

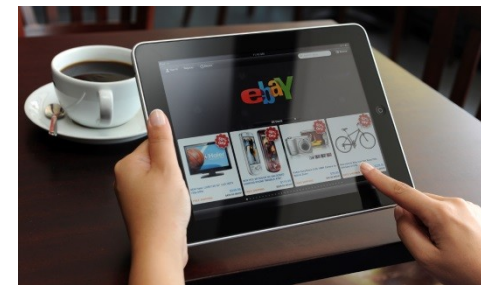
# Big Data – Beyond the Hype

Tushar B. Kute,  
<http://tusharkute.com>



# The war of data

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network
  - Cloud



# How much data we have?

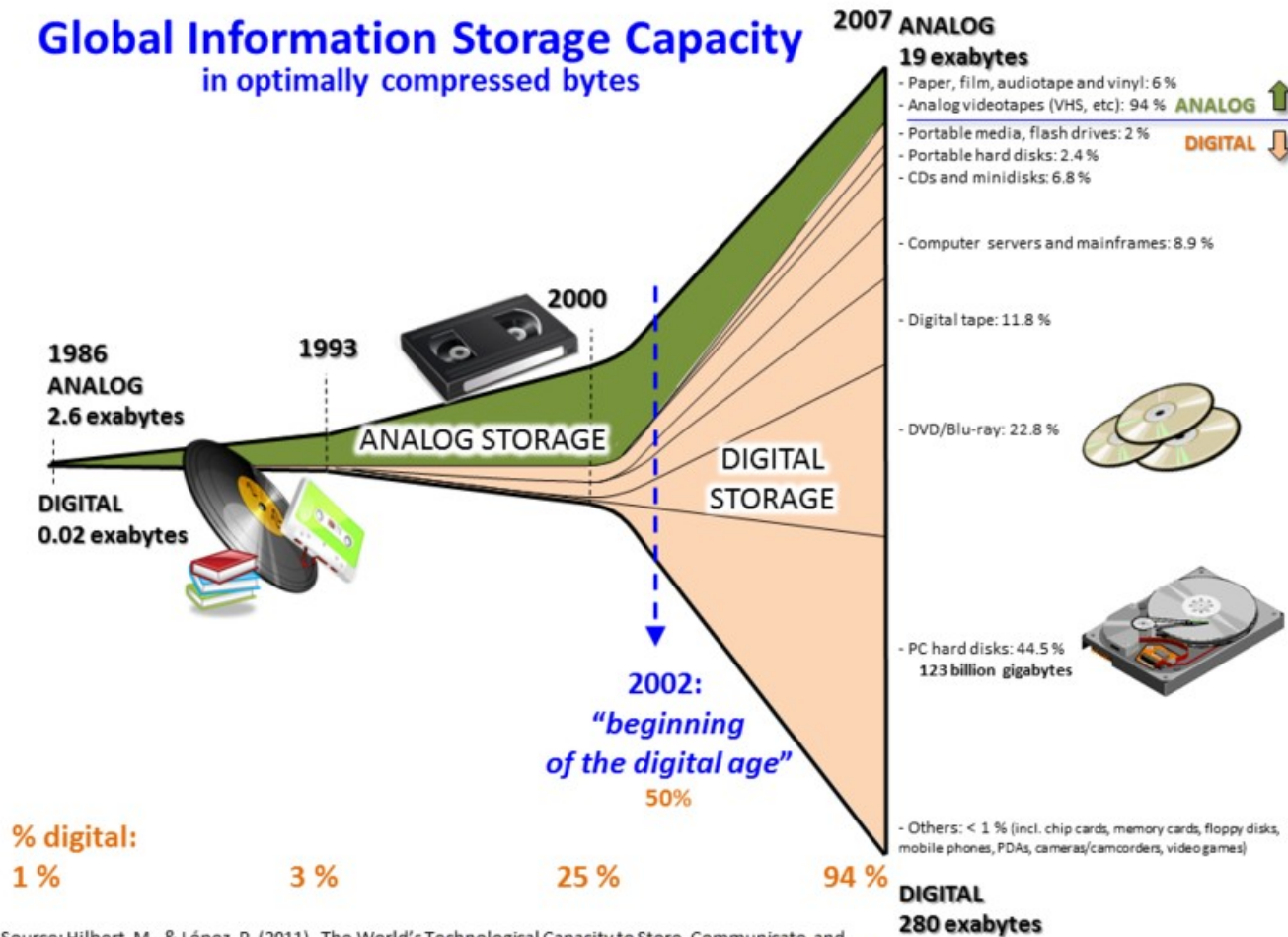
- Google processes 130+ PB a day
- Facebook has 100TB+ of daily logs
- eBay has 16.5 PB of user data + 150 TB/day
  - Cost of 1 TB of disk: Rs.4000
  - Time to read 1 TB disk:  
3 hrs (100 MB/s)

