



UPRAVLJANJE VELIKIM KOLIČINAMA PODATAKA

Predavanje 1.

BUZZ WORD *BIG DATA*

☆
CITE
A+あ
f
t
g+

buzzword

[buhz-wurd]
Spell Syllables

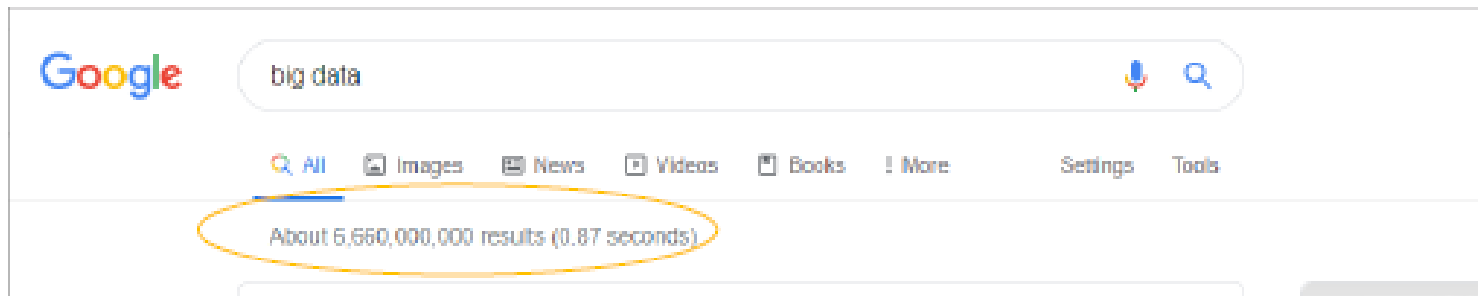
Examples Word Origin
[See more synonyms on Thesaurus.com](#)



noun

1. a word or phrase, often sounding authoritative or technical, that is a vogue term in a particular profession, field of study, popular culture, etc.

Origin of buzzword 

1965-1970

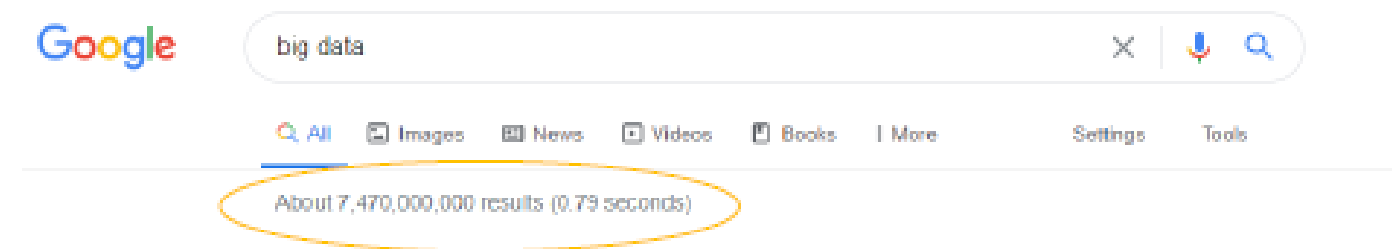





Google big data  

[All](#) [Images](#) [News](#) [Videos](#) [Books](#) [More](#) [Settings](#) [Tools](#)

About 6,660,000,000 results (0.87 seconds)

2019-10-28



Google big data   

[All](#) [Images](#) [News](#) [Videos](#) [Books](#) [More](#) [Settings](#) [Tools](#)

About 7,470,000,000 results (0.79 seconds)

