Data Warehousing

Concepts

PG em DS&BA – 2ª Edição Blended

Ana Lucas

Parts of this presentation were taken from the backing material of the book

Modern Database Management, 13 Edition, 2019
Jeffrey A. Hoffer, V. Ramesh, Heikki Topi

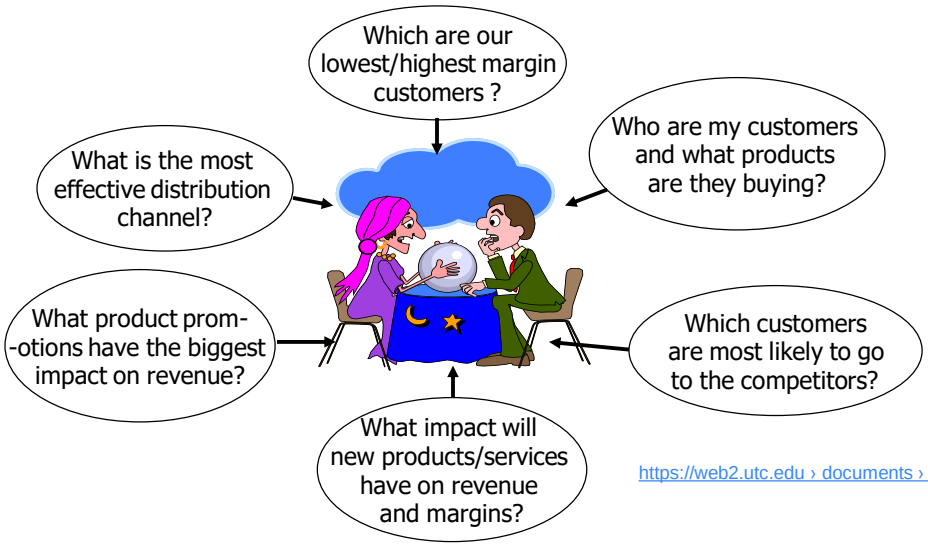
Data, Data everywhere yet …

- **I can't find the data I need**
  - data is scattered over the network
  - many versions, subtle differences
- **I can't get the data I need**
  - need an expert to get the data
- **I can't understand the data I found**
  - available data poorly documented
- **I can't use the data I found**
  - results are unexpected
  - data needs to be transformed from one form to another

https://web2.utc.edu › documents › dataWarehouse

3

Why Data Warehousing?



https://web2.utc.edu › documents › dataWarehouse

4

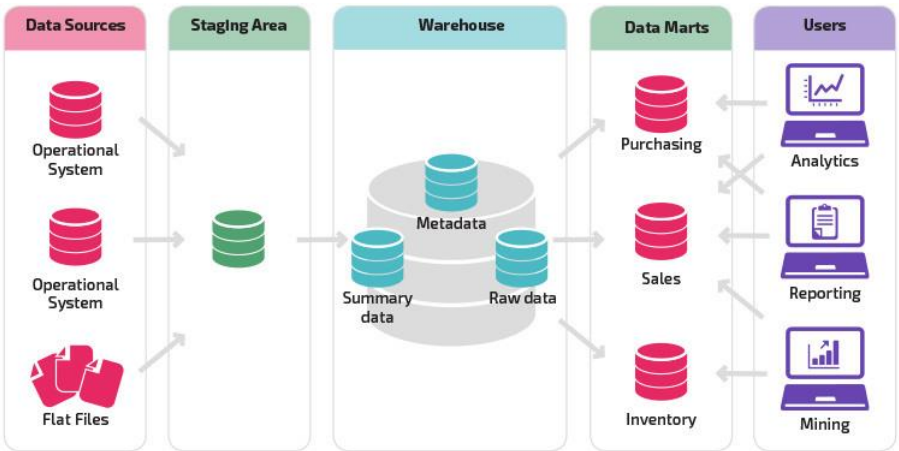
History

1988 – The IBM researchers Barry Devlin and Paul Murphy published the article “An architecture for a business and information system” where they introduced the term **"business data warehouse“**

1992 – Bill Inmon publishes the book ”Building the Data Warehouse”

.

