Introduction to Big Data

Introduction to Big Data

- Big data is an all-inclusive term that refers to extremely large, very fast, highly diverse and complex data that cannot be managed with traditional data management tools
- It includes all kind of data, which helps deliver the right information, to right person, in right quantity, at right time, to make right decision
- Big data can be harnessed by developing infinitely scalable, flexible, and evolutionary data architectures, coupled with the use of cost-effective computing machines.

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

By Gartner

Understanding Big Data

- Big data is data that exceeds the processing capacity of conventional database systems.
- The data is too big, moves too fast, or doesn’t fit the structures of your database architectures.
- To gain value from this data, you must choose an alternative way to process it.
- At the fundamental level it is just another collection of data that can be analyzed and utilized for the benefit of the business and on another level, it is a special kind of data that poses unique challenges and offers unique benefits.

Life Cycle of Big Data

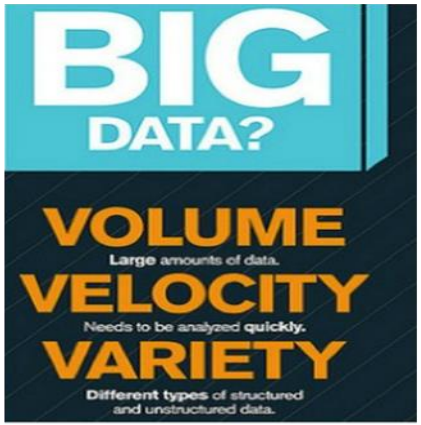
Big data is mostly, over 90%, unstructured data
There are huge opportunities for technology providers to innovate and manage the entire life cycle of Big Data

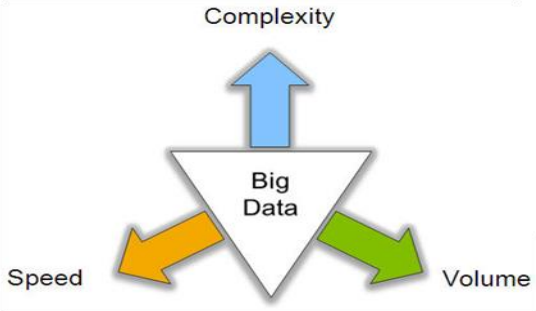
- Generate
- Gather
- Store
- Organize
- Analyze
- Visualize

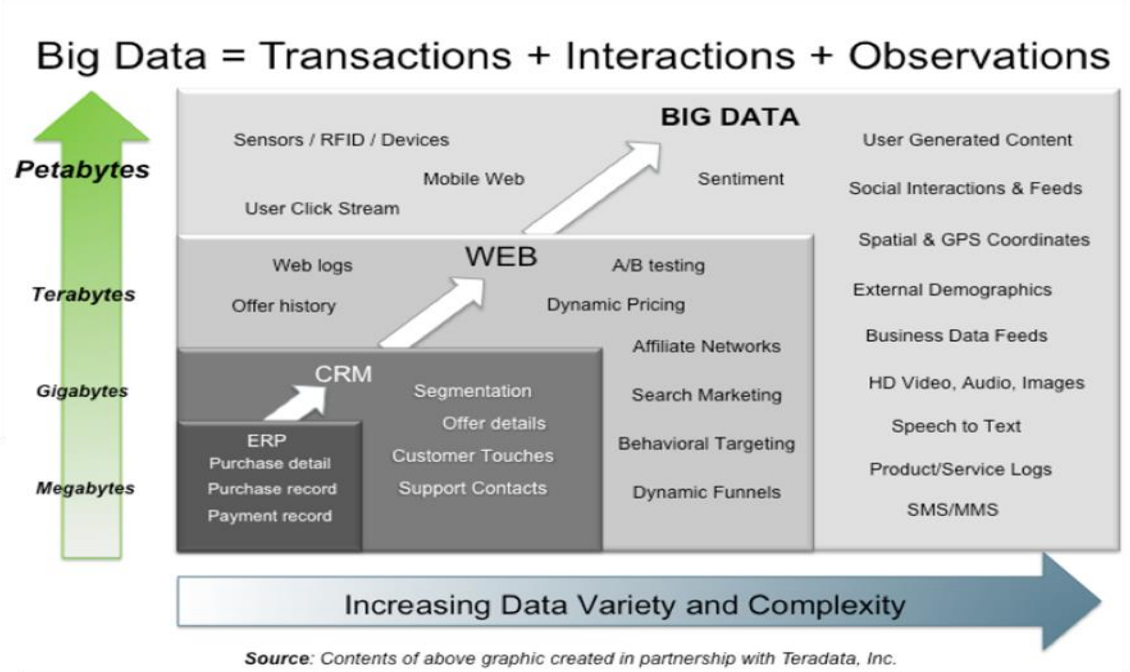
The Three *V*-s

Volume and Velocity are driven by variety
The varying varacity and value of data complicates the situation

