

Big Data

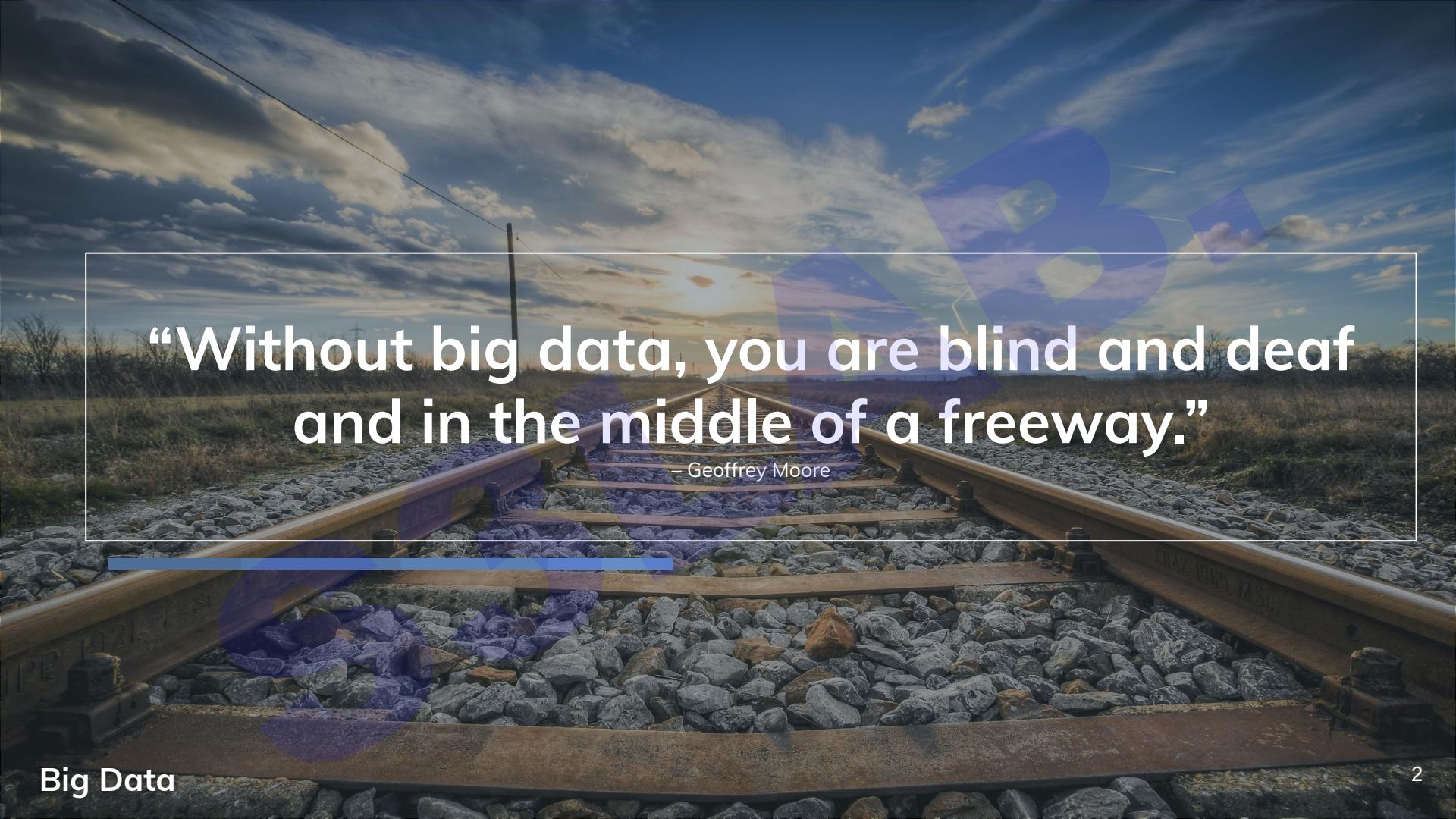
(Understanding about Big data)

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

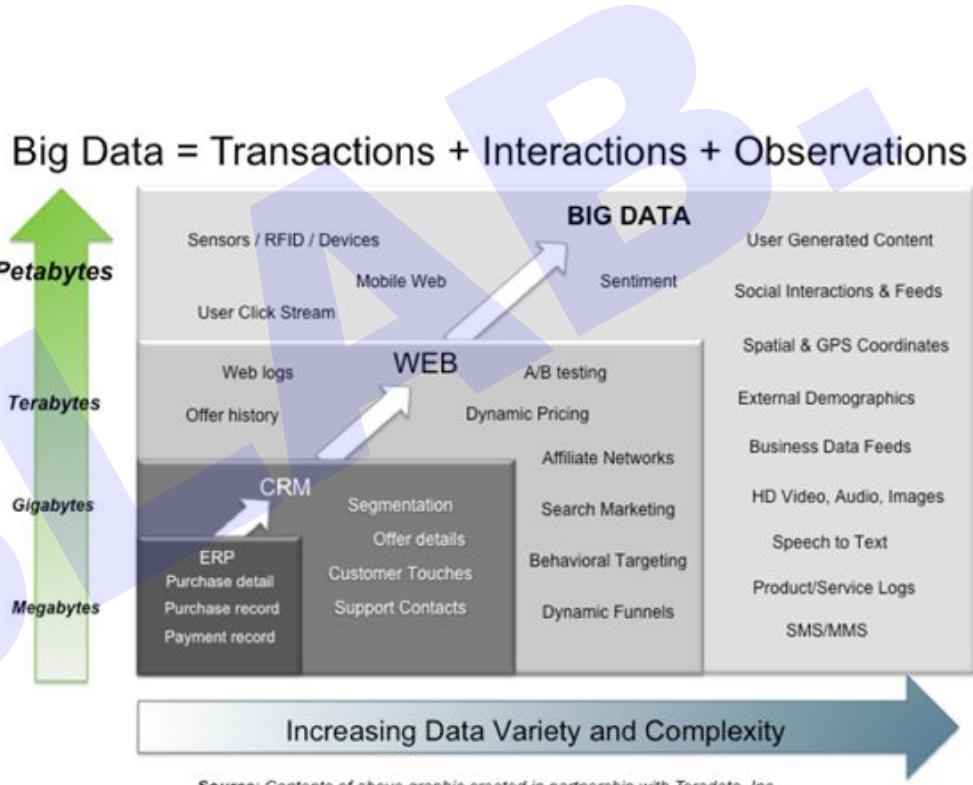


What is Big Data

- 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

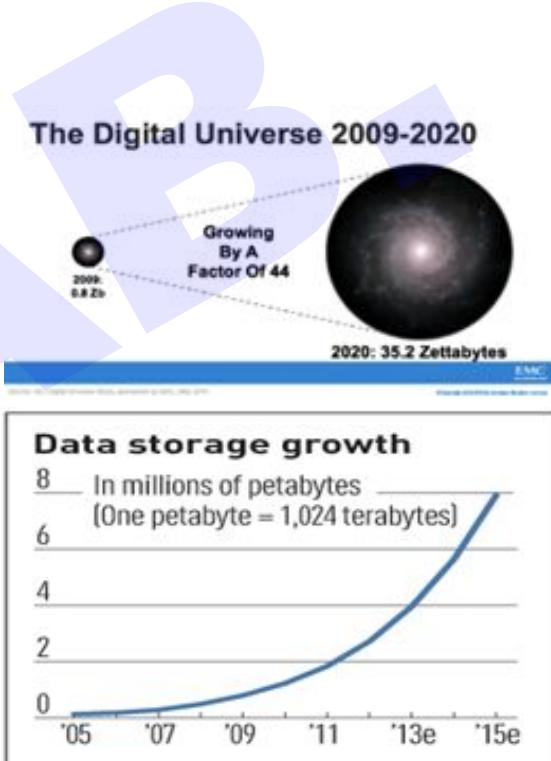
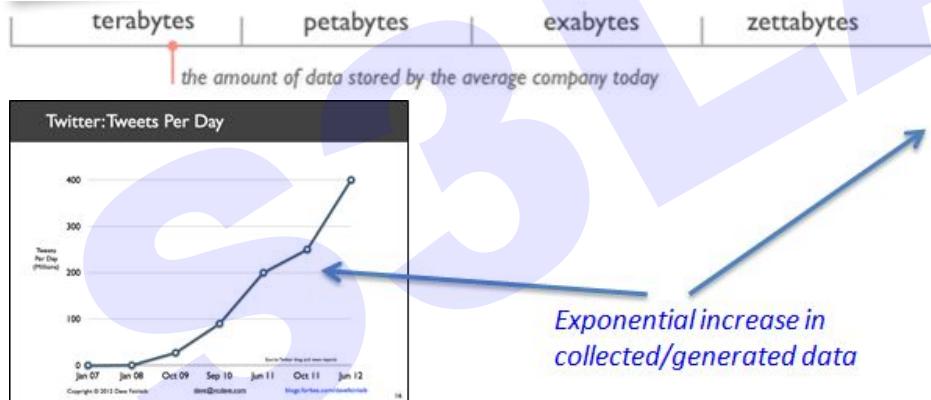




