LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Disciplina de Gestão de Dados e de Bases de Dados

Ano Letivo 2020/2021

Data Warehousing

Concepts

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

History

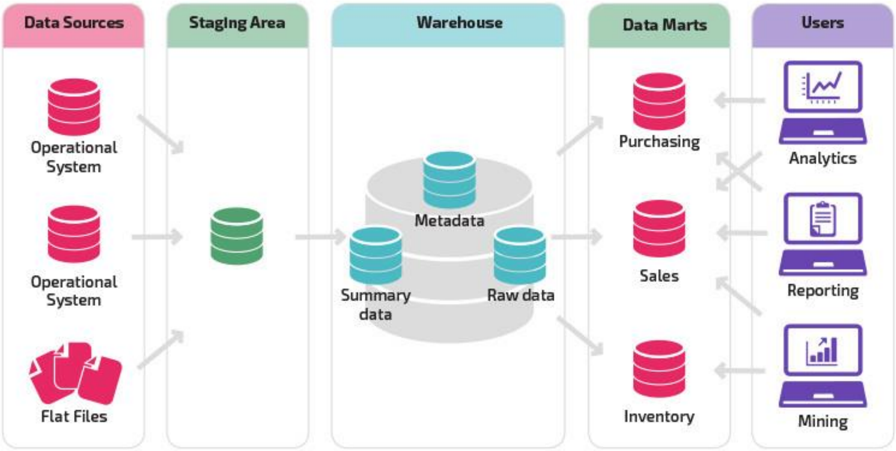
1988 – The IBM researchers Barry Devlin and Paul Murphy publish the article "An architecture for a business and information system" where they introduced the term **"business data warehouse"**

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

.

https://en.wikipedia.org/wiki/Data_warehouse

Data Warehouse Overview



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

Concepts (1/2)

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

Data Mart

- A data warehouse that is limited in scope

Concepts (2/2)

Data Warehousing

Is the process whereby organizations **create and maintain data warehouses** and **extract meaning and inform decision making** from their informational assets through these data warehouses

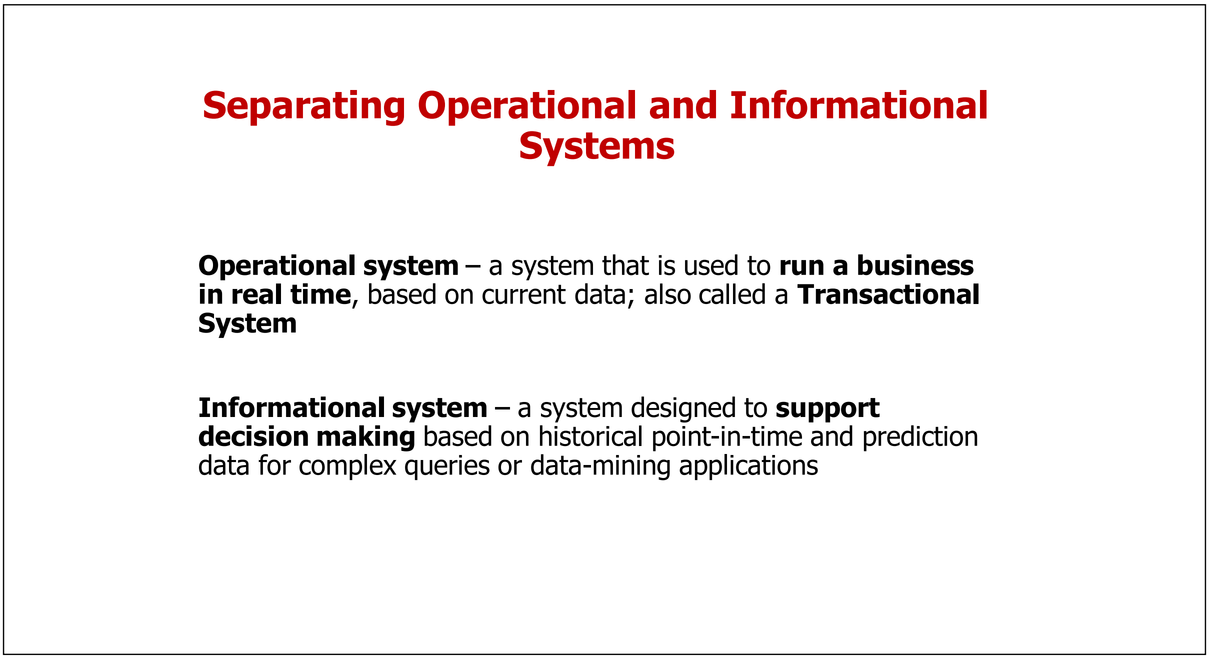
https://quizlet.com/32435342/itm-4271-11-flash-cards/

https://www.tutorialspoint.com/dwh/dwh_tutorial.pdf

Need for Data Warehousing

Integrated, company-wide view of high-quality information (from disparate databases)