Big Data: 3V's

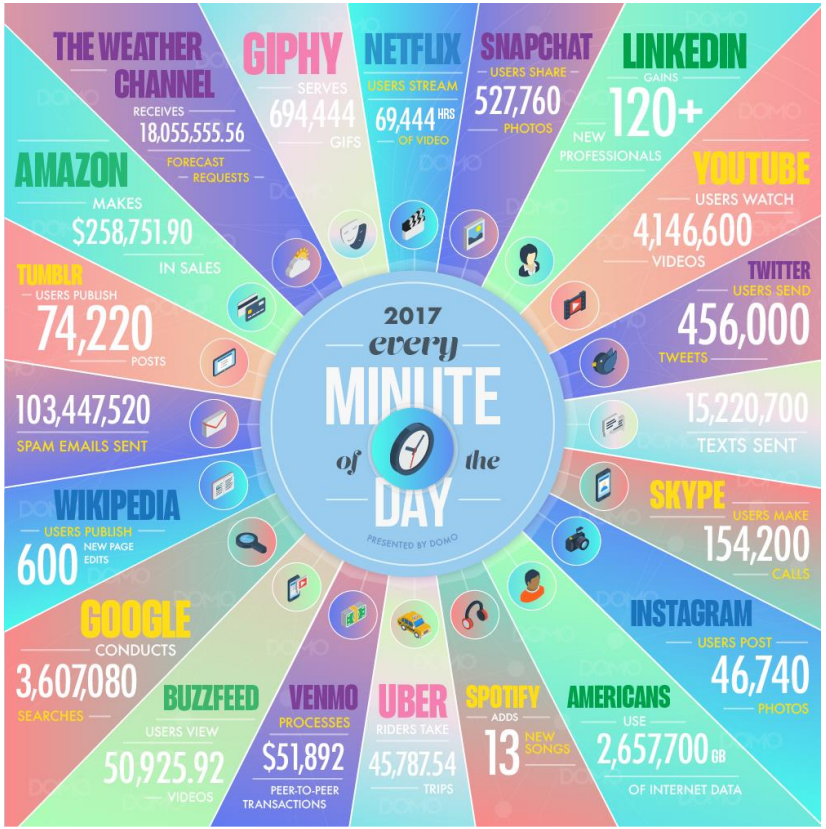






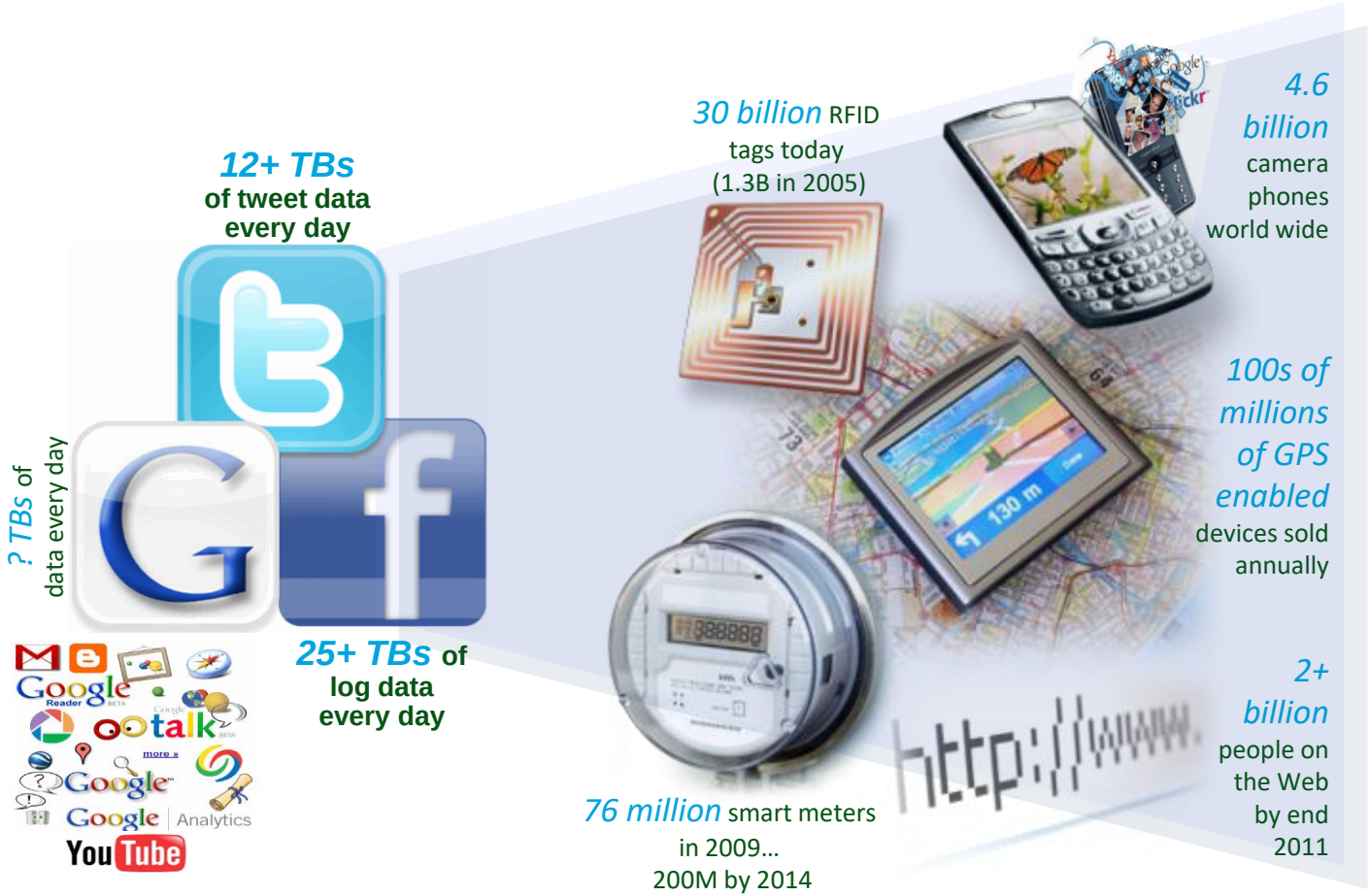
Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Volume



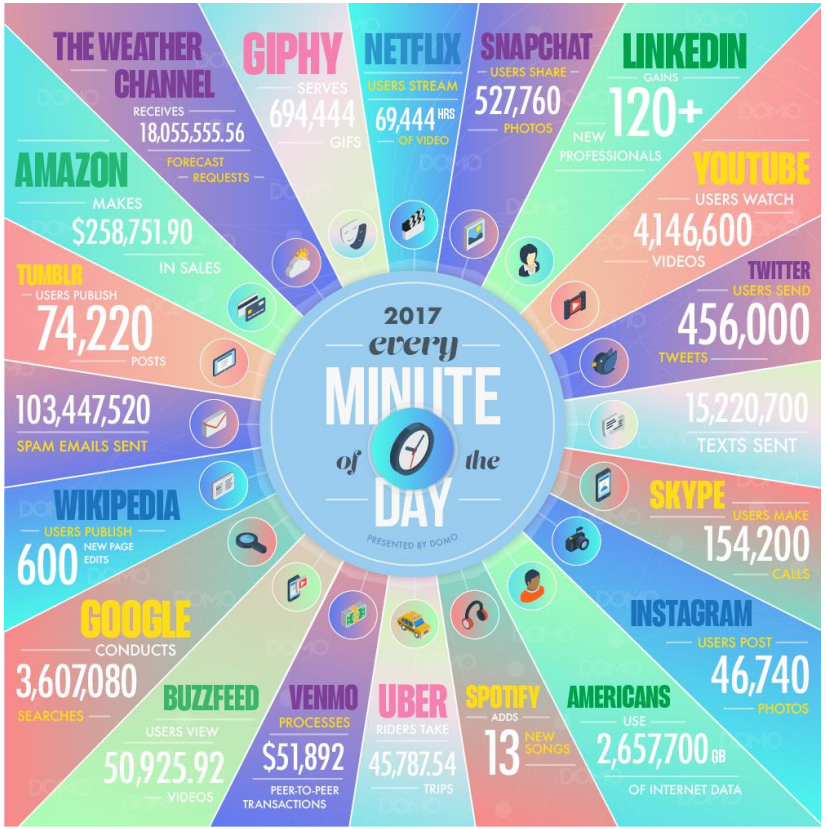
- Quantity of data generated in the world is doubling every 12-18 months
- Data sets too large to store and analyse using traditional databases and finding something from it in a reasonable period of time is like a miracle. (Petabytes and Exabytes) (1 Exabyte = 1 Million TB)

Image Source: https://www.domo.com/blog/data-never-sleeps-5/


12+ TBs
of tweet data
every day
? TBs of
data every day
25+ TBs of
log data
every day
30 billion RFID
tags today
(1.3B in 2005)
4.6
billion
camera
phones
world wide
100s of
millions
of GPS
enabled
devices sold
annually
2+
billion
people on
the Web
by end
2011
76 million smart meters
in 2009…
200M by 2014
Google
Reader
talk
more »
Google
Google Analytics
YouTube
http://www.

