ETL – Extraction, Transformation and Loading

1. Data Warehouse Concepts

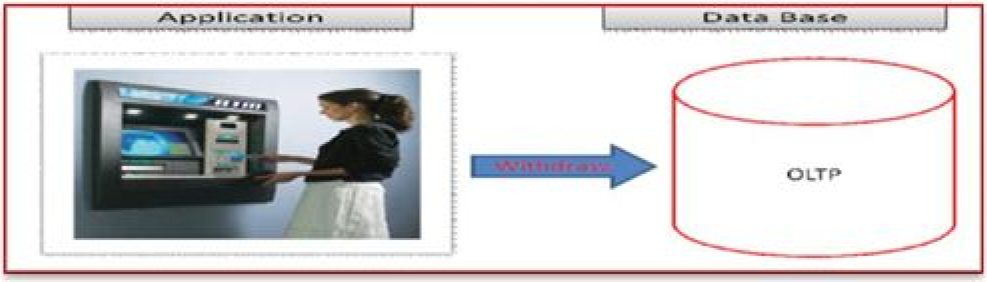
1.1 Data

1. Anything which giving some meaningful information is known as data 2. Data is two types
1. Transactional Data
2. Analytical Data

1.2 Transactional Data

1. Is run time data or day to day data
2. is current and detail
3. Is useful to run the business
4. Is stored in OLTP(Online Transaction Processing)
5. Source of transactional data is Applications
6. Example: ATM Transactions , Share market transactions..etc

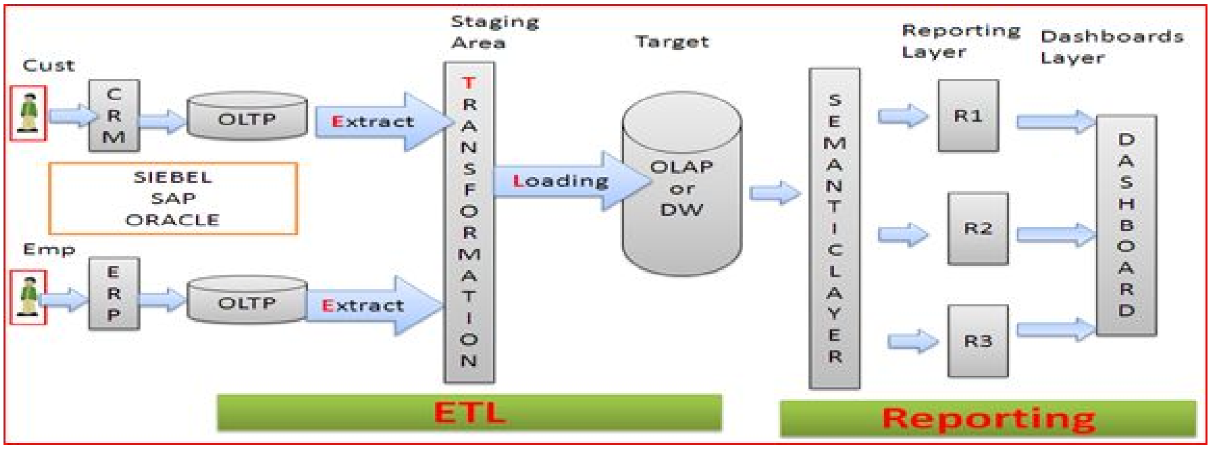
Transaction Example Diagram :



1.3 Analytical Data

1. is useful to ANALYSE the business
2. is Historical and summarized
3. Is stored in OLAP(Online Analytical Processing) or DW(Data Warehouse)
4. Source of Analytical data is OLTP

1.4 DW Architecture



1.5 DW Tools

1. DW tools are divided into two types. some of those tools are:

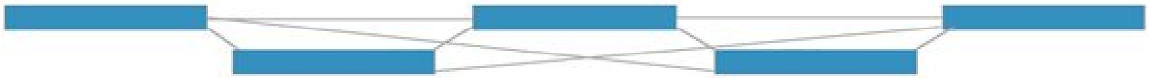
ETL	Reporting
Informatica	OBIEE
Data Stage	BI Publisher
Abintio	Cognos
SSIS	SAP-BO
ODI	DOMO
OWB	Qlick View
BODI	MSTR

1.6 OLTP Vs OLAP

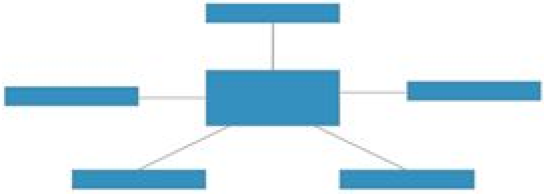
OLTP	OLAP
1. Is useful to store Transactional data	1. Is useful to store Analytical data
2. Is useful to run the business	2. Is useful to Analyze the business
3. The nature of data is current and Detail	3. The nature of data is historical and summarized
4. OLTP Supports CRUD(Create , Partially read, update and delete)	4. OLAP supports only read
5. It is a application oriented DB	5. It is subject oriented DB
6. It is volatile	6. It is nonvolatile
7. In OLTP data storage time is fixed	7. In OLAP data storage time is variant
8. OLTP DB are isolated as Applications	8. OLAP is integrated as per subject area
9. No of users are more(customers + emp)	9. No of users are less (MM+HM)
10. In OLTP we will use normalizes schema	10. In OLAP we will use Denormalized Schema

Transactional Vs Analytical Systems

1. Transactional schema optimized for Partial read/write—multiple joins



2. Analytics schema optimized for querying large datasets—few joins



SAP HANA ETL – Extraction, Transformation and Loading

SAP HANA ETL based replication uses SAP Data Services to migrate data from SAP or non-SAP source system to target HANA database. It enables to read the business data at Application layer. You need to define data flows in Data Services, scheduling a replication job and defining source and target system in data store in Data Services designer.

