DataWarehouses: motivations, definition

Chapter content:

- Motivations
- Definitions: DW, OLTP vs OLAP
- DW industrial landscape

What is a DataWarehouse?

widespread feeling that "society is data rich but information poor"

Business intelligence (BI): set of techniques and tools that enable a company to transform business data into into meaningful and useful information for decision making.

DataWarehouse (DW): repository that stores the data and infrastructure to support analysis.

according to R.Kimball:

> *"a copy of transaction data specifically structured for query and analysis"*

cf also Bill Inmon's more precise definition (later).

Table des matières

2018-2019

Motivation story

(source: `http://philip.greenspun.com/sql/data-warehousing.html`)

Let's *imagine* a conversation back in the 90s, between the Chief Information Officer of WalMart and a sales guy from Sybase.

Walmart: *"I want to keep track of sales in all of my stores simultaneously."*
Sybase: *"You need our wonderful RDBMS software. You can stuff data in as sales are rung up at cash registers and simultaneously query data out right here in your office. That's the beauty of concurrency control."*

So Walmart buys a $1 million HP multi-CPU server and a $500,000 Sybase license, and builds a normalized database:

`Sales`(product id, store id, quantity sold, date/time of sale)
`Products`(product id, product name, product category, manufacturer id)
`Stores`(store id city id, store location, phone number)
`Cities`(city id, city name, state, population)

Motivation story (continued)

Some time after, a Walmart executive asks: *"I noticed that there was a Colgate promotion recently, directed at people who live in small towns. How much Colgate toothpaste did we sell in those towns yesterday? And how much on the same day a month ago?"*

Her query looks like:

```
SELECT sum(sales.quantitysold)
FROM sales, products, stores, cities
WHERE products.manufacturer_id = 68 -- restrict to Colgate-Palmolive
 and products.product_category = 'toothpaste'
 and cities.population < 40000
 and sales.datetime_of_sale::date = 'yesterday'::date -- restrict to yesterday
 and sales.product_id = products.product_id
 and sales.store_id = stores.store_id
 and stores.city_id = cities.city_id
```

The query returns after 20mins. But dbadmins realize cash registers cannot process the sales when the toothpaste query is run.

Motivation story (continued)

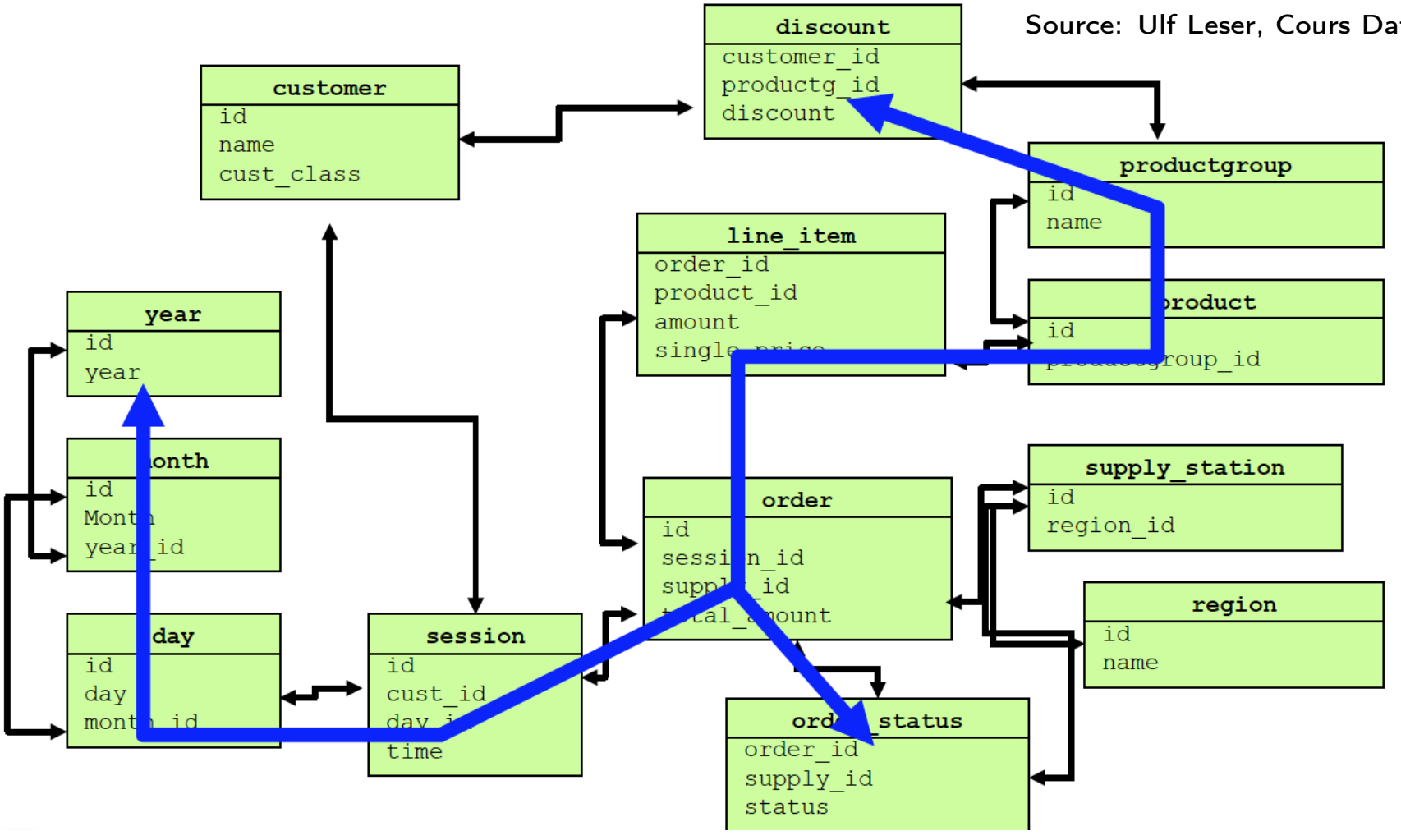
Walmart: *"We type in the toothpaste query and our system wedges."*
Sybase: *"Of course it does! You built an on-line transaction processing (OLTP) system. You can't feed it a decision support system (DSS) query and expect things to work!"*
Walmart: *"But I thought the whole point of SQL and your RDBMS was that users could query and insert simultaneously."*
Sybase: *"Uh, not exactly. The system prevents simultaneous Writes and Reads to guarantee coherent information: this is called "pessimistic locking".*
Walmart: *"Can you fix your system so that it doesn't lock up?"*
Sybase: *"No. But we made this great loader tool so that you can copy everything from your OLTP system into a separate DSS system at 100 GB/hour."*

