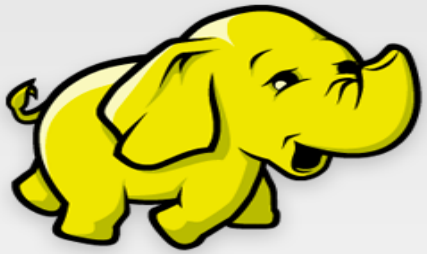
University of Rwanda

College of Business and Economics

Business Information Technology – **BIT**

Year 3 - HUYE

BIT 3233: Big Data and Social Media



Lecturer: Marc SENTWALI

TOPICS

- Part#1: Big data and Machine Learning apps (Theory &Practice)
  - Apache Spark & Python
- Part#2: Social Media (Theory &Practice)
  - Blogging

MS

Information and Introduction to Big Data

MS

Aims

1. This unit aims to introduce you to the concepts behind Big Data, the core technologies used in managing large-scale data sets, and a range of technologies for developing solutions to large-scale data analytics problems.

2. This unit is intended for students who want to understand modern large-scale data analytics systems. It covers a wide range of topics and technologies, and will prepare students to be able to build such systems as well as use them efficiently and effectively address challenges in big data management.

Not possible to cover every aspect of big data

MS

Learning outcomes

- After completing this unit, you are expected to:
  - elaborate on the important characteristics of Big Data
  - develop an appropriate storage structure for a Big Data repository
  - utilize the machine learning algorithms to manipulate Big Data

MS

What is Big Data?

- Big data is like teenage sex:
  - everyone talks about it
  - nobody really knows how to do it
  - everyone thinks everyone else is doing it
  - so everyone claims they are doing it...

--Dan Ariely, Professor at Duke University

MS

Dan Ariely

January 7, 2013 ·

Follow

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Le Bigdata,

c'est comme le sexe chez les adolescents :

tout le monde

en parle, personne ne sait vraiment

comment le faire, tout le monde

pense que tout le monde le fait, donc

tout le monde prétend le faire.

What is Big Data?

No standard definition! here is from Wikipedia:

A. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.

B. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

C. Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."

MS

Definition and Characteristics of Big Data

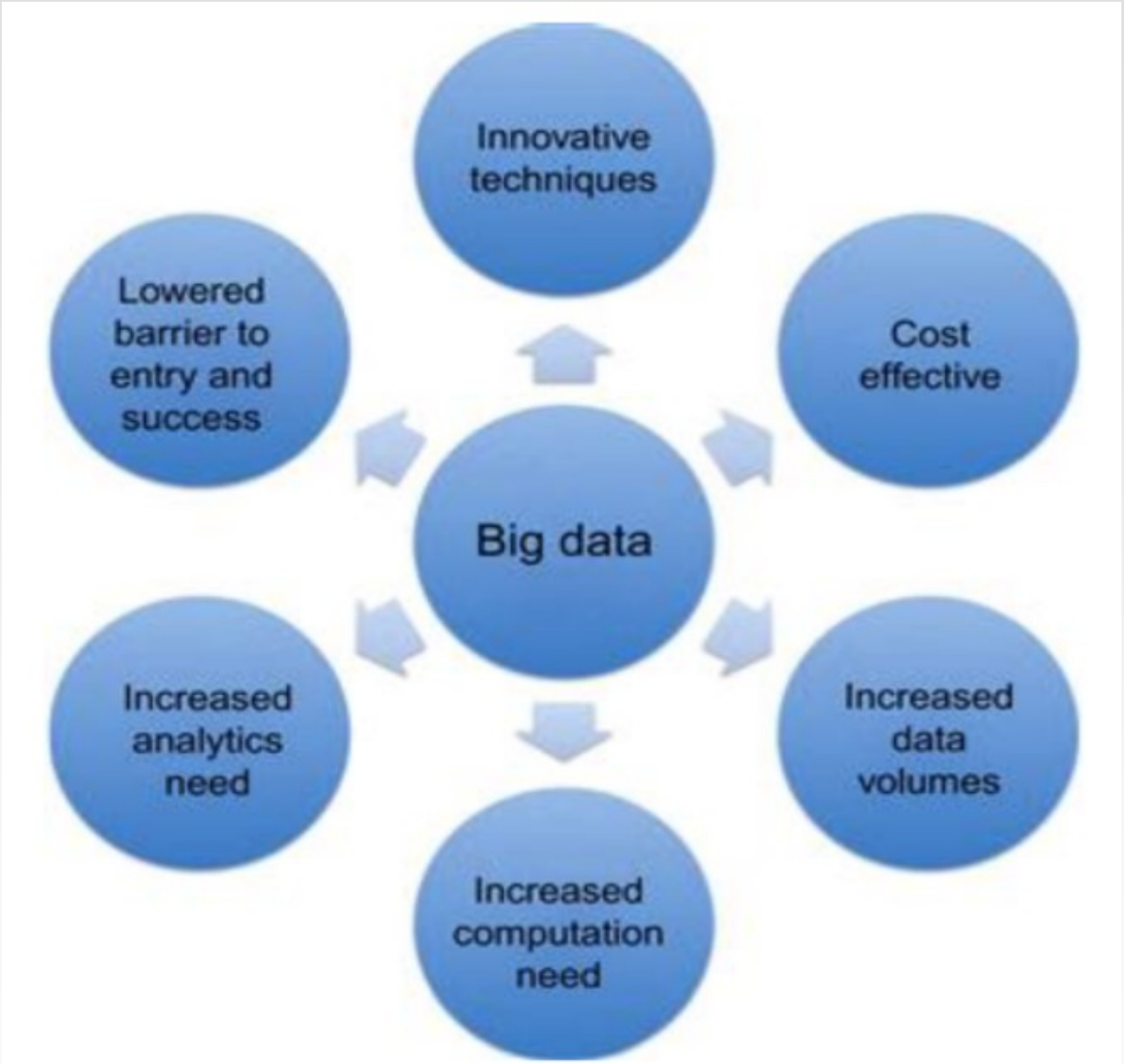
"*Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.*" -- Gartner

which was derived from:

"*While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety**. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.*" – Doug Laney

MS

What made Big Data needed?



WHAT MADE BIG DATA NEEDED?

Big Data became a necessity due to several factors and evolving trends in technology, business, and society. The following factors have contributed to the increasing need for Big Data:

1. **Data Growth:** The digital revolution has led to explosive growth in the amount of data generated. The proliferation of the Internet, social media, mobile devices, sensors, and IoT (Internet of Things) devices has resulted in an unprecedented volume of data being created and collected.

2. **Variety of Data:** Data comes in various forms, including structured data (traditional databases), semi-structured data (XML, JSON), and unstructured data (text, images, videos). Traditional data management systems were not equipped to handle this diverse range of data types.

3. **Real-time Data:** Many applications require the ability to process and analyze data in real-time or near-real-time. Traditional batch processing systems couldn't meet these demands, necessitating the development of real-time data processing solutions.
4. **Value in Data:** Organizations recognized the potential value hidden within their data. Extracting insights from data can lead to better decision-making, improved customer experiences, and competitive advantages. Businesses have increasingly sought to harness this value.

6. **Competition and Innovation:** In a rapidly changing business landscape, companies have turned to data analytics to gain a competitive edge. Analyzing Big Data can lead to innovation, the discovery of market trends, and the development of new products and services.
7. **Improved Decision-Making:** Big Data analytics provides the means to make data-driven decisions. Whether in healthcare, finance, marketing, or any other field, decision-makers rely on data analysis to make informed choices and optimize processes.

7. **Customer Insights:** Understanding customer behavior and preferences is crucial for businesses. Big Data analytics allows organizations to gain a deeper understanding of their customers, tailor their offerings, and provide a more personalized experience.
8. **Scientific Discovery:** In fields like genomics, astronomy, climate research, and particle physics, Big Data plays a pivotal role in scientific discovery. Researchers use Big Data to analyze and derive insights from large datasets, leading to breakthroughs and advancements in these domains.
9. **Fraud Detection and Security:** Big Data analytics is essential for identifying patterns of fraudulent activities in areas such as credit card transactions, cybersecurity, and insurance claims. It helps in proactive threat detection and prevention.

10. **Healthcare and Life Sciences:** In healthcare, Big Data aids in disease diagnosis, drug discovery, and patient care optimization. Large volumes of patient data, genetic information, and medical records are analyzed to improve healthcare outcomes.
11. **Urban Planning and Smart Cities:** Big Data is used to improve city infrastructure, transportation, and resource allocation in smart city initiatives. Data from sensors and IoT devices help city planners make informed decisions.
12. **Social and Environmental Impact:** Big Data can be leveraged to address social and environmental issues. For example, it can be used to track and respond to environmental changes, natural disasters, and humanitarian crises.

In brief, the need for Big Data has arisen from the massive growth in data volume, variety, and velocity, as well as the potential to extract valuable insights and drive innovation. It has become an essential tool for businesses, scientific research, healthcare, and various other sectors to address complex challenges and make data-informed decisions.

MS

Key Computing Resources for Big Data

- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

MS

Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

MS

Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

MS

Who is generating Big Data?

