

www.questpond.com

The Software Interview Question Bank

-- Maintained by Shivprasad Koirala shiv_koirala@yahoo.com

Looking for a job but do not know where to start buy my interview question series books from bpb@bol.net.in

Are you looking for a job mail your resume at jobatyourdoortstep@yahoo.co.in

Do not have time to prepare for interview its on head join our one day course at mumbai and feel the confidence call 9892966515 for more details.

Do you have a question which can cost somebody a good job mail me at shiv_koirala@yahoo.com

Do you have a suggestion / tips and tricks which can make job searcher easier mail me at shiv_koirala@yahoo.com.

How to buy the book

BPB has done a great job of making this book reach to places where i can hardly imagine. But just incase its not near to your place mail bpb@bol.net.in.

If you are from India you can contact one of the shops below:-

MUMBAI-22078296/97/022-22070989

KOLKATA-22826518/19

HYDERABAD-24756967,24756400

BANGALORE-25587923,25584641

AHMEDABAD-26421611

BHATINA(PUNJAB)-2237387,

CHENNAI-28410796,28550491

DELHI/NEW DELHI-23254990/91,23325760,26415092,24691288

Pakistan

M/s. Vanguard Books P Ltd, 45 The Mall, Lahore, Pakistan (Tel: 0092-42-7235767, 7243783 and 7243779 and Fax: 7245097)

E-mail: vbl@brain.net.pk

If you are not from india or pakistan :-

Ray McLennan, director,Motilal (UK) Books of India,367 High Street.

London Colney,

St.Albans, Hertfordshire,AL2 1EA, U.K.

Tel. +44 (0)1727 761 677,Fax.+44 (0)1727 761

357,info@mlbduk.com,www.mlbduk.com

If you want to purchase the book directly through BPB Publication's delhi , India :-

bpb@bol.net or bpb@vsnl.com

Contents

How to buy the book	2
From the Author	4
Career Path Institute	5
Data Warehousing/Data Mining	6
What is “Data Warehousing”?	6
What are Data Marts?	6
What are Fact tables and Dimension Tables?	6
What is Snow Flake Schema design in database?	9
What is ETL process in Data warehousing?	10
How can we do ETL process in SQL Server?	11
What is “Data mining”?	11
Compare “Data mining” and “Data Warehousing”?	12
What is BCP?	13
How can we import and export using BCP utility?	14
In BCP we want to change field position or eliminate some fields how can we achieve this?	15
What is Bulk Insert?	17
What is DTS?	18
Can you brief about the Data warehouse project you worked on?	19
What is an OLTP (Online Transaction Processing) System?	20
What is an OLAP (On-line Analytical processing) system?	20
What is Conceptual, Logical and Physical model?	21
What is Data purging?	21
What is Analysis Services?	21
What are CUBES?	22
What are the primary ways to store data in OLAP?	22
What is META DATA information in Data warehousing projects?	23
What is multi-dimensional analysis?	23
What is MDX?	25
How did you plan your Data ware house project?	25
What are different deliverables according to phases?	28
Can you explain how analysis service works?	29
What are the different problems that “Data mining” can solve?	45
What are different stages of “Data mining”?	46
What is Discrete and Continuous data in Data mining world?	49
What is MODEL in Data mining world?	49
DB)How are models actually derived?	50
What is a Decision Tree Algorithm?	50
Can decision tree be implemented using SQL?	52
What is Naïve Bayes Algorithm?	52
Explain clustering algorithm?	53
Explain in detail Neural Networks?	54

What is Back propagation in Neural Networks?	57
What is Time Series algorithm in data mining?	58
Explain Association algorithm in Data mining?	58
What is Sequence clustering algorithm?	58
What are algorithms provided by Microsoft in SQL Server?	58
How does data mining and data warehousing work together?	60
What is XMLA?	61
What is Discover and Execute in XMLA?	62
Distribution Partner	63

From the Author

First thing thanks to all those who have sent me complaints and also appreciation for what ever titles i have written till today. But interview question series is very near to my heart as i can understand the pain of searching a job. Thanks to my publishers (BPB) , readers and reviewers to always excuse all my stupid things which i always do.

So why is this PDF free ?. Well i always wanted to distribute things for free specially when its a interview question book which can fetch a job for a developer. But i am also bounded with publishers rules and regulations. And why not they have a whole team of editor, printing guys, designers, distributors, shopkeepers and including me. But again the other aspect, readers should know of what they are buying , the quality and is it really useful to buy this book. So here are sample free questions which i am giving out free to the readers to see the worth of the book.

I can be contacted at shiv_koirala@yahoo.com its bit difficult to answer all answers but as i get time i do it.

We have recently started a career counselling drive absolutely free for new comers and experienced guys. So i have enlisted the following guys on the panel. Thanks to all these guys to accept the panel job of consulting. Feel free to shoot them questions just put a title in the mail saying “Question about Career”. I have always turned up to them when i had some serious career decision to take.

Shivprasad Koirala :- Not a great guy but as i have done the complete book i have to take up one of the positions. You can contact me at shiv_koirala@yahoo.com for technical career aspect.

Tapan Das :- If you think you are aiming at becoming a project manager he is the right person to consult. He can answer all your questions regarding how to groom your career as a project manager tapand@vsnl.com.

Kapil Siddharth :- If you are thinking to grow as architect in a company then he is a guy. When it comes to role model as architect i rate this guy at the top. You can contact him at kapilsiddharth@hotmail.com

Second if you think you can help the developers mail me at shiv_koirala@yahoo.com and if i find you fitting in the panel i will display your mail address. Please note there are no financial rewards as such but i am sure you will be proud of the work you are doing and whos knows what can come up.

Lets make Software Industry a better place to work Happy Job Hunting and Best of Luck

Career Path Institute

Author runs the “Softwar Career Path Insitute” personally in mumbai. If you are interested you can contact him regarding admissions at shiv_koirala@yahoo.com. Our courses are mainly targetting from how to get a job perspective.

Below are some of the courses offered :-

- Interview preparation course two days (Saturday and Sunday Batch). (C# , SQL Server)
- Full one year course for C# , SQL Server

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.

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.

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 which are spanned across the globe. Head-office of the company decides to collect data from all these branches for anticipating market. So 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: - Data Mart in action

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. But 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 which will be needed in a business process. While dimension tables revolve around fact tables. They describe the attributes of the fact tables. Let's try to understand these two conceptually.

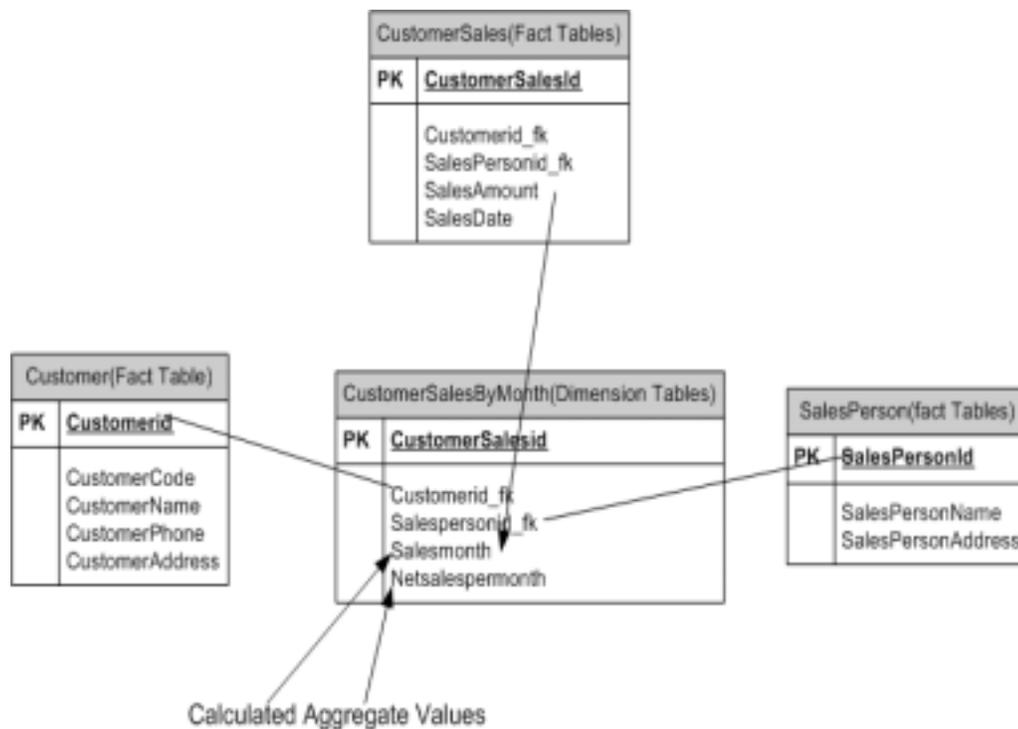


Figure 8.2 : - Dimensional Modeling

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. But 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 design is also called 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's the difference between Star and Snow flake schema?