Big Data: 3V's

Volume (scale)

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





Big Data: 3V's

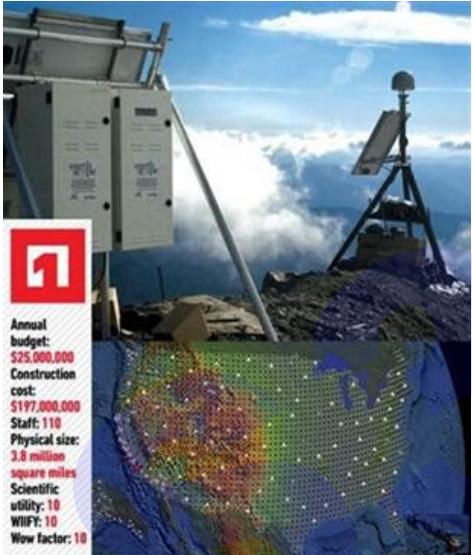
Volume (scale)



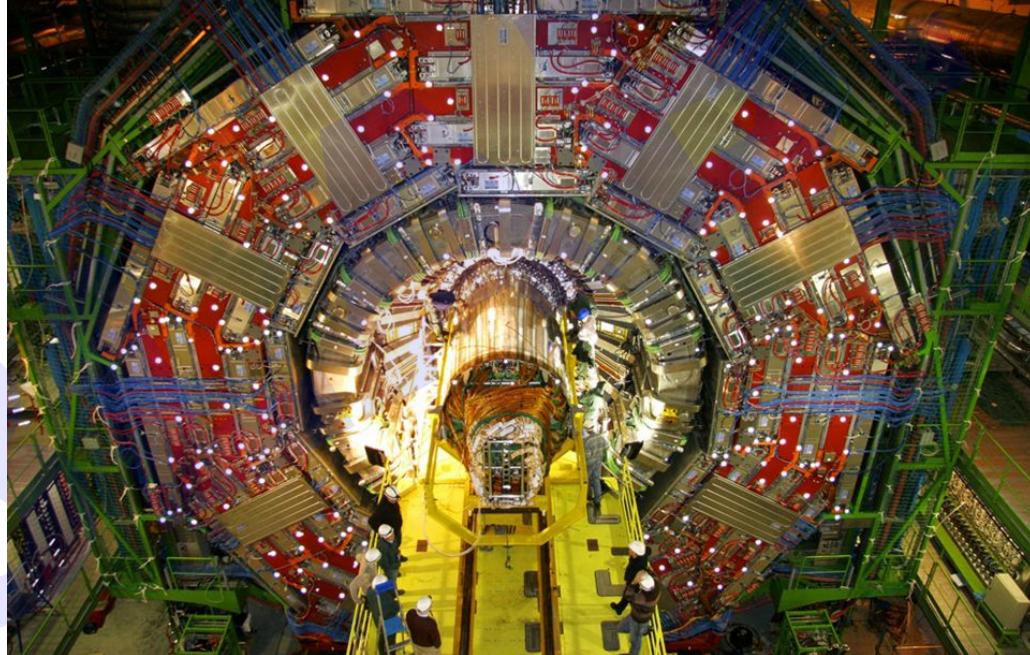


Big Data: 3V's

Volume (scale)



EarthScope - 67 terabytes of data



CERN's Large Hadron 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.

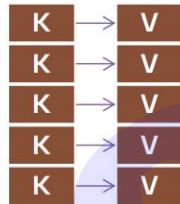


Big Data: 3V's

Variety (Complexity)

- Semi-Structured, NoSQL

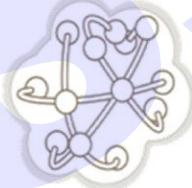
Key-Value Stores



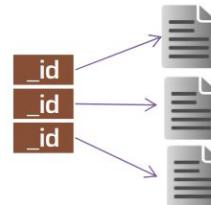
Column Stores



Graph Databases



Document Stores

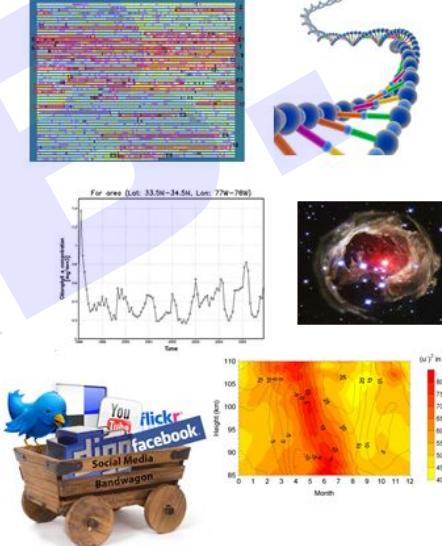




Big Data: 3V's

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.)

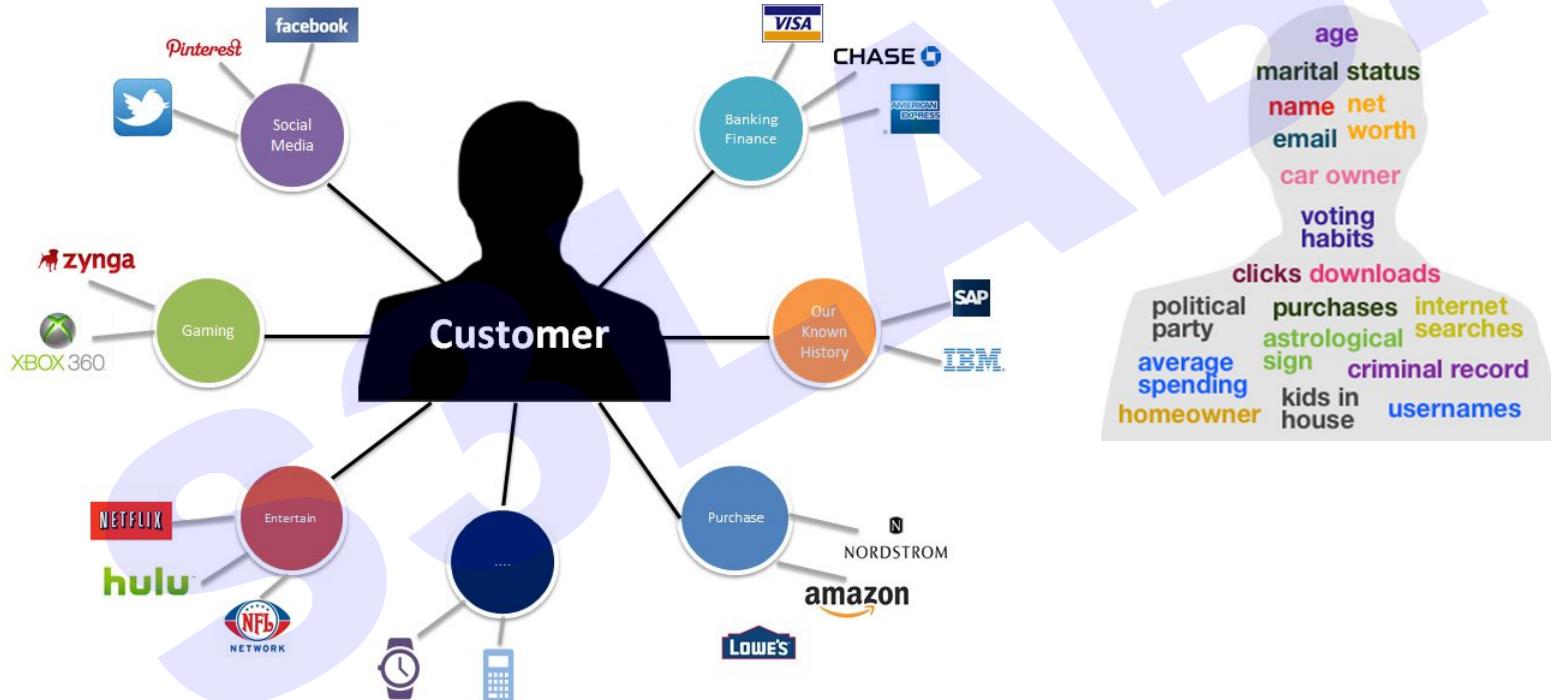


To extract knowledge ➡ all these types of data need to linked together



Big Data: 3V's

Variety (Complexity)





