

Big Data and Data Analytics

Richard Lui

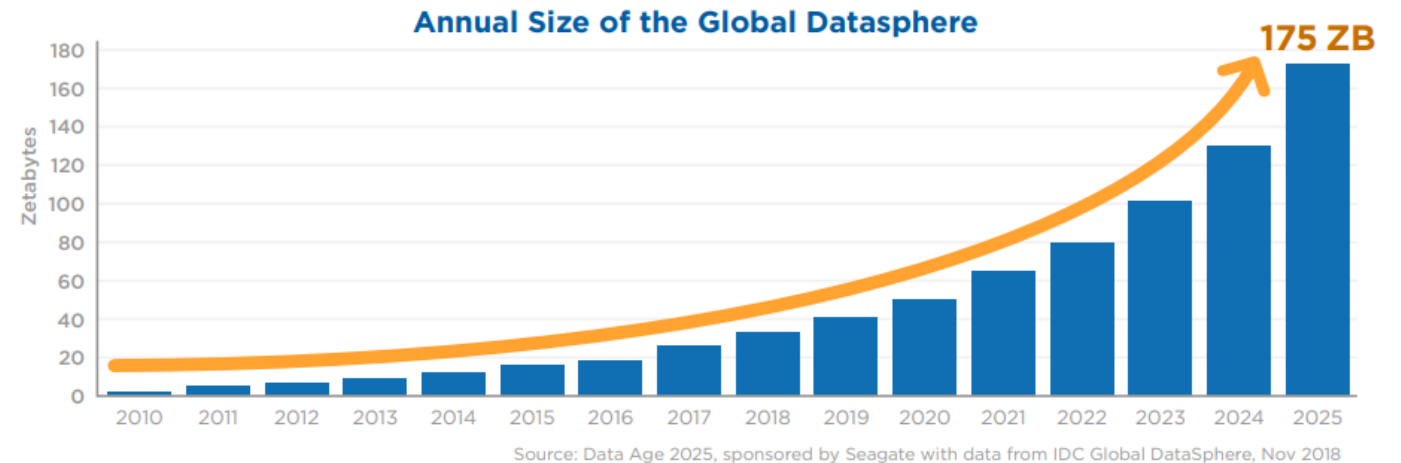
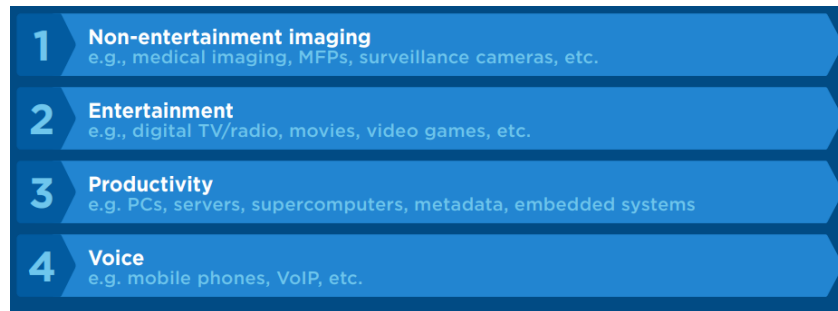
The Big Data Era

- Companies/Organizations are generating and keeping more and more data
- Enables the computer to learn on their own by analysing a vast amount of data
- Develop machines with the ability that are usually done by us human with our natural intelligence
- Big data: why should you care?
 - https://www.youtube.com/watch?v=ji18sDbWI_k



Global Datasphere

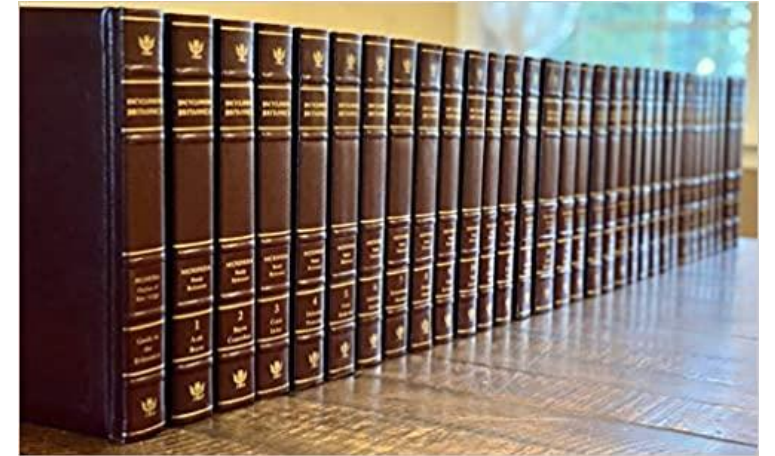
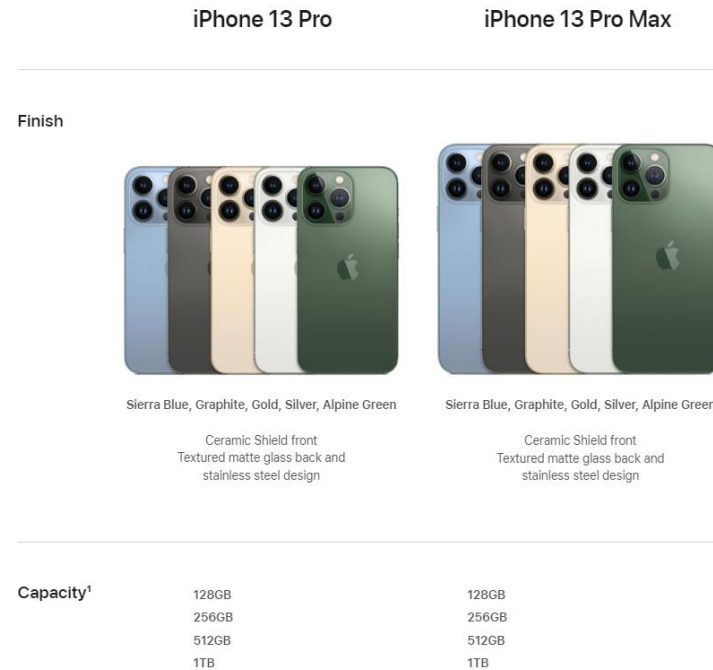
- *Global Datasphere* is a measure of all new data that is captured, created, and replicated in any given year across the globe.
- International Data Corporation (IDC) predicts the global datasphere will grow from 33 Zettabytes to 175 Zettabytes



How “large” is the data?

- One *Terabyte (TB)* = 1,000 *Gigabytes (GB)*.
 - A single TB could hold 1,000 copies of the Encyclopedia Britannica
 - All the X-rays in a large hospital

Name	Symbol	Value
Kilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}



- One *Petabyte (PB)* = 1,000 *Terabytes*
 - If you took all of the information from all US academic research libraries and lumped it all together, it would add up to 2 petabytes.
 - Google was said to process around 20 petabytes of data a day in 2008


Google Processing 20 Petabytes a Day

Google's MapReduce jobs are executed daily day, processing more than 20 petabytes of data per day.

Rich Miller | Jan 09, 2008



Google (GOOG) engineers Jeffrey Dean and Sanjay Ghemawat have published a new paper providing some details on MapReduce, the company's technology for processing the huge datasets for its web index. The paper notes that more than 100,000 MapReduce jobs are executed daily day, processing more than 20 petabytes of data per day. The paper was published in the January 2008 issue of *Communications of the ACM*. Full text is limited to members (a copy has been posted elsewhere on the web, but we won't link it here).

