DISTRIBUTED FILE SYSTEM ΚΑΙ MAPREDUCE

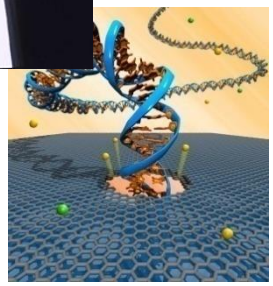
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Γεράσιμος Ραζής (razis@uth.gr)

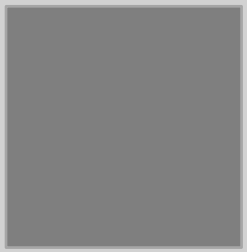
Big Data

- 90% των σημερινών δεδομένων δημιουργήθηκαν τα τελευταία 2 χρόνια
- Νόμος του Moore: Διπλασιασμός δεδομένων κάθε 18 μήνες
- YouTube: 13 εκατ. ώρες και 700 δις αναπαραγωγές το 2010
- Facebook: 10TB/ημέρα συμπιεσμένα
- CERN/LHC: 40TB/ημέρα (15PB/έτος)
- Πολλά, πολλά ακό
- Web logs, αρχεία



Πρόβλημα: έκρηξη δεδομένων

1 EB (Exabyte=10^{18}bytes) = 1000 PB (Petabyte=10^{15}bytes)
Κίνηση δεδομένων κινητής τηλεφωνίας στις ΗΠΑ για το 2010

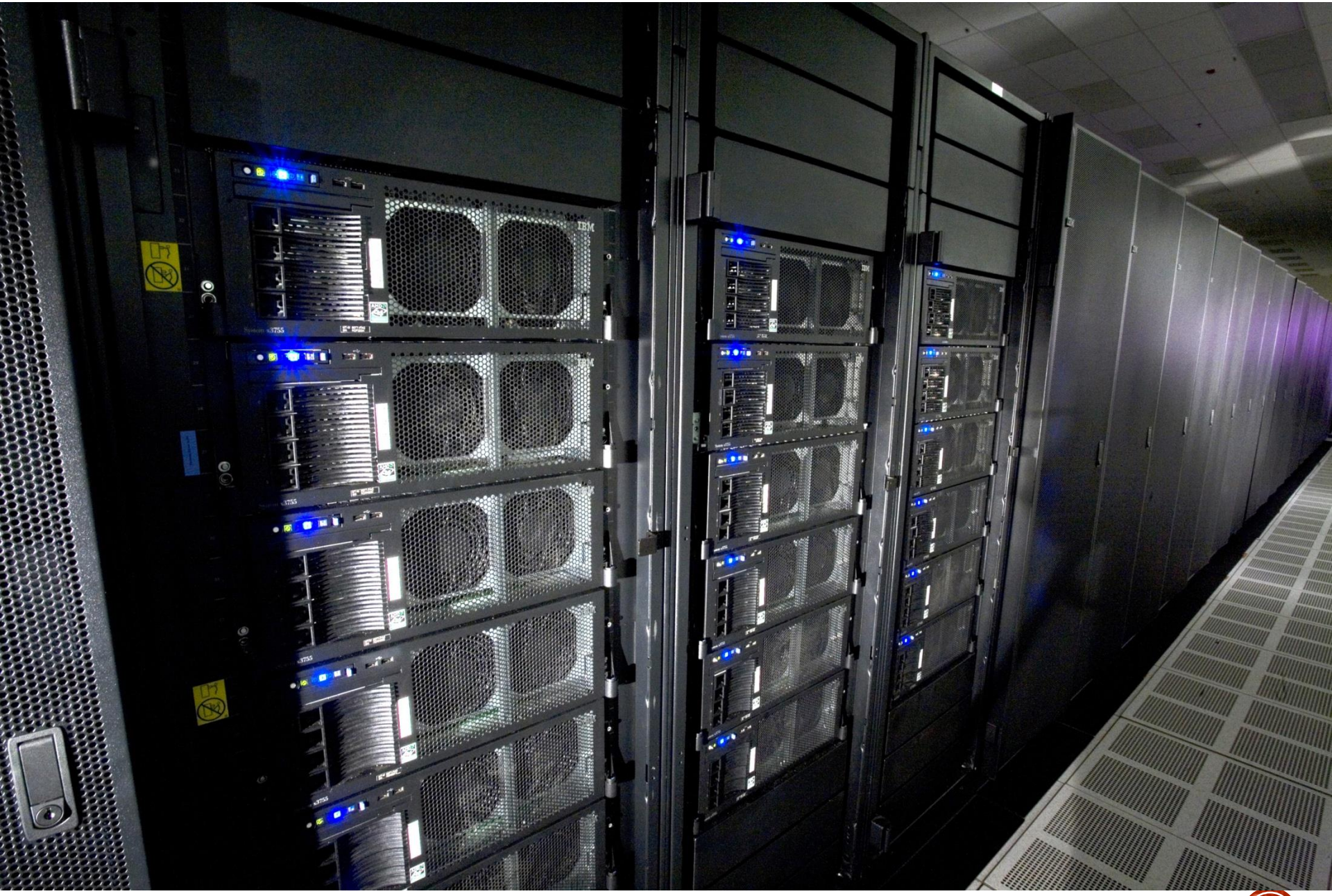


1.2 ZB (Zettabyte) = 1200 EB
Σύνολο ψηφιακών δεδομένων το 2010

75!!! ZB (ZettaByte = 10^{21} bytes)
Εκτίμηση για σύνολο ψηφιακών δεδομένων το 2020