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**, and in every ten minutes in **2013**. This rate is still growing enormously.

Velocity



- If traditional data is like a drop of water, Big Data is like flowing river
- Speed at which data is generated by billions of devices, and communicated at the speed of light. (High speed Internet availability)
- Mobile devices can generate and communicate data from anywhere, at any time
- Processing should be faster than generation
- Analyse data while it is being generated without even putting it into databases

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Velocity (Speed)

- Data is begin generated fast from different sources 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

Sources of Real-time/Fast Data

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

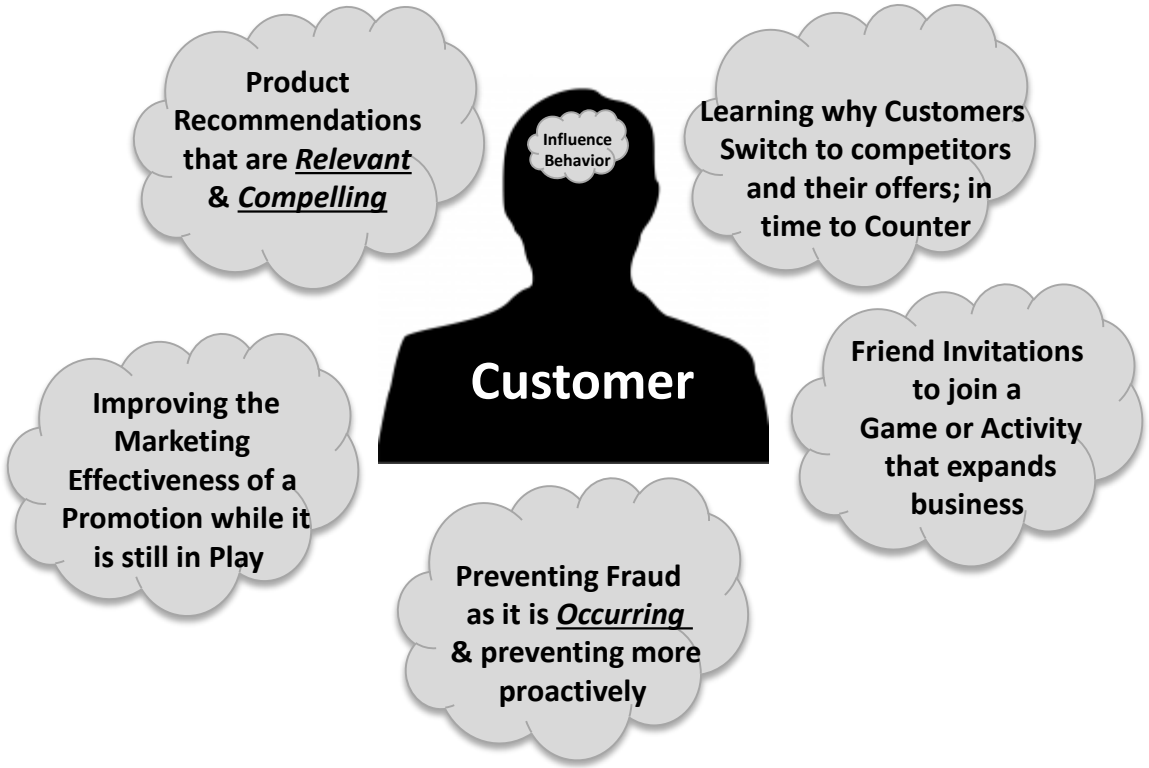
Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

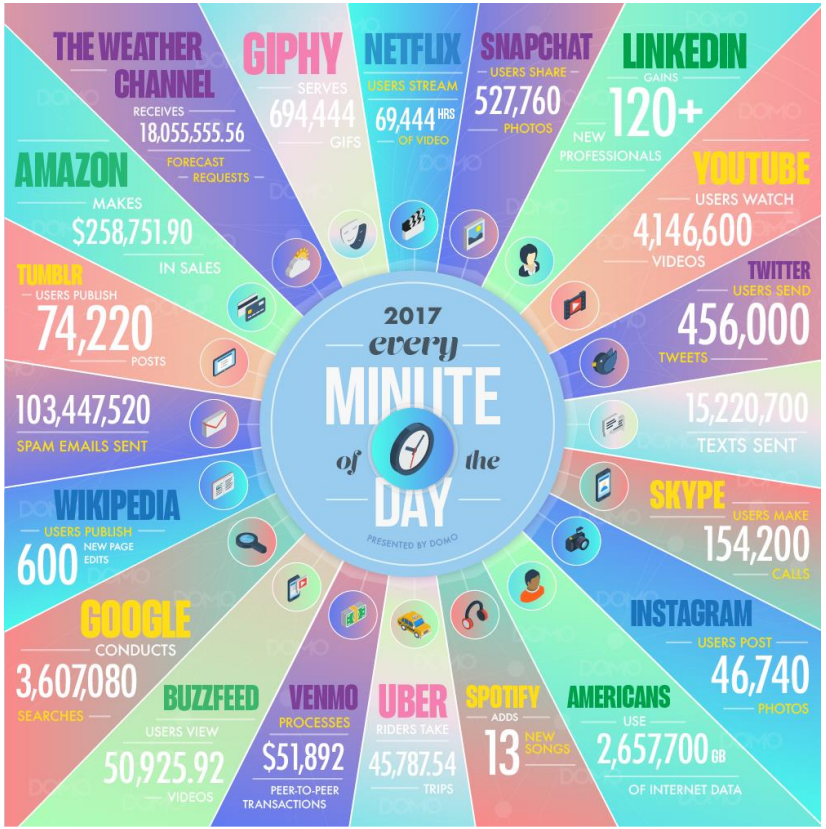
Sensor technology and networks
(measuring all kinds of data)

- 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 needs to be handled

Real-Time Analytics/Decision Requirement



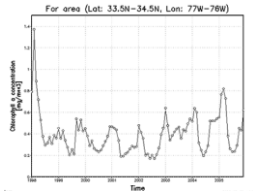
Variety

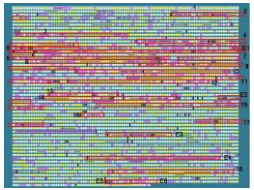


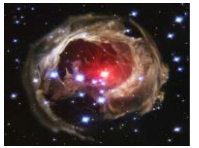
- Different types of data that we can use is inclusive of all forms of data, for all kinds of functions, from all sources and devices
- Generated by different entities
  - Humans
  - Machines (HW + SW)
  - Sensors

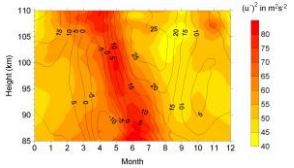
Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Variety (Complexity)









- 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 ,
- Form Data type range in numbers to text, graph, map, audio, video and others
- Function: human conversation, songs and movies, business transaction records, machine operation performance data, new product design data, old archived data, etc.
- Source of data: Mobile phone and tablets, web access and search logs , Business transactional information, temperature and pressure sensors on machines, RFID tags on assets generate incessant and repetitive data.
- Big Public Data (online, weather, finance, detecting human faces from pictures, compare voice to identify the speaker, comparing handwritings to identify the writer, etc.)

