Lecture 02: Intro Big Data Analytics and applications

KD44103 Big Data Analytics
Faculty of Computing & Informatics,
Universiti Malaysia Sabah

UMS
UNIVERSITI MALAYSIA SABAH

Learning Outcomes

At the end of the lecture, students are able to:

1. Explain the basic concepts of Big data analytics
2. Differentiate the various Big data analytics
3. Explain the basic concepts and techniques, tools and applications

Defining Big Data

- No single standard definition…

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

(Gartner, 2013)

Defining Big Data

(Fisher et. Al.)

- Big data means that the data is unable to be handled and processed by most current information system or methods
- Most of the traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to big data.

Defining Big Data (Cont…)

(Laney et. Al.)

A well known definition of Big Data known as 3Vs

- Volume (Data is Huge)
- Velocity (Data is changing with time and coming with a velocity)
- Variety (Data is coming from multiple sources in multiple forms)


BIG
DATA
ANALYTICS
STORAGE
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
DEFINITION
CURRENT
MAY
MOVING
WITHIN
EXAMPLES
EVERY
RECONSIDER
PRACTITIONERS
CAPTURE
ZETTABYTES
THOUGHT
INTERNET
BUSINESS
SETS
GENOMICS
DISTRIBUTED
DESCRIBING
RADIO-FREQUENCY
COMPLEXITY
MANAGEMENT
TERABYTES
CASE
ABILITY
PETABYTES
INCLUDE
ELAPSED
SYSTEMS
TOLERABLE
SIZE
MPP
SAN
QUALITIES
PARALLEL
SINCE
MASSIVELY
GROW
SET
CONNECTOMICS
ORGANIZATIONS
RELATIONAL
USE
LARGER
CONTINUES
SOCIAL
INDEXING
TENS
COMPLEX
CITATION
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
CAPACITY
BURIED
WORLD'S
HUNDREDS
DESKTOP
RECORDS
NETWORKS
DATABASES
SEARCH
SOLID
TYPES
CURRENTLY
GARTNER
OPPORTUNITIES
WORKING
AMOUNT
DIFFICULTY

The Power of Big Data

- Big Data can bring "**big values**" to our life in almost every aspects.
- Technologically, Big Data allows **diverse and heterogeneous data to be fully integrated and analyzed to help us make decisions**.
- Today, with the Big Data technology, **thousands of data from seemingly unrelated areas can help support important decisions**.

Examining Big Data Types

- New data sources like the data generated from sensors, smartphone, and tablets. Previously produced data hadn't been captured or stored and analyzed in a usable way.

Types of Data

- Data can be generated in tow main ways:
- **Human-generated:** This is data that humans, in interaction with computers, such as online services and digital devices.
- **Machine-generated:** refers to data that is created by a machine without human intervention, (software and hardware) in response to real-world events.
- **Example:** log file captures an authorization decision made by a security service and point-of-sale system generates a transaction against inventory to reflect items purchased by a customer.

Types of Data

Data can be divided into four categorize:

- Structured Data
- Unstructured Data
- Semi-Structured Data
- Metadata

Structured Data

- Data model or schema, and stored in tabular form
- Most often stored in a relational database
- Normally generated by enterprise applications and information systems
- Needs special consideration regarding processing or storage.
- **Example:** Banking transactions, invoices, customer records.

Unstructured Data

- *Unstructured data* is data that does not follow a specified format. If 20 percent of the data available to enterprises is structured data, the other 80 percent is unstructured.
- Faster growth rate than structured data.
- This form of data is either textual or binary and often conveyed via files that are self-contained and non-contained and nonrelation
- Example: Audio, Video, and Images file.

Semi-Structured Data

- Level of structure and consistency but is not relational in nature.
- Instead, is hierarchical or graph based.
- Commonly stored in files that contain text (XML, JSON).
- Examples: Electronic data interchange (EDI), spreadsheet, storage requirements.
- An example of pre-processing of semi-structured data would be the validation of an XML file to ensure that it conformed to its schema definition.