2020-10-14



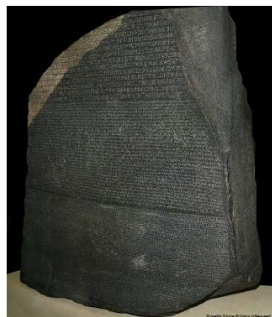
VELIKI PODACI

VELIKI **PODACI**

BAZE PODATAKA ISTORIJA



Komunikacija / Zapisi



Pismo



Račun



Štampa



BAZE PODATAKA ISTORIJA



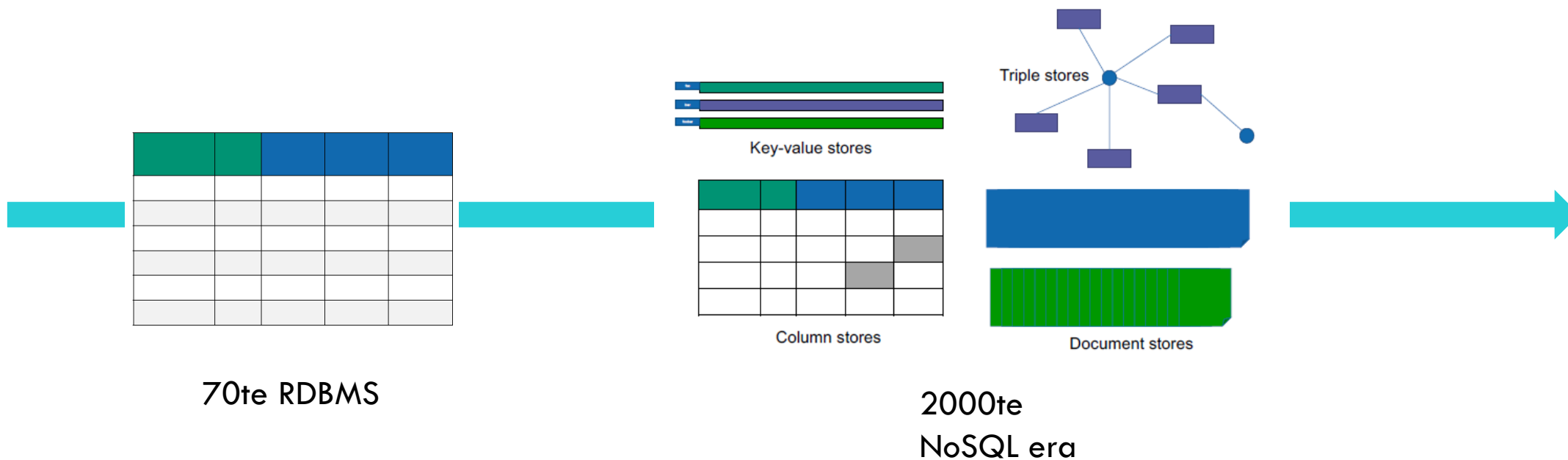
Računari

- ▼ Lorem Ipsum
- ▼ Dolor sit amet
- ▼ Consectetur
 - Adipiscing
 - Elit. In
- ▼ Imperdiet
 - Ipsum ante

Fajl sistem (60te)

70te
Era baza podataka

BAZE PODATAKA ISTORIJA



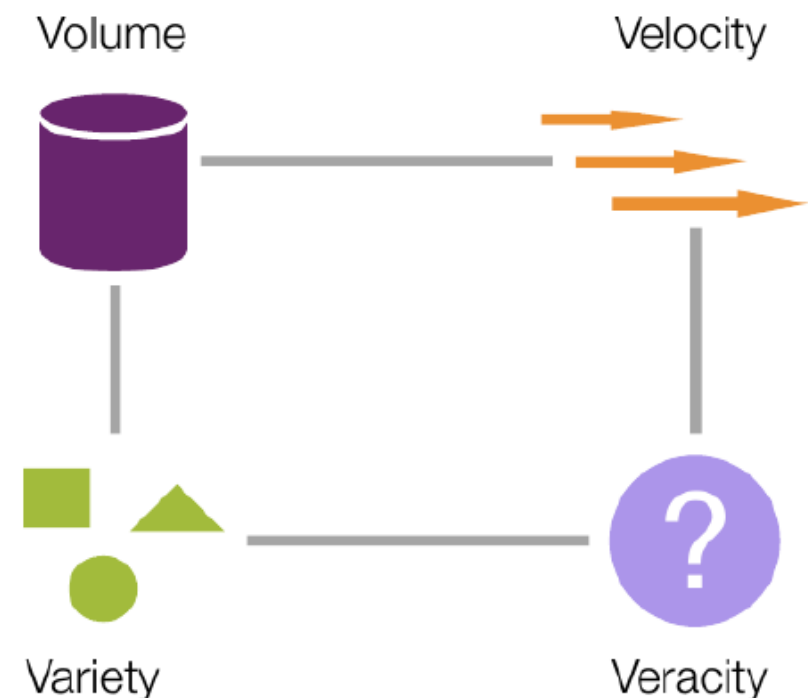
VELIKI PODACI

VELIKI PODACI

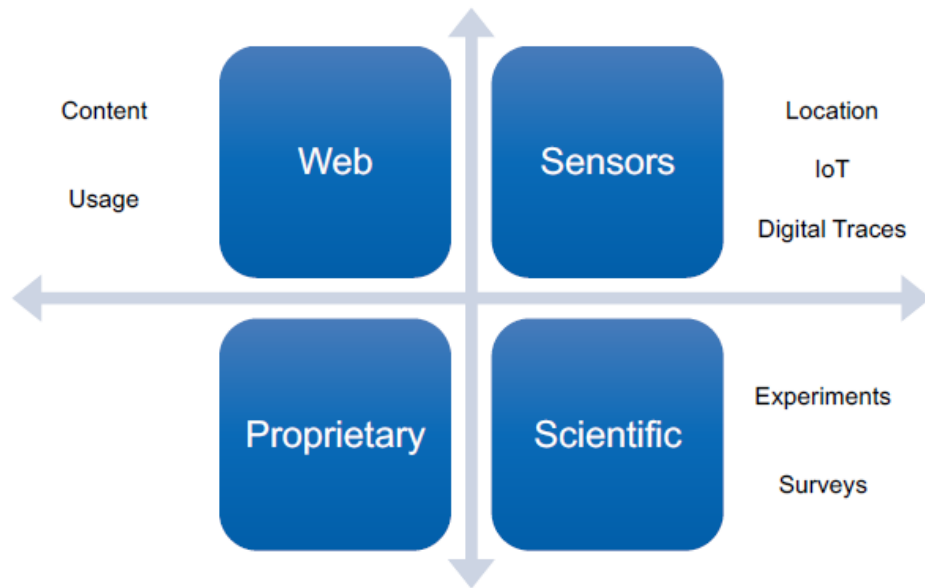
ŠTA JE VELIKO?

