Chameli Devi Group of Institutions Department of
Artificial Intelligence and Data Science
AD 603 (A) Data Mining and Warehousing
B. Tech VI Semester
Unit -II

Syllabus: Unit-II: OLAP Systems: Basic concepts, OLAP queries, Types of OLAP servers, OLAP Operations etc. Data Warehouse Hardware and Operational Design: Security, Backup and Recovery.

OLAP System: Basic Concepts-

OLAP (Online Analytical Processing) is the technology support the multidimensional view of data for many Business Intelligence (BI) applications. OLAP provides fast, steady and proficient access, powerful technology for data discovery, including capabilities to handle complex queries, analytical calculations, and predictive "what if" scenario planning. OLAP is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making.

Characteristics of OLAP System-

Th**e** need for more intensive decision support prompted the introduction of a new generation of tools. Generally used to analyze the information where huge amount of historical data is stored. Those new tools, called online analytical processing (OLAP), create an advanced data analysis environment that supports decision making, business modeling, and operations research.

Its four main characteristics are:

1. Multidimensional data analysis techniques
2. Advanced database support
3. Easy to use end user interfaces
4. Support for client/server architecture.

1. **Multidimensional Data Analysis Techniques**: Multidimensional analysis is inherently representative of an actual business model. The most distinctive characteristic of modern OLAP tools is their capacity for multidimensional analysis (for example actual vs budget). In multidimensional analysis, data are processed and viewed as part of a multidimensional structure. This type of data analysis is particularly attractive to business decision makers because they tend to view business data as data that are related to other business data.

2. Advanced Database Support:

- For efficient decision support, OLAP tools must have advanced data access features. Access too many different kinds of DBMSs, flat files, and internal and external data sources.
- Access to aggregated data warehouse data as well as to the detail data found in operational databases.
- Advanced data navigation features such as drill-down and roll-up.
- Rapid and consistent query response times.
- The ability to map end-user requests, expressed in either business or model terms, to the appropriate data source and then to the proper data access language (usually SQL).
- Support for very large databases. As already explained the data warehouse can easily and quickly grow to multiple gigabytes and even terabytes.

3. Easy-to-Use End-User Interface: Advanced OLAP features become more useful when access to them is kept simple. OLAP tools have equipped their sophisticated data extraction and analysis tools with easy-to-use graphical interfaces. Many of the interface features are "borrowed" from previous generations of data analysis tools that are already familiar to end users. This familiarity makes OLAP easily accepted and readily used.

4. Client/Server Architecture: Conform the system to the principals of Client/server architecture to provide a framework within which new systems can be designed, developed, and implemented. The client/server environment enables an OLAP system to be divided into several components that define its architecture. Those components can then

be placed on the same computer, or they can be distributed among several computers. Thus, OLAP is designed to meet ease-of-use requirements while keeping the system flexible.

V Motivation for using OLAP

I) Understanding and improving sales: For an enterprise that has many products and uses a number of channels for selling the products, OLAP can assist in finding the most popular products and the most popular channels. In some cases it may be possible to find the most profitable customers.

II) Understanding and reducing costs of doing business: Improving sales is one aspect of improving a business, the other aspect is to analyze costs and to control them as much as possible without affecting sales. OLAP can assist in analyzing the costs associated with sales.

Guidelines for OLAP Implementation

Following are a number of guidelines for successful implementation of OLAP. The guidelines are, somewhat similar to those presented for data warehouse implementation.

1. Vision: The OLAP team must, in consultation with the users, develop a clear vision for the OLAP system. This vision including the business objectives should be clearly defined, understood, and shared by the stakeholders.

2. Senior management support: The OLAP project should be fully supported by the senior managers and multidimensional view of data. Since a data warehouse may have been developed already, this should not be difficult.

3. Selecting an OLAP tool: The OLAP team should familiarize themselves with the ROLAP and MOLAP tools available in the market. Since tools are quite different, careful planning may be required in selecting a tool that is appropriate for the enterprise. In some situations, a combination of ROLAP and MOLAP may be most effective.

