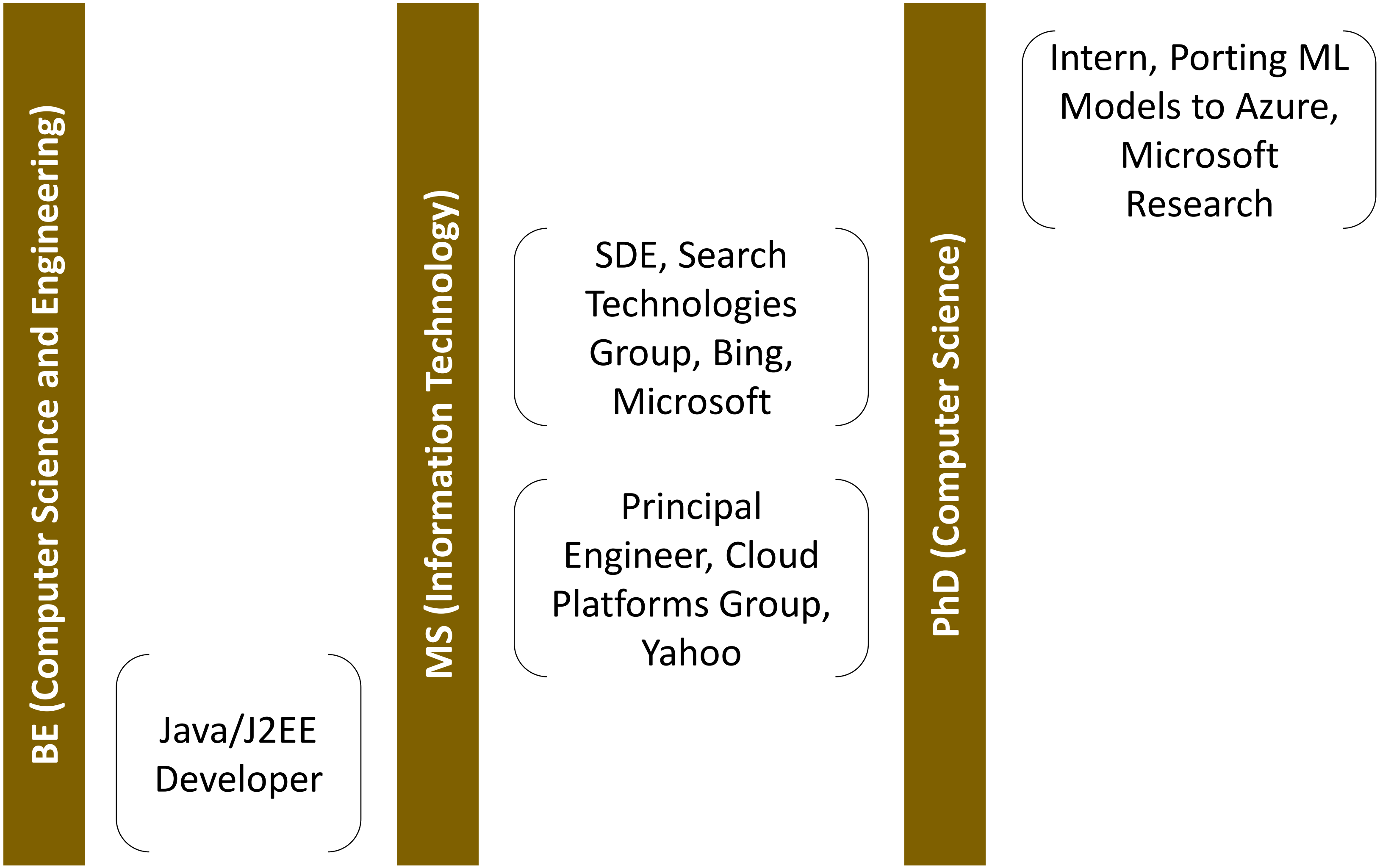
Big Data and Hadoop

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

Data is the new oil. - Clive Humby, 2006.

Know Your Instructor



BE (Computer Science and Engineering)

Java/J2EE Developer

MS (Information Technology)

SDE, Search Technologies Group, Bing, Microsoft

Principal Engineer, Cloud Platforms Group, Yahoo

PhD (Computer Science)

Intern, Porting ML Models to Azure, Microsoft Research

Agenda

- Introduction to Big Data
- Course Dynamics
- Evolution of Systems and Technologies
    - Data Storage
    - Data Processing

What Comes Next?

byte

kilobyte

megabyte

gigabyte

??

???

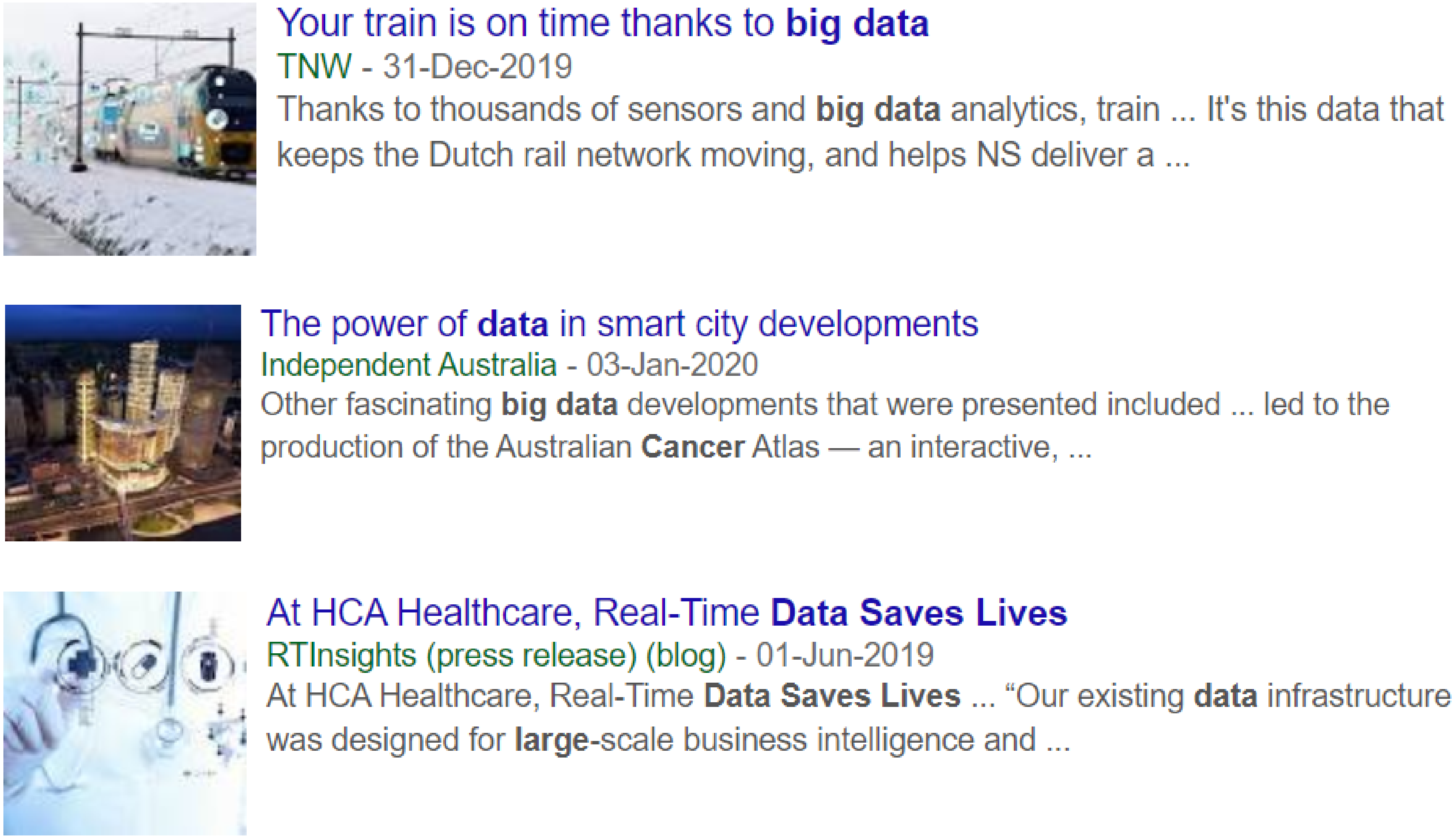
????

