ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
UNIVERSITY OF MACEDONIA

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Ηλεκτρονικό Επιχειρείν & Τεχνολογία Καινοτομίας

Τα Δεδομένα Μεγάλης Κλίμακας: Τεχνικές και Εργαλεία Ανάλυσής τους, και η προσφορά τους ως υπηρεσία του υπολογιστικού νέφους

ΠΟΛΥΜΕΝΗΣ ΙΟΡΔΑΝΗΣ

Περιεχόμενα

1. Εισαγωγή στα δεδομένα Μεγάλης Κλίμακας (Big Data)
2. Big Data Techniques
3. Big Data Tools
4. Cloud Computing
5. Big Data and Cloud Computing
6. Big Data as a Service (BDaaS)
7. E2E Big Data as a Service Solutions
8. Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για δεδομένα μεγάλης κλίμακας
9. Συμπεράσματα

Σκοπός

- **Να γίνει κατανοητό το θεωρητικό πλαίσιο των Big Data μέσα από τη παρουσίαση τεχνικών και εργαλείων**
- **Να δειχθεί πως το υπολογιστικό νέφος μπορεί να βοηθήσει στο θέμα της διαχείρισης των Big Data**
- **Να παρουσιαστεί το πλαίσιο του Big Data as a Service και End to End υπηρεσίες που προσφέρονται από εταιρείες**
- **Να γίνει σύγκριση της επεξεργασίας σε τοπικό και cloud περιβάλλον**

Εισαγωγή στα δεδομένα Μεγάλης Κλίμακας (Big Data)

Τί είναι τα Big Data

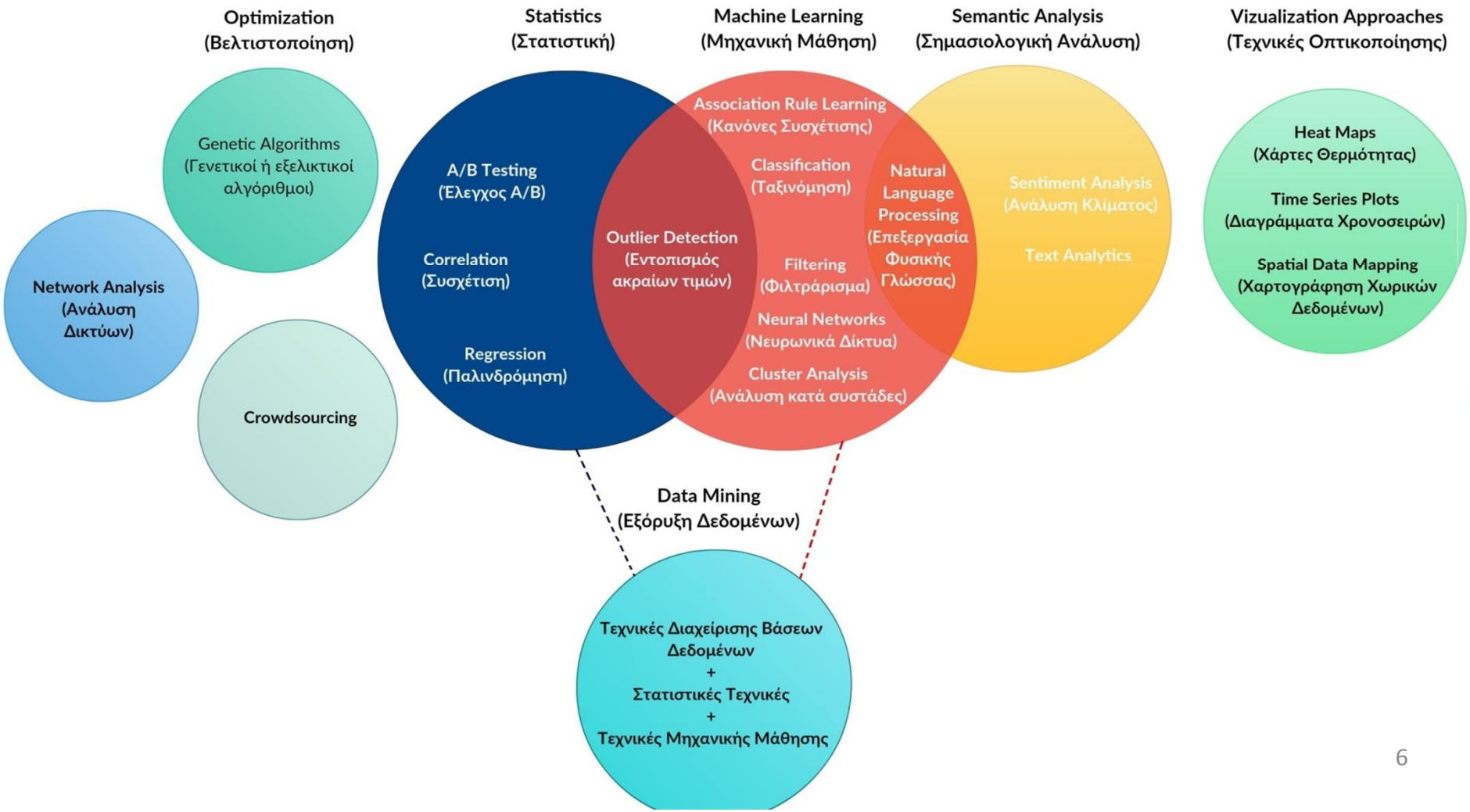
'' Τεράστια σύνολα δεδομένων που έχουν μια πιο μεγάλη, πιο ποικίλη και πολύπλοκη δομή, που δημιουργεί δυσκολίες στην αποθήκευση, την ανάλυση και την οπτικοποίησή τους ώστε να είναι εκμεταλλεύσιμα για περαιτέρω διαδικασίες σε μια επιχείρηση''. (McKinsey Global Institute)

