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”

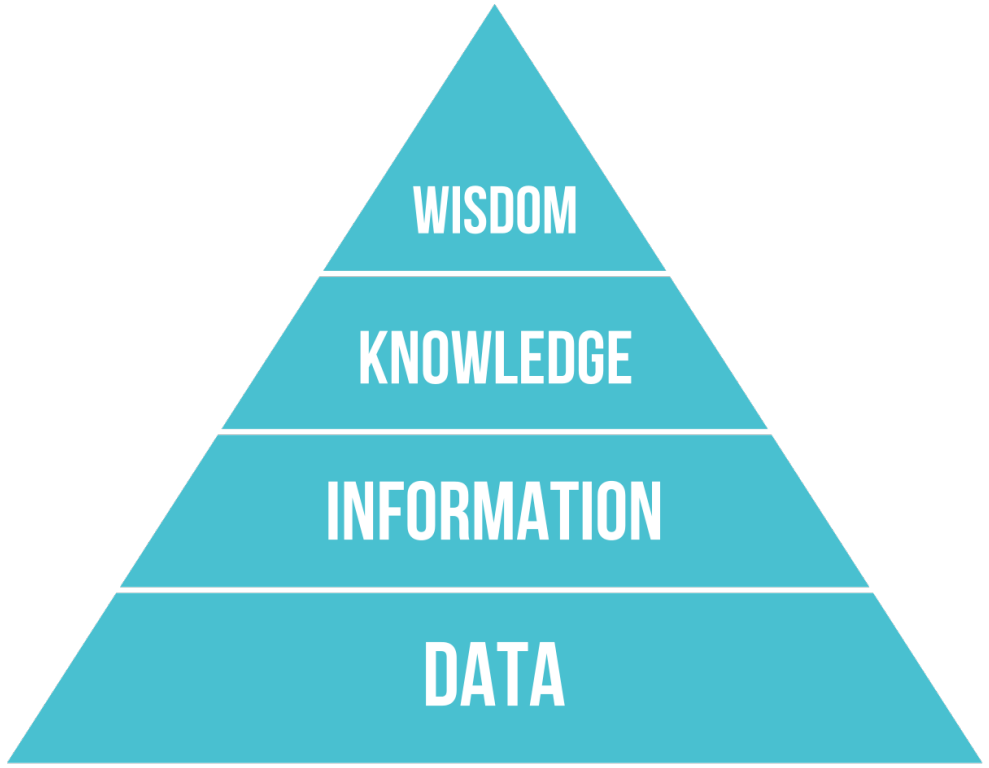
W. Edwards Deming, American Statistician

New Business Model: Instagram's Fable

(xkcd.com)

