# Sisteme Distribuite

Cursul 12

Mihai Zaharia

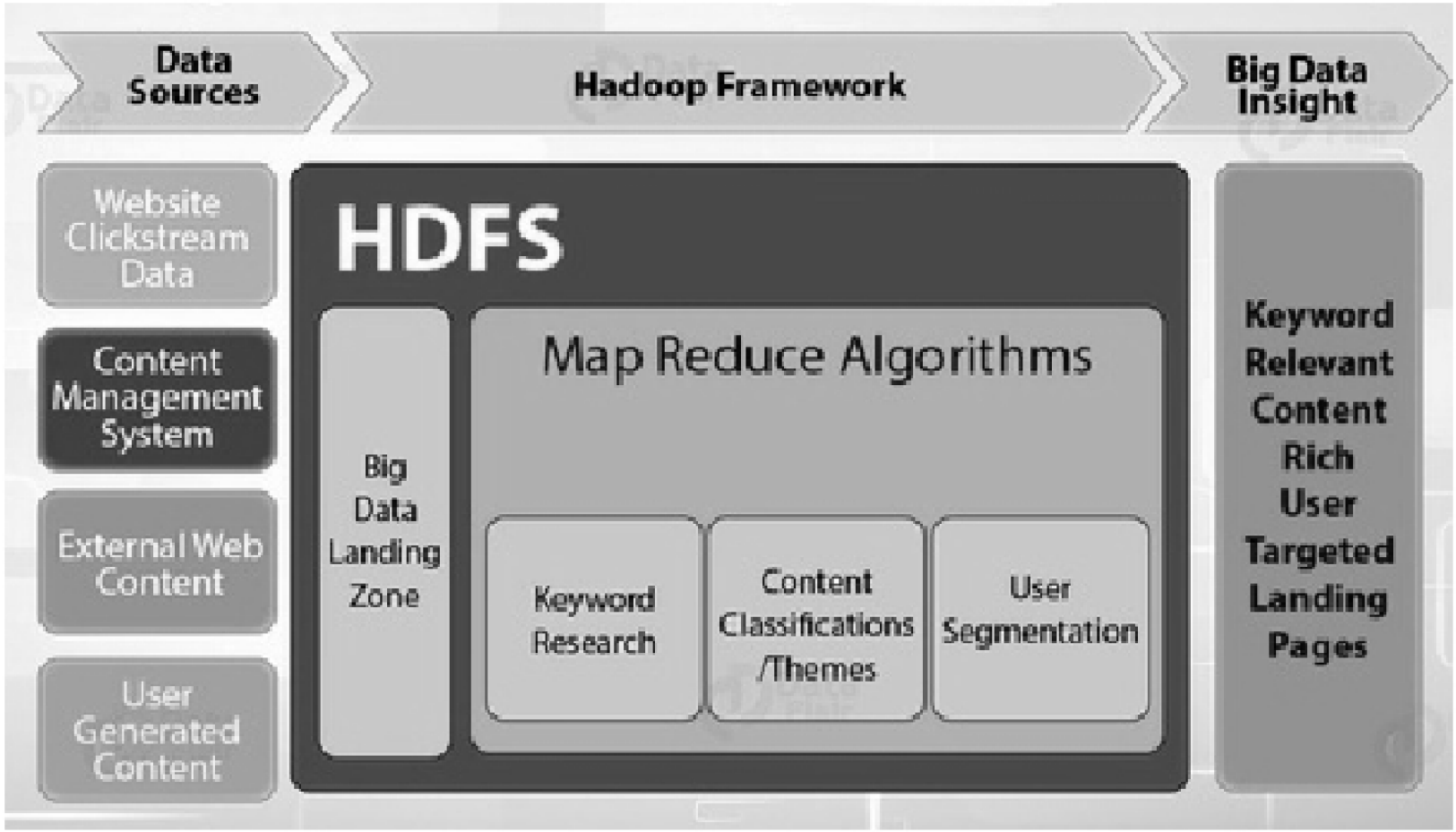
# Big data governance


bazată pe Hadoop

# Hadoop?

studenți sau profesori?

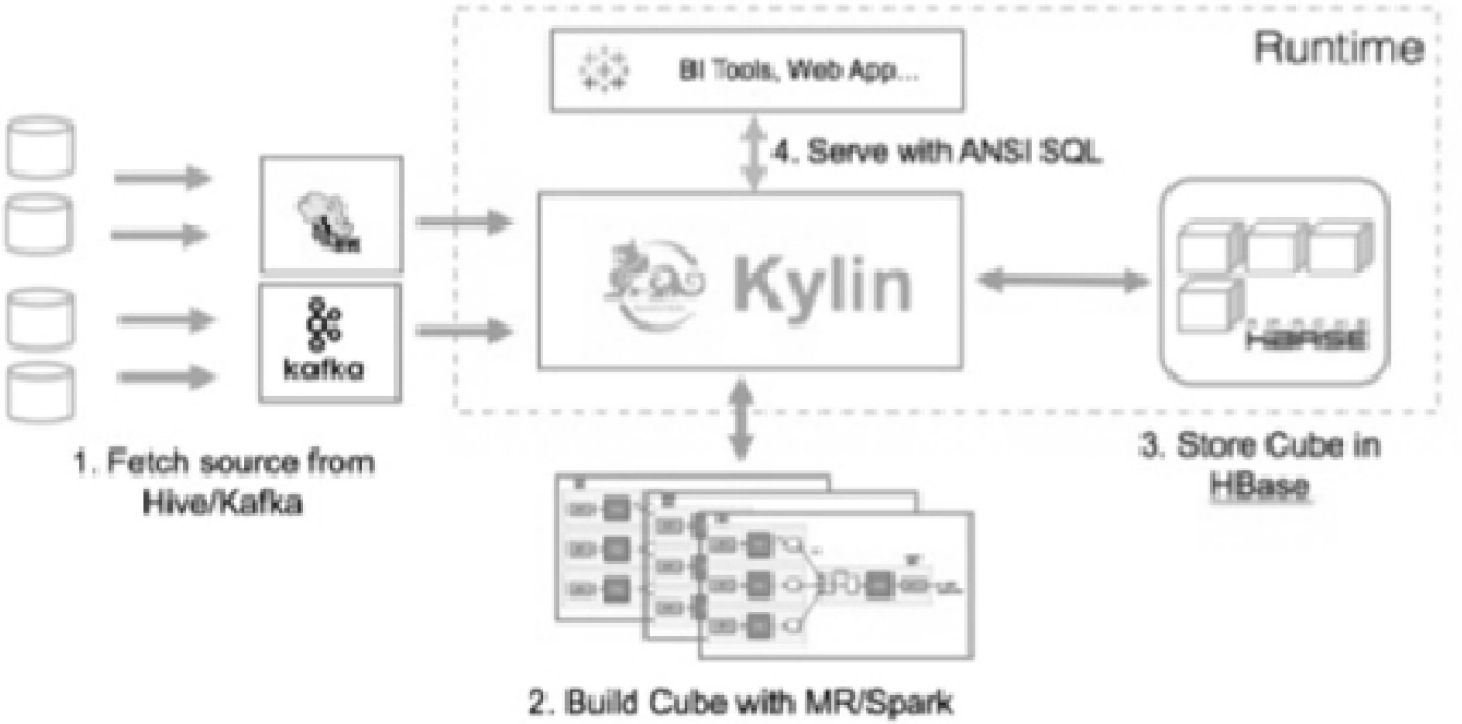
# Hadoop


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?