Big Data: 3V's

Velocity (Speed)

- Data is being 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

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)



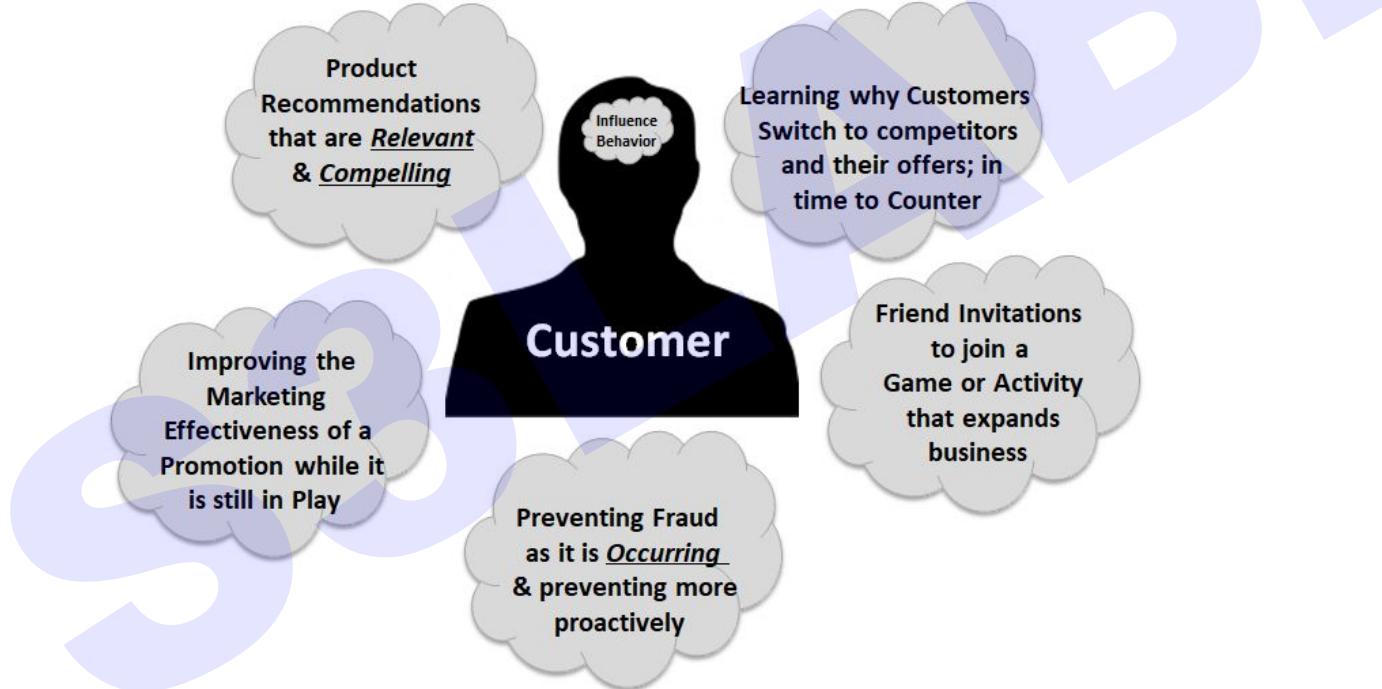
Sensor technology and networks
(measuring all kinds of data)

- 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



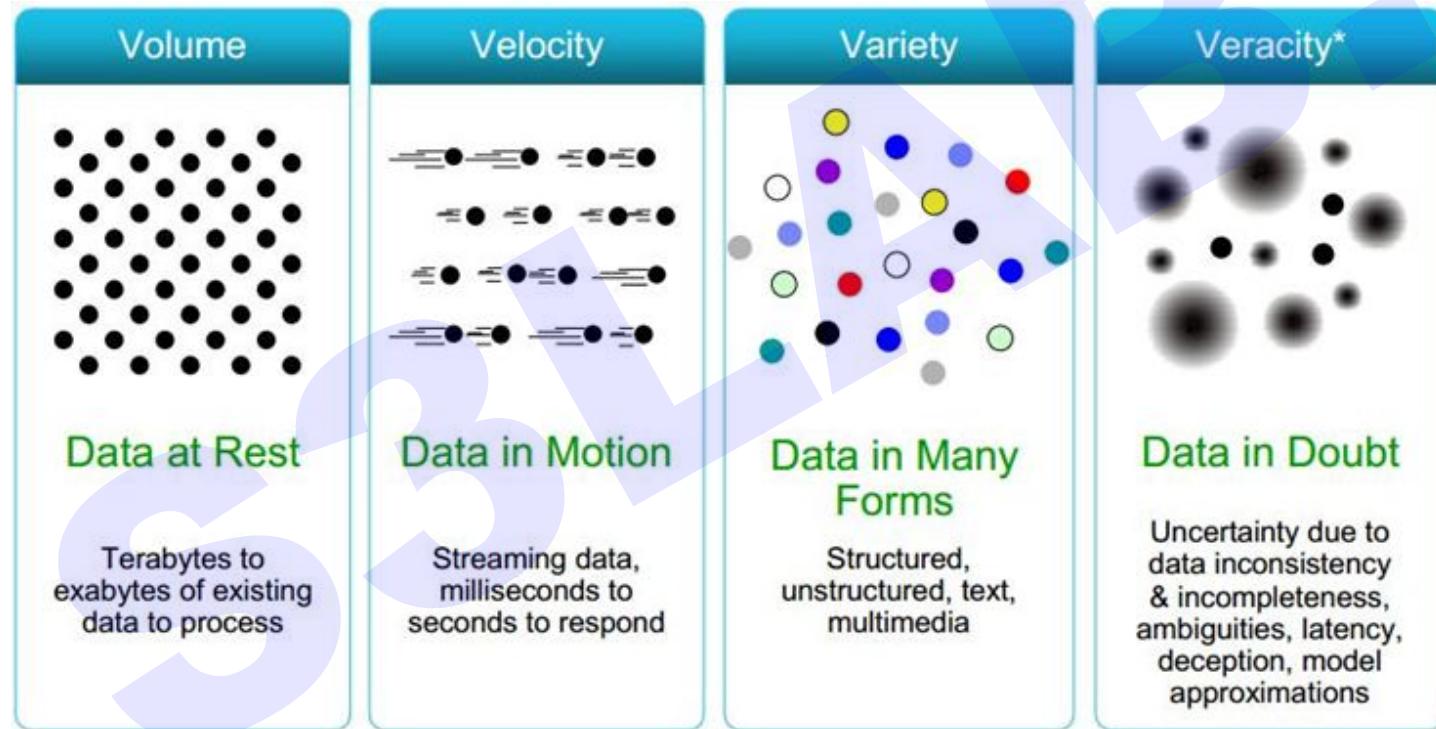
Big Data: 3V's

Velocity (Speed)



