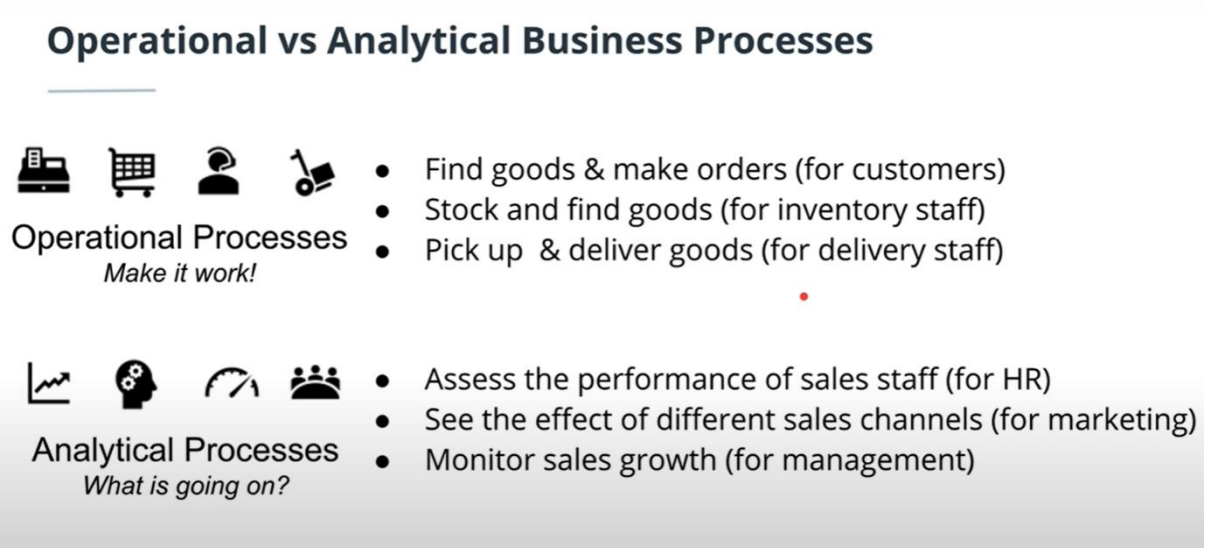
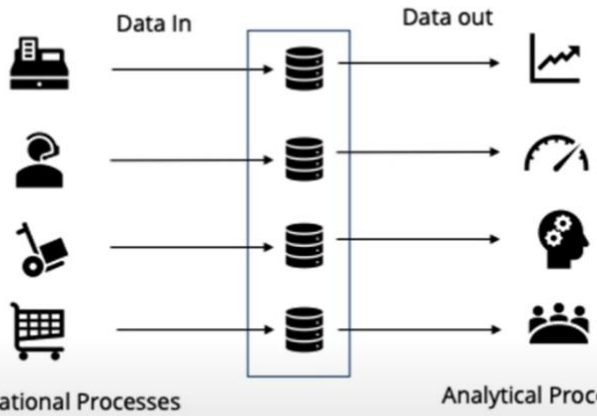
What Is A Data Warehouse? A Business Perspective

You are in charge of a retailer's data infrastructure. Let's look at some business activities.

- Customers should be able to find goods & make orders
- Inventory Staff should be able to stock, retrieve, and re-order goods
- Delivery Staff should be able to pick up & deliver goods
- HR should be able to assess the performance of sales staff
- Marketing should be able to see the effect of different sales channels
- Management should be able to monitor sales growth
  Ask yourself: Can I build a database to support these activities? Are all of the above questions of the same nature?
  Let's take a closer look at details that may affect your data infrastructure.
- Retailer has a nation-wide presence → **Scale?**
- Acquired smaller retailers, brick & mortar shops, online store → **Single database? Complexity?**
- Has support call center & social media accounts → **Tabular data?**
- Customers, Inventory Staff and Delivery staff expect the system to be fast & stable → **Performance**
- HR, Marketing & Sales Reports want a lot information but have not decided yet on everything they need → **Clear Requirements?**
  Ok, maybe one single relational database won't suffice :)