# Data and Big Data

- “90% of the world’s data was generated in the last few years.”
- Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year.
- The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field.
- The same amount was created in every two days in 2011, in every six minutes in 2016 and in every 52 seconds in 2022. This rate is still growing enormously.

# Data and Big Data

## Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

# Big Data Definition

- No single standard definition...

*“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...*

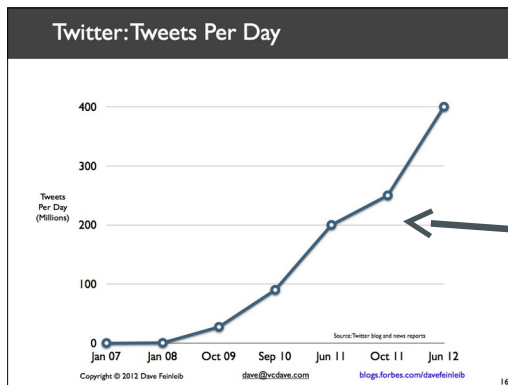
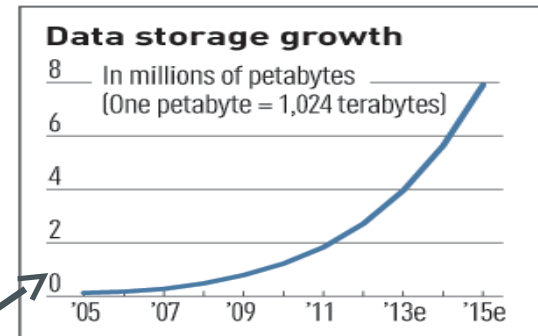
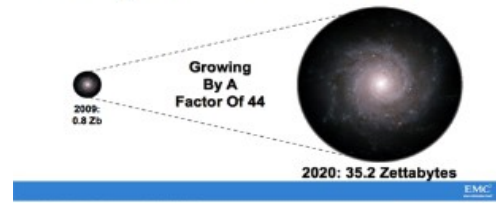
# What is Big Data

- Big Data is a collection of large datasets that cannot be processed using traditional computing techniques.
- It is not a single technique or a tool, rather it involves many areas of business and technology.

