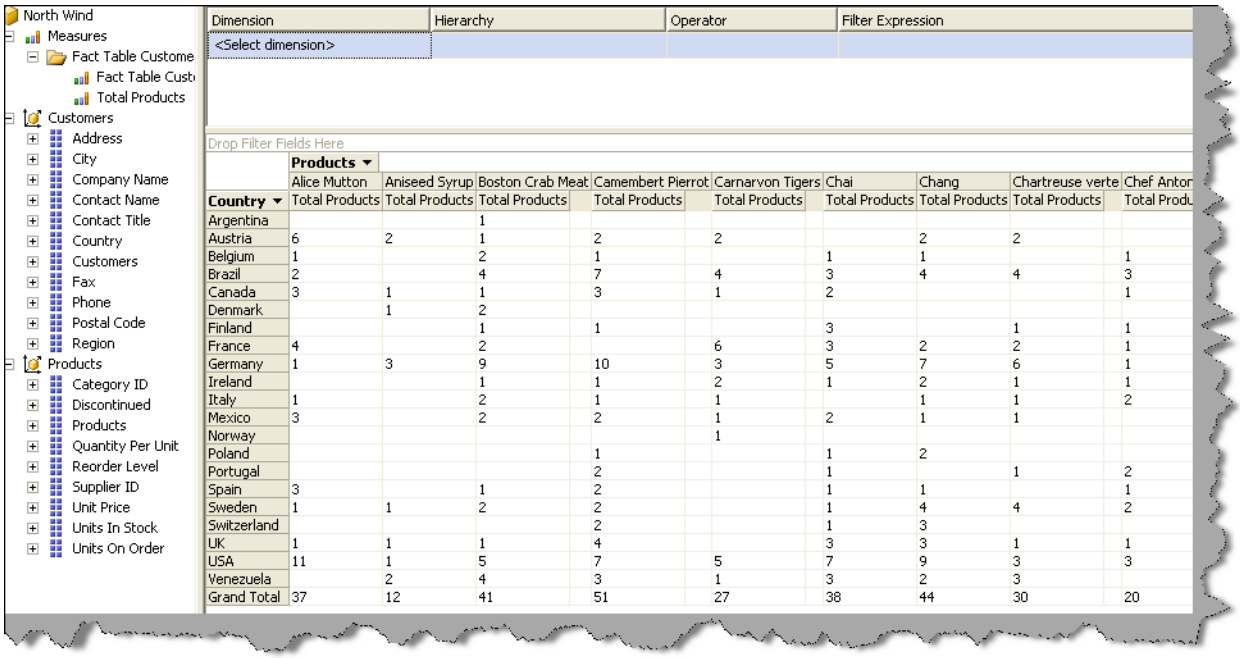


Figure 8.35: - Product Sales by country

Note: - I do not want my book to increase pages just because of images but sometimes the nature of the explanation demands it. Now you can just summarize to the interviewer from the above steps how you work with analysis services.

(Q) What are the different problems that "Data mining" can solve?

There are basically four problems that "Data mining" can solve:-

Analyzing Relationships

This term is also often called as "Link Analysis". For instance, one of the companies who sold adult products did an age survey of his customers. He found his entire products where bought by customers between age of 25 – 29. He further became suspicious that all of his customers must have kids around 2 to 5 years as that's the normal age of marriage. He analyzed further and found that maximum of his customers where married with kids. Now the company can also try selling kid products to the same customer as they will be interested in buying it, which can tremendously boost up his sales. Now here the link analysis was done between the "age" and "kids" decide a marketing strategy.

Choosing right Alternatives

If a business wants to make a decision between choices data mining can come to rescue. For example one the companies saw a major resignation wave in his company. So the HR decided to have a look at employee's joining date. They found that major of the resignations have come from employee's who have stayed in the company for more than 2 years and there where some resignation's from fresher. So the HR made decision to motivate the fresher's rather than 2 years completed employee's to retain people. As HR, thought it has easy to motivate fresher's rather than old employees.

Prediction

Prediction is more about forecasting how the business will move ahead. For instance company has sold 1000 Shoe product items, if the company puts a discount on the product sales can go up to 2000.

Improving the current process.

Past data can be analyzed to view how we can improve the business process. For instance for past two years company has been distributing product "X" using plastic bags and product "Y" using paper bags. Company has observed closely that product "Y" sold the same amount as product "X" but has huge profits. Company further analyzed that major cost of product "X" was due to packaging the product in plastic bags. Now the company can improve the process by using the paper bags and bringing down the cost and thus increasing profits.

(Q) What are different stages of "Data mining"?

Problem Definition.

This is the first step in "Data mining" define your metrics by which the model will be evaluated. For instance if it's a small travel company he would like to measure his model on number of tickets sold , but if it's a huge travel companies with lot of agents he would like to see it with number of tickets / Agent sold. If it's a different industry together like bank they would like to see actual amount of transactions done per day.

There can be several models which a company wants to look into. For instance in our previous travel company model, they would like to have the following metrics:-

- Ticket sold per day
- Number of Ticket sold per agent
- Number of ticket sold per airlines
- Number of refunds per month

So you should have the following checklist:-

- What attribute you want to measure and predict?
- What type of relationship you want to explore? In our travel company example you would like to explore relationship between number of tickets sold and Holiday patterns of a country.

Preprocessing and Transforming Data

This can also be called as loading and cleaning of data or to remove unnecessary information to simplify data. For example you will be getting data for title as "Mr.", "M.r.", "Miss", "Ms" etc ... Hmm can go worst if these data are maintained in numeric format "1", "2", "6" etc...This data needs to be cleaned for better results.

You also need to consolidate data from various sources like EXCEL, Delimited Text files; any other databases (ORACLE etc).

Microsoft SQL Server 2005 Integration Services (SSIS) contains tools, which can be used for cleaning and consolidating from various services.

Note: - Data warehousing ETL process is a subset of this section.

Exploring Models

Data mining / Explore models means calculating the min and max values, look in to any serious deviations that are happening, and how is the data distributed. Once you see the data you can look in to if the data is flawed or not. For instance normal hours in a day is 24 and you see some data has more than 24 hours, which is not logical. You can then look in to correcting the same.

Data Source View Designer in BI Development Studio contains tools, which can let you analyze data.

Building Models

Data derived from Exploring models will help us to define and create a mining model. A model typically contains input columns, an identifying column, and a predictable column. You can then define these columns in a new model by using the Data Mining Extensions (DMX) language or the Data Mining Wizard in BI Development Studio.

After you define the structure of the mining model, you process it, populating the empty structure with the patterns that describe the model. These are known as training the model. Patterns are found by passing the original data through a mathematical algorithm. SQL Server 2005 contains a different algorithm for each type of model that you can build. You can use parameters to adjust each algorithm.

A mining model is defined by a data mining structure object, a data mining model object, and a data mining algorithm.

Verification of the models.

By using viewers in Data Mining Designer in BI Development Studio, you can test / verify how well these models are performing. If you find you need any refining in the model you have to again iterate to the first step.

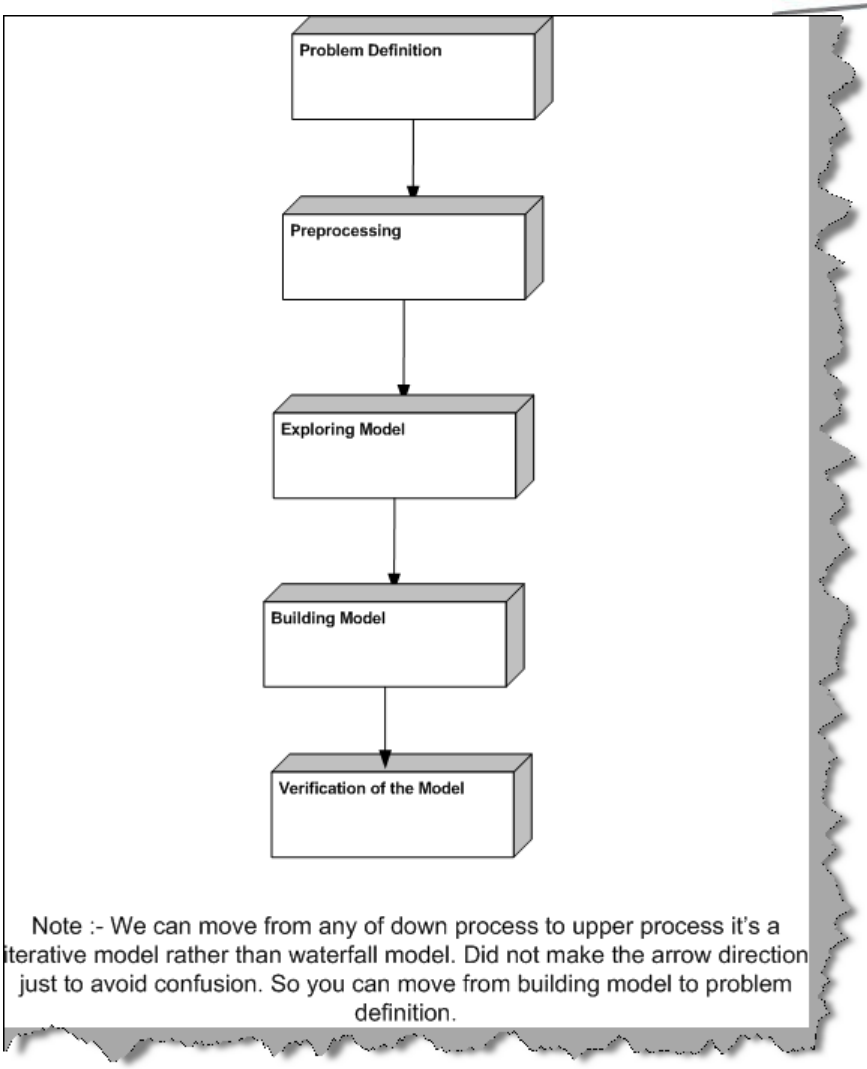


Figure 8.36: - Data mining life Cycle.

(DB) What is Discrete and Continuous data in Data mining world?

Discrete: - A data item that has a finite set of values. For example Male or Female.

Continuous: - This does not have finite set of value, but rather continuous value. For instance sales amount per month.

(DB) What is MODEL is Data mining world?

MODEL is extracting and understanding different patterns from a data. Once the patterns and trends of how data behaves are known we can derive a model from the same. Once these models are decided we can see how these models can be helpful for prediction / forecasting, analyzing trends, improving current process etc.

DB) How are models actually derived?

Twist: - What is Data Mining Algorithms?

Data mining Models are created using Data mining algorithm's. So to derive a model you apply Data mining algorithm on a set of data. Data mining algorithm then looks for specific trends and patterns and derives the model.

Note : - Now we will go through some algorithms which are used in "Data Mining" world. If you are looking out for pure "Data Mining" jobs, these basic question will be surely asked. Data mining algorithm is not Microsoft proprietary but is old math's which is been used by Microsoft SQL Server. The below section will look like we are moving away from SQL Server but trust me…if you are looking out for data mining jobs these questions can be turning point.

(DB) What is a Decision Tree Algorithm?

Note: - As we have seen in the first question that to derive a model we need algorithms. The further section will cover basic algorithms which will be asked during interviews.

"Decision Tree" is the most common method used in "data mining". In a decision tree structure, leaves determine classification and the branches represent the reason of classifications.

For instance below is a sample data collected for an ISP provider who is in supplying "Home Internet Connection".

A	B	C	D
Customer	**Age**	**Marketing Way**	**Internet Connection**
1000-2000	32-40	Direct	Did not Buy
1000-2000	18-25	Direct	Bought
2000-5000	32-40	By Phone	Did not Buy
2000-5000	18-25	By Phone	Bought
5000 and Above	32-40	By Phone	Bought
5000 and Above	18-25	By Phone	Bought

Figure 8.37: - Sample Data for Decision Tree

Based on the above data we have made the following decision tree. So you can see decision tree takes data and then start applying attribute comparison on every node recursively.

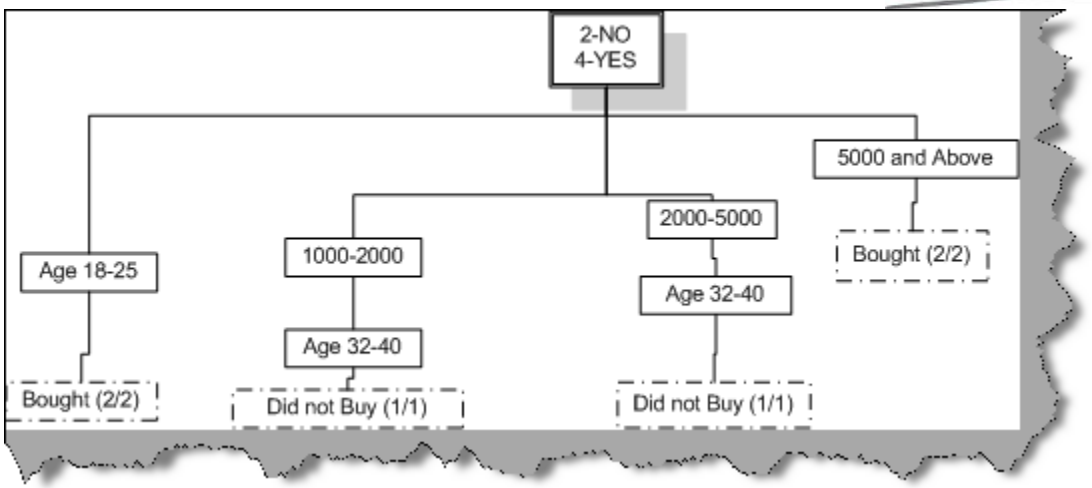


Figure 8.38: - First Iteration Decision Tree

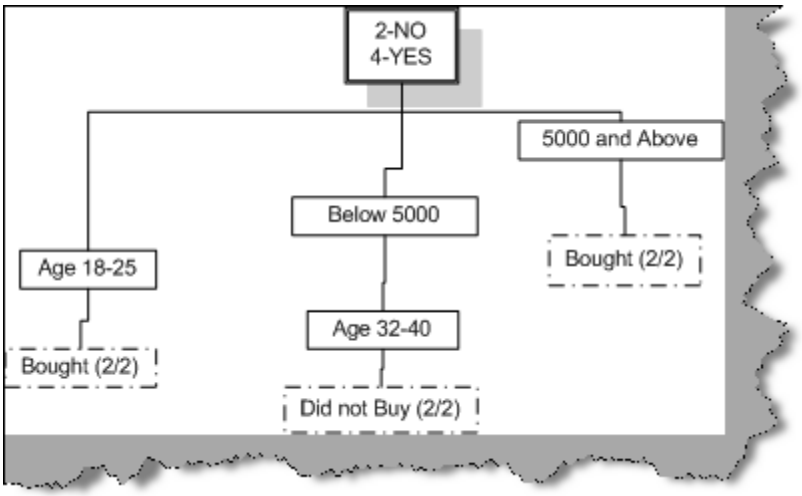


Figure 8.39: - Conclusion from the Decision Tree

From the "Decision Tree" diagram, we have concluded following predictions "-

- Age 18-25 always buys internet connection, irrelevant of income.
- Income drawers above 5000 always buy internet connection, irrelevant of age.

Using this data we have made predictions that if we market using the above criteria's we can make more "Internet Connection" sales.

Therefore, we have achieved two things from "Decision tree":-

Prediction

- If we market to age groups between 32-40 and income below 5000 we will not have decent sales.

- If we target customer with Age group 18-25 we will have good sales.
- All income drawers above 5000 will always have sales.

Classification

- Customer classification by Age.
- Customer classification depending on income amount.

(DB) Can decision tree be implemented using SQL?

With SQL, you can only look through one angle point of view. But with decision tree as you traverse recursively through all data you can have multi-dimensional view. For example give above using SQL you could have made the conclusion that age 18-25 has 100 % sales result. But "If we market to age groups between 32-40 and income below 5000 we will not have decent sales." Probably a SQL can not do (we have to be too heuristic).

(DB) What is Naïve Bayes Algorithm?

"Bayes' theorem can be used to calculate the probability that a certain event will occur or that a certain proposition is true, given that we already know a related piece of information."

Ok that is a difficult things to understand lets make it simple. Let us take for instance the sample data down.

A	B	C	D	E
Customer	**Pants**	**Shirts**	**Shoes**	**Socks**
Cust1	1	x	x	x
Cust2	x	1	x	x
Cust3	x	x	1	x
Cust4	x	x	x	1
Cust5	1	1	x	x
Cust6	1	1	x	x
Cust7	x	x	1	1
Cust8	x	x	1	1

Figure 8.40: - Bayesian Sample Data

If you look at the sample, we can say that 80 % of time customer who buy pants also buys shirts.

```
P (Shirt | Pants) = 0.8
```

Customer who buys shirts are more than who buys pants , we can say 1 of every 10 customer will only buy shirts and 1 of every 100 customer will buy only pants.

```
P (Shirts) = 0.1

P (Pants) = 0.01
```

Now suppose we a customer comes to buys pants how much is the probability he will buy a shirt and vice-versa. According to theorem:-

```
Probability of buying shirt if bought pants = 0.8-0.01 / 0.1=7.9

Probability of buying pants if bought shirts = 0.8-0.1 / 0.01=70
```

So you can see if the customer is buying shirts there is a huge probability that he will buy pants also. So you can see naïve bayes algorithm is use for predicting depending on existing data.

(DB) Explain clustering algorithm?

"Cluster is a collection of objects which have similarity between then and are dissimilar from objects different clusters."

Following are the ways a clustering technique works:-

- **Exclusive:** A member belongs to only one cluster
- **Overlapping:** A member can belong to more than one cluster
- **Probabilistic:** A member can belong to every cluster with a certain amount of probability.
- **Hierarchical**: Members are divided into hierarchies, which are sub-divided into clusters at a lower level.

(DB) Explain in detail Neural Networks?

Humans always wanted to beat god and neural networks is one of the step towards that. Neural network was introduced to mimic the sharpness of how brain works. Whenever human see something, any object for instance an animal. Many inputs are sent to his brains for example it has four legs, big horns, long tail etc etc. With these inputs, your brain concludes that it's an animal. From childhood, your brain has been trained to understand these inputs and your brain concludes output depending on that. This all happens because of those 1000 neurons, which are working inside your brain inter-connected to decide the output.

That is what human tried to devise neural network. So now, you must be thinking how it works.

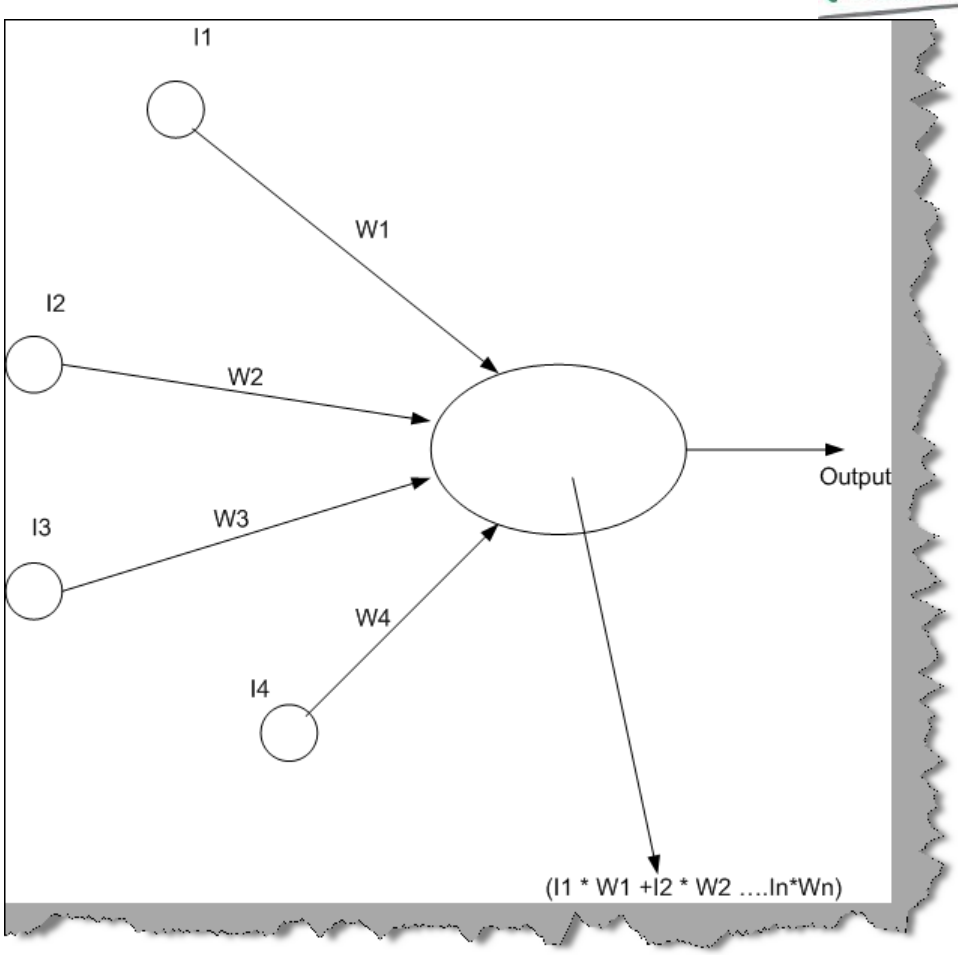


Figure 8.41: - Artificial Neuron Model

Above is the figure, which shows a neuron model. We have inputs (I1, I2 … IN) and for every input there are weights (W1, W2 …. WN) attached to it. The ellipse is the "NEURON". Weights can have negative or positive values. Activation value is the summation and multiplication of all weights and inputs coming inside the nucleus.

```
Activation Value = I1 * W1 + I2 * W2+ I3 * W3+ I4 * W4…… IN * WN
```

There is threshold value specified in the neuron, which evaluates to Boolean or some value, if the activation value exceeds the threshold value.

So probably feeding a customer sales records we can come out with an output is the sales department under profit or loss.

	Input	Weight	Input * Weight
Description	Number of Customer	Sales Amount per customer	NetSales
London	12	200	2400
India	10	100	1000
Germany	13	150	1950
Greece	5	40	200
		Total Sales figure	5550

Figure 8.42: - Neural Network Data

For instance, take the case of the top customer sales data. Below is the neural network defined for the above data.

