Big Data

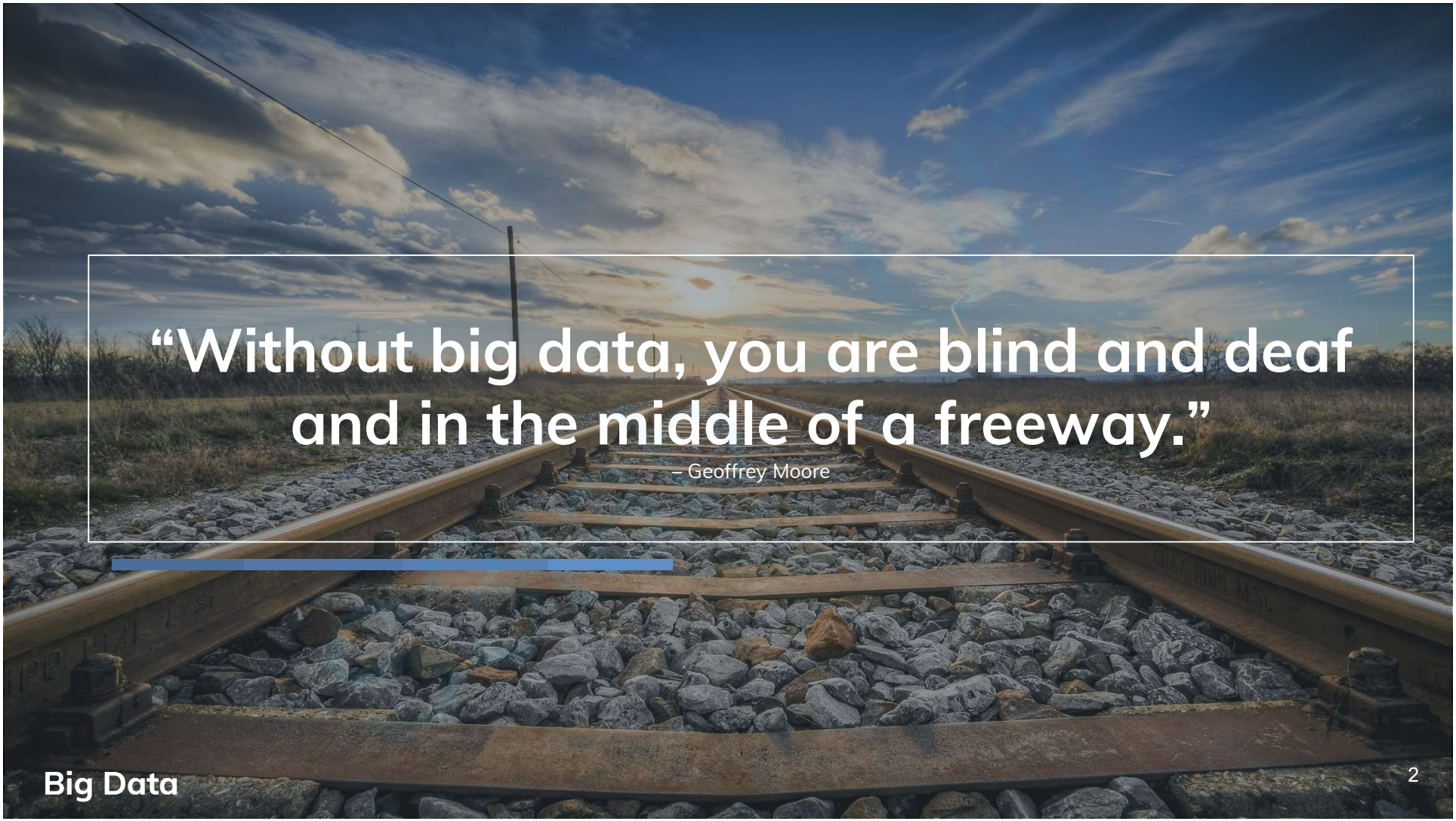
(Understanding about Big data)

Trong-Hop Do

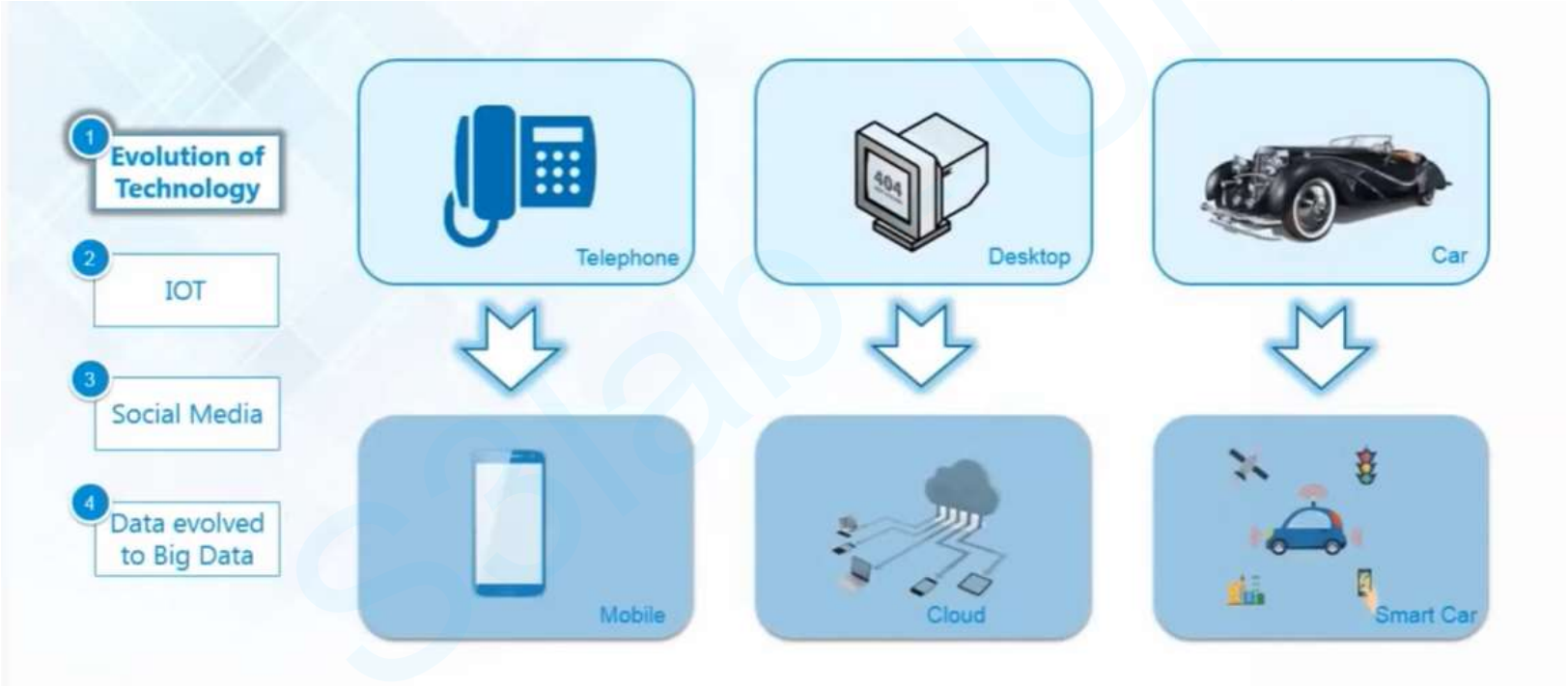
September 8th, 2020

S3Lab
Smart Software System Laboratory

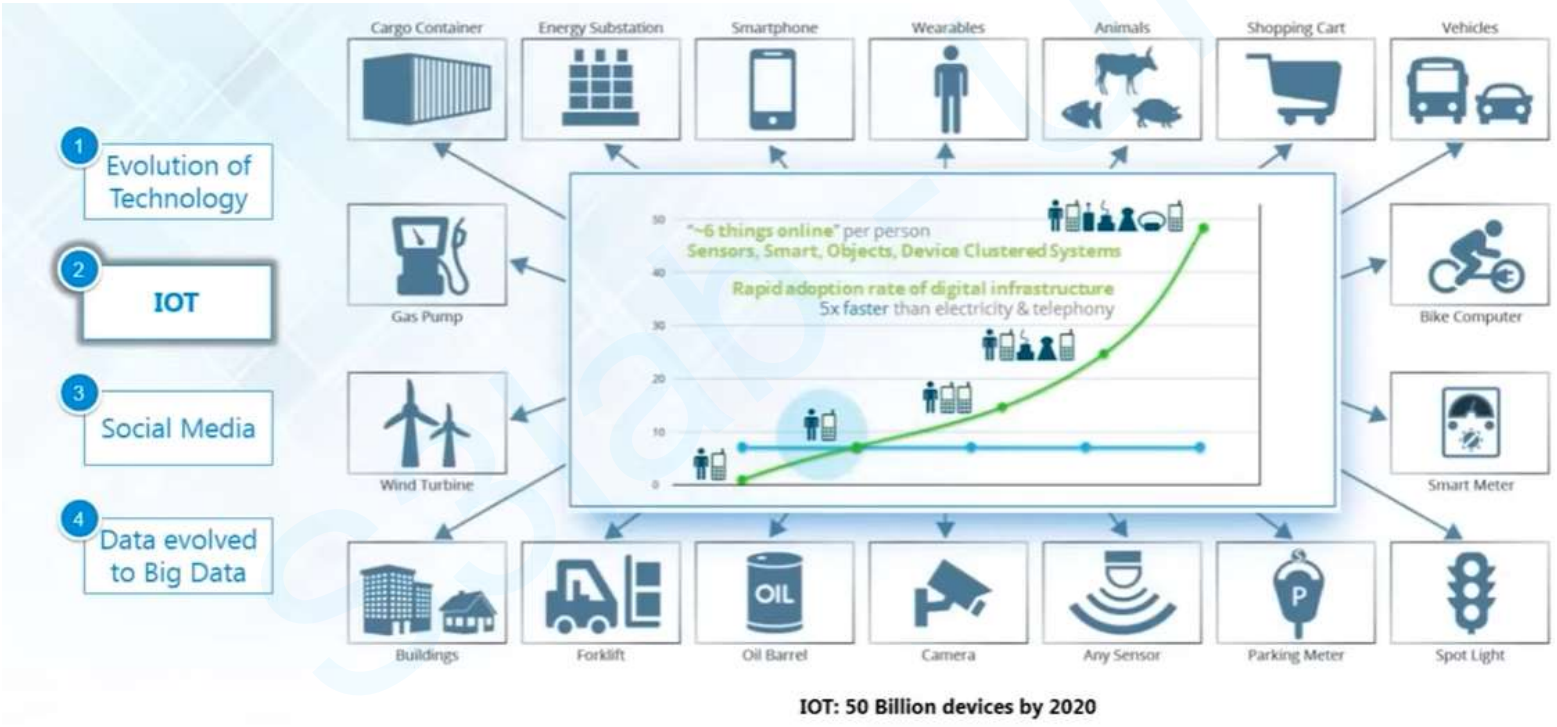
s3lab - UIT


“Without big data, you are blind and deaf and in the middle of a freeway.”
– Geoffrey Moore

Evolution of Technology



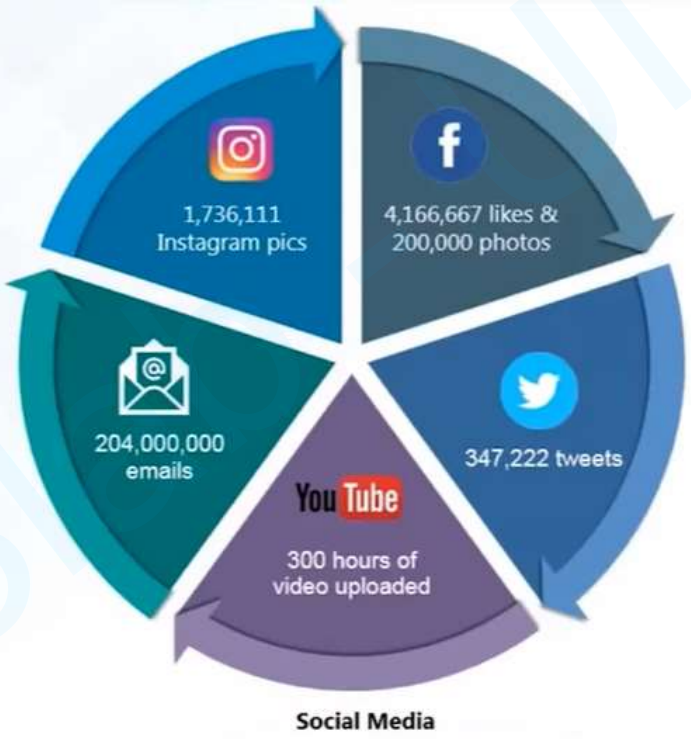
IOT



IOT: 50 Billion devices by 2020

Social media

1. Evolution of Technology
2. IOT
3. **Social Media**
4. Data evolved to Big Data



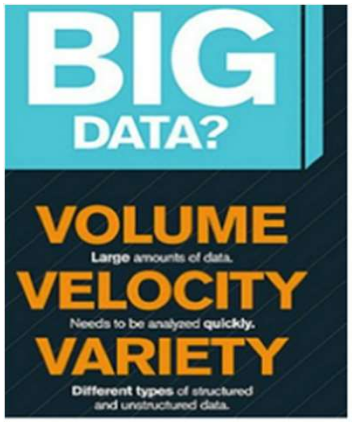
Other factors

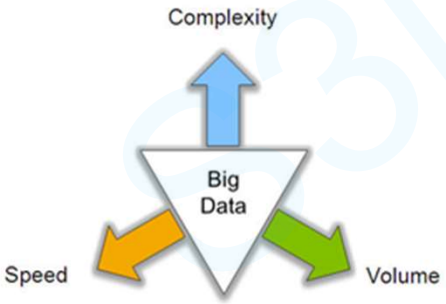
What is Big Data

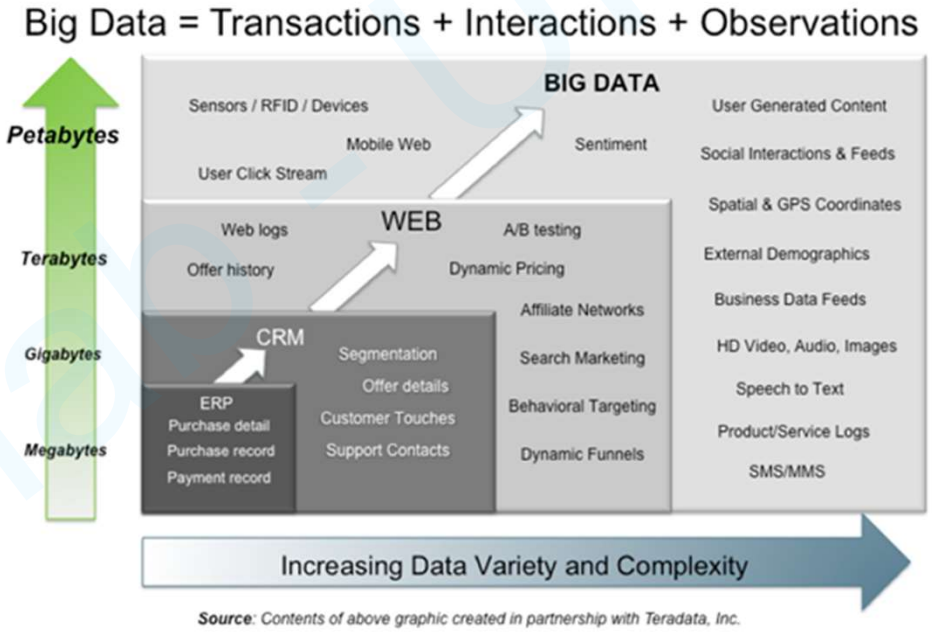
- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Challenges: Capture, Curation, Storage, Search, Sharing, Transfer, Analysis, and Visualization.