# Characteristics of Big Data: Volume

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

The Digital Universe 2009-2020



*Exponential increase in collected/generated data*



# Computer Memory Units

## ***UNITS OF COMPUTER MEMORY***

<b>1 Bit</b>	<b>Binary Digit</b>
<b>1 Nibble</b>	<b>4 Bits</b>
<b>8 Bits</b>	<b>1 Byte</b>
<b>1024 Bytes</b>	<b>1 KB (Kilo Byte)</b>
<b>1024 Kilo Bytes</b>	<b>1 MB (Mega Byte)</b>
<b>1024 Mega Bytes</b>	<b>1 GB (Giga Byte)</b>
<b>1024 Giga Bytes</b>	<b>1 TB (Tera Byte)</b>
<b>1024 Tera Bytes</b>	<b>1 PB (Peta Byte)</b>
<b>1024 Peta Bytes</b>	<b>1 EB (Exa Byte)</b>
<b>1024 Exa Bytes</b>	<b>1 ZB (Zetta Byte)</b>
<b>1024 Zetta Bytes</b>	<b>1 YB (Yotta Byte)</b>
<b>1024 Yotta Bytes</b>	<b>1 BB (Bronto Byte)</b>
<b>1024 Bronto Bytes</b>	<b>1 GB* (Geop Byte)</b>
<b>1024 Geop Bytes</b>	<b>1 SB (Sagan Byte)</b>
<b>1024 Sagan Bytes</b>	<b>1 PB (Piya Byte)</b>
<b>1024 Piya Bytes</b>	<b>1 AB (Alpha Byte)</b>
<b>1024 Alpha Bytes</b>	<b>1 KB* (Kryat Byte)</b>
<b>1024 Kryat Bytes</b>	<b>1 AB* (Amos Byte)</b>
<b>1024 Amos Bytes</b>	<b>1 PB* (Pectrol Byte)</b>
<b>1024 Pectrol Bytes</b>	<b>1 BB* (Bolger Byte)</b>
<b>1024 Bolger Bytes</b>	<b>1 SB* (Sambo Byte)</b>
<b>1024 Sambo Bytes</b>	<b>1 QB (Quesa Byte)</b>
<b>1024 Quesa Bytes</b>	<b>1 KB** (Kinsa Byte)</b>
<b>1024 Kinsa Bytes</b>	<b>1 RB (Ruther Byte)</b>
<b>1024 Ruther Bytes</b>	<b>1 BB** (Bubni Byte)</b>
<b>1024 Bubni Bytes</b>	<b>1 SB** (Seaborg Byte)</b>
<b>1024 Seaborg Bytes</b>	<b>1 BB*** (Bohr Byte)</b>
<b>1024 Bohr Bytes</b>	<b>1 HB (Hassiu Byte)</b>
<b>1024 Hassiu Bytes</b>	<b>1 MB* (Meitner Byte)</b>
<b>1024 Meitner Bytes</b>	<b>1 DB (Darmstad Byte)</b>
<b>1024 Darmstad Bytes</b>	<b>1 RB* (Roent Byte)</b>
<b>1024 Roent Bytes</b>	<b>1 CB (Coper Byte)</b>

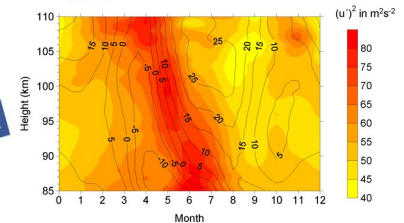
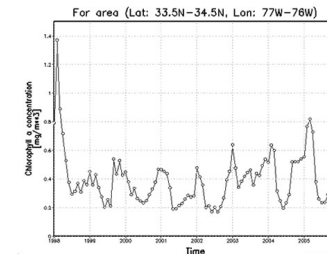
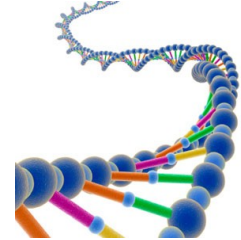
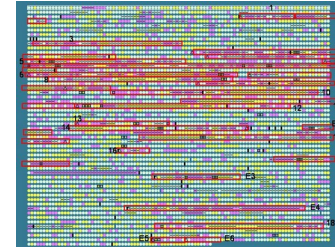
*www.raj-gigaworld.blogspot.com*

R.D

*www.facebook.com/raj.dev/36generalknowledge*

# Characteristics of Big Data: Variety

- 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



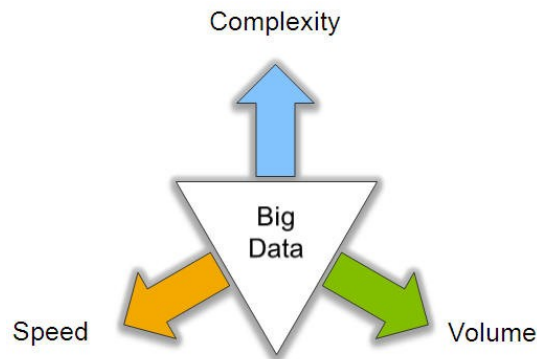
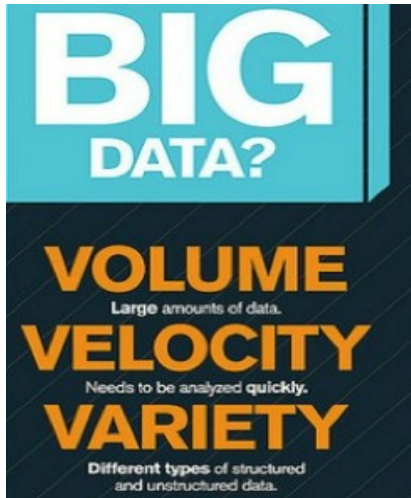
To extract knowledge all these types of data need to be linked together