This new system is the *data warehouse*. This replication of data will

- allow sales to be recorded in the database without interruption
- allow to reorganize the copied data to optimize analytical queries (e.g., better schema)
- allow to integrate smoothly data from different stores (e.g., after a merger with Kmart if Kmart records transactions in a Teradata system).

Motivations (2)

Analyst: *"How many sales completed in dec. before Christmas per group of product and discount?"*



Source: Ulf Leser, Cours Data Warehouse

Large relations (millions of orders,sessions), many joins ⇒ hard query.

Motivations (2)

Analyst: *"How many sales completed in dec. before Christmas per group of product and discount?"*

```
SELECT Y.year, PG.name, DI.disc, count(*)
FROM year Y, month M, day D, session S,
  line_item I, order O, product P, productgroup PG,
  discount DI, order_status OS
WHERE M.year_id = Y.id and
  D.month_id = M.id and
  S.day_id = D.id and
  O.session_id = S.id and
  I.order_id = O.id and
  I.product_id = P.id and
  P.productgroup_id = PG.id and
  DI.productgroup_id = PG.id and
  O.id = OS.order_id and
  D.day < 24 and
  M.month = 12
  and OS.status='FINISHED'
GROUP BY Y.year, PG.name, DI.discount
ORDER BY Y.year, DI.discount
```

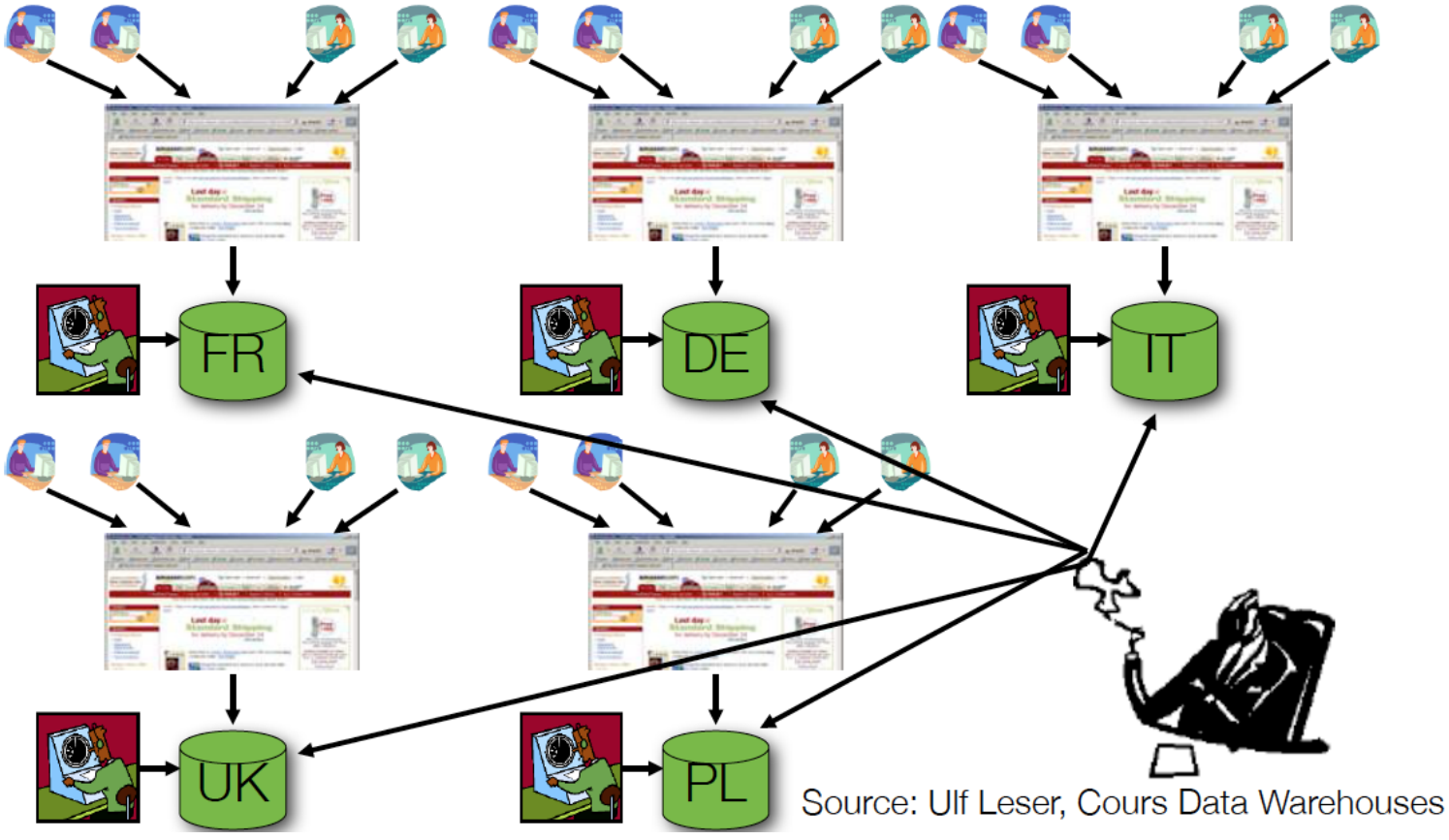
Source: Ulf Leser, Cours Data Warehouses