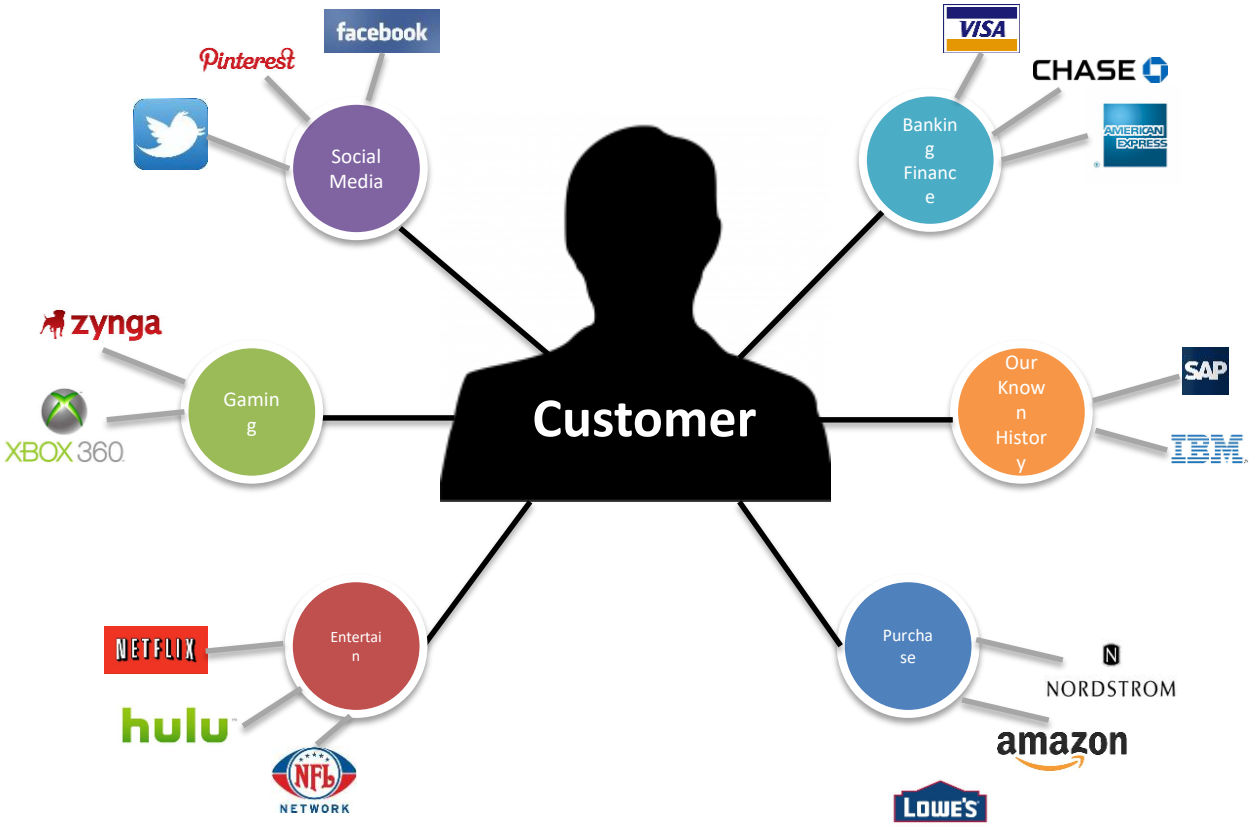
Broadly speaking there are three broad types of sources of data

Human to Human Communication, Human-Machine communication and

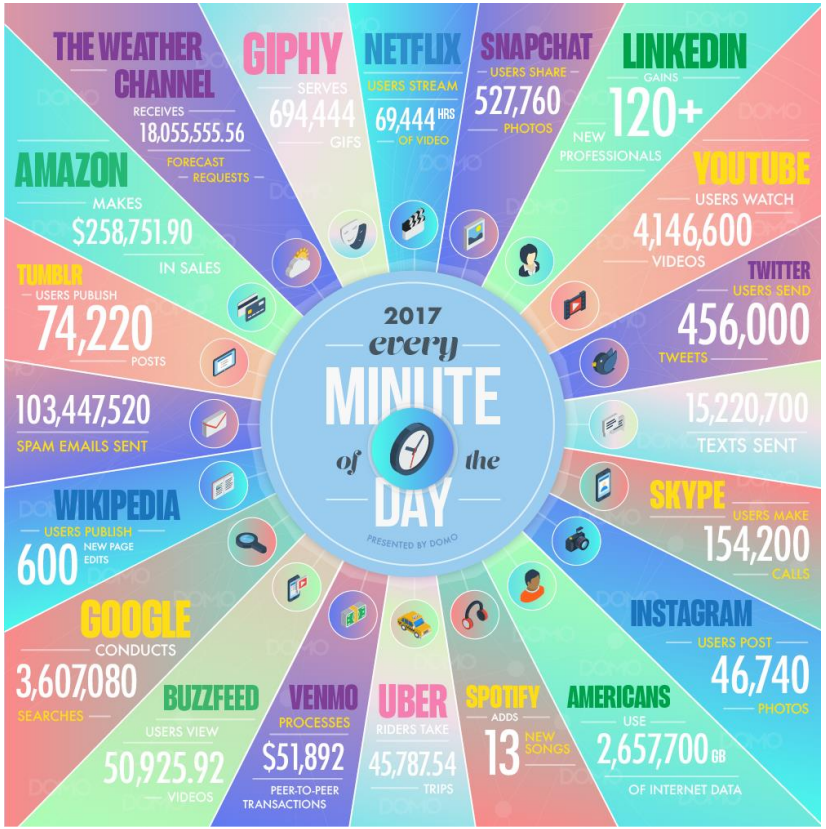
Machine to Machine communication.