# Characteristics of Big Data: Velocity

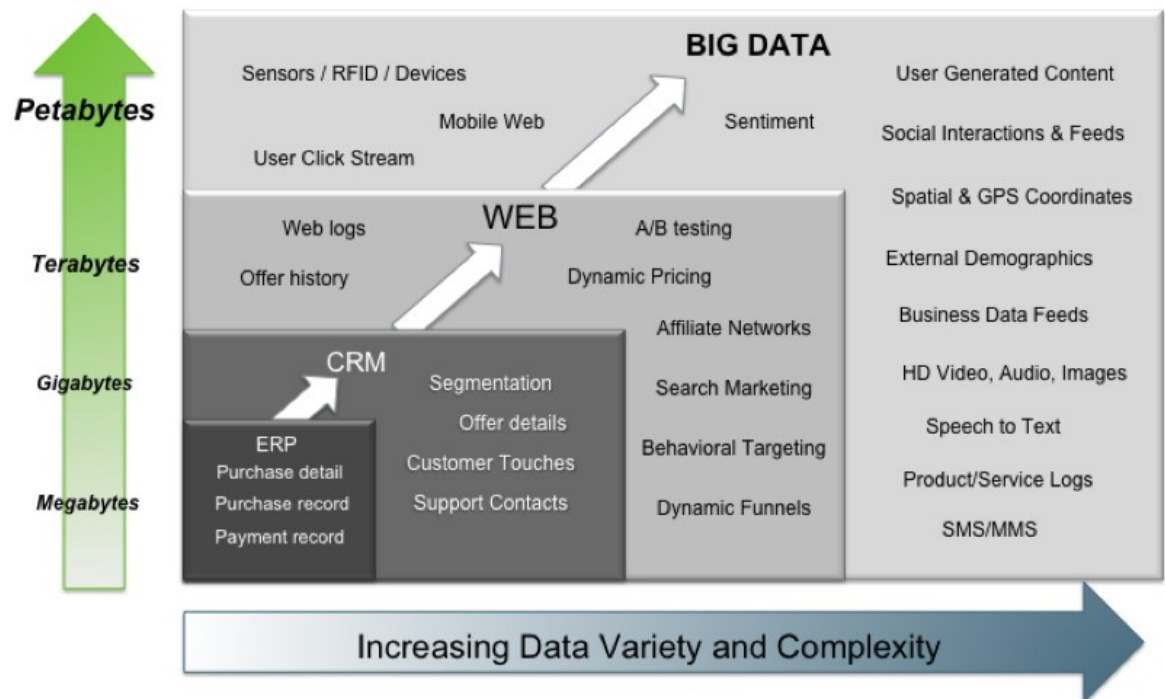
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions, missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like send promotions right now for store next to you.
  - **Healthcare monitoring:**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