Large relations (millions of orders,sessions), many joins $\Rightarrow$ hard query.

Motivations (2)

Analyst: *"How many sales issued in dec. before Christmas per group of product and discount?"*

Amazon.fr, Amazon.de, ...



Source: Ulf Leser, Cours Data Warehouses

Motivations (2)

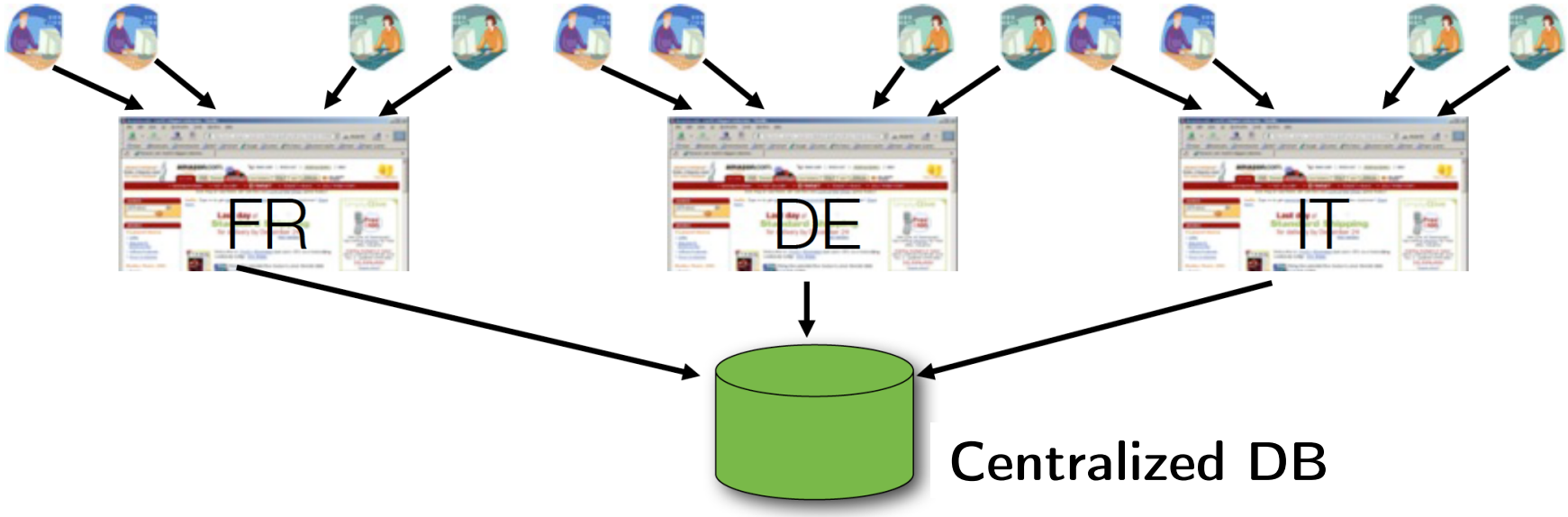
Analyst: *"How many sales issued in dec. before Christmas per group of product and discount?"*

Amazon.fr, Amazon.de, ...

1. Heterogeneity
   - the database schemas are modified from time to time
   - some properties specific to each country (TVA, expedition cost,etc.)
   - different semantics for data (measures, etc.)
2. Amount of data
   - hard query $\Rightarrow$ can re-use for similar queries if view "christmasOrders"
   - Network usage: transfer large amounts of data through web
   - Diverging requirements:
     - Operations do not need historical data (purge past orders)
     - The analyst has no need for details (customer name, supplier...)

