Semantic Data Management

(Erasmus Mundus Master in Big Data Management and Analytics)

Open Data

(Master in Innovation and Research in Informatics – Data Science)

Introduction and Motivation

VARIETY IN COMPLEX DATA ECOSYSTEMS

“WITHOUT DATA, YOU’RE JUST ANOTHER PERSON WITH AN OPINION”

W. Edwards Deming, American Statistician

Data As The New Cornerstone

We have witnessed the bloom of a new business model based on data analytics: *Data is not a passive but an active asset*

- «Data is the new oil!» - Clive Humby, 2006
- «No! Data is the new soil» - David McCandless, 2010

The effective use of data to make decisions gave rise to the**data-driven society** concept

The confluence of three major socio-economic and technological trends makes data-driven innovation a new phenomenon today. These three trends include (OECD):

- The exponential growth in data generated and collected,
- the widespread use of data analytics including start-ups and small and medium enterprises (SMEs), and
- the emergence of a paradigm shift in knowledge (subjective predictions vs. evidence-based predictions)

Organizations must adapt their **infrastructures** to benefit from the data deluge

- Digital data doubling every 18 months (IDC)

Innovation is mandatory!

New Business Model: Instagram's Fable

(xkcd.com)

Current Trends

- The overkilling approach (end-to-end monitoring)
  - Facebook: Facebook + Instagram + Whatsapp + LinkedIn + …
  - Google: Android + Google Search + Calendar + Gmail + Doodle + …

- Domain-oriented approaches
  - Cross available data (companies fusion, buying data, open data, agreements, etc.)
  - Digitalise the organization processes to measure business metrics (e.g., the national health system, tax collection, e-banking, etc.)
  - Monitor the user to learn habits (phone apps, internet navigation, service usage, wearables, RFID, etc.)
    - Sometimes, in an indirect way (e.g., provide (free) services to learn habits; such as free wi-fi to geolocate the user)
  - Sensor monitoring (Smart Cities, Internet of Things, etc.)
  - …

Bottom line: most of the times, the most interesting data is not available (innovation comes to play!)

Same Purpose; Different Means

The economic and social role of data is not new

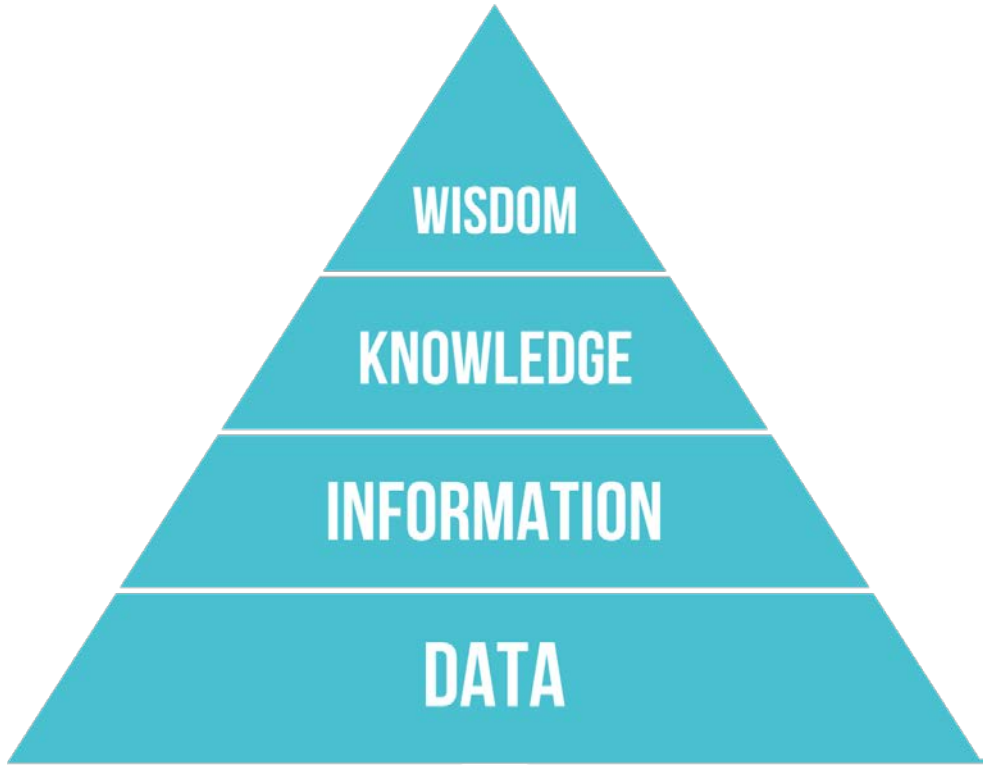
- Economic and social activities have long evolved around the analysis and use of data
- In business, concepts such as **Business Intelligence** and **Data Warehousing** already emerged in the 1960s and became popular in the late 1980s

Decision support systems refer to any IT system supporting decision making

- The **Business Intelligence Lifecycle** as cornerstone
- Later, **Big Data** broadened the principles in which **Business Intelligence** is built

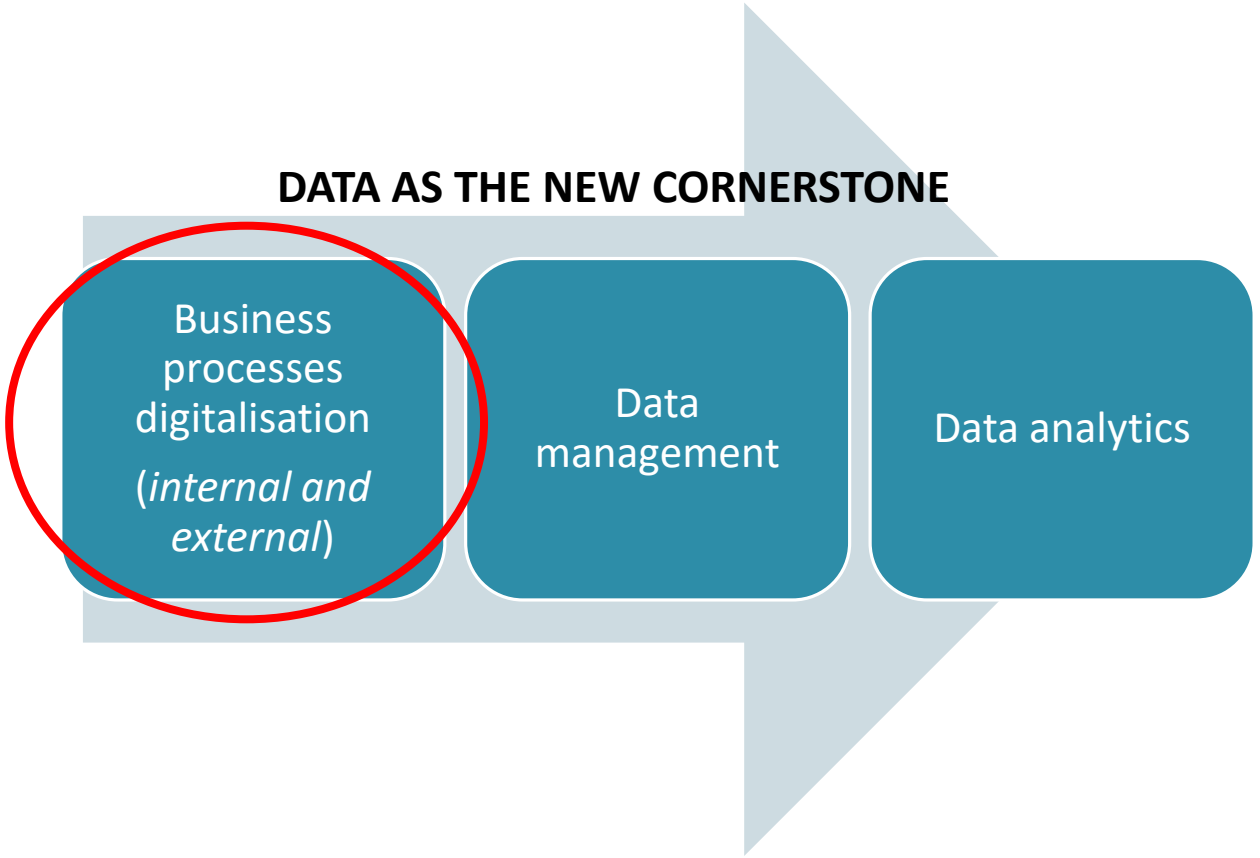
Decision Support Systems

IT systems aimed at exploiting data and transform it into information and knowledge