Data Warehousing Overview



https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Concepts (1/3)

Data Warehouse

A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes

- ***Subject-oriented:*** **e.g. customers, patients, students, products**
- ***Integrated:*** **consistent naming conventions, formats, encoding structures; from multiple data sources**
- ***Time-variant:*** **can study trends and changes**
- ***Non-updatable:*** **read-only, periodically refreshed**

Inmon (1992)

Concepts (2/3)

Data Warehouse (another definition)

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a format they can understand and use in a business context

Devlin and Murphy (1988)

Concepts (3/3)

Data Mart

- A data warehouse that is limited in scope (usually related to a **business process**)

Data Warehousing – it is a Process

- Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previously possible
- A decision support database maintained separately from the organization's operational database

https://web2.utc.edu › documents › dataWarehouse

Need for Data Warehousing

- **Consolidation of information resources**
- **Improved query performance**
- **Separate research and decision support functions from the operational systems**
- **Foundation for data mining, data visualization, advanced reporting and OLAP (On Line Analytical Processing) tools**

https://web2.utc.edu › documents › dataWarehouse

Issues with Company-Wide Operational View

- **Inconsistent key structures**
- **Synonyms**
- **Free-form vs. structured fields**
- **Inconsistent data values**
- **Missing data**

Examples of heterogeneous data

STUDENT DATA

StudentNo	LastName	MI	FirstName	Telephone	Status	• • •
123-45-6789	Enright	T	Mark	483-1967	Soph	
389-21-4062	Smith	R	Elaine	283-4195	Jr	

STUDENT EMPLOYEE

StudentID	Address	Dept	Hours	• • •
123-45-6789	1218 Elk Drive, Phoenix, AZ 91304	Soc	8	
389-21-4062	134 Mesa Road, Tempe, AZ 90142	Math	10	

STUDENT HEALTH

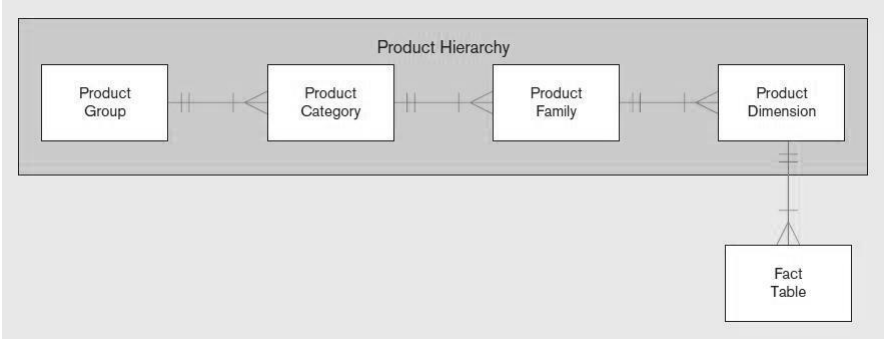
StudentName	Telephone	Insurance	ID	• • •
Mark T. Enright	483-1967	Blue Cross	123-45-6789	
Elaine R. Smith	555-7828	?	389-21-4062	

Comparison of Operational and Data Warehouse Systems

Data warehouse	Operational system
Subject oriented	Transaction oriented
Large (hundreds of GB up to several TB to Petabytes)	Medium (several GB to TB)
Historic data	Current data
De-normalized table structure (few tables, many columns per table)	Normalized table structure (many tables, few columns per table)
Batch updates	Continuous updates
Usually very complex queries	Simple to complex queries