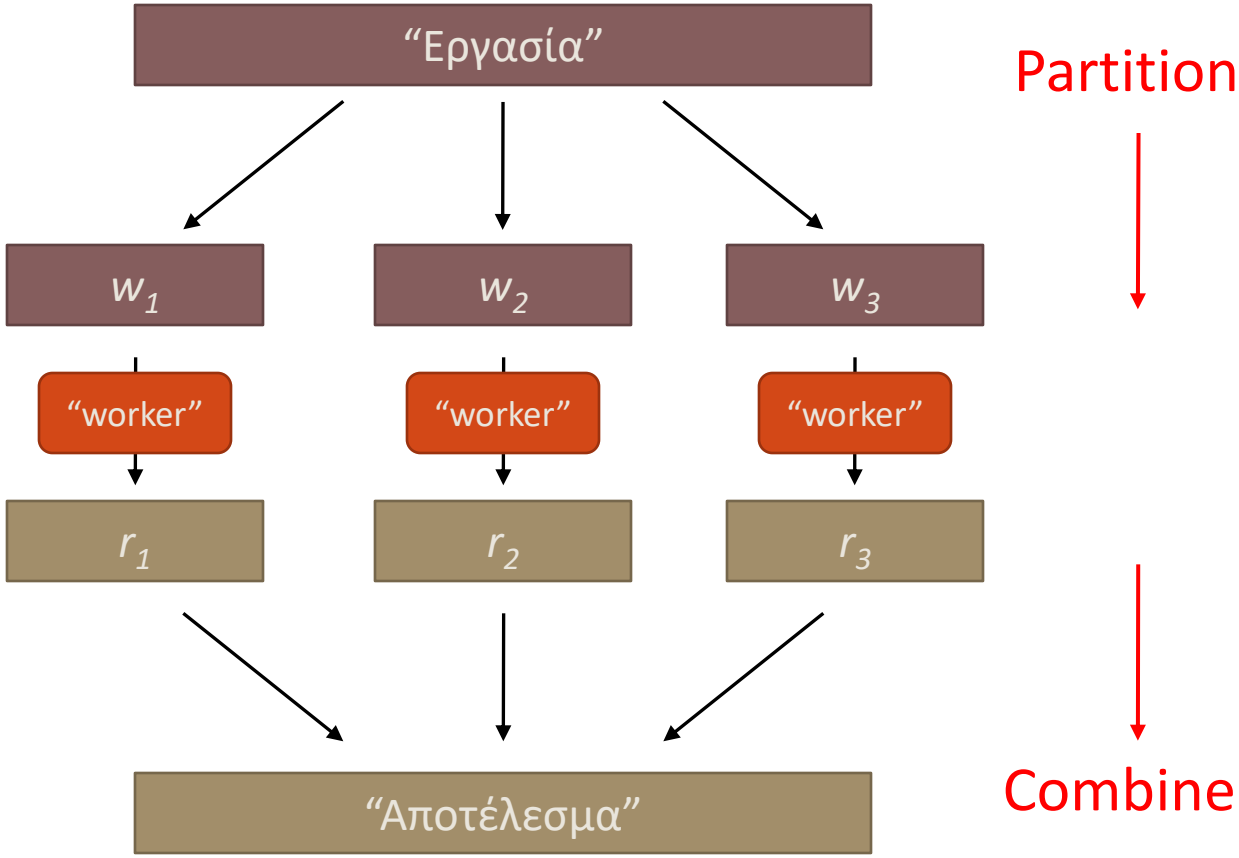
Λύση: Κλιμακωσιμότητα (Scalability)

Πώς;



Source: Wikipedia (IBM Roadrunner)

Διαίρει και Βασίλευε (Divide and Conquer αγγλιστί!)



Προκλήσεις Παραλληλοποίησης

- Πώς αναθέτουμε μονάδες εργασίας σε workers;
- Αν έχουμε περισσότερες μονάδες εργασίας από workers;
- Εάν οι workers χρειαστεί να μοιραστούν ενδιάμεσα ημιτελή δεδομένα;
- Πώς συνοψίζουμε τέτοιου είδους ενδιάμεσα δεδομένα;
- Πώς ξέρουμε ότι όλοι οι workers τελείωσαν;
- Τι γίνεται εάν κάποιοι workers διακόπηκαν;

Τι το κοινό έχουν όλα αυτά τα προβλήματα;

Συγχρονισμός

- Τα προβλήματα παραλληλοποίησης προκύπτουν από:
  - Επικοινωνία μεταξύ workers
  - Πρόσβαση σε κοινόχρηστους πόρους (πχ, δεδομένα)
- Επομένως χρειαζόμαστε μηχανισμούς συγχρονισμού

Source: Ricardo Guimarães Herrmann

Τι είναι το MapReduce;

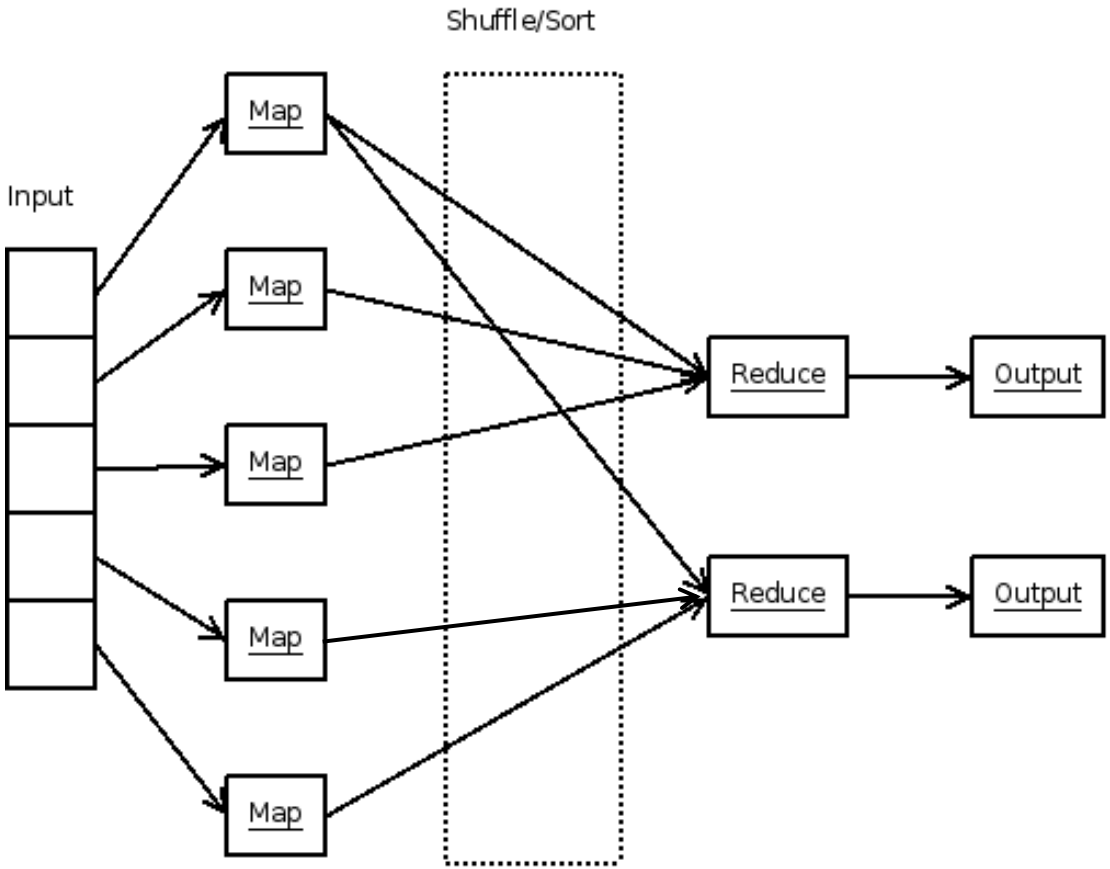
- Ένα προγραμματιστικό μοντέλο
- Για την ανάπτυξη εφαρμογών οι οποίες
  - επεξεργάζονται γρήγορα και παράλληλα τεράστιες ποσότητες δεδομένων
  - σε συστοιχίες (clusters) υπολογιστών

