DATA WAREHOUSING

Introduction and Overview

What is a Data Warehouse?

A complete repository of corporate data extracted from transaction systems that is available for ad-hoc access by “knowledge workers.”

What is a Data Warehouse? (2)

- The data warehouse is an information environment that
  - Provides an integrated and total view of the enterprise (data)
  - Makes the enterprise's current and historical data easily available for decision making
  - Makes decision-support transactions possible without hindering operational systems
  - Renders the organization's information consistent
  - Presents flexible and interactive source of strategic information

What is a Data Warehouse? (3)

- A DW is a simple concept
  - Take all the information in the organization, clean and transform it, and then provide useful strategic information based on it
  - This concept was born out of need, and realization that large quantities of data exists in disintegrated chunks within an organization
- A DW is a computing environment, not a product
  - Not a single hardware or software product; rather it is an environment built with different hardware, software, and people connected by various processes
  - It is a user-centric environment, driven by the needs of the decision maker
  - It is a flexible environment for data analysis

What is a Data Warehouse? (4)

- A blend of technologies
  - Data acquisition
  - Data modeling
  - Data management
  - Data cleaning
  - Metadata management
  - Storage management
  - Applications
  - Management tools

- Data warehousing is new kind of computing environment geared towards strategic information

The Need for DW

- The need for *strategic* information
  - Competitive edge
  - Improve performance (revenue, profits, etc)
- Characteristics of strategic information
  - Integrated
  - Data integrity
  - Accessible
  - Credible
  - Timely

Data Glut…

- We are drowning in data, but we have little knowledge
  - The data is not accessible for strategic information and decision making
  - Many enterprises have separate databases for sales, human resources, payroll, products and services, etc
- Operational systems
  - They maintain record of events for day-to-day operations
  - They are not accessible easily for analysis and strategic information

Strategic Information Scarcity…

- Executives are interested in strategic information that can help them make decisions regarding their business's direction and growth
  - Strategic information is extracted or discovered from large quantities of data; it requires analysis of easily accessible and clean data
- Data warehousing is a solution for the 'data glut, knowledge scarcity' problem; it is essentially a kind of decision-support system

Failure of Earlier Decision-Support Systems

The need for strategic information has existed from the earliest days of competitive business

- Ad hoc reports
- Special extraction programs
- Small applications
- Decision-support systems
- Executive information systems
- Data warehousing

Data Warehouse and Operational Systems

- Operational systems – OLTP
  - Making the wheels of business turn

- Data warehouse
  - Watching the wheels of business turn

- Different scope, different purposes

How are they Different? (1)

- **Consolidates operational and historical data.**

- **Usually (but not always) periodic or batch updates rather than real time.**

- **Starts out with a 6x12 availability requirement...but 7x24 usually becomes the goal.**

How are they Different? (2)

- **Operational systems run the business -- DW gives insight into how to <u>improve</u> the business.**
- **Data warehousing goes beyond traditional MIS by allowing interactive data exploration by end-users.**
- **Database structures designed to support DSS: star schema, denormalized tables, sampling, etc.**
  - **Tradeoffs must be carefully evaluated.**

How are they Different? (3)

	Operational	Informational
Data content	Current values	Archieved, derived, summarized
Data structure	Optimized for transactions	Optimized for complex queries
Access frequency	High	Medium to low
Access type	Read, update, delete	Read
Usage	Predictable, repetitive	Ad-hoc, random, heuristic
Response time	Sub-seconds	Several seconds to minutes
User	Large number	Relatively small number

Typical Applications

- Impact on organization’s core business is to streamline and maximize profitability.
  - Fraud detection.
  - Profitability analysis.
  - Direct mail/database marketing.
  - Customer retention modeling.
  - Credit risk prediction.
  - Inventory management.
  - Yield management.
- ROI on any one of these applications can justify HW/SW costs in most organizations.

Typical Early Adopters

- Financial service/insurance.
- Retailing and distribution.
- Telecommunications.
- Transportation.
- Government.

Common thread: lots of customers and transactions.

What Are End User Expectations?

- Point and click access to data.
- Insulation from DBMS structures.
  - Want semantic data model - not 3rd normal form.
- Integration with existing tools: SAS, Excel, etc.
- Interactive response times for on-line analysis...but batch is important, too.

Quantification of Response Times

- On-line analytical processing (OLAP) queries must be executed in a small number of seconds.
  - Often requires denormalization and/or sampling.
- Complex query scripts and large list selections can generally be executed in a small number of minutes.
- Sophisticated modeling algorithms (e.g., data mining) can generally be executed in a small number of hours (even for millions of customers).