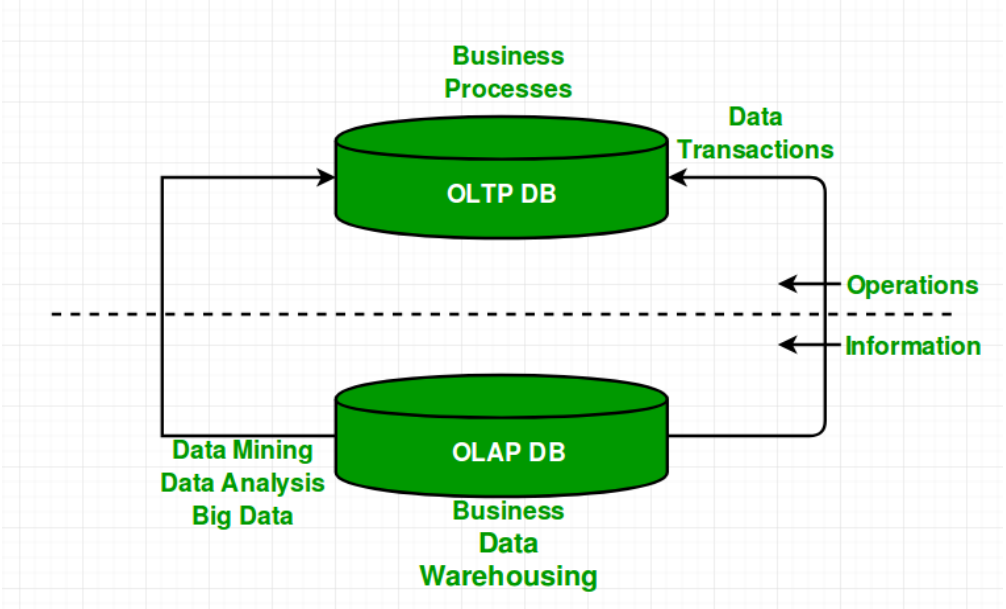
4. Corporate strategy: The OLAP strategy should fit in with the enterprise strategy and business objectives. A good fit will result in the OLAP tools being used more widely.

5. Focus on the users: The OLAP project should be focused on the users. Users should, in consultation with the technical professional, decide what tasks will be done first and what will be done later. Attempts should be made to provide each user with a tool suitable for that person's skill level and information needs. A good GUI user interface should be provided to non-technical users

6. Joint management: The OLAP project must be managed by both the IT and business professionals. Many other people should be involved in supplying ideas. An appropriate committee structure may be necessary to channel these ideas.

7. Review and adapt: As noted in last chapter, organizations evolve and so must the OLAP systems. Regular reviews of the project may be required to ensure that the project is meeting the current needs of the enterprise.

OLTP vs. OLAP



Key-Point	OLAP	OLTP
Full Form	Online Analytical Processing System	Online Transaction Processing System
Primary objective	Data Analysis	Data processing
System used	Information based system	Operations based system
Orientation	Analysis oriented/Column oriented	transaction oriented/Row oriented
Data characteristics	Volume	Transactions
Management system	Database query management system	Database modification system
Data store	Datawarehouse	Traditional RDBMS
SQL query used	Select	Select, Insert, Update, Delete
Backup activity	time to time and not as such important	Incremental and complete backup
Function	decision support	day to day operation
Data used	Summary/Historical	Current data
Schema	Start-schema	Entity model in 3NF
Examples	Reporting and analysis system	online banking, online message, online tickets etc..

Fig. Difference between OLAP and OLTP

OLAP Queries

OLAP queries are complex queries that:

➢Touch large amounts of data

➢Discover patterns and trends in the data

➢Typically expensive queries that take long time

➢Also called decision-support queries

Example-I:

Query Syntax: SELECT ...GROUP BY ROLLUP (grouping_Column_reference_list);

Example-II: SELECT Time, Location, product, sum (revenue) AS Profit FROM sales GROUP BY ROLLUP (Time, Location, product);

The Query calculates the standard aggregate values specified in the GROUPBY clause. Then, it creates progressively higher-level subtotals, moving from right to left through the list of grouping columns. Finally, it creates a grand total.

OLAP Servers

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

Types of OLAP Servers

We have four types of OLAP servers –

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

Relational OLAP: ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following –

- Implementation of aggregation navigation logic
- Optimization for each DBMS back end
- Additional tools and services
- Can handle large amounts of data
- Performance can be slow

Multidimensional OLAP: MOLAP uses array-based multidimensional storage engines for multidimensional views of data.

- Multidimensional data stores
- The storage utilization may be low if the data set is sparse.
- MOLAP server uses two levels of data storage representation to handle dense and sparse datasets.

Hybrid OLAP: Hybrid OLAP technologies attempt to combine the advantages of MOLAP and ROLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allow storing the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers: Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

OLAP Operations:

Four types of analytical operations in OLAP are
1. Roll-up
2. Drill-down
3. Slice and dice

4. Pivot (rotate)

Roll-up: Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways:

1. Reducing dimensions

2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

Consider the following diagram:

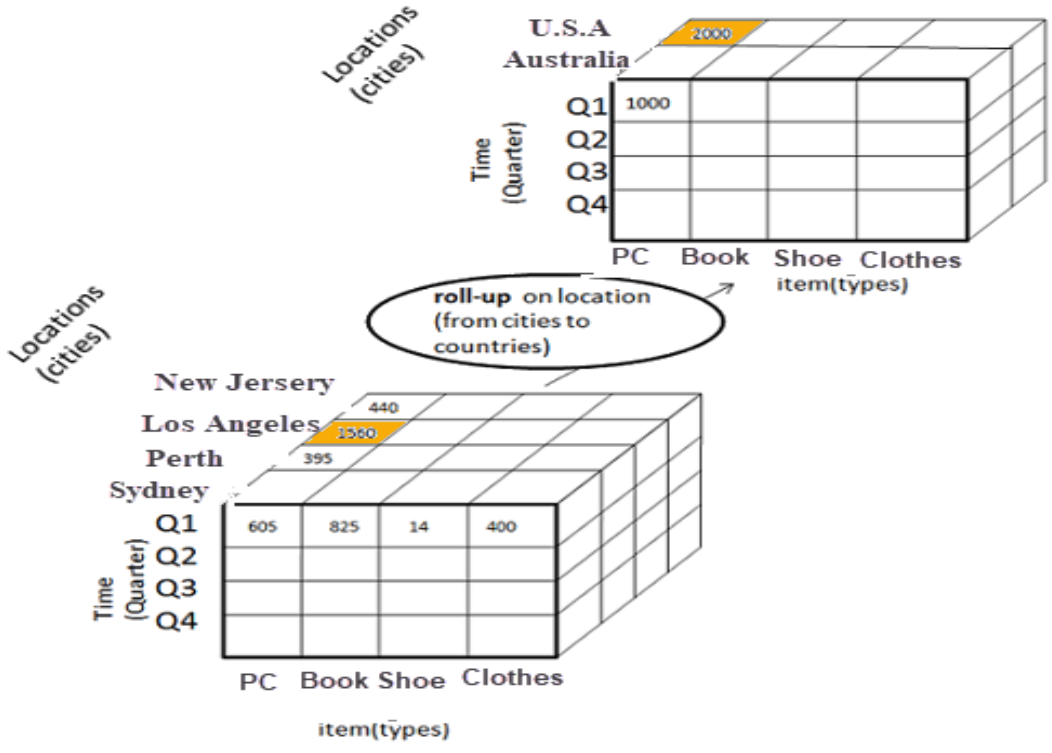


Fig: Roll-up operation

•In this example, cities New jersey and Lost Angles and rolled up into country USA.
•The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up
•In this aggregation process, data is location hierarchy moves up from city to the country.
•In the roll-up process at least one or more dimensions need to be removed. In this example, Quarter dimension is removed.

2) Drill-down: In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via

•Moving down the concept hierarchy
•Increasing a dimension

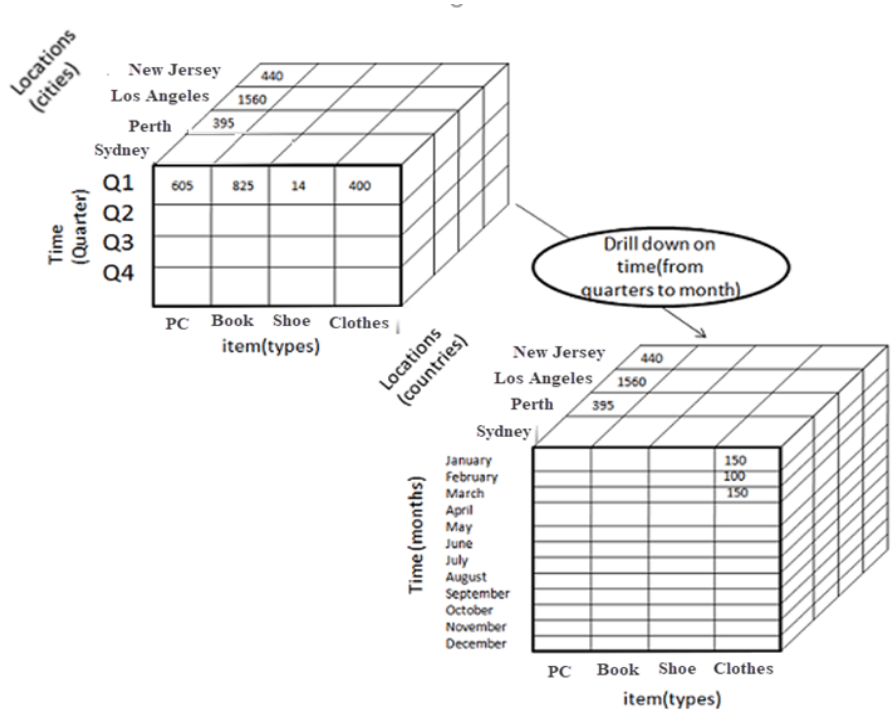


Figure: Drill-down operation

Consider the diagram above:

•Quarter Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.

•In this example, dimension months are added.