... κυρίως για non time critical προβλήματα !!

MapReduce

- Το πρόβλημα “σπάει” σε 2 φάσεις, την Map και την Reduce
- **Map**: Μη αλληλεπικαλυπτόμενα κομμάτια από δεδομένα εισόδου (εγγραφές <key,value>) ανατίθενται σε διαφορετικές διεργασίες (mappers) οι οποίες βγάζουν ένα σετ από ενδιάμεσα <key,value> αποτελέσματα
- **Reduce**: Τα δεδομένα της Map φάσης τροφοδοτούνται σε ένα συνήθως μικρότερο αριθμό διεργασιών (reducers) οι οποίες “συνοψίζουν” τα αποτελέσματα εισόδου σε μικρότερο αριθμό <key,value> εγγραφών

MapReduce


Shuffle/Sort
Input
Map
Map
Map
Map
Map
Reduce
Reduce
Output
Output

Πότε είναι χρήσιμο;

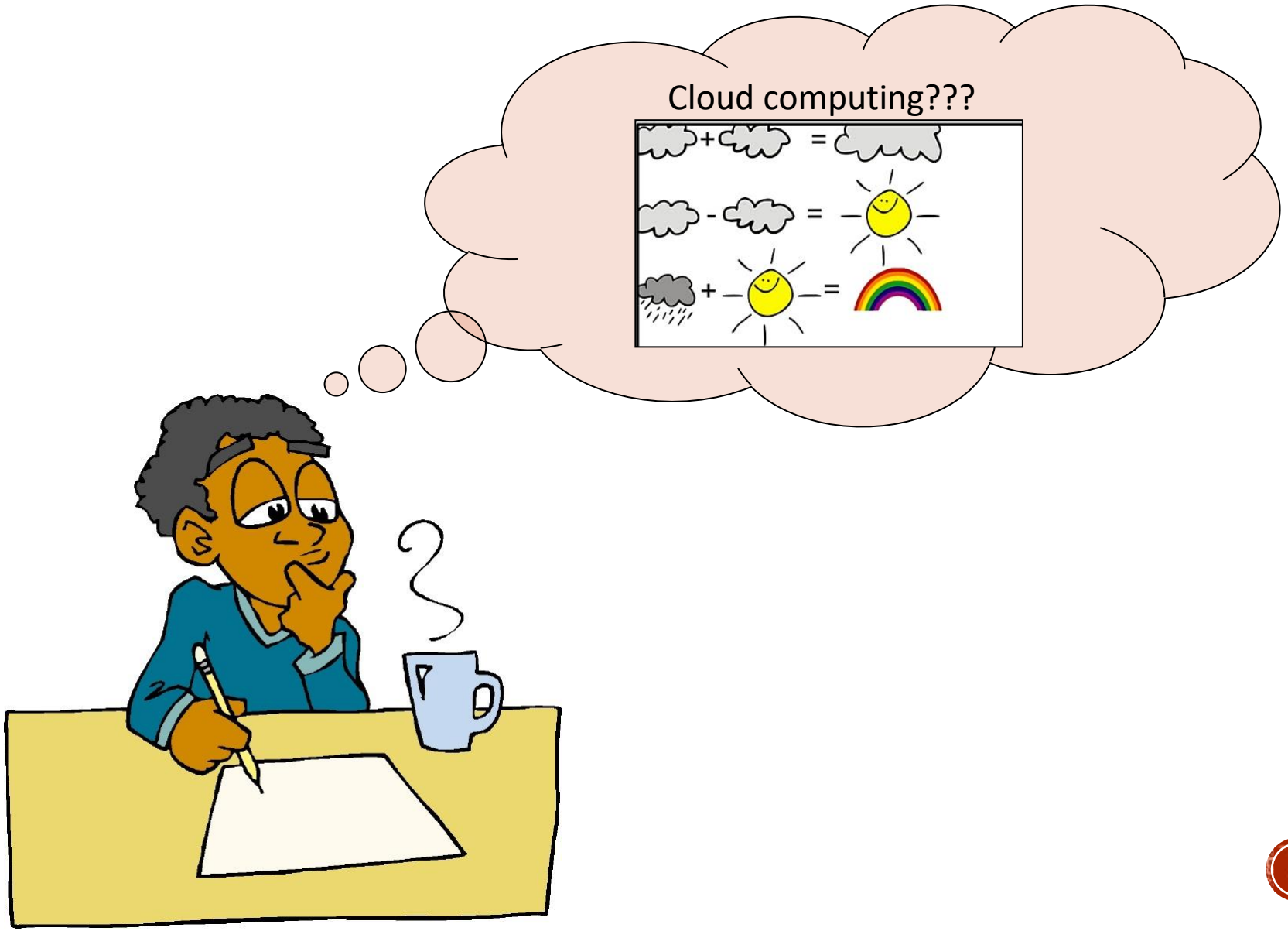
- Καλή επιλογή για:
  - Δεικτοδότηση/ανάλυση log αρχείων
  - Ταξινόμηση μεγάλου όγκου δεδομένων
  - Ανάλυση εικόνων

- Απαγορευτική επιλογή για:
  - Επεξεργασία real-time critical data (π.χ. ΣΑΕ για έλεγχο κίνησης ρομποτικού βραχίονα σε χειρουργική επέμβαση)