 **VERTIV.**

Vertiv Modular Solutions

GET STARTED NOW

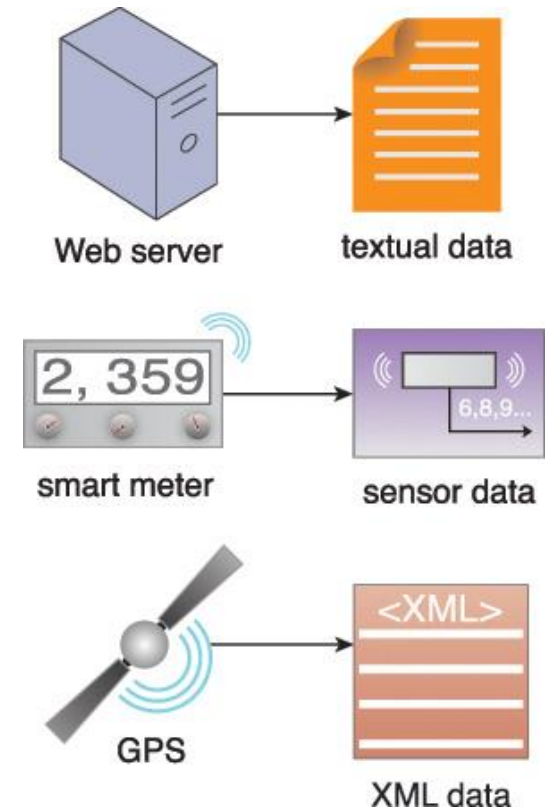
Scale With Confidence.

Rapidly Deploy

- One *Exabyte (EB)* = 1,000 Petabytes, or one billion gigabytes (GB)
 - Some technologists have estimated that all the words ever spoken by mankind would be equal to five Exabytes.
- One *Zettabytes (ZB)* = 1,000 Exabytes
 - As much information as there are grains of sand on all the world's beaches
 - If you were to store 175 zettabytes on DVDs, your stack of DVDs would be long enough to circle Earth 222 times.
 - If you attempted to download 175 zettabytes at the average current internet connection speed, it would take you 1.8 billion years to download.

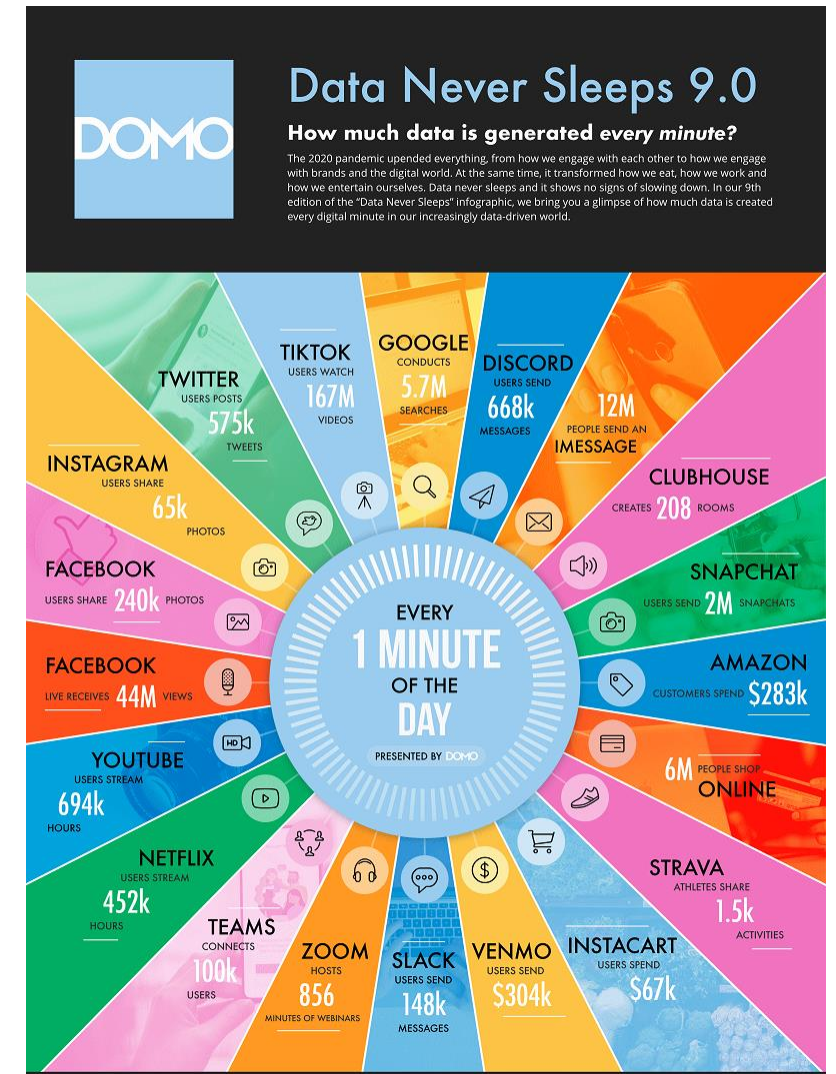
Where are the data coming from?

- Your every interaction with your computer or phone
- Your every interaction on social media
- Every time you walk down the street with a phone in your pocket, it's tracking your location through GPS sensors
- Every time you buy something with your credit cards or octopus card
- Every time you read an article online
- Every time you stream a song, movie or podcast
- ...



Explosion of data

- Exponential growth of the Internet and World Wide Web
- Transactions and interaction of users with e-commerce and mobile applications
- Social network activities
 - E.g. YouTube, Facebook, Instagram, Twitter
- Companies collect and store a large volume of data from different types of users
 - E.g. Google, Baidu, Netflix, Uber
- Internet of Things (IoT) and wireless sensors
 - Smart watch, thermostat, water heaters, smoke detectors, ...



A chart which provides an overview of what happens online every minute
<https://www.socialmediatoday.com/news/what-happens-on-the-internet-every-minute-2021-version-infographic/607586/>

Internet of Things (IoT)

- The Internet of things (IoT) describes physical objects with sensors, processing ability, software, and other technologies
- Connect and exchange data with other devices and systems over the Internet or other communications networks.



A smart toilet seat that measures blood pressure, weight, pulse and oxygen levels.



How Facebook track your data?

- How Facebook Tracks Your Data
 - https://www.youtube.com/watch?v=JAO_3EvD3DY
- Facebook has 2.89 billion active users, as of the second quarter of 2021 (Source: Statistica)
- Collect, store and analyze users data and behavior
 - e.g. the posts and pages liked by users, tags used for tagging photo, posted images
- Suggest posts and advertisement which match the users' preference

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.

Example: Facebook advertising

01:01

facebook

The Hong Kong Polytechnic University

Sponsored · 🌐

#理大創科開放日 告訴你創科如何改變世界! 立即報名, 參加開放日

polyu.edu.hk

理大創科開放日 - 創

21

YOUTUBE.COM

理大主校門揭幕典禮 PolyU Main Entrance Unveiling Ceremony

Learn more

授 and 243 others 9 Comments 23 Shares

Save video
Add this to your saved videos.

Hide ad
Never see this ad again.

Report ad
Tell us about a problem with this ad.

View edit history

Why am I seeing this ad?

Turn on notifications for this post

Why you're seeing this ad

The Hong Kong Polytechnic University wants to reach people like you, who may have:

- Shown interest in Technology, Science and more
- Communicated in Traditional Chinese (Taiwan), English (UK) or English (US)
- Set their age to 18 and older
- A primary location in Hong Kong

What else influences your ads

Your personalized ads may be based on other advertiser choices, your profile and activities—like websites you visit and ads you interact with—as well as other information not listed here. [Learn more about how ads work](#)

Facebook or Instagram added topics to your account based on your activity on Meta technologies, such as your engagement with certain Facebook pages and ads, as well as information you've provided on Meta technologies. [Learn more](#)

What you can do

You can review the full list of topics below that are related to this ad, and choose to see less of any topic.

If you choose to see less of a topic, you won't get as many related ads and advertisers can't target ads to you based on this topic. Changes you make to these topics will apply to all the Facebook and Instagram accounts in your Account Center.