S3lab - UIT

Big Data: 3V's



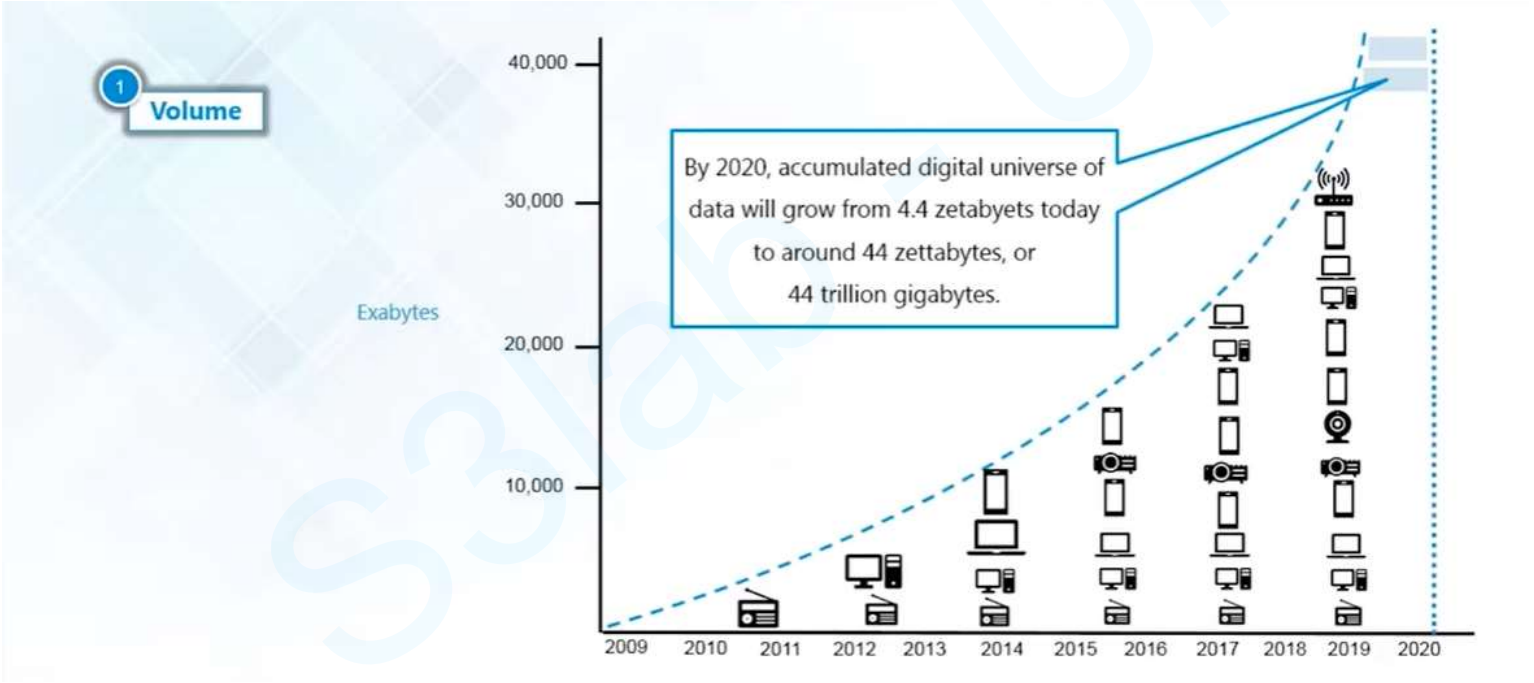




Source: Contents of above graphic created in partnership with Teradata, Inc.

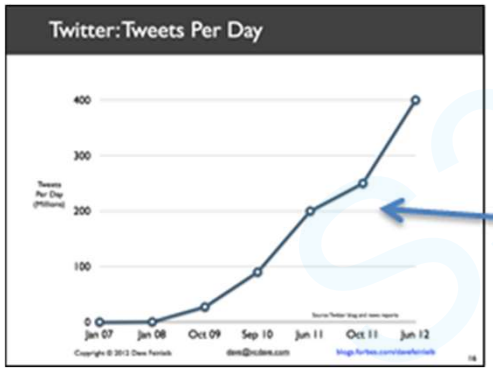
Big Data: 3V's

Volume (scale)

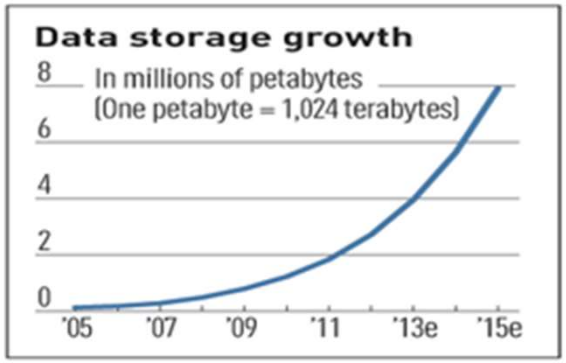


Big Data: 3V's

Volume (scale)



Exponential increase in collected/generated data



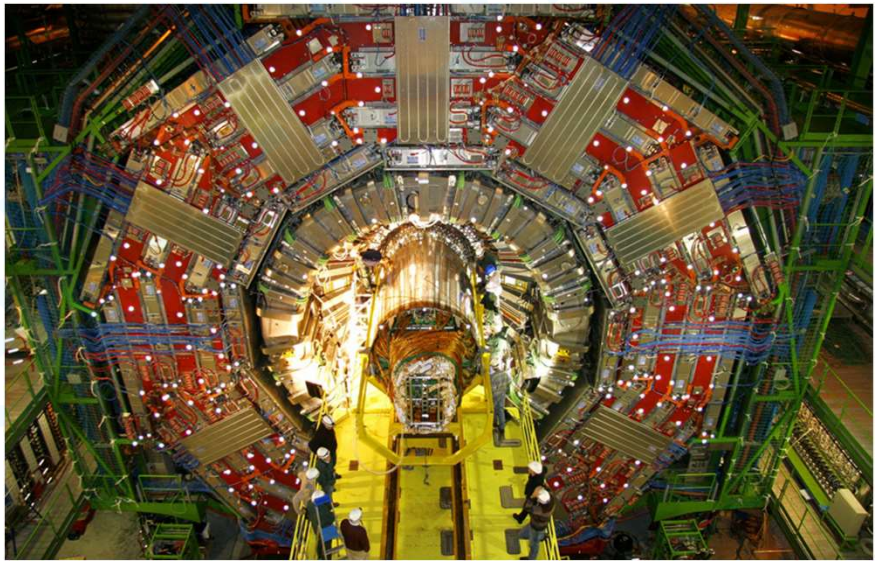
Big Data: 3V's

Volume (scale)

Big Data: 3V's

Volume (scale)

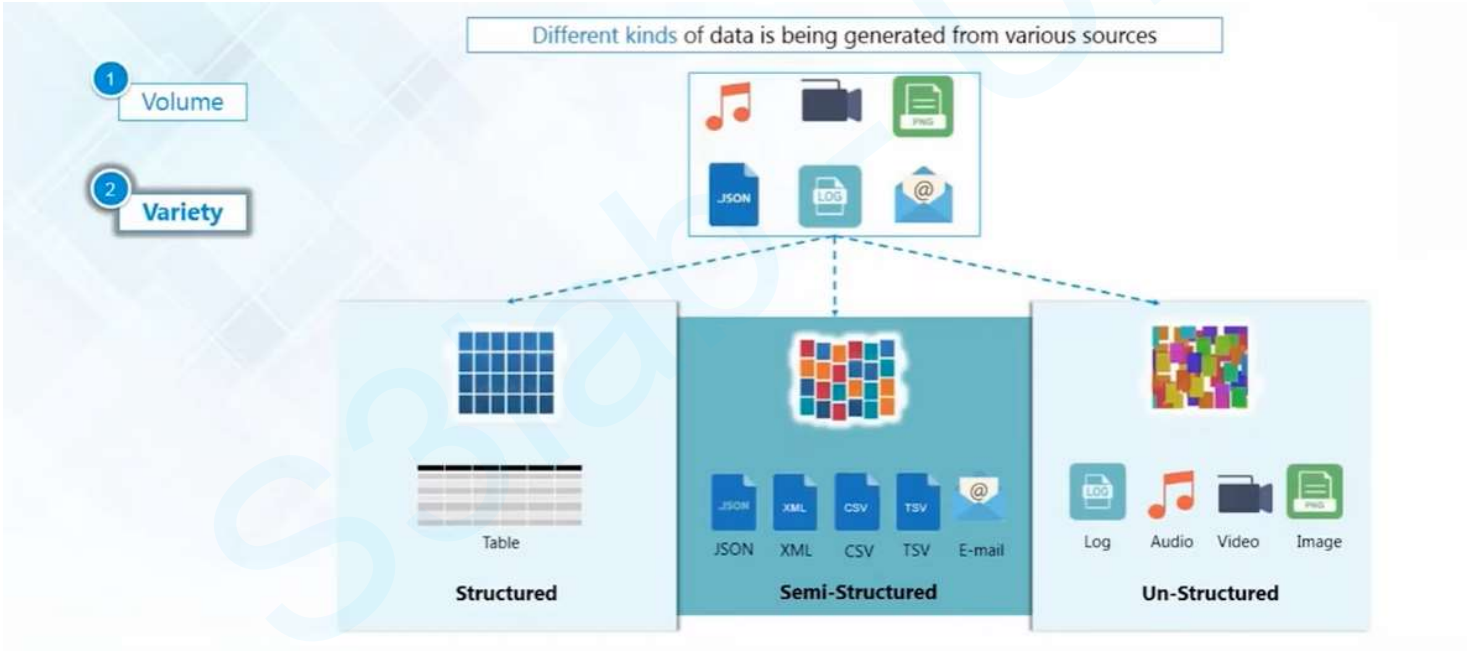
Earthscope - 67 terabytes of data



CERN's Large Hydron Collider (LHC) generates 15 PB a year

Big Data: 3V's

Variety (Complexity)



Big Data: 3V's

Variety (Complexity)

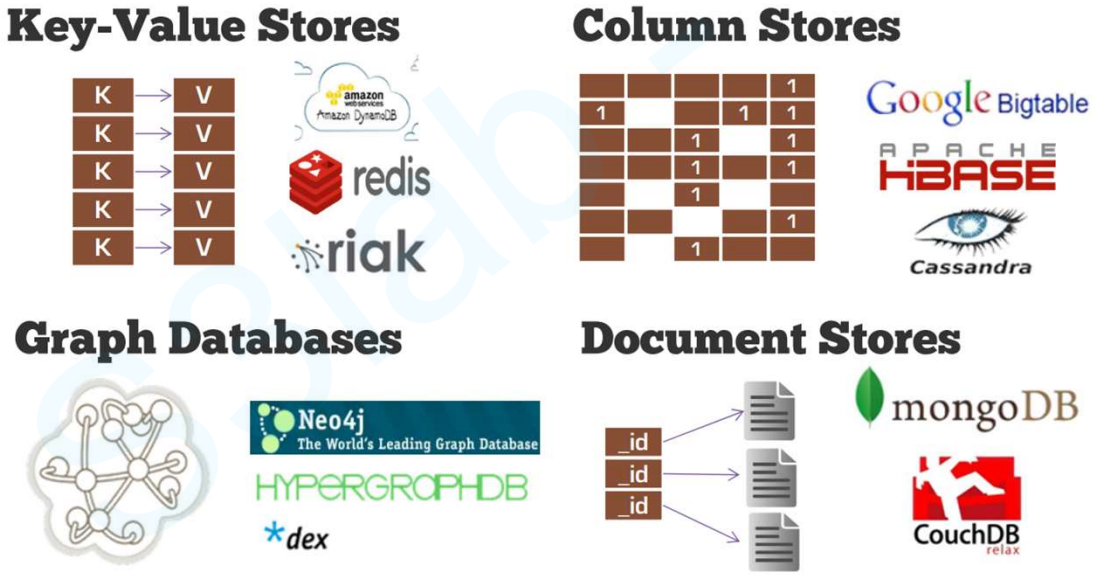
- Big data could be of three types
  - **Structured**: The data that can be stored and processed in a fixed format (fixed schema) is called as Structured Data. Ex. RDBMS
  - **Semi-Structured**: not have a formal structure of a data model, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze. Ex. XML files or JSON documents.
  - **Unstructured**: Text Files and multimedia contents like images, audios, videos are example of unstructured data. The unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization are unstructured.