Separation of *operational* and *informational* systems and data (for improved performance)



Separating Operational and Informational Systems

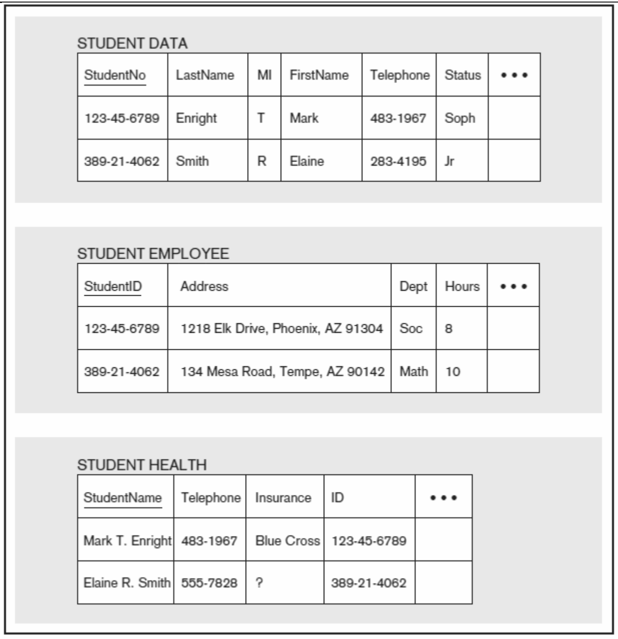
Operational system – a system that is used to **run a business in real time**, based on current data; also called a **Transactional System**

Informational system – a system designed to **support decision making** based on historical point-in-time and prediction data for complex queries or data-mining applications

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 Informational Systems

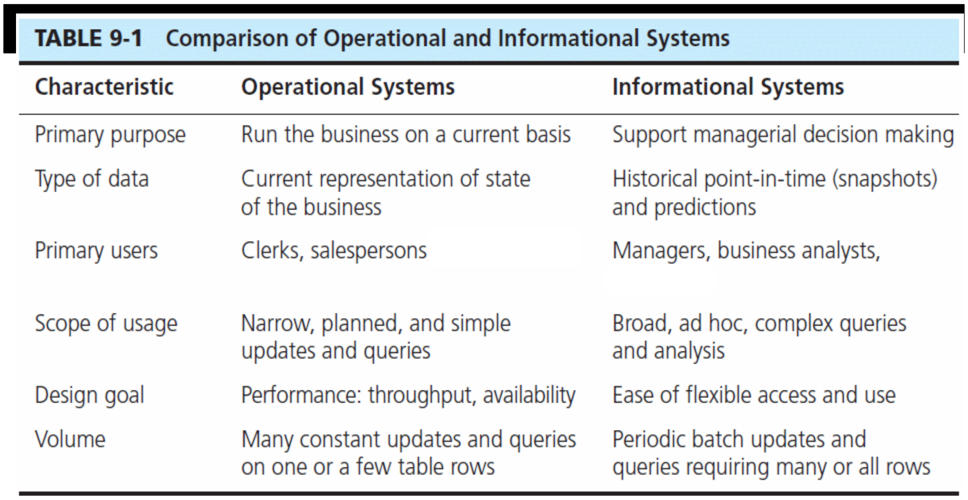
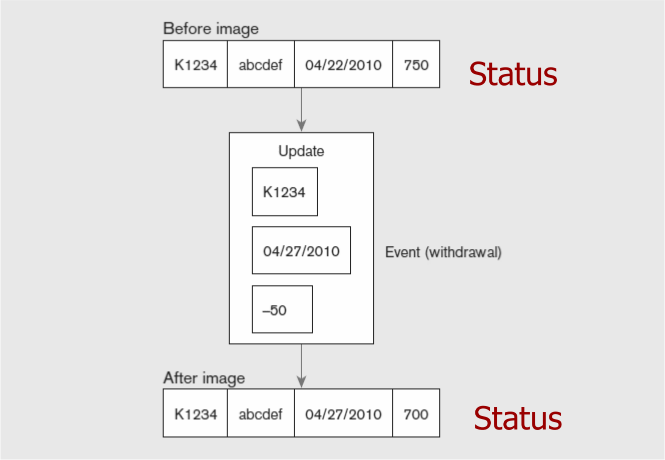