Adapted from https://web2.utc.edu › documents › dataWarehouse

Normalization

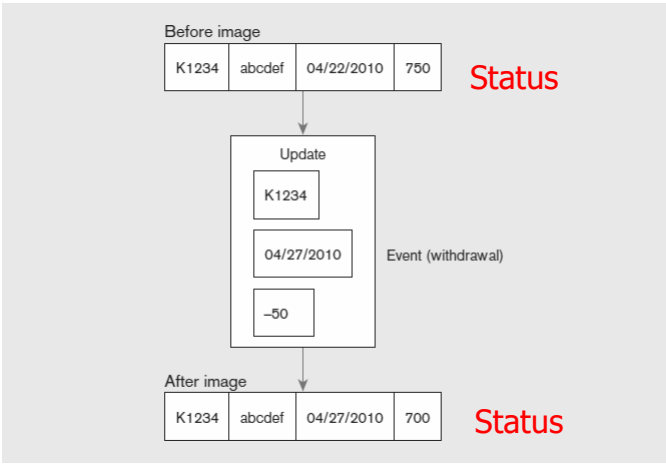


Denormalization

Product Key	Product Description	Brand Name	Category Name
1	PowerAll 20 oz	PowerClean	All Purpose Cleaner
2	PowerAll 32 oz	PowerClean	All Purpose Cleaner
3	PowerAll 48 oz	PowerClean	All Purpose Cleaner
4	PowerAll 64 oz	PowerClean	All Purpose Cleaner
5	ZipAll 20 oz	Zippy	All Purpose Cleaner
6	ZipAll 32 oz	Zippy	All Purpose Cleaner
7	ZipAll 48 oz	Zippy	All Purpose Cleaner
8	Shiny 20 oz	Clean Fast	Glass Cleaner
9	Shiny 32 oz	Clean Fast	Glass Cleaner
10	ZipGlass 20 oz	Zippy	Glass Cleaner
11	ZipGlass 32 oz	Zippy	Glass Cleaner

Kimball, Ross (2013)

Data Characteristics
Status vs. Event Data



Example of DBMS log entry

Event = a database action (create/ update/ delete) that results from a transaction

Transient Data – Operational Data

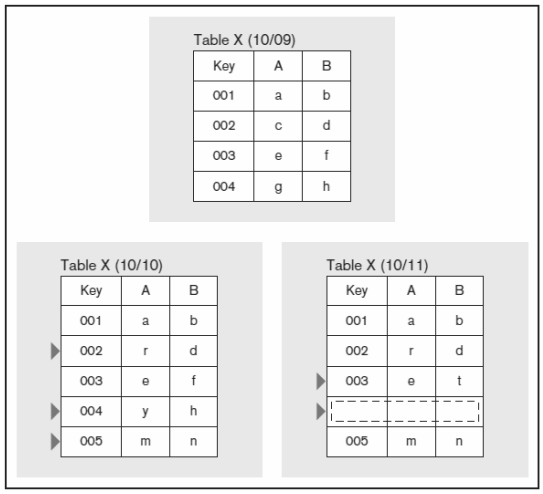


Table X (10/09)

Key	A	B
001	a	b
002	c	d
003	e	f
004	g	h

Table X (10/10)

Key	A	B
001	a	b
002	r	d
003	e	f
004	y	h
005	m	n

Table X (10/11)

Key	A	B
001	a	b
002	r	d
003	e	t
005	m	n

With **transient data**, **changes to existing records are written over previous records**, thus destroying the previous data content

Periodic Data – Warehouse Data

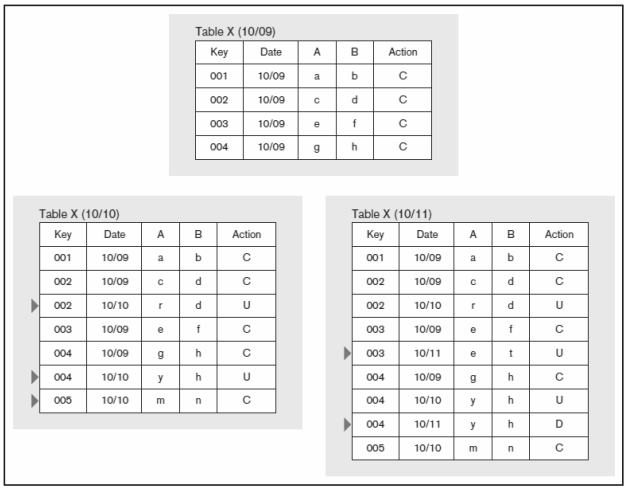


Table X (10/09)

Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
003	10/09	e	f	C
004	10/09	g	h	C

Table X (10/10)

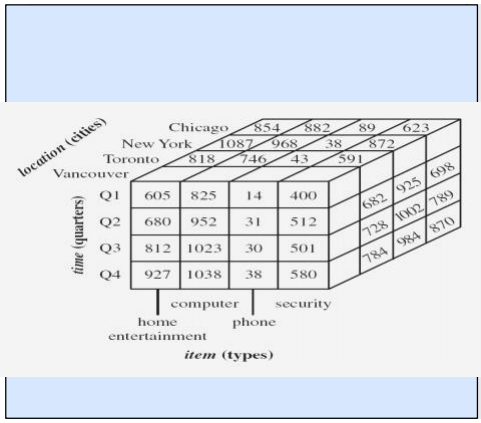
Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
002	10/10	r	d	U
003	10/09	e	f	C
004	10/09	g	h	C
004	10/10	y	h	U
005	10/10	m	n	C

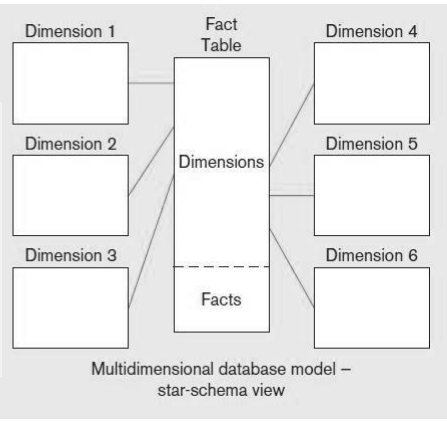
Table X (10/11)

Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
002	10/10	r	d	U
003	10/09	e	f	C
003	10/11	e	t	U
004	10/09	g	h	C
004	10/10	y	h	U
004	10/11	y	h	D
005	10/10	m	n	C

Periodic data are never physically altered or deleted once they have been added to the store

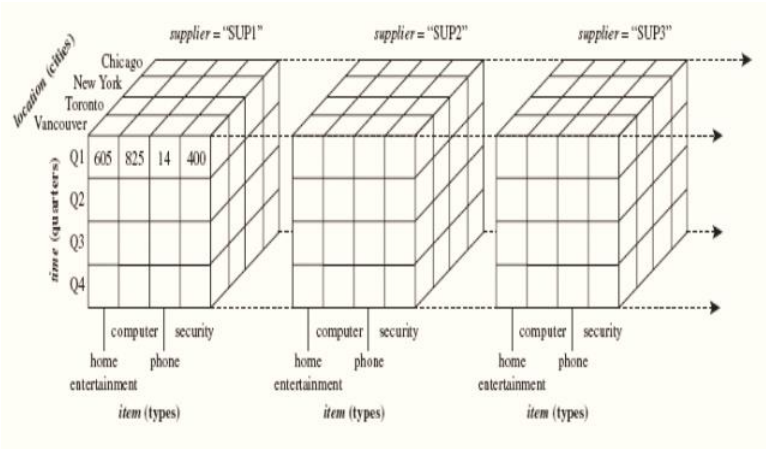
Dimensional Model





Multidimensional database model – star-schema view

Dimensional Model



Data Lake

Data Lake

Pentaho CTO James Dixon has generally been credited with coining the term "data lake" on October, 2010.