?????

Sizes

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

The Impact of Big Data



Your train is on time thanks to **big data**

TNW - 31-Dec-2019

Thanks to thousands of sensors and **big data** analytics, train ... It's this data that keeps the Dutch rail network moving, and helps NS deliver a ...

The power of **data** in smart city developments

Independent Australia - 03-Jan-2020

Other fascinating **big data** developments that were presented included ... led to the production of the Australian **Cancer** Atlas — an interactive, ...

At HCA Healthcare, Real-Time **Data Saves Lives**

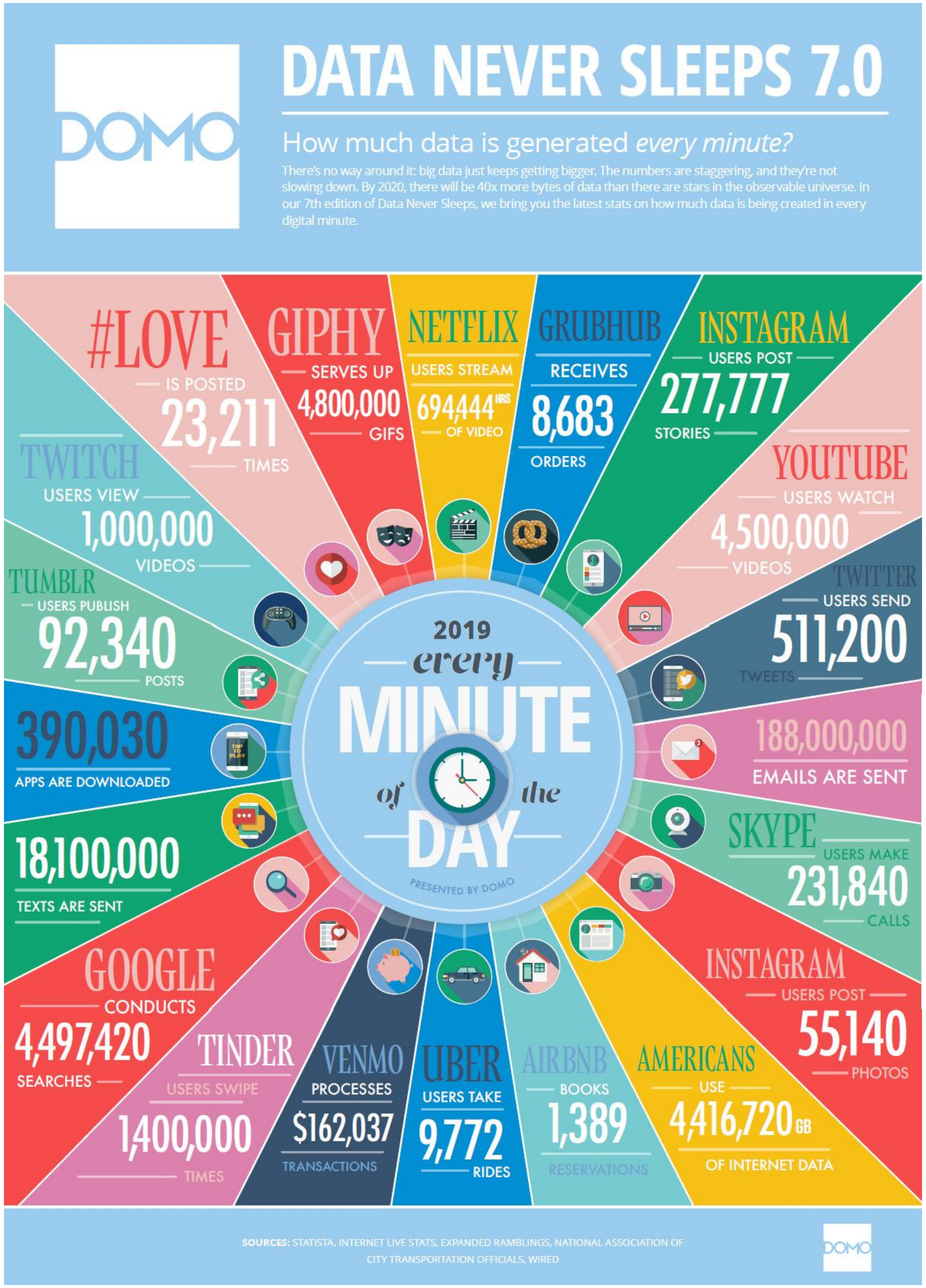
RTInsights (press release) (blog) - 01-Jun-2019

At HCA Healthcare, Real-Time **Data Saves Lives** ... "Our existing **data** infrastructure was designed for **large**-scale business intelligence and ...

Big Data is Ubiquitous

- Facebook (**per day** statistics)
  - 1.5 billion people are active on Facebook **daily!**
  - More than 300 million photos get uploaded **per day**!
  - Totally, more than 2.5 Trillion posts!
- Facebook (per minute statistics)
  - **Every minute** there are 510,000 comments posted and 293,000 statuses updated!
- Youtube (**per minute** statistics)
  - Users watch 4,146,600 YouTube videos!

Source: Forbes



DATA NEVER SLEEPS 7.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute.