Star schema is good when you do not have big tables in data warehousing. But 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.

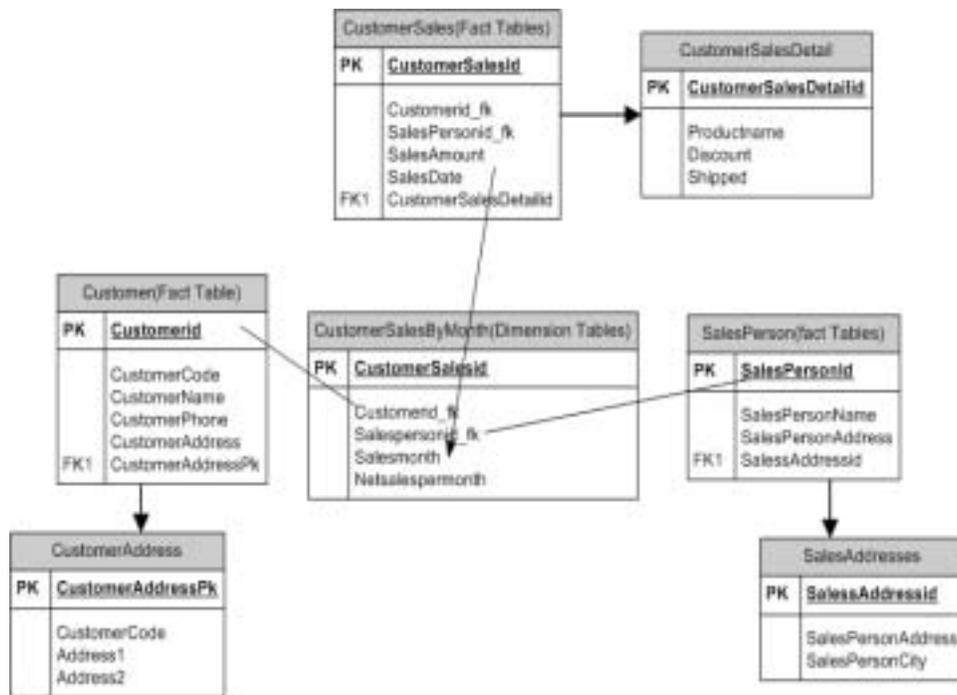


Figure 8.4 : - Snow Flake Schema

(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. So extraction process handle's 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's 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. So you can have some kind of stored procedure which 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.

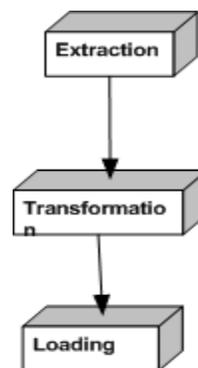


Figure 8.5 : - ETL stages

(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.

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’s 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?

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’s 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 quite 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”.

What is BCP?

Note: - It's 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's where this old and cryptic friend will come to use. It's 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. But 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.

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 which are used.

-T: - signifies that we are using windows authentication

-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.

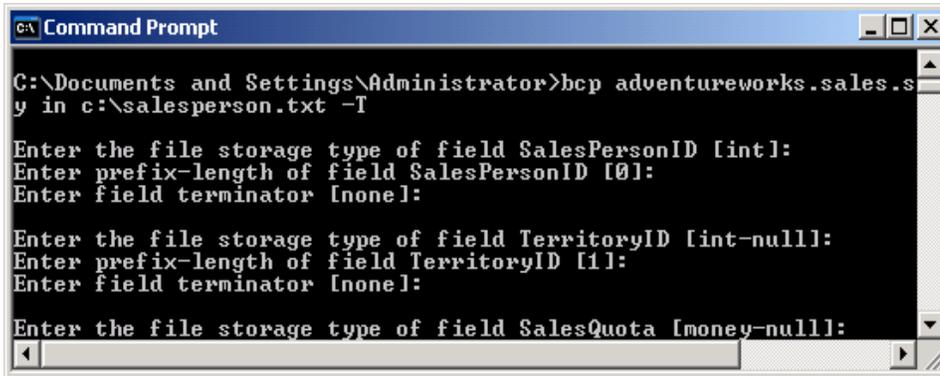


Figure 8.7 : - After executing BCP command prompts for some properties

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. Or you want the sequence to be changed as field2, field1 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.

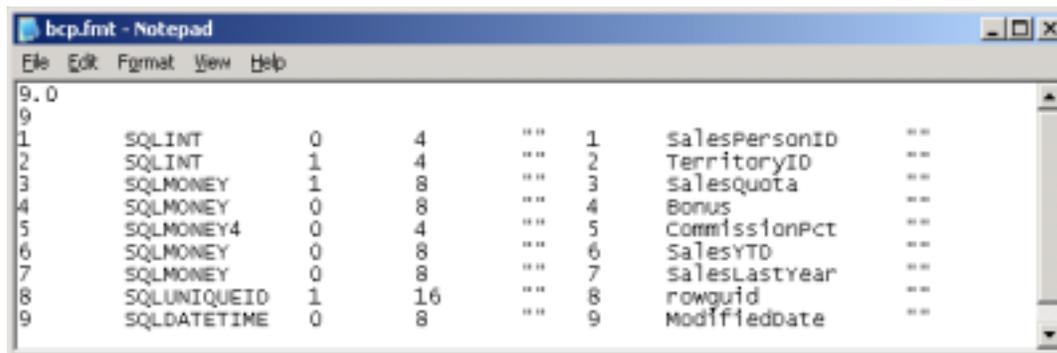


Figure 8.8 : - Format file generated due to BCP.

FMT file is basically the format file for BCP to govern how it should map with tables. Lets say, in from our salesperson table we want to eliminate commissionpct, salesytd and saleslastyear. So you have to modify the FMT file as shown below. We have made the values zero for the fields which has to be eliminated.

```

9.0
9|
1      SQLINT      0      4      ""      1      SalesPersonID  ""
2      SQLINT      1      4      ""      2      TerritoryID    ""
3      SQLMONEY    1      8      ""      3      SalesQuota     ""
4      SQLMONEY    0      8      ""      4      Bonus          ""
5      SQLMONEY4   0      0      ""      0      CommissionPct  ""
6      SQLMONEY    0      0      ""      0      SalesYTD       ""
7      SQLMONEY    0      0      ""      0      SalesLastYear  ""
8      SQLUNIQUEID 1      16     ""      8      rowguid        ""
9      SQLDATETIME 0      8      ""      9      ModifiedDate   ""

```

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.

```

9.0
9|
1      SQLINT      0      4      ""      1      SalesPersonID  ""
2      SQLINT      1      4      ""      2      TerritoryID    ""
3      SQLMONEY    1      8      ""      3      SalesQuota     ""
4      SQLMONEY    0      8      ""      4      Bonus          ""
5      SQLMONEY4   0      4      ""      9      CommissionPct  ""
6      SQLMONEY    0      8      ""      8      SalesYTD       ""
7      SQLMONEY    0      8      ""      7      SalesLastYear  ""
8      SQLUNIQUEID 1      16     ""      6      rowguid        ""
9      SQLDATETIME 0      8      ""      5|      ModifiedDate   ""

```

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.

