

Αναλυτική μεγάλων δεδομένων στον okeanos-knossos



Βασίλης Καφφές
ΙΠΣΥ, Ε.Κ. «ΑΘΗΝΑ»

Μεγάλα Δεδομένα και Επιστήμη Δεδομένων

- ▶ Τα **Μεγάλα Δεδομένα** (Big Data) αφορούν τη συλλογή, αποθήκευση, επεξεργασία και ανάλυση μεγάλου όγκου ετερογενών δεδομένων.
 - Πρόκειται για δεδομένα που χαρακτηρίζονται από ένα ή περισσότερα από τα λεγόμενα Vs: *Volume*, *Velocity*, *Variety* (+ *Variability* και *Complexity*)
- ▶ Η **Επιστήμη των δεδομένων** (Data Science) αποτελεί μια σύνθεση τεχνικών εφαρμοσμένης στατιστικής και μηχανικής μάθησης με τεχνικές προγραμματισμού, μοντελοποίησης και διαχείρισης δεδομένων.

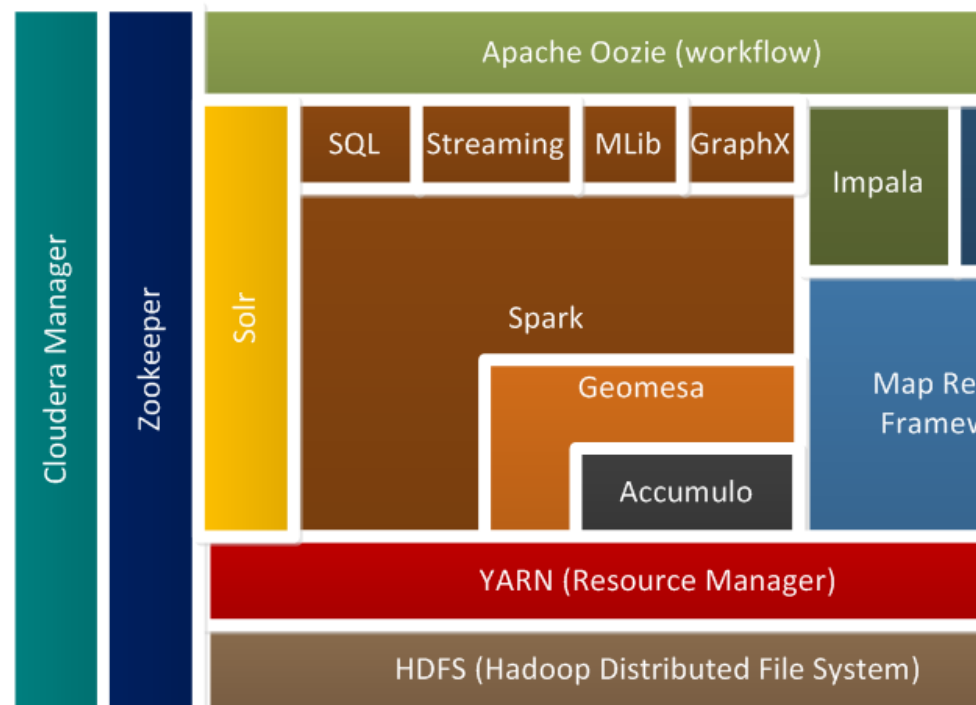
Το Έργο ΜΕΔΑ²

- ▶ ΜΕΔΑ²: «Από τα Μεγάλα Δεδομένα στην Επιστήμη Δεδομένων»
 - Ερευνητικό έργο στα πλαίσια της Πράξης «Υπολογιστικές Επιστήμες και Τεχνολογίες Δεδομένων, Περιεχομένου και Αλληλεπίδρασης» της Δράσης για την «Στρατηγική Ανάπτυξη Ερευνητικών και Τεχνολογικών Φορέων»
 - Αποτελεί συνέχεια και επέκταση του έργου ΜΕΔΑ: «Μεγάλα Δεδομένα – Προσκλήσεις, Μέθοδοι και Τεχνικές Αποδοτικής Διαχείρισης» που υλοποιήθηκε από το ΙΠΣΥ / Ε.Κ. Αθηνά στο πλαίσιο της Δράσης ΚΡΗΠΙΣ της ΓΓΕΤ.