He describes a **data mart** (a subset of a data warehouse) as akin to a bottle of water**…"cleansed, packaged and structured for easy consumption"** while a **data lake is more like a body of water in its natural state**. Data flows from the streams (the source systems) to the lake. Users have access to the lake to examine, take samples or dive in.

https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses

Data Lake

A storage repository, that holds a vast amount of raw data in its native format until it is needed

- A place to store unlimited amounts of data in any format inexpensively, especially for archive purposes
- Allows collection of data that you may or may not use later: "just in case"
- A way to describe any large data pool in which the schema and data requirements are not defined until the data is queried: "just in time" or "schema on read"
- **Complements EDW and can be seen as a data source for the EDW – capturing all data but only passing relevant data to the EDW**
- **Allows for data exploration to be performed without waiting for the EDW team to model and load the data (quick user access)**

https://pt.slideshare.net/jamserra/big-data-architectures-and-the-data-lake

23

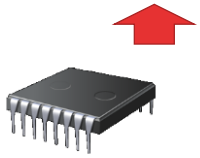
Data Warehouse vs Data Lake

	Data Lake	Data Warehouse
Data Structure	Raw	Processed
Purpose of Data	Not Yet Determined	Currently In Use
Users	Data Scientists	Business Professionals
Accessibility	Highly accessible and quick to update	More complicated and costly to make changes

https://www.talend.com/resources/data-lake-vs-data-warehouse/

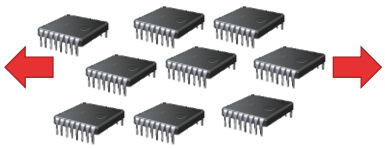
24

Data Warehouse vs. Data Lake Architectures
Scale Up vs. Scale Out



Scale Up (DW)

Scaling vertically means adding resources to a single node, typically involving the addition of CPUs, memory or storage to a single computer



Scale Out (DL)

Uses Clusters of Commodity PCs

Make Many CPUs work together

Learn how to divide your problems into independent threads

Characteristics of Big Data

Schema on Read, rather than Schema on Write

- Schema on Write– preexisting data model, how traditional databases are designed (relational databases)
- Schema on Read – data model determined later, depends on how you want to use it (XML, JSON)
- Capture and store the data, and worry about how you want to use it later

Examples of JSON and XML

JSON Example

```
{"products": [
        {"number": 1, "name": "Zoom X", "Price": 10.00},
        {"number": 2, "name": "Wheel Z", "Price": 7.50},
        {"number": 3, "name": "Spring 10", "Price": 12.75}
]}
```

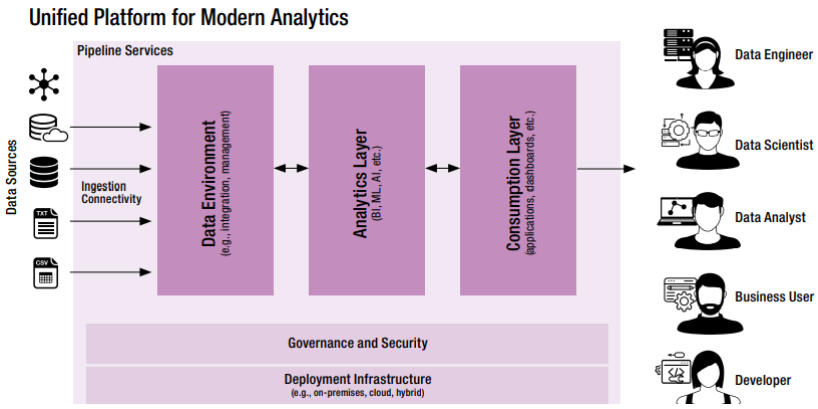
JavaScript Object Notation

XML Example

```
<products>
        <product>
                <number>1</number> <name>Zoom X</name> <price>10.00</price>
        </product>
        <product>
                <number>2</number> <name>Wheel Z</name> <price>7.50</price>
        </product>
        <product>
                <number>3</number> <name>Spring 10</name> <price>12.75</price>
        </product>
</products>
```

eXtensible Markup Language

Unified Platform for Modern Analytics



Halper, F. (2021). *Unified Platforms for Modern Analytics* (Best Practices Report Q3 2021). TWDI - Transforming Data with Intelligence.

