

Data Warehousing

▼ In class ideas

Consolidated data in place

Application used for operational purpose of a company

ERP or FICO - **Enterprise resource planning**

HRMS - **Human Resources Management System**

CRM - **Customer relationship management**

TMS - **Transportation management system**

EAI - **Enterprise application integration**

Transactional systems or OLTP - Online transactional process

Delivery application - Used by

Integration - Pushing data from one application to another application

EAI - used for automated exchange of information between enterprise applications.

OLAP - Online analytical processing

OLTP - Online Transaction Processing

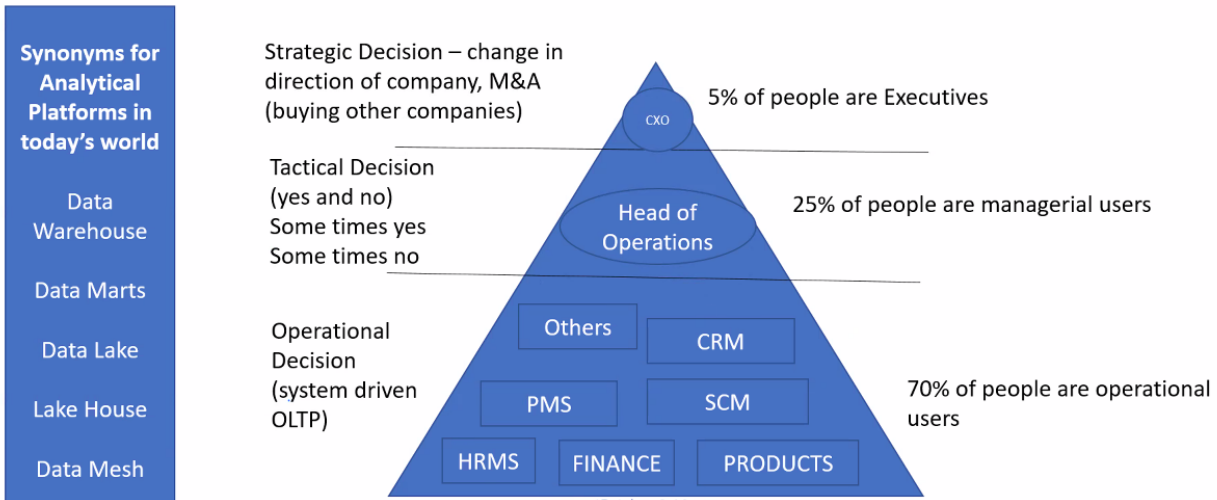
Entry level -- operational users -- operational decision — OLTP

Manager -- more data points than just that transaction alone

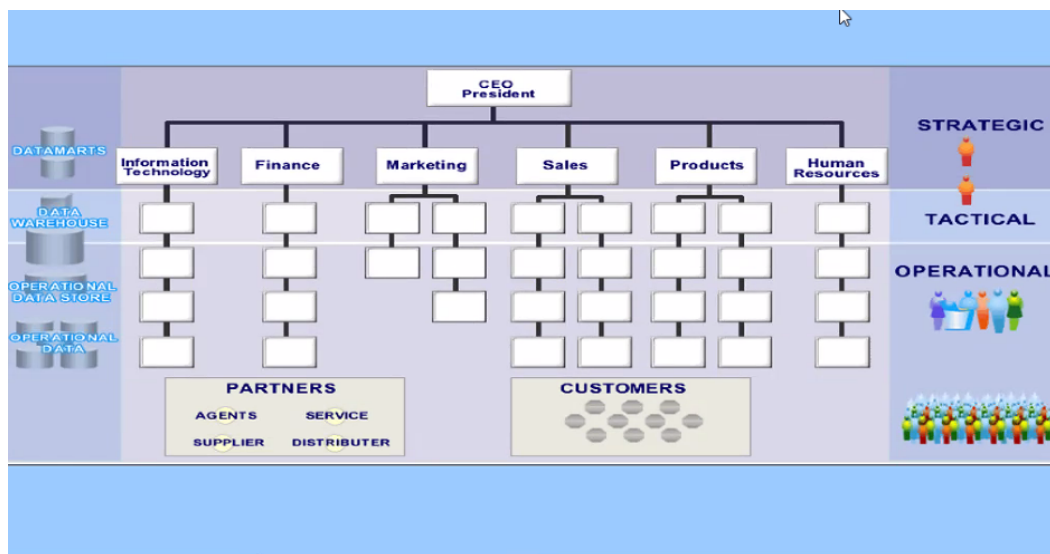
Executives -- Strategic Decisions

OLAP is used to get insights

Datawarehouse is an application which is built to help managers and executives to take decisions



Data Analytical platforms



Operational data -- OLTP -- Normalized DATABASE

DATA which gets generated -- RDBMS (oracle, SQL Server)

Data Warehouse - analytical platform

Data Mart - problem of dept but follows same concept of analytical platform

Data gets captured will be stored in OLTP but it is fragmented

OLTP systems were data gathered

- ERP
 - Web apps
 - mobile apps
 - IOT
 - Legacy (old systems)
 - API
-
- KPO - Problem Statement
 - BPO - Says about Problem Statement and

1) What is normalization

- 2) Diff between OLTP and DW
- 3) Business Process / Analytical Process
- 4) Information and Data association
- 5) Analytics end user?