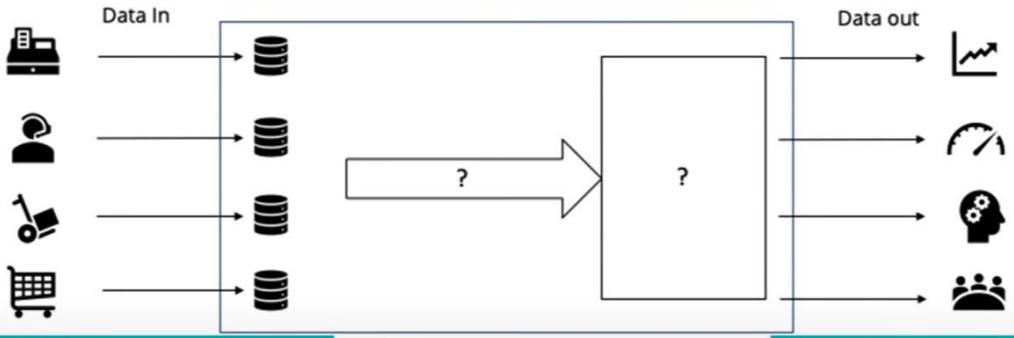
Same data source for operational & analytical processes?



Operational Databases
- Excellent for operations
- No redundancy, high integrity

Operational Databases
- Too slow for analytics, too many joins
- Too hard to understand

Solution: Create 2 processing modes, Create a system for them to co-exist



OLTP: online **transactional** processing

OLAP: online **analytical** processing

Data Warehouse is a system (including processes, technologies & data representations) that enables us to support analytical processes

What is Data Ware Housing?

Tech Perspective: DWH Definition 1

A data warehouse is a copy of transaction data specifically structured for query and analysis.

[REF:KIMBALL]

Tech Perspective: DWH Definition 2

A data warehouse is a **subject-oriented**, **integrated**, **nonvolatile**, and **time-variant** collection of data in support of management's decisions.

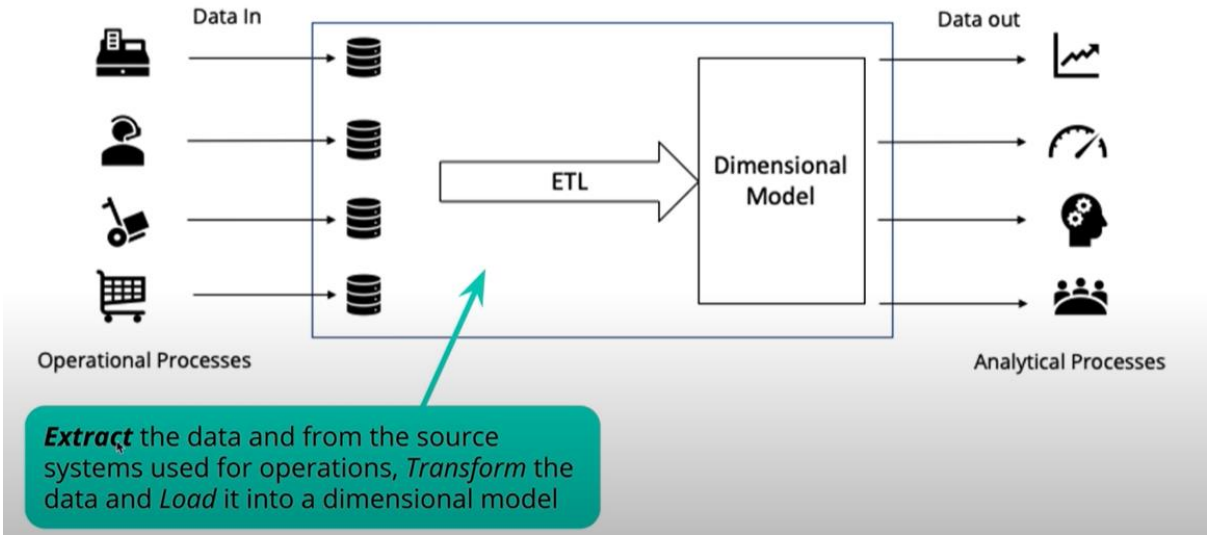
[REF:INMON]

