Web Data Mining
Εργαστήριο 1

Ανδρέας Παπαδόπουλος

andpapad+epl451@gmail.com
andpapad@cs.ucy.ac.cy

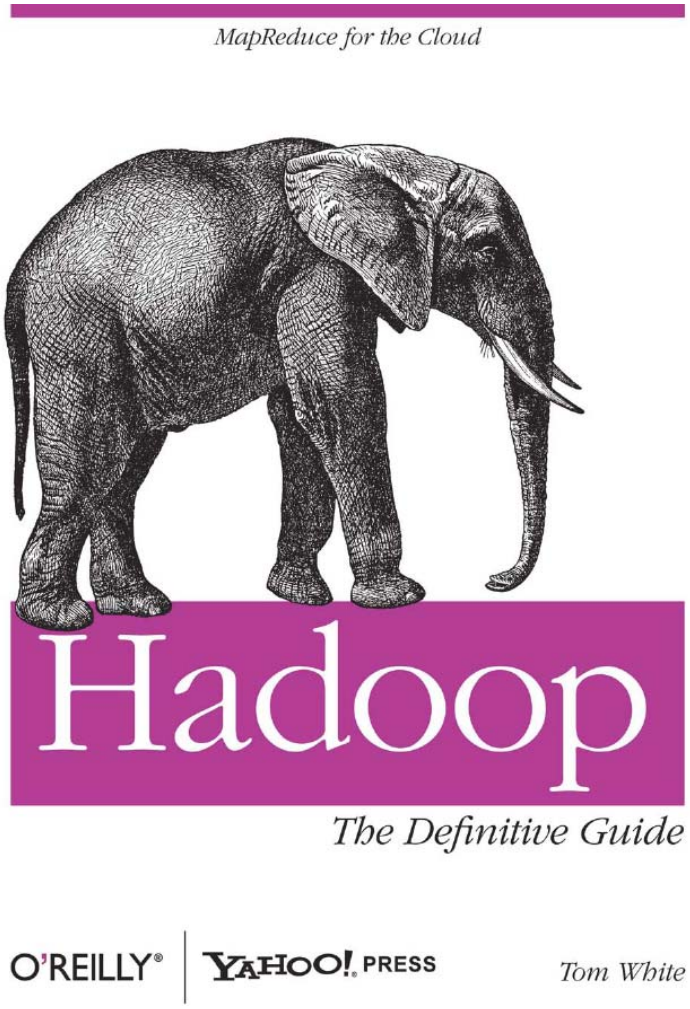
Εισαγωγή

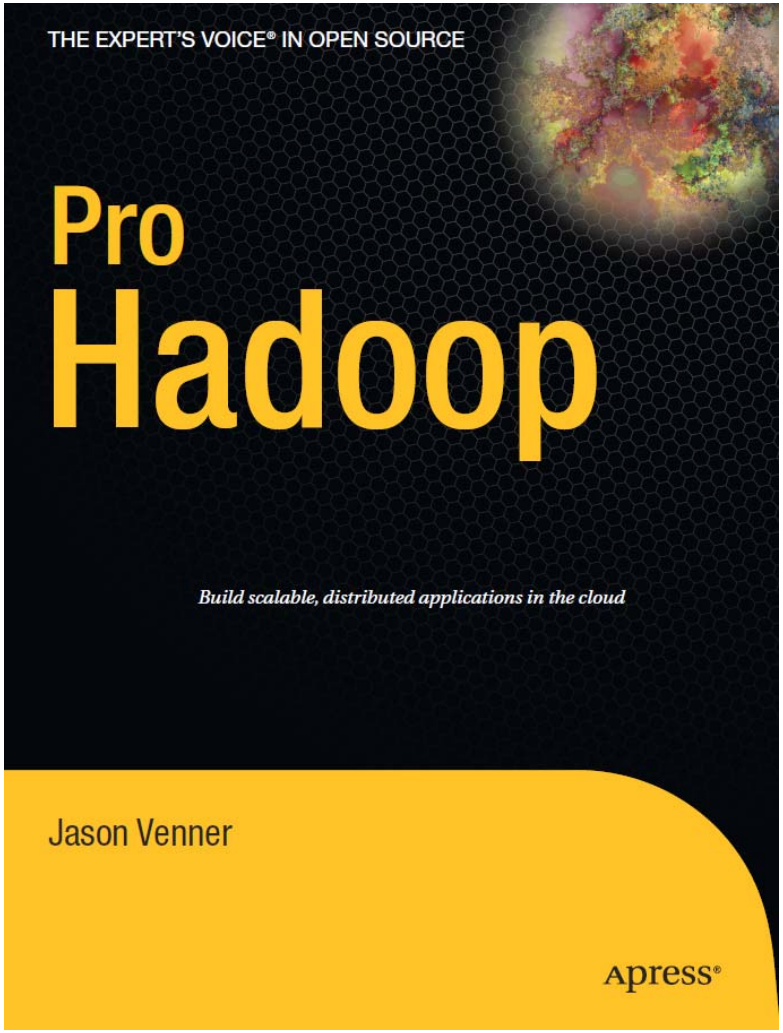
- Παρασκευή 18:00 – 19:30
- Αίθουσα 101, Κτήριο Θετικών και εφαρμοσμένων επιστημών.
- Ώρες Γραφείου
  - Παρασκευή 17:00 – 18:00
  - ΘΕΕ 01 LAB 217

