03. Data Warehousing Fundamentals

1

What is Data Warehousing?

- A data warehouse is a database designed to enable business intelligence activities:
  - It exists to help users understand and enhance their organization's performance.
  - It is designed for query and analysis rather than for transaction processing, and usually contains historical data derived from transaction data
- A data warehouse environment can include an
  - Extraction, transportation, transformation, and loading (ETL) solution
  - Statistical analysis
  - Reporting
  - Data mining capabilities and
  - Applications that manage the process of gathering and transforming data into useful, actionable information, and delivering it to business users.

2

Data warehousing:

- The data warehouse works with data collected from multiple sources:
  - Internally developed systems
  - Purchased applications
  - Third-party data syndicators and other sources
- A data warehouse key goal:
  - To become the organization's "single source of truth";
  - To become the consistent source of data that all users can look to
- The data warehouse
  - Holds many months/years of data to support historical analysis.
  - Acquires the data from multiple data sources through an ETL process.
- Defining the ETL process is a very large part of the design effort of a data warehouse

3

Data Warehouse vs Database

- DW constitutes the entire information base for ALL time
- DW supports Data Mining and Business Intelligence
- Database relates to Real Time Information, running the business

4

Questions from the Business:

5

Common Data 'Situations'

- There is data everywhere … But:
  - I can't find the data I need
    - Data is scattered all over the network
    - Many versions of the data exist … with some differences
  - I can't get the data I need
    - I need an expert to get the data
  - I can't understand the data I found
    - The data is poorly documented
  - I can't use the data I found
    - The results are unexpected
    - Data needs to be transformed from one form to the other

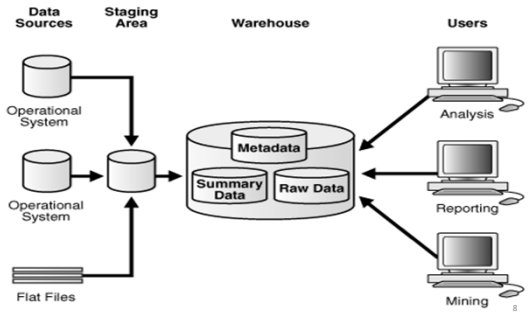
6

… then, What's a Data Warehouse?

- A single, complete and consistent store of data obtained from a variety of different sources made available to users in a form understandable and usable in a business context
- A copy of transactional data specifically structured for querying and reporting
- Data Warehousing = a process of transforming data into information and making it available to users in a timely enough manner to make a difference

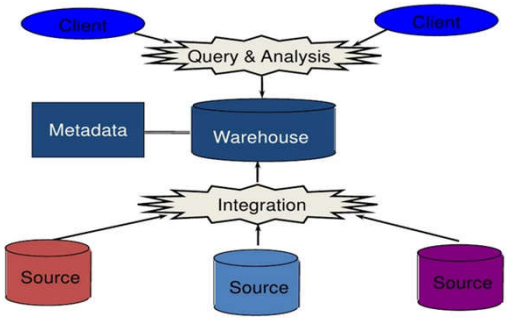
7

Data Warehouse Architecture



8

Data Warehouse Architecture



9

Data Warehouse Metadata

- 'Data about data', important for orderly construction, retrieval and control of warehouse data

- **Technical Metadata**
  - Where the data comes from
  - How the data changes
  - How the data is organized
  - Who owns the data
  - Who's responsible for the data
  - Who can access the data
  - …
- **Business Metadata:**
  - What data are available
  - Where the data are
  - What the data mean
  - How to access
  - Predefined reports and queries

10

Data Warehouse Characteristics

- Subject Oriented
  - Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a data warehouse that concentrates on sales. Using this data warehouse, you can answer questions such as "Who was our best customer for this item last year?" or "Who is likely to be our best customer next year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented
- Integrated
  - Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

11

Data Warehouse Characteristics

- Nonvolatile
  - Nonvolatile means that, once entered into the data warehouse, data should not change. This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred
- Time Variant
  - A data warehouse's focus on change over time is what is meant by the term time variant. In order to discover trends and identify hidden patterns and relationships in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive.

12

DW environment characteristics

- Data is structured for simplicity of access and high-speed query performance
- End users are time-sensitive and desire speed-of-thought response times
- Large amounts of historical data are used
- Queries often retrieve large amounts of data, perhaps many thousands of rows
- Both predefined and ad hoc queries are common
- The data load involves multiple sources and transformations.

13

Data Warehouse Example Analyses:

- Consolidation of last year's sales figures
- Inventory analysis
- Profit by product and by customer
- Slice and Dice operations
- Highly aggregated data … with possibility to drill down
- Trend analyses
- Data Mining
- Serving reports, dashboards and other interfaces to end-users

14

Related Ideas

- OLTP: Systems tuned for known transactions and workloads
- Data Mart: A subset of organizational data store usually oriented to a specific purpose or a major data subject
- Data Mining: application of tools and processes on data to yield: associations, sequences, classifications, clusters, forecasting … often revealing unexpected, valuable results
- OLAP: Sifting through data to analyze complex relationships in search of patterns, trends and exceptions. Term frequently associated to with multidimensional databases and terms such as Decision support, Business Intelligence, Executive Information Systems

15

OLTP vs Data Warehousing

OLTP

- Application Oriented
- Used to run business
- Detailed data
- Current; up-to-date
- Isolated data
- Clerical user
- Read/update access
- No redundancy
- Performance metric: transaction throughput

