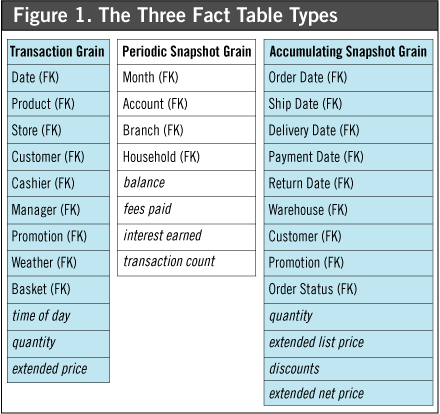
I am working on creation of datalake (subset of data warehouse) as of now. Our DW data resides on HDFS. my work involves creating facts and dimensions of datalake. Apart from that I write procedures which talks directly with tables in data warehouse. My task is to integrate multiple sources and write procedures to retrieve data from those.

**Differences between Top-down Approach and Bottom-up Approach**

* Top-down approach analyzes risk by aggregating the impact of internal operational failures while bottom-up approach analyzes the risks in individual process using models
* Top-down approach doesn’t differentiate between high frequency low severity and low frequency high severity events while bottom-up approach does
* Top-down approach is simple and not data intensive whereas bottom-up approach is complex as well as very data intensive
* Top-down approaches are backward-looking while bottom-up approaches are forward-looking

**Top-down Approach**  
In simple terms, top-down approach is an investment strategy that selects various sectors or industries and tries to achieve a balance in an investment portfolio. The top-down approach analyzes the risk by aggregating the impact of internal operational failures. It measures the variances in the economic variables that are not explained by the external macro-economic factors. As such, this approach is simple and not data-intensive. Top-down approach relies mainly on historical data. This approach is opposite to bottom-up approach.  
  
**Bottom-up Approach**  
A bottom-up approach on the other hand is an investment strategy that depends on the selection of individual stocks. It observes the performance and management of companies and not general economic trends. The bottom-up approach analyzes individual risk in the process by using mathematical models, and is thus data-intensive. This method does not rely on historical data. It is a forward-looking approach unlike the top-down model, which is backward-looking.



A row in an **accumulating snapshot** fact table summarizes the measurement events occurring at predictable steps between the beginning and the end of a process.

Slowly Changing Dimensions (SCD) - dimensions that change slowly over time, rather than changing on regular schedule, time-base. In Data Warehouse there is a need to track changes in dimension attributes in order to report historical data. In other words, implementing one of the SCD types should enable users assigning proper dimension's attribute value for given date. Example of such dimensions could be: customer, geography, employee.

**Incremental Aggregation in Informatica**

Incremental Aggregation is the process of capturing the changes in the source and calculating the aggregations in a session. This process makes the integration service to update the target incrementally and avoids the process of calculating the aggregations on the entire source. Consider the below sales table as an example and see how the incremental aggregation works.

Source:

YEAR PRICE

----------

2010 100

2010 200

2010 300

2011 500

2011 600

2012 700

For simplicity, I have used only the year and price columns of sales table. We need to do aggregation and find the total price in each year.  
  
When you run the session for the first time using the incremental aggregation, then integration service process the entire source and stores the data in two file, index and data file. The integration service creates the files in the cache directory specified in the aggregator transformation properties.  
  
After the aggregation, the target table will have the below data.

Target:

YEAR PRICE

----------

2010 600

2011 1100

2012 700

Now assume that the next day few more rows are added into the source table.

Source:

YEAR PRICE

----------

2010 100

2010 200

2010 300

2011 500

2011 600

2012 700

2010 400

2011 100

2012 200

2013 800

Now for the second run, you have to pass only the new data changes to the incremental aggregation. So, the source will contain the last four records. The incremental aggregation uses the data stored in the cache and calculates the aggregation. Once the aggregation is done, the integration service writes the changes to the target and the cache. The target table will contains the below data.

Target:

YEAR PRICE

----------

2010 1000

2011 1200

2012 900

2013 800

**Points to remember**

1. When you use incremental aggregation, first time you have to run the session with complete source data and in the subsequent runs you have to pass only the changes in the source data.
2. Use incremental aggregation only if the target is not going to change significantly. If the incremental aggregation process changes more than hhalf of the data in target, then the session perfromance many not benfit. In this case go for normal aggregation.

**Note:** The integration service creates a new aggregate cache when

* A new version of mapping is saved
* Configure the session to reinitialize the aggregate cache
* Moving or deleting the aggregate files
* Decreasing the number of partitions

**Configuring the mapping for incremental aggregation**  
  
Before enabling the incremental aggregation option, make sure that you capture the changes in the source data. You can use lookup transformation or stored procedure transformation to remove the data which is already processed. You can also create a trigger on the source database and can read only the source changes in the mapping.