Slice: Here, one dimension is selected, and a new sub-cube is created. Following diagram explain how slice operation performed:

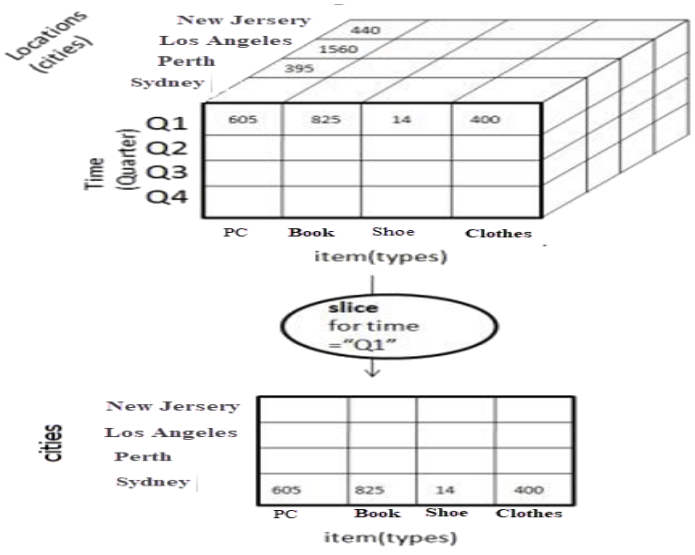


Figure 2.4: Slice operation

•Dimension Time is Sliced with Q1 as the filter.

•A new cube is created altogether.

Dice: This operation is similar to a slice. The difference in dice is to select 2 or more dimensions that result in the creation of a sub-cube.

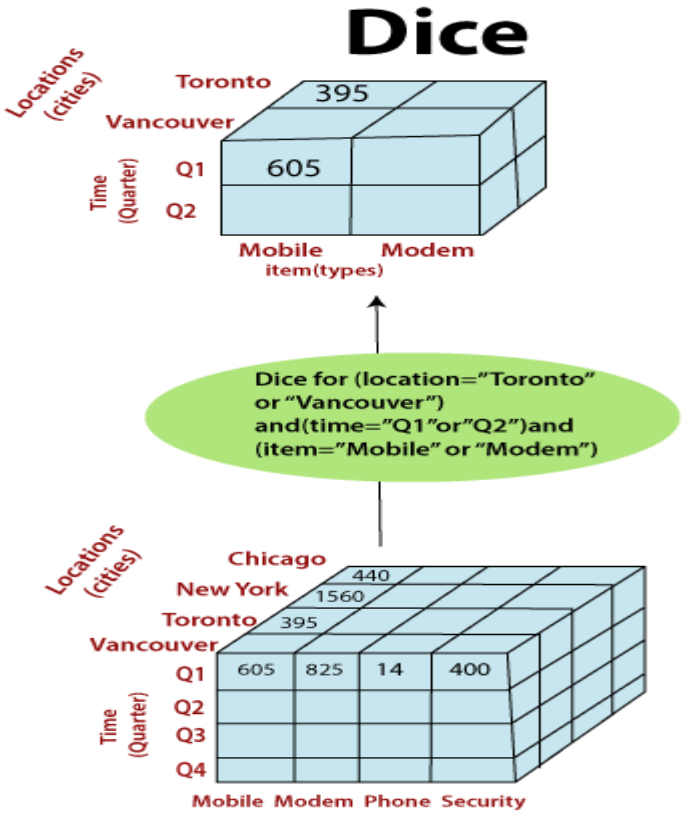


Figure 2.5: Dice

Pivot: In Pivot, rotate the data axes to provide a substitute presentation of data. In the following example, the pivot is based on item types.