TABLE 9-1 Comparison of Operational and Informational Systems

Characteristic	Operational Systems	Informational Systems
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons	Managers, business analysts,
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance: throughput, availability	Ease of flexible access and use
Volume	Many constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

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

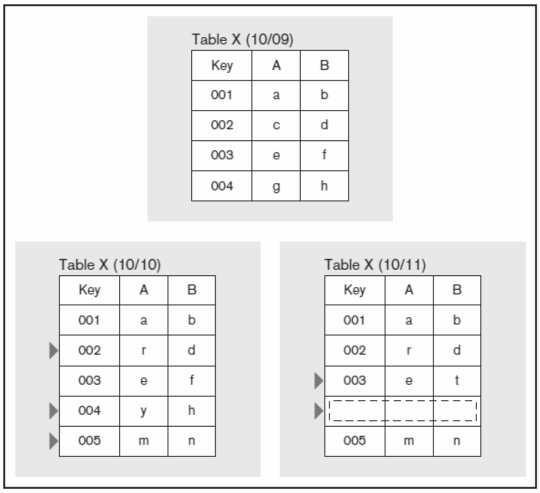


Table X (10/09)

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

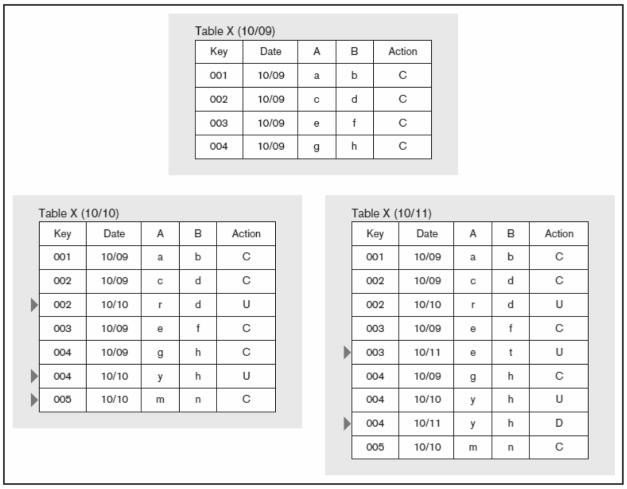


Table X (10/09)

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)

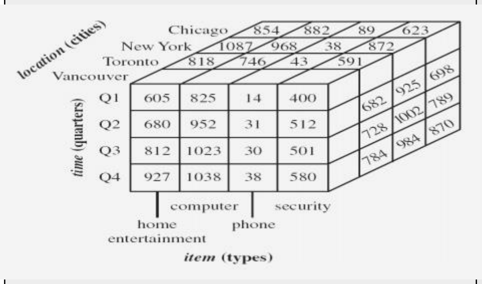
Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
▶ 002	10/10	r	d	U
003	10/09	e	f	C
004	10/09	g	h	C
▶ 004	10/10	y	h	U
▶ 005	10/10	m	n	C

Table X (10/11)

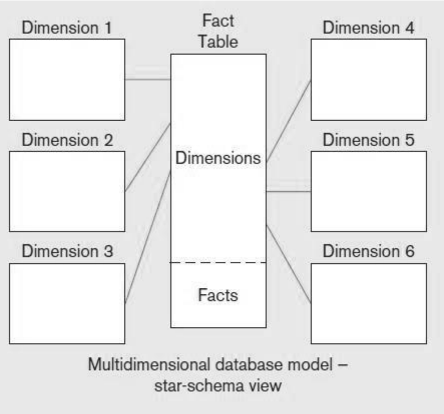
Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
002	10/10	r	d	U
003	10/09	e	f	C
▶ 003	10/11	e	t	U
004	10/09	g	h	C
004	10/10	y	h	U
▶ 004	10/11	y	h	D
005	10/10	m	n	C