What is Bulk Insert?

Bulk insert is very similar to BCP command but we can not 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 “SalesPersonDummy”. 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'
)
```

What is DTS?

Note :- It's 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.

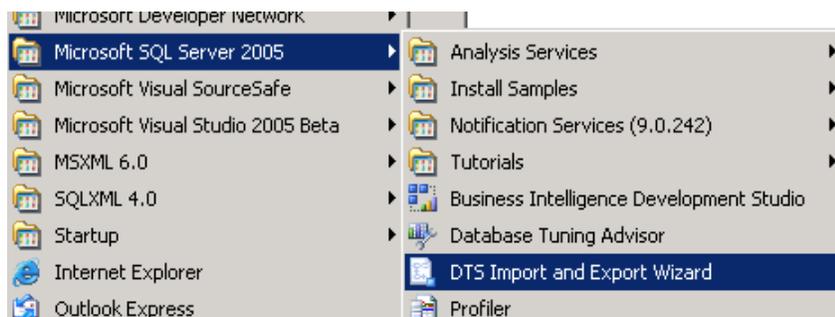


Figure 8.11 : - DTS Import Export

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.

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.

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 which is executed during night hours.
- √ As OLAP systems are mainly used for reporting or batch processing, so “Denormalization” designs are encouraged.

-
- √ Transactions are mainly batch transactions which are running so there are no huge volumes of transaction.
 - √ Do not need to have recovery process as such until the project specifies specifically.

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's 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's 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's 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. So this cleaning process is called as "Data Purging".

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 to 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?

There are primary three ways in which we store information in OLAP:-

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 doesn't 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's 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's not an enough definition for interviews we need something more than that to tell the interviewer. It's the complete documentation of a data warehouse project. From perspective of SQL Server all Meta data is stored in Microsoft repository. It's all about way the structure is of data ware house, 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.