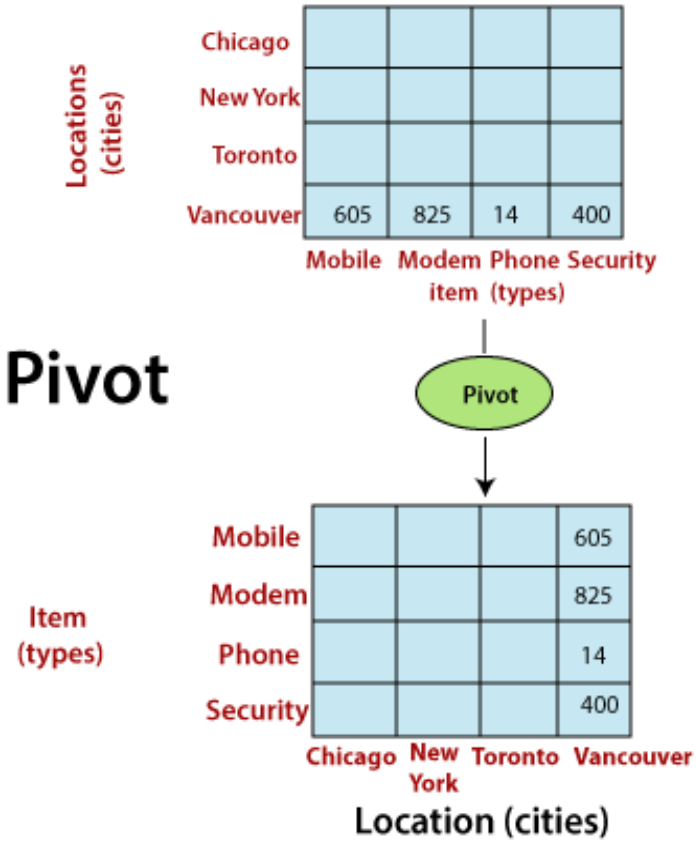


Figure 2.6: Pivot operation

Hardware and operational design

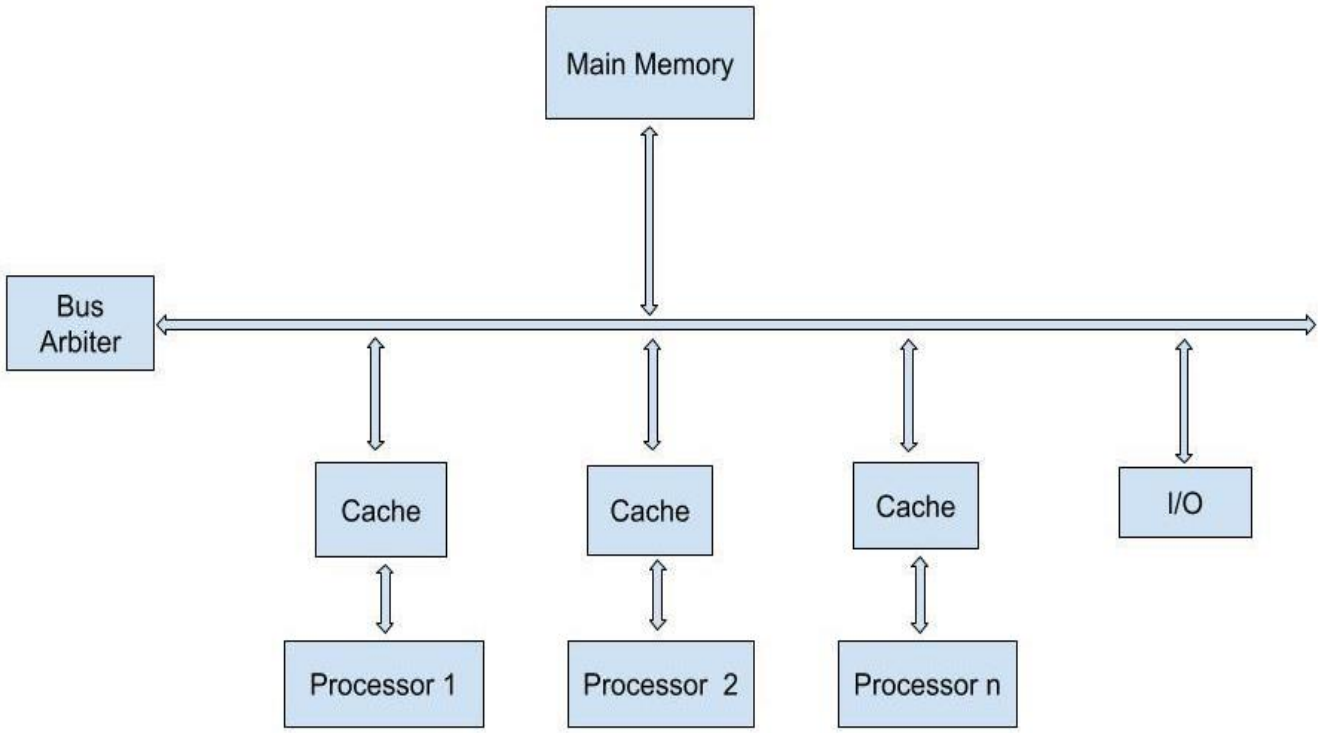
Server Hardware: Two main hardware architectures

- Symmetric Multi-Processing (SMP)
- Massively Parallel Processing (MPP)

-SMP machine is a set of tightly coupled CPUs that share memory and disk.

-MPP machine is a set of loosely coupled CPUs, each of which has its own memory and disk.

Symmetric Multi-Processing (SMP): An SMP machine is a set of CPU s that share memory and disk. This is sometimes called a shared-everything environment the CPUs in an SMP machine are all equal a process can run on any CPU in the machine, run on different CPUs at different times.