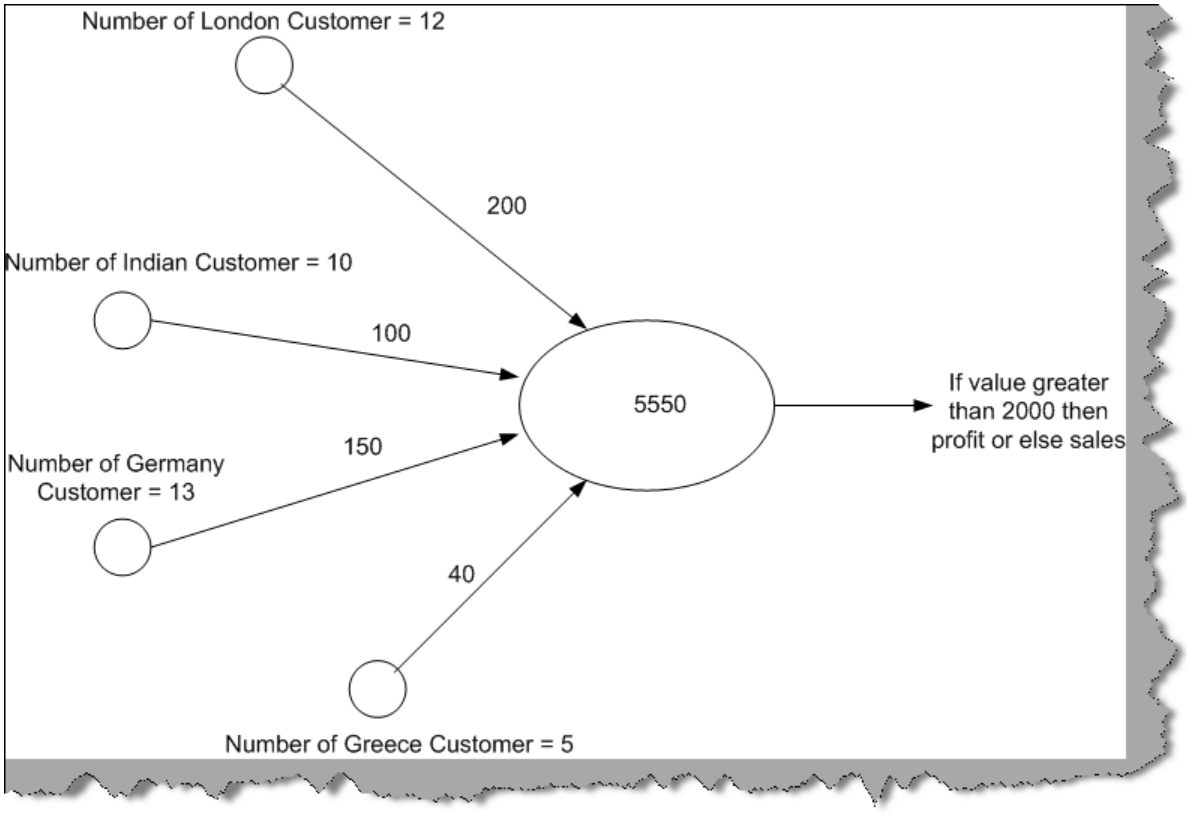


Figure 8.43: - Neural Network for Customer Sales Data

You can see neuron has calculated the total as 5550 and as it is greater than threshold 2000 we can say the company is under profit.

The above example was explained for simplification point of view. However, in actual situation there can many neurons as shown in figure below. It's a complete hidden layer from the data miner perspective. He only looks in to inputs and outputs for that scenario.

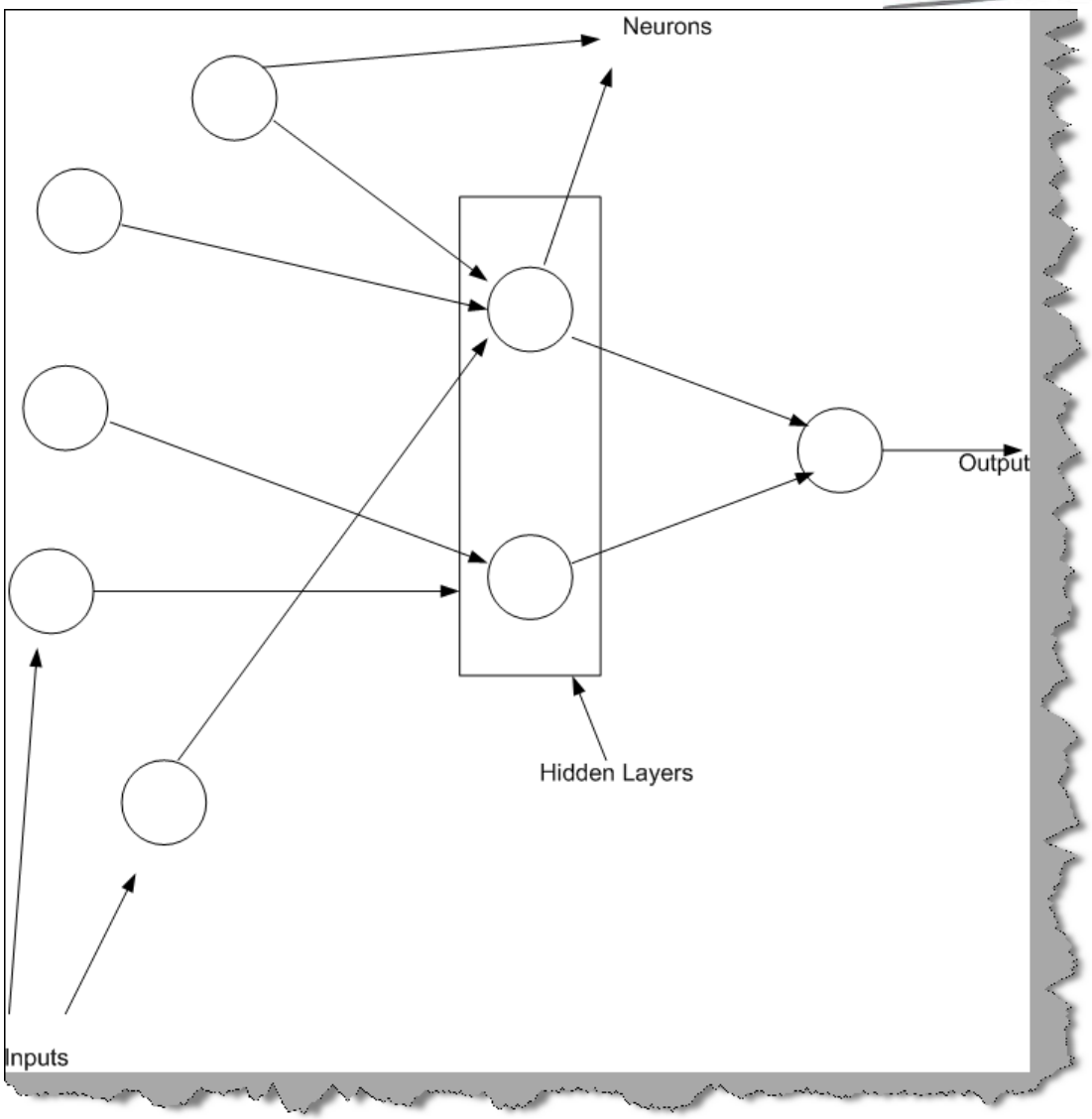


Figure 8.44: - Practical Neural Network

(DB) What is Back propagation in Neural Networks?

Back propagation helps you minimize error and optimize your network. For instance in our top example we get neuron summation as 80000000000, which is a weird figure (as you are expecting values between 0 to 6000 maximum). So you can always go back and look at whether you have some wrong input or weights. So the error is again Fed back to the neural network and the weights are adjusted accordingly. This is also called training the model.

(DB) What is Time Series algorithm in data mining?

The Microsoft Time Series algorithm allows you to analyze and forecast any time-based data, such as sales or inventory. So the data should be continuous and you should have some past data on which it can predict values.

(DB) Explain Association algorithm in Data mining?

Association algorithm tries to find relation ship between specific categories of data. In Association first it scans for unique values and then the frequency of values in each transaction is determined. For instance if lets say we have city master and transactional customer sales table. Association algorithm first find unique instance of all cities and then see how many city occurrences have occurred in the customer sales transactional table.

(DB) What is Sequence clustering algorithm?

Sequence clustering algorithm analyzes data that contains discrete-valued series. It looks for how the past data is transitioning and then makes future predictions. It's a hybrid of clustering and sequencing algorithm

Note: - UUUh I understand algorithm are dreaded level question and will never be asked for programmer level job, but guys looking for Data mining jobs these questions are basic. It's difficult to cover all algorithms existing in data mining world, as its complete area by itself. As been an interview question book I have covered algorithm which are absolutely essential from SQL Server point of view. Now we know the algorithms we can classify where they can be used. There are two important classifications in data mining world Prediction / Forecasting and grouping. So we will classify all algorithms which are shipped in SQL server in these two sections only.

(DB) What are algorithms provided by Microsoft in SQL Server?

Predicting an attribute, for instance how much will be the product sales next year.

- Microsoft Decision Trees Algorithm
- Microsoft Naive Bayes Algorithm
- Microsoft Clustering Algorithm
- Microsoft Neural Network Algorithm

Predicting a continuous attribute, for example, to forecast next year's sales.

- Microsoft Decision Trees Algorithm
- Microsoft Time Series Algorithm

Predicting a sequence, for example, to perform a click stream analysis of a company's Web site.

- Microsoft Sequence Clustering Algorithm

Finding groups of common items in transactions, for example, to use market basket analysis to suggest additional products to a customer for purchase.

- Microsoft Association Algorithm

- Microsoft Decision Trees Algorithm

Finding groups of similar items, for example, to segment demographic data into groups to better understand the relationships between attributes.

- Microsoft Clustering Algorithm
- Microsoft Sequence Clustering Algorithm

Why we went through all these concepts is when you create data mining model you have to specify one the algorithms. Below is the snapshot of all SQL Server existing algorithms.

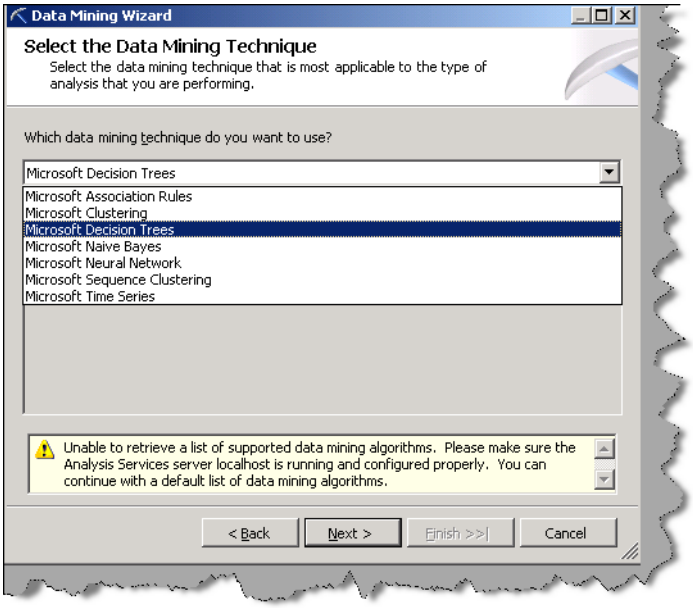


Figure 8.45: - Snapshot of the algorithms in SQL Server

Note: - During interviewing it's mostly the theory that counts and the way you present. For datamining I am not showing any thing practical as such probably will try to cover this thing in my second edition. But it's a advice please do try to run make a small project and see how these techniques are actually used.

(DB) How does data mining and data warehousing work together?

Twist: - What is the difference between data warehousing and data mining?

This question will be normally asked to get an insight how well you know the whole process of data mining and data warehousing. Many new developers tend to confuse data mining with warehousing (especially fresher's). Below is the big picture which shows the relation between "data warehousing" and "data mining".

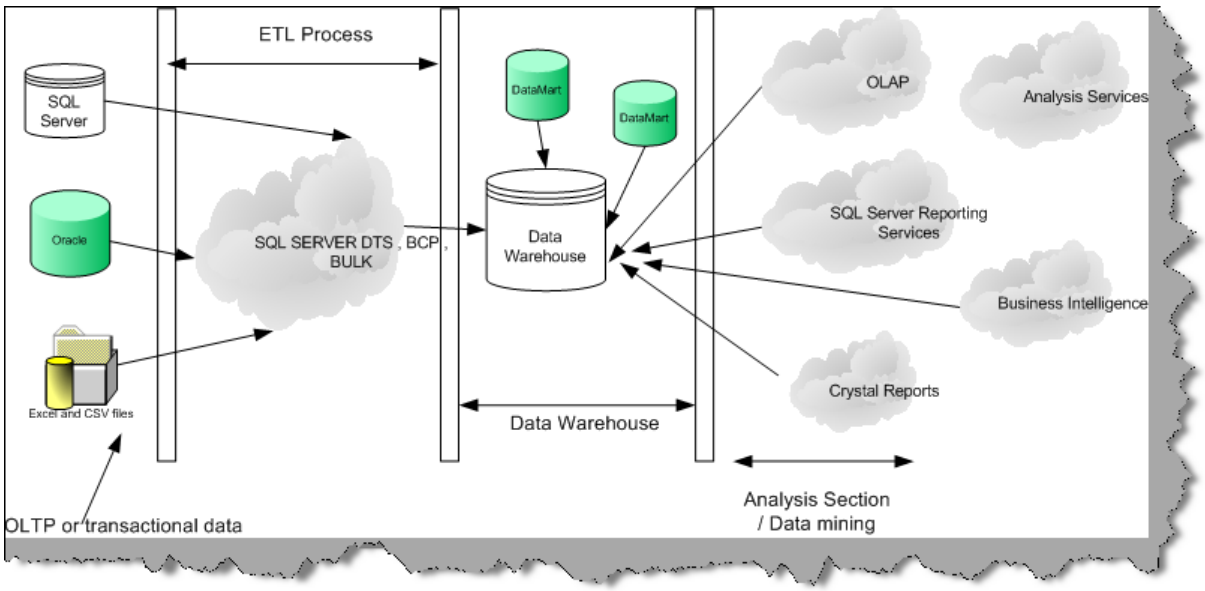


Figure 8.46: - Data mining and Data Warehousing

Let us start from the most left hand side of the image. First section comes is the transaction database. This is the database in which you collect data. Next process is the ETL process. This section extracts data from the transactional database and sends to your data warehouse, which is designed using STAR or SNOW FLAKE model. Finally, when your data warehouse data is loaded in data warehouse, you can use SQL Server tools like OLAP, Analysis Services, BI, Crystal reports or reporting services to finally deliver the data to the end user.

Note: - Interviewer will always try goof you up saying why should not we run OLAP, Analysis Services, BI, Crystal reports or reporting services directly on the transactional data. That is because transactional database are in complete normalized form which can make the data mining process complete slow. By doing data warehousing we denormalize the data which makes the data mining process more efficient.

(Q) What is XMLA?

XML for Analysis (XMLA) is fundamentally based on web services and SOAP. Microsoft SQL Server 2005 Analysis Services uses XMLA to handle all client application communications to Analysis Services.

XML for Analysis (XMLA) is a Simple Object Access Protocol (SOAP)-based XML protocol, designed specifically for universal data access to any standard multidimensional data source residing on the Web. XMLA also eliminates the need to deploy a client component that exposes Component Object Model (COM) or Microsoft .NET Framework.

(Q) What is Discover and Execute in XMLA?

The XML for Analysis open standard describes two generally accessible methods: Discover and Execute. These methods use the loosely-coupled client and server architecture supported by XML to handle incoming and outgoing information on an instance of SSAS.

The Discover method obtains information and metadata from a Web service. This information can include a list of available data sources, as well as information about any of the data source providers. Properties define and shape the data that is obtained from a data source. The Discover method is a common method for defining the many types of information a client application may require from data sources on Analysis Services instances. The properties and the generic interface provide extensibility without requiring you to rewrite existing functions in a client application.

The Execute method allows applications to run provider-specific commands against XML for Analysis data sources.

Chapter 9: Integration Services / DTS

Note: - We had seen some question on DTS in the previous chapter "Data Warehousing". But in order to just make complete justice with this topic I have included them in integration services.

(Q) What is Integration Services import / export wizard?

Note :- What is DTS import / Export Wizard ?

Note: - Try to do this practically as it can be useful if the interviewer wants to visualize the whole stuff.

DTS import / export wizard lets us import and export from external data sources. There are seven steps, which you can just go through of how to use the wizard.

You can find DTS import and export wizard as shown below.

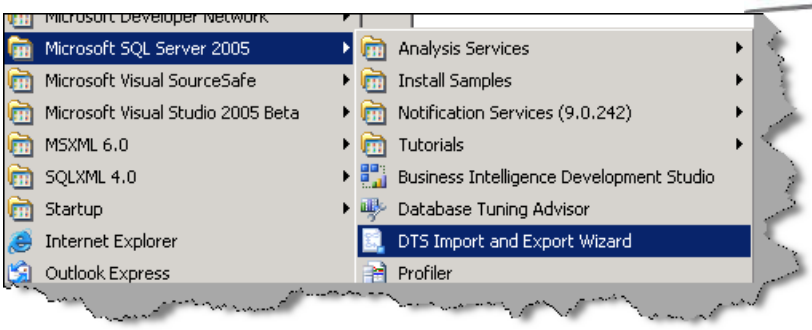


Figure 9.1: - Location of DTS Import and Export Wizard

You will be popped with a screen as below click "Next"

Figure 9.2 : - Import / Export Wizard

Next step is to specify from which source you want to copy data. You have to specify the Data source name and server name. For understanding purpose we are going to move data between "AdventureWork" databases. I have created a dummy table called as "SalesPersonDummy" which has the same structure as that of "SalesPerson" table. But the only difference is that "SalesPersonDummy" does not have data.

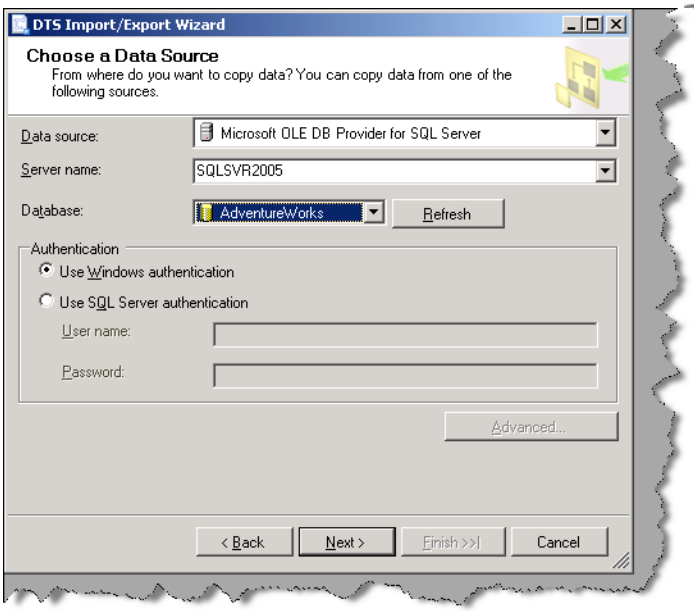


Figure 9.3: - Specify the Data Source.

Next step is to specify the destination where the source will be moved. At this moment, we are moving data inside "AdventureWorks" itself so specify the same database as the source.

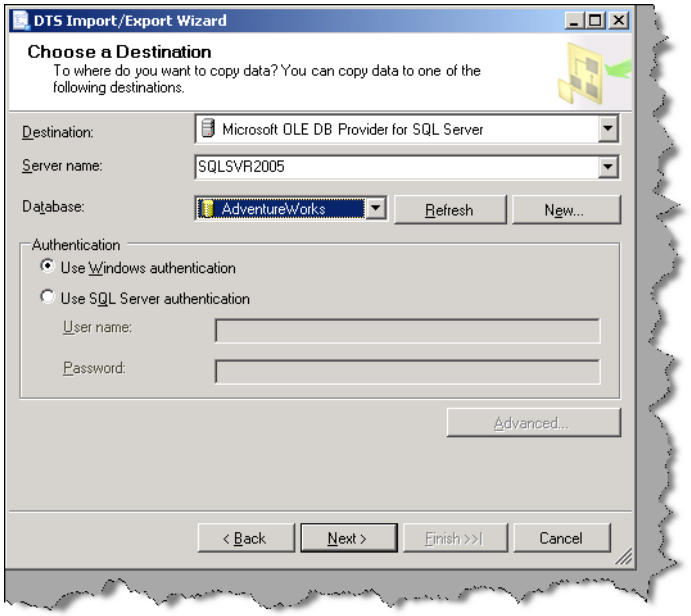


Figure 9.4: - Specify Destination for DTS

Next step is to specify option from where you want to copy data. For the time being we going to copy from table, so selected the first option.

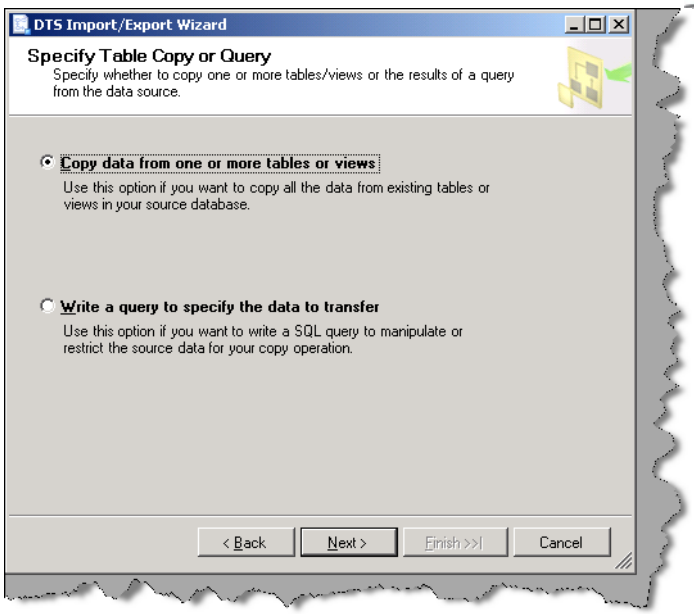


Figure 9.5: - Specify option

Finally choose which object you want to map where. You can map multiple objects if you want.

