











Big Data

Χρήστος Γκόγκος

19/5/2022

https://github.com/chgogos/big_data

Ψηφιακή επανάσταση (digital disruption)

 UBER World's largest taxi company Owns NO Taxis	 airbnb World's largest Accommodation provider Owns NO Real estate	 skype WeChat World's largest Phone companies Owns NO Telco infra	 Alibaba Group World's most Valuable retailer Owns NO Inventory
 facebook Most popular Media owner Owns NO Content	 SocietyOne World's fastest Growing bank Owns NO Actual money	 NETFLIX World's largest movie house Owns NO Cinemas	 Apple Google World's largest Software vendors Owns NO Apps

Το λογισμικό «τρώει» τον κόσμο

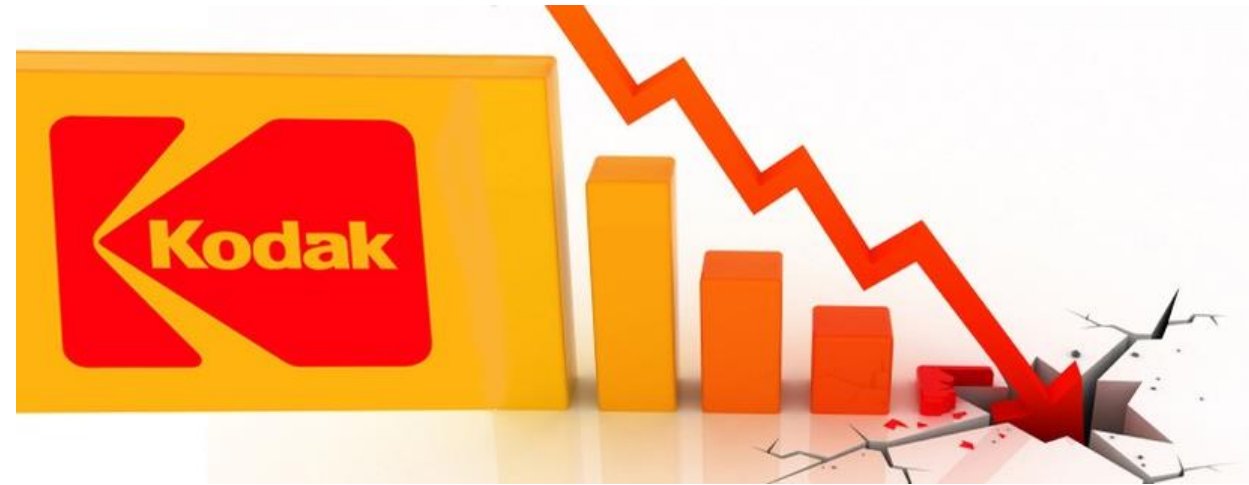
- “[Why Software Is Eating the World](#)”, **Marc Andreessen**, The Wall Street Journal, August 20, 2011
 - «Οι εταιρείες όλων των ειδών πρέπει να γνωρίζουν ότι η επανάσταση του λογισμικού έρχεται»
 - «Αυτό αφορά ακόμα και τις εταιρείες οι οποίες στηρίζουν τη λειτουργία τους σε λογισμικό σήμερα»
- “[10 years later, software really did eat the world](#)”, CIODIVE.com, April 19, 2021



