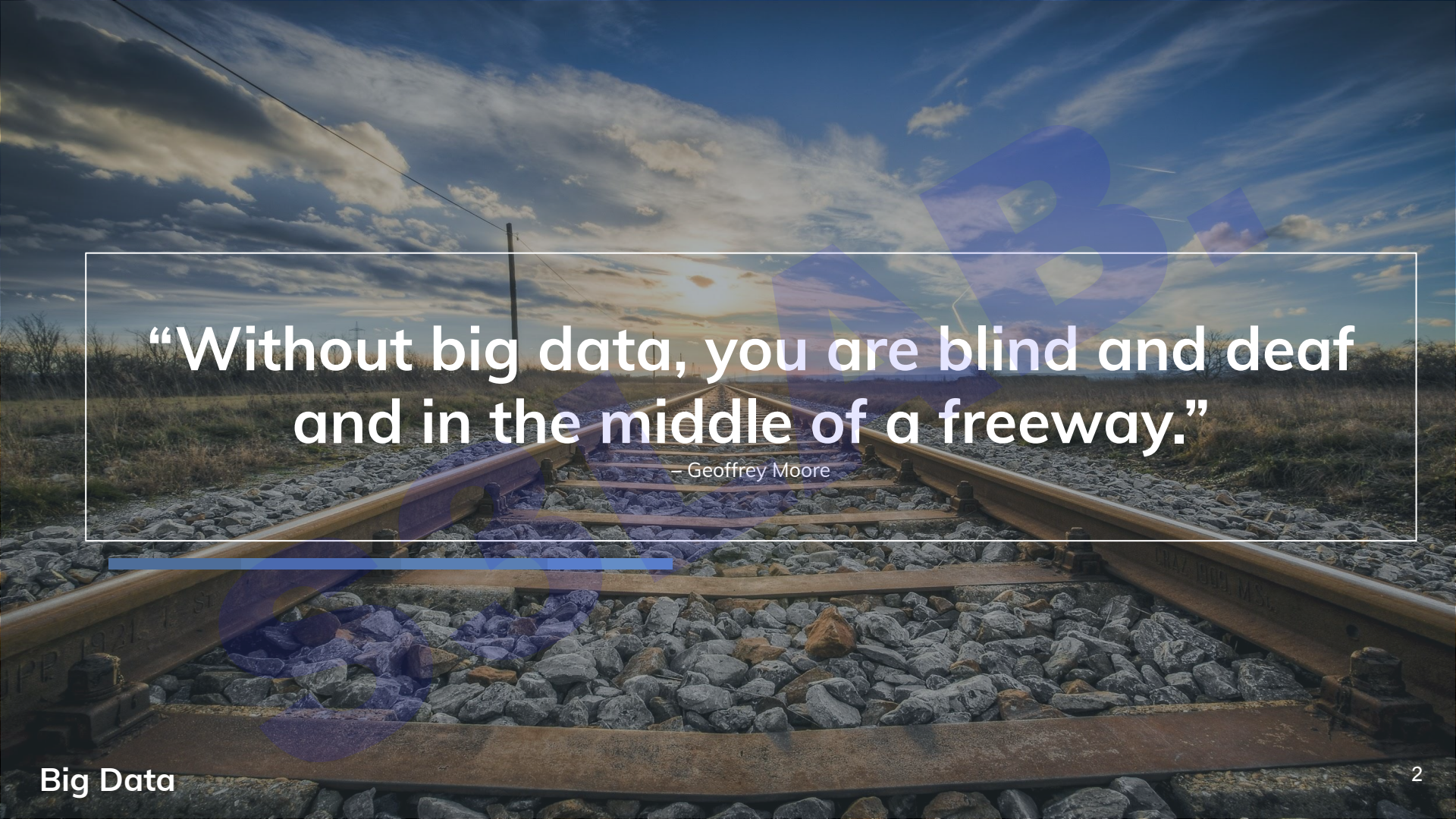
Big Data

(Understanding about Big data)

Instructor: Thanh Binh Nguyen

September 1st, 2019

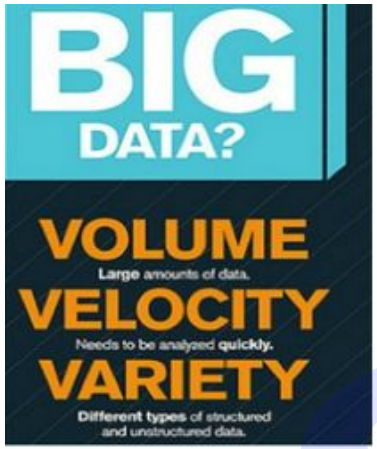
S³Lab
Smart Software System Laboratory

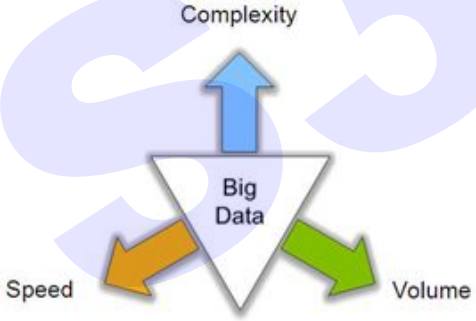

“Without big data, you are blind and deaf and in the middle of a freeway.”
– Geoffrey Moore

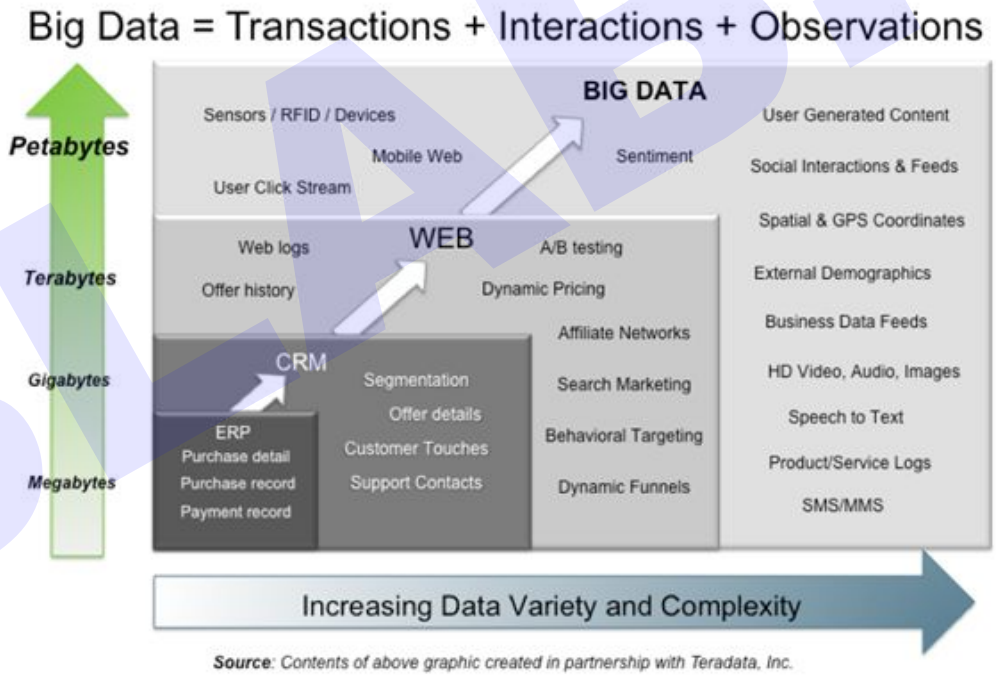
What is Big Data

- 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**.

Big Data: 3V's





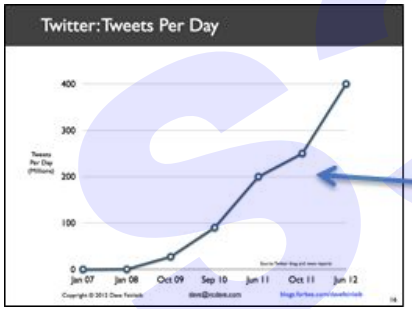


Source: Contents of above graphic created in partnership with Teradata, Inc.