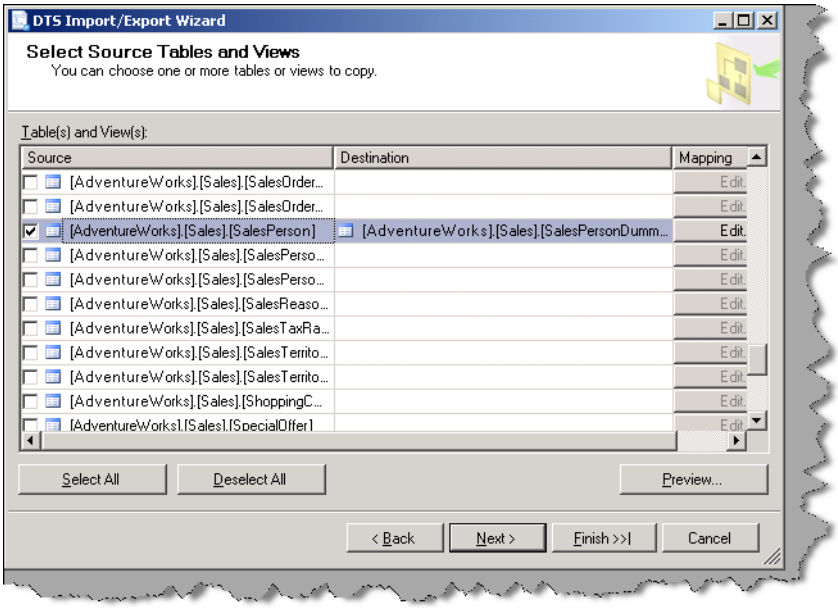


Figure 9.6: - "Salesperson" is mapped to "SalesPersonDummy"

When everything goes successful you can see the below screen, which shows the series of steps DTS has gone through.

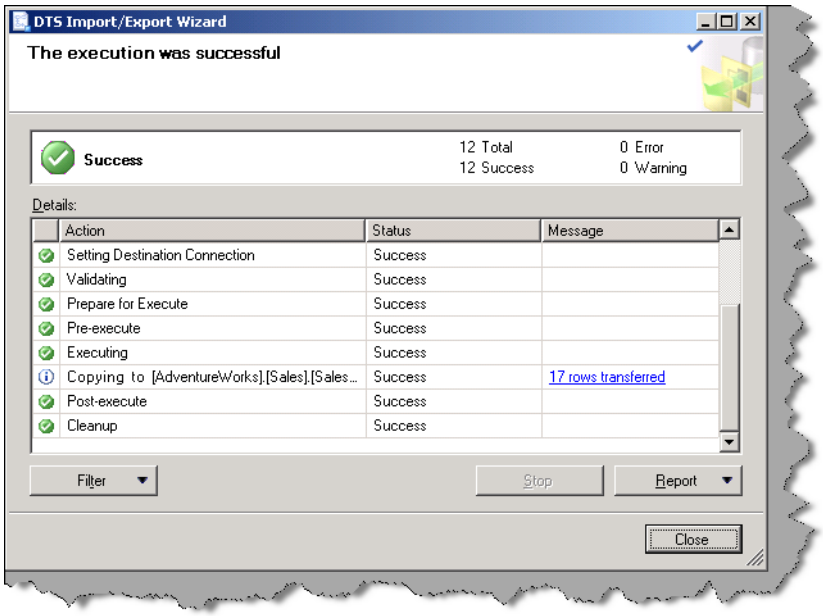


Figure 9.7: - Successful execution after series of checks

(Q) What are prime components in Integration Services?

There are two important components in Integration services:-

- **DTP (Data transformation pipeline)**

DTP is a bridge, which connects the source (CSV, any other Database etc) and the destination (SQL Server Database). While DTP moves data between source and destination, transformation takes place between input and output columns. Probably some column will go as one to one mapping and some with some manipulations.

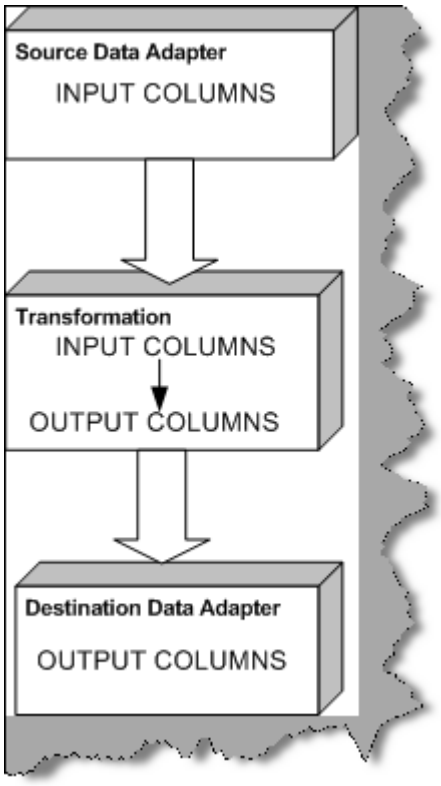


Figure 9.8: - Data Transformation Pipeline

- **DTR (Data transformation runtime)**

While DTP acts as bridge DTR controls you integration service. They are more about how will be the workflow and different components during transformation. Below are different components associated with DTR:-

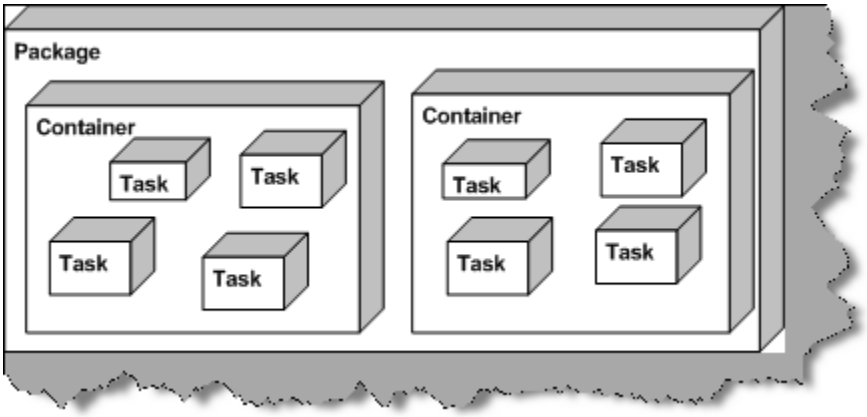


Figure 9.9 : - Data Transformation Runtime

- **Task: -** It is the smallest unit, which you want to execute.

- **Container: -** Container logically groups task. For instance you have a task to load CSV file in to database. So you will have two or three task probably :-
    - Parse the CSV file.
    - Check for field data type
    - Map the source field to the destination.

So you can define all the above work as task and group them logically in to a container called as Container.

- **Package: -** Package are executed to actually do the data transfer.

DTP and DTR model expose API, which can be used in .NET language for better control.

Note: -I can hear the shout practical.. Practical. I think I have confused you guys over there. So let us warm up on some practical DTS stuff. 1000 words is equal to one compiled program – Shivprasad Koirala? I really want to invent some proverbs if you do not mind it.

(Q) How can we develop a DTS project in Integration Services?

Twist: - Can you say how have you implemented DTS in your project and for what?

Note: - We had visited DTS import / export wizard in previous section of this chapter. But for a real data transformation or a data warehousing (ETL process) it's not enough. You will need to customize the project, there's where we can use this beautiful thing called as "BI development project". If possible just try to go step by step in creating this sample project.

You can get the development studio as shown below.

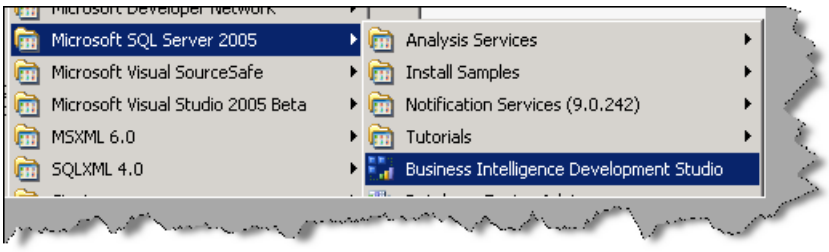


Figure 9.10: - Location of BI development studio

Click File—New – Project and select "Data Transformation Project".

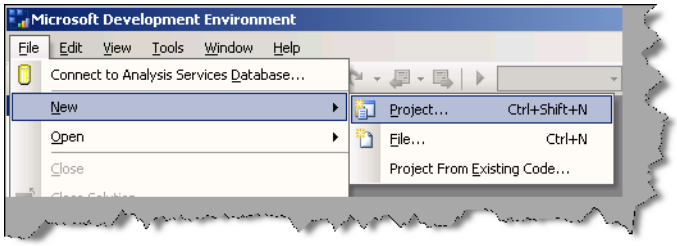


Figure 9.11: - New Project DTS

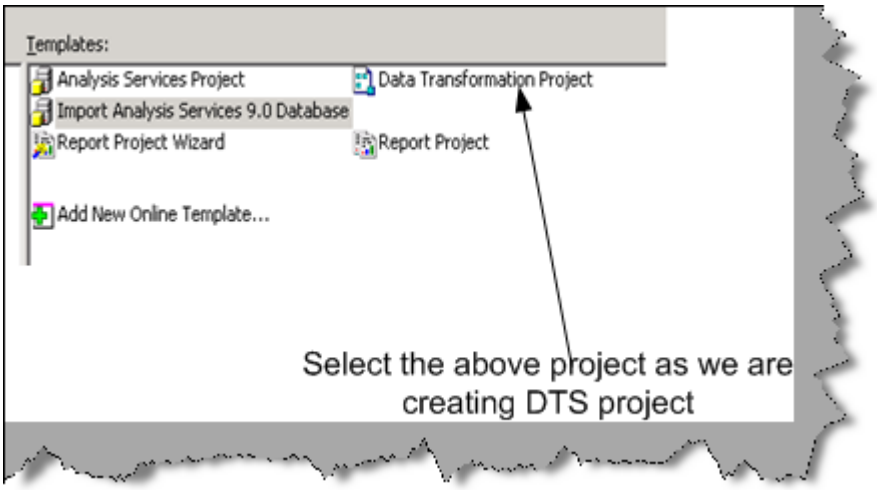


Figure 9.12: - Dialog DTS project

Give name to the project as "Salesperson" project. Before moving ahead let me give a brief about what we are trying to do. We are going to use "Sales.SalesPerson" table from the "adventureworks" database. "Sales.Salesperson" table has field called as "Bonus". We have the following task to be accomplished:-

Note: - These both tables have to be created manually by you. I will suggest to use the create statements and just make both tables. You can see in the image below there are two tables "SalesPerson5000" and "SalesPersonNot5000".

- Whenever "Bonus" field is equal to 5000 it should go in"Sales.Salesperson5000".
- Whenever "Bonus" field is not equal to 5000, it should go in"Sales.SalespersonNot5000".

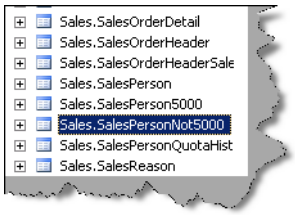


Figure 9.13 : - Snapshot of my database with both tables

One you selected the "Data transformation project" , you will be popped with a designer explorer as show below. I understand you must be saying its cryptic…it is. But let's try to simplify it. At the right hand, you can see the designer pane, which has lot of objects on it. At right hand side you can see four tabs (Control flow, Data Flow, Event handlers and Package Explorer).

Control flow: - It defines how the whole process will flow. For example if you loading a CSV file. Probably you will have task like parsing, cleaning and then loading. You can see lot of control flow items, which can make your data mining task easy. First we have to define a task in which we will define all our data flows. Therefore, you can see the curve arrow, which defines what you have to drag and drop on the control flow designer. You can see the arrow tip, which defines the output point from the task.

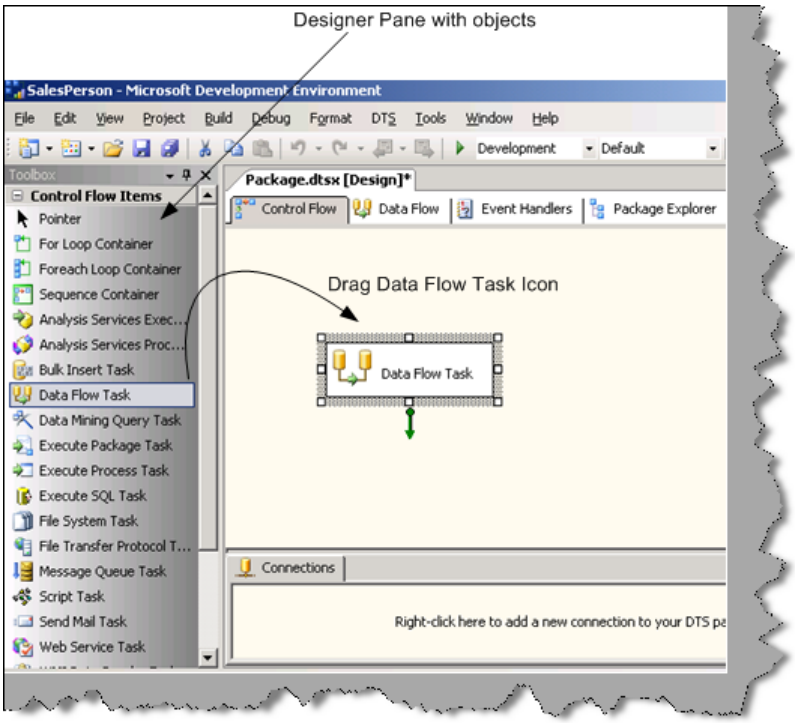


Figure 9.14: - Data Flow Task

In this project I have only define one task, but in real time project something below like this can be seen (Extraction, Transformation and Loading: - ETL). One task points as a input to other task and the final task inputs data in SQL Server.

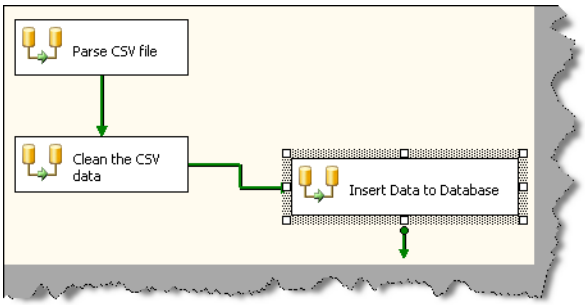


Figure 9.15: - Multiple Task CSV

Data Flow: - Data flow say how the objects will flow inside a task. So Data flow is subset of a task defining the actual operations.

Event Handlers: - The best of part of DTS is that we can handle events. For instance if there is an error what action do you want it to do. Probably log your errors in error log table, flat file or be more interactive send a mail.

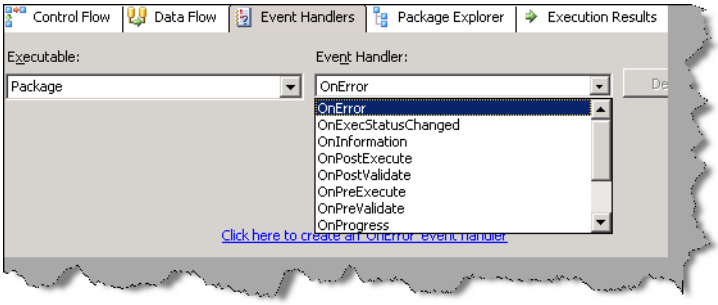


Figure 9.16: - Event Handlers

Package Explorer: - It shows all objects in a DTS in hierarchical way.

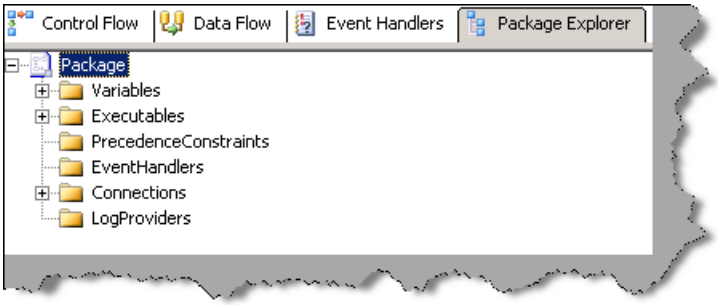


Figure 9.17:- Package Explorer

Now that you have defined your task its time to define the actual operation that will happen with in the task. We have to move data from "Sales.SalesPerson" to "Sales.SalesPerson5000" (if their "Bonus" fields are equal to 5000) and "Sales.SalesPersonNot5000" (if their "Bonus" fields are not equal to 5000). In short, we have "Sales.SalesPerson" as the source and other two tables as

Destination. So click on the "Data Flow" tab and drag the OLEDB Source data flow item on the designer, we will define source in this item. You can see that there is some error which is shown by a cross on the icon. This signifies that you need to specify the source table that is "Sales.Salesperson".

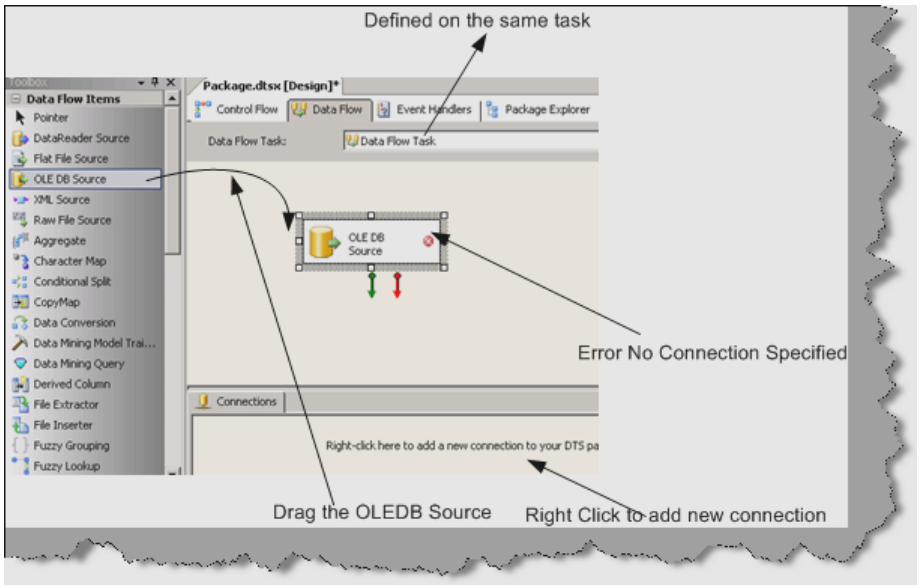


Figure 9.18: - Adding OLEDB Source

In order to specify source tables we need to specify connections for the OLEDB source. So right click on the below tab "Connections" and select "New OLEDB Connection". You will be popped up with a screen as show below. Fill in all details, specify the database as "AdventureWorks", and click "OK".

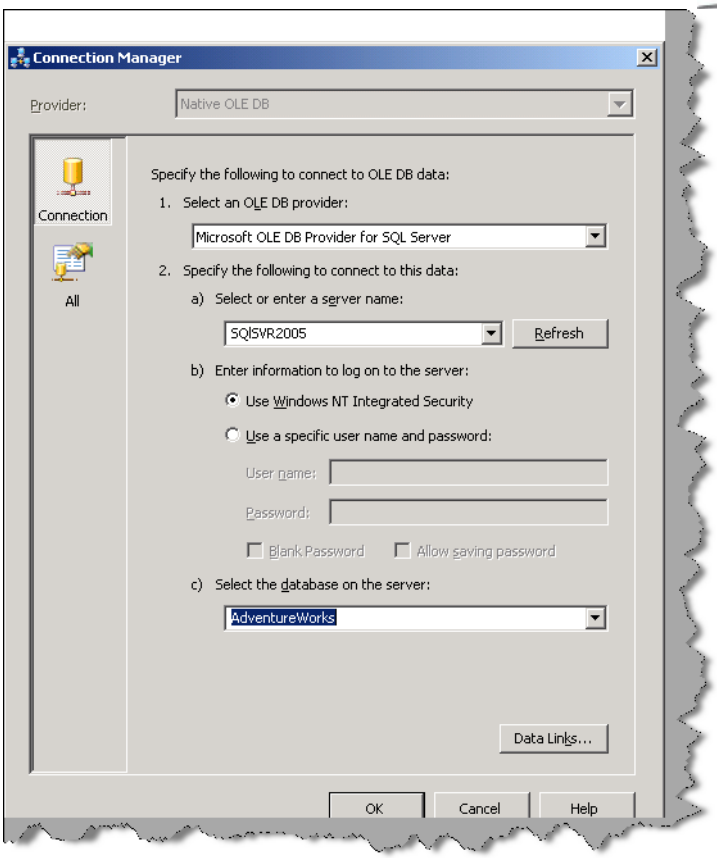


Figure 9.19: - Connection Manager

If the connection credentials are proper you can see the connection in the "Connections" tab as shown in below figure.

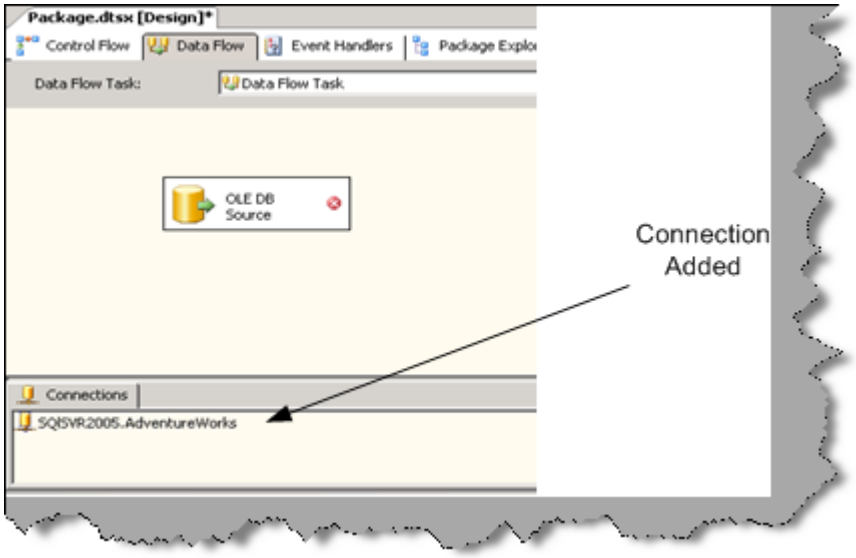


Figure 9.20: - Connection Added Successfully

Now that we have defined the connection, we have to associate that connection with the OLE DB source. So right click and select the “Edit” menu.