<https://a16z.com/>

Παράδειγμα αποτυχίας προσαρμογής στη νέα πραγματικότητα

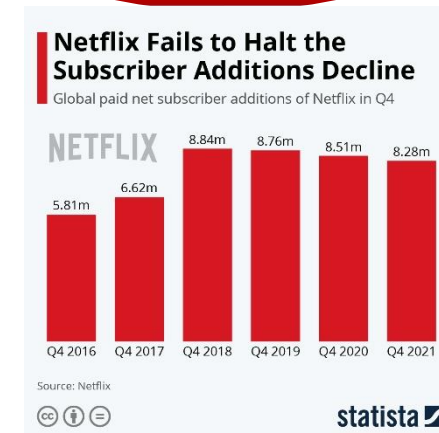
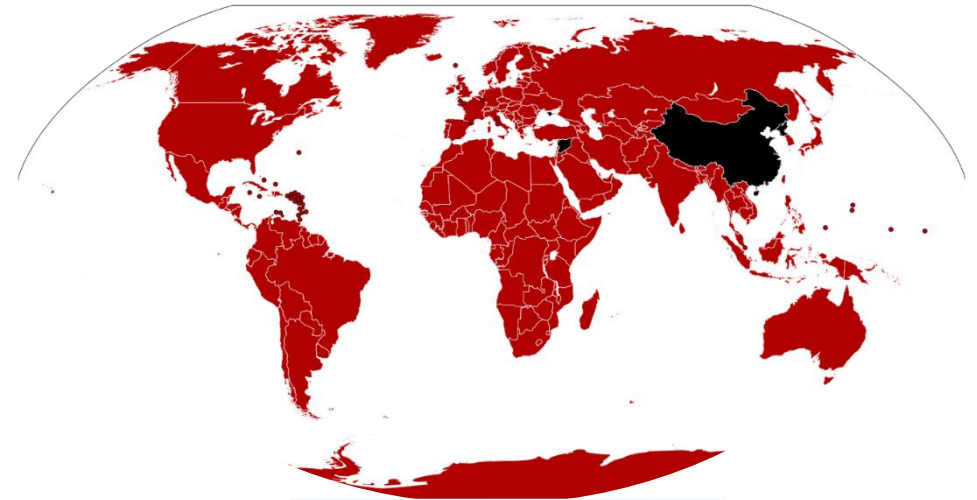
- Η εταιρία Kodak δεν προσαρμόστηκε στο ψηφιακό μετασχηματισμό που επέφερε η χρήση των κινητών και smartphones ως καμερών
- Οδηγήθηκε σε πτώχευση το 2012



<https://startsmallscalefast.com/agilebusiness/real-disruption-lessons-kodak/>

Παράδειγμα επιτυχίας (;) προσαρμογής στη νέα πραγματικότητα (Netflix)

- 1999: μηνιαίες συνδρομές, αποστολή DVDs με courier, προτάσεις για νέες ταινίες
- 2007: Video on demand μέσω του διαδικτύου
- 2009: Netflix prize (\$1.000.000)
- 2013: Παραγωγή περιεχομένου, σειρά House of Cards
- 2016: Παραγωγή 126 σειρών
- 2019: 137 εκατομμύρια συνδρομητές (με πληρωμή)
- 2021(Απρίλιος): 208 εκατομμύρια συνδρομητές
- 2022(Μάρτιος): 221,6 εκατομμύρια συνδρομητές