Technology

See less

Science

See less

Innovation

See less

https://www.facebook.com/help/794535777607370?ref=learn_more_ip

Data Analytics

- **Data:** Any piece of information stored and/or processed by a computer or mobile device.
- **Data Analytics** refers to the technologies and processes that turn raw data into insight for making decisions and facilitates drawing conclusion from data



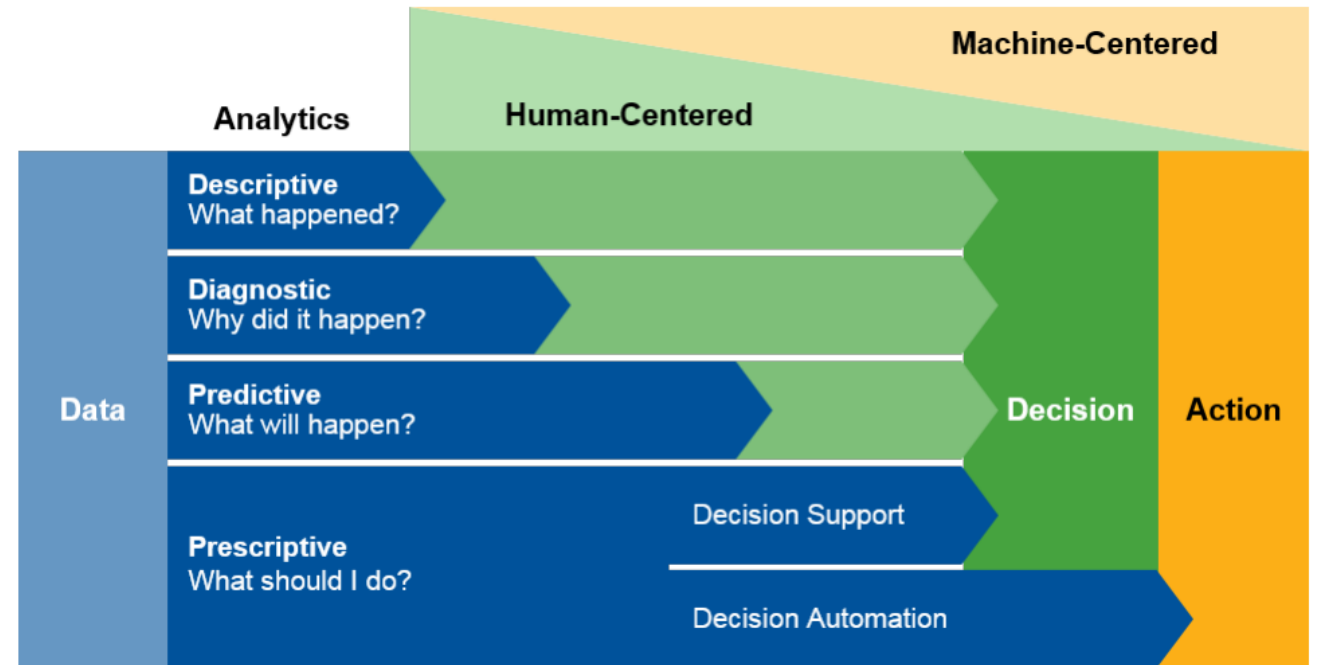
Clickstreams in an e-commerce website

```
{  
  "timestamp":"2022-08-12 03:01:58.732726",  
  "user_id":"35",  
  "click_id":"15cf179b9c9d483a...",  
  "event_name":"Search",  
  "user_ip":"11.22.33.44",  
  "additional_data":{  
    "engagement_time":40,  
    "product_id":12345  
  }  
}
```



Four data analytic capabilities

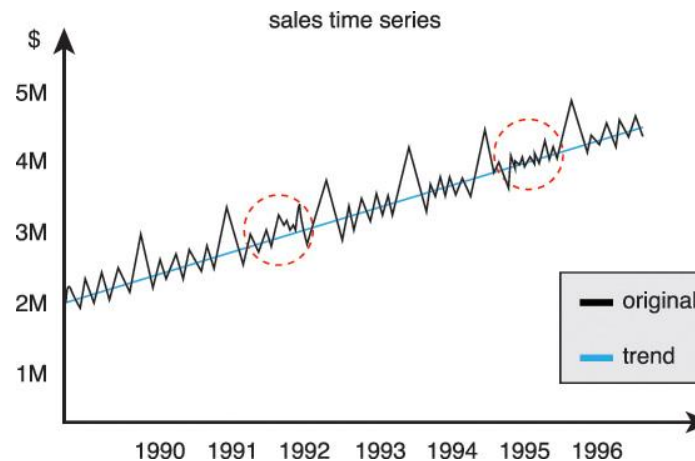
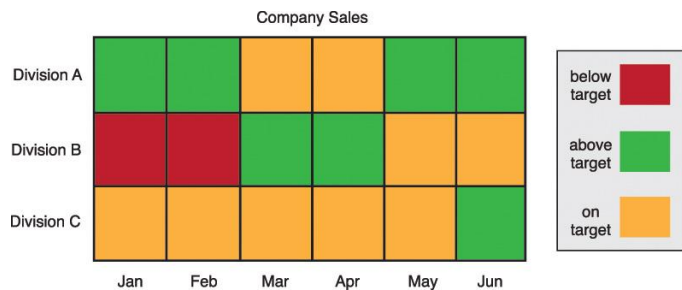
- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics



Source: Gartner's 2017 Planning Guide for Data and Analytics.

Descriptive analytics

- “*What has happened?*”
- Examples
 - What was the sales volume over the past 12 months?
 - What is the number of support calls received as categorized by severity and geographic location?
- It is estimated that 80% of generated analytics results are descriptive in nature.
- Descriptive analytics are often carried out via ad-hoc reporting or dashboards



Diagnostic analytics

- Diagnostic analytics aim to determine the *cause* of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
- Sample questions
 - Why were Q2 sales less than Q1 sales?
 - Why have there been more support calls originating from the Eastern region than from the Western region?

Predictive Analytics

- Generate *future predictions* based upon past events.
- Sample questions
 - What are the chances that a customer will default on a loan if they have missed a monthly payment?
 - What will be the patient survival rate if Drug B is administered instead of Drug A?

Prescriptive Analytics

- What should I do if “x” happens?
- Prescriptive analytics provide specific (prescriptive) recommendations to the user.
 - e.g. When is the best time to trade a particular stock?
- Various outcomes are calculated, and the best course of action for each outcome is suggested

Skills needed to become a data analyst

- Mathematical and statistical ability
- Programming languages
 - R, python, SQL
- Analytical mindset, problem solving and communication skills



4V of Big Data

- **Volume**

- A huge amount of data

- **Velocity**

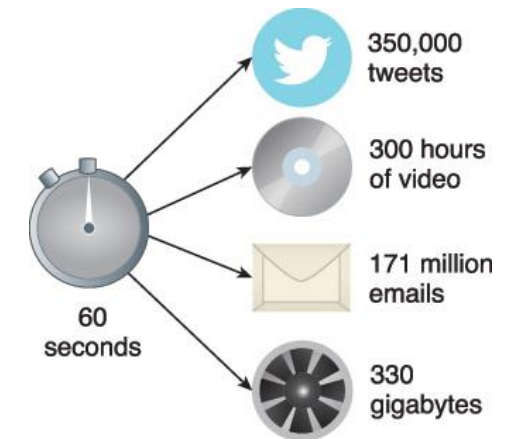
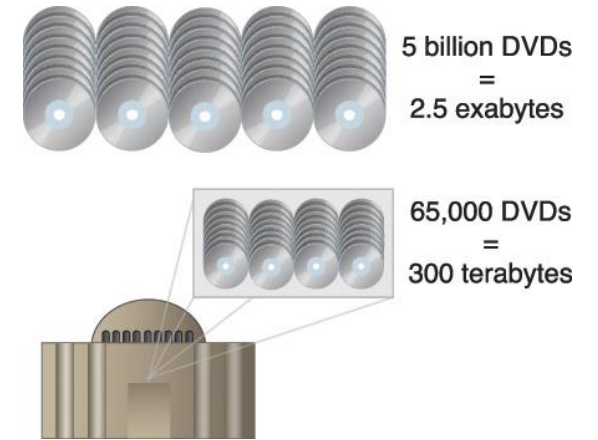
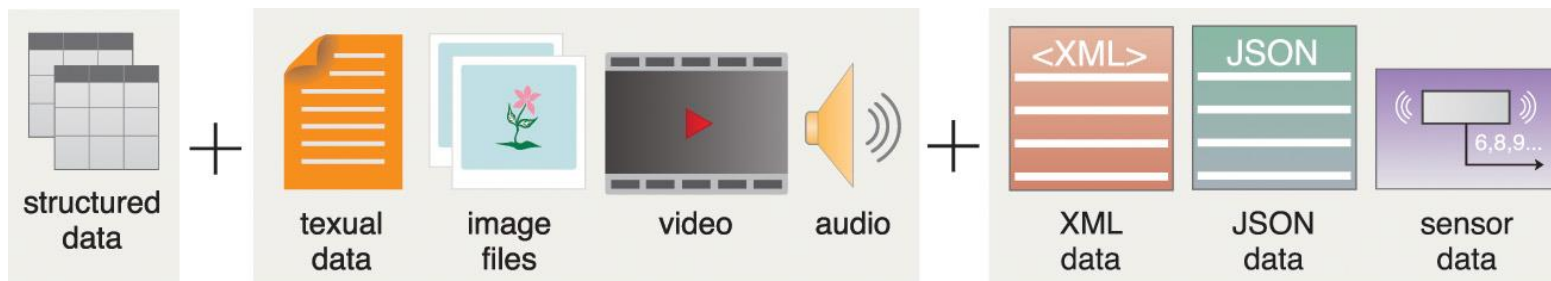
- High speed and continuous flow of data