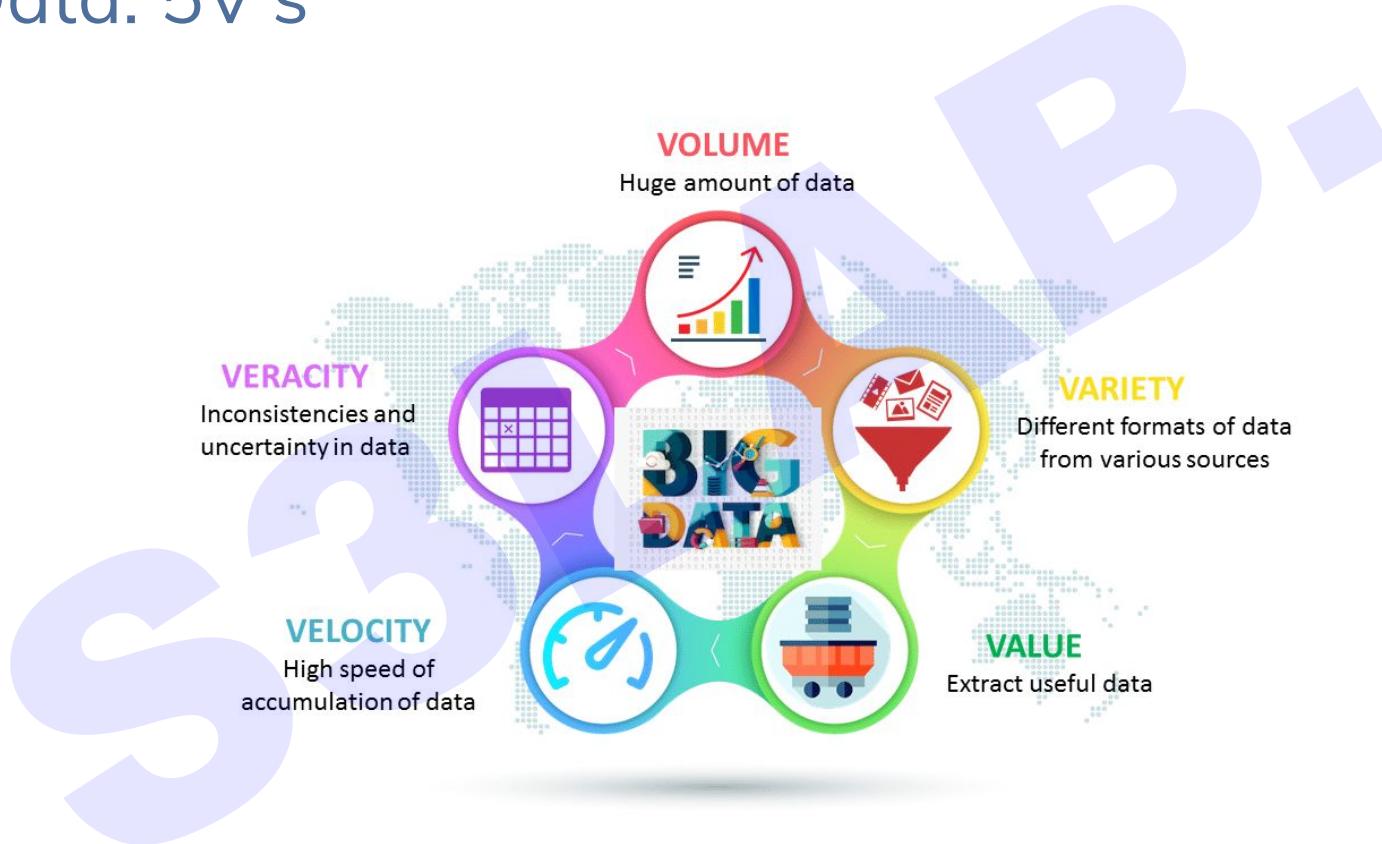


Big Data: 4V's





Big Data: 5V's





Big Data: NV'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

1

BANKING AND SECURITIES

Challenges:

- Early warning for Securities fraud and Trade visibility.
- Card fraud detection and audit trails.
- Enterprise credit risk reporting.
- Customer data transformation and analytics.

The Securities Exchange Commission (SEC) is using big data to monitor financial market activity by using network analytics and natural language processors. This helps to catch illegal trading activity in the financial markets.

2

COMMUNICATIONS, MEDIA & ENTERTAINMENT

Challenges:

- Collecting, analyzing and utilizing consumer insights.
- Leveraging mobile and social media content.
- Understanding patterns of real-time, media content usage.

Wimbledon Championships leverages big data to deliver detailed sentiment analysis on the tennis matches to TV, mobile and web users in real-time.



Big Data: Applications

3

HEALTHCARE PROVIDERS

Challenges:

- Rising Medical costs.
- Unavailability/inadequate/unusable Data.

Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

4

EDUCATION

Challenges:

• Incorporating data from varied sources.	• Untrained Staff and Institutions about Big Data	• Issues of privacy and data protection.
---	---	--

The University of Tasmania, Australia with over 26000 students has deployed a Learning and Management System that tracks, log time, time spent on different pages and the overall progress of a student over time.



Big Data: Applications

MANUFACTURING & NATURAL RESOURCES

Challenges:

- Increase in the volume, complexity and velocity of data due to rising demands of Natural resources.
- Large volumes of untapped data from the manufacturing industry.
- Underutilization of data prevents improved quality, energy efficiency, reliability and better profit margins.

Enhancement in Supply chain capabilities from big data being used to increase productivity

GOVERNMENT

Challenges:

- Integration and
- Interoperability of big data.

The Food and Drug Administration (FDA) is using big data to detect and study patterns of food-related illnesses and diseases, allowing for faster response to treatments.



Big Data: Applications

The infographic illustrates the applications of Big Data across several industries, each accompanied by a numbered icon and a small icon representing the industry.

INSURANCE (Icon: House in hands)

Challenges:

- Lack of personalized services, pricing, targeted services to new market segments.
- Underutilization of data gathered by loss adjusters.
- Hunger for better insight.

Customer insights for transparent and simpler products.

Predicting customer behavior through data derived from social media, GPS-enabled devices and CCTV footage.

Claims management, predictive analytics from big data has been used to offer faster service

RETAIL & WHOLESALE TRADE (Icon: Bar chart with arrow)

Utilized Data derived from customer loyalty cards, POS scanners, RFID etc.

Optimized staffing through data from shopping patterns, local events etc.

Reduced fraud and

Timely analysis of inventory



Big Data: Applications



9

TRANSPORTATION

Challenges:

- Data from location-based social networks and high speed data from telecoms have affected travel behavior.
- Transport demand models are still based on poorly understood new social media structures.

Some applications of big data by governments, private organizations and individuals include:

Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)

Private sector use of big data in transport:

revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)

Individual use of big data includes:

route planning to save on fuel and time, for travel arrangements in tourism etc.



10

ENERGY & UTILITIES

Challenges:

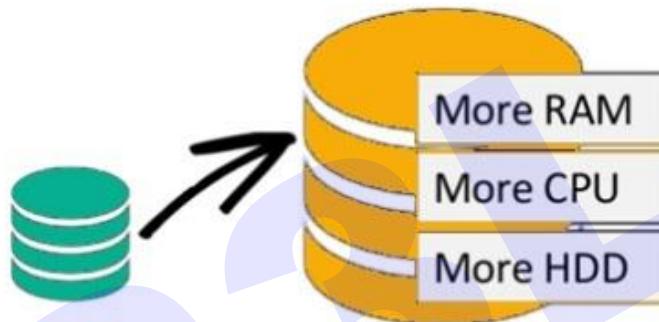
- 60% of electric grid assets will need replacement in this decade.
- Global installed wind capacity increased by 12.4%.
- Smart meters become mainstream, while consumers want more control & insights into energy consumption.

Smart meter readers allow data to be collected almost every 15 minutes. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use.

