Q.1. Explain backroom architecture model.

→

fig from ppt (page no - 36).

A] Source systems -
It is rare data warehouse at the enterprise level that does not pull data from multiple sources.

① Enterprise Resource Planning (ERP) system -
The ERP systems are made up of dozens of modules that cover major functional areas of business, such as order entry, human resources, purchasing & manufacturing.

② Operational Data stores (ODS) -
③ Master Data management (MDM) system
④ Message queues, log files & redo files
⑤ proprietary formats.
⑥ external suppliers.

B] Extract

① data profiling -
Data profiling is the process of reviewing source data, understanding structure, content & interrelationships & identifying potential for data projects.

② change data capture -
change data capture quickly identifies & processes only the data that has changed, not entire tables & makes the change data available for further use.

③ extract system

C] clean & conform -
① data cleansing -
② error event tracking
③ audit dimension creation
④ depulicating
⑤ conforming.

D] Deliver -
① hierarchy manager
② surrogate key pipeline
③ surrogate key generator
④ ~~aggrega~~ fact table builders.

E] ETL management services -
① job scheduler
② backup system
③ recovery & restart
④ version control
⑤ workflow migration
⑥ workflow monitor
⑦ sorting
⑧ lineage & dependency.

F] additional backroom services & trends -
① data service providers
② functional service providers
③ data delivery services
④ ETL data stores
⑤ ETL system data stores
⑥ data quality data stores.

G] ETL metadata -
① process metadata - ETL operations statistic
- audit results

PAGE NO.:
DATE : / /

- quality screen results.

② technical metadata - system inventory
including version numbers
- source descriptions
- source access methods
- ETL job logic, extract & transforms
- Exception handling logic
- Processing schedules.
- batch parameters.

③ business metadata - data dictionary
- logical data map
- business rule ~~to~~ logic
- data quality screen specifications.

Q.2. Explain technical, process & business metadata in DWBI architecture.

→ A] Technical ~~me~~ metadata -

① It includes the system metadata that defines the data structures themselves like tables, fields, data types, indexes & partitions in the relational engine & databases, dimensions, measures & data mining models.

② In the ETL process, technical metadata defines the sources & targets for a particular task, the transformations & their frequency.

③ Some technical metadata elements are ~~the~~ useful for the business users like tables & columns names.

④ It describes the structure, format & rules for storing data.

B] Process metadata -

① It describes the results of various operations in the warehouse.

② In the ETL process, each task logs key data about its execution such as start time, end time, CPU seconds used, disk reads, disk writes & rows processed.

③ Similar process metadata is generated when users query the warehouse. This data is initially valuable for troubleshooting ETL or query process.

④ Business people at an information provider

would be interested in analysing this process data because it tells them who is using their products, what products they are using & what service level they are receiving.

c] Business metadata-

① It describes the contents of data warehouse in more user accessible terms.

② It tells you what data you have, where it comes from, what it means & what is its relationship to other data in warehouse.

③ Ese - display name, content description fields.

Q.13. Explain presentation server metadata.

→ In our technical architecture, the ETL metadata contains all the processes & declarations necessary to populate the presentation server. At the other end, the BI application metadata contains all the processes & declarations necessary to fetch data from the presentation server on behalf of the BI environment. That leaves relatively little metadata uniquely owned by presentation server.

a] Process metadata-

① database monitoring system tables - contain info about the usage of tables.

PAGE NO.:
DATE : / /

② aggregate usage statistics - include OLAP usage.

b] technical metadata-

① database system tables - contain std RDBMS table, column, view, index.

② partition setting - include partition definition.

③ stored procedures & SQL scripts - for creating partitions, indexes & aggregates

④ OLAP system definitions - contain system info specific to OLAP databases.

⑤ Target data policies & procedures - include retention, backup, archives, recovery & ownership.

c] business metadata-

Business metadata regarding the presentation server is provided by the BI application semantic layer, the OLAP definitions or the database system table & column definitions directly.

PAGE NO.:
DATE : / /

Q.4. Discuss in brief parallel processing hardware architecture.

→ There are three basic parallel processing hardware architectures in the server market-
① symmetric multiprocessing (SMP)
② massively parallel processing (MPP)
③ non-uniform memory architecture (NUMA).

A] symmetric multiprocessing (SMP) -
The SMP architecture is a single machine with multiple processors all managed by one operating system & all accessing the same disk and memory area. The use of single shared communications channel to communicate bet^ CPU, memory & I/O can become a bottleneck to system performance as the number of CPU & their clock speeds increase.

B] Massively parallel processing (MPP) -
MPP systems are basically a string of relatively independent computers, each with its own operating system, memory & disk all coordinated by passing messages back & forth. The strength of MPP is the ability to connect hundreds of machine nodes together & apply them to a problem using a brute-force approach.
MPP systems are basically found in large scale, multi-terabyte data warehouses.