Social

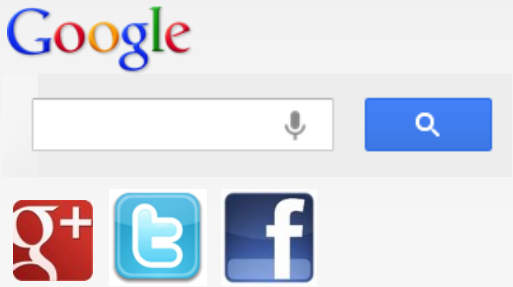
User Tracking & Engagement

Homeland Security

eCommerce

Financial Services

Real Time Search



MS

WHO IS GENERATING BIG DATA?

Big Data is generated by a wide range of sources and entities across various sectors and industries. Here are some of the key generators of Big Data:

1. **Internet Users**: Individuals generate Big Data through their online activities, such as social media interactions, web searches, online shopping, and content consumption. User-generated content, like text, images, videos, and comments, contributes to the vast volume of online data.
2. **Social Media Platforms:** Social media platforms like Facebook, Twitter, Instagram, and LinkedIn generate enormous amounts of data every second as users post updates, share content, and engage with others. This data includes posts, comments, likes, shares, and user profiles.
3. **E-commerce and Retail:** Online retailers and traditional retail businesses that use electronic point-of-sale systems generate extensive data on customer transactions, preferences, and buying behavior. This data is used for market analysis, inventory management, and personalized recommendations.

4. **IoT Devices:** The Internet of Things (IoT) includes a wide array of devices like smart thermostats, wearables, connected vehicles, and industrial sensors. These devices constantly collect data on temperature, location, movement, and other variables, generating vast streams of real-time data.
5. **Healthcare:** Healthcare facilities generate Big Data through electronic health records (EHRs), medical imaging, patient monitoring devices, and genetic sequencing. This data is used for patient care, research, and healthcare management.
6. **Financial Services:** Financial institutions, including banks and credit card companies, generate data through financial transactions, fraud detection systems, and market trading. Analyzing this data is essential for fraud prevention, risk assessment, and investment strategies.

7. **Transportation and Logistics:** Transportation companies collect data from GPS trackers, telematics, and sensors on vehicles, ships, and airplanes. This data helps optimize routes, reduce fuel consumption, and enhance supply chain management.
8. **Energy and Utilities:** The energy sector generates data from smart meters, sensors, and monitoring systems in power plants and electrical grids. This data aids in efficient energy distribution, demand forecasting, and maintenance.
9. **Manufacturing and Industry:** Manufacturing facilities generate data through sensors on machines, production lines, and quality control systems. This data helps improve production processes, reduce downtime, and enhance quality control.

10. **Agriculture:** Modern agriculture relies on data from sensors and satellite imagery to monitor soil conditions, crop health, and weather patterns. This information is used to optimize farming practices and increase crop yields.
11. **Research and Academia:** Research institutions, universities, and laboratories generate Big Data in scientific experiments, simulations, and research studies. Data is collected and analyzed to make discoveries and advance knowledge.

MS

12. **Government and Public Services:** Government agencies collect data on demographics, public health, transportation, and various other domains. This data is used for policy-making, urban planning, and disaster response.
13. **Entertainment and Media:** Streaming services, like Netflix and Spotify, generate data on user preferences and content consumption. This data is used to recommend content and create personalized experiences.

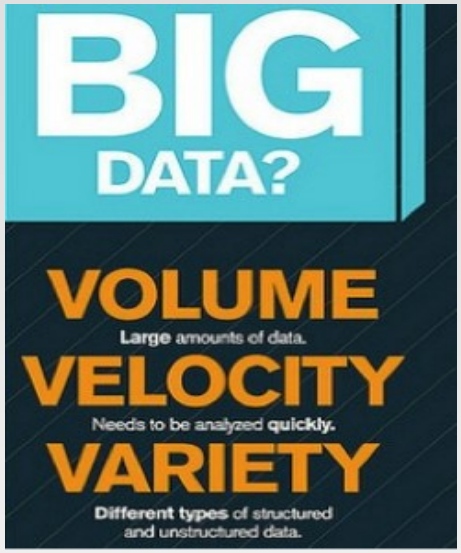
14. **Environmental Monitoring:** Environmental agencies and organizations collect data on weather, air quality, water quality, and ecosystems. This information is critical for understanding and addressing environmental issues.

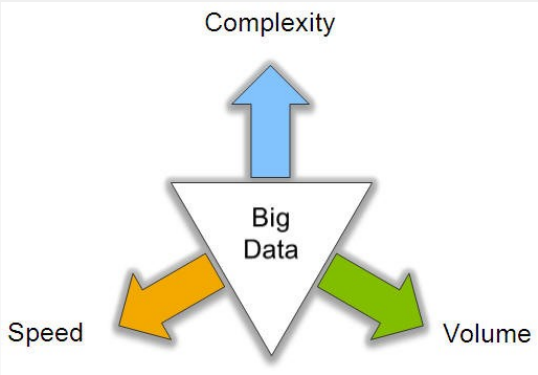
15. **Social and Civic Engagement:** Crowdsourced data from citizen reporting, surveys, and social initiatives contribute to data on various social and civic issues, enabling data-driven decision-making.

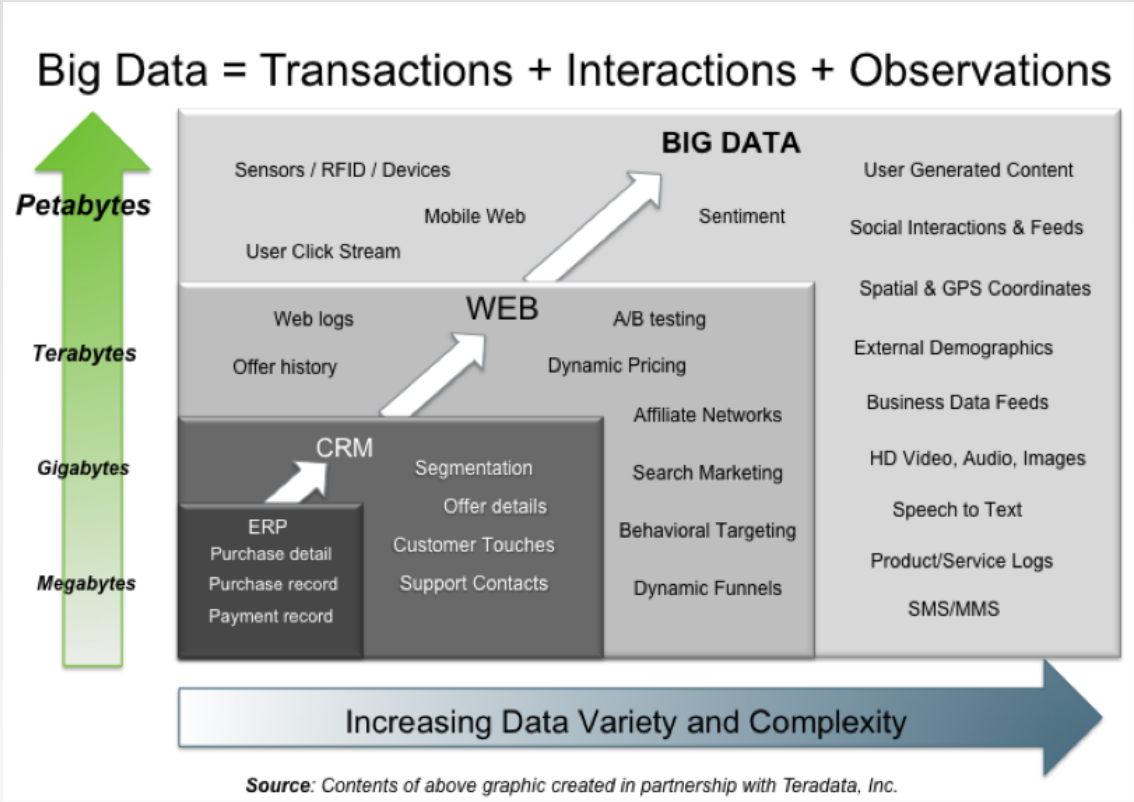
These are just a few examples of the many entities and sources that generate Big Data. The rapid growth of digital technologies and the increasing connectivity of devices continue to expand the scope and diversity of data generation across various sectors.