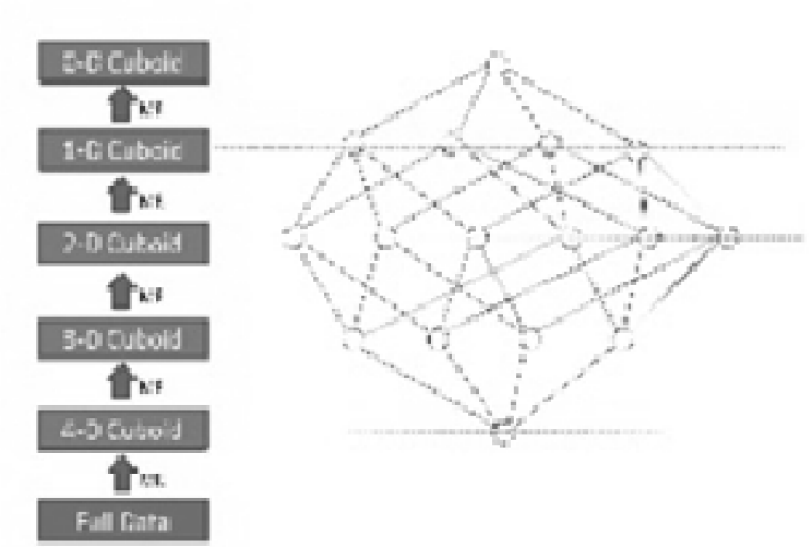
# OLAP & Hadoop




- un exemplu bazat pe kylin

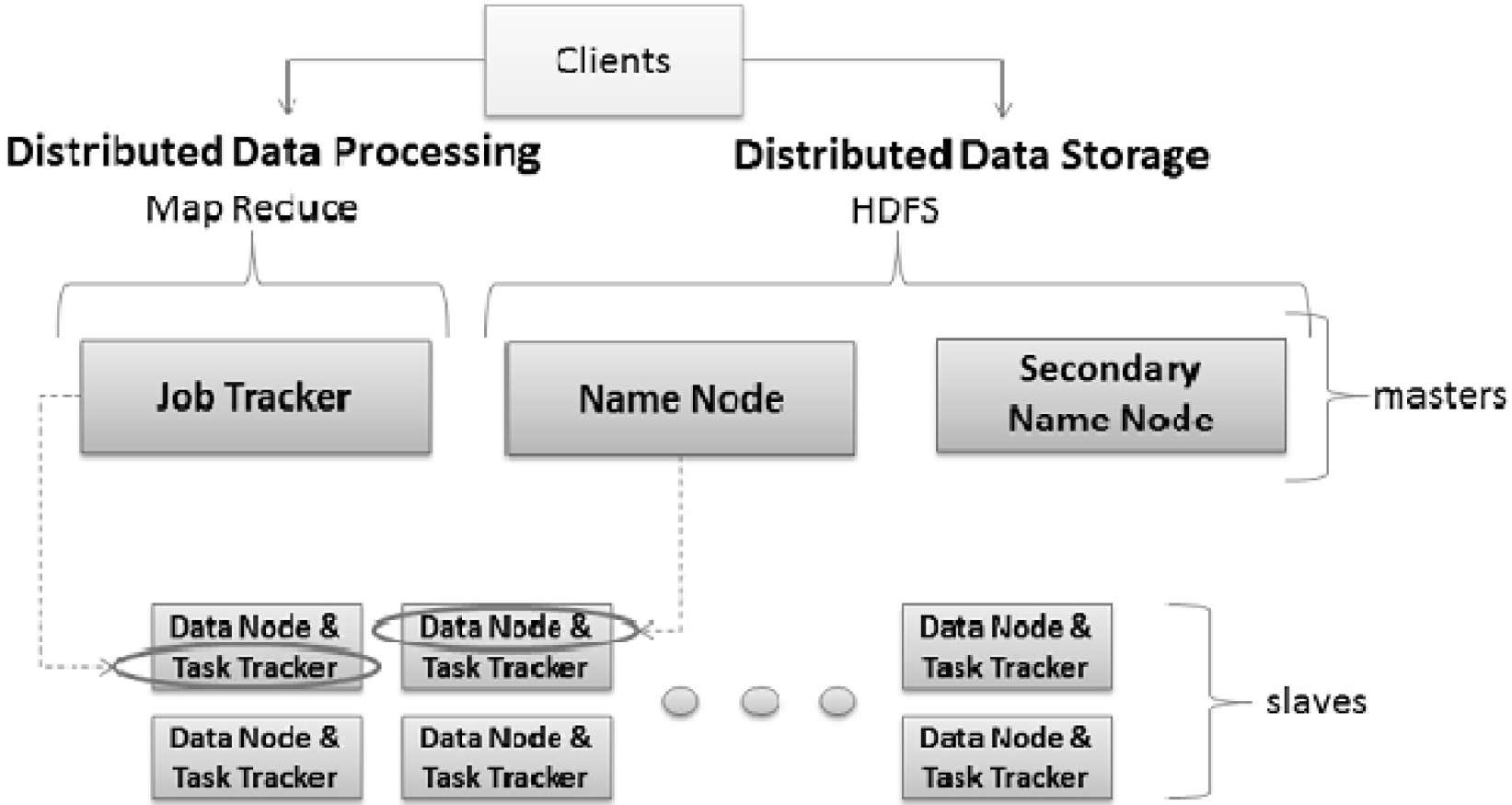
# OLAP & Hadoop

- 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;