C] Non-uniform memory architecture (NUMA) -
NUMA is hybrid of SMP & MPP, combining

PAGE NO.:
DATE: / /

the shared disk flexibility of SMP with the parallel speed of MPP. NUMA is similar to clustering SMP machines but with tighter connections, more bandwidth & greater coordination among nodes.

D] Clusters -
It is possible to connect several computers together in a cluster that acts like a single server. There are two major uses for clusters - high availability & extending server capacity, called scale out. The goal of scale out cluster is to provide an environment where multiple systems can concurrently access the same data. The term scale up means adding capacity by getting a bigger machine.

E] shared cache -
When storage is shared among all the machines in a cluster environment, the system must keep track of uncommitted transactions.

F] Federated data -
The federated data environment requires that data is split or federated across multiple nodes and that queries are satisfied by collecting data from multiple nodes & joining the data to satisfy the queries.

G] Replicated data -
The data replication environment requires that data be replicated to different nodes &

PAGE NO.:
DATE: / /

that queries are satisfied by local relational engine.

Q.5. Explain presentation server system architecture in BIS.

→

Fig. from ppt (page no - 46)

① Presentation servers are the target platforms where the data is stored for direct querying by business users, reporting systems & other BI applications.

② Business requirements for info :-
- access to data from all major business processes
- access to both summary & atomic data
- single source for analytical data

③ Detail atomic data :-
- analytic queries require detail or summary data.
- atomic data are built with conformed dimensions as per enterprise bus matrix
- stored in ~~*~~ RDBMS than in OLAP because of data management capabilities, flexibility & broad accessibility that relational databases provide.

④ Aggregates -
- In order to improve ~~performance~~ performance at summary levels, we add the second or element of the presentation server

PAGE NO :
DATE / /

layer :- aggregate.

- Pre-aggregating data during the load process is one of the primary tools available to improve performance for analytic queries.
- Aggregates occupy a separate logical layer.
- Aggregates are like indexes. They will be built & rebuilt on a daily basis; the choice of aggregates will change over time based on analysis of actual query usage.
- usage based optimization.

⑤ Aggregate navigation.

The aggregate navigator receives a user query & examines to see if it can be answered using a smaller, aggregate table. If so, the query is rewritten to work against the aggregate table & submitted to database engine.

Technologies to provide aggregate navigation functionality – ① OLAP engines
② relational OLAP (ROLAP) services
③ BI application servers or query tools.

PAGE NO.
DATE : / /

DWBI - 2

Q.1 Explain use of Enterprise Data Warehouse Bus Architecture in dimensional modeling.

→

① The enterprise data warehouse bus matrix is the overall data architecture for the DWBI system.

② Multiple development teams can work on components of the matrix fairly independently & asynchronously with confidence that ~~that~~ they will fit together like the pieces of a puzzle.

③ Developing the data architecture via the bus matrix is a rational approach to decomposing the daunting task of planning an enterprise data warehouse.

④ A bus is a common structure that everything connects to and derives power from. The bus in computer is a std interface that allows many different kinds of devices to connect.

⑤ Planning crisis -

- The manager is supposed to understand the content & location of most complicated asset owned by enterprise; the source data.
- Every data element in every system must be understood - The DWBI manager must be able to retrieve any requested element of data & if necessary, clean it up & correct it.
- If building the data warehouse all at once is too discouraging and building it as isolated pieces defeats the overall goal, what is to be done?

PAGE NO.:
DATE: / /

- To solve above problem we need bus architecture.