MS

Big Data Characteristics: 3V





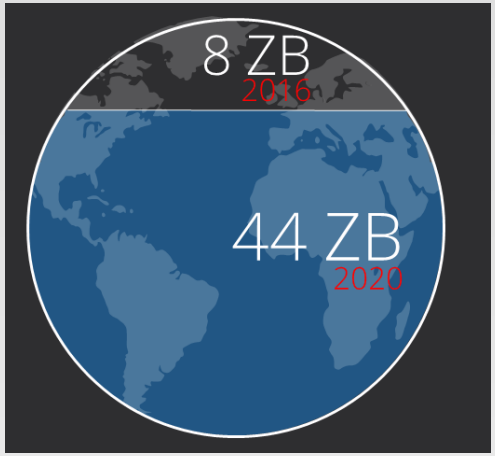


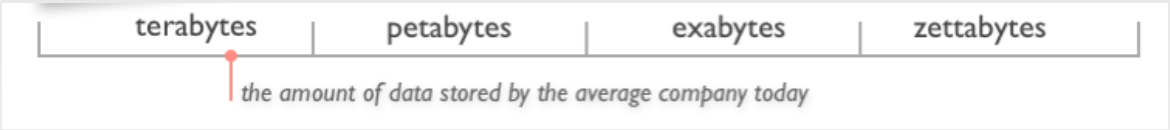
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

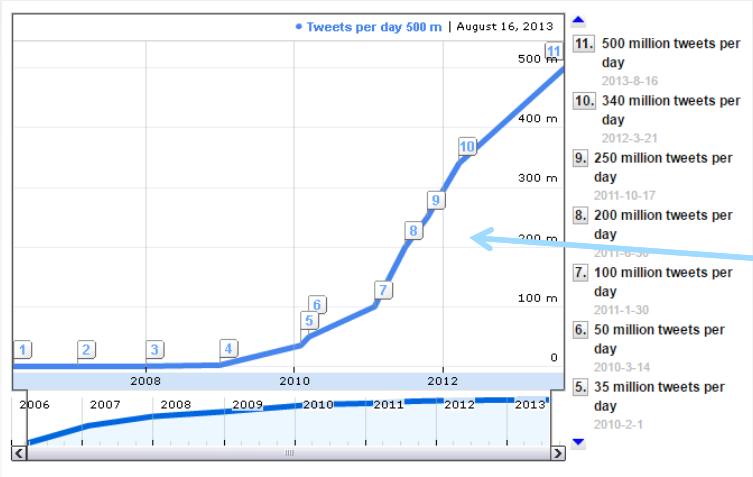
MS

Volume (Scale)

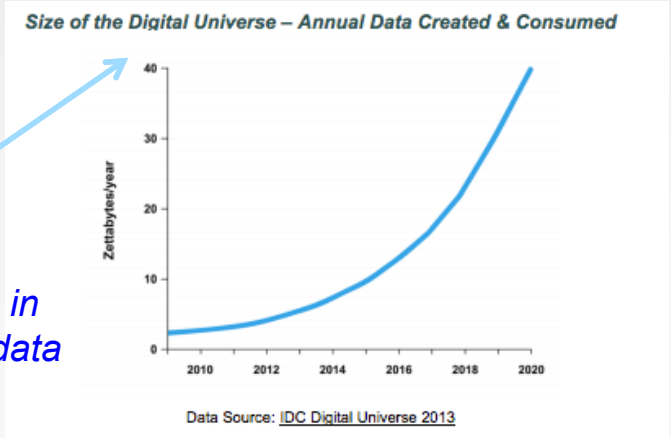
- Data Volume
  - Growth 40% per year
  - From 8 zettabytes (2016) to 44zb (2020)
- Data volume is increasing exponentially







Exponential increase in collected/generated data



MS

Google

Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

INTERNET ARCHIVE

400B pages, 10+ PB (2/2014)

YAHOO!

Hadoop: 365 PB, 330K nodes (6/2014)

JPMorganChase

150 PB on 50k+ servers running 15k apps (6/2011)

ebay

Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

300 PB data in Hive + 600 TB/day (4/2014)

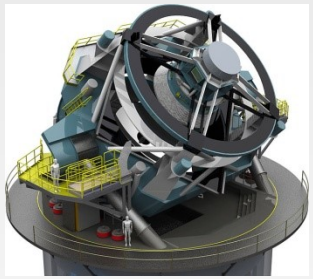
facebook

LHC: ~15 PB a year

CERN

amazon web services™

S3: 2T objects, 1.1M request/second (4/2013)



LSST: 6-10 PB a year (~2020)

SKA: 0.3 – 1.5 EB per year (~2020)

MS

How much data?

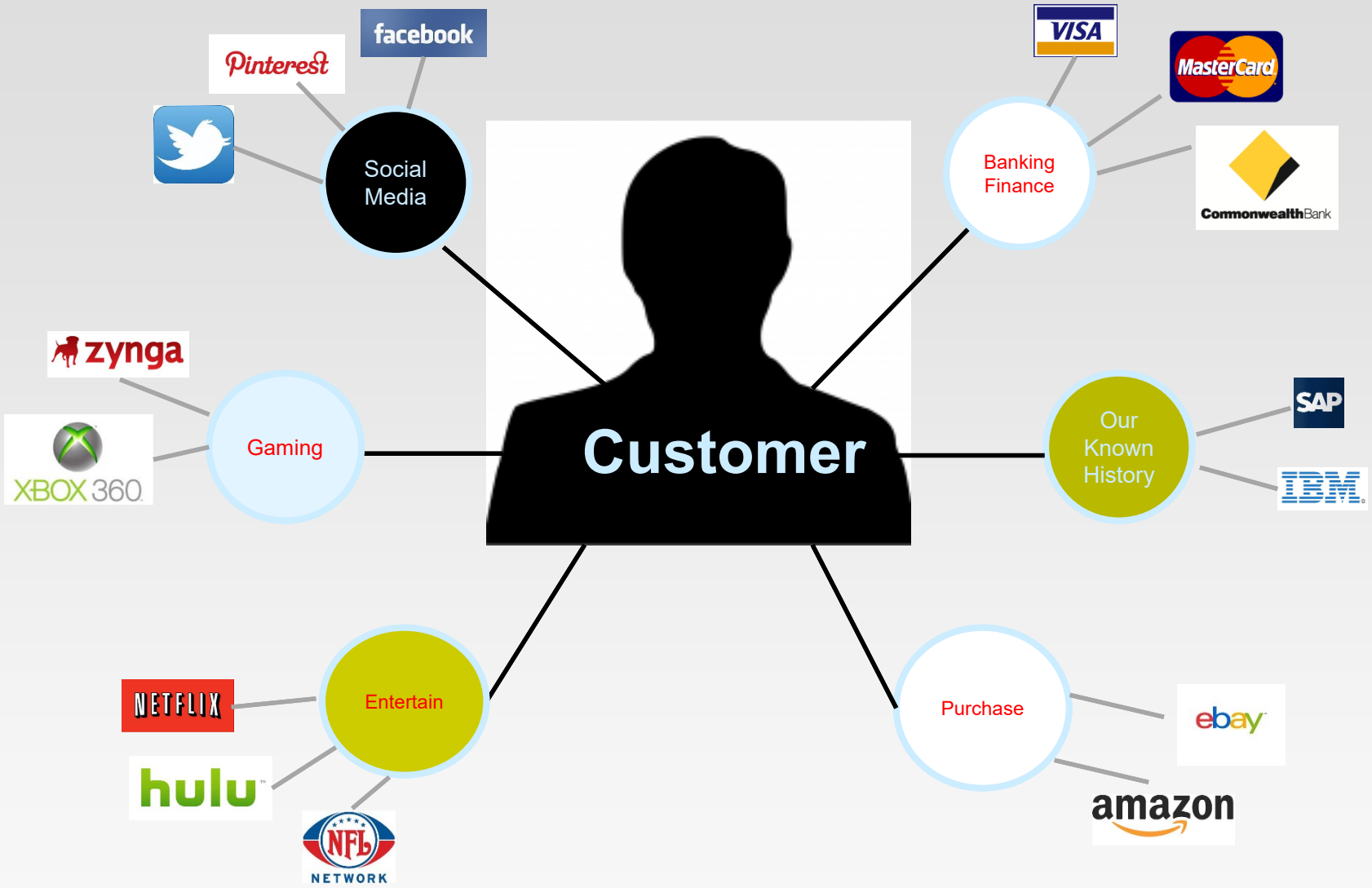
Variety (Complexity)

- Different Types:
  - Relational Data (Tables/Transaction/Legacy Data)
  - Text Data (Web)
  - 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
- Different Sources:
  - Movie reviews from IMDB and Rotten Tomatoes
  - Product reviews from different provider websites

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

MS

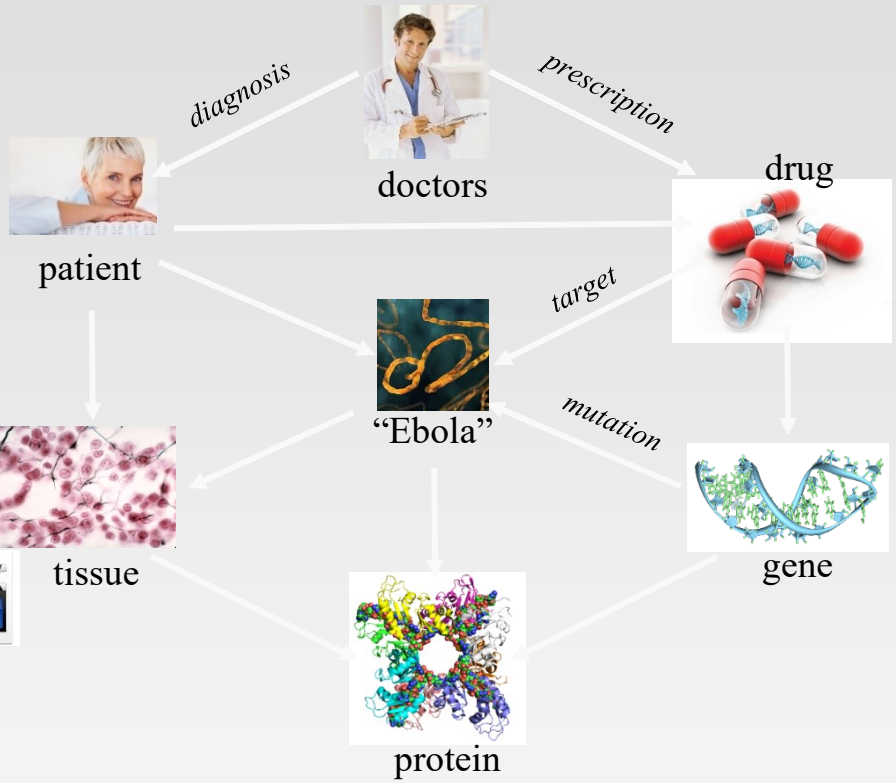
A Single View to the Customer



MS

A Global View of Linked Big Data

Diversified social network



Heterogeneous information network

MS

Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - **E-Promotions:** Based 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
  - **Disaster management and response**

MS

TASK#1

1. What are other available **V**s for Big Data Characteristics
2. Why big data is nowadays sounded (everyone talks about it)
3. Why are structured data 20% compared to unstructured data 80%?