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

A Single View to the Customer



Veracity



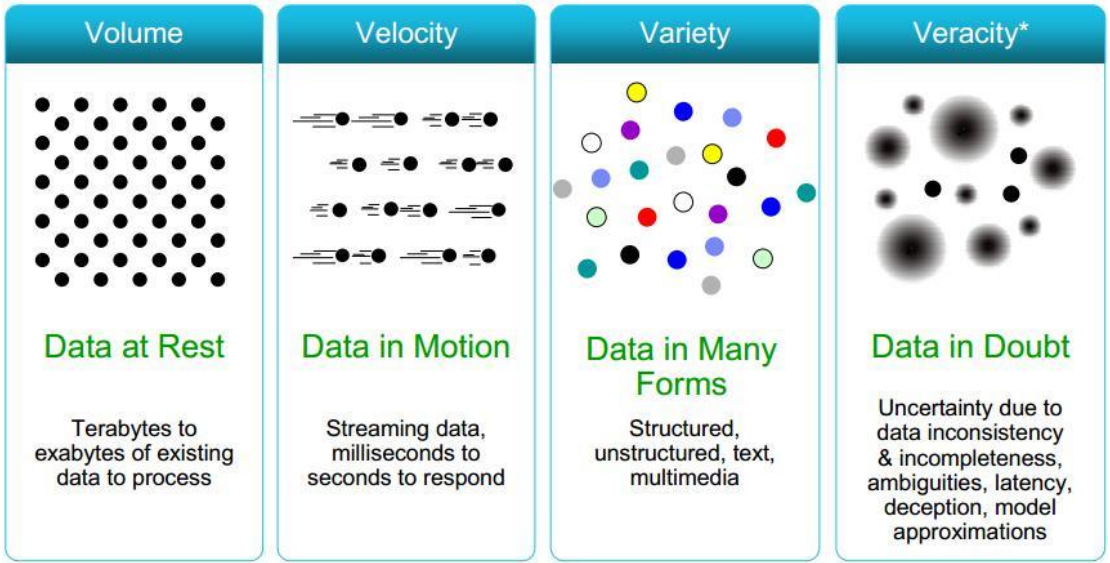
- Relates to the truthfulness, believability and quality of data
- Big Data is messy and there is considerable misinformation out there.
- The reasons for poor quality of data can range from technical error to human error to malicious intent.
- Volumes makes up for quality
  - Eg. Tweets with spelling mistakes, short words
  - u→you, thr→there, teh→the

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Veracity

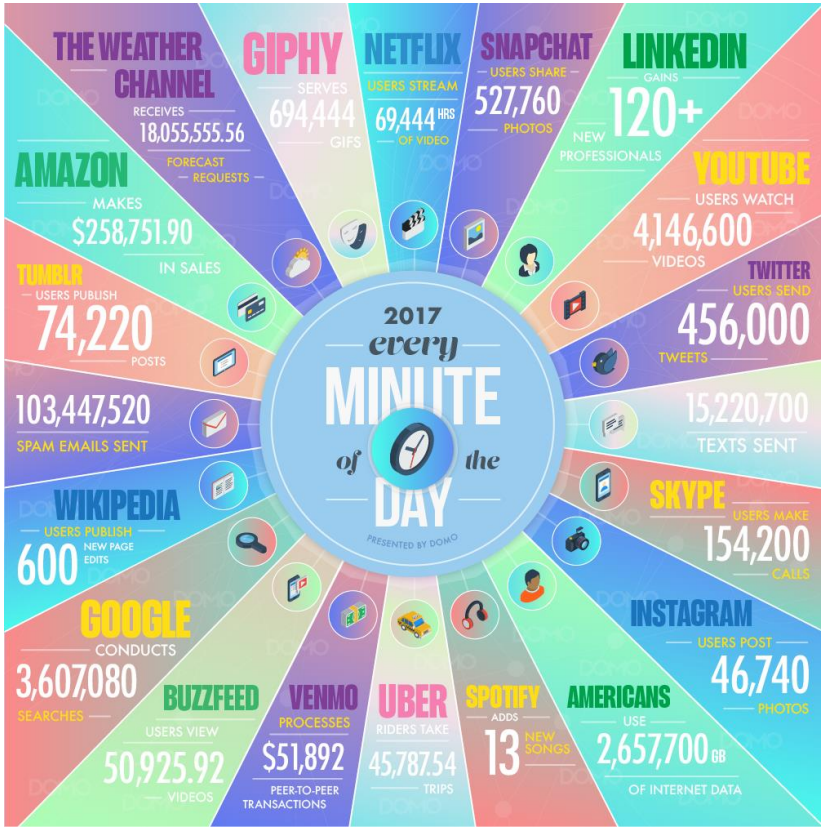
- The source of information may not be authoritative
  - Whitehouse.gov or nytimes.com is authentic and complete, but Wikipedia data may not be equally reliable.
- The data may not be communicated and received correctly because of human or technical failure
  - Sensors and communication machines may malfunction and may record and transmit incorrect data.
  - Urgency may require the transmission of the best data available at a point in time. Such data makes reconciliation with later, accurate, records more problematic.
- The data provided and received may, however also be intentionally wrong for competitive or security reasons.
  - There could be disinformation and malicious information spread for strategic reasons.

Some Make it 4V's



Additional *V*-s

Value



Getting value out of Big Data!!!

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

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

By Gartner

Definition

“Big data” is

high-volume, -velocity and -variety information assets

that demand cost-effective, innovative forms of information processing

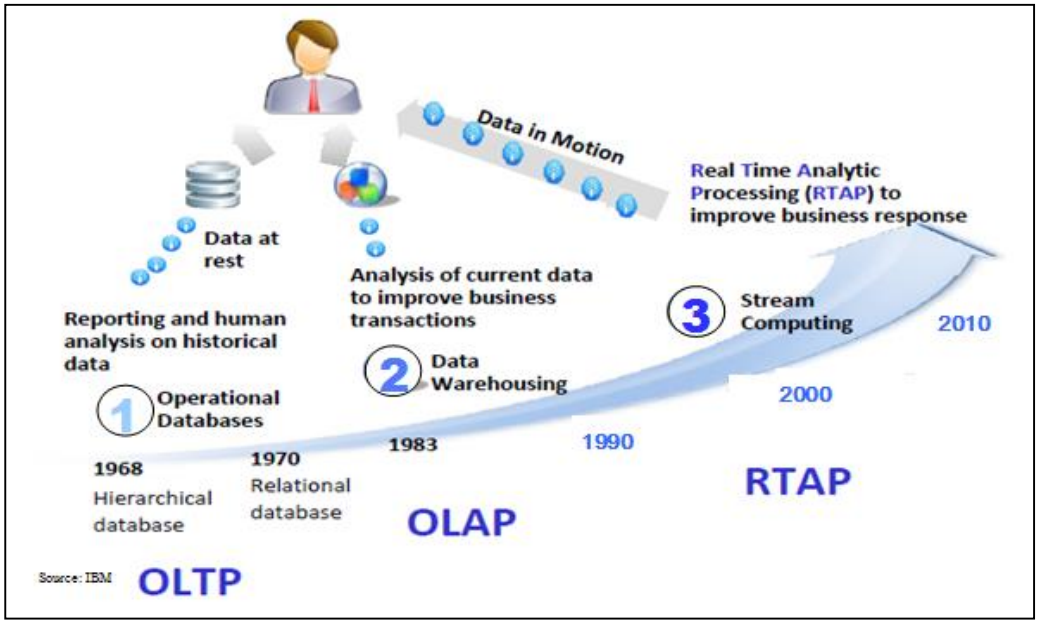
for enhanced insight and decision making

By Gartner

Wikipedia Definition

- Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate…
- Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. …
- The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. …
- Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Harnessing Big Data