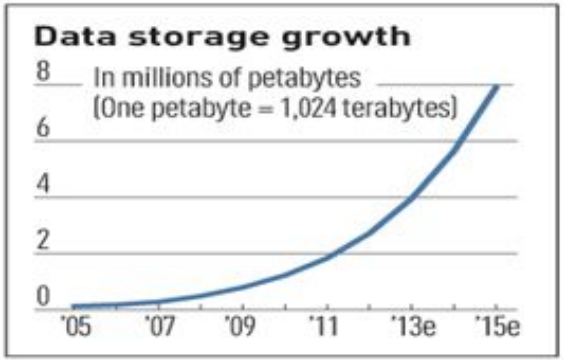
Big Data: 3V's

Volume (scale)

- Data Volume is increasing exponentially:
  - 44x increase from 2009 - 2020.
  - From 0.8 zettabytes to 35zb



Exponential increase in collected/generated data



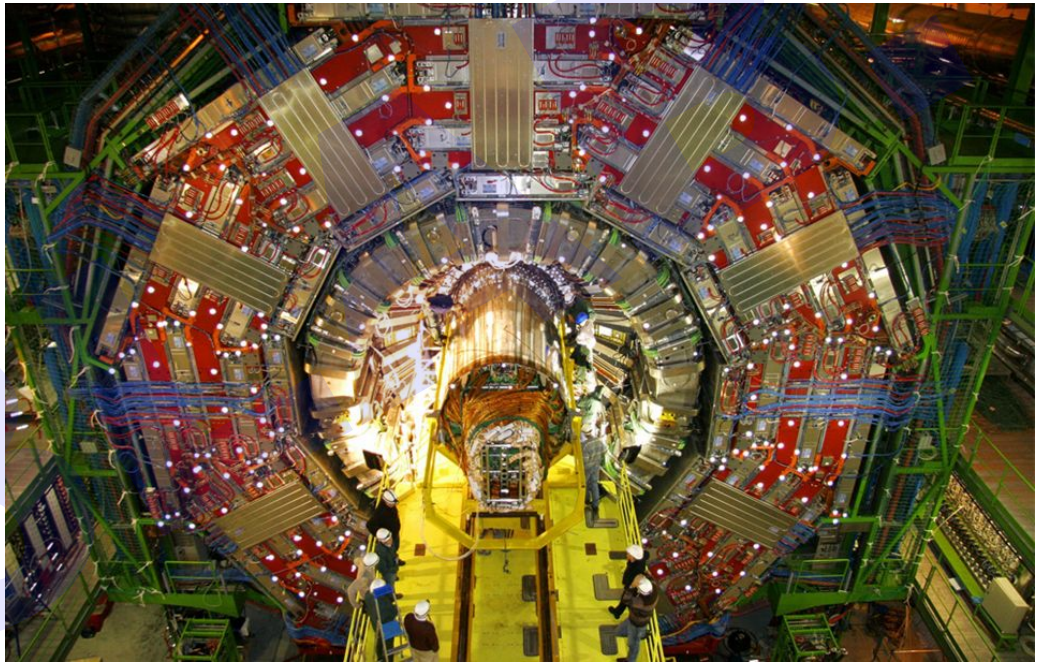
Big Data: 3V's

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

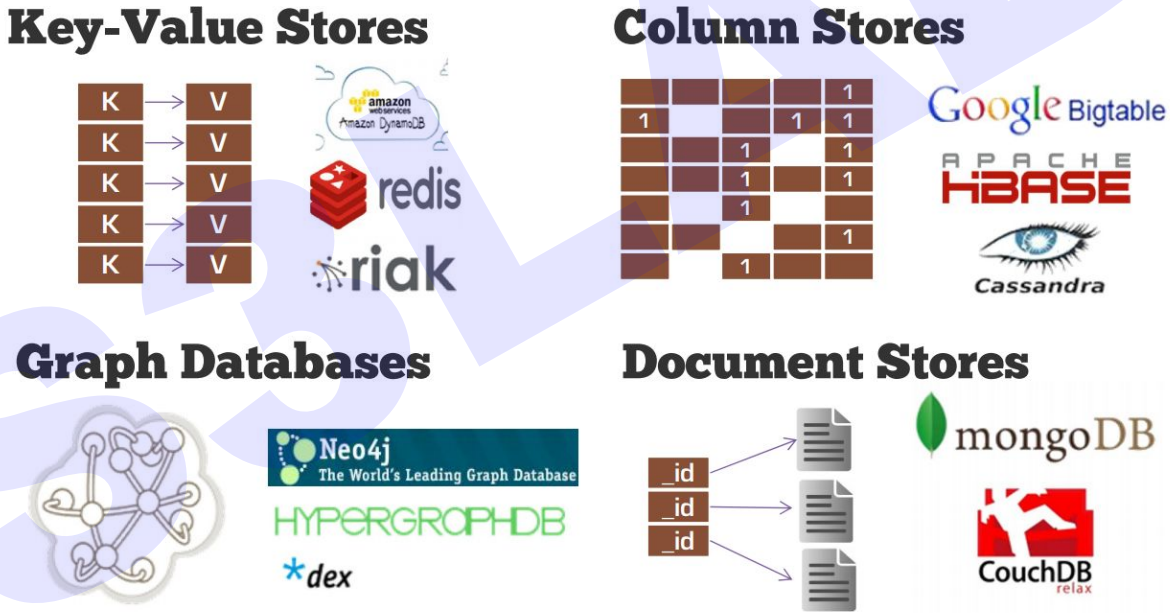
Variety (Complexity)

- Big data 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.

Big Data: 3V's

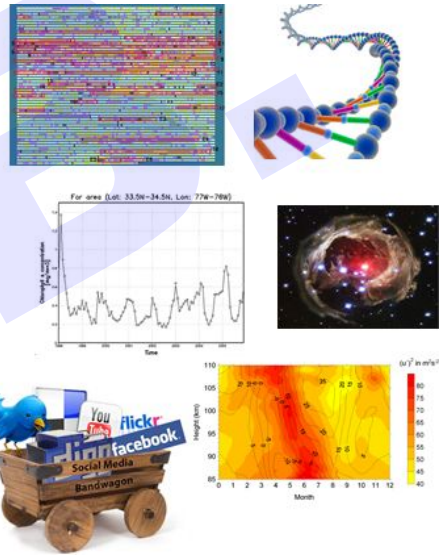
Variety (Complexity)

- Semi-Structured, NoSQL



Big Data: 3V's

Variety (Complexity)

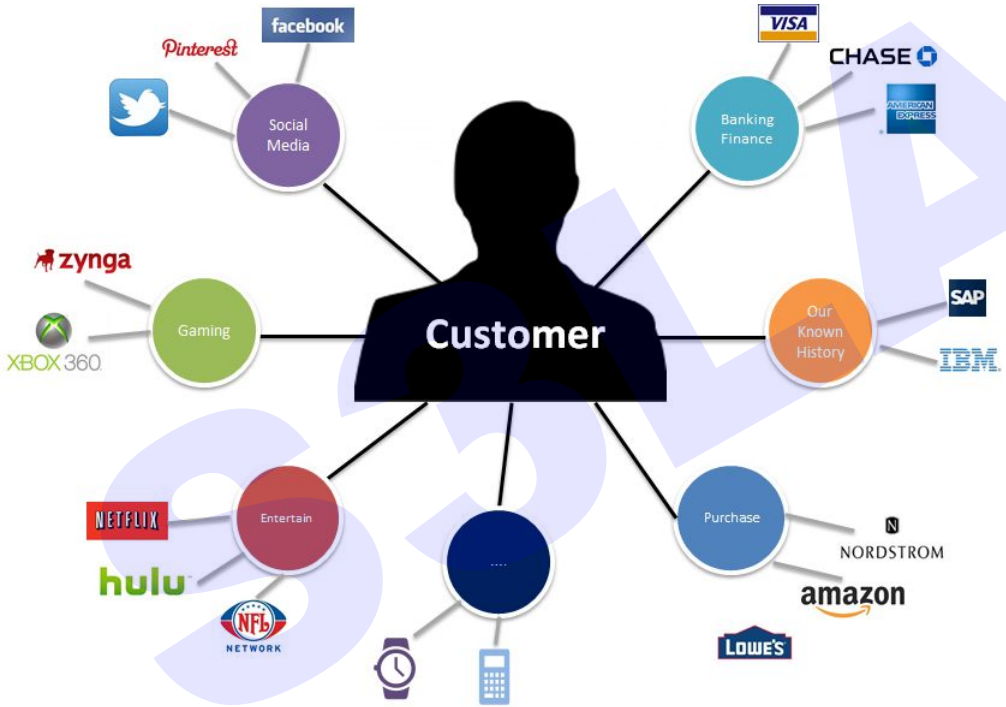


- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web, log)
- Semi-structured Data (XML)
- Graph Data: Social network, Semantic web (RDF - Resource Description Framework)…
- Streaming Data: You can only scan the data once
- A single application can be generating / collecting many types 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)

- 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

Big Data: 3V's

Velocity (Speed)

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

Scientific instruments
(collecting all sorts of data)

