



















Big Data

Χρήστος Γκόγκος

11/5/2020

https://github.com/chgogos/big_data

Digital disruption

 UBER World's largest taxi company Owns NO 	 airbnb World's largest Accommodation provider Owns NO 	 skype WeChat World's largest Phone companies Owns NO 	 Alibaba Group World's most Valuable retailer Owns NO 
 facebook Most popular Media owner Owns NO 	 SocietyOne World's fastest Growing bank Owns NO 	 NETFLIX World's largest movie house Owns NO 	 Apple Google World's largest Software vendors Owns NO 

Τι είναι τα Big Data;

- Τα Big Data είναι τεράστιες ποσότητες δομημένων, ημιδομημένων και αδόμητων δεδομένων
- Τα Big Data εμφανίζονται ως αποτέλεσμα τεχνολογικών εξελίξεων που συνέβησαν τα τελευταία 50+ έτη
- Δεδομένα τα οποία δεν μπορούν να αποθηκευτούν και να επεξεργαστούν από ένα μόνο (μεγάλο) υπολογιστικό σύστημα κατηγοριοποιούνται ως Big Data

2017 *This Is What Happens In An Internet Minute*



2018 *This Is What Happens In An Internet Minute*



2019 *This Is What Happens In An Internet Minute*



2020 *This Is What Happens In An Internet Minute*



Σημεία καμπής

- 1991: Internet
- 1996: κόστος ψηφιακής αποθήκευσης < κόστος έντυπης αποθήκευσης
- 1997: Google
- 1999: Big Data, Internet of Things
- 2003: Google File System paper (by Google)
- 2004: MapReduce paper (by Google)
- 2005: Hadoop
- 2006: Cloud Computing
- 2014: κίνηση στο διαδίκτυο από κινητά > κίνηση από σταθερούς Η/Υ
- 2015: Deep Learning

Το κόστος της υπολογιστικής επεξεργασίας και αποθήκευσης έφτασε σε κομβικό σημείο ανάμεσα στο 2008 και το 2010