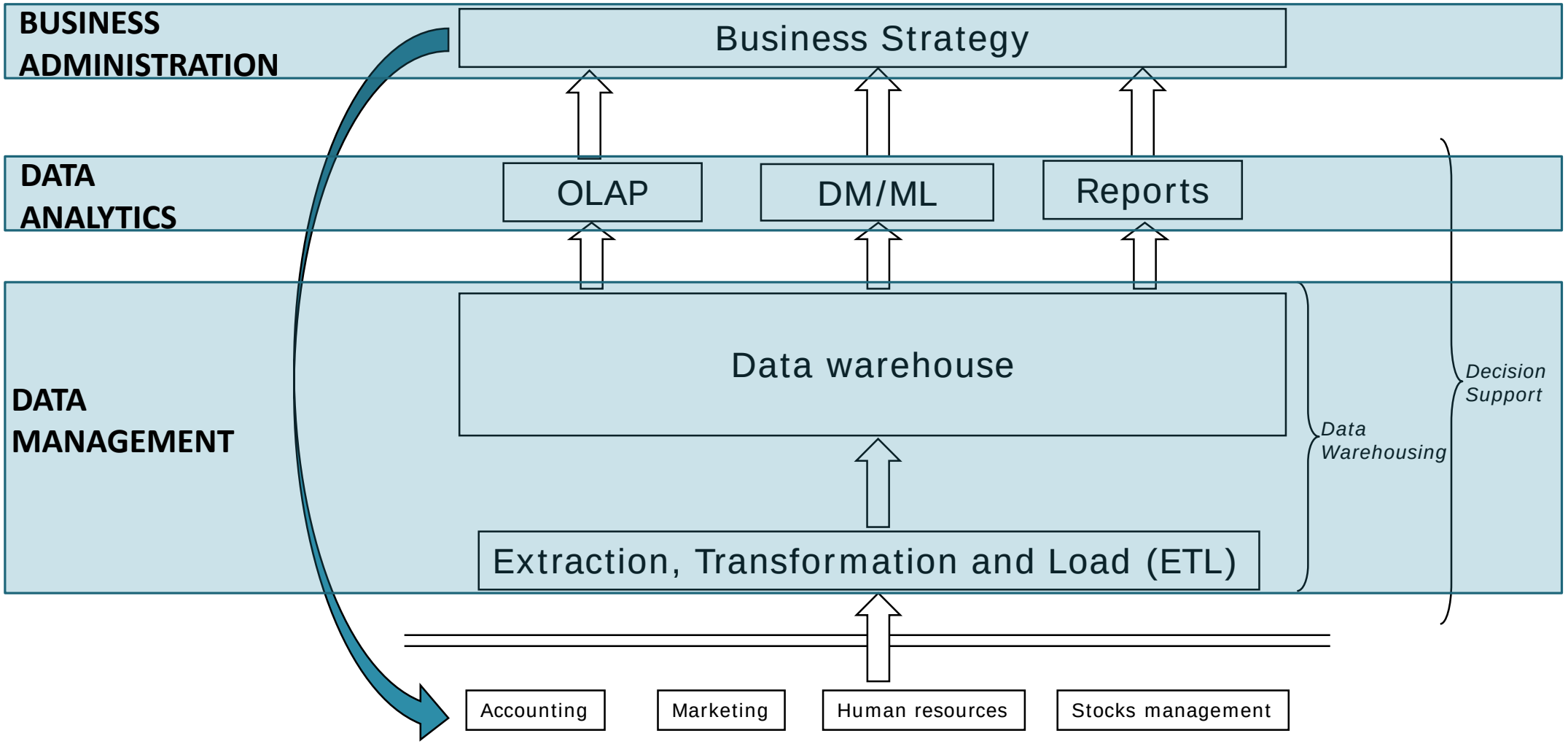
DIKW pyramid

Decision Support Systems



Business Intelligence

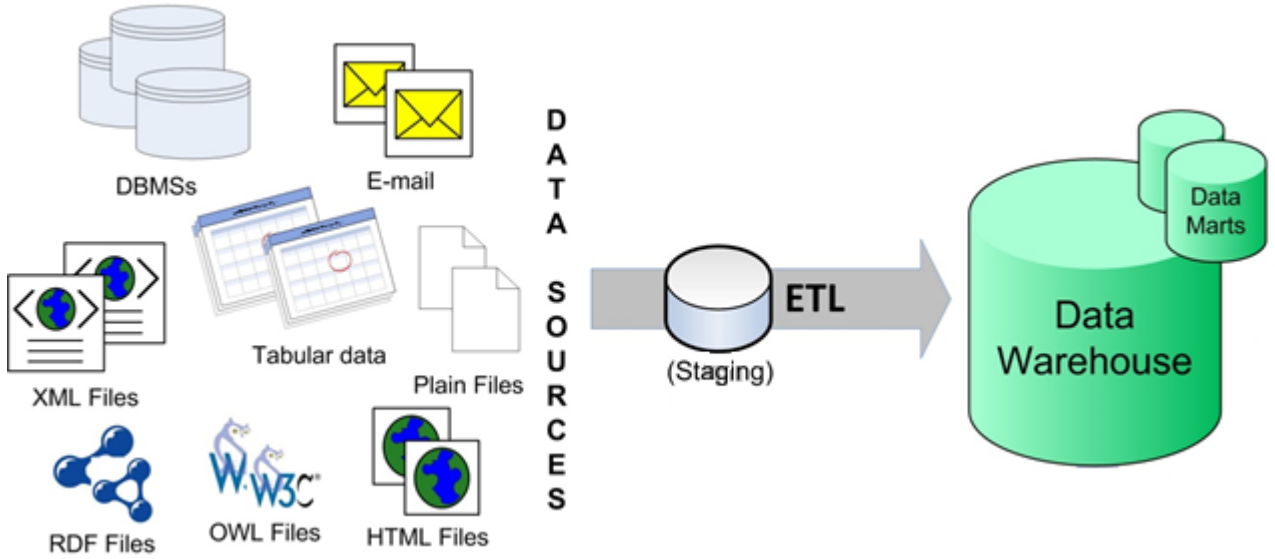
Business Intelligence Lifecycle



Business Intelligence: Data Management

Well-established de facto standards:

- Architecture: Data warehousing
- Modeling: Multidimensional modeling

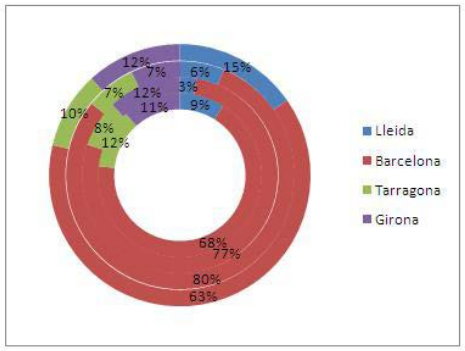


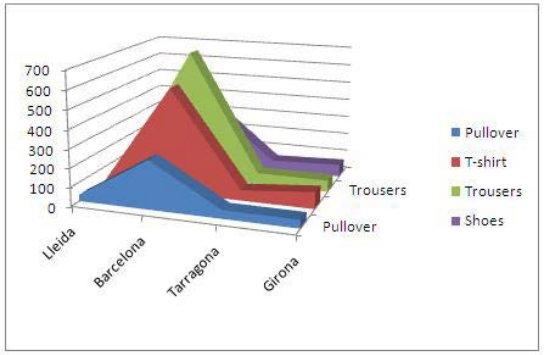
Business Intelligence: Analytics

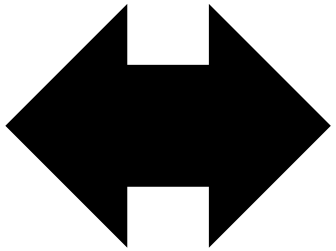
Typically, the analysis of the data has been considered at three different levels of detail

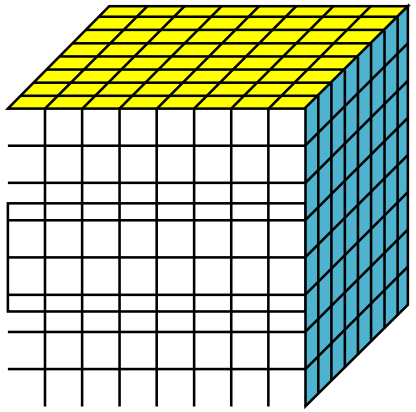
- Querying & Reporting: Static report generation
- OLAP: Dynamic summarizations of data
- Data Mining and Machine Learning: Inference of hidden patterns or trends

	Lleida	Barcelona	Tarragona	Girona
Pullover	34	260	45	44
T-shirt	20	564	56	89
Trousers	55	700	63	62
Shoes	87	360	54	67





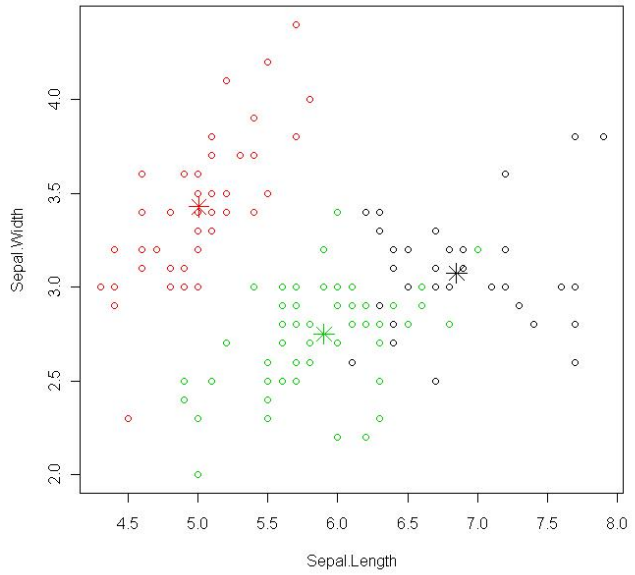




Business Intelligence: Analytics

Typically, the analysis of the data has been considered at three different levels of detail