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

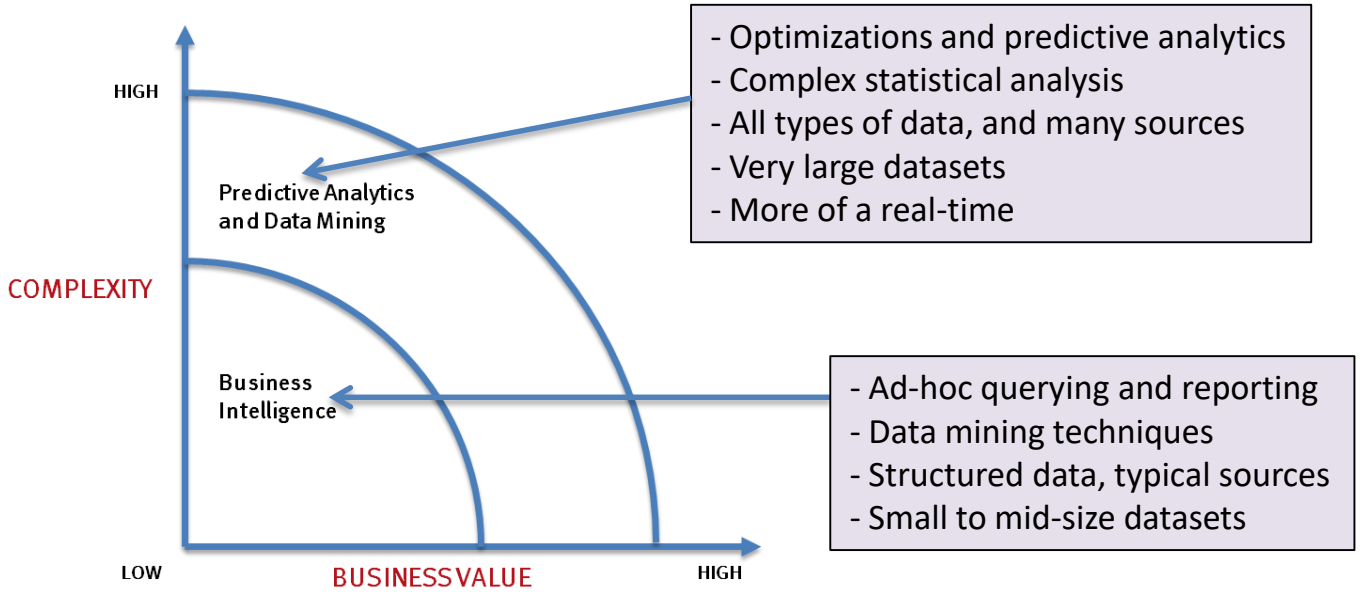
The Model Has Changed…

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

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

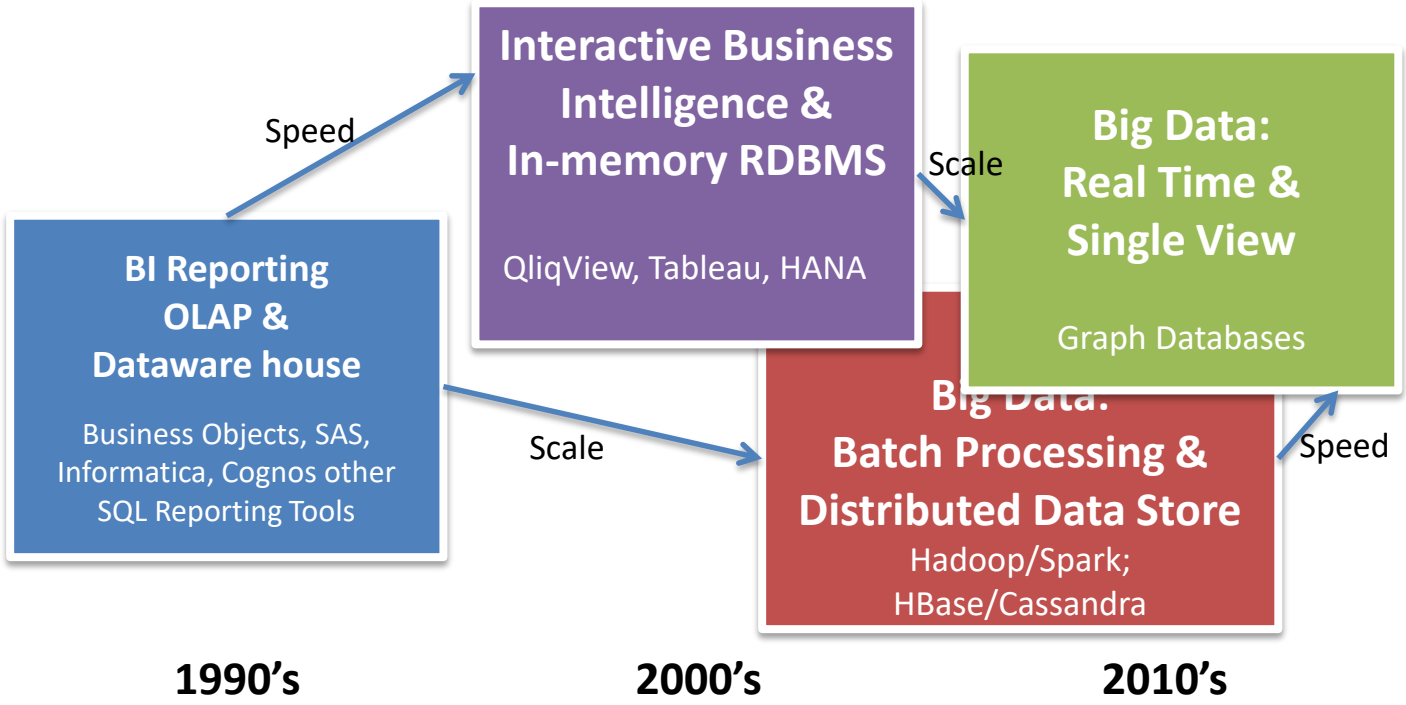
New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data



Organizations that do not learn to engage with Big Data, could find themselves left far behind their competitors, landing in the bustbin of history.

THE EVOLUTION OF BUSINESS INTELLIGENCE



What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of 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.

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.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