Tech Perspective: DWH Definition 3

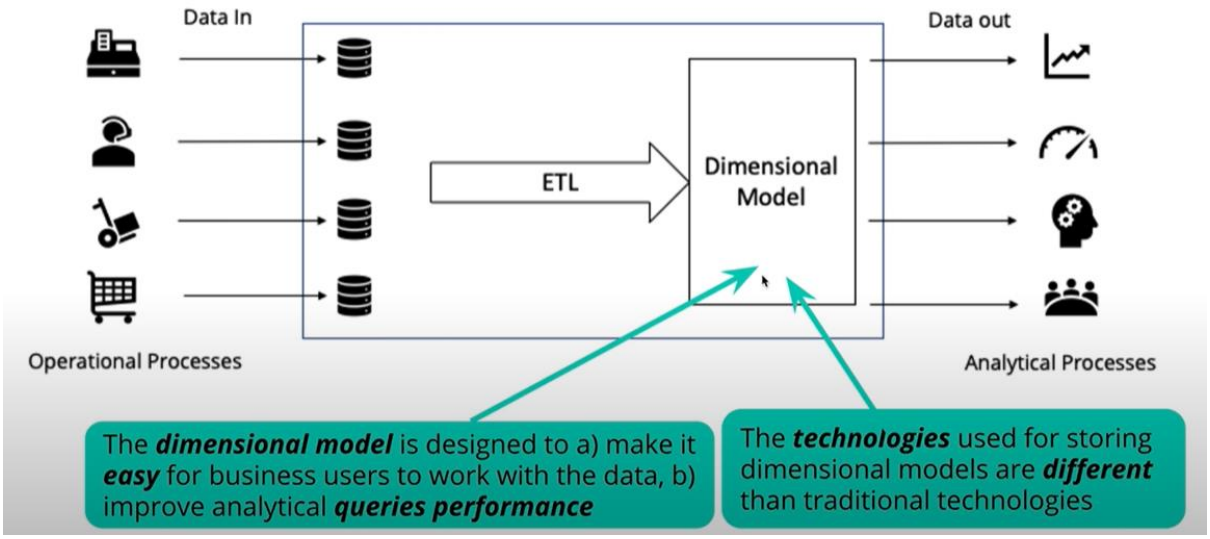
A data warehouse is a system that **retrieves** and **consolidates** data **periodically** from the source systems into a **dimensional** or **normalized** data store. It usually **keeps years of history** and is **queried for business intelligence** or other **analytical activities**. It is typically **updated in batches**, not every time a transaction happens in the source system.

[REF:RAINARDI]