# Big Data: 3 Vs

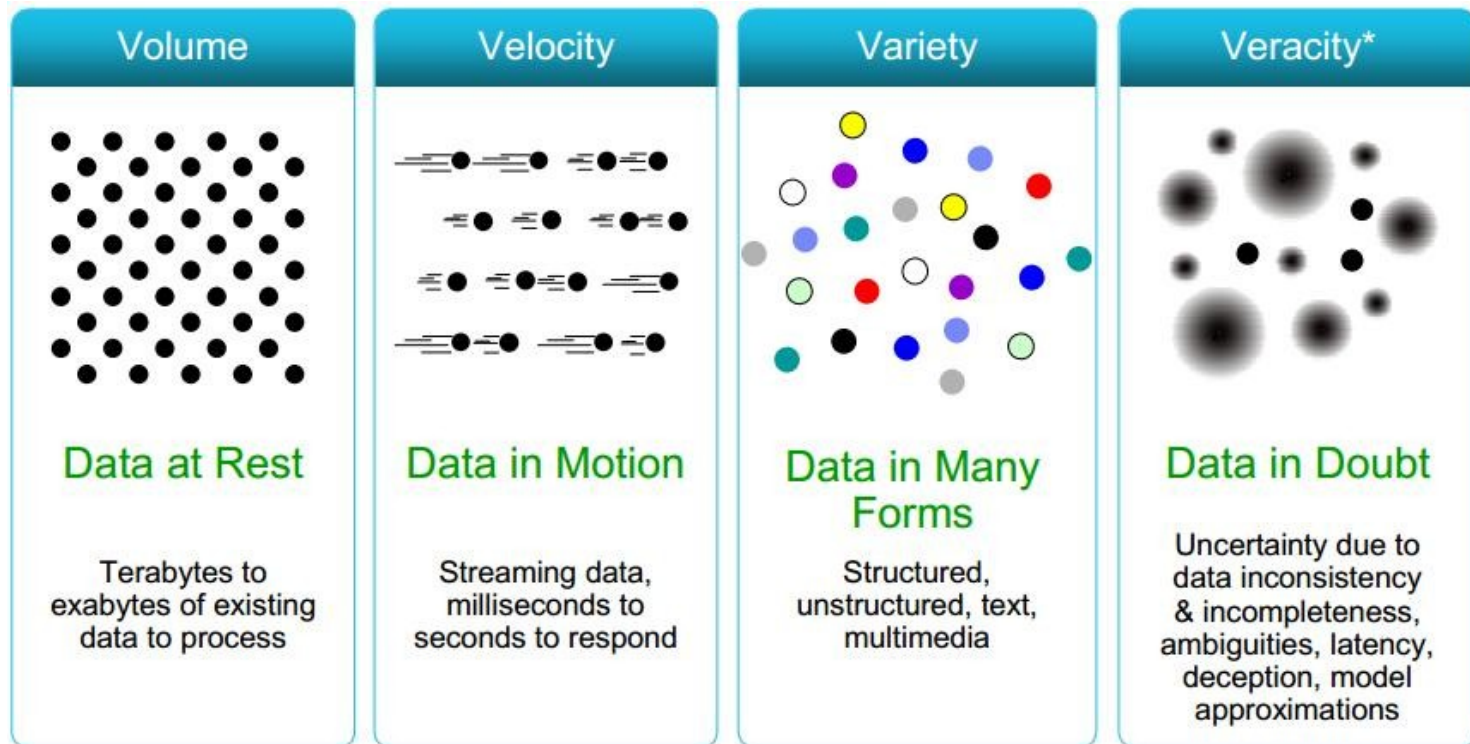


Big Data = Transactions + Interactions + Observations



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

# Big Data: The 4<sup>th</sup> V



# Veracity and Value

- Veracity
  - The truthfulness or reliability of the data, which refers to the data quality and the data value.
  - Big data must not only be large in size, but also must be reliable in order to achieve value in the analysis of it.
  - The data quality of captured data can vary greatly, affecting an accurate analysis.

# Veracity and Value

- Value
  - The worth in information that can be achieved by the processing and analysis of large datasets.
  - Value also can be measured by an assessment of the other qualities of big data.
  - Value may also represent the profitability of information that is retrieved from the analysis of big data.



# Variability

- Variability
  - The characteristic of the changing formats, structure, or sources of big data.
  - Big data can include structured, unstructured, or combinations of structured and unstructured data.
  - Big data analysis may integrate raw data from multiple sources.
  - The processing of raw data may also involve transformations of unstructured data to structured data.



# Other features

- Exhaustive
  - Whether the entire system (i.e.,  $n=all$ ) is captured or recorded or not. Big data may or may not include all the available data from sources.
- Fine-grained and uniquely lexical
  - Respectively, the proportion of specific data of each element per element collected and if the element and its characteristics are properly indexed or identified.

