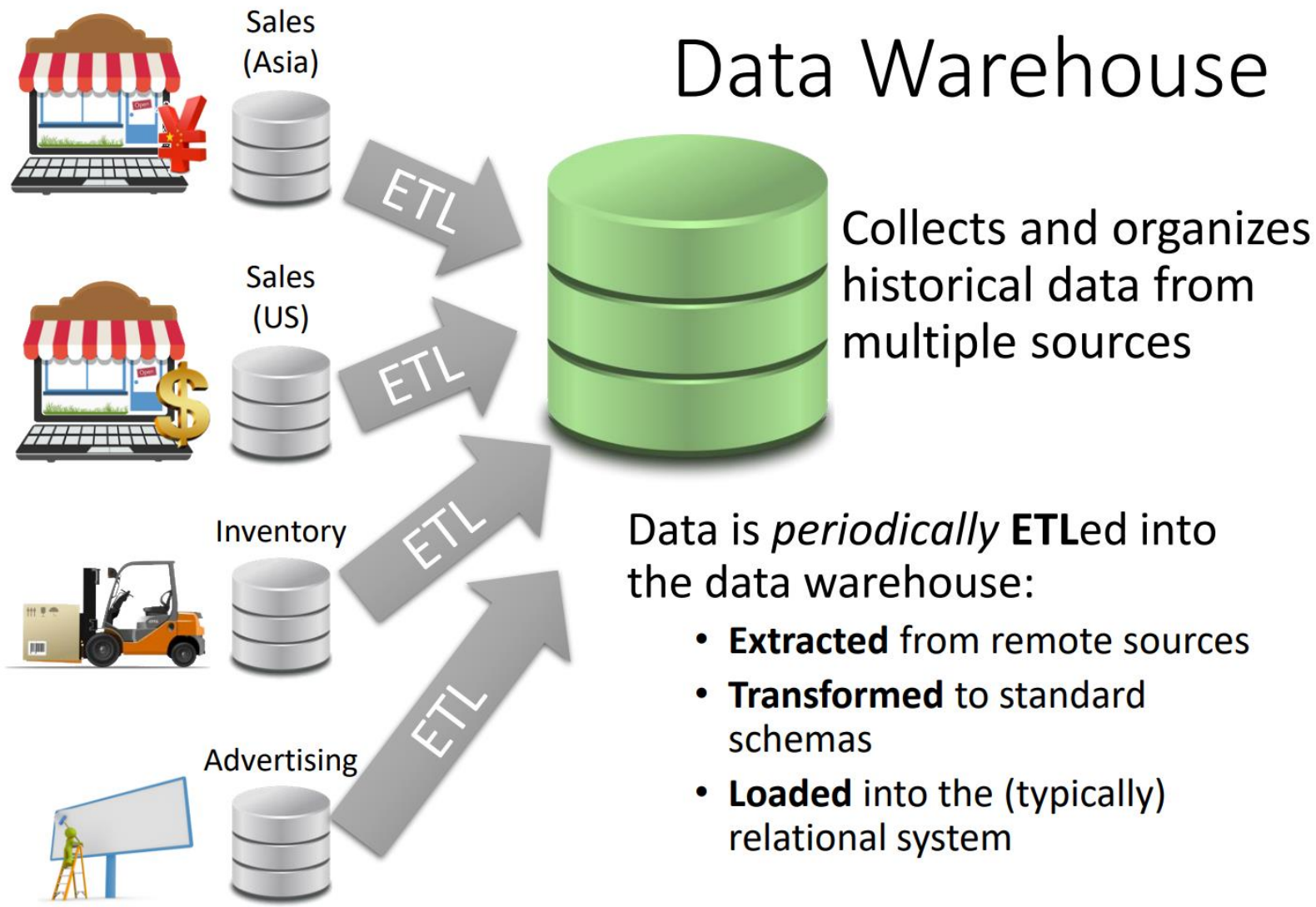
Discussion 13

Data Mining and ML

Outline:

- Data Warehouse
- Clustering
- Regression

Data Warehouse



Data Warehouse

Multidimensional Data Model

Sales **Fact Table**

pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
12	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35
11	2	2	22
11	3	2	10
12	1	2	26

Locations

locid	city	state	country
1	Omaha	Nebraska	USA
2	Seoul		Korea
5	Richmond	Virginia	USA

Dimension Tables

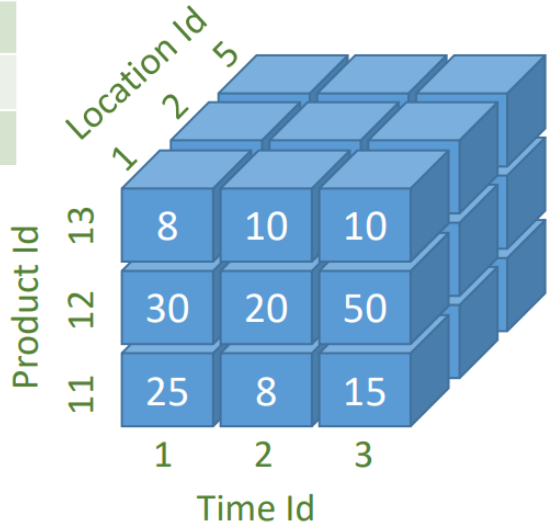
Products

pid	pname	category	price
11	Corn	Food	25
12	Galaxy 1	Phones	18
13	Peanuts	Food	2

Time

timeid	Date	Day
1	3/30/16	Wed.
2	3/31/16	Thu.
3	4/1/16	Fri.

- Multidimensional "Cube" of data



- Sales Fact Table
  - Contains only foreign keys → Efficient
- Easy to manage Dimensions
  - Galaxy1 → Phablet: no need to update **Fact Table**
- Normalization
  - Minimizing redundancy
  - More on this later …

Data Warehouse

Cross Tabulation (Pivot Tables)

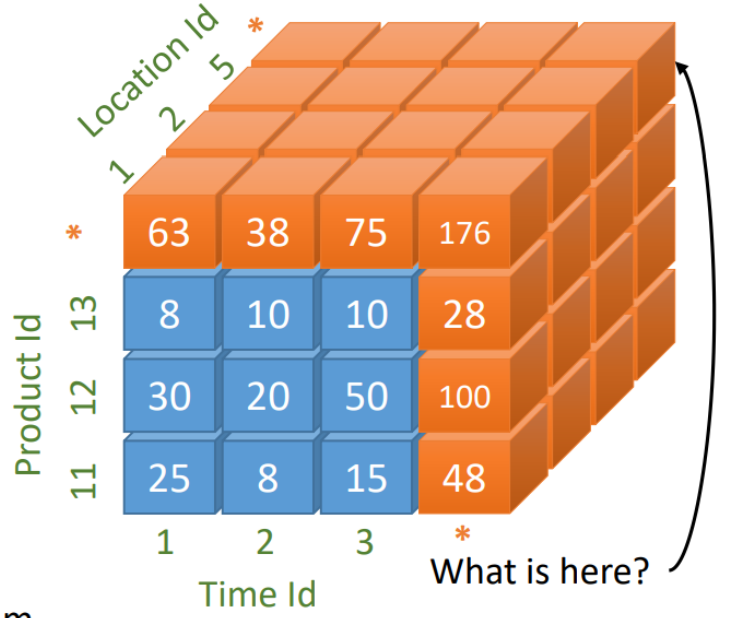
Item	Color	Quantity
Desk	Blue	2
Desk	Red	3
Sofa	Blue	4
Sofa	Red	5

		Item		
		Desk	Sofa	*Sum*
Color	Blue	2	4	***6***
	Red	3	5	***8***
	Sum	***5***	***9***	***14***

Data Warehouse

Cube Operator

- Generalizes cross-tabulation to higher dimensions.