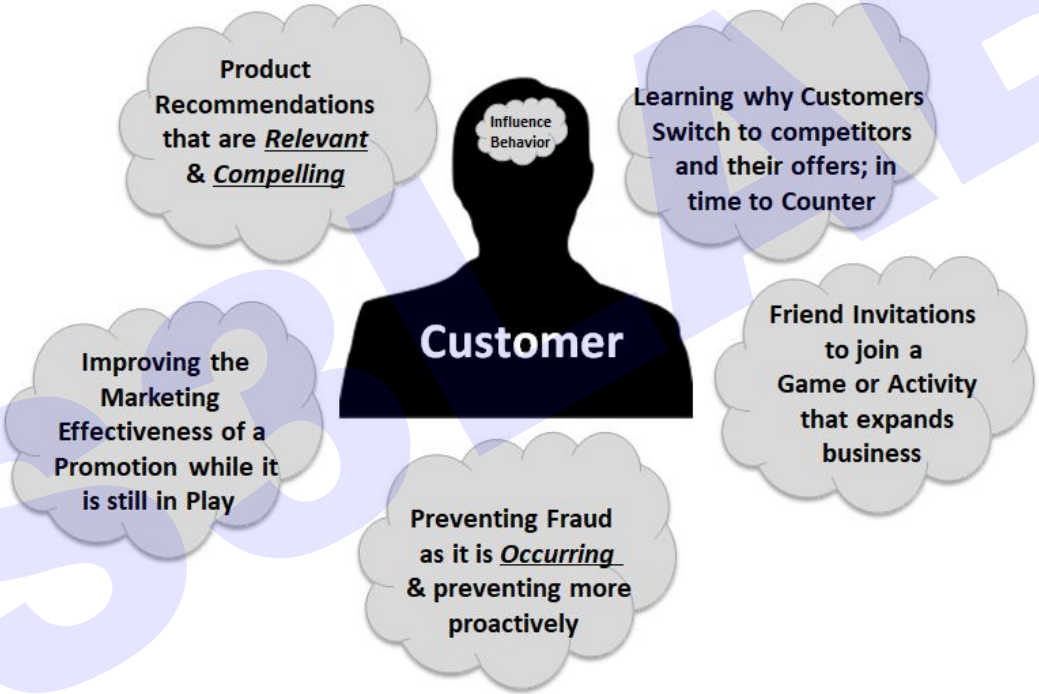
Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

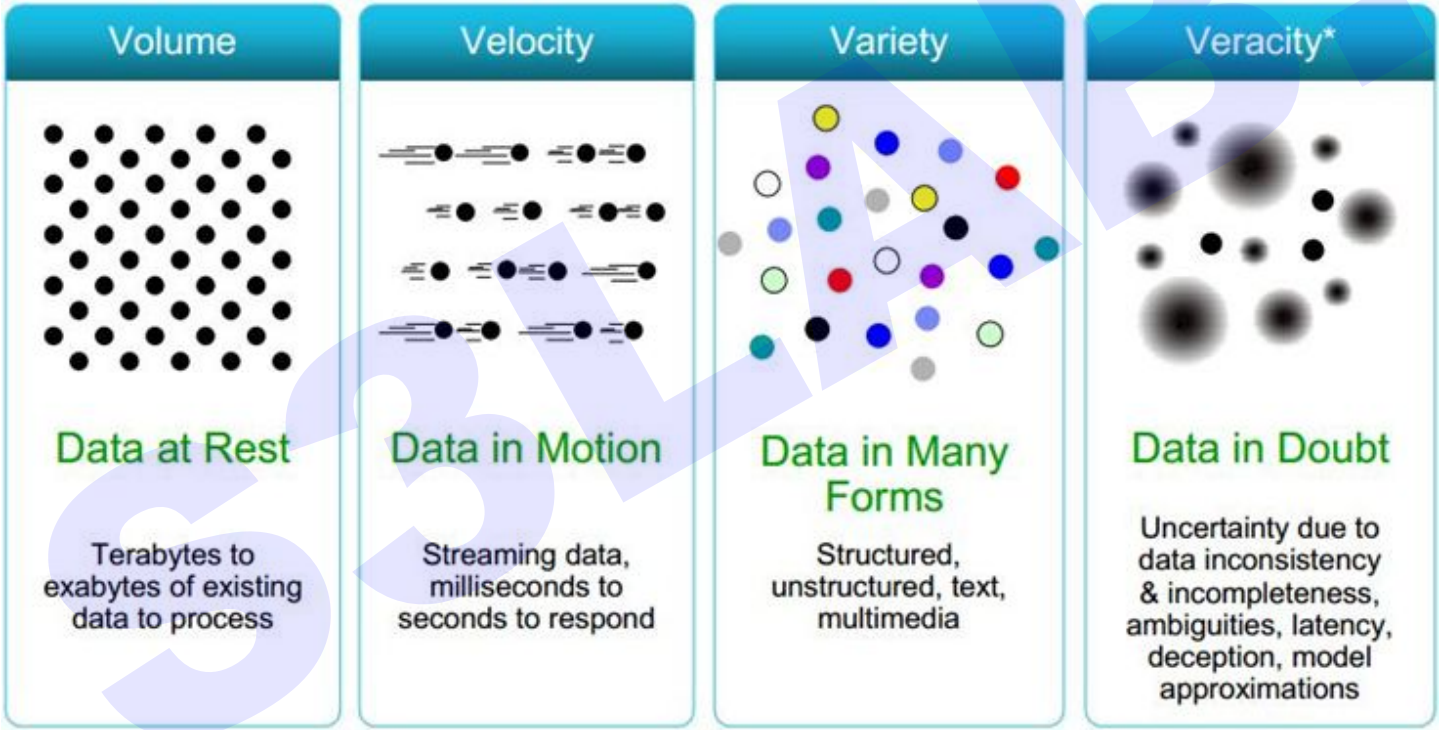
- The progress and innovation is no longer hindered by the ability to **collect data**. But, by the ability to **manage**, **analyze**, **summarize**, **visualize**, and **discover knowledge** from the collected data in a timely manner and in a scalable fashion

Big Data: 3V's

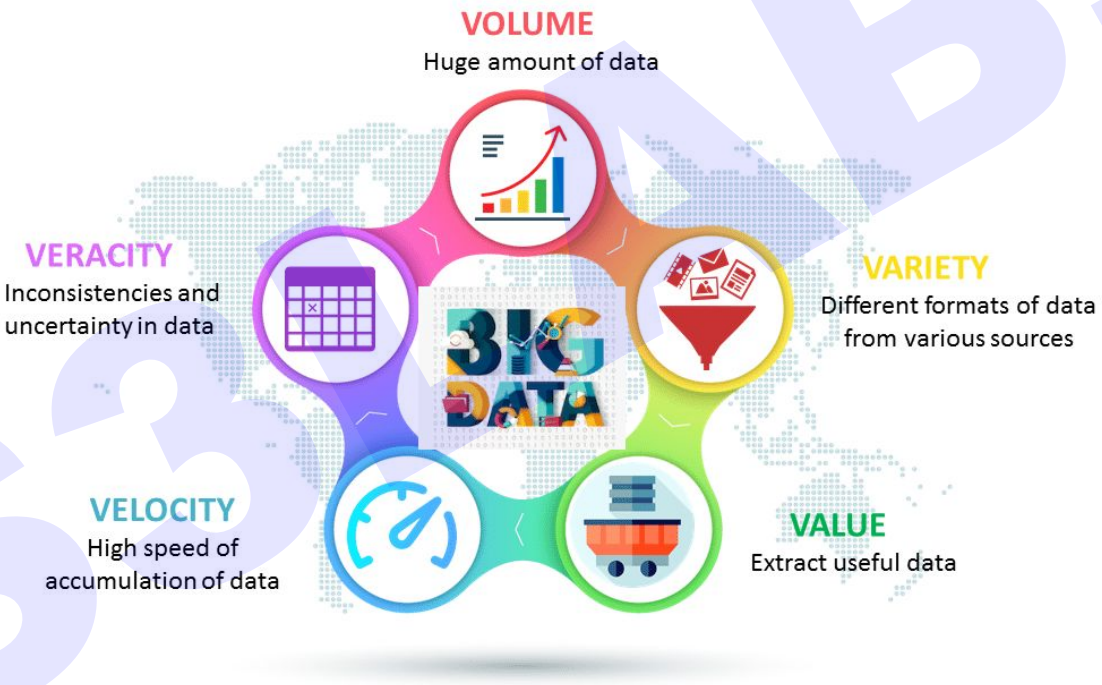
Velocity (Speed)