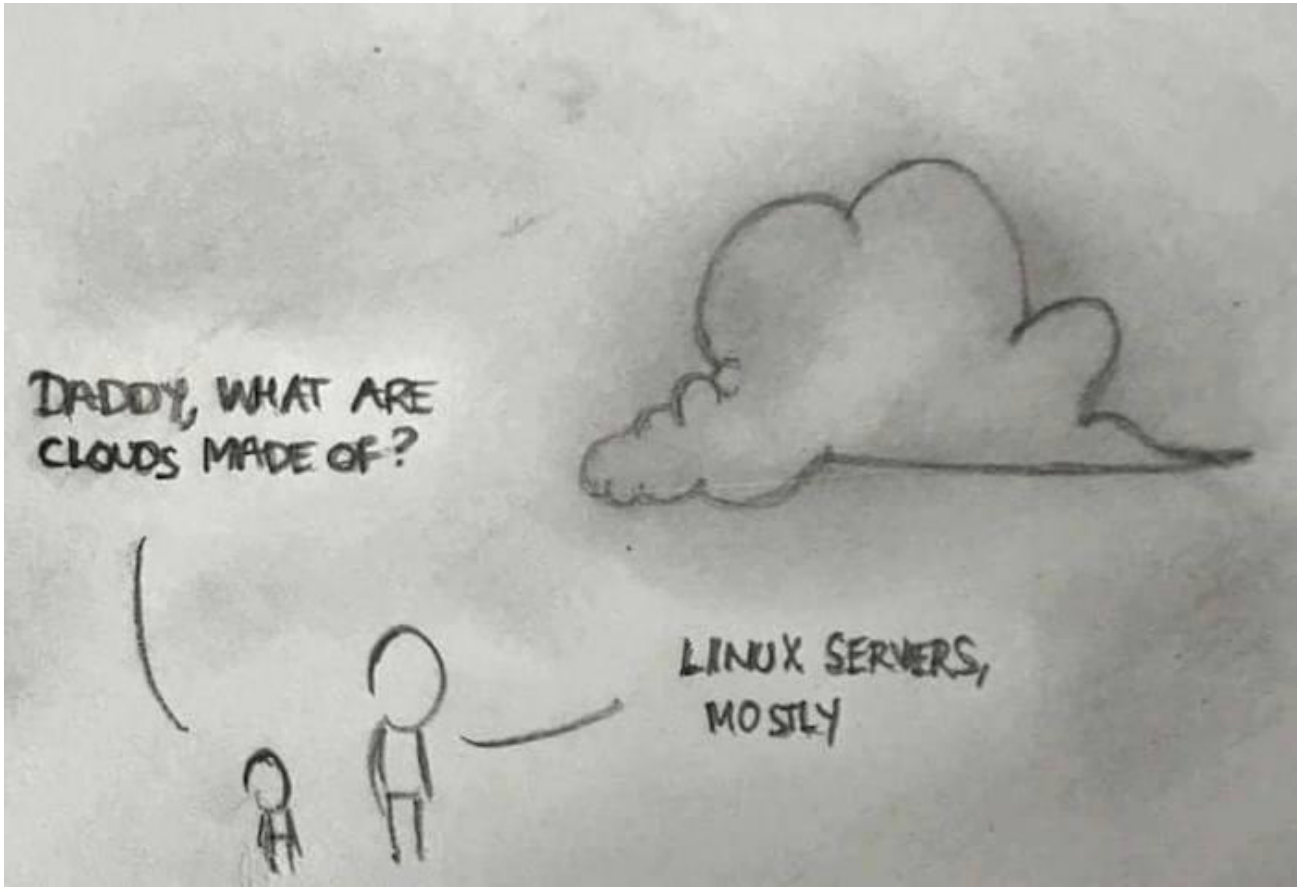
Τυπικές Εφαρμογές

- Log and/or clickstream analysis of various kinds
- Marketing analytics
- Machine learning and/or sophisticated data mining
- Image processing
- Processing of XML messages
- Web crawling and/or text processing
- General archiving, including of relational/tabular data