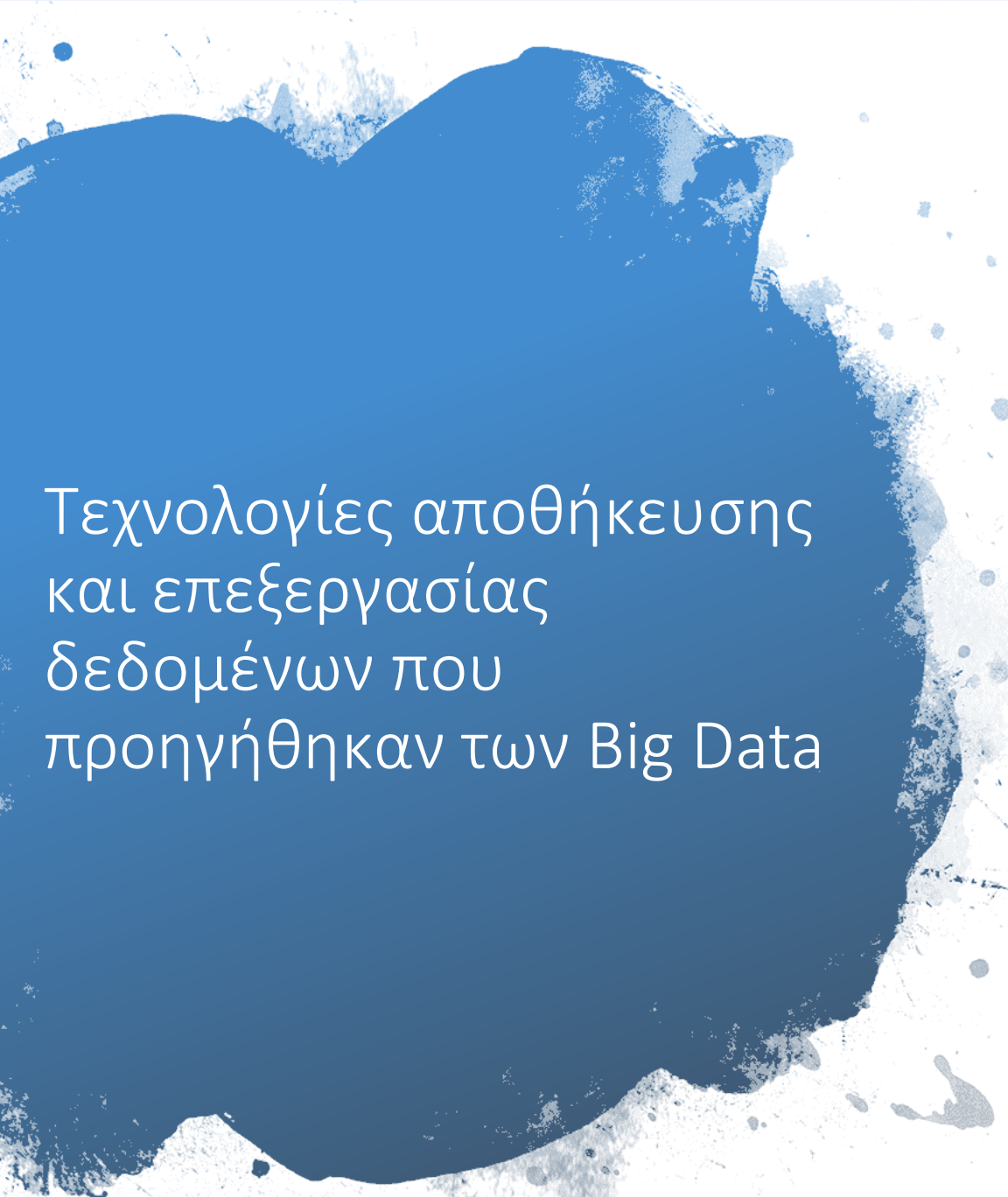


Περισσότερες επιχειρήσεις έχουν πλέον τη δυνατότητα να διαχειρίζονται Big Data

Σημαντικές αλλαγές των τελευταίων ετών

- Πολλαπλασιασμός ψηφιακών δεδομένων που παράγουν οι χρήστες μέσω της συμπεριφοράς τους (κοινωνικά δίκτυα)
- Μείωση κόστους αποθήκευσης - επεξεργασίας
- Μείωση τιμών αισθητήρων
- Σημαντική πρόοδος σε αλγόριθμους μηχανικής μάθησης





Τεχνολογίες αποθήκευσης
και επεξεργασίας
δεδομένων που
προηγήθηκαν των Big Data

- Ιεραρχικές Βάσεις Δεδομένων
- Δικτυωτές Βάσεις Δεδομένων
- Σχεσιακές Βάσεις Δεδομένων
- Object Oriented Βάσεις Δεδομένων
- Data warehouses και data marts
(ημερήσια ή εβδομαδιαία ενημέρωση)
- Online Transactional Processing (OLTP)
- Online Analytical Processing (OLAP)
- Επιχειρηματική Ευφυΐα (Business Intelligence)

Τα 3V των big data (1:Volume=όγκος)

Volume

Velocity

Variety

- Ο όγκος στα Big Data αφορά την ανάγκη αποθήκευσης και επεξεργασίας Terabytes ή και Petabytes δεδομένων
- Το 2020 αναμένεται να υπάρχουν 50 δισεκατομμύρια συσκευές συνδεδεμένες στο διαδίκτυο
- Η ποσότητα των δεδομένων που συλλέγονται αυξάνεται συνεχώς

Big Data are **high-volume**, **high-velocity**, and/or **high-variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization (Gartner 2012)

Τα 3V των big data (2:Velocity=ταχύτητα)

Volume

Velocity

Variety

- Η ταχύτητα στα Big Data αφορά τον υψηλό ρυθμό με τον οποίο παράγονται νέα δεδομένα
- Ορισμένες εφαρμογές απαιτούν λήψη αποφάσεων σε πραγματικό χρόνο (online):
 - ανίχνευση απάτης (credit card fraud detection)
 - recommendation systems (ηλεκτρονικές αγορές – θέαση περιεχομένου)

