

Vilnius universitetas Matematikos ir informatikos fakultetas Informatikos katedra



DIDIEJI DUOMENYS (BIG DATA). KAS TAI?

doc. dr. Olga Kurasova

<u>Olga.Kurasova@mii.vu.lt</u>

2017



Didieji duomenys. Jų atsiradimo šaltiniai

- Modernios technologijos leidžia generuoti milžiniškus duomenų kiekius.
- Pagrindiniai didžiųjų duomenų šaltiniai:
 - Astronomija,
 - Meteorologija,
 - Genų inžinerija,
 - Medicina,
 - Bankinės ir finansinės sistemos,
 - Socialiniai tinklai,
 - Telekomunikacija,
 - Daiktų internetas (Internet of Things, IOT)
 - Saitynas (web).
 - Kiti



Didieji duomenys



H2020-ICT-2015

- "big data" is when the size of the data itself becomes part of the problem
- "big data" is data that becomes large enough that it cannot be processed using conventional methods

Didžiųjų duomenų ištakos

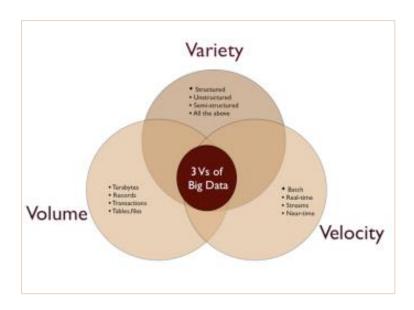


- o Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources (1997).
- Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Didieji duomenys (3V)

Big data can be described by the following characteristics:

- o Volume (didžiulis duomenų kiekis).
- Variety (plati duomenų įvairovė).
- Velocity (nuolat atsirandantys nauji duomenys).



Didieji duomenys (3V)

Gartner

Application Delivery Strategies



Date: 6 February 2001

File: 949 Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

Very difficult to define (precisely):

- o data is "**big**" if it defies traditional processing & storage paradigms bigness becomes part of the problem.
- o the "3Vs": volume (size), velocity (bytes/s), variety (database, jpeg, video, numbers, text in language X...).
- ...to which we add the 4th V to denote creation of Value (by linking, aggregating, analysing, visualizing...).



Didieji duomenys (4V)



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

TRM

Didieji duomenys (5V)

- Volume from terabytes to petabytes and up.
- Variety an expanding universe of data types and sources.
- Velocity accelerated data flow in all directions.
- Variability inconsistent data flows with periodic peaks.
- Complexity the need to correlate and share data across entities.

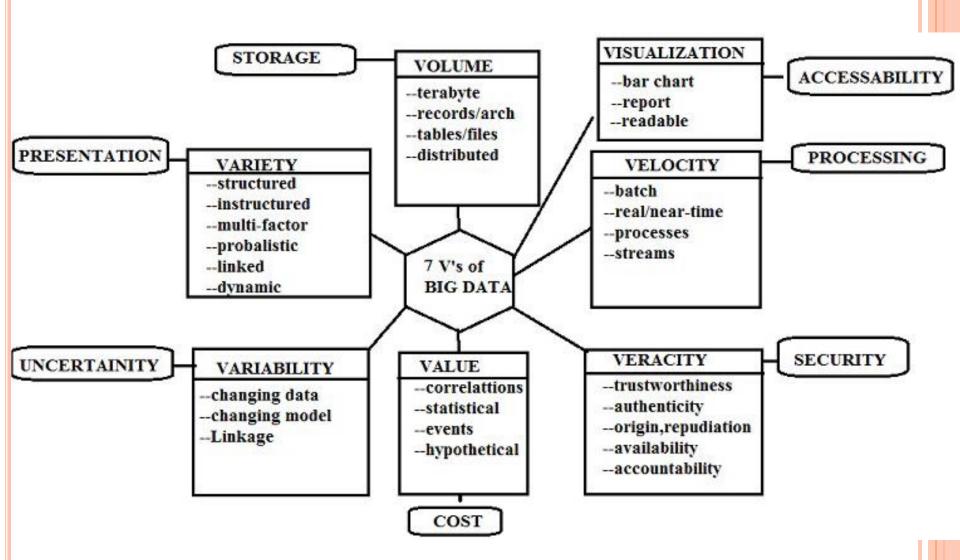




"Big data is not only about analytics, it's about the whole pipeline. So when you think about big data solutions, you have to think about all the different steps: collect, store, organize, analyze, and share," said Amazon CTO Werner Vogels.