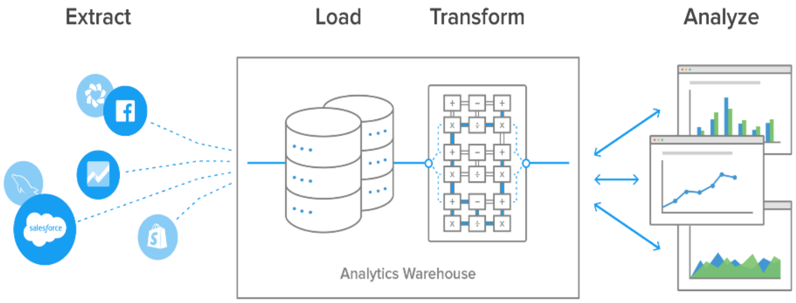


Fig. 1 ETL Process Overview : Image credit - *https://www.stitchdata.com/etldatabase/etl-process/*

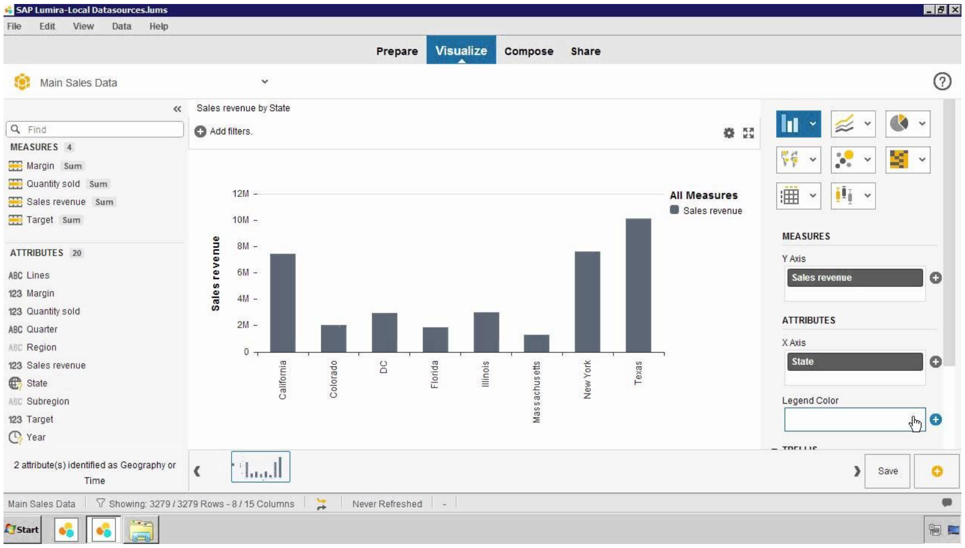
Steps:

1. **Extraction**: As part of this project, we will be provided data source.
   a) For the data source, create logical data map which explain the tables, table type, column names, datatypes, relationship constraints between the tables.
   b) Create test data in the SAP HANA database, covering all boundary conditions that the data can take.

2. **Transformation**:
   a) Cleaning: Mapping Null to 0 or Male to M, date format consistency.
   b) Identifying and removing duplicate records
   c) Key restructuring: Establish key relationships across tables
   d) Derivation: Apply rules to data that derive the new calculated values from existing data – for example, creating a revenue metric that subtracts taxes
   e) Filtering: Selecting only certain rows or columns
   f) Joins: Linking multiple tables
   g) Splitting: split a column into multiple column

3. **Load**: The loading phase is the last step of the ETL process. The information from data sources are loaded and stored in a form of tables. We can use Incremental Loading method which is widely used to load data in data warehouses from the respective source systems. This technique is employed to perform faster load in less time utilizing less system resources. But since, we are creating only enough data to cover all the boundary conditions so it will be better to use Full Load method.

4. **Visualization**: SAP HANA includes new capabilities in transaction capture and playback, new tools to help users analyze and visualize data, and a number of enhancements intended to make HANA a more compelling option for organizations, both midsize and large.


SAP Lumira-Local Datasources.lums
File
Edit
View
Data
Help
Prepare
Visualize
Compose
Share
Main Sales Data
Find
MEASURES 4
Margin Sum
Quantity sold Sum
Sales revenue Sum
Target Sum
ATTRIBUTES 20
Lines
Margin
Quantity sold
Quarter
Region
Sales revenue
State
Subregion
Target
Year
2 attribute(s) identified as Geography or Time
Sales revenue by State
Add filters.
Sales revenue
12M
10M
8M
6M
4M
2M
0
California
Colorado
DC
Florida
Illinois
Massachusetts
New York
Texas
All Measures
Sales revenue
MEASURES
Y Axis
Sales revenue
ATTRIBUTES
X Axis
State
Legend Color
TRELLIS
Save
Main Sales Data
Showing: 3279 / 3279 Rows - 8 / 15 Columns
Never Refreshed
Start