#LOVE IS POSTED 23,211 TIMES

GIPHY SERVES UP 4,800,000 GIFS

NETFLIX USERS STREAM 694444 HRS OF VIDEO

GRUBHUB RECEIVES 8,683 ORDERS

INSTAGRAM USERS POST 277,777 STORIES

YOUTUBE USERS WATCH 4,500,000 VIDEOS

TWITTER USERS SEND 511,200 TWEETS

188,000,000 EMAILS ARE SENT

SKYPE USERS MAKE 231,840 CALLS

INSTAGRAM USERS POST 55,140 PHOTOS

AMERICANS USE 4,416,720 GB OF INTERNET DATA

AIRBNB BOOKS 1,389 RESERVATIONS

UBER USERS TAKE 9,772 RIDES

VENMO PROCESSES $162,037 TRANSACTIONS

TINDER USERS SWIPE 1,400,000 TIMES

GOOGLE CONDUCTS 4,497,420 SEARCHES

18,100,000 TEXTS ARE SENT

390,030 APPS ARE DOWNLOADED

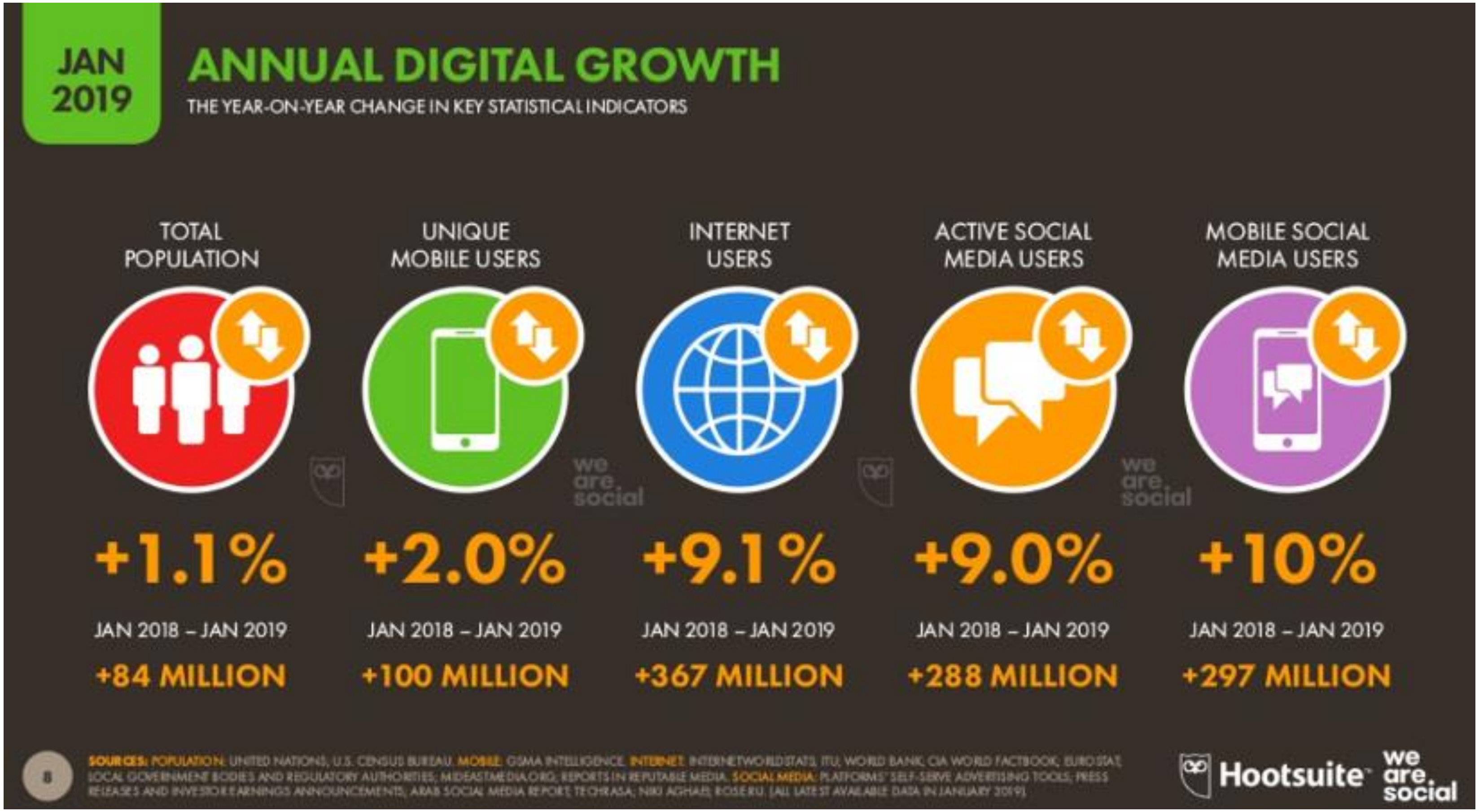
TUMBLR USERS PUBLISH 92,340 POSTS

TWITCH USERS VIEW 1,000,000 VIDEOS

SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED

Source: https://www.visualcapitalist.com/big-data-keeps-getting-bigger/

And, It is Growing!