Didieji duomenys (7V)

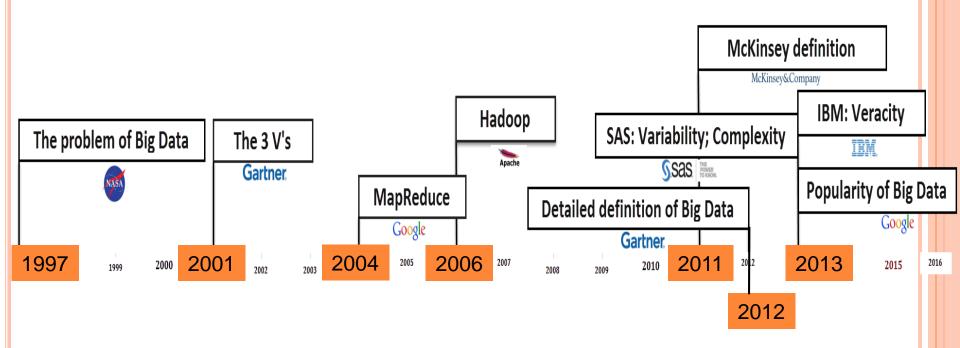


Didieji duomenys

o Big data refers to data being collected in ever-escalating volumes, at increasingly high velocities, and for a widening variety of unstructured formats and variable semantic contexts. Big data can be historical (meaning stored data) or real-time (meaning streamed directly from the source).

• Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Didieji duomenys: anksčiau ir dabar



What Types of Data are in Big Data?

- Structured Data
 - Tables, Relational Data ...
- Unstructured Data
 - Raw text, Images, Video, Audio ...
- Semi-structured
 - Hybrid data, such as documents with tables ...
- Metadata
 - Structured data about data
- Streaming Data
 - Data that moves across networks at high speed
- Temporal Data
 - Data including trends/activities in time
- Geospatial Data
 - Data that includes information on positions in space
- Many others ...



Matavimo vienetai

Reikšmė	Žymėjimas	Pavadinimas
1000	kB	kilobyte
1000^2	MB	megabyte
1000^{3}	GB	gigabyte
1000^{4}	TB	terabyte
1000^5	PB	petabyte
1000^{6}	\mathbf{EB}	exabyte
1000^{7}	ZB	zettabyte
1000^{8}	YB	youttabyte

How Big is Big Data?

- o Internet traffic is now ∼5 Zettabyte per year (IBM)
- Visa processes 150 Million transactions per day (VISA)
- Library of Congress holds 3.2 Petabytes of data
- 207 Terabytes of video loaded daily on YouTube (2012)
- 50 billion devices connected to the Internet by 2020 (IDC)
- 50 Billion photos on Facebook in 2010
- 400 Million Tweets per day (Washington Post)

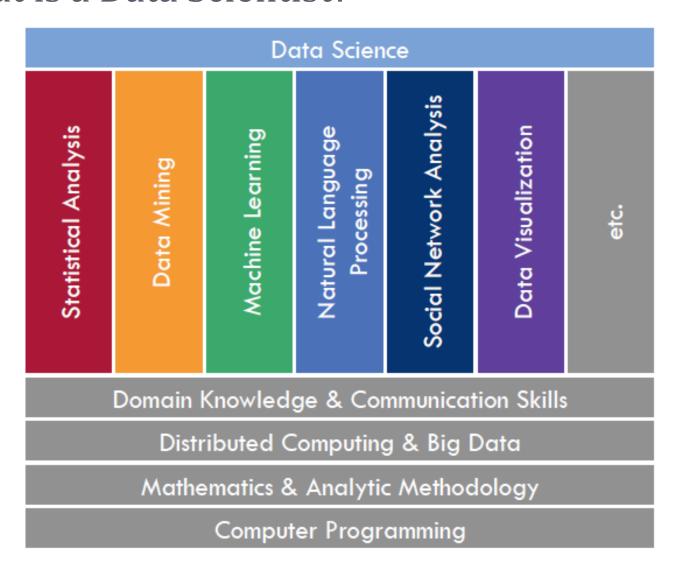


How Big is Big Data?

- Seagate sold 330 Exabytes of hard drives in 2011
- LHC produces 500 Exabytes of particle collision data per day CERN!
- o iPhone 5s: 76 Gigaflops
- Fastest supercomputer: 50 Petaflops
- Interesting Comparison: Human Brain has 100 Billion Neurons (100 Giga-Neurons), 100 Trillion Synapses (100 Tera-Synapses), neurons "fire" 1-1000 times/second (100 Giga-fires to 100 Tera-fires per second)



What is a Data Scientist?