Motivations (2)

Limitations of some possible solutions

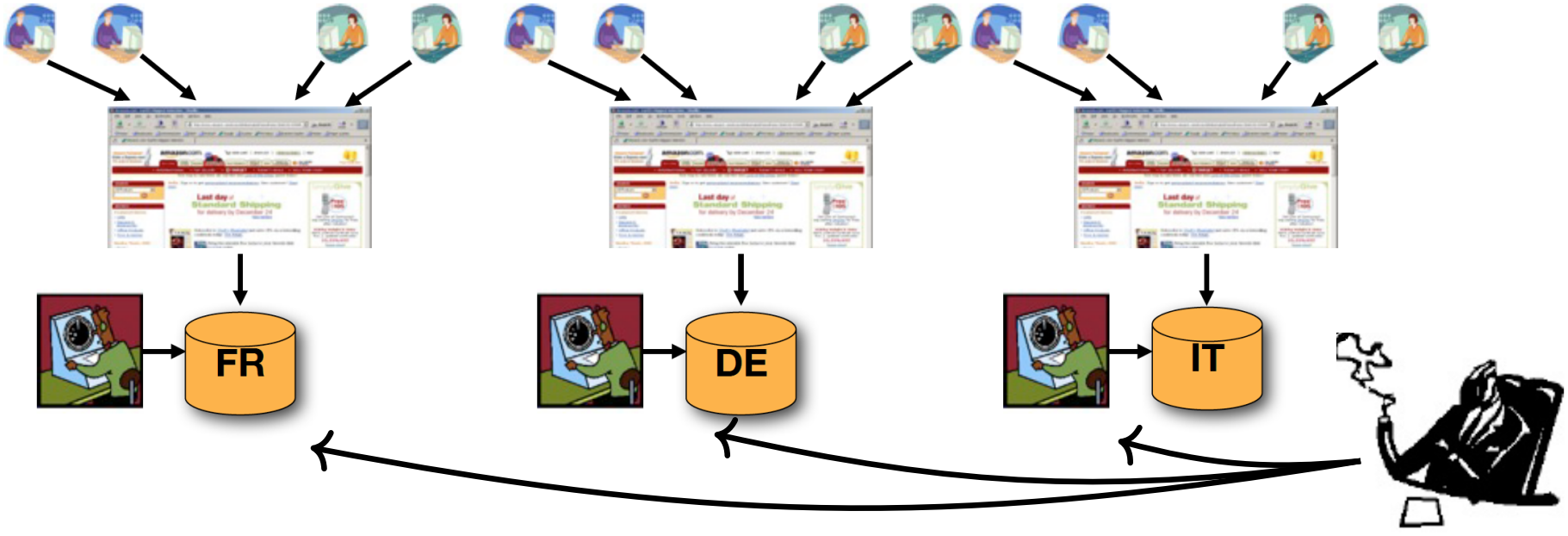


Solution to heterogeneity but

- each branch must connect through the network
- long delay for operations
- does not help with amount of data

Motivations (2)

Limitations of some possible solutions

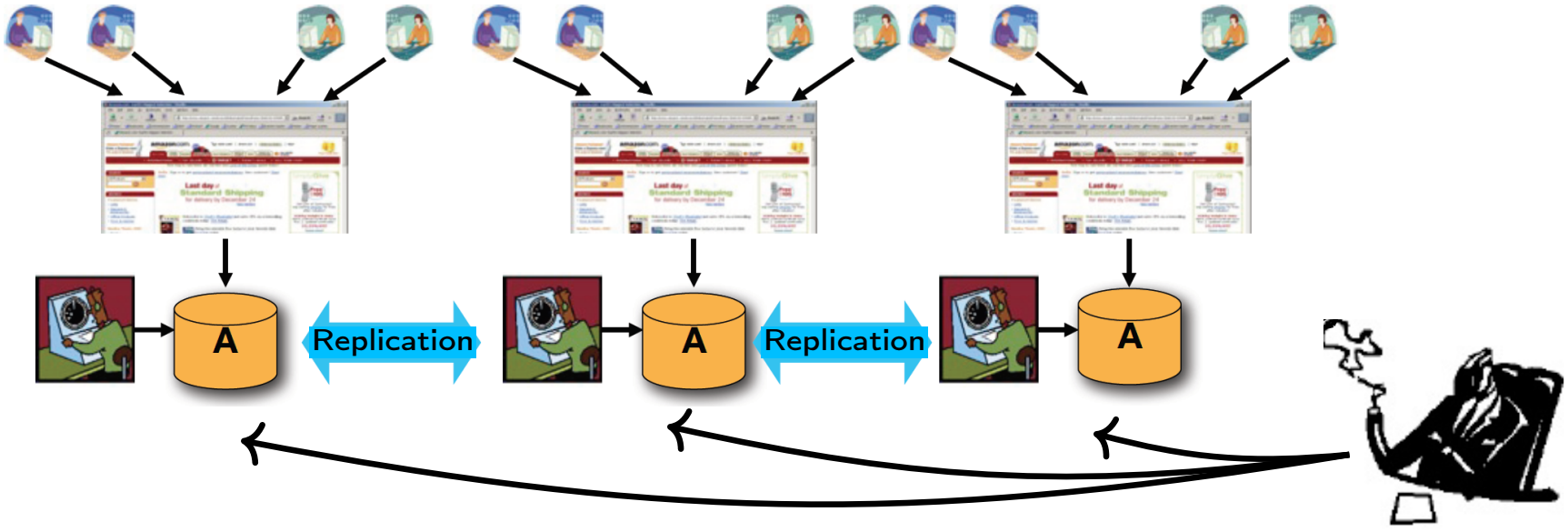


Short delay on operational transactions but

- does not help with heterogeneity issue
- long delay for analytical queries

Motivations (2)