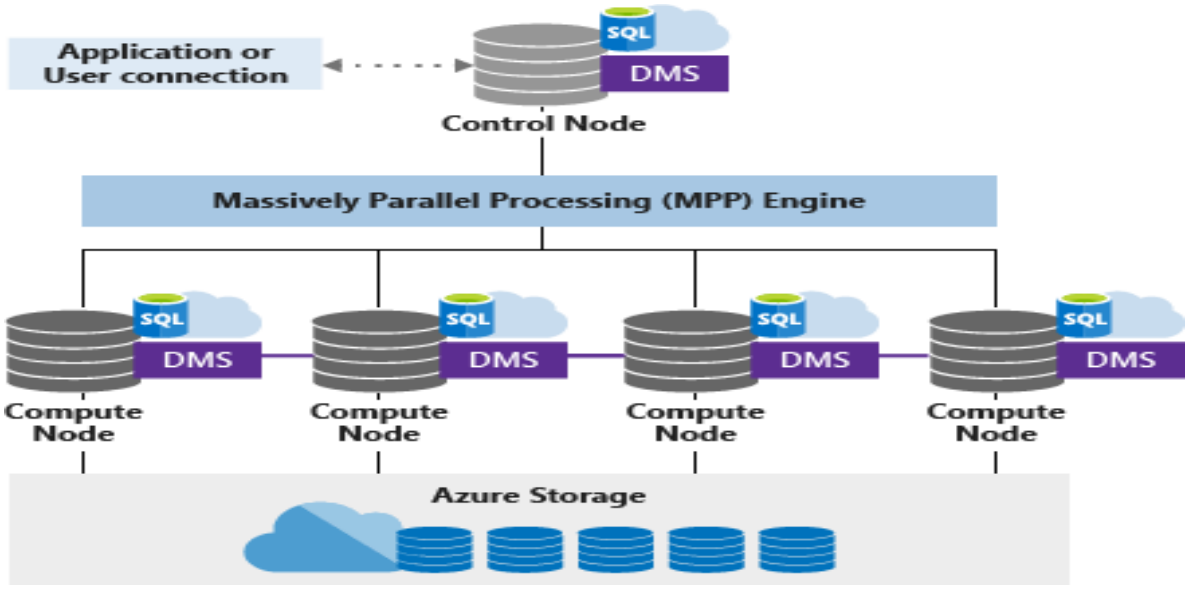
Main memory and data bus or I/O bus being shared among multiple processors in SMP

Scalability of SMP machines:

Length of the communication bus connecting the CPUs is a natural limit. As the bus gets longer the inter-process or communication speeds become slower, each extra CPU imposes an extra, band with load on the bus, increases memory contention, and so on.

Massively Parallel Processing (MPP):

Made up of many loosely coupled nodes linked together by a high-speed connection. Each node has its own memory, and the disks are not shared. Most MPP system sallow a disk to be dual connected between two nodes.



Shared Nothing Environment

A) Nothing is shared directly. Many different models of MPP

- Nodes with two CPUs
  -One CPU is dedicated to handling I/O
  -Nodes that are genuine SMP machines

- Single CPU node
  -Some will be dedicated to handling I/O and others to processing

b) Level of shared nothing varies from vendor to vendor
Example: IBM SP/2 is a. fully shared nothing system

Virtual Shared Disk (VSD)

- An extra layer of software to be added to allow disks to be shared across nodes
- System will suffer overheads if an I/O is issued and data has to be shipped from node to node.

Distributed Lock Manager

- MPP machines require this to maintain the integrity of the distributed resources
- Track which node's memory holds the current copy of each piece of information
- Getting the data rapidly from node to node

Network Hardware

Network Architecture

- Sufficient bandwidth to supply the data feed and user requirements .

Impact to design

- User access via WAN –impacts the design of Query Manager
- Source system data transfer
- Data extractions

Example: Problem of getting the data from the source systems.

It may not get the data to the warehouse system early enough to allow it to be loaded, transformed, processed and backed up within the overnight time window.

Guideline

- Ensure that the network architecture and bandwidth are capable of supporting the data transfer and any data extractions in an acceptable time.
- The transfer of data to be loaded must be complete quickly enough to allow the rest of the overnight processing to complete.

Client Hardware

Client Management

- Those responsible for client machine management will need to know the requirements for that machine to access the data warehouse system.
- Details such as the network protocols supported on the server, andthe server's Internet address, will need to be supplied.
- If multiple access paths to the server system exist this information needs to be relayed to those responsible for the client systems.
- During node fall over users may need to access a different machine address.