Su didžiaisiais duomenimis susiję sprendimai ir technologijos























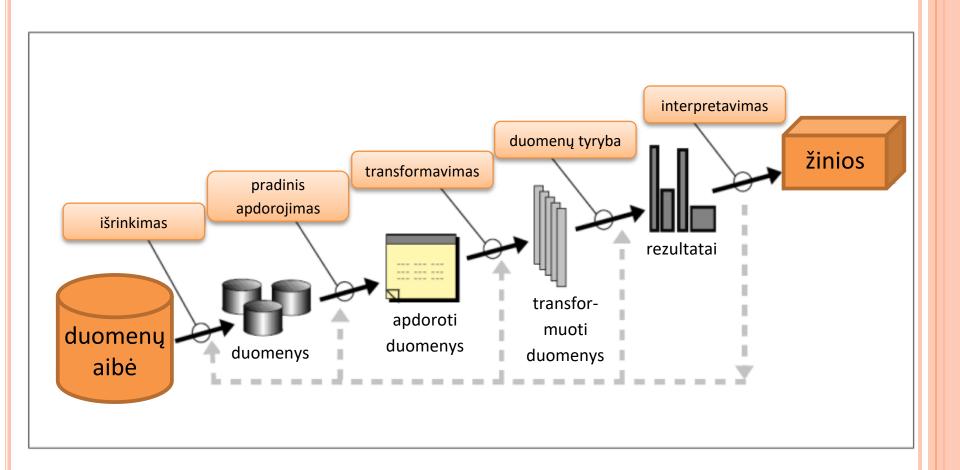




Duomenų tyryba didžiųjų duomenų eroje

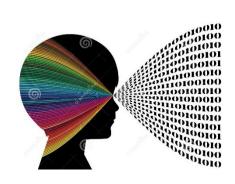
1,11091 7,04704 2,68999 3,51433 6,24851 -7,70285 3,98986 -15,0384 5,9192 5,87187	
-5,15074 1,14707 1,49299 2,55285 -2,27442 -1 1,11091 7,04704 2,68999 3,51433 6,24851 -7,70285 3,98986 -15,0384 5,9192 5,8	7187
1,91532 -4,58991 -1,90508 4,07452 -11,9971 -7 -5,15074 1,14707 1,49299 2,55285 -2,27442 -14,6506 7,86902 -7,36946 9,63875 7,1	1384
0,020048 0,630172 1,08796 2,26984 -4,79242 -8 1,91532 -4,58991 -1,90508 4,07452 -11,9971 -7,92136 9,69161 0,222057 10,8834 5,8	7123
	5558
	3966
-3,13074 1,14707 1,4 -5 15074 1 14707 1 49399 2 1,72396 0,97497 5,85485 5,75813 1,33302 -4 -0,10323 -10,4471 -3,17563 5,31002 0,14175 -9,40686 9,52215 1,64615 -1,88292 6,6	4164
2,22032 7,0 7,0 7,0 7,0 7,0 7,0 7,0 7,0 7,0 7,0	7983
-5,15074 1,14707 0,020048 0,630172 1,0 0,020048 0,630172 1,0 0,020048 0,630172 1,08796 2 -0,83436 1,05809 7,03569 2,13677 -4,65407 -8 4,52856 11,1303 -0,52264 3,58449 -5,35348 -9,28854 9,6102 -3,04153 5,72042 5,2	8749
1,51352 -4,58351 0,48333 10,4433 0,4 -0,89494 8,11653 -6,95531 6, 2,58662 5,18565	2663
0,020048 0,630172 -0,10325 -10,4471 -3,1 -3,1 -3,17563 5, -1,00736 0,836051 -0	6401
-0,8945 1,11091 7,04704 2,68999 3,51 1,72396 0,97497 5,85485 5, ^{-2,78432} -0,30631 -5	8779
-7,	3681
1,51352 -4,36351 -1 -1	4618
4,526 0,020048 0,6301/2 1-3,15074 1,14707 1,45255 2,55265	37187 11384
-0,89494 8,11653 -{ 1,71552 -4,36571 -1,70506 1,71552 -4,36571 -1,70506 4,07452	37123
2,3662 -0,10323 -10,4471 -3,0,02046 0,030172 1,06750 0,02046 0,030172 1,06750 2,20364	15558
-1,007: 1,72396 0,97497 5 -0,63434 6,11033 -0,53331 -0,63331 0,01701	33966
22,764. 4,52856 11,1303 -(-0,10323 -10,4471 -3,17503 -0,10323 -10,4471 -3,17503 -3,31002	54164