Big Data: 4V's



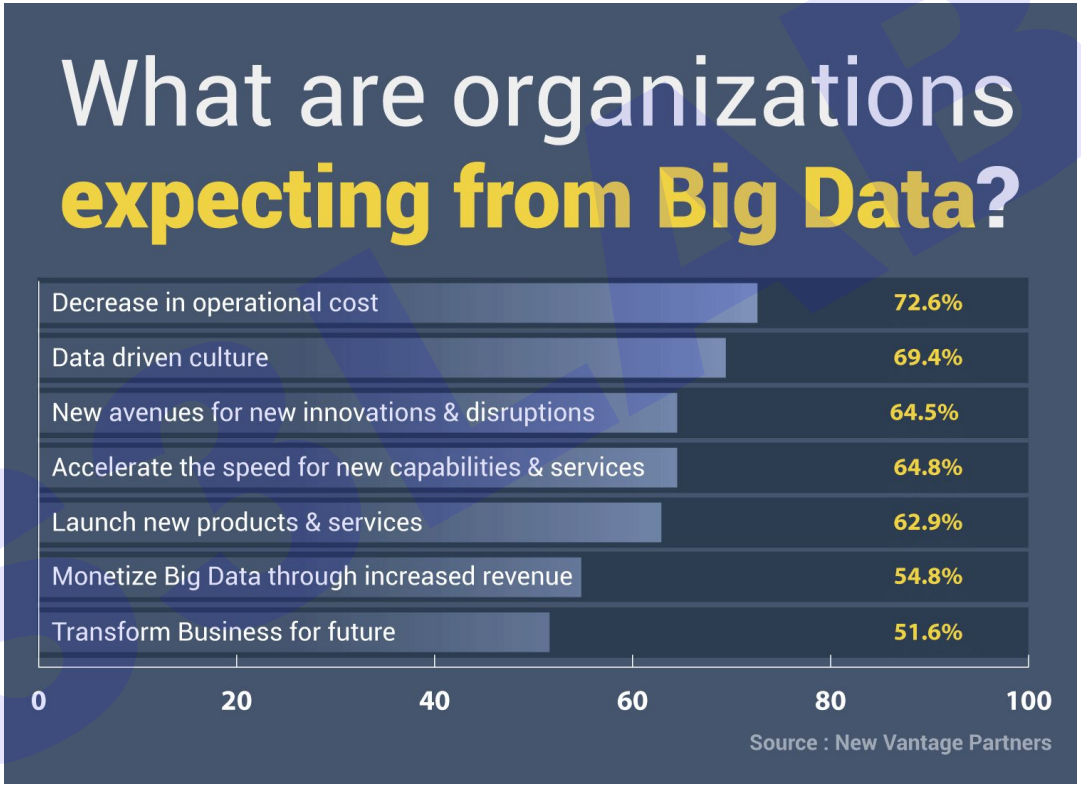
Big Data: 5V's


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

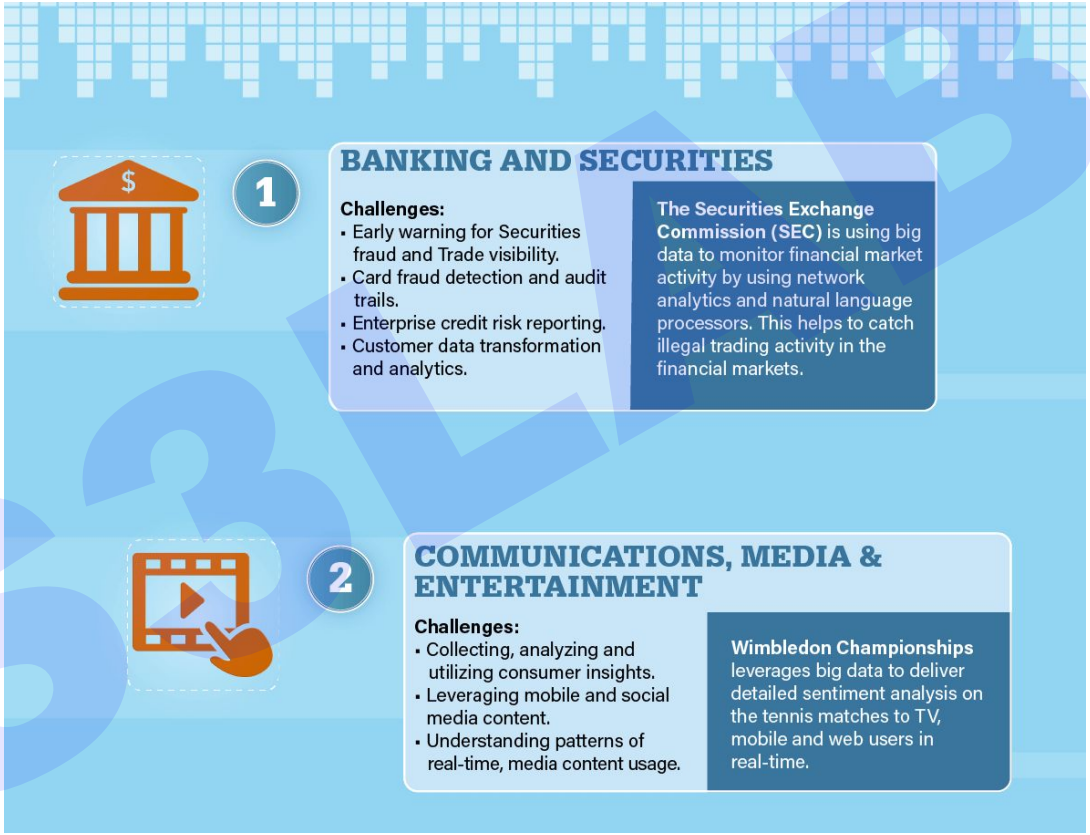
Big Data: **N**V’s

- 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

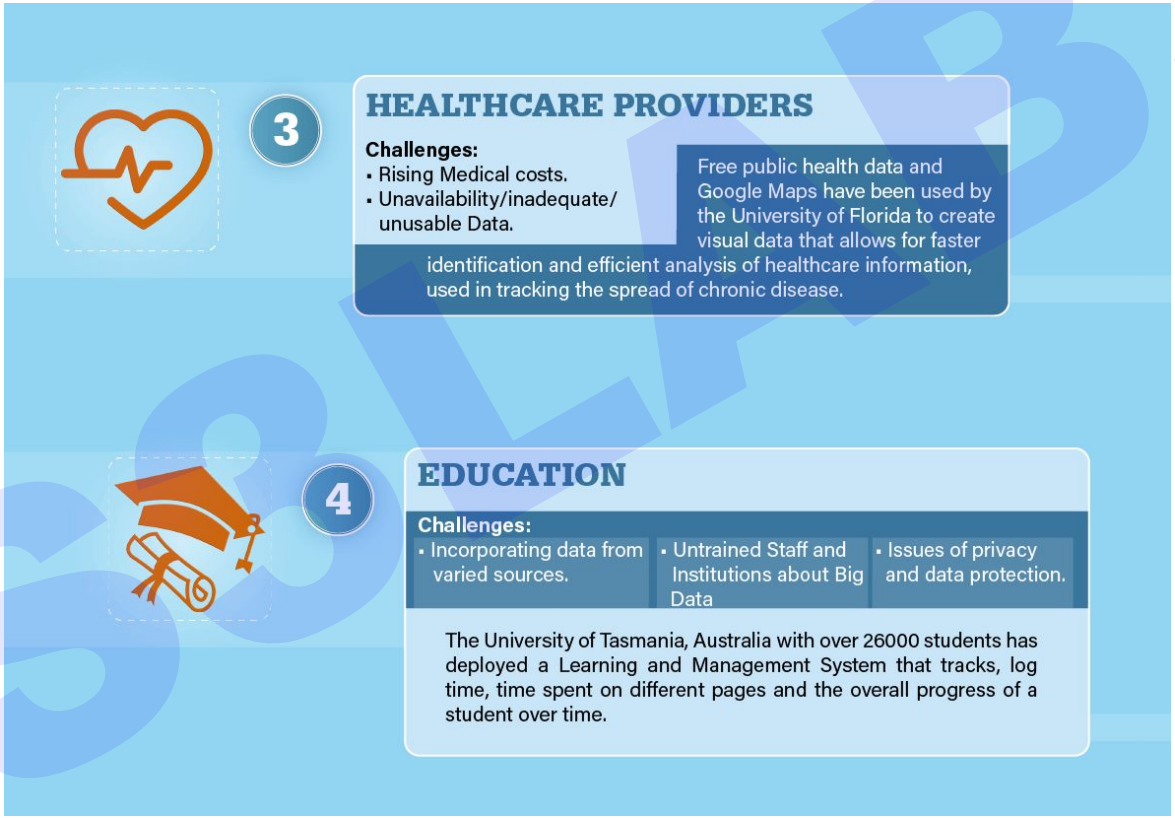
Big Data: Applications



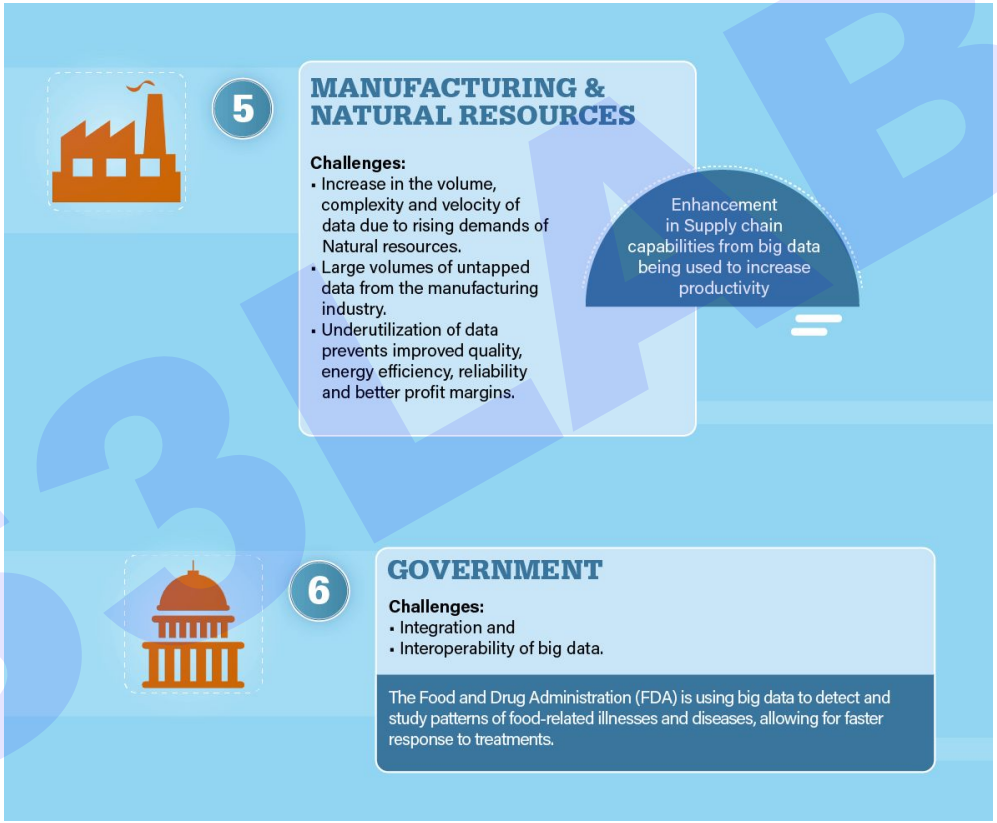
Big Data: Applications