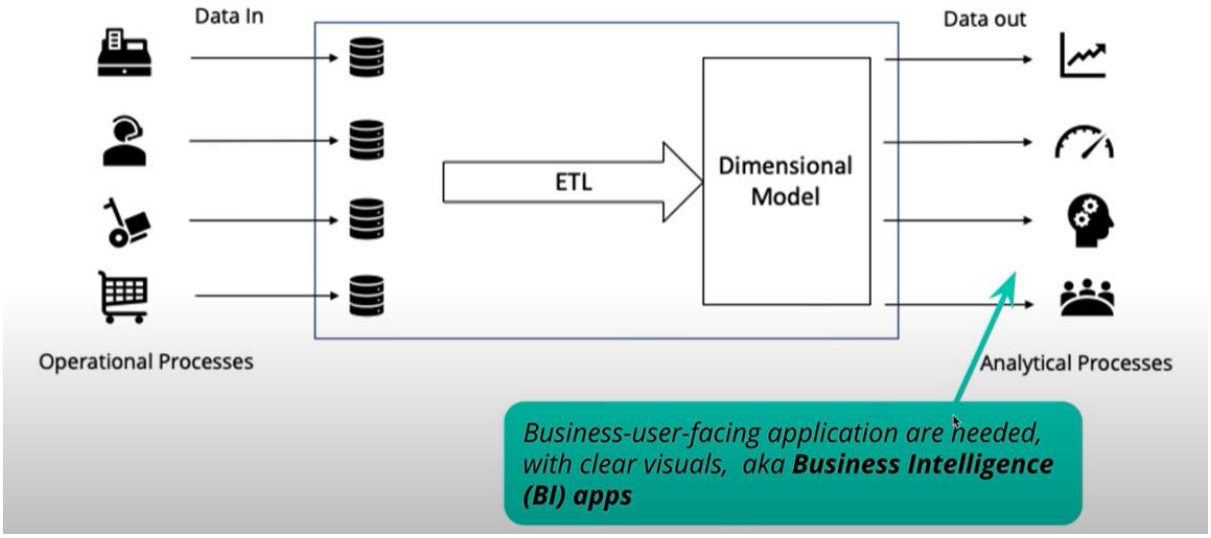
DWH: Tech Perspective



DWH: Tech Perspective



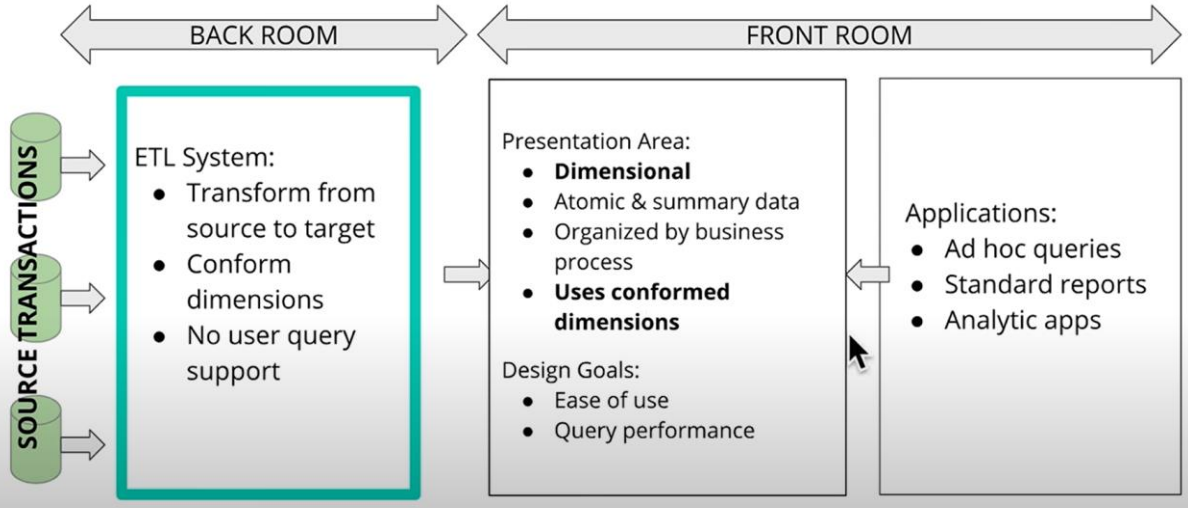
DWH: Tech Perspective



Data Warehouse Goals

- Simple to understand
- Performant
- Quality Assured
- Handles new questions well
- Secure

Kimball's Bus Architecture



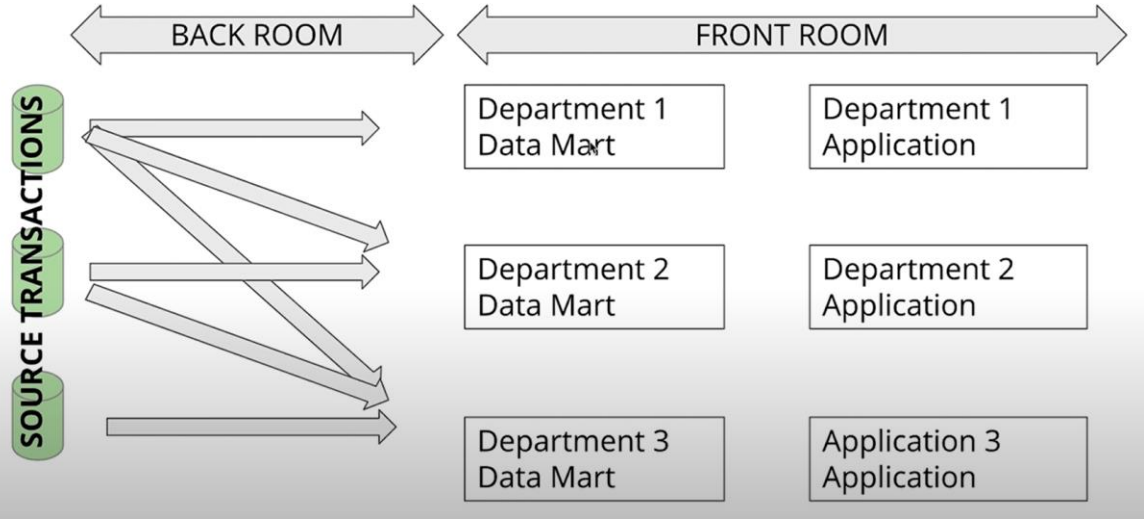
Kimball's Bus Matrix

Business processes	Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issue purchase order	X	X	X				
Receive warehouse deliveries	X	X	X				X
Store inventory	X	X	X	X			
Retail sales	X	X	X	X	X	X	X
....							