- Querying & Reporting: Static report generation
- OLAP: Dynamic summarizations of data
- Data Mining and Machine Learning: Inference of hidden patterns or trends




```
> (kc <- kmeans(newiris, 3))
K-means clustering with 3 clusters of sizes 38, 50, 62

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.006000    3.428000     1.462000    0.246000
3     5.901613    2.748387     4.393548    1.433871

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [30] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3
 [59] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3
 [88] 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1 1 1 3 3 1
[117] 1 1 1 3 1 3 1 3 1 1 3 3 1 1 1 1 1 3 1 1 1 1 3 1 1 1 3 1 1
[146] 1 3 1 1 3

Within cluster sum of squares by cluster:
[1] 23.87947 15.15100 39.82097

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

Big Data

Data Warehousing Vs. Big Data

Big Data can be seen as the evolution of Data Warehousing ecosystems after the **complete digitalisation of the business processes** and the **incorporation of relevant external data** to the decision making processes of the organization

- Open data (external) Vs. organization data (internal)
  - Semi-structured and unstructured data as first-class citizens
- Adding *plug-and-play* data sources Vs. monolithic sources
  - Not in control Vs. Well-controlled sources
- *On-demand* data quality threshold
  - Lightweight transformations Vs. heavy transformations
- *On-demand* analysis Vs. traditional analysis
  - Load-first model-later (Data Lake) Vs. schema-fixed approach (DW)
- And many others consequences…
  - Semantic-aware solutions
  - Privacy
  - Etc.

What is Big Data?

VOLUME

Veracity

Velocity

Value

vArIaBiLiTy

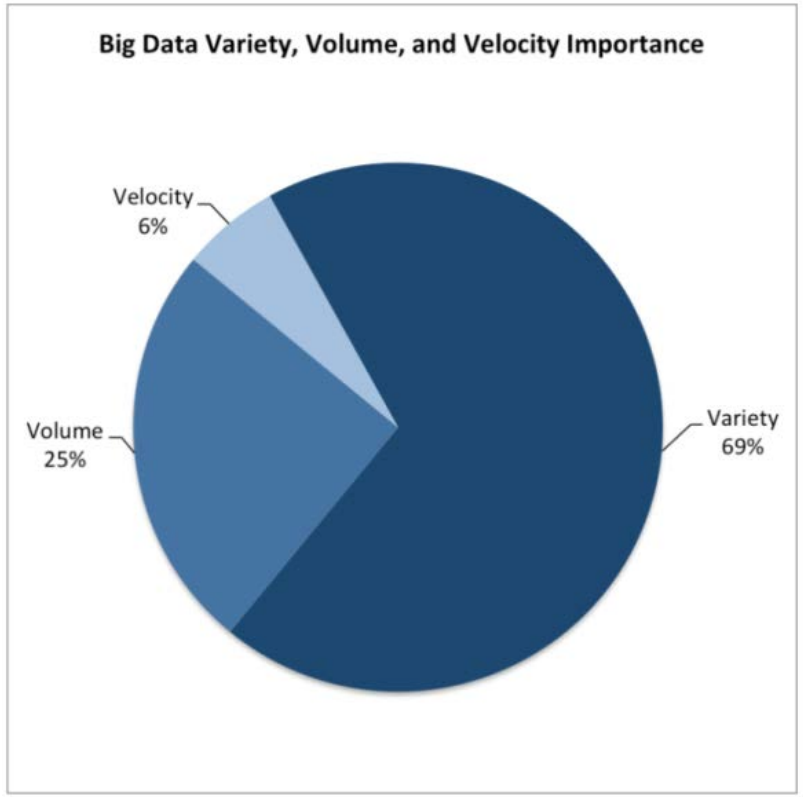
Variety

Today, the Focus is on Variety

That Big Data is synonymous with large volumes of data is a **myth**

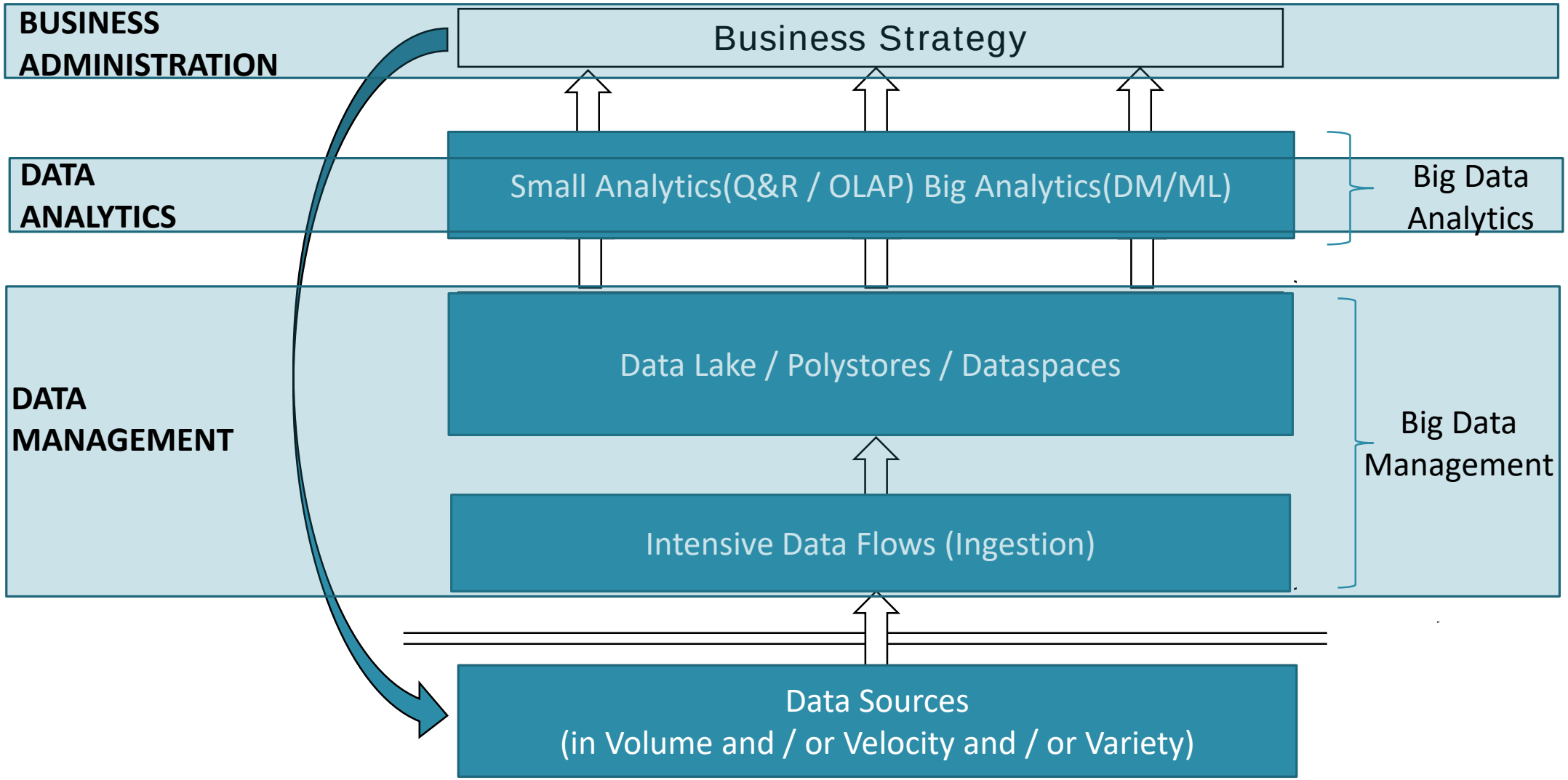
"*Rather, it is the ability to **integrate** more sources of data than ever before — new data, old data, big data, small data, structured data, unstructured data, social media data, behavioral data, and legacy data*"

The Variety Challenge



MIT Sloan Management Review (2016): http://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/

The Big Data Lifecycle



The Long Tail of Big Data

The ultimate goal is…

- **Integrate** new data sources **on-demand**,
  - Legacy Systems
  - External Data (typically, semi-structured or unstructured data)
  - Social Media and Behavioural Data Sources
- Provide the required flexibility for conducting **on-demand data analysis** techniques
  - Data preparation