Desired Features of DW

- Database designed for analytical tasks
- Data from multiple sources
- Easy to use and conducive to long interactive sessions by users
- Read-intensive data usage
- Direct interaction of the user with the system
- Content updated periodically and stable
- Ability for users to run queries and get results online
- Ability for users to initiate reports

Business Intelligence

- Data warehousing supports business intelligence
- What is BI?
- Business Intelligence is a process that adds value to your business processes through monitoring performance indicators about business environment and their impact on business strategy to help define, refine and improve business model for Profitable Operations
- In lay terms, BI entails
  - Ability run simple queries
  - Ability to perform ‘what if’ analyses in different ways
  - Ability to interactively analyze results
  - Ability to discover trends and apply them to future results

Data Warehouse High-level Implementation Steps

1. **Identify key business requirements.**
2. **Identify key data sources and volumes.**
3. **Identify phased deliverables with quantifiable business benefits.**
4. **Software/hardware selection.**
5. **Data warehouse construction.**

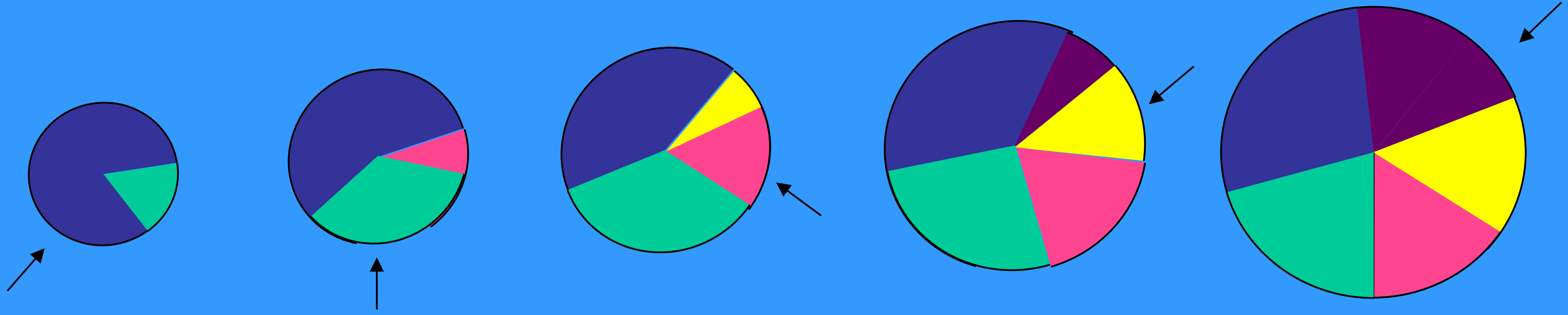
   -Data extraction and cleansing.

   -Logical and physical design.

   -Software integration.
6. **Productionalize.**
7. **Go to step one for next deliverable.**

Information Evolution in a Data Warehouse Environment

STAGE 1:	STAGE 2:	STAGE 3:	STAGE 4:	STAGE 5:
REPORT	ANALYZE	PREDICT	OPERATIONALIZE	ACTIVATE
WHAT happened?	WHY did it happen?	WHAT will happen?	What IS happening?	What do you WANT to happen?



Primarily Batch	Increase in Ad Hoc Queries	Analytical Modeling Grows	Continuous Update & Time Sensitive Queries Become Important	Event Based Triggering Takes Hold

🟦 Batch 🟩 Ad Hoc 🟥 Analytics 🟨 Continuous Update/Short Queries 🟪 Event-Based Triggering

Data Warehouse and Data Marts

	Data Warehouse	Data Mart
Scope	♦ Application –Neutral ♦ Centralized, shared ♦ Cross LOB/enterprise ♦ Multiple subject areas	♦ Specific application requirements ♦ Multiple databases with redundant data ♦ LOB, departmental or user area ♦ Partial-subject area
Data Perspective	♦ Historical, detailed data ♦ Some summary ♦ Lightly denormalized	♦ Detailed (some history) ♦ Summarized ♦ Highly denormalized
Characteristics	♦ Flexible, extensible ♦ Strategic, durable ♦ Data orientation	♦ Restrictive, non-extensible ♦ Tactical, short life ♦ Project, business-process orientation