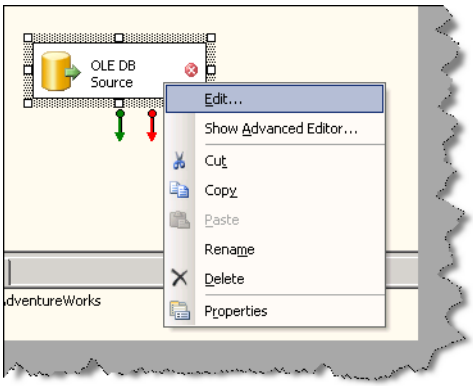


Figure 9.21: - Edit OleDB

Once you click edit, you will see a dialog box as shown below. In data access mode select “Table or View” and select the “Sales.Salesperson” table. To specify the mapping click on “Columns” tab and then press ok.

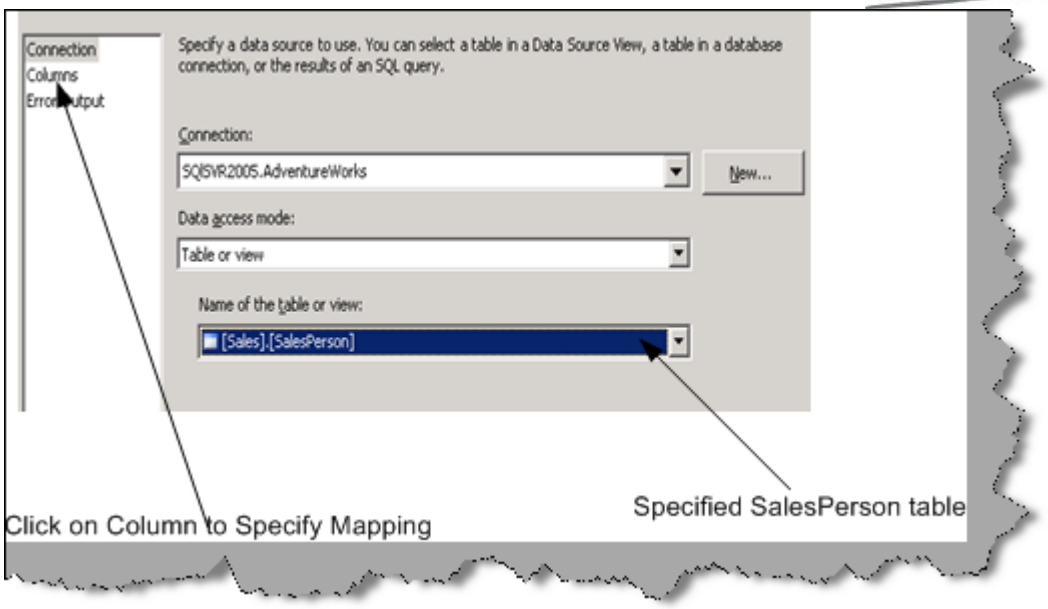


Figure 9.22: - Specify Connection Values

If the credentials are ok you can see the red Cross is gone and the OLE DB source is not ready to connect further. As said before we need to move data to appropriate tables on condition that "Bonus" field value. So from the data flow item drag and drop the "Conditional Split" data flow item.

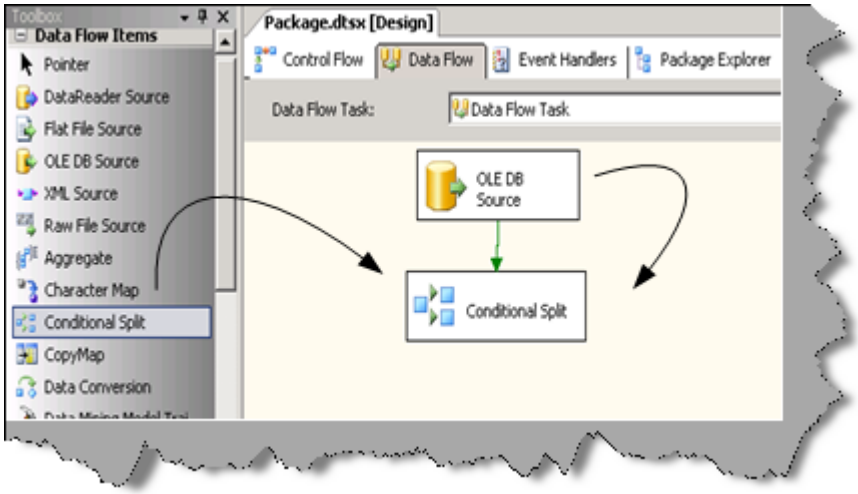


Figure 9.23: - Conditional Split

Right click on the "Conditional Split" data flow item so that you can specify the criteria. It also gives you a list of fields in the table which you can drag drop. You can also dragdrop the operators and specify the criteria. I have made two outputs from the conditional split one, which is equal to 5000 and second, not equal to 5000.

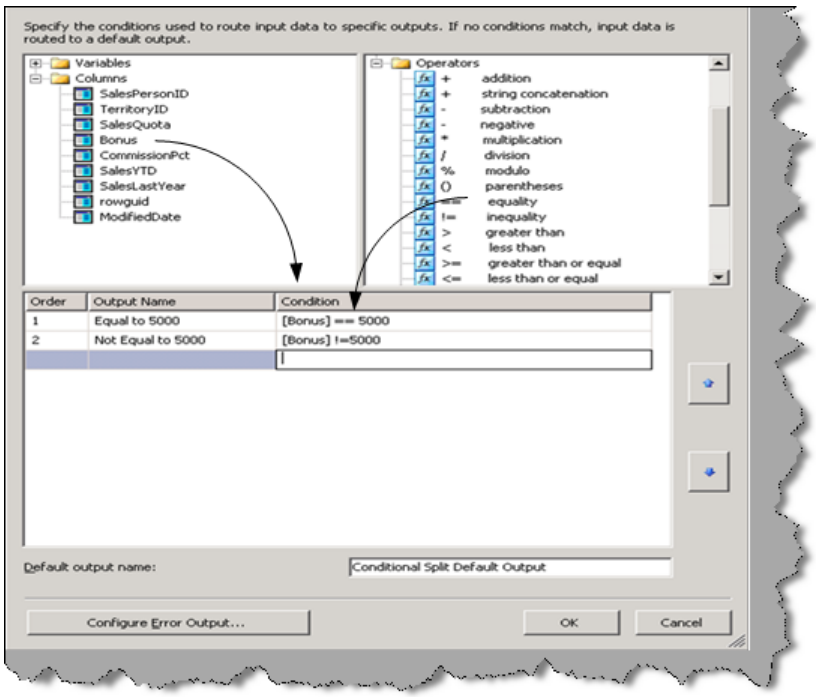


Figure 9.24: - Specifying Conditional Split Criteria

Conditional split now has two outputs one which will go in “Sales.SalesPerson5000” and other in “Sales.SalesPersonNot5000”. Therefore, you have to define two destination and the associate respective tables to it. So drag two OLE DB destination data flow items and connect it the two outputs of conditional split.

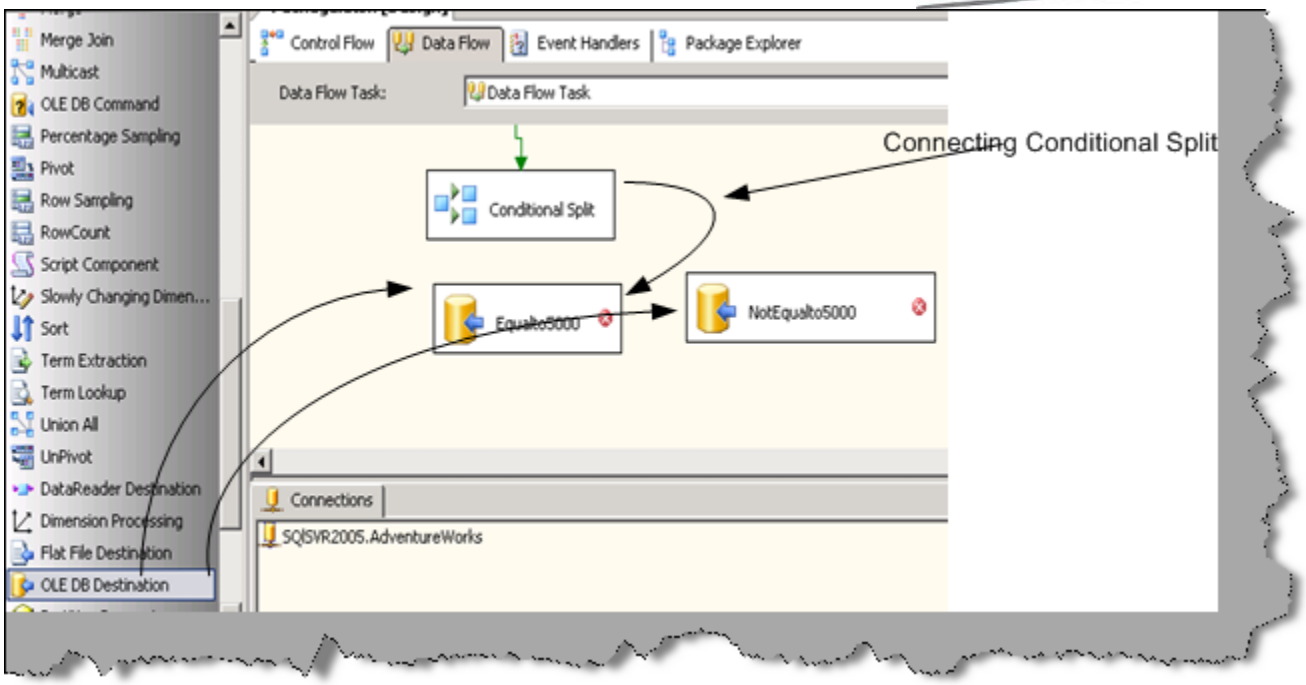


Figure 9.25: - Specify Destination

When you drag from the conditional split items over OLEDB destination items it will pop up a dialog to specify which output this destination has to be connected. Select the one from drop down and press ok. Repeat this step again for the other destination object.

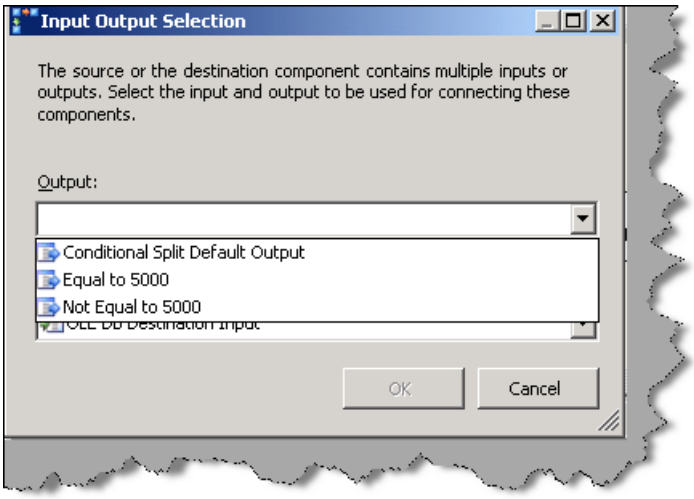


Figure 9.26: - Specify Input and output Selection

That's the final Data flow structure expected.

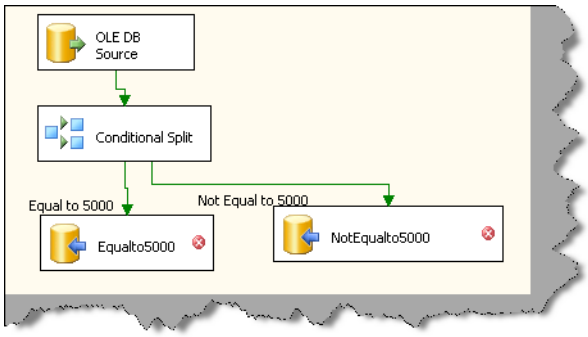


Figure 9.27: - Final DTS

Its time to build and run the solution, which you can do from the drop down. To run the DTS you press the green icon as pointed by arrow in the below figure. After you run query both the tables have the appropriate values or not.

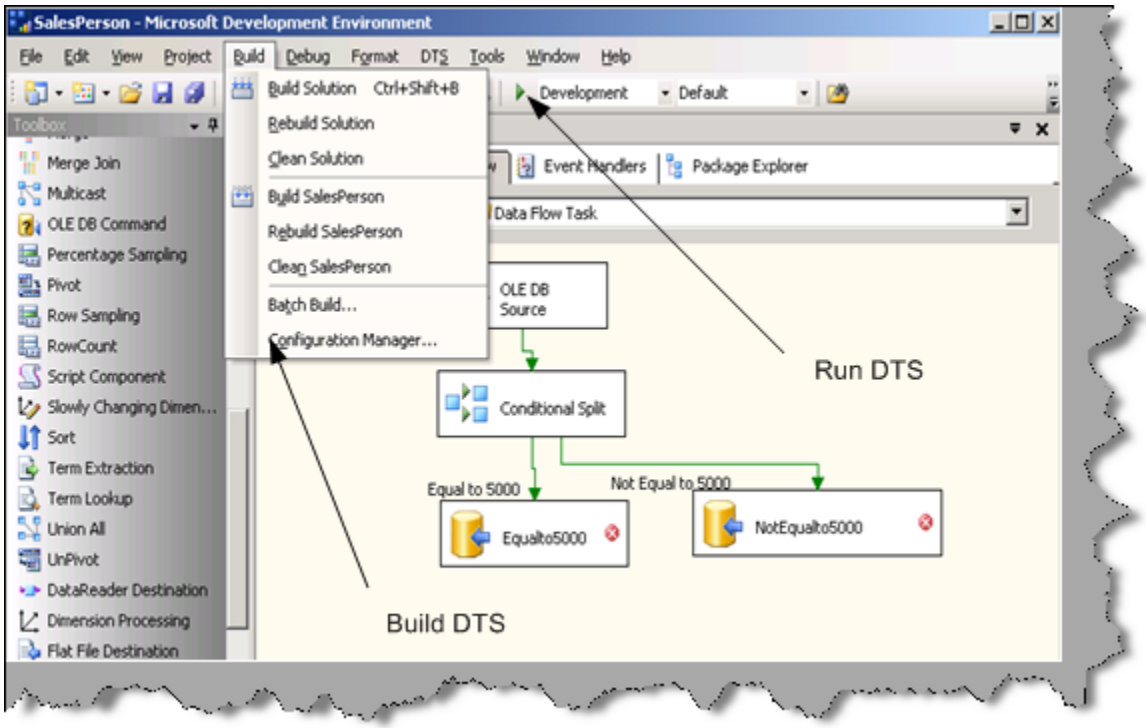


Figure 9.28: - Build and Run

Note: - You can see various data flow items on the right hand side; it's out of the scope to cover all items (You must be wondering how much time this author will say out of scope , but its fact guys something you have to explore). In this sample project we needed the conditional split so we used it. Depending on projects you will need to

explore the toolbox. It's rare that any interviewer will ask about individual items but rather ask fundamentals or general overview of how you did DTS.

Chapter 10: Replication

(Q) Whats the best way to update data between SQL Servers?

By using Replication, we can solve this problem. Many of the developers end up saying DTS, BCP or distributed transaction management. But this is one of the most reliable ways to maintain consistency between databases.

(Q) What are the scenarios you will need multiple databases with schema?

Following are the situations you can end up in to multi-databases architecture:-

24x7 Hours uptime systems for online systems

This can be one of the major requirements for duplicating SQL Server's across network. For instance, you have a system, which is supposed to be 24 hours online. This system is hosted in a central database, which is far away in terms of Geographic's. As said first that this system should be 24 hours online, in case of any break over from the central server we hosted one more server, which is inside the premises. So the application detects that it cannot connect to the online server so it connects to the premises server and continues working. Later in the evening using replication all the data from the local SQL Server is sent to the central server.

License problems

SQL Server per user usage has a financial impact. So many of the companies decide to use MSDE, which is free, so that they do not have to pay for the client licenses. Later every evening or in some specific interval this all data is uploaded to the central server using replication.

Note: - MSDE supports replication.

Geographical Constraints

It is if the central server is far away and speed is one of the deciding criteria.

Reporting Server

In big multi-national sub-companies are geographically far away and the management wants to host a central reporting server for the sales, which they want to use for decision making and marketing strategy. So here, the transactional SQL Server's entire database is scattered across the sub-companies and then weekly or monthly we can push all data to the central reporting server.

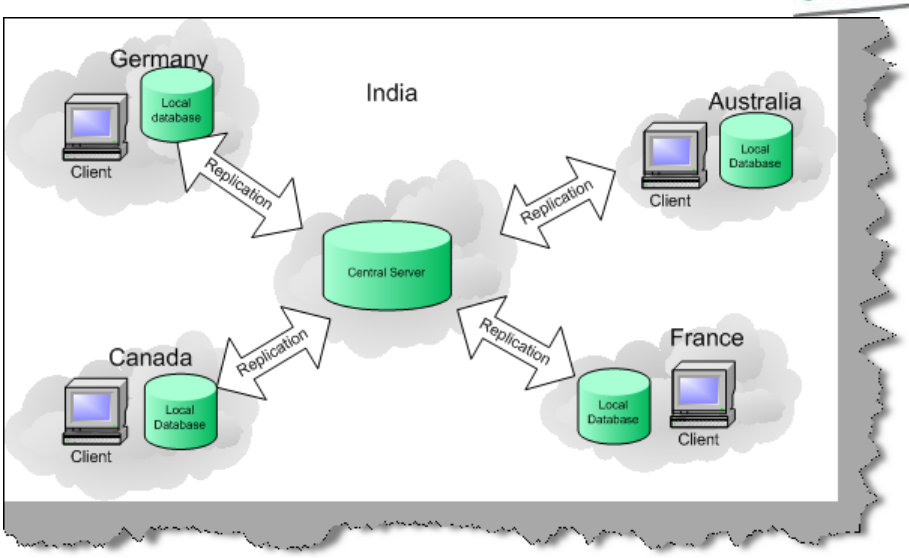


Figure 10.1: - Replication in Action

You can see from the above figure how data is consolidated in to a central server, which is hosted in India using replication.

(DB) How will you plan your replication?

Following are some important questions to be asked before going for replication.

Data planning

It is not necessary that you will need to replicate the complete database. For example, you have Sales database, which has Customer, Sales, Event logs, and History tables. You have requirement to host a centralized reporting server, which will be used by top management to know "Sales by Customer". To achieve this you do not need the whole database on reporting server, from the above you will only need "Sales" and "Customer" tables.

Frequency planning

As defined in the top example let's say management wants only "Sales by Customer weekly", so you do not need to update every day , rather you can plan weekly. But if the top management is looking for "Sales by Customer per day" then probably your frequency of updates would be every night.

Schema should not have volatile "baseline"

Note: - I like this word "baseline" it really adds weight while speaking as a project manager. It's mainly used to control change management in projects. You can say "Baseline" is a process by which you can define a logical commit to a document. For example you are coding a project and you have planned different versions for the project. So after every version you do a baseline and create a setup

```
and deploy to the client side. Any changes after this will be a new
version.
```

One of the primary requirements of a replication is that the schemas, which should be replicated across, should be consistent. If you are keeping on changing schema of the server then replication will have huge difficulty in synchronizing. So if you are going to have huge and continuous changes in the database schema rethink over replication option. Or else a proper project management will help you solve this.

(Q) What are publisher, distributor and subscriber in "Replication"?

Publisher is the one who owns the database and is the main source for data. Publisher identifies what data should be distributed across.

Distributor is a bridge between publisher and subscriber. Distributor gathers all the published data and holds it until it sends it across to all subscriber. So as it is a bridge who sits in between publisher and subscriber, it supports multiple publisher and subscriber concept.

Subscriber is the end source or the final destination to which data has to be transmitted.

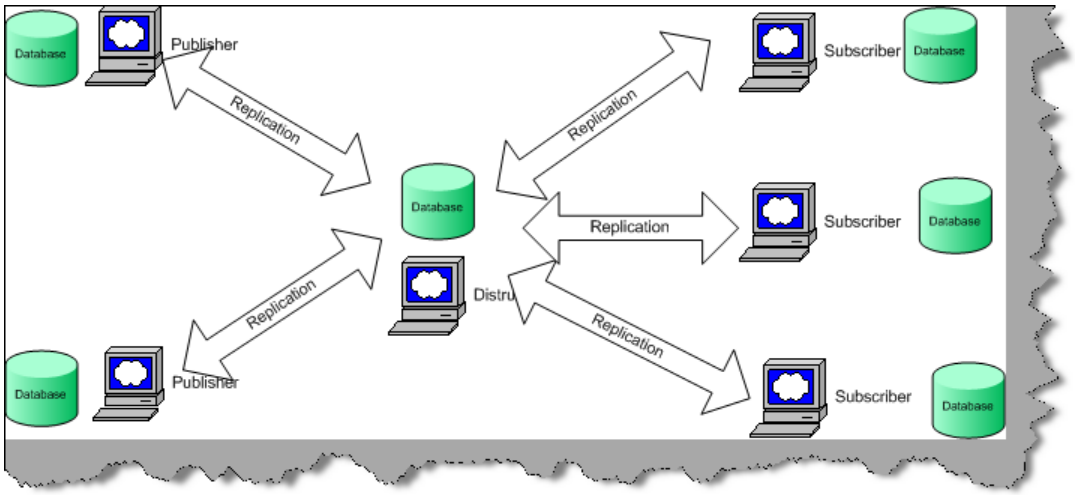


Figure 10.2: - Publisher, Distributor, and Subscriber in action

(Q) What is "Push" and "Pull" subscription?

Subscription can be configured in two ways:-

- **Push subscription**

In push subscription, the publisher has full rights when to update data to subscriber. Subscriber completely plays a passive role in this scenario. This model is needed when we want full control of data in hand of publisher.

- **Pull subscription**

In pull subscription, the subscriber requests for new or changed data. Subscriber decides when to update himself. This model is needed when we want the control to be on hands of subscriber rather than publisher.

(DB) Can a publication support push and pull at one time?

A publication mechanism can have both. But a subscriber can have only one model for one publication. In short, a subscriber can either be in push mode or pull mode for a publication, but not both.

(Q) What are different models / types of replication?