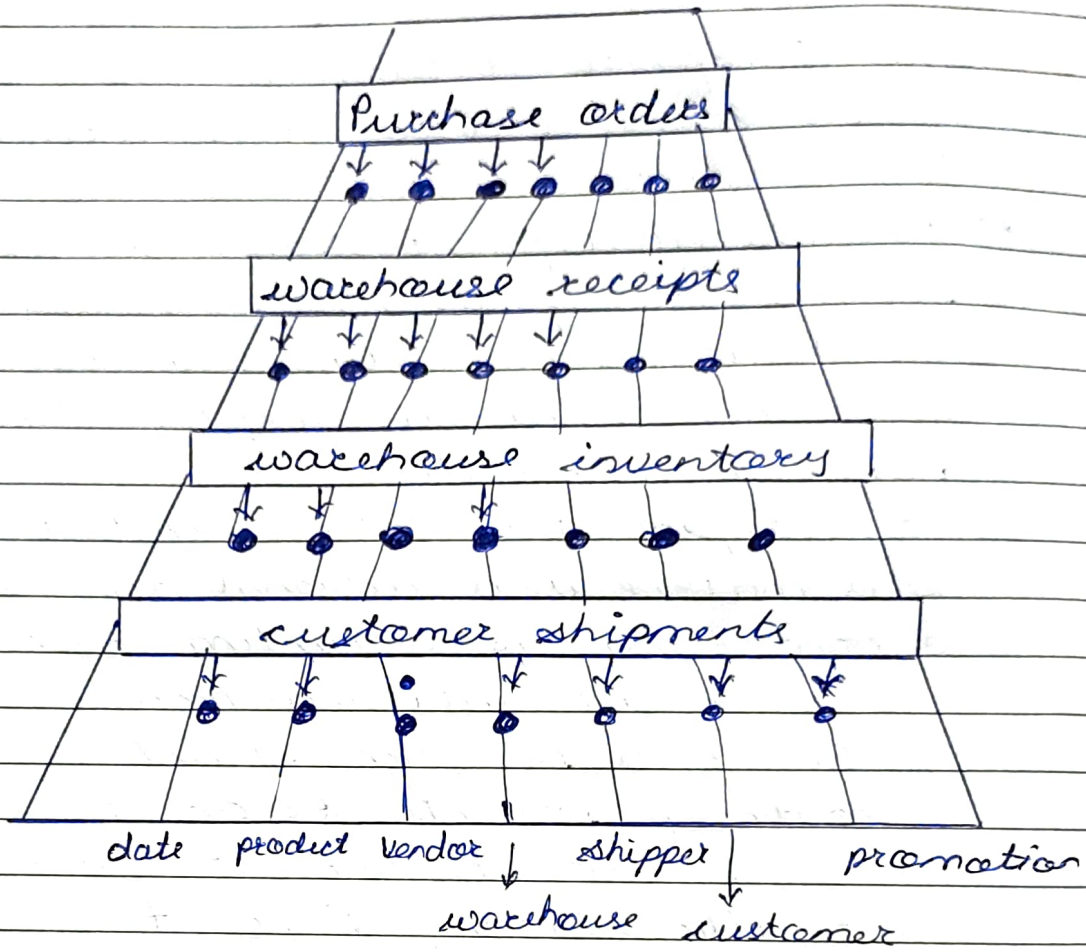


diagram of data warehouse bus with conformed dimension interfaces

Q. 2. → Explain four step dimensional design process.

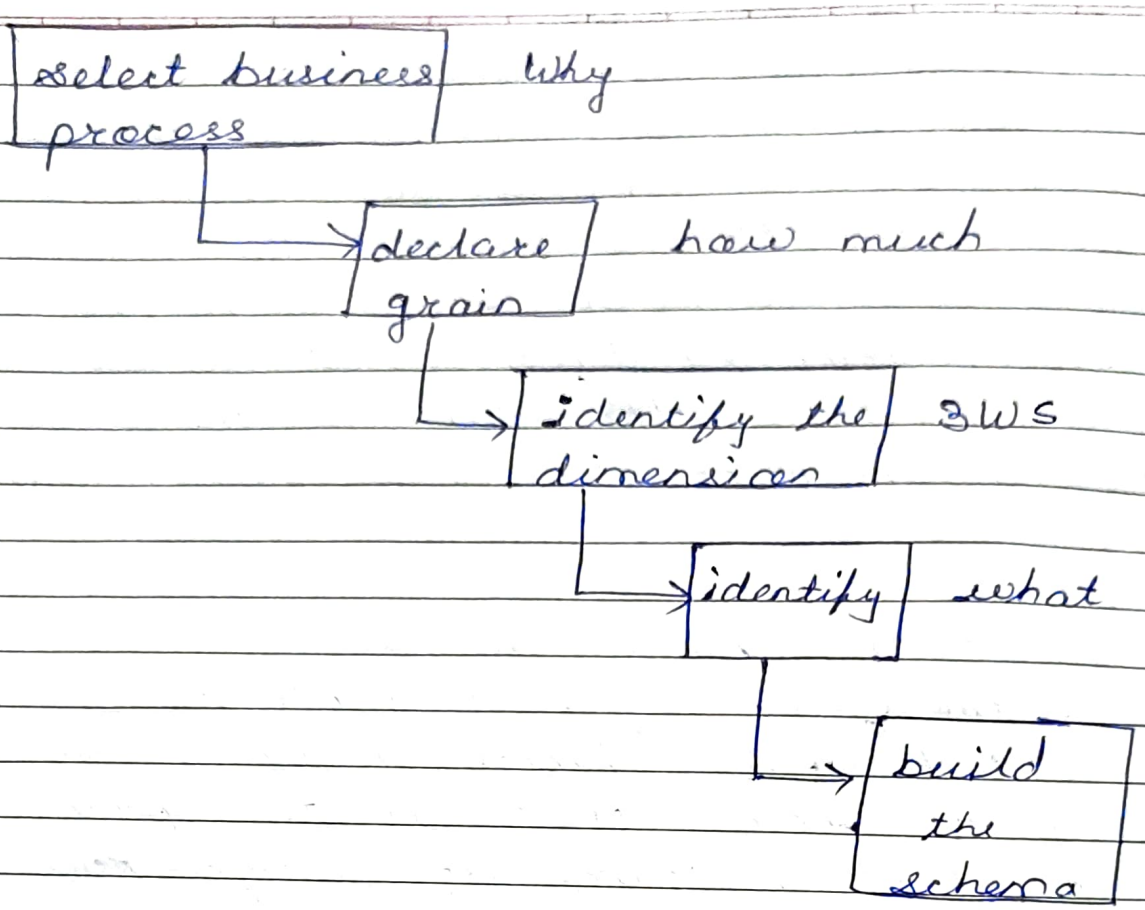
step 1 :- Choose the business process.
Each row of the enterprise data warehouse bus matrix corresponds to a candidate business process identified while gathering the business requirements.

