

Το Τεχνολογικό Υπόβαθρο

Δεδομένα, Εργαλεία και Υπολογιστικές Υποδομές

Δημήτρης Σκούτας

Κύριος Ερευνητής

Ινστιτούτο Πληροφοριακών Συστημάτων



ΑΘΗΝΑ

Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας

Τα προαναφερθέντα έργα υλοποιήθηκαν στο πλαίσιο της Πρόσκλησης με κωδικό 031 με τίτλο «ΔΡΑΣΗ ΣΤΡΑΤΗΓΙΚΗΣ ΑΝΑΠΤΥΞΗΣ ΕΡΕΥΝΗΤΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΩΝ ΦΟΡΕΩΝ», που χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία», ΕΣΠΑ 2014-2020 ΕΣΠΑ 2014-2020, με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης). Αναλυτικά: "Υπολογιστικές Επιστήμες και Τεχνολογίες Δεδομένων, Περιεχομένου και Αλληλεπίδρασης" (MIS 5002437) με 3 υποέργα: Υποέργο 1 «Γλωσσικές Τεχνολογίες Ανάλυσης Περιεχομένου και Αλληλεπίδρασης», Υποέργο 2 «Τεχνολογίες Ανάλυσης Περιεχομένου στον Πολιτισμό», Υποέργο 3 «ΜΕΔΑ: Μεγάλα Δεδομένα – Προκλήσεις, Μέθοδοι και Τεχνικές Αποδοτικής Διαχείρισης» και "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα" (MIS 5002434) με Υποέργο 1: "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα".



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο
Περιφερειακής Ανάπτυξης



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΔΙΑΡΘΡΩΤΙΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΙΧΕΙΡΗΣΙΑΚΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

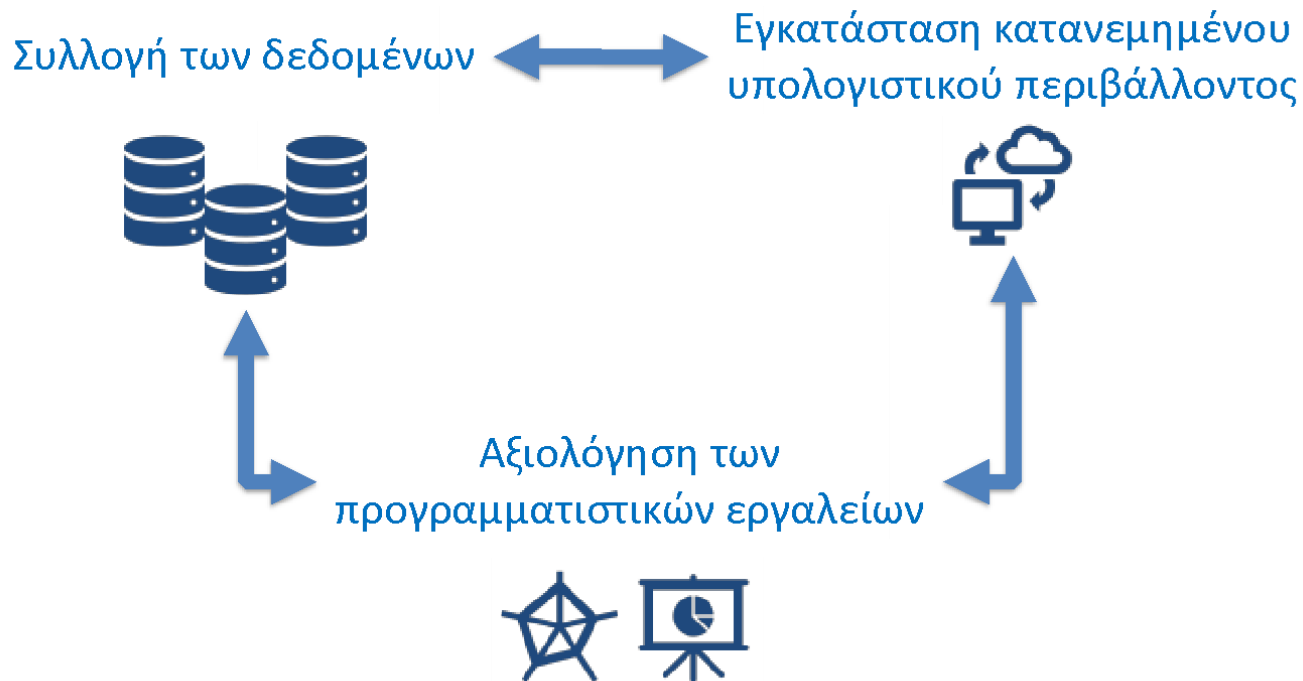
ΕΠΑΝΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ



ανάπτυξη - εργασία - αλληλεγγύη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στόχοι και Εργασίες





Αποθετήριο Δεδομένων

Μεγάλη ποικιλία διαθέσιμων συνόλων δεδομένων με ευρεία γκάμα ως προς:

- είδος
 - π.χ. γεωχωρικά, κοινωνικά, κλπ.
- μορφή
 - π.χ. χρονοσειρές, γράφοι, κλπ.
- μέγεθος
 - από μερικά MB έως μερικές εκατοντάδες GB
- περιεχόμενο
 - ειδήσεις, μηνύματα, φωτογραφίες, κλπ.





Περιεχόμενα Αποθετηρίου

- Χωρο-χρονικά: 15 σύνολα δεδομένων
 - OpenStreetMap, Geonames, κλπ.
- Κοινωνικά: 2 σύνολα δεδομένων
 - Greek Tweeto-Sphere, Socioscope Data
- Γονιδιακά: 8 σύνολα δεδομένων
 - GO, KEGG, MeSH, κλπ.
- Βιβλιογραφικά: 13 σύνολα δεδομένων
 - DBLP, Pubmed, Microsoft Academic Graph, κλπ.
- Περιβαλλοντικά: 3 σύνολα δεδομένων
 - Energy, Biomass





Μεταδεδομένα

- Short Title
- Full Title
- Contact Person
- Description
- Owner
- Source
- Subject
- Tags
- RepoPath
- Spatial Coverage
- Temporal Coverage
- Frequency
- Publication Year
- Size
- Format
- License
- ReadMe





Υπολογιστικό Περιβάλλον

Οι απαιτούμενοι πόροι εξασφαλίστηκαν μέσω της υπηρεσίας ~okeanos-knossos του ΕΔΥΤΕ.

Δυνατότητα ανάπτυξης έως:

- 40 κόμβων / εικονικών μηχανών (αρχική τιμή: 32)
- 600 πυρήνων (αρχική τιμή: 200)
- 1 TB RAM στους προαναφερθέντες κόμβους (αρχική τιμή: 340GB)
- 50 TB αποθηκευτικού χώρου στους προαναφερθέντες κόμβους (αρχική τιμή: 17 TB)
- 15 ιδιωτικών δικτύων και μέχρι 15 δημοσίων διευθύνσεων IP





Χρήση των Πόρων

Κατασκευάστηκε μία υπολογιστική συστάδα 16 εικονικών μηχανών

- Οι 12 εικονικές μηχανές χρησιμοποιούνται για τη φιλοξενία των υπολογιστικών εργαλείων. Η καθεμία από τις εν λόγω εικονικές μηχανές διαθέτει τα παρακάτω χαρακτηριστικά:
 - 16 πυρήνες
 - 32 GB μνήμης RAM
 - 1 TB αποθηκευτικού χώρου (σκληρός δίσκος) για το κατανεμημένο σύστημα αρχείων
- Η 1 εικονική μηχανή χρησιμοποιείται αποκλειστικά για τη διαχείριση των παραπάνω εικονικών μηχανών. Η εν λόγω εικονική μηχανή διαθέτει τα παρακάτω χαρακτηριστικά:
 - 8 πυρήνες
 - 16 GB μνήμης RAM
- Οι υπόλοιπες 3 εικονικές μηχανές καλύπτουν επιμέρους / ad hoc ανάγκες
- Το λειτουργικό σύστημα που χρησιμοποιήθηκε στις εικονικές μηχανές είναι το Ubuntu Server 16.04 LTS