- Snapshot replication
- Merge replication
- Transactional replication

Note: - Below I will go through each of them in a very detail way.

(Q) What is Snapshot replication?

A complete picture of data to be replicated is taken at the source. Depending on the schedule defined when the replication should happen, destination data is completely replaced by this. So over a period of time changes are accumulated at the publisher end and then depending on the schedule it is sent to the destination.

Note: - In snapshot you will also be sending data which has not changed.

(Q) What are the advantages and disadvantages of using Snapshot replication?

Advantages:-

- Simple to setup. If the database is small or you only want to replicate master data (State code, Pin code etc), it is the best approach, as these values do not change heavily.
- If you want to keep a tight control over when to schedule the data this is the best approach. For example, you will like to replicate when the network traffic is low (probably during Saturday and Sunday).

Disadvantages:

Note: - This is probably the least used approach. So definitely the interviewer is expecting the disadvantages points to be clearer, rather than advantages.

- As data start growing the time taken to complete the replication will go on increasing.

(Q) What type of data will qualify for "Snapshot replication"?

- Read-only data are the best candidates for snapshot replication.
- Master tables like zip code, pin code etc are some valid data for snapshot replication.

(Q) What is the actual location where the distributor runs?

You can configure where the distributor will run from SQL Server. But normally if it's a pull subscription it runs at the subscriber end and for push subscription it runs on the publisher side.

(Q) Can you explain in detail how exactly "Snapshot Replication" works?

Following are the basic steps for "Snapshot Replication" to work:-

Note: - There are two important components "Snapshot Agent" and "Distribution Agent" which we will define first. Snapshot agent creates image of the complete published data and copies it to the distributor. Distribution Agent sends the copied image and replaces the data on the subscriber side.

- Snapshot agent places a shared lock on the data to be published.
- Whole snapshot is then copied to the distributor end. There are three files, which are created one for database schema, BCP files and the index data.]
- Finally the snapshot agent releases lock over the published data.
- Distribution agent then replaces the subscriber data using files created by snapshot agent.

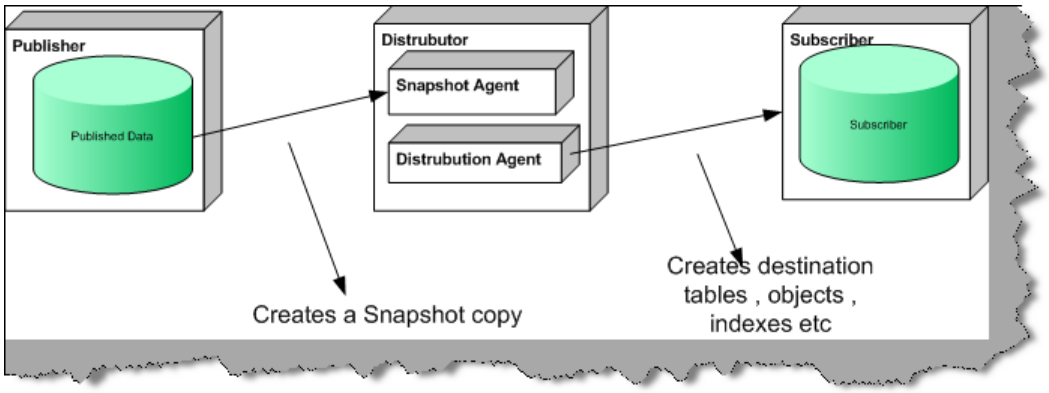


Figure 10.3: - Snapshot replication in Action

(Q) What is merge replication?

If you are looking forward to manage changes on multiple servers, which need to be consolidated, merge replication is the best design.

(Q) How does merge replication works?

Merge Agent component is one of the important components which makes merge replication possible. It consolidates all data from subscriber and applies them to publisher. Merge agent first copies all data from publishers to the subscribers and then replicates them vice-versa so that all stakeholders have consistent data.

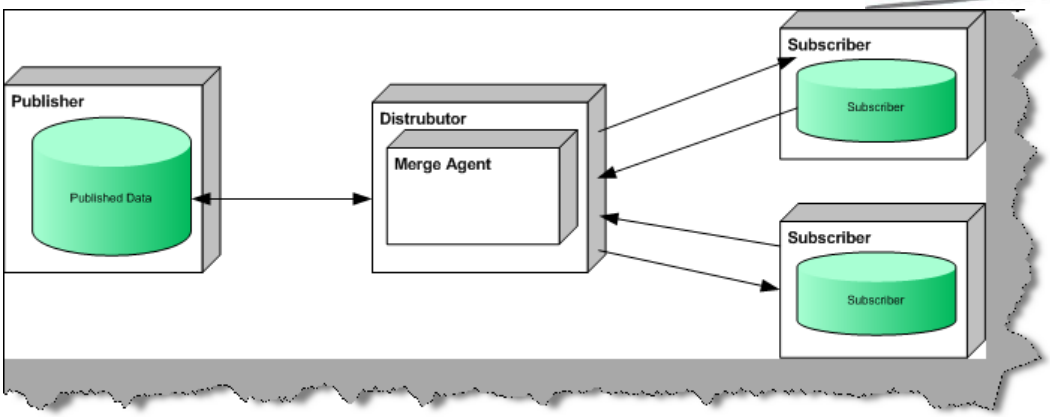


Figure 10.4: - Merge Replication

Merge agent stands in between subscriber and publisher. Any conflicts are resolved through merge agent in turn, which uses conflict resolution. Depending how you have configured the conflict resolution the conflicts are resolved by merge agent.

(Q) What are advantages and disadvantages of Merge replication?

Advantages:

- This is the only way you can manage consolidating multiple server data.

Disadvantage:

- It takes lot of time to replicate and synchronize both ends.
- There is low consistency, as lot of parties has to be synchronized.
- There can be conflicts while merge replication if the same rows are affected in more than once subscriber and publisher. Definitely, there is conflict resolution in place but that adds complication.

(Q) What is conflict resolution in Merge replication?

There can be practical situations where same row is affected by one or many publishers and subscribers. During such critical times Merge agent will look, what conflict resolution is defined and make changed accordingly.

SQL Server uniquely identifies a column using globally unique identifier for each row in a published table. If the table already has a uniqueidentifier column, SQL Server will automatically use that column. Else, it will add a rowguid column to the table and create an index based on the column.

Triggers will be created on the published tables at both the Publisher and the Subscribers. These are used to track data changes based on row or column changes.

(Q) What is a transactional replication?

Transactional replication as compared to snapshot replication does not replicate full data, but only replicates when anything changes or something new is added to the database. So whenever on publisher side we have INSERT, UPDATE and DELETE operations, these changes are tracked and only these changes are sent to the subscriber end. Transactional Replication is one of the most preferred replication methodologies as they send least amount of data across network.

(Q) Can you explain in detail how transactional replication works?

- Any change made to the publisher's database is logged in to a log file.
- Later log reader agent reads the changes and sends it to the distribution agent.
- Distribution agent sends the data across to the subscribers.

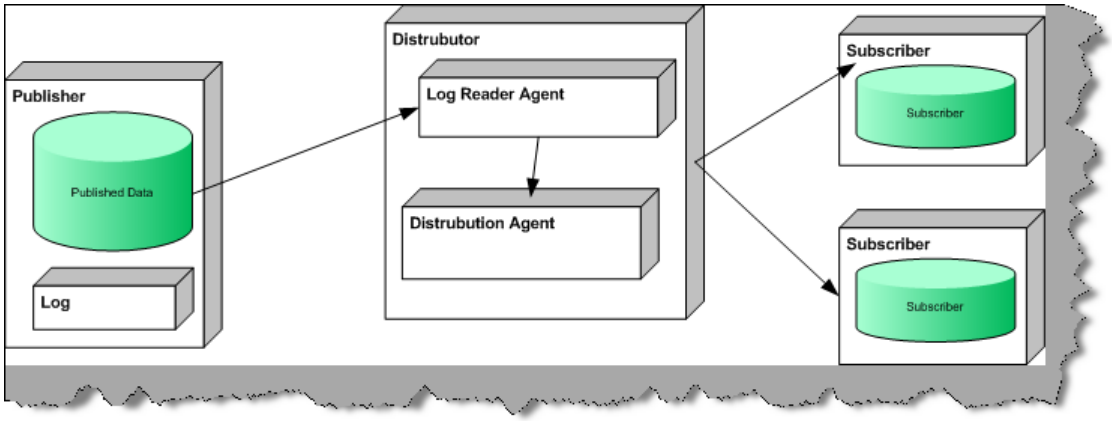


Figure 10.5: - Transactional Replication

(Q) What are data type concerns during replications?

- If it is a transactional replication, you have to include a "Timestamp" column.
- If it is merge replication, you will need a "uniqueidentifier" column. If you do not have one replication creates one.

Note: - As this is an interview question book we will try to limit only to theoretical basis. The best way is to practically do one sample of replication with a sample project. But just for your knowledge I will show some important screen's of replication wizard.

In the "SQL Server Management studio", you can see the publication folder. When you right click on it, you can see the "New Publication" menu.

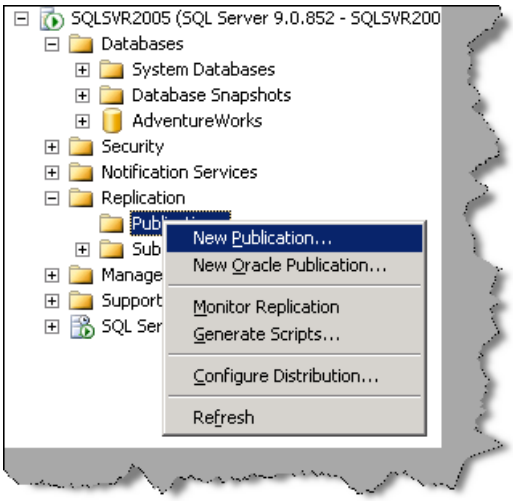


Figure 10. 6: - Create new publication

This wizard will be used to specify the "Distributor".

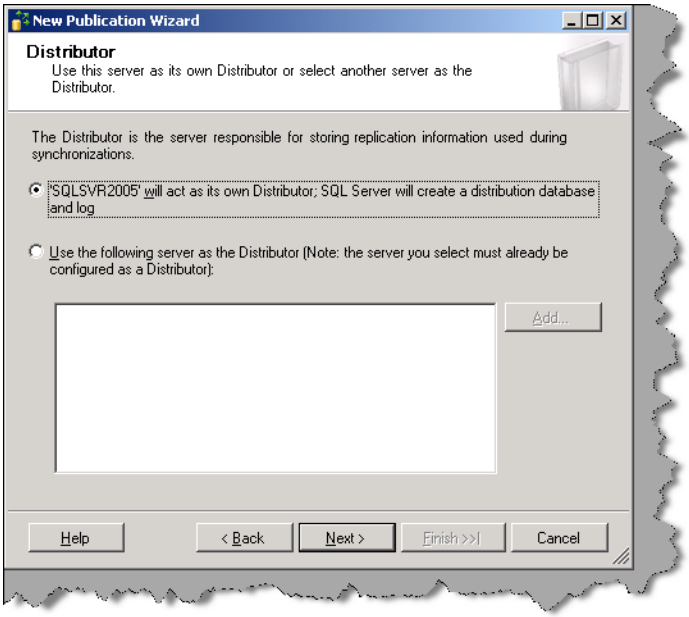


Figure 10.7: - Specify the Server as distributor

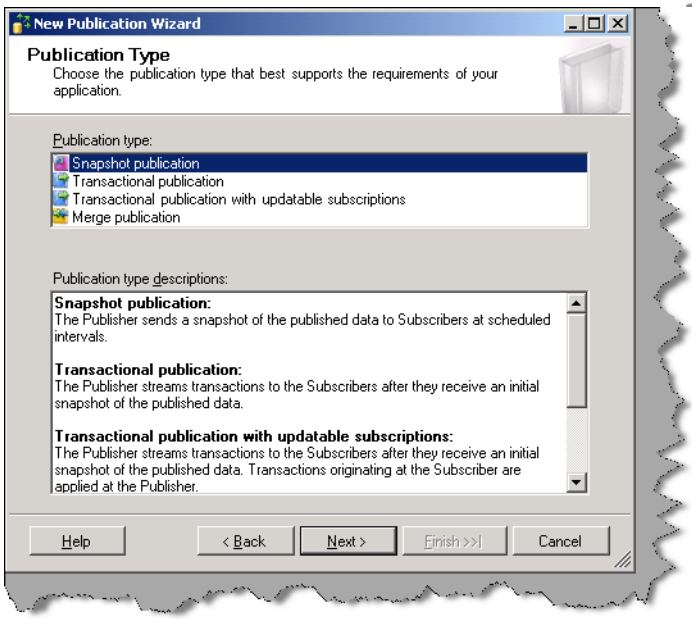


Figure 10.8: - Specify Type of replication

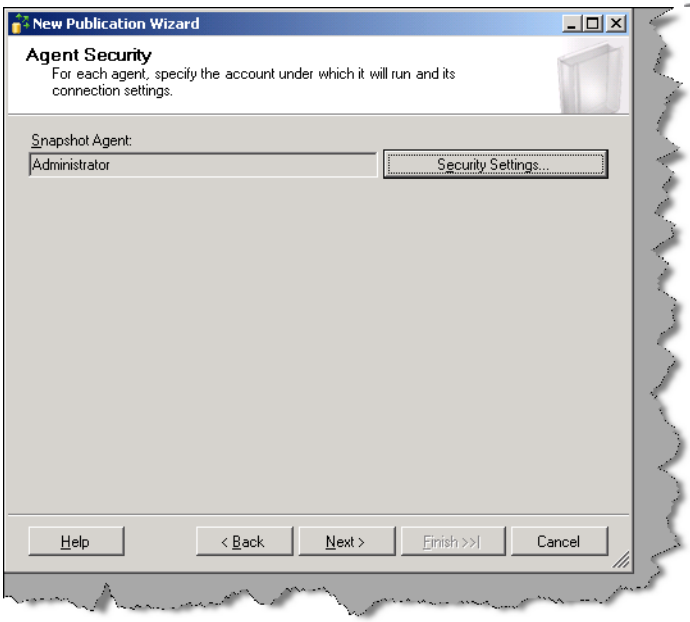


Figure 10.9: - Specify under which agent it will run under

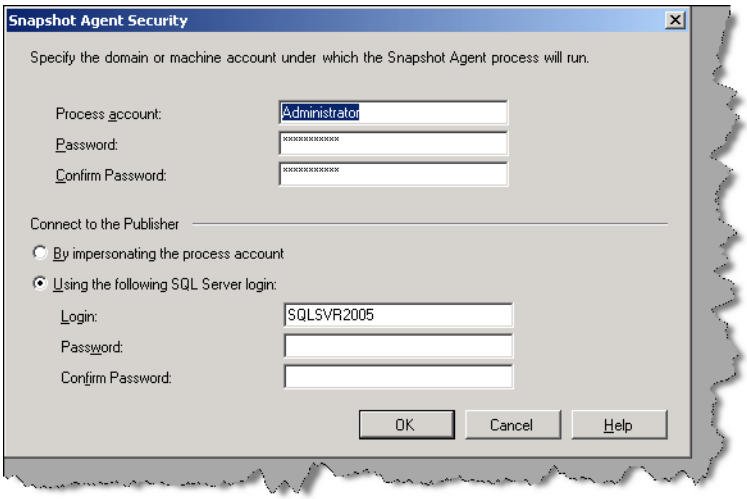


Figure 10.10: - Security Details

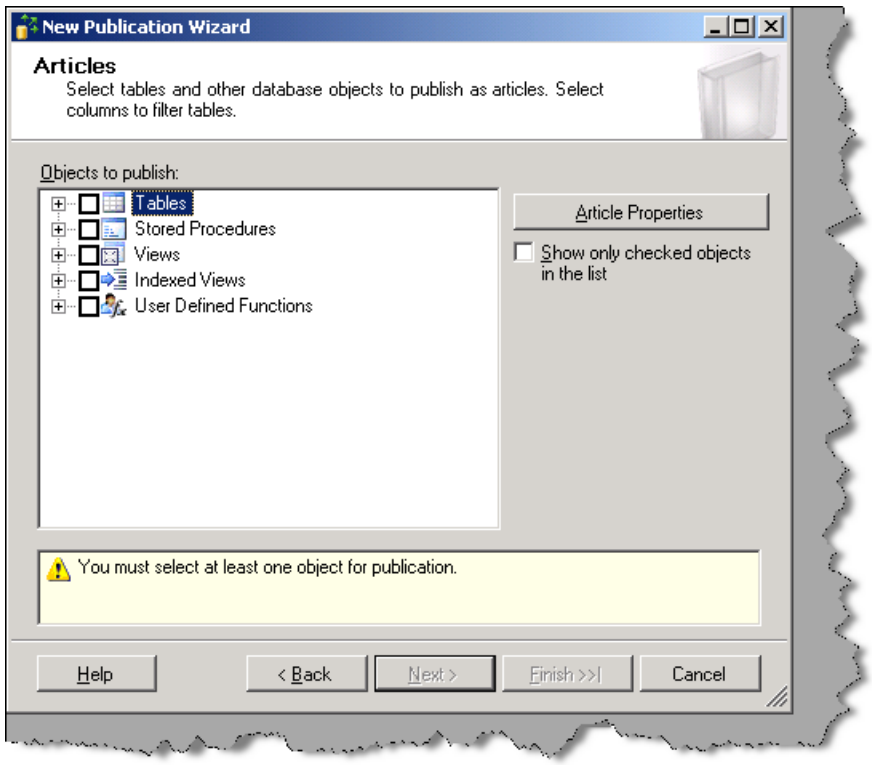


Figure 10.11: - Specify which objects you want to replicate

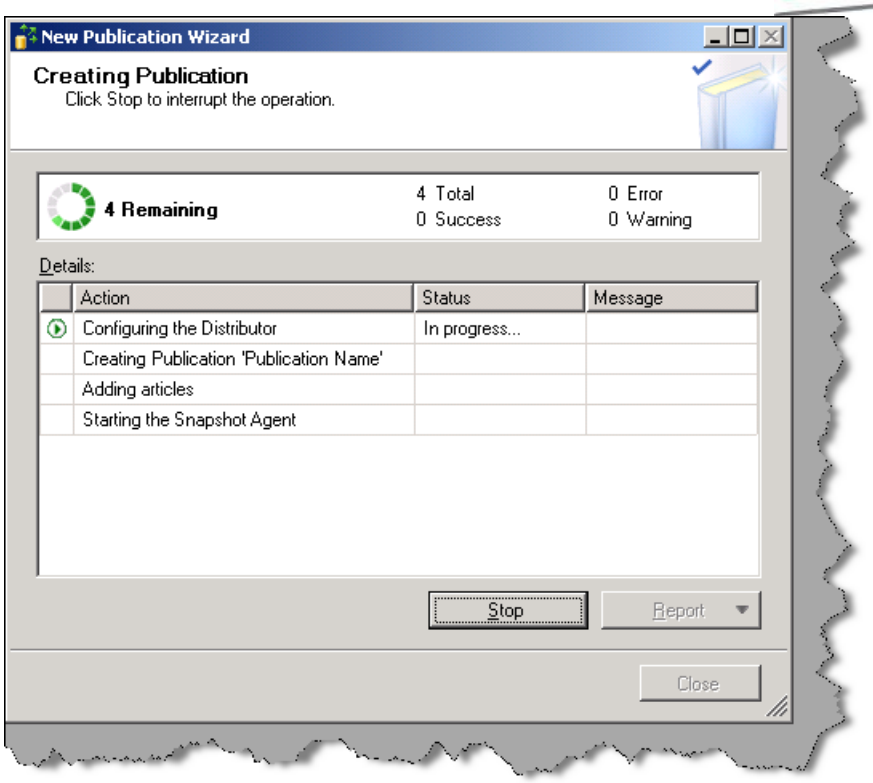


Figure 10.12: - Replication in Action

Chapter 11: Reporting Services

Note: - I know every one screaming this is a part of Data mining and warehousing. I echo the same voice with you my readers, but not necessarily. When you want to derive reports on OLTP systems this is the best way to get your work done. Secondly reporting services is used so much heavily in projects now a day that it will be completely unfair to discuss this topic in a short way as subsection of some chapter.

(Q) Can you explain how can we make a simple report in reporting services?

We will be using "AdventureWorks" database for this sample. We would like to derive a report how much quantity sales were done per product. For this sample we will have to refer three tables Salesorderdetails, Salesorderheader and product table. Below is the SQL which also shows what the relationship between those tables is:-

```
select production.product.Name  as ProductName, count(*) as TotalSales
from  sales.salesorderdetail
```

```
inner join Sales.Salesorderheader
on Sales.Salesorderheader.salesorderid=
Sales.Salesorderdetail.Salesorderid
inner join production.product
on production.product.productid=sales.salesorderdetail.productid
group by production.product.Name
```

Therefore, we will be using the above SQL and trying to derive the report using reporting services.

First click on business intelligence studio menu in SQL Server 2005 and say File --> New --> Project. Select the "Report" project wizard. Let's give this project name "TotalSalesByProduct". You will be popped with a startup wizard as shown below.

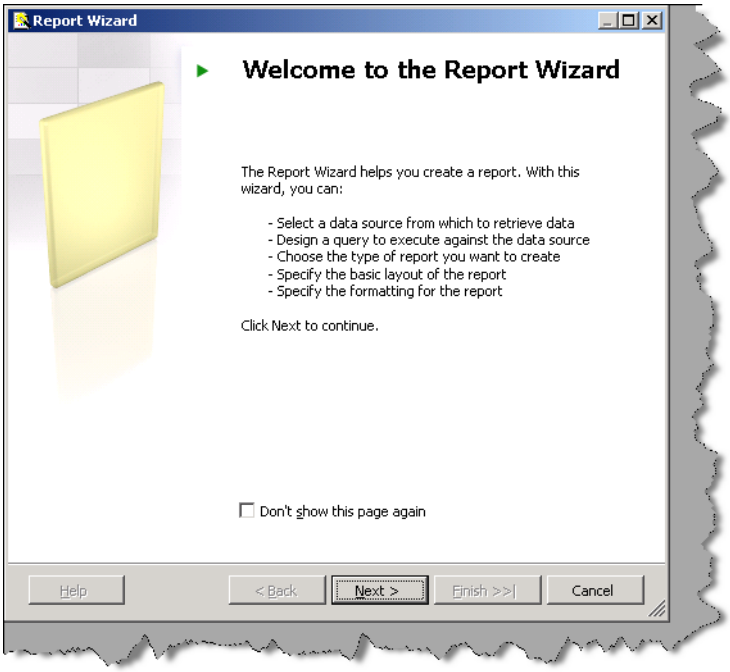


Figure 11. 1: - Welcome reporting services wizard

Click next and you will be prompted to input data source details like type of server, connection string and name of data source. If you have the connection string just paste it on the text area or else click edit to specify connection string values through GUI.