Data Warehousing

- Subject oriented
- Used to analyze business
- Summarized and refined
- Snapshot data
- Integrated data
- Knowledge user (manager)
- Mostly read access
- Redundancy present
- Performance metric: query throughput

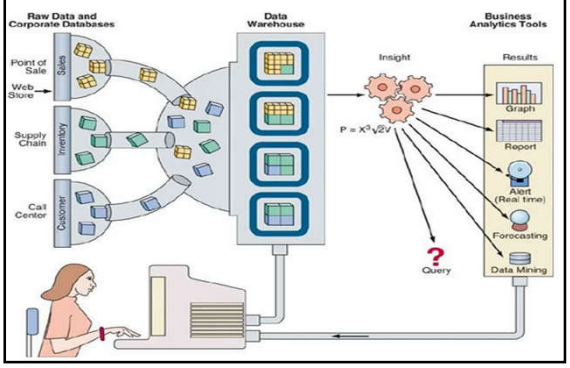
16

Data Mart

- A scaled down version of the data warehouse
- An analytical data store to focus on specific business functions or for a specific group within the organization
- Justification:
  - Provides easy access to frequently needed data
  - Features improved end-user response time
  - Ease of creation … in less time
  - Lower cost than implanting a full DW
  - Potential users are more clearly defined

17

Business Intelligence and DW



What is … BI

- Skills, knowledge, technologies, applications, quality, risks, security issues and practices used to help a business acquire better understanding of market behavior and commercial context.
- Includes:
  - collection
  - integration
  - analysis
  - interpretation and
  - presentation

of business information

19

What is BI?

- "The processes, technologies and tools needed to turn data into information and information into knowledge and knowledge into plans that drive profitable business action. BI encompasses data warehousing, business analytics and knowledge management."
  - The Data Warehouse Institute
- Business Intelligence is defined as "knowledge gained about a business through the use of various hardware/software technologies which enable organizations to turn data into information"
  - Data Management Review

20

Just What is BI?

- Business Intelligence (BI) usually refers to technologies, applications, and practices for the collection, integration, analysis, and presentation of business information
- BI systems provide historical, current, and predictive views of business operations, most often using data that has been gathered into a data warehouse or a data mart and occasionally working from operational data

21

Why BI?

Business Intelligence provides managers with the facts they need to make better decisions for competitive advantage:

Before	After
•Lots of data, few insights, information silos	•Insights delivered from a consistent source
•Information architecture consists of a 1000+ spreadsheets	•Fit for purpose information architecture that reflects key business questions
•Sending around spreadsheets to communicate performance	•Easy access to personalised on
•Lengthy process to pull together data from different systems into a coherent report	•line performance information
•There isn't a single version of the truth, managers don't trust reports	•Rapid process to pull together data from different systems into a coherent report
•Decisions made largely on gut feel	•Advancing towards a single version of the truth, managers trust reports

BI is about recognising that a company's data is an important strategic asset that can yield management information and the business capturing, storing, accessing, analysing and using that information to improve decision making

22

BI Capabilities

23

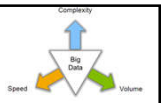
… And… What then is … Big Data?

A collection of data sets that are large and complex in nature.

- **Both structured and unstructured**
- **Data volumes grow very fast to the point of rendering traditional relational database systems and conventional statistical tools inadequate**

High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation

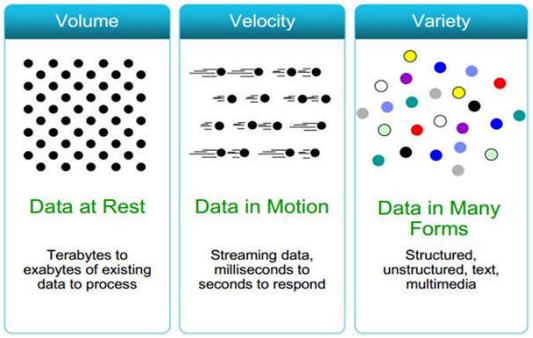
Big Data 3 Vs



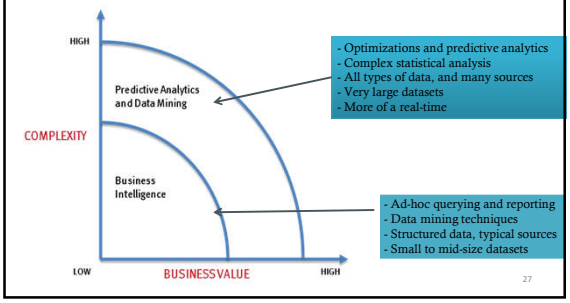
- Volume.
  - Especially With organizations collecting data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.
- Velocity. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.
- Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions

25

Big Data: 3 Vs

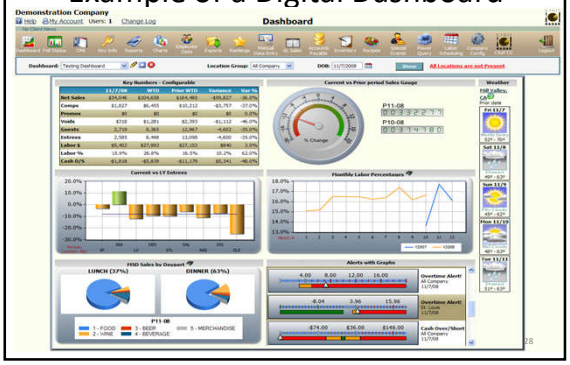


What's Driving Big Data?



27

Example of a Digital Dashboard



28