# Other features

- Relational
  - If the data collected contains common fields that would enable a conjoining, or meta-analysis, of different data sets.
- Extensional
  - If new fields in each element of the data collected can be added or changed easily.
- Scalability
  - If the size of the big data storage system can expand rapidly.

# What is Data Science ?

- Data science is an **interdisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply **knowledge** and **actionable insights** from data across a broad range of application domains.
- Data science is related to data **mining**, **machine learning** and **big data**.
- Data science (DS) is a **multidisciplinary** field of study with goal to address the challenges in big data.
- Data science **principles** apply to all data – big and small.

# What is Data Science ?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

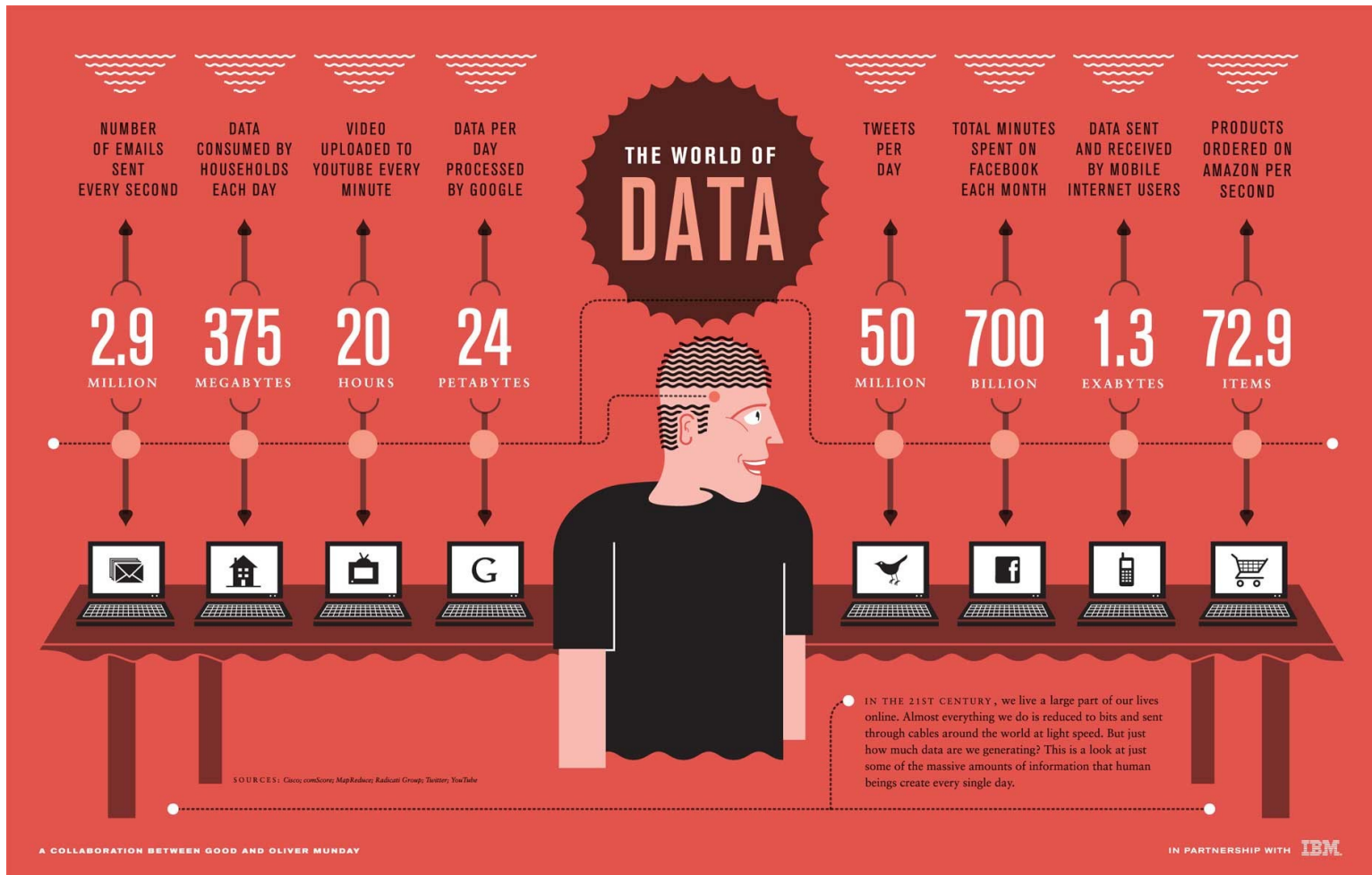
# What Comes Under Big Data?