Τυπικές εφαρμογές … και Cloud



Cloud services



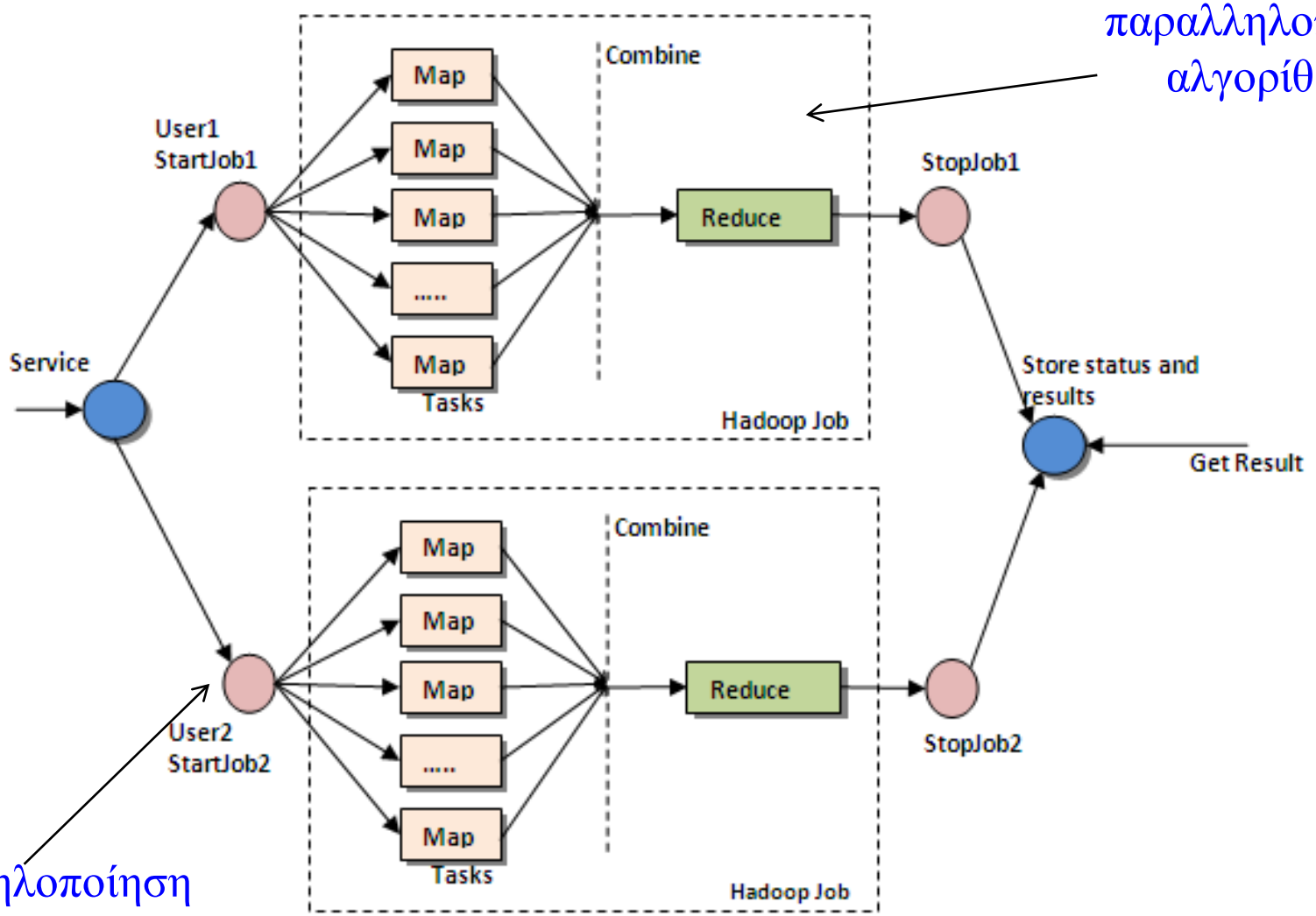
Amazon

Amazon services

Αλλά και… Amazon Web Services

- Amazon Simple Storage Service (Amazon S3)
  - Χώρος αποθήκευσης
- Amazon Elastic Compute Cloud (Amazon EC2)
  - Virtual computing enironment/application server
- Amazon SimpleDB
  - Light ΒΔ: διαχείριση δομημένων δεδομένων
- Amazon CloudFront
  - Κατανεμημένη αποθήκευση και διάθεση (cache services)
- Amazon Simple Queue Service (Amazon SQS)
  - Διαχείριση ουρών μηνυμάτων
- Amazon Comprehend
  - NLP υπηρεσίες βασισμένες στην AI
- Amazon Elastic MapReduce
  - Χρήση MapReduce

Amazon Elastic MapReduce



MapReduce

- Πώς δουλεύει;
- Τι χρειάζεται;

DFS – Κατανεμημένο Σύστημα Αρχείων

- Ένα κατανεμημένο κλιμακώσιμο σύστημα αρχείων για εφαρμογές που διαχειρίζονται μεγάλα κατανεμημένα σύνολα δεδομένων
- Χρησιμοποιείται σαν είσοδος και έξοδος από το MapReduce
- Αρχιτεκτονική Master/Slave

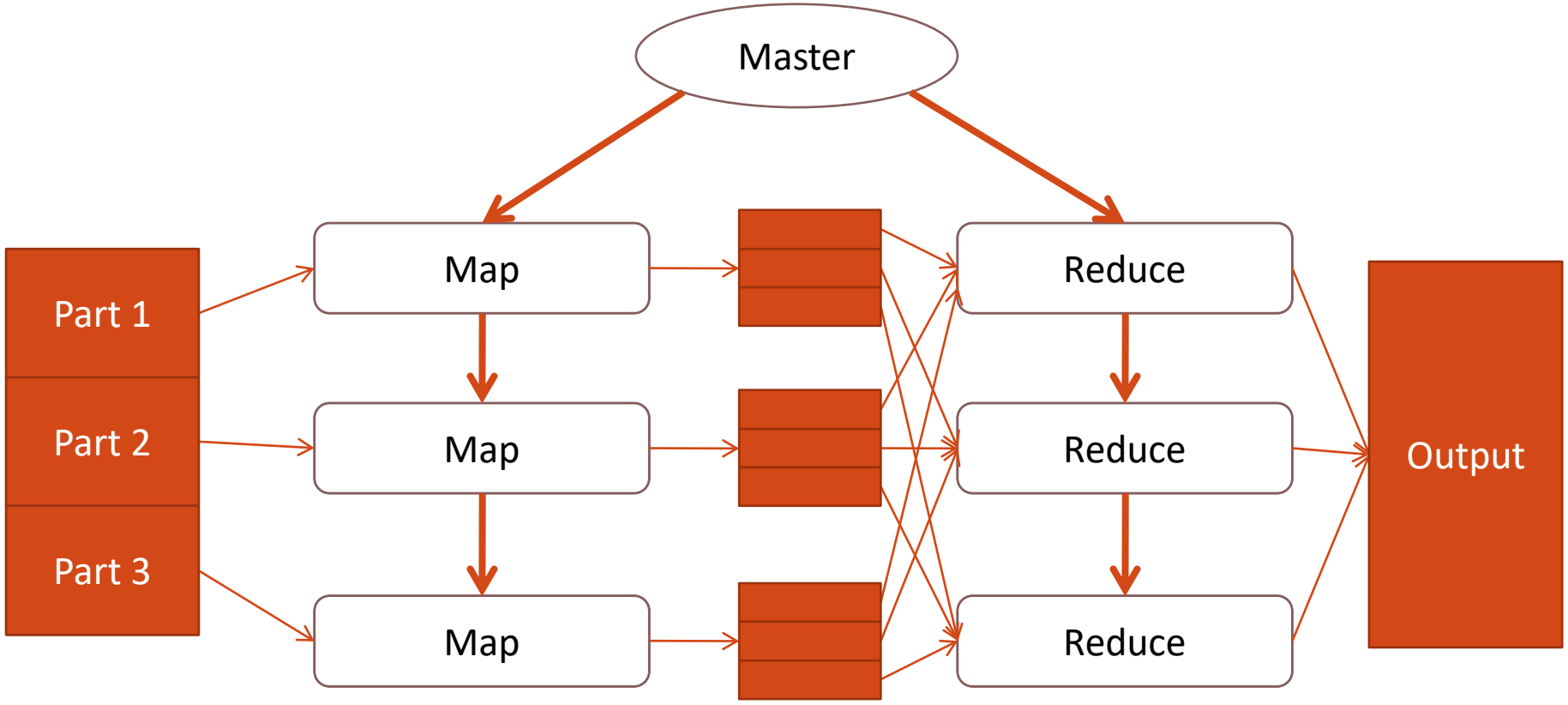
Πριν ξεκινήσουμε

- Η είσοδος ανεβαίνει στο DFS “χωρίζεται” σε M κομμάτια, μεγέθους 16 έως 64 MB
- Κάθε μηχάνημα που συμμετέχει στον υπολογισμό εκτελεί ένα αντίγραφο του προγράμματος
- Ένα από όλα τα μηχανήματα αναλαμβάνει το ρόλο του Master. Αυτός αναθέτει εργασίες στα υπόλοιπα (εργάτες)
  - Αυτές μπορεί να είναι map ή reduce εργασίες