But let's 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's where MDX queries are useful. MDX queries are fired against OLAP data bases. 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 ware house 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 loses 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 work flow, 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. But 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 which 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 bottle neck for optimization. So you can probably look for loops and you also 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?”. It’s a testing phase which can be done either 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 which was gathered during the requirement phase.

- √ Rolling out to Production

Once the customer has approved your UAT, its time to roll out the data ware house 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.

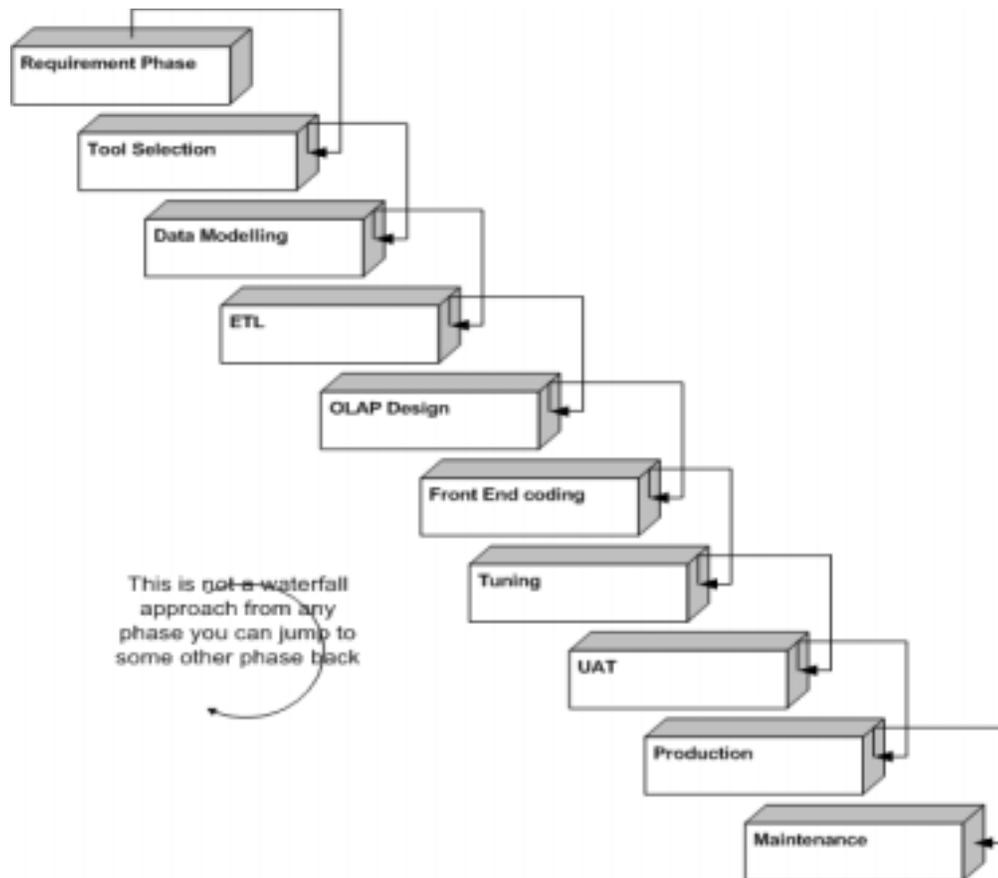


Figure 8.14 : - Data ware house project life cycle

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 eye balls 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 which 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 “NorthWind” database for showing analysis server demo.

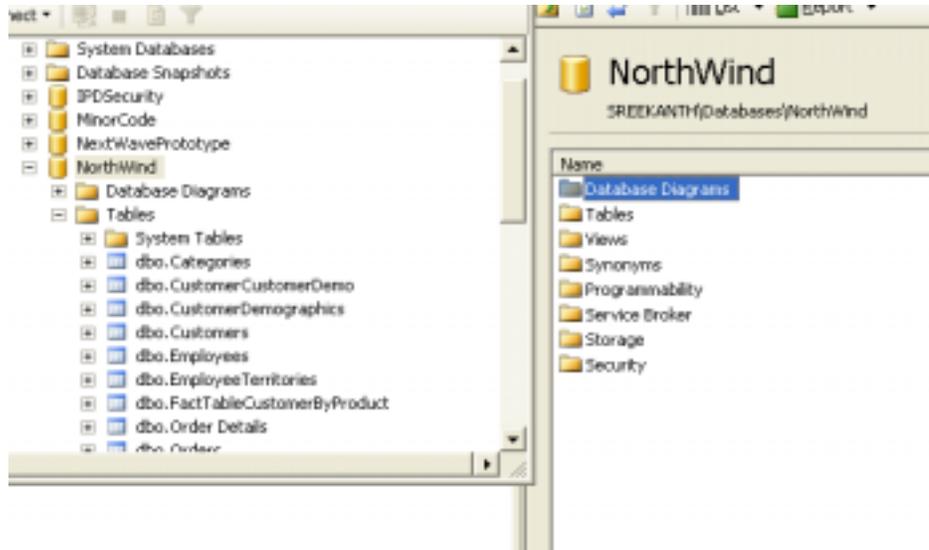


Figure 8.16 : - NorthWind Snapshot.

We are not going to use all tables from “NorthWind”. 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 on 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 TotalProducts sold. All the data in the Fact table I have loaded from “Orders” and “Order Details”. Means I have taken all customerid and productid with their respective totals and made entries in the Fact table.

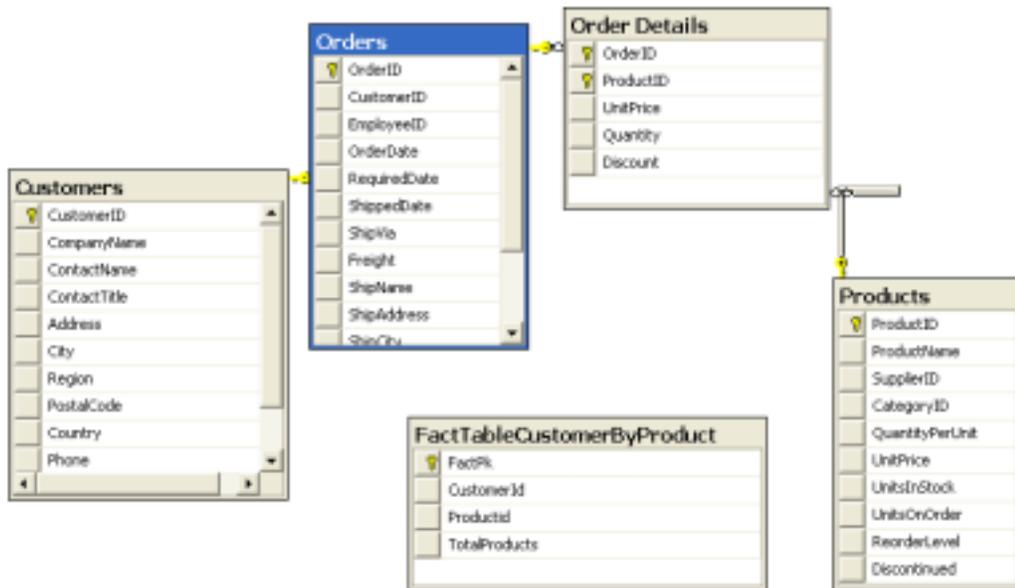


Figure 8.17: - Fact Table

Ok I have created my fact table and also populated using our ETL process. Now its time to use this fact table to do analysis.