JAN 2019

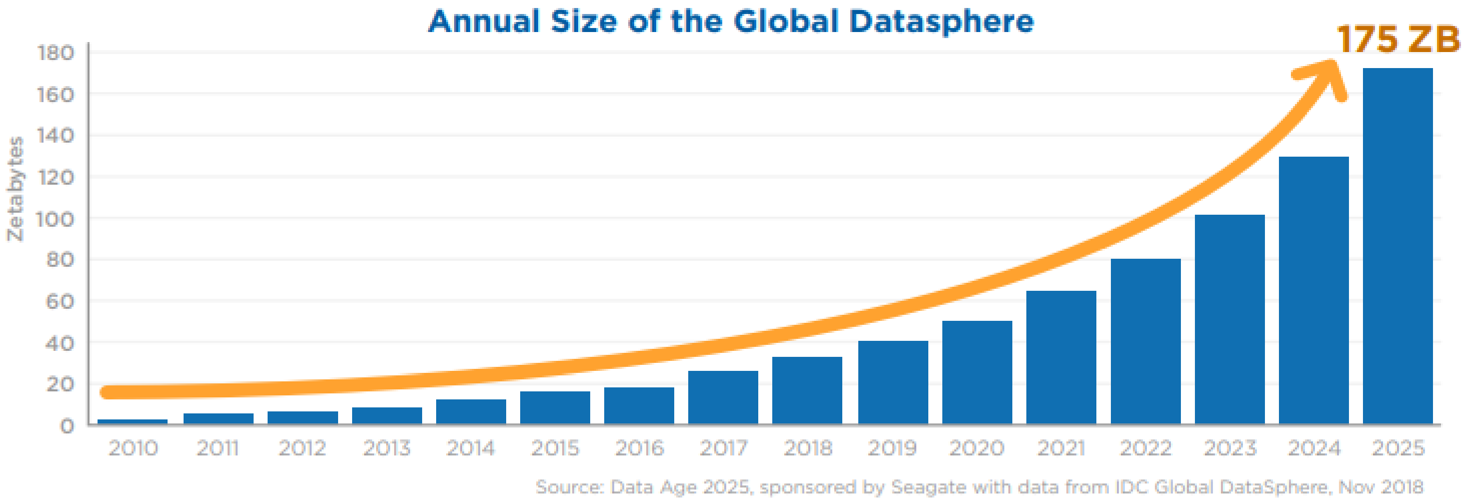
ANNUAL DIGITAL GROWTH

THE YEAR-ON-YEAR CHANGE IN KEY STATISTICAL INDICATORS

TOTAL POPULATION	UNIQUE MOBILE USERS	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	MOBILE SOCIAL MEDIA USERS
+1.1%	+2.0%	+9.1%	+9.0%	+10%
JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019
+84 MILLION	+100 MILLION	+367 MILLION	+288 MILLION	+297 MILLION

SOURCES: POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS; ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT; LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS RELEASES AND INVESTOR EARNINGS ANNOUNCEMENTS; ARAB SOCIAL MEDIA REPORT; TECHRASA; NIKI AGHAEI; ROSE.RU. (ALL LATEST AVAILABLE DATA IN JANUARY 2019).

Data Growth



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Mankind’s quest to digitize the world!
33 ZB (2018) → 175 ZB (2025)
size of global datasphere*

*Source: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Global datasphere is growing!

How have the computers evolved to capture, process and analyze these data?

Course Dynamics

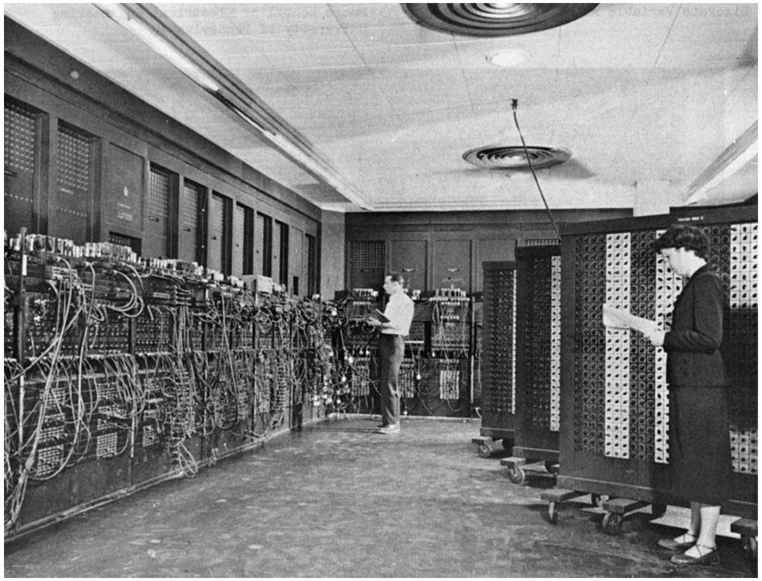
http://vvtesh.co.in/teaching/bdh-2020.html

Student Presentations

- Please register
  - your choice of presentation topic,
  - your team details (Team size: 3 or 4)

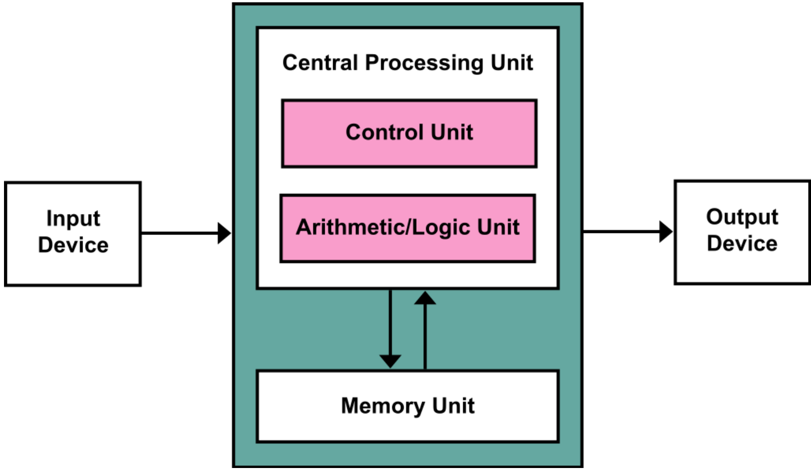
with the TA before 13th Jan 2020.