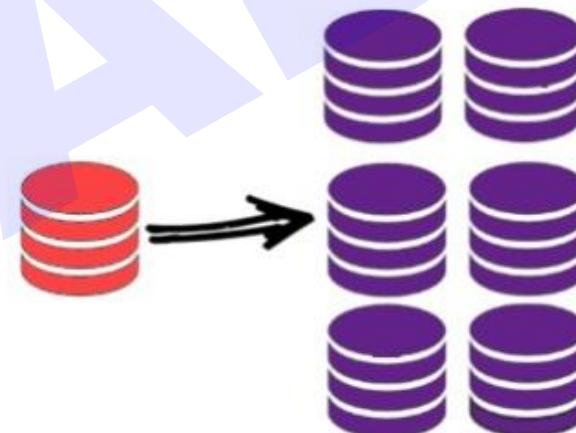


Big Data: Scale

Scale-Up (*vertical* scaling):



Scale-Out (*horizontal* scaling):





Big Data: Evolution

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

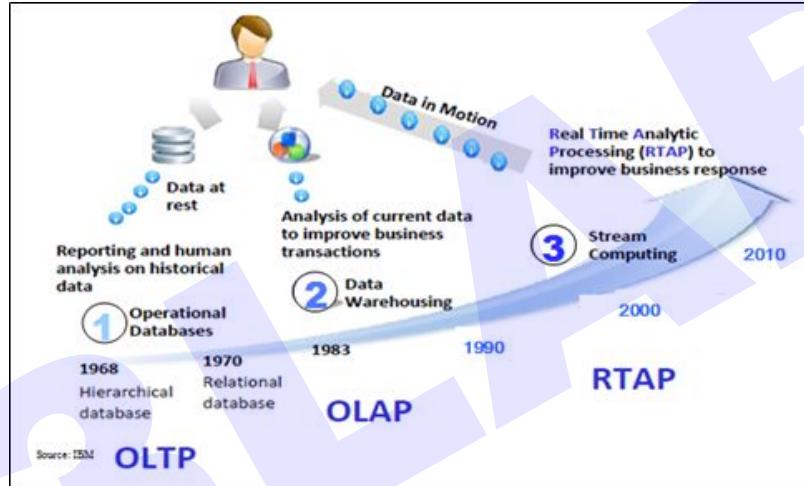


- New Model: All of us are generating data, and all of us are consuming data





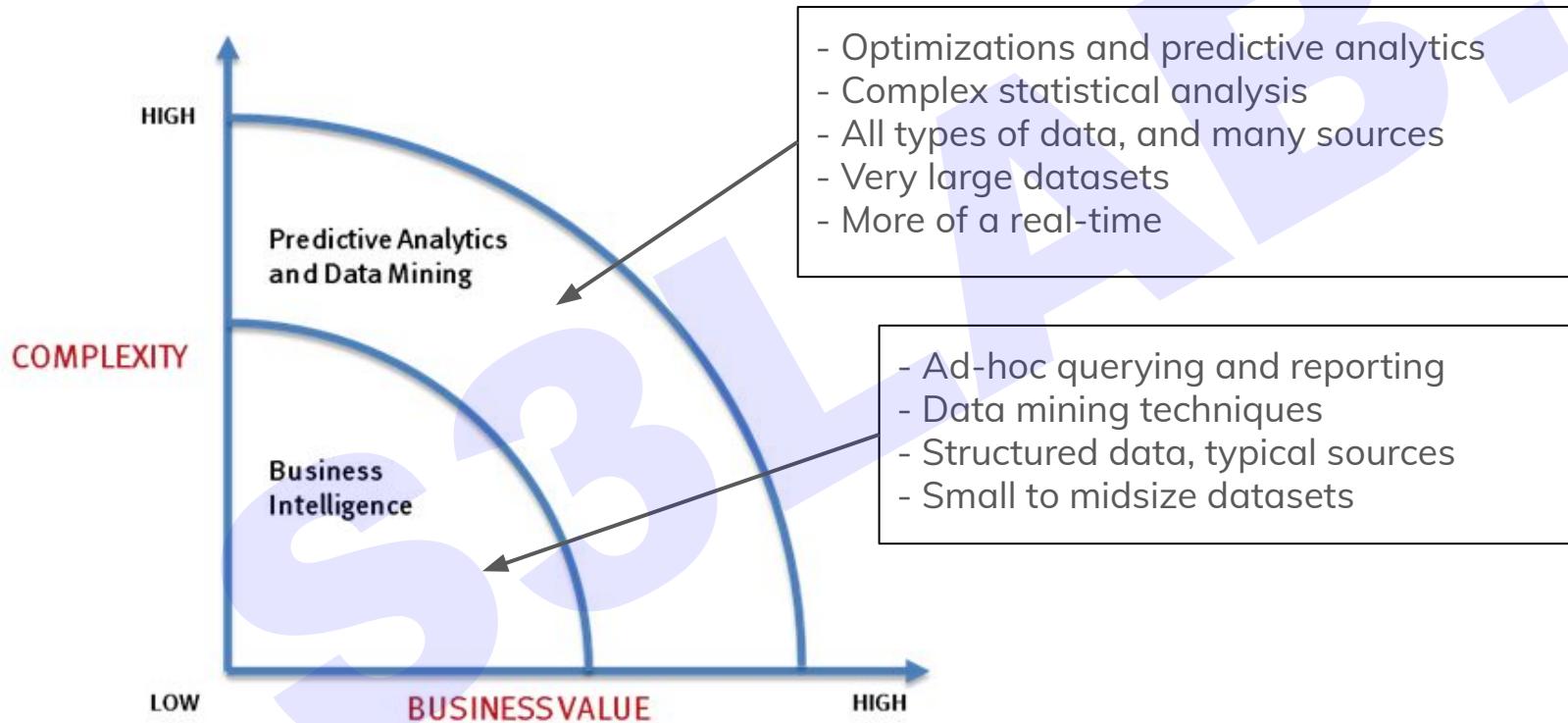
Big Data: Evolution



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

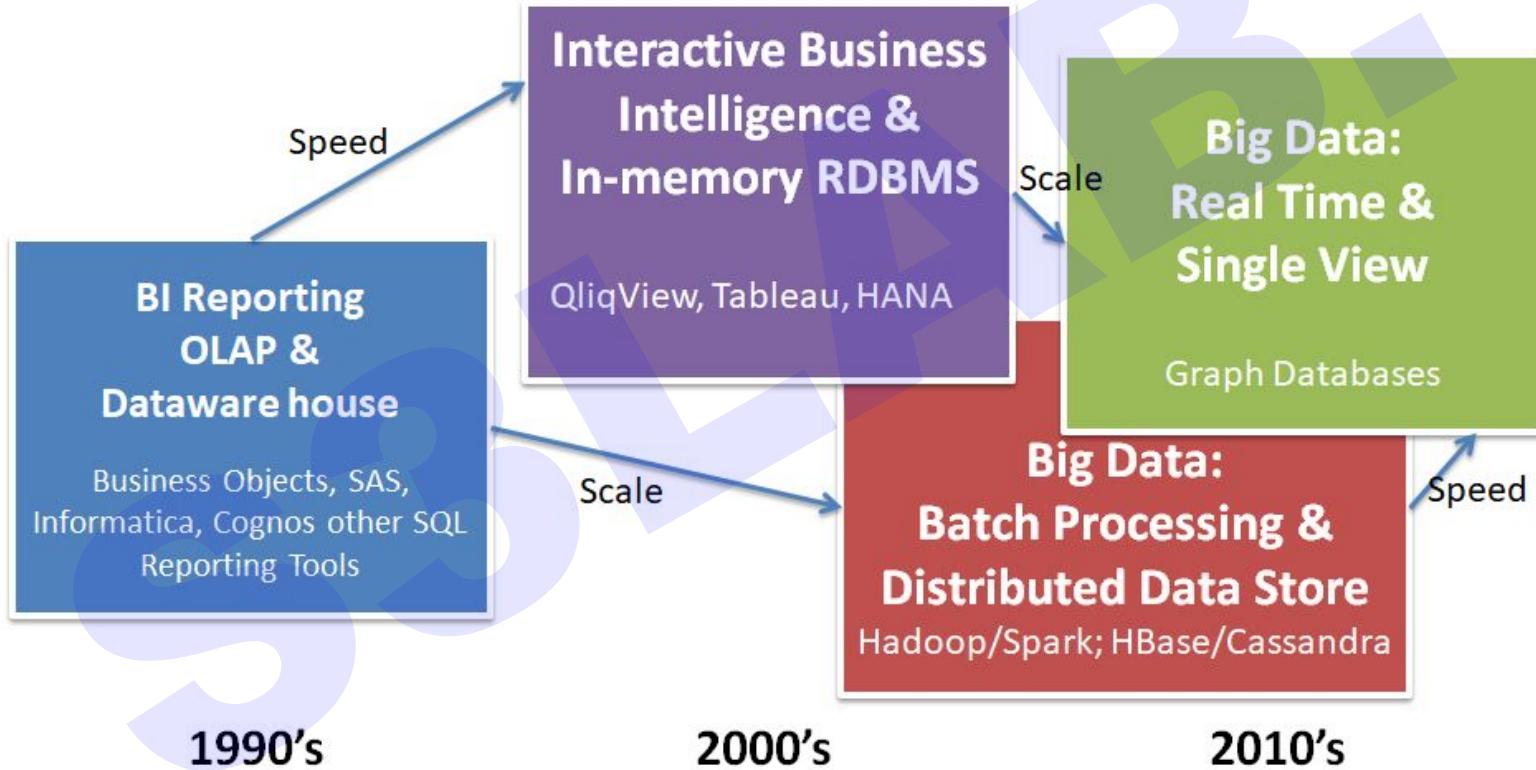


Big Data: Evolution





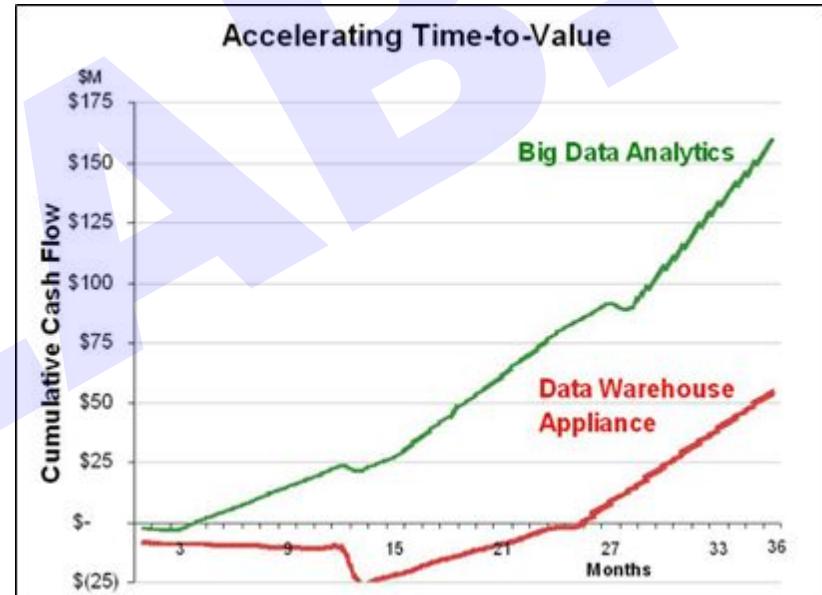
Big Data: Evolution





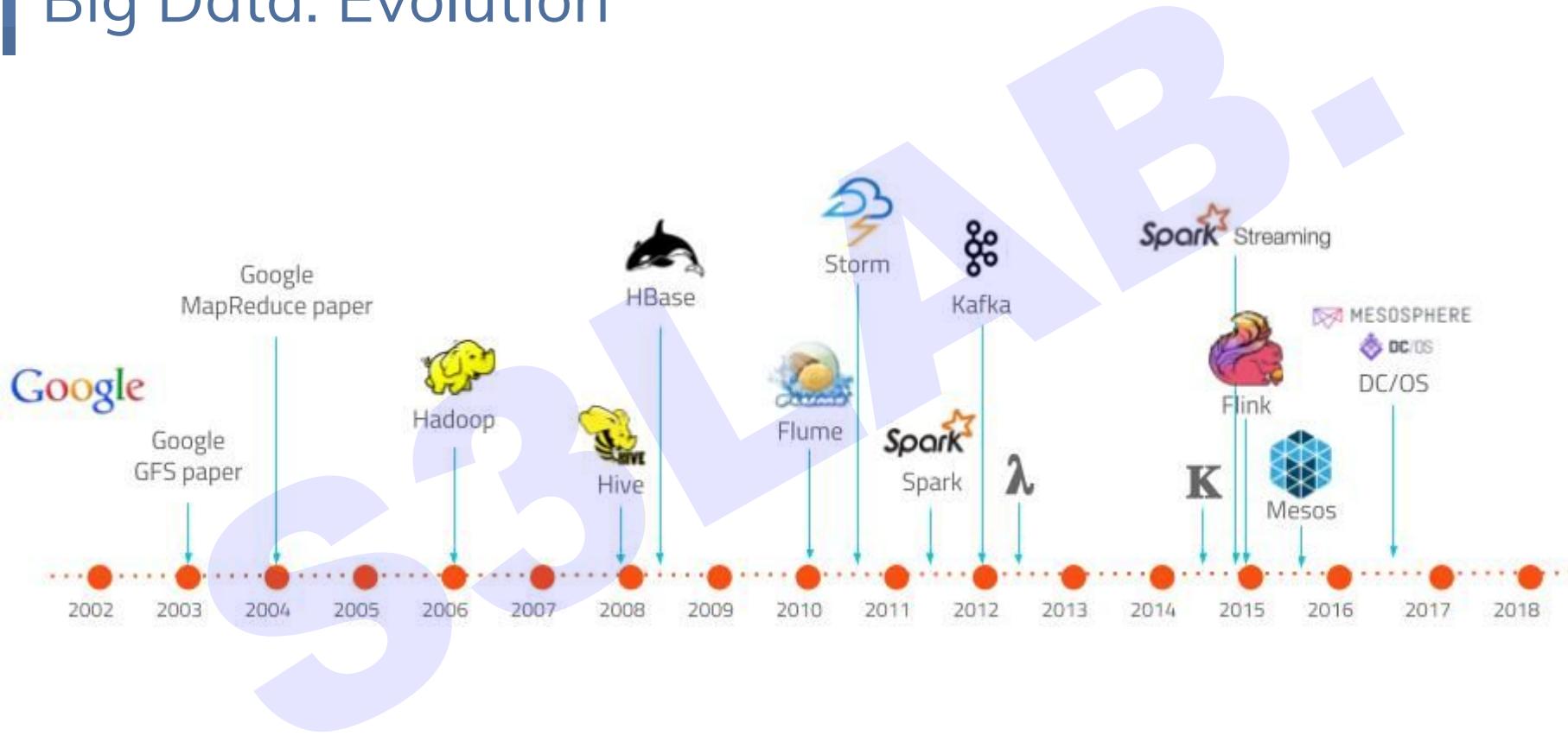
Big Data: Evolution

- 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

FUN FACTS?

Initially, decoding human genome took **10 years**, now it can be achieved in **1 week**.

Big Data market is projected to grow from **\$42 B** in 2018 to **\$103B** in 2027 attaining a 10.48% CAGR.

★ IN 2020 ★

44 Zettabytes data	1.7 MB data created by each user/sec	6.1 Billion Smartphone users	21 Billion IOT Devices	1/3rd data passing through Cloud
--------------------	--------------------------------------	------------------------------	------------------------	----------------------------------