So let's start our BI studio as shown in figure below.

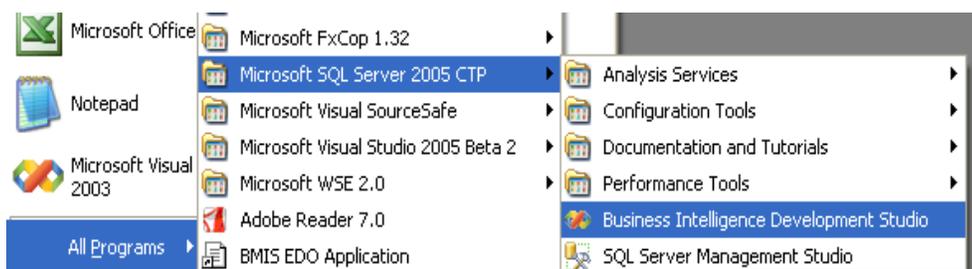


Figure 8.18 : - Start the Business Development Studio

Select "Analysis" project from the project types.

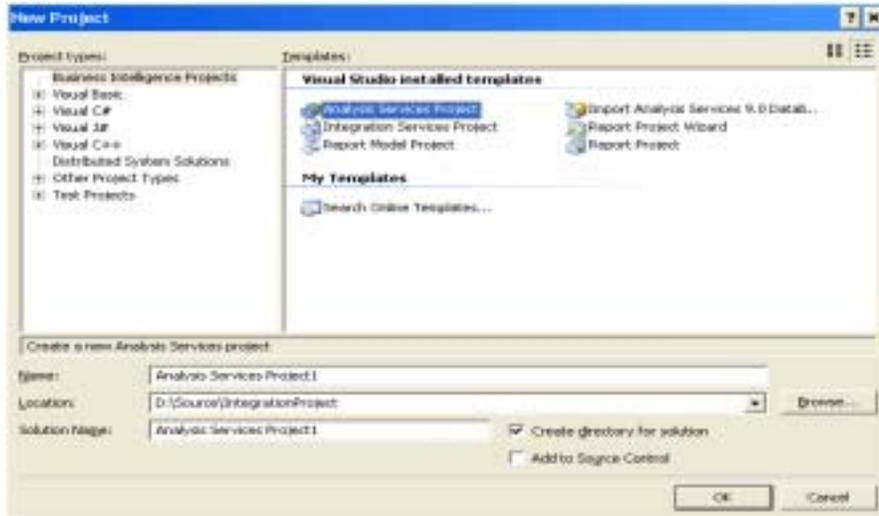


Figure 8.19 : - Select Analysis Services Project

I have name the project as “AnalysisProject”. You can see the view of the solution explorer.
Data Sources :- This is where we will define our database and connection.

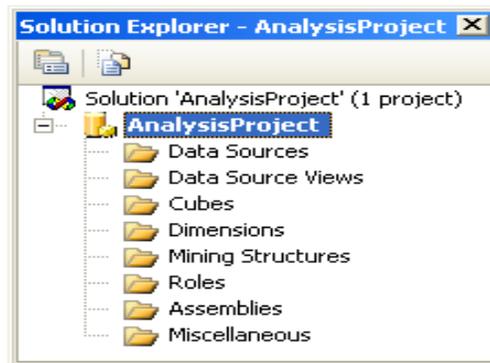


Figure 8.20 : - Solution Explorer

To add a new “data Source” right click and select “new Data Source”.

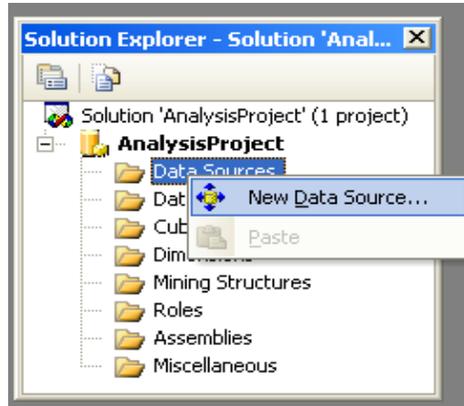


Figure 8.21 : - Create new data Source

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.

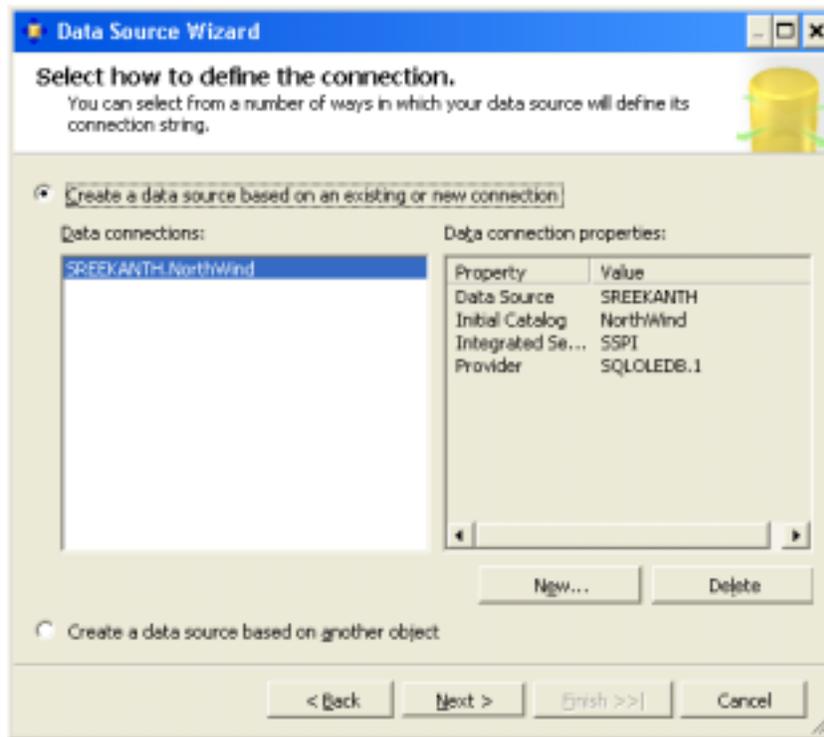


Figure 8.22 : - Define Data source connection details

After that its time to define view.

Data Source View: - It's an abstraction view of data source. Data source is the complete database. It's rare that we will need the complete database at any moment of time. So 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.

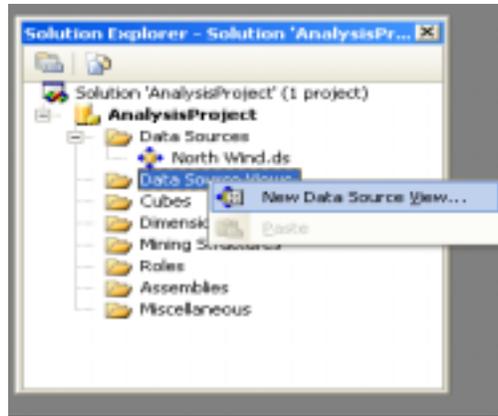


Figure 8.23 : - Create new Data source view

So here we will select only two tables “Customers”, “Products” and the fact table.

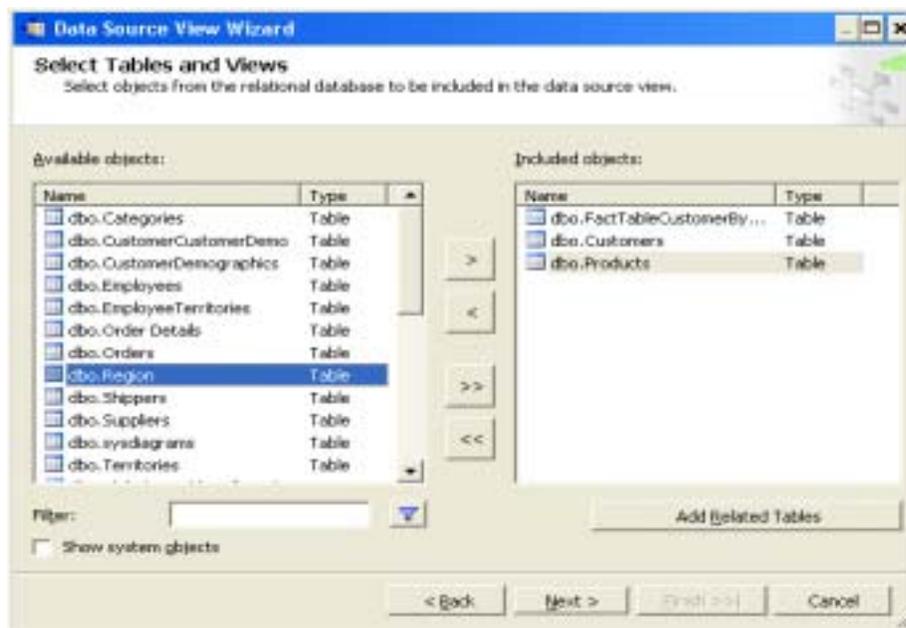


Figure 8.24 : - Specify tables for the view

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.

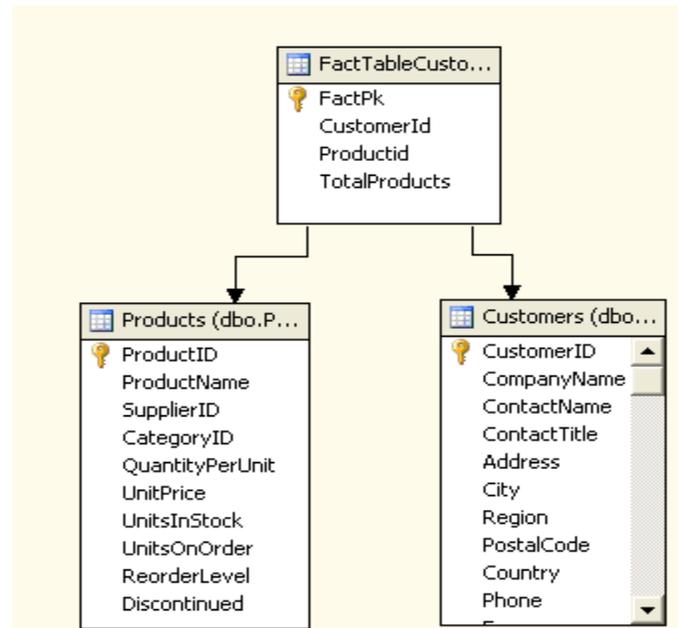


Figure 8.25 : - Final Data Source view

Check “Autobuild” as we are going to let the analysis service decide which tables he want to decide as “fact” and “Dimension” tables.

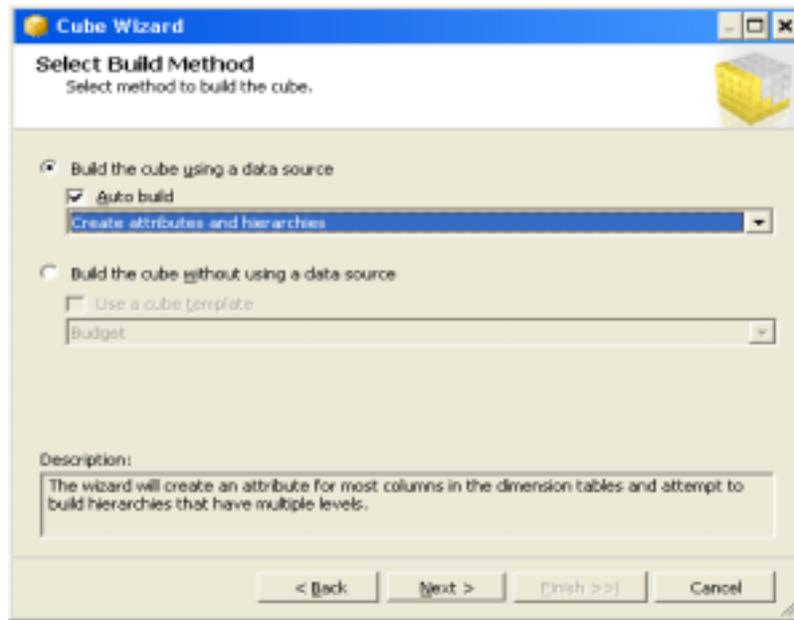


Figure 8.26 : - Check Auto build

After that comes the most important step which are the fact tables and which are dimension tables. SQL Analysis services decides by itself, but we will change the values as shown in figure below.