3+1 Vs – četiri dimenzije velikih podataka

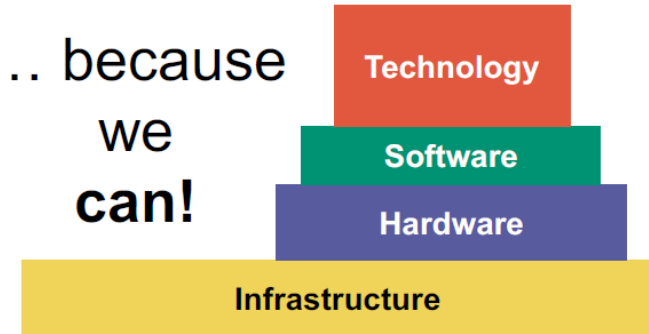
1. Volume - količina
2. Velocity – brzina
3. Variety – raznovrsnost
4. Veracity – vrednost
 - Veracity – verodostojnost
 - Variability - promenljivost
 - Value - vrednost



KOLIČINA



... because
we
can!





... because
data
carries
value



Utility( + )

>

Utility() + Utility()

JEDINICE

kilo (k)	1,000 (3 zeros)
Mega (M)	1,000,000 (6 zeros)
Giga (G)	1,000,000,000 (9 zeros)
Tera (T)	1,000,000,000,000 (12 zeros)
Peta (P)	1,000,000,000,000,000 (15 zeros)
Exa (E)	1,000,000,000,000,000,000 (18 zeros)
Zetta (Z)	1,000,000,000,000,000,000,000 (21 zeros)
Yotta (Y)	1,000,000,000,000,000,000,000,000 (24 zeros)

kibi (ki)	1,024 (2^{10})
Mebi (Mi)	1,048,576 (2^{20})
Gibi (Gi)	1,073,741,824 (2^{30})
Tebi (Ti)	1,099,511,627,776 (2^{40})
Pebi (Pi)	1,125,899,906,842,624 (2^{50})
Exbi (Ei)	1,152,921,504,606,846,976 (2^{60})
Zebi (Zi)	1,180,591,620,717,411,303,424 (2^{70})
Yobi (Yi)	1,208,925,819,614,629,174,706,176 (2^{80})

KOLIČINA

<https://techjury.net/stats-about/big-data-statistics/>

Do 2020 - 40 ZB

<https://followthedata.wordpress.com/2014/06/24/data-size-estimates>

- Google: 15 EB
- Facebook: 300 PB
- Spotify: 100 PB

Bytes

Megabyte **1,000,000**

Gigabyte **1,000,000,000**

Terabyte **1,000,000,000,000**

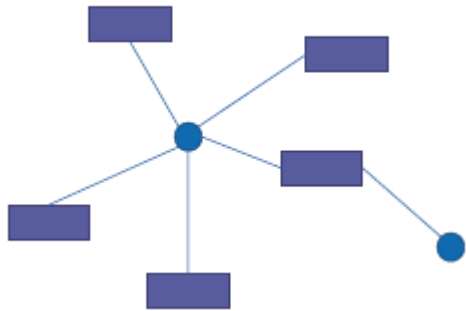
Petabyte **1,000,000,000,000,000**

Exabyte **1,000,000,000,000,000,000**

Zettabyte **1,000,000,000,000,000,000,000**

Yottabyte **1,000,000,000,000,000,000,000,000**

RAZNOVRSNOST



Lorem ipsum dolor sit amet, consectetur aliquet vulputate sed quis nulla. Donec e Nullam sed uma nec nisl rhoncus ullam ornare libero quis consequat. Lorem adipiscing elit. Aenean eu efficitur orci commodo turpis.

<https://followthedata.wordpress.com/2014/06/24/data-size-estimates>

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

BRZINA

- Brzina pristizanja
Brzina procesiranja
- Podaci:
 - Se generišu automatski
 - Predstavljaju nusproizvod aktivnosti ljudi
- Google: procesira 100 PB/dnevno
- Facebook: procesira 600 TB/ dnevno
- Spotify: procesira 2.2 TB/ dnevno

<https://followthedata.wordpress.com/2014/06/24/data-size-estimates>

http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

TRI KLJUČNA FAKTORA

Kapacitet (Capacity)



"How much data can we store?"

Propusnost (Throughput)



"How fast can we transmit data?"

Čekanje (Latency)



"When do I start receiving data?"