Unified Platform for Modern Analytics

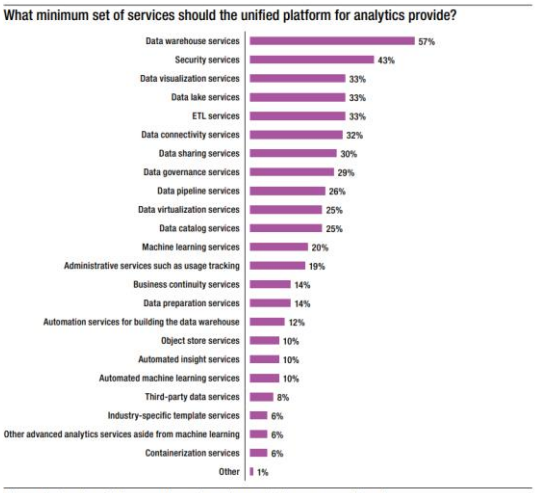


Figure 4. Based on 398 respondents. A maximum of six responses allowed.

Halper, F. (2021). *Unified Platforms for Modern Analytics* (Best Practices Report Q3 2021). TWDI - Transforming Data with Intelligence.

Data Monetization

Data monetization is the act of measuring the economic benefit of corporate data. The benefits can be in the form of actual euros, but they can also pave the way to new products, services and even process improvements

Laskowski, N. (n.d.). Data Monetization. Retrieved November 26, 2021, from https://searchcio.techtarget.com/definition/data-monetization

Data Monetization

Data monetization strategies include some combination of three moneymaking approaches:

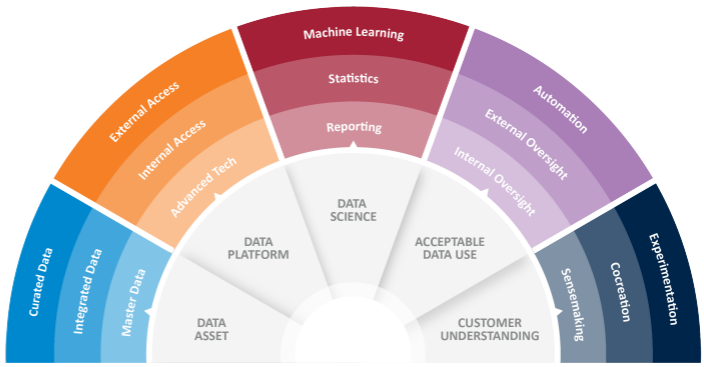
1. **Improving core business processes using data—making money from doing things better, cheaper, and faster**
2. **"Wrapping" analytics around offerings—making money by distinguishing offerings with features and experiences**
3. **Selling information solutions—making money by deploying new information offerings**

Wixom. B.H. & Farrell, K.(2019). *Building Data Monetization Capabilities that Pay Off* (Research Briefing Volume 19, Number 11, November 2019). MIT Sloan Center for Information Systems Research.

Data Monetization

Figure 1: Data Monetization Capabilities Are Evolutionary

In this diagram, data monetization capabilities are represented by the grey segments and practices by the colored bands. Companies evolve data monetization capabilities, with more advanced practices building on foundational predecessors.



Wixom. B.H. & Farrell, K.(2019). *Building Data Monetization Capabilities that Pay Off* (Research Briefing Volume 19, Number 11, November 2019). MIT Sloan Center for Information Systems Research.

Teradata University

https://academics.teradata.com/

Teradata Community

https://support.teradata.com/community

ana.lucas@iseg.ulisboa.pt

www.isegexecutive.education