0,3421 -0,83436 1,05809 7 1,72520 0,57437 3,63463 1,72520 0,57437 3,63463 3,73813	.7983
2,535. 2,58882 8,18563 4 4,32630 11,1305 -0,32204 4,32630 11,1505 -0,32204 5,36445	28749
-3,1335 -1,00736 0,836051 -(0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03305 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03430 1,03605 7,03505 0,03605 1,03605 1,03605 7,03505 0,03605 1,	12663
-2,70432 -0,30031 -2	16401
	58779
	73681
-3,2335	54618
-0.024 5 14504 1 5093 5 -3.13531 -3.0484 3,19325 -3.13531 -3.0484 3,19325 3,82446	19727
-0,024e -6,14694 11,5993	52903
-0,2634 -2,42808 -2,359686,14694 11,5993 5,03532 -6,14694 11,5993 5,03532 3,68279	51749
0,93500 -3,23398 0,739948 1 1,66801 -1,53136 2,95147 1,66801 -1,53136 2,95147 0,441002	02664
-7	17958
-0,426 ³ -1,7316 0,14579 -8 -3,23398 0,739948 1,61084 -3,23398 0,739948 1,61084 3,85228	34149
1,0704 -0,26349 2,65934 -(-0,02463 3,72305 -3,05367 -0,02463 5,72305 -3,05367 -0,735423 6,3526 0,735423 6,3526	55706
0,4220 0,935001 7,52704 -5 1,7310 0,14575 0,70205 1,7310 0,14575 7,77002	20848
-1,13148 -3,6351 -0,20349 2,00334 -0,30402 -0,20349 2,00334 -0,20349 2,00334 -0,20349 2,00334 -0,20349 2,00334 -0,20349 2,00334 -0,20349 2,00334 -0,20349 2,00349	16366
-0,42632 10,1015 -2,0753001 7,02704 -3,6726 0,753001 7,02704 -3,6726 0,045108 -6,57357 -17,6352 3,57077	06282
1,67647 4,57368 2 1,13146 -3,0331 3,4333 -1,13146 -3,0331 3,4331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 3,4331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0331 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13146 -3,0318 -1,13148 -1,13	41031 41212
0,42202 5,5/122 3 0,4202 10,1013 4,23203 0,4202 10,1013 4,23203 2,3344 0,33443 0,0432 0,0443 0,0444 0	.8634
1,07047 4,37300 2,22432 1,07047 4,37300 2,22432 2,00033 -7,03132 -0,23503 0,32712 1,07047 4,37300 2,22432 3,00047 4,000	75276
0,42202	59255
0,74202	,5255