Limitations of some possible solutions

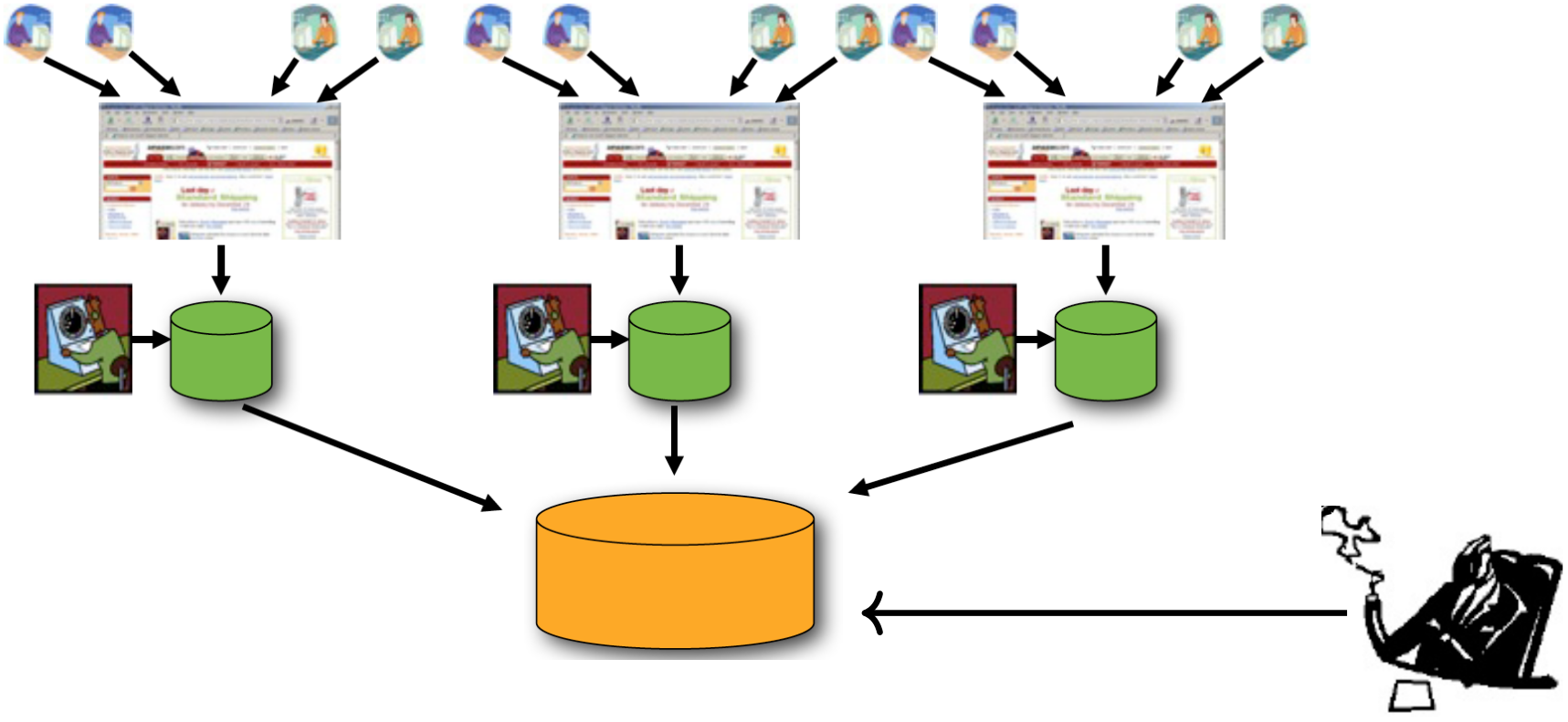


Short delay on analytical queries but

- huge relations in local databases
- longer delay on operational transactions

Motivations (2)

The DataWarehouse approach



- redundant data
- DW stores selected and transformed data
- specific modelization
- asynchronous update of the DW data
- set of specific tools (data preparation, data visualization)

Table des matières

2018-2019

DataWarehouses: motivations, definition

- Motivations
- Definitions: DW, OLTP vs OLAP
- DW industrial landscape

Definition

William H.Inmon's definition ['92]

A data warehouse is a *subject oriented*, *integrated*, *time varying*, *non-volatile* collection of data in support of management's decision making process.

- **subject oriented**: the DW organized according to one or several subjects determined by the analysts' requirements
- **integrated**: the content results from the integration of data from multiple sources
- **time-varying**: keeps track of data changes so that reports show evolution over time
- **non-volatile**: new data can be added, but data is ±never deleted nor updated

OLTP vs OLAP

OLTP: *Online Transaction Processing*
OLAP: *Online Analytical Processing*

Aspect	Operational DB	DW
User	clerk	manager
Concurrency	huge (thousands)	limited (hundreds)
Interaction	short (s)	long analyses (min,h)
Type of interaction	Insert, Update, Delete	Read,periodically (bulk) inserts
Type of query	many simple queries	few, but complex queries (typically drill-down, slice...)
Query scope	a few tuples (often 1)	many tuples (range queries)
Data source	single DB	multiple independant DB...
Schema	query-independant (3NF)	based on queries
Data	original, detailed, dynamic	derived,consolidated, integrated,historicized,partially aggregated,stable
Size	MB,GB	TB,PB
Availability	crucial	not so crucial
Architecture	3-tier (ANSI-SPARC)	adapted to data integration

Table des matières

Application domains

- Retail
- e-business
- Banks, finance
- Insurances
- Telecoms
- Logistics, Travels, Hotels
- Health
- (Life,. . .) Science
- Public Administrations
- . . .