Client Tools

- The tool should not be allowed to affect the basic design of the warehouse itself.
- Multiple tools will be used against the data warehouse.
- Should be thoroughly tested and trialed to ensure that they are suitable for the users.
- Testing of the tools should ideally be performed in parallel with the data warehouse design:
- Usability issues to be exposed,
- Drive out any requirements that the tool will place on the data warehouse

Disk Technology:

RAID Technology

Redundant Array of Inexpensive Disks

- The purpose of RAID technology is to provide resilience against disk failure, so that the loss of an individual disk does not mean loss of data.
- Striping is a technique in which data is spread across multiple disks.
- RAID levels 0, 1 and 5 are commercially viable and thus widely available

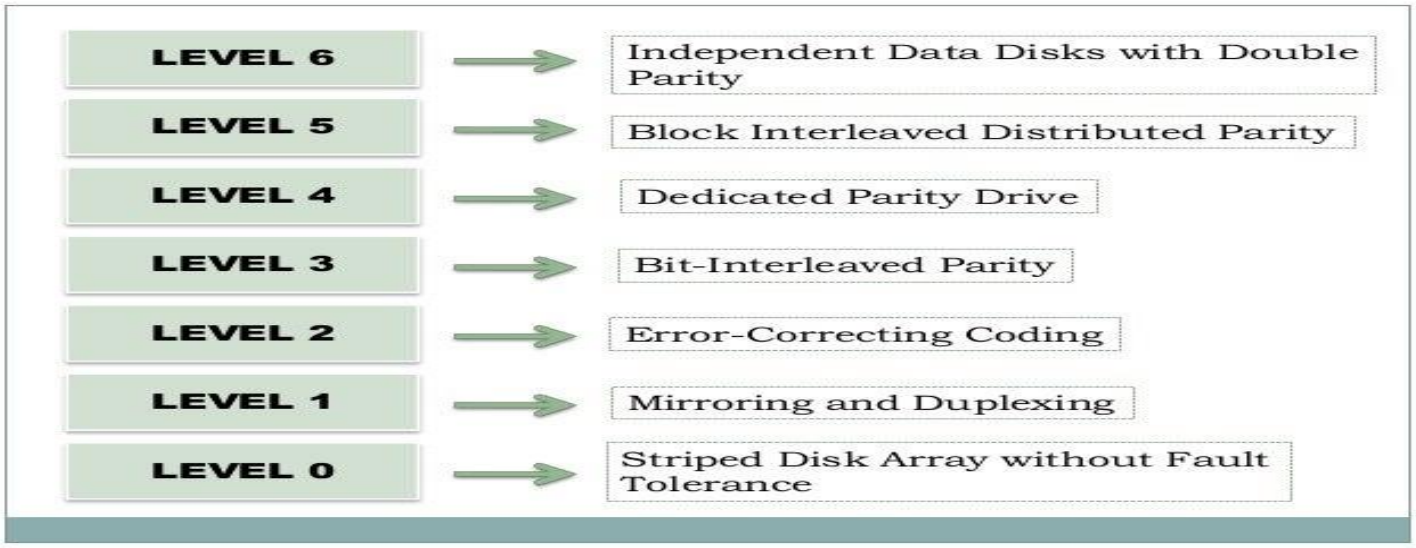
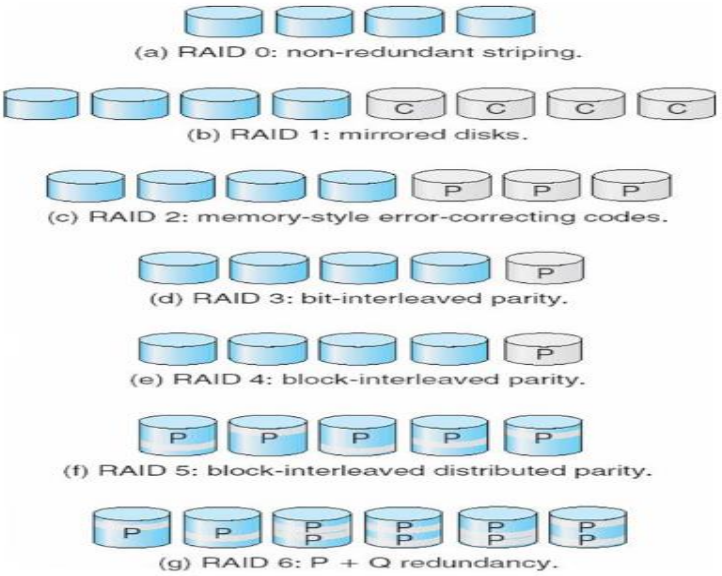


Table: RAID Level with Descriptions



What is Data Warehouse Security?

Data warehouses pull data from many different sources, and warehouses have many moving parts. Security issues arise every time data moves from one place to another. Data warehouse security entails taking the necessary steps to protect information so that only authorized personnel can access it.

Data warehouse security should involve the following:

- Strict user access control so that users of the warehouse can only access data they need to do their jobs.
- Taking steps to secure networks on which data is held.
- Carefully moving data and considering the security implications involved in all data movement.