Εισαγωγή στα δεδομένα Μεγάλης Κλίμακας (Big Data)

Δεδομένα μεγάλης κλίμακας σε σύγκριση με την παραδοσιακή εξόρυξη δεδομένων

	Παραδοσιακή Εξόρυξη Δεδομένων	Big Data Analytics
Volume (Όγκος)	Gigabytes, Terabytes	Terabytes, Petabytes, Exabytes
Velocity (Ταχύτητα)	Μικρότερη Ταχύτητα Αλλαγής Δεδομένων	Μεγάλη Ταχύτητα Δημιουργίας Νέων Δεδομένων
Variety (Ποικιλία)	Δομημένα Δεδομένα	Ημιδομηνένα, Αδόμητα Δεδομένα
Veracity (Ορθότητα)	Μεγάλη αξιοπιστία δεδομένων	Κίνδυνος αναξιόπιστων δεδομένων
Value (Αξία)	Μικρότερες δυνατότητες για αξία σε σύγκριση με τους ανταγωνιστές	Η σωστή εκμετάλλευση δημιουργεί ανταγωνιστικό πλεονέκτημα

Big Data Techniques



Big Data Tools

Συστήματα επεξεργασίας γενικού σκοπού

- Apache Hadoop
- Spark

Big Data SQL Systems

- Hive
- IBM Big SQL

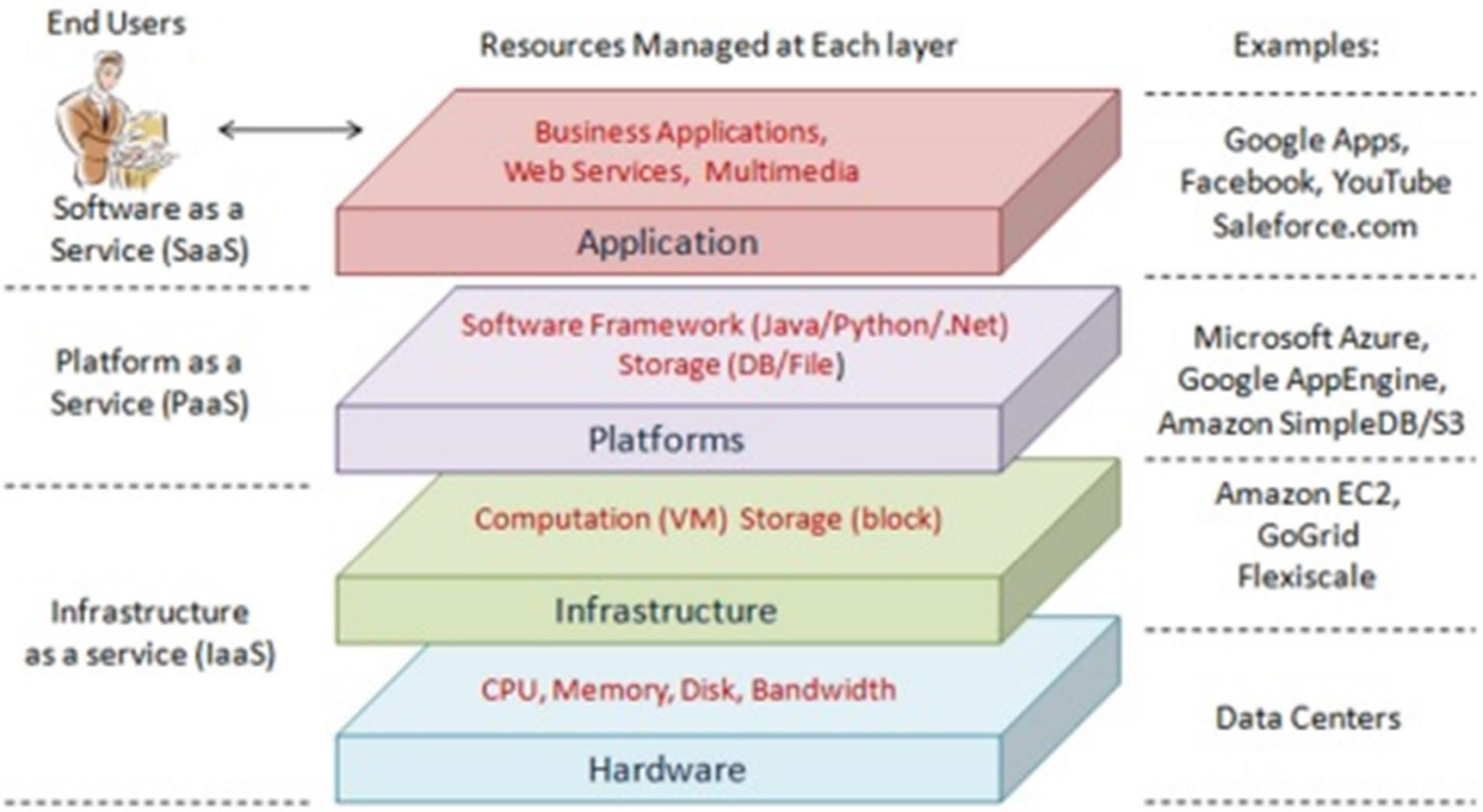
Big Data Graph Processing Tools