Sclab - UIT

Big Data: 3V's

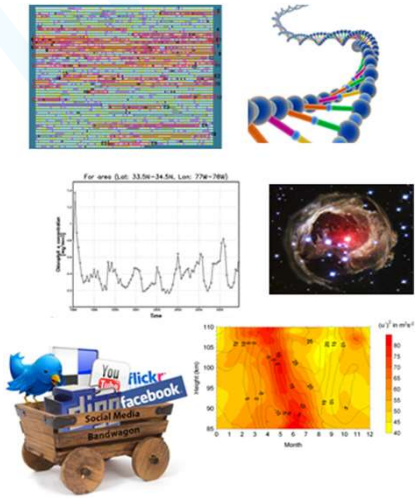
Variety (Complexity)

- Semi-Structured, NoSQL



Big Data: 3V's

Variety (Complexity)



- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web, log)
- Semi-structured Data (XML)
- Graph Data: Social network, Semantic web (RDF)…
- Streaming Data: You can only scan the data once
- A single application can be generating / collecting many types of data
- Big Public Data (online, weather, finance, etc.)

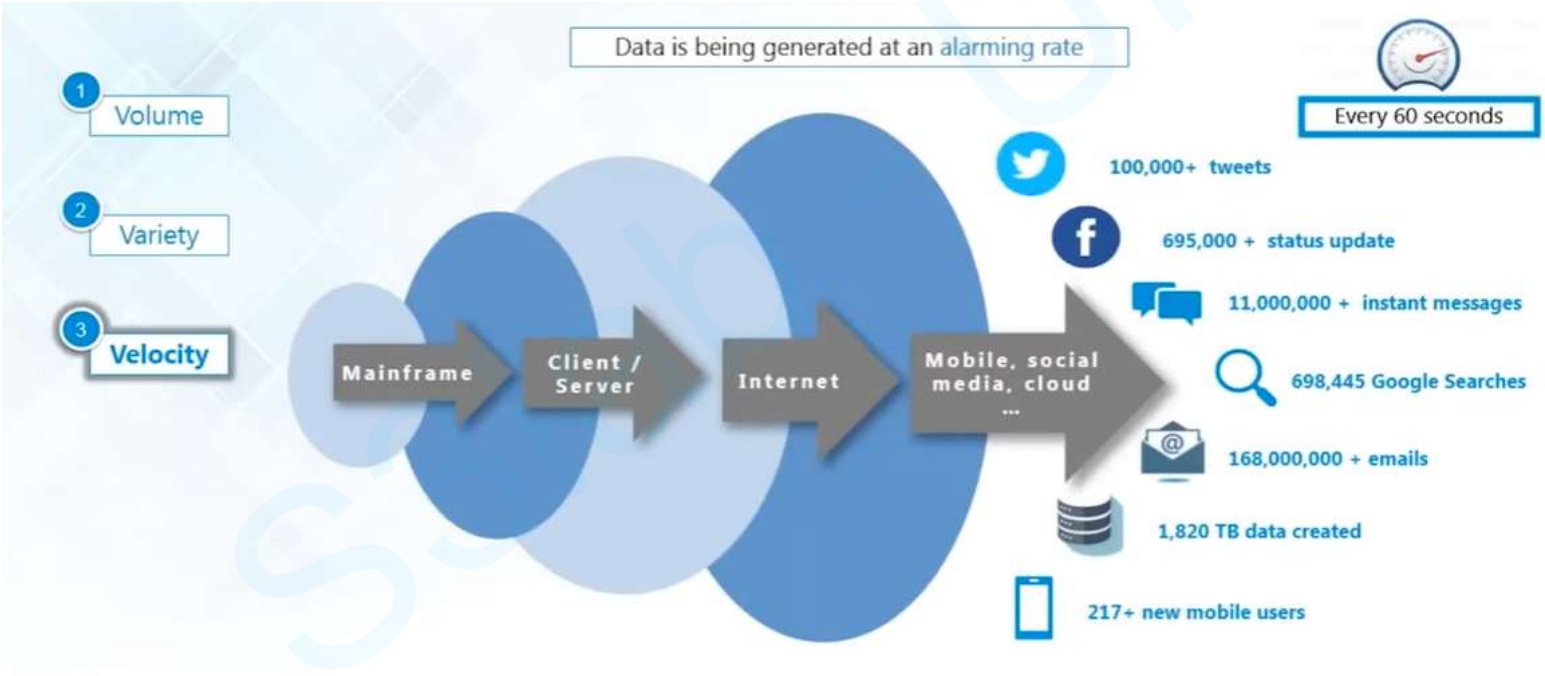
To extract extract knowledge ➠ all these types of data need to linked together

Big Data: 3V's

Variety (Complexity)

Big Data: 3V's

Velocity (Speed)



Big Data: 3V's

Velocity (Speed)

- Data is begin generated fast & need to be processed fast
- Online Data Analytics
- Late decisions ➟ missing opportunities
- Examples
  - **E-Promotions**: Base on your current location, your purchase history, what you like ➟ send promotions right now for store next to you
  - **Healthcare monitoring**: sensors monitoring your activities and body ➟ any abnormal measurements require immediate reaction

Slab - UIT

Big Data: 3V's

Velocity (Speed)

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

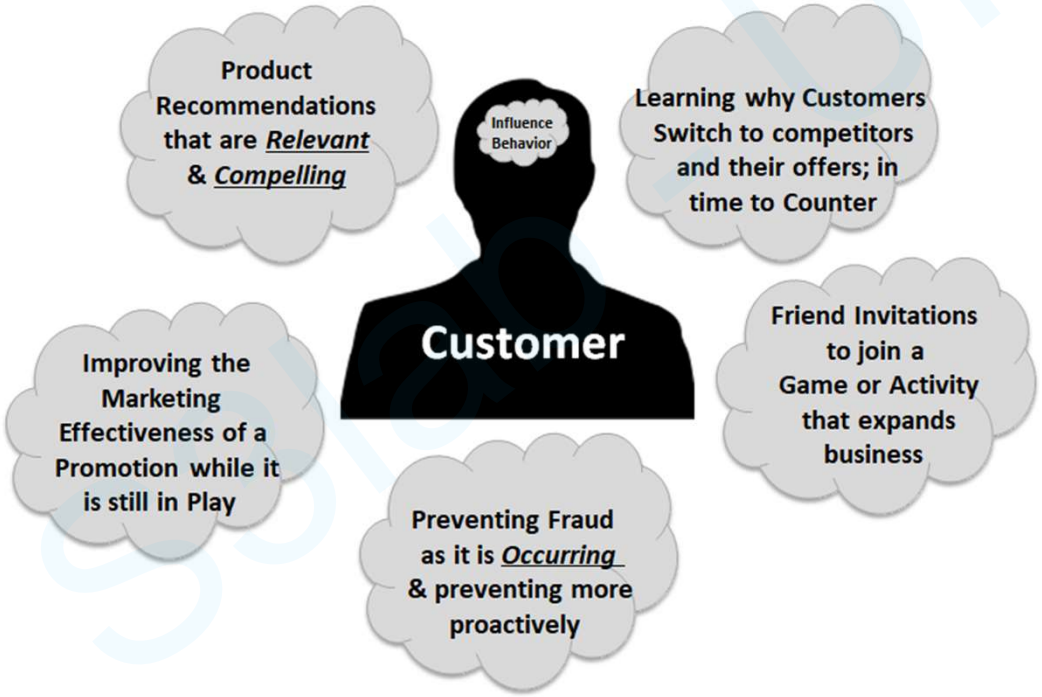
Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data. But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

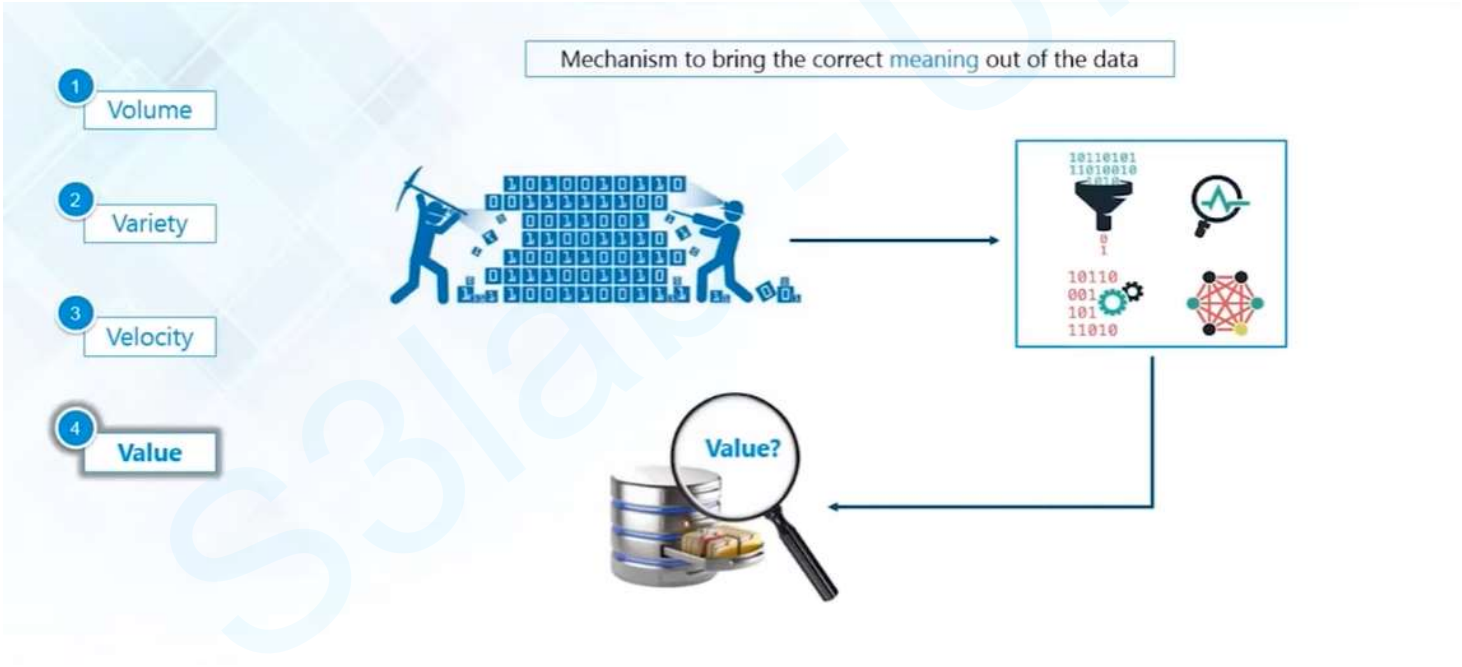
Big Data: 3V's

Velocity (Speed)