ETL: A Closer Look

- ETL:
  - Extracting:
    - Get the data from its source
    - Possibly deleting old state
  - Transforming:
    - Integrates many sources together
    - Possibly cleansing: inconsistencies, duplication, missing values, etc..
    - Possibly producing diagnostic metadata
  - Loading:
    - Structuring and loading the data into the dimensional data model

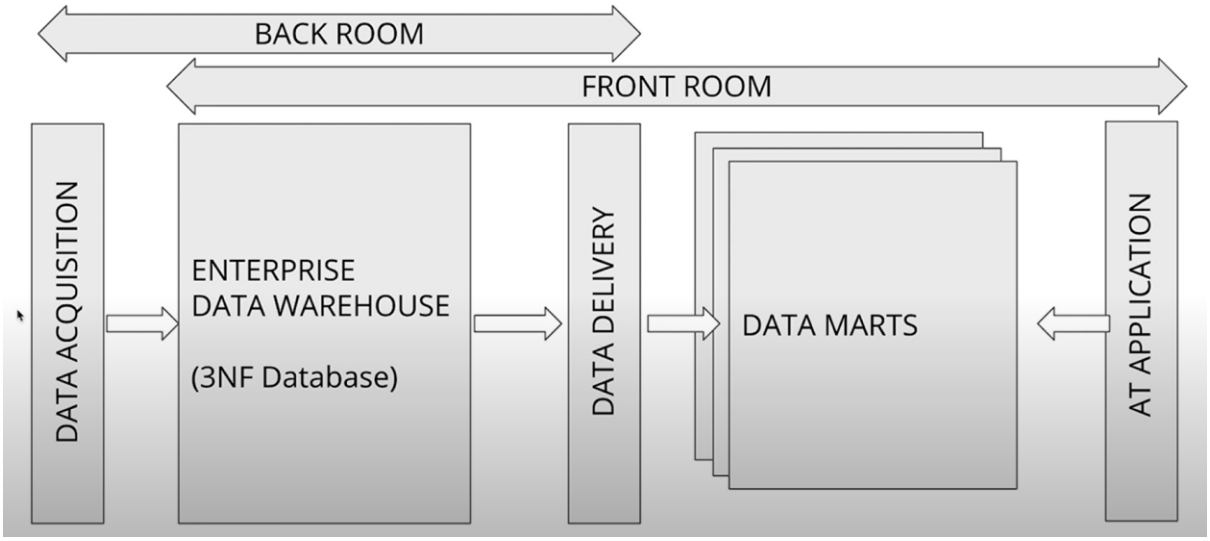
Independent Data Marts



Independent Data Marts

- Departments have independent ETL processes & dimensional models
- These **separate & smaller** dimensional models are called "Data Marts"
- Different fact tables for the same events, **no conformed dimensions**
- Uncoordinated efforts can lead to **inconsistent views**
- Despite awareness of the emergence of this architecture from departmental autonomy, it is generally discouraged