- GraphX

Stream Processing Big Data Tools

- Storm
- S4
- InfoSphere Streams

Αρχιτεκτονική και είδη υπηρεσιών του Cloud Computing

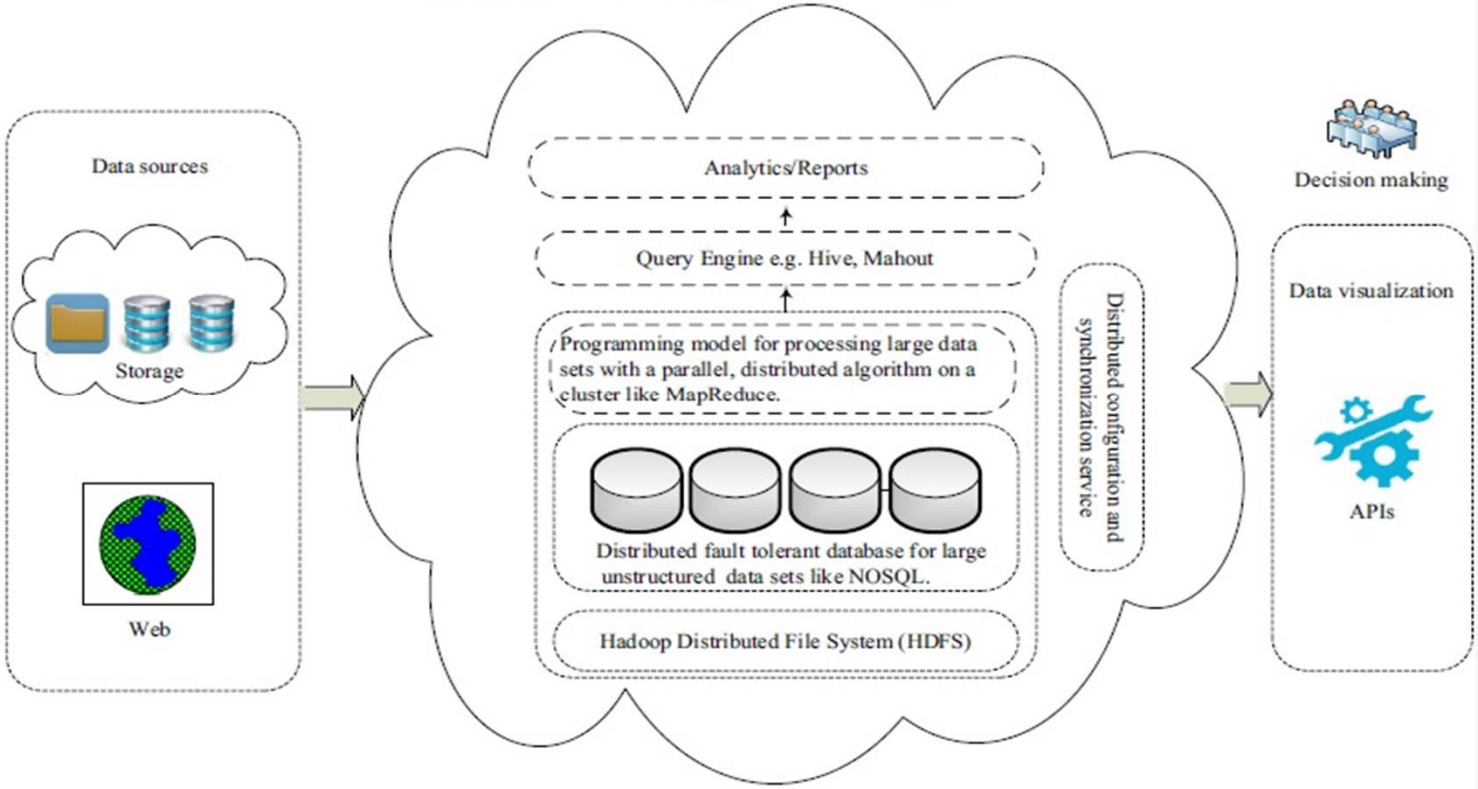


Μοντέλα Cloud Computing

- **Public cloud:** Στο συγκεκριμένο μοντέλο οι υπολογιστικοί πόροι προσφέρονται από τον πάροχο και γίνονται διαθέσιμοι μέσω διαδικτύου στον χρήστη. Ο χρήστης πληρώνει συνήθως ανάλογα με την χρήση (pay as you go).
- **Private cloud:** Στο συγκεκριμένο μοντέλο η παροχή cloud είναι για την αποκλειστική χρήση ενός πελάτη. Μερικές φορές αποτελεί και ιδιοκτησία του πελάτη.
- **Community cloud:** Στο συγκεκριμένο μοντέλο έχουμε τους πελάτες που έχουν παρόμοιες ανάγκες και αποφασίζουν για διάφορους λόγους να μοιραστούν τις φυσικές υποδομές
- **Hybrid cloud:** Αναφέρεται στον οποιοδήποτε συνδυασμό των παραπάνω μοντέλων cloud computing.