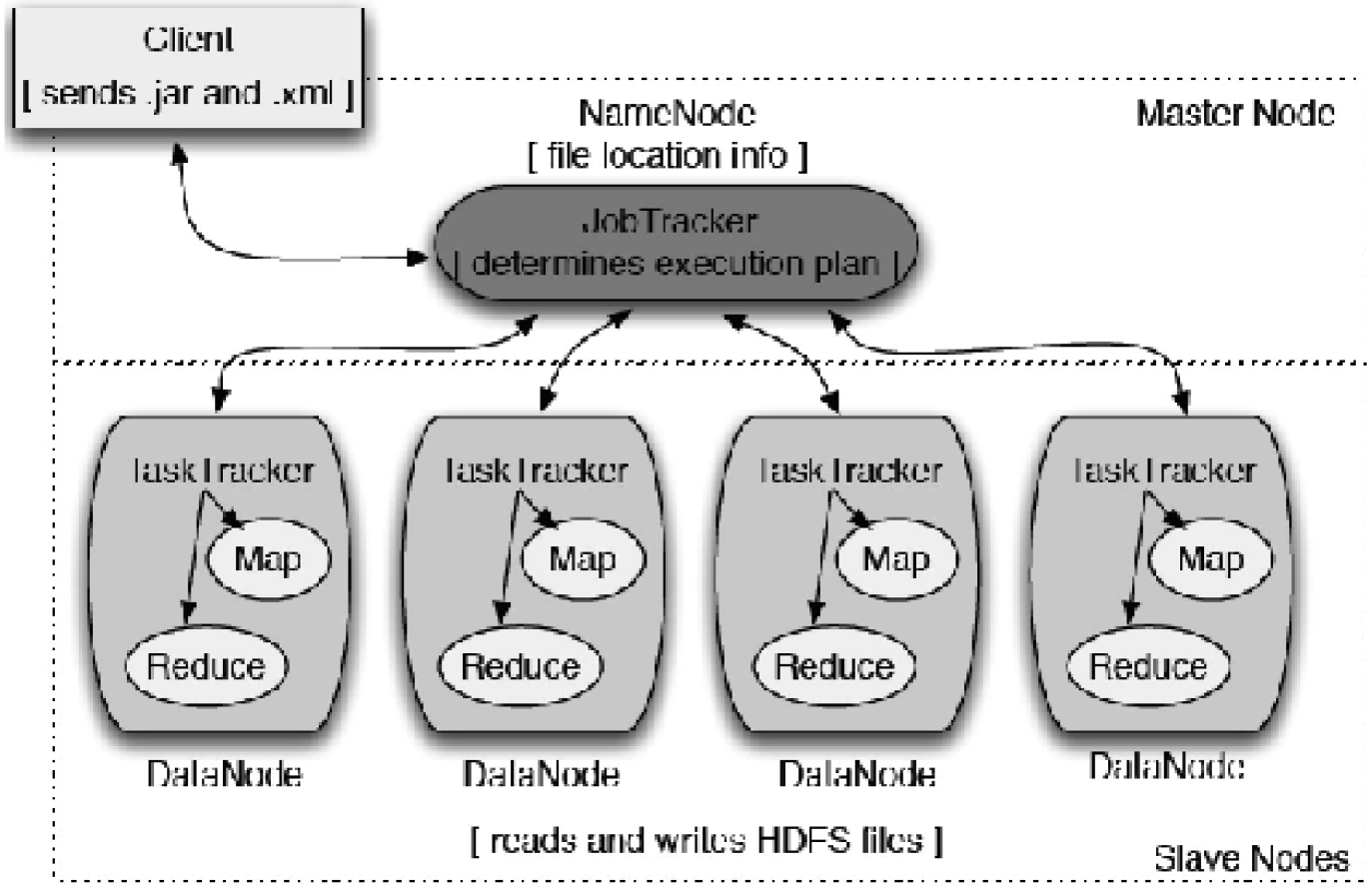

- generarea cubului cu mapreduce

# HDSF



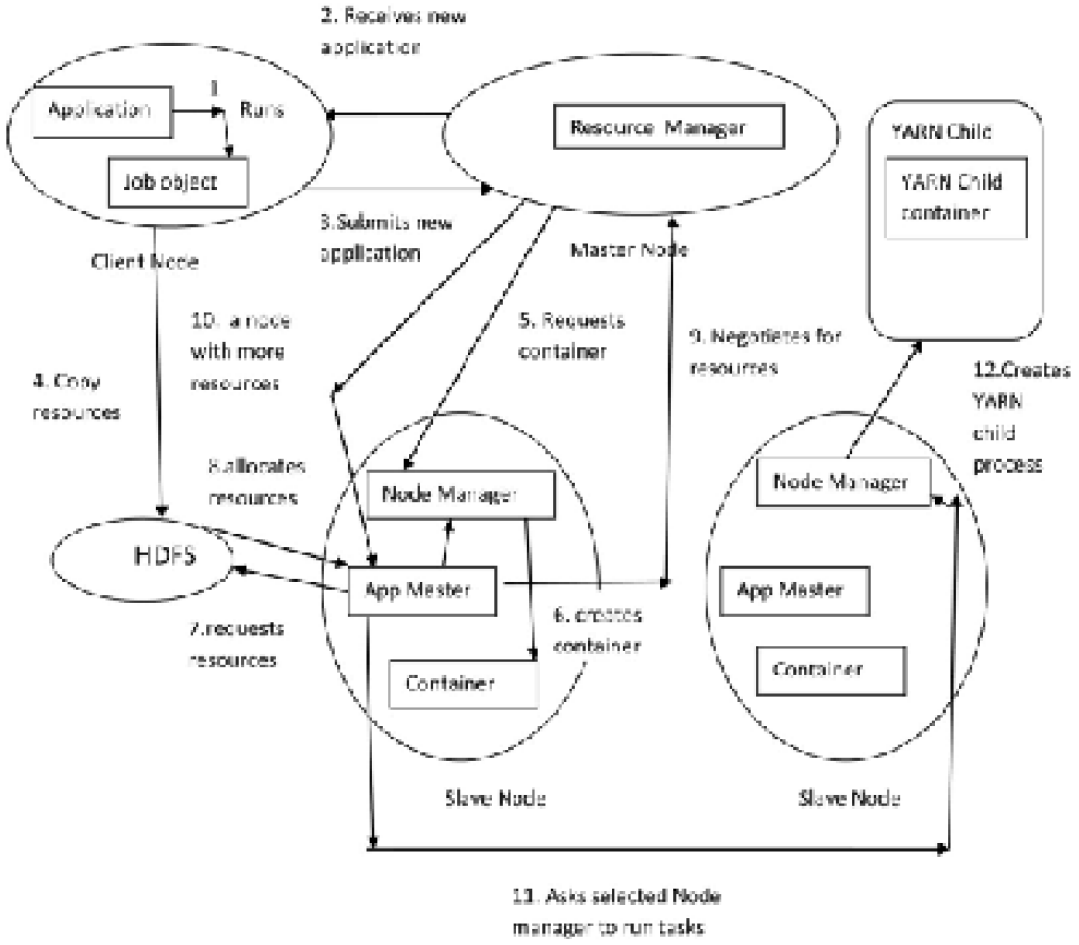

- roluri server in hadoop

# MapReduce



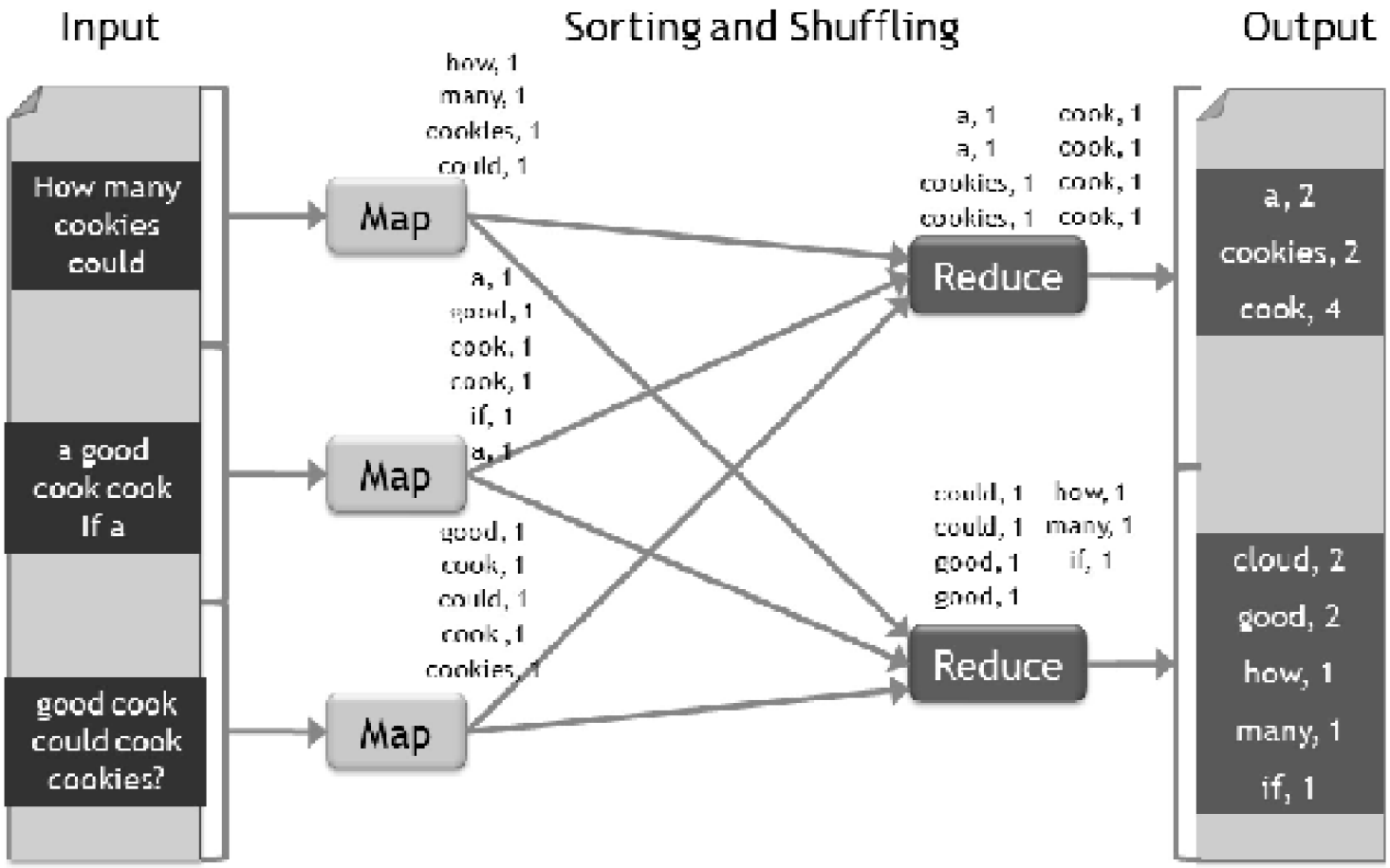

un instrument!