Στόχοι του Έργου

- ▶ Απόκτηση τεχνογνωσίας στην Επιστήμη των Δεδομένων
 - Απόκτηση θεωρητικών βάσεων, τεχνογνωσίας καθώς και εξοικείωσης με εργαλεία και υποδομές της Επιστήμης Δεδομένων
 - Εγκατάσταση του αναγκαίου υπόβαθρου σε υποδομές, προγραμματιστικά εργαλεία και συλλογές δεδομένων
- ▶ Ανάπτυξη καινοτόμων μοντέλων, μεθόδων και αλγορίθμων για την αντιμετώπιση των προκλήσεων που θέτουν τα Μεγάλα Δεδομένα
- ▶ Εφαρμογή σε πιλοτικές εφαρμογές για κλάδους που θέτει σε προτεραιότητα η Στρατηγική Έρευνας και Τεχνολογίας για την Έξυπνη Εξειδίκευση

Τεχνολογίες στο ΜΕΔΑ²



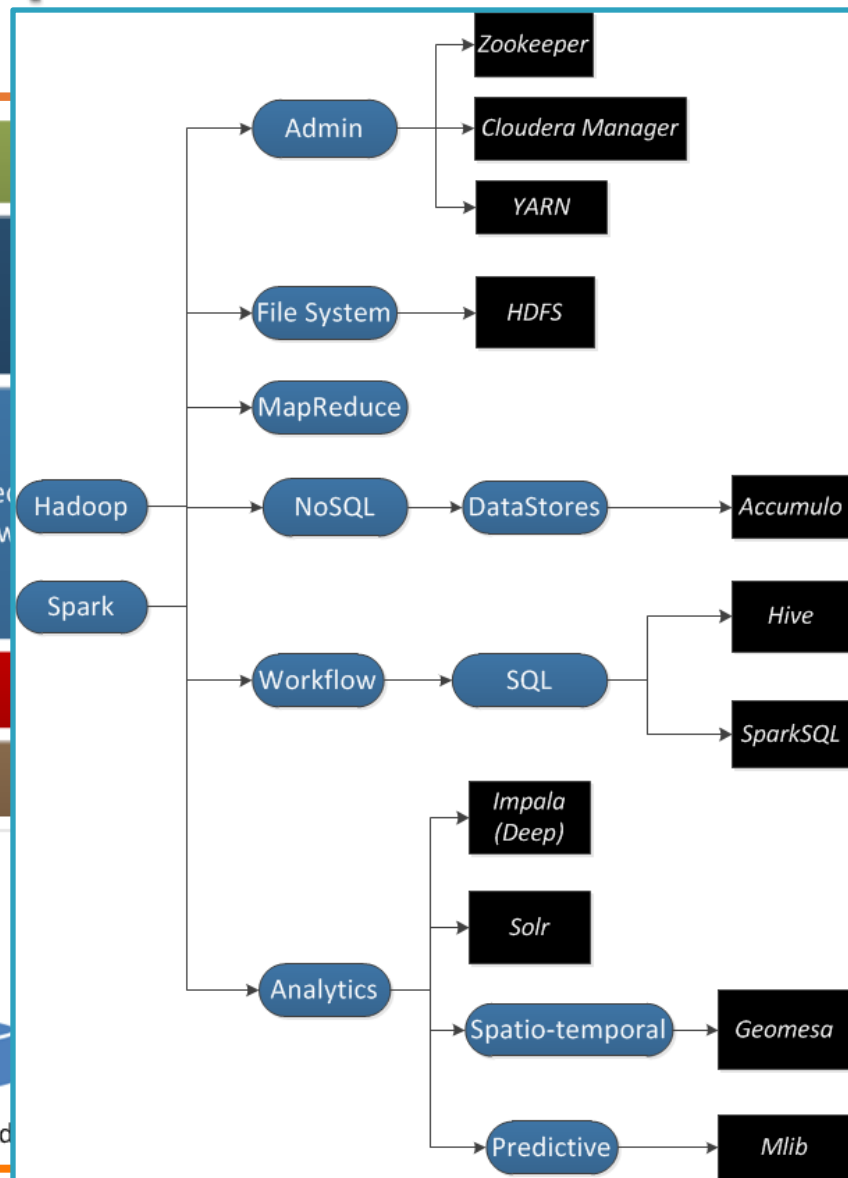
Import or export



Unstructured or semi structured Data



Structured



Απαιτήσεις Συστήματος

Υλικό	Απαιτήσεις	Σημειώσεις
CPU	16+ CPU cores	Συστήνεται από Cloudera για Data Science – https://www.cloudera.com/documentation/enterprise/release-notes/topics/hardware_requirements_guide.html#concept_zmr_fyb_hbb
Μνήμη	32+ GB	Όπως παραπάνω
Δίσκος	>1TB	• 41 σύνολα δεδομένων πάνω από 2TB • x3 – παράγοντας replication στο HDFS • Λειτουργικό σύστημα