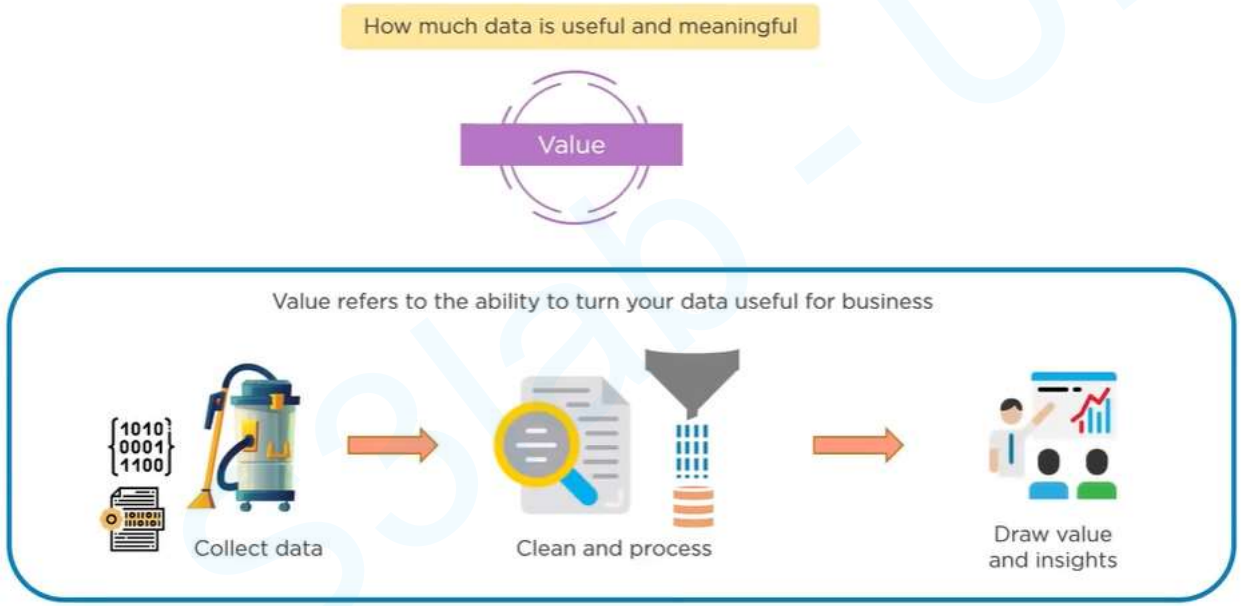
Big Data: 5V's

Value



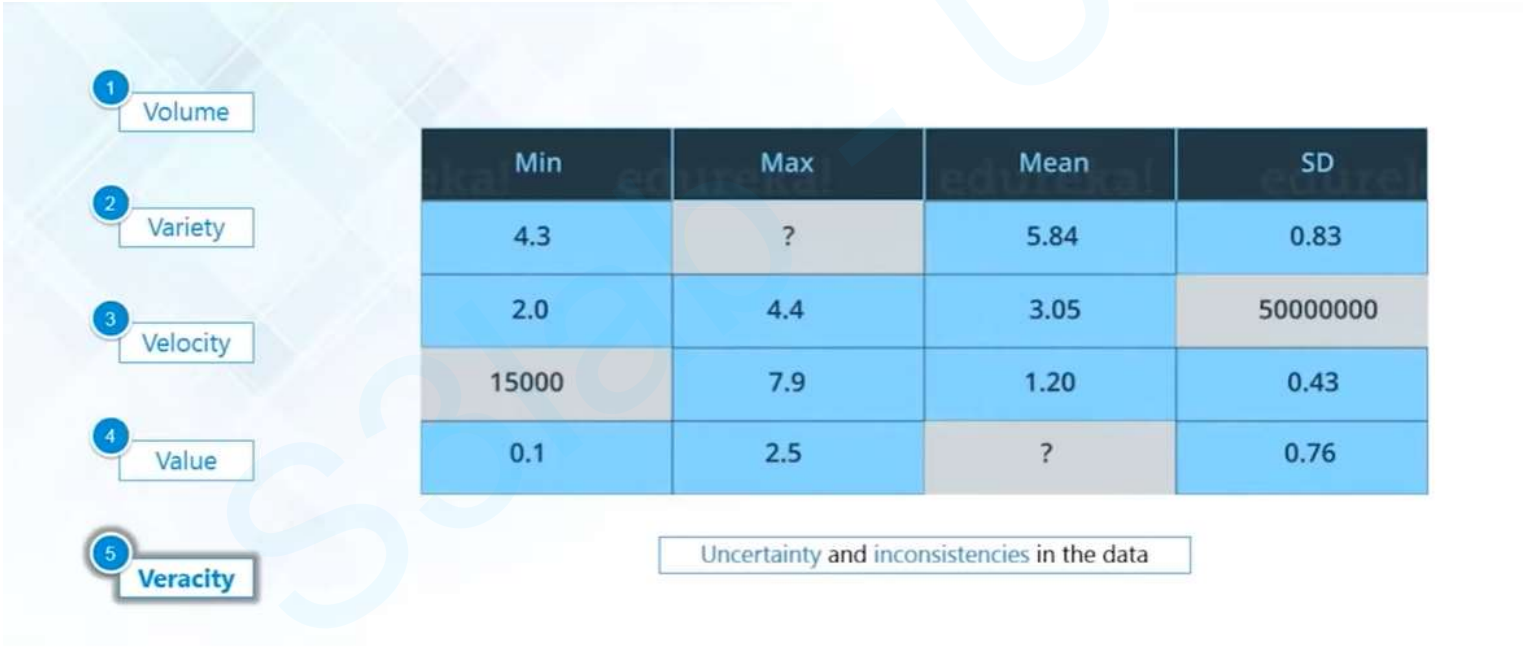
Big Data: 5V's

Value



Big Data: 5V's

Veracity

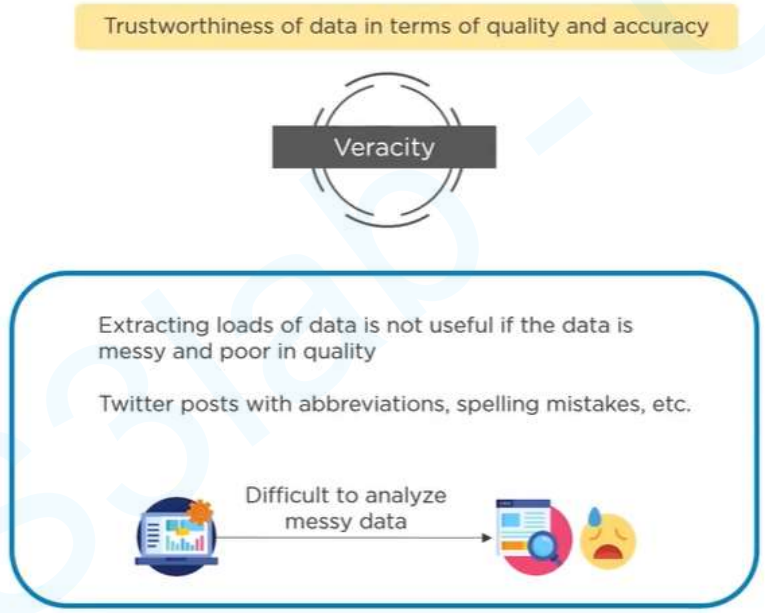


Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

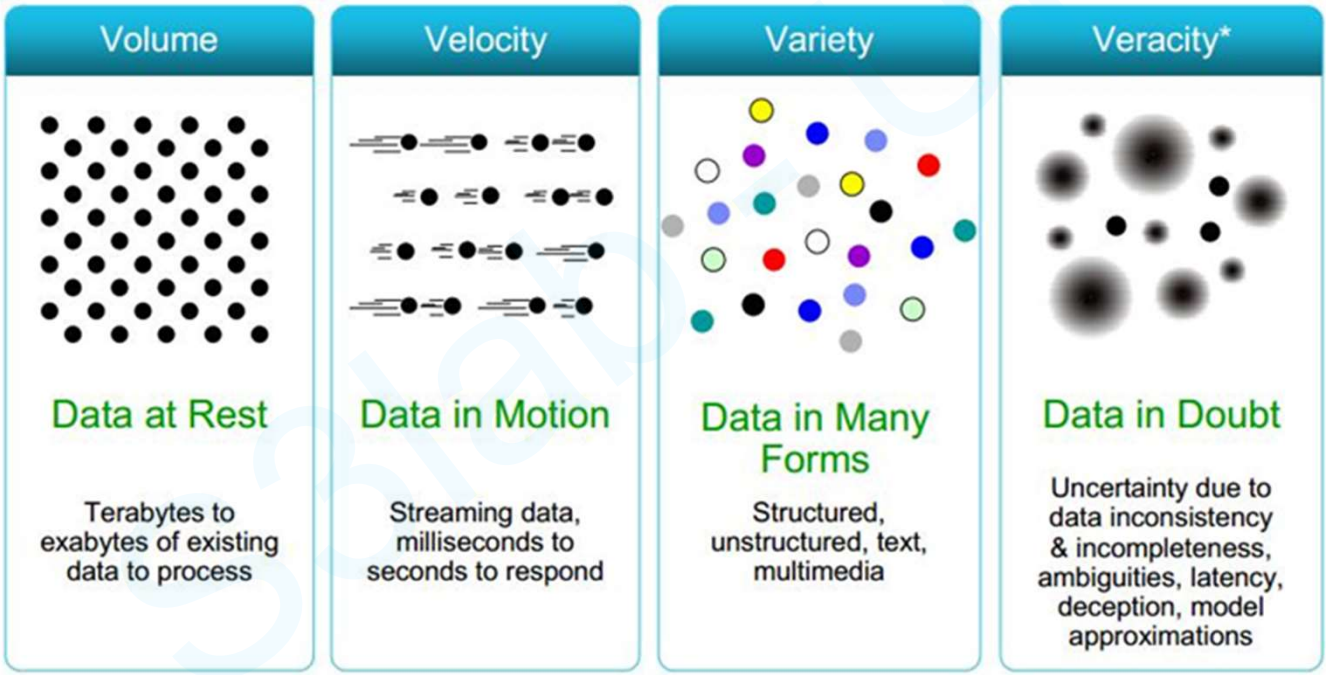
Uncertainty and inconsistencies in the data

Big Data: 5V's

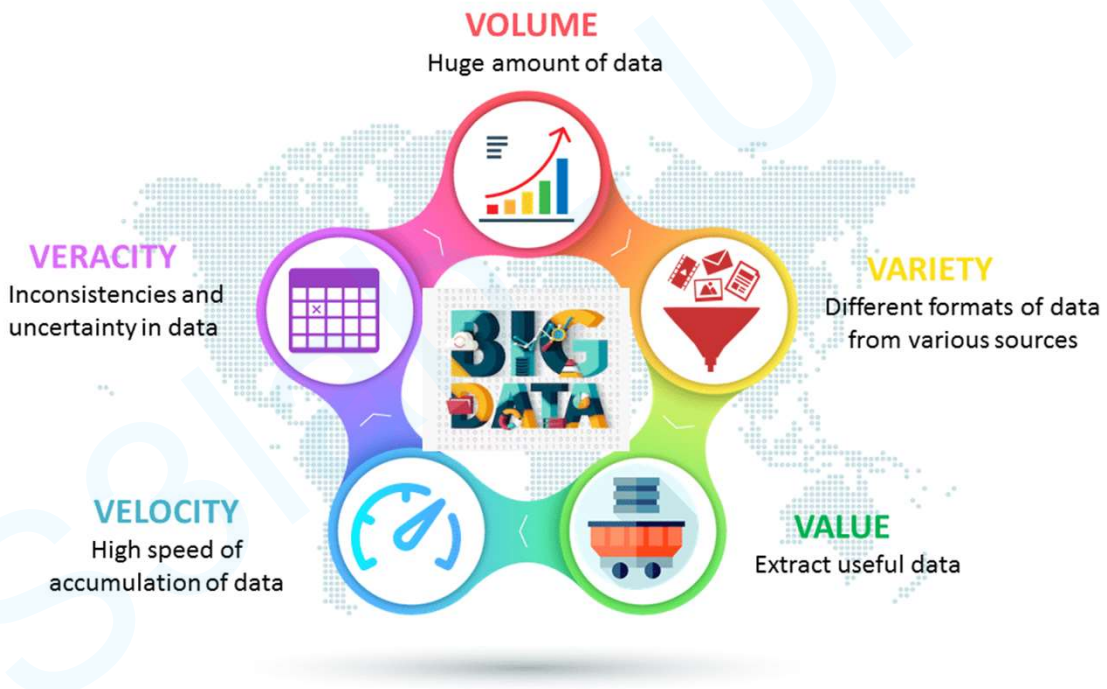
Veracity



Big Data: 4V's


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data: 5V's


VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

Big Data: **N**V's

- The above image depicts the five V's of Big Data but as and when the data keeps evolving so will the V's. I am listing five more V's which have developed gradually over time:
  - Validity: correctness of data
  - Variability: dynamic behaviour
  - Volatility: tendency to change in time
  - Vulnerability: vulnerable to breach or attacks
  - Visualization: visualizing meaningful usage of data

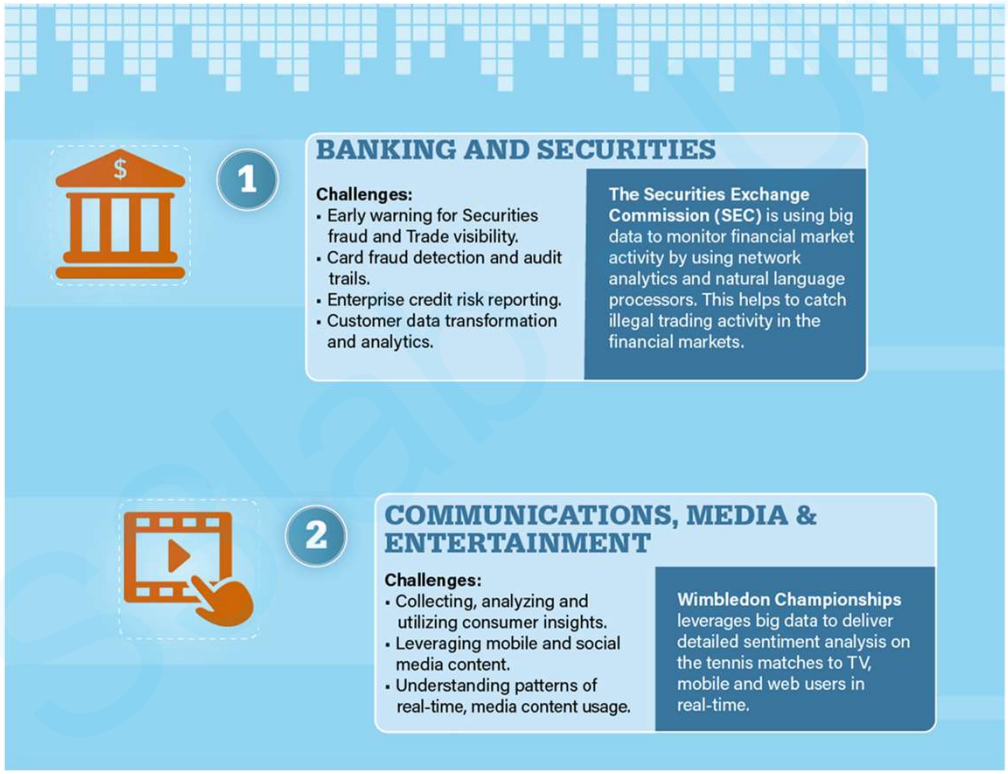
Silab - IUT

Big Data: Applications

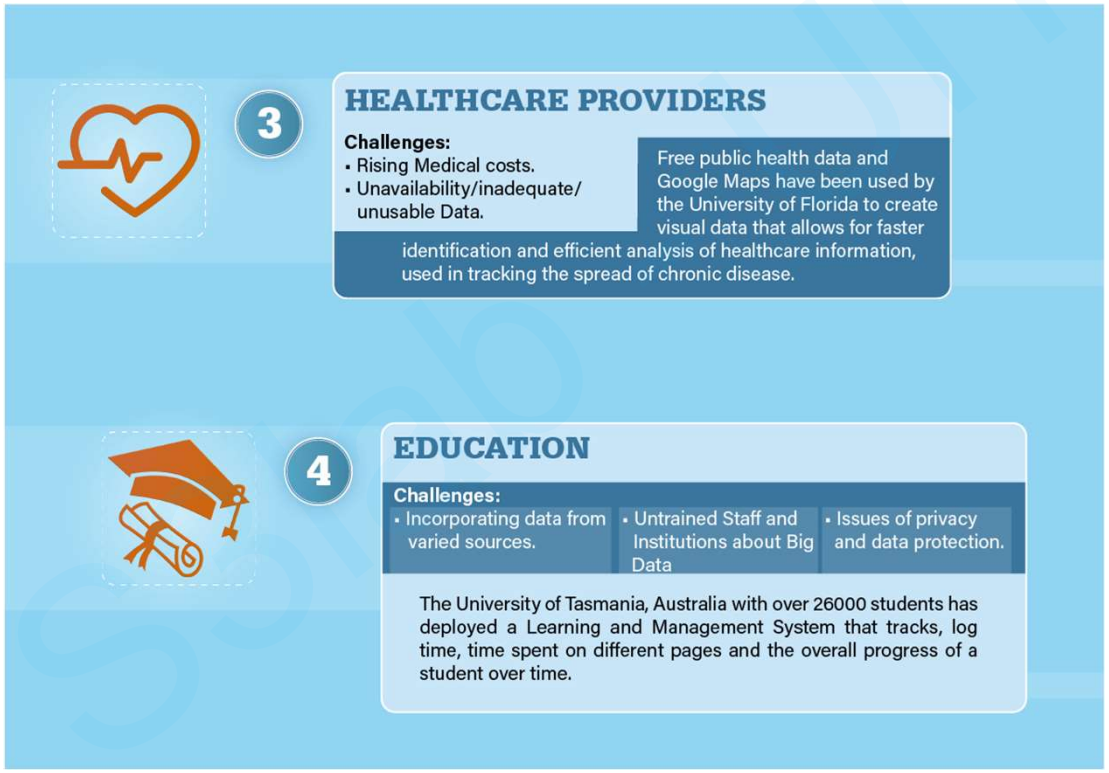
As you see, big data revenue has grown rapidly through the years and will continue to