# MapReduce



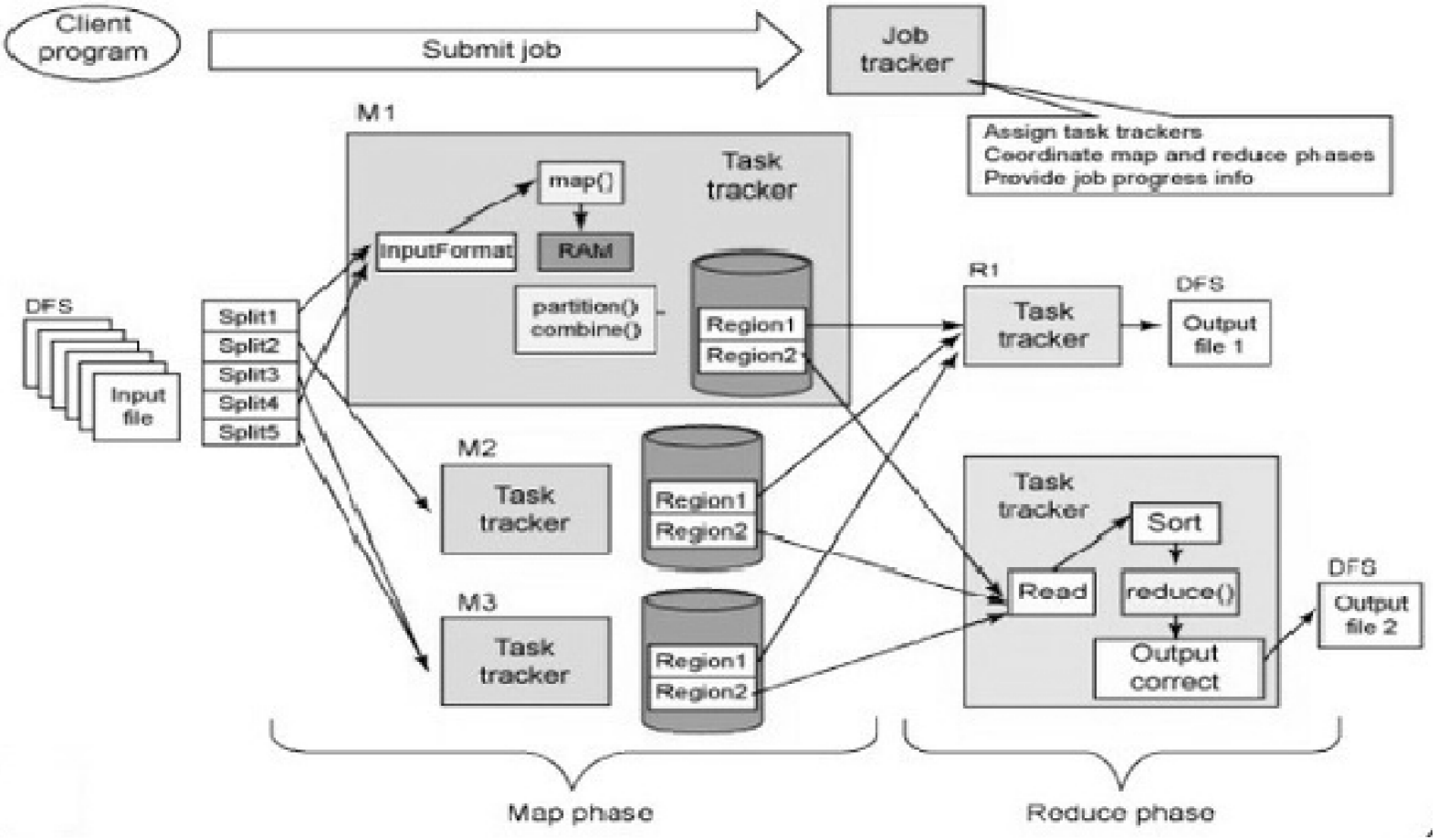

execuția unui job în map reduce

# Map Reduce - exemplu




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

# Map Reduce



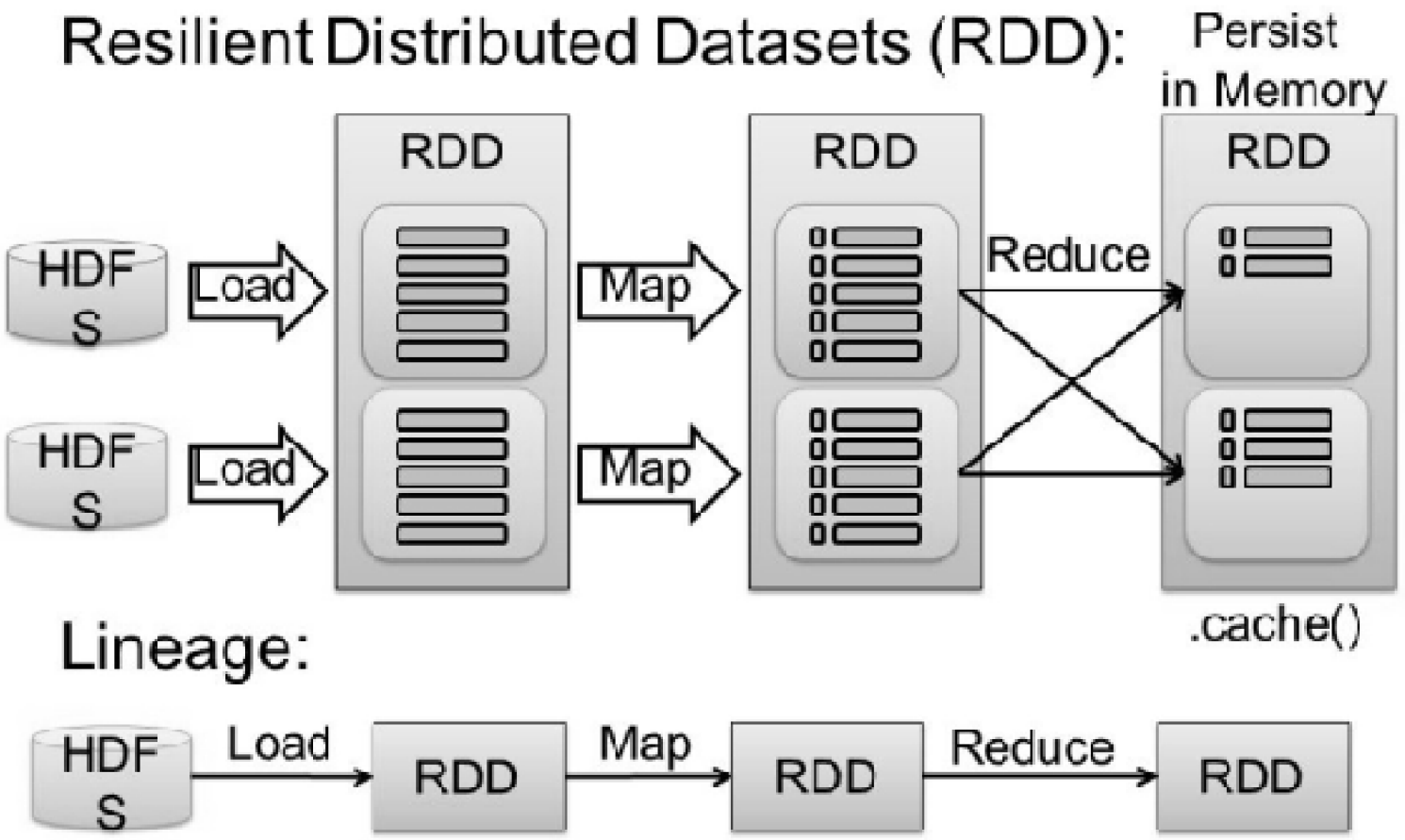

Pașii

# Hadoop sau ElasticSearch?

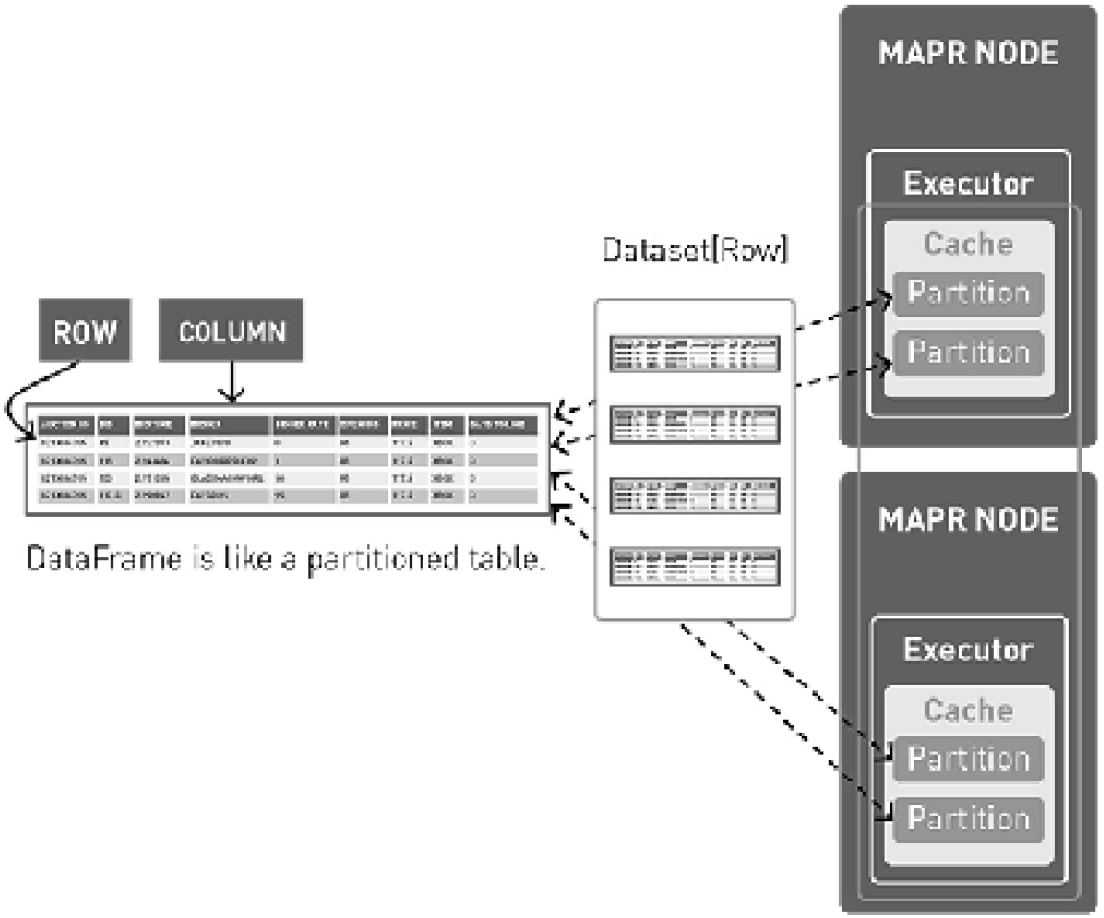
Hadoop

# Spark

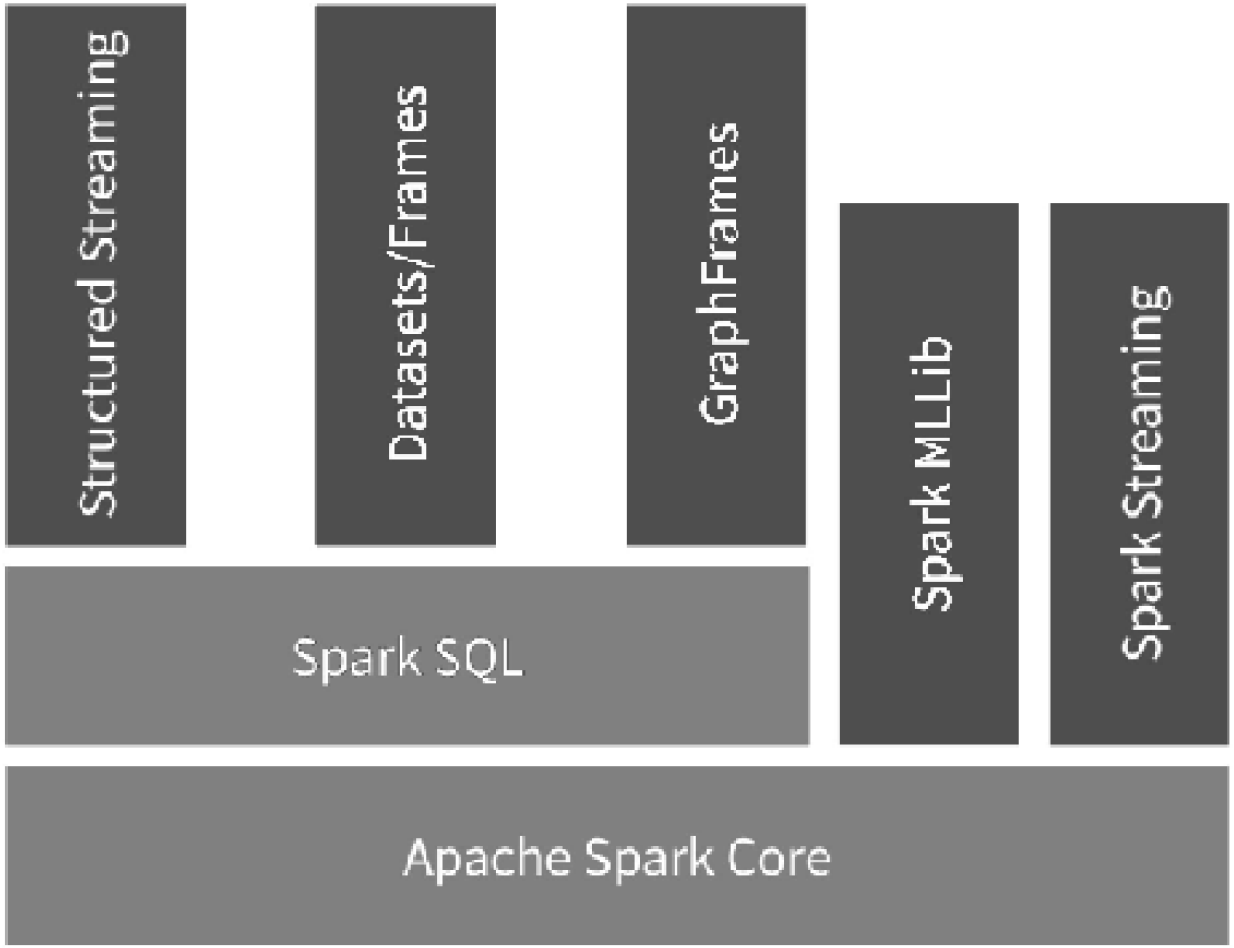