Τύποι Κόμβων

▶ Manager/Utility

- εκτελεί διαδικασίες του cluster που δεν είναι master processes όπως τα Cloudera Manager και Hive Metastore

▶ Master

- εκτελεί Hadoop master processes όπως τα HDFS NameNode και YARN Resource Manager

▶ Gateway

- σημεία πρόσβασης για την εκκίνηση jobs στο cluster όπως Spark2 gateway

▶ Slave/Worker

- κυρίως για την εκτέλεση HDFS DataNodes και άλλων κατανεμημένων διαδικασιών όπως τα Impalad ή Accumulo Tablet Servers

Κατανομή Διαθέσιμων Πόρων

- ▶ Μέσω του okeanos-knossos παρέχεται η δυνατότητα ανάπτυξης έως:
 - 32 κόμβων.
 - 200 πυρήνων.
 - 340GB RAM.
 - 17TB αποθηκευτικού χώρου.
 - 15 ιδιωτικών δικτύων και 15 δημοσίων διευθύνσεων IP.

Name	CPUs	RAM (GB)	Disks (GB)	Public IPs
manager	8	8	30	✓
master	16	32	2500	✓
slaves[1-9]	16	32	1000-2000	✗
Total	168	328	16800	2

Πιλοτικά Σενάρια Χρήσης και Αξιολόγησης

- ▶ Στόχος σεναρίων αξιολόγησης
 - Μεταφορά σύγχρονων μεθόδων για επεξεργασία δεδομένων που χρησιμοποιούνται στο Ινστιτούτο Πληροφοριακών Συστημάτων (ΙΠΣΥ) στο οικοσύστημα των Μεγάλων Δεδομένων
 - Αναγνώριση εργαλείων που είναι πιο κατάλληλα για κάθε κατηγορία δεδομένων
- ▶ Στόχος πιλοτικών εφαρμογών
 - Αποτελεσματική υλοποίηση των βασικών αλγορίθμων και της τεχνογνωσίας του ΙΠΣΥ
 - Ευκολία χρήσης (αλλά μη εμπορικά εργαλεία)
 - Ανταπόκριση σε πραγματικά προβλήματα
- ▶ Σενάρια αξιολόγησης που έχουν σχεδιαστεί
 - Χωρική Ευρετηρίαση και Ανάλυση
 - Πρόβλεψη Ψευδών Ειδήσεων
 - Ανάστροφη Ευρετηρίαση Βιβλιογραφικών Δεδομένων

Χωρική Ευρετηρίαση και Ανάλυση

- ▶ Τύποι ερωτημάτων: *Range Query*, *kNN query*, *Distance join query*.
- ▶ Εργαλεία **GeoSpark** και **GeoMesa**.
- ▶ Χωρικά Ευρετήρια: *R-tree*, *Quadtrees*, *KDB-tree*, *space-filling curves*.
- ▶ Σύνολα δεδομένων: *NY taxis*, *OSM*, *Flickr*, *GDELT*
 - Δείγμα από το καθένα (20%, 40%, 60%, 80%, 100%) για κλιμάκωση.
 - 200 τυχαία σημεία για κάθε τύπο ερωτημάτων.
- ▶ Αξιολόγηση κάθε τύπου ερωτήματος και για κάθε σύνολο
 - range queries – μεταβολή *radius* (10 , 100, 1000 μέτρα)
 - kNN queries – μεταβολή *k* (1, 10, 100)
 - distance joins – μεταβολή ϵ (10 , 100, 1000 μέτρα)
- ▶ Αξιολόγηση κλιμάκωσης των διαφόρων ερωτημάτων
 - Μεταβολή μεγέθους δεδομένων (βάσει των παραπάνω δειγμάτων)
 - Μεταβολή αριθμού κόμβων της συστάδας υπολογιστών (1, 3, 6, 9)

Πρόβλεψη Ψευδών Ειδήσεων (1 / 2)

- ▶ Αξιοποίηση μεγάλου αριθμού διαθέσιμων ελληνικών tweets
 - Παρακολούθηση χαρακτηριστικών διάδοσης παλαιότερων ειδήσεων στο Twitter
 - Αυτόματη κατηγοριοποίηση ειδήσεων ως ψευδών με βάση τα χαρακτηριστικά διάδοσής τους
- ▶ Σενάρια αξιολόγησης
 - Διασταύρωση του περιεχομένου των tweets με ιστοσελίδες, δημιουργία ενός συνόλου εκπαίδευσης από tweets
 - Πειραματισμός με μεγάλο αριθμό χαρακτηριστικών και εύρεση των πιο κατάλληλων
 - Ανάπτυξη αλγορίθμων μηχανικής μάθησης και αξιολόγησή τους

