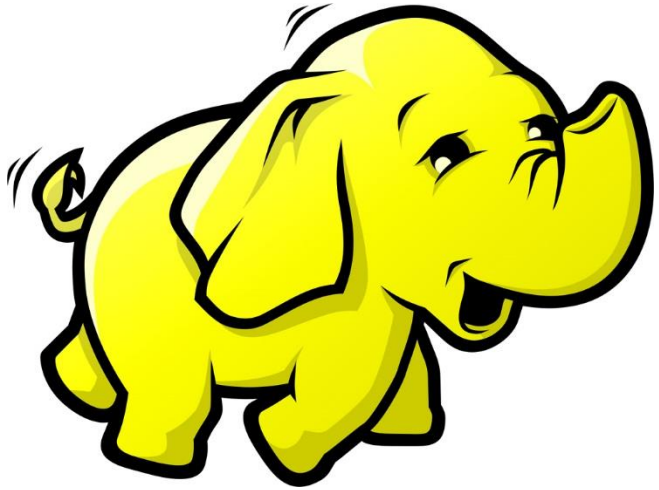
HADOOP și MAPREDUCE – SOLUȚIE PENTRU PROCESAREA DISTRIBUITĂ A VOLUMELOR MARI DE DATE ÎN REȚELE DE CALCULATOARE



Claudia Chitu

Master IISC, an 2

2015

Agenda

- **Decriere subiect: de ce Hadoop și MapReduce**
- HDFS (arhitectură și comparație cu NFS)
- MapReduce prezentare
- Exemplu
- Concluzii

De ce Hadoop și MapReduce

- Hadoop
  - Framework foarte răspândit pentru Cloud Computing
  - Procesare distribuită a volumelor mari de date
  - Folosește paralelizarea operațiilor
  - Este scalabil, este fezabil (gestionează dezastrele foarte bine)
- Componență
  - Pachetul Common
  - HDFS
  - MapReduce

De ce Hadoop și MapReduce

- Pachetul Common
  - Biblioteci
  - Utilitare

- HDFS (Hadoop Distributed File System)
  - Sistem de fișiere scris în Java
  - Cuprinde:
    - NodNume
    - Cluster de noduri de date (NodDate)

De ce Hadoop și MapReduce

- MapReduce
  - Scop: împărțirea seturilor de date de intrare în seturi independente pentru procesare paralelă
  - Roluri:
    - Programarea sarcinilor de lucru
    - Monitorizarea sarcinilor de lucru
    - Reexecutarea în cazul eșuării operațiilor de lucru