Big Data: Applications

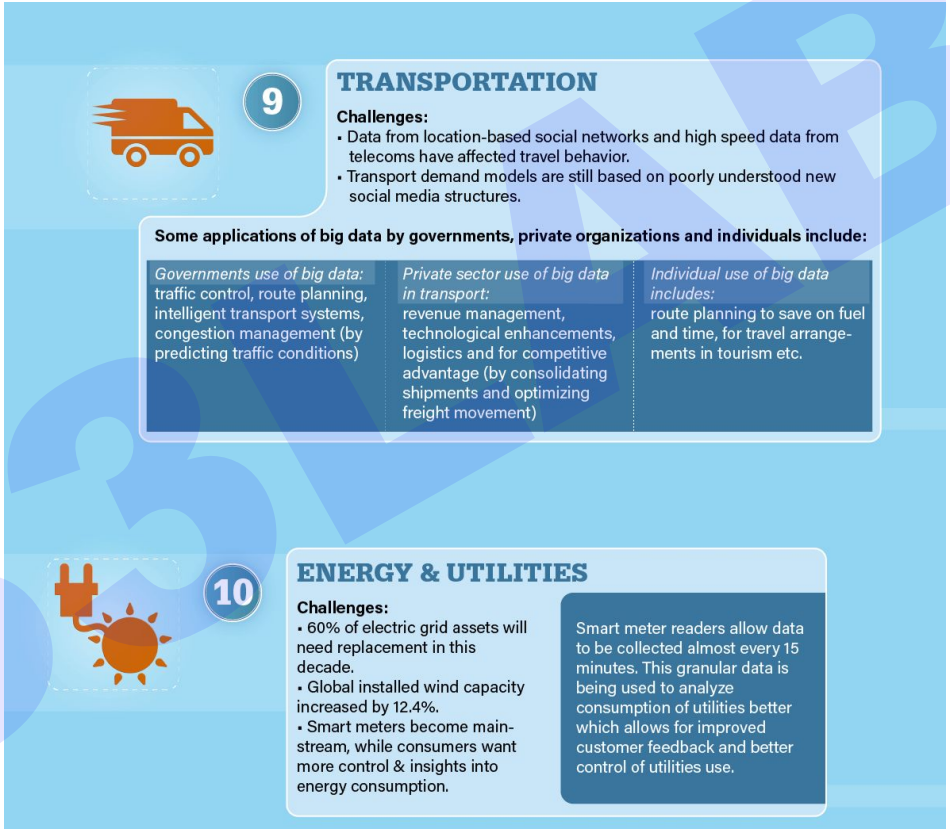


Big Data: Applications

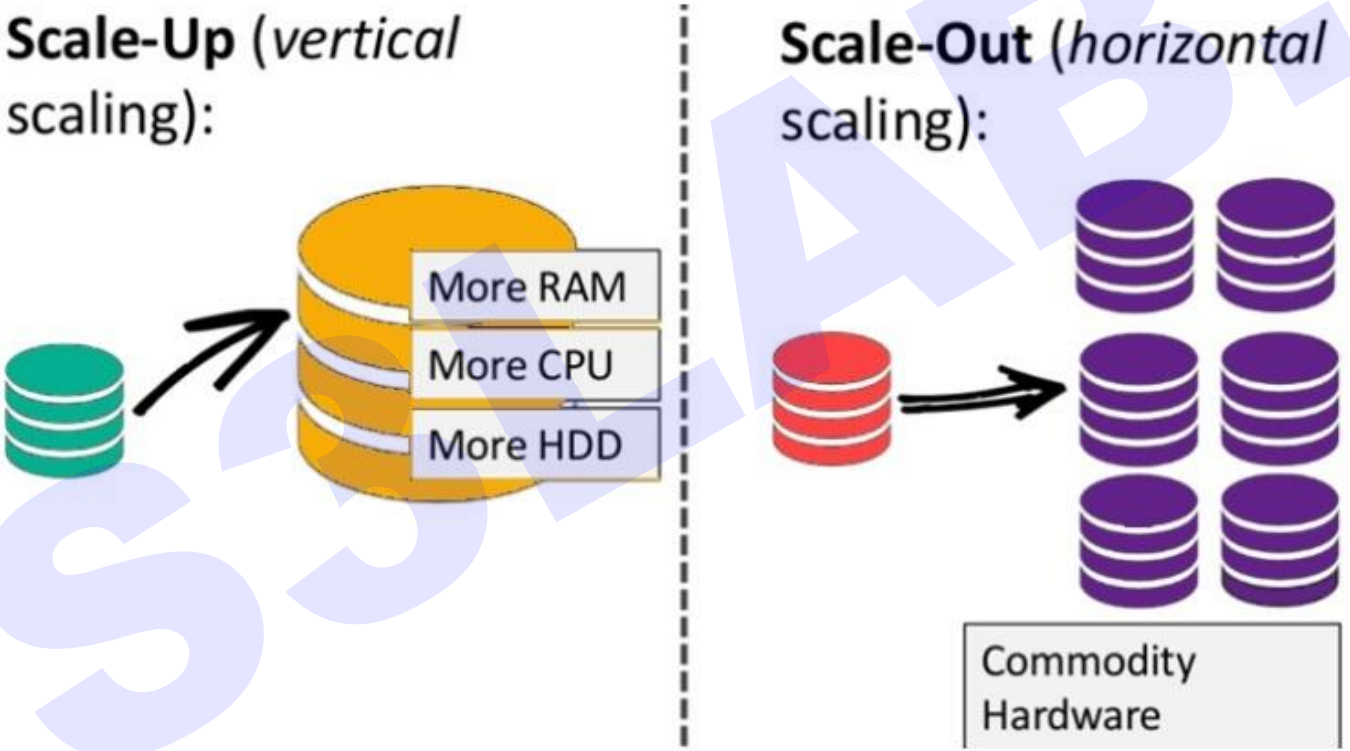


Big Data: Applications

Big Data: Applications



Big Data: Scale

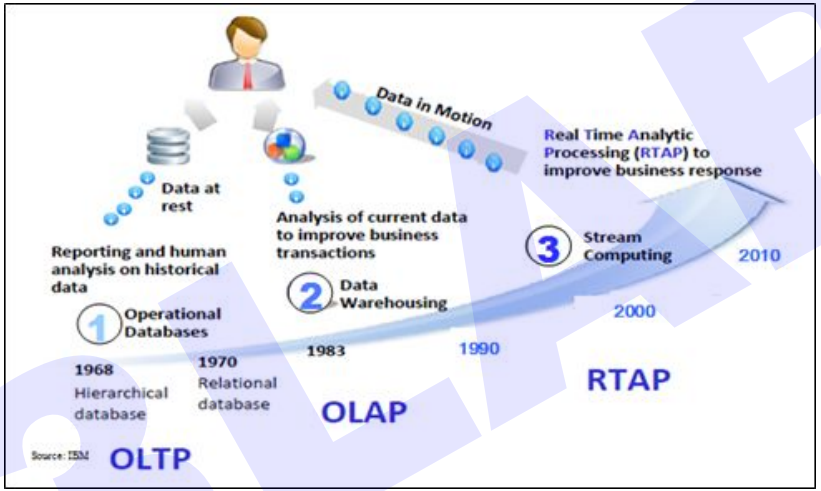


Big Data: Evolution

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

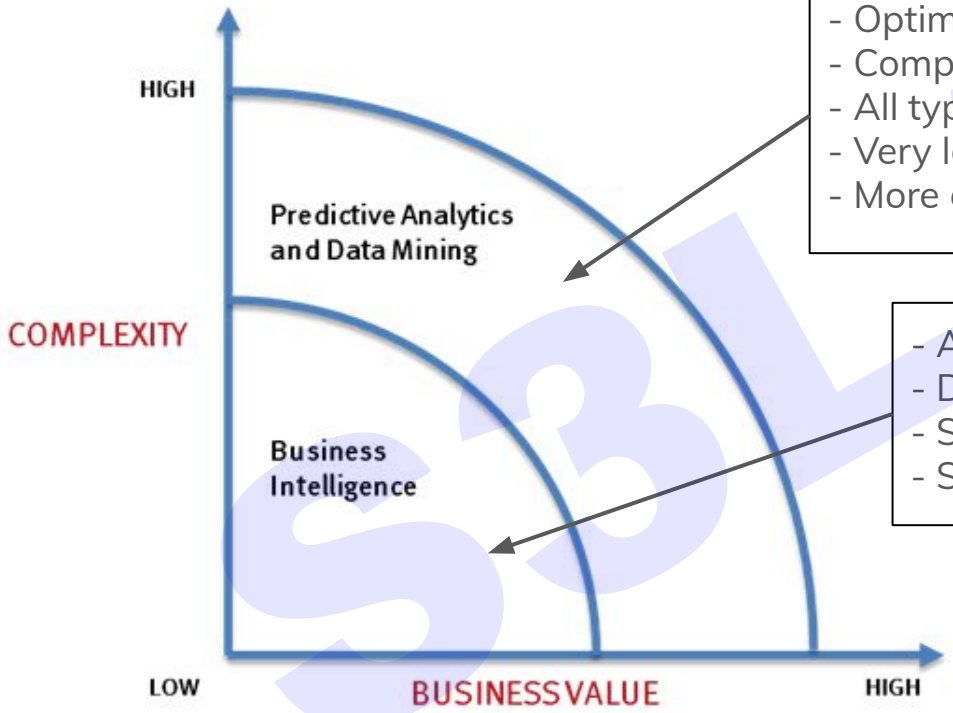