MS

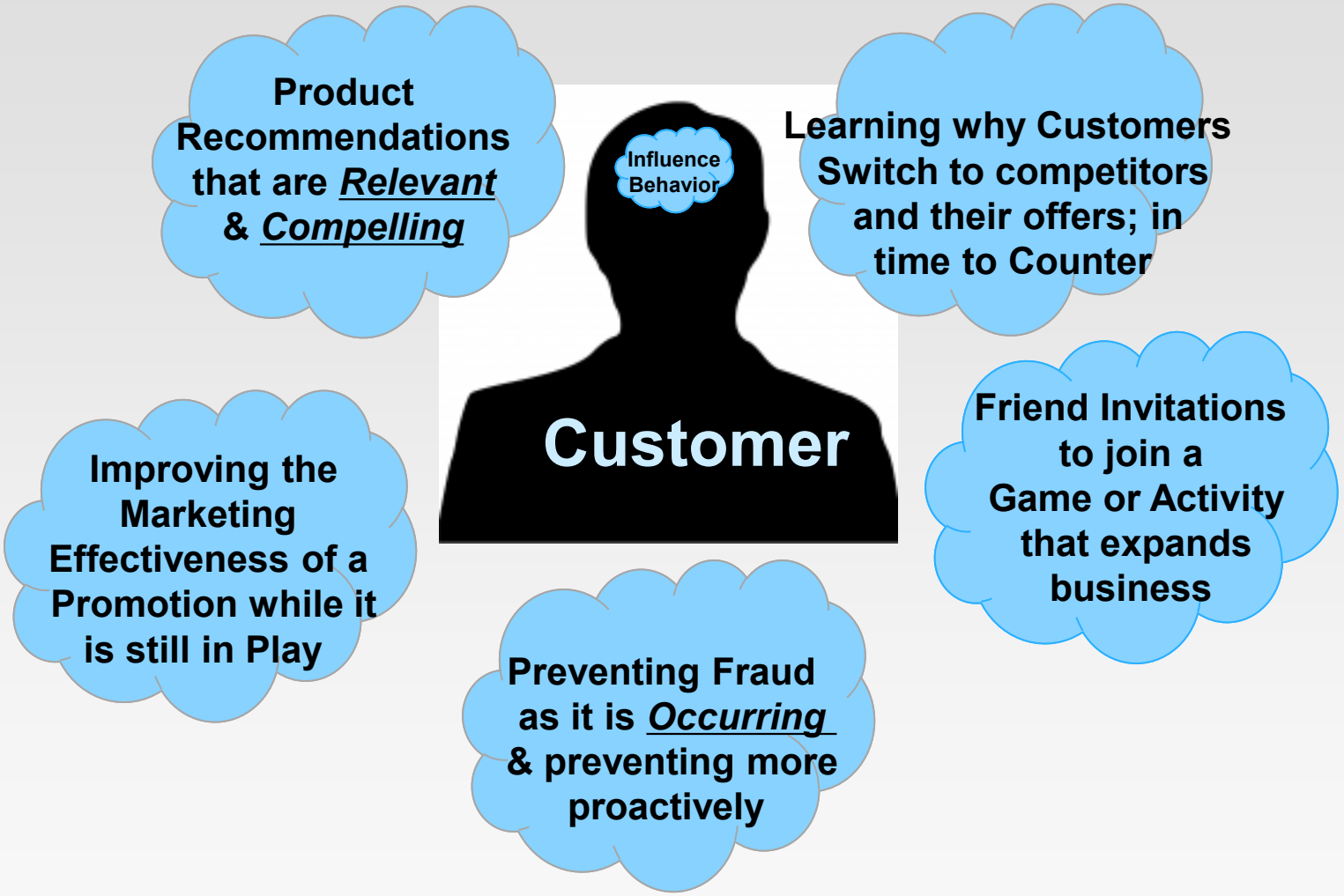
Discuss

- Is it true that a “Huge volume” of data can create “Huge business opportunities?
  - If yes, HOW?
  - If No, WHY?

  Note: Support your ideas/views

MS

Real-Time Analytics/Decision Requirement



MS

Extended Big Data Characteristics: 6V

1. **Volume**: In a big data environment, the amounts of data collected and processed are much larger than those stored in typical relational databases.
2. **Variety**: Big data consists of a rich variety of data types.
3. **Velocity**: Big data arrives to the organization at high speeds and from multiple sources simultaneously.
4. **Veracity**: Data quality issues are particularly challenging in a big data context.
5. **Visibility/Visualization**: After big data being processed, we need a way of presenting the data in a manner that's readable and accessible.
6. **Value**: Ultimately, big data is meaningless if it does not provide value toward some meaningful goal.

MS

Veracity (Quality & Trust)

- *Data = quantity + quality*
- When we talk about big data, we typically mean its quantity:
  - What capacity of a system provides to cope with the sheer size of the data?
  - Is a query feasible on big data within our available resources?
  - How can we make our queries tractable on big data?
  - . . .
- Can we trust the answers to our queries?
  - Dirty data routinely lead to misleading financial reports, strategic business planning decision $\Rightarrow$ loss of revenue, credibility and customers, disastrous consequences
- *The study of data quality is as important as data quantity*

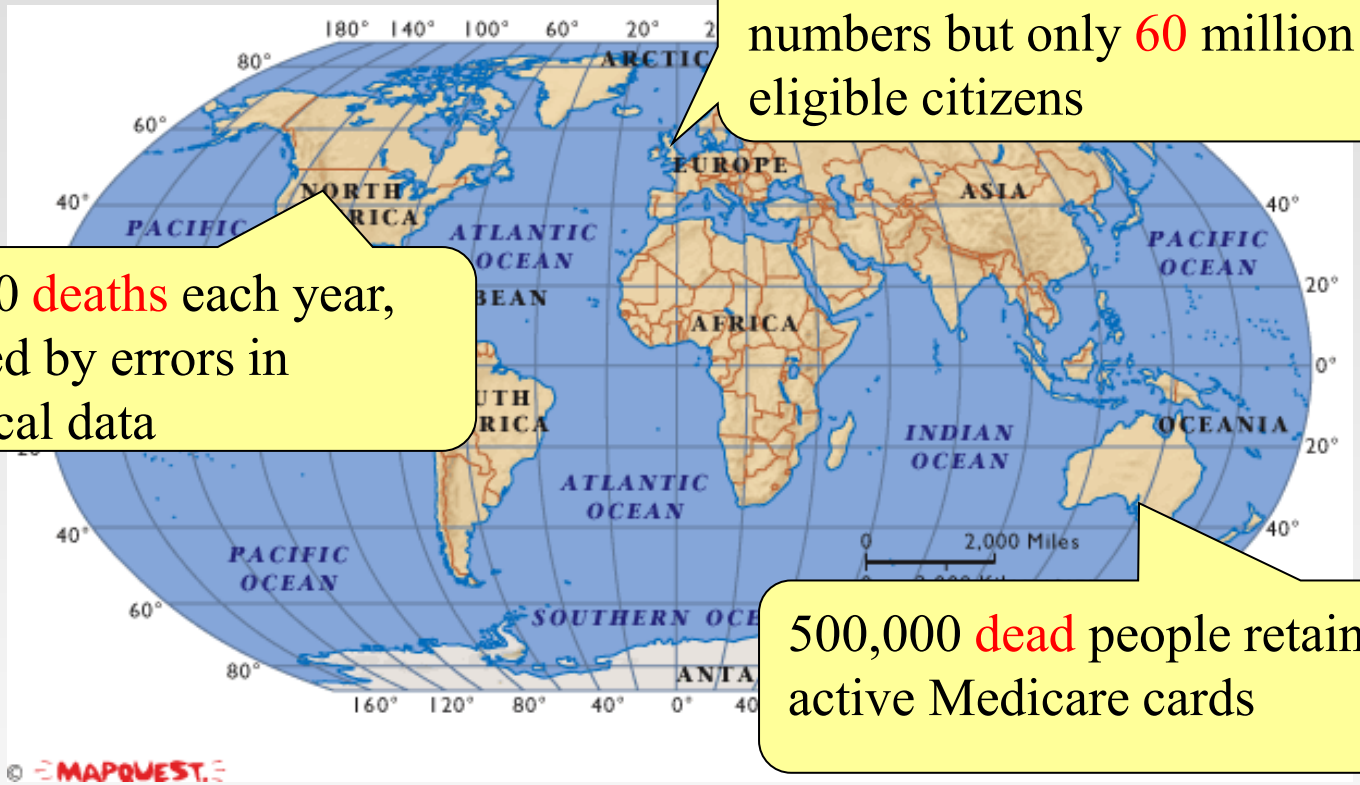
MS

Data in real-life is often dirty

81 million National Insurance numbers but only 60 million eligible citizens

98000 deaths each year, caused by errors in medical data

500,000 dead people retain active Medicare cards



MS

Visibility/Visualization

- Visible to the process of big data management
- Big Data – visibility = Black Hole?