Evolution of Computers



ENIAC
Early 1900s

Stored-program
Von Neumann
Architecture
1940



Two Kinds of Problems

Storage

Processing

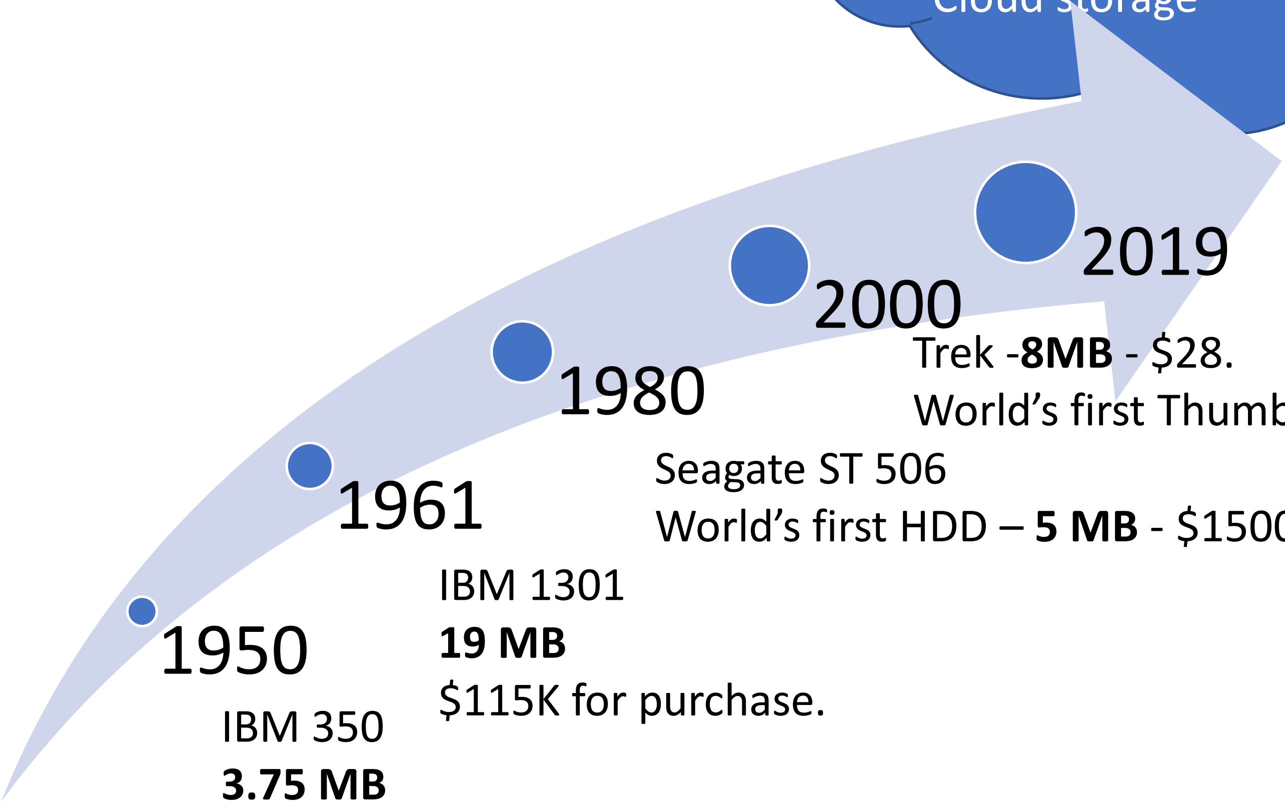
Data Storage

Evolution

Seagate **4 TB** external HDD - $89.99

Exadrive DC100 - SSD – **100 TB**

Cloud storage



2019

2000

Trek -**8MB** - $28.
World's first Thumbdrive

1980

Seagate ST 506
World's first HDD – **5 MB** - $1500.

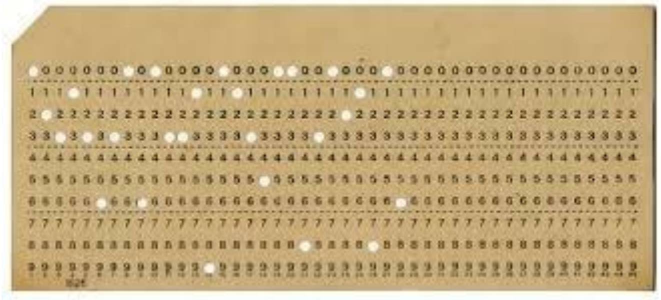
1961

IBM 1301
19 MB
$115K for purchase.

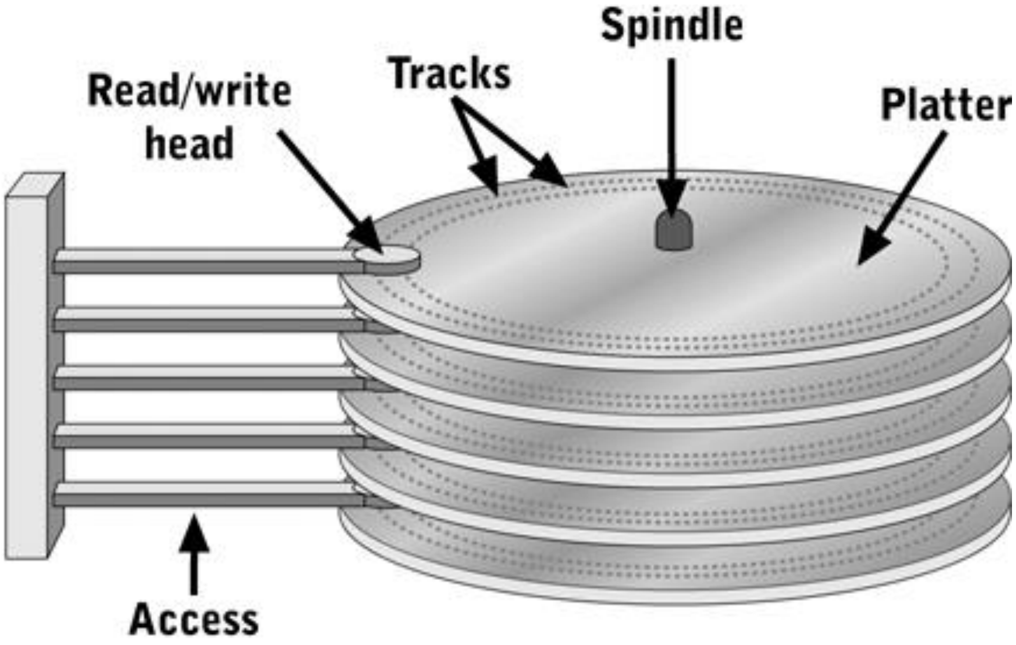
1950

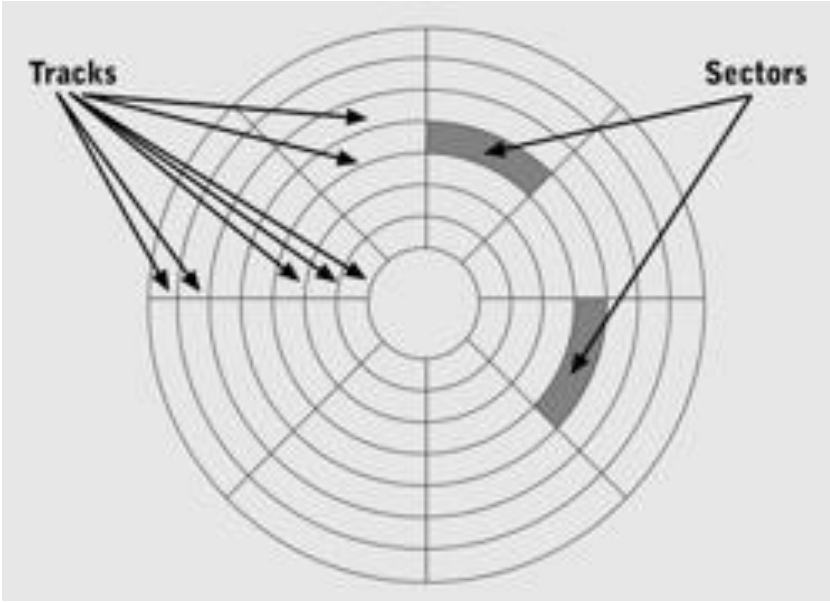
IBM 350
3.75 MB
Leased out for $3200 per month!