Zaharia et al., NSDI'12



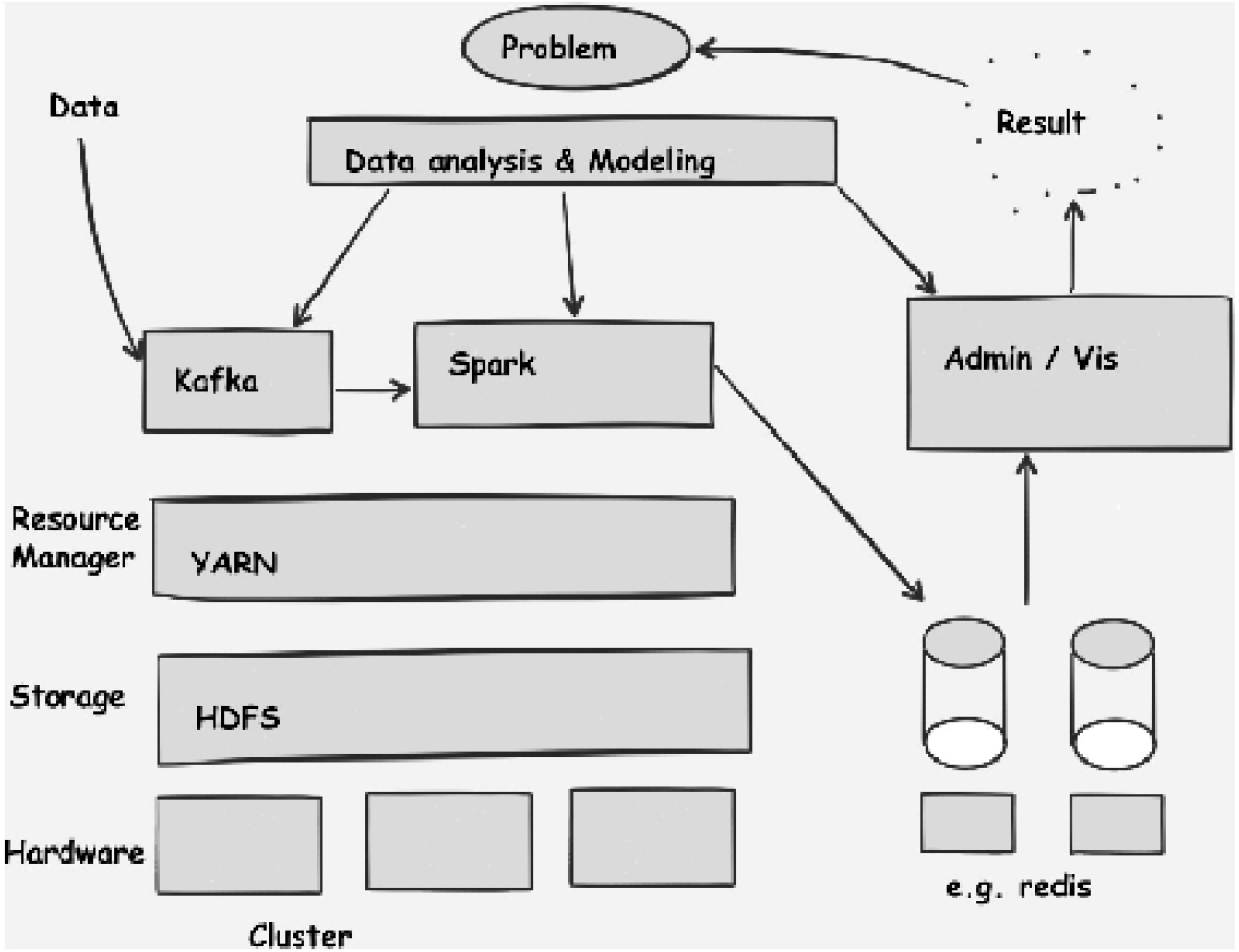
# Spark SQL


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

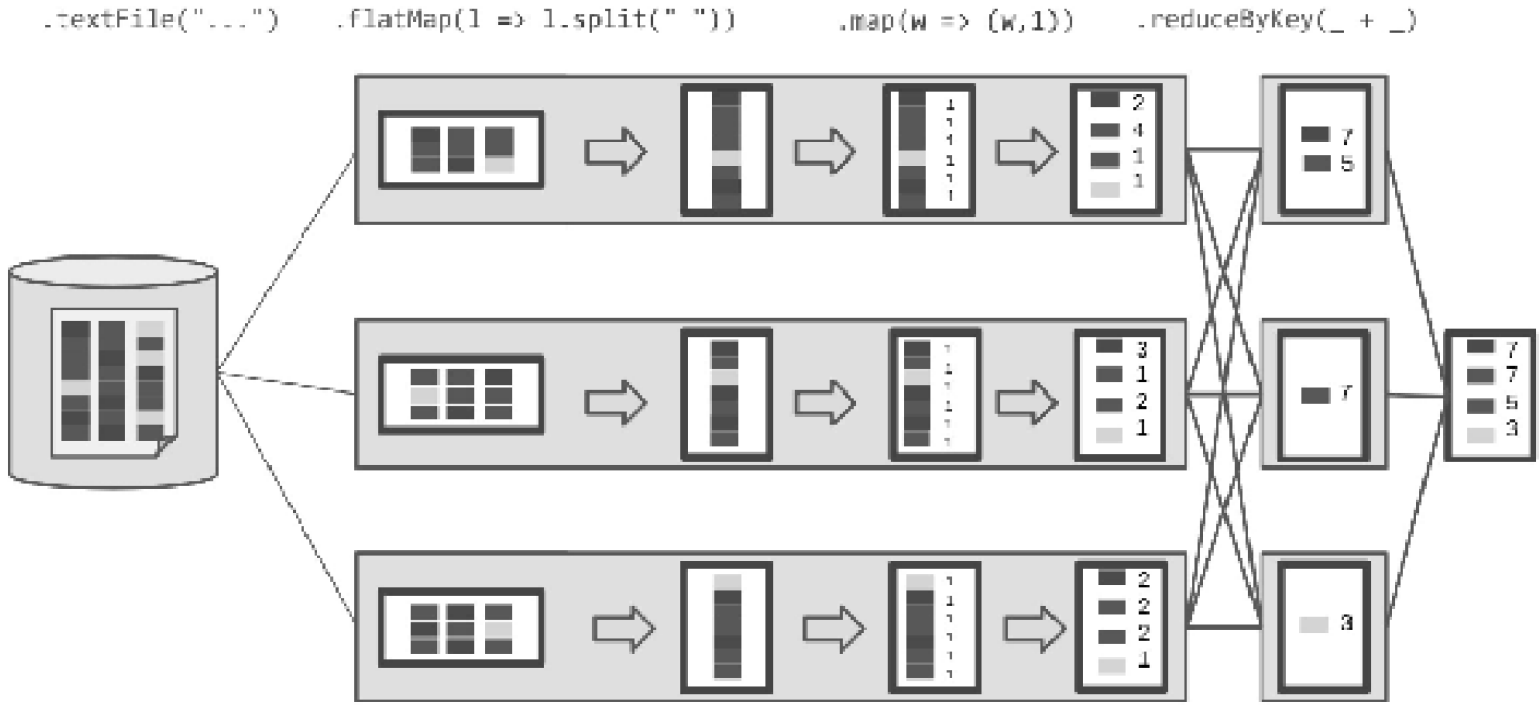
# Structura Spark



# O posibilă platformă de date bazată pe Spark

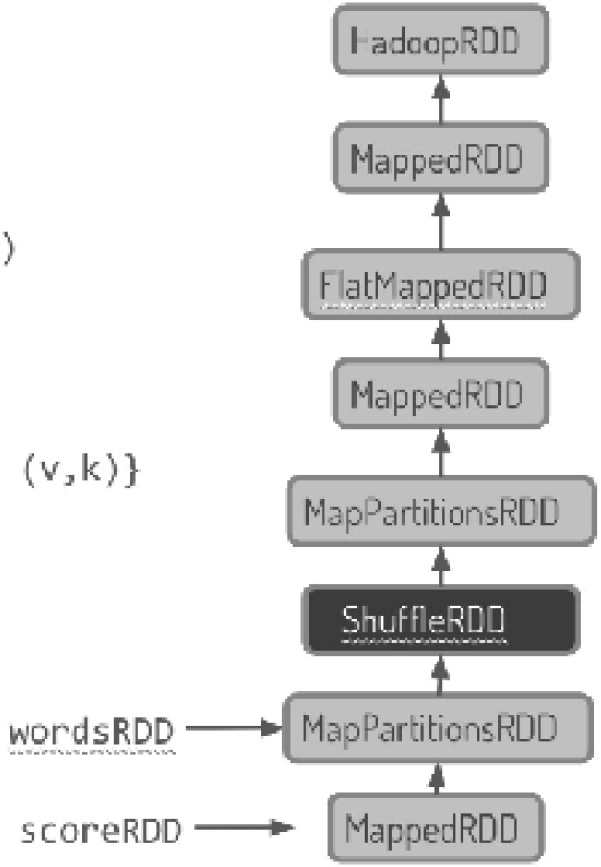


# RDD

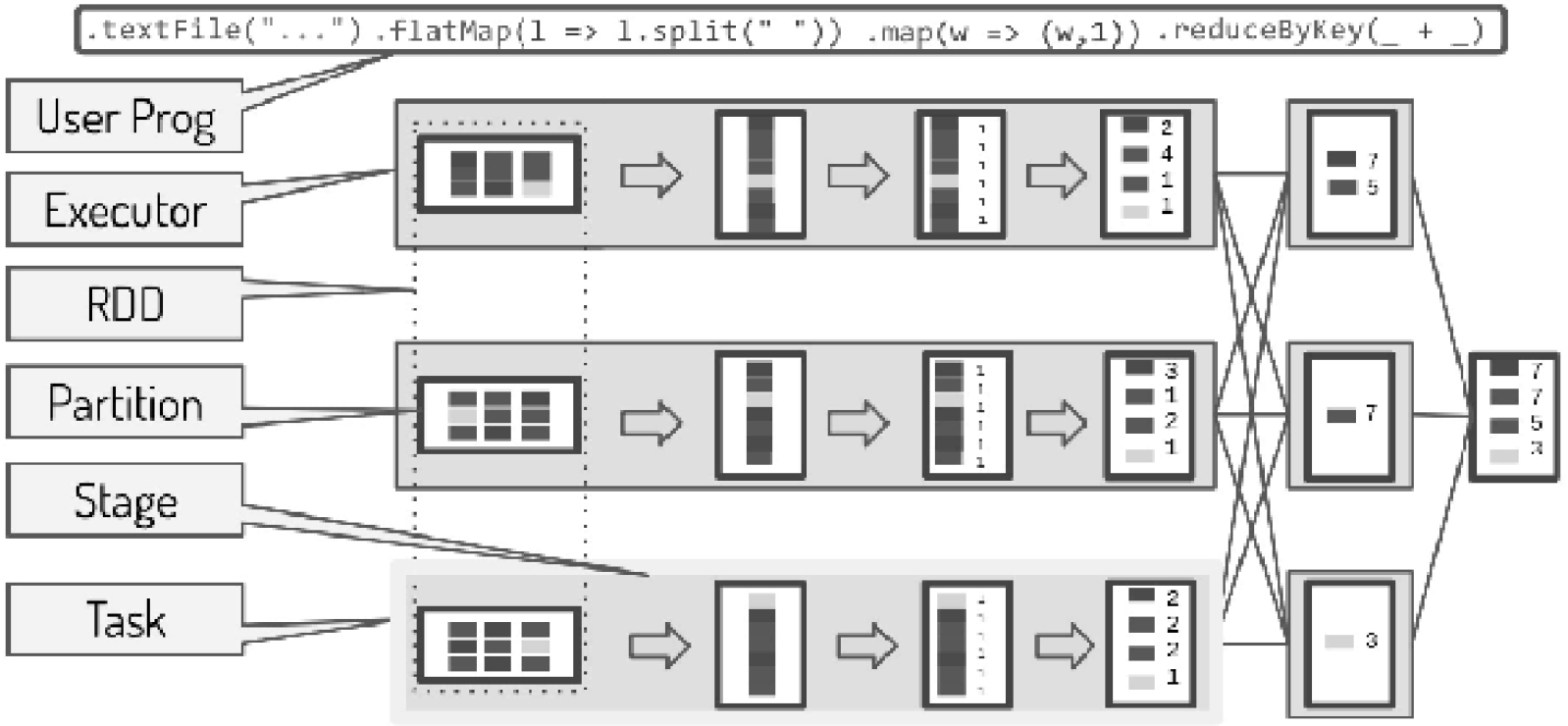


# RDD & DAG

```