KAPACITET

1956. IBM RAMAC 350



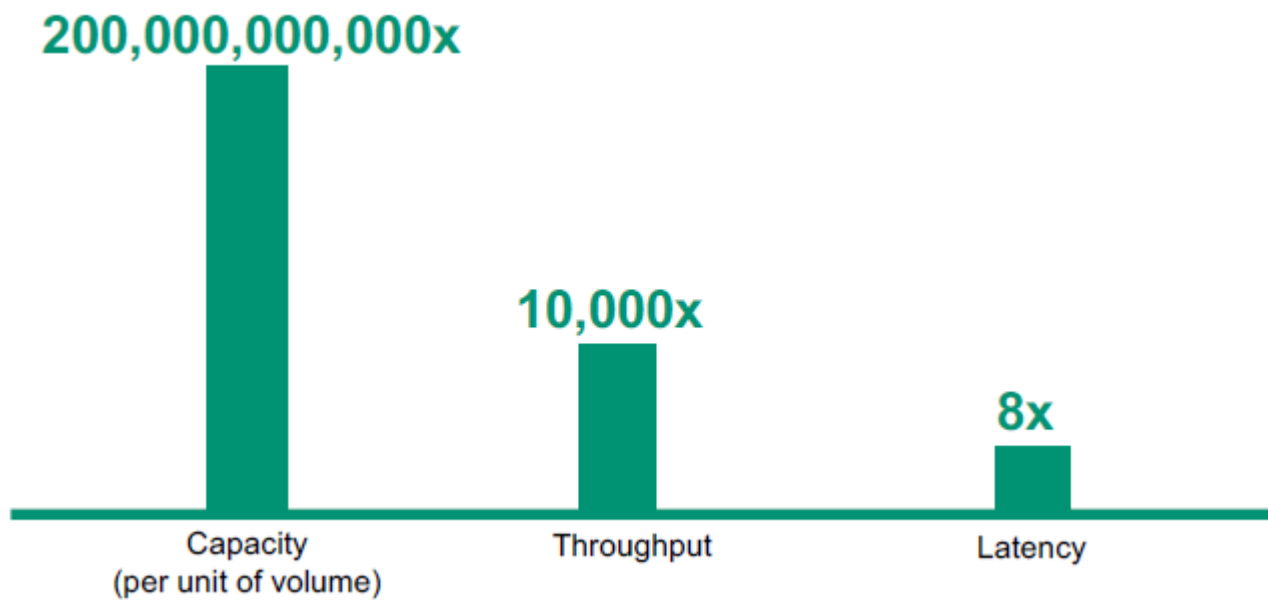
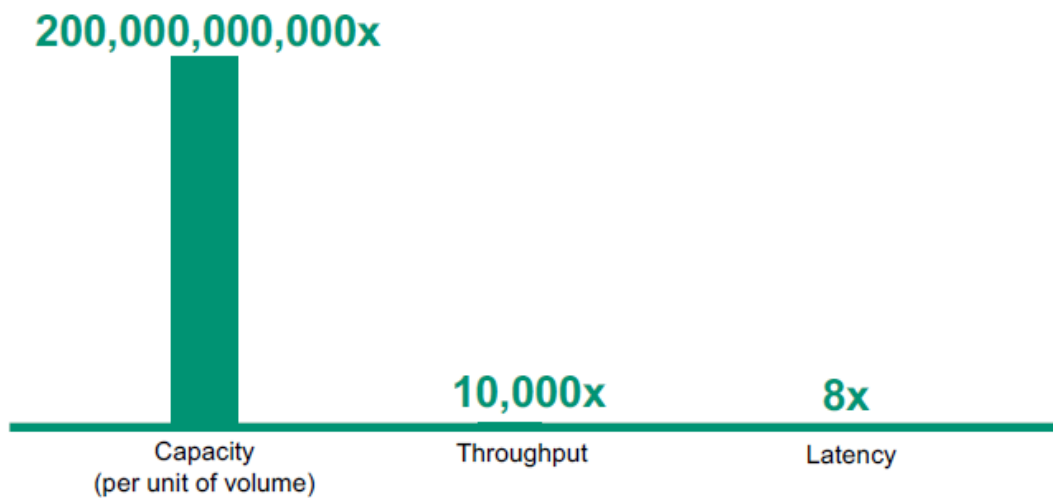
3.75 Megabytes

2020. WD HC650



20TB

1956-2020

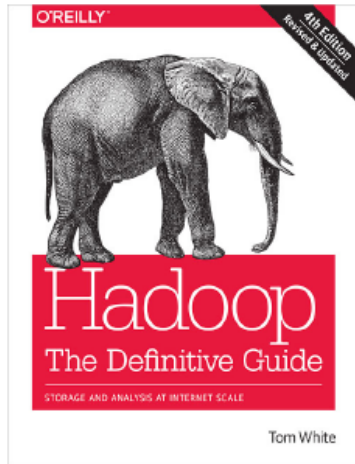


PARALELA



~1,000 word
per minute

Throughput:



~ 600,000 words.

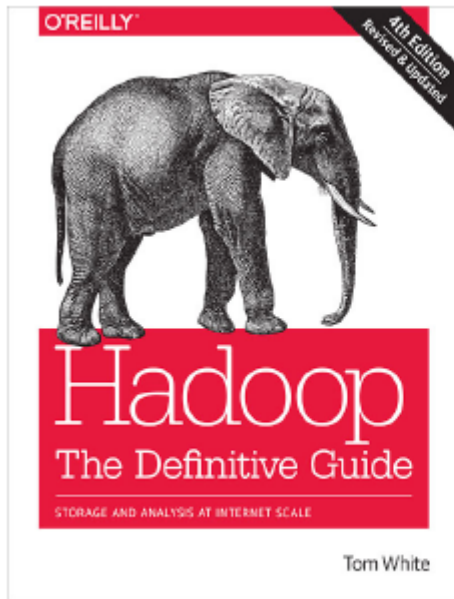
Capacity:



~ 1 minute to
stand up,
go to the shelf,
pick the book,
find the page.