Big Data as a Service

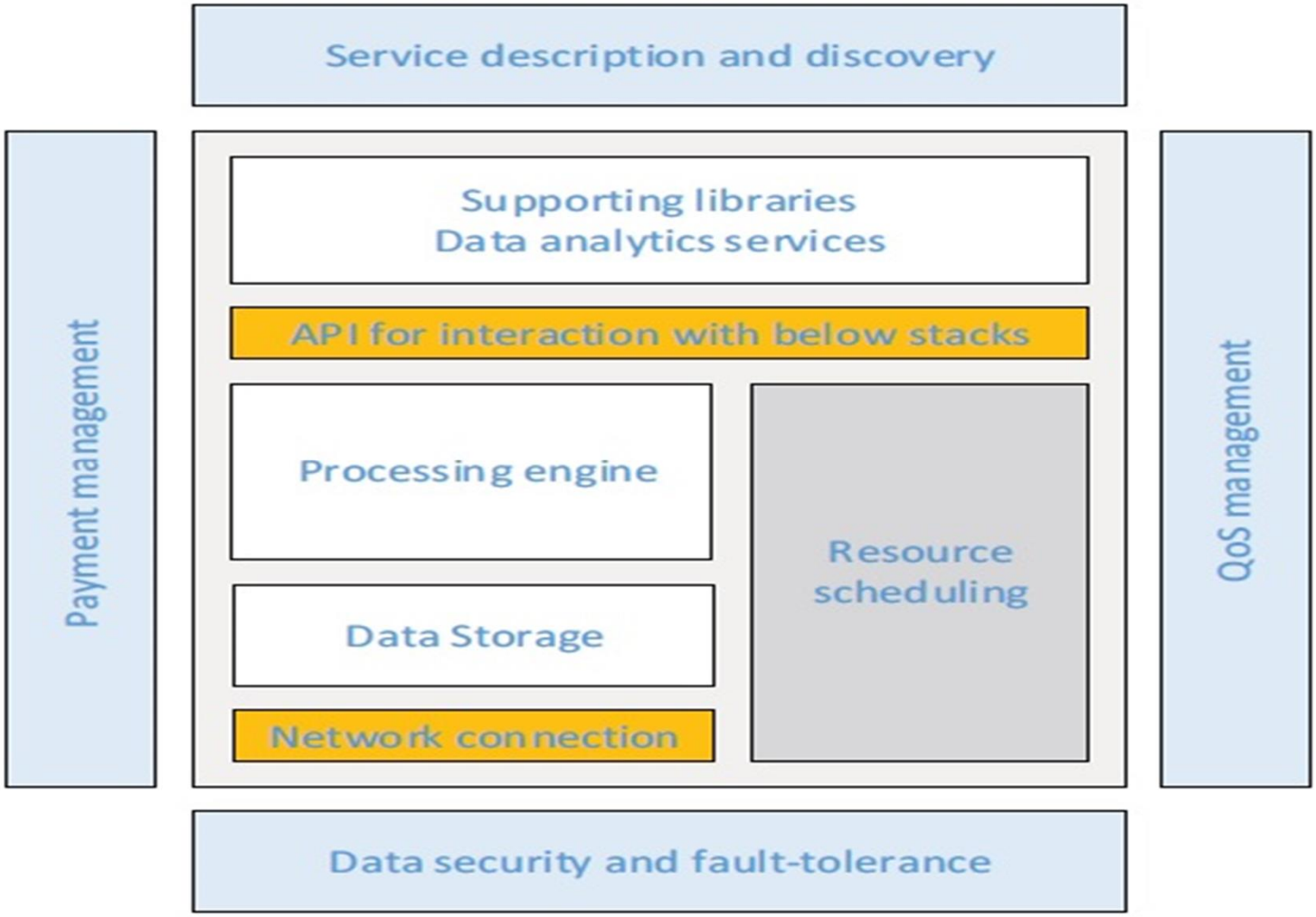
Το Big Data as a Service (BDaaS) προέρχεται από το συνδυασμό κάποιων από τα SaaS (Software as a Service), PaasS (Platform as a Service) και IAAS (Infrastructure as a Service)