- **Variety**

- Different types of structured, semi-structured and unstructured data coming from heterogeneous sources

- **Veracity**

- Data may be inconsistent, incomplete and messy

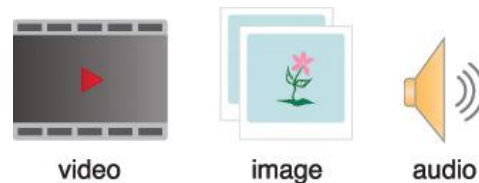


Structured vs. Unstructured data

- Structured data
 - Data conforms to a data model or schema and is often stored in tabular form.
- Unstructured data
 - Data that does not conform to a data model or data schema is known as unstructured data.
 - Estimated to makes up 80% of the data within any given enterprise.
- Semi-structured data
 - Non-tabular structure, but conform to some level of structure

course_id	title	dept_name	credits
BIO-101	Intro. to Biology	Biology	4
BIO-301	Genetics	Biology	4
BIO-399	Computational Biology	Biology	3
CS-101	Intro. to Computer Science	Comp. Sci.	4
CS-190	Game Design	Comp. Sci.	4
CS-315	Robotics	Comp. Sci.	3
CS-319	Image Processing	Comp. Sci.	3
CS-347	Database System Concepts	Comp. Sci.	3
EE-181	Intro. to Digital Systems	Elec. Eng.	3
FIN-201	Investment Banking	Finance	3
HIS-351	World History	History	3
MU-199	Music Video Production	Music	3
PHY-101	Physical Principles	Physics	4

Structured data



Unstructured data



Semi-structured data

Are the data structured/unstructured?


**Berghotel
Grosse Scheidegg**
3818 Grindelwald
Familie R. Müller

Rech. Nr. 4572	30.07.2007/13:29:17
Bar	Tisch 7/01
2xLatte Macchiato	à 4.50 CHF 9.00
1xGloki	à 5.00 CHF 5.00
1xSchweinsnitzel	à 22.00 CHF 22.00
1xChässpätzli	à 18.50 CHF 18.50

Total :	CHF 54.50
Incl. 7.6% MwSt	54.50 CHF: 3.85
Entspricht in Euro 36.33 EUR	
Es bediente Sie: Ursula	
MwSt Nr.: 430 234	
Tel.: 033 853 67 16	
Fax.: 033 853 67 19	
E-mail: grossescheidegg@bluewin.ch	

Starbucks updated their cover photo.
June 22 · 🌐

Taste the tropics. 🍍🥥 The all new Starbucks® Paradise Drink has arrived—a blissful combination of pineapple, passionfruit, and coconutmilk (US & Canada).



10K 4.7K Comments 1.6K Shares

Like Comment Share

Write a comment...

Christine Cappuccia
I wish Starbucks would eliminate the sugar in these types of drinks. I don't touch any of the cold drinks, as refreshing as they sound, because of the high sugar content.
Like Reply 4w 53

Barbara Dias
Christine Cappuccia I just posted a comment wishing the same thing! I'm a Bariatric patient and can't have it but loved them before!
Like Reply 4w 8

Mel Parsons
agreed- I wish Starbucks would cut down on the sugar, it is not good for anyone. I limit what i get here.
Like Reply 4w 7

Starbucks
June 22 · 🌐

Taste the tropics. 🍍🥥 The all new Starbucks® Paradise Drink has arrived—a blissful combination of pineapple, passionfruit, and coconutmilk (US & Canada).

10K 4.7K Comments 1.6K Shares

Like Comment Share

Most relevant ▾

Isabella Jordan
Sounds yummy, why is it not available in Britain?
Like Reply 10w 10

15 Replies

Marcey Nicole O'Malley Westerholm
Sounds like the beach. 🌊
Like Reply 9w 3

Starbucks Author
We can hear the waves of paradise with every sip.
Like Reply 9w

Most Relevant is selected, so some replies may have been filtered out.

Briana Alexandria Gleason
Got it as my birthday drink today! So delicious and refreshing 🌟
Like Reply 8w 8

It's estimated that 90% of the big data we generate is unstructured!

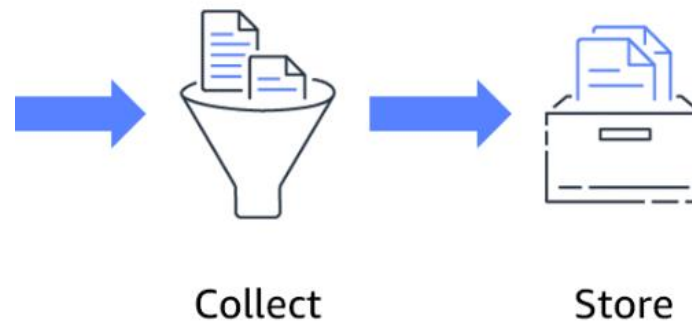
The big data processing cycle

Collect

- Collecting the raw data—such as transactions, logs, and mobile devices
- Permits developers to ingest a wide variety of data

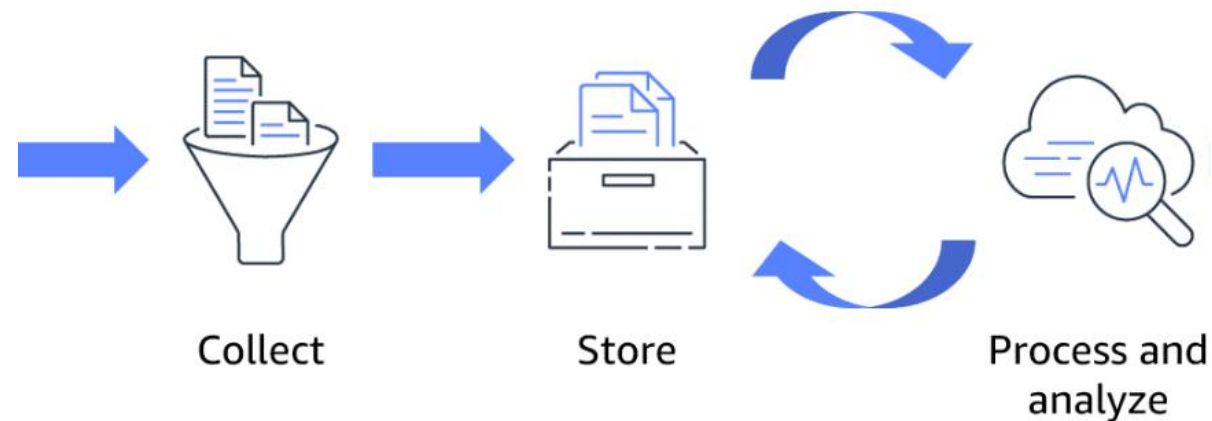
Store

- Requires a secure, scalable, and durable repository to store data before or after the processing tasks.