Application: Retail(1)

Walmart

Pioneer: one of the largest (retail) warehouses since. Teradata solution.

	92	2001	2004	2008
size	1 TB	70 TB	>500 TB	2.5 PB

Innovative practices... and secretive.

$20M Prototype DW to analyze sales history launched in 1990 while still unstable (40% queries rejected). A study over selected analysts estimated ROI per query at $12,000.

[*Data Warehousing: Using the Wal-Mart Model*,Westman]

I'll spare you the legendary data mining story about "beer and diapers" ;)
[http://www.dssresources.com/newsletters/66.php]

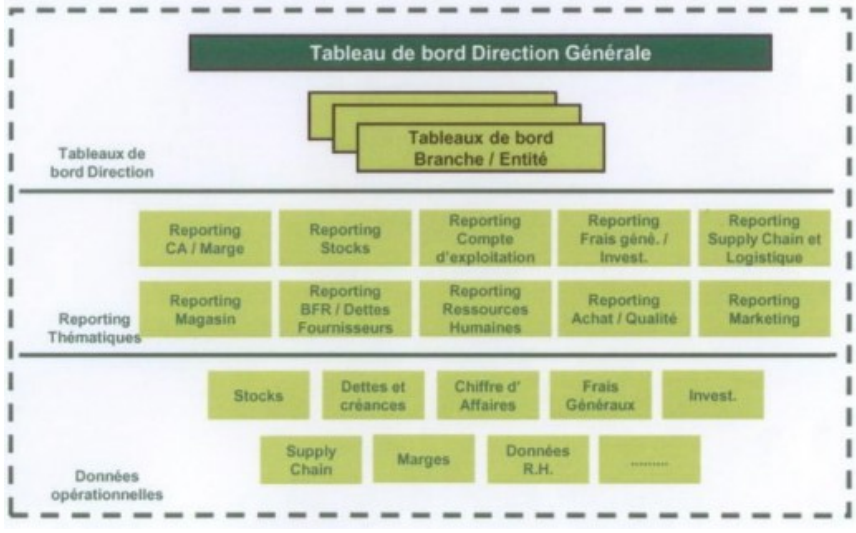
Application: Retail(2)

Casino group

One of the earliest DW in France. Teradata solution

	94	2002	2009
size	80 GB	10 TB	18 TB
users	50	1,500	3500
queries/day		25,000	600,000

Saved millions when realized that Coca-Cola stocks were often low [Marcel][Espinasse]



Sources: `http://www.mycustomer.com/topic/technology/casino-group-upgrades-teradata-data-warehouse`
`http://fr.teradata.com/newsrelease.aspx?id=12338`
`http://www.lemagit.fr/actualites/2240197570/Pour-harmoniser-ses-calculs-de-marge-Casino-sengage-dans-la-refonte-de-son-decisionnel`

Application: Retail(3)

Goals:

- improve sales, product offerings
- optimize supply chain
- merchandising (store layouts. . .)
- optimize promotions
- customer retention
- compliance

Application: Telecom

Bouygues Telecom

Back in 2007: due to market saturation, crucial shift from acquiring new customers to hanging on to the really valuable ones. Bouygues had disintegrated architecture with 3-4 data warehouses and 300 data marts $\Longrightarrow$ *BI results were unreliable.*

Restructured into a single Teradata DW:

1. consolidate analytics into a single DW (CRM, call detail records)
2. shorten latency: closer to real-time
3. support temporary "sandbox" access to the DW
4. linear scalability

E.g. compute customer churn scores in 4h.
Reduced maintenance overhead $\Longrightarrow$ 33% cost savings for the DW.

[http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB6365.pdf
Sandbox: an independant part of the DWH where people can experiment with new data (check quality) and tools.]

Application: Telecom (2)

France Telecom

Situation in 2002: Consolidates 50 Databases, 80 TB capacity, migration from hourly update to continuous.
Oracle solution on HP machine.

Performances: 500M CDR/day (100GB). 180 Billion CDR stored. 8000 end users, 600 peak concurrent access.
4s for standard queries.