Τύποι του Big Data as a Service

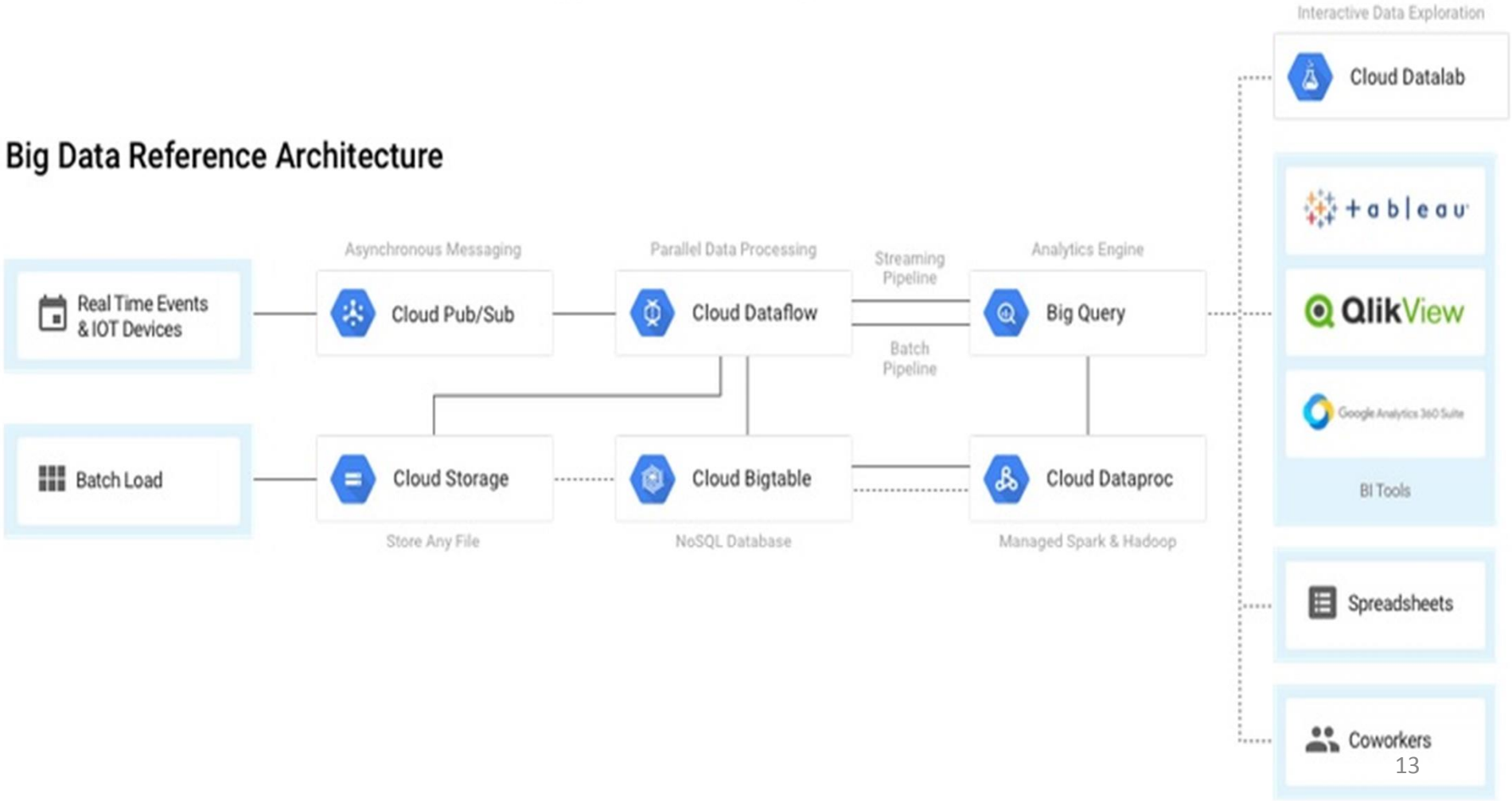
- ✓ **Core BDaaS:** Στη συγκεκριμένη επιλογή παρέχεται μια πλατφόρμα Hadoop με το Yarn και το HDFS ή και το Spark μαζί με μερικές υπηρεσίες από τα οικοσυστήματά τους.
- ✓ **Performance BDaaS:** Η συγκεκριμένη υπηρεσία επικεντρώνεται στη παροχή βοήθειας σε εταιρείες που ήδη εργάζονται με το Hadoop με στόχο να ευθυγραμμίσει τις υποδομές τους και να μεγιστοποιήσουν την απόδοση του Hadoop.
- ✓ **Feature BDaaS:** Εστιάζει στη παραγωγικότητα και σχεδιάζεται για να δώσει ταχύτητα στα μεγάλα δεδομένα με εύκολο και οικονομικό τρόπο. (πχ μέσα από προγραμματιστικές διεπαφές και προσαρμογείς βάσεων δεδομένων που βάζουν σε δεύτερη μοίρα τις τεχνολογίες του Hadoop, αυξάνοντας και σταματώντας τις υπολογιστικές συστάδες του Hadoop όπως απαιτεί ο φόρτος εργασίας)

Αρχιτεκτονική ενός πλαισίου του BDaaS



ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Ηλεκτρονικό Επιχειρείν & Τεχνολογία Καινοτομίας