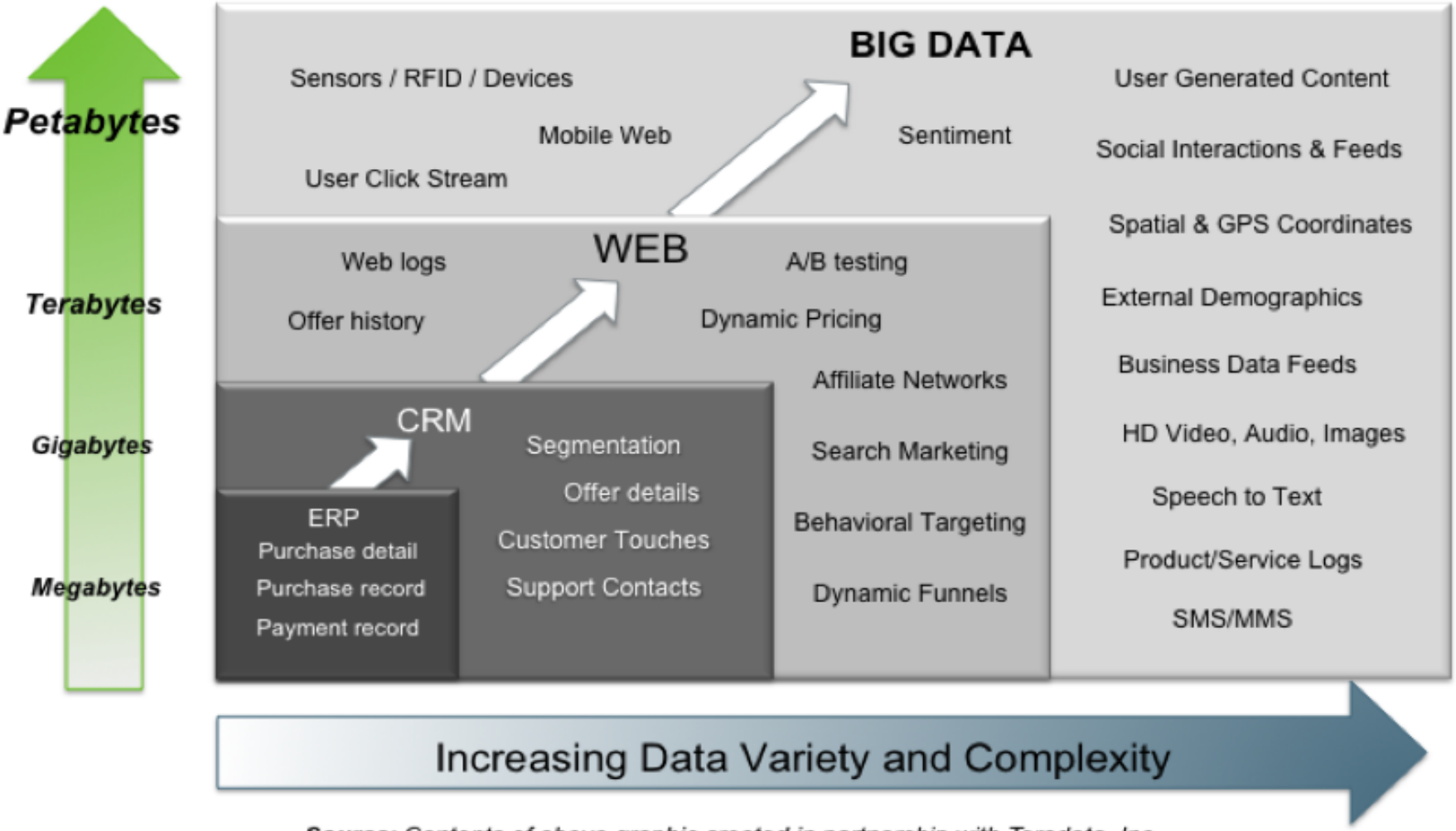
Metadata

- Provide information about a dataset's characteristics and structure.
- Mostly it is machine-generated and can be appended to data.
- The tracking of metadata is crucial to Big Data processing, storage and analysis because it provides information about the pedigree of the data and its provenance during processing.
- **Examples:** XML tags providing the author and creation date of a document
- Big Data solutions rely on metadata, particularly when processing semi-structured and unstructured data.

Sources of Big Data

- (Big data analytics)

Big Data = Transactions + Interactions + Observations



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

https://www.google.de/search?q=evolution+of+business+intelligence&newwindow=1&tbm=isch&tbo=u&source=univ&sa=X&ei=gEGoU5KXBuTb4QSGsoH4BQ&ved=0CDsQsAQ&biw=1366&bih=64

What is Big Data?

What makes data, “Big” Data?