Source: Gartner Group, Kevin Strange

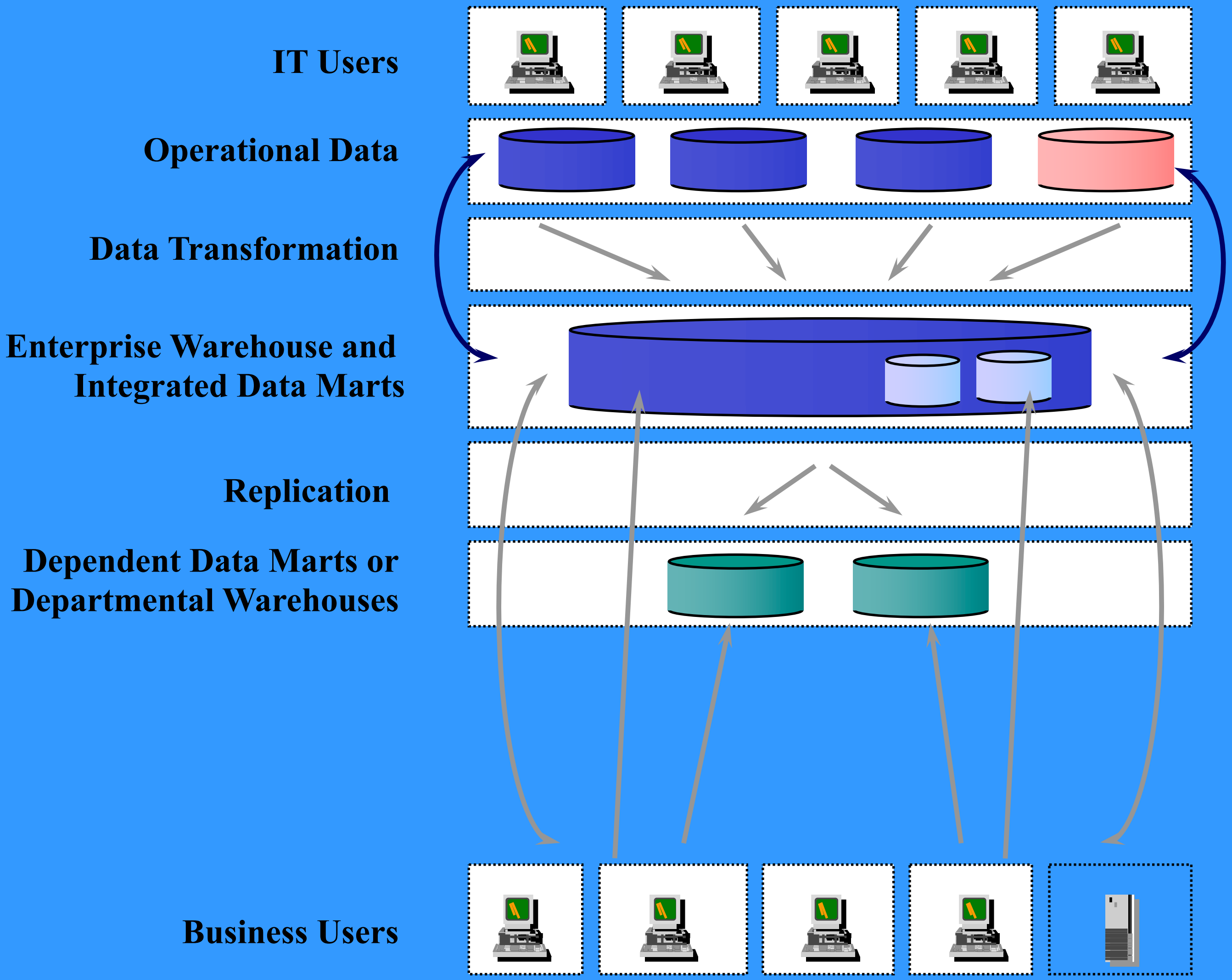
Which One First?

- Top-down approach or bottom-up approach?
- Enterprise-wide or departmental?
- What first – one data warehouse or multiple data marts?
- Build pilot or go with a full-fledged implementation
- Dependent or independent data marts

A Practical Approach

- Chief proponent of this approach is Kimball
- The practical approach
  - Plan and define requirements at the overall corporate level
  - Create the architecture for a complete warehouse
  - Conform and standardize the data content
  - Implement the data warehouse as a series of marts, one at a time
- In this approach, a data mart is a logical subset of the entire data warehouse (dependent data marts)

A Typical Data Warehouse Environment



IT Users
Operational Data
Data Transformation
Enterprise Warehouse and Integrated Data Marts
Replication
Dependent Data Marts or Departmental Warehouses
Business Users

DW Building Blocks or Components

- Key architectural components of a DW
  - Data source
  - Data staging
  - Data storage
  - Information and delivery
  - Management and control