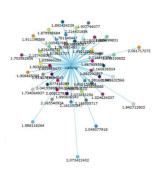
Duomenų tyryba žinių radimo procese



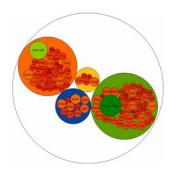
Duomenų vizuali analizė

- Vizualizuoti (lot. visualis regimas) nematomą atvaizdą, daiktą, reiškinį, daryti matomą (Tarptautinių žodžių žodynas, 1985).
- Vizualizavimas tai informacijos kodavimas į regimuosius vaizdus.
- Vizualizavimas tai grafinis informacijos pateikimas, palengvinantis informacijos suvokimą.





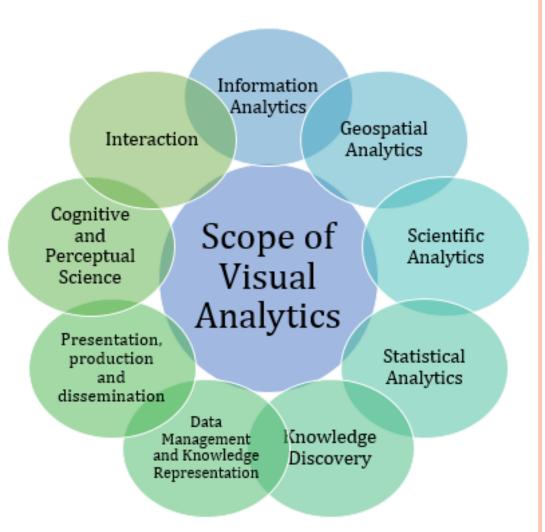




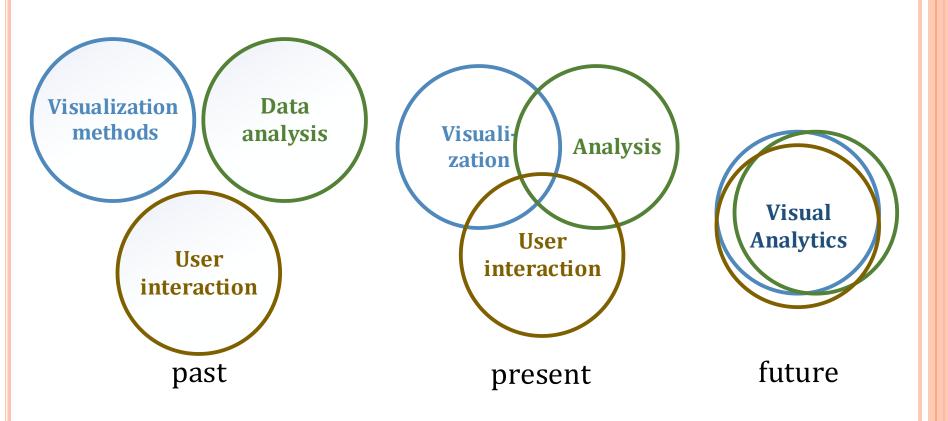
Visual Analytics

Visual Analytics

combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets.

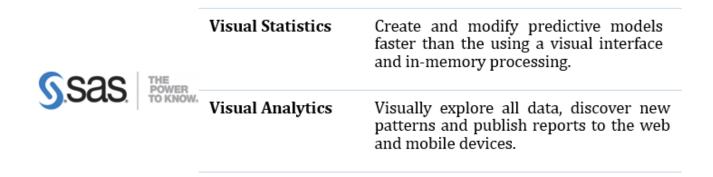


Vizualizavimas – anksčiau, dabar, ateityje



Visual Analytics as a Web Services

 SAS is an integrated system of software solutions for data management, graphics design, statistical and mathematical analysis, business forecasting and decision support.

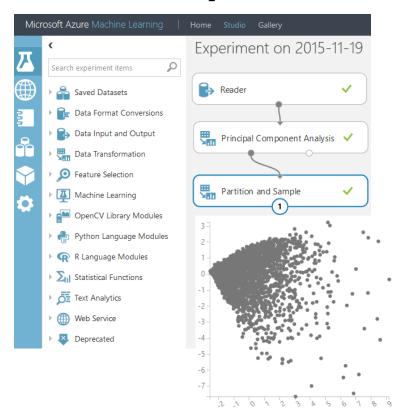


 Oracle Business Analytics is analytics for insight and innovation. Oracle Data Visualization Cloud Service get instant clarity with stunningly visual analysis and selfservice discovery.



Visual Analytics as a Web Services

Microsoft Azure is an open, flexible, enterprise-grade cloud computing platform. Microsoft Azure Machine Learning is a service to build predictive analytics models and to easily deploy those models for consumption as cloud web services.



Platforms for Visual Analytics

Pentaho, a Hitachi Group Company, is a leading data integration and business analytics company with an enterprise-class, open source-based platform for diverse big data deployments.

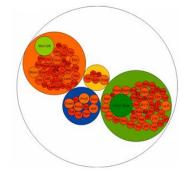


Blended big data analytics Discovery and visualizations

Highvolume data processing

Adaptive big data layer Predictive analytics & data mining



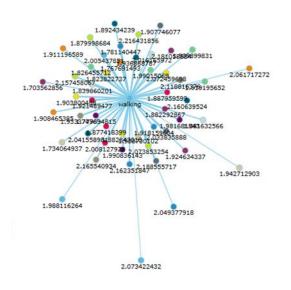




Platforms for Visual Analytics

Amazon Web Services offers a broad set of global compute, storage, database, analytics, application, and deployment services.





Datameer is big data analytics platform for Hadoop that empowers business users to directly integrate, analyze, and visualize any data.



Cloud based Data Exploration

- O IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data.
- IBM Watson Analytics offers the benefits of advanced analytics without the complexity.



Skaitmeninis intelektas ir didieji duomenys

- Skaitmeninio intelekto metodai suteikia galingus įrankius sprendžiant didžiųjų duomenų iššūkius.
- Evoliuciniai skaičiavimai, neuroniniai tinklai, neraiškiosios sistemos geba susidoroti su dideliais kiekiais neapibrėžtumų, susijusių su didžiųjų duomenų įvairumu ir nepastovumu.
- Tačiau didžiųjų duomenų kiekiai kelia didelius iššūkius ir skaitmeninio intelekto metodų efektyviam taikymui.