Big Data: Applications

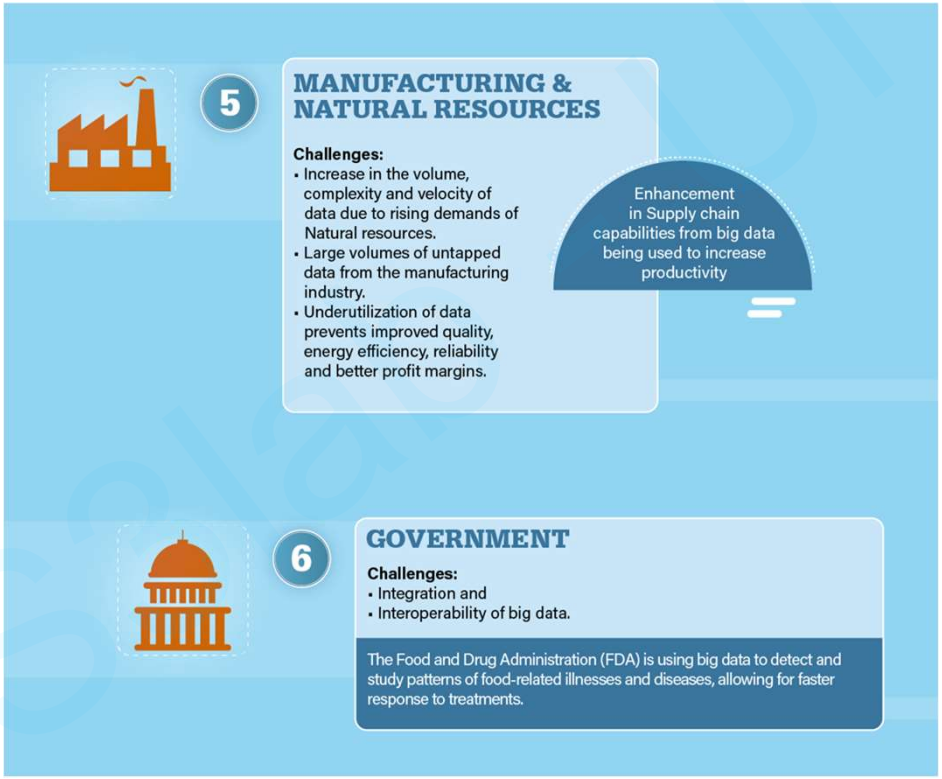
Big Data: Applications



Big Data: Applications

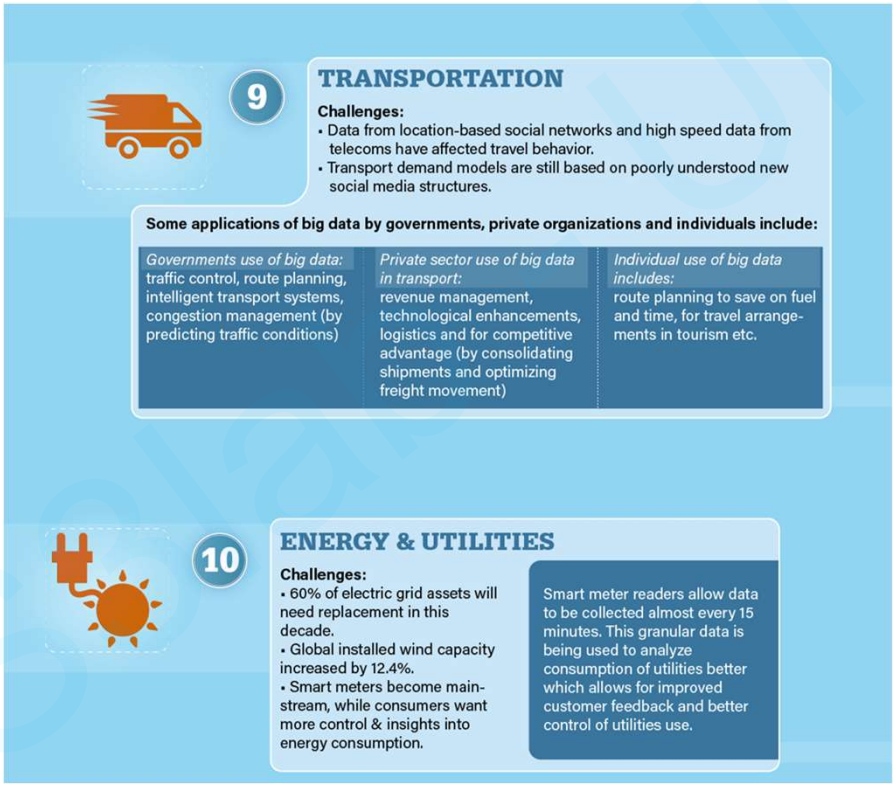


Big Data: Applications



Big Data: Applications

Big Data: Applications



Big Data: Applications

Weather forecast

Challenge

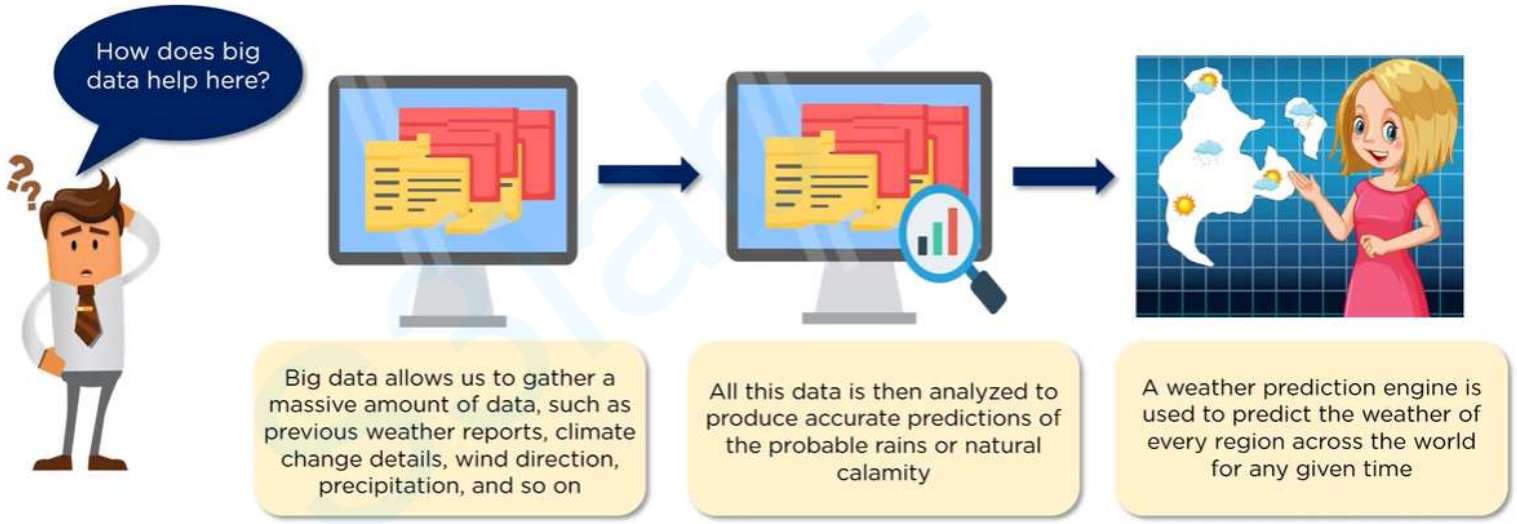
It is very inconvenient when the weather changes suddenly. Imagine having storms, hurricanes, floods without a warning.

Solution

The solution is to design a system which predicts the weather conditions accurately. Big data is used in designing such a system

Big Data: Applications

Weather forecast



Big Data: Applications

Weather forecast

Predicting a landslide is very difficult with just the basic warning signs

The University of Melbourne developed an advanced tool which predicts the boundary where a landslide is likely to occur two weeks before

The predictions are made two weeks before the landslide occurs. The tool works on big data and applied mathematics

Big Data: Applications

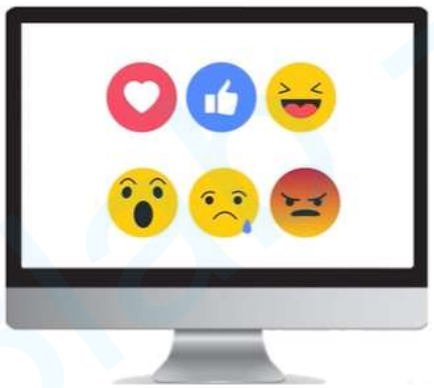
Media and entertainment

Publishers are now able to gather more data about their visitors, which in turn enables them to serve relevant advertising

- Used for targeted advertisement
- Customer sentiment analysis
- Recommendation engines
- An insight into customer churn analysis

Big Data: Applications

Media and entertainment



The media and entertainment companies know how their customers feel by analyzing their emotions in posts, messages, and conversations

- Used for targeted advertisement
- Customer sentiment analysis
- Recommendation engines
- An insight into customer churn analysis

Big Data: Applications

Media and entertainment

- Used for targeted advertisement
- Customer sentiment analysis
- **Recommendation engines**
- An insight into customer churn analysis

Recommendation engines can accurately predict what a customer would like to watch next depending on their past watch and the browse history

Big Data: Applications

Media and entertainment

Used for targeted advertisement

Customer sentiment analysis

Recommendation engines

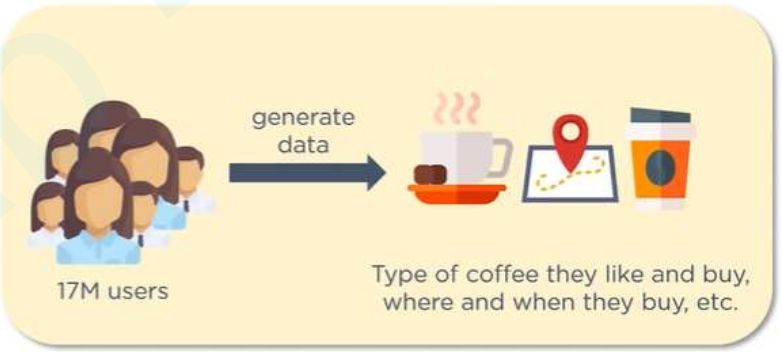
An insight into customer churn analysis

Analyzing the behavioral patterns of previously-churned customers, a marketer can identify the warning signs from current customers and prevent losing customers

Big Data: Applications

Media and entertainment

Did you know that Starbucks app uses big data?



s3lab - UIT

Big Data: Applications

Media and entertainment

Did you know that Starbucks app uses big data?

Based on the customer's ordering history, the app suggests new products, sends personalized offers such as a birthday discount and so on

Big Data: Applications

Health care

With the available big data, medical researches are done very efficiently, and new treatments and medicines are discovered

- Medical research
- Personalized treatment
- Cost reduction
- Health of the population

Big Data: Applications

Health care

Hyper-personal care is provided to each depending on their past medical history

- Medical research
- Personalized treatment
- Cost reduction
- Health of the population

Big Data: Applications

Health care

Readmissions account for a significant reason for increased healthcare costs. If all the available data is analyzed well, accurate insights regarding the patient's health are obtained

- Medical research
- Personalized treatment
- Cost reduction
- Health of the population