Εγκατάσταση Εργαλείων

Εναλλακτικές διαδικασίες εγκατάστασης:

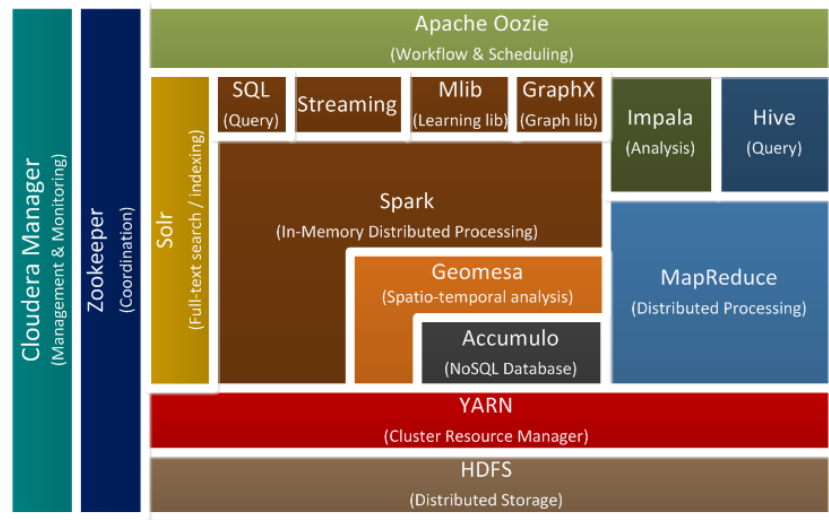
- Χειροκίνητη μέθοδος
- Διανομή Cloudera
- Διανομή Hortonworks
- Διανομή MapR

Επιλογή:

- Η Cloudera εμφανίζει την εντονότερη παρουσία στην αγορά, με την Hortonworks να έρχεται δεύτερη και την MapR τρίτη. Αποτελεί δοκιμασμένη λύση, με τη μεγαλύτερη κοινότητα χρηστών και το περισσότερο πληροφοριακό υλικό διαθέσιμο διαδικτυακά.



Αρχιτεκτονική & Εργαλεία της Υποδομής

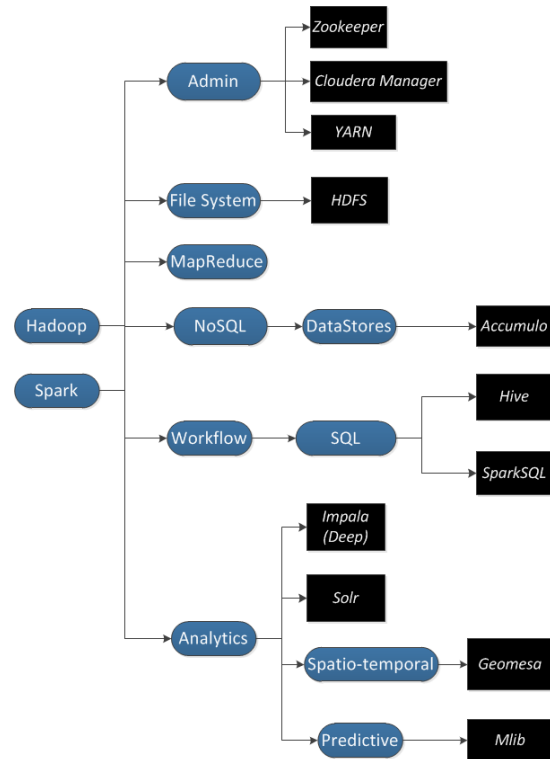


Import or export

Unstructured or semi structured Data



Structured Data



Παρακολούθηση Εγκατεστημένων Εργαλείων

Cluster 1 (CDH 5.14.2, Parcels)