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

Dimensional Model

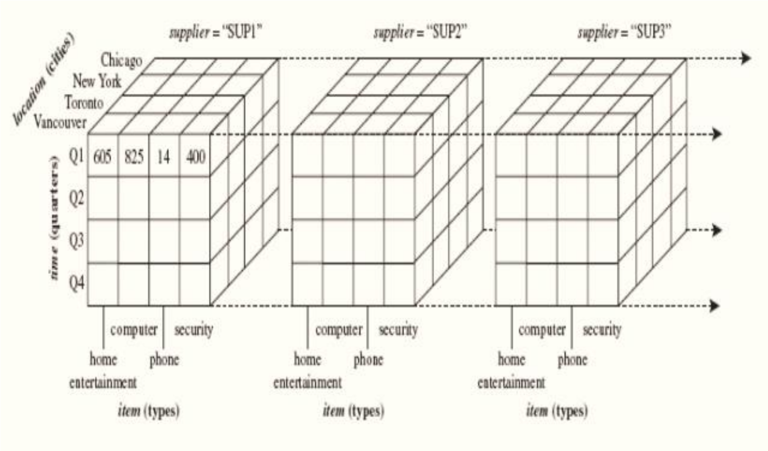


	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580



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

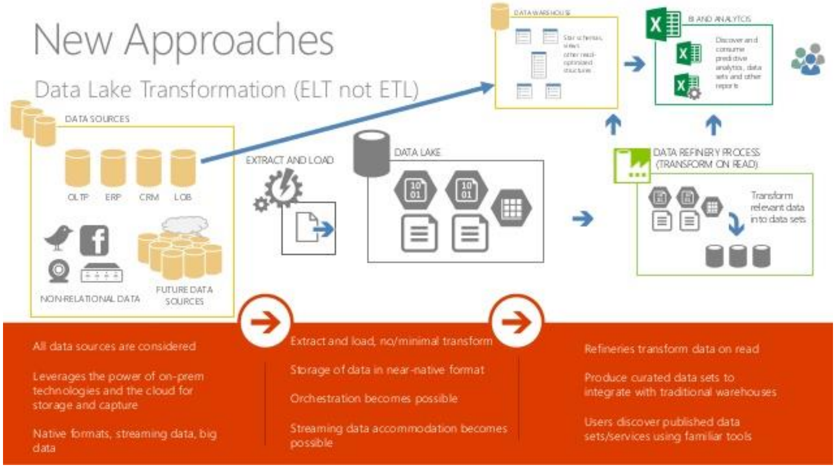
Data Lake

A storage repository, usually Hadoop, 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

Data Lake + Data Warehouse Better Together


New Approaches
Data Lake Transformation (ELT not ETL)
DATA SOURCES
OLTP
ERP
CRM
LOB
NON-RELATIONAL DATA
FUTURE DATA SOURCES
EXTRACT AND LOAD
DATA LAKE
10 01
DATA WAREHOUSE
Star schemas, views other read-optimized structures
BI AND ANALYTICS
Discover and consume predictive analytics, data sets and other reports
DATA REFINERY PROCESS (TRANSFORM ON READ)
Transform relevant data into data sets

All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data

Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

Streaming data accommodation becomes possible

Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

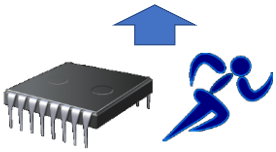
Users discover published data sets/services using familiar tools

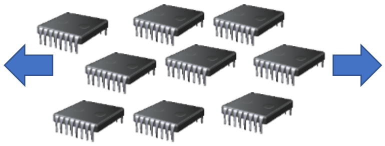
Data Warehouse vs Data Lake

	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/

Data Warehouse vs. Data Lake
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)

Make Many CPUs work together

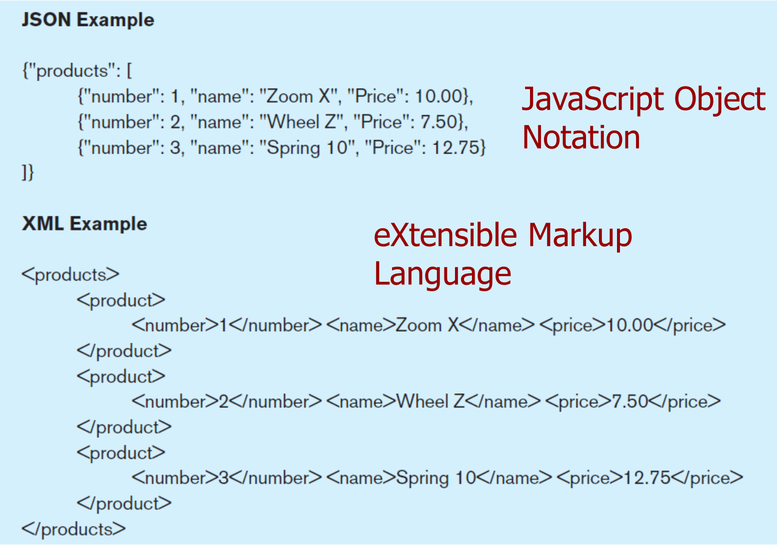
Learn how to divide your problems into independent threads