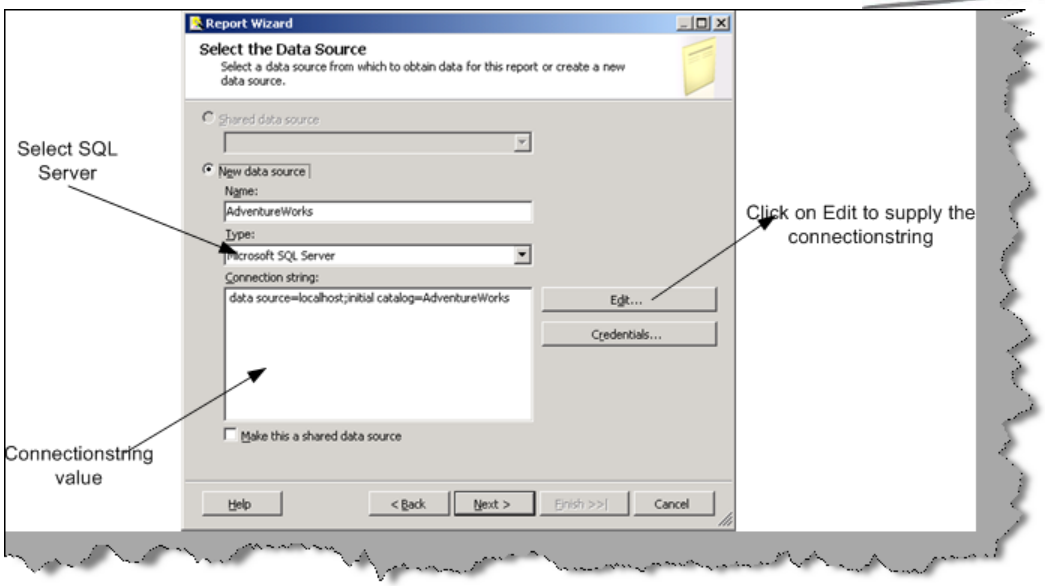


Figure 11.2: - Specify Data Source Details

As we are going to use SQL Server for this sample specify OLEDB provider for SQL Server and click next.

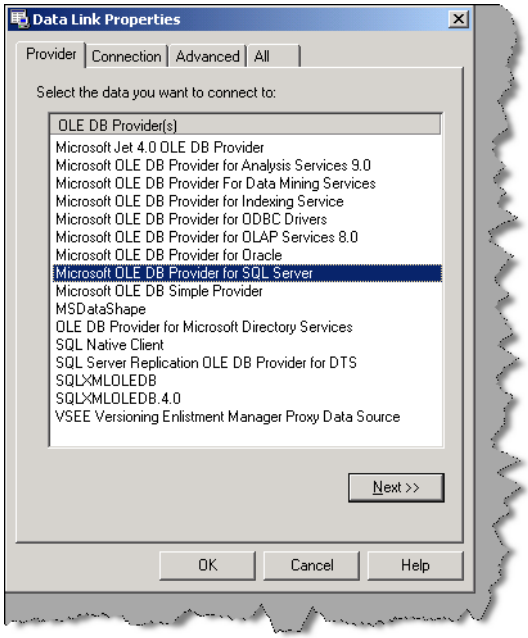


Figure 11.3: - Specify provider

After selecting the provider specify the connection details, which will build your connection string. You will need to specify the following details Server Name, Database name and security details.

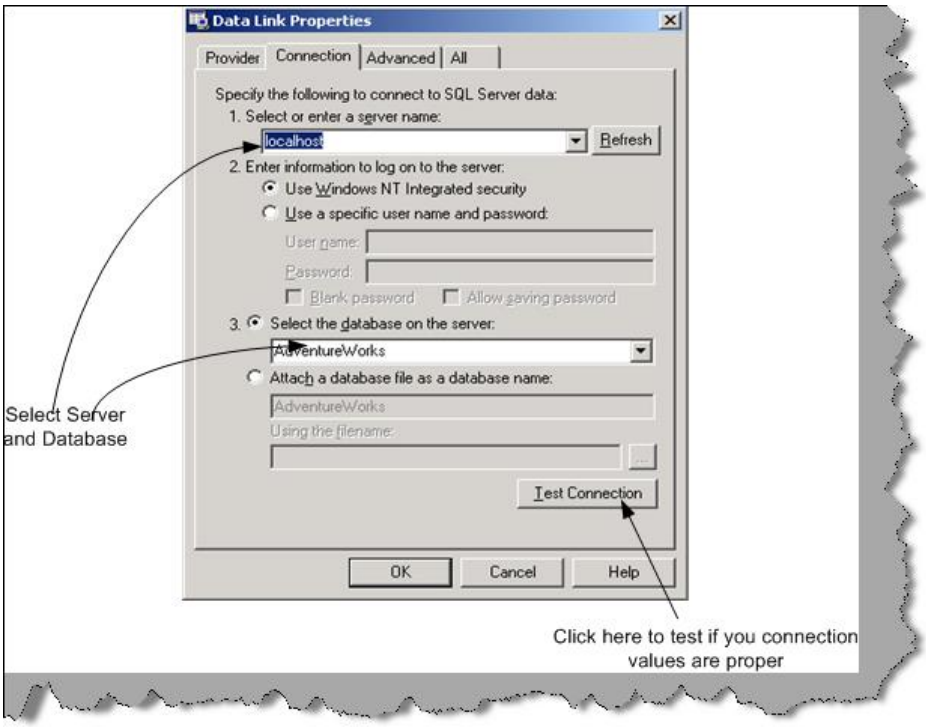


Figure 11.4: - Specify Connection Details

This is the most important step of reporting services, specifying SQL. You remember the top SQL we had specified the same we are pasting it here. If you are not sure about the query, you can use the query builder to build your query.

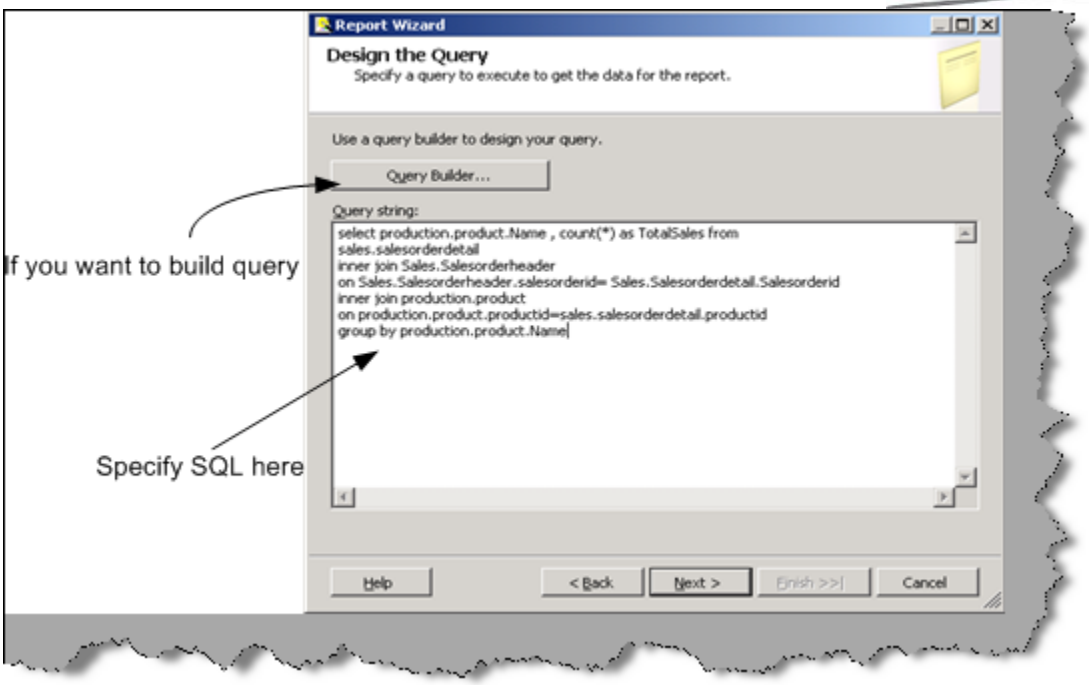


Figure 11.5: - SQL Query

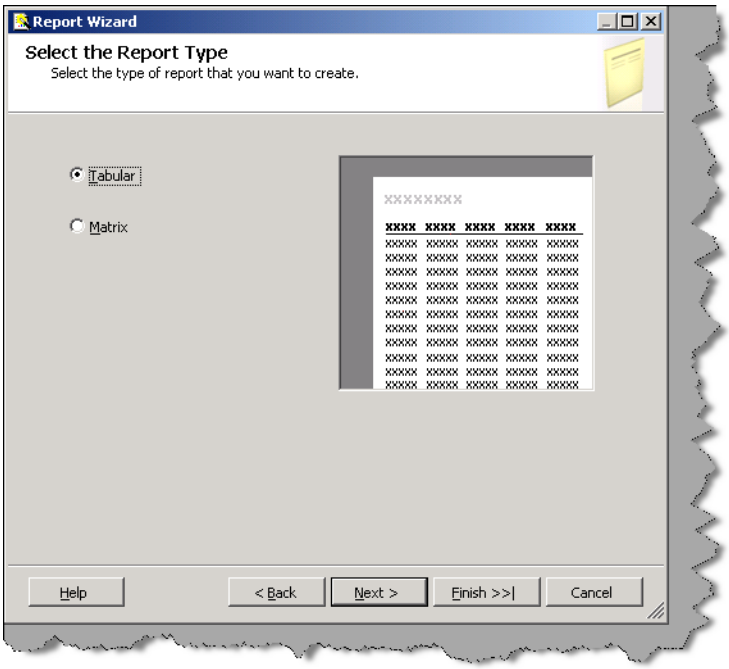


Figure 11.6: - Type of report

Now it is the time to include the fields in reports. At this moment, we have only two fields name of product and total sales.

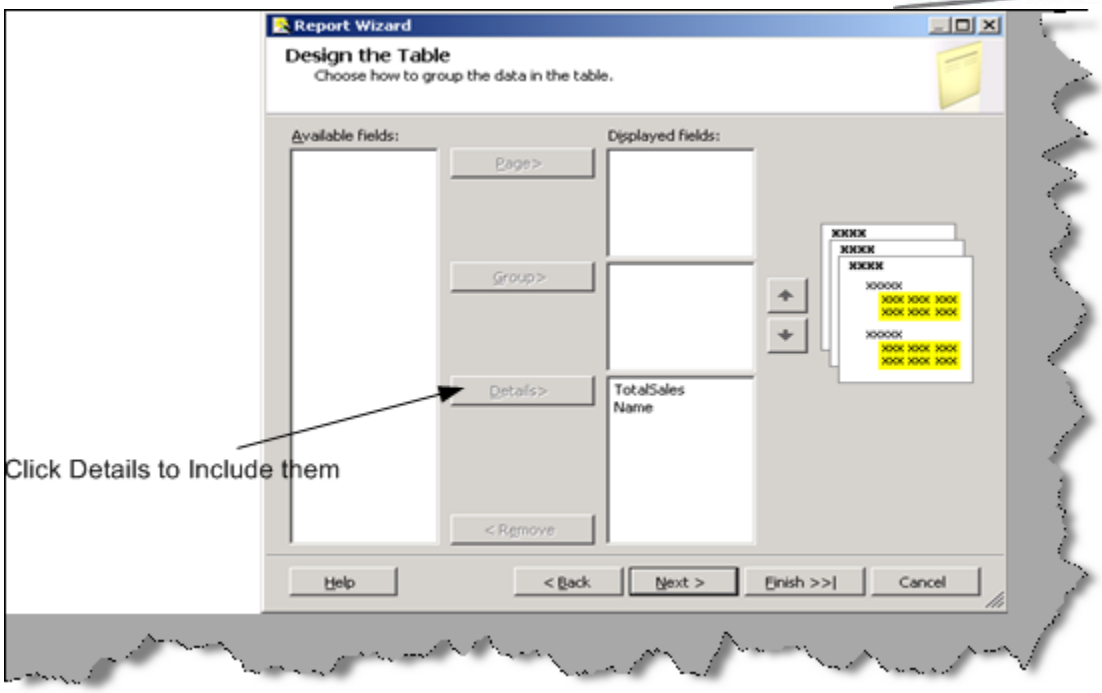


Figure 11.7: - Specify field positions

Finally, you can preview your report. In the final section, there are three tabs data, layout and preview. In data tab, you see your SQL or the data source. In layout tab, you can design your report most look and feel aspect is done in this section. Finally below is the preview where you can see your results.

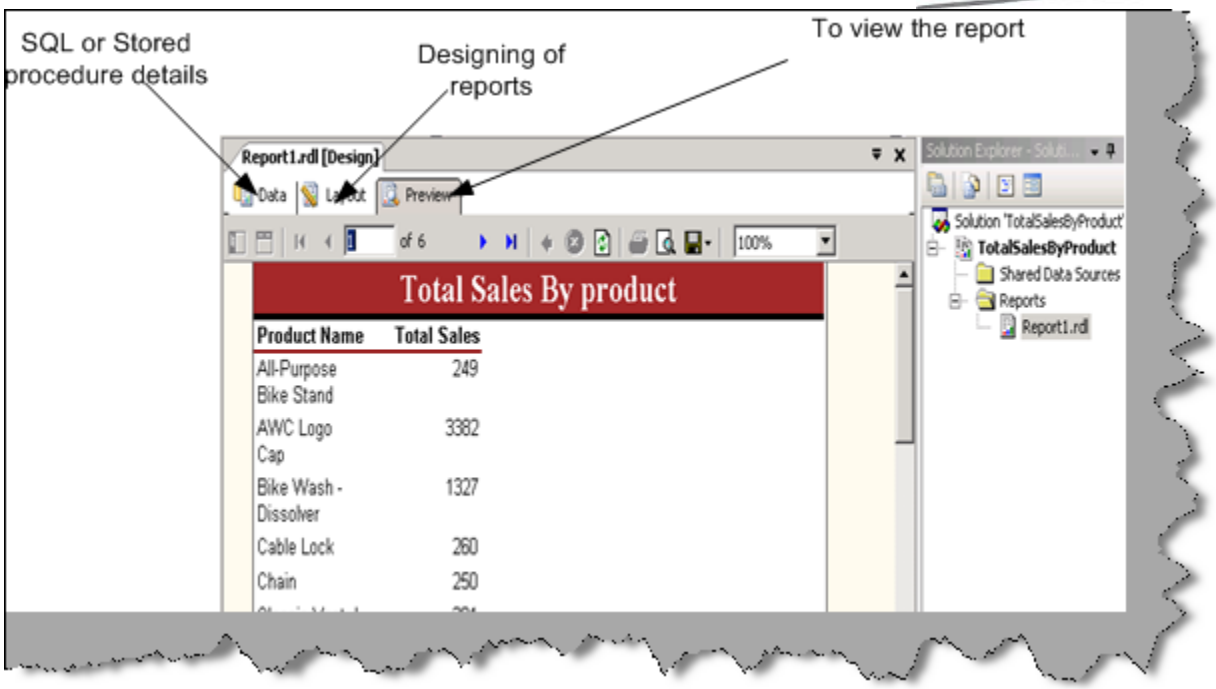


Figure 11.8: - Final view of the report

(Q) How do I specify stored procedures in Reporting Services?

There are two steps to specify stored procedures in reports of reporting services:-

Specify it in the query string.

For instance, we have a stored procedure "GettotalSalesofproductsbySales" which has "@ProductSold" as the input parameter.

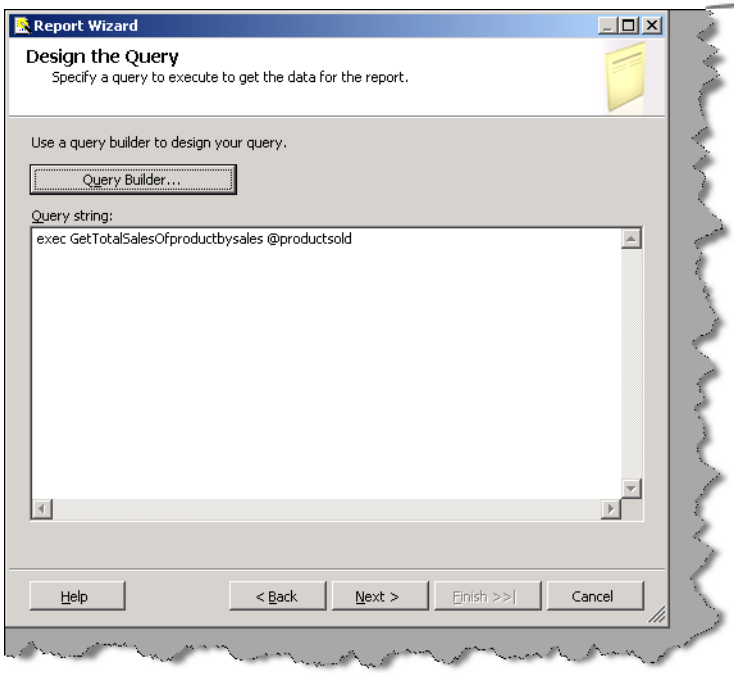


Figure 11.9: - stored procedure in the query builder

You have to also specify the command type from the data tab.

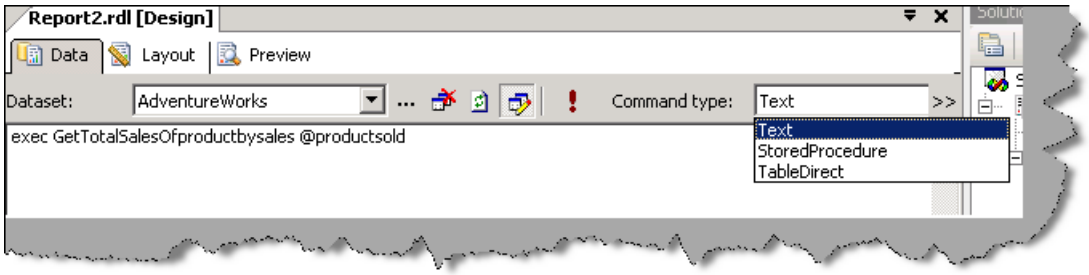


Figure 11.10: - Specify the command type from the Data tab.

(Q) What is the architecture for "Reporting Services "?

"Reporting Services" is not a stand-alone system but rather a group of server sub-system, which work together for creation, management, and deployment of reports across the enterprise.

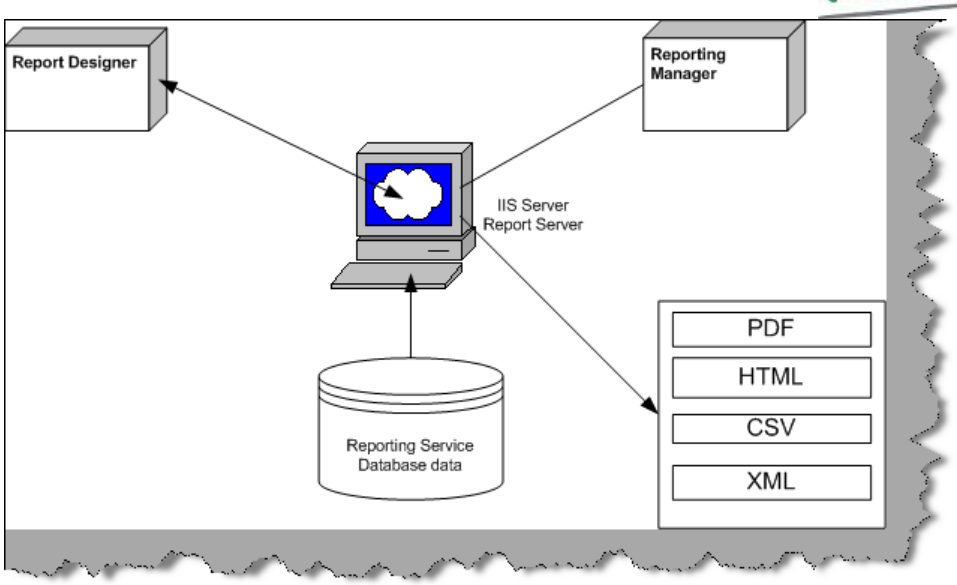


Figure 11.11: - Reporting Services Architecture

Report designer

This is an interactive GUI, which will help you to design and test your reports.

Reporting Service Database

After the report is designed, they are stored in XML format. These formats are in RDL (Report Design Layout) formats. These entire RDL format are stored in Report Service Database.

Report Server

Report Server is nothing but an ASP.NET application running on IIS Server. Report Server renders and stores these RDL formats.

Report Manager

It's again an ASP.NET web based application, which can be used by administrators to control security and managing reports. From administrative perspective that have the authority to create the report, run the report etc...

You can also see the various formats, which can be, generated XML, HTML etc using the report server.

Chapter 13: Transaction and Locks

(Q) What is a "Database Transactions "?

It is a unit of interaction within a database, which should be independent of other transactions.

(Q) What is ACID?

"ACID" is a set of rule which are laid down to ensure that "Database transaction" is reliable. Database transaction should principally follow ACID rule to be safe. "ACID" is an acronym, which stands for:-

- **Atomicity**

A transaction allows for the grouping of one or more changes to tables and rows in the database to form an atomic or indivisible operation. That is, either all of the changes occur or none of them do. If for any reason the transaction cannot be completed, everything this transaction changed can be restored to the state it was in prior to the start of the transaction via a rollback operation.

- **Consistency**

Transactions always operate on a consistent view of the data and when they end always leave the data in a consistent state. Data may be said to be consistent as long as it conforms to a set of invariants, such as no two rows in the customer table have the same customer id and all orders have an associated customer row. While a transaction executes these invariants may be violated, but no other transaction will be allowed to see these inconsistencies, and all such inconsistencies will have been eliminated by the time the transaction ends.