Change of paradigm

Data Analysis Democratisation

From Model-First (Load-Later) to Load-First Model-Later

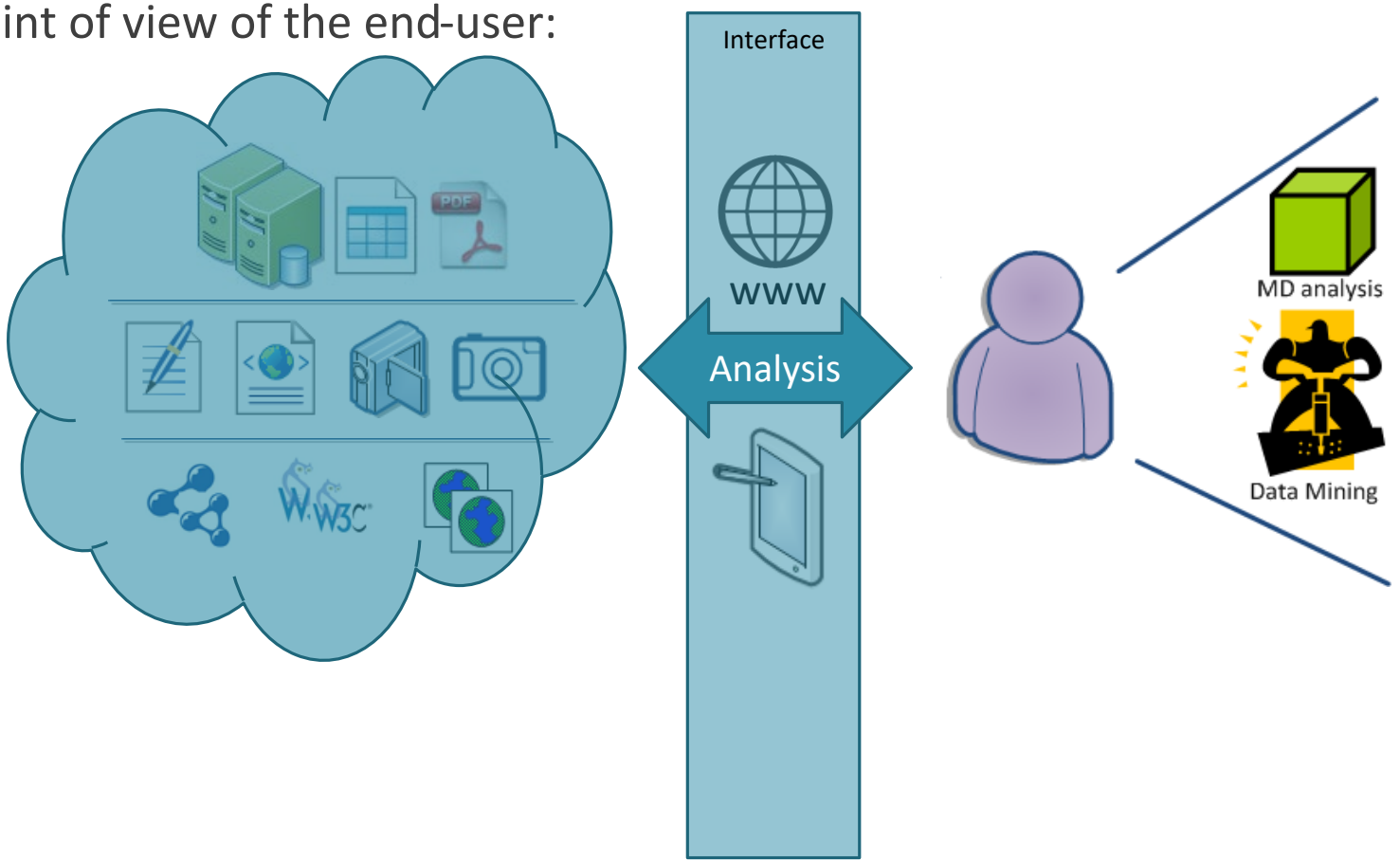
Full automation of the data pipeline

Challenges

FROM THE END-USER POINT OF VIEW

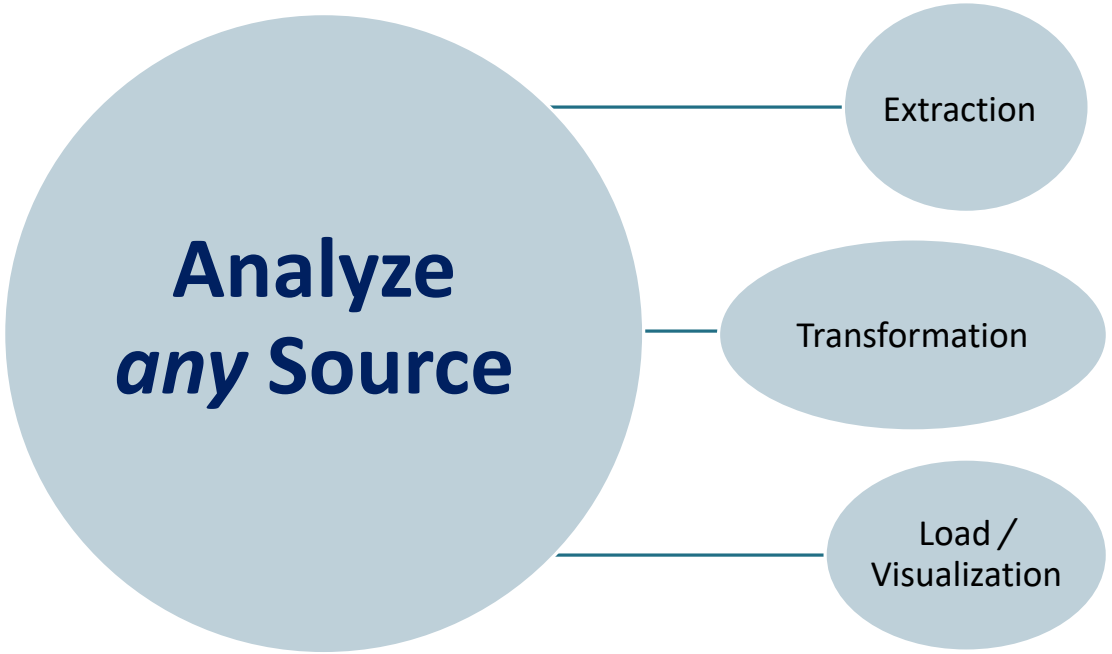
Data Analysis Democratisation

From the point of view of the end-user:



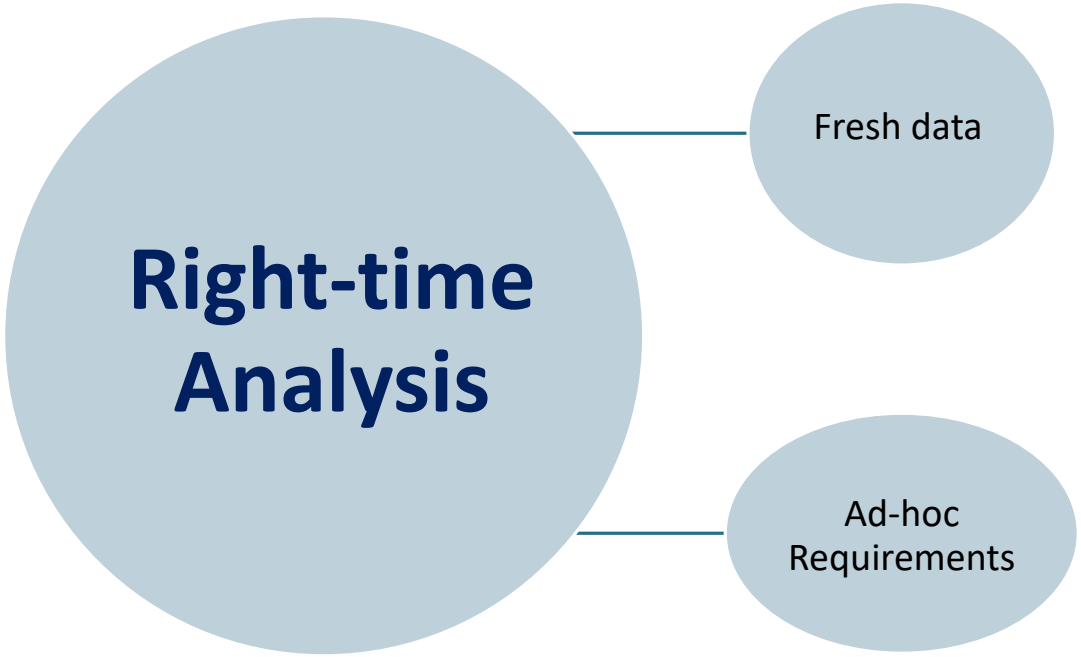
Data Analysis Democratisation

From the point of view of the end-user:



Data Analysis Democratisation

From the point of view of the end-user:



Examples

- *A company wants to enrich their DW data (products, customers, shops) with external data coming from Twitter (opinion about their products), logs generated by their web and feedback received by phone, web or the company app. The constraint is that external data must be loosely coupled but be completely integrated from an analysis point of view*
- *A journalist wants to analyze the evolution of asylum demands in Europe in the last 10 years. What data is available in the EU Open Data portals available? How data should be fetched and crossed in order to answer her questions?*
- *An organization wants to develop a Big Data platform to fulfill the need to deal with massive heterogeneous and unstructured data (new data sources with unknown schema may appear at any moment). For now, they are happy to ingest data in batch mode but they would like to process the data in real-time in the forthcoming months*