▼ Credit and AML

What's a Credit Score?

- A credit score is a calculated value that represents an individual's creditworthiness. It's mainly used by lenders, such as banks and credit card companies, to determine the likelihood that the individual will repay their debts.
- 3 digits ranges from 350 to 850 FICO **Fair Isaac Corporation**
 - Excellent: 800–850

- Very Good: 740–799
- Good: 670–739
- Fair: 580–669
- Poor: 300–579

How Credit Score gets calculated?

- History, types of loans, length of credit history, debt utilization, and whether you've applied for new accounts
 1. Payment history (35%)
 2. Amounts owed (30%)
 3. Length of credit history (15%)
 4. Types of credit (10%)
 5. New credit (10%)

What is AML anti money laundering?

- • Anti-Money Laundering (AML) laws, regulations, and procedures reduce the ease of hiding profits from crime.
- Financial institutions combat money laundering with Know Your Customer (KYC) and Customer Due Diligence (CDD) measures
 - KYC determines the identity of new clients and whether their funds originated from a legitimate source.
 - Maintain accurate and up-to-date records of transactions and customer information for regulatory compliance and potential investigations
 -

How bank identify aml

- Unusual transaction patterns
- Cash-intensive activities

▼ Bank Tran example

- Transaction cannot be changed
 - Phone number of customer changed
 - Sensitive columns will have old value (history info)
original will become history and current will remain current
-

OLTP - Requires Current data to operate better

OLAP requires history data, current data, and changed data

- Old data is called **HISTORICAL**
 - Current data is called **ACTIVE**
 - Master data is called Dimension
 - Transaction table is called Facts table
-

02/09/24 Thu

What is DW?

- A process of transforming data into information making it available to user in a timely enough manner to make a difference.
 - Data → Information
 - Info is also called as insights
- Collected from various source made to end users in what they can understand and use in a business context.
- Same data but perspective wise should make sense to them that is business context

Typical problem **WITHOUT** DW

- Data will be scattered over network
- many versions of data
- Need expert to get data
- Poorly documented
- results are unexpected
- transferring from one form to other

What management wants?

- Should be integrated across enterprise
 - Data engineers job to analyze from diff source and integrate
- Summary should have real value to organization
 - OLTP have all data's
 - Most tactical and strategic decs using aggregated value.
- Historical data holds the key to understand data over time
 - Key to know the before and after analysis
- what if capabilities required
 - If remove products from a line what is the impact on revenue, what will the upcoming reactions.

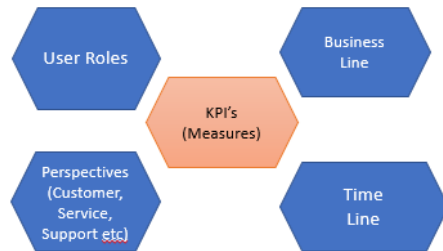
OLTP used to push faster

OLAP used to pull data faster

Master rec have changed and current data's

Analysis view from positions

Analysis View from Managers & Above

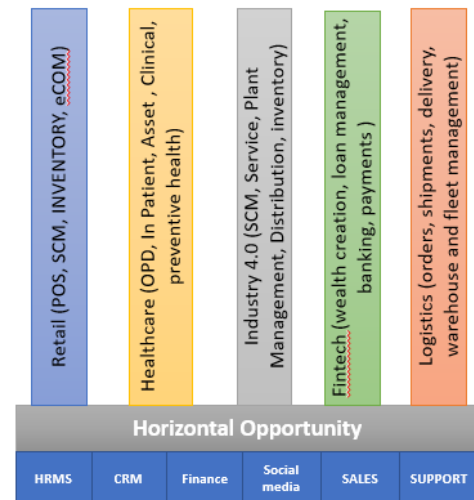


World of Analysis

Viewing Metrics / Measures based on different perspectives (Dimensions)

Analysing Sales Revenue by Customer Segment
 Analysing Sales Revenue by Time Period
 Analysing Sales Revenue by Product Line
 Analysing Sales Revenue by Location
 Analysing Sales Revenue by Payment Type
 Analysing Sales Revenue by Channel (Shop, Online, TV, Call Centre)

www.iBridae.com



▼ Rajesh sir's notes on historical data

customer

cust_id | cust_nm | cust_city | cust_dob | cust_occupation | cmail | cust_phone
 cust_gender

1500 | UMA, R | CHENNAI | 2/12/2000 | Student | uma.r@gmail.com
 4545454545 F (Current)

OLTP -- Current DATA which is required to operate better

OLAP -- History data for analysis (3 records of UMA, original, changed, current)

1500 UMA, R BLR 12/12/2000 Student
uma.r@gmail.com 4545454545 F (His)

1500 UMA, R BLR 12/12/2000 Student
uma.r@gmail.com 1212121212 F (History)

bank_txn

txn_id	txn_date	cust_id	txn_type	Amount
100001	12-jan-24	1500	DEP	4000

100002	13-jan-24	1500	WD	2000
100003	14-jan-24	1500	WD	1000
100004	08-feb-24	1500	DEP	8000