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}
]}
```

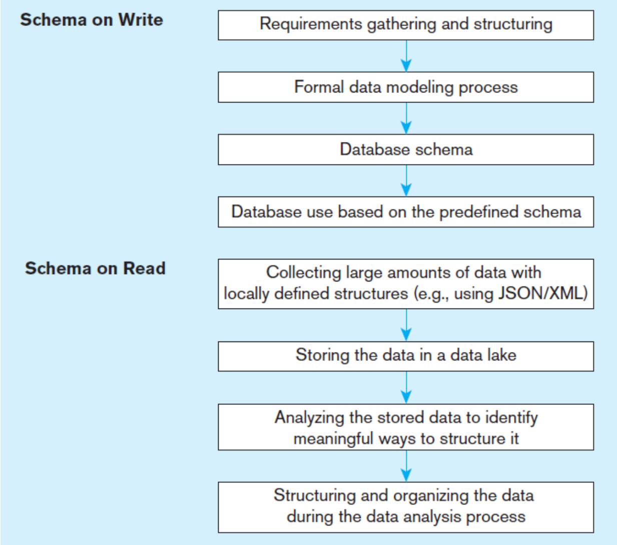
JavaScript Object Notation

XML Example

eXtensible Markup Language

```
<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>
```

Schema on write vs. schema on read



Schema on Write

Requirements gathering and structuring

Formal data modeling process

Database schema

Database use based on the predefined schema

Schema on Read

Collecting large amounts of data with locally defined structures (e.g., using JSON/XML)

Storing the data in a data lake

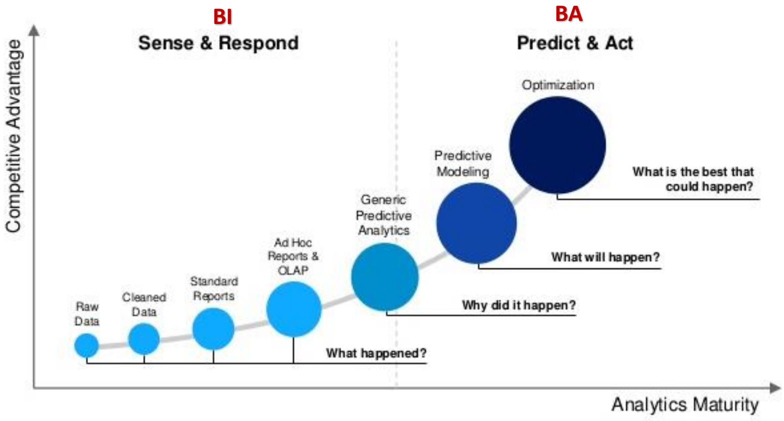
Analyzing the stored data to identify meaningful ways to structure it

Structuring and organizing the data during the data analysis process

Traditional database design

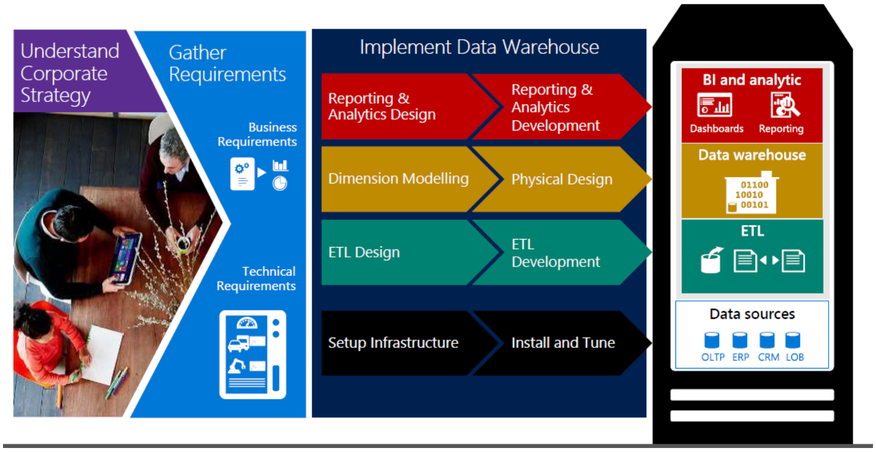
The big data approach

Evolution of BI



Source: Adapted from Delaware Consulting

Data Warehousing uses a Top-Down Approach



Understand Corporate Strategy

Gather Requirements

Business Requirements

Technical Requirements

Implement Data Warehouse

Reporting & Analytics Design

Reporting & Analytics Development

Dimension Modelling

Physical Design

ETL Design

ETL Development

Setup Infrastructure

Install and Tune

BI and analytic

Dashboards

Reporting

Data warehouse

01100
10010
00101