- **Black Box Data:** It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data:** Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data:** The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data:** The power grid data holds information consumed by a particular node with respect to a base station.

# What Comes Under Big Data?

- Transport Data: Transport data includes model, capacity, distance and availability of a vehicle.
- Search Engine Data: Search engines retrieve lots of data from different databases.
- Structured data: Relational data.
- Semi Structured data: XML data.
- Unstructured data: Word, PDF, Text, Media Logs.

# Big Data Today



# Types of Big Data

- Following are the types of Big Data:
  - Structured
  - Unstructured
  - Semi-structured



# Structured Data

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.





# Structured Data


Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

# Unstructured Data

- Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

# Unstructured Data

Google    +Prafulla  

**Web** News Images Videos Maps More ▾ Search tools 

---

About 3,15,00,000 results (0.37 seconds)

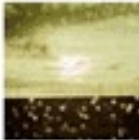
**IBM Hadoop & Enterprise - IBM.com**  
**Ad** [www.ibm.com/HadoopInEnterprise](http://www.ibm.com/HadoopInEnterprise) ▾  
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!  
IBM has 28,706 followers on Google+

**100% Uptime for Hadoop - wandisco.com**  
**Ad** [www.wandisco.com/hadoop](http://www.wandisco.com/hadoop) ▾  
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

**Hadoop Big Data - Simplilearn.com**  
**Ad** [www.simplilearn.com/BigData\\_Training](http://www.simplilearn.com/BigData_Training) ▾  
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!









---

**News for hadoop big data**

 **What you missed in Big Data: Hadoop applications Watson ...**  
SiliconANGLE (blog) - 19 hours ago  
**big data** cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for **hadoop big data** on Google

Sponsored ⓘ

 <b>Big Data Big Analytics: ...</b> Rs. 348.00 Amazon.in	 <b>Oracle Big Data ...</b> Rs. 549.00 Amazon.in	 <b>Big Data Analytics With ...</b> Rs. 455.00 Amazon.in	 <b>Hadoop Beginner's ...</b> Rs. 595.00 Amazon.in
 <b>Hadoop in Action</b> Rs. 460.00 Flipkart	 <b>Big Data Analytics with ...</b> Rs. 3,100.00 Amazon.in	 <b>Hadoop Mapreduce ...</b> Rs. 468.00 Amazon.in	 <b>Hadoop: The Definitive ...</b> Rs. 553.00 Amazon.in

# Semi-structured Data

- Semi-structured data can contain both the forms of data.
- We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
- Example of semi-structured data is a data represented in an XML file.