Source Data

- Production data - data from operational systems
  - Major issue: disparity (different formats, management systems, interfaces, hardware, software, etc)
- Internal data – data from ‘private’ (not widely available) sources like spreadsheets, access databases, files, records, etc
  - Major issues: selecting internal data, deciding on whether to incorporate internal data,
- Archived data – backup data from operational systems
  - Major issues – deciding how much archived data to include
- External data – data from data analysis agencies
  - Major issue: selecting relevant data, integrating the data

Data Staging

- Area where data is prepared for the data warehouse
- Major functions
  - Extraction
  - Transformation
  - Loading

Data Storage

- The data storage is a separate repository
- Being a separate repository, it has its own
  - DBMS (using a RDBMS or a MDDBMS)
  - Hardware and software
  - Administration tools

Information Delivery

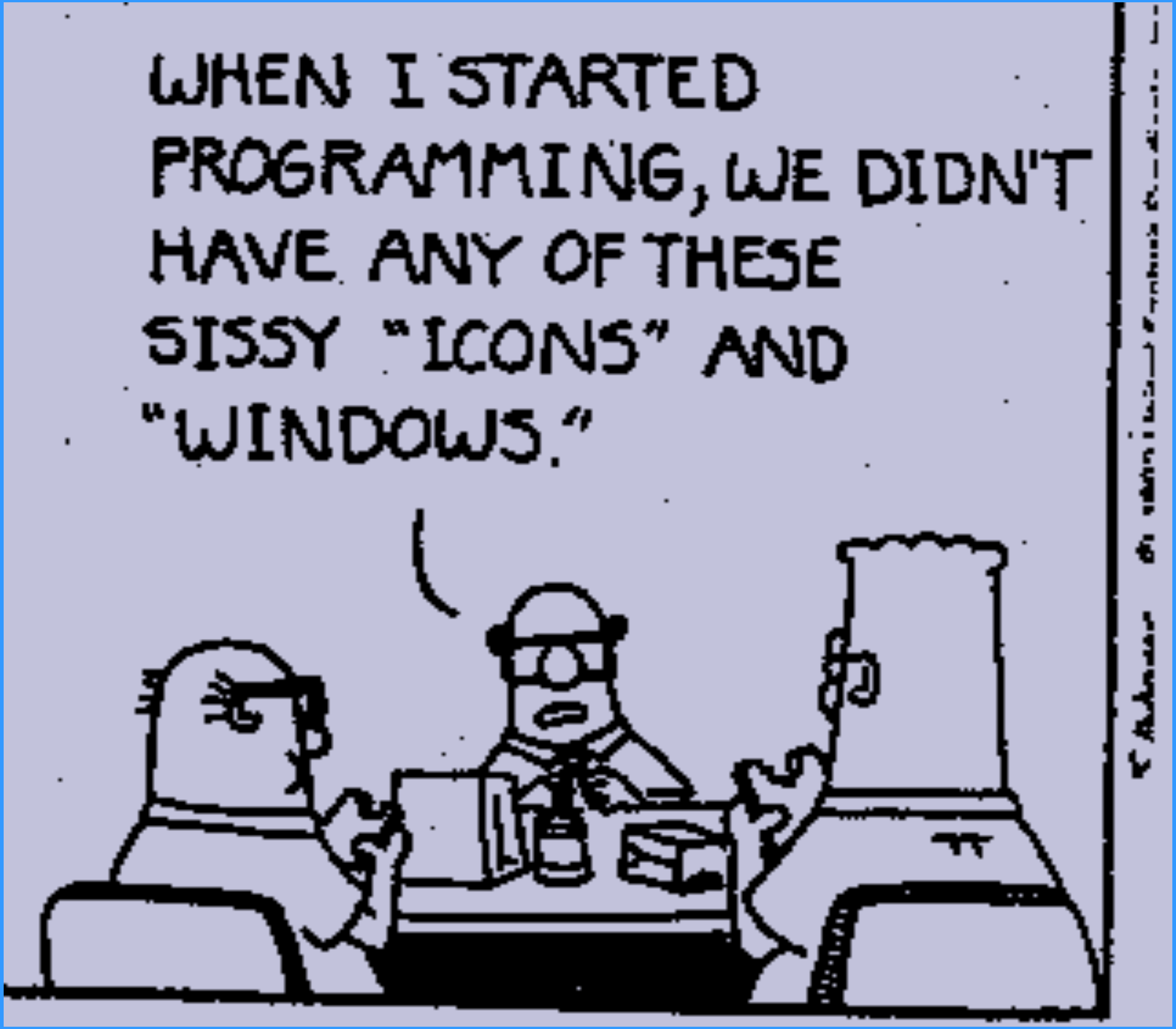
Metadata

- Data about data in the DW
- Similar to the data dictionary and data catalog in database management system