Master

- Ο Master διατηρεί δομές δεδομένων, όπως:
  - Κατάσταση μίας εργασίας
  - Τοποθεσίες των δεδομένων εισόδου, εξόδου και ενδιάμεσων αποτελεσμάτων
- Ο Master είναι υπεύθυνος για το χρονοπρογραμματισμό της εκτέλεσης των εργασιών

Σενάριο Χρήσης



Ολοκλήρωση Εργασιών

- Όταν ένας εργάτης ολοκληρώσει την εργασία του ενημερώνει τον Master
  - Έτσι ενδεχομένως να λάβει και άλλη εργασία (συνήθως ίδια σε άλλο μέρος του partition big data file)
- Όταν όλοι ενημερωσουν τον Master ότι ολοκληρώθηκαν οι εργασίες τους, τότε αυτός τερματίζει και επιστρέφει τη λειτουργία στο αρχικό πρόγραμμα του χρήστη

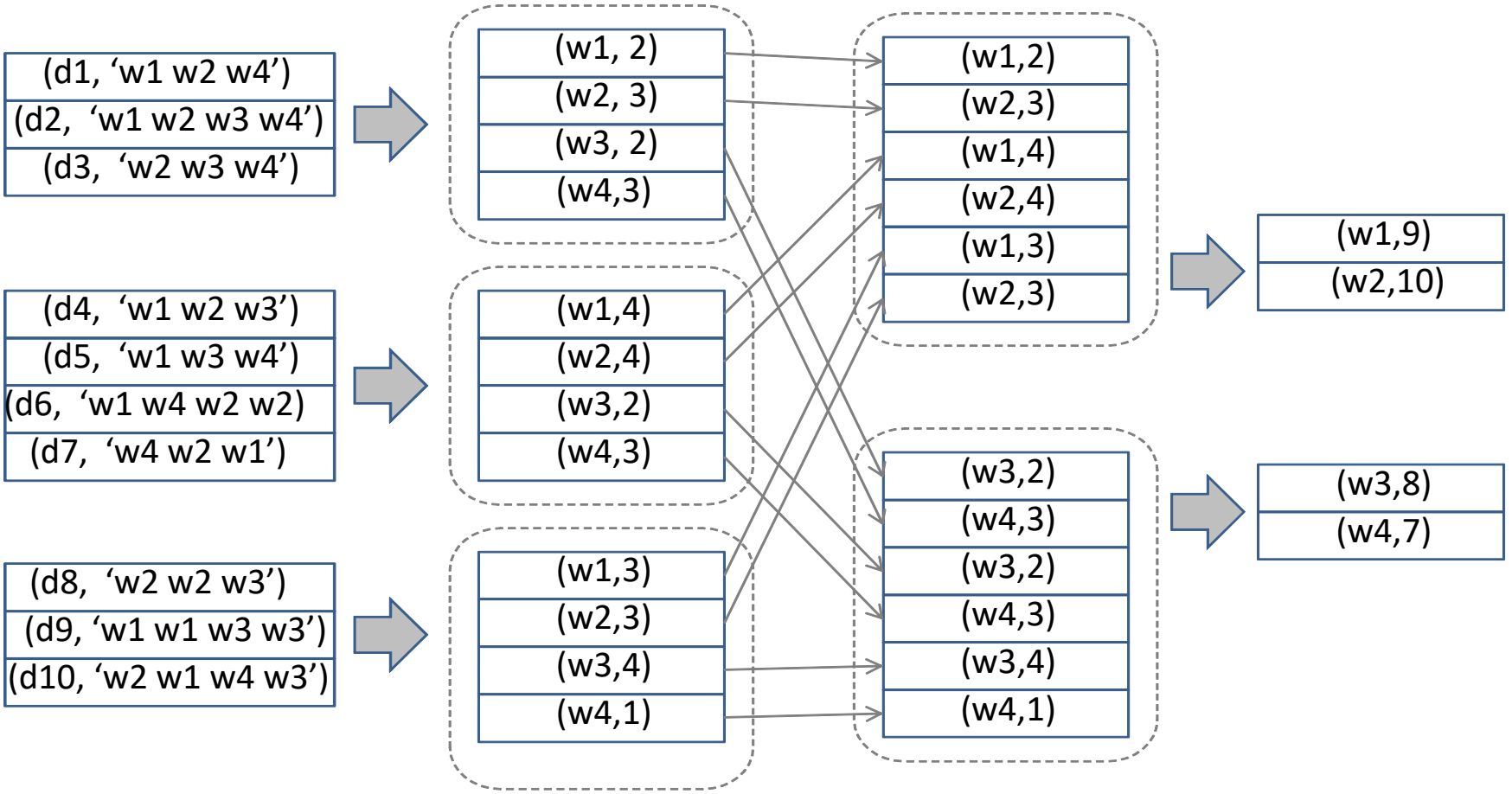
Παράδειγμα: Μέτρηση Λέξεων 1/4

- Στόχος: μέτρηση της συχνότητας εμφάνισης λέξεων σε ένα μεγάλο σύνολο κειμένων
- Πιθανή χρήση:
  - Εύρεση *tf* του όρου *i* βάσει του *tf(i)*idf(i)*
  - Εύρεση δημοφιλών URL σε webserver logfiles
- Πλάνο υλοποίησης:
  - "Ανέβασμα" των κειμένων στο MapReduce
  - Γράφω μια map και μια reduce συνάρτηση
  - Τρέχω μια MapReduce εργασία
  - Παίρνω πίσω τα αποτελέσματα