(Secondary) Storage Technologies

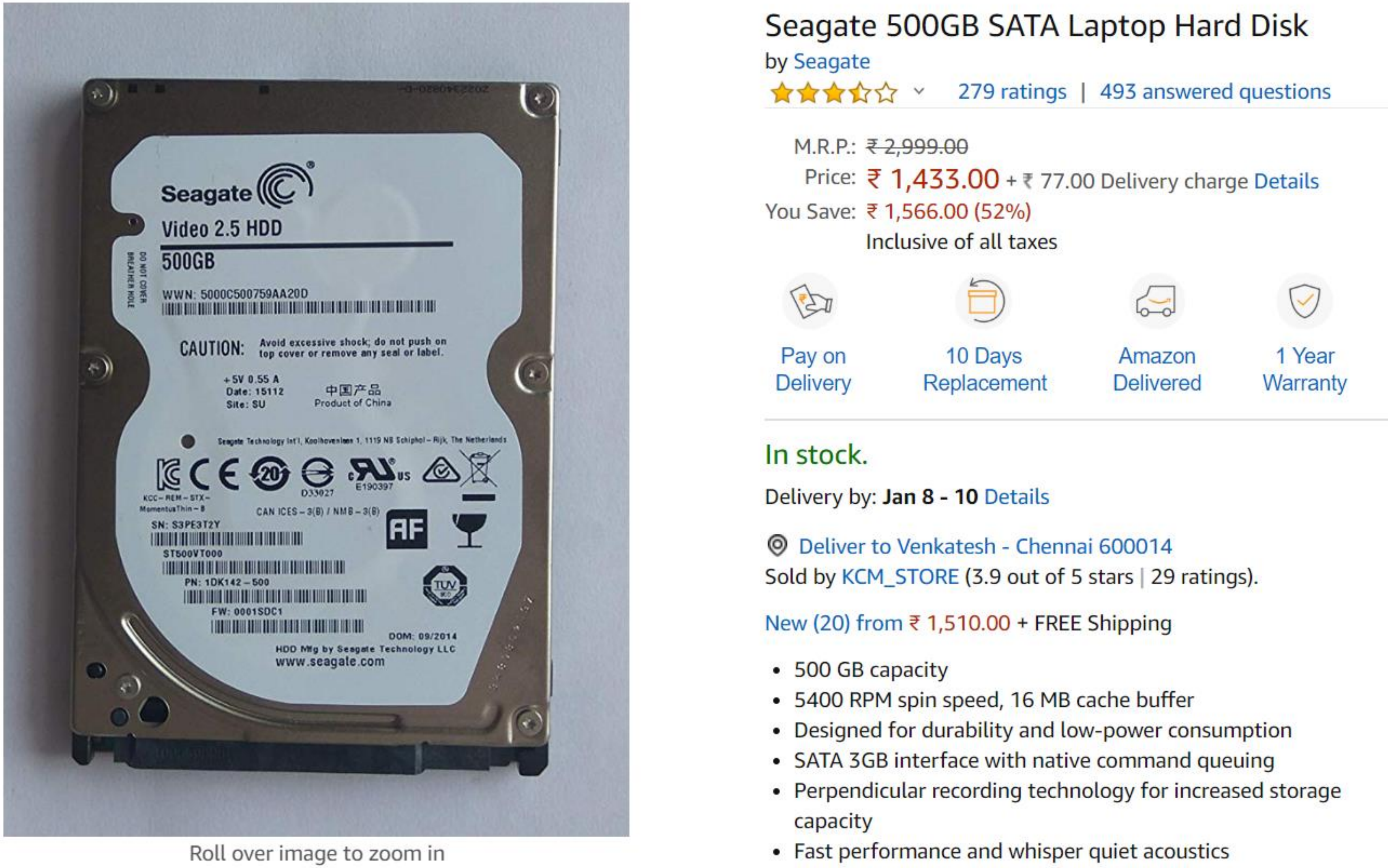


Disk Drive and Access Time





Source: Systems Architecture, Fifth Edition



Roll over image to zoom in

Seagate 500GB SATA Laptop Hard Disk

by Seagate

★★★⯪☆ ⌄ 279 ratings | 493 answered questions

M.R.P.: ~~₹ 2,999.00~~

Price: ₹ 1,433.00 + ₹ 77.00 Delivery charge Details

You Save: ₹ 1,566.00 (52%)

Inclusive of all taxes

Pay on Delivery | 10 Days Replacement | Amazon Delivered | 1 Year Warranty

In stock.

Delivery by: **Jan 8 - 10** Details

Deliver to Venkatesh - Chennai 600014

Sold by KCM_STORE (3.9 out of 5 stars | 29 ratings).

New (20) from ₹ 1,510.00 + FREE Shipping

- 500 GB capacity
- 5400 RPM spin speed, 16 MB cache buffer
- Designed for durability and low-power consumption
- SATA 3GB interface with native command queuing
- Perpendicular recording technology for increased storage capacity
- Fast performance and whisper quiet acoustics

Average Access Time

- Head switching time is considered negligible (H)
- Head seek time (S)
- Rotational delay = Time taken for ½ a rotation (average) (R)
- Read time = time to spin an entire sector (T)
- Average Access Time = H + S + R + T

*Sector is a minimum storage unit

Quiz

- If disk spins at 6000 RPM, compute the rotational delay.

Quiz

- If disk spins at 6000 RPM, compute the rotational delay.
  - One turn takes 1/6000 min or 1/100 sec = 10ms
  - ½ a turn takes 5ms.

Read Time

- If the drive spins at 6000RPM and the disk has 20 sectors per track, what is the read time?
  - Time for 1 full spin is $\frac{1}{6000}\text{min} = \frac{1}{100}\text{sec} = 10\text{ms}$
  - Time for 1/20 of a spin is $10\text{ms} \times \frac{1}{20} = 0.5\text{ms}$

Average Access Time

- Drive spins at 7200RPM and has average seek time of 8ms. The disk has 24 sectors per track. What is the average access time?

Head seek time	0.008 sec (Given)
Rotational delay	1/120 * (1/2) = 0.0042 sec
Read time	0.0084 (full spin) / 24 sectors = 0.00035 sec
Avg Seek Time	**= 0.008 + 0.0042 + 0.00035** **= 0.01255 sec or 12.55 ms**

Characteristics

Attribute	Description
Speed	Time to read/write
Volatility	Data persistence even when powered off
Access Method	Serial, Parallel
Portability	Internal, External
Capacity	Volume of data storage