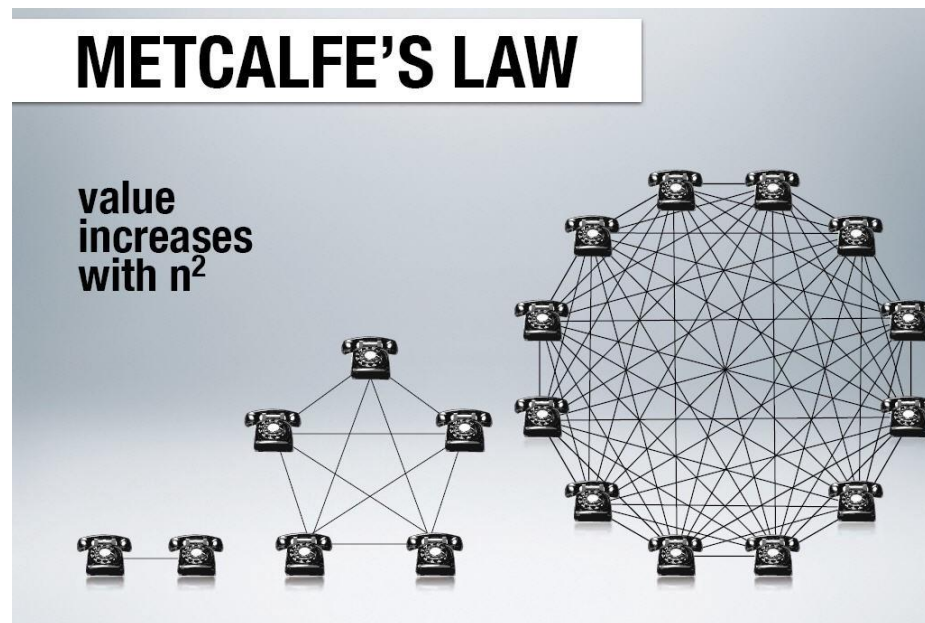
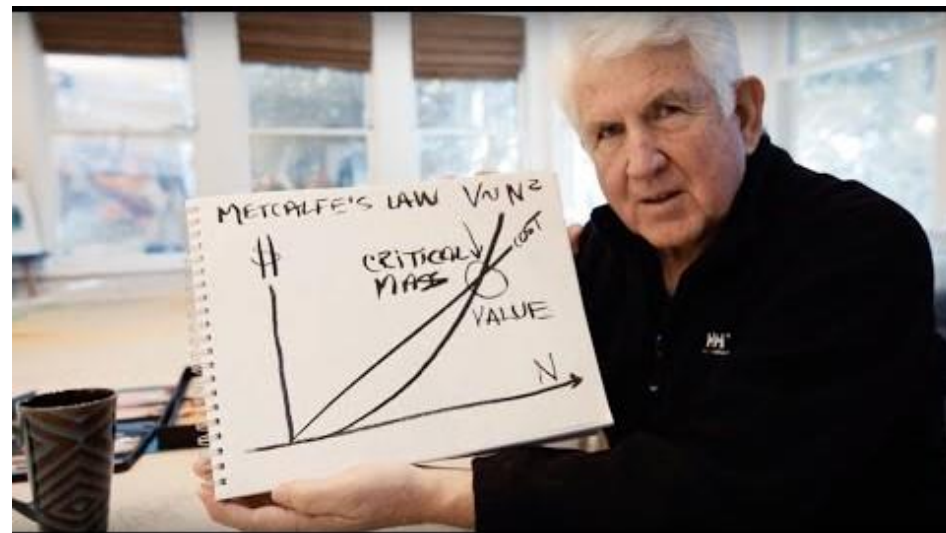