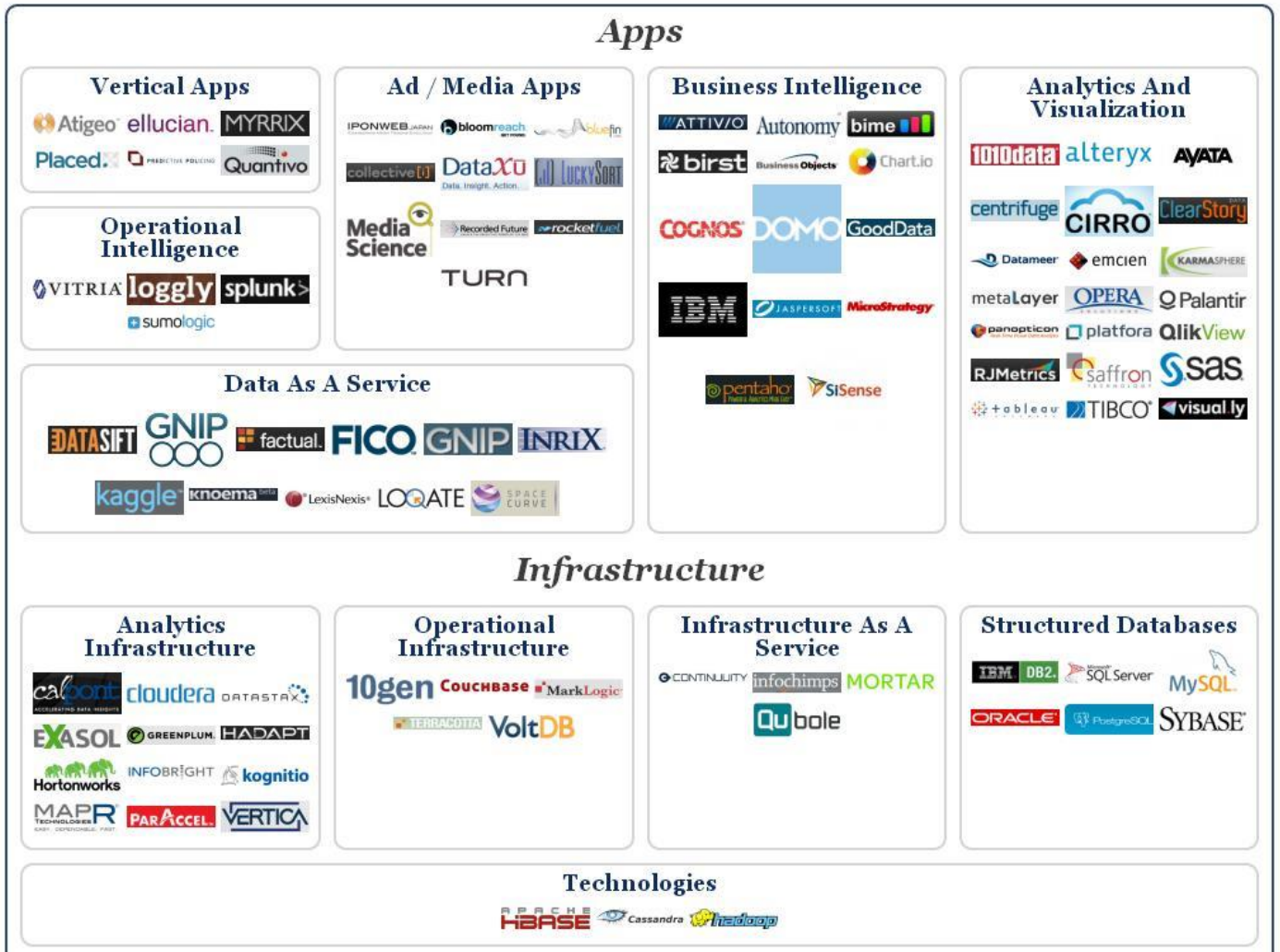
Structured data − Relational data, **Semi Structured data** − XML data, **Unstructured data** − Word, PDF, Text, Media Logs

Benefitting from Big Data

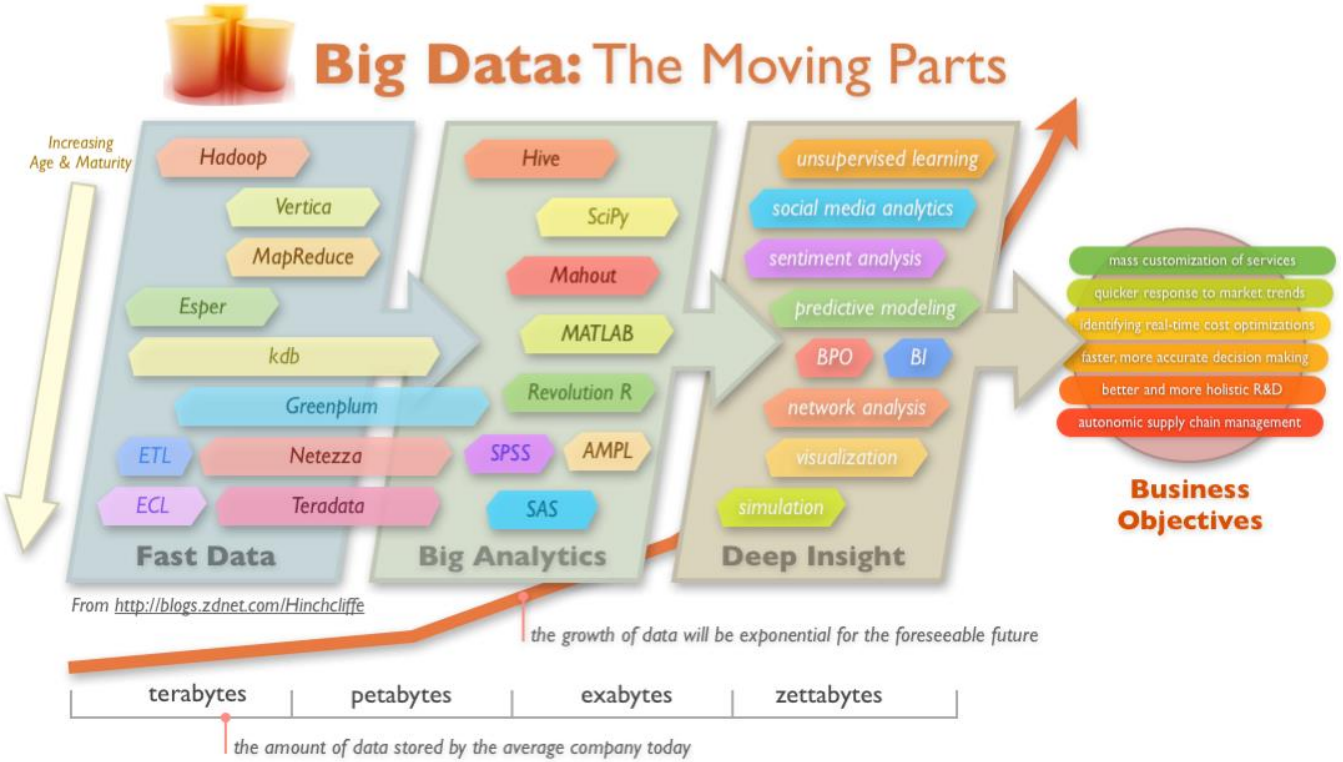
There are 3 major types of Big Data Applications

- Monitoring and Tracking
  - Consumer goods producers : Sentiments and needs of their customers
  - Industrial organizations : Track inventory in massive interlinked global supply chains
  - Factory owners: Monitor machine performance and do preventive maintenance
  - Utility companies: predict energy consumption, manage demand and supply
  - Information Technology: Track website performance and improve its usefulness
  - Financial organizations : to project trends better and make more effective and profitable bets, etc.
- Analysis and Insight
  - Political organizations: to micro-target voters and win elections
  - Police: To predict and prevent crimes
  - Hospitals : to better diagnose diseases and make medicine prescriptions
  - Advertising Agencies: To design more targeted marketing campaigns more quickly
  - Fashion Designers: To track trends and create more innovative products.
- Digital Product Development
  - Stock market feeds could be a digital product, Imagination is the limit

The Big Data Landscape



Big Data Technology



Big Data Technologies

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real-time and can protect data privacy and security.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology

– Operational and Analytical Systems

Operational Big Data

This 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. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

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.