- *A company with a large dataset of information from their activities want to deploy a flexible Machine Learning pipeline. They want to avoid data analysts spend 80% of their time pre-processing data. There are many reasons: (i) data analysts should not repeat the same transformations once and again, (ii) their should collaboratively share their knowledge pre-processing data and (iii) they want to avoid lock-in knowledge (i.e., analysts keep their knowledge in personal scripts not shared with the company)*

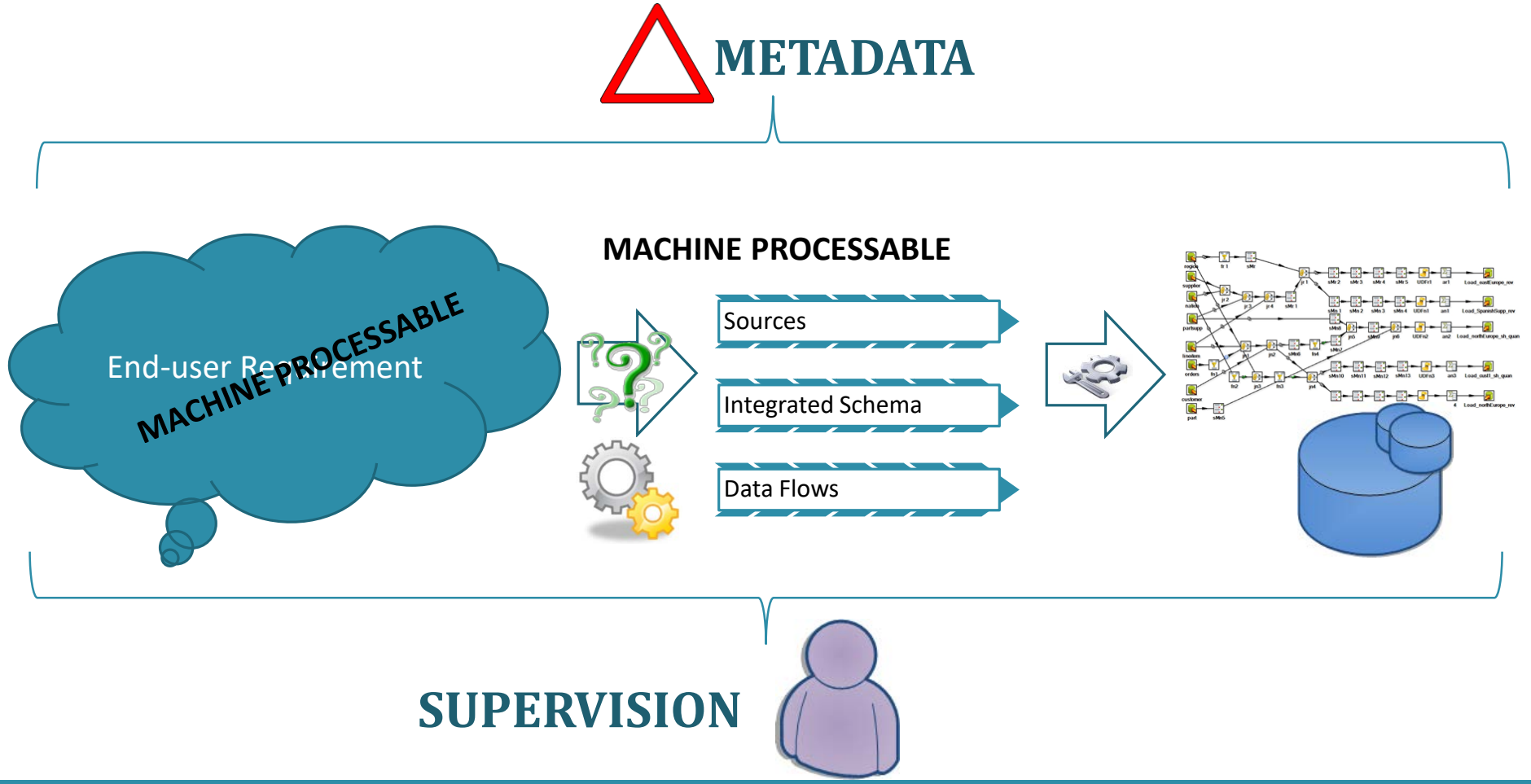
Yet Another Example

How Our Likes Helped Trump to Win

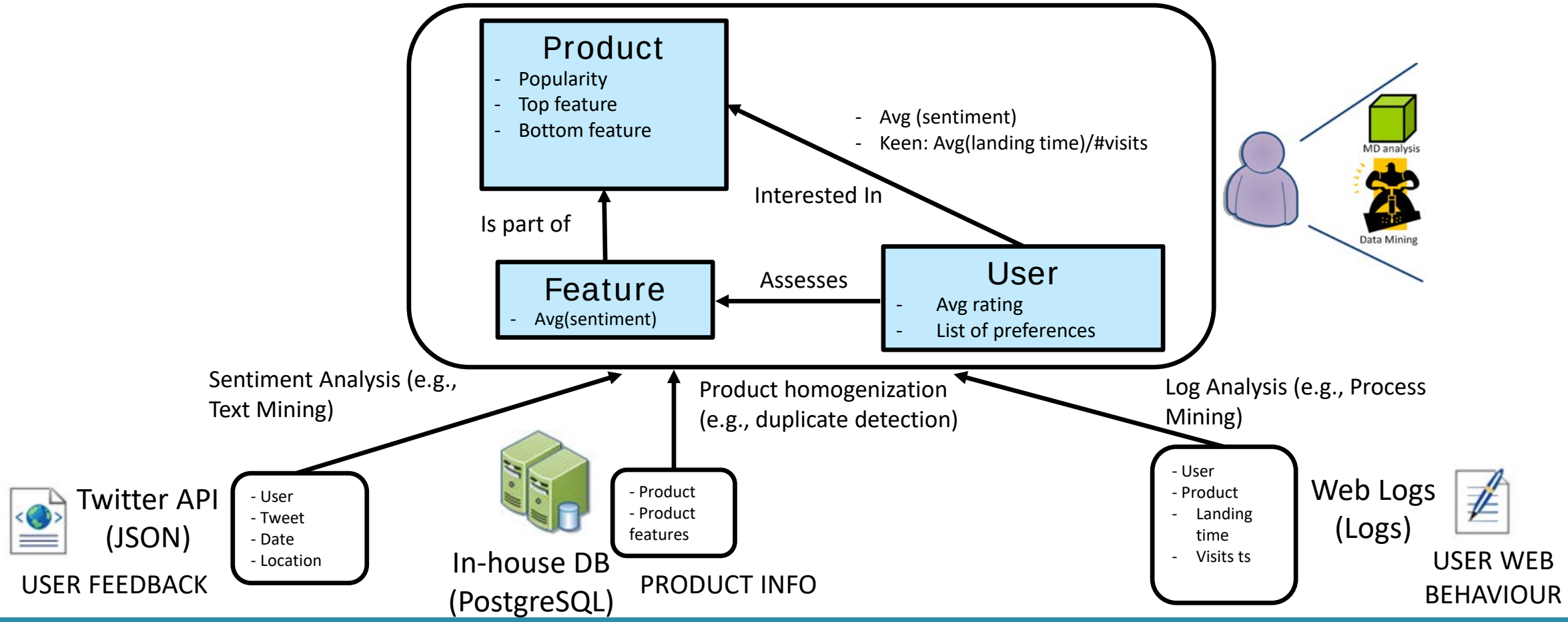
Challenges

FROM THE IT POINT OF VIEW

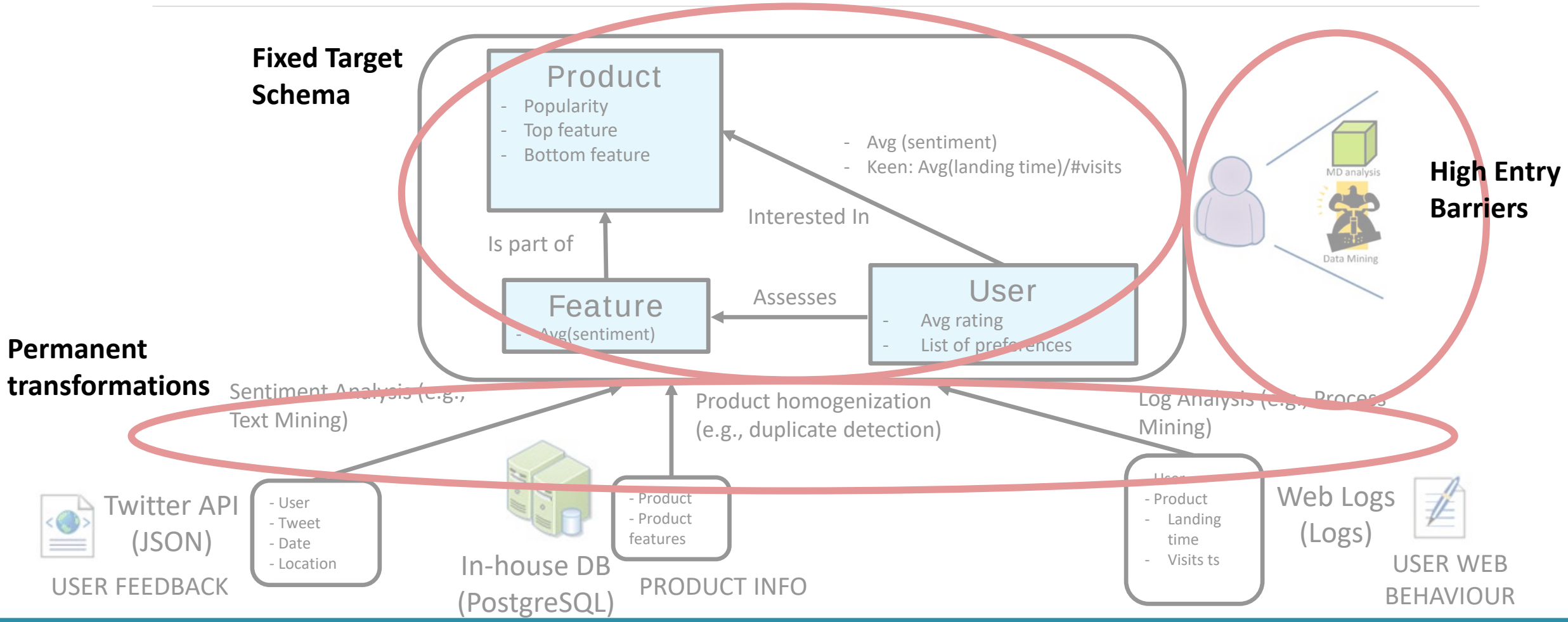
The Missing Link: Metadata



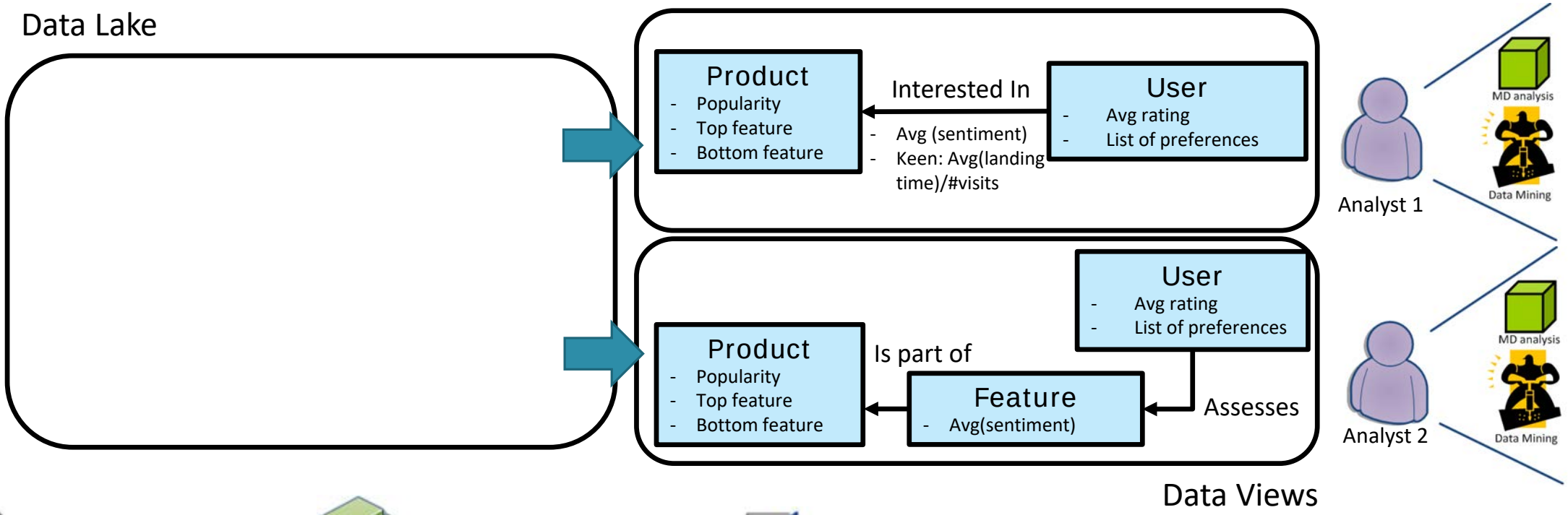
Model-First (Load-Later)



Drawbacks


Fixed Target Schema
Product
- Popularity
- Top feature
- Bottom feature
- Avg (sentiment)
- Keen: Avg(landing time)/#visits
Interested In
Is part of
Feature
- Avg(sentiment)
Assesses
User
- Avg rating
- List of preferences
MD analysis
Data Mining
High Entry Barriers
Permanent transformations
Sentiment Analysis (e.g., Text Mining)