Εισαγωγή

- Hadoop
  - Κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων
  - http://hadoop.apache.org/
- Mahout
  - Scalable Machine Learning and Data Mining
  - On top of Apache Hadoop
  - Use the map/reduce paradigm
  - http://mahout.apache.org/

Εργαστήριο 1 - Hadoop





Επισκόπηση Hadoop

- Λογισμικό ανοιχτού κώδικα
- Υποστηρίζει κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων (Petabytes)
- Παρέχει μια υλοποίηση του μοντέλου Map Reduce
- Βασίστηκε στο Google Map Reduce framework και το Google File System (GFS)
- Έργο του Apache Software Foundation
- Αναπτύσσετε και χρησιμοποιείτε κυρίως από την Yahoo!
- Χρησιμοποιείται σε μεγάλους οργανισμούς ανά το παγκόσμιο π.χ. Facebook, Alibaba, New York Times

Hadoop Distributed File System HDFS

- Κατανεμημένο σύστημα αρχείων
- Ιδανικό για αποθήκευση μεγάλων αρχείων (ιδανικά μεγέθους πολλαπλάσιο του 64Mb)
- Παρόμοιο με το Google File System (GFS)
- Αξιόπιστο

- Οι κόμβοι επικοινωνούν μεταξύ τους για εξισορρόπηση των δεδομένων (replication)
- Ο κεντρικός master server είναι ο *namenode*
- Οι υπόλοιποι κόμβοι ονομάζονται *datanodes*

Ο μηχανισμός του Map Reduce

- Ακολουθεί το μοντέλο αρχιτεκτονικής master/slave.
- Ο κεντρικός master server είναι ο *jobtracker*
- Αναλαμβάνει τον διαμερισμό των εργασιών map reduce στους υπόλοιπους κόμβους, slave servers – *tasktrackers*, για εκτέλεση
- Οι *tasktrackers* απλά εκτελούν τις εργασίες που του αναθέτει ο *jobtracker*.

	Master	Slave
Map Reduce	jobtracker	tasktracker
HDFS	namenode	datanode

Εργαστήριο 1 - Hadoop

- Hadoop API
  - http://hadoop.apache.org/common/docs/r0.20.0/api/index.html
  - web UI for MapReduce job tracker(s): http://hadoopmaster:50030/jobtracker.jsp
  - web UI for task tracker(s) http://hadoopmaster:50060/tasktracker.jsp
  - web UI for HDFS name node(s) http://hadoopmaster:50070/dfshealth.jsp

Εργαστήριο 1 - Hadoop

- Παράδειγμα 1 - WordCount

Map	***Reduce***
Input: a document Output: key=word value=1	Input: key=word values=list of values (1) Output: key=word value=occurrences (SumOfInputValues)

Map(void *input) {
for each word w in input
EmitIntermediate(w,1);
}
	Reduce(String key, Iterator values) {
	int result = 0;
	for each v in values
	result += v;
	Emit(w , result);
	}

Εργαστήριο 1 - Hadoop

- Παράδειγμα 2 – Inverted Index

Map	***Reduce***
Input: a document Output: key=word value=filename	Input: key=word value=list of values(filenames) Output: key=word value=files

Map(void *input) {
for each word w in input
EmitIntermediate(w, filename);
}
	Reduce(String key, Iterator values) {
	String files =” ”;
	for each v in values
	files += v+”-“;
	Emit(w , files);
	}

Εργαστήριο 1 – Εκτέλεση του WordCount

- Αντιγράψτε το κώδικα του wordcount
- Κάντε compile και δημιουργήστε ένα αρχείο jar
- Ανεβάστε το στο server hadoopmaster
- Τρέξτε το παράδειγμα

Εργαστήριο 1 – Εργασία

- Αλλάξτε το κώδικα του wordcount έτσι ώστε να λύνει το πρόβλημα στην διαφάνεια 47 της διάλεξης 2
- Ακολουθήστε τα ίδια βήματα με πριν