Storing/Managing/Processing Data

- RDBMS
- ETL
- OLTP
- Data Warehouse
- Data Lake
- Cloud Storage
- STaaS

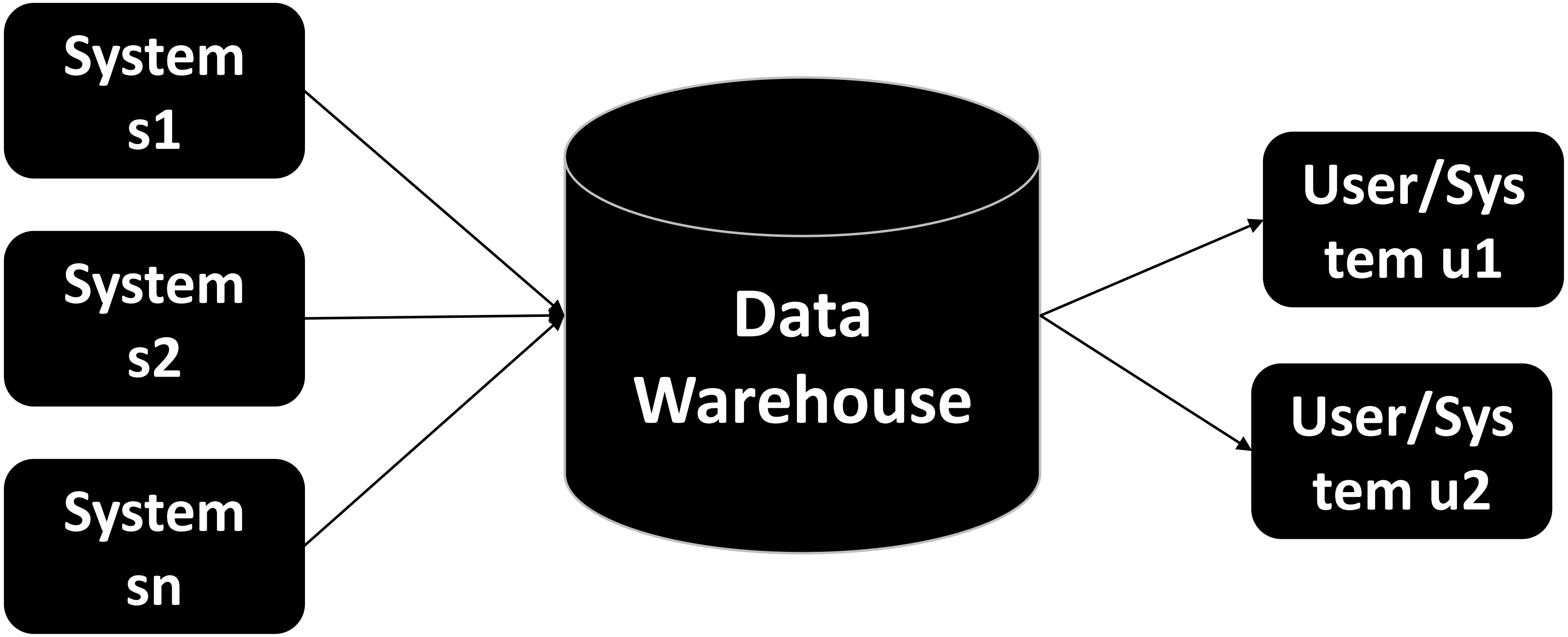
ETL

- Extract, Transform and Load (ETL)
  - general procedure followed to address data variety
- Variety of data sources
  - Tabs, Sensors, Desktops, Bots, Multiple databases, Files,…
- Variety of data formats
  - Text, PDFs, XML, JSON, Images, Videos, …

Fast OLTP

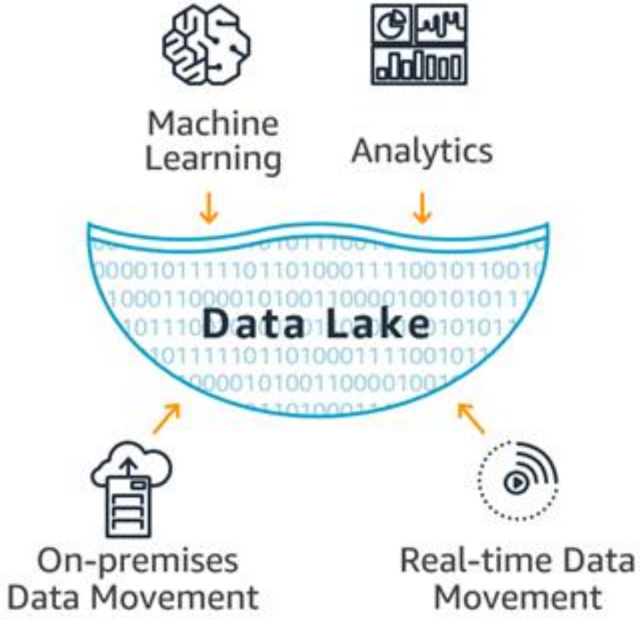
- Online Transaction Processing (OLTP)
- Real-time/Near Real-time Performance. Finds application in:
  - Banking
  - Railway Reservations
  - Stock Market Trading
    - Handle transactions in milliseconds.
      - VoltDB, MemSQL, …

Data Warehouse



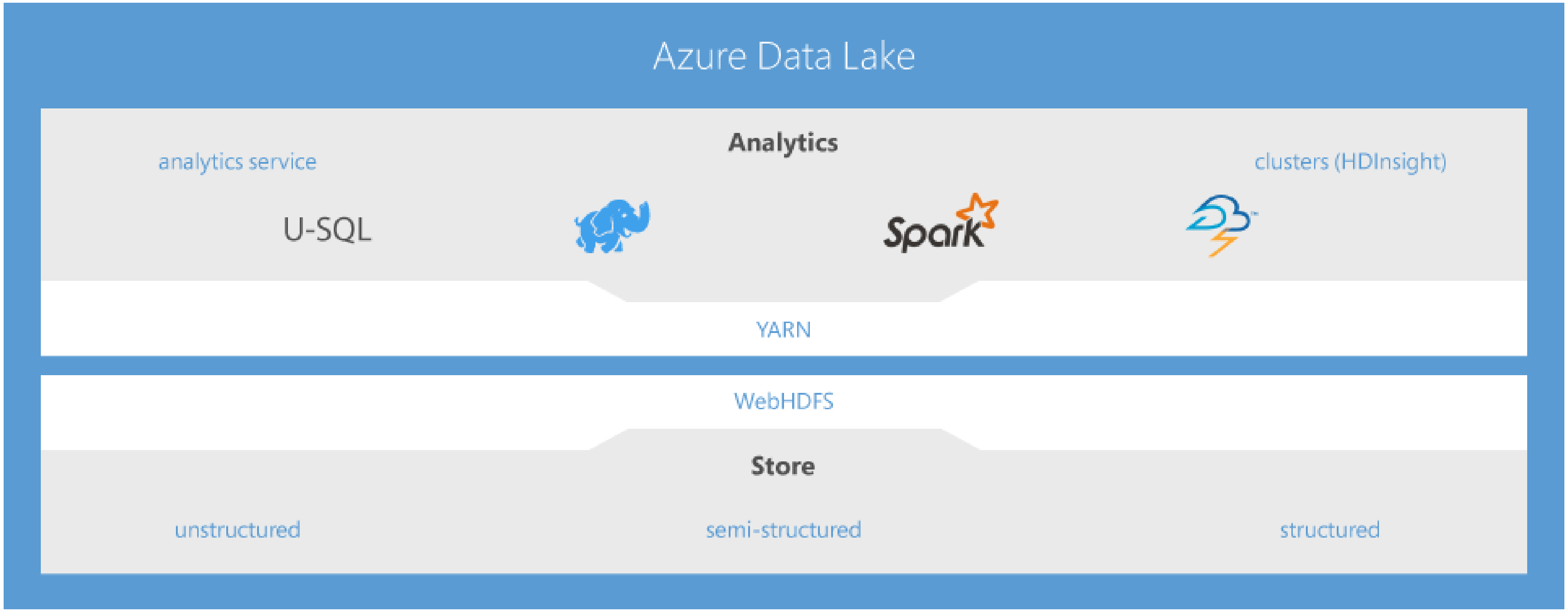
Data Lakes

- No schema definition.
- Store everything
  - often without or with very little pre-processing, /cleaning.
- Use ML, analytics to query, or gather insights.