<https://finance.yahoo.com/quote/NFLX>

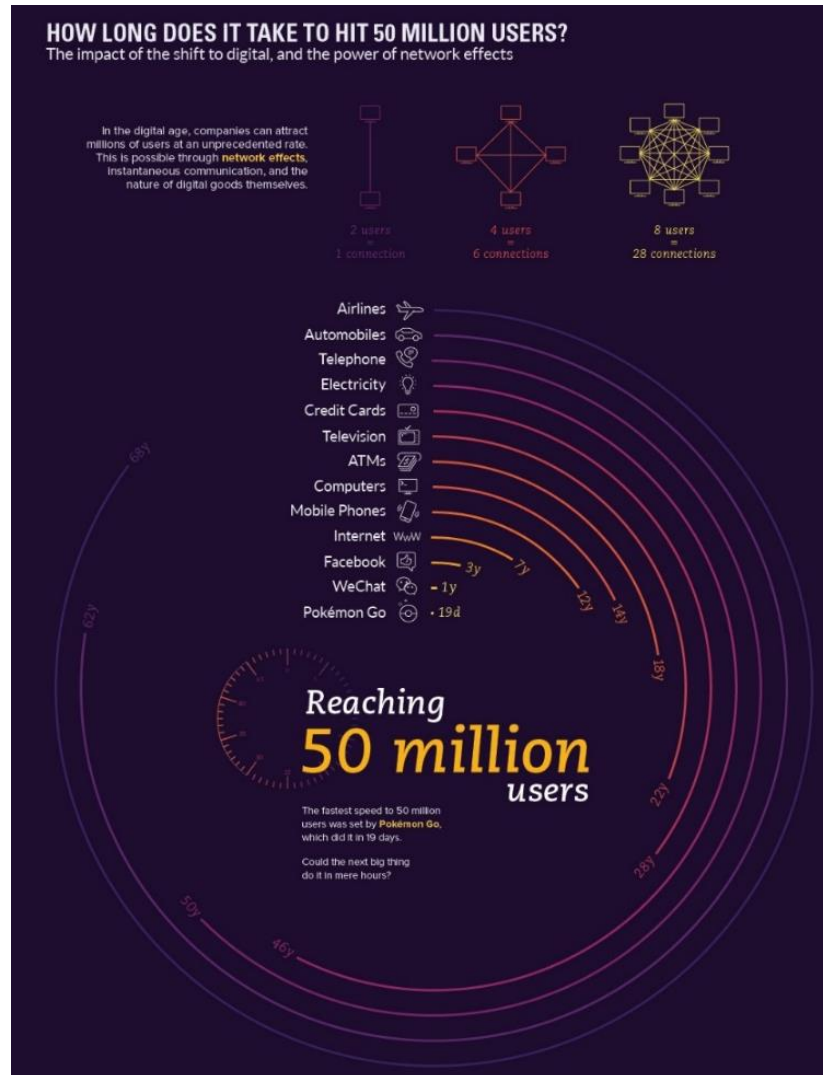
Ο νόμος του Metcalfe

Ο νόμος του [Metcalfe](#) αφορά την αύξηση του αριθμού των συνδέσεων αλλά και της αξίας που δημιουργείται καθώς ένα δίκτυο μεγαλώνει. Ένα δίκτυο με n κόμβους αποκτά αξία της τάξης του n^2

$$\frac{n(n - 1)}{2}$$



Σε πόσο χρόνο αποκτήθηκαν 50 εκατομμύρια πελάτες;



Product / Technology	Time it Took to Hit 50 Million Users
Airlines	64 years
Automobiles	62 years
Telephone	50 years
Electricity	46 years
Credit Cards	28 years
Television	22 years
ATMs	18 years
Computers	14 years
Mobile Phones	12 years
Internet	7 years
Facebook	4 years
WeChat	1 year
Pokemon Go	19 days

<https://www.visualcapitalist.com/how-long-does-it-take-to-hit-50-million-users/>

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
- Ο όρος Big Data χρησιμοποιείται ήδη από τη δεκαετία του 90, η ευρεία χρήση του όρου αποδίδεται στον [John Mashley](#)

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*



2021 *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
- 2016: VR/AR (Oculus Rift)
- 2017: AlphaZero
- 2020: Proof of Stake
- 2021: Metaverse