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)}
```

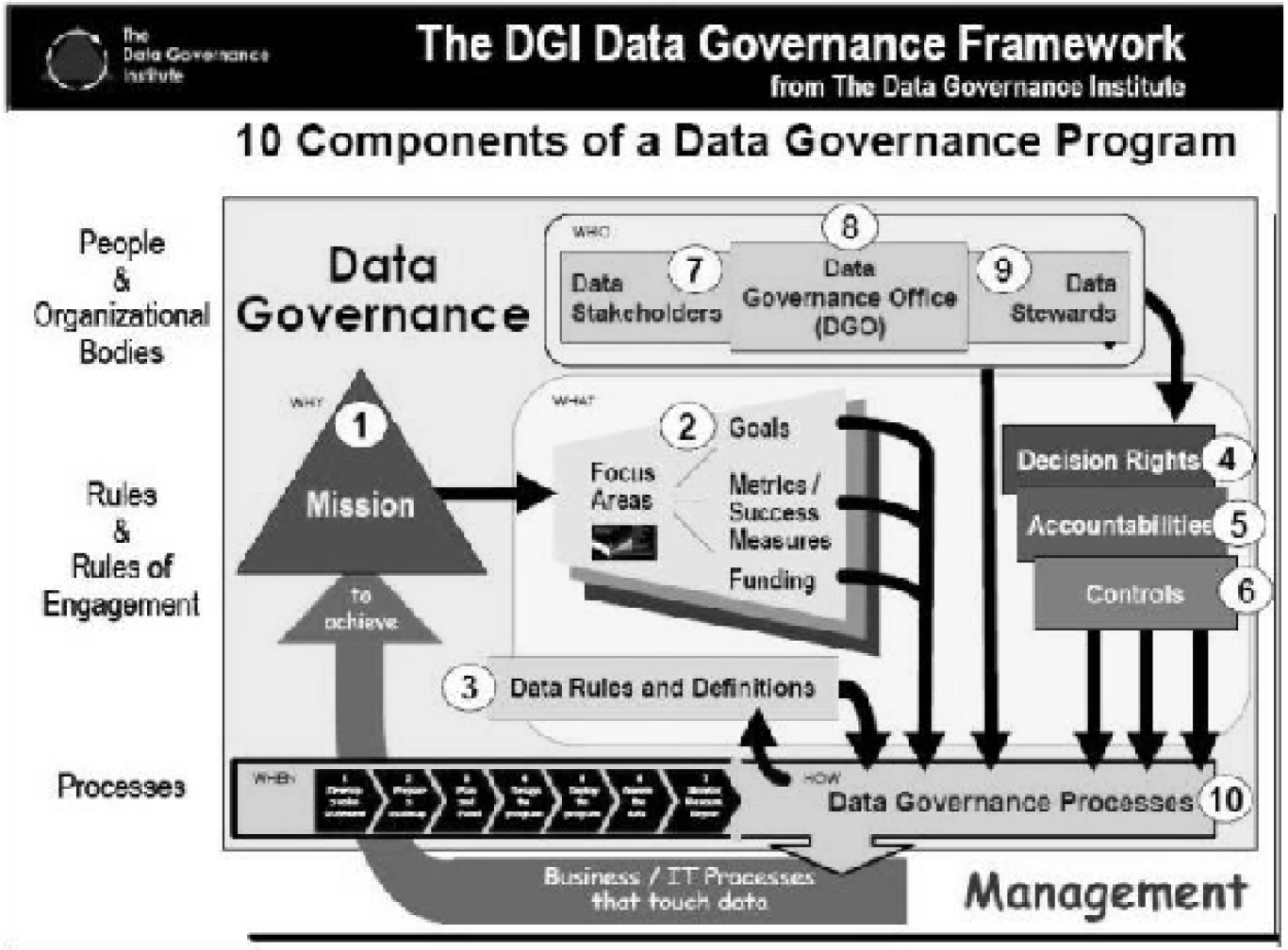


# Spark Components




- X

# Guvernarea datelor



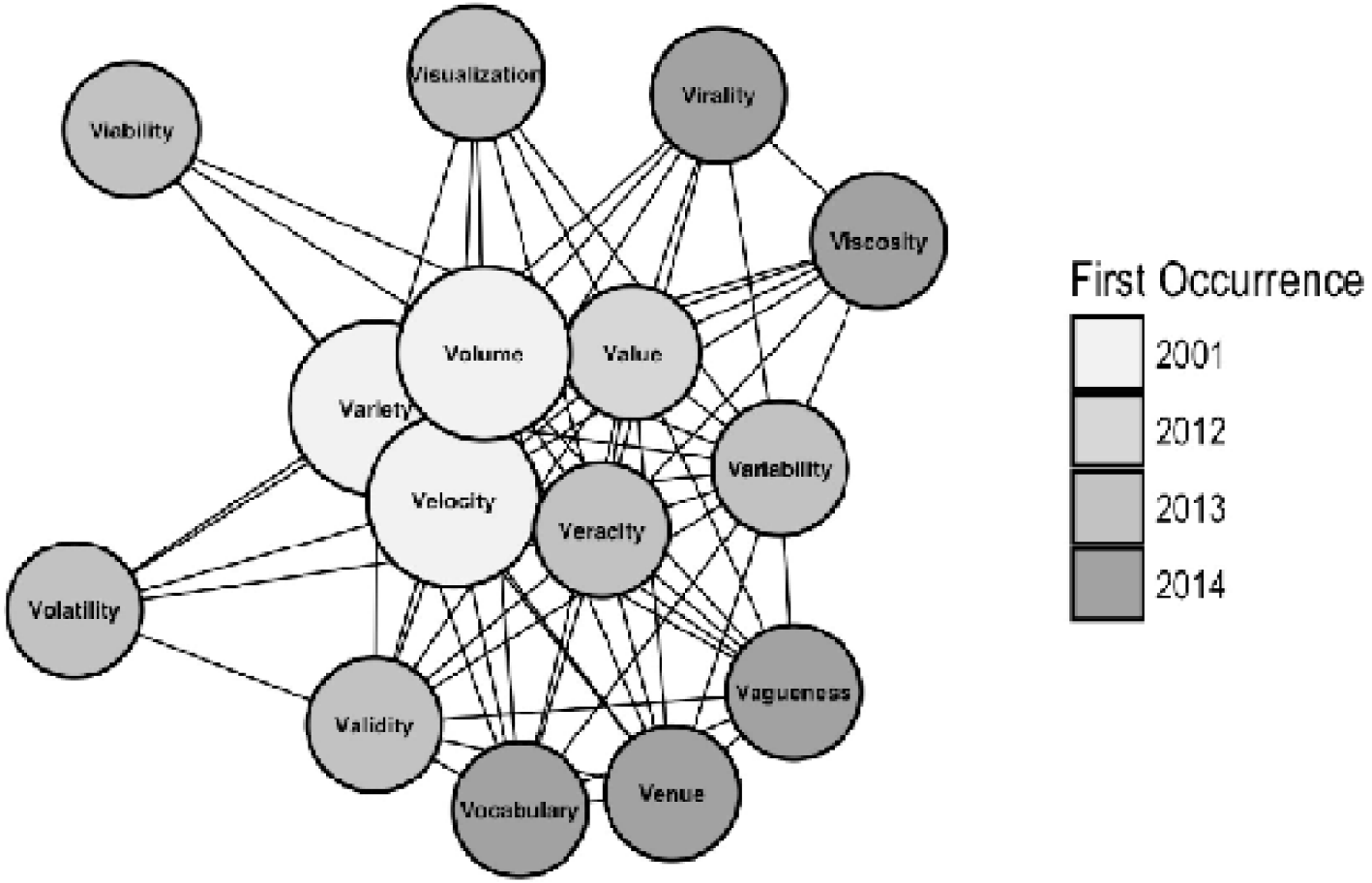

- X

# Big Data

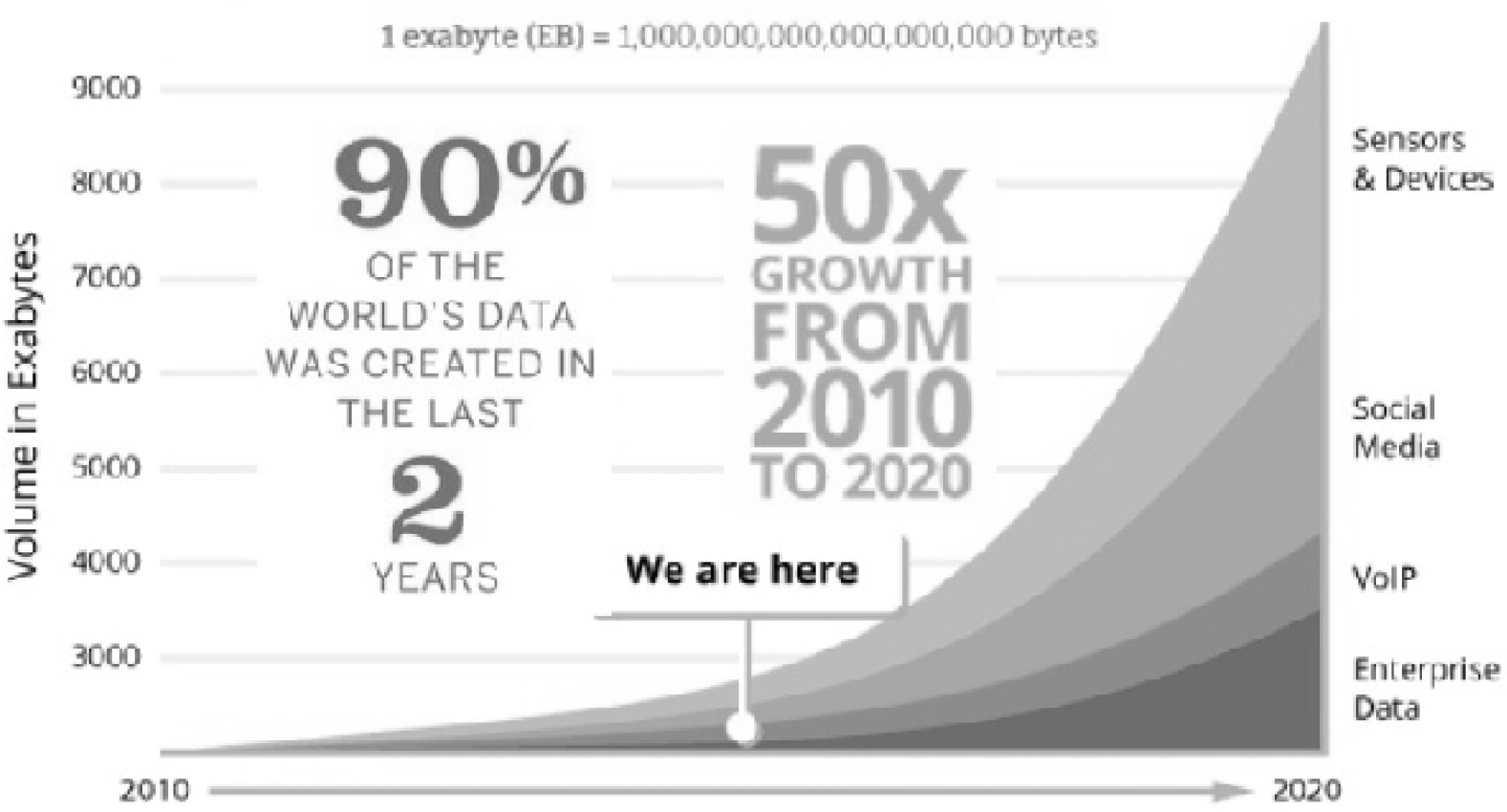
- *"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even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