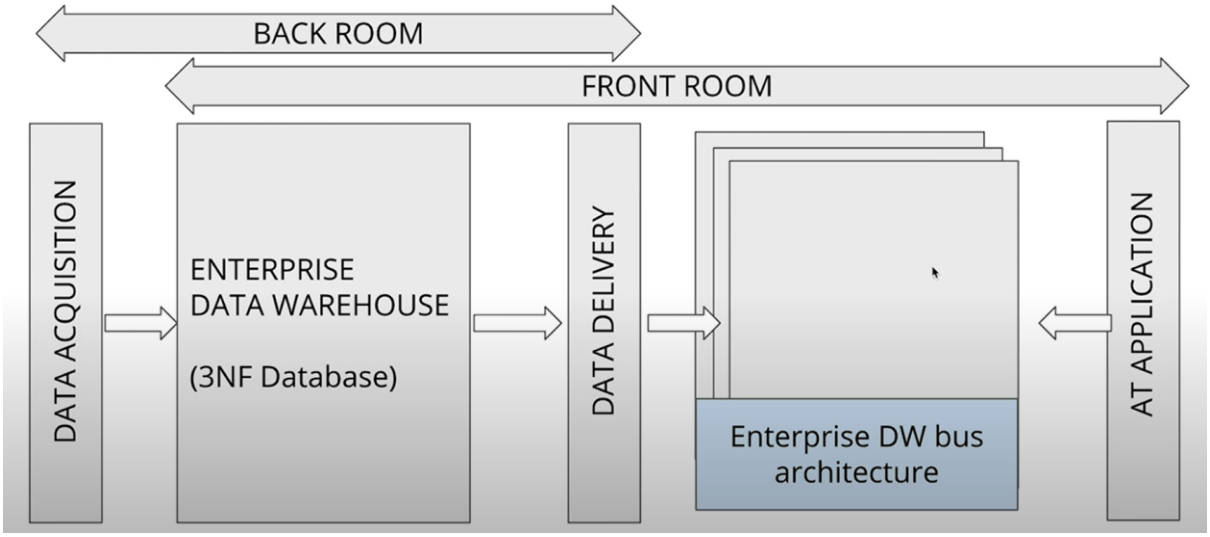
Inmon's Corporate Information Factory (CIF)



Inmon's Corporate Information Factory (CIF)
Data Marts

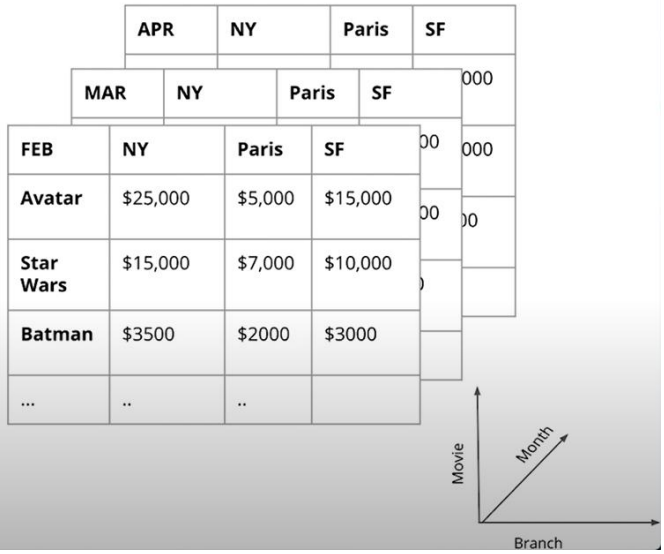
- 2 ETL Process
  - Source systems → 3 NF DB
  - 3 NF DB → Departmental Data Marts
- The 3NF DB acts an enterprise wide data store.
  - Single integrated source of truth for data-marts
  - Could be accessed by end-users if needed
- Data marts dimensionally modelled & unlike Kimball's dimensional models, they are mostly aggregated

Hybrid Kimball Bus & Inmon CIF



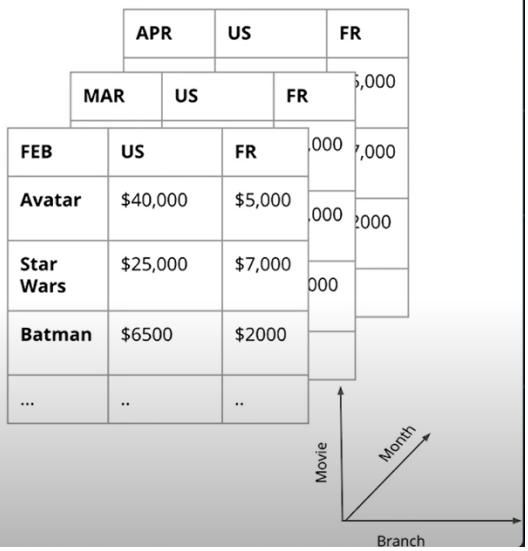
OLAP Cubes

- An OLAP cube is an aggregation of a fact metric on a number of dimensions
- E.g. Movie, Branch, Month
- Easy to communicate to business users
- Common OLAP **operations** include: **Rollup, drill-down, slice, & dice**