Παράδειγμα: Μέτρηση Λέξεων 2/4

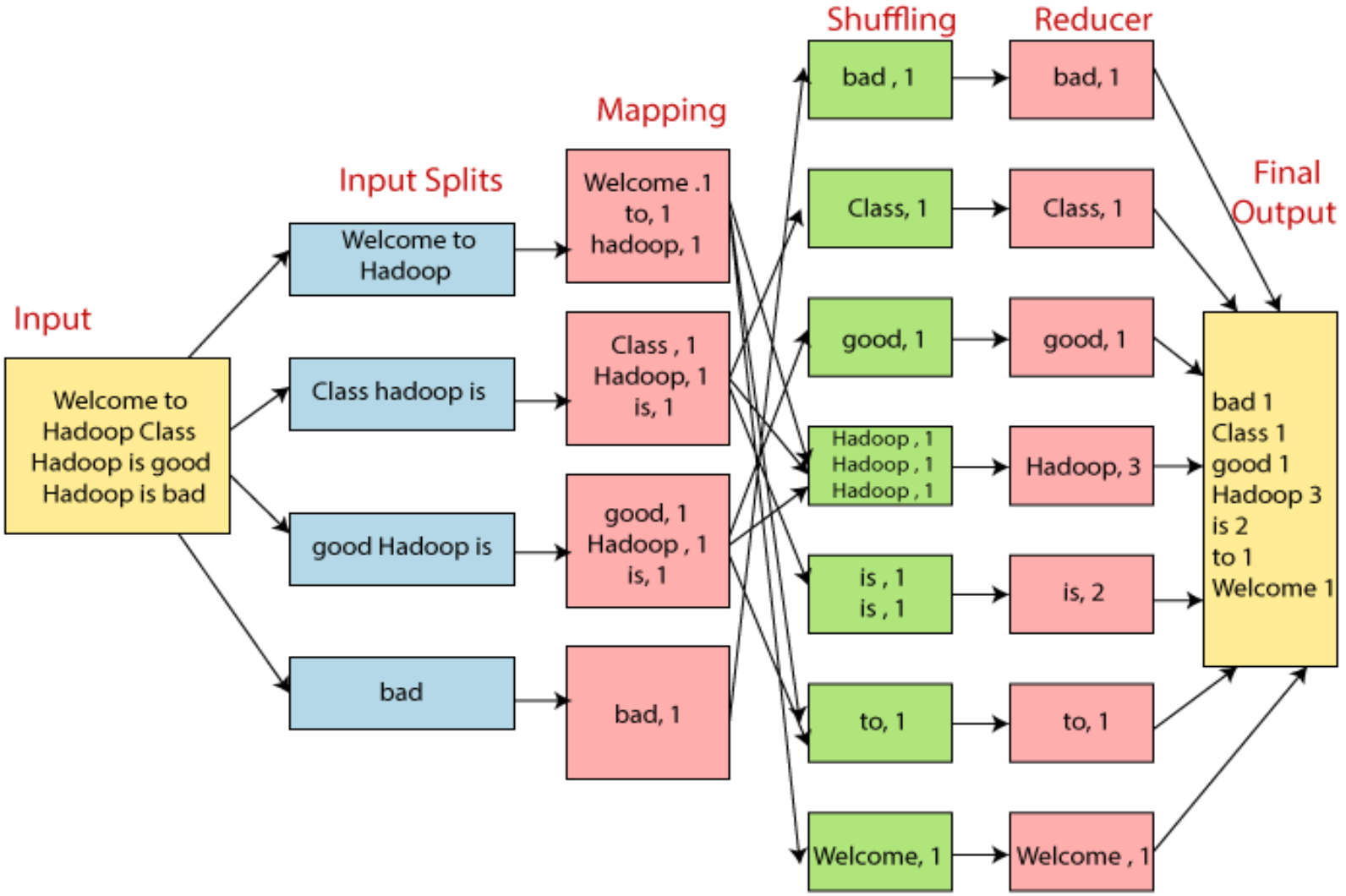
```
map(key, value):
// key: document name; value: text of document
    for each word w in value:
     emit(w, 1)

reduce(key, value):
// key: a word; value: an iterator over counts
     result = 0
     for each count v in values:
          result += v
     emit(result)
```