Latency:

PARALELA

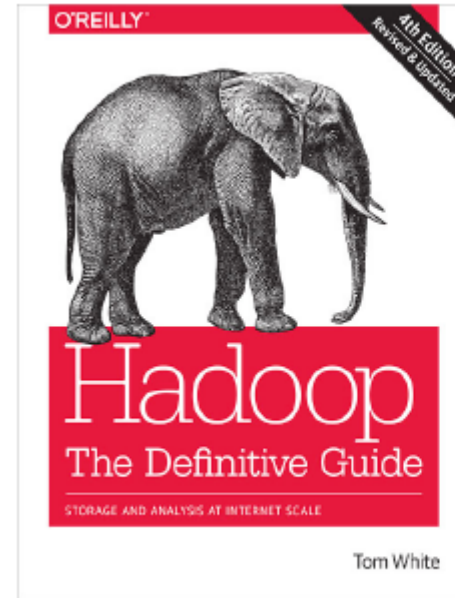


600,000 words

1,000 words per minute



10 hours



120,000,000,000,000,000 words

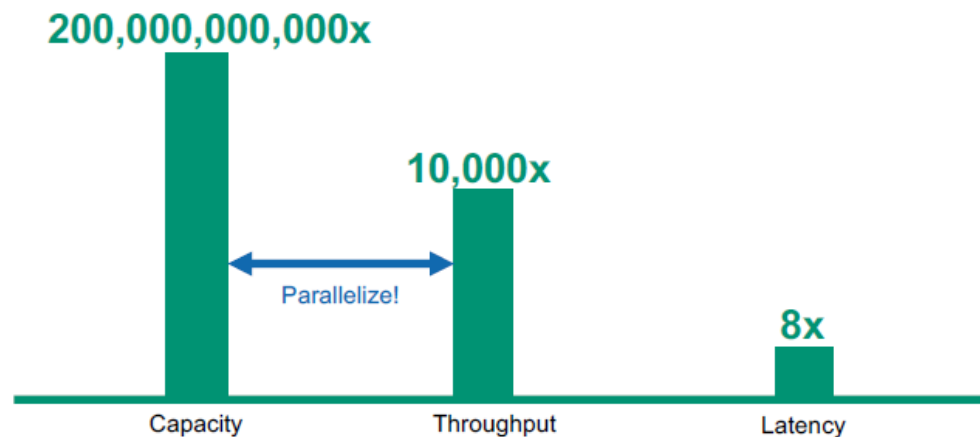
10,000,000 words per minute



22,800 years

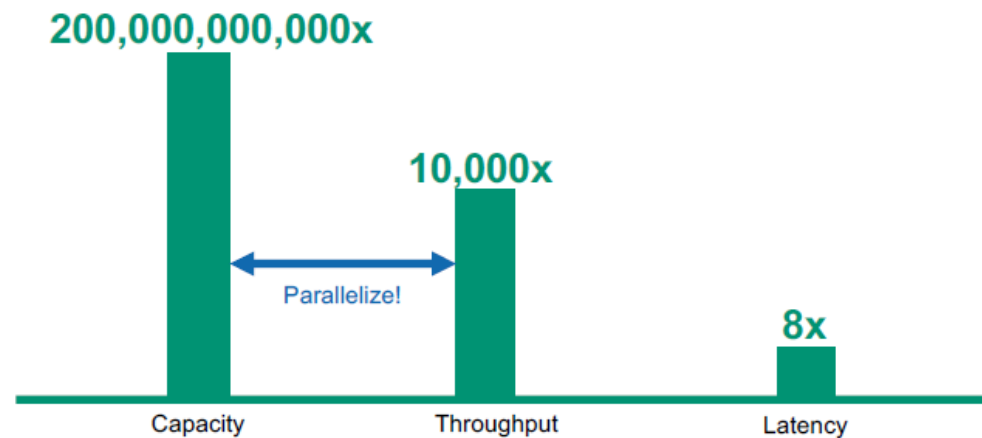
REŠENJA?

2220: 200,000,000 persons could read it all in 10 hours.