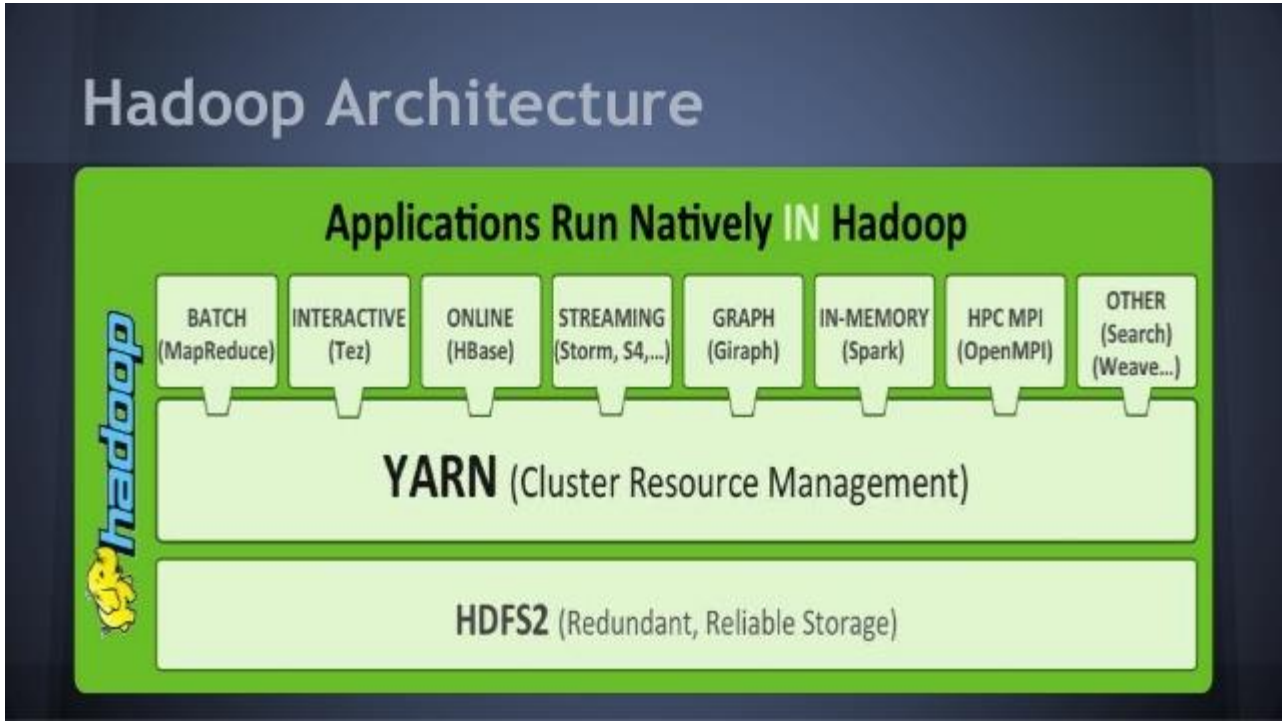
De ce Hadoop și MapReduce

- Comparație între Hadoop și Spark
  - Hadoop:
    - Este superior Spark pentru sarcini de ETL
    - MapReduce este specific unor situații de lucru
    - Este folosit pentru procesare de volume mari de date
  - Spark:
    - Motor rapid și **general** pentru Big Data
    - Generalizează modelul MapReduce
    - Este superior Hadoop pentru sarcini de machine learning