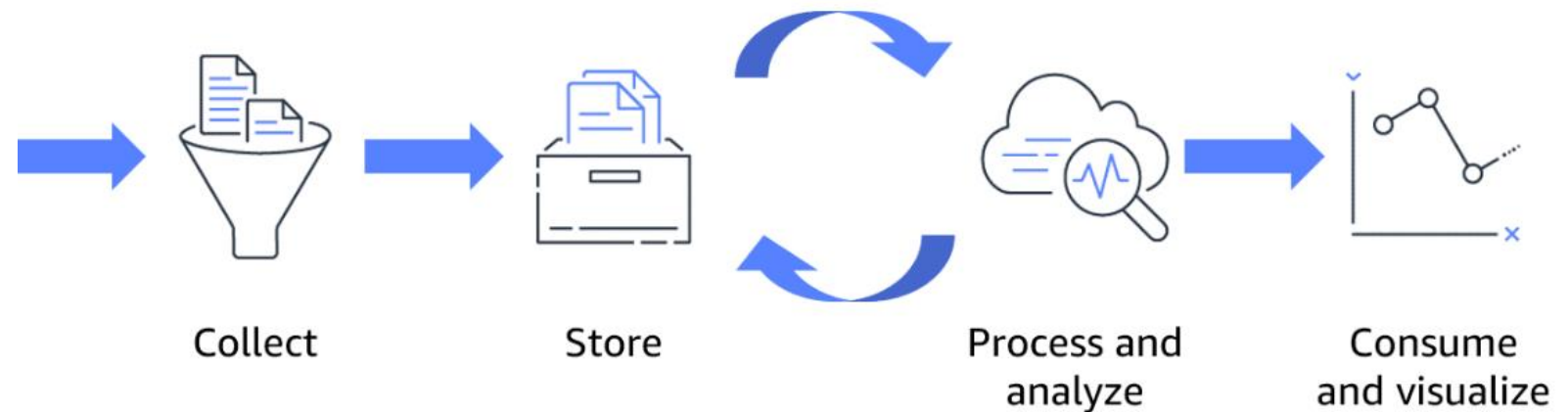
Process and analyze

- Data is transformed from its raw state into a consumable format
- Usually by means of sorting, aggregating, joining, and performing more advanced functions and algorithms.
- The resulting datasets are then stored for further processing or made available for consumption with business intelligence and data visualization tools.



Consume and visualize

- Data is made available to stakeholders through self-service business intelligence and data visualization tools to allow fast and easy exploration of datasets.
- Users might also consume the resulting data in the form of statistical *predictions* (in the case of predictive analytics) or recommended actions (in the case of prescriptive analytics)



Databases

- Designed to store and handle transaction data (live, real time data)
- *Relational databases* (e.g. Mysql) store data in tables with fixed rows and columns.
- *Non-relational databases (NoSQL)* store data in a variety of data models (e.g. JSON)
 - More flexible schema (how the data is organized)



Relational

ID	first_name	last_name	cell	city	year_of_birth	location_x	location_y
1	'Mary'	'Jones'	'516-555-2048'	'Long Island'	1986	'-73.9876'	'40.7574'

ID	user_id	profession
10	1	'Developer'
11	1	'Engineer'

ID	user_id	name	version
20	1	'MyApp'	1.0.4
21	1	'DocFinder'	2.5.7

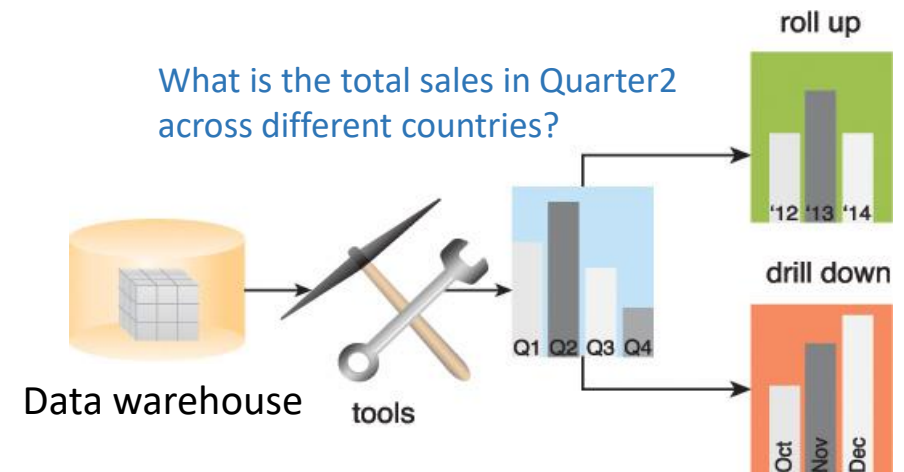
ID	user_id	make	year
30	1	'Bentley'	1973
31	1	'Rolls Royce'	1965

MongoDB

```
{
  first_name: "Mary",
  last_name: "Jones",
  cell: "516-555-2048",
  city: "Long Island",
  year_of_birth: 1986,
  location: {
    type: "Point",
    coordinates: [-73.9876, 40.7574]
  },
  profession: ["Developer", "Engineer"],
  apps: [
    { name: "MyApp",
      version: 1.0.4 },
    { name: "DocFinder",
      version: 2.5.7 }
  ],
  cars: [
    { make: "Bentley",
      year: 1973 },
    { make: "Rolls Royce",
      year: 1965 }
  ]
}
```

Data Warehouse

- Data warehouse is a *giant* database storing *highly structured information* that is *optimized for analytics*.
- Typically store *current* and *historical data* from one or more systems and disparate data sources
 - May not reflect the most up-to-date state of the data.
 - Business analysts and data scientists can connect data warehouses to explore the data, look for insights, and generate reports for business stakeholders.
- Examples
 - Google BigQuery, Amazon Redshift.



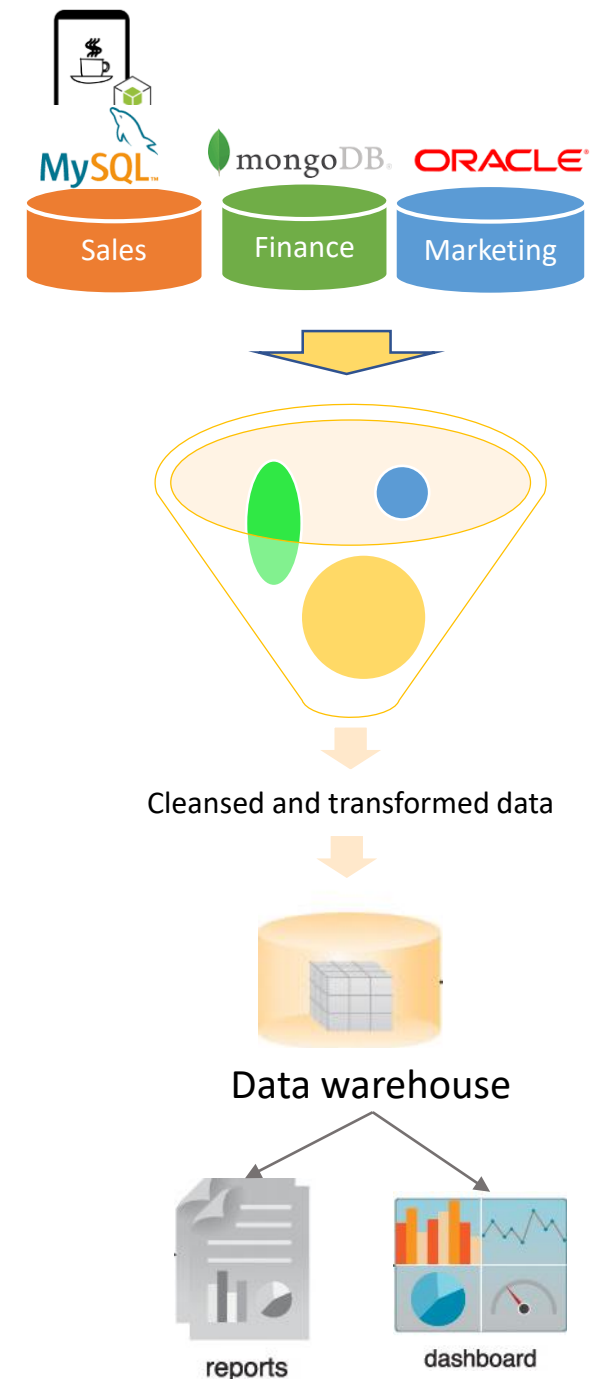
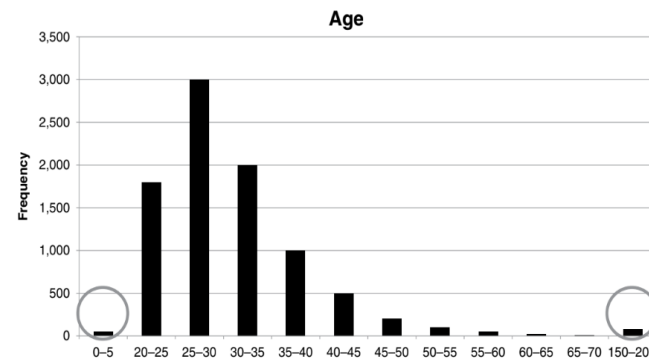
Extract, transform, load (ETL)