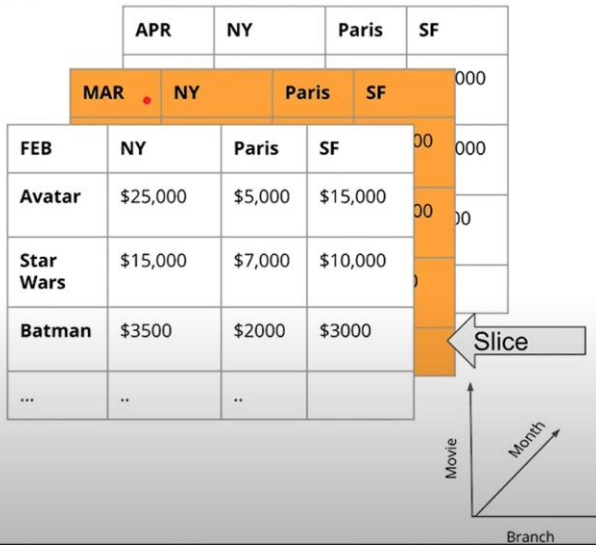
OLAP Cubes Operations: Roll-up & Drill Down

- **Roll-up**: Sum up the sales of each city by Country: e.g. US, France (less columns in branch dimension)
- **Drill-Down**: Decompose the sales of each city into smaller districts (more columns in branch dimension)
- The OLAP cubes should store the finest grain of data (atomic data), in case we need to drill-down to the lowest level, e.g. Country → City → District → Street, etc..



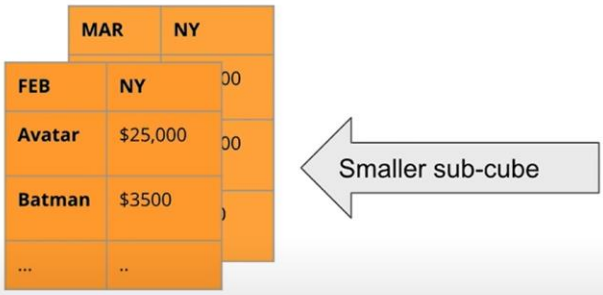
OLAP Cubes Operations: Slice

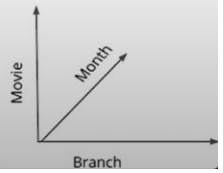
- Reducing N dimensions to N-1 dimensions by restricting one dimension to a single value
- E.g. month='MAR'



OLAP Cubes Operations: Dice

- Same dimensions but computing a sub-cube by restricting, some of the values of the dimensions
- E.g. month in ['FEB', 'MAR'] and movie in ['Avatar', 'Batman'] branch = 'NY'





OLAP Cubes query optimization

- Business users will typically want to slice, dice, rollup and drill-down all the time
- Each such combination will potentially go through all the facts table (suboptimal)
- The "**GROUP by CUBE (movie, branch, month)**" will make **<u>one</u> pass through the facts** table and will aggregate all possible combinations of groupings, of length 0, 1, 2 and 3 e.g:
  - Total revenue
  - Revenue by movie
  - Revenue by branch
  - Revenue by month
  - Revenue by movie, branch
  - Revenue by branch, month
  - Revenue by movie, month
  - Revenue by movie, branch, month
- Saving/Materializing the output of the CUBE operation and using it is usually enough to answer all forthcoming aggregations from business users without having to process the whole facts table again

The Last Mile: Delivering the analytics to users

Data is available...

- In an understandable & performant dimensional model
- With *Conformed Dimensions* or separate *Data Marts*
- For users to report and visualize
  - By interacting directly with the model
  - Or in most cases, through a BI application

The Last Mile: Delivering the analytics to users

OLAP cubes is a very convenient way for slicing, dicing and drilling down

How do we serve these OLAP cubes?

OLAP cubes technology

Approach 1: **Pre-aggregate** the OLAP cubes and saves them on a special purpose non-relational database (**MOLAP**)

Approach 2: Compute the OLAP cubes **on the fly** from the existing relational databases where the dimensional model resides (**ROLAP**)

Demo: Column format in ROLAP

- Use a postgresql with a columnar table extension
- Load a dataset in a normal table
- Load the same dataset in a columnar table
- Compare the performance of the fact-aggregating query performance in both tables