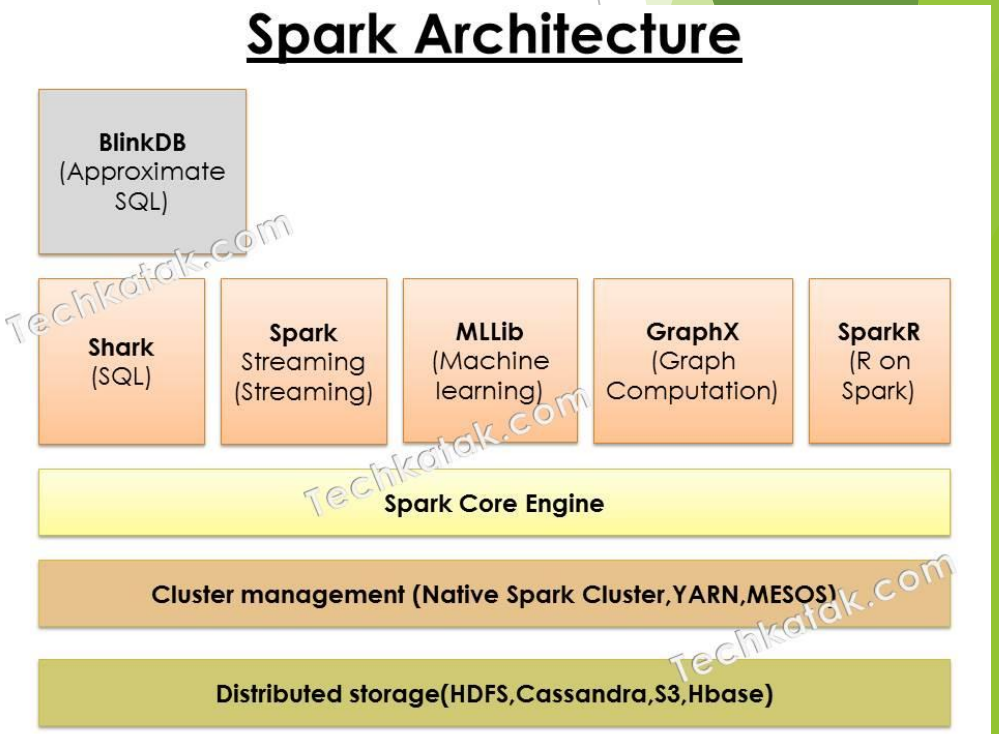
De ce Hadoop și MapReduce

- Spark:
  - RDDs(resilient distributed datasets)
    - Colecții distribuite de obiecte pot fi păstrate în memorie (cache) pe nodurile din clustere
    - Ele pot fi manipulate cu operatori paraleli
    - În caz de eșec, se reconstruiește automat
  - Interfața: API Scala, Scala console
  - Berkeley, Princeton, Yahoo research sunt utilizatori de Spark
  - Spark este mai potrivit pentru programatori foarte experimentați, iar Hadoop pentru utilizatori de SQL.

De ce Hadoop și MapReduce