S3lab - UIT

Big Data: Applications

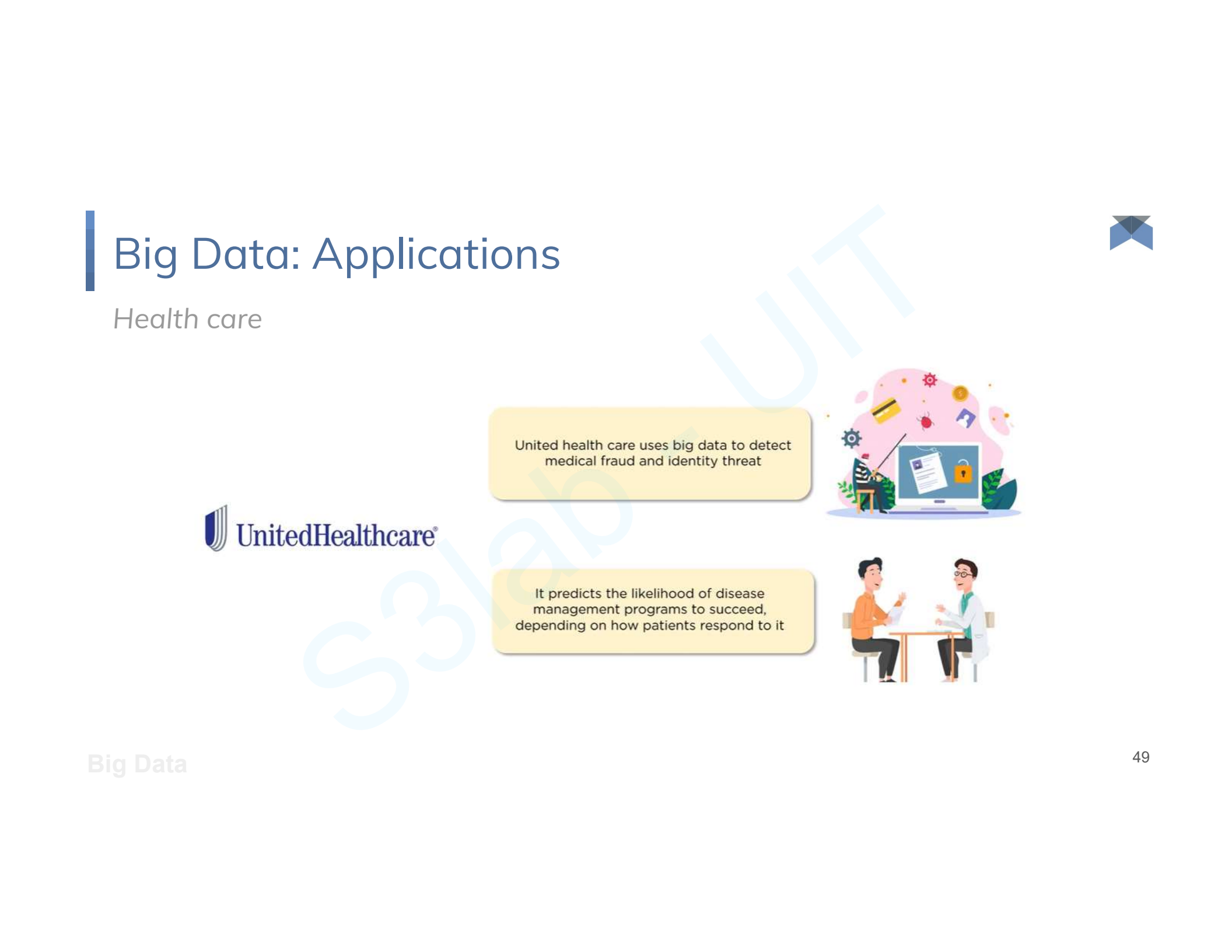
Health care

Analyzing big data helps in identifying disease trends based on geography and demographics

- Medical research
- Personalized treatment
- Cost reduction
- Health of the population

Big Data: Applications

Health care

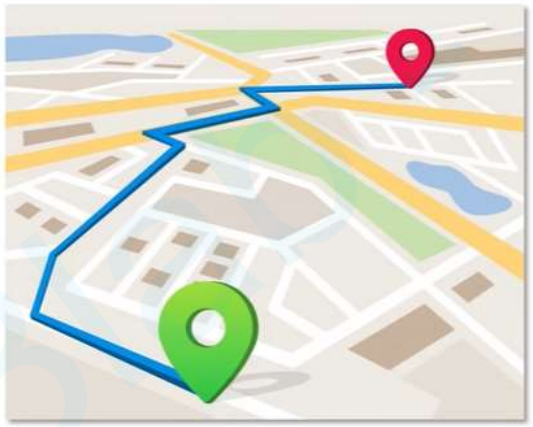


United health care uses big data to detect medical fraud and identity threat

It predicts the likelihood of disease management programs to succeed, depending on how patients respond to it

Big Data: Applications

Logistic



- Flexible routing
- Capacity planning
- Smart warehousing
- Customer satisfaction

The sensors within the vehicles analyze the fastest route to reach the destination. This dynamic routing plan will take into consideration the weather, traffic, and orders

Big Data: Applications

Logistic

Big Data: Applications

Logistic

- Flexible routing
- Capacity planning
- Smart warehousing
- Customer satisfaction

Using big data analytics and tracking sensors, warehouse robotics is improved resulting in efficient resource allocation and reduced cost

Big Data: Applications

Logistic

- Flexible routing
- Capacity planning
- Smart warehousing
- Customer satisfaction

Semantic analysis and text processing helps to analyze customer reactions which will eventually create an instant feedback loop

Big Data: Applications

Logistic

UPS is one of the biggest package shipping company in the world. They use a wide variety of data every moment

They use big data to optimize routes dynamically. The system can automatically change routes in real time. By doing so, UPS can anticipate traffic jams, bad weather, and so on

Big Data: Applications

Travel and tourism

- Revenue management
- Market research
- Personalized offers
- Investment opportunities

©Simplilearn. All rights reserved.

Big Data: Applications

Travel and tourism

Big data is used by companies to analyze information about their competitors which gives an understanding of what other hotels are offering their customers

- Revenue management
- Market research
- Personalized offers
- Investment opportunities

Big Data: Applications

Travel and tourism

Based on a tourists' past travel history and likes, they can receive personalized experiences that are focused on their needs

- Revenue management
- Market research
- **Personalized offers**
- Investment opportunities

S3lab UIT

Big Data: Applications

Travel and tourism

Few countries use Big Data to examine tourism flows and discover investment opportunities in their country. Investment in such areas is profitable

- Revenue management
- Market research
- Personalized offers
- **Investment opportunities**

Big Data: Applications

Travel and tourism

Airbnb is one of the best online homestay networks. To help travelers around the world find the best properties, Airbnb analyzes all the big data to produce this result

It lists properties through customers preferences, keywords, and pricing. By doing so, it delivers its customers with the best stay

Big Data: Applications

Government and law enforcement

Predictive policing uses big data to forecast criminal activity before it happens based on the information given to them through big data analysis

- Predictive policing
- Tackling unemployment
- Poverty
- Improving social and economic policies

Big Data: Applications

Government and law enforcement

By analyzing the number of students graduating every year and the number of job openings, governments can have an idea of the unemployment in the country

- Predictive policing
- Tackling unemployment
- Poverty
- Improving social and economic policies

Big Data: Applications

Government and law enforcement

Predictive policing

Tackling unemployment

Poverty

Improving social and economic policies

Governments use big data tools to discover areas which fall under the poverty line and does the needful

S3ind UIT

Big Data: Applications

Government and law enforcement

By analyzing the data collected from public surveys, governments can design efficient services and policies which will benefit the citizens

- Predictive policing
- Tackling unemployment
- Poverty
- Improving social and economic policies

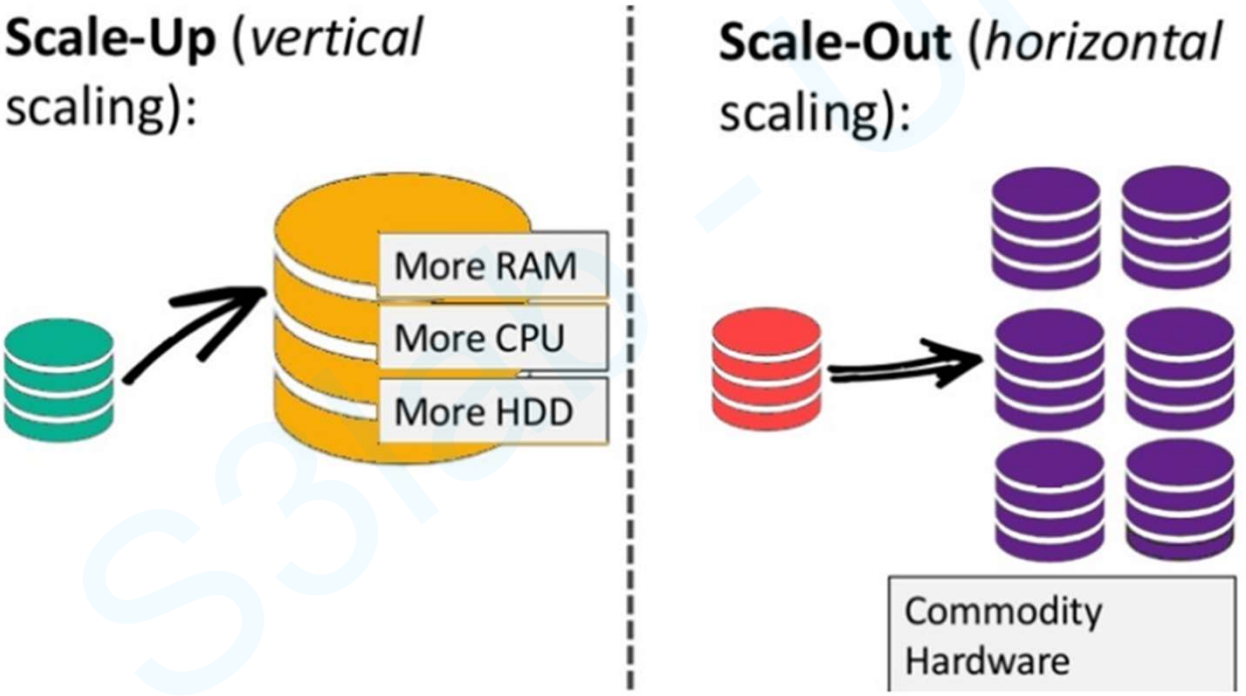
Big Data: Applications

Government and law enforcement

The New York Police Department uses big data analytics to protect its citizens

The department identifies crime trends, threats, and prevent crimes by analyzing data such as certain emails, fingerprints, and records from police investigations, and other public databases

Big Data: Scale



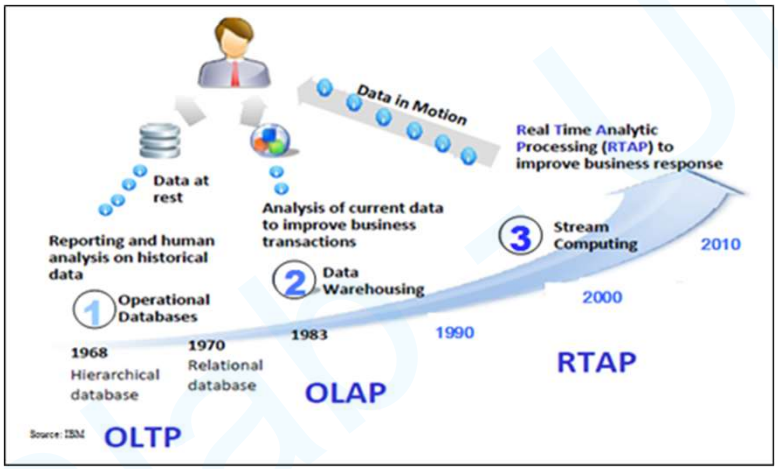
Big Data: Evolution

- The Model of Generating / Consuming Data has changed
  - Old Model: a few companies are generation data, all others are consuming data

  - New Model: All of us are generating data, and all of us are consuming data

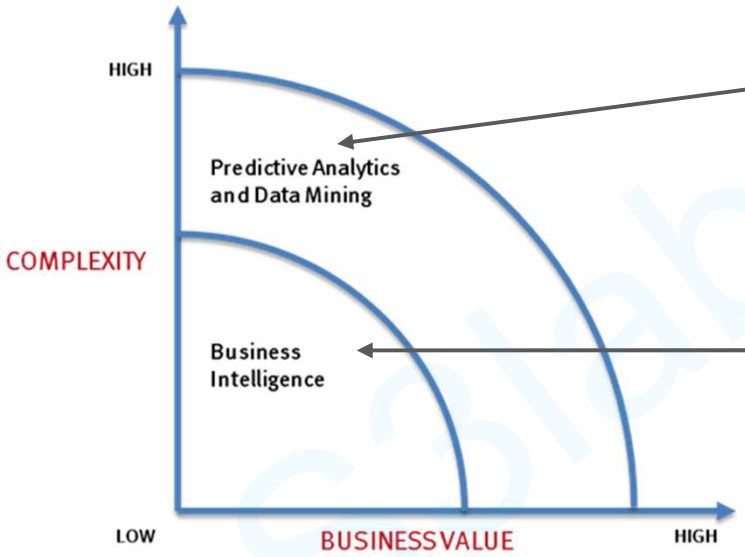
Solab UIT

Big Data: Evolution



- **OLTP**: Online Transaction Processing (DBMSs)
- **OLAP**: Online Analytical Processing (Data Warehousing)
- **RTAP**: Real-time Analytics Processing (Big Data Architecture & Technology)