Πρόβλεψη Ψευδών Ειδήσεων (2/2)

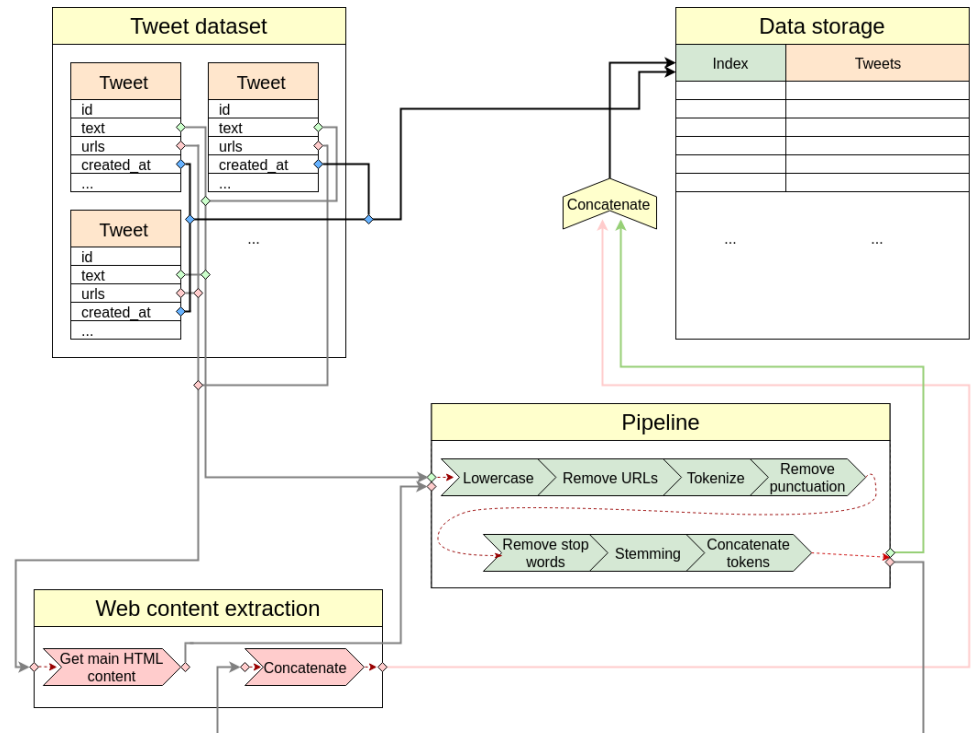
► Επεξεργασία tweets

- Apache Spark/PySpark
- Εξαγωγή δεδομένων ιστού (web scraping) για URLs
- Επεξεργασία κειμένου



► Apache Solr στα extended tweets με περιεχόμενα ιστού (web contents)

- Κατανεμημένη ευρετηρίαση
- Κατανεμημένη κειμένο-χωρική αναζήτηση



Ανάστροφη Ευρετηρίαση Βιβλιογραφικών Δεδομένων

- ▶ Ευρετηρίαση των συνόλων δεδομένων επιστημονικής βιβλιογραφίας σε λέξεις-κλειδιά. Σύγκριση των μηχανών:
 - Sphinx (μη-κατανεμημένο περιβάλλον)
 - Apache Solr (κατανεμημένο περιβάλλον)
 - Elasticsearch (κατανεμημένο περιβάλλον)
- ▶ Πειράματα αποτίμησης για:
 - Χρόνο κατασκευής του ευρετηρίου
 - Χρόνο απόκρισης των μηχανών αναζήτησης
 - Διεκπεραιωτικής ικανότητας (throughput) των μηχανών αναζήτησης
- ▶ Αξιολόγηση πειραμάτων
 - Παραμετροποίηση μηχανών (π.χ., replication, sharding)
 - Κλιμάκωση του μεγέθους της υπολογιστικής συστάδας

Βασίλης Καφφές – vkaffes@imis.athena-innovation.gr
Ε.Κ. ΑΘΗΝΑ' – <https://www.athena-innovation.gr/>
ΙΠΣΥ: Ινστιτούτο Πληροφοριακών Συστημάτων –
www.imis.athena-innovation.gr