[http://www.oracle.com/technetwork/testcontent/vldw-cases-winter-132893.pdf]

Goal summary for Telecom:

- Analyze traffic
- understand customer behavior (churn, product lifecycle,profiling)
- finance/billing operations

Other typical goals

(Finance, Insurance, Health)

- risk management
- claims analysis
- asset& liability management
- compliance

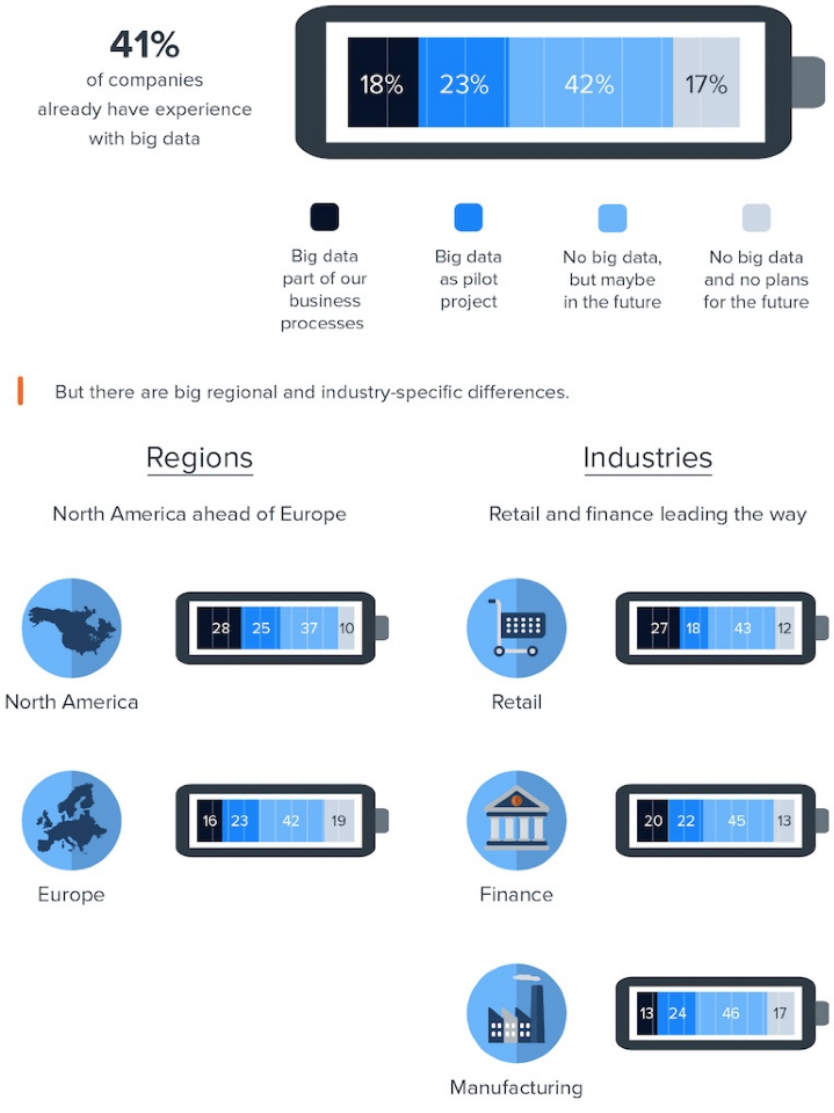
Product (suppliers)

Commercial:

Open Source:

600 M US $ acq. 06/2015

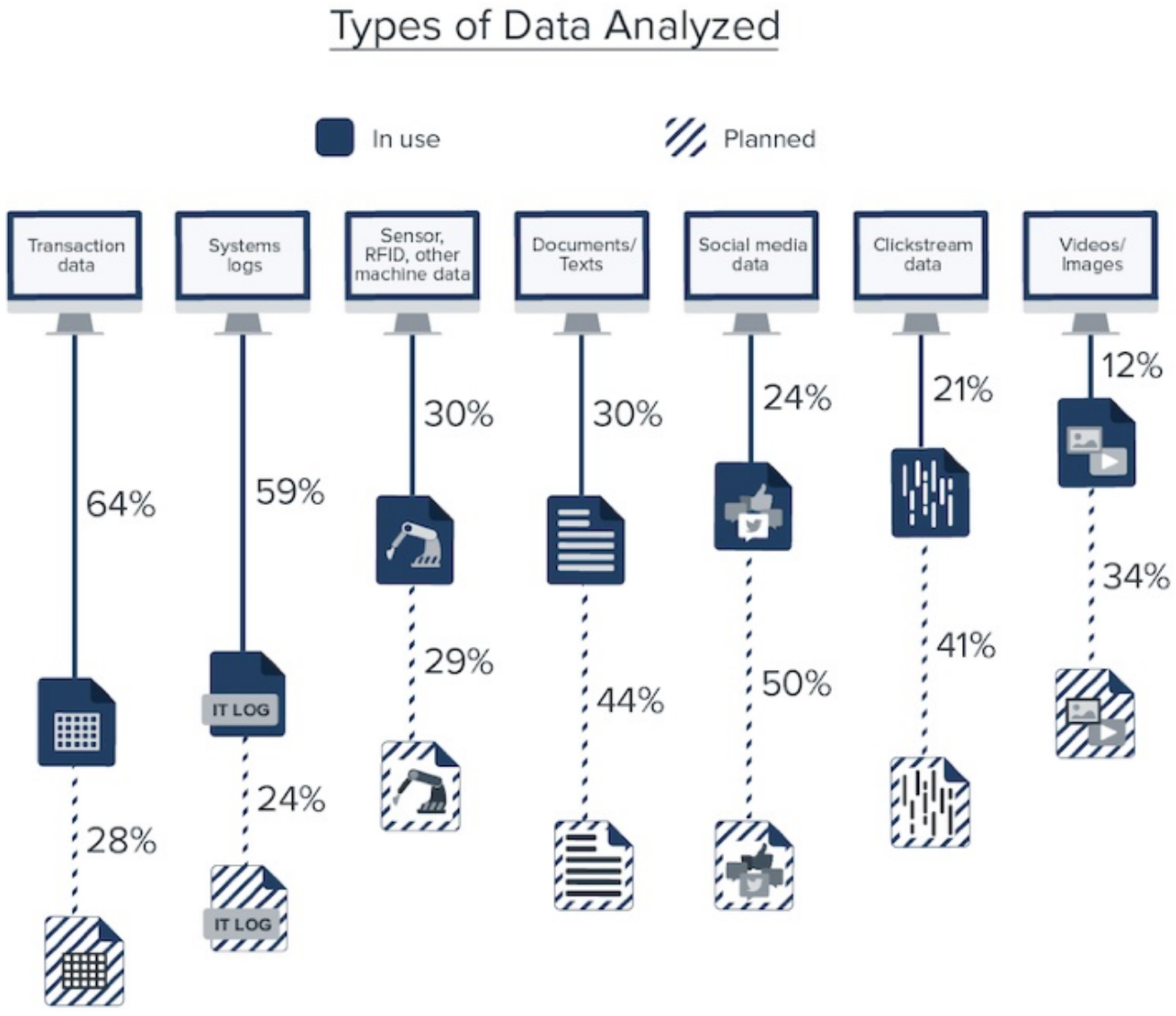
"Big data" BI market



[http://barc-research.com/big-data-use-cases-2015-infographic/]

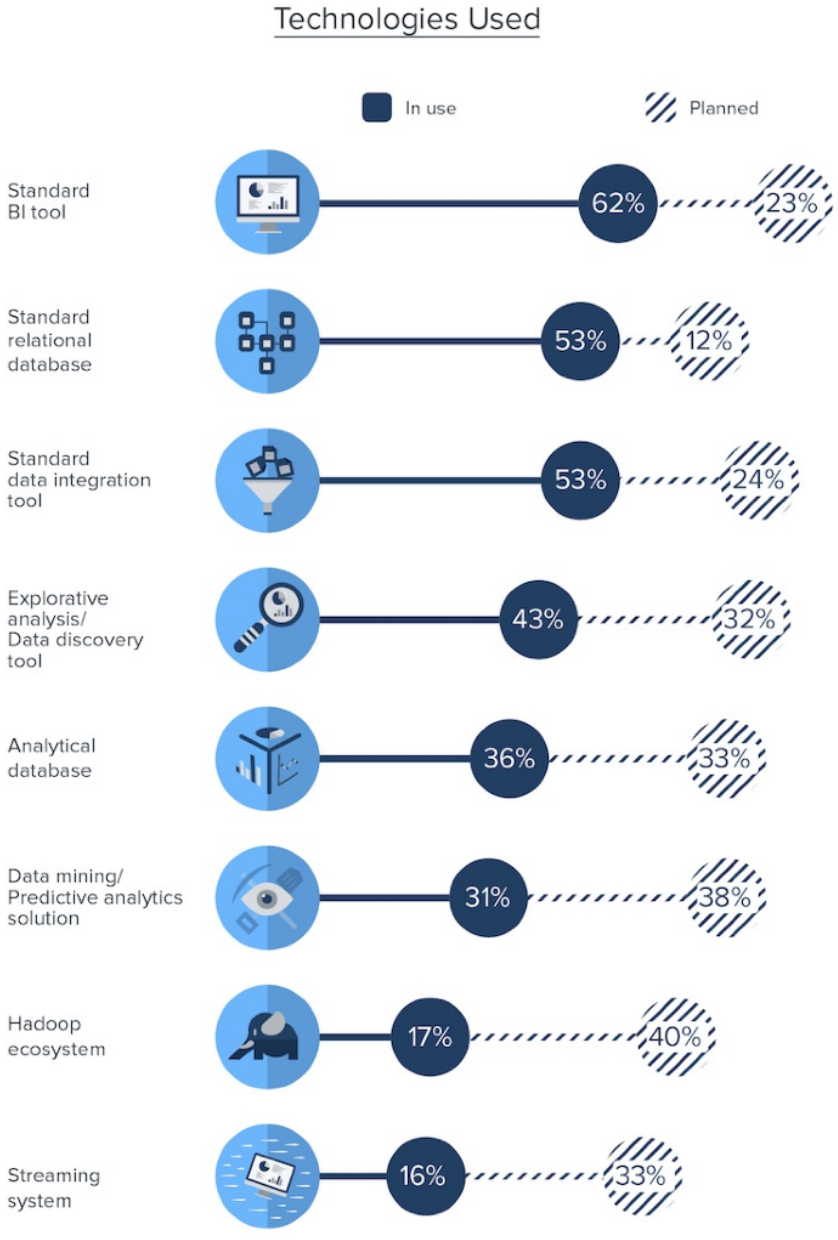
2015 survey conducted over 550 firms by Barc research: market analysis firm sponsored by hp, teradata, tableau, cloudera, etc.

"Big data" evolution



[http://barc-research.com/big-data-use-cases-2015-infographic/]

Practices are evolving quickly



These lectures focus on traditional SQL-based architectures (no Cloud. See Manolescu and Biffet).

[http://barc-research.com/big-data-use-cases-2015-infographic/]