Transactions cannot be changed

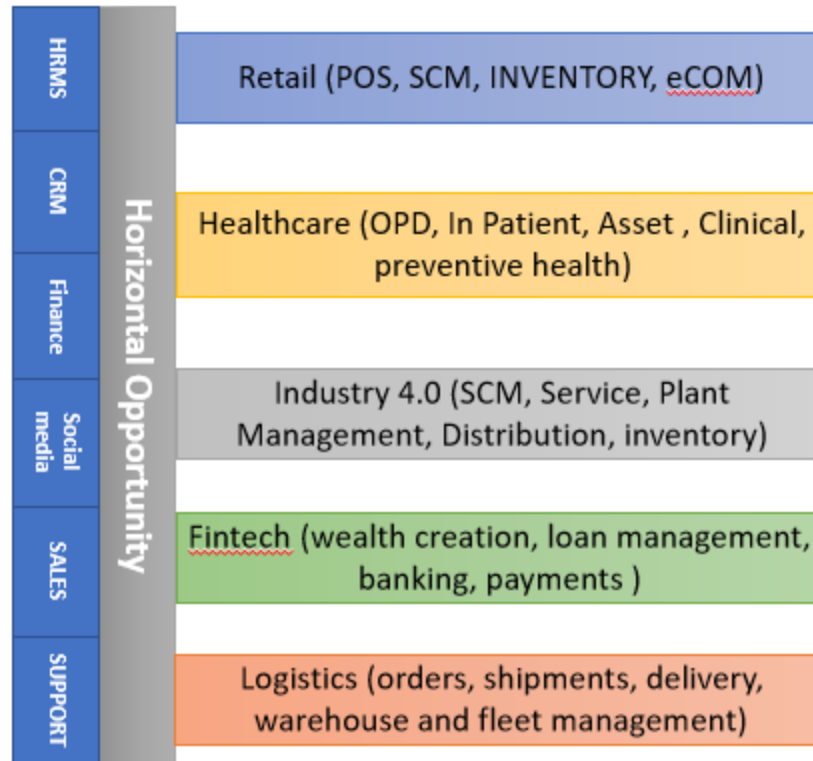
MASTER -- Dimension

TXNS -- Fact

09/02/24 Fri

KPI - Key performance indicator

- Store
 - Sales revenue
 - Cost of running a store
 - profit
 - Revenue of a month
- Any company big have will at least 10 dept
- DW was built only for big companies



- Basics of all business are horizontal opportunity

World of analysis

- Viewing metrics or measures based on different perspective(dimension) (products selling , price)
 - Analyzing sales revenue by customer segment e.g.(kids, teenager, old age people)
 - Analyzing sales revenue by time period
 - Analyzing sales revenue by product line
 - Analyzing sales revenue by payment type
 - Analyzing sales revenue by channel(shop, online, tv, call center)

OLTP (Operational) Vs OLAP (Analysis)



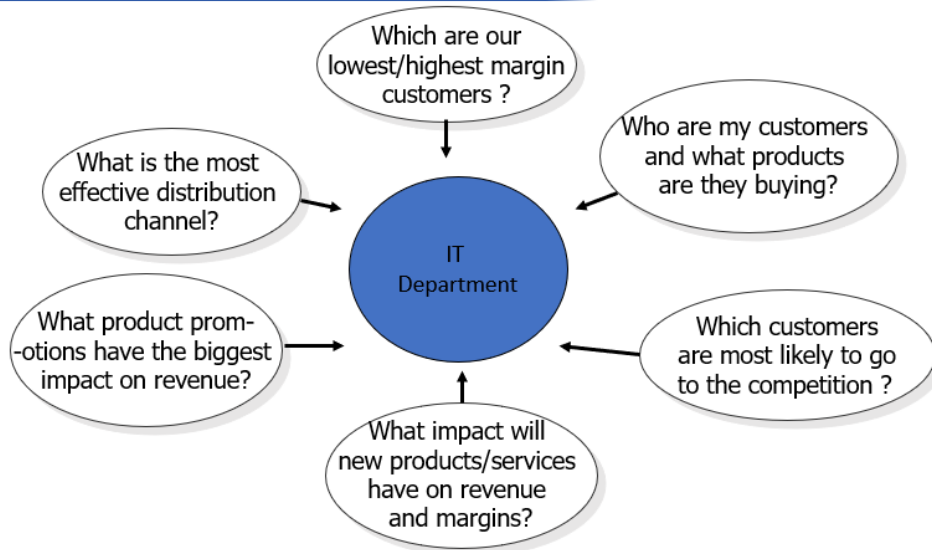
- | | |
|---|---|
| <ul style="list-style-type: none">• More DML operations (Update, Delete, Inserts)• Point Queries (usually we go to database based on the PK or FK to get data)• Very specific while issuing queries• We typically store 2 years to 3 years in operational systems• Used for day today activities (must to run the business). We don't (except for audit) store changes in the operational systems.• Focus to implement the operations part of the business.• Follow the 3rd normal form data modeling to create the model. | <ul style="list-style-type: none">■ Typically we don't change the data in DW (less or no updates and deletes)■ Queries based on time period, set of products, set of customers etc (Range Queries)■ We typically store 3+ years of data (for analysis)■ We track the changes in the DW as part of historical management.■ Used mainly for analytics (trend analysis, customer behavior etc)■ De-normalization / Dimensional Model is the way design the DW storage |
| <ul style="list-style-type: none">• Small amount of data retrieved• Only 2 to 3 years data available• Daily activities or audit purpose• Focus on operational part of business• 3rd normal form data modelling
More tables | <ul style="list-style-type: none">• Large amount of data retrieved• we have 3+ years data will be available• Historical management analysis• mainly for analytics• De norm or dimensional modelling |