How did data become “Big”

Source: http://as.wiley.com/WileyCDA/WileyTitle/productCd-111887613X.html

Recent Big Data Trends

- The world's data volume is expected to grow 40% per year, and 50 times by 2020.
- The market value of big data in 2010 was $3.2 billion, and this value was expected to increase to $16.9 billion in near future.
- Huge increase in demand for Big Data skills between now and 2020.

Current Big Data Trends in 2022

- The Big Data industry has seen tremendous growth in just a few years. It shot up from $169 billion in 2018 to $274 billion in 2022 — a 62% increase.
- The global Big Data market is projected to generate $103 billion in revenue by 2027 (SiliconANGLE, Wikibon)
- Approximately 2.5 quintillion bytes of data are created each day (LinkedIn)

30+ Incredible Big Data Statistics (2022) (explodingtopics.com)

Key Big Data Statistics 2022

- In 2022 the global big data industry is worth $274.3 billion.
- Google receives more than 3.5 billion searches every day.
- 100 billion messages are exchanged on WhatsApp every day.
- 95% of businesses struggle to manage unstructured data.
- Big data in healthcare will be worth $67 billion by 2025.
- 79 zettabytes of data were generated in 2021.
- 180 zettabytes of data will be generated in 2025.
- Data interactions have increased by 5000% since 2010.
- More than 1.2 billion years have been spent online.
- The big data analytics market is set to reach a whopping $103 billion by the year 2027.
- Due to poor data quality, the US economy loses approximately $3.1 trillion annually.
- With the help of big data, Netflix saves over $1 billion annually with customer retention.

Characteristics of Big Data

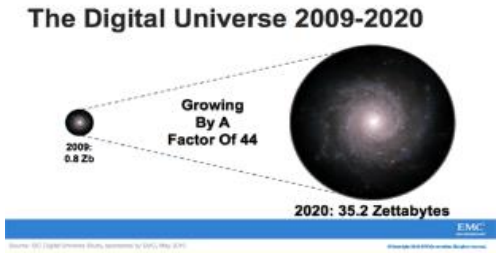
Big data is typically broken down by three main characteristics:

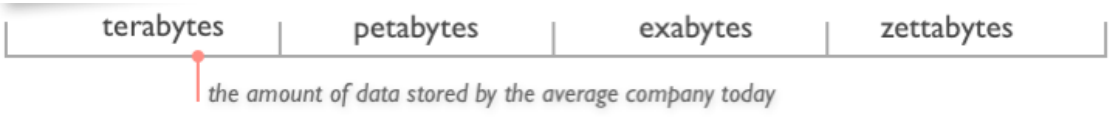
- **Volume:** How much data: Huge financial institutions generate terabytes of data daily.
- **Velocity:** How fast that data is processed: Hitting the threshold of 100 transactions per minute is easy for a respectable bank.
- **Variety:** The various types of data from transaction details and history to credit scores and risk assessment reports — the banks have troves of such data.

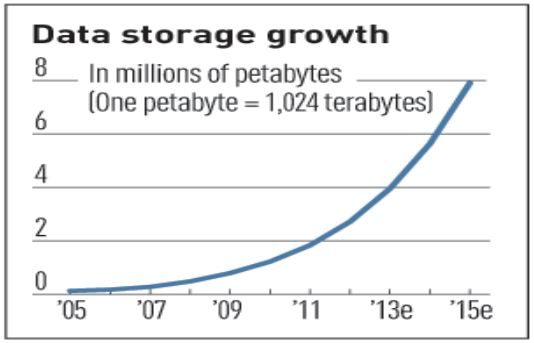
Characteristics of Big Data:

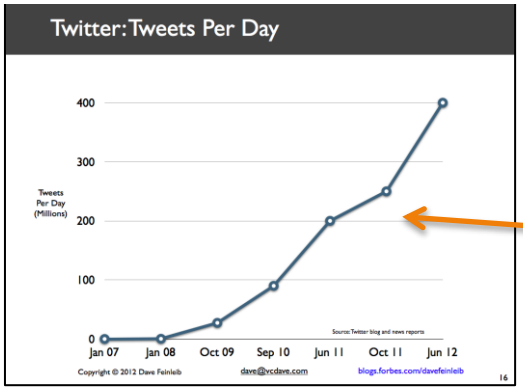
1-Scale (Volume)