http://image.slidesharecdn.com/dinatechtalk-150201035119-conversion-gate01/95/josa-techtalks-big-data-on-hadoop-7-638.jpg?cb=1440319902



http://techkatak.com/spark-for-fast-batch-processing/

HDFS (Hadoop Distributed File System)

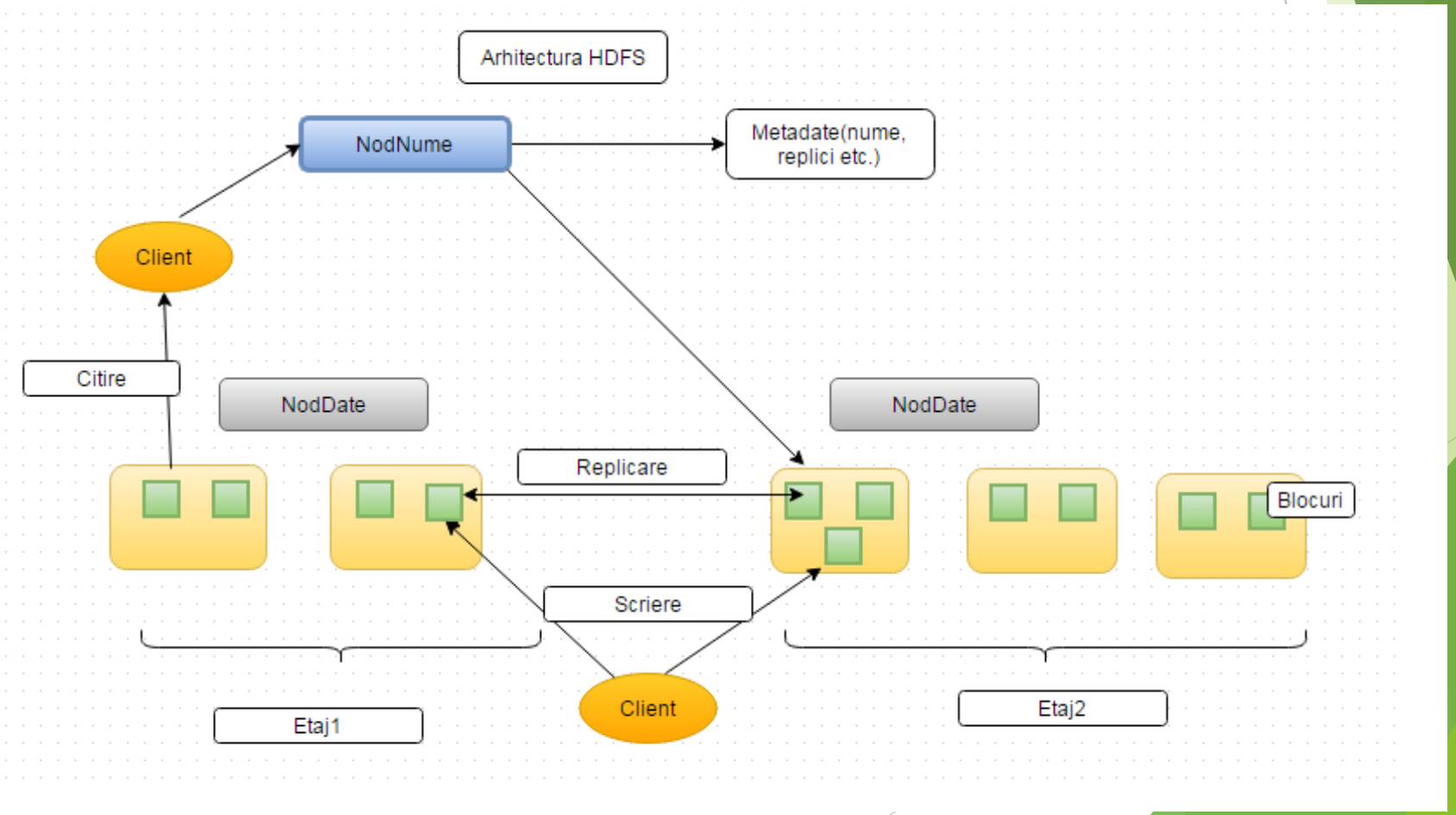
- Caracteristici
  - Se cunosc informații despre server = > reducere trafic în rețea
  - Gestionare de tip FIFO (sau Fair scheduler folosit de Facebook)
  - Replicarea:
    - Replicile sunt stocate separat pentru o disponibilitate mai mare
    - Timp de execuție mai bun prin arhitectura sa
    - Toleranță la erori