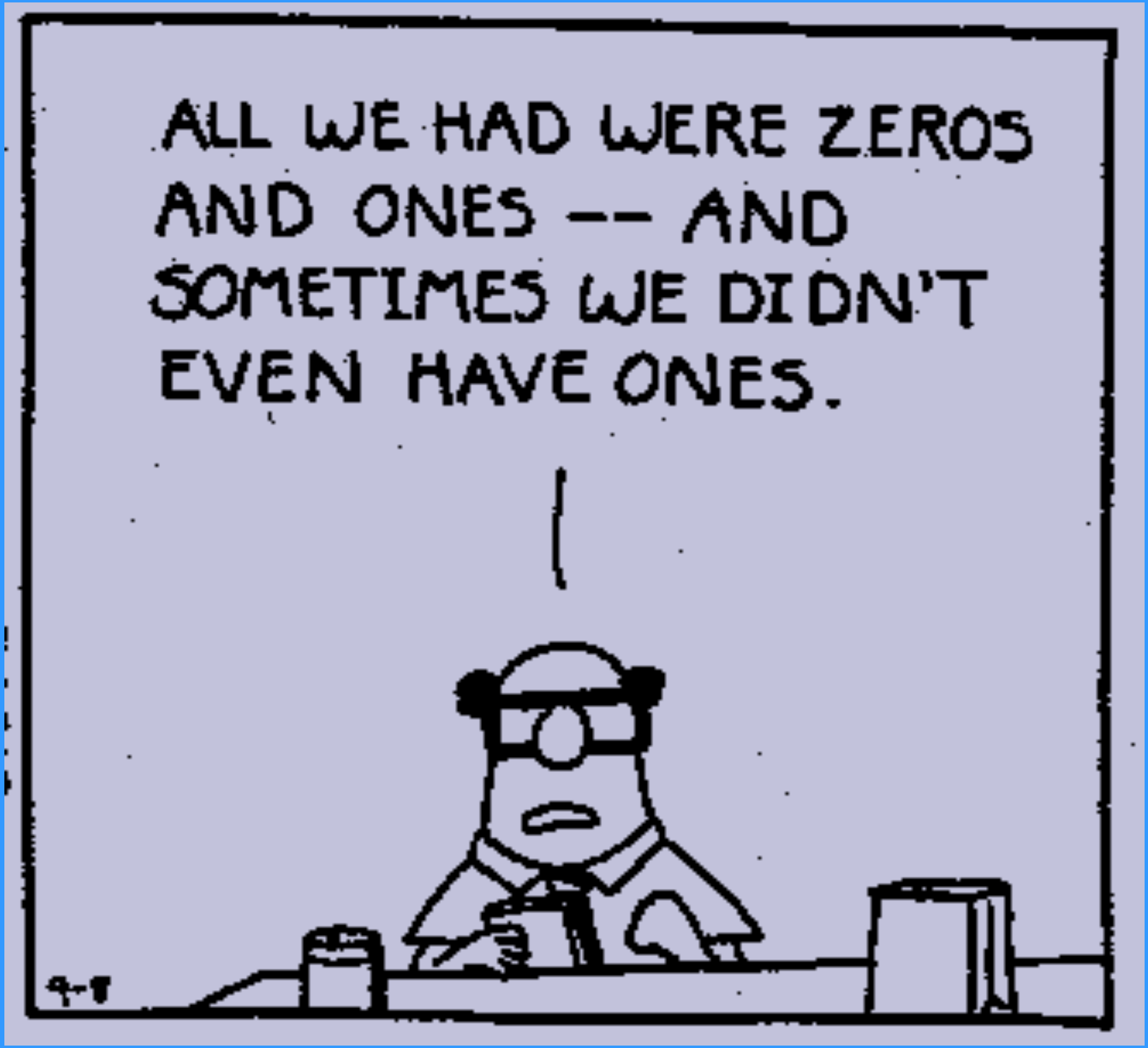
Management and Control

- Manages and controls all other components of the data warehouse
  - ETL
  - Information delivery
  - User accounts and access
  - Batch updating
- Management and control component interacts with the metadata