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







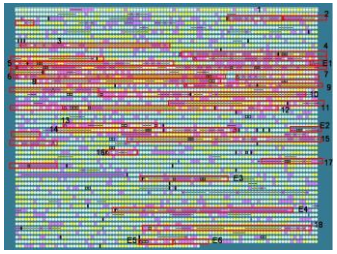


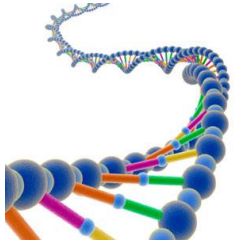
Exponential increase in collected/generated data

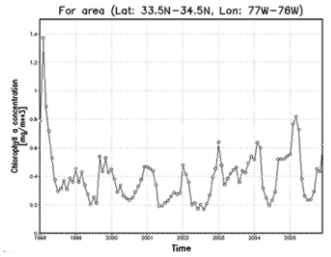
Characteristics of Big Data:

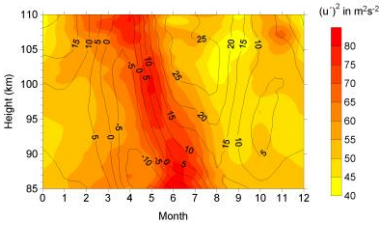
2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









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

Characteristics of Big Data:

3-Speed (Velocity)

- 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

Characteristics of Big Data

(Latest Enhanced in Big Data)

The 3Vs definition was incomplete so following dimensions to the data are added more sub-characteristics :

- Veracity
- Validity
- Value
- Variability
- Venue
- Vocabulary and Vagueness

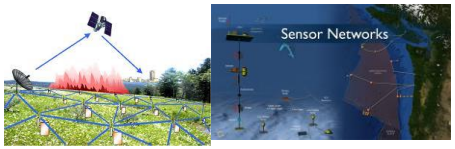
The data satisfying set of all these properties is known as Big Data.

Real-time/Fast Data

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

The importance of Big Data

- The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable:
  - Cost reductions
  - Time reductions
  - New product development and optimized offerings
  - Smarter business decision making.

The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

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

Big Data Areas of Applications

- Health and Well being Policy
- making and public opinions
- Smart cities and more efficient society
- New online educational models: MOOC and Student-Teacher modeling
- Robotics and human-robot interaction
- Many ..Many more

Big Data – Challenges

- Difficult in recognizing the right data and determining how to best use it
- Struggling to find the right talent
- Data access and connectivity obstacle
- Data technology landscape is evolving extremely fast
- Finding new ways of collaborating across functions and businesses
- Security concerns

Big Data Analytics

- Big Data analytics is a technique used to uncover important insights such unobserved correlations, hidden patterns, market trends, and consumer preferences. Big Data analytics offers a number of benefits, including the ability to use it to improve decision-making and stop fraud.

Examples of Big Data Analytics

- Big Data analytics can be applied in a variety of ways to enhance businesses and organizations. Here are a few instances:
  - ✓ Help improve the customer experience, analytics are used to understand customer behavior.
  - ✓ Future trend forecasting to aid in better commercial decision-making.
  - ✓ Understanding what works and what doesn't in marketing initiatives to improve them.
  - ✓ Understanding bottlenecks and how to remove them will increase operational effectiveness.
  - ✓ Increasing the speed of fraud and other types of abuse detection.

Big Data Industry Applications

- Ecommerce - Predicting customer trends and optimizing prices are a few of the ways e-commerce uses Big Data analytics
- Marketing - Big Data analytics helps to drive high ROI marketing campaigns, which result in improved sales
- Education - Used to develop new and improve existing courses based on market requirements
- Healthcare - With the help of a patient's medical history, Big Data analytics is used to predict how likely they are to have health issues
- Media and entertainment - Used to understand the demand of shows, movies, songs, and more to deliver a personalized recommendation list to its users
- Banking - Customer income and spending patterns help to predict the likelihood of choosing various banking offers, like loans and credit cards
- Telecommunications - Used to forecast network capacity and improve customer experience
- Government - Big Data analytics helps governments in law enforcement, among other things

Big Data Analytics Tools

Here are some of the key big data analytics tools :