➢In SQL:

SELECT Item, Color, **SUM**(Quantity) **AS** QtySum
FROM Furniture
GROUP BY ***CUBE*** (Item, Color);

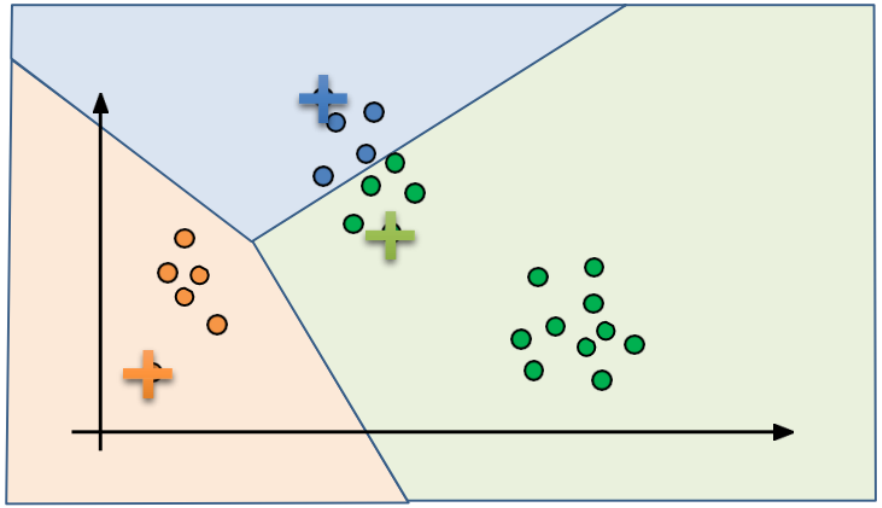
Item	Color	Quantity
Desk	Blue	2
Desk	Red	3
Sofa	Blue	4
Sofa	Red	5

Item	Color	QtySum
Desk	Blue	2
Desk	Red	3
Desk	*	5
Sofa	Blue	4
Sofa	Red	5
Sofa	*	9
*	*	14
*	Blue	6
*	Red	8

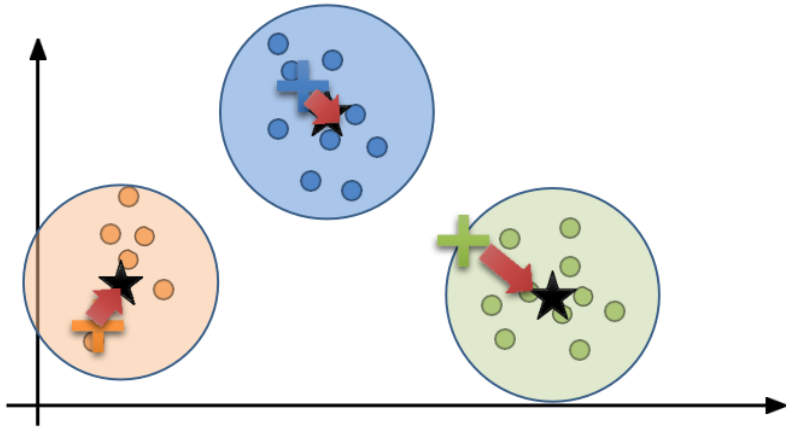
Clustering

K-Means Clustering

Compute Assignments



Update Centers



Clustering

Res-A: weighted reservoir sampling

- **Goal:** *Sample **k** records from a stream where record **i** is included in the sample with probability proportional to w_i*
- **Algorithm:**
  - For each record ***i*** draw a uniform random number:

$$u_i \sim \mathbf{Unif}(0, 1)$$

  - Select the top-k records ordered by: u_i^{1/w_i}

- **Common ML Pattern?**
  - **Query Function:** `[pow(rand(), 1 / record.w), record]`
  - **Agg. Function:** *top-k heap*

Regression

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$