Analysis (data platform stage)

- what happened ? (static report)
- Why it happened? (Ad hoc reports)
- What if analysis? (what may be the impact if an event happens?)
- What will happen(ML ,SAS,SPSS,MATLIB,H2O)
- Real time analytics and responses (What's happening & automation of designs)

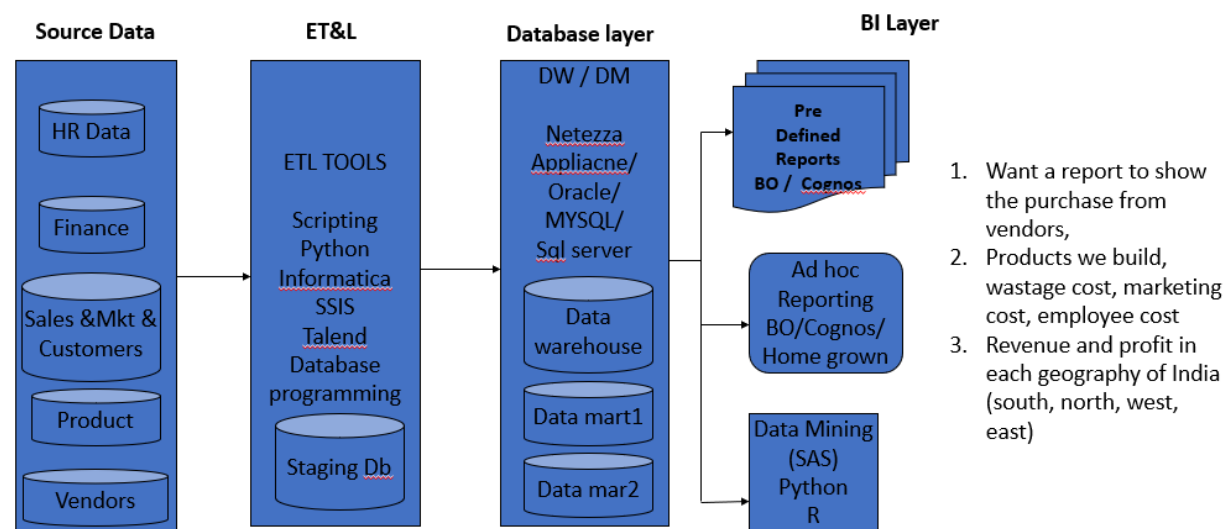
- What happened? (static report)
 - What was the sales of last year?
 - What is the sale of current year?
 - How many people we hired in India this month?
 - How many projects we executed last month?
 - ROI does not exists
 - Reports will be generated to show the summary values and KPI (key performance indicators)
- Why it happened? (Ad hoc reports)
 - The sales in the 3rd quarter is above than the normal quarterly sales. I am interested to see the catalyst which caused this extra sales.
 - Normal attrition in the company is 11.5%, but last quarter we have an attrition of 18%? What's going on? → How do you find the reason? If you are going to become a data analyst you will assist the companies to find the reasons of that exception data. Get the reason behind the actual impact.
 - If we identify the good, we will change the current process to become better.
 - If we identify the bad, we will change the process to be **proactive**.
 - ROI – return on investment becomes higher
 - Reports & dashboards will be built to do better ad hoc analysis (drill down, drill up)
 - With this knowledge, we modified out business process (BPR == Business Process Reengineering)
- What If Analysis
 - What if we move the department of Customer service from US(1.2 million\$) to India (445K)?
 - If we have 100 products we sell, out of 100 products, two products are slow moving products. What if I completely shut down selling those two products?
 - Impact on factory
 - Impact on employees (22)
 - Impact on Revenue
 - Impact on Suppliers
- What will happen? (ML – Python, SAS, SPSS, R, Matlib, H2O.ai)
 - What will be the sales for next year?
 - What will be the attrition rate of IT department next year?
 - What products will give more problems from customer perspective?
- Real Time Analytics and Responses (What's happening & automation of Decisions)
 - Streaming Analytics
 - Fraud detection in financial transactions (being proactive rather than reactive using AI)