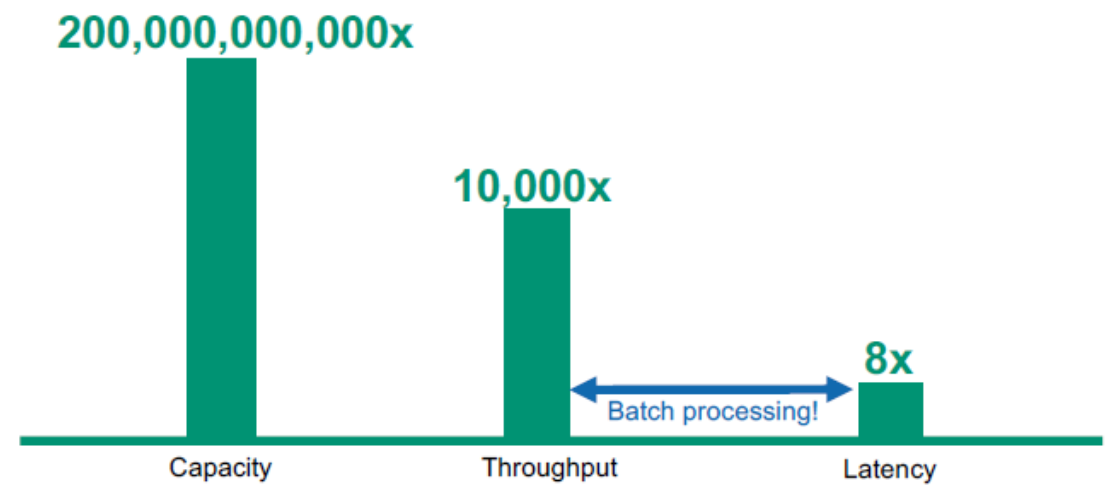
REŠENJA?

clusters of machines (10,000s)



REŠENJA?

clusters of machines (10,000s)



PRILAGOĐAVANJE VELIČINI - SKALIRANJE

- Vertikalno – dodavanje resursa jednom nodu u sistemu
 - Skuplje od horizontalnog
- Horizontalno – dodavanje više nodova
 - Teže za kontrolu grešaka/otkaza
 - Teže za razvoj



BIG DATA DEFINICIJA

Big Data is a portfolio of technologies that were designed to

store, manage and analyze data that is too **large** to fit on a single machine

while accommodating for the issue of **growing discrepancy between capacity, throughput and latency.**

ŠTA SA PODACIMA?

