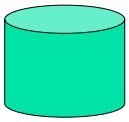
# Business Intelligence (BI) Overview

1

## What is Business Intelligence?

Business Intelligence enables the business to make intelligent, fact-based decisions

| **Aggregate Data** | **Present Data** | **Enrich Data** | **Inform a Decision** |
|---|---|---|---|
|  |  |   |  |
| Database, Data Mart, Data Warehouse, ETL Tools, Integration Tools | Reporting Tools, Dashboards, Static Reports, Mobile Reporting, OLAP Cubes | Add Context to Create Information, Descriptive Statistics, Benchmarks, Variance to Plan | Decisions are Fact-based and Data-driven |

2

## CPU – Content, Performance, Usability

- Content
  - The business determines the "what", BI enables the "how"
- Performance
  - Minimize report creation and collection times (near zero)
- Usability
  - Delivery Method →Push vs Pull
  - Medium → Excel, PDF, Dashboard, Cube, Mobile Device
  - Enhance Digestion → "A-ha" is readily apparent, fewer clicks
  - Tell a Story → Trend, Context, Related Metrics, Multiple Views

3

## How Important is BI?

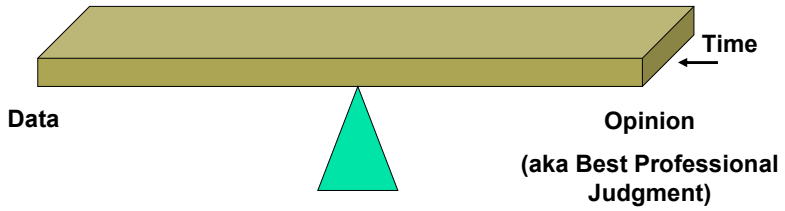
Top 10 Business and Technology Priorities for 2011:

1. Cloud computing
2. Virtualization
3. Mobile technologies
4. IT Management
5. **Business Intelligence**
6. Networking, voice and data communications
7. Enterprise applications
8. Collaboration technologies
9. Infrastructure
10. Web 2.0

**Source: Gartner's 2011 CIO Agenda (aka "Reimagining IT: The 2011 CIO Agenda").**

4

## Why is Business Intelligence So Important?



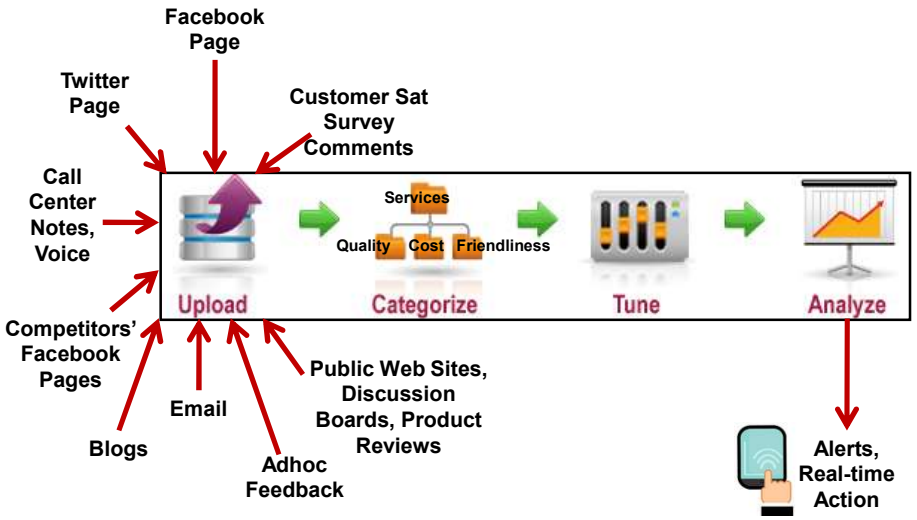

Making Business Decisions is a Balance

In the absence of data, business decisions are often made by the HiPPO.