A visualization of Divvy bike rides across Chicago

- Big data visualization tools:

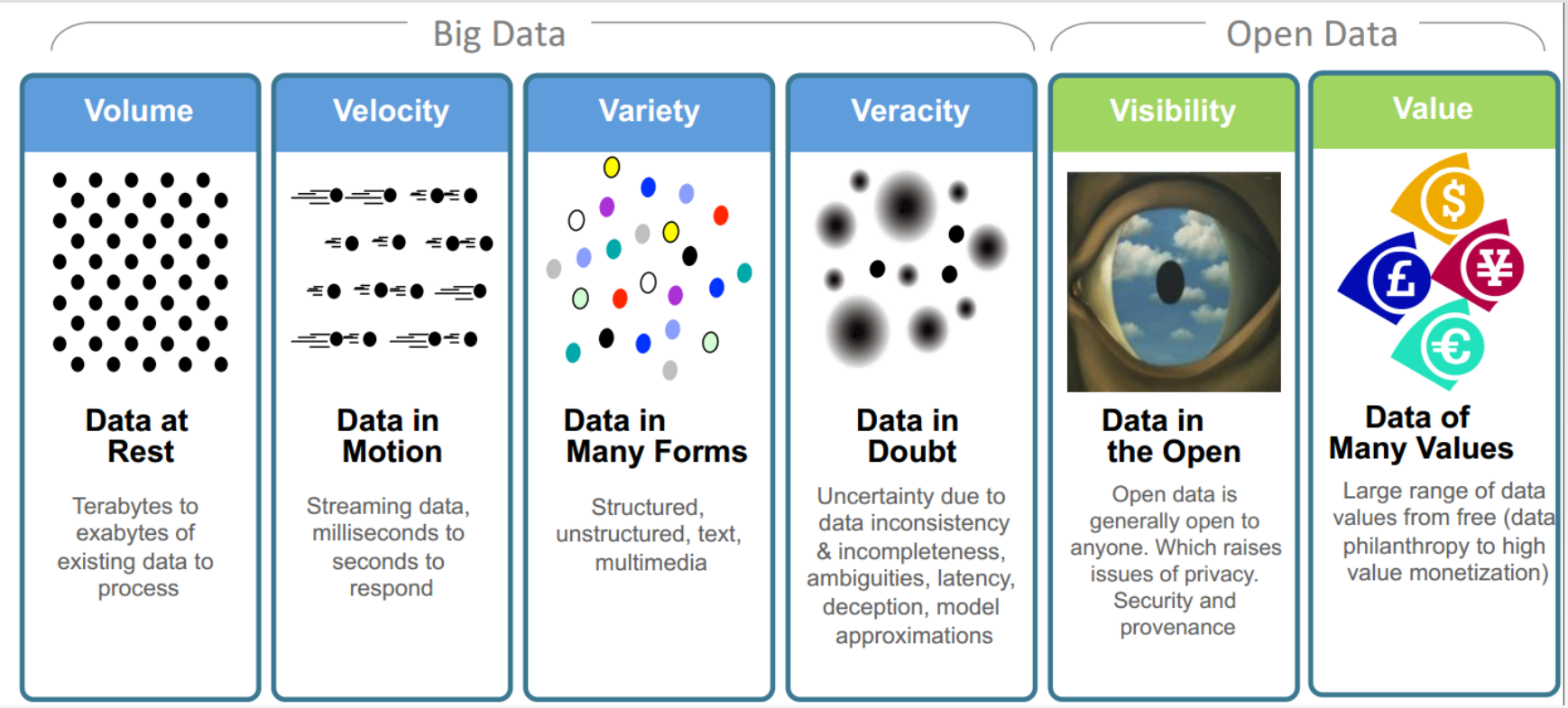
MS

Value

- Big data is meaningless if it does not provide value toward some meaningful goal

MS

Big Data: 6V in Summary



Transforming Energy and Utilities through Big Data & Analytics. By Anders Quitzau@IBM

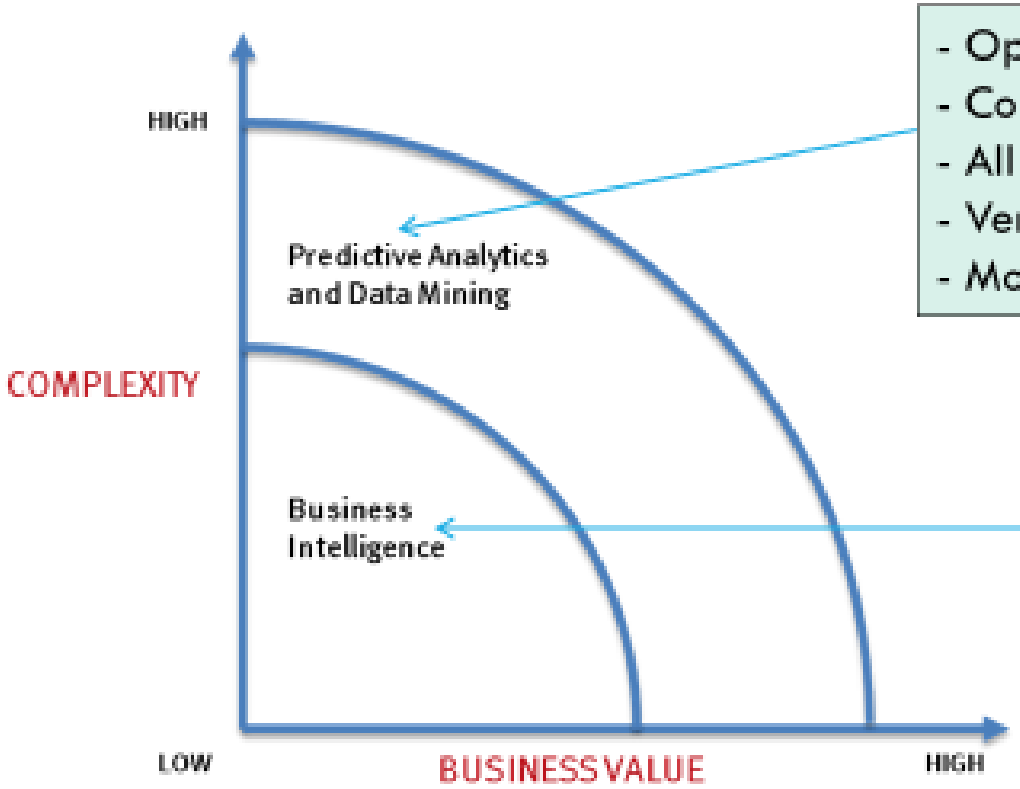
MS

Other V's

- Variability
  - Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
- Viscosity
  - This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
- Virality
  - Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.
- Volatility
  - Big data volatility refers to how long is data valid and how long should it be stored. You need to determine at what point is data no longer relevant to the current analysis.
- More V's in the future …

MS

WHAT'S DRIVING BIG DATA

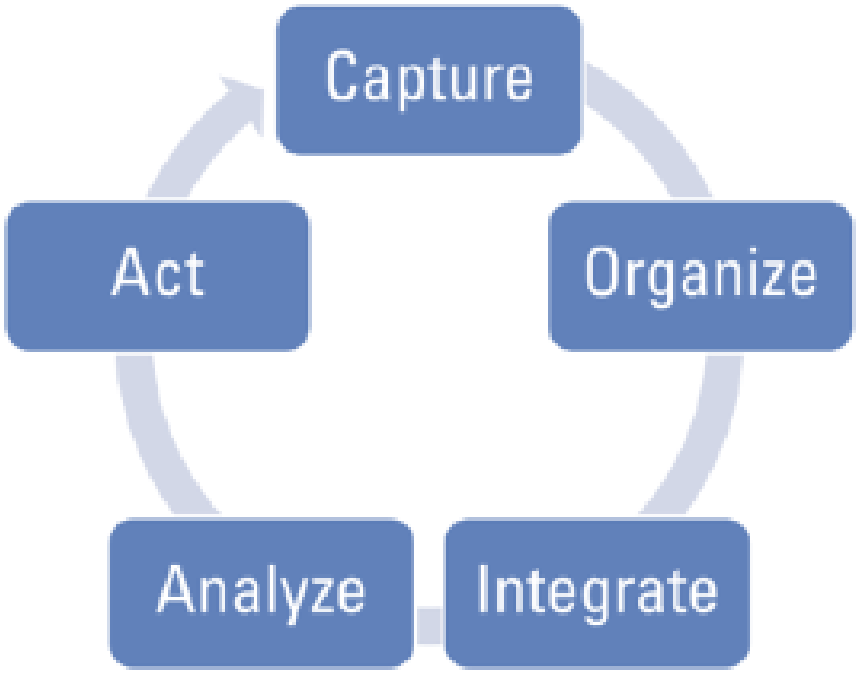


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

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

MS

THE CYCLE OF BIG DATA MANAGEMENT.