Ekstrakcija informacija iz podataka i donošenje odluka

Sensors
Measurements
Events
Logs



Raw Data

Aggregated data
Intermediate data



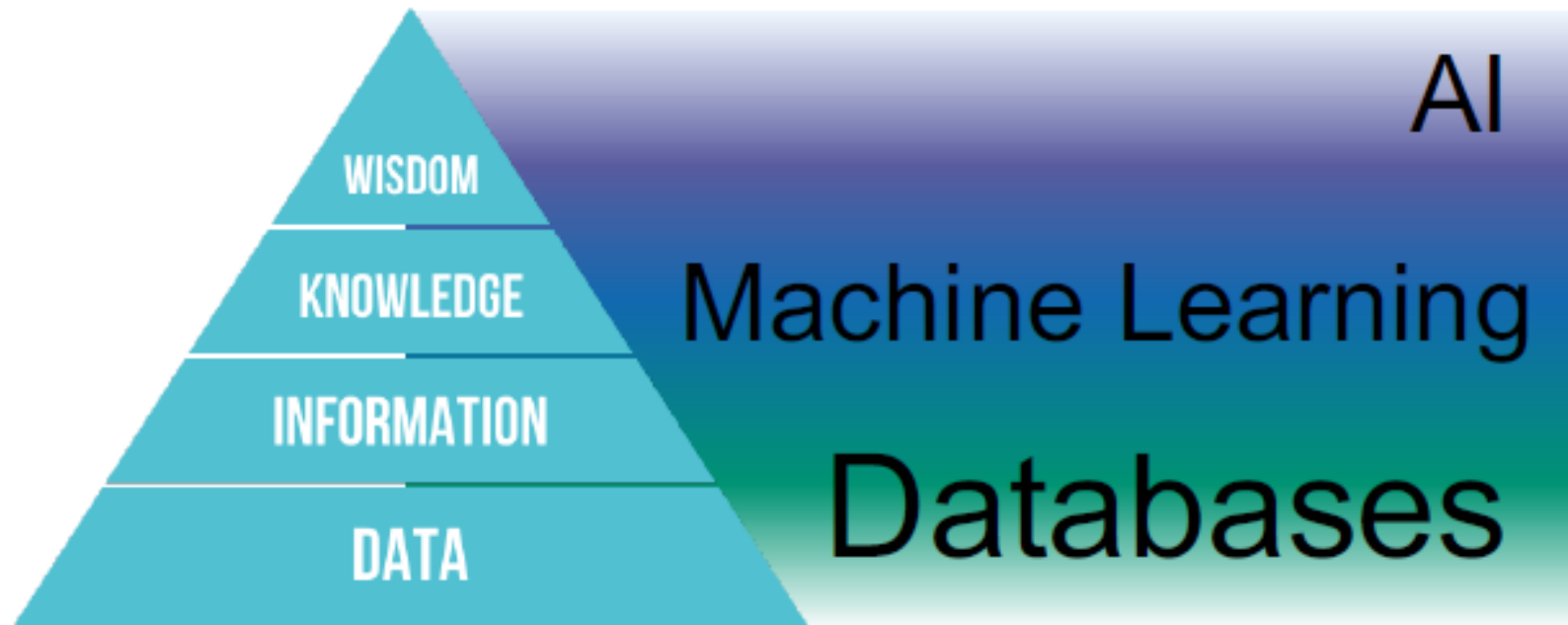
Artem Kabanov / iStock Photo

Derived Data

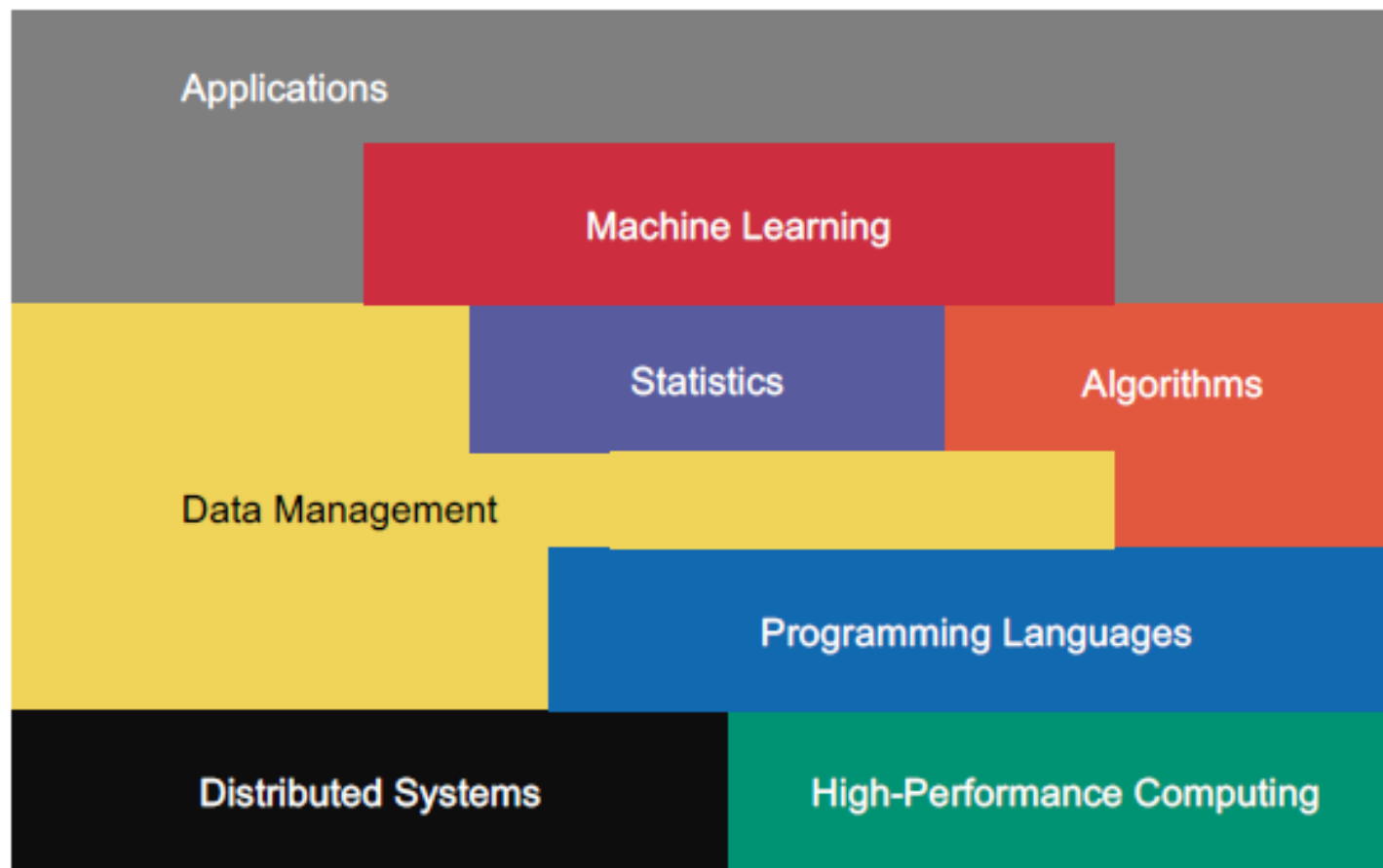


big data

ŠTA SA PODACIMA?



VELIKI PODACI KROZ DISCIPLINE



TEHNOLOGIJE

Tehnologije Velikih podataka - fokus na osmišljavanju efikasnih tehnika:

- reprezentacije i skladištenja;
- procesiranja i donošenja zaključaka!

	Koncepti	Tehnologije
Skladišta	Object storage Distributed file systems Syntax	S3, Azure Blob Storage HDFS XML, JSON
Modeli	Wide column stores Data models and schemas Graphs	Hbase XML/JSON Schema neo4j, Cypher
Procesiranje	2-step distributed query processing DAG-based distributed query processing	Hadoop MapReduce Spark
Upravljanje	Document storage Query languages	DB