Cox si Ellsworth, IEEE, 1997

# Dimensiunile blg data

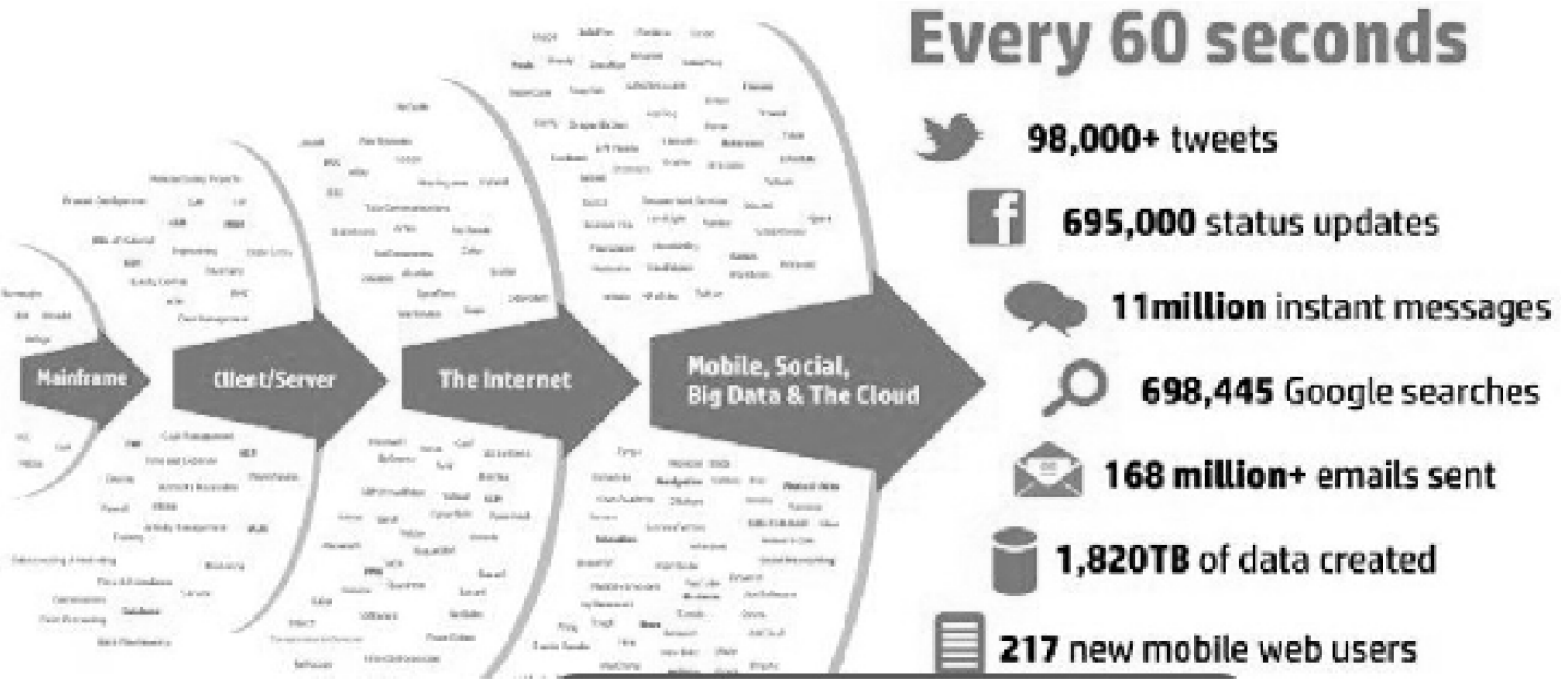


# Dimensiunile Big Data 1. Volumul



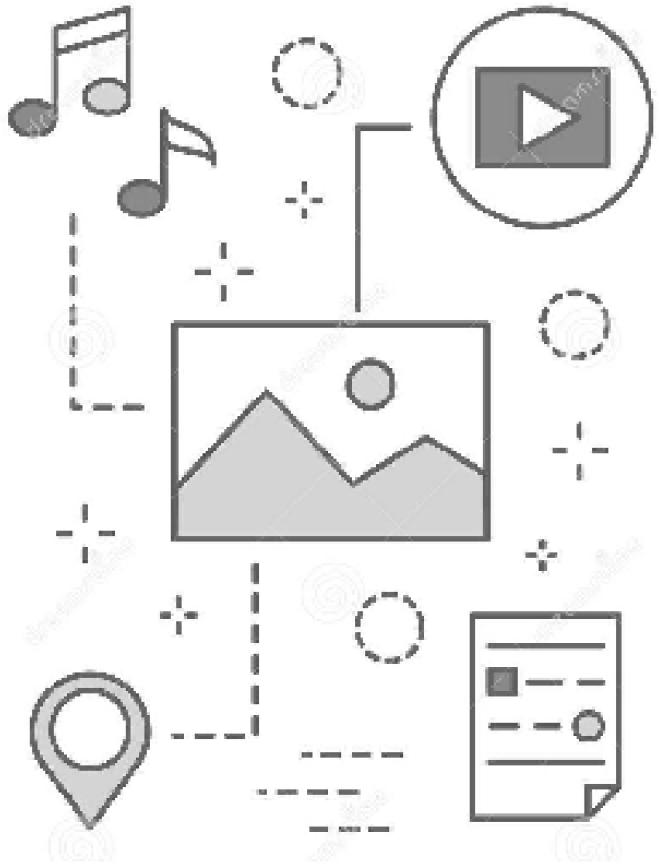

creștere exponențială a datelor generate

# Dimensiunile Big Data 2. Velocitate



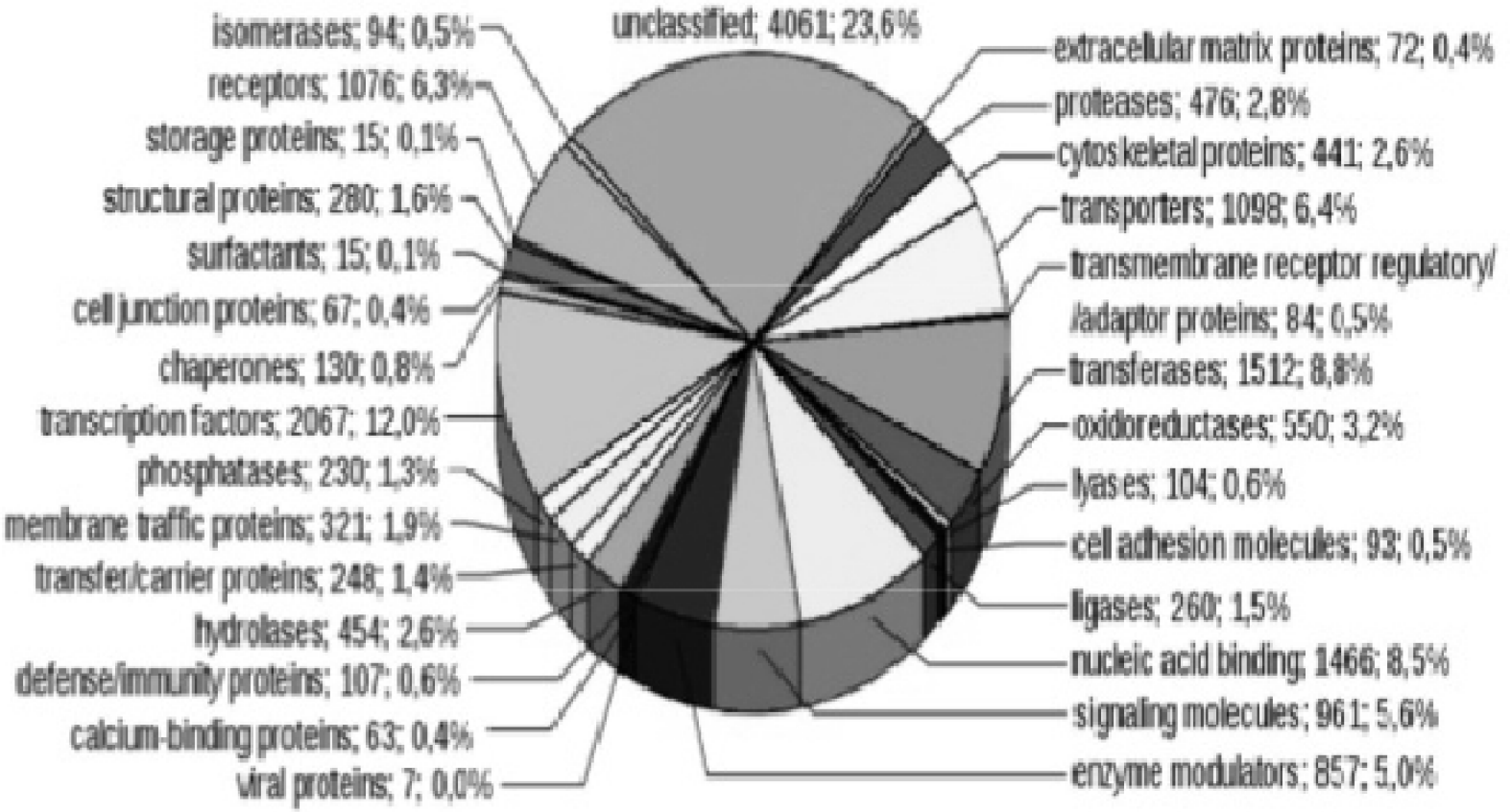

viteza de producție a noilor date

# Dimensiunile Big Data 3. Varietate



- Datele structurate
- Datele semistructurate
- Date nestructurate