Physical Security Practices

1. Restricting and controlling physical access to data warehouses has been made easy thanks to tech innovations like biometric readers, anti-tailgating systems and other physical access control mechanisms. These might look excessive and expenditure overhead, but they play a crucial role in ensuring the integrity and safety of the precious enterprise data.
2. Imparting information about security protocols and ensuring all the personnel in the proximity of the data warehouse religiously obey and adhere to these rules is one of the keys to success. It's understandable that an employee can be used by intruders to gain access, but if the employee in question is ardently following the specified guidelines it makes a world of difference.

Software-Based Security Measures

Data Encryption: Data encryption is one of the primary safeguards against data theft. All data should be encrypted using algorithms like AES (advanced encryption standard) or FIPS 140-2 certified software for data encryption, whether it's in the transactional database or the data warehouse. Some proponents would argue that data encryption adversely affects the performance and data access speed of data centers, but considering the alternative, it is preferred.

Data Segmenting and Partitioning: Data encryption although an added security measure can be quite cumbersome if applied without segmenting and partitioning. Segmenting and partitioning entail classifying or splitting data into sensitive and non-sensitive information. After going through partitioning the data should be accordingly encrypted and put into separate tables ready for consumption.

Securing On-The-Move: Data Securing data in a single place and transit are two different ball games. Here data in transit means the one which is being relayed from transactional databases in real time to the data warehouse. These transactional databases can be anywhere geographically, therefore using protective protocols, such as SSL or TSL is highly recommended. Cloud-based data warehouses nowadays provide a secure and impenetrable tunnel between the database and the cloud storage which should be leveraged.

Trusted Witness Server: As mentioned earlier, hackers and intruders nowadays have become as skilled and sophisticated as the security measures they are up against. Implementing a trusted witness server is akin to hiring a watchdog that avidly and quite tenaciously keeps vigil on your data access points.

It can detect an unwarranted and suspicious attempt at accessing data and generate an alert immediately. This allows the people responsible for the data warehouse security to stop the intruders dead in their tracks.

Backup and recovery of Data Warehouse

Backup and recovery are among the most important tasks for an administrator, and data warehouses are no different. However, because of the sheer size of the database, data warehouses introduce new challenges for an administrator in the backup and recovery area. Data warehouses are unique in that the data can come from a myriad of resources and it is transformed before finally being inserted into the database; but mostly because it can be very large. Managing the recovery of a large data warehouse can be a daunting task and traditional OLTP backup and recovery strategies may not meet the needs of a data warehouse.

Strategies and Best Practices for Backup and Recovery

Devising a backup and recovery strategy can be a daunting task. And when you have hundreds of gigabytes of data that must be protected and recovered in the case of a failure, the strategy can be very complex.

The following best practices can help you implement your warehouse's backup and recovery strategy:

- Best Practice A: Use ARCHIVELOG Mode
- Best Practice B: Use RMAN
- Best Practice C: Use Read-Only Table spaces
- Best Practice D: Plan for NOLOGGING Operations
- Best Practice E: Not All Table spaces are Equally Important

Best Practice-A: Use ARCHIVELOG Mode Archived redo logs are crucial for recovery when no data can be lost, because they constitute a record of changes to the database.

Oracle can be run in either of two modes:
•ARCHIVELOG --Oracle archives the filled online redo log files before reusing them in the cycle.
•NOARCHIVELOG --Oracle does not archive the filled online redo log files before reusing them in the cycle.

Running the database in “ARCHIVELOG” mode has the following benefits:

- The database can be completely recovered from both instance and media failure.
- The user can perform backups while the database is open and available for use.
- Oracle supports multiplexed archive logs to avoid any possible single point of failure on the archive logs
- The user has more recovery options, such as the ability to perform table space-point-in-time recovery (TSPITR).
- Archived redo logs can be transmitted and applied to the physical standby database, which is an exact replica of the primary database.
- The database can be completely recovered from both instance and media failure.

Running the database in “NOARCHIVELOG” mode has the following consequences:

- The user can only back up the database while it is completely closed after a clean shutdown.
- Typically, the only media recovery option is to restore the whole database, which causes the loss of all transactions since the last backup.

Best Practice-B: Use RMAN There are many reasons to adopt RMAN. Some of the reasons to integrate RMAN into your backup and recovery strategy are that it offers:

- Extensive reporting
- Incremental backups
- Downtime free backups
- Backup and restore validation
- Backup and restore optimization
- Easily integrates with media managers
- Block media recovery

- Archive log validation and management
- Corrupt block detection

Best Practice C: Use Read-Only Table spaces

Best Practice D: Plan for NOLOGGING Operation

Best Practice E: Not All Table spaces are Equally Important