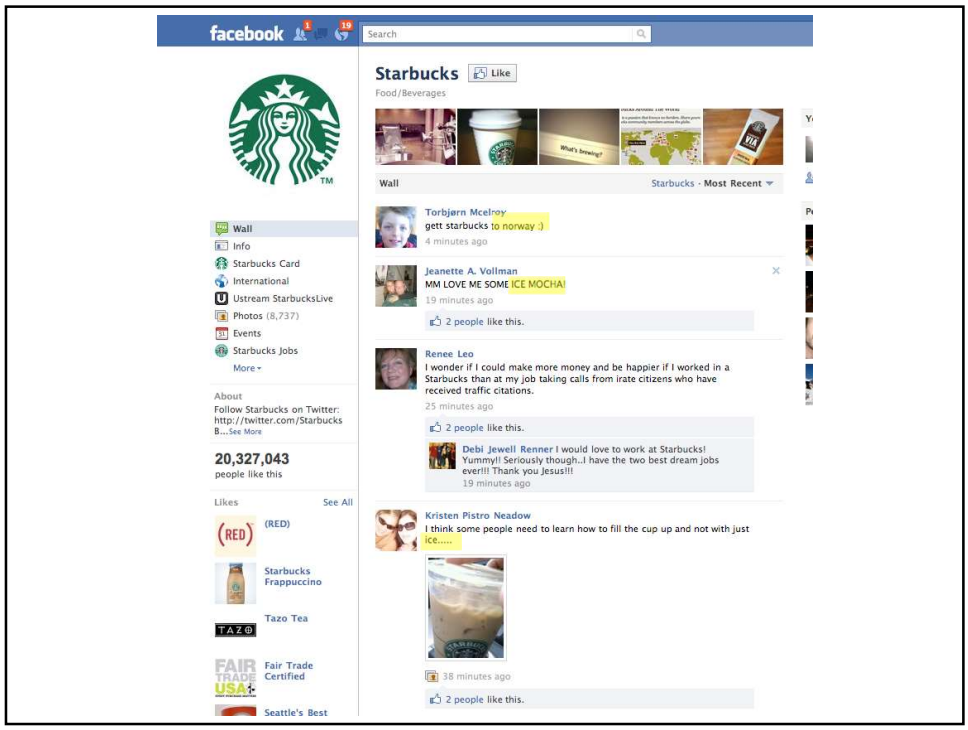
With Business Intelligence, we can get data to you in a timely manner.

5

## Unstructured Text Processing




6


facebook
Search
Starbucks
Like
Food/Beverages
Wall
Starbucks - Most Recent
Wall
Info
Starbucks Card
International
Ustream StarbucksLive
Photos (8,737)
Events
Starbucks Jobs
More
About
Follow Starbucks on Twitter: http://twitter.com/Starbucks
B...See More
20,327,043
people like this
Likes
See All
(RED)
Starbucks Frappuccino
Tazo Tea
Fair Trade Certified
Seattle's Best
Torbjørn Mcelroy
gett starbucks to norway :)
4 minutes ago
Jeanette A. Vollman
MM LOVE ME SOME ICE MOCHA!
19 minutes ago
2 people like this.
Renee Leo
I wonder if I could make more money and be happier if I worked in a Starbucks than at my job taking calls from irate citizens who have received traffic citations.
25 minutes ago
2 people like this.
Debi Jewell Renner I would love to work at Starbucks! Yummy!! Seriously though..I have the two best dream jobs ever!!! Thank you Jesus!!!
19 minutes ago
Kristen Pistro Neadow
I think some people need to learn how to fill the cup up and not with just ice.....
38 minutes ago
2 people like this.

7


facebook
Search
McDonald's
Like
Food/Beverages
Wall
i'm lovin' it
new McCafé SHAKES
Wall
Info
Local
LATEST
McDonald's
Today is Respect Your Cat Day!
Celebrate today
Source: community.mcdonalds.com
And in honor of your furry friend (or even just the neighbor's), celebrate with a new McCafé Shake – now with real whipped cream and a cherry on top.
23 hours ago via McDonald's · Share
1,433 people like this.
View all 341 comments
McDonald's
It's Make Up Your Own Holiday Day! What will you celebrate?
Celebrate today
Source: community.mcdonalds.com
Whatever holiday you create, celebrate it with a new McCafé Shake – now topped off with real whipped cream and a cherry.
Saturday at 9:00am via McDonald's · Share
1,211 people like this.
View all 471 comments