HDFS (Hadoop Distributed File System)

- Structură
  - Prezintă checkpointuri (protejează metadatele fișierelor de sistem)
  - Ierarhie de fișiere și foldere reprezentate de inodes
    - Conțin atribute de permisiuni, modificări, timp de acces, spațiu pe disc
    - Împărțire pe blocuri a fișierelor (128 megabytes)
    - Blocurile sunt replicate în datanoduri

HDFS (Hadoop Distributed File System)

- Arhitectura
  - De tip master/sclav



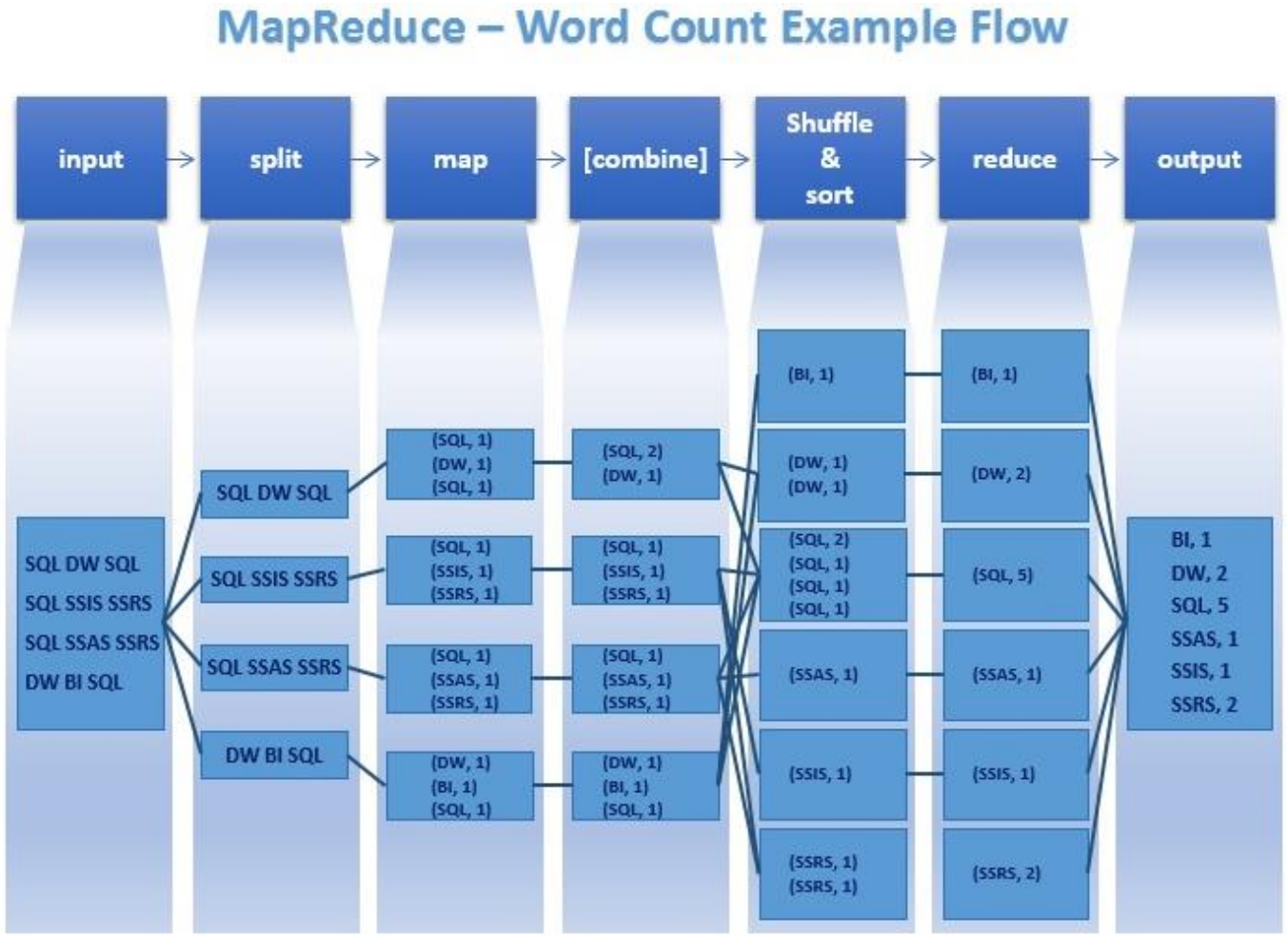
HDFS (Hadoop Distributed File System)