Data must first be captured, and then organized and integrated.

After this phase is successfully implemented, data can be analyzed based on the problem being Addressed

Finally, management takes action based on the outcome of that analysis.

Variety: Types of Big Data

- Big Data could be of three types:
  1. Structured
  2. Semi-Structured
  3. Unstructured

MS

1. Structured

- The data that can be stored and processed in a fixed format is called as Structured Data. Data stored in a relational database management system (RDBMS) is one example of 'structured' data. It is easy to process structured data as it has a fixed schema. Structured Query Language (SQL) is often used to manage such kind of Data.

Examples of Structured Data

An 'Employee' table in a database is an example of Structured Data

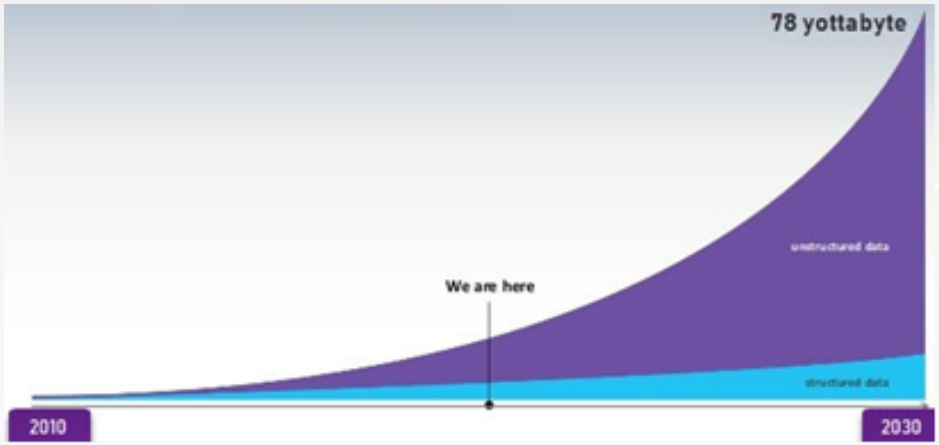
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Joseph	Male	Finance	650000
3398	Jane	Female	Admin	650000
7465	Mary	Male	Admin	500000
7500	Peter	Male	Finance	500000
7699	Joly	Female	Finance	550000

2. Semi-Structured

- Semi-Structured Data is a type of data which does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze. XML files or JSON documents are examples of semi-structured data.

Examples of Semi-structured Data, Personal data stored in an XML file:

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

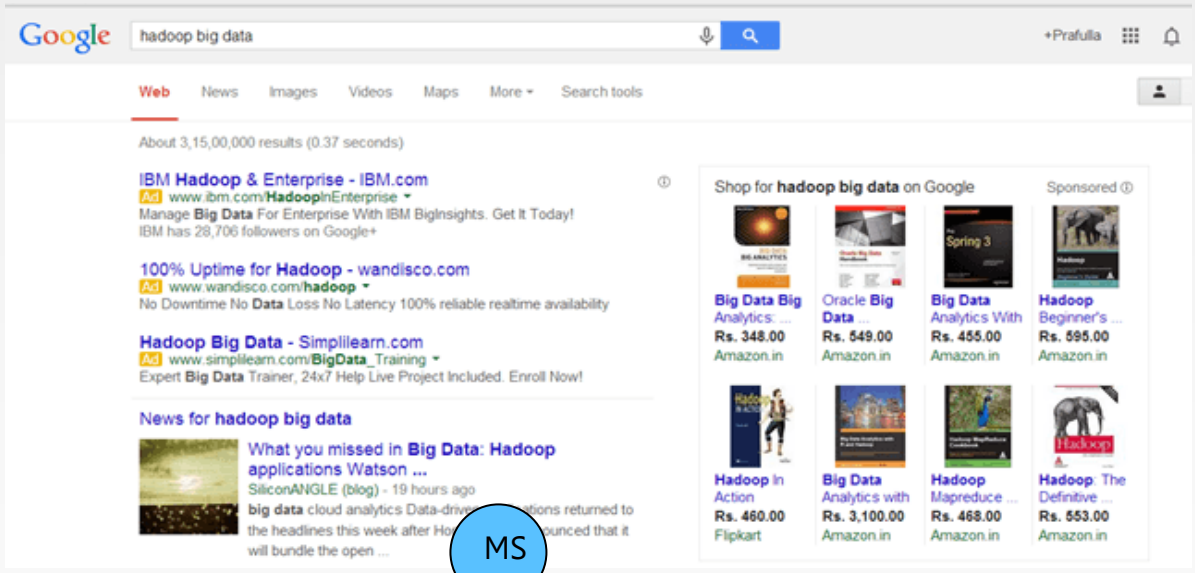


3. Unstructured

- The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data. 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.

Examples of Un-structured Data

The output returned by 'Google Search'



Cloud Computing and Big Data

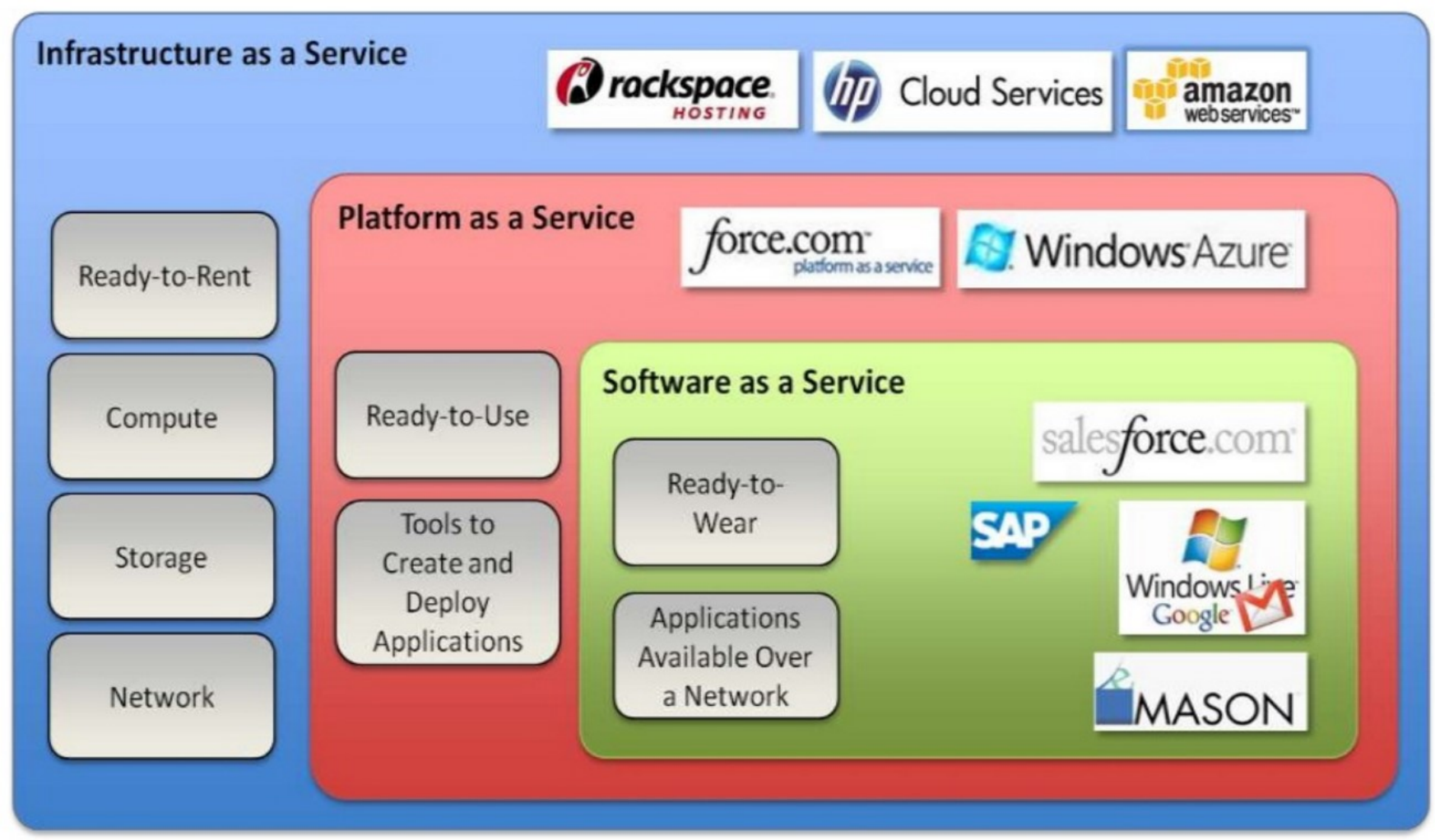
- Cloud Computing provides the infrastructure, resources, and services needed to store, process, and analyze Big Data efficiently.
- The scalability, cost-efficiency, and accessibility of cloud platforms have made them a preferred choice for many organizations looking to leverage Big Data for insights and decision-making.