- The ETL processes move data from its original source (e.g. database or other sources) to the data warehouse on a regular schedule (e.g., hourly or daily)
- **Extract:** Extract data from homogeneous/heterogeneous sources.
- **Transform:** Clean the data and transform the data into appropriate format
- **Load:** Insert data into the target data warehouse

Missing data

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800	?	620	Churner
2	28	1,200	Single	?	Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married	?	Nonchurner
6	44	?	?	?	Nonchurner
7	22	1,200	Single	?	Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34	?	Single	?	Churner
10	50	2,100	Divorced	?	Nonchurner

Outliers

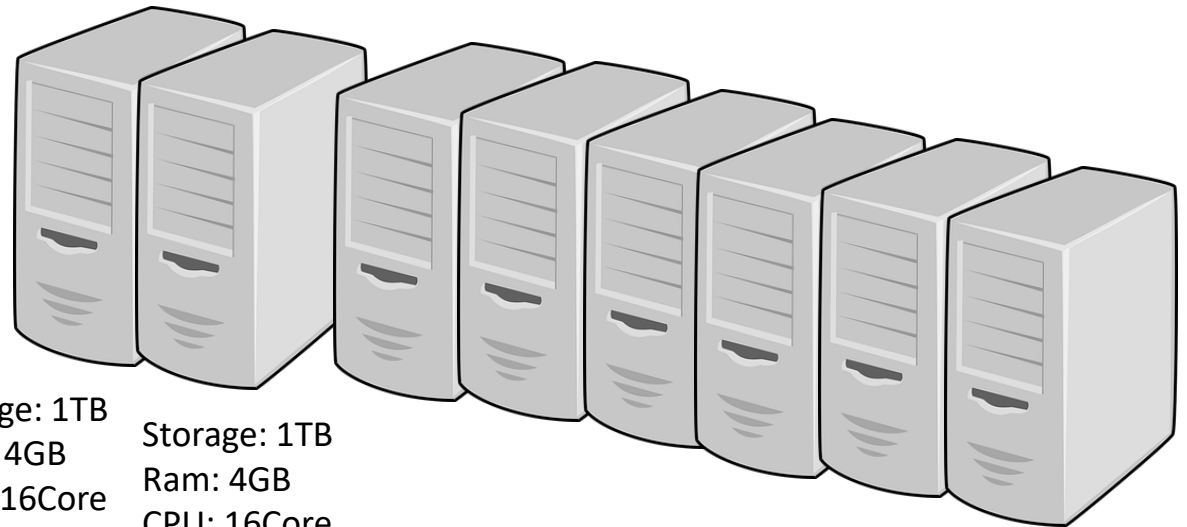


Solving the big data challenges

- How Facebook Manage 1.2 Billion Users Data
 - https://www.youtube.com/watch?v=_1uQf8dVLZY
- **Scaling up (Vertical scaling)**
 - Have a supercomputer with enormous amounts of storage attached to an extremely fast network.
- **Scaling out (Horizontal scaling)**
 - Have a lot of smaller computers, each with a modest amount of storage, connected by networking.



Storage: 1TB→2TB
Ram: 4GB→**8**GB
CPU: 16Core →32 Core



Storage: 1TB
Ram: 4GB
CPU: 16Core

Storage: 1TB
Ram: 4GB
CPU: 16Core

Which approach works better for solving the big data challenges?

Processing of Big Data

- The challenges of Big Data cannot be handled easily by traditional storage technology, e.g. databases
- *Hadoop*
 - A framework that allows for storing a large amount of data and the distributed processing of large data sets across clusters of computers
- *MapReduce*
 - a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster.
- *Apache Spark*
 - An open-source unified analytics engine for large-scale data processing



The Hadoop Eco-system

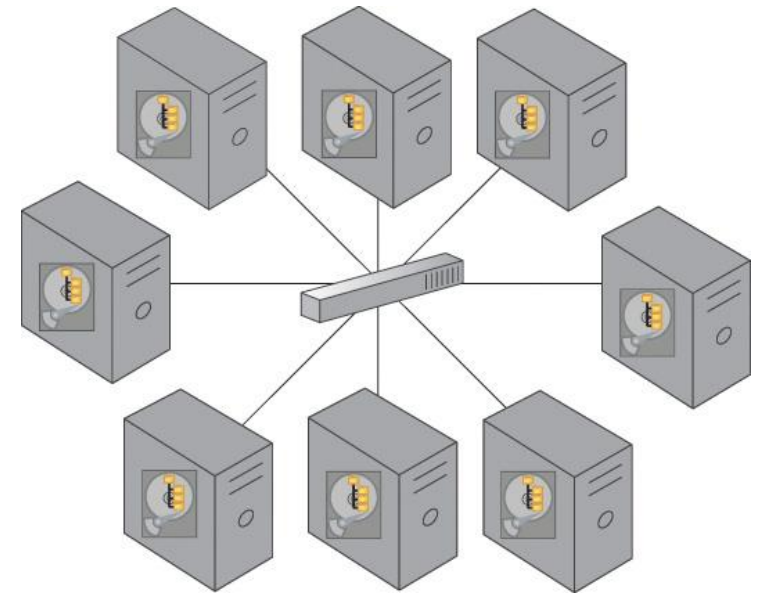


Source: Simplylearn

<https://www.simplilearn.com/tutorials/hadoop-tutorial/hadoop-ecosystem>

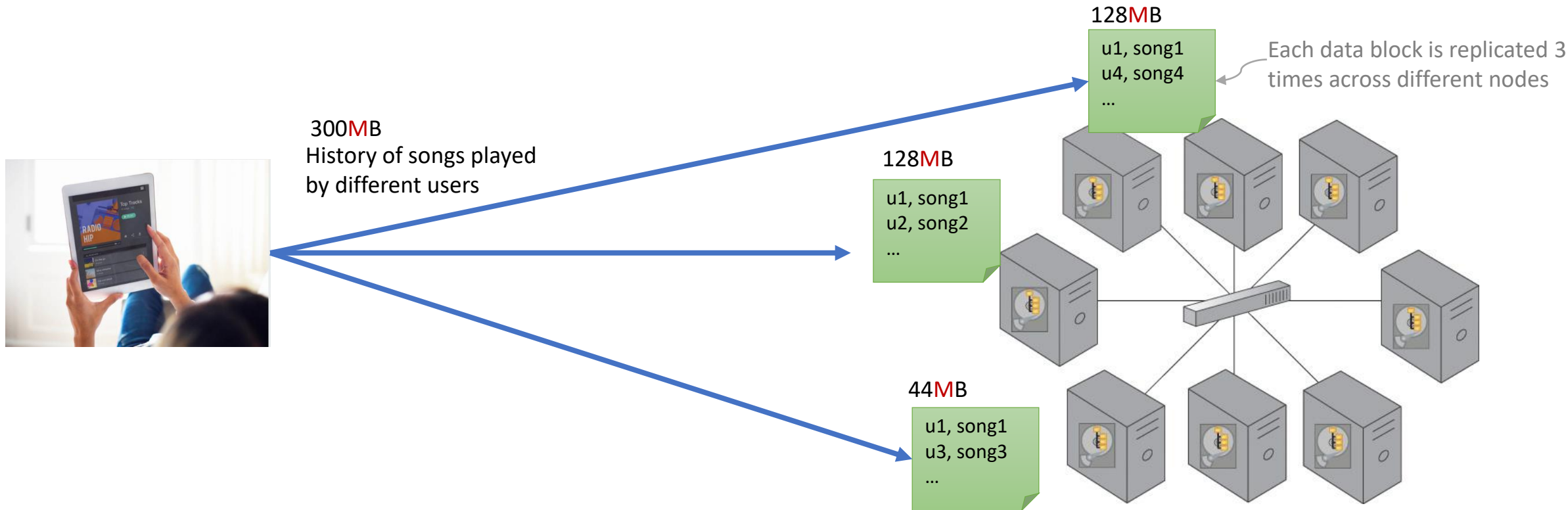
Distributed File Systems

- A *cluster* is a tightly coupled collection of servers, or nodes.
- A *distributed file system* can allow us to store large files which spread across the nodes of a cluster
 - E.g. Hadoop Distributed File System (HDFS).