Management Wants to KNOW



- Only analytical systems can answer these questions

Data Warehouse Architecture



- Source of data can be anywhere
- To extract from source we use - ETL

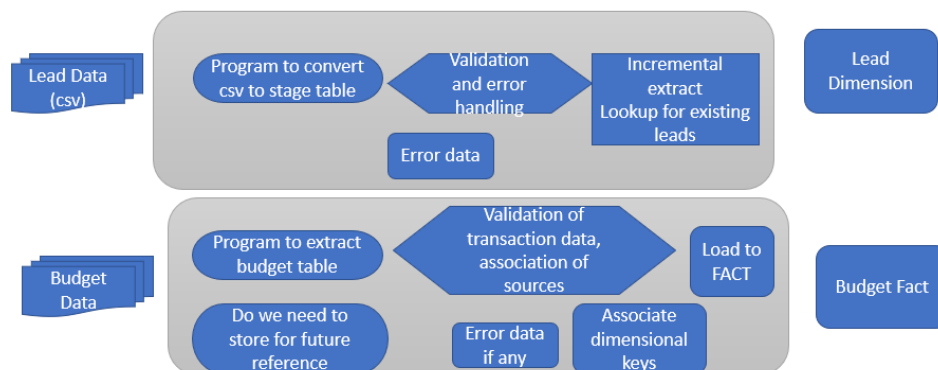
- Database layer - different RDBMS storage available to create data mesh or data mart
- BI layer - give report to end user
 - Ad-hoc will be second layer
 - Data mining or prediction layer

12/02/24

Visualization of ETL jobs/ pipeline

Moving data from one place to another

Visualization of ETL jobs / Pipeline



- Take csv and convert to stage table
- Check for validation or errors
- Incremental extract and lookup for existing table
- Look for budget data
- Program and extract

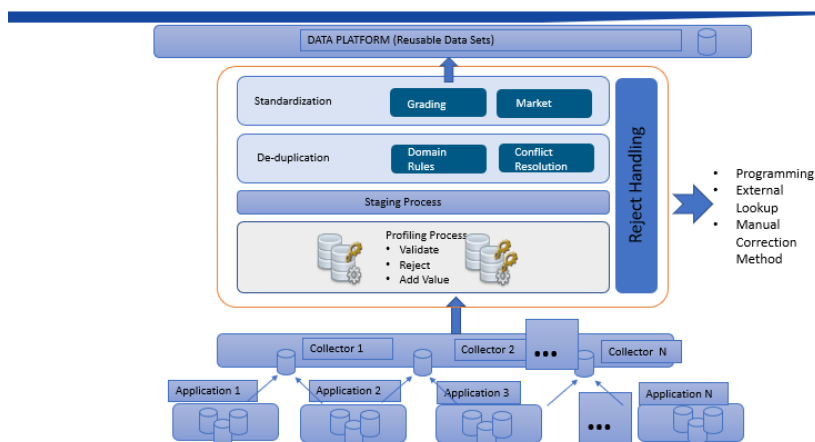
▼ Data Analytics Platform – different views

- Data generated from origin

- What is the origin of employee record (HRMS) – ROO (record of origin)
 - Where is this data? – operational system
 - What is the format which you have data? -- RDBMS
 - Do I have access? – user security
 - Does technology can read / process the data we have? – Informatica
- Types of data
 - Structured Data -- RDBMS
 - Semi Structured Data – NoSQL (mongodb), emails, log files
 - Un Structured Data – documents (pdf), word, audio, video
- When ever we built tech to handle Structured , Semi Structured and Un Structured it is called as data lakes.
- Semi-structured + unstructured data – Data Lake (all types of data)
- Data Engineer (data analysis, data processing (data pipeline), data reporting/dashboards)
- The data in the data warehouse / MDM / Curated Zone is called as record of reference (ROR) – single version of truth (is place where we can see enterprise data can be understood)
- **Source systems**
 - Existing application data in RDBMS format, NO SQL Format
 - Files (csv, json, excel, xml)
 - External Data (Market data)
- **Extract, Transform and Load (ETL)**
 - Extract, Transform and Load – this is the process we follow to cleanse / integrate and process data. (tool should be capable of doing transformation in memory)

- ETL tools are Informatica, Data Stage, SSIS, Ab Initio etc.
- Ingest, Decode, Transform, Curated Zone, streaming
- Extract Sing\$, US\$, GBP, transform Sing\$ as well as US\$ into GBP and store the values.
- **Data Storage for processed data**
 - RDBMS, Flat files (s3)
 - Snowflake (cloud Data warehouse) – cheaper in storage compared to high end databases like oracle, db2
 - Big Data (HDFS)
- **Business Intelligence**
 - Then we deliver valuable information to our end users
 - BI tools are BO, Cognos, MicroStrategy, OLAP tools etc.
- **ML / AI**
 - Prediction of the future event / events
 - Changing the operations by dynamic learning (deep learning)