Cloud Computing

- The buzzword before “Big Data”
- Cloud Computing is a general term used to describe a new class of network based computing that takes place over the Internet
  - A collection/group of integrated and networked hardware, software and Internet infrastructure (called a platform).
  - Using the Internet for communication and transport provides hardware, software and networking services to clients
  - These platforms hide the complexity and details of the underlying infrastructure from users and applications by providing very simple graphical interface or API
- A technical point of view
  - Internet-based computing (i.e., computers attached to network)
- A business-model point of view
  - Pay-as-you-go (i.e., rental)

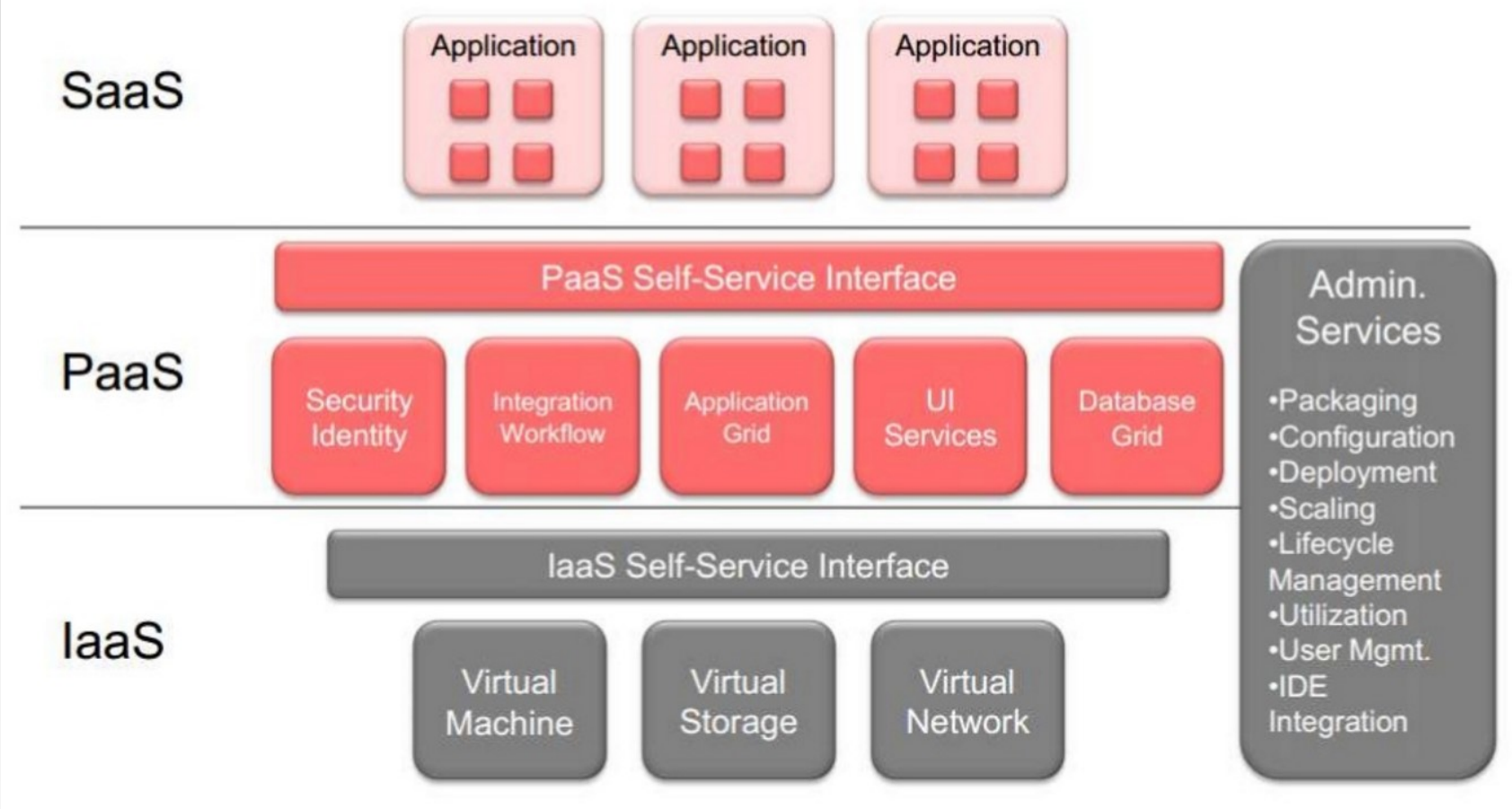
MS

Cloud Computing Architecture



MS

Cloud Computing Services



MS

Cloud Computing Services

- Infrastructure as a service (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, Amazon S3
- Platform as a Service (PaaS)
  - Offering a development platform on the cloud.
  - Google's Application Engine, Microsofts Azure
- Software as a service (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Googles gmail and Microsofts hotmail, Google docs

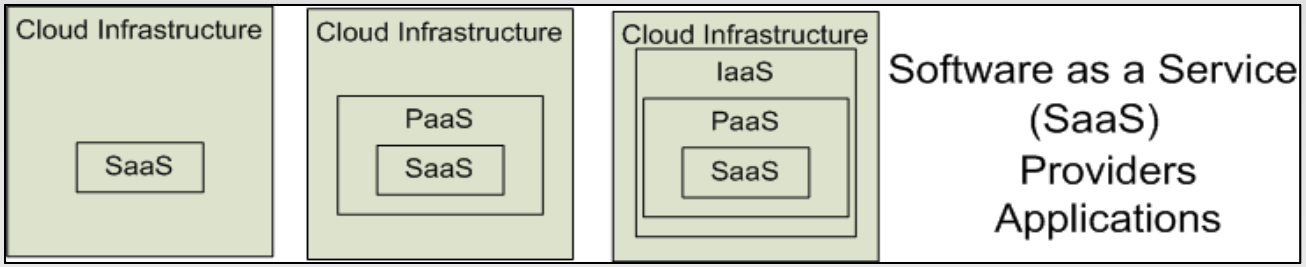
MS

Cloud Services

Software as a Service (SaaS) | Platform as a Service (PaaS) | Infrastructure as a Service (IaaS)

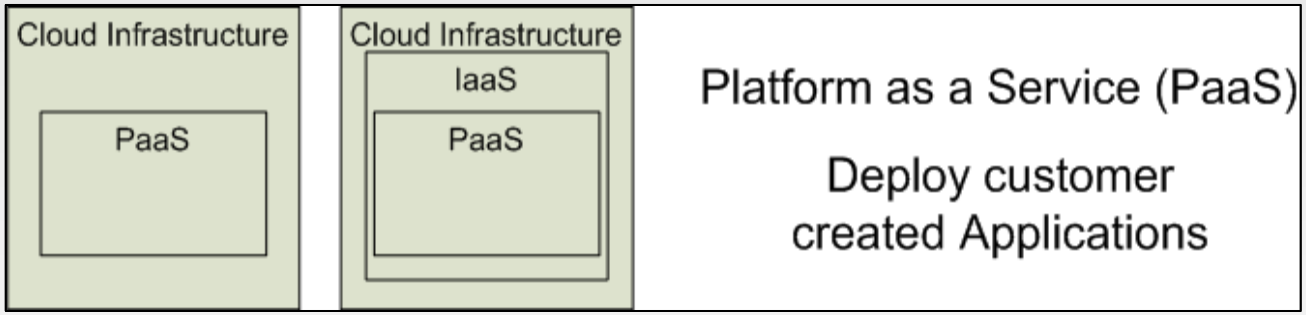
SalesForce CRM

LotusLive



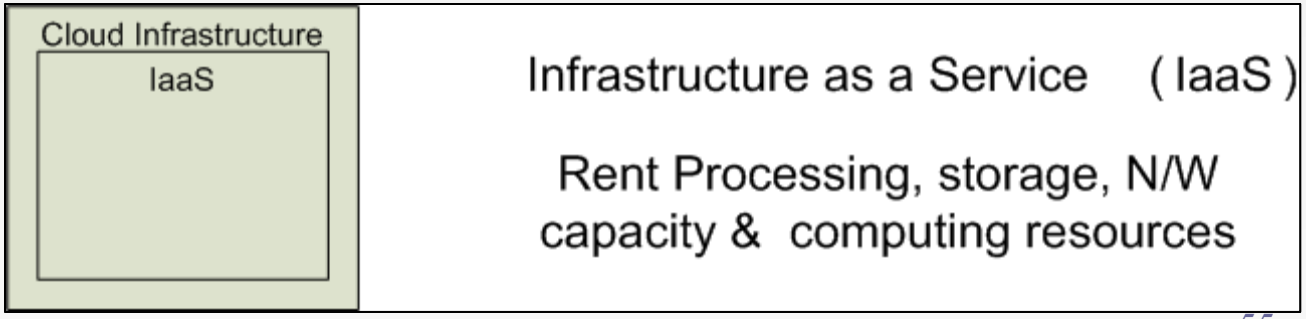
Google App

Windows Azure
The Future Made Familiar



amazon web services

rackspace HOSTING



MS

Why Study Big Data Technologies?