Source: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Microsoft’s Azure Data Lake



More details at https://azure.microsoft.com/en-us/resources/videos/azure-data-lake-making-big-data-easy/

Data Warehouse Vs. Data Lake

Characteristics	Data Warehouse	Data Lake
Data	**Relational** from transactional systems, operational databases, and line of business applications	**Non-relational and relational** from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed **prior** to the DW implementation (schema-on-write)	Written **at the time of analysis** (schema-on-read)
Price/Performance	Fastest query results using **higher cost storage**	Query results getting faster using **low-cost storage**
Data Quality	Highly **curated data** that serves as the central version of the truth	Any data that may or may not be curated (ie. **raw data**)
Users	**Business analysts**	**Data scientists**, Data developers, and Business analysts (using curated data)
Analytics	**Batch** reporting, BI and visualizations	Machine **Learning**, Predictive analytics, data discovery and profiling

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Storing on the Cloud

- Gmail: Gives 15 GB of free storage (as of 2020)
- Several online sites for storing images, apps, files, …
  - Security
  - Ease of sharing
  - Backups
  - Availability

Google Drive

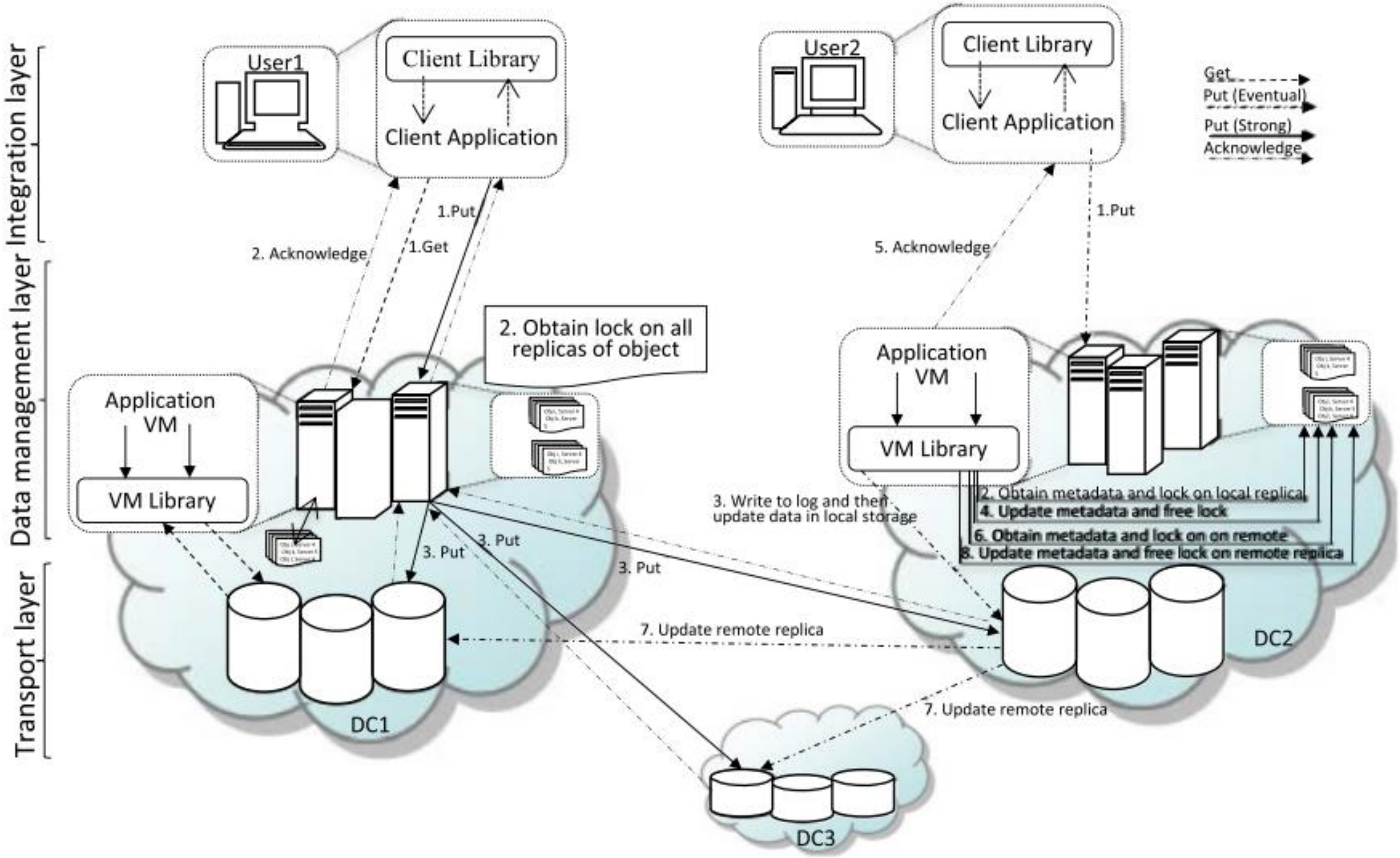
OneDrive

Dropbox

Storage as a Service (STaaS)

- What is it?
  - A business model in which a company rents space in their storage infrastructure to another company or individual.
- How does it work?
  - STaaS provider rents space
  - cost-per-gigabyte-stored and cost-per-data-transfer basis.
- Benefits
  - Shifting from Capital Expenditure to Operational Expenditure
  - Scale up/down at will (temporarily)

Cloud Storage



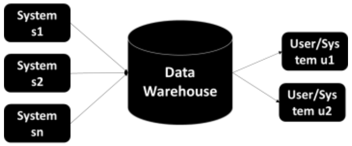
Mansouri et al., **Data Storage Management in Cloud Environments**, ACM CSUR 2018.

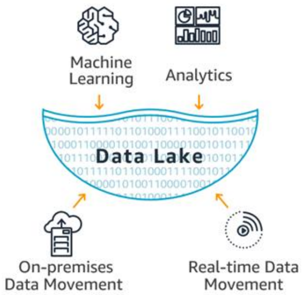
Cloud Computing

A data center

Source: https://resources.sei.cmu.edu/asset_files/WhitePaper/2010_019_001_28877.pdf

Summary





Data Storage - Summary