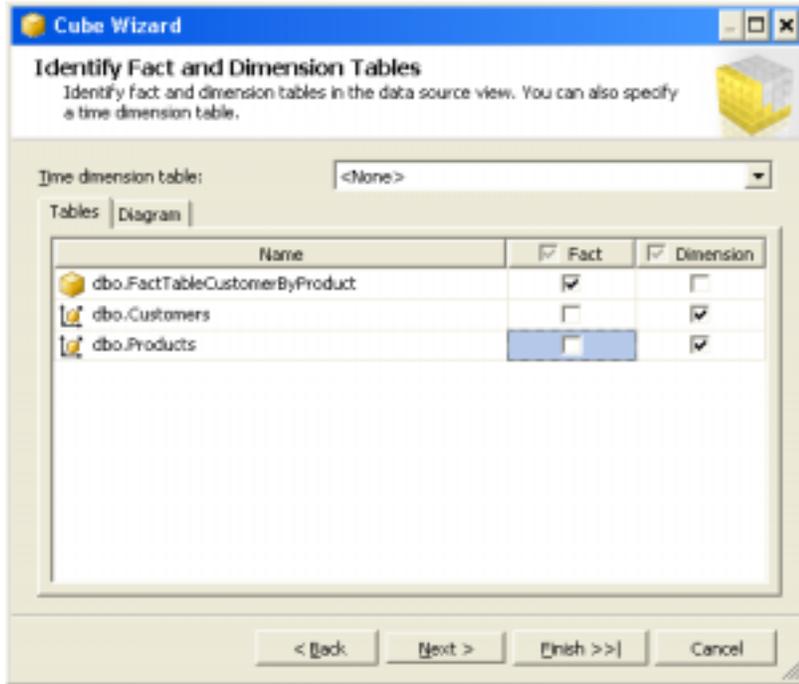


Figure 8.27 : - Specify Fact and Dimension Tables

This screen defines measures.

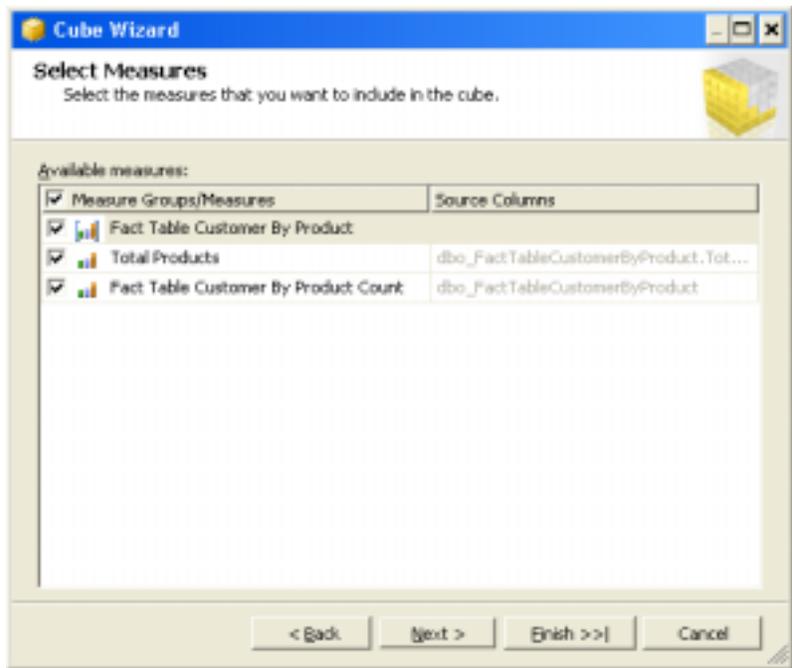


Figure 8.28 : - Specify measures

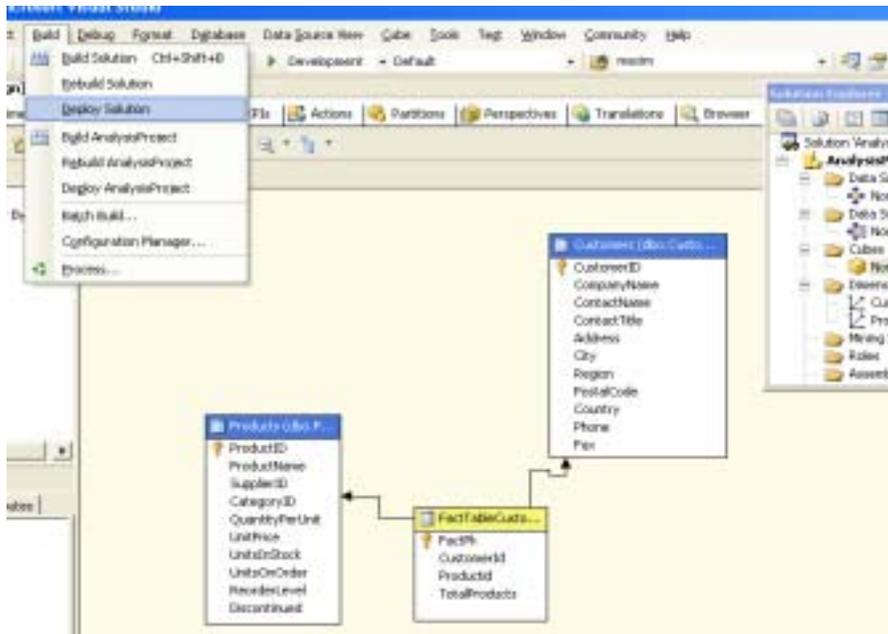


Figure 8.29 : - Deploy Solution

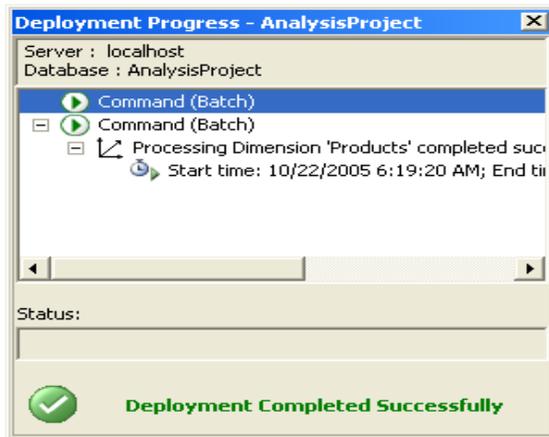


Figure 8.30 : - Deployment Successful

- √ 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.

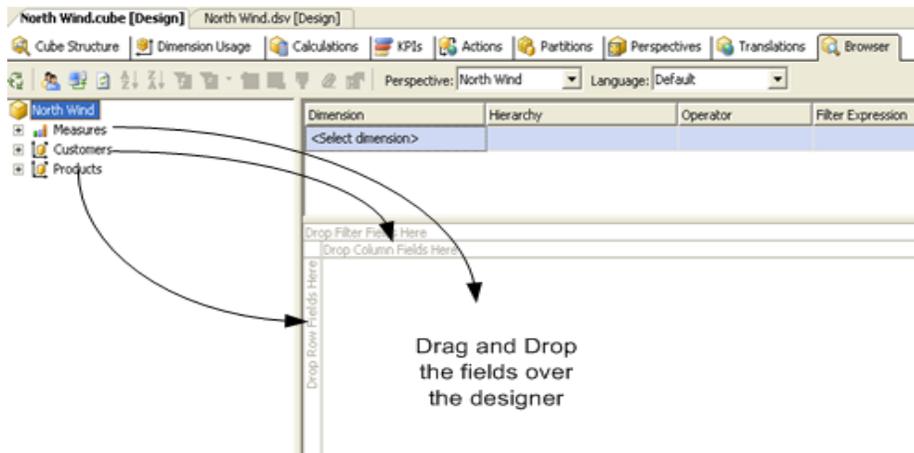


Figure 8.32: - Drag and Drop the fields over the designer



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.

Dimension	Hierarchy	Operator	Filter Expression
<Select dimension>			

Drop Filter Fields Here										
	Products ▾									
	Alice Mutton	Aniseed Symp	Boston Crab Meat	Canenberg	Remat	Carmanon	Tiger's Cha	Chang	Charboux verbe	Chief A
Customers ▾	Total Products	Total Products	Total Products	Total Products	Total Products	Total Products	Total Products	Total Products	Total Products	Total P
ALFKI		1							1	
ANATR				1						
ANTON	1		1					1		
AROUT				1				1		
BERGS	1	1	2	2			1	1	2	1
BLAUS				1	1				1	
BOLID	1					1	1		1	
BONAP	1		2			2			1	1
BOTTN	2	1	1	2			1			
BSBEV		1	1							
CACTU										
CENIC										
CHOPS				1			1	1		
COMM										
CONSH								1		
DRACD										
DUMON	1							1		
EASTC				1			1			1
ERNGH	4	2	1	2	1			2	2	
FAMSA			1	1					1	
FOUJG						2				
FOUJL								3	2	1
FRANK			1	2				1		
FRAMR										
FRANS			1	1	1					

Figure 8.34: - Product and Customer Report

This is the second report which says in which country I have sold how many 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 were 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 were some resignations from fresher. So the HR made decision to motivate the freshers rather than 2 years completed employee's to retain people. As HR thought it's easy to motivate freshers 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.