Product homogenization (e.g., duplicate detection)
Log Analysis (e.g., Process Mining)
Twitter API (JSON)
- User
- Tweet
- Date
- Location
USER FEEDBACK
In-house DB (PostgreSQL)
- Product
- Product features
PRODUCT INFO
- Product
- Landing time
- Visits ts
Web Logs (Logs)
USER WEB BEHAVIOUR

Load-First Model-Later



USER FEEDBACK

Twitter API
(JSON)

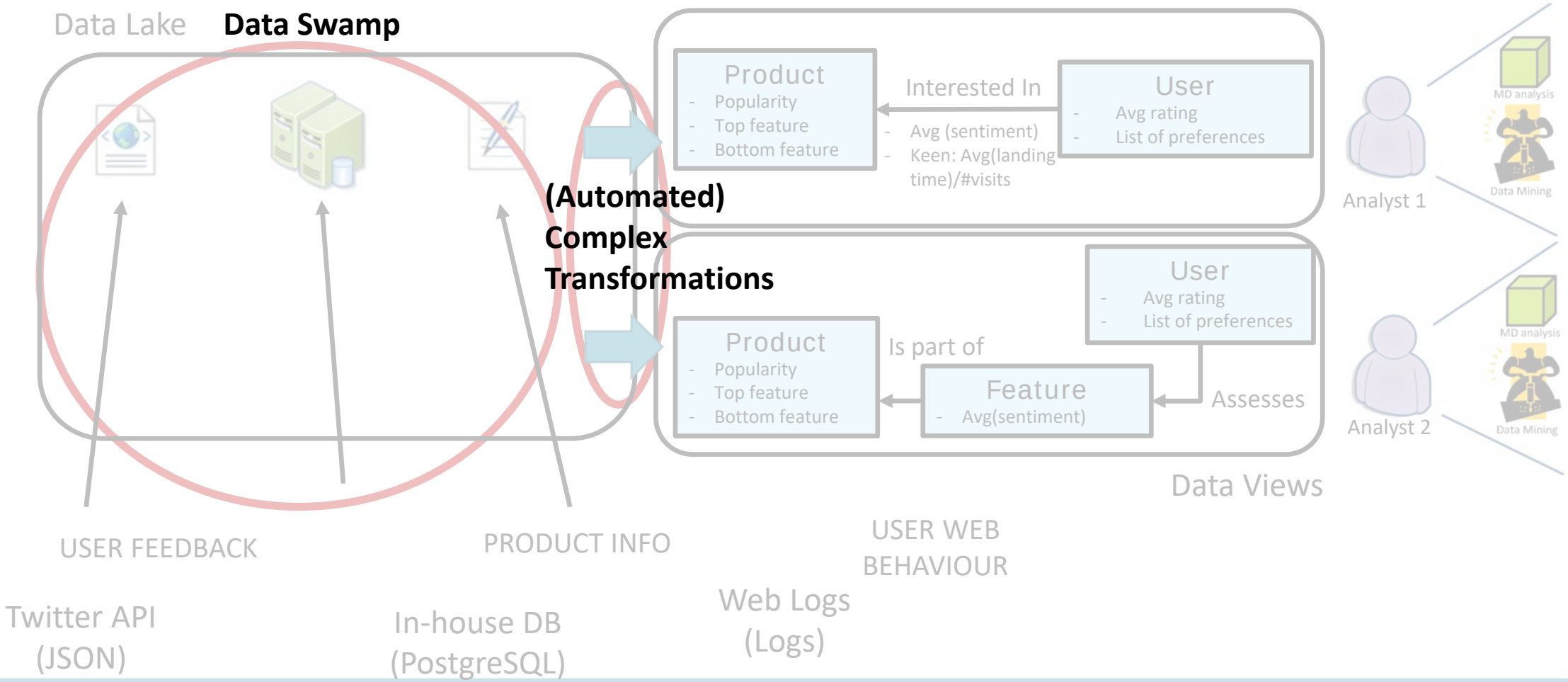
PRODUCT INFO

In-house DB
(PostgreSQL)

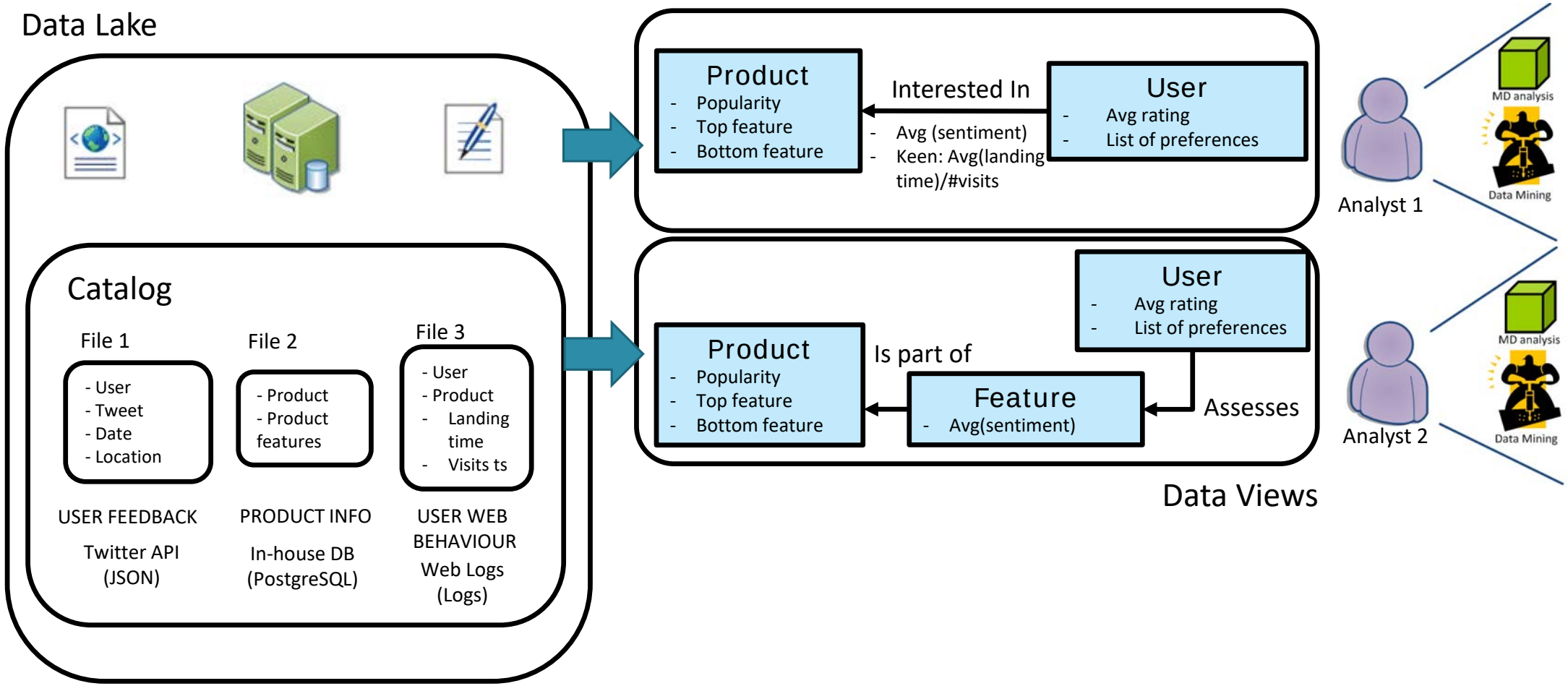
USER WEB BEHAVIOUR

Web Logs
(Logs)

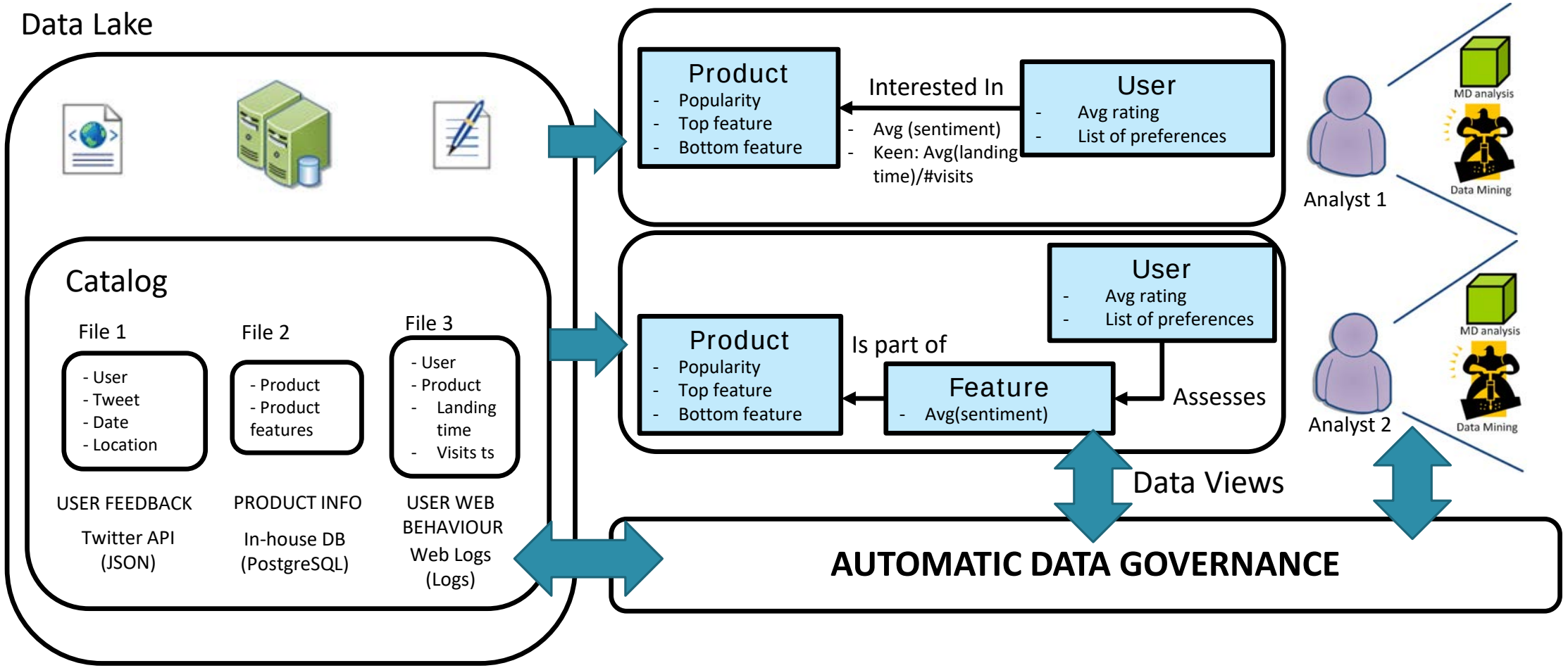
Drawbacks


Data Lake
Data Swamp
(Automated)
Complex
Transformations
Product
- Popularity
- Top feature
- Bottom feature
Interested In
- Avg (sentiment)
- Keen: Avg(landing time)/#visits
User
- Avg rating
- List of preferences
Product
- Popularity
- Top feature
- Bottom feature
Is part of
Feature
- Avg(sentiment)
Assesses
User
- Avg rating
- List of preferences
Analyst 1
Analyst 2
MD analysis
Data Mining
MD analysis
Data Mining
Data Views
USER FEEDBACK
PRODUCT INFO
USER WEB BEHAVIOUR
Twitter API (JSON)
In-house DB (PostgreSQL)
Web Logs (Logs)

From Data Swarms to Semantic Data Lakes



From IT-Centered to User-Centered



Thanks! Any Question?

OROMERO@ESSI.UPC.EDU

HOMEPAGE: HTTP://WWW.ESSI.UPC.EDU/DTIM/PEOPLE/OROMERO

TWITTER: @ROMERO_M_OSCAR