Splitting large dataset

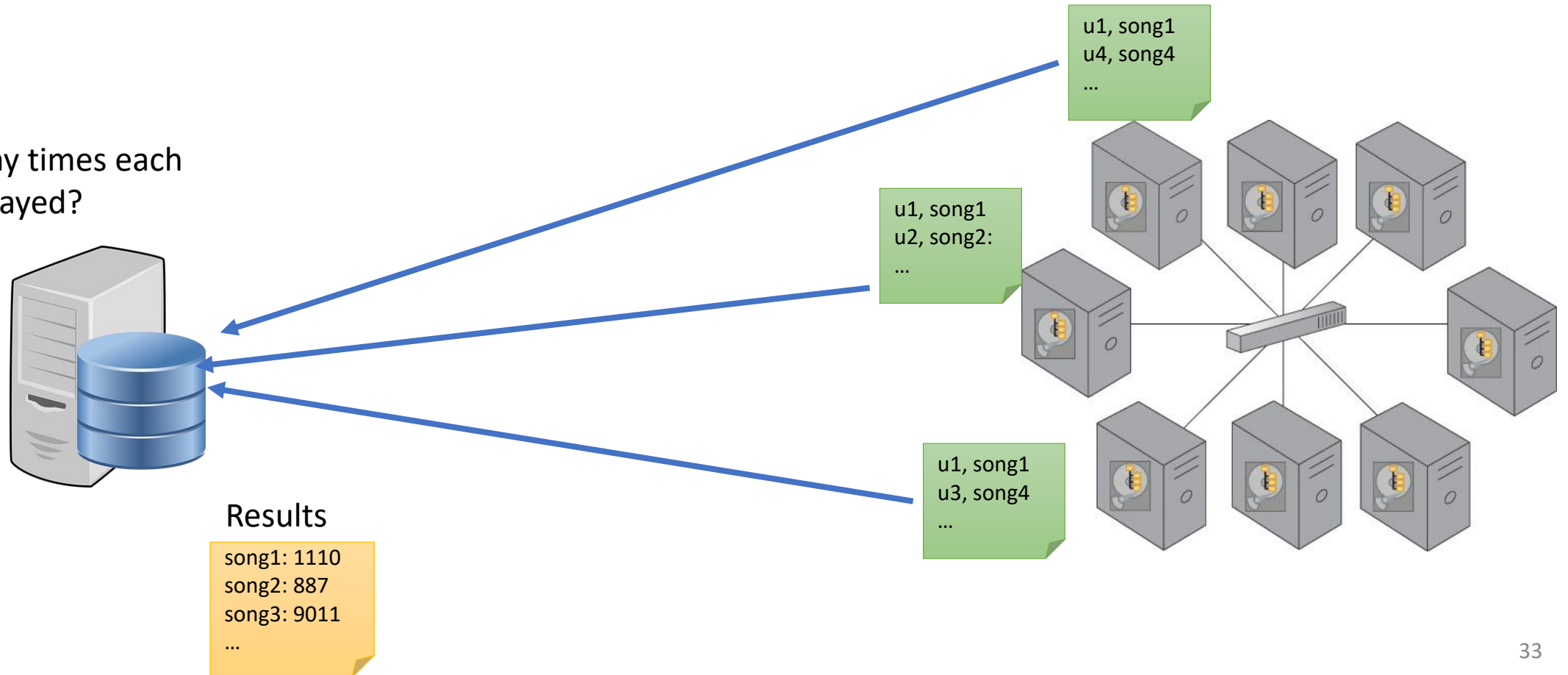
- Split large dataset into smaller data blocks and stored in different nodes
- In Hadoop, each block contains 128 MB of data and replicated three times by default
 - Replication Factor: The number of times Hadoop framework replicate each and every data block.



Traditional approach

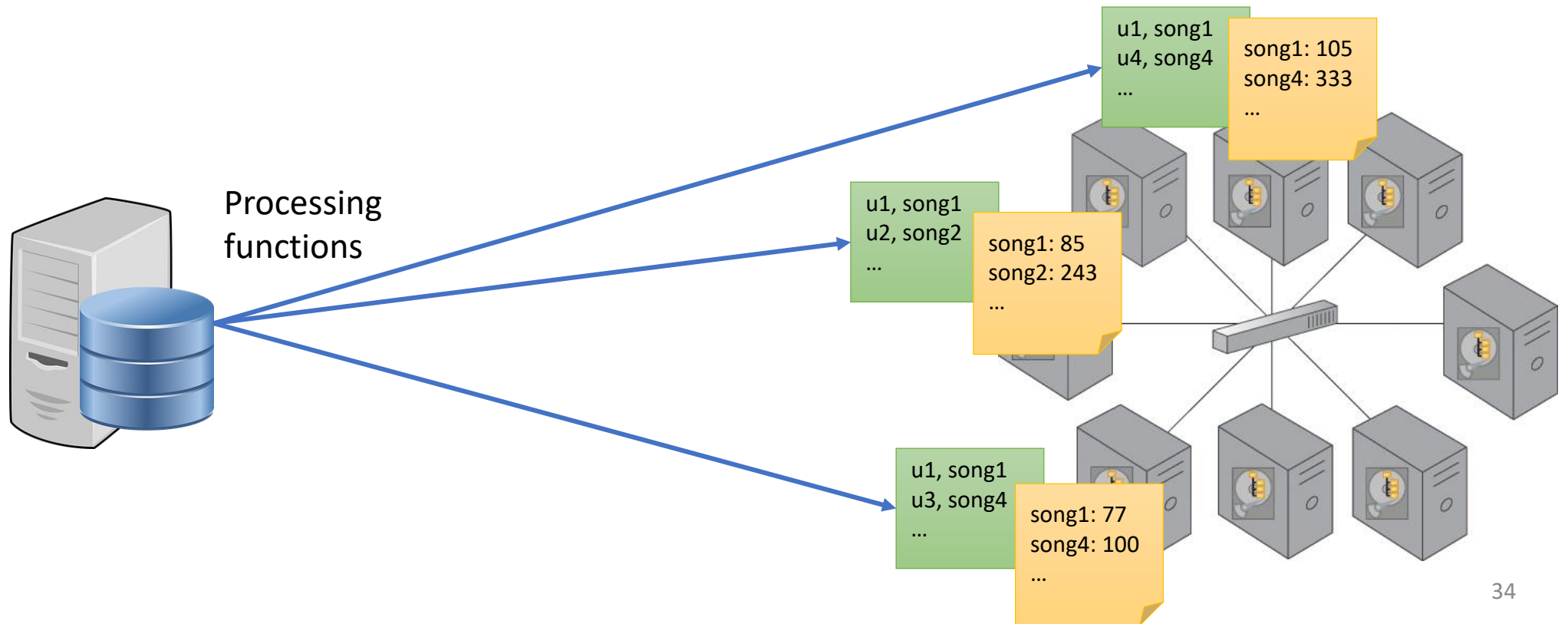
- Moving huge amount data to the processing unit is costly
- The processing unit becomes the bottleneck

How many times each song is played?