**Incremental load** is defined as the activity of **loading** only new or updated records from the database into an established QVD. **Incremental loads** are useful because they run very efficiently when compared to full **loads**, particularly so for large data sets

An **OLAP** cube is a multidimensional database that is optimized for **data warehouse** and online analytical processing (**OLAP**) applications. ... In **OLAP** cubes, **data** (measures) are categorized by dimensions. **OLAP** cubes are often pre-summarized across dimensions to drastically improve query time over relational databases.

It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. This chapter cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

## OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

* Roll-up
* Drill-down
* Slice and dice
* Pivot (rotate)

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. This chapter cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

## 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.

## Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

## Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store 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

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

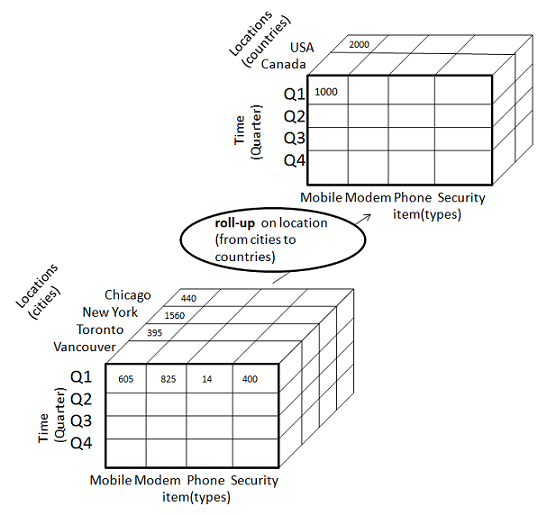
* Roll-up
* Drill-down
* Slice and dice
* Pivot (rotate)

### Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

* By climbing up a concept hierarchy for a dimension
* By dimension reduction

The following diagram illustrates how roll-up works.



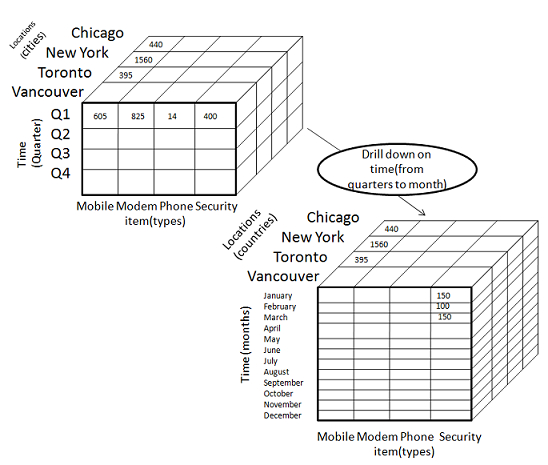
* Roll-up is performed by climbing up a concept hierarchy for the dimension location.
* Initially the concept hierarchy was "street < city < province < country".
* On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
* The data is grouped into cities rather than countries.
* When roll-up is performed, one or more dimensions from the data cube are removed.

### Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways −

* By stepping down a concept hierarchy for a dimension
* By introducing a new dimension.

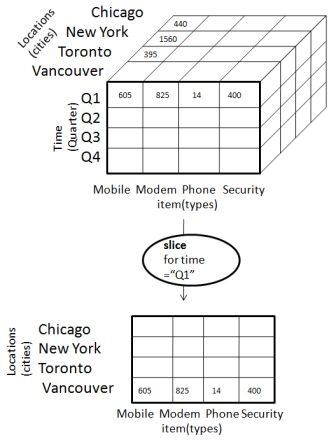
The following diagram illustrates how drill-down works −



* Drill-down is performed by stepping down a concept hierarchy for the dimension time.
* Initially the concept hierarchy was "day < month < quarter < year."
* On drilling down, the time dimension is descended from the level of quarter to the level of month.
* When drill-down is performed, one or more dimensions from the data cube are added.
* It navigates the data from less detailed data to highly detailed data.

### Slice

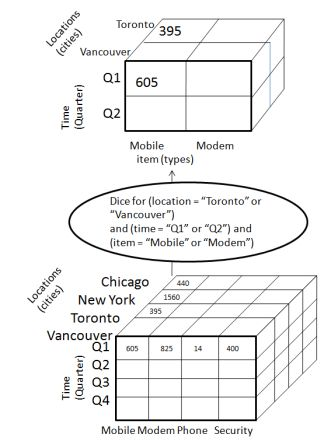
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



* Here Slice is performed for the dimension "time" using the criterion time = "Q1".
* It will form a new sub-cube by selecting one or more dimensions.

### Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

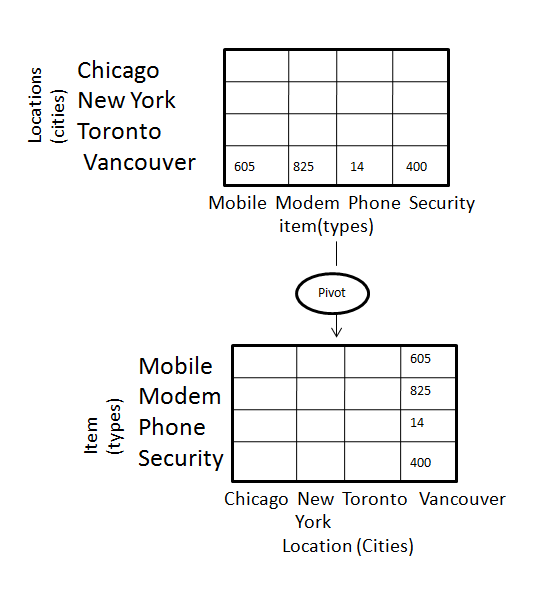


The dice operation on the cube based on the following selection criteria involves three dimensions.

* (location = "Toronto" or "Vancouver")
* (time = "Q1" or "Q2")
* (item =" Mobile" or "Modem")

### Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



## OLAP vs OLTP

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **Data Warehouse (OLAP)** | **Operational Database (OLTP)** |
| 1 | Involves historical processing of information. | Involves day-to-day processing. |
| 2 | OLAP systems are used by knowledge workers such as executives, managers and analysts. | OLTP systems are used by clerks, DBAs, or database professionals. |
| 3 | Useful in analyzing the business. | Useful in running the business. |
| 4 | It focuses on Information out. | It focuses on Data in. |
| 5 | Based on Star Schema, Snowflake, Schema and Fact Constellation Schema. | Based on Entity Relationship Model. |
| 6 | Contains historical data. | Contains current data. |
| 7 | Provides summarized and consolidated data. | Provides primitive and highly detailed data. |
| 8 | Provides summarized and multidimensional view of data. | Provides detailed and flat relational view of data. |
| 9 | Number or users is in hundreds. | Number of users is in thousands. |
| 10 | Number of records accessed is in millions. | Number of records accessed is in tens. |
| 11 | Database size is from 100 GB to 1 TB | Database size is from 100 MB to 1 GB. |
| 12 | Highly flexible. | Provides high performance. |

## 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.

## Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

## Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store 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.

Multidimensional OLAP (MOLAP) uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the dataset is sparse. Therefore, many MOLAP servers use two levels of data storage representation to handle dense and sparse datasets.

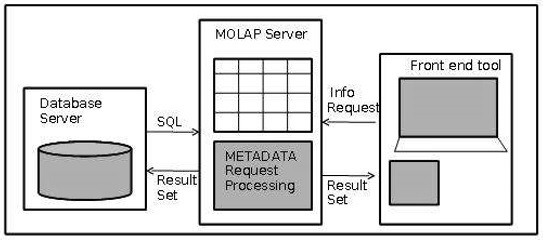
## Points to Remember −

* MOLAP tools process information with consistent response time regardless of level of summarizing or calculations selected.
* MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
* MOLAP tools need fastest possible performance.
* MOLAP server adopts two level of storage representation to handle dense and sparse data sets.
* Denser sub-cubes are identified and stored as array structure.
* Sparse sub-cubes employ compression technology.

## MOLAP Architecture

MOLAP includes the following components −

* Database server.
* MOLAP server.
* Front-end tool.



## Advantages

* MOLAP allows fastest indexing to the pre-computed summarized data.
* Helps the users connected to a network who need to analyze larger, less-defined data.
* Easier to use, therefore MOLAP is suitable for inexperienced users.

## Disadvantages

* MOLAP are not capable of containing detailed data.
* The storage utilization may be low if the data set is sparse.

## MOLAP vs ROLAP

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **MOLAP** | **ROLAP** |
| 1 | Information retrieval is fast. | Information retrieval is comparatively slow. |
| 2 | Uses sparse array to store data-sets. | Uses relational table. |
| 3 | MOLAP is best suited for inexperienced users, since it is very easy to use. | ROLAP is best suited for experienced users. |
| 4 | Maintains a separate database for data cubes. | It may not require space other than available in the Data warehouse. |
| 5 | DBMS facility is weak. | DBMS facility is strong. |