- **Isolation**

To a given transaction, it should appear as though it is running all by itself on the database. The effects of concurrently running transactions are invisible to this transaction, and the effects of this transaction are invisible to others until the transaction is committed.

- **Durability**

Once a transaction is committed, its effects are guaranteed to persist even in the event of subsequent system failures. Until the transaction commits, not only are any changes made by that transaction not durable, but are guaranteed not to persist in the face of a system failure, as crash recovery will rollback their effects.

The simplicity of ACID transactions is especially important in a distributed database environment where the transactions are being made simultaneously.

(Q) What is "Begin Trans", "Commit Tran", "Rollback Tran" and "SaveTran"?

Begin Tran: - It's a point which says that from this point onwards we are starting the transaction.

Commit Tran: - This is a point where we say we have completed the transaction. From this point the data is completely saved in to database.

Rollback Tran: - This point is from where we go back to the start point that i.e. "Begin Tran" stage.

Save Tran: - It is like a bookmark for rollback to come to some specified state. When we say "rollback Tran", we go back directly to "Begin Tran", but what if we want to go back to some specific point after "Begin Tran". Therefore, "Save Tran" is like bookmarks, which can be used to come back to that state rather than going directly to the start point.

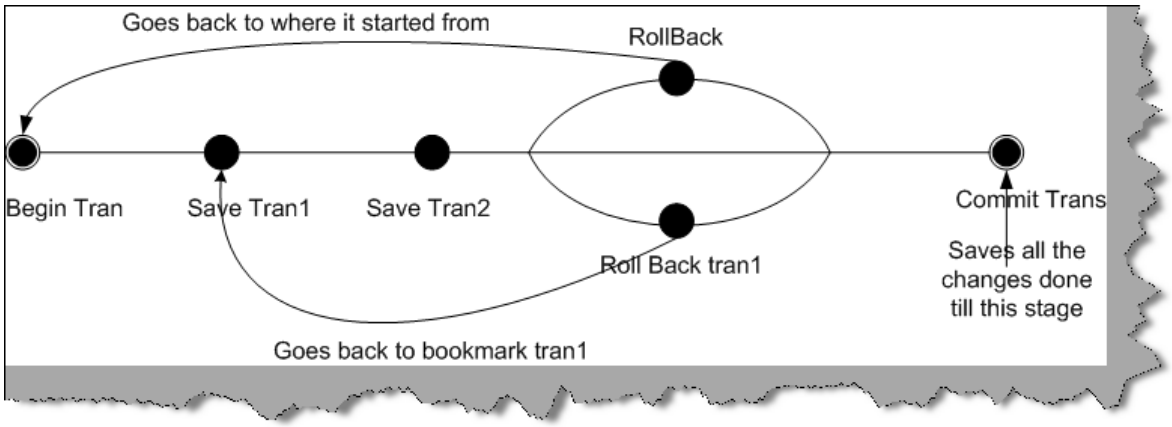


Figure 13.1: - Different Types of Transaction Points

There are two paths defined in the transaction one which rollbacks to the main state and other which rollbacks to a "tran1". You can also see "tran1" and "tran2" are planted in multiple places as bookmark to roll-back to that state.

Brushing up the syntaxes

To start a transaction

BEGIN TRAN Tran1

Creates a book point

SAVE TRAN PointOne

This will roll back to point one

ROLLBACK TRAN PointOne

This commits complete data right when Begin Tran point

COMMIT TRAN Tran1

(DB) What are "Checkpoint's" in SQL Server?

In normal operation everything that is done by SQL Server is not committed directly to database. All operation is logged in to "Transaction Log" first. "CheckPoint" is a point which signals SQL Server to save all data to main database. If there are no "CheckPoints" then the log file will get full.

You can use the "CHECKPOINT" command to commit all data in to SQL SERVER. "CheckPoint" command is also fired when you shut the SQL Server, that's why it takes long time to shut down.

(DB) What are “Implicit Transactions”?

In order to initiate a transaction we use “Begin Tran Tran1” and later when we want to save complete data we use “Commit Tran <TransactionName>”. In SQL Server you can define to start transaction by default i.e. with out firing “Begin Tran Tr1”. You can set this by using:-

SET IMPLICIT_TRANSACTIONS ON

So after the above command is fired any SQL statements that are executed will be by default in transaction. You have to only fire “Commit Tran <Transaction Name>” to close the transaction.

(DB) Is it good to use “Implicit Transactions”?

No. If in case developer forgets to shoot the “Commit Tran” it can open lot of transaction’s which can bring down SQL Server Performance.

(Q) What is Concurrency?

In multi-user environment if two users are trying to perform operations (Add, Modify and Delete) at the same time is termed as “Concurrency”. In such scenarios, there can be lot of conflicts about the data consistency and to follow ACID principles.

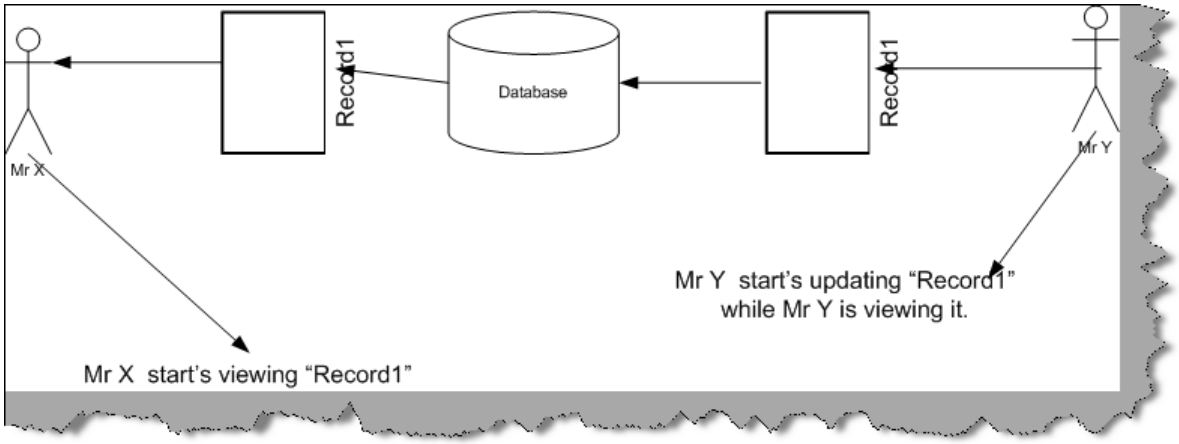


Figure 13.2: - Concurrency Problem

For instance, the above figure depicts the concurrency problem. “Mr X” started viewing “Record1” after some time “MR Y” picks up “Record1” and starts updating it. So “Mr X” is viewing data which is not consistent with the actual database.

(Q) How can we solve concurrency problems?

Concurrency problems can be solved by implementing proper “Locking strategy”. In short by “Locking”. Locks prevent action on a resource to be performed when some other resource is already performing some action on it.

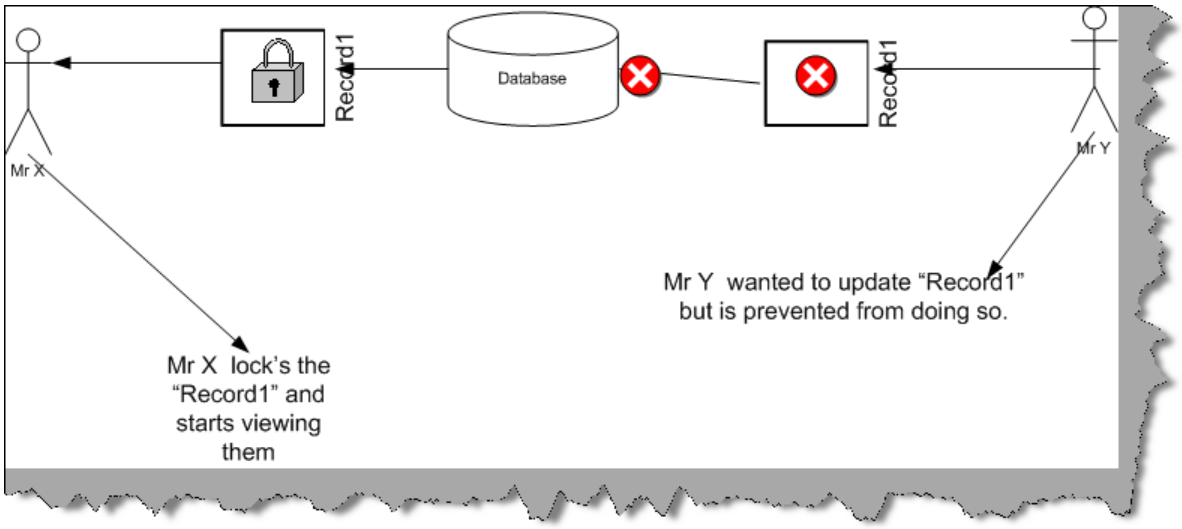


Figure 13.3: - Locking implemented

In our first question we saw the problem above is how locking will work. "Mr. X" retrieves "Record1" and locks it. When "Mr Y" comes in to update "Record1" he can not do it as it's been locked by "Mr X".

Note: - What I have showed is small glimpse, in actual situations there are different types of locks we will going through each in the coming questions.

(Q) What kind of problems occurs if we do not implement proper locking strategy?

There are four major problems that occur:-

- Dirty Reads
- Unrepeatable reads
- Phantom reads
- Lost updates

(Q) What are "Dirty reads"?

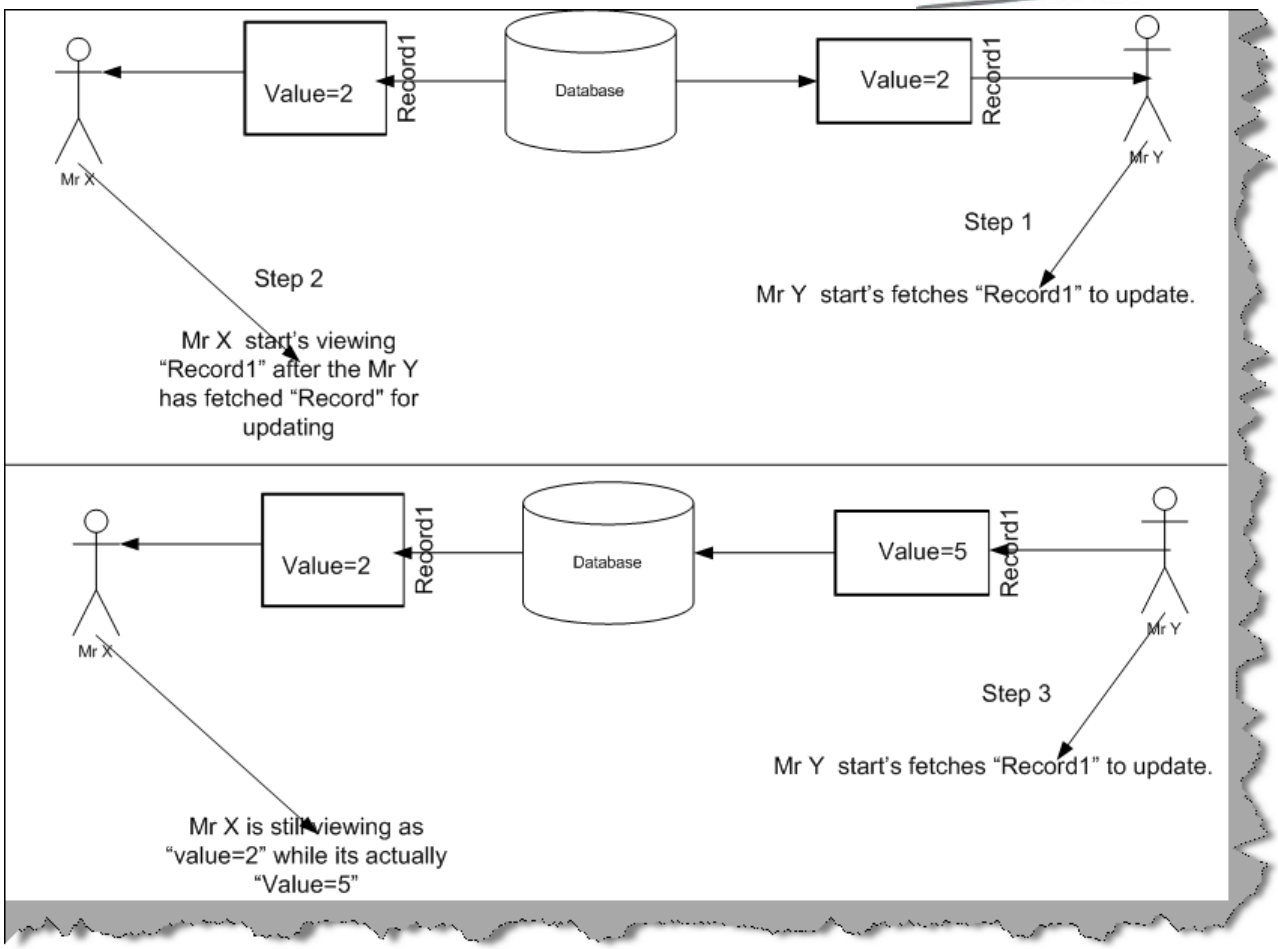


Figure 13.4: - Dirty Reads

"Dirty Read" occurs when one transaction is reading a record, which is part of a half, finished work of other transaction. Above figure defines the "Dirty Read" problem in a pictorial format. I have defined all activities in Step's which shows in what sequence they are happening (i.e. Step1, Step 2 etc).

- **Step1: -** "Mr. Y" Fetches "Record" which has "Value=2" for updating it.
- **Step2:-** In Mean Time, "Mr. X" also retrieves "Record1" for viewing. He also sees it as "Value=2".
- **Step3:-** While "Mr. X" is viewing the record, concurrently "Mr. Y" updates it as "Value=5". Boom… the problem "Mr. X" is still seeing it as "Value=3", while the actual value is "5".

(Q) What are "Unrepeatable reads"?

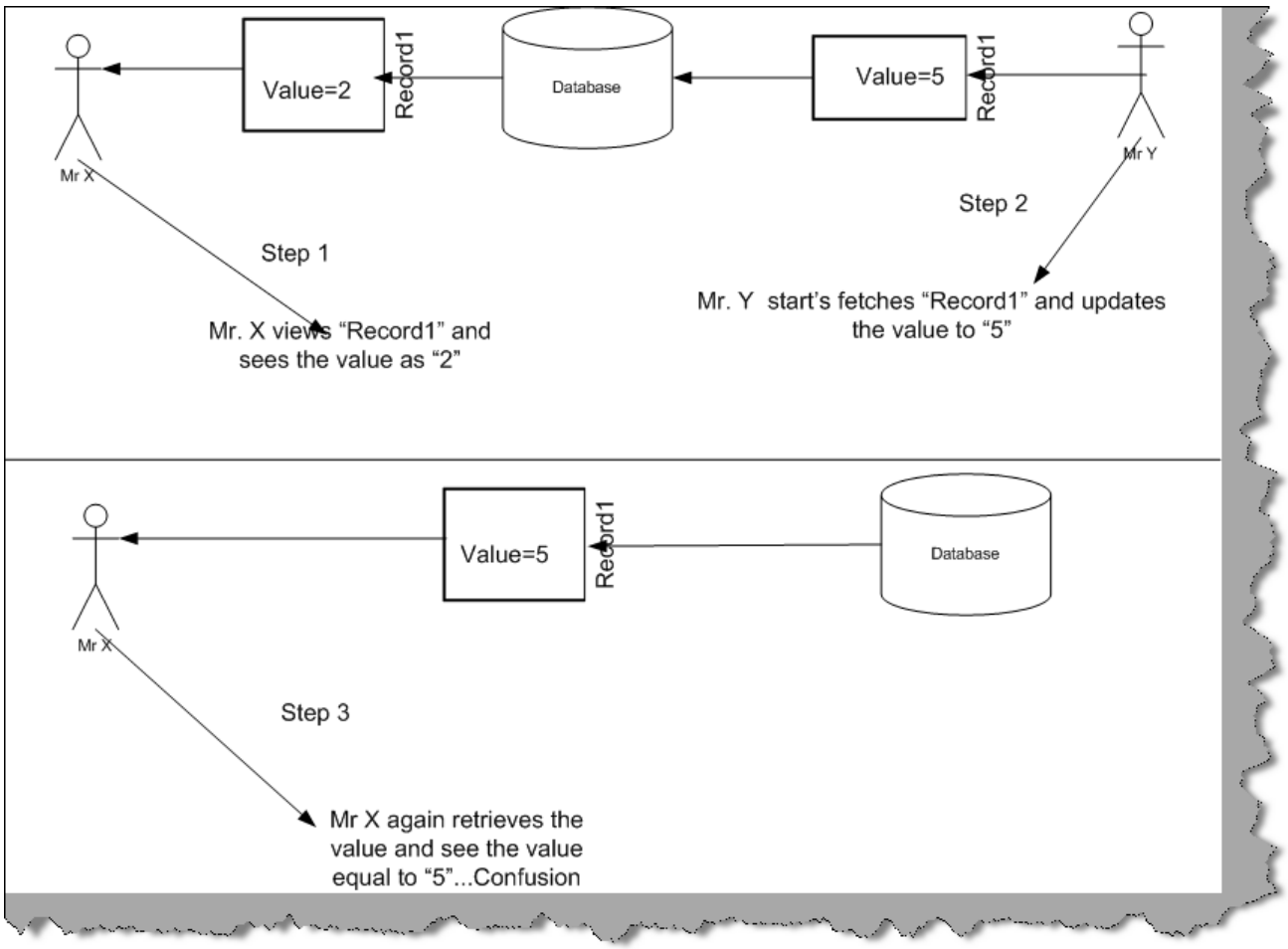


Figure 13.5: - Unrepeatable Read

In every data read if you get different values then it's an "Unrepeatable Read" problem. Lets try to iterate through the steps of the above given figure:-

- **Step1:-** "Mr. X" get "Record" and sees "Value=2".
- **Step2**:- "Mr. Y" meantime comes and updates "Record1" to "Value=5".
- **Step3:-** "Mr. X" again gets "Record1" ohh... values are changed "2" … Confusion.

(Q) What are "Phantom rows"?

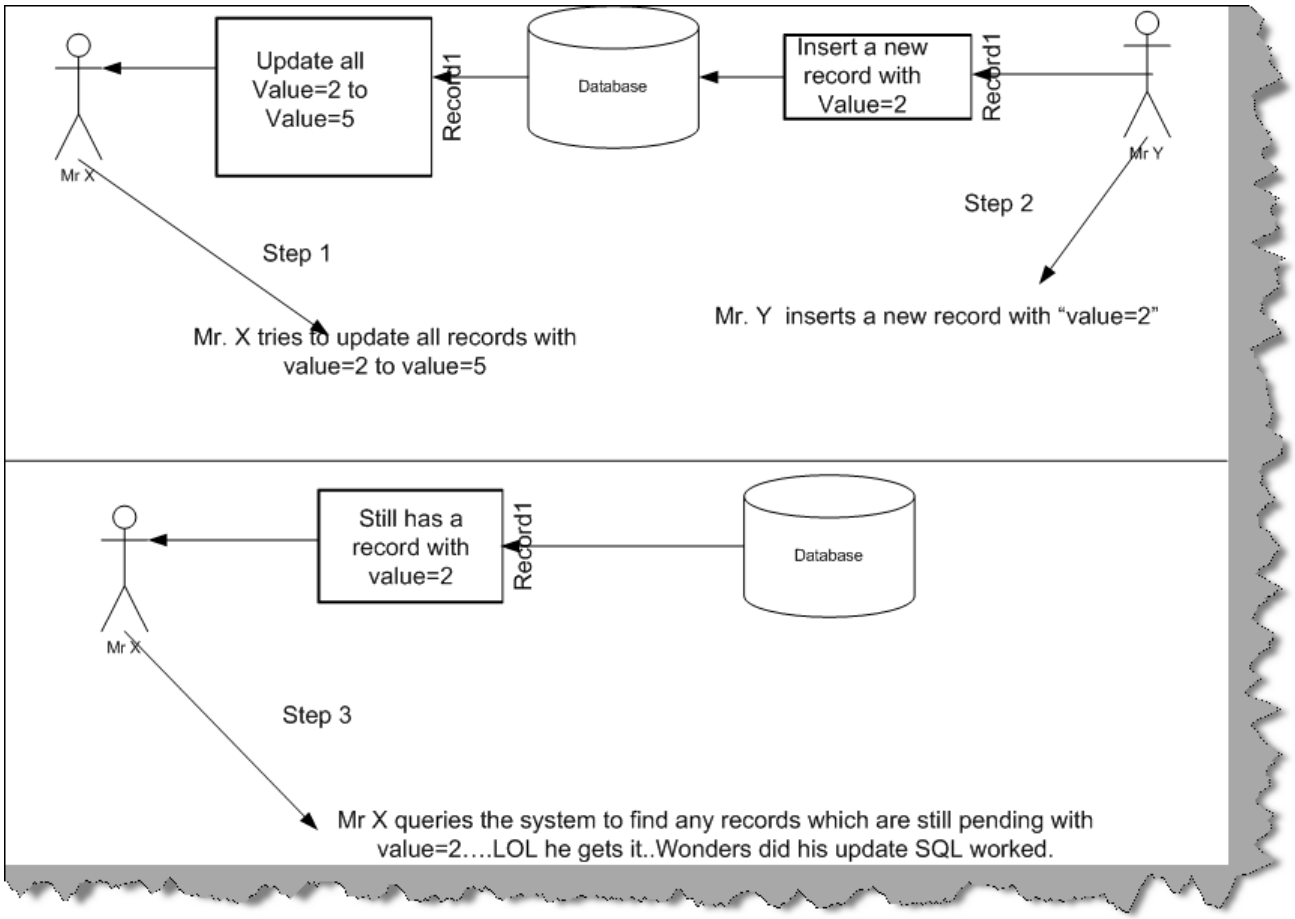


Figure 13.6: - Phantom Rows

If "UPDATE" and "DELETE" SQL statements seems to not affect the data then it can be "Phantom Rows" problem.

- **Step1:-** "Mr. X" updates all records with "Value=2" in "record1" to "Value=5".
- **Step2:-** In mean time "Mr. Y" inserts a new record with "Value=2".
- **Step3**:- "Mr. X" wants to ensure that all records are updated, so issues a select command for "Value=2"….surprisingly find records which "Value=2"…

So "Mr. X" thinks that his "UPDATE" SQL commands are not working properly.