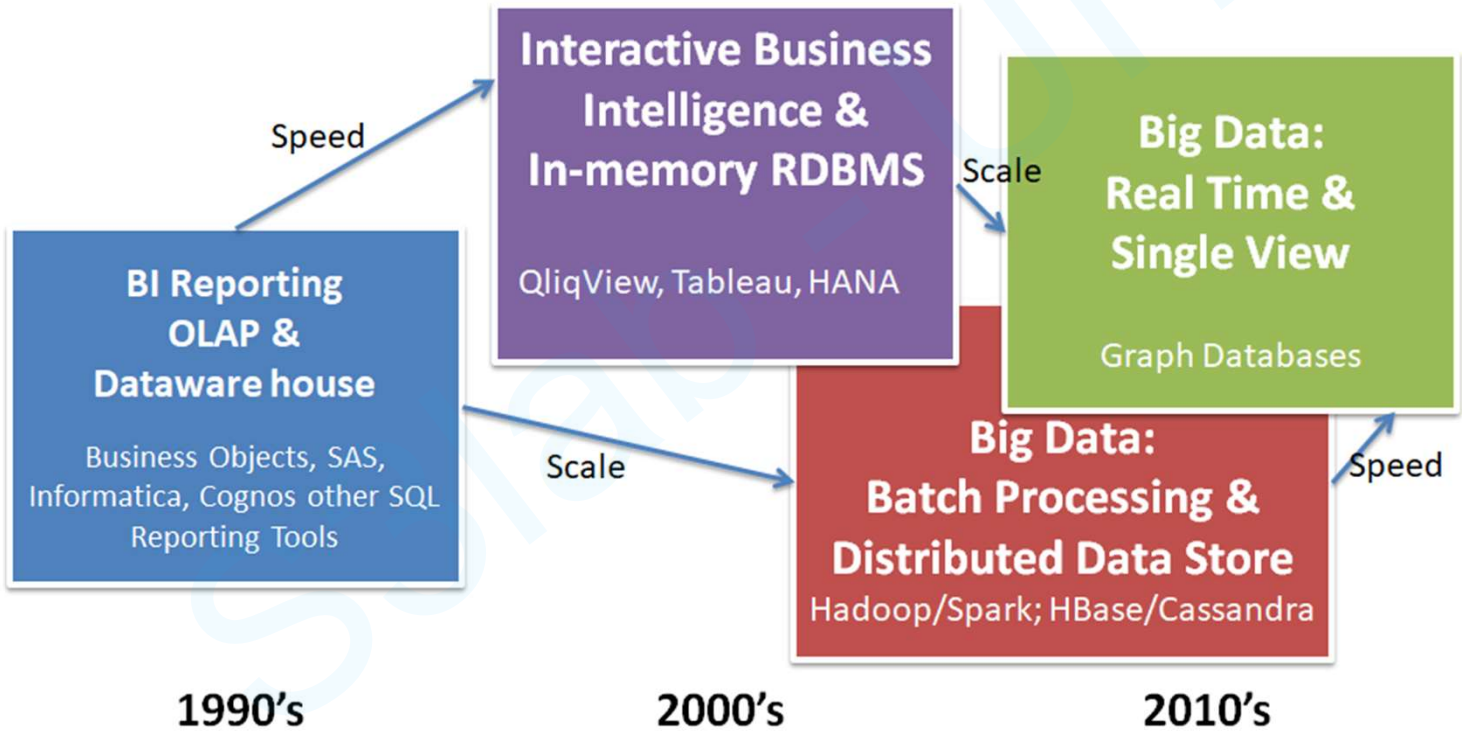
Big Data: Evolution



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

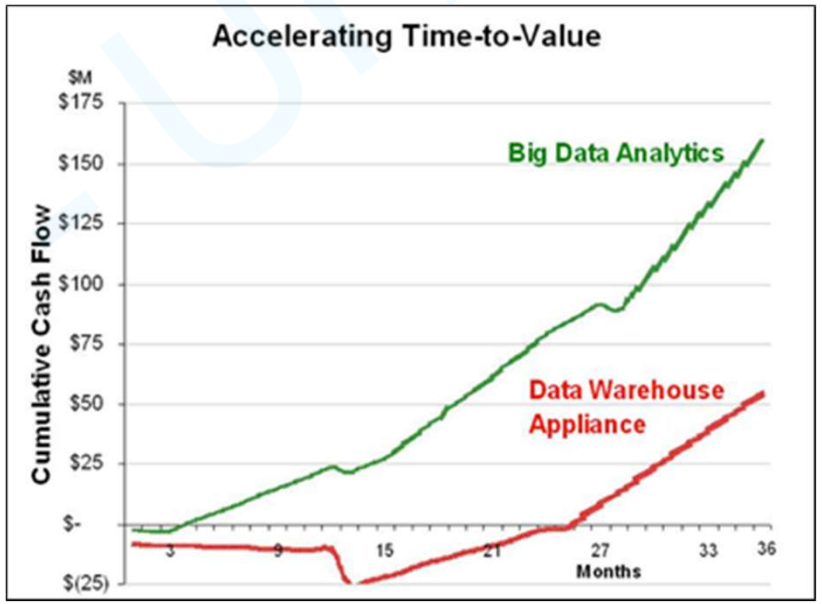
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to midsize datasets

Big Data: Evolution



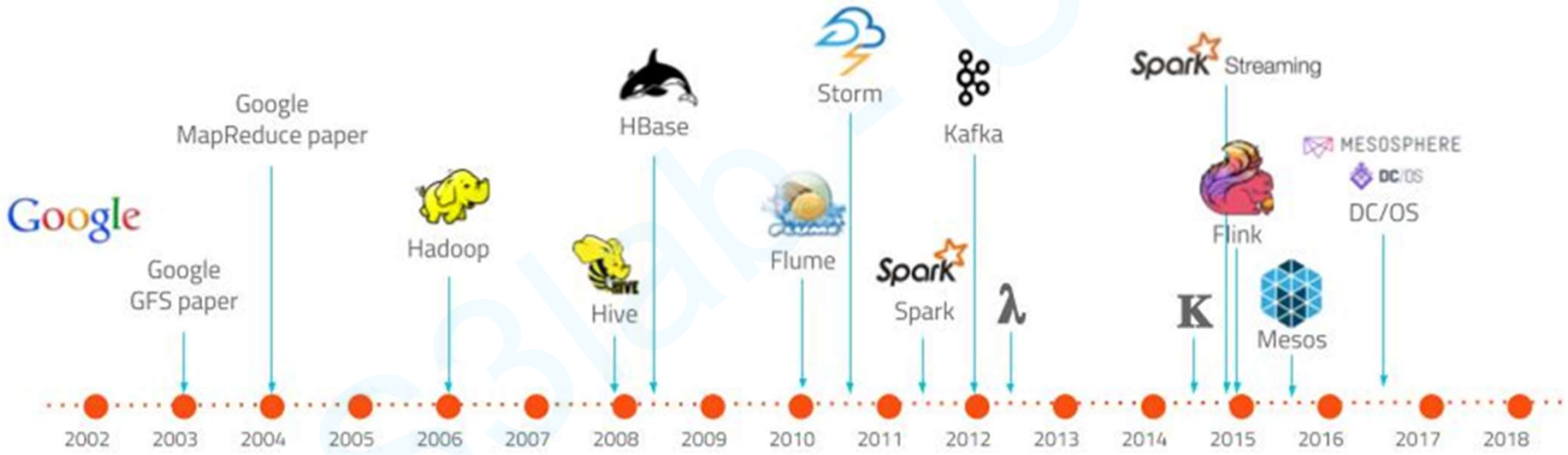
Big Data: Evolution

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



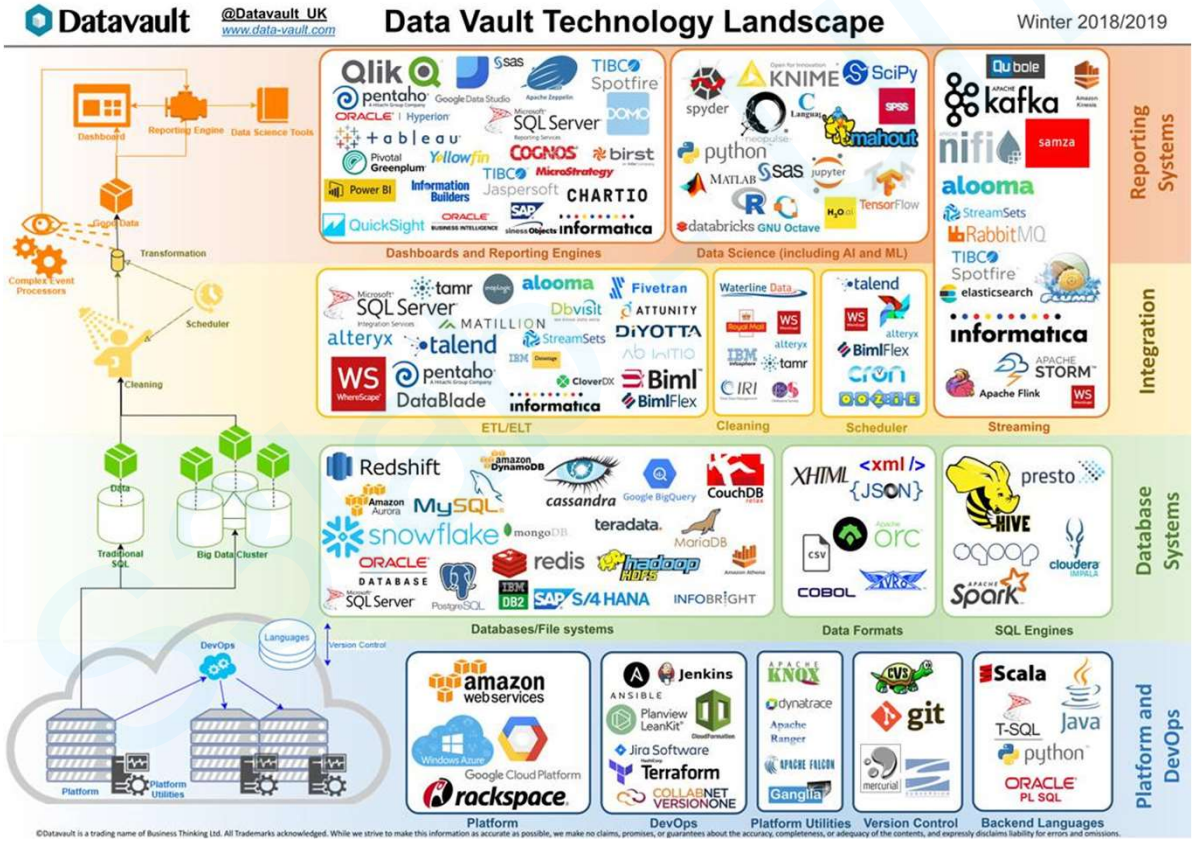
Sajad UT

Big Data: Evolution

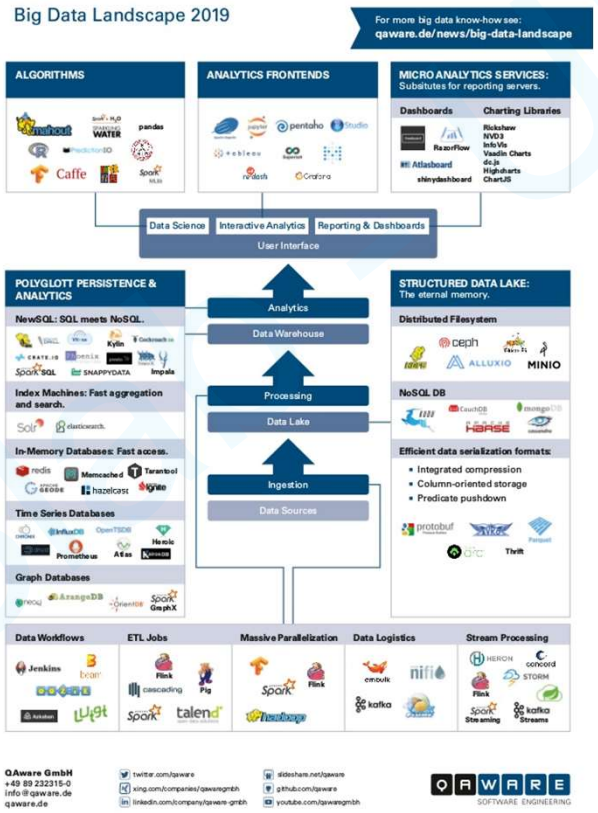


Big Data: Evolution

Big Data: Landscape



Big Data: Landscape



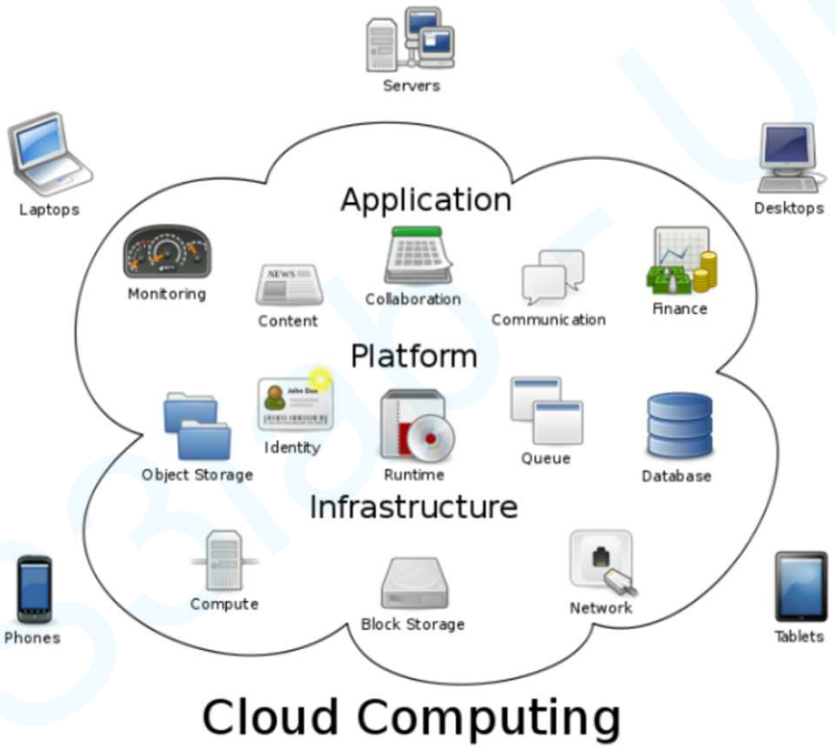
Big Data: Landscape (Open sources)

Big Data: Cloud Computing

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth network & multi-core processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …

Sajib - UIT

Big Data: Cloud Computing



Big Data: Cloud Computing

Benefits

- Cost & Management
  - Economies of scale, “outsourced” resource management
- Reduced time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability: Hardware not owned

Sslab - UIT

Big Data: Cloud Computing

Benefits

Big Data: Cloud Computing

Issues

- Data Security
  - Agree with the cloud service provider ensure data security.
- Performance
  - Service-Level Agreement (SLA) should be clear
- Compliance
  - Depend on the service provider
- Legal Issues
  - Data stored in multiple **locations**
- Cost: pay as per usage, use services in a controlled manner

Sslab - UIT

Big Data: Cloud Computing

Deployment Models

- **Public**: computing infrastructure is hosted at the vendor's premises
- **Private**: Computing architecture is dedicated to customer and is not shared with other organizations.
- **Hybrid**: Host some critical, secure applications in private clouds. The not so critical applications are hosted in the public cloud
  - Cloud bursting: the organization uses its own infrastructure for normal usage, but cloud is used for peak loads.