8


9

# BI Technologies

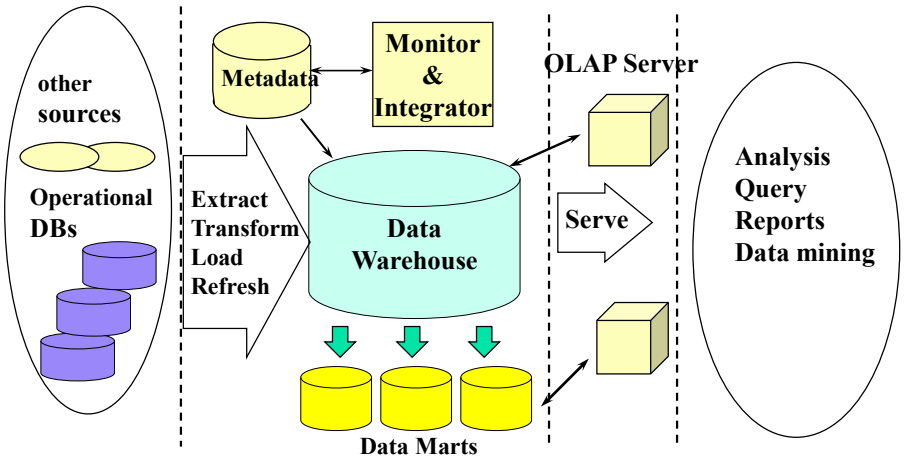
- Analytic Databases

**DB2**
**Oracle**
**SQL Server** → **Teradata**
**Netezza** → **Vertica**
**Aster Data**
**Par Accel**
**Greenplum** → **Semantic Databases (TIDE)**

- BI is a consolidating industry
  - Oracle: Siebel, Hyperion, Brio, Sun
  - SAP: Business Objects, Sybase
  - IBM: Cognos, SPSS, Coremetrics, Unica, **Netezza**
  - EMC: **Greenplum**
  - HP: **Vertica**
  - Teradata: **Aster Data**
  - **.....**
- Independent vendors: MicroStrategy, Informatica, SAS
- Reporting standards determined mainly by Microsoft, Apple and Adobe

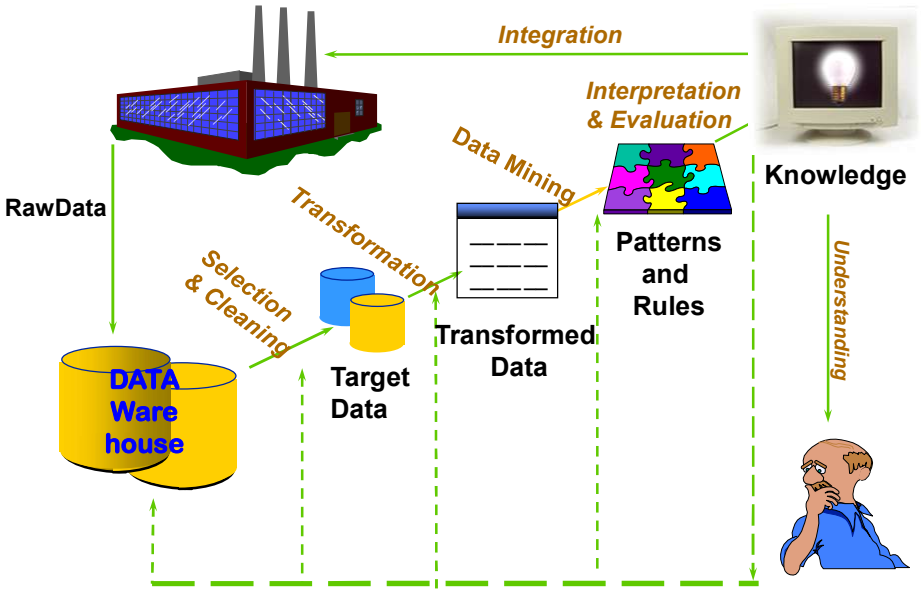
10

## Multi-Tiered Architecture




11

## Knowledge Discovery Process




12

# DATA PREPROCESSING

13

## Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking *attribute values*, lacking certain *attributes of interest*, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
  - Required for both OLAP and Data Mining!

14

## Why can Data be Incomplete?

- Attributes of interest are not available (e.g., customer information for sales transaction data)
- Data were not considered important at the time of transactions, so they were not recorded!
- Data not recorder because of misunderstanding or malfunctions
- Data may have been recorded and later deleted!
- Missing/unknown values for some data

15

## Why can Data be Noisy/Inconsistent?

- Faulty instruments for data collection
- Human or computer errors
- Errors in data transmission
- Technology limitations (e.g., sensor data come at a faster rate than they can be processed)
- Inconsistencies in naming conventions or data codes (e.g., 2/5/2002 could be 2 May 2002 or 5 Feb 2002)
- Duplicate tuples, which were received twice should also be removed