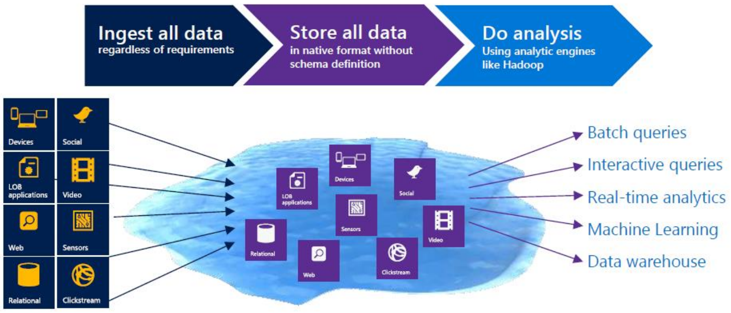
ETL

Data sources

OLTP ERP CRM LOB

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

The "data lake" uses a Bottom-Up Approach



Ingest all data regardless of requirements

Store all data in native format without schema definition

Do analysis Using analytic engines like Hadoop

Devices

Social

LOB applications

Video

Web

Sensors

Relational

Clickstream

Batch queries

Interactive queries

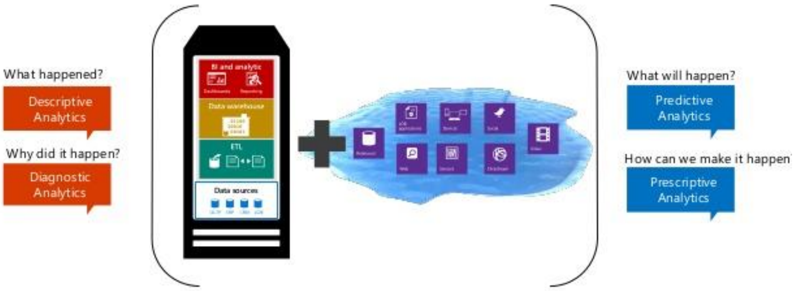
Real-time analytics

Machine Learning

Data warehouse

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

Data Lake + Data Warehouse Better Together



What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics

What will happen?

Predictive Analytics

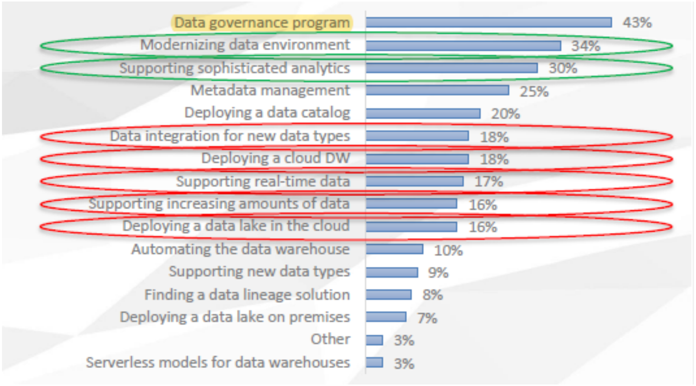
How can we make it happen

Prescriptive Analytics

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

Cloud Data Warehouse Trends For 2019
A Survey From TDWI and Talend (October 2018)

What are your organization's biggest priorities for data management in 2019? Please select up to 3 responses



Teradata University

https://academics.teradata.com/

Teradata Community

https://support.teradata.com/community

Executive Education

www.isegexecutive.education
Rua do Quelhas, 6
1200-781 Lisboa

(+351) 213 922 891
info@executive.education