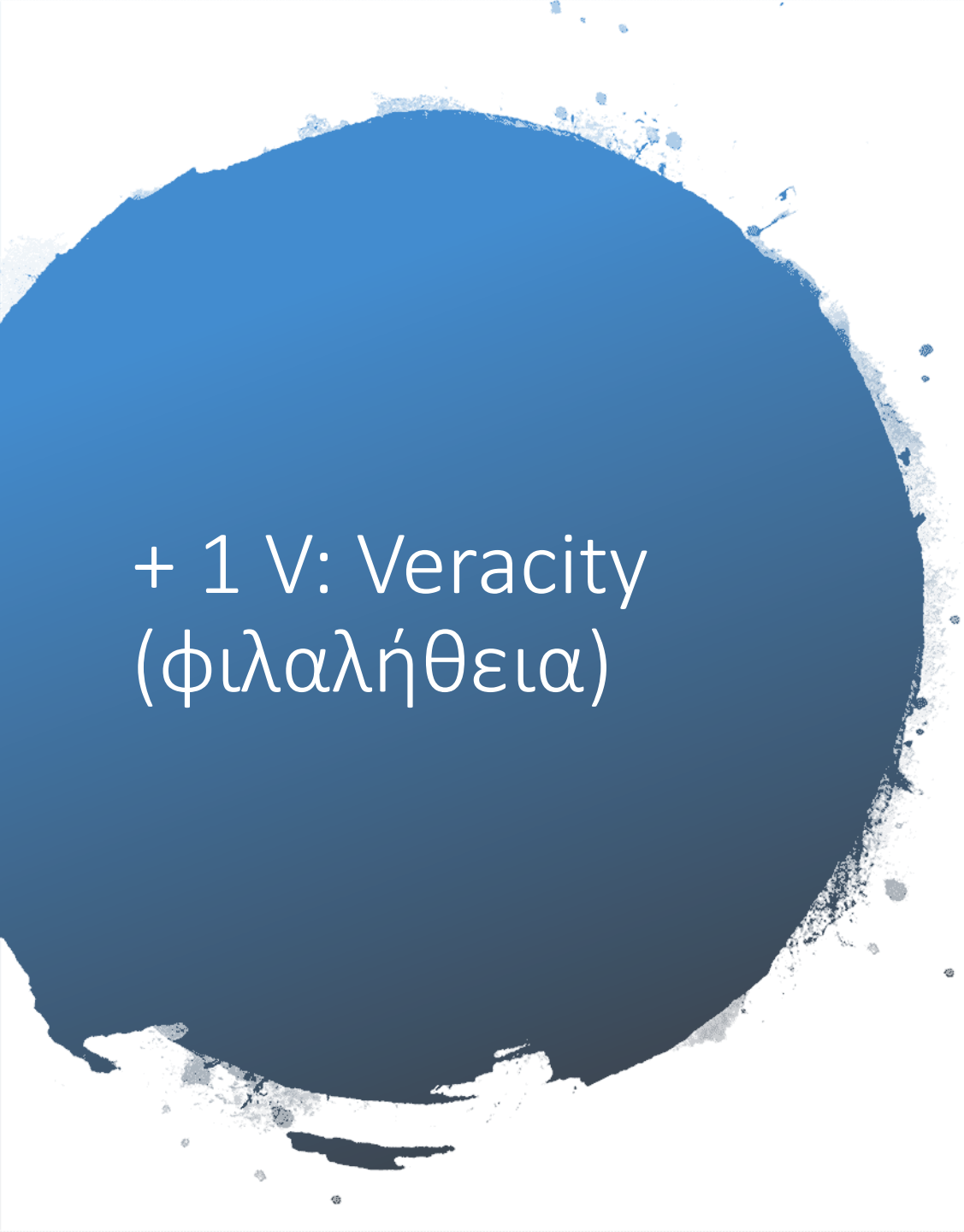
Τα 3V των big data (3:Variety=ποικιλομορφία)

Volume

Velocity

Variety

- Η ποικιλομορφία στα Big Data αφορά τη συγκέντρωση δεδομένων από διάφορες πηγές και σε διάφορες μορφές
- Μορφές δεδομένων:
 - Κείμενο
 - Δεδομένα πλοήγησης στο διαδίκτυο
 - Tweets - Posts
 - Δεδομένα αισθητήρων
 - Ήχος
 - Βίντεο
 - Αρχεία καταγραφής (logs)
 - ...



+ 1 V: Veracity
(φιλαλήθεια)

Η φιλαλήθεια αφορά τη συντεταγμένη προσπάθεια αποφυγής συσσώρευσης μη έγκυρων δεδομένων καθώς και δεδομένων που απλά προσθέτουν «θόρυβο» στο σύστημα.

Σχετίζεται με την απομάκρυνση δεδομένων που είναι: ασυνεπή, σε πολλαπλά αντίτυπα, που έχει λήξει ο κύκλος ζωής τους κ.α.



Big Data και επιχειρήσεις

- Δεν είναι όλα τα δεδομένα Big Data
- Το σύστημα Big Data που θα προταθεί θα πρέπει:
 - να παρέχει προστιθέμενη αξία για την επιχείρηση
 - να είναι οικονομικό
 - να λειτουργεί με αποδεκτή ταχύτητα



Τεχνολογίες που σχετίζονται με τα Big Data

- Hadoop
- Map Reduce
- Compute Engines (Spark, Tez, Flink, Storm, ...)
- In-memory technology
- NoSQL
- Cloud computing
- Containers
- Serverless
- IoT
- Edge computing (fog computing)
- Machine learning – analytics
- Deep learning – business intelligence