  - New Model: All of us are generating data, and all of us are consuming data

Big Data: Evolution



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

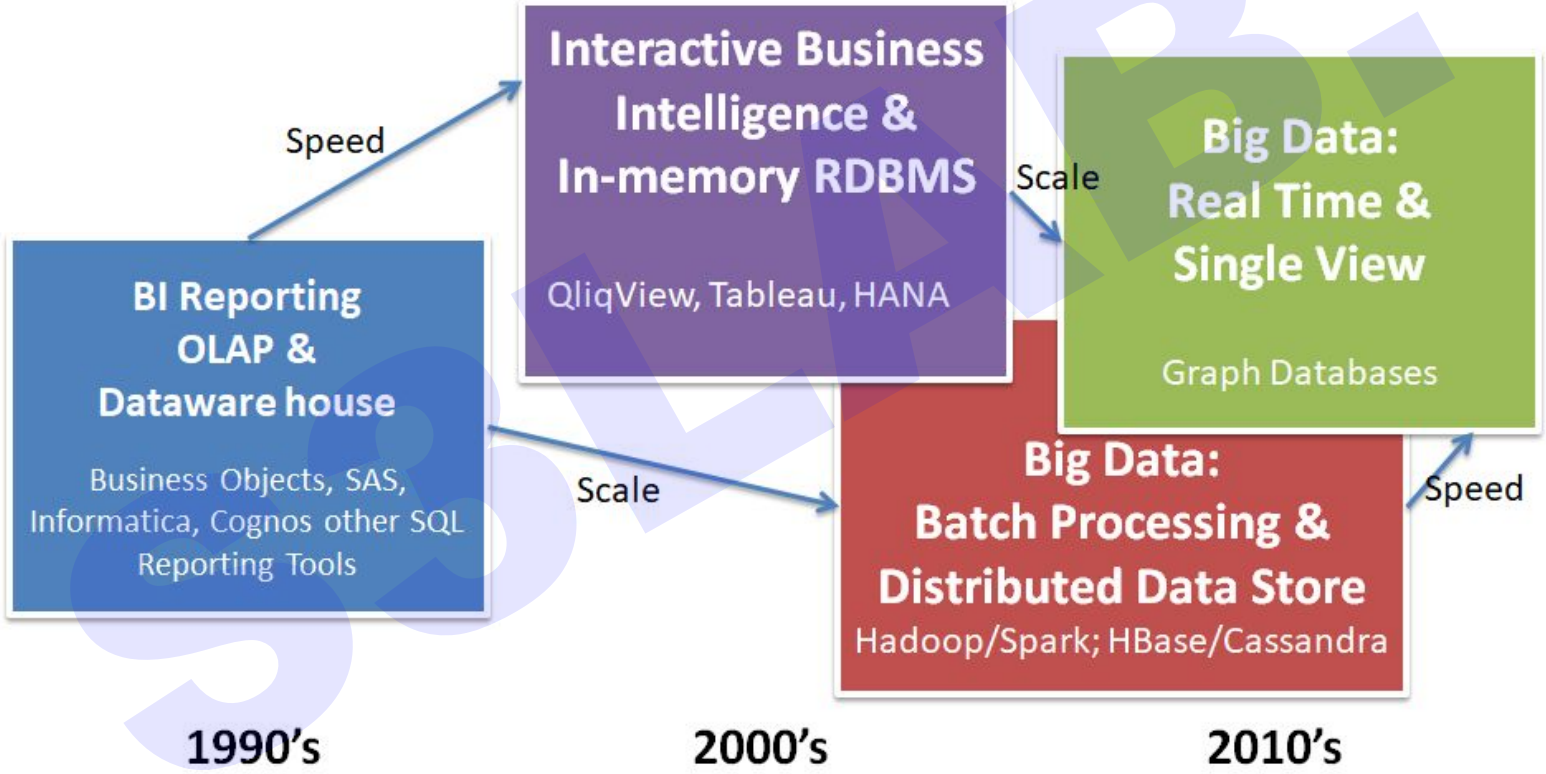
Big Data: Evolution



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

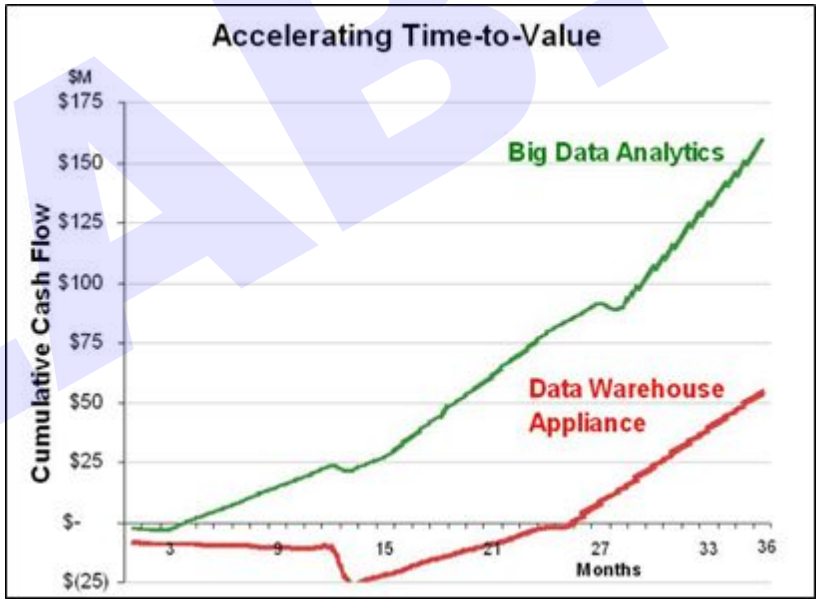
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to midsize datasets