Παράδειγμα: Μέτρηση Λέξεων 3/4

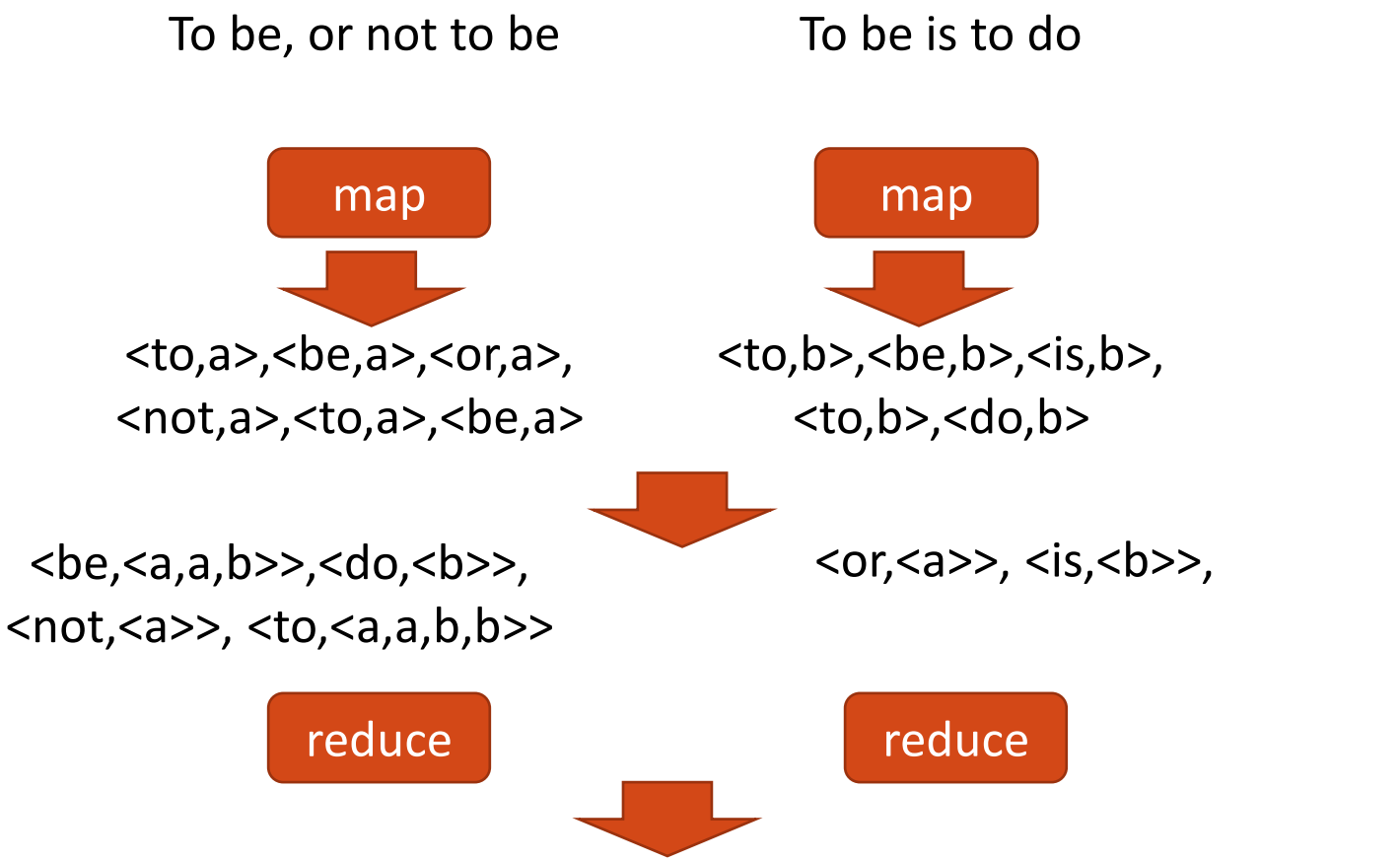


M=3 mappers R=2 reducers

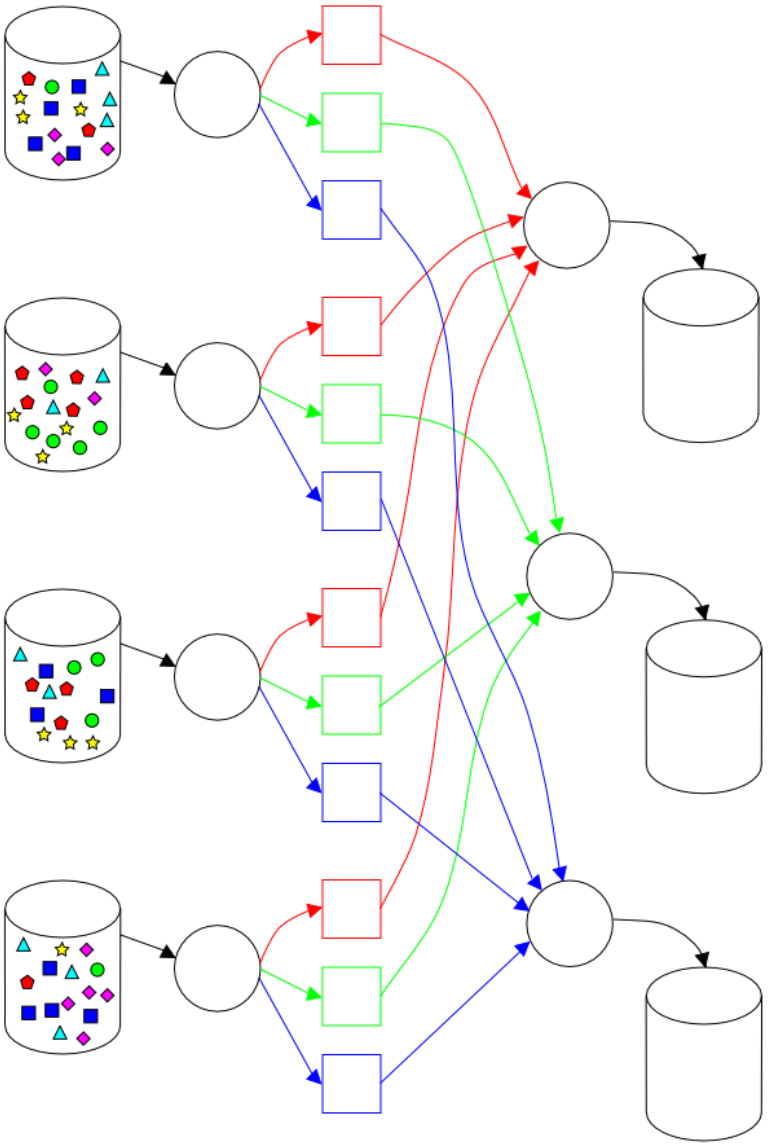
Παράδειγμα: Μέτρηση Λέξεων 4/4



Παράδειγμα: Ανεστραμμένο Ευρετήριο (Inverted Index)



Παράδειγμα: Μέτρηση Σχεδίων



http://www.systems-deployment.com/animation.html

Ανοχή στα σφάλματα

- Ο Master επικοινωνεί με τους εργάτες περιοδικά. Εάν κάποιος δεν ανταποκριθεί για ένα χρονικό διάστημα τότε αναθέτει την εργασία του σε κάποιον άλλο.
- Ο Master περιοδικά διατηρεί αντίγραφα ελέγχου των δομών του. Σε περίπτωση βλάβης ένας νέος μπορεί να ξεκινήσει άμεσα.
- Τα ενδιάμεσα αποτελέσματα που παράγονται από τις map και reduce εργασίες διατηρούνται σε προσωρινά αρχεία σε τοπικά συστήματα αρχείων έως όλη η είσοδος να έχει επεξεργαστεί. Στη συνέχεια ενημερώνεται ο Master και η πληροφορία γίνεται διαθέσιμη σε όλους.

Περισσότερες πληροφορίες

Dean, Jeff and Ghemawat, Sanjay. **MapReduce: Simplified Data Processing on Large Clusters** http://labs.google.com/papers/mapreduce-osdi04.pdf

http://netcins.ceid.upatras.gr/OpSys-II/
διαφάνειες Network-Centric Information Systems (NetCINS) Lab,
Πανεπιστήμιο Πατρών