(Q) What are “Lost Updates”?

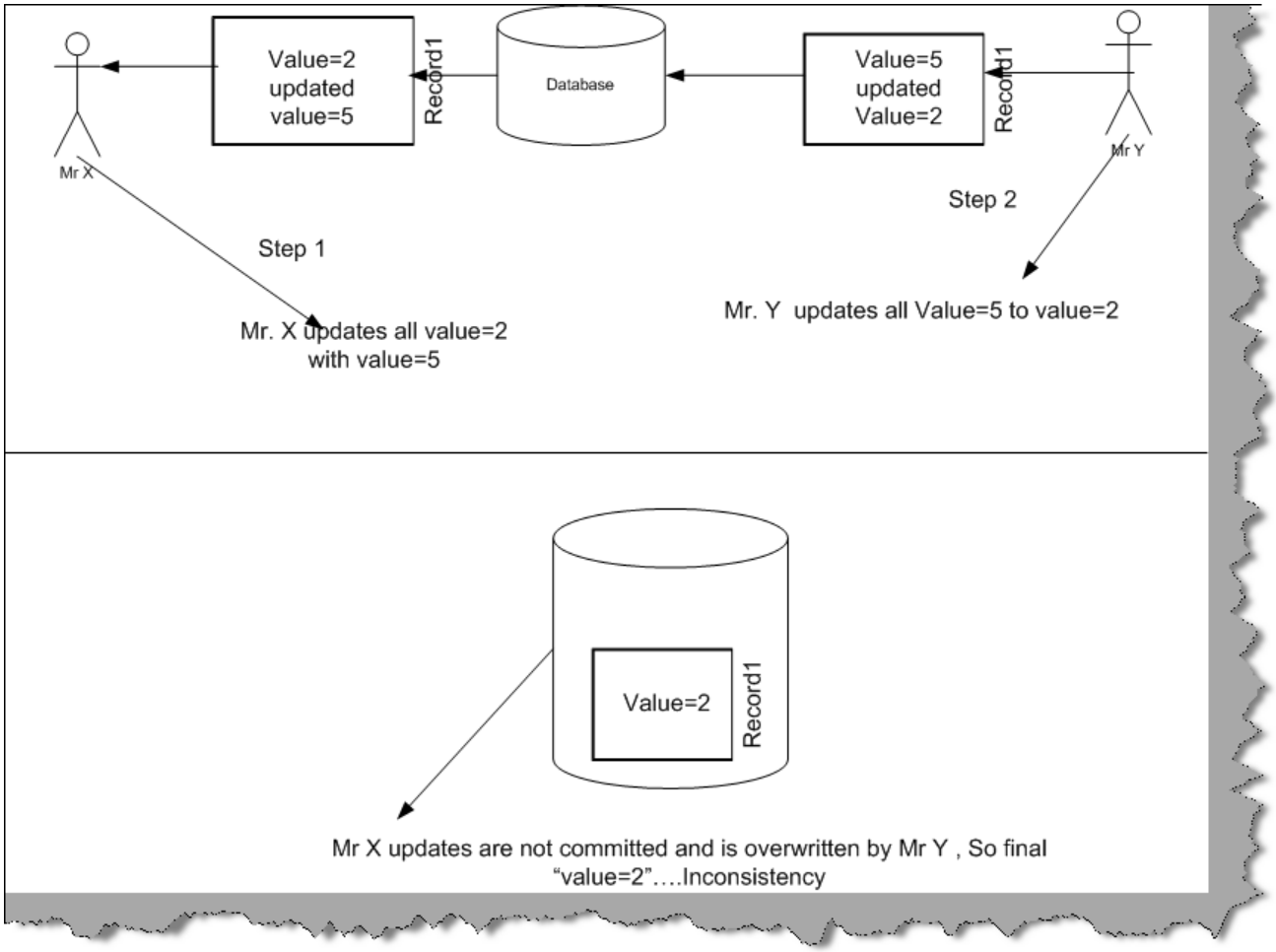


Figure 13.7: - Lost Updates

“Lost Updates” are scenario where one updates which is successfully written to database is over-written with other updates of other transaction. So let’s try to understand all the steps for the above figure:-

- **Step1:-** “Mr. X” tries to update all records with “Value=2” to “Value=5”.
- **Step2**:- “Mr. Y” comes along the same time and updates all records with “Value=5” to “Value=2”.

- **Step3 :-** Finally the "Value=2" is saved in database which is inconsistent according to "Mr. X" as he thinks all the values are equal to "2".

(Q) What are different levels of granularity of locking resources?

Extent:-Extent is made of one or more pages. So all pages are locked and data inside those pages are also locked.

Page: - Page lock puts lock on all data, table and indexes in the page.

Database:-If you are making database structure changes then whole database will be locked.

Table:-We can also lock object at a table level. That means indexes related to it also are locked.

Key: - If you want to lock a series a key of indexes, you can place lock on those group of records.

Row or Row Identifier (RID:-This is the lowest level of locking. You can lock data on a row level.

(Q) What are different types of Locks in SQL Server?

Below are the different kinds of locks in SQL Server:-

- **Shared Locks (S):** - These types of locks are used while reading data from SQL Server. When we apply a Shared lock on a record, then other users can only read the data, but modifying the data is not allowed. Other users can add in new records to the table but can not modify the row which has shared lock applied to it.
- **Exclusive Locks (X):-** These types of lock are not compatible with any other type of locks. As the name suggests any resource, which is having exclusive, locks will not allow any locks to take over it. Nor it can take over any other type of locks. For instance if a resource is having "Shared" lock on a resource you cannot make an "Exclusive lock" over the resource. They are specially used for "Insert", "Update" and "Delete" operations.
- **Update Locks (U):-** "Update" locks are in a mid-level between "Shared" and "Exclusive" locks. When SQL Server wants to modify data and later promote the "Update" locks to "Exclusive" locks then "Update" locks are used. "Update" locks are compatible with "Shared" locks.Ok just to give a brief of how the above three locks will move in actual environment.

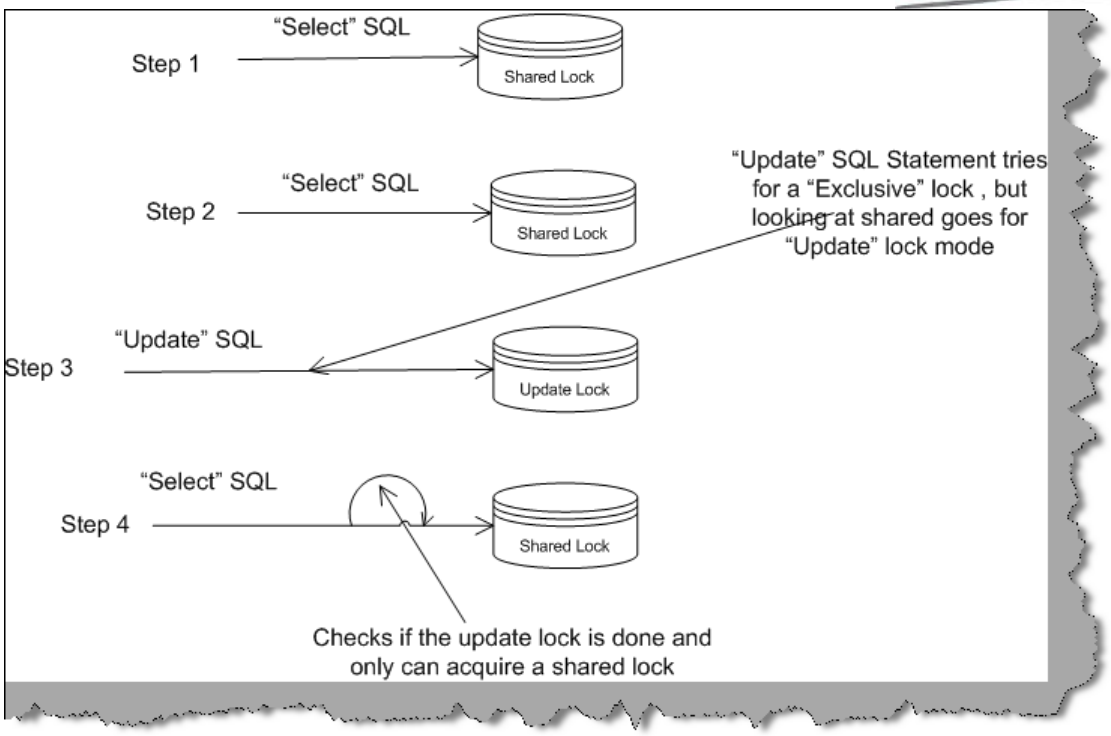


Figure 13.8: - Different Lock sequence in actual scenarios

- **Step1:-** First transaction issues a "SELECT" statement on the resource, thus acquiring a "Shared Lock" on the data
- **Step2:-** Second transaction also executes a "SELECT" statement on the resource, which is permitted as "Shared" lock is honored by "Shared" lock.
- **Step3:-** Third transaction tries to execute an "Update" SQL statement. As it's a "Update" statement it tries to acquire an "Exclusive". But because we already have a "Shared" lock on it, it acquires a "Update" lock.
- **Step4:-** The final transaction tries to fire "Select" SQL on the data and try to acquire a "Shared" lock. But it can not do until the "Update" lock mode is done.

So first "Step4" will not be completed until "Step3" is not executed. When "Step1" and "Step2" is done "Step3" make the lock in to "Exclusive" mode and updates the data. Finally, "Step4" is completed.

- **Intent Locks: -** When SQL Server wants to acquire a "Shared" lock or an "Exclusive" lock below the hierarchy you can use "Intent" locks. For instance one of the transactions has acquired as table lock and you want to have row level lock you can use "Intent" locks. Below are different flavors of "Intent" locks but with one main intention to acquire locks on lower level:-
    - Intent locks include:

- Intent shared (IS)
- Intent exclusive (IX)
- Shared with intent exclusive (SIX)
- Intent update (IU)
- Update intent exclusive (UIX)
- Shared intent update (SIU)

- **Schema Locks:** - Whenever you are doing any operation, which is related to "Schema" operation, this lock is acquired. There are basically two types of flavors in this :-
  - **Schema modification lock (Sch-M):-** Any object structure change using ALTER, DROP, CREATE etc will have this lock.
  - **Schema stability lock (Sch-S)** – This lock is to prevent "Sch-M" locks. These locks are used when compiling queries. This lock does not block any transactional locks, but when the Schema stability (Sch-S) lock is used, the DDL operations cannot be performed on the table.
  - **Bulk Update locks:-**Bulk Update (BU) locks are used during bulk copying of data into a table. For example when we are executing batch process in mid-night over a database.
  - **Key-Range locks:** - Key-Range locks are used by SQL Server to prevent phantom insertions or deletions into a set of records accessed by a transaction.

Below are different flavors of "Key-range" locks

- RangeI_S
- RangeI_U
- RangeI_X
- RangeX_S
- RangeX_U

(Q) What are different Isolation levels in SQL Server?

Twist: - What is an Isolation level in SQL Server?

Locking protects your data from any data corruption or confusion due to multi-user transactions. Isolation level determines how sensitive are your transaction in respect to other transactions. How long the transaction should hold locks to protect from changes done by other transactions. For example if you have long exclusive transaction, then other transactions who want to take over the transaction have to wait for quiet long time. So, isolation level defines the contract between two transactions how they will operate and honor each other in SQL Server. In short how much is on transaction isolated from other transaction.

(Q) What are different types of Isolation levels in SQL Server?

Following are different Isolation levels in SQL Server:-

- **READ COMMITTED**
- **READ UNCOMMITTED**
- **REPEATABLE READ**
- **SERIALIZABLE**

```
Note: - By default SQL Server has "READ COMMITTED" Isolation level.
```

Read Committed

Any "Shared" lock created using "Read Committed" will be removed as soon as the SQL statement is executed. So if you are executing several "SELECT" statements using "Read Committed" and "Shared Lock", locks are freed as soon as the SQL is executed.

However, when it comes to SQL statements like "UPDATE / DELETE AND INSERT" locks are held during the transaction.

With "Read Committed" you can prevent "Dirty Reads" but "Unrepeatable" and "Phantom" still occurs.

Read Uncommitted

This Isolation level says, "Do not apply any locks". This increases performance but can introduces "Dirty Reads". So why is this Isolation level in existence?. Well sometimes when you want that other transaction do not get affected and you want to draw some blurred report , this is a good isolation level to opt for.

Repeatable Read

This type of read prevents "Dirty Reads" and "Unrepeatable reads".

Serializable

It is the king of everything. All concurrency issues are solved by using "Serializable" except for "Lost update". That means all transactions have to wait if any transaction has a "Serializable" isolation level.

```
Note: - Syntax for setting isolation level:-
```

```
SET TRANSACTION ISOLATION LEVEL <READ COMMITTED|READ
UNCOMMITTED|REPEATABLE READ|SERIALIZABLE>
```

(Q) If you are using COM+ what "Isolation" level is set by default?

In order to maintain integrity COM+ and MTS set the isolation level to "SERIALIZABLE".

(Q) What are “Lock” hints?

This is for more control on how to use locking. You can specify how locking should be applied in your SQL queries. This can be given by providing optimizer hints. “Optimizer” hints tells SQL Server that escalate me to this specific lock level. Example the below query says to put table lock while executing the SELECT SQL.

```
SELECT * FROM MasterCustomers  WITH (TABLOCKX)
```

(Q) What is a “Deadlock”?

Deadlocking occurs when two user processes have locks on separate objects and each process is trying to acquire a lock on the object that the other process has. When this happens, SQL Server ends the deadlock by automatically choosing one and aborting the process, allowing the other process to continue. The aborted transaction is rolled back and an error message is sent to the user of the aborted process. Generally, the transaction that requires the least amount of overhead to rollback is the transaction that is aborted.

(Q) What are the steps you can take to avoid “Deadlocks”?

Below are some guidelines for avoiding “Deadlocks”:-

- Make database normalized as possible. As more small pieces, the system is better granularity you have to lock which can avoid lot of clashing.
- Do not lock during user is making input to the screen, keep lock time as minimum as possible by good design.
- As far as possible avoid cursors.
- Keep transactions as short as possible. One way to help accomplish this is to reduce the number of round trips between your application and SQL Server by using the stored procedures or keeping transactions with a single batch. Another way of reducing the time a transaction takes to complete is to make sure you are not performing the same reads repeatedly. If you do need to read the same data more than once, cache it by storing it in a variable or an array, and then re-reading it from there.
- Reduce lock time. Try to develop your application so that it grabs locks at the latest possible time, and then releases them at the very earliest time.
- If appropriate, reduce lock escalation by using the ROWLOCK or PAGLOCK
- Consider using the NOLOCK hint to prevent locking if the data being locked is not modified often.
- If appropriate, use as low of isolation level as possible for the user connection running the transaction.
- Consider using bound connections.

(DB) How can I know what locks are running on which resource?

In order to see the current locks on an "object" or a "process" expand the management tree and right click on "Activity" tab. So in case you want to see "dead locks" or you want to terminate the "dead lock" you can use this facility to get a bird-eye view.

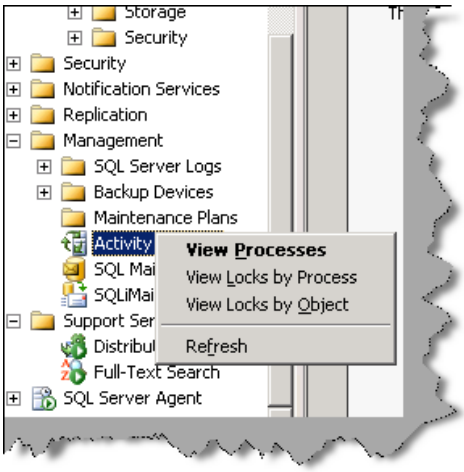


Figure 13.9: - View current locks in SQL Server

What is the use of SQL Server governor?

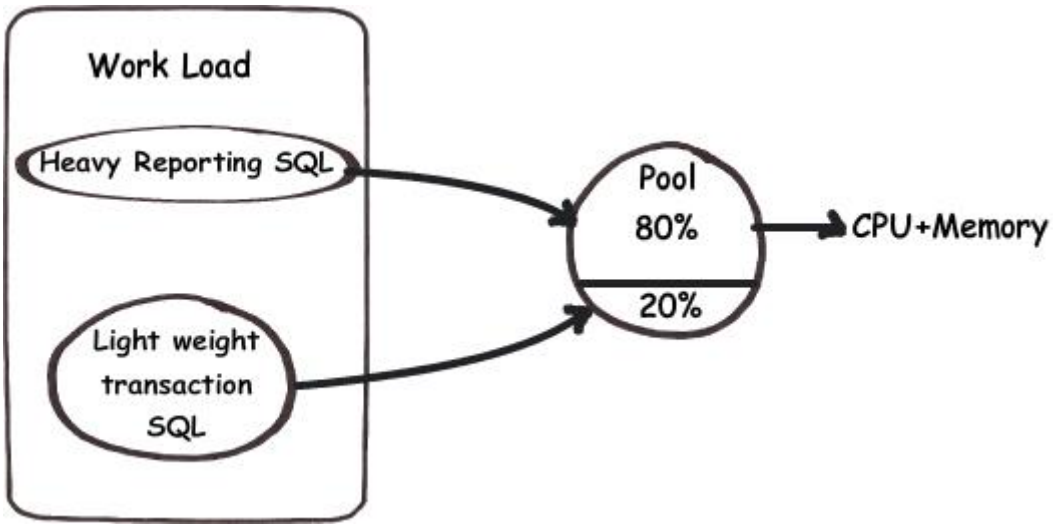
SQL Server is a giant processing engine which processes various kinds of workloads like SQL queries, transactions etc. In order to process these workloads appropriate CPU power and RAM memory needs to be allocated.

Now workloads are of different nature some are light workloads while some are heavy. You would never like that heavy SQL operations hijacking the complete CPU and memory resources thus affecting other operations.

So one of the ways to achieve this is by identifying those SQL queries and putting restriction on the maximum CPU and memory resource for those queries. So for example as shown in the below figure if you have some heavy SQL which does reporting you

would like to allocate 80 % of the CPU and memory resources. While for light weight SQL you would like to allocate only 20%.

This is achieved by using SQL Server governor.



See the video :- What is the use of SQL Server governor to see to how configure SQL Server governor?

Others

How to combine table row in to a single column / variable ?

```
declare @Combine varchar(200)
set @Combine=''
SELECT @Combine = @Combine + [Column1] FROM
[Customer].[dbo].[SomeTable]
print @Combine
```

What is hashing?

Hashing is a process of converting string of characters into a shorter fixed-length value or key which represents the original string. It has many uses like encryption (MD5 , SHA1 etc) or creating a key by which the original values can be retrieved easily (Hash tables collections in .NET).

What is CDC (Change data capture) in SQL Server?

Change data capture helps to capture insert, update and delete activity in SQL Server.

How to enable CDC on SQL Server ?

Enabling CDC is a two-step process:-

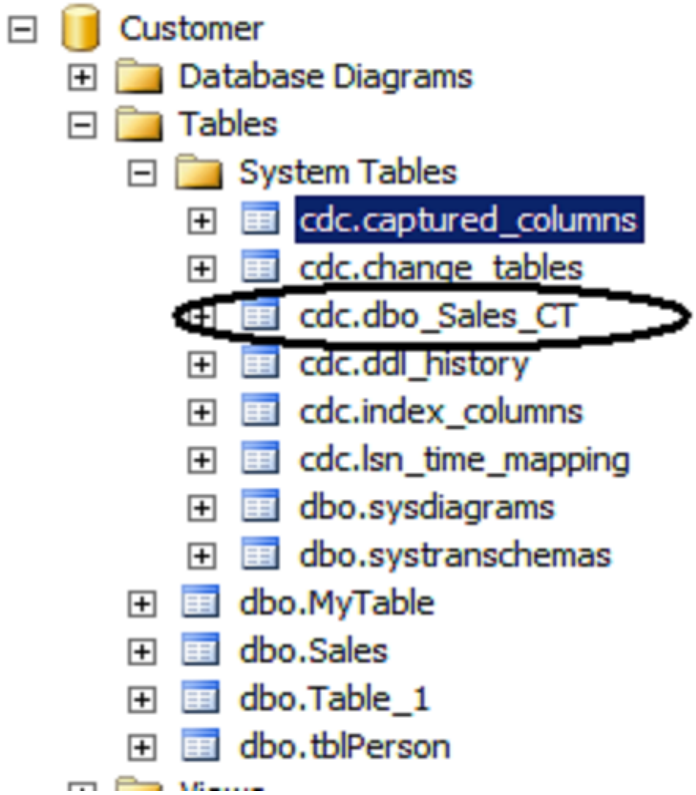
The first step is to fire exec “sp_cdc_enable_db” and enable CDC on database level.

```
EXEC sys.sp_cdc_enable_db
```

Once CDC is enabled on a database level, we need to also specify which tables needs to be enabled for CDC. Below is a simple code snippet which shows how “Sales” table has been enabled for CDC.

```
EXEC sys.sp_cdc_enable_table
@source_schema=N'dbo',
@source_name=N'Sales',
@role_name=NULL
```

Once CDC is enabled, you will find the below tables created for CDC. The most important table is _CT table. For example you can see the below image, for the sales table it has created “dbo_Sales_CT” table.



Now if we modify any data in the "Sales" table the "Sales_CT" table will be affected. After any modification on the "Sales" table, in "Sales_CT" table we will get two rows one with the old value and the other with new value. Below image shows that "Rajendra" has been modified to "Raju" in the "Sales" table.

WIN-BQBERHWW...bo_Sales_CT* | SQLQuery7.sql -...inistrator (66))

	__$operation	Custom...	ProductName	Amount	Vend...
▶	3	Rajendra	Books	119.0000	Archies
	4	Raju	Books	119.0000	Archies

How can we know in CDC what kind of operations have been done on a record?

If you see the _CT table it has a column called as"_$operation". This field will help us identify what kinds of transactions are done with the data. Below are the possible values depending on operation done on the data:-

- Delete Statement = 1

- Insert Statement = 2
- Value before Update Statement = 3
- Value after Update Statement = 4

Will CDC work if SQL Server Agent is not running ?

No, CDC needs SQL Server agent. Without it CDC will not function.