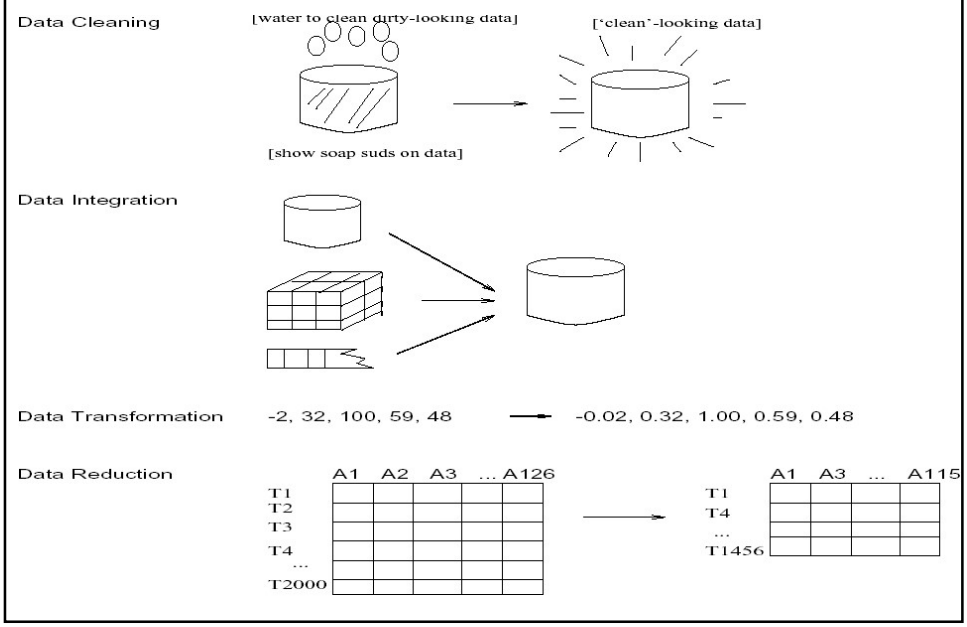
16

# Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

17

# Forms of data preprocessing




18

## Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

19

## How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification)—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

20

# How to Handle Missing Data?

| Age | Income | Team | Gender |
|---|---|---|---|
| 23 | 24,200 | Red Sox | M |
| 39 | ? | Yankees | F |
| 45 | 45,390 | ? | F |

**Fill missing values using aggregate functions (e.g., average) or probabilistic estimates on global value distribution**

**E.g., put the average income here, or put the most probable income based on the fact that the person is 39 years old**

**E.g., put the most frequent team here**

21

# How to Handle Noisy Data? Smoothing techniques

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - computer detects suspicious values, which are then checked by humans
- Regression
  - smooth by fitting the data into regression functions

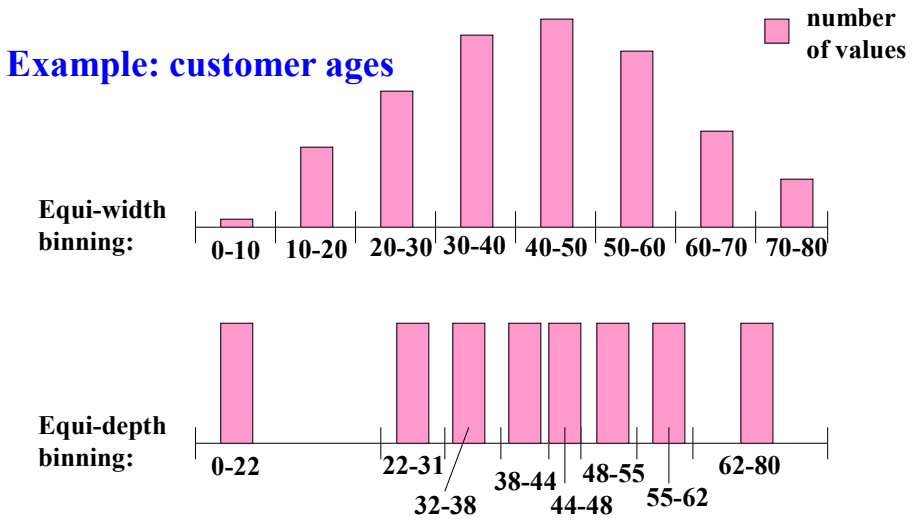
22

## Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling – good handing of skewed data

23

## Simple Discretization Methods: Binning




24

# Smoothing using Binning Methods

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries: [4,15],[21,25],[26,34]
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34