Big Data: Evolution

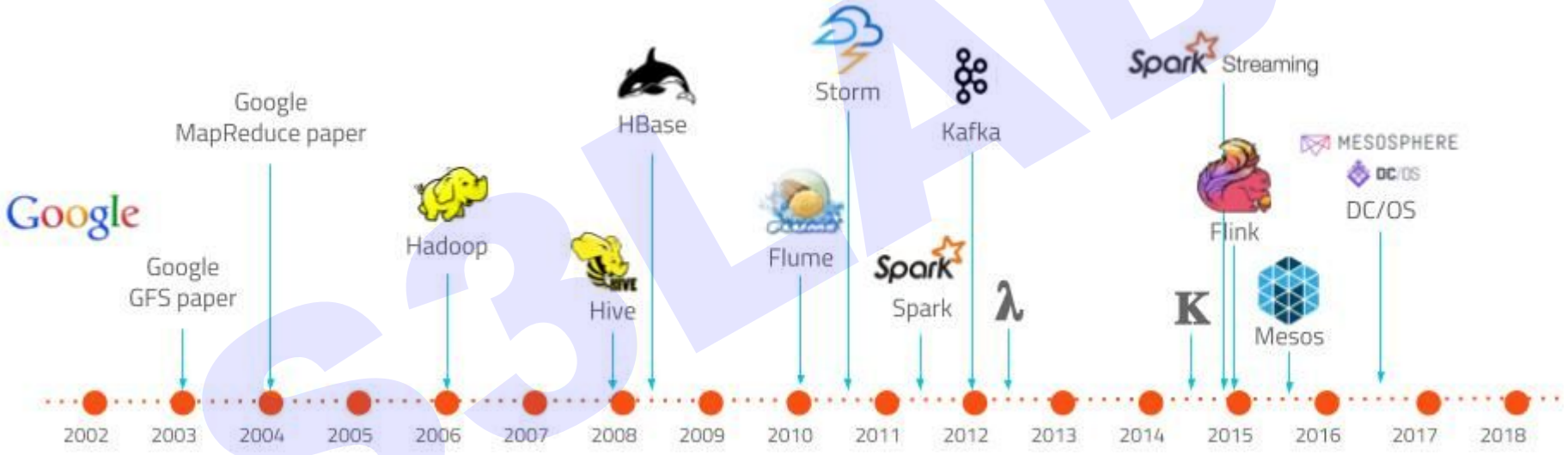


Big Data: Evolution

- 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

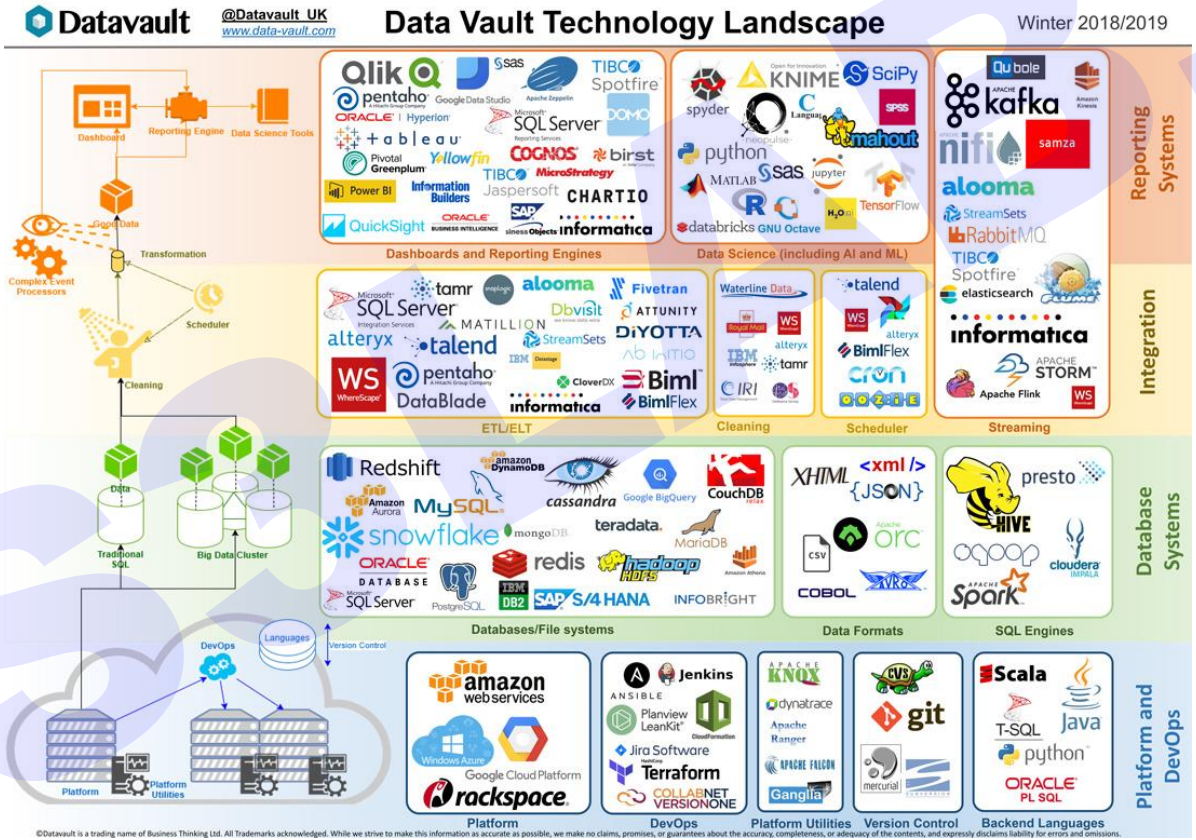


Big Data: Evolution

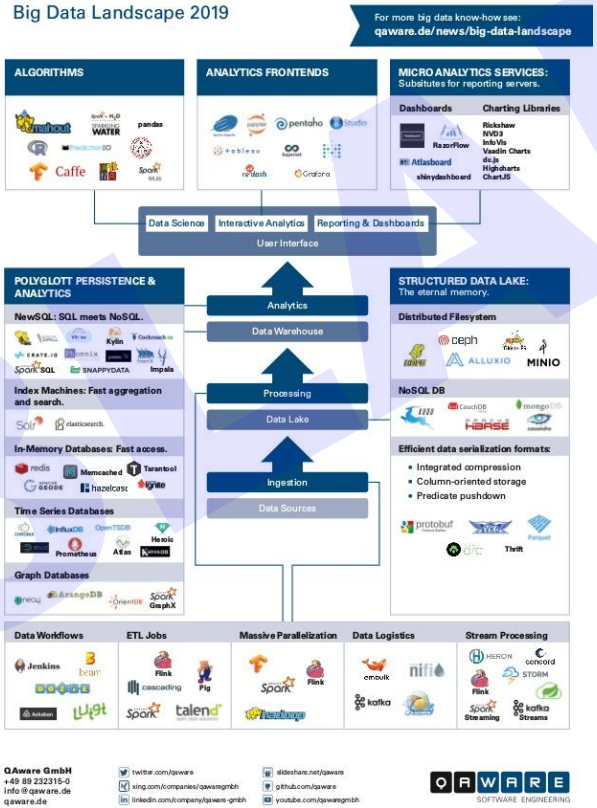


Big Data: Evolution

Big Data: Landscape



Big Data: Landscape

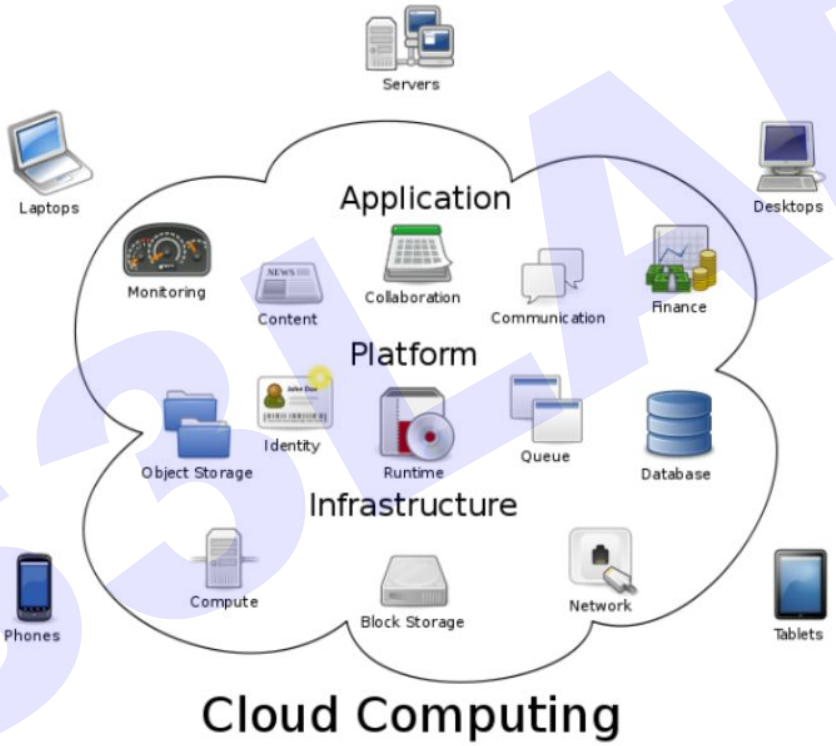


Big Data: Landscape (Open sources)

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, …

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

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

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.

