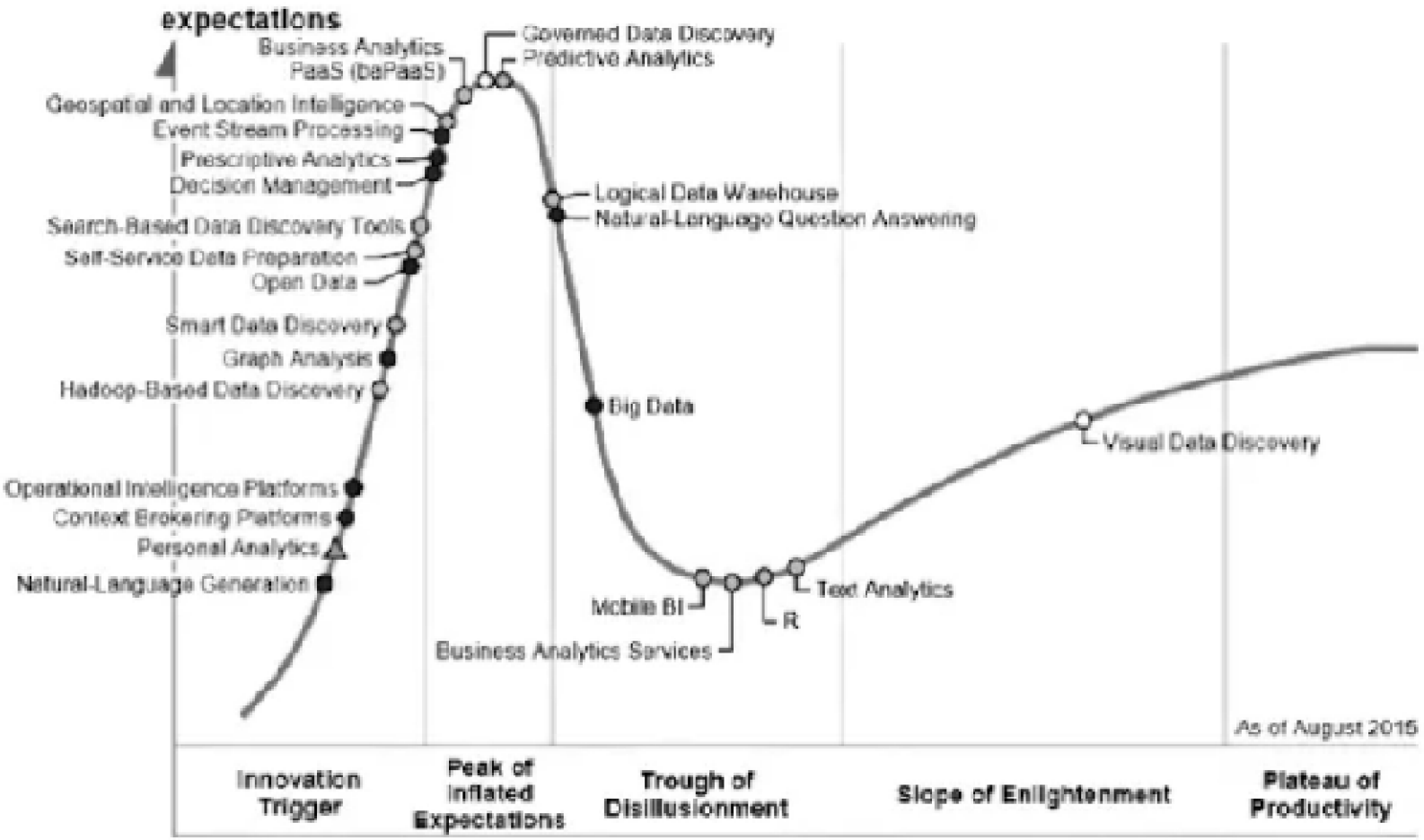
Sisteme Distribuite

Cursul 12

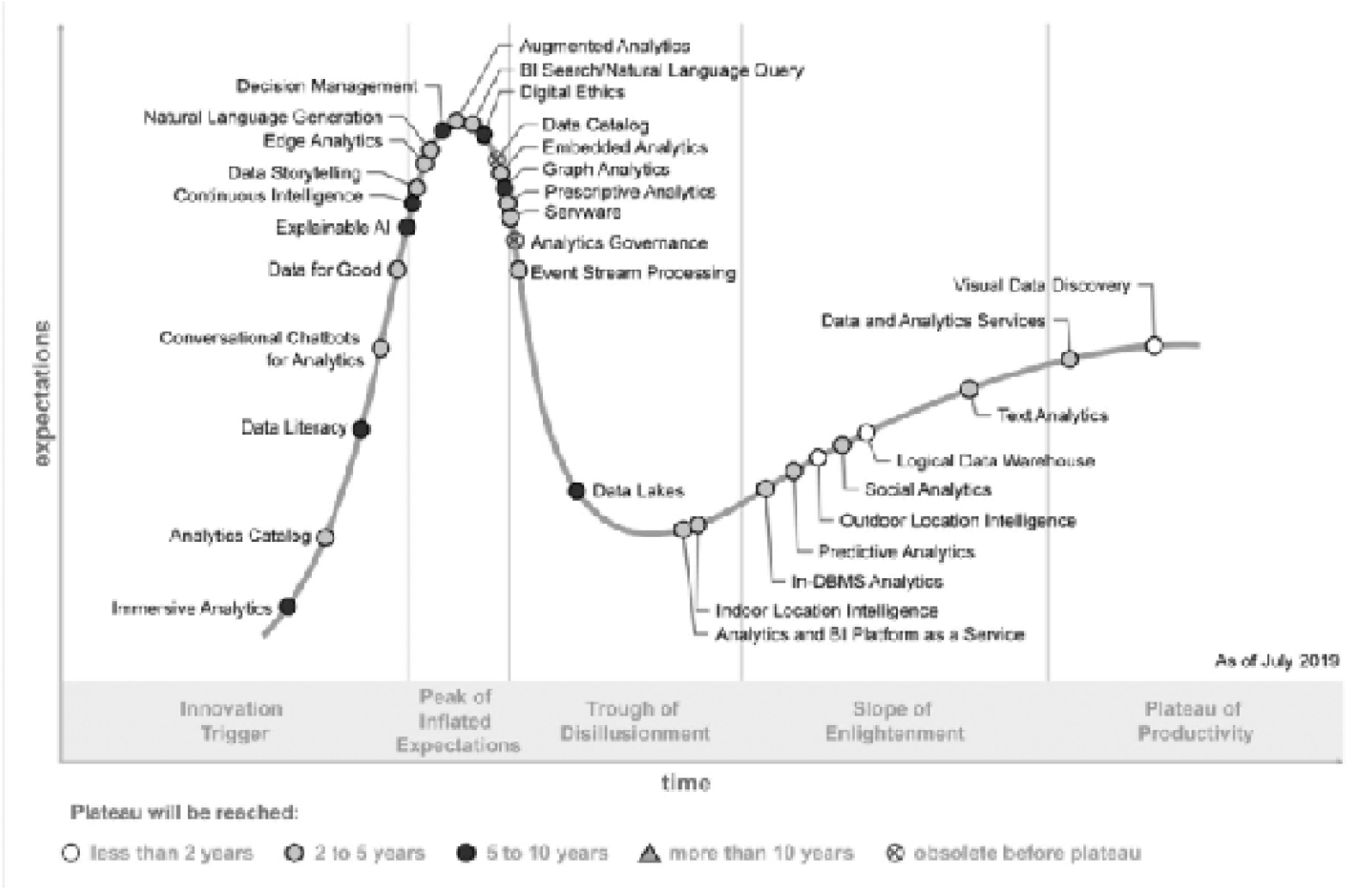
Mihai Zaharia

De ce?



Gartner - hyper cicle for bussiness inteligence and analitics 2015

De ce?



Gartner - hyper cicle for bussiness inteligence and analitics 2019

Bussiness Intelligence

- X

Bussiness Analitics

user≡Corporate Manager

Datele în business analitics&intelligence

- x

Data governance -> bigdata

- x

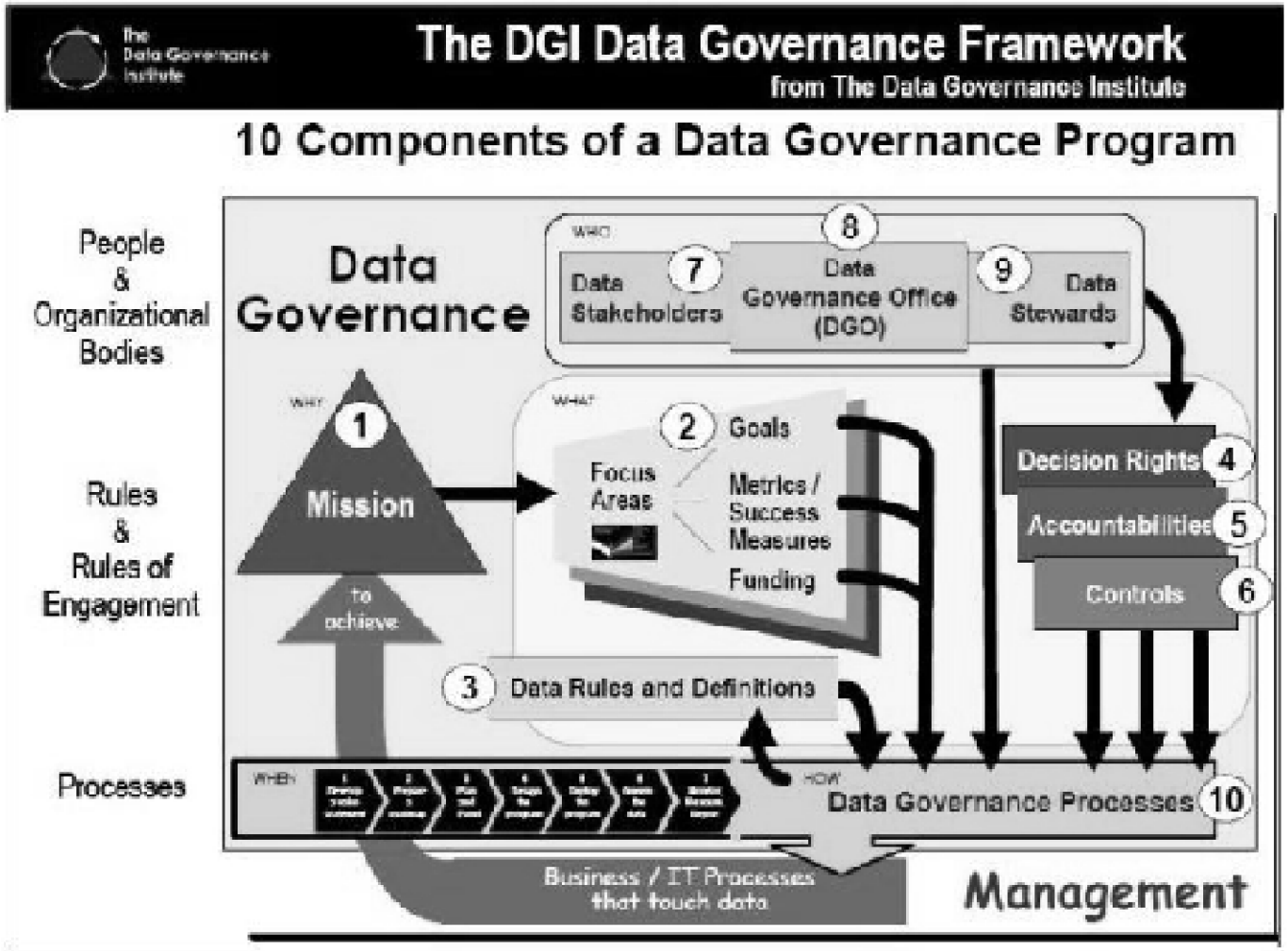
Guvernarea datelor - istoric

- X

Guvernarea datelor

- x

Guvernarea datelor



- X

Importanța guvernării datelor

- x

Scopurile guvernării datelor

- x

GDPR vs guvernarea datelor

- x

Responsabilii pentru guvernarea datelor

cauza morții CDO-ului!

Responsabilii pentru guvernarea datelor

Managerul și echipa pentru guvernarea datelor

Responsabilii pentru guvernarea datelor

Comitetul

Responsabilii pentru guvernarea datelor

Data stewards

Guvernarea big data

depinde de nivelul din care este analizată

Calitatea datelor

cum se face?!

Gestiunea metadatelor

- x

Chiar este înțeleasă guvernarea datelor?

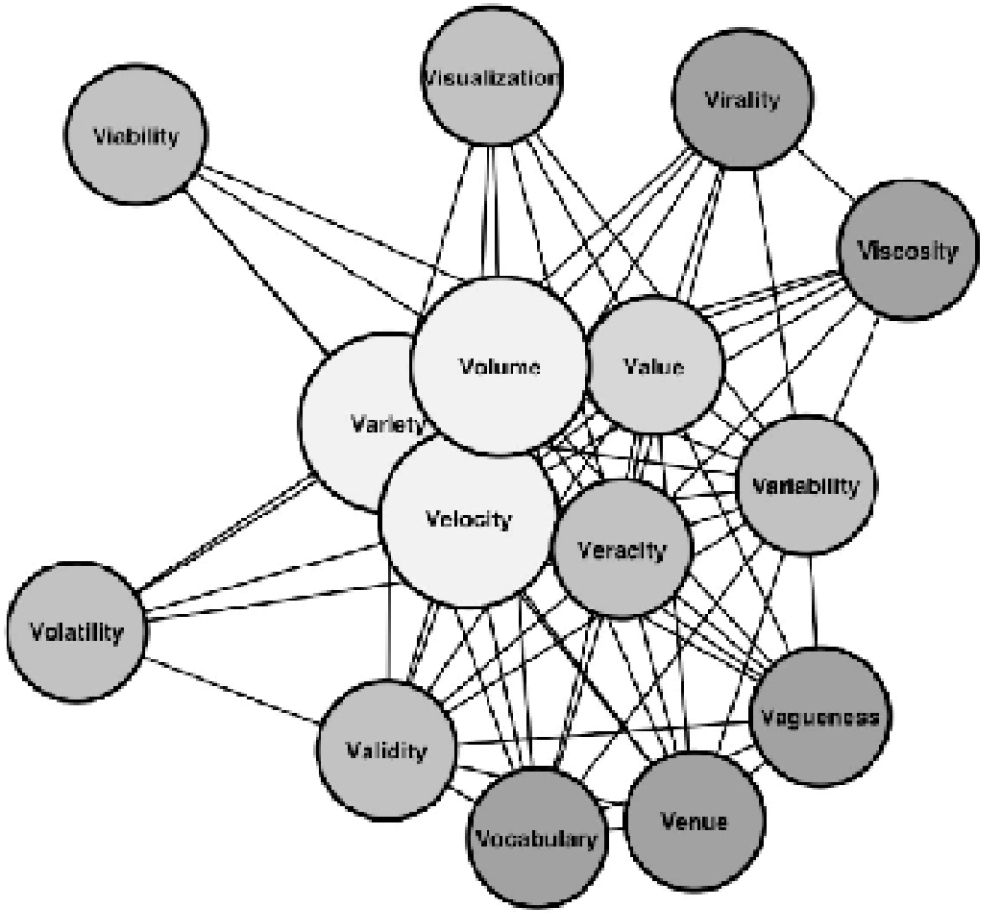
Oh my God It's full of stars!

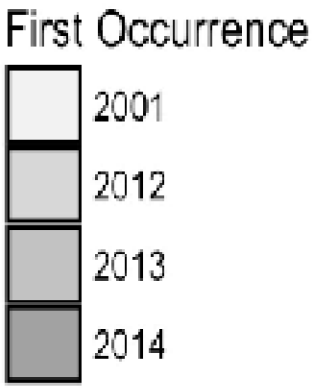
Big Data

- *"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources."*

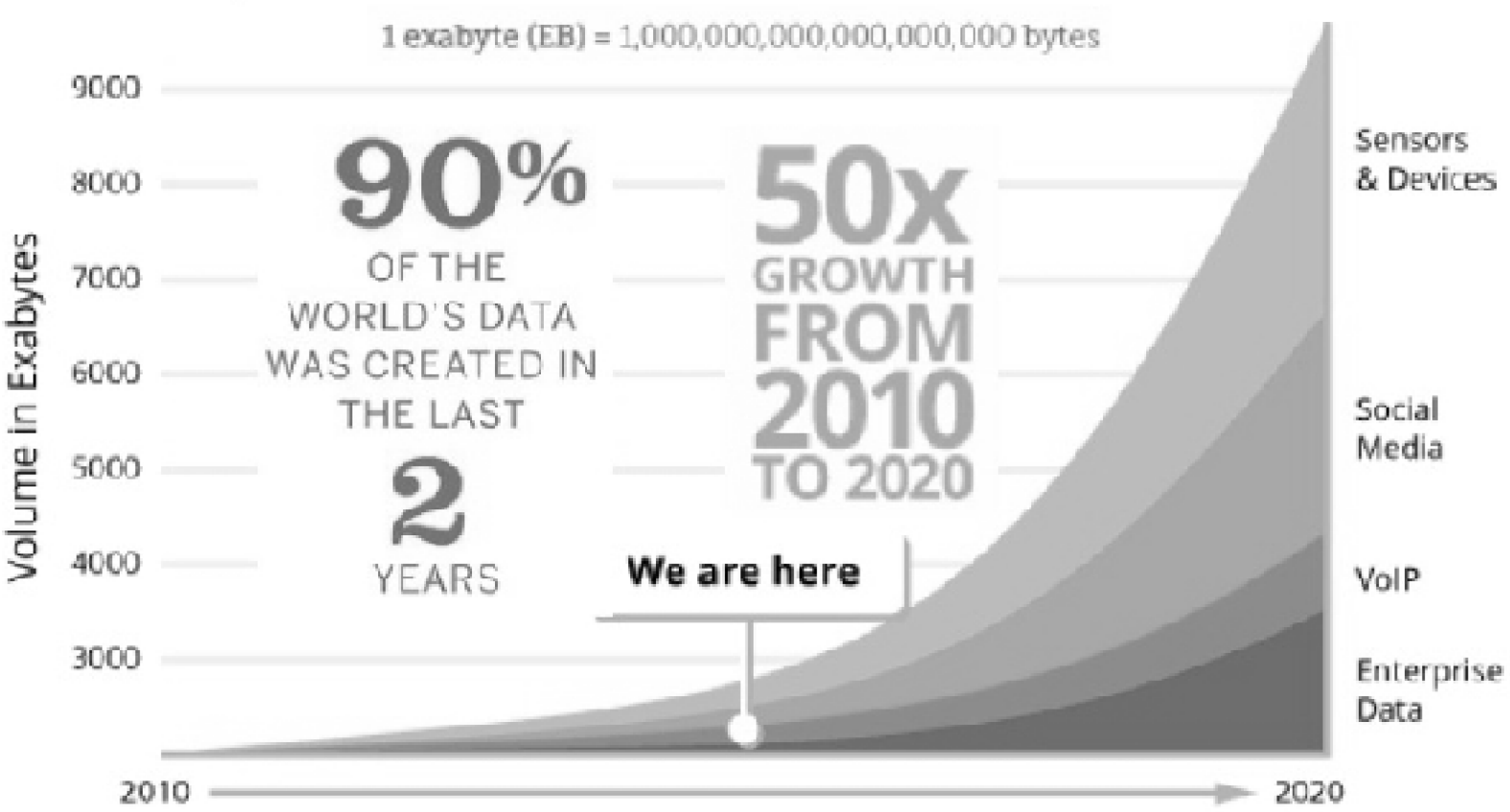
Cox si Ellsworth, IEEE, 1997

Dimensiunile blg data



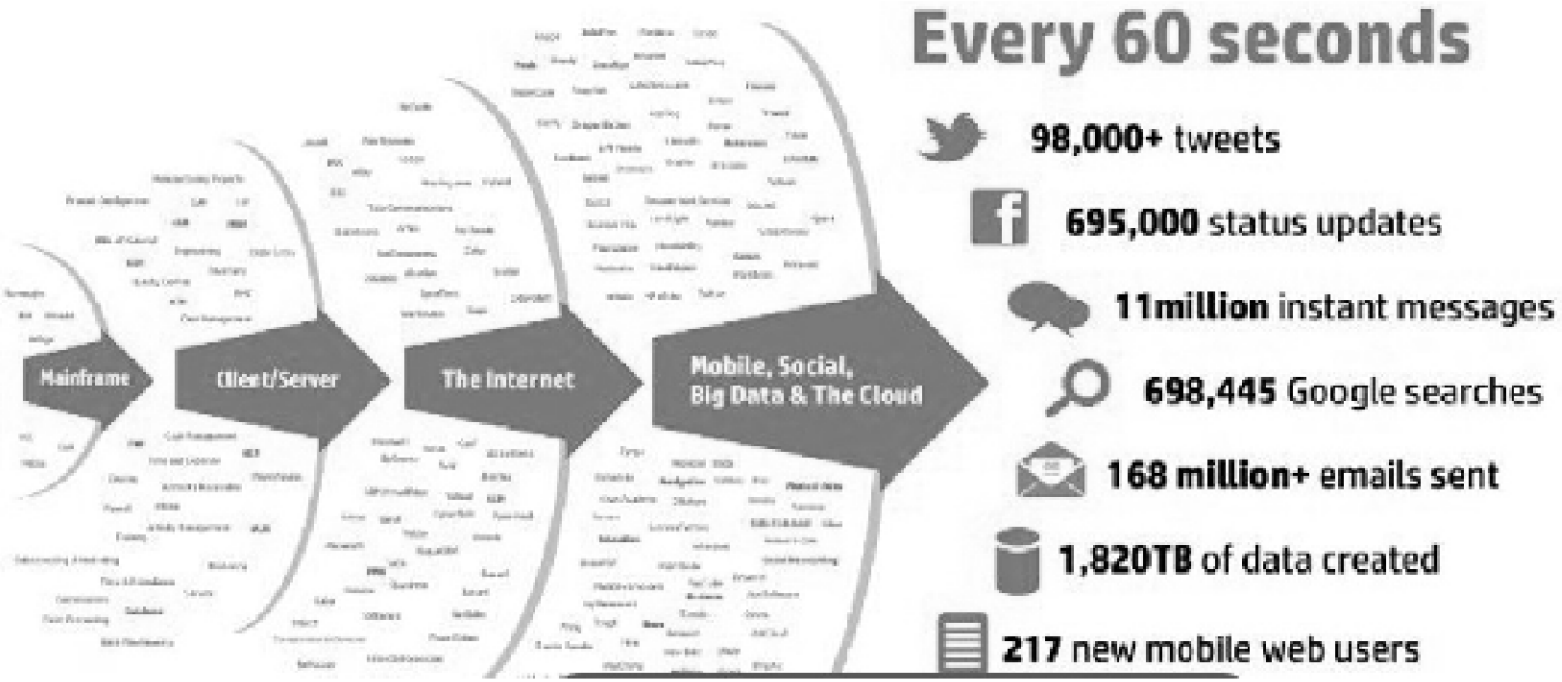


Dimensiunile Big Data 1. Volumul



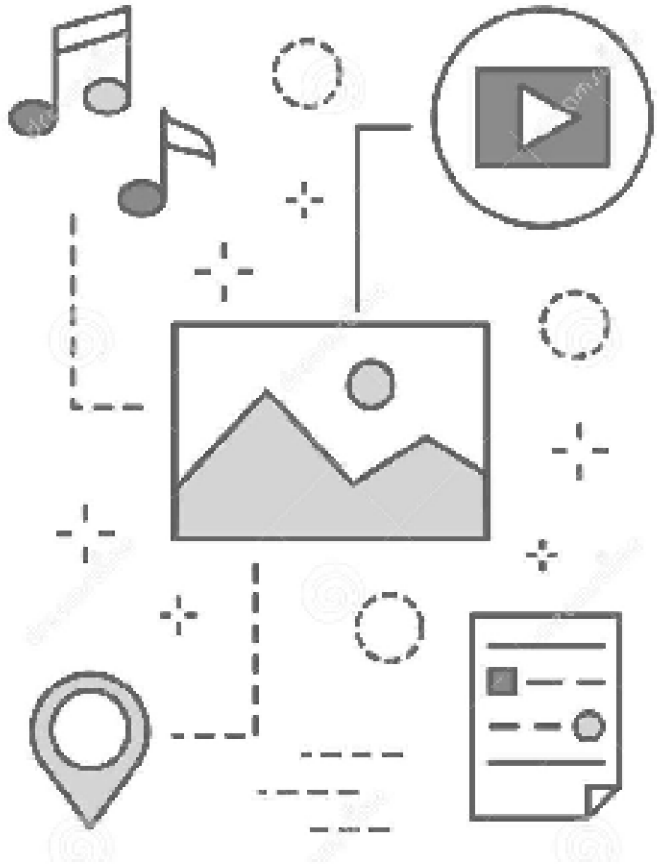
creștere exponențială a datelor generate

Dimensiunile Big Data 2. Velocitate



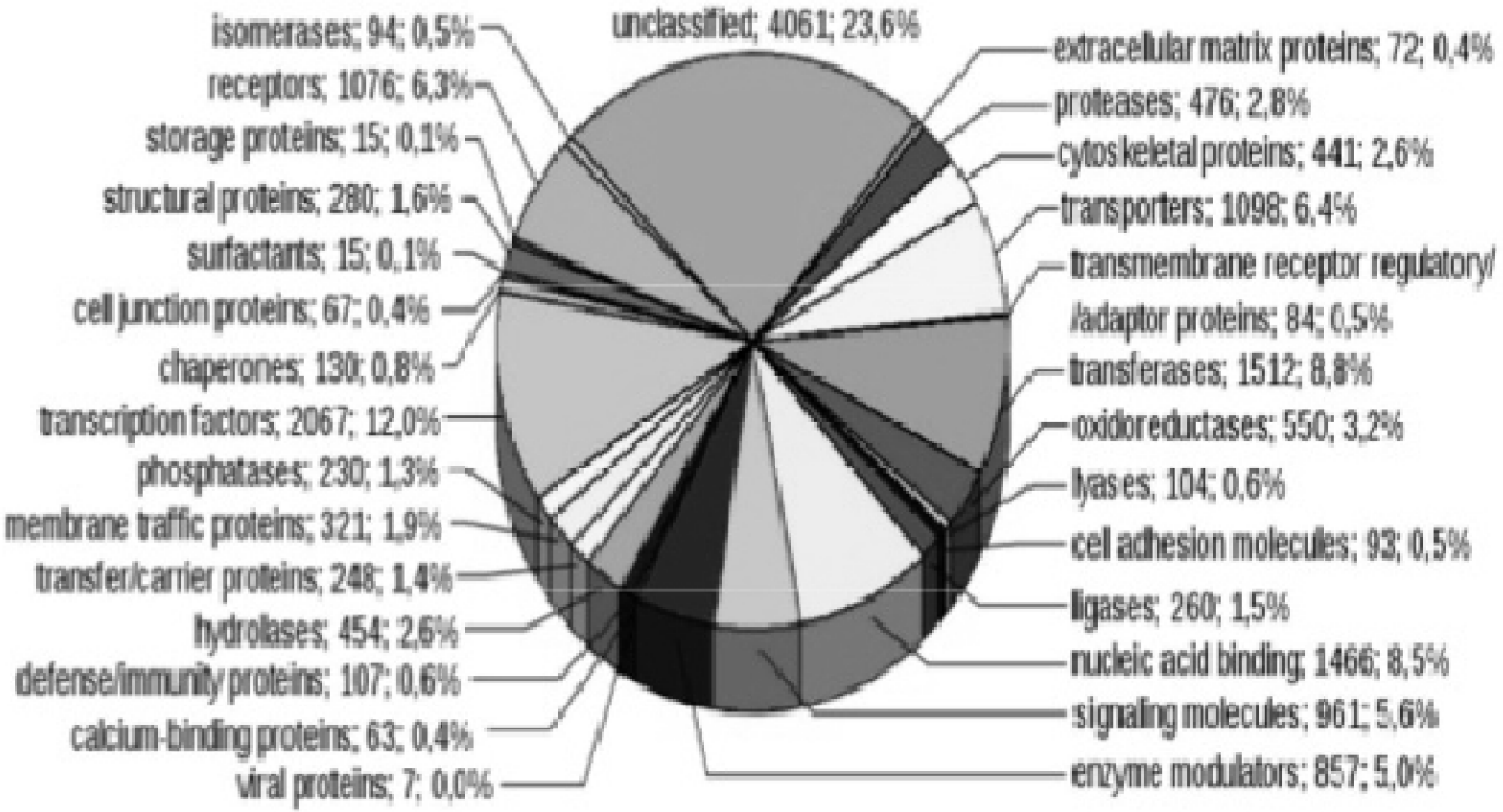
viteza de producție a noilor date

Dimensiunile Big Data 3. Varietate



- Datele structurate
- Datele semistructurate
- Date nestructurate

Dimensiunile Big Data 4. Variabilitate

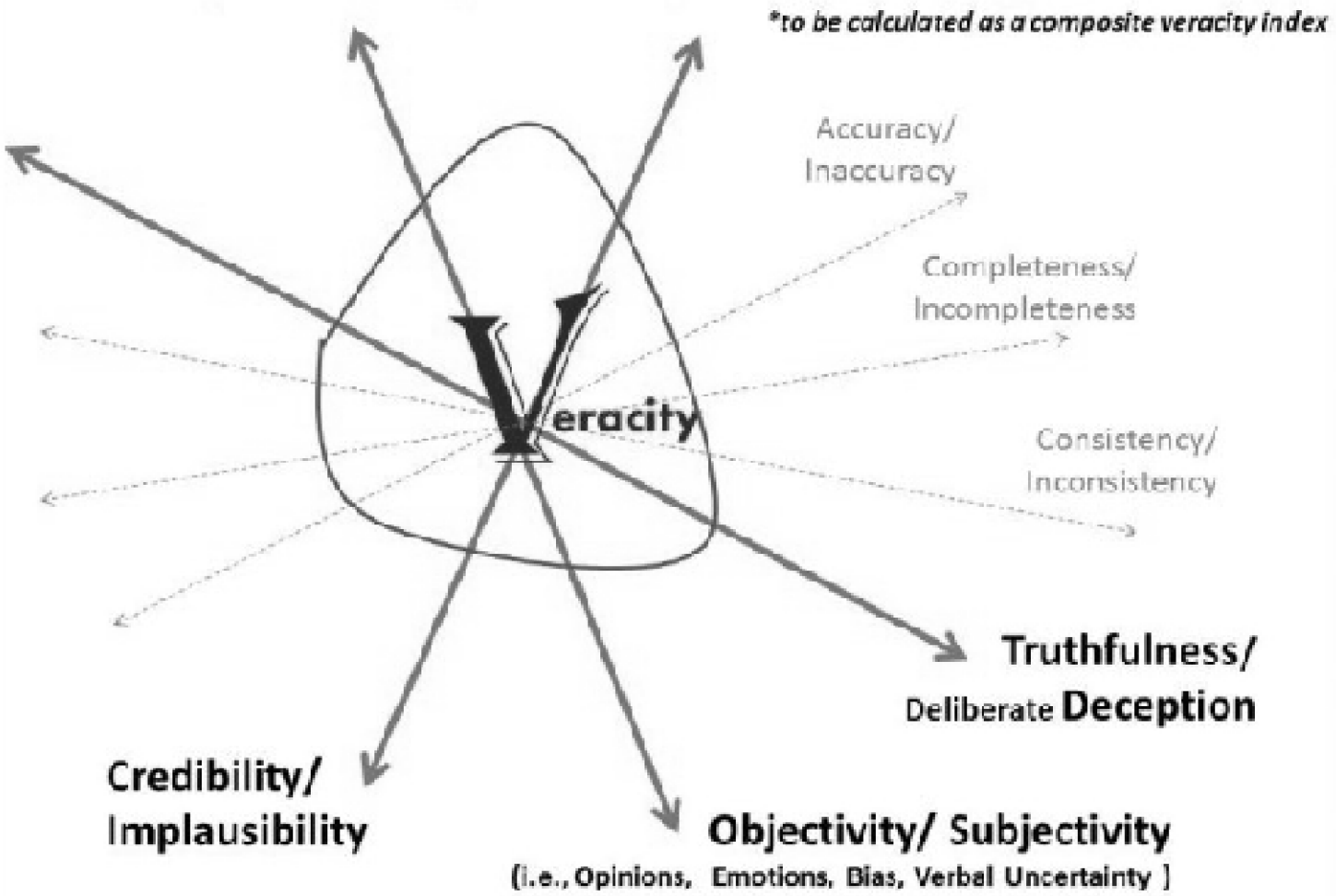


Functions of 17,209 Genes

Neomogenitatea sau

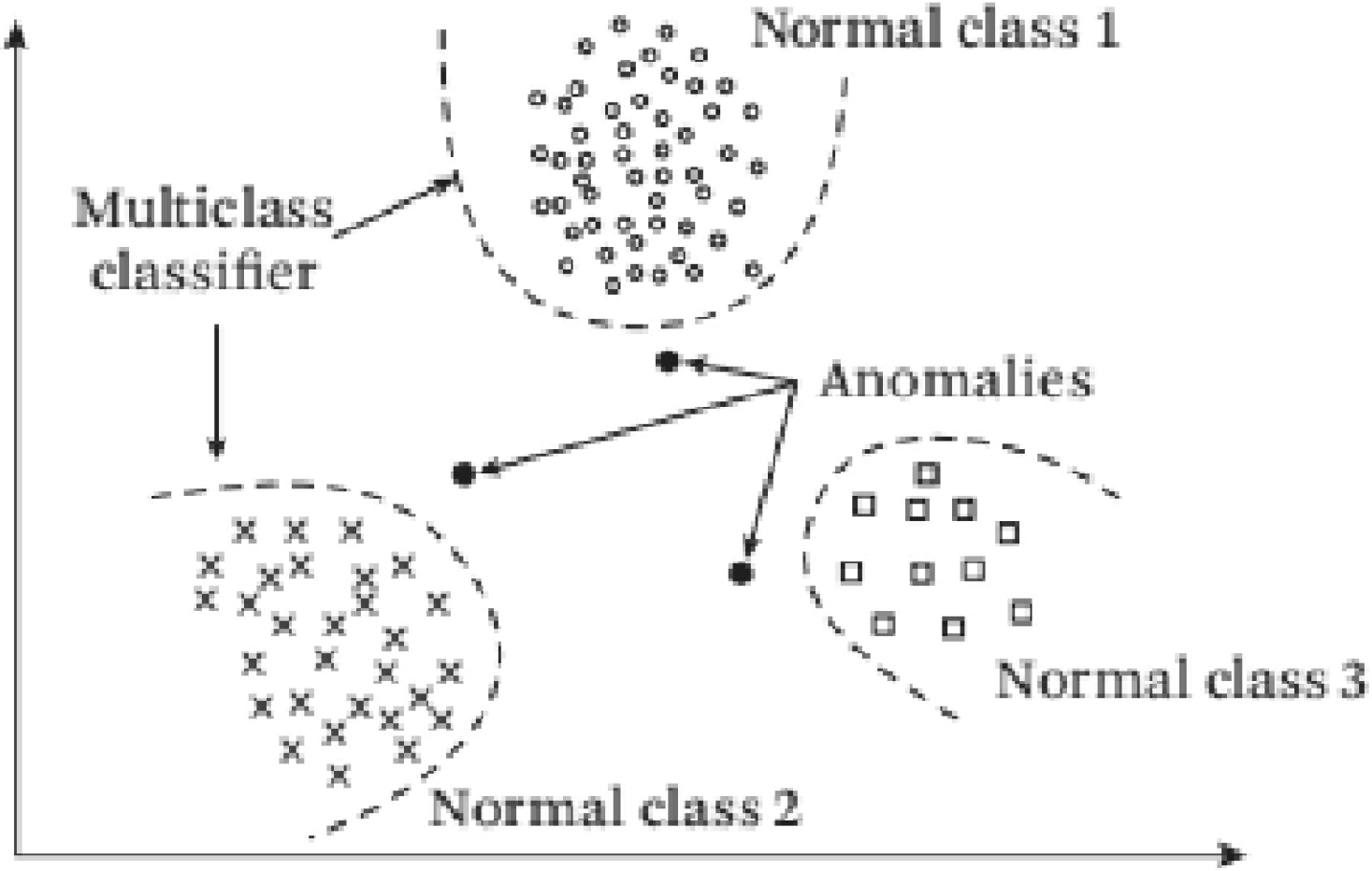
Variatia dimensională și compozițională

Dimensiunile Big Data 5. Veracitate



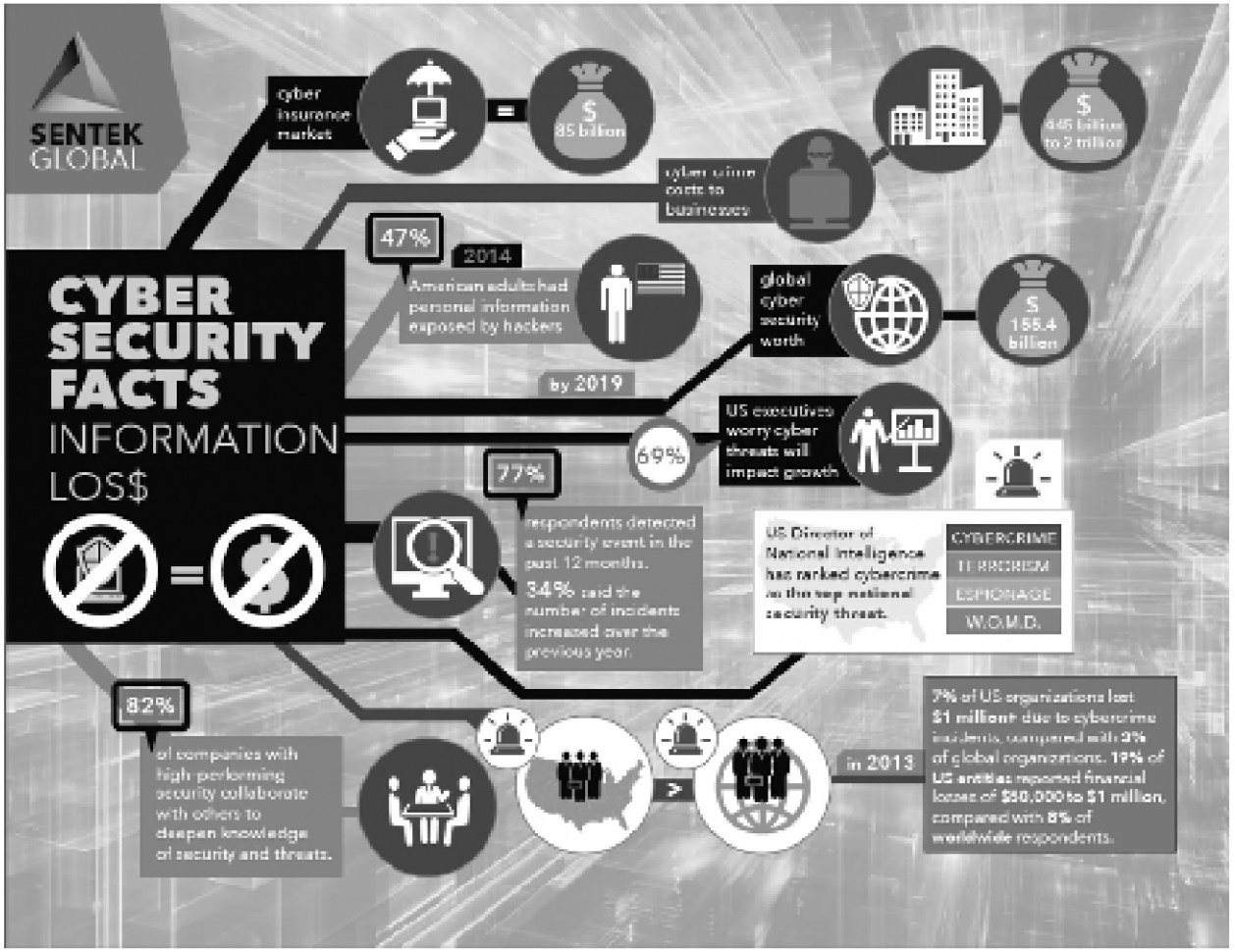
veridicitate sau corectitudine

Dimensiunile Big Data 6. Validitatea



Detectarea anomaliilor după gruparea pe categorii

Dimensiunile Big Data 7. Vulnerabilitate



după Forbes

Dimensiunile Big Data 8. Volatilitate

date noi vs. date vechi?

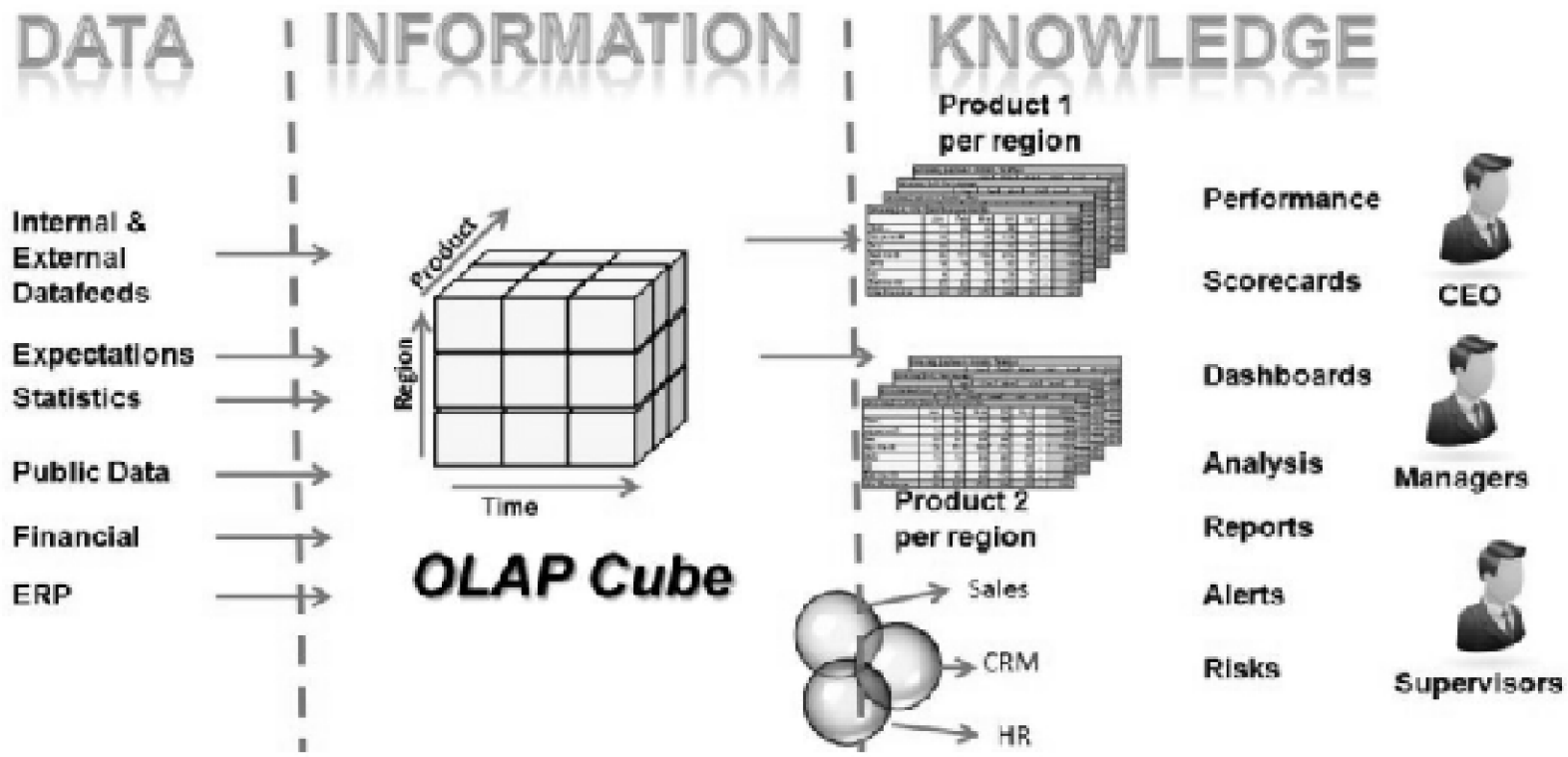
Dimensiunile Big Data 9. Vizualizarea

Googke Facets

Dimensiunile Big Data 10. Valoare

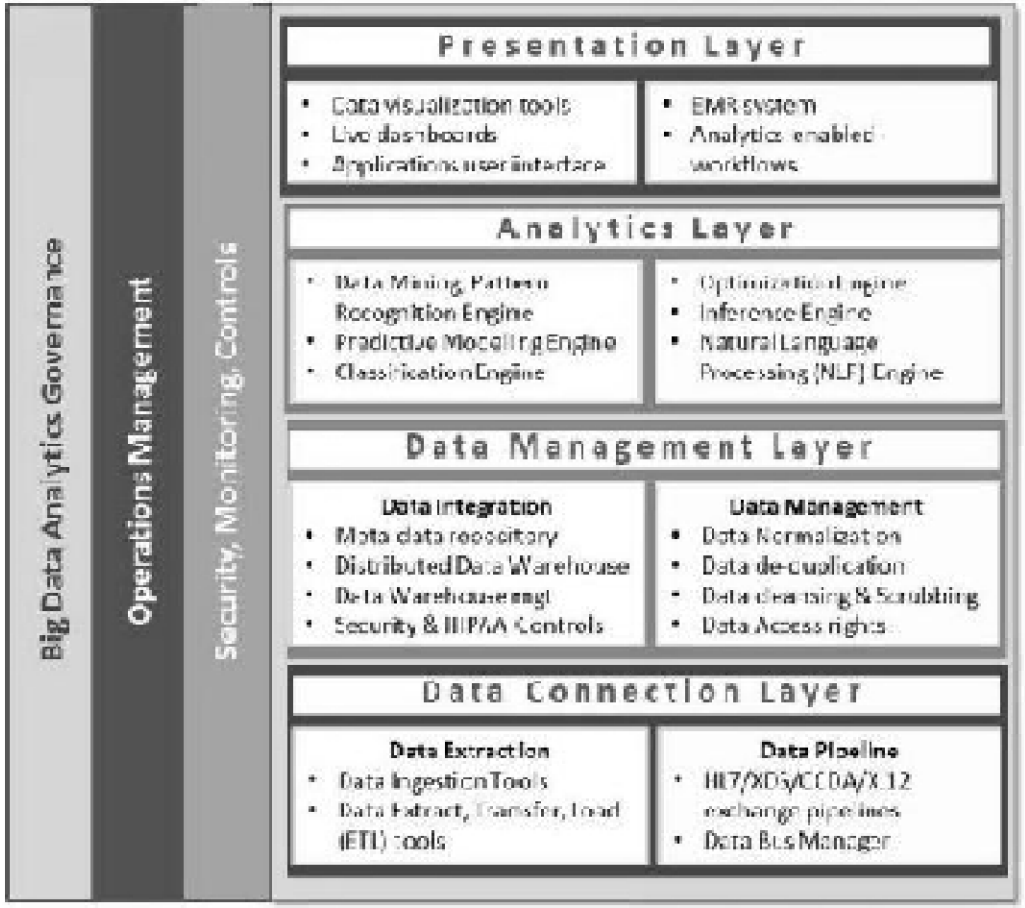
guvern, corporații, IMM, PFA

Cubul OLAP


DATA
INFORMATION
KNOWLEDGE
Internal & External Datafeeds
Expectations
Statistics
Public Data
Financial
ERP
Product
Region
Time
OLAP Cube
Product 1 per region
Product 2 per region
Sales
CRM
HR
Performance
Scorecards
Dashboards
Analysis
Reports
Alerts
Risks
CEO
Managers
Supervisors

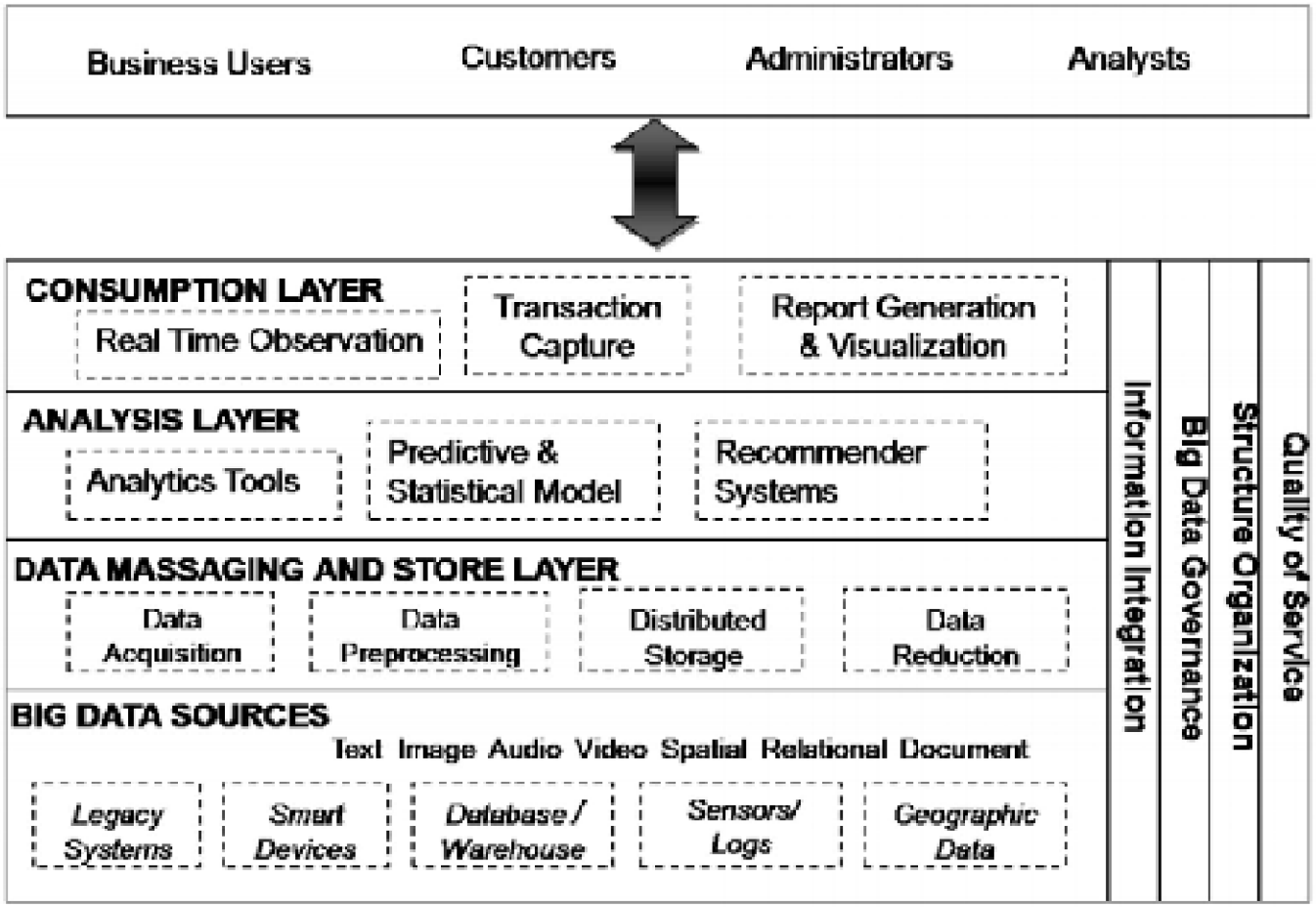
- d

Arhitectura unei platforme



pentru data analitics

Arhitectura unei platforme



- X

Big data governance

Organization	Establish Data Council, Data Stewards, Data Governance Steering Committee
Metadata	Data definitions, data lineage, technical metadata, data registration
Compliance	Regulatory audits, policies, compliance with security/privacy policies
Data Quality	Ensure data is compete & correct. Measure, improve, certify data
Business Process Integration	Data procurement, ownership, polices around data frequency, availability, etc.
Master Data Management	Establish data & usage taxonomy; business critical hierarchy; Admins, Members, Providers, Consumers, etc.
Information Lifecycle Management (ILM)	Data retention, regulatory compliance with data/model snapshots; purge Schedule, storage/archiving

componente

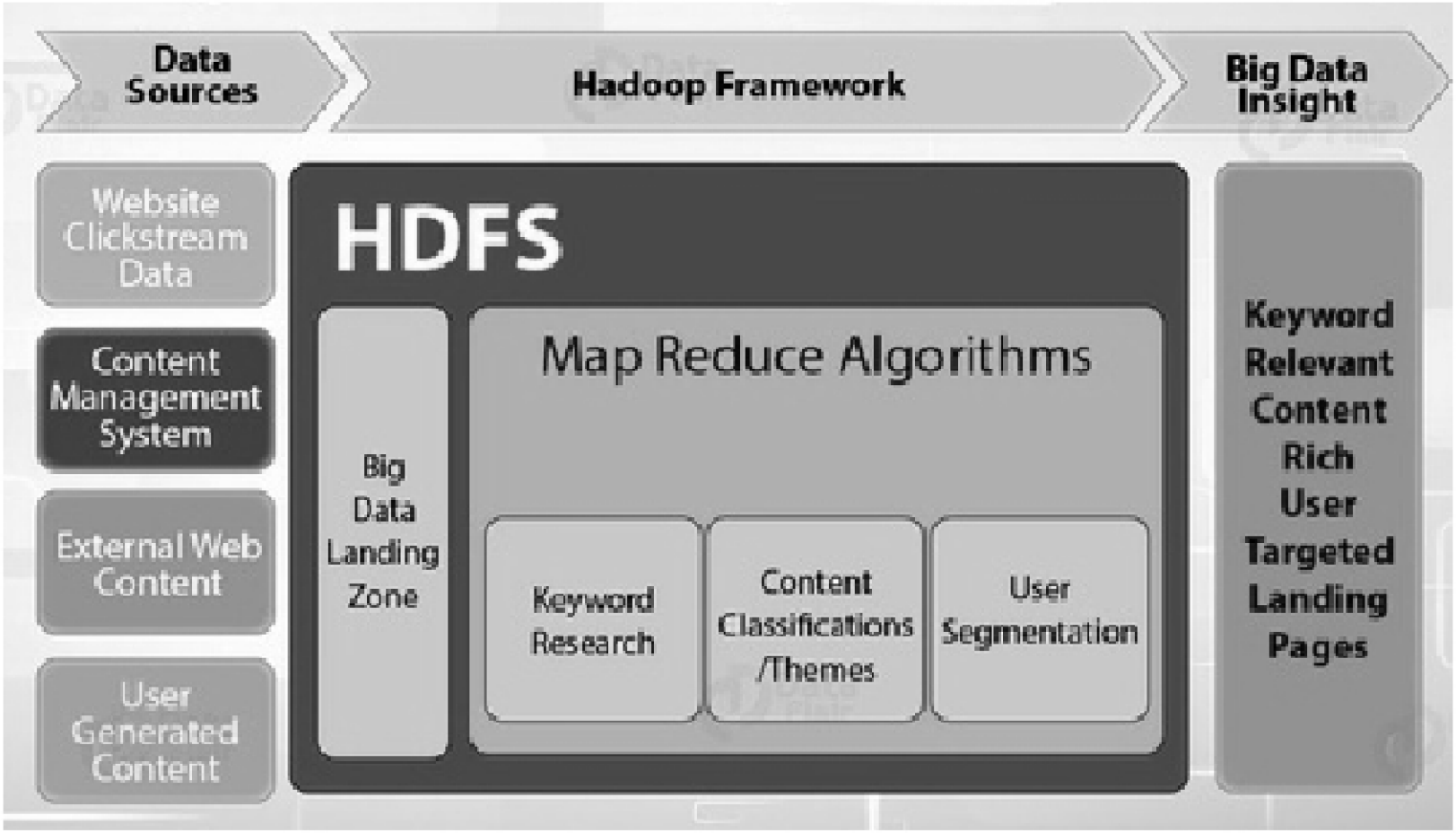
Big data governance

bazată pe Hadoop

Hadoop?

studenți sau profesori?

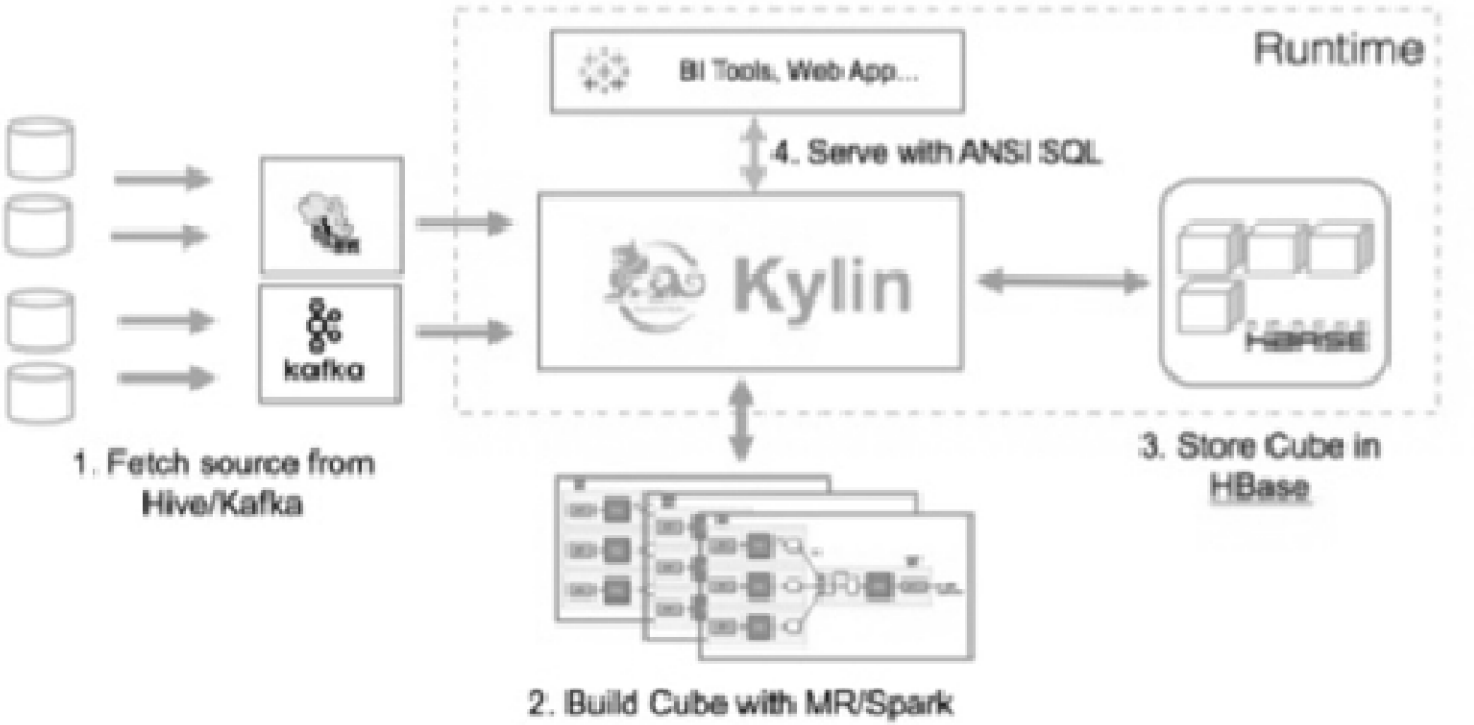
Hadoop


Data Sources
Hadoop Framework
Big Data Insight
Website Clickstream Data
Content Management System
External Web Content
User Generated Content
HDFS
Big Data Landing Zone
Map Reduce Algorithms
Keyword Research
Content Classifications /Themes
User Segmentation
Keyword Relevant Content Rich User Targeted Landing Pages

- X

Guvernarea datelor: Tactică sau Strategică?

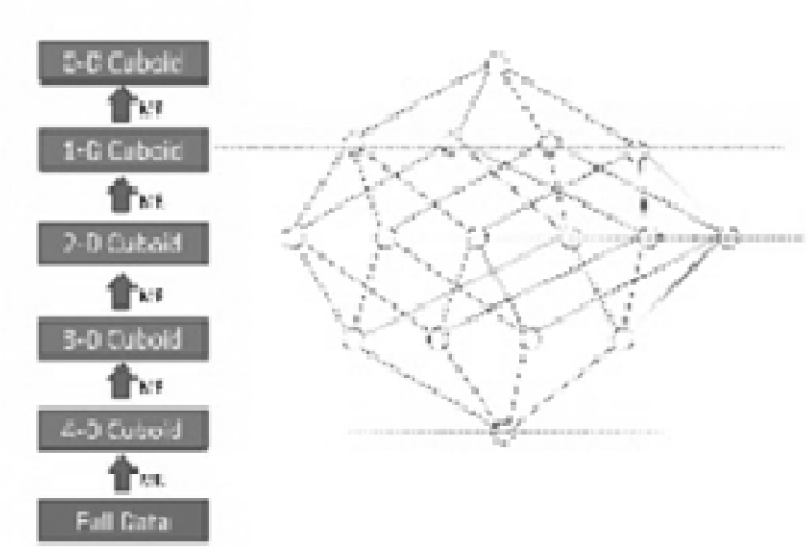
OLAP & Hadoop



- un exemplu bazat pe kylin

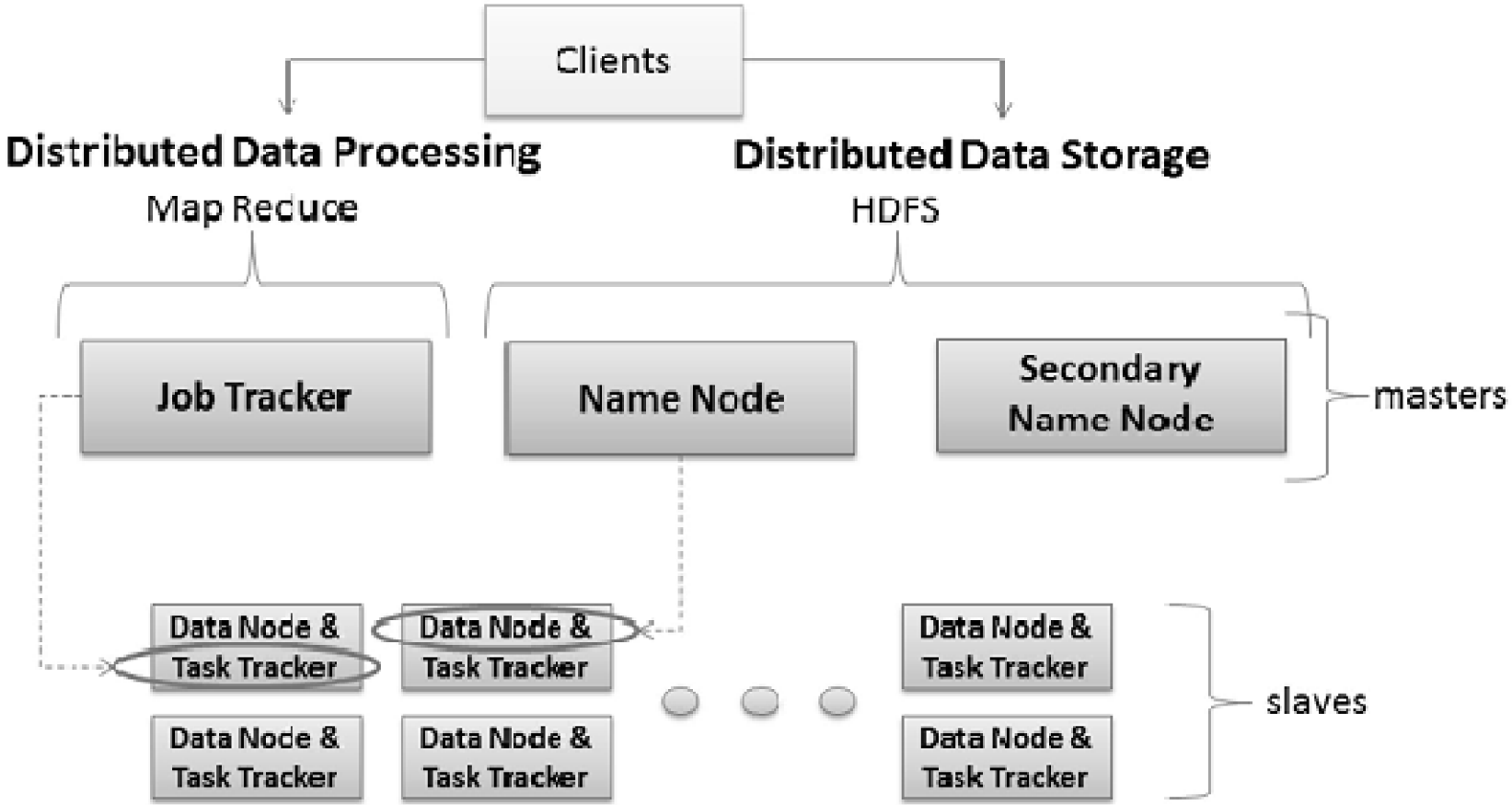
OLAP & Hadoop

- Calculate Cuboids by layer :N dim (Base cuboid), N-1 dim, N-2…, 1, 0
- Reuse previous layer's result
- HDFS used for data sharing
- Totally need N round MR;



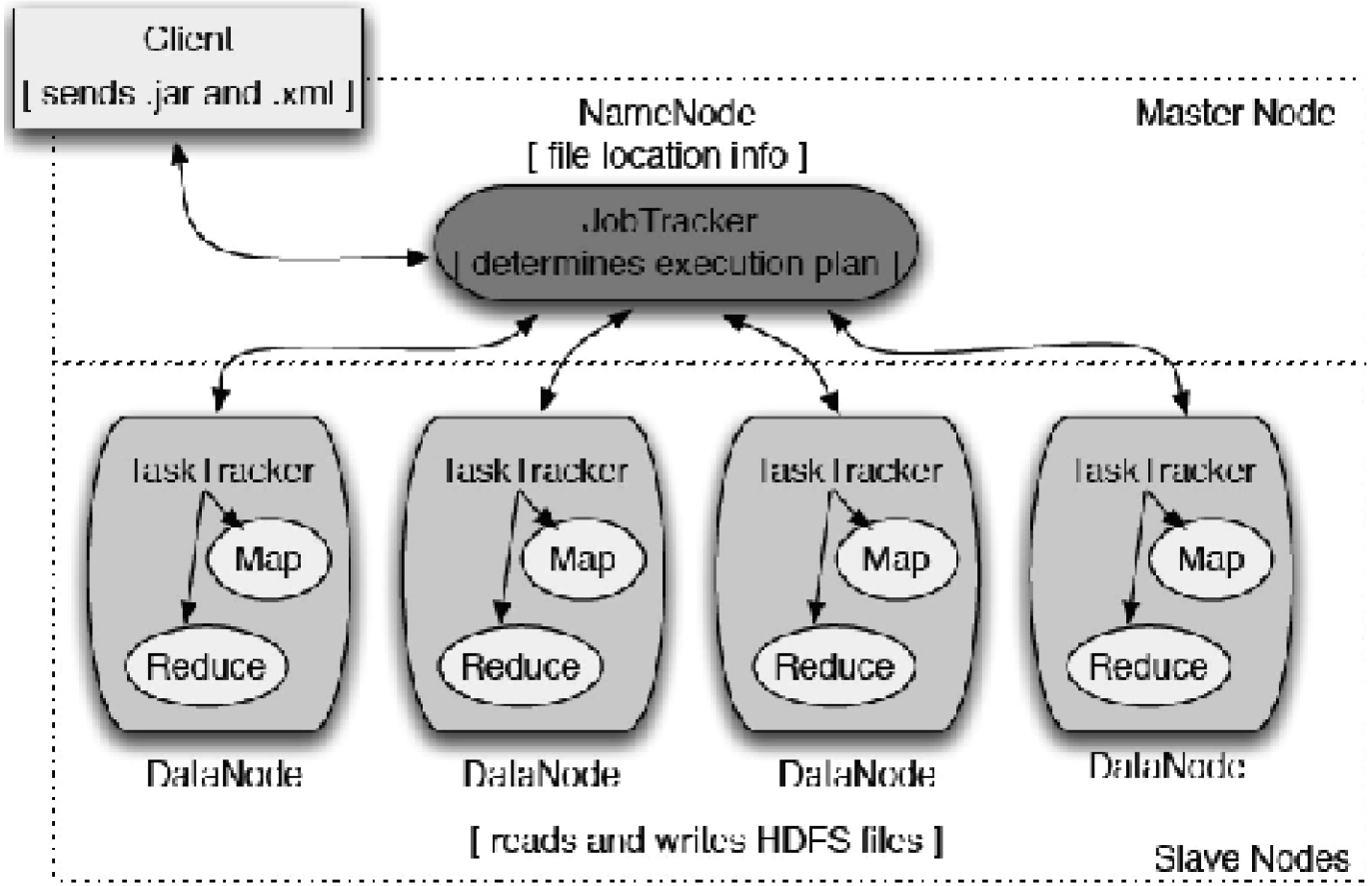
- generarea cubului cu mapreduce

HDSF



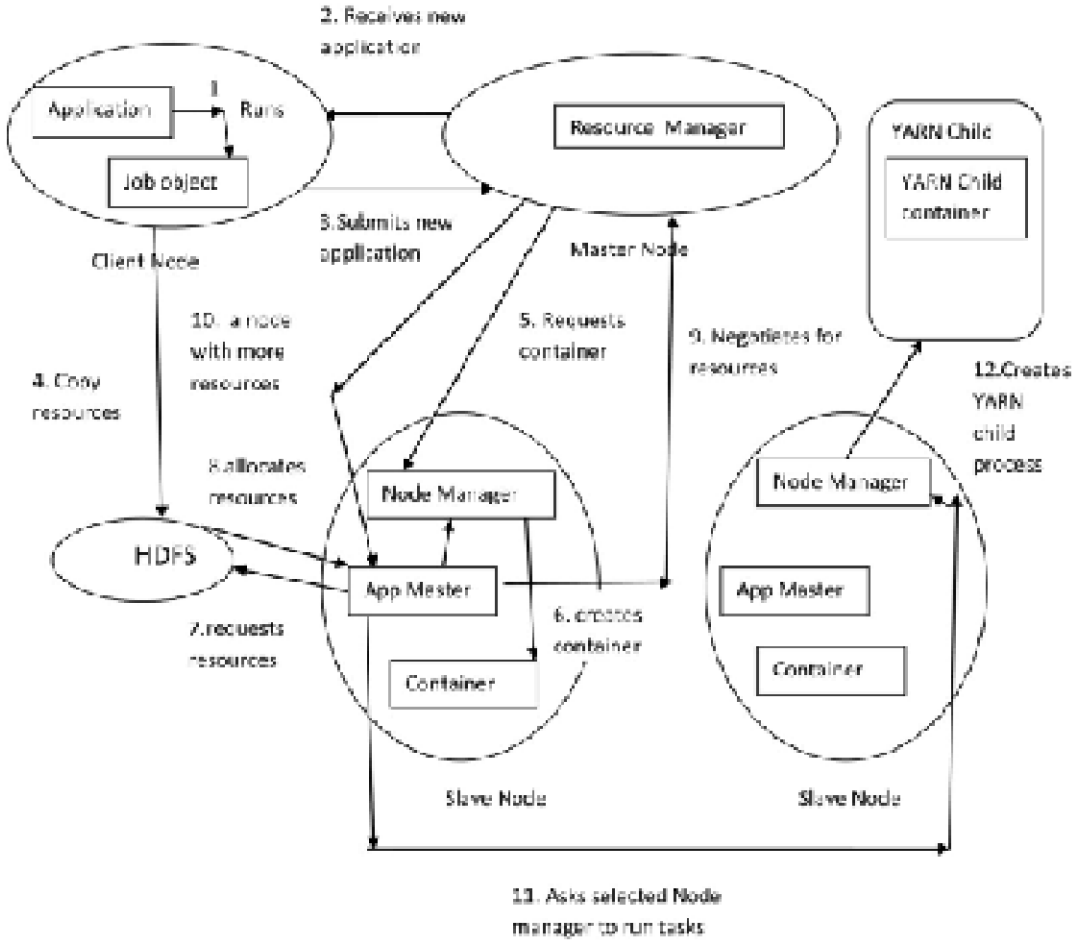
- roluri server in hadoop

MapReduce



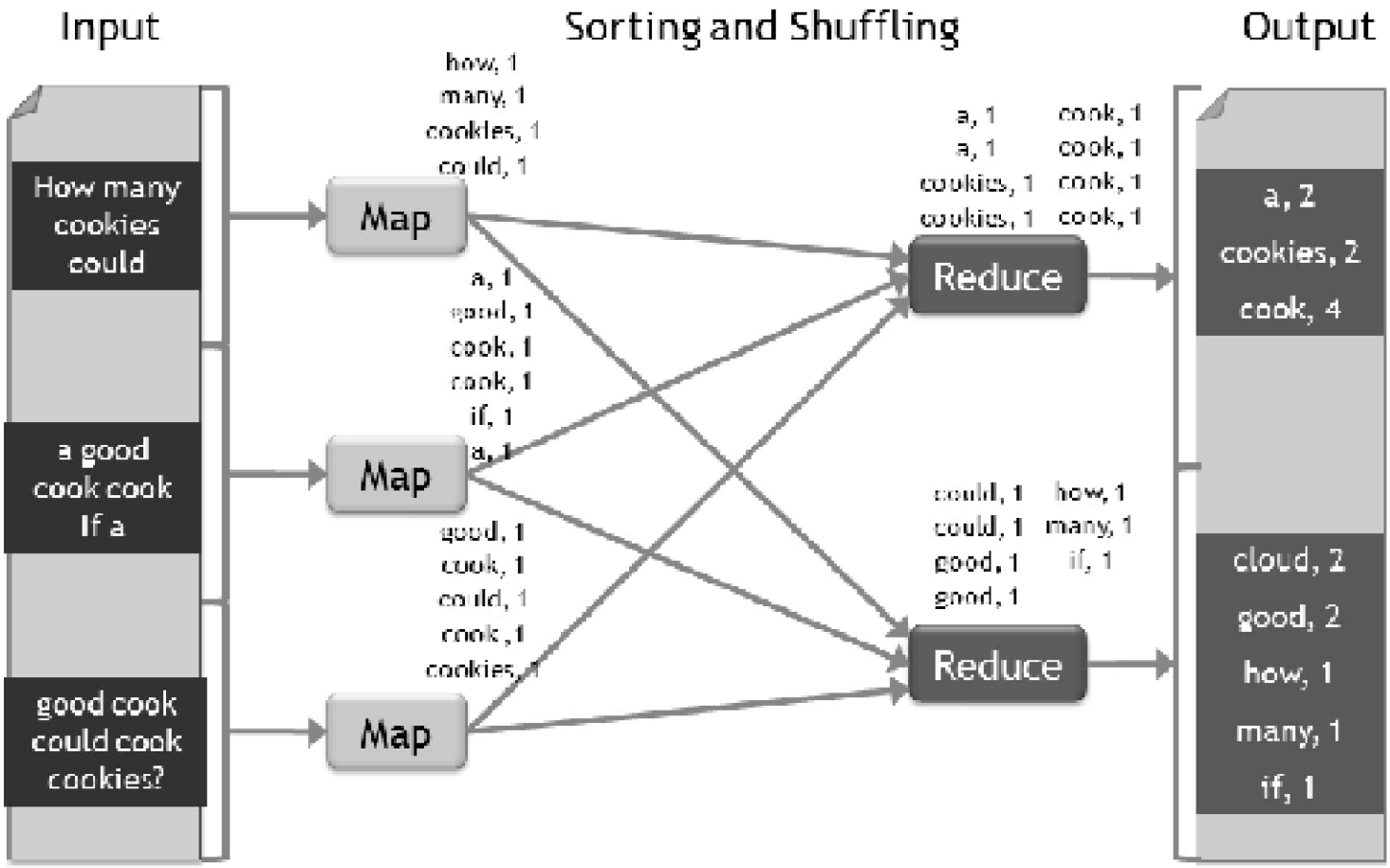
un instrument!

MapReduce



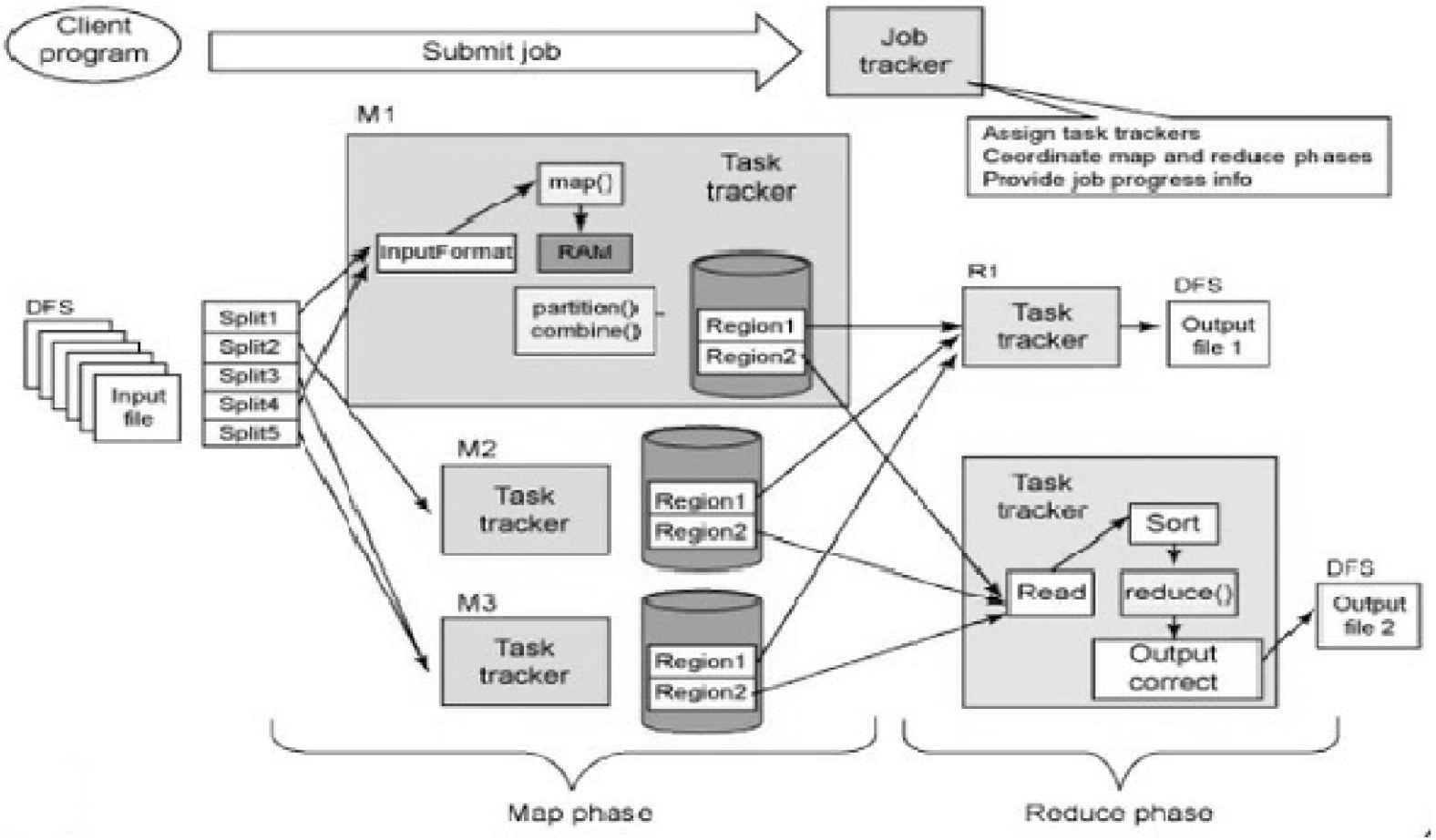
execuția unui job în map reduce

Map Reduce - exemplu



<k1, v1> ->transformare -> <k2, v2> -> reducere -> <k3, v3>

Map Reduce



Pașii

Securitatea Hadoop

subțirică rău

Măsuri minime securizare hadoop

ca de obicei!

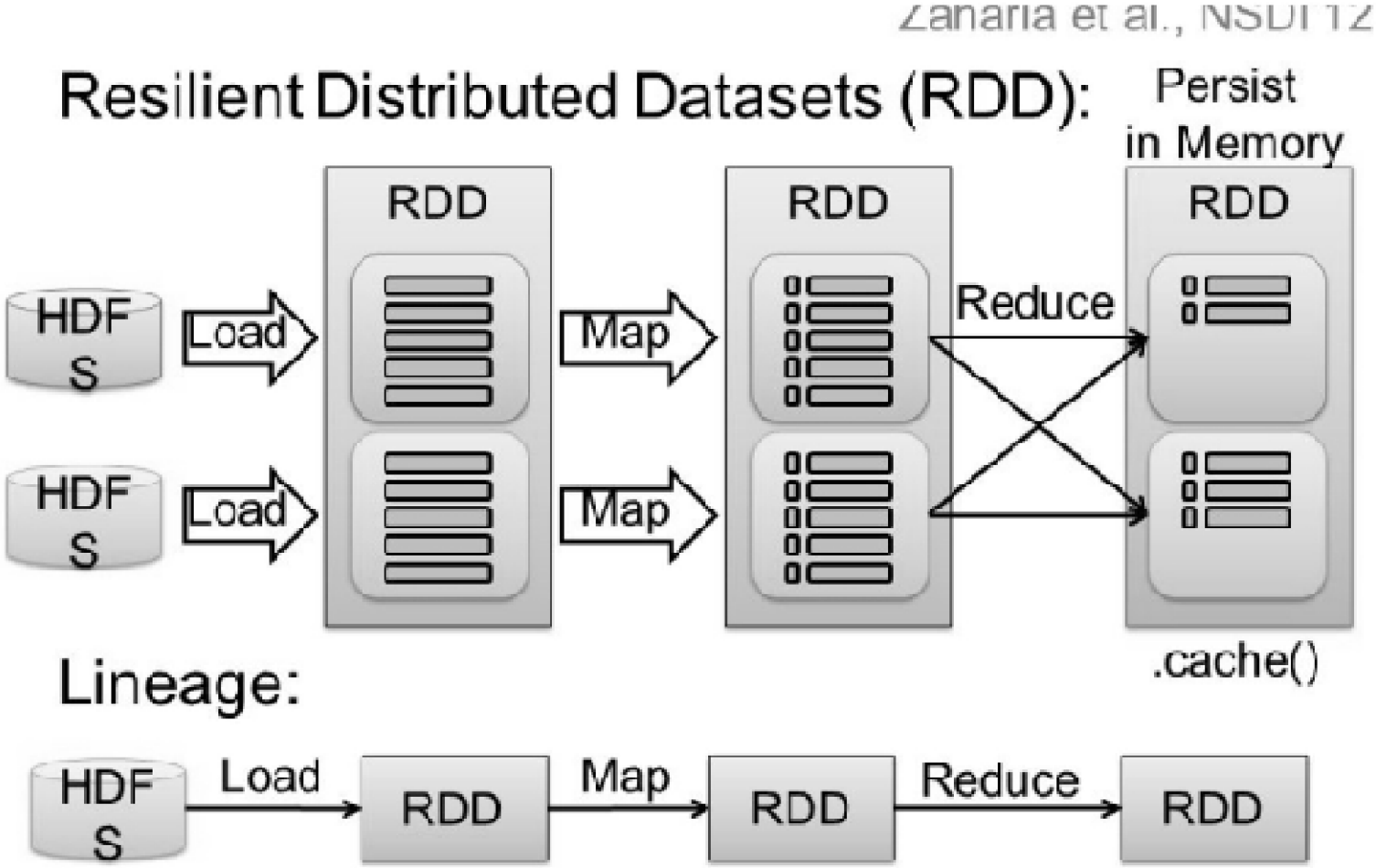
Hadoop sau ElasticSearch?

Hadoop

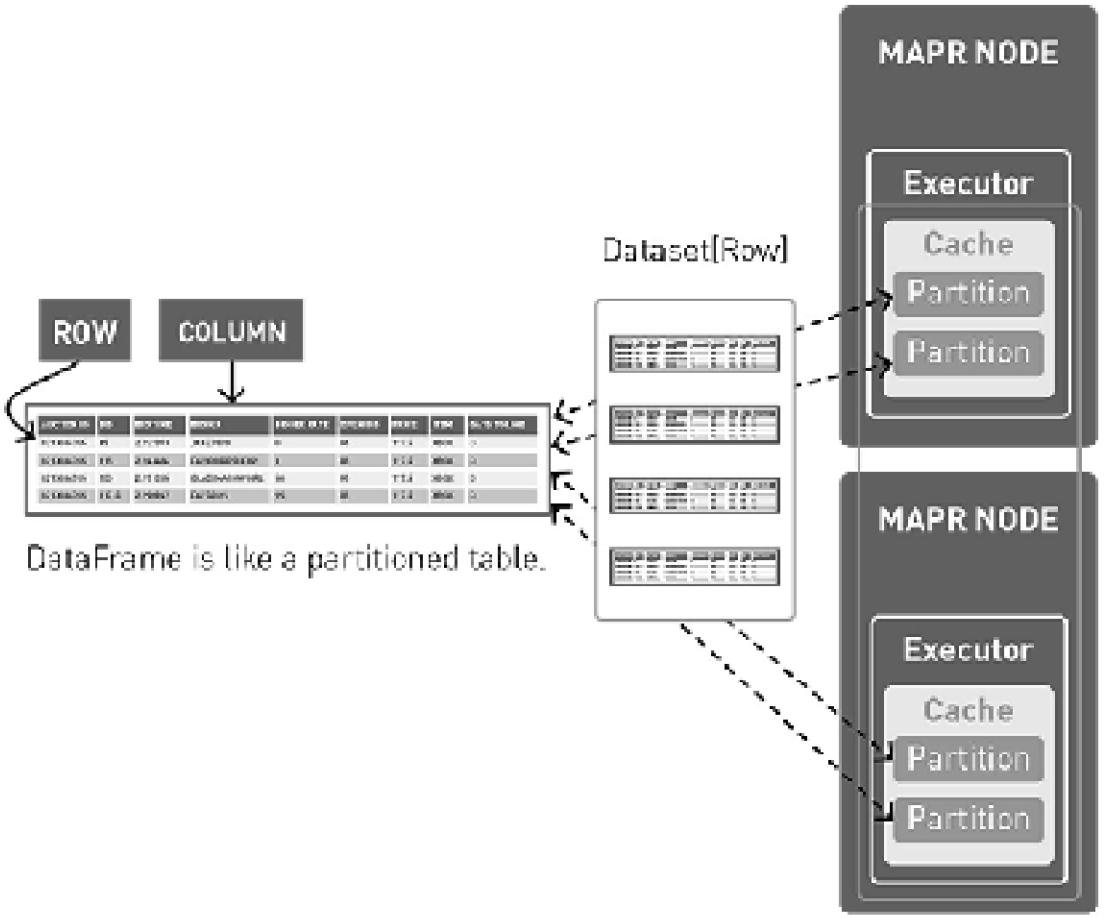
Exemple aplicare Map Reduce

cel mai cunoscut!

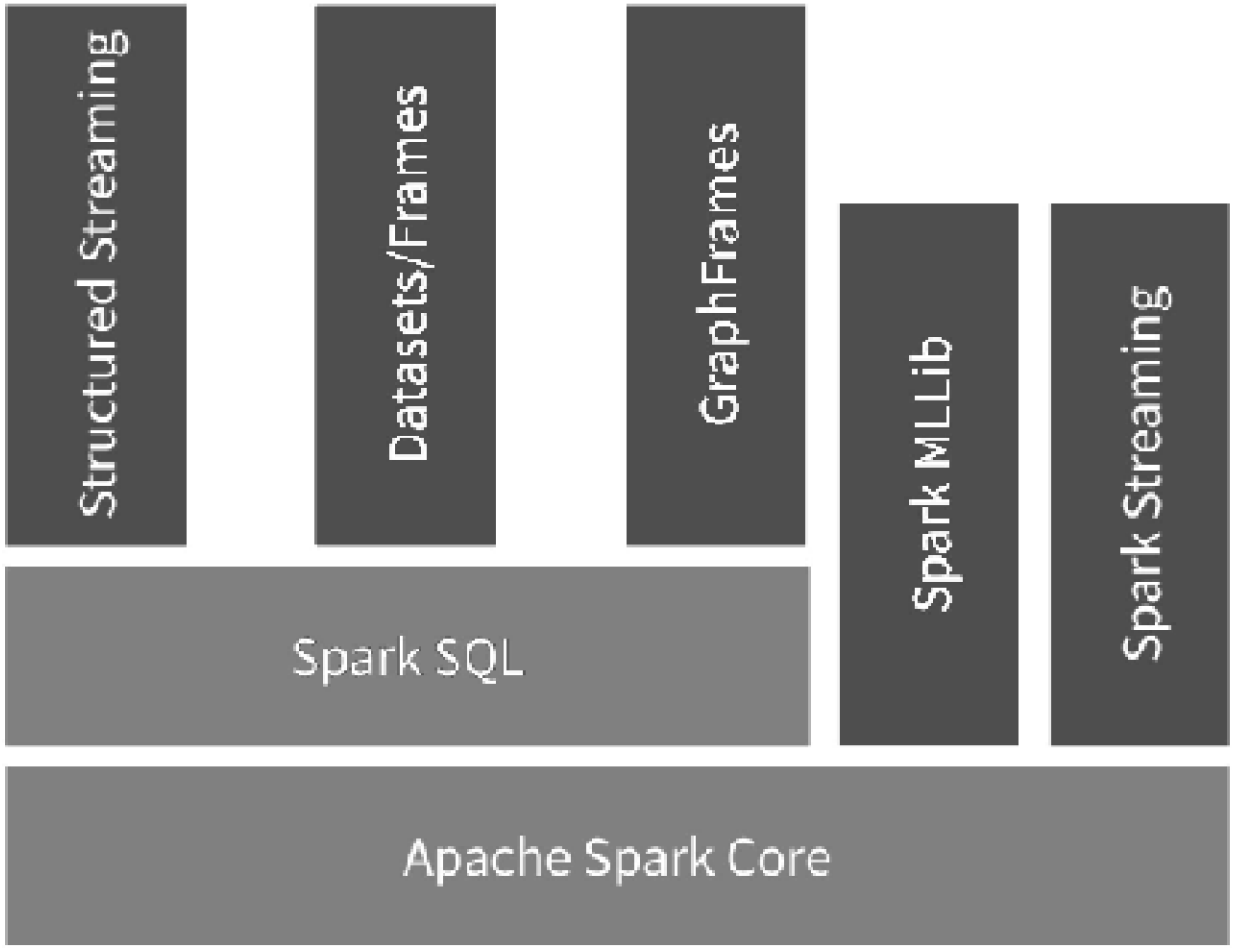
Spark


Zaharia et al., NSDI'12
Resilient Distributed Datasets (RDD):
Persist in Memory
RDD
RDD
RDD
HDFS
Load
Map
Reduce
HDFS
Load
Map
.cache()
Lineage:
HDFS
Load
RDD
Map
RDD
Reduce
RDD

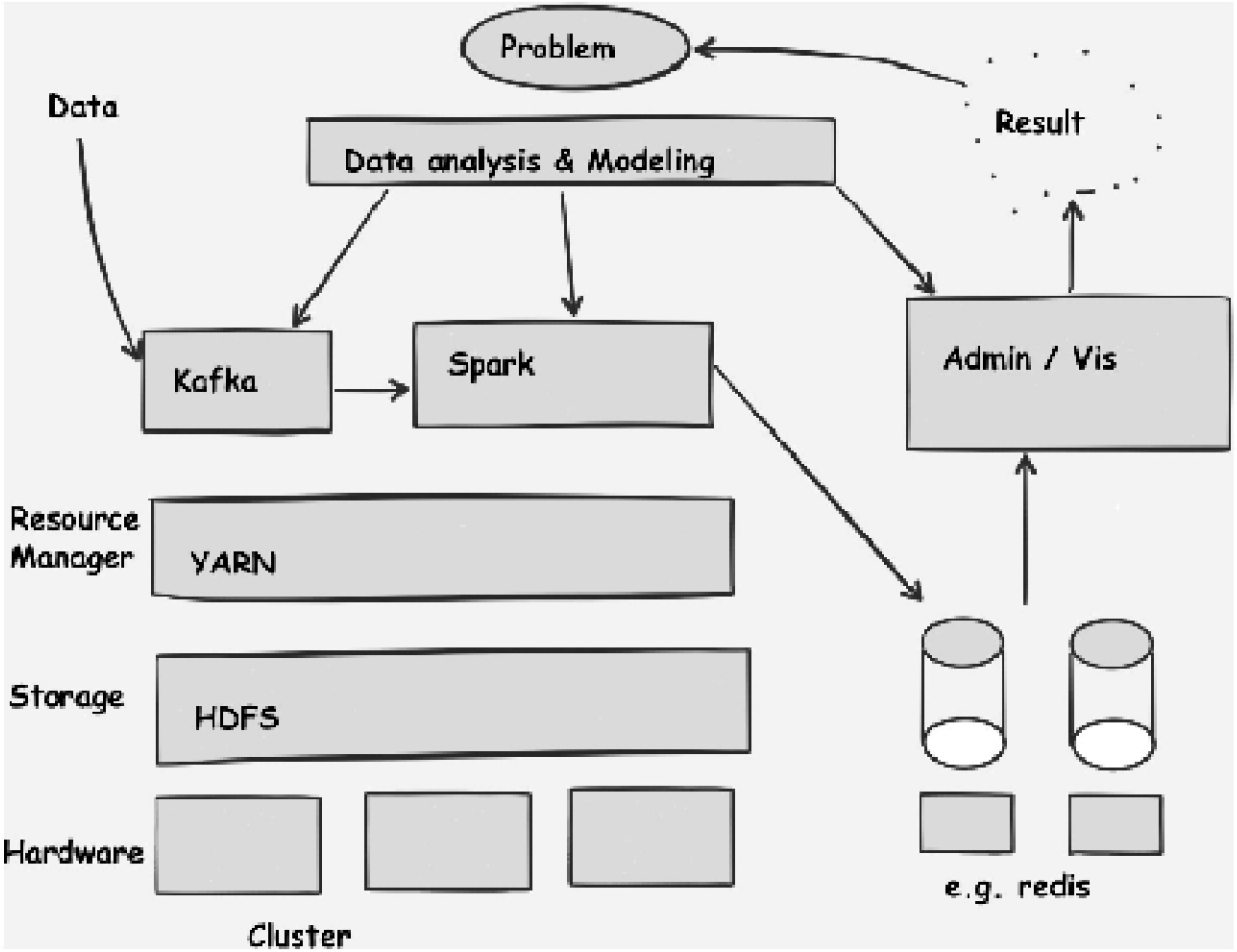
Spark SQL


MAPR NODE
Executor
Cache
Partition
Partition
Dataset[Row]
ROW
COLUMN
DataFrame is like a partitioned table.
MAPR NODE
Executor
Cache
Partition
Partition

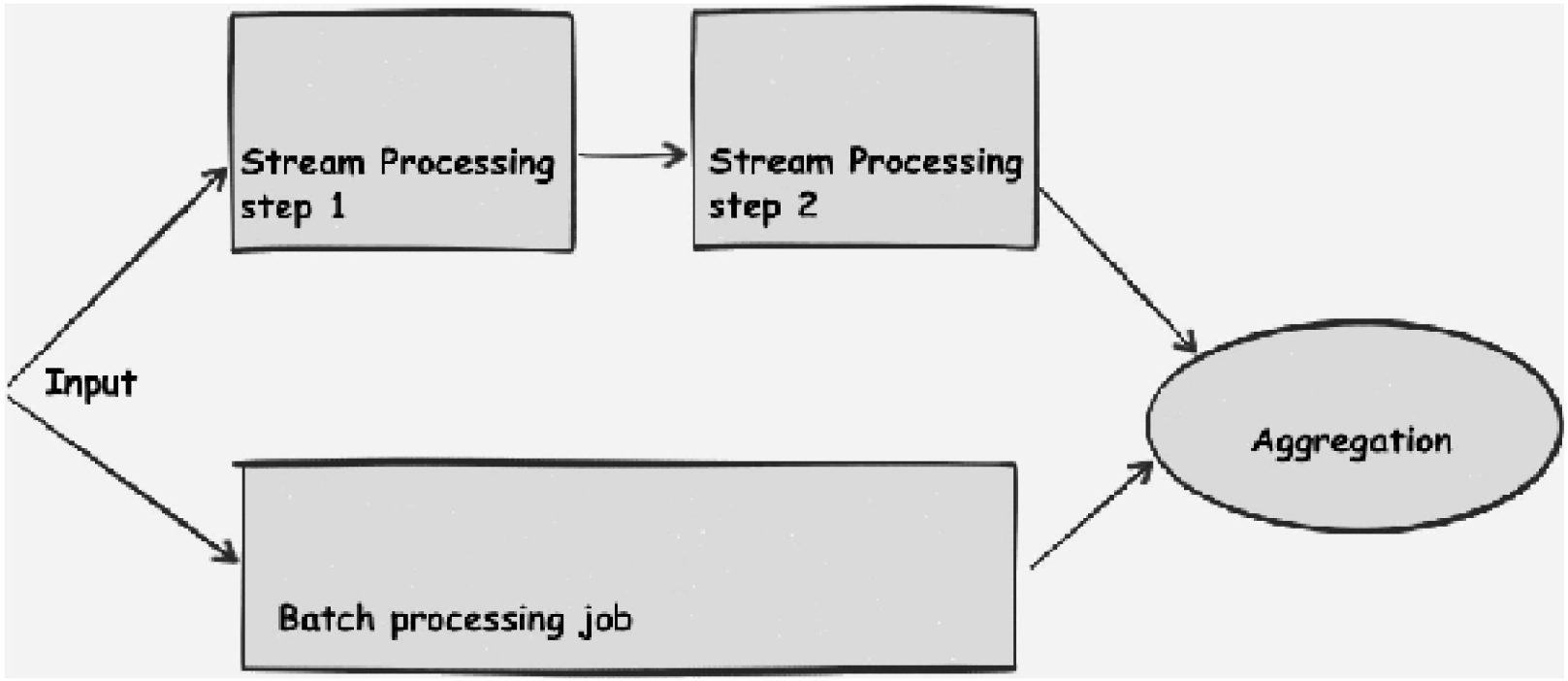
Structura Spark



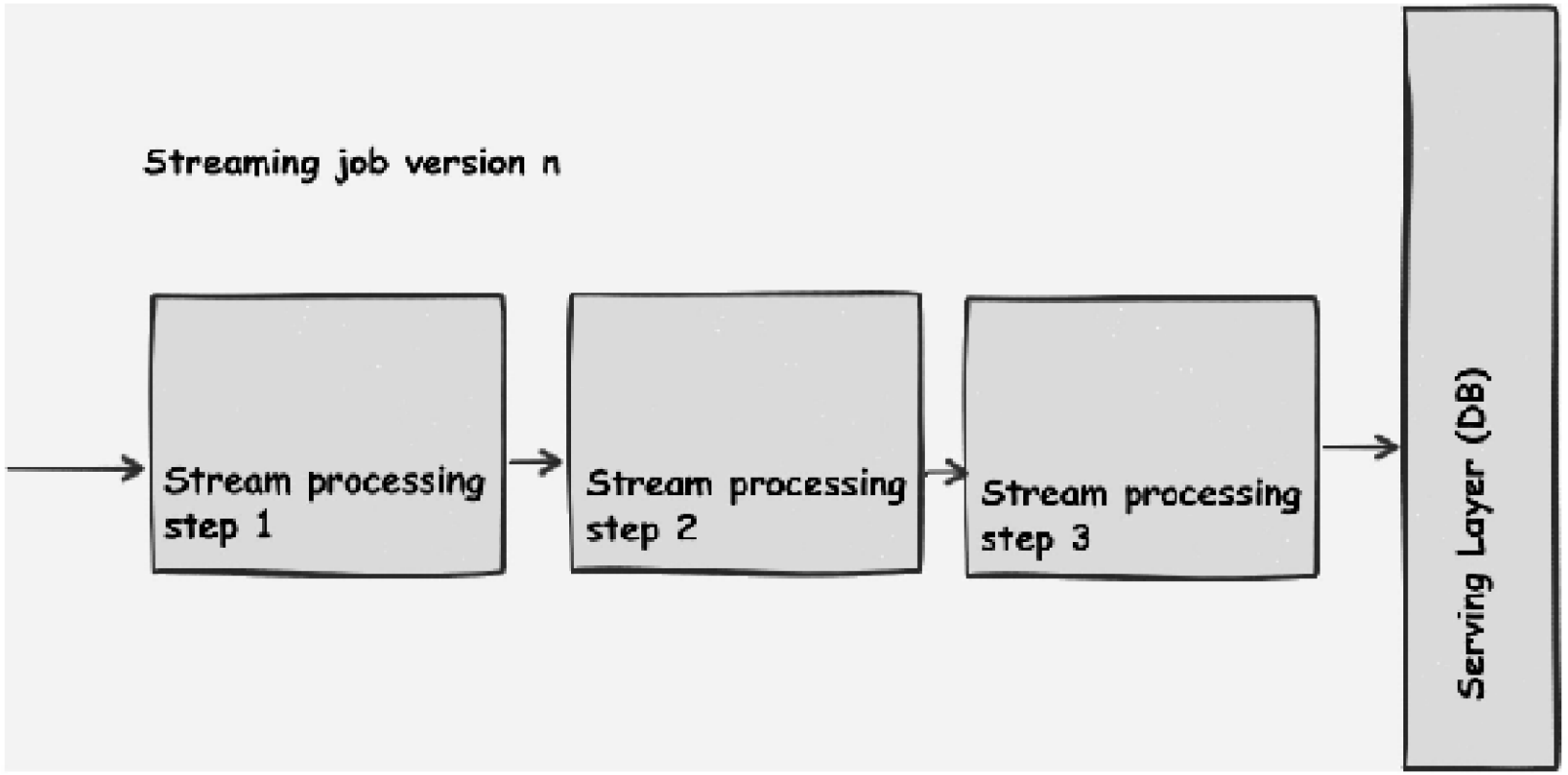
O posibilă platformă de date bazată pe Spark



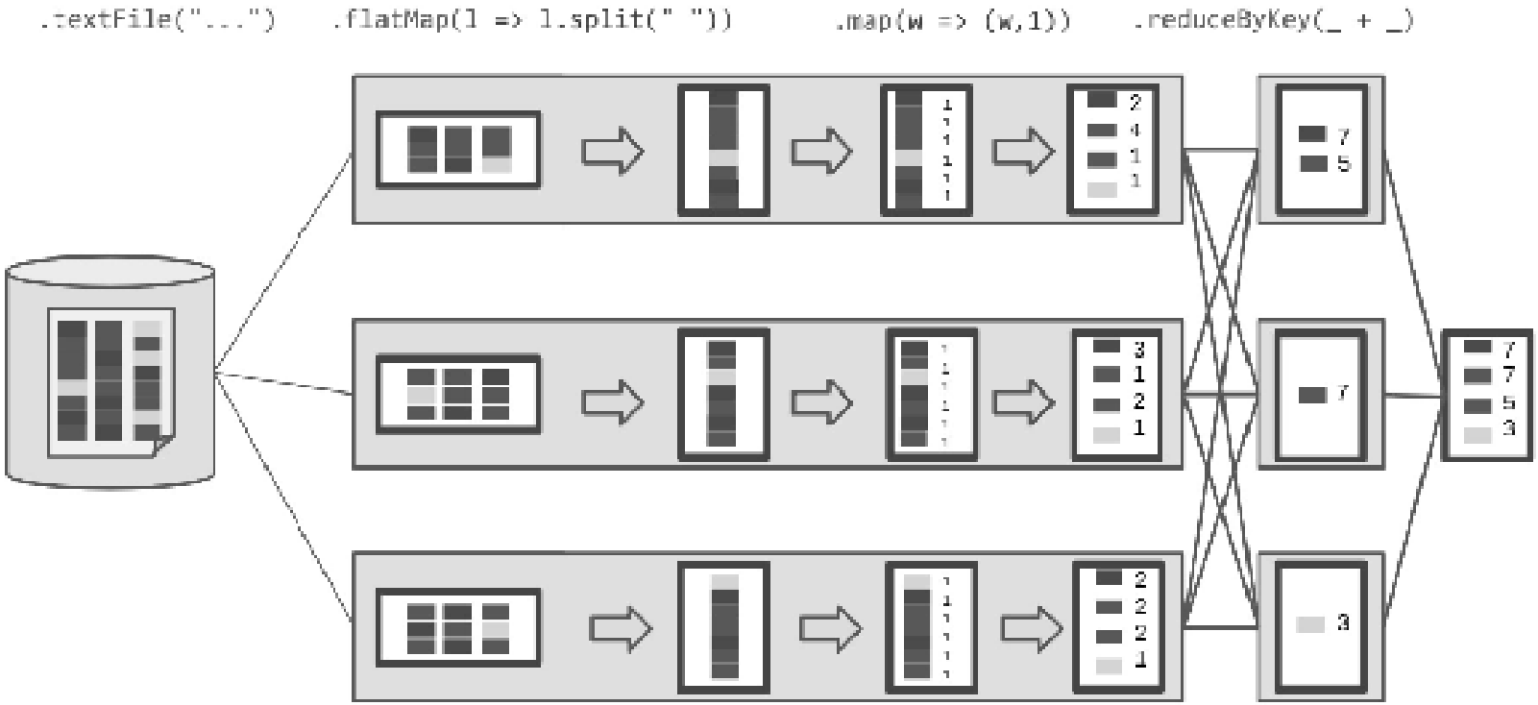
Arhitectura Lambda



Arhitectura Kappa

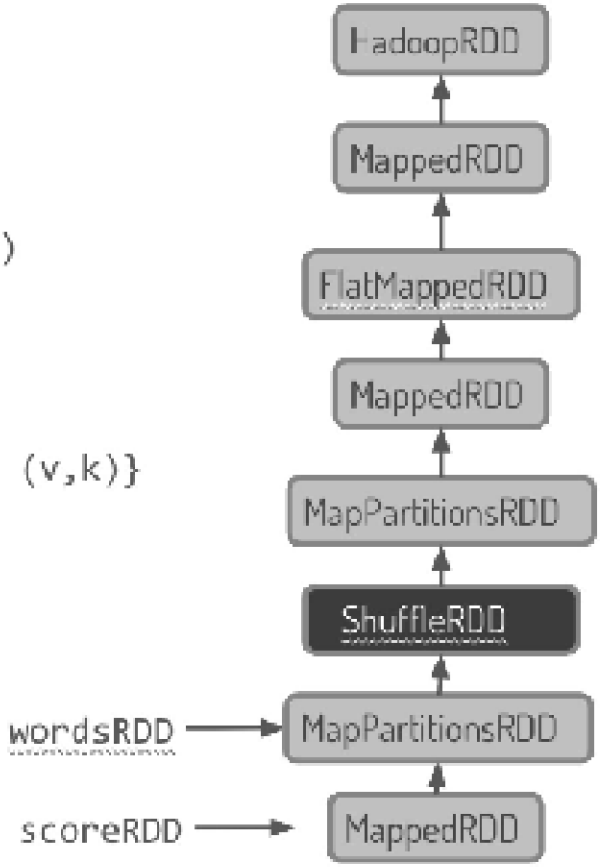


RDD

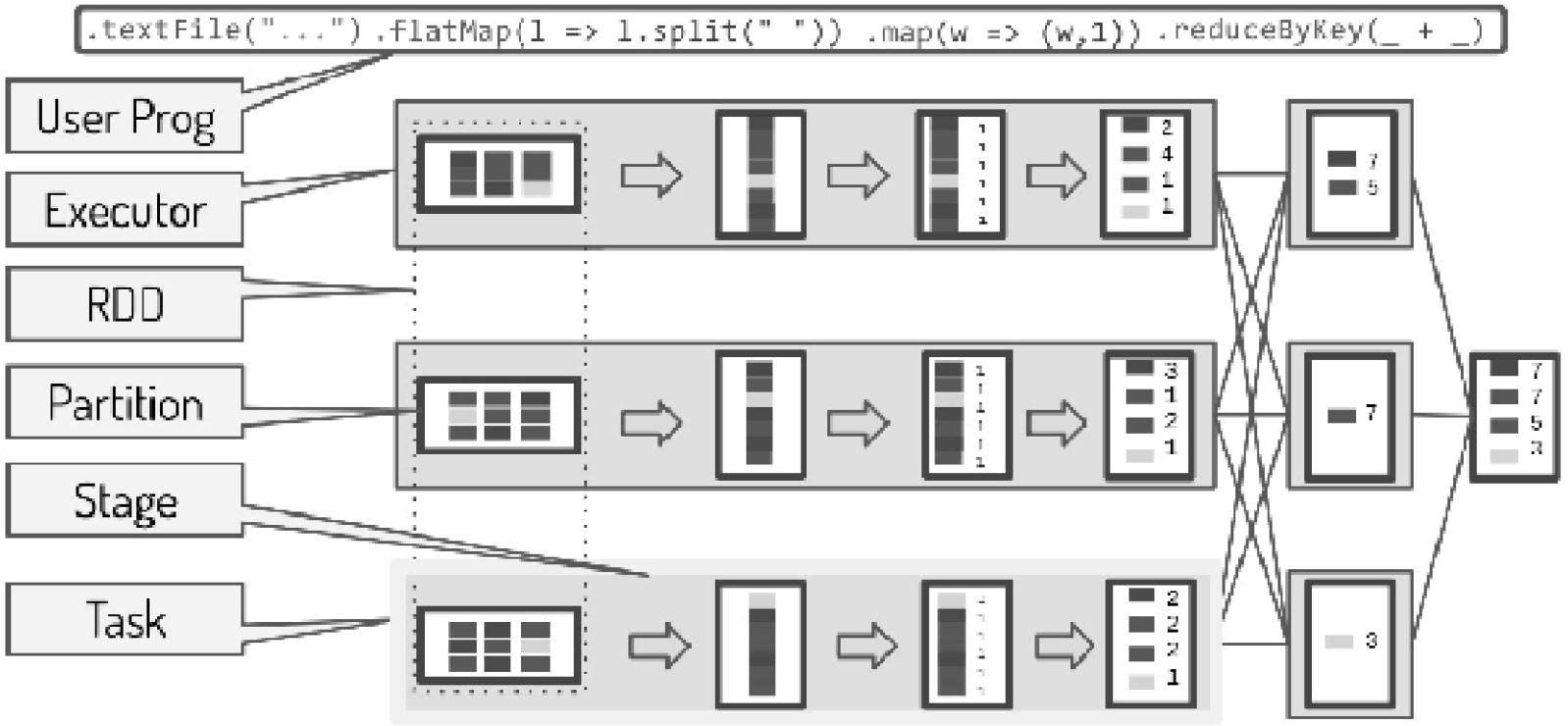


RDD & DAG

```
val file = spark.textFile("hdfs://...")
val wordsRDD = file.flatMap(line =>
line.split(" "))
.map(word => (word, 1))
.reduceByKey(_ + _)
val scoreRdd = words.map{case (k,v) => (v,k)}
```



Spark Components



- X