- Comparație HDFS cu NFS
  - HDFS:
    - Stochează un volum mare de date
    - E scalabil
    - Nu este foarte genral
    - Proiectat să optimizeze performanța fluxurilor de citire
  - NFS:
    - Cel mai răspândit sistem de fișiere
    - Funizează acces la un volum de date stocat pe o singură mașină
    - Putere de procesare destul de limitată(o singură mașină)
    - Transparență

MapReduce prezentare

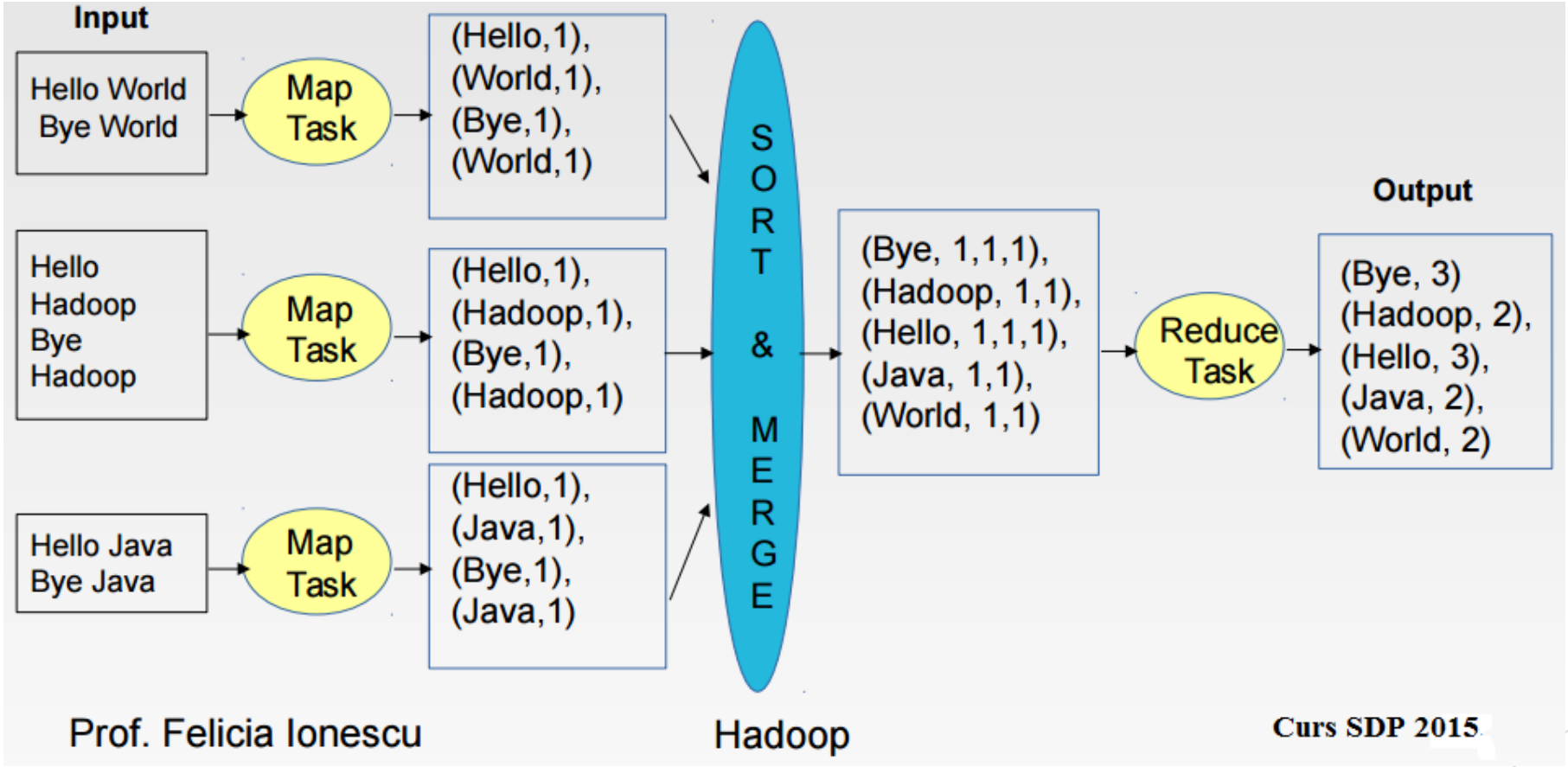
- Descriere
  - Model de programare implementat pentru procesarea datelor
  - Constă în două transformări: **Map** și **Reduce**
  - **Aplicații:**
    - Sortare distribită
    - Descompunere în valori singulare
    - Clasificare de documente
    - Machine learning
    - Statistical machine translation
    - Indexare inversă
    - Web link graph reversal