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

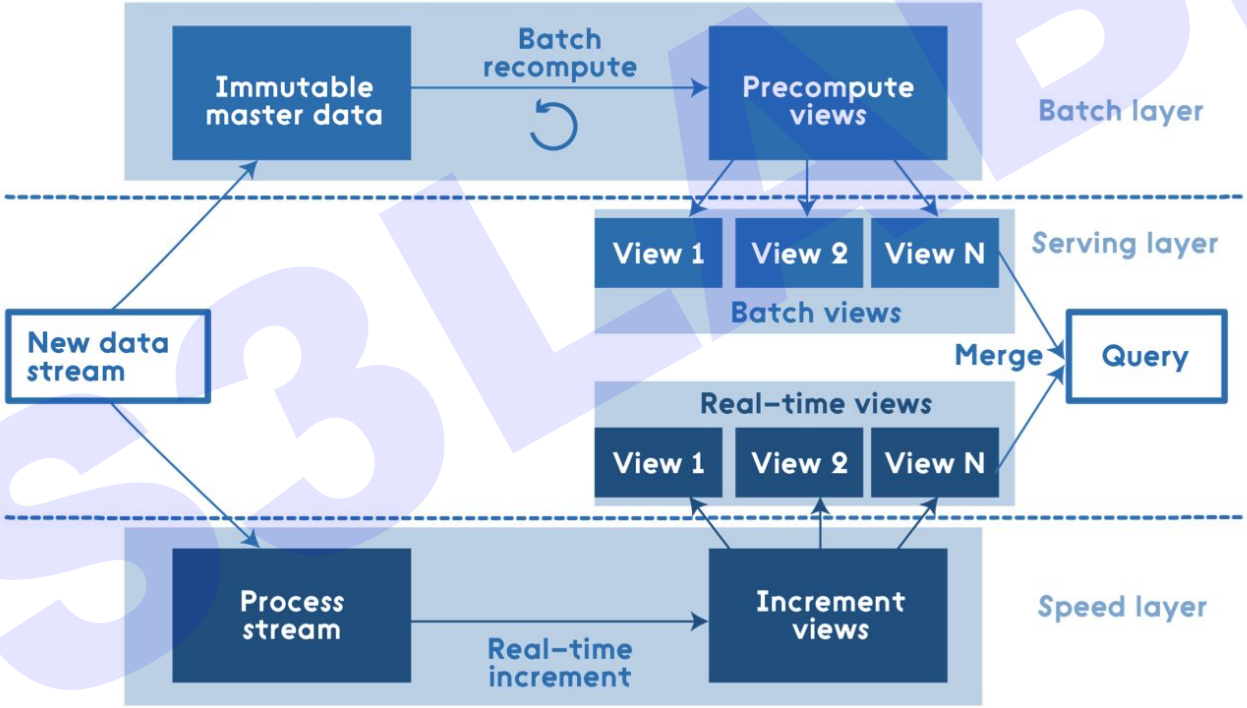
Big Data: Lambda Architecture

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.

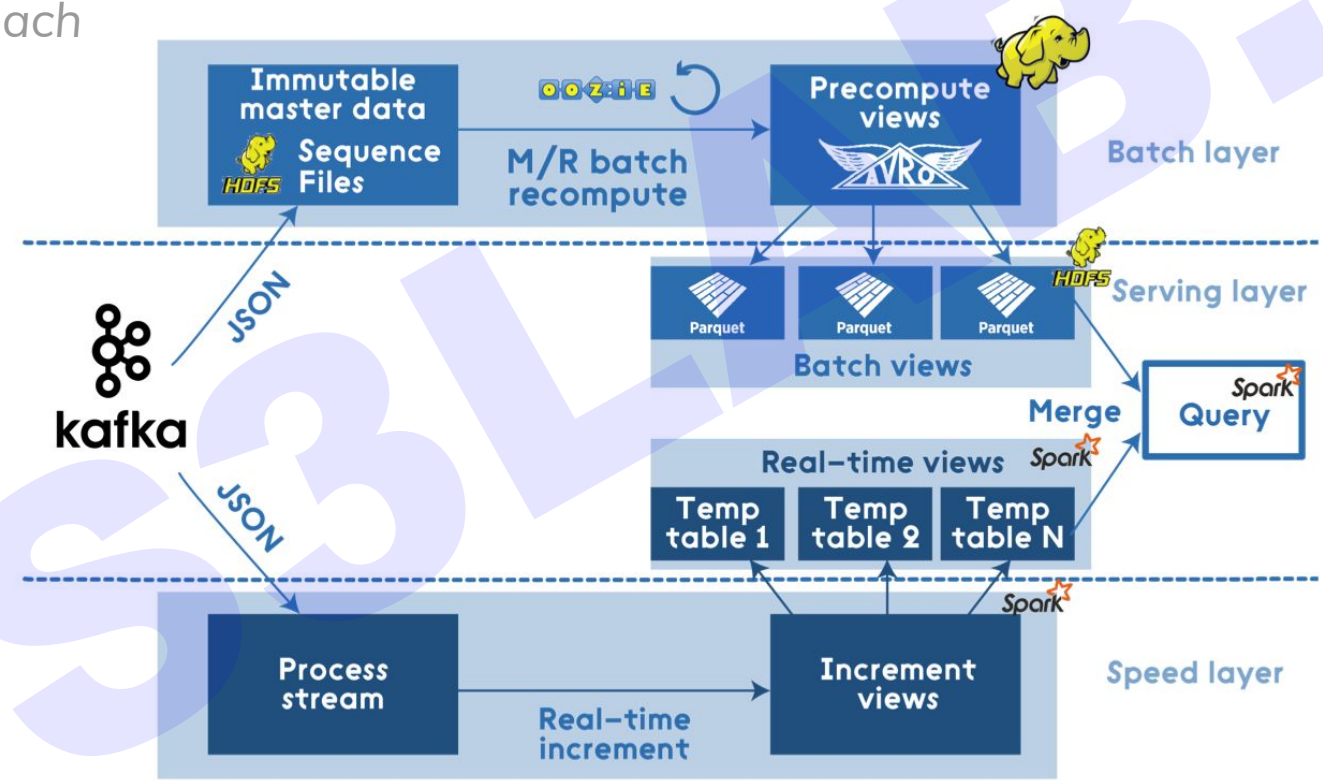
Big Data: Lambda Architecture

What is Lambda Architecture?



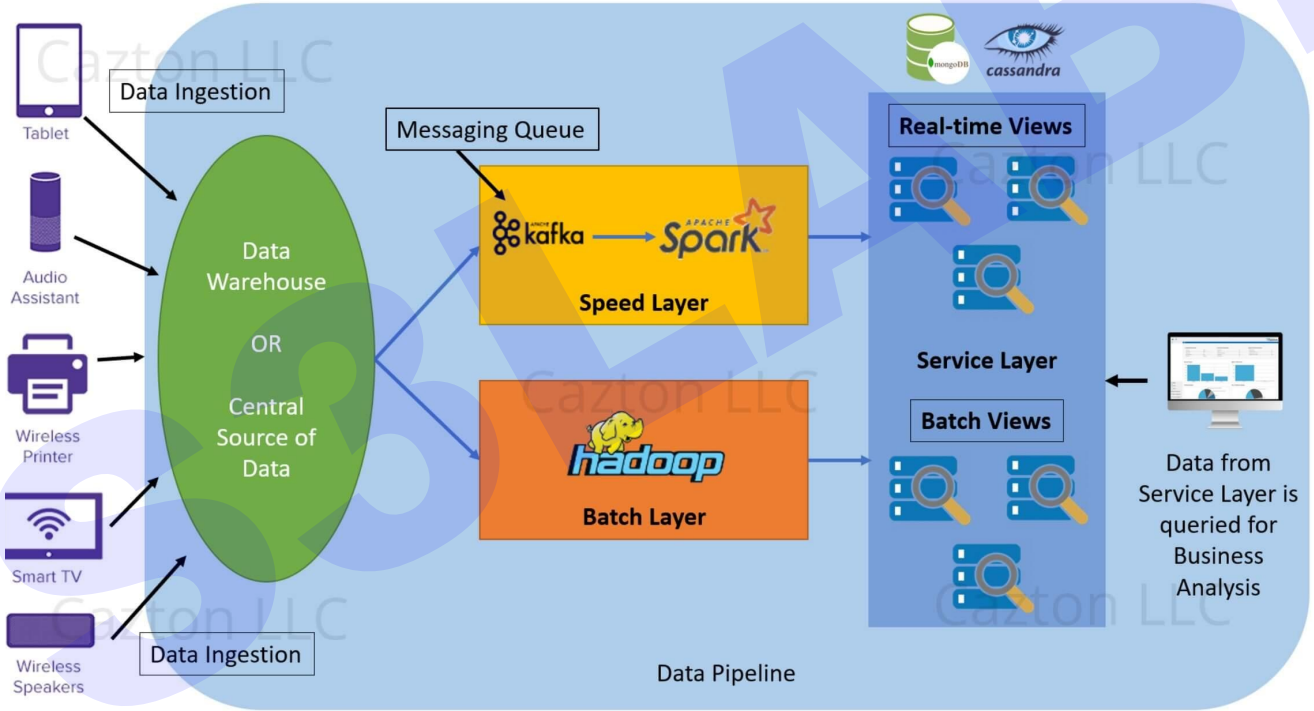
Big Data: Lambda Architecture

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.