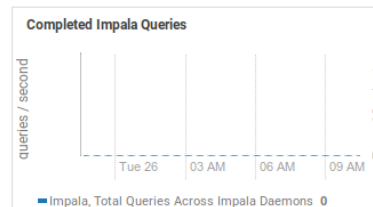
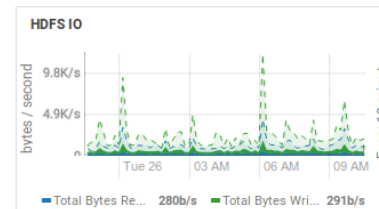
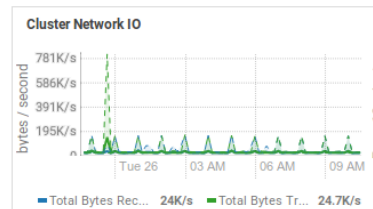
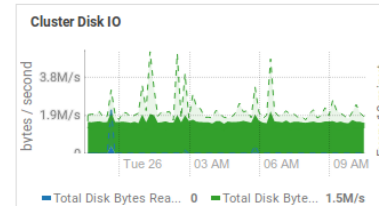
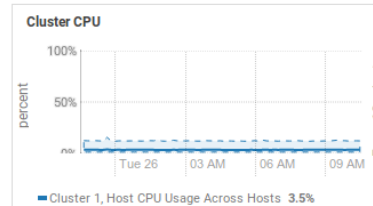
✓	8 Hosts	5
✓	Accumulo	
✓	HDFS	
✓	Hive	
✓	Hue	
✓	Impala	
✓	Oozie	
✓	Solr	
✓	Spark 2	
✓	YARN (MR2 In...	
✓	ZooKeeper	1

Cloudera Management Service

✓	Cloudera Man...	
---	-----------------	--

Charts

30m 1h 2h 6h 12h 1d 7





Αξιολόγηση Εργαλείων (I)

Ευρετηρίαση και ανάλυση γεωχωρικών δεδομένων

- Τύποι ερωτημάτων:
 - range queries
 - kNN queries
 - distance joins
- Εργαλεία
 - GeoSpark
 - GeoMesa
- Σύνολα δεδομένων:
 - GDELT
 - Flickr
 - OpenStreetMap

Σύνοψη αποτελεσμάτων:

- Σε σύνολα δεδομένων που είναι στο μέγεθος του Flickr, ο χρόνος απόκρισης του Geospark είναι τουλάχιστον μια τάξη μεγέθους μικρότερος από αυτόν του Geomesa.
- Η κατάσταση διαφοροποιείται κάπως όταν ο όγκος των δεδομένων αυξάνεται, όπως στην περίπτωση του GDELT.





Αξιολόγηση Εργαλείων (II)

Επεξεργασία tweets,
δημιουργία
ανάστροφου
ευρετηρίου, αναζήτηση

- Εργαλεία
 - Apache Spark
 - Apache Solr
- Σύνολα δεδομένων:
 - Greek Tweeto-Sphere

Σύνοψη αποτελεσμάτων:

- Η επεξεργασία με τη χρήση του Apache Spark παρουσιάζει εξαιρετικές δυνατότητες κλιμάκωσης.
- Η δημιουργία αναστροφου ευρετηρίου με τη χρήση του Apache Solr δεν παρουσιάζει σημαντική δυνατότητα κλιμάκωσης.
- Η αναζήτηση με τη χρήση του Apache Solr παρουσιάζει δυνατότητες κλιμάκωσης.





Αξιολόγηση Εργαλείων (III)

Ανάστροφη
ευρετηρίαση
βιβλιογραφικών
δεδομένων

- Εργαλεία
 - Apache Solr
 - Elasticsearch
 - Sphinx
- Σύνολα δεδομένων:
 - DBLP, PubMed

Σύνοψη αποτελεσμάτων:

- Σε ένα κόμβο, το Sphinx σταθερά επιτυγχάνει τους χαμηλότερους χρόνους κατασκευής ευρετηρίου, ακολουθεί το Apache Solr και μετά το Elasticsearch
- Το Elasticsearch εμφανίζει πλεονεκτήματα σε δυνατότητες κλιμάκωσης και ευκολία παραμετροποίησης.





VIRTUAL EVENT

Ημέρες Δράσης

23-24-25
ΙΟΥΝΙΟΥ
2021



Thank you!