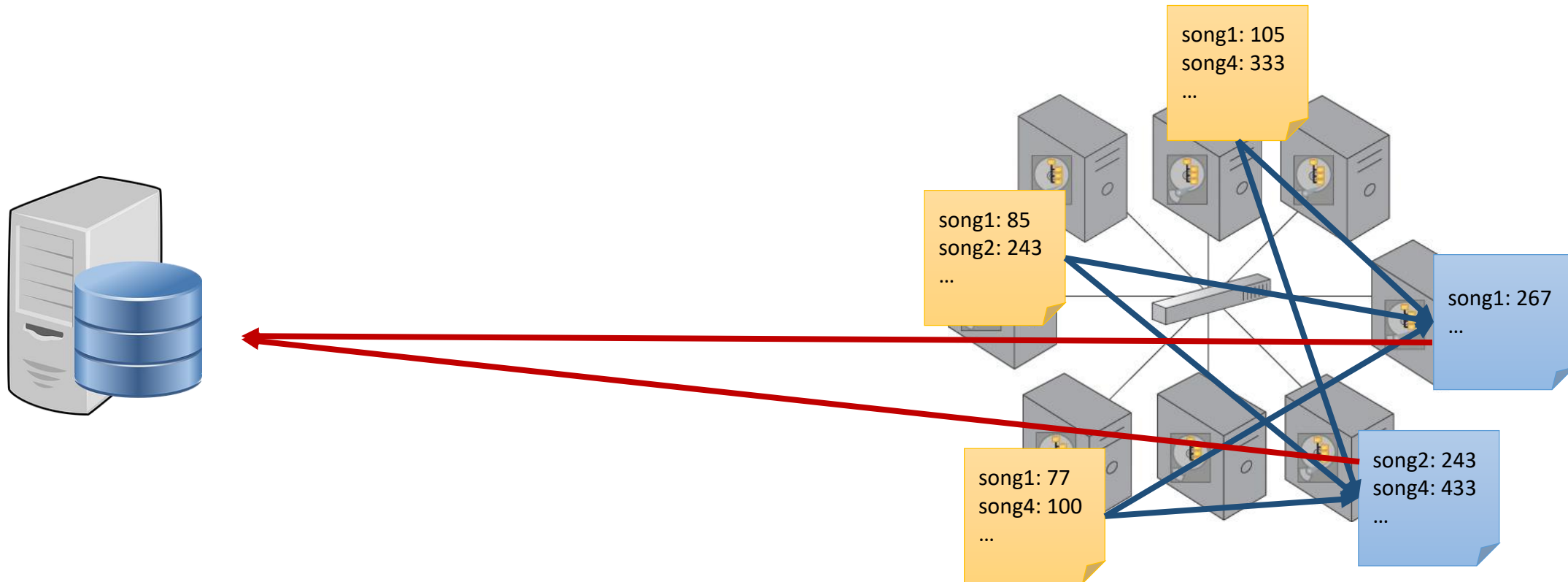
Map function

- Instead of moving data to the processing unit, we are moving the processing unit to the data
- MapReduce consists of two distinct tasks — Map and Reduce.
 - **Map:** process data to create key-value pairs in parallel



Reduce function

- MapReduce consists of two distinct tasks — Map and Reduce.
 - **Map:** process data by workers based on where data is stored
 - **Reduce:** Aggregate results by the “reduce workers”



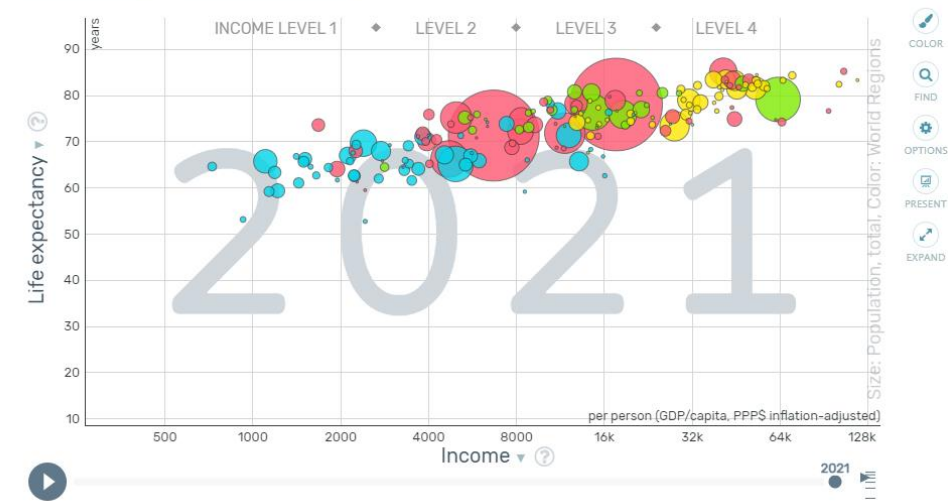
Visualization

- Creation and study of the visual representation of data
- One of the most important tools for data analytics/science.

World Health Chart

This graph is like a world map for health and wealth. Push play to see it change over time. Scroll down to see Hans Rosling explain it in a video.

[Printable PDF](#) — [Fullscreen version](#)



<https://www.gapminder.org/fw/world-health-chart>

Importance of visualization

- A good visualization will clearly answer your question about a data set of interest.
- A *great* visualization will suggest even what the question was itself without additional explanation.

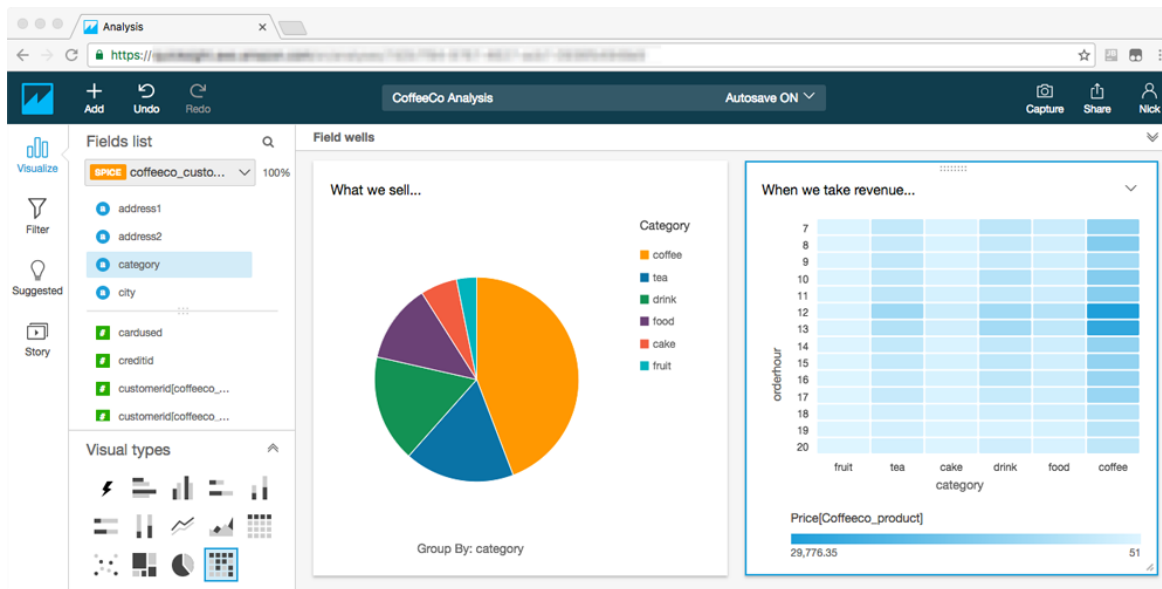
"The greatest value of a picture is when it forces us to notice what we never expected to see."

"The simple graph has brought more information to the data analyst's mind than any other device. "

- John Tukey

Dashboards

- Dashboard is a read-only snapshot of an analysis that you can share with other users for reporting purposes.



AWS QuickSight

<https://aws.amazon.com/quicksight>