MapReduce prezentare

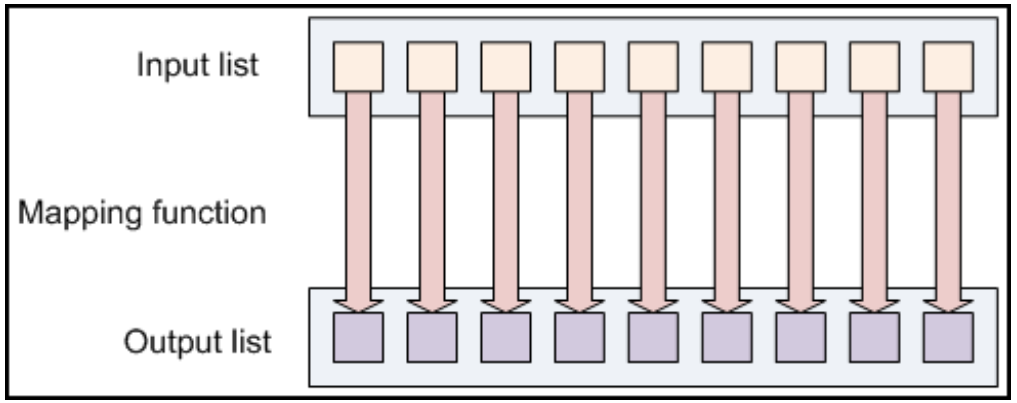


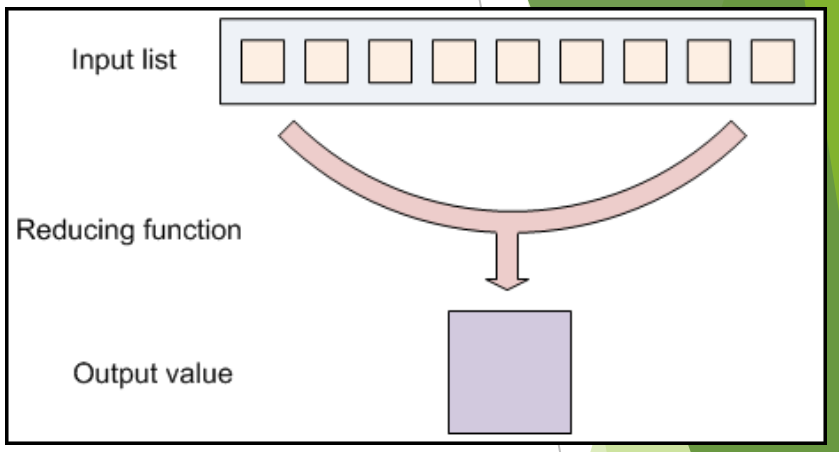
Sursa: https://www.mssqltips.com/sqlservertip/3222/big-data-basics--part-5--introduction-to-mapreduce/

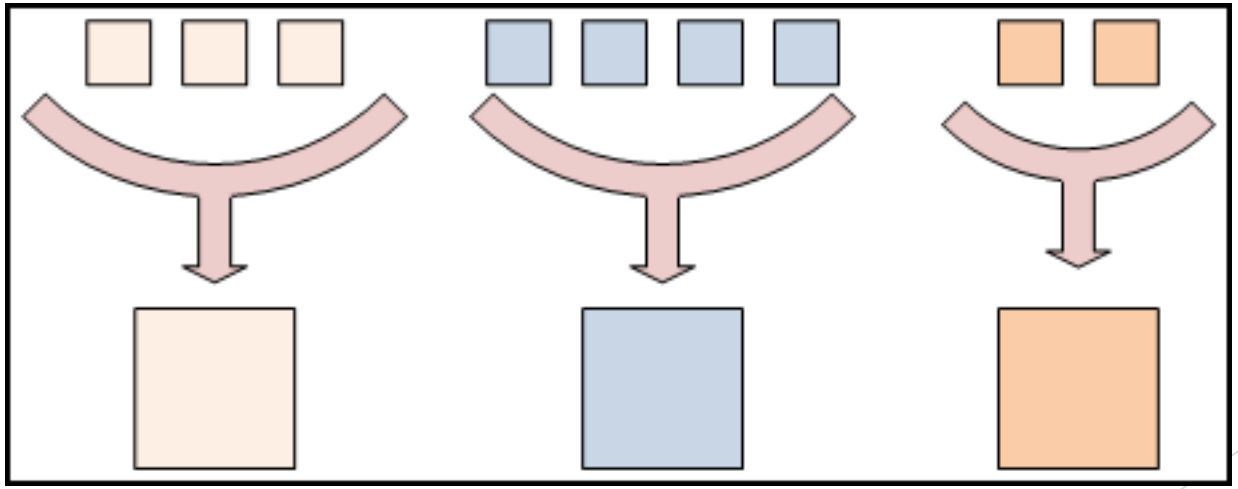
Exemplu


Input
Hello World
Bye World
Hello
Hadoop
Bye
Hadoop
Hello Java
Bye Java
Map Task
Map Task
Map Task
(Hello,1),
(World,1),
(Bye,1),
(World,1)
(Hello,1),
(Hadoop,1),
(Bye,1),
(Hadoop,1)
(Hello,1),
(Java,1),
(Bye,1),
(Java,1)
SORT & MERGE
(Bye, 1,1,1),
(Hadoop, 1,1),
(Hello, 1,1,1),
(Java, 1,1),
(World, 1,1)
Reduce Task
Output
(Bye, 3)
(Hadoop, 2),
(Hello, 3),
(Java, 2),
(World, 2)
Prof. Felicia Ionescu
Hadoop
Curs SDP 2015

Exemplu







Sursa : https://developer.yahoo.com/hadoop/tutorial/module4.html

Concluzii

- Hadoop este un mediu de lucru gratuit foarte răspândit folosit pentru Cloud Computing
- Este folosit pentru procesarea distribuită a volumelor mari de date (petabytes de date)
- Datele sunt împărțite pe mii de mașini și procesate în paralel
- Hadoop prezintă o scalabilitate foarte mare
- Este foarte cunoscut în lumea Big Data
- Soluția MapReduce nu este generală, ci specifică unor clase de aplicații, mai ales dedicate domeniului ETL