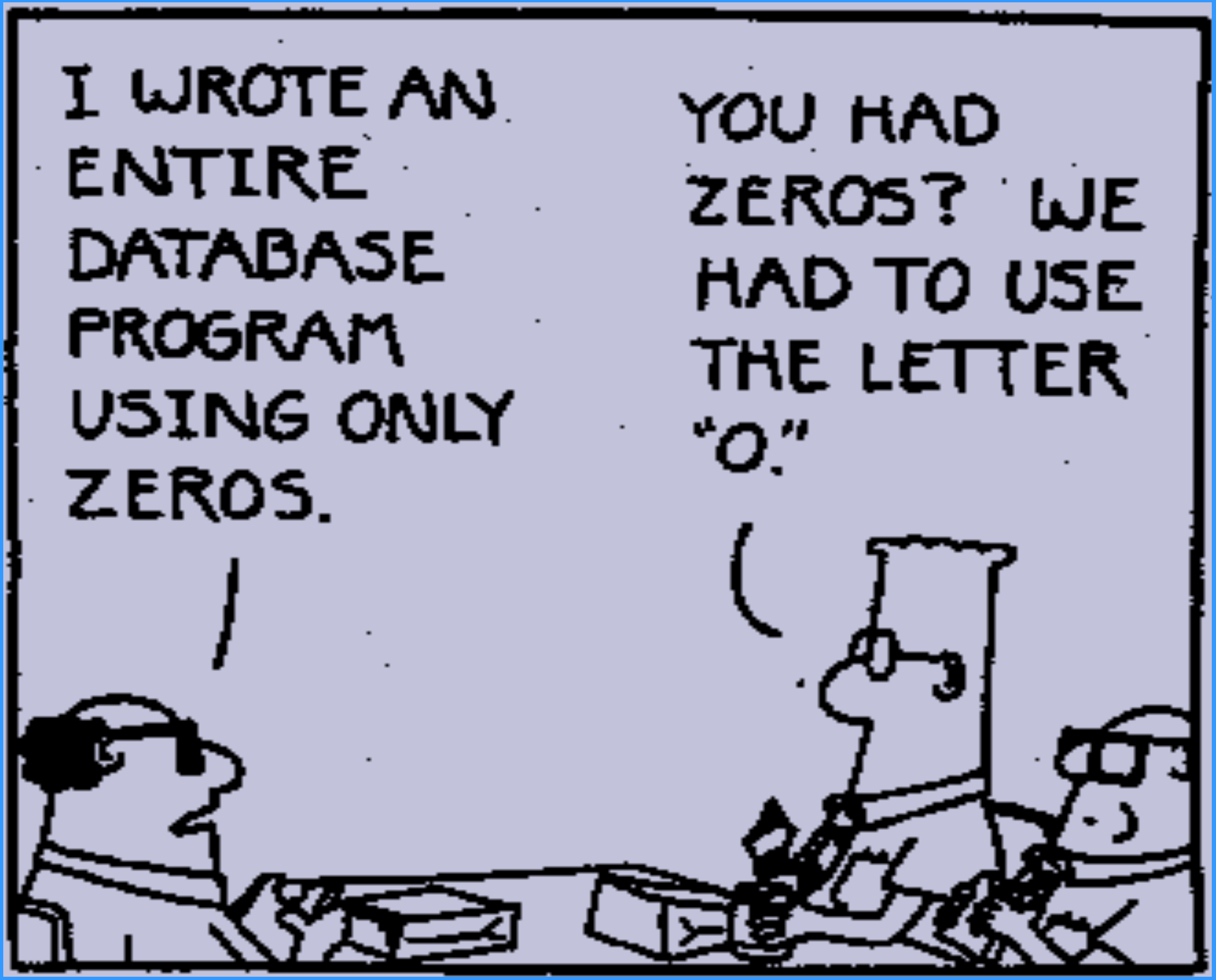
Why is this Hard?


WHEN I STARTED PROGRAMMING, WE DIDN'T HAVE ANY OF THESE SISSY "ICONS" AND "WINDOWS."

Why is this Hard?


ALL WE HAD WERE ZEROS AND ONES -- AND SOMETIMES WE DIDN'T EVEN HAVE ONES.
9-8

Why is this Hard?


I WROTE AN ENTIRE DATABASE PROGRAM USING ONLY ZEROS.
YOU HAD ZEROS? WE HAD TO USE THE LETTER "O."

Why is this Hard?

- **There are no stable requirements in a data warehouse environment.**
- **Familiar database techniques break down in DSS at large scale.**
- **The scale factor in VLDB implementations is difficult to comprehend.**
- **Performance impacts are often non-linear.**
- **Complex architectures for deployment.**
- **Rapidly changing product characteristics.**
- **And so on...**

Approach

- Develop an understanding of underlying RDBMS implementation techniques.
- Apply these techniques to VLDB DSS environments and understand where they break down.
- Provide a “toolkit” of design techniques for maximizing performance in a variety of data warehouse implementation scenarios.
- Place particular emphasis on harnessing parallel technology as a means of overcoming scale.