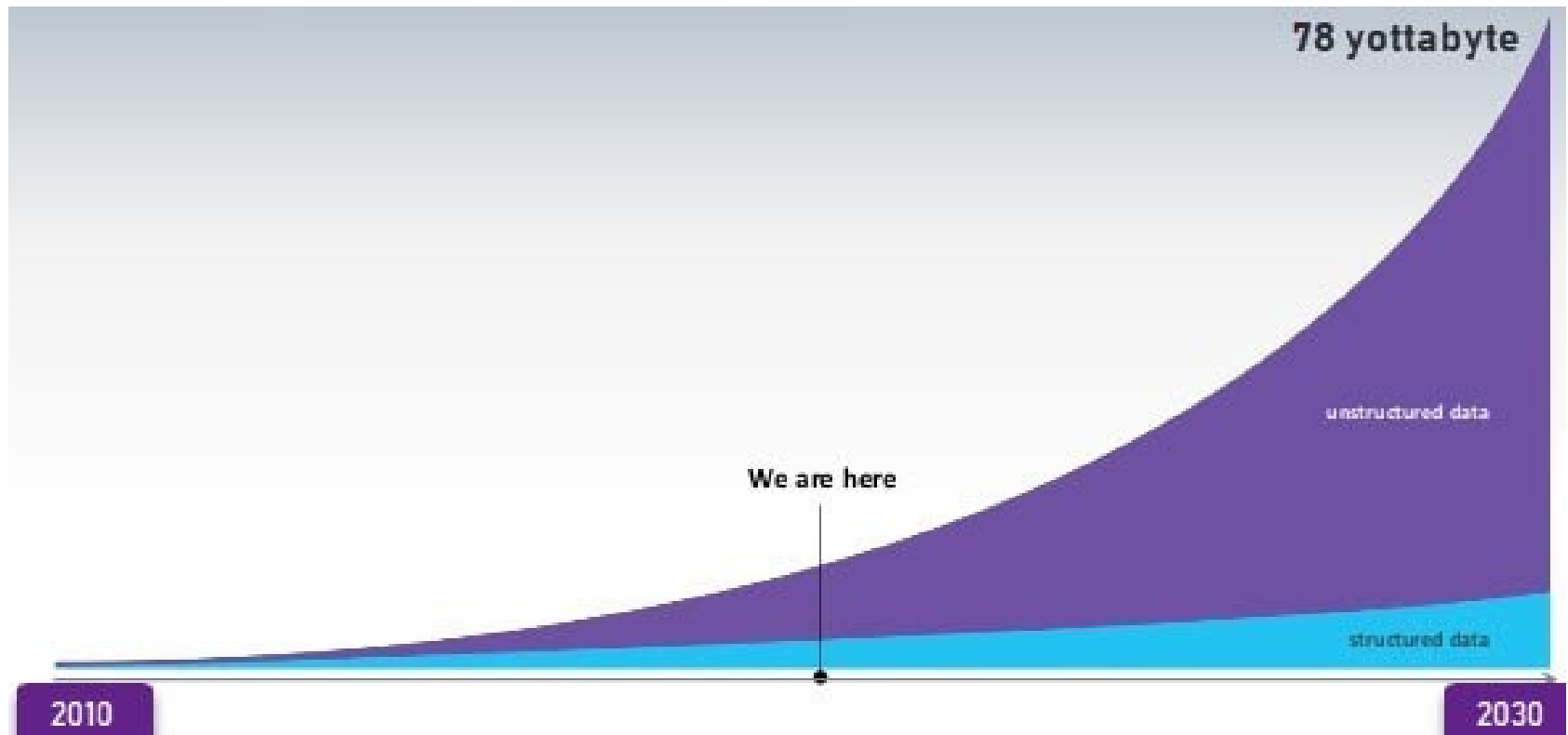
# Semi-structured Data

- 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>`

# Growth of data over years



# Benefits of Big Data

- Businesses can utilize outside intelligence while taking decisions
- Improved customer service
- Early identification of risk to the product/services, if any
- Better operational efficiency



# Big Data Technologies

- Operational Big data
- Analytical Big data

# Operational Big Data

- These include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.
- NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently.

# Analytical Big Data

- These includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.
- MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

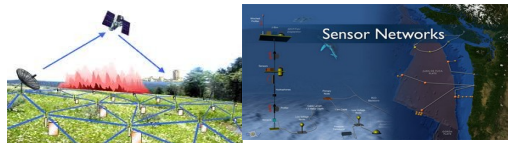
# Example case studies



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



**Scientific instruments**  
(collecting all sorts of data)



**Sensor technology and networks**  
(measuring all kinds of data)



**Mobile devices**  
(tracking all objects all the time)

# Big Data generation models

- 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



# Challenges in Big Data

- The major challenges associated with big data are as follows:
  - Capturing data
  - Curation
  - Storage
  - Searching
  - Sharing
  - Transfer
  - Analysis
  - Presentation

# Thank you

*This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License*



@mitu\_skillologies



/miTuSkillologies



@mitu\_group



/company/mitu-  
skillologies



MITUSkillologies

## Web Resources

<https://mitu.co.in>  
<http://tusharkute.com>

[contact@mitu.co.in](mailto:contact@mitu.co.in)  
[tushar@tusharkute.com](mailto:tushar@tusharkute.com)