BIG DATA SLUČAJEVI

There is

Big Data



Anna Lindstedt / iStock Photo

and

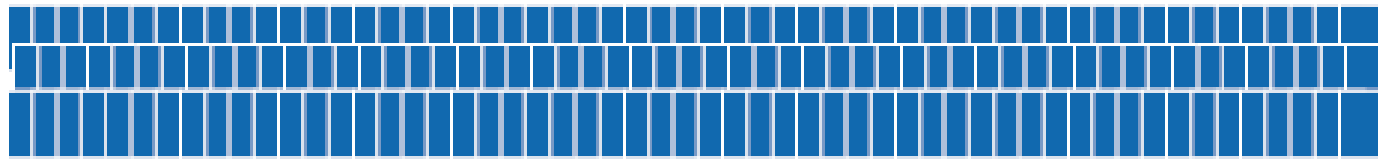
Big Data



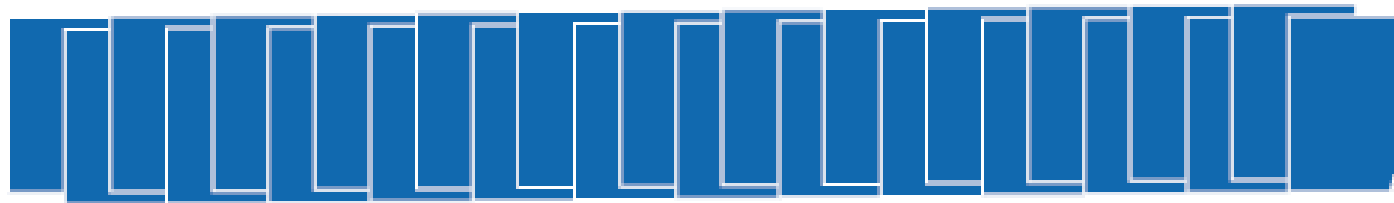
Yelena Kuznetsov / iStock Photo

BIG DATA SLUČAJEVI

A **huge** amount of **large** files?



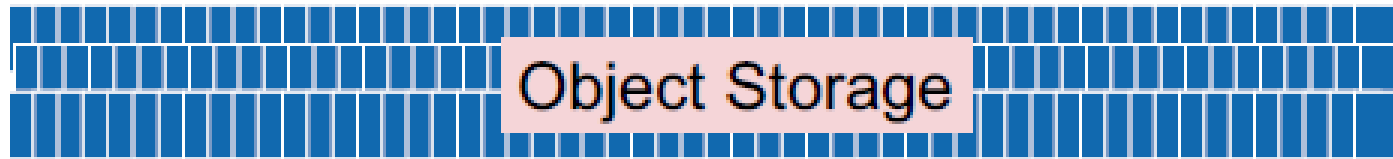
vs.



A **large** amount of **huge** files?

BIG DATA SLUČAJEVI

Billions of TB files



VS.



Millions of PB files

BIG DATA SLUČAJEVI

Key-Value Model

Object Storage

File System

Block Storage

Billions of
<TB files

VS.

Millions of
<**PB** files

DISTRIBUIRANI FS



Local disk

It **might** fail



Cluster with 100s to 10,000s of machines

nodes **will** fail

FAULT TOLERANCE AND ROBUSTNESS

Fault tolerance

Automatic Recovery

Error detection

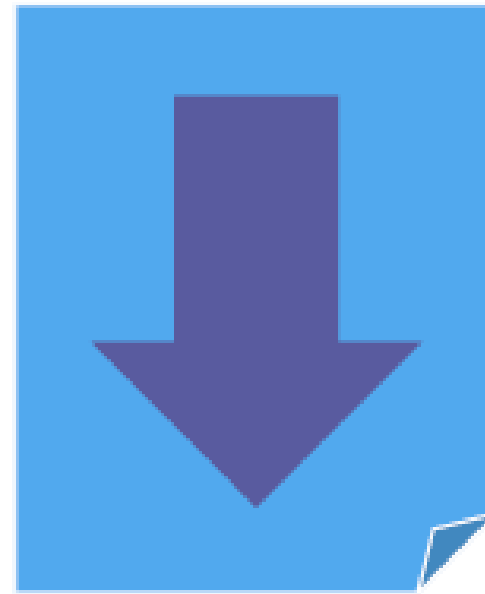
Monitoring

FILE READ MODEL



Random access

VS.



Scan the file

FILE UPDATE MODEL

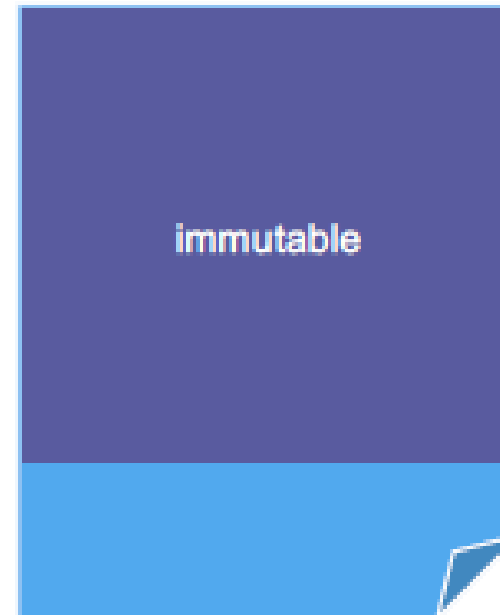
suitable for

Sensors



Logs

Intermediate data



Append

PERFORMANCE

Top priority:

Throughput



Secondary:

Latency