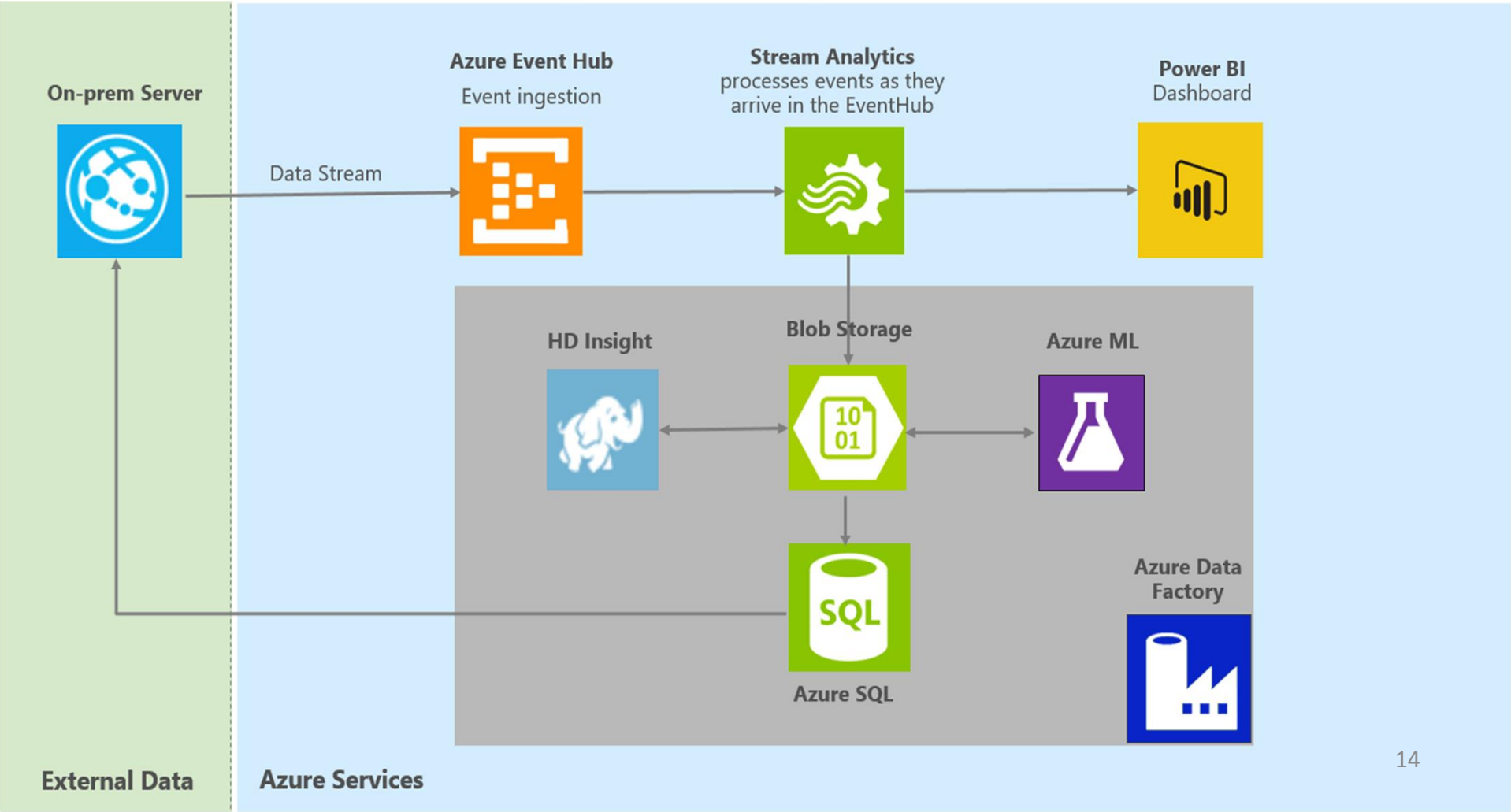
Google Big Data E2E Solution



ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
UNIVERSITY OF MACEDONIA

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Ηλεκτρονικό Επιχειρείν & Τεχνολογία Καινοτομίας

Microsoft Big Data E2E Solution



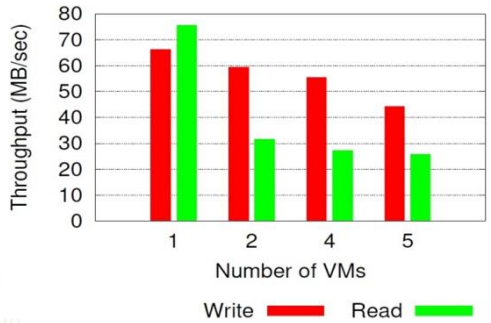
Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

Απόδοση (1)

- ✓ **Σύμφωνα με τους Feller, Ramakrishnan και Morin (2015):**

1. **Όσο αυξάνεται ο αριθμός των εικονικών μηχανών, μειώνεται η απόδοση ανάγνωσης και εγγραφής από το δίσκο.**
2. **Το Hadoop σε εικονικές μηχανές έχει σημαντική πτώση απόδοσης σε αυξανόμενης κλίμακας δεδομένα και στην επεξεργασία αλλά και στη διαχείρισή τους.**

- ✓ **Η προσθήκη επιπλέον κόμβων μειώνει το χρόνο εκτέλεσης μιας διεργασίας και αυξάνει την απόδοση (Gohil, Garg, Panchal 2014) .**



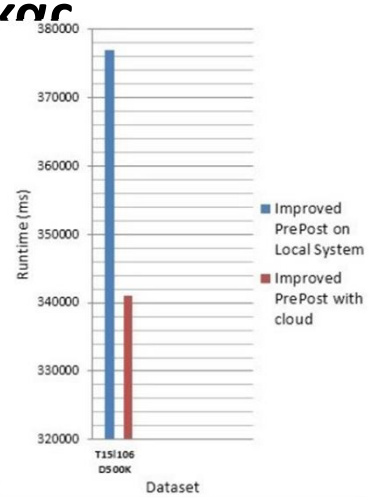
No. of nodes in cluster	Execution time to finish WordCount Job (sec)	Average Execution time to finish WordCount Job (sec)
1	699	684.33
	673	
	681	
3	365	365.7
	364	
	368	
5	192	192.33
	189	
	196	
7	165	163.67
	169	
	157	

Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

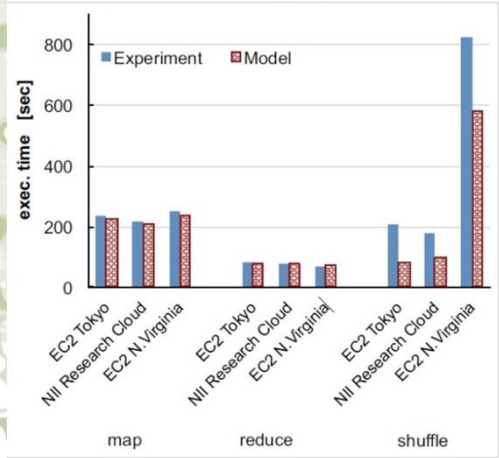
Απόδοση (2)