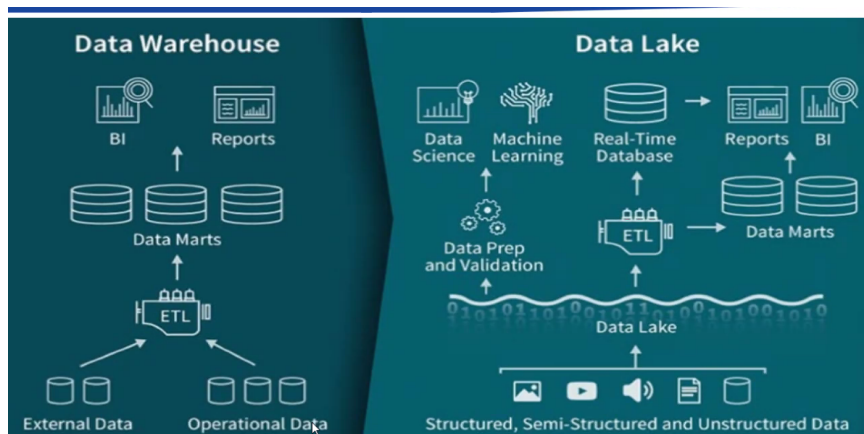
Data Processing Visualization



14/12/24 wed

Data pipeline world-

- As/is layer (data captured in source system)
- Validation & rejection layer(remove values which have no values)
- Integration, Merging and Aggregation
- Reusable data tables
- Dimension and facts



- Data Warehouse
 - Enterprise-wide
 - Industry follows E-R Model or Star schemas
 - Structure for corporate view of data
- Data Mart
 - Departmental
 - Mostly Star Schema based (Facts and dimensions)
 - Quick turn around (up and running as there are less stakeholders)
- Data Lakes
 - Big Data (HDFS) -- Hadoop
 - AWS S3 or Azure Object Storage
 - Store mostly unprocessed data in the data lakes
 - On consumption you do transformation (compute is heavy)

Data Transformation Terms



- | | |
|---|--|
| <ul style="list-style-type: none"> • Extracting DATA from source <ul style="list-style-type: none"> • SQL, File Read, API Calls • Conditioning <ul style="list-style-type: none"> • Files (data type issue) US, APAC – mm/dd/yyyy, dd/mm/yyyy, dd-mon-yy • Scrubbing <ul style="list-style-type: none"> • Errors in spelling mistakes • BANGALORE, BENGALURU, BLR, BANGALURU • Merging <ul style="list-style-type: none"> • Merge two different sets of data • India sales, SL sales, BL Sales (+++) • House holding (Cluster) <ul style="list-style-type: none"> • Get HR Data (FT, Consultant, Intern, Others) • Aggregation <ul style="list-style-type: none"> • Sum, min, max, avg, mean, count • Rejection <ul style="list-style-type: none"> • Reject the invoices which are having amount = 0 • Tax registration number is a must for regulation, if you don't have the registration number, reject that data (error records) | <ul style="list-style-type: none"> • Enrichment <ul style="list-style-type: none"> • <u>City name, state name (ip)</u> • (target – <u>city name, zipcode, state name</u>) (filling missing fields) • Scoring <ul style="list-style-type: none"> • Activity, process where you can grade a customer / employees / supplier. If cust is with you for more than 5 years then assign grade A, >2 years and < 5 years then assign B, >1 < 2 years then assign C, Else D • Deduplication – removing duplicate entries in the source data set. • Loading (SQL, File Write, HDFS, RedShift, Snowflake) • Validating – we reject data (accept / reject) • Delta – difference – incremental extract <ul style="list-style-type: none"> • We read all the timesheet data (up until yesterday) this morning. Tomorrow you get only the data which are new or modified rather than extracting full |
|---|--|

www.iBridae360.com

19/02 Monday

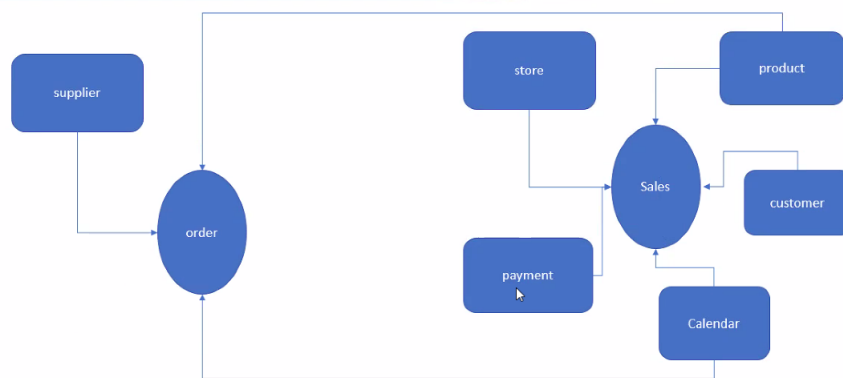
Building a DW model

- Normalization - 3rd NF
- De normalization - Dimensional modelling
 - master table - dimension
 - tnx table - fact table

De norm

- Star schema - completely de normalized
- Snow Flake scheme - Partially de normalized - Flexible

Star Schema



OLTP is subjective

Dimensional table

Dimension table attributes in the DW

- Define business in terms already familiar to users
- Wide rows with lots of descriptive text
- Small tables (about a million rows)
- Joined to fact table by a foreign key
- heavily indexed
- hierarchical elements exists in the dimensional
- typical dimensions
 - time periods, geographic region (markets, cities), products, customers, salesperson, etc.