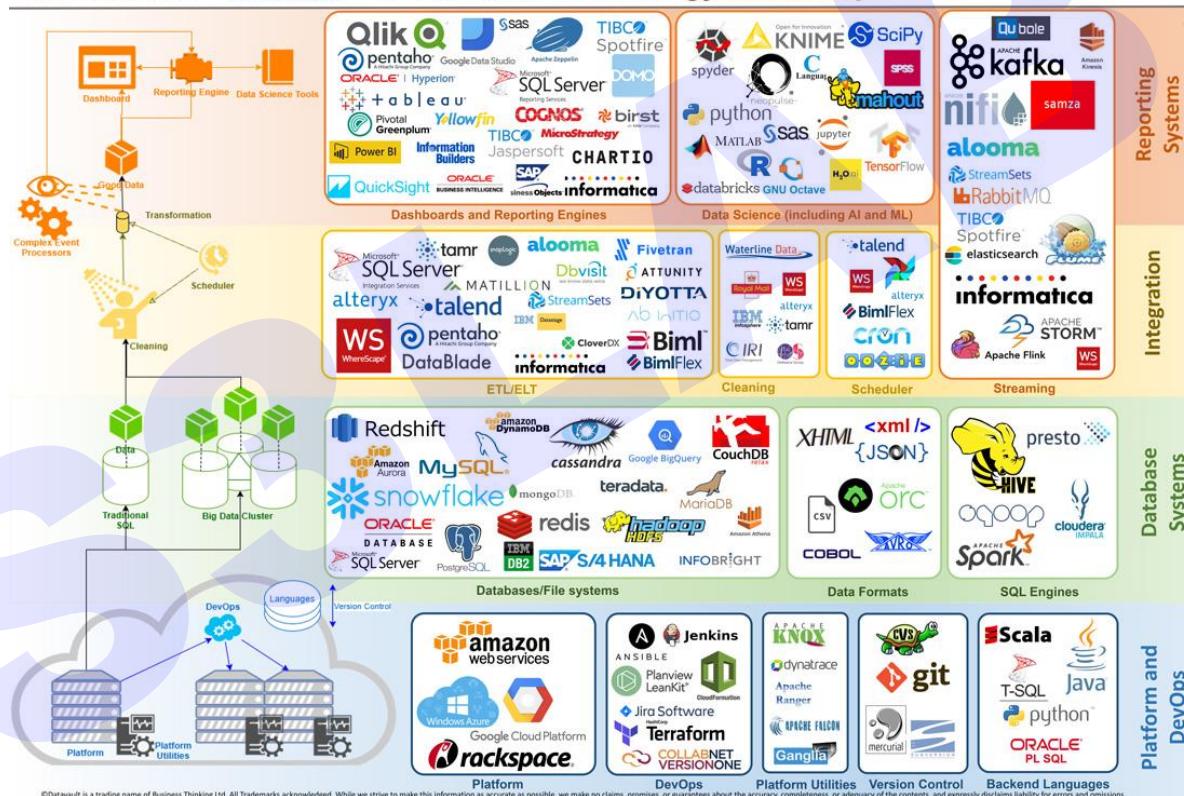
Big Data: Landscape



@Datavault_UK
www.data-vault.com

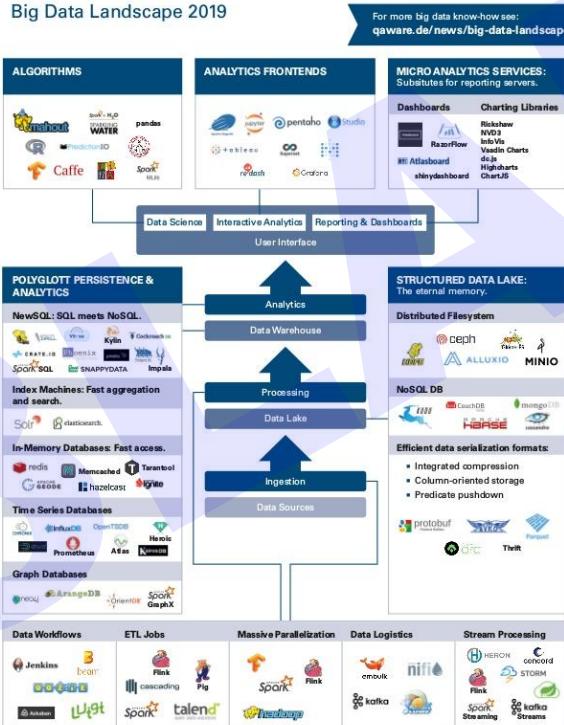
Data Vault Technology Landscape

Winter 2018/2019





Big Data: Landscape





Big Data: Landscape (Open sources)



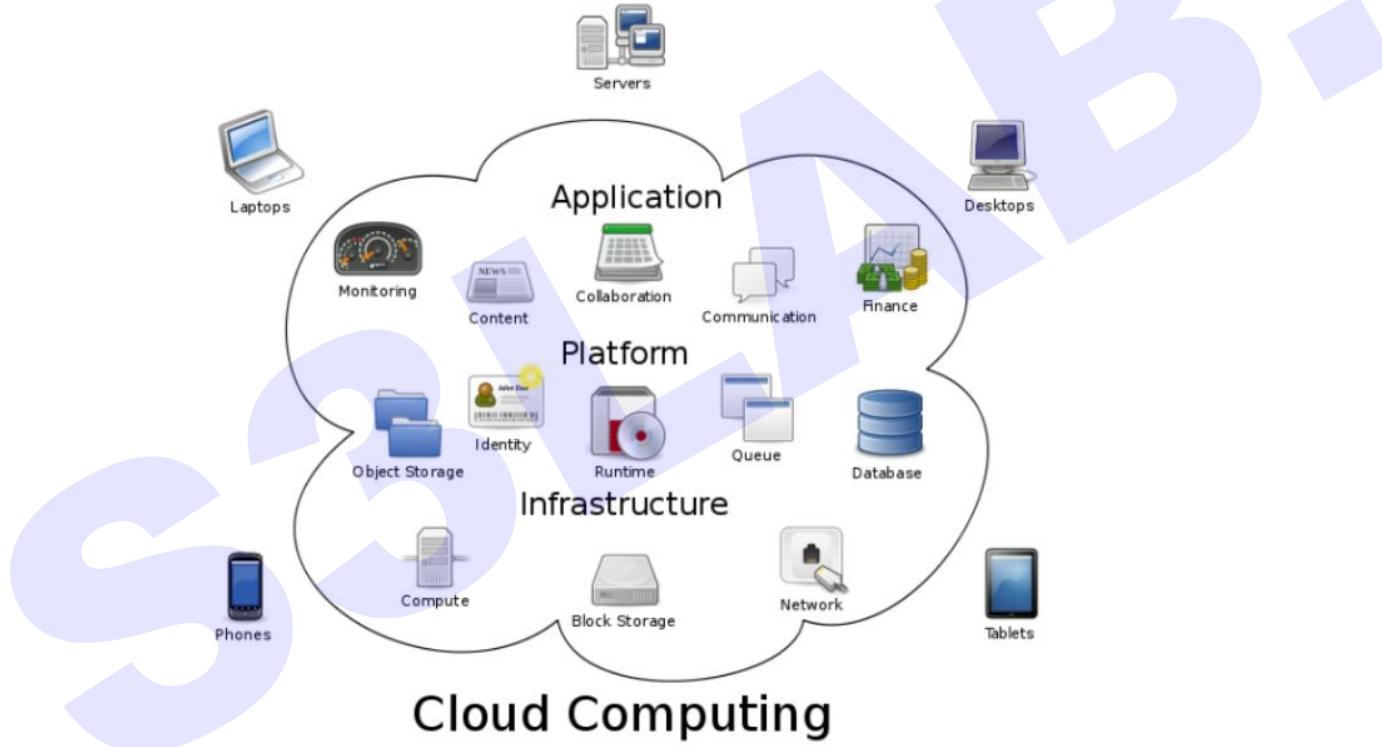


Big Data: Cloud Computing

- 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

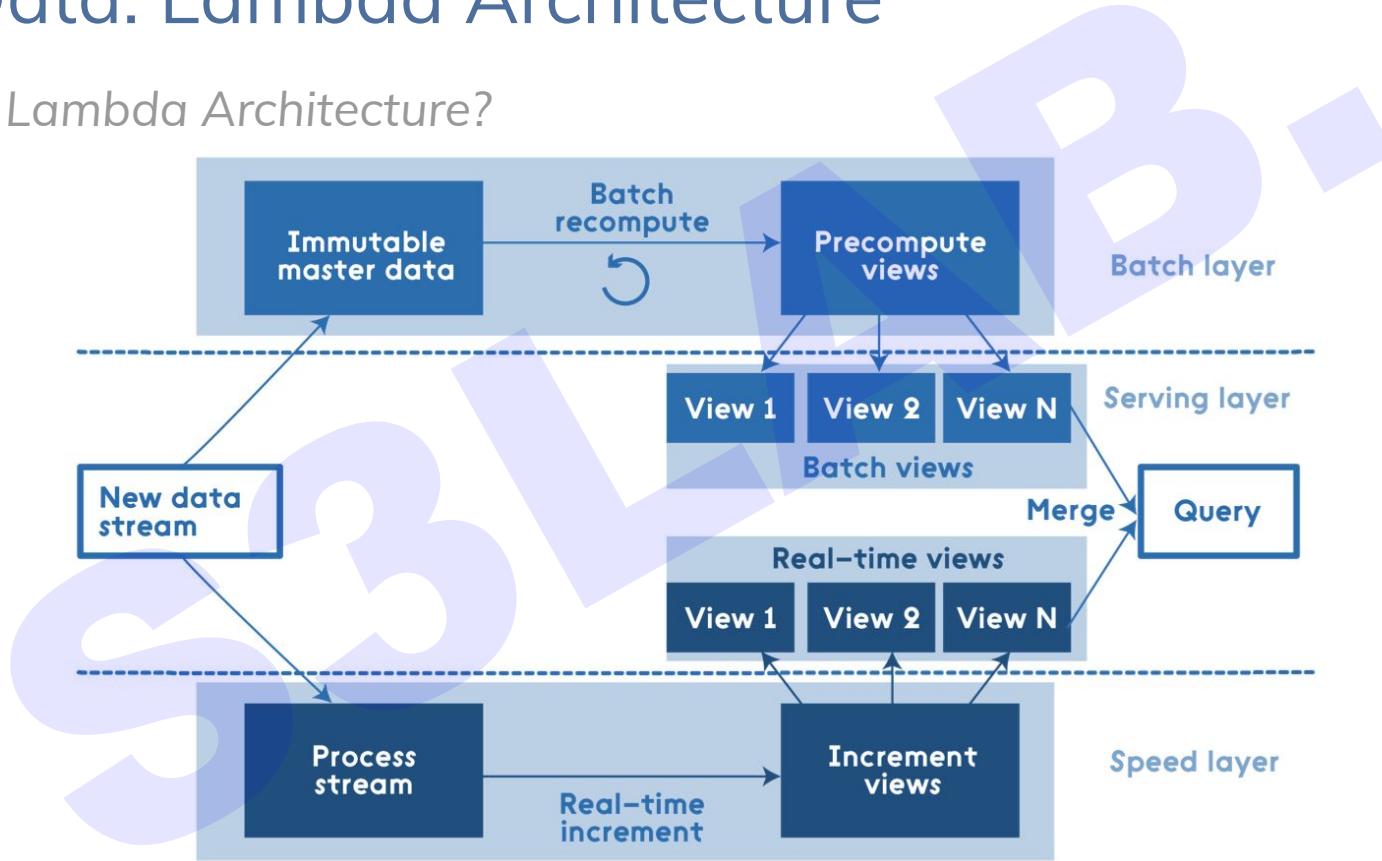
What is Lambda Architecture?