Considerations

- Logical and physical data modeling.
- OLAP implementation techniques.
- Extract, transform, and loading of data.
- Indexing structures.
- Join algorithms.
- Parallel processing deployment.
- Data mining.
- Data quality management.
- Capacity planning and service level agreements.
- Platform configuration.
- Data warehouse architecture.

Reality Check

Hardware is the easyware…

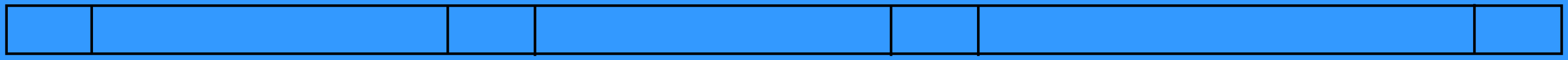
…software is the hardware.

Reality Check

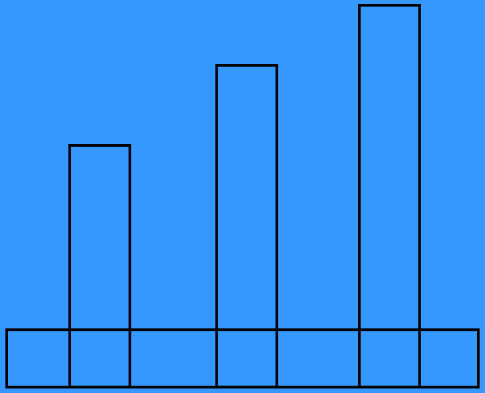
If the software doesn’t scale, it doesn’t matter how much your hardware can scale up!

Amdahl’s Law

Sequential Execution:



Ideal Parallel Execution:



Let “s” be the fraction of the program that must be performed sequentially.

Amdahl’s Law

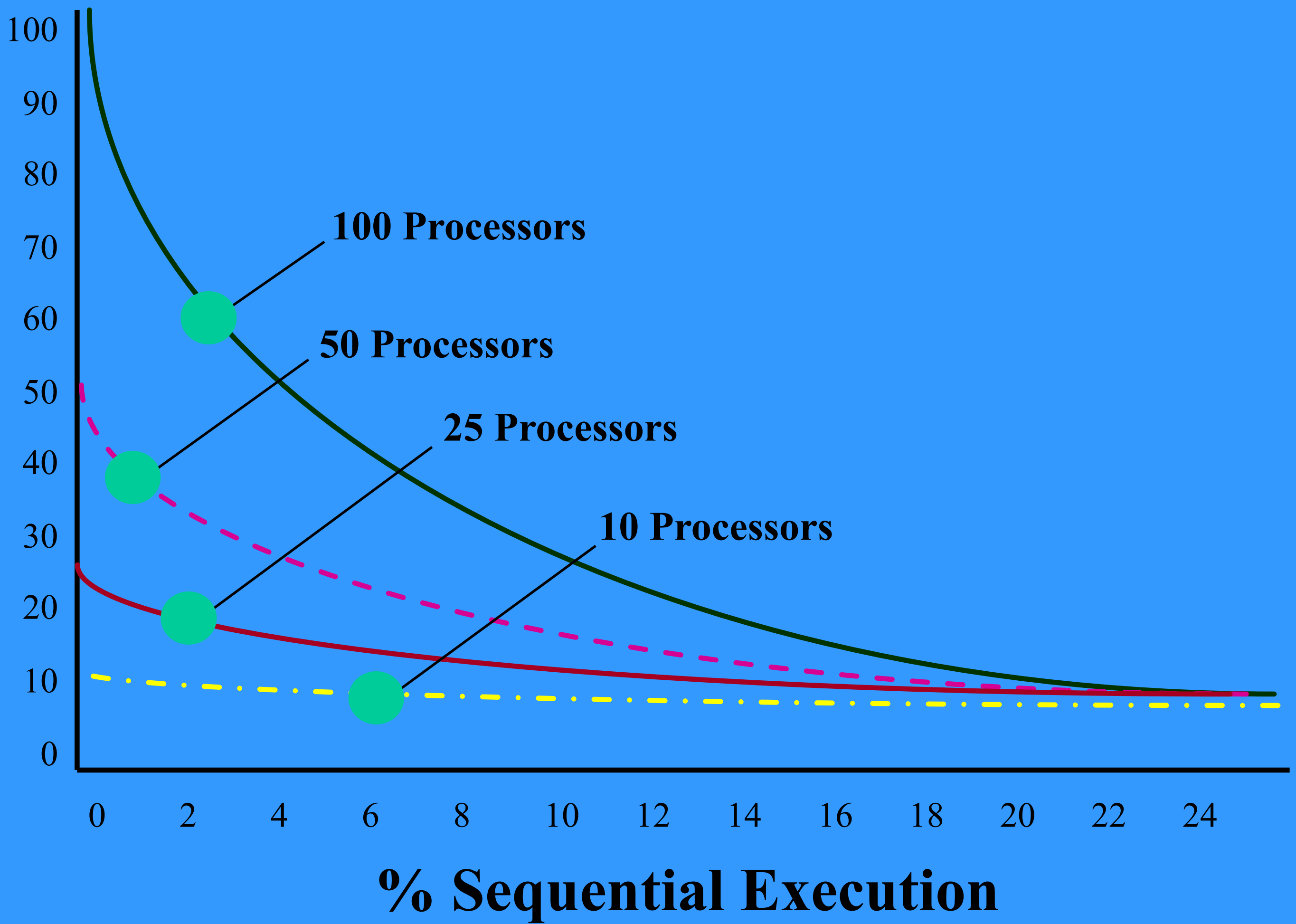
Quantitative representation:

- SpeedUp = 1 / (Percent Sequential + (Percent Parallel / # of Processors) + Overhead)
- SpeedUp = 1 / (0.1 + (0.9 / 4) + 0) = 3.08

The keys to parallel performance:

- Amount of parallelism
- Overhead management

Amdahl’s Law

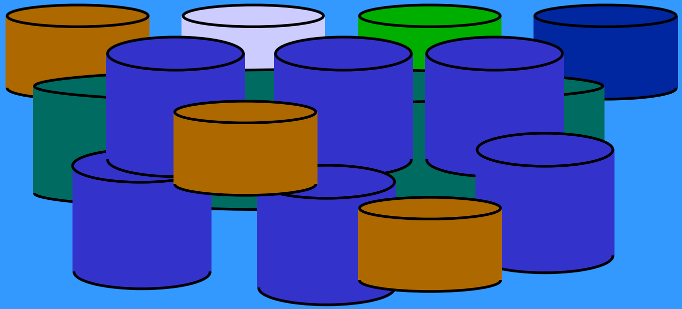

100
90
80
70
60
50
40
30
20
10
0
100 Processors
50 Processors
25 Processors
10 Processors
0
2
4
6
8
10
12
14
16
18
20
22
24
% Sequential Execution

Parallel Processing: The Impact

- How long to read a Terabyte of data?
  - Question posed in Information Week article on VLDB implementations.
  - Answer provided: 1.2 days, serially.
- Parallel Processing can speed-up
  - 0.6 Days with 2 parallel tasks
  - Less than 18 minutes with 100 parallel tasks, provided that:
    - Software has even distribution of tasks.
    - Hardware can sustain I/O levels.

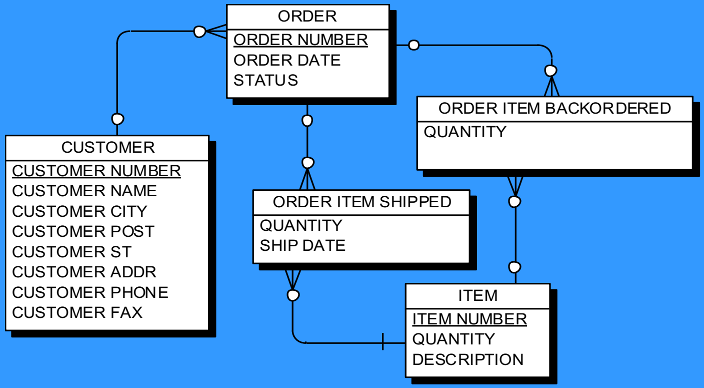
Scalability - It Is Not Just About Size

Amount of Detailed Data



Concurrent Users

Complexity of Data Model



ORDER
ORDER NUMBER
ORDER DATE
STATUS

ORDER ITEM BACKORDERED
QUANTITY

CUSTOMER
CUSTOMER NUMBER
CUSTOMER NAME
CUSTOMER CITY
CUSTOMER POST
CUSTOMER ST
CUSTOMER ADDR
CUSTOMER PHONE
CUSTOMER FAX

ORDER ITEM SHIPPED
QUANTITY
SHIP DATE

ITEM
ITEM NUMBER
QUANTITY
DESCRIPTION

Query Complexity

- Simple Direct at the start
- Moderate Multi-table Join
- Regression analysis
- Query tool support