Sslab - UIT

Big Data: Cloud Computing

Type of Services

- Infrastructure as a service (**Iaas**):
  - Why buy machines when you can rent cycles?
  - Amazon's EC2, Rackspace
- Platform as a service (**PaaS**):
  - Give me nice API and take care of the maintenance, upgrades, ...
  - Google App Engine (GAE), Windows Azure
- Software as a service (**SaaS**):
  - Just run it for me
  - Gmail, Salesforce, dropbox

eSlab - UIT

Big Data: Lambda Architecture

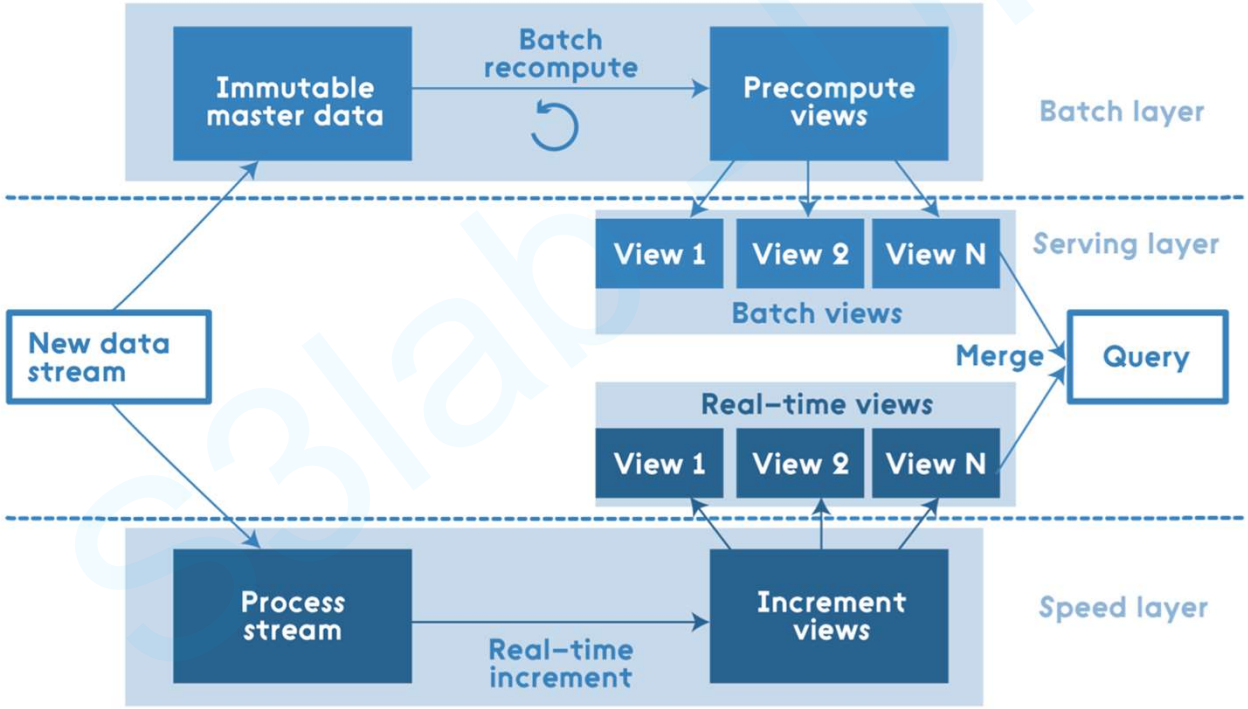
What is Lambda Architecture?

- This is the new big data architecture.
  - Designed to ingest and process
  - Query both fresh and historical (batch) data in a single architecture.
  - Solve the problem of computing arbitrary functions, contains 3 layers:
    - **Batch layer** (Data lake): historical archive, batch query, batch processes for general analytics or ad hoc.
    - **Serving layer**: handles serving up results. Also, combined with both the speed and batch layer.
    - **Speed layer**: queuing, stream, and do the same analytics as batch but in real-time on only the most recent data.

Sglab - UIT

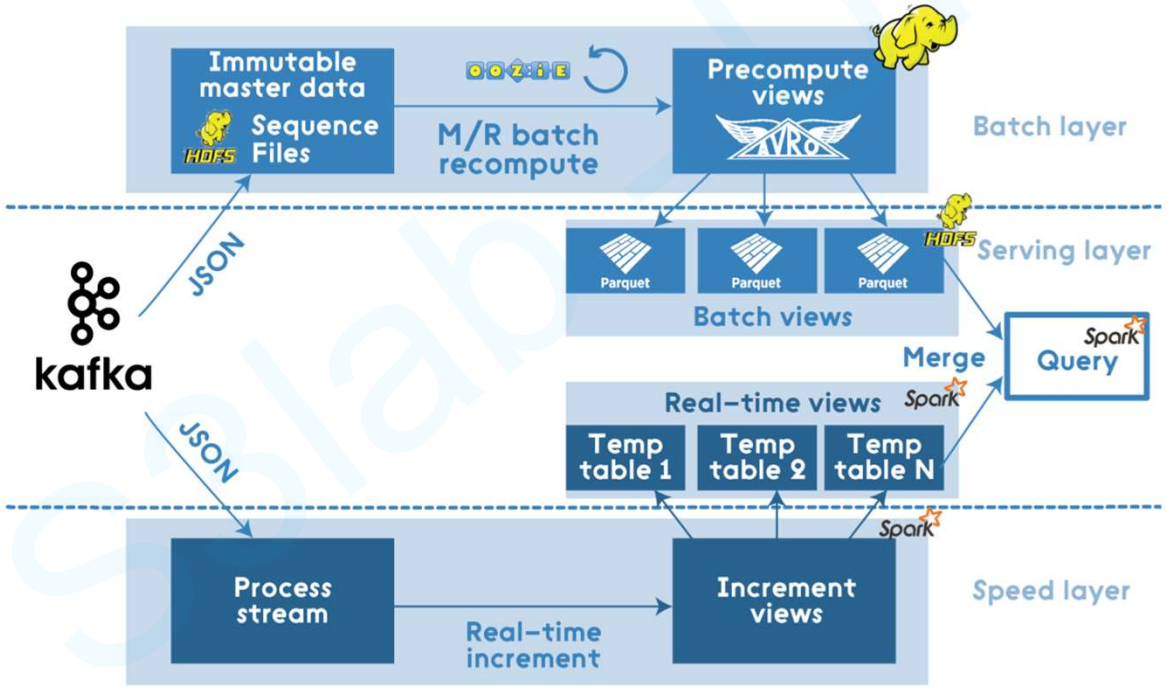
Big Data: Lambda Architecture

What is Lambda Architecture?



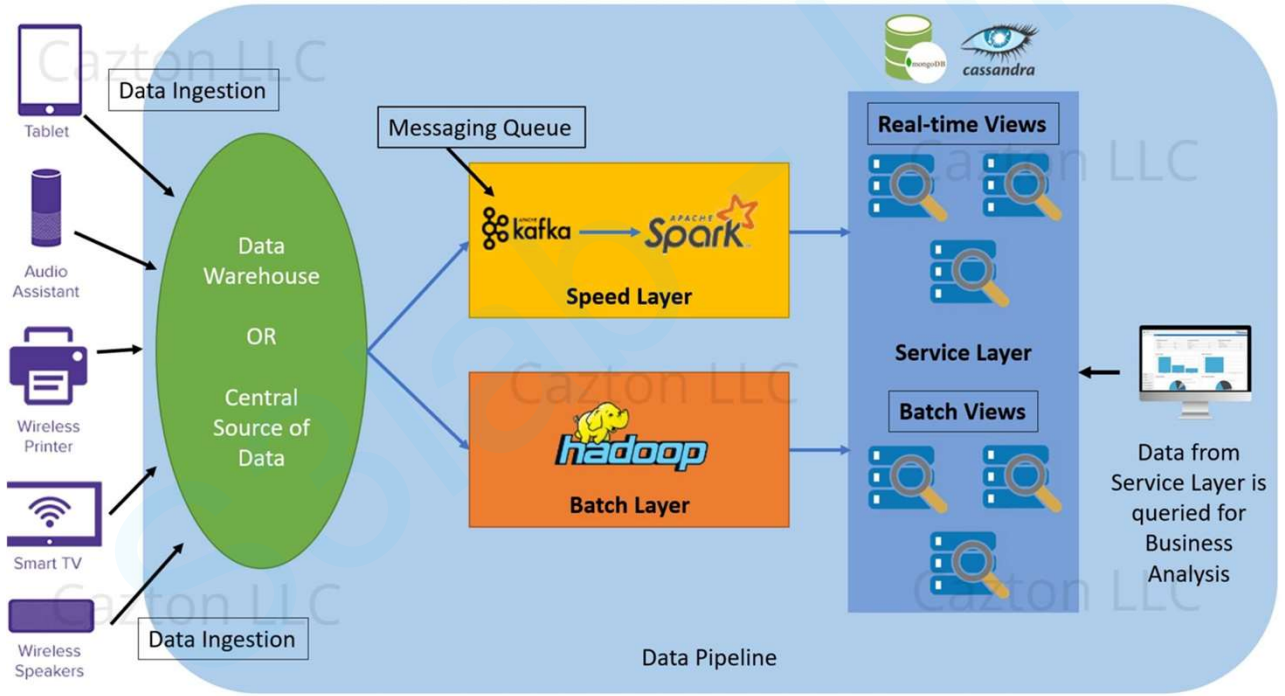
Big Data: Lambda Architecture

real-approach



Big Data: Lambda Architecture

real-approach

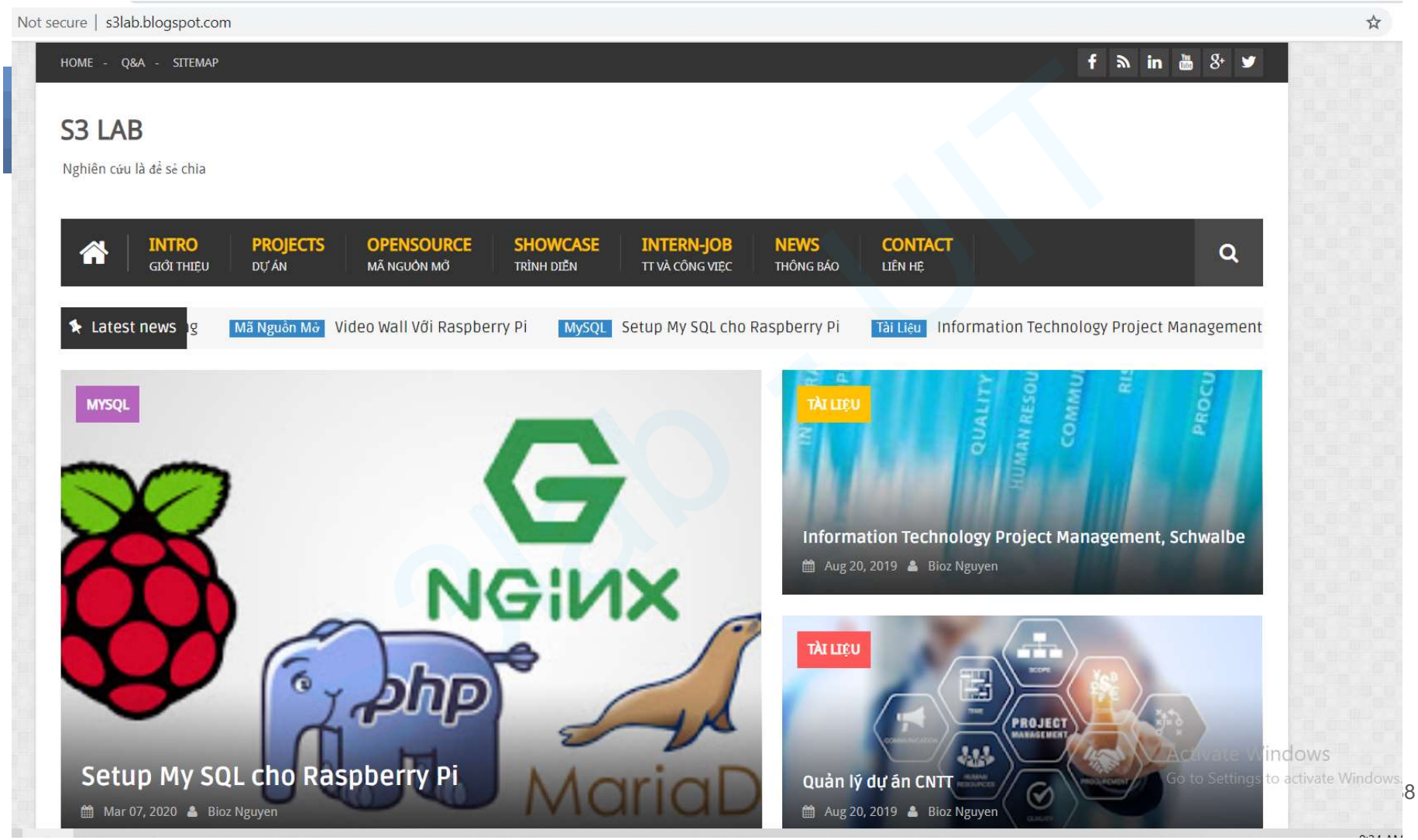


Q & A

Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.

S3lab UIT


Not secure | s3lab.blogspot.com
HOME - Q&A - SITEMAP
S3 LAB
Nghiên cứu là để sẻ chia
INTRO GIỚI THIỆU
PROJECTS DỰ ÁN
OPENSOURCE MÃ NGUỒN MỞ
SHOWCASE TRÌNH DIỄN
INTERN-JOB TT VÀ CÔNG VIỆC
NEWS THÔNG BÁO
CONTACT LIÊN HỆ
Latest news
Mã Nguồn Mở Video Wall Với Raspberry Pi
MySQL Setup My SQL cho Raspberry Pi
Tài Liệu Information Technology Project Management
MYSQL
Setup My SQL cho Raspberry Pi
Mar 07, 2020 Bioz Nguyen
TÀI LIỆU
Information Technology Project Management, Schwalbe
Aug 20, 2019 Bioz Nguyen
TÀI LIỆU
Quản lý dự án CNTT
Aug 20, 2019 Bioz Nguyen
Activate Windows
Go to Settings to activate Windows.