- Historical is associated with only Dimension tables
- information in Dimension changes
- information in fact table does not change
- many to many have 2 diff dimension
- 1 to many have same dimension
- More structured handling data ware house
- data analytical platform and cloud handling called as data lake

Important types

- Least bothered in star schema is called type 1
- core is called type 2
 - Before and after analysis can be done in type 2

- As discussed, over the period of time the master data gets changed in the source system. When that change happens, in DW we capture the same in one of the following ways.
- TYPE 1 or SCD1
 - No history stored (We do this, if business does analysis based on the current data alone)
- TYPE 2 or SCD2
 - Complete history stored. (when ever product price gets changed, the system should track it to see the price fall pattern)
 - We have to generate a unique key called surrogate ID's in this type of tables
- TYPE 3 or SCD3
 - Only current and previous values are stored.
 - We design the same in such a way that the current and previous values are stored in one physical key record. (storing customer_city and customer_pre_city)

Surrogate Id

Current product price

Previous product price

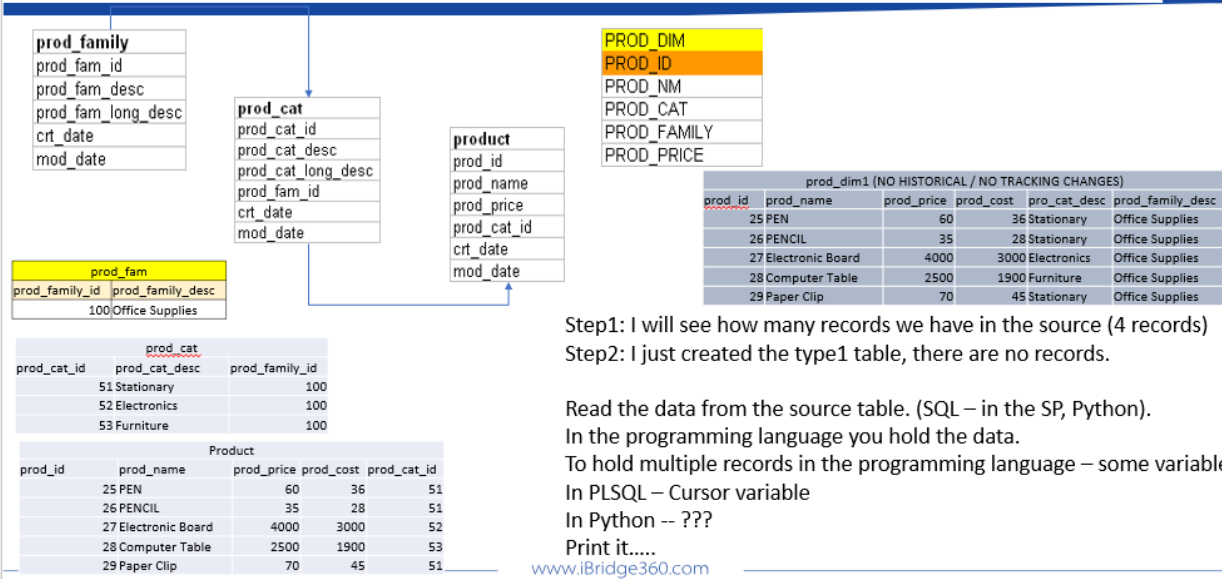
Expecting two IDs for the same product

Therefore, create a surrogate ID. This is an alternate ID that represents multiple versions of the same ID.

What type of type 2

- Surrogate id
- Have start and end date

Type1 dimension



Type2 Data Visualization (tracking changes)



prod_fam

prod_family_id	prod_family_desc
100	Office Supplies

prod_cat

prod_cat_id	prod_cat_desc	prod_family_id
51	Stationary	100
52	Electronics	100
53	Furniture	100

product

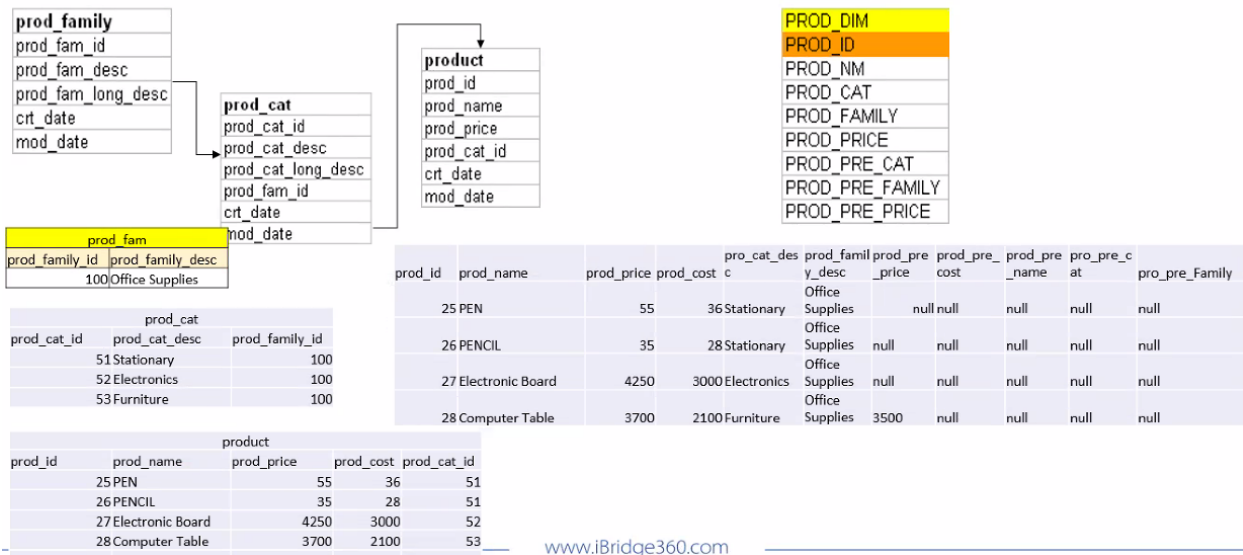
prod_id	prod_name	prod_price	prod_cost	prod_cat_id
25	PEN	60	36	51
26	PENCIL	50	28	51
27	Electronic Board	4000	3000	52
28	Computer Table	2100	1900	53
29	Paper Clip	55	35	51
30	Pen holder	100	80	51
31	Wireless Mouse	850	800	52