These two classes of technology are complementary and frequently deployed together.

Operational vs. Analytical Systems

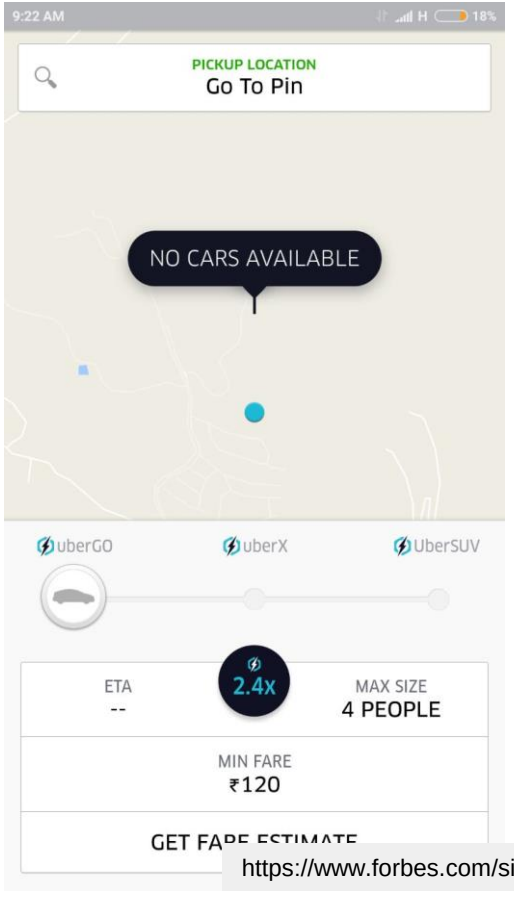
	Operational	Analytical
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce, MPP Database

Use cases

Use Case: Big Data in Oil & Gas Drilling

http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/

Use Case: Uber - Pay Surge Pricing if Battery is Low



https://www.forbes.com/sites/amitchowdhry/2016/05/25/uber-low-battery/#19762c0474b3

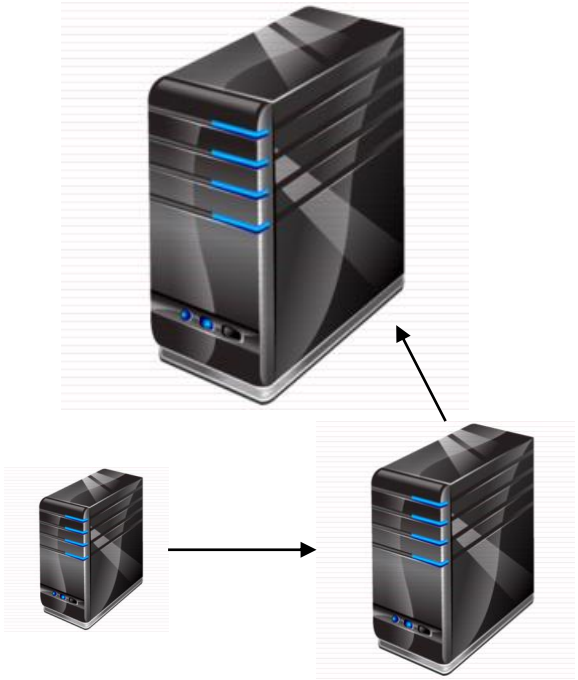
Big Data Challenges

Big Data Challenges: Size does matter

1KB	Kilobyte
1MB	Megabyte
1GB	Gigabyte
1TB	Terabyte
1PB	Petabyte
1EB	Exabyte
1ZB	Zettabyte
1YB	Yottabyte

1 GB = 1 hr
1 TB = 1024 hrs = 102 days
1 PB = 286 yrs **> 1 lifetime**
1 EB = 293K yrs

Big Data Challenges: Vertical Vs Horizontal Scaling



Vertical Scaling

Horizontal Scaling

Big Data Challenges

Scaling

Source: https://s-media-cache-ak0.pinimg.com/736x/10/0c/d0/100cd0da1c19e5d6f850ed23c3633714.jpg

Big Data Challenges: Scale of infrastructure

Image Source: https://datacenter.legrand.com

Further Reading

- A Brief History of Big Data Everyone Should Read
- Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity
- What is big data? - OpenSource.com & O'Reilly
- Uber Use Case
- 5 Big Data Use Cases To Watch
- Best Big Data Analytics Use Cases
- The 5 game changing big data use cases
- Big Data - The 5 Vs Everyone Must Know
- Top SlideShare Presentations on Big Data
- Google Data Center 360° Tour

How to store *huge* files?

Requirements?

- Efficient Access
- Effective utilization of space
- Redundancy (Failsafe)
  - Given: probability of 1 disk failing is 1% per year
  - What are the chances that 1 out of 10^3 disk fails at a data center?

HDFS

Hadoop distributed File System

HDFS

- Data storage system used by Hadoop
  - Hadoop: Project to develop open-source software for reliable, scalable, distributed computing★
  - Will discuss Hadoop later
- Components
- Architecture
- Tasks / Services

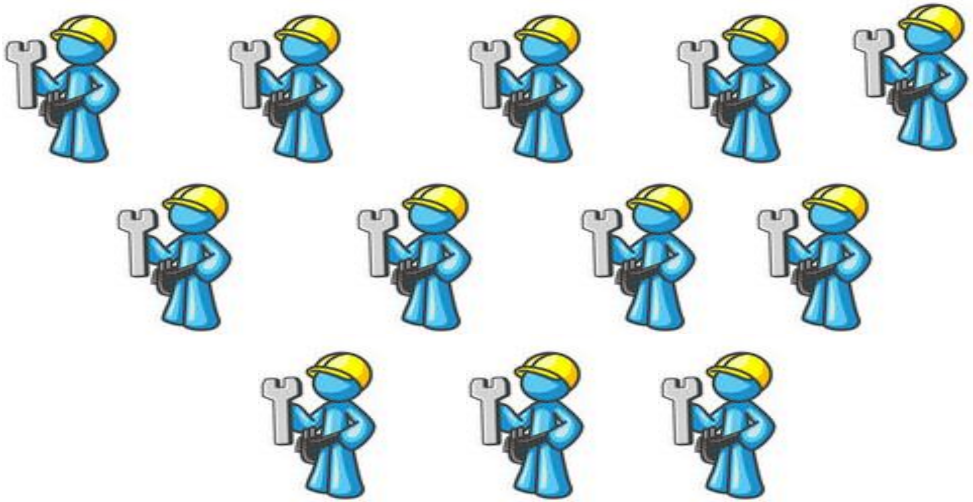
★ http://hadoop.apache.org/

Components of **HDFS**

Secondary NameNode

Active NameNode

Standby NameNode



DataNodes