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

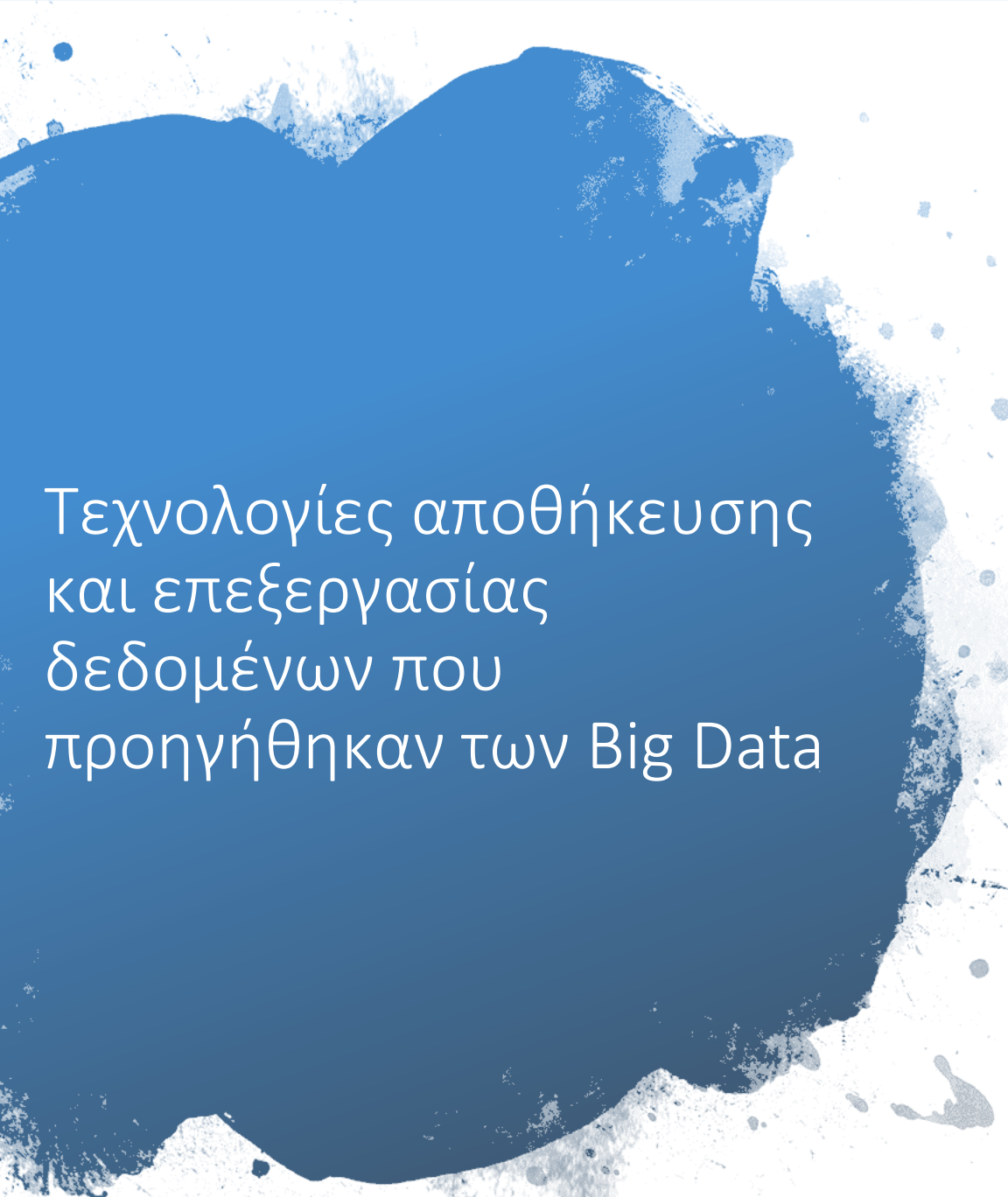


Περισσότερες επιχειρήσεις έχουν πλέον τη δυνατότητα να διαχειρίζονται Big Data

Σημαντικές αλλαγές των τελευταίων ετών

- Πολλαπλασιασμός ψηφιακών δεδομένων που παράγουν οι χρήστες μέσω της συμπεριφοράς τους (επισκέψεις σε ιστοσελίδες, ηλεκτρονικές συναλλαγές, κοινωνικά δίκτυα κ.α.)
- Μείωση κόστους αποθήκευσης - επεξεργασίας
- Μείωση τιμών αισθητήρων
- Φορέσιμες συσκευές
- 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
- Αντικειμενοστραφείς 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 2030 αναμένεται να υπάρχουν 500 δισεκατομμύρια συσκευές συνδεδεμένες στο διαδίκτυο
- Η ποσότητα των δεδομένων που συλλέγονται αυξάνεται συνεχώς

Τα μεγάλα δεδομένα είναι δεδομένα υψηλού όγκου (**high-volume**), υψηλής ταχύτητας (**high-velocity**), και/ή υψηλής ποικιλότητας (**high-variety**) που απαιτούν νέες μορφές επεξεργασίας οι οποίες επιτρέπουν βελτιωμένες διαδικασίες λήψης αποφάσεων, δημιουργία διαίσθησης και βελτιστοποίηση διαδικασιών (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 κατά περίπτωση προτείνεται 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