prod_dim2 (Historical changes -- track all the changes what happens in the OLTP)

prod_dim_key	prod_id	prod_name	prod_price	prod_cost	prod_cat_desc	prod_family_desc
100	25	PEN	60	36	Stationary	Office Supplies
101	26	PENCIL	40	28	Stationary	Office Supplies
102	27	Electronic Board	4000	3000	Electronics	Office Supplies
103	28	Computer Table	2100	1900	Furniture	Office Supplies
104	29	Paper Clip	55	35	Furniture	Office Supplies
108	30	Pen Holder	100	80	Stationary	Office Supplies
109	26	PENCIL	50	28	Stationary	Office Supplies
110	31	Wireless Mouse	850	800	Electronics	Office Supplies

Show me all the prod_names and the current price of the product

Type 3 (implementation)



Fact Table

- Consists of measures
- Central table in the dimensional model
 - Most of the transaction table in OLTP systems are fact table
 - most are numeric column
 - narrow row few columns
 - Large number of rows in fact table
 - Access via dimensions
 - All the dim keys or sur id will be foreign key
 - In retail examples of fact table are sales_fact, return_fact, purcha_order_fact, invoice_fact etc.
 - In HRMS payroll , timehseet.taining session ,

- In customer service outbound calls, inbounds calls, service calls.

Measures

- Measure is column in fact table column in which we can know some case of transaction
- other than pk and fk all are measures
- 3 types
- Fully additive - can apply aggregate fun with any dimension in star schema
- semi additive - cant apply all aggregate fun in this type(ex percentage) age is semi additive
- non additive - cant apply any aggregate function in this type(ex trans_type = "S" or "R") eg gender (M or F)

-
- Fully additive measure
 - Where we can apply any aggregate functions with any dimensions in that star schema
 - Semi additive measure
 - We wont be able to apply all aggregate functions in this type of column (Ex: percentage)
 - Non additive measure
 - We wont be able to apply any aggregate functions in this type of column (Ex: Tran_type = 'S' or 'R')
 - Sales_fact
 - Qty -- min, max, avg, count, sum →
 - Price_Per_Qty -- min, max, avg, count → making sense (sum → not making sense)
 - Discount
 - Profit_percentage -- sum (profit_percentage) → does not makes sense
 - Profit --
 - Amount -- Fully aggregate
 - Call center business
 - Call_id
 - Call_date
 - Customer
 - Case_type
 - Comment (.....) – these kind of measures in the fact table is called as non additive measure

- Foot fall is physical count or record of a store

The Snowflake schema

- A normalized version of star schema
- dimension + fact (does not change) + look up
- won't have much of redundant data mostly have a loop up
- Number of joins become more (impacts performance of query)
- Dimension minimum have one hierarchy, Some dimension have multiple hierarchy
- Analyze end user req and space constraints to pick the best
- Storing historical data becomes a problem in snowflake

Fact Table


- Consists of pk and fk
- pk - unique
- FK -- dimensional table (PK)
- remaining columns are called as measures

Fact less Fact table

You are viewing Ibridge 360's screen

View Options

Fact Less Fact tables



- Usually fact table has all the measures through which you measure the performance of the business.
- Typical fact table has qty, cost, price, discount, margin columns where you can apply aggregate functions.
- In some cases, we won't be having these kind of measures ex: any event tracking DW
- Web click analysis, attendance system etc are the examples
 - Website, User, Time, Page (Dim)
 - Webclick_id, website_id, user_id, period_id, page_id (webclick_fact)
 - Employee, time, subject, student, schedule
 - att_id, employee_id, time_id, sub_id, stud_id, sche_id (att_fact)
 - fee_id, stud_id, time_id, course_id, amount (Fee_fact) - fact tables with measures
- In the above mentioned systems, the existence of the record becomes the fact. So in web click analysis the combination of user, date and webpage is the fact.

OLAP

What is OLAP ?

- viewing data in multi dim way

Why OLAP ?

- Slice and dice for data warehouse
- Pick what we want slice , is by applying where clause
- Dice is grouping mechanism
- OLAP is multi dim way of storing and viewing data
- RDBMS is a 2 dim way of storing and viewing data
- Analysis purpose OLAP
- Consolidated data to be created before OLAP is data ware house and database comes in picture