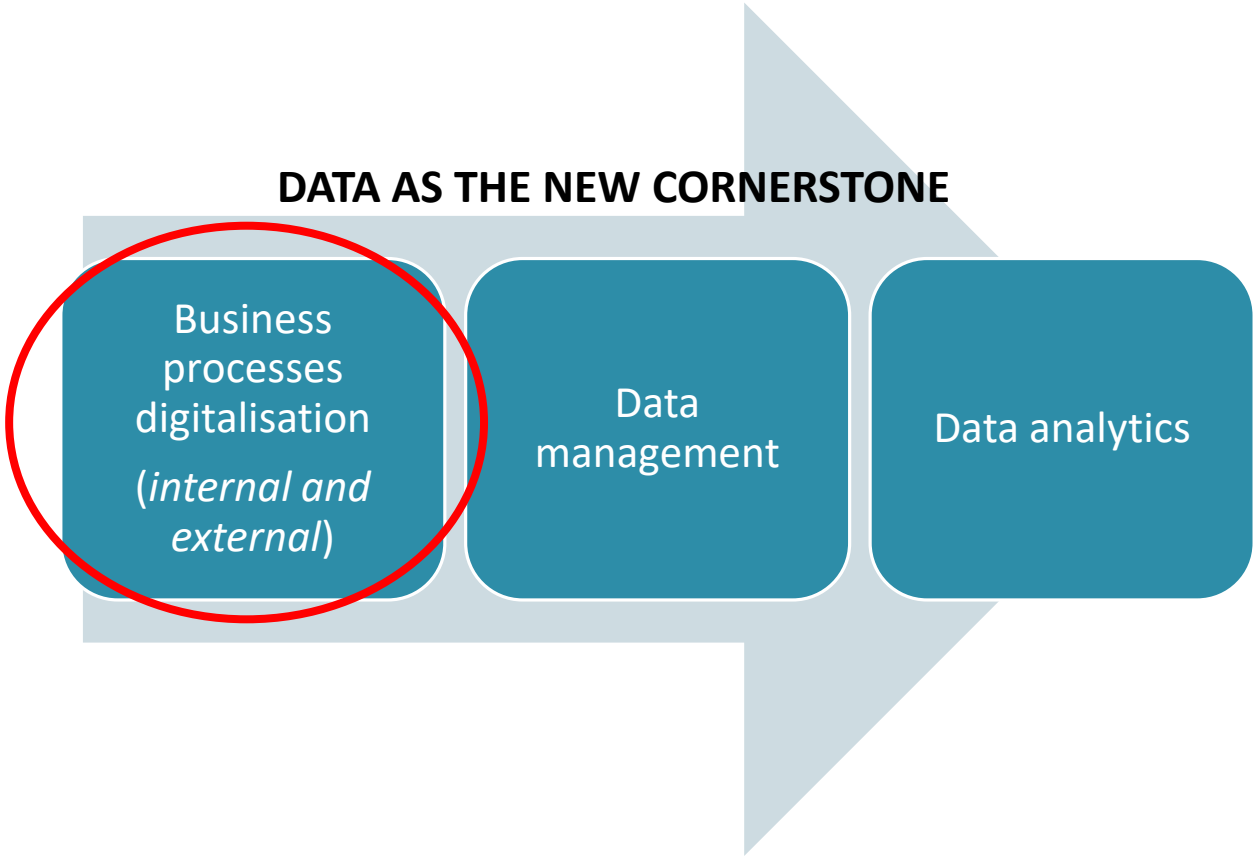
Decision Support Systems

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



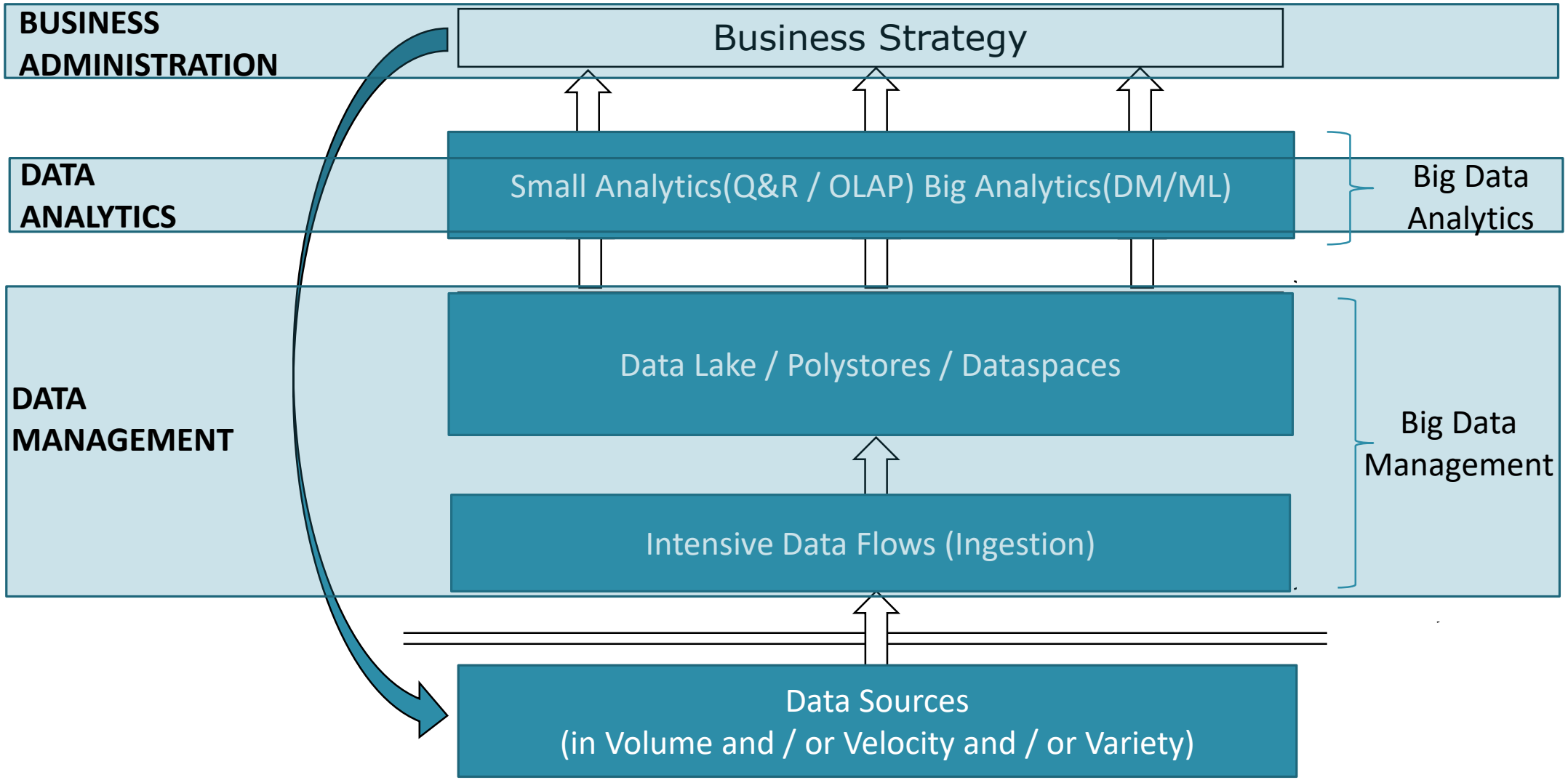
DIKW pyramid

Decision Support Systems



The IBM Quant Crunch report, 2017 https://www.ibm.com/downloads/cas/3RL3VXGA

The Data Value Creation Chain

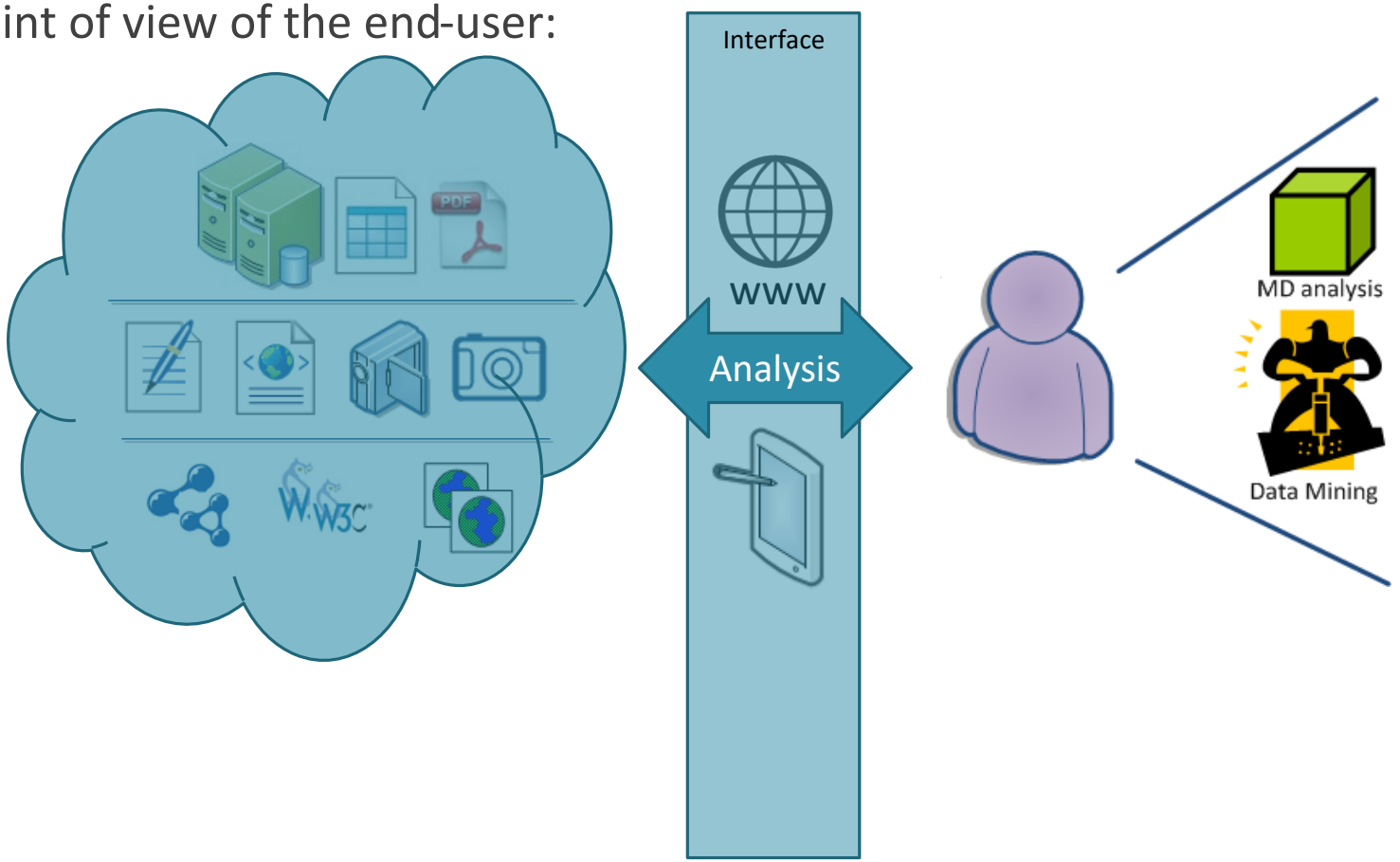


Challenges

FROM THE END-USER POINT OF VIEW

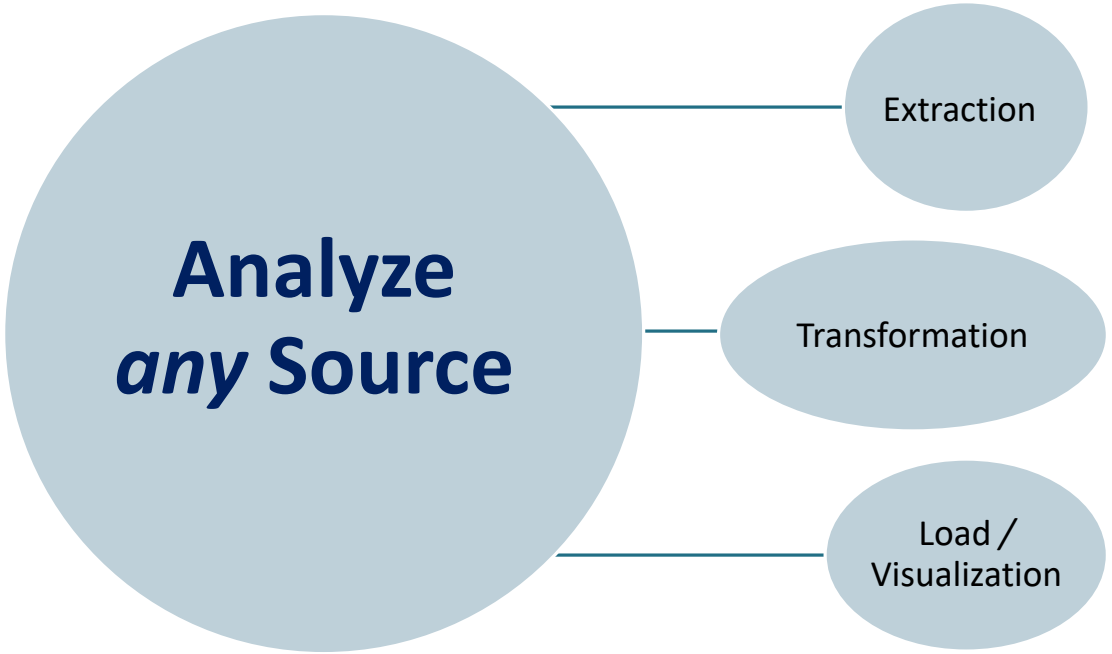
Data Analysis Democratisation

From the point of view of the end-user:



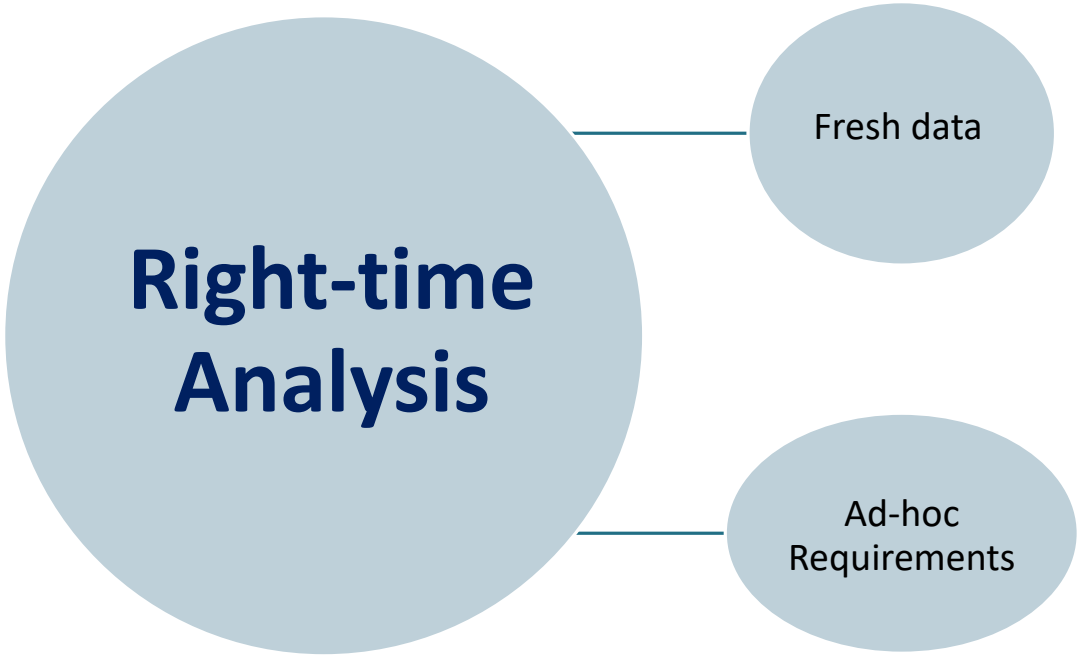
Data Analysis Democratisation

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?*
- *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)*

Linking Data, Analysis and Business

This is why AI has yet to reshape most businesses

For many companies, deploying AI is slower and more expensive than it might seem.

by **Brian Bergstein**

February 13, 2019

link: https://www.technologyreview.com/2019/02/13/137047/this-is-why-ai-has-yet-to-reshape-most-businesses/

Challenges in Data Management

FROM THE IT POINT OF VIEW

What is Big Data?

VOLUME

Veracity

Velocity

Value

vArIaBiLiTy

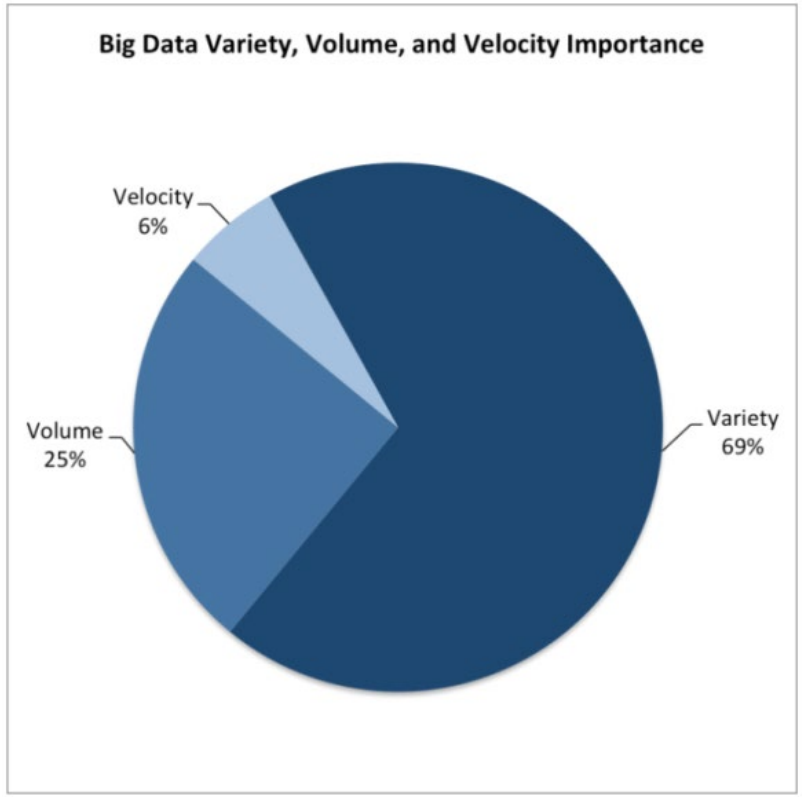
Variety

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



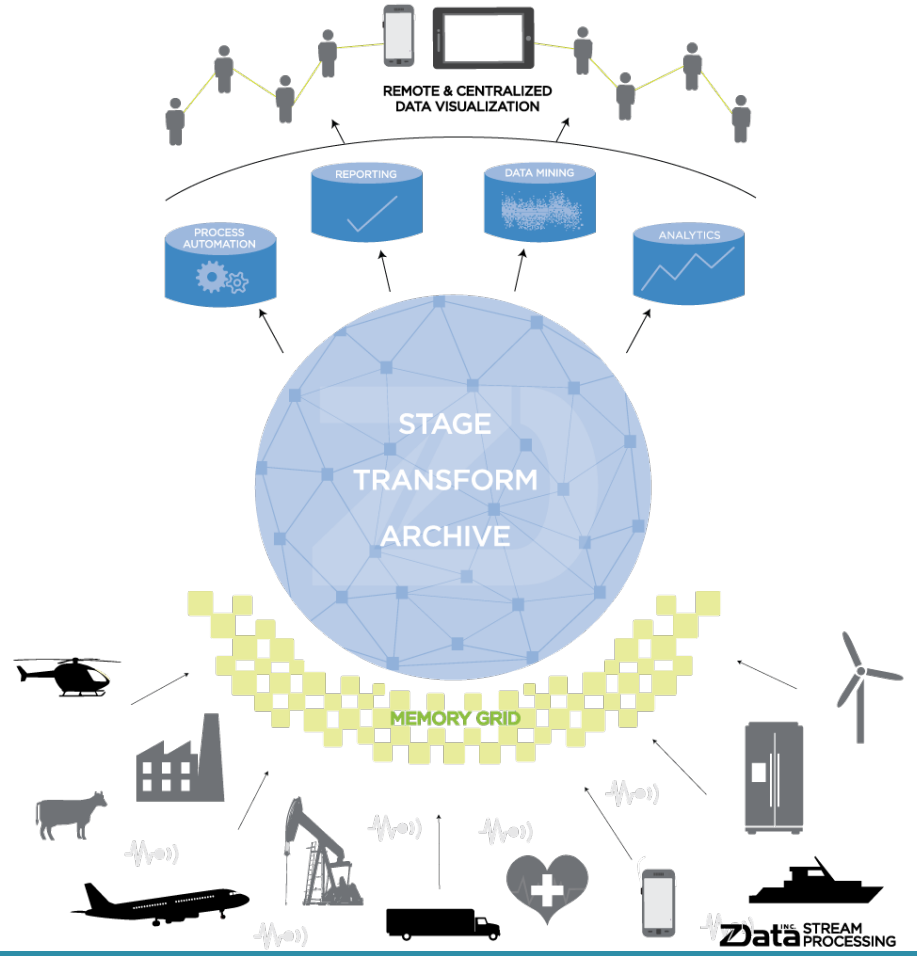
MIT Sloan Management Review (2016): http://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/

The Data Lake

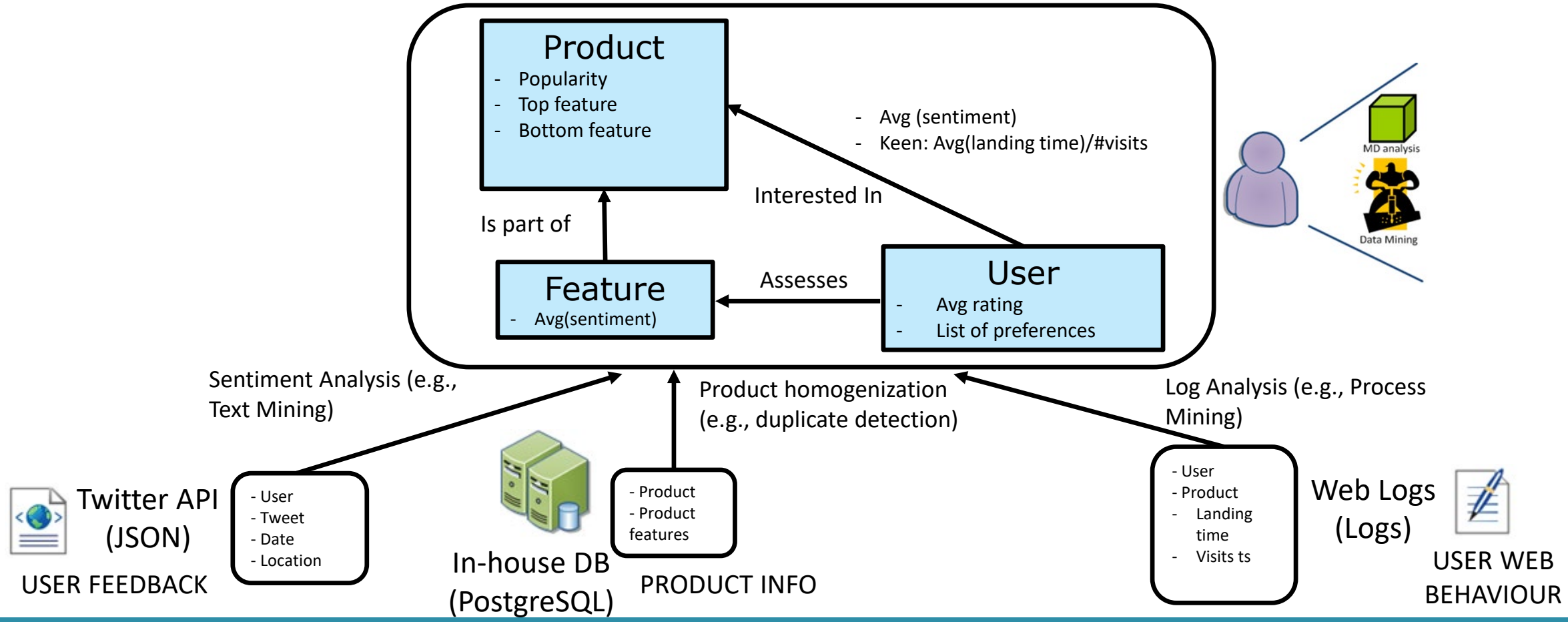
IDEA: Load-first, Model-Later

Modeling at load time restricts the potential analysis that can be done later (Big Analytics)

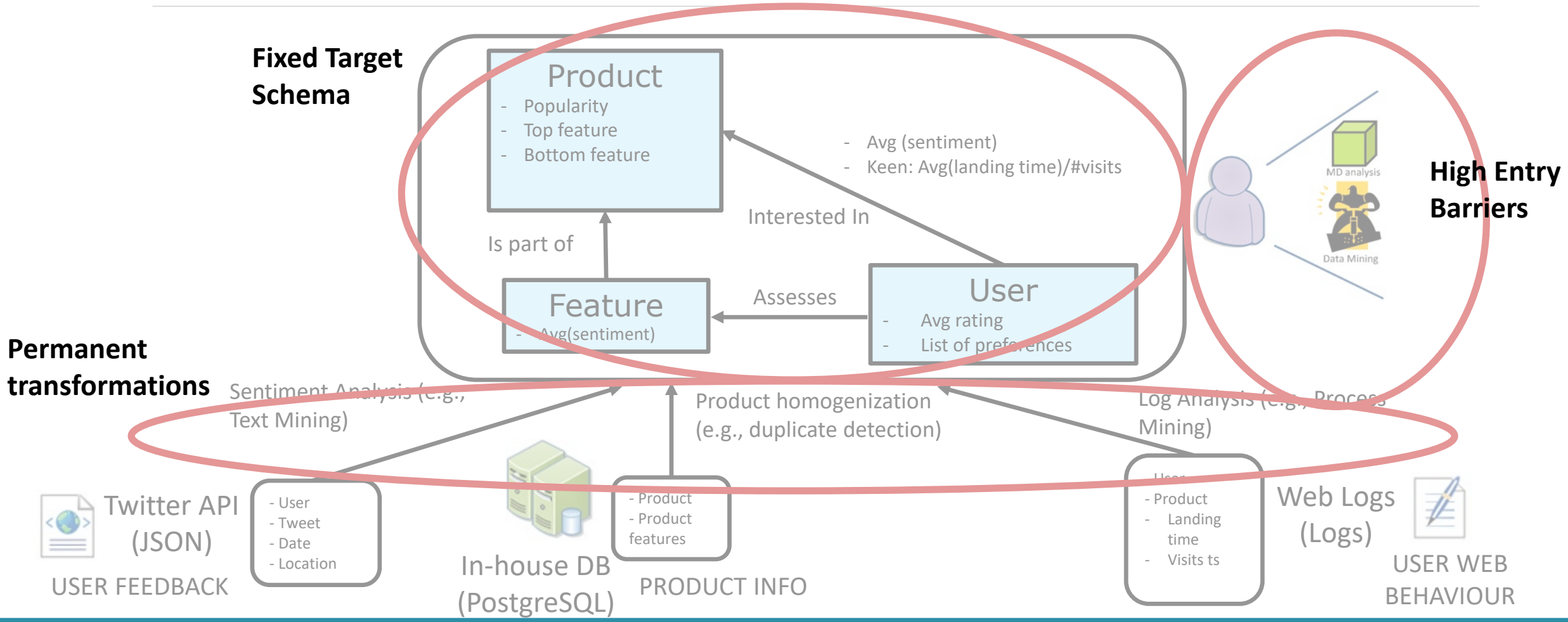
Store raw data and create on-demand views to handle with precise analysis needs



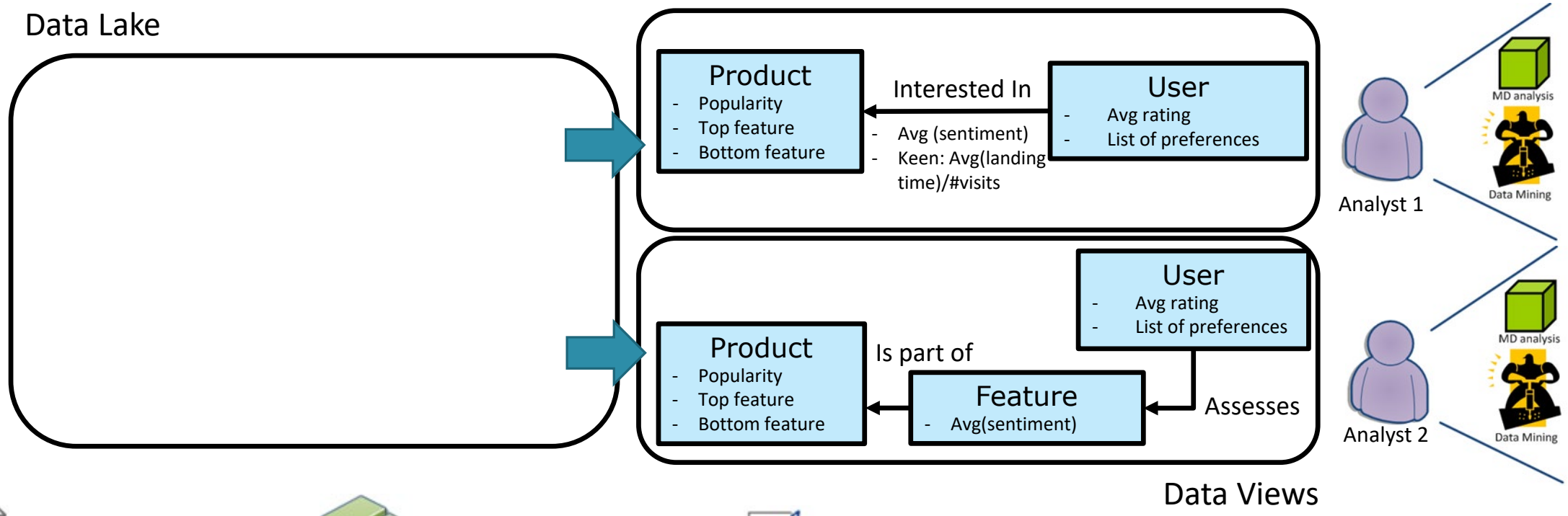
Model-First (Load-Later)



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 FEEDBACK
- User
- Tweet
- Date
- Location
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)

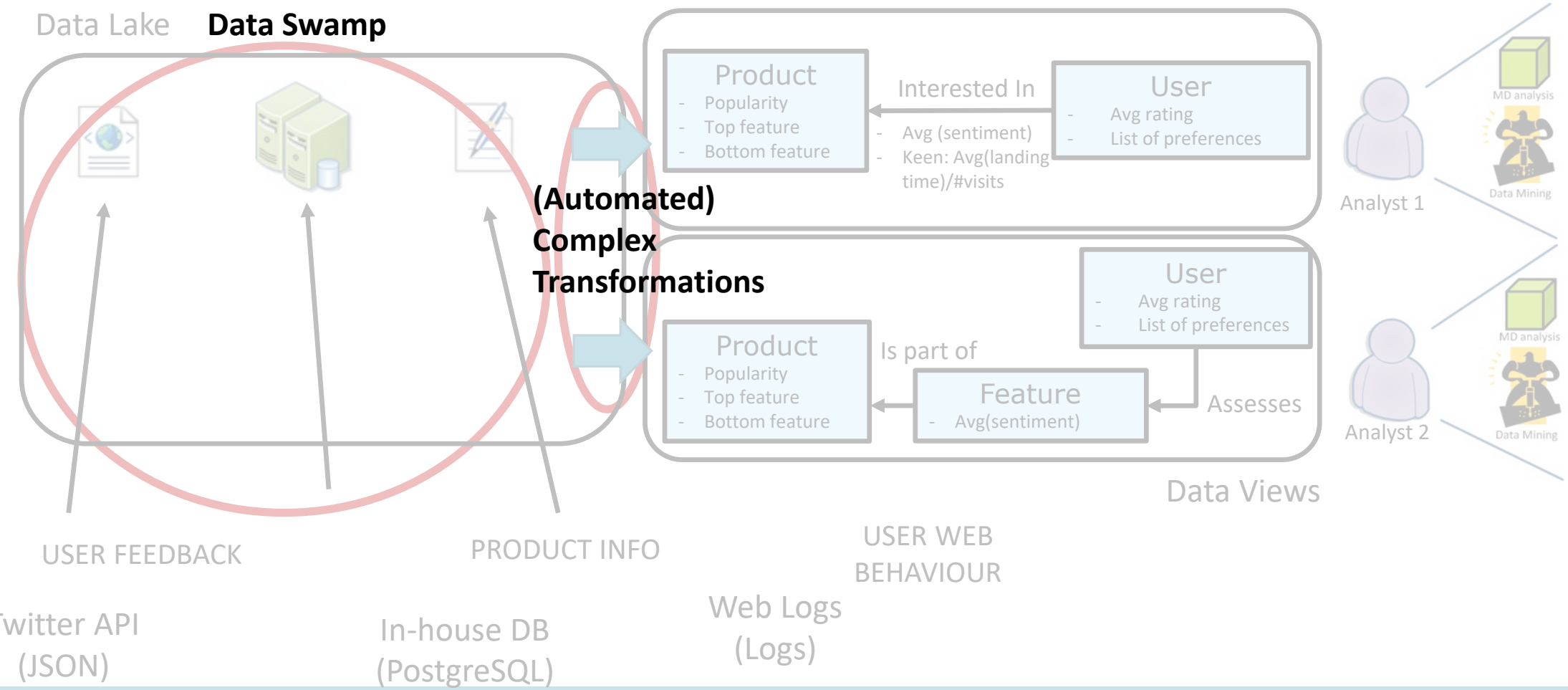
PRODUCT INFO

In-house DB
(PostgreSQL)

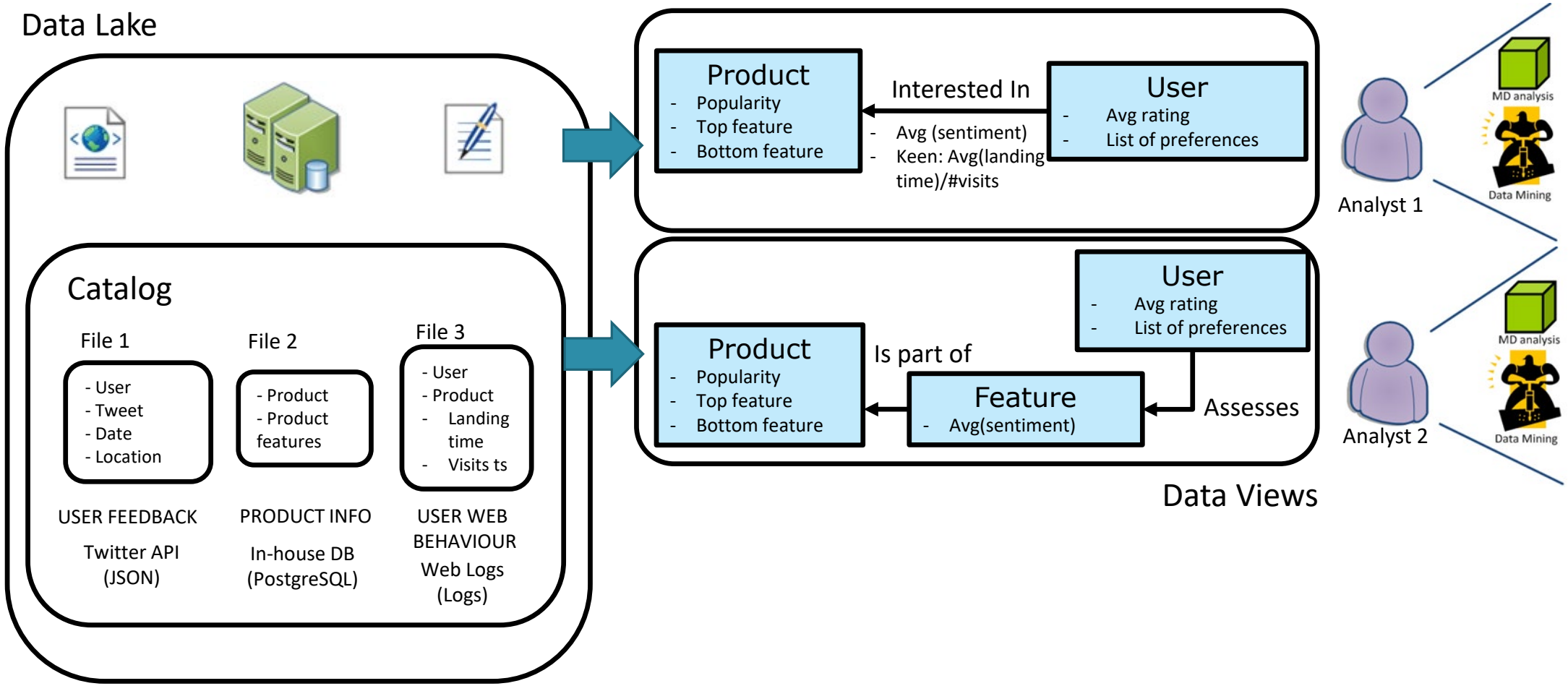
USER WEB BEHAVIOUR

Web Logs
(Logs)

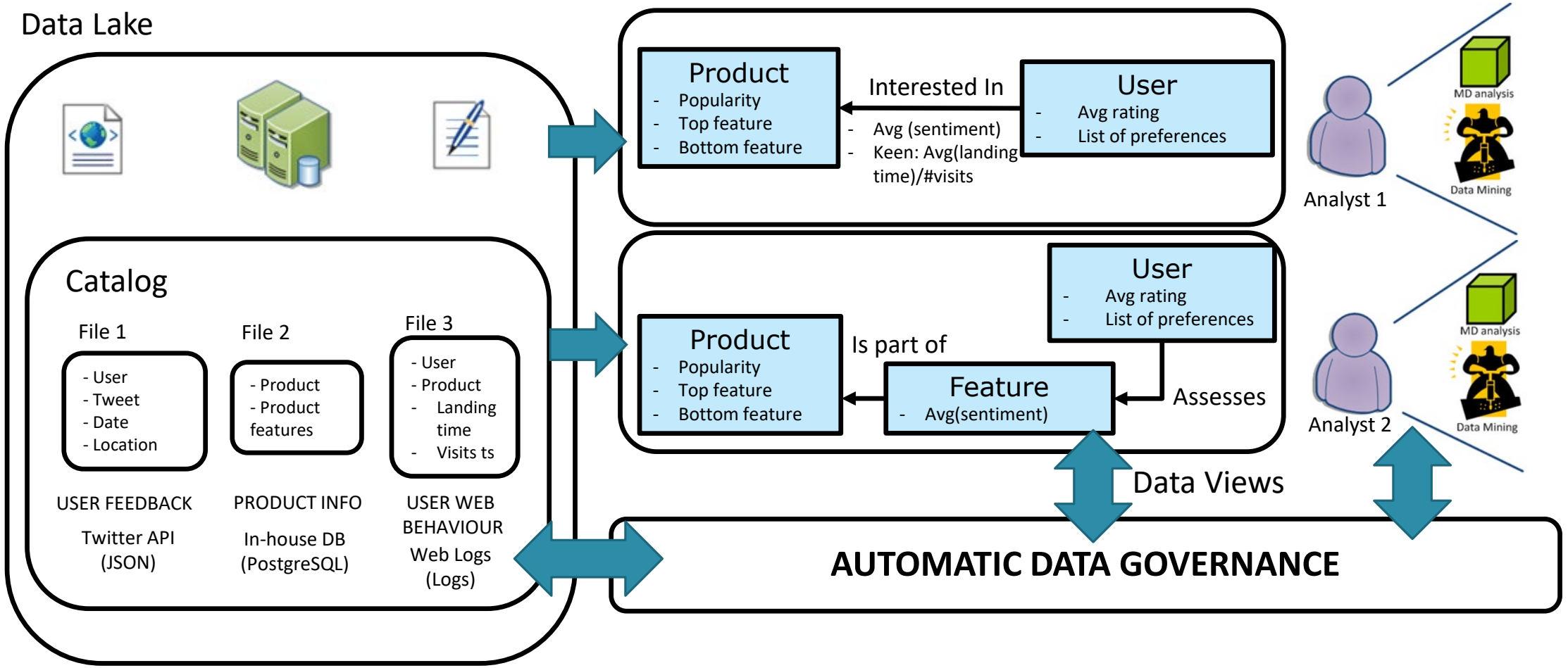
Drawbacks


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
Data Views
Analyst 1
Analyst 2
MD analysis
Data Mining
MD analysis
Data Mining
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



Challenges in Data Analytics

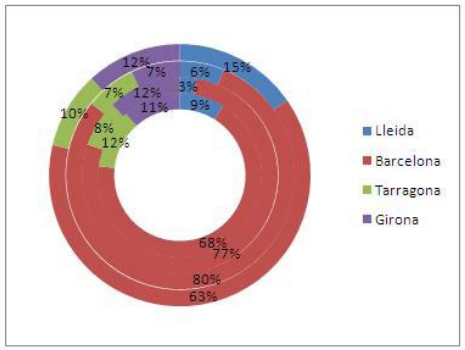
FROM THE IT POINT OF VIEW

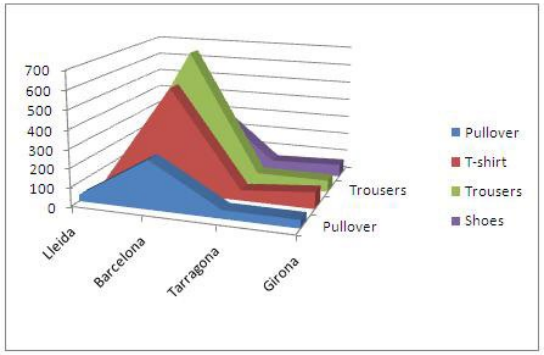
Analytics

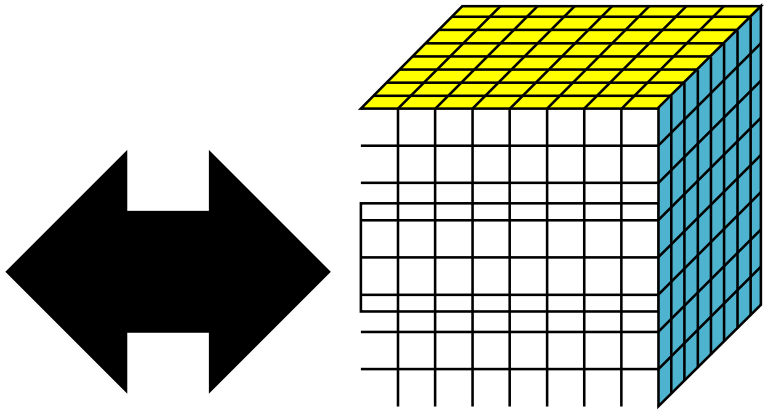
Typically, the analysis of the data has been considered at three different levels of detail

- 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



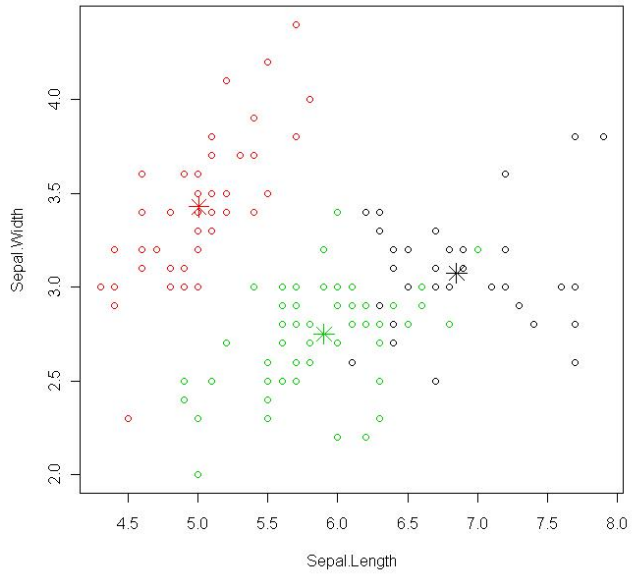




Analytics

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
 [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 1 3 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"
```

Current Challenges

Data analytics rely on well-defined data schemata

- Descriptive data analytics rely on the **underlying multidimensional schema** to automate OLAP operations
- Probabilistic approaches require fixed input data structures (e.g., matrixes) but the problem is how to **create such input data structure**
  - Variables are typically extracted from large data repositories
  - It may require complex data transformations (cleansing and homogenization)
  - Data analysts may require a new variable at any moment

The current challenge in Data Analytics boils down to two aspects:

- How to incorporate on-demand new variables that contextualize the analysis and bring new evidences from where to discover patterns
- Develop new analytical frameworks that facilitate and democratize the access to the data deluge

A strict **data governance** policy is required in order to automate data analysis (in whatever form)

Thanks! Any Question?

OROMERO@ESSI.UPC.EDU

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

TWITTER: @ROMERO_M_OSCAR