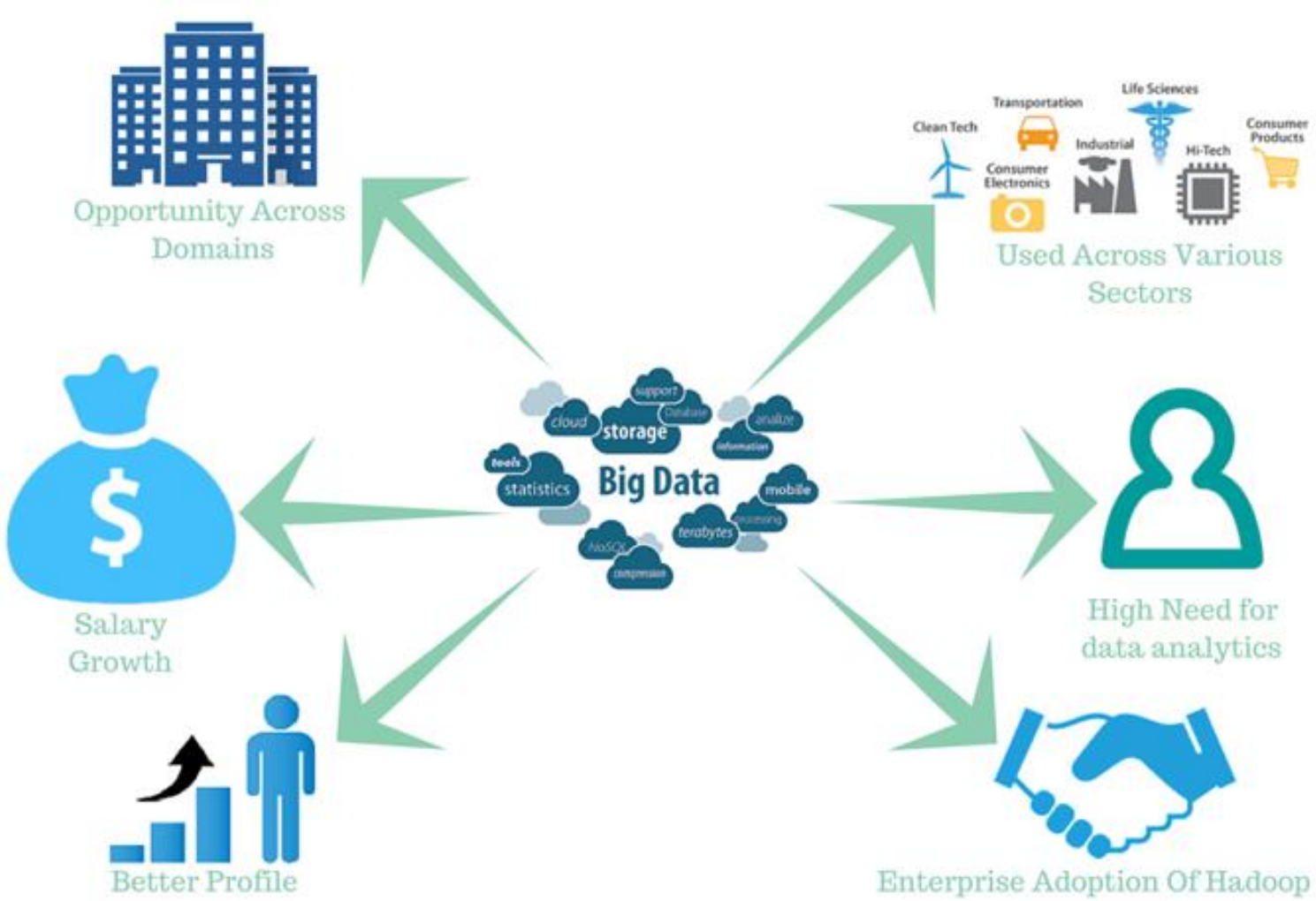
- Hadoop - helps in storing and analyzing data
- MongoDB - used on datasets that change frequently
- Talend - used for data integration and management
- Cassandra - a distributed database used to handle chunks of data
- Spark - used for real-time processing and analyzing large amounts of data
- STORM - an open-source real-time computational system
- Kafka - a distributed streaming platform that is used for fault-tolerant storage

Visualization tools

With the help of visualization tools:

- Tableau
- JupyteR
- Zoho Reports/ analytics
- Power Bi
- Google Charts
- IBM Watson
- Plotly
- Qlikview
- Dundas BI
- A user can easily analyze the data and present a new marketing strategy.

Job Opportunities and Big Data Analytics:



The End

Quick Review Question

- Describe the Dimension Characteristics of Big Data
- Differentiate between Structured and Unstructured data
- Discuss the V's of Big Data?
- Who are generating big data?