- The hottest topic in both research and industry
- Highly demanded in real world
- A promising future career
  - Research and development of big data systems:

    distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:

    social marketing, healthcare, …
  - Data analysis: to get values out of big data

    discovering and applying patterns, predicative analysis, business intelligence, privacy and security, …

MS

Why study Big Data Technologies (Cont..)

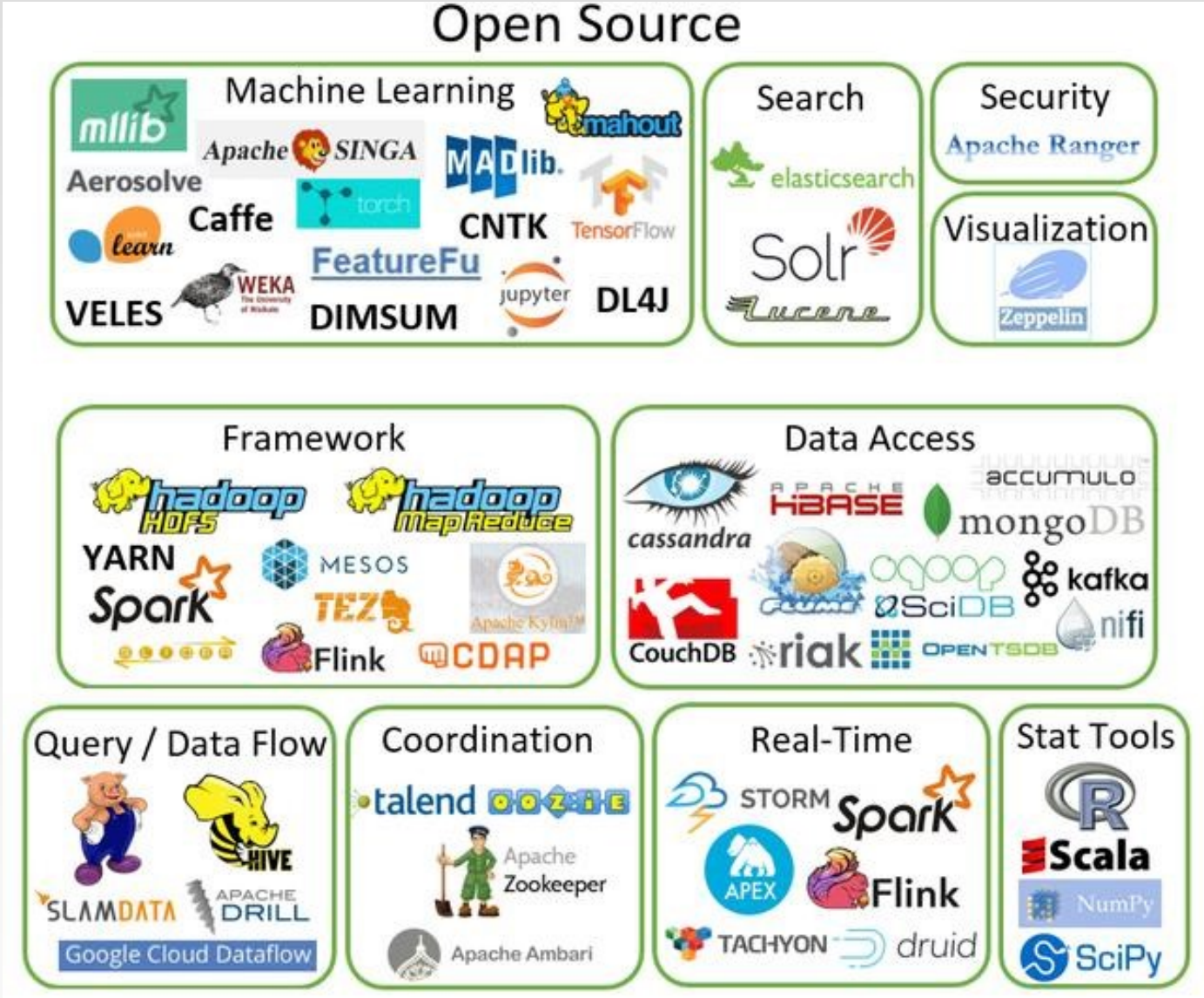
- Studying Big Data technologies not only offers career prospects but also provides an opportunity to engage in meaningful work, drive innovation, and address complex challenges.
- It equips you with skills that are in high demand across industries and sets you on a path to a rewarding and impactful career.

Task#2

- **Explain:**

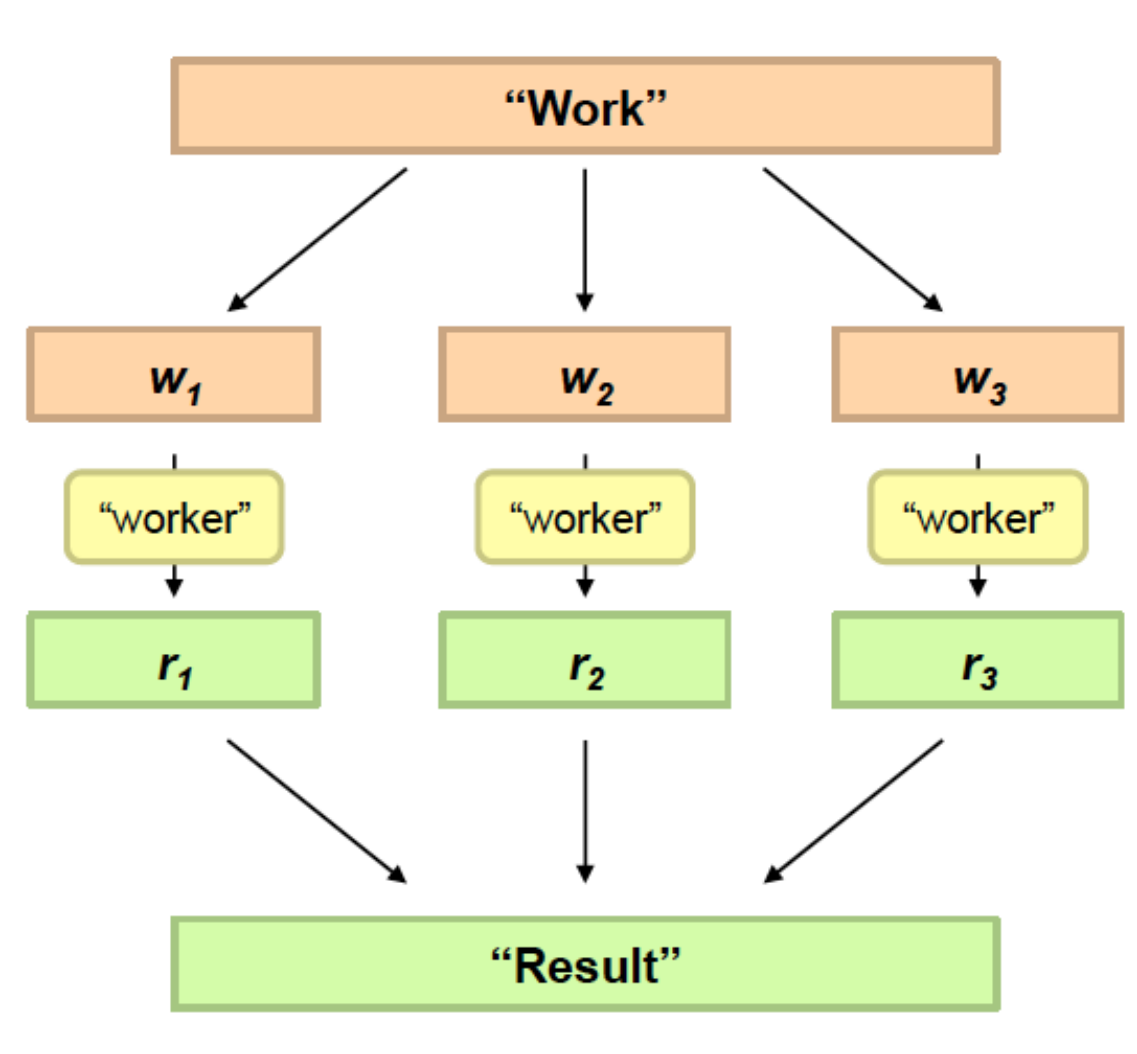
How different businesses can benefit from Big Data and what they need to do in order to use it effectively?

Big Data Open Source Tools



MS

Philosophy to Scale for Big Data Processing



Divide Work

Combine Results

MS

Distributed processing is non-trivial

- How to assign tasks to different workers in an efficient way?
- What happens if tasks fail?
- How do workers exchange results?
- How to synchronize distributed tasks allocated to different workers?

MS

Big data storage is challenging

- Data Volumes are massive
- Reliability of Storing PBs of data is challenging
- All kinds of failures: Disk/Hardware/Network Failures
- Probability of failures simply increase with the number of machines …

MS

END!

MS