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 check list:-

- √ 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

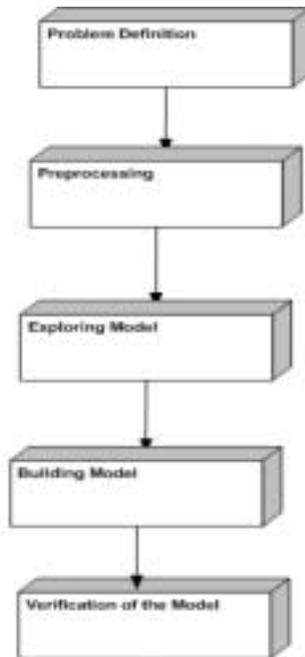
Data derived from Exploring models will help us to define and create a mining model. A model typically contains input columns, an identifying column, and a predictable column. You can then define these columns in a new model by using the Data Mining Extensions (DMX) language or the Data Mining Wizard 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. This is 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.



Note :- We can move from any of down process to upper process it's a iterative model rather than waterfall model. Did not make the arrow direction just to avoid confusion. So you can move from building model to problem definition.

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.

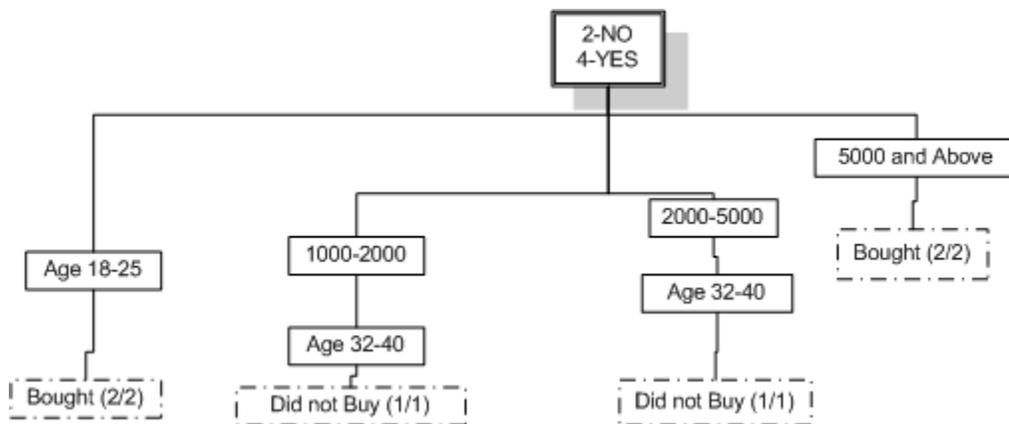


Figure 8.38 : - First Iteration Decision Tree

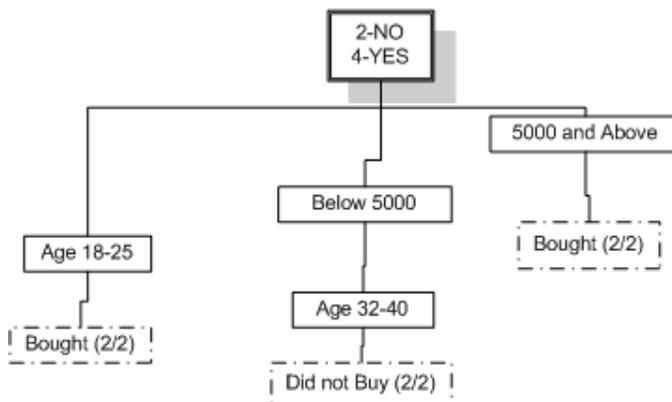


Figure 8.39 : - Conclusion from the Decision Tree

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.

So 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's a difficult things to understand lets make it simple. Let's 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.

$$P(\text{Shirt} | \text{Pants}) = 0.8$$

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.

$$P(\text{Shirts}) = 0.1$$

$$P(\text{Pants}) = 0.01$$

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:-

$$\text{Probability of buying shirt if bought pants} = 0.8 \cdot 0.01 / 0.1 = 7.9$$

$$\text{Probability of buying pants if bought shirts} = 0.8 \cdot 0.1 / 0.01 = 70$$

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's what human tried to devise neural network. So now you must be thinking how it works.

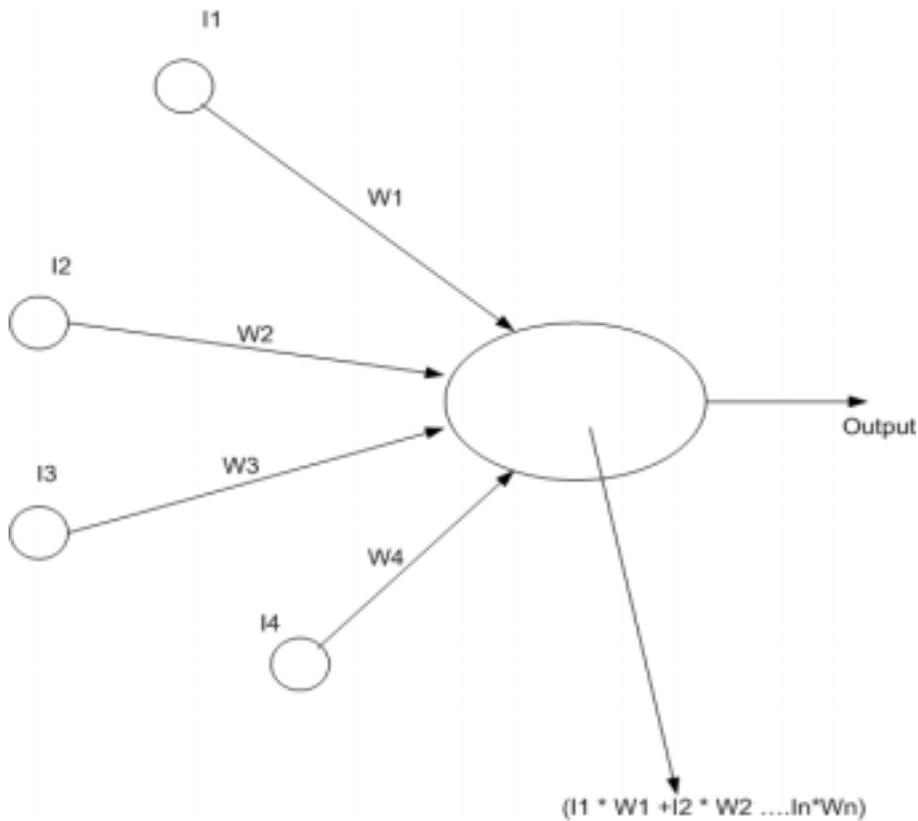


Figure 8.41 : - Artificial Neuron Model

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.

$$\text{Activation Value} = I1 * W1 + I2 * W2 + I3 * W3 + I4 * W4 + \dots + IN * WN$$

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.

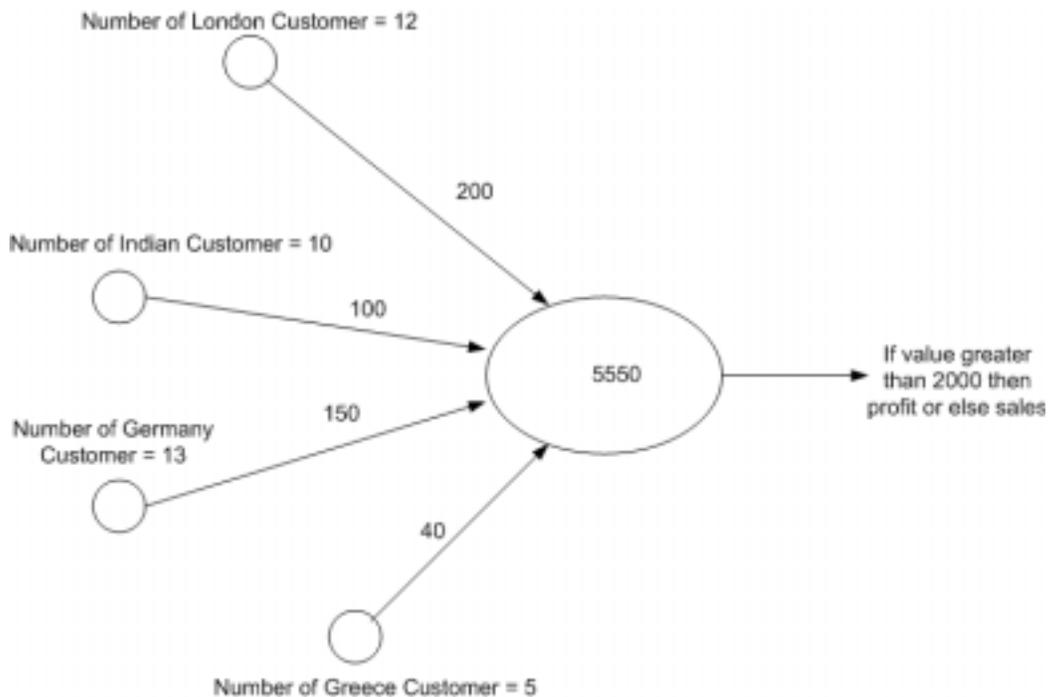


Figure 8.43: - Neural Network for Customer Sales Data

You can see neuron has calculated the total as 5550 and as it's greater than threshold 2000 we can say the company is under profit.

The above example was explained for simplification point of view. But 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.

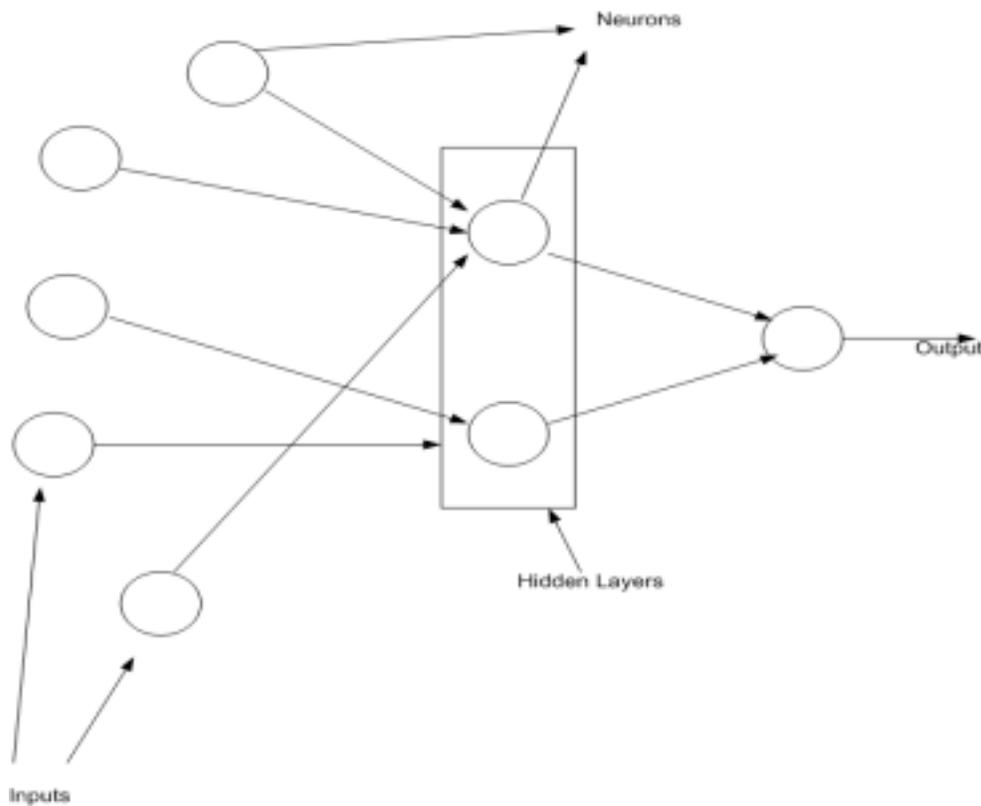


Figure 8.44: - Practical Neural Network

(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.

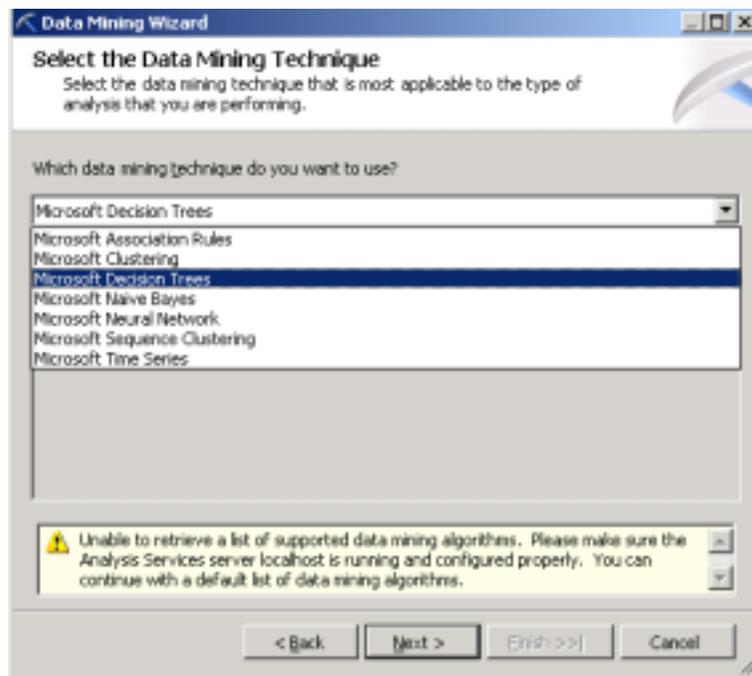


Figure 8.45: - Snapshot of the algorithms in SQL Server

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 freshers). Below is the big picture which shows the relation between “data warehousing” and “data mining”.

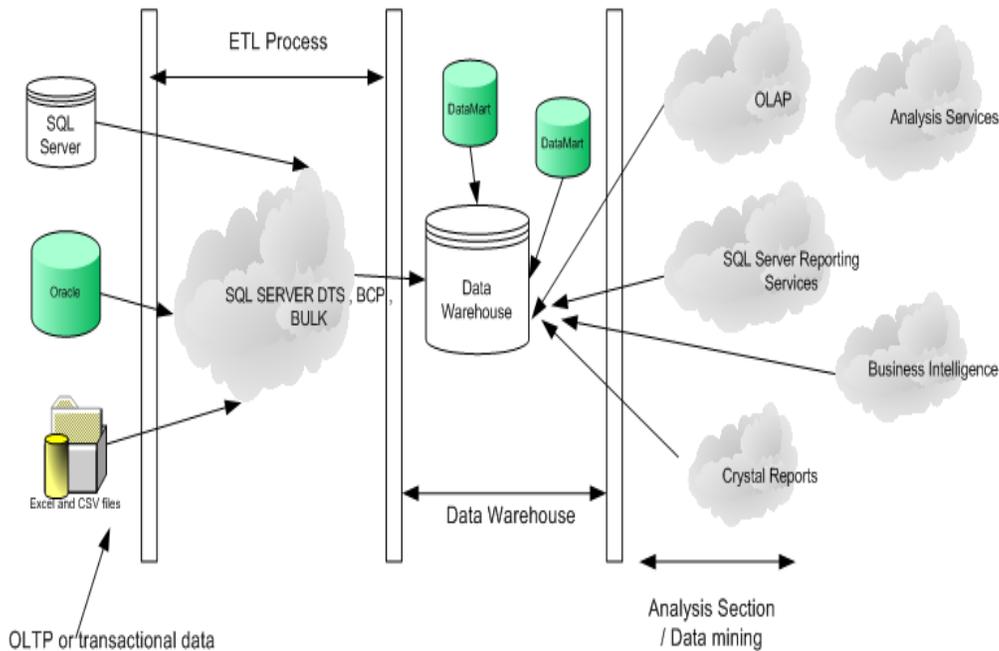


Figure 8.46 : - Data mining and Data Warehousing

Let's 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.

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.

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.

Distribution Partner

Do you have a news group or website where you want to distribute this PDF free. Contact at shiv_koirala@yahoo.com we will put your logo and send you a complete zipped file which you can host at your site to increase user visits. Be our partner and increase your user visits in your website. We do not take any charge from our partner to distribute these sample copies. But yes the contents of this PDF can not be modified. If you are our partner you get regular updates of our interview question releases.

Illegal distributions of this PDF will be taken seriously. Just send a mail and be our distribution partner with your website logo proudly do not distribute illegally.

www.questpond.com