- ✓ **Thakare, Rathi, S., Sedamkar 2016:**

1. **Το cloud έχει καλύτερη απόδοση και με μερικούς αλγόριθμους εξόρυξης δεδομένων**
2. **Ο PrePost τρέχει σε περιβάλλον cloud αρκετά πιο γρήγορα για το ίδιο σετ δεδομένων**



- ✓ **Η απόδοση του Hadoop σε υβριδικό cloud εξαρτάται από τον τύπο των εφαρμογών που εκτελούνται και μπορεί να υπάρχει ελπίδα για βελτίωση της απόδοσης μόνο αν χρησιμοποιείται αυτός ο συνδυασμός υποδομής για συγκεκριμένες εφαρμογές και μετά από επιλογή του σωστού παρόχου cloud ώστε να μειωθεί η επίδραση των χρόνων επικοινωνίας ανάμεσα στη φάση shuffle του Hadoop. (Ohnaga, Aida, , Abdul-Rahman, 2015)**



Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

Απόδοση (3)

- ✓ **Στον αλγόριθμο μηχανικής μάθησης Bayesian, τα αποτελέσματα ήταν εντυπωσιακά καθώς το Spark σε περιβάλλον cloud ήταν περίπου 2,3 φορές πιο γρήγορο από ότι σε τοπικό περιβάλλον (Ahmed, Ismail, Hyder, Sheraz, Fouq, 2016)**

Spark HiBench Execution Results on Conventional Cluster

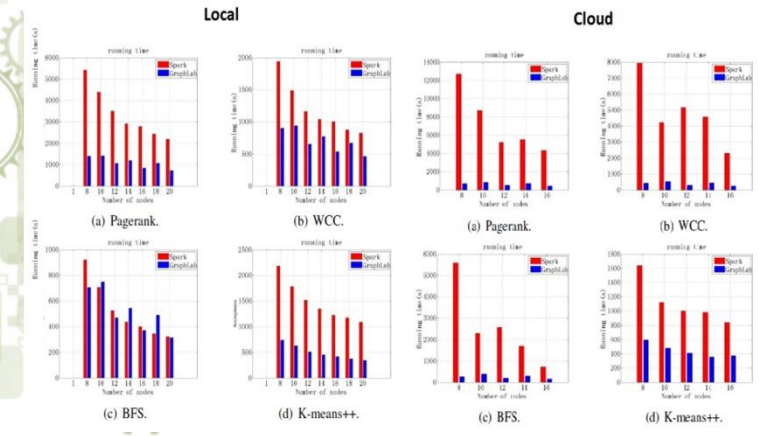
Benchmarks	Input Data Size(bytes)	Duration(s)	Throughput(bytes/s)	Throughput/node
Bayesian	450822628	231.789	1944969	648323
PageRank	259928115	576.603	450792	150264
TeraSort	3200000000	461.694	6930997	2310332
WordCount	2204456452	142.487	15471281	5157093

Spark HiBench Execution Results on Cloud

Benchmarks	Input Data Size(bytes)	Duration(s)	Throughput(bytes/s)	Throughput/node	Speedup
Bayesian	375706036	99.165	3788695	1262898	2.3374
PageRank	259928115	346.854	749387	249795	1.6623
TeraSort	3200000000	136.431	23455079	7818359	3.3840
WordCount	2204457594	68.753	32063438	10687812	2.0724

- ✓ **Οι Wei, Chen, Zhou, Zhou, Jianhua, (2016) παρατήρησαν στις δοκιμές τους ότι:**

1. **Ο χρόνος εκτέλεσης το Spark, που είναι γενικής χρήσης πλατφόρμα, για αλγόριθμους γράφων σε εικονικές μηχανές σε περιβάλλον cloud, ανεβαίνει σχεδόν 100% σε σύγκριση με την εκτέλεση σε τοπικές εικονικές μηχανές.**



Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

Τοποθεσία Δεδομένων

- ✓ **Αποδείχθηκε ότι η απομακρυσμένη αποθήκευση των δεδομένων αύξησε την απόδοση σε σχέση με το HDFS σε τοπικό δίσκο (Accenture, 2014)**

- ✓ **Οι Ma, Fan, Liu, Jiang, Peng, K., (2016) δίνουν μεγάλη σημασία στη τοποθεσία των δεδομένων και θεωρούν ότι είναι μια σημαντική πρακτική για τη βελτίωση της επίδοσης του MapReduce.**

1. **Προτείνουν το vLocality που απαιτεί αναθεώρηση της τοποθεσίας δεδομένων τριών επιπέδων όπως και του προγραμματισμού των εργασιών.**
2. **Αποδείχθηκε ότι βελτιώνει σημαντικά όλες τις εφαρμογές που εκτελέστηκαν κατά τις δοκιμές με ποσοστό στη χειρότερη περίπτωση λίγο κάτω από το 10%.**

Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

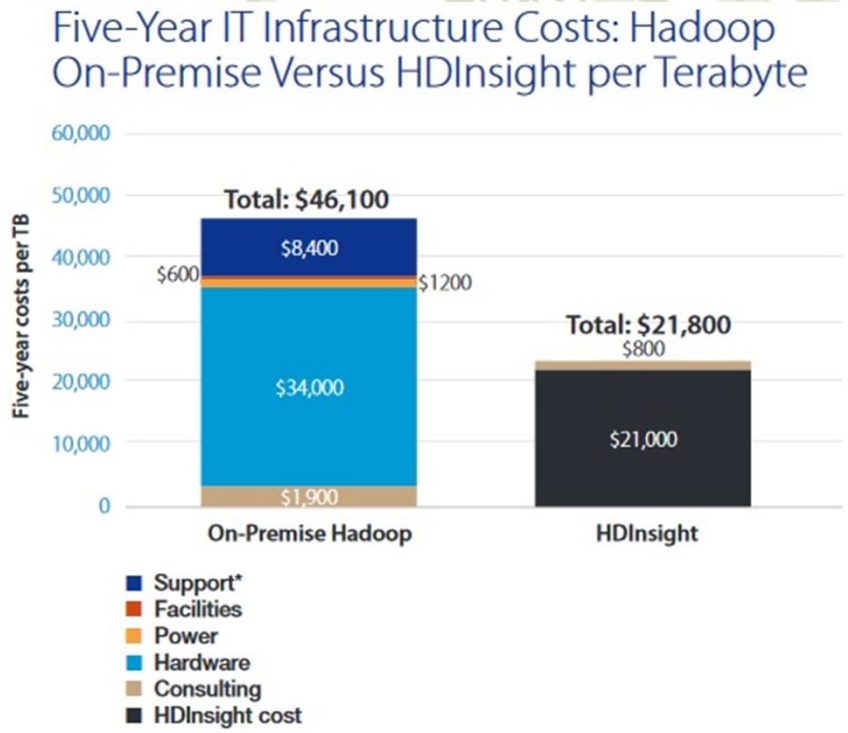
Ασφάλεια

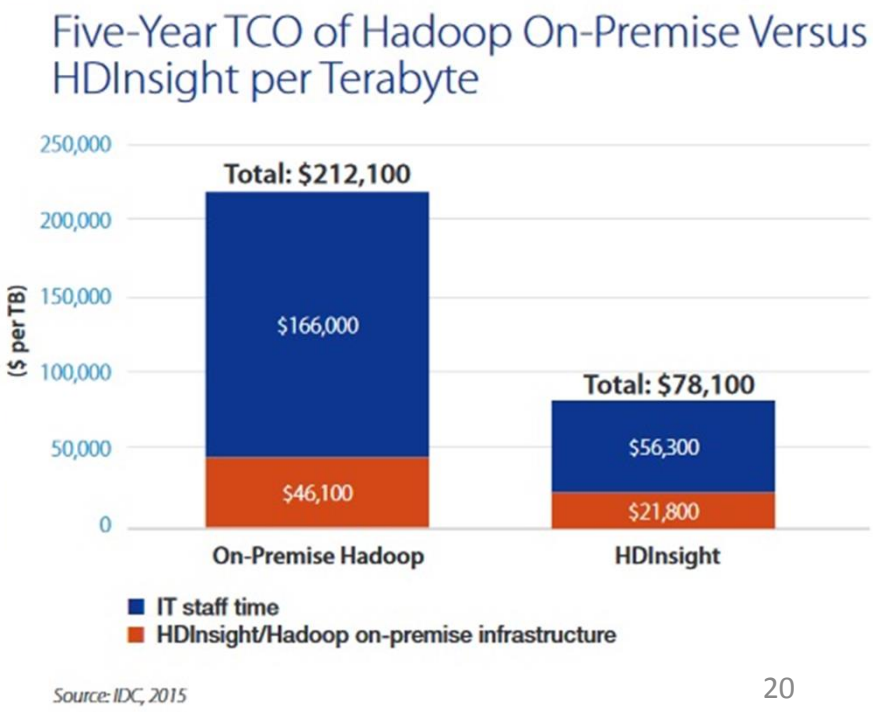
- ✓ **Ο κυριότερος λόγος μη υιοθέτησης των υπηρεσιών cloud απο οργανισμούς**
- ✓ **Προβλήματα σε επίπεδο δικτύου, ταυτοποίησης και δεδομένων (Dewangan & Negi, 2016)**
- ✓ **Συνεχώς προτείνονται νέοι τρόποι για μεγαλύτερη ασφάλεια.**

Σύγκριση του τοπικού περιβάλλοντος και του υπολογιστικού νέφους για εφαρμογές δεδομένων μεγάλης κλίμακας

Συνολικό Κόστος

- ✓ **Αποδείχθηκε ότι το περιβάλλον υπολιγστικού νέφους παρέχει καλύτερη αναλογία Απόδοσης – Συνολικού Κόστους (Accenture, 2014)**





Συμπεράσματα

- ✓ **Τα Big Data είναι παντού και είναι απαραίτητα για κάθε οργανισμό**
- ✓ **Το Cloud Computing είναι μια αποτελεσματική λύση στο θέμα των αναγκαίων πόρων αποθήκευσης και επεξεργασίας των Big Data**
- ✓ **Οι εταιρείες προσφέρουν υπηρεσίες που κάνουν τη διαχείριση των Big Data προσιτή ακόμη και σε μικρούς οργανισμούς δίνοντας τους τη δυνατότητα για επιχειρηματική εξέλιξη**

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
UNIVERSITY OF MACEDONIA

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Ηλεκτρονικό Επιχειρείν & Τεχνολογία Καινοτομίας