PAGE NO.:
DATE: / /

step 2 :- Declare the grain.
Once the business process has been identified, the design team must declare the grain of fact table. It is crucial to define exactly what a fact table row is in the proposed business process dimensional model. Without aggreement on the grain of the fact table, the design process cannot successfully move forward.

step 3:- identify the dimensions
Once the grain of fact table is firmly established, the choice of dimensions is fairly straightforward. It is at this point you can start thinking of foreign keys.

step 4:- identify the facts.
The final step in four step design process is to carefully select the facts or metrics that are applicable to the business process. The facts may be physically captured by the measurement event or derived from these measurements.



Q. 3. What is dimension table? Explain surrogate keys in dimension table.

→ A] Dimension table -

(1) Dimension tables are filled with big & bulky descriptive fields.

(2) Attributes serve two purposes -
   1) query filtering
   2) query result set labeling

(3) Power of data warehouse is proportional to quality & depth of dimension attributes

(4) Attributes are textual fields.

(5) Dimension attributes should be -
   1) verbose (labels consisting of full words)
   2) descriptive
   3) complete
   4) discretely valued

5) quality assured
6) most dimensional models have 8 to 15 dimensional tables.

(6) Dimension tables represent hierarchical relationships.

(7) Dimension tables consists of highly correlated clusters of attributes grouped to represent the key objects of a business, such as its products, customers, employees or facilities.

8] Surrogate keys

(1) Primary keys are called as surrogate keys
(2) Primary keys are simple integers assigned in sequence starting with 1.
(3) A surrogate key uniquely identifies each entity in the dimension table.
(4) Surrogate keys are necessary to handle changes in dimension table attributes.
(5) Advantages of using surrogate keys -
1) performance → compact surrogate keys translate into better performance.
2) buffer from operational key management practices
3) mapping to integrate different sources → same entity is assigned to different natural keys by different source systems.

Q.4. Explain different methods for implementing slowly changing dimension.

→ A] Type 1 - Overwrite the dimension attribute-

① When the attribute value changes, the old value is merely updated or overwritten with the most current value. The dimension attributes reflects the latest state, but any historical values are lost.

② It is also used when the old value has no business significance, understanding that any historically accurate associations are lost.

B] Type 2 - Add a new dimension row.

① advantage - This allows us to accurately keep all historical information.

② disadvantages - This will cause the size of table to grow fast. In cases where the number of rows for the table is very high to start with, storage & performance can become a concern.

③ Type 2 is used when it is necessary for the data warehouse to track historical changes.

c] Type 3 - Add a new dimension attribute -

① advantage - This does not increase the size of the table, since new info is updated. This allows us to keep some part of history.

② disadvantage - It will not be able to keep all history where an attribute is

changed more than once.

③ Type 3 should only be used when it is necessary for the data warehouse to track historical changes & when such changes will only occur for a finite no. of time.

Q. 5. Explain transaction fact tables, periodic snapshot fact tables.

→ A] Transaction fact tables -

① The most basic & common fact tables are transactional.

② The grain is specified to be one row per transaction, or one row per line on a transaction.

③ Rows are inserted into fact tables only when a transaction activity occurs.

④ The transaction grain is a point in space & time; the measurements at a transaction grain must be true at that moment.

⑤ Transactional fact tables capture the measurement at its most atomic dimensional level.

B] Periodic snapshot fact table

① Periodic snapshot fact tables capture the state of metrics at a specified point-in time.

② Periodic snapshot fact tables are used to quickly assess & review the performance of the measures over specified time intervals.