- 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

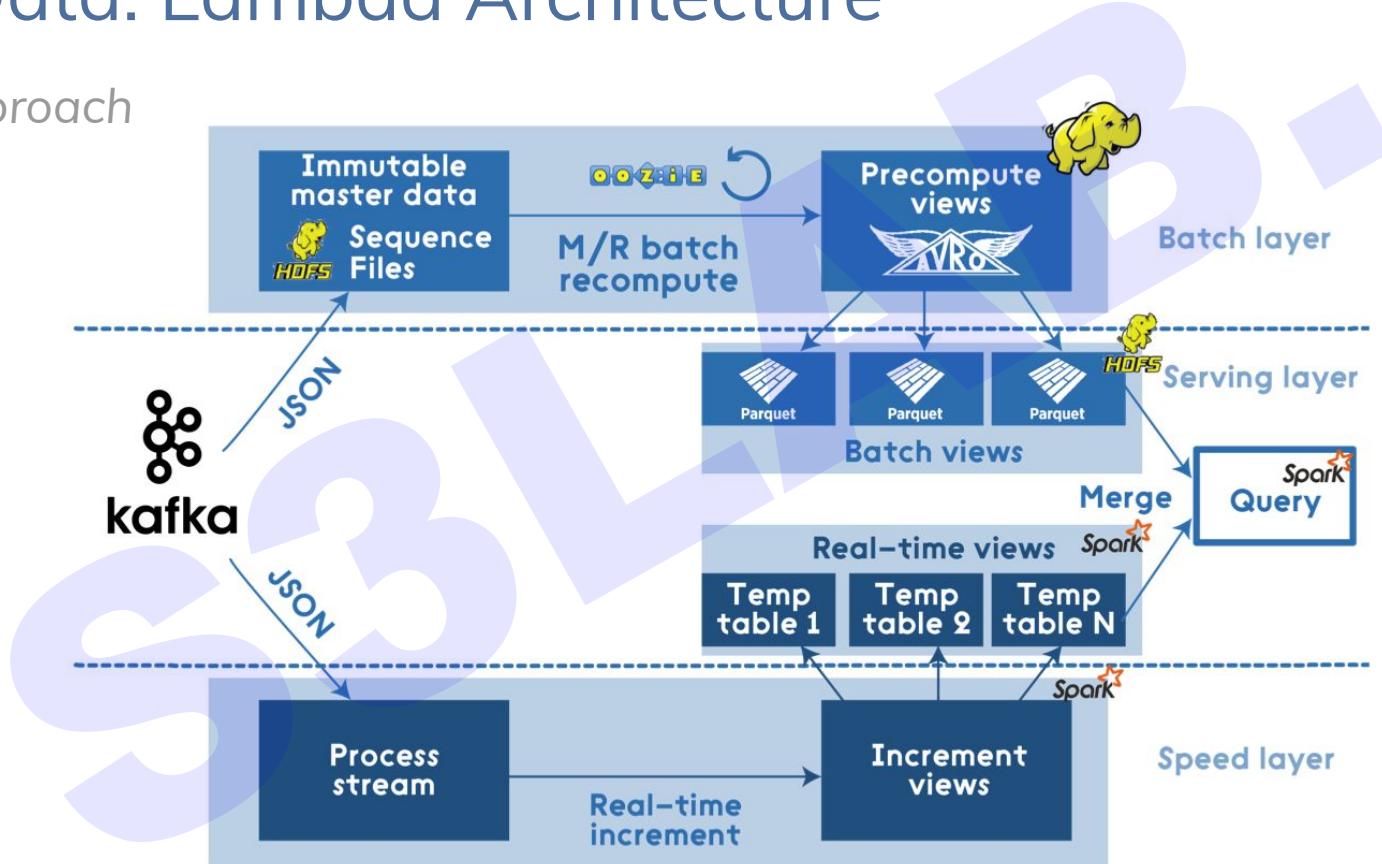
What is Lambda Architecture?





Big Data: Lambda Architecture

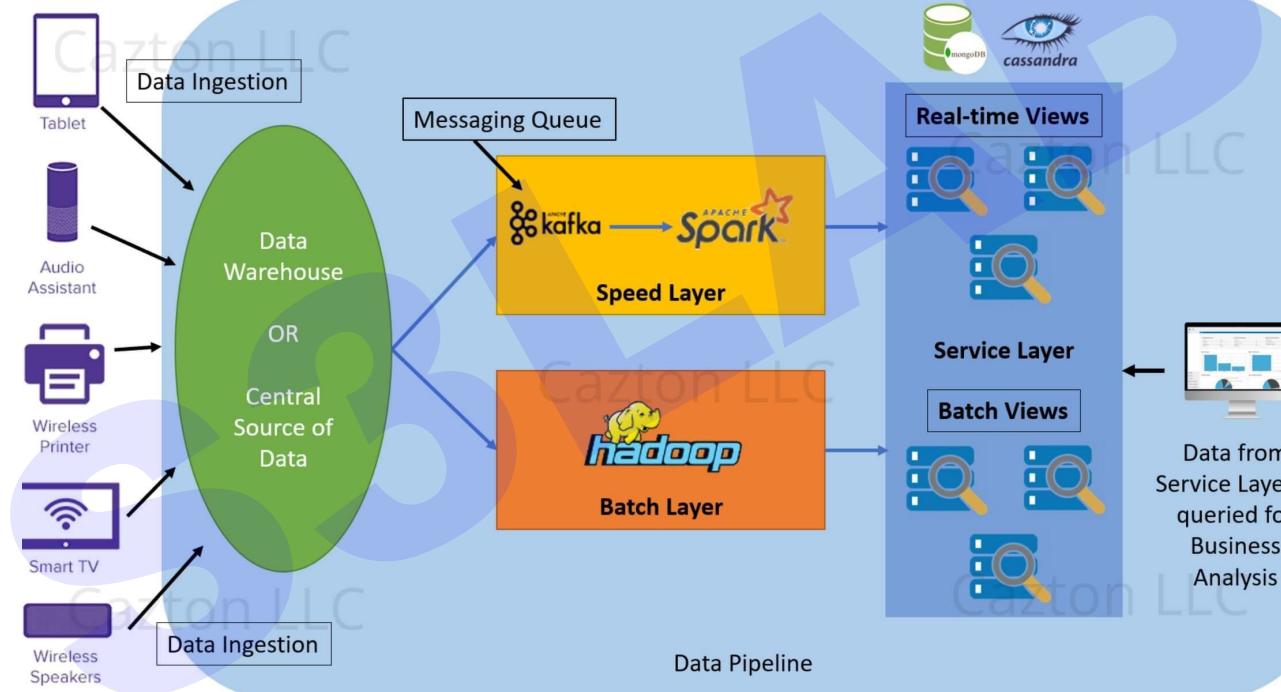
real-approach





Big Data: Lambda Architecture

real-approach





Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.