# Dimensiunile Big Data 4. Variabilitate

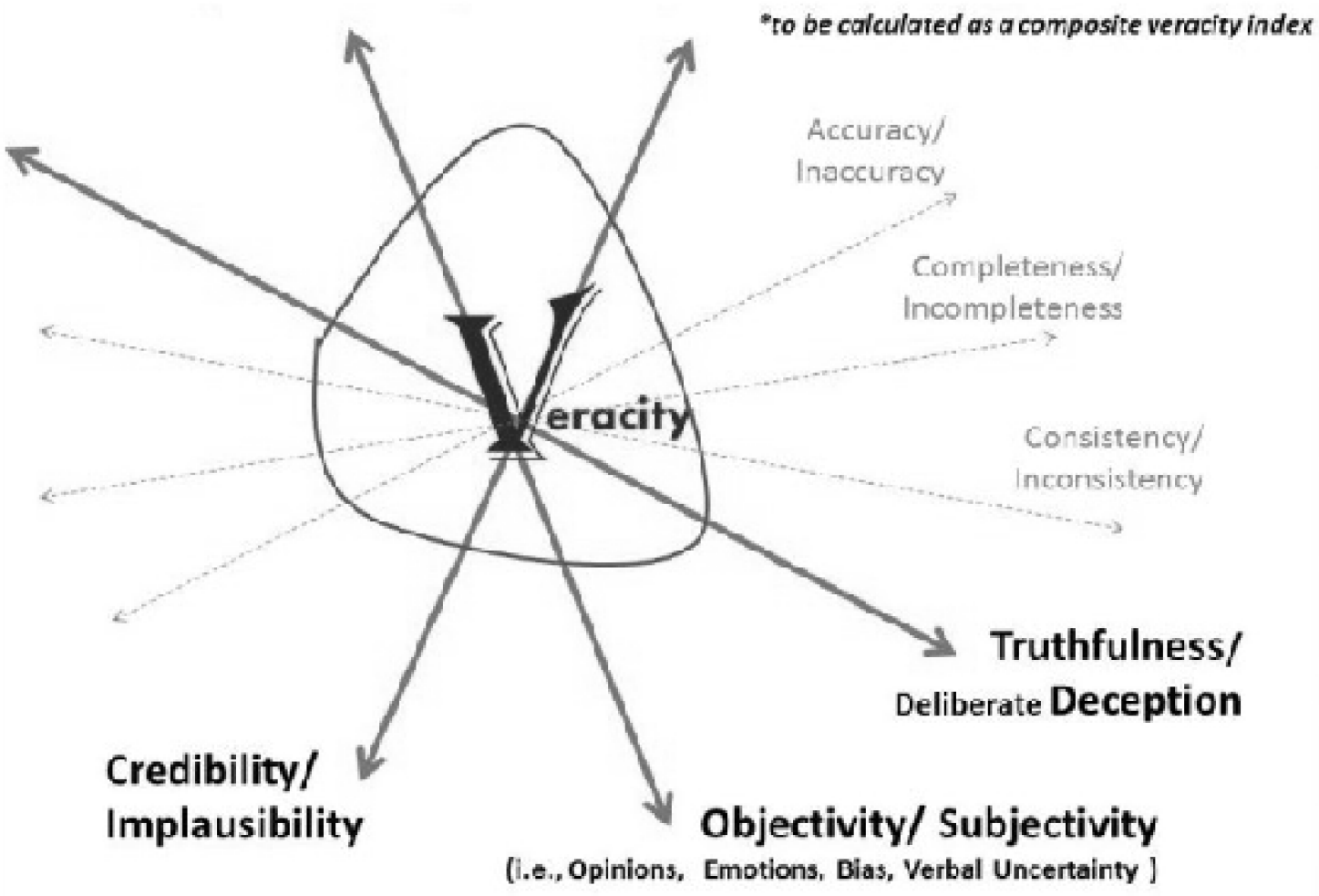



Functions of 17,209 Genes

Neomogenitatea sau

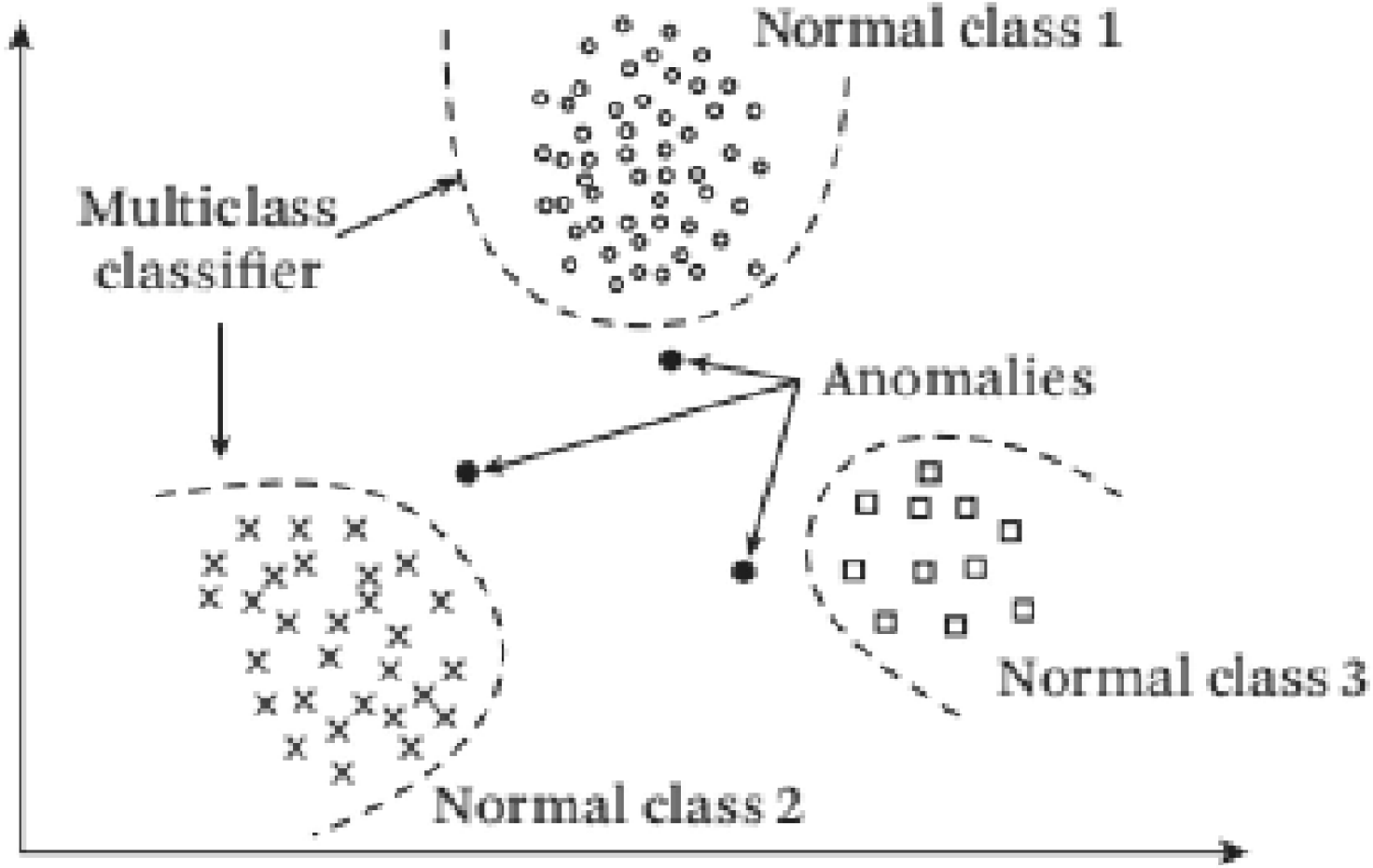
Variatia dimensională și compozițională

# Dimensiunile Big Data 5. Veracitate



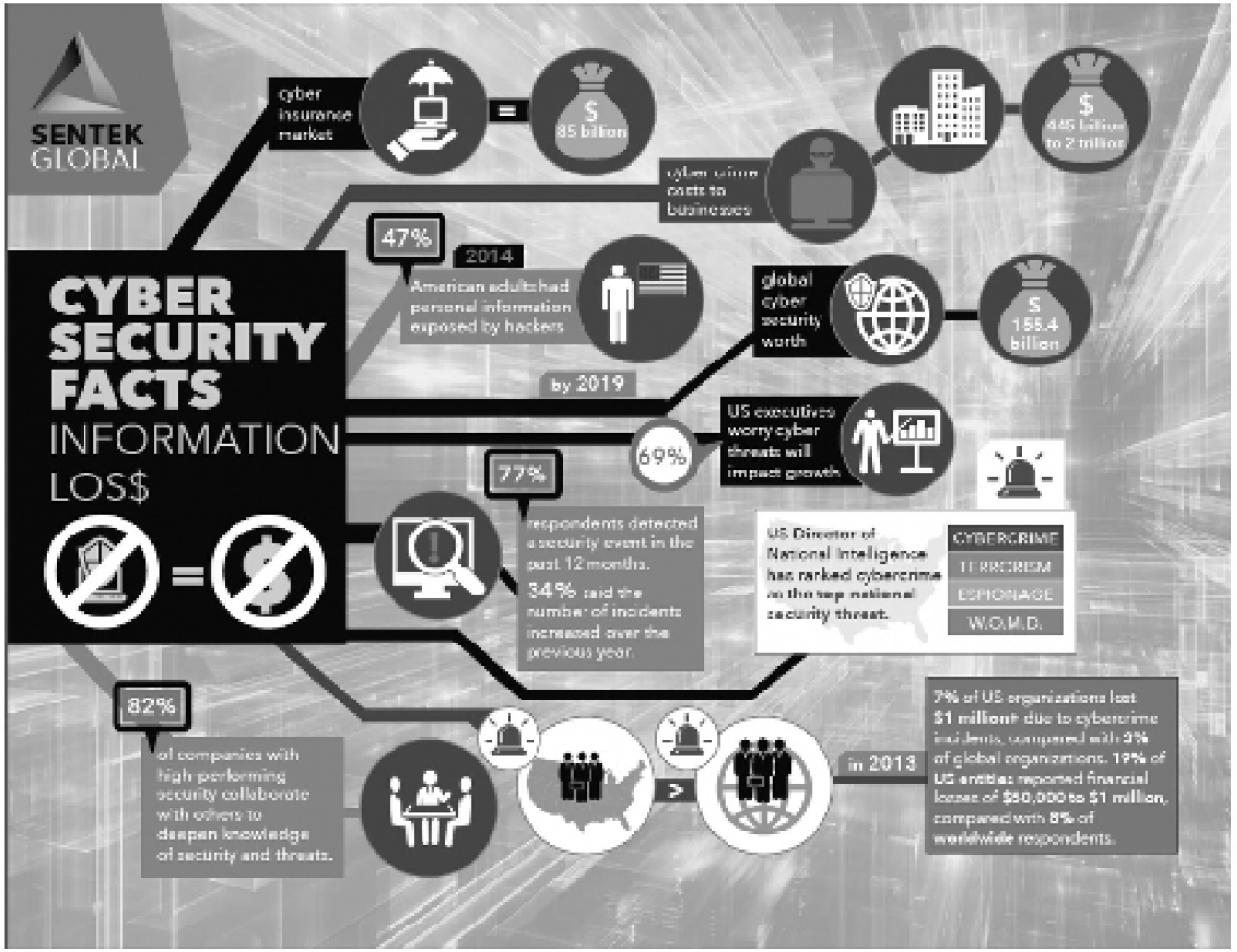

veridicitate sau corectitudine

# Dimensiunile Big Data 6. Validitatea




Detectarea anomaliilor după gruparea pe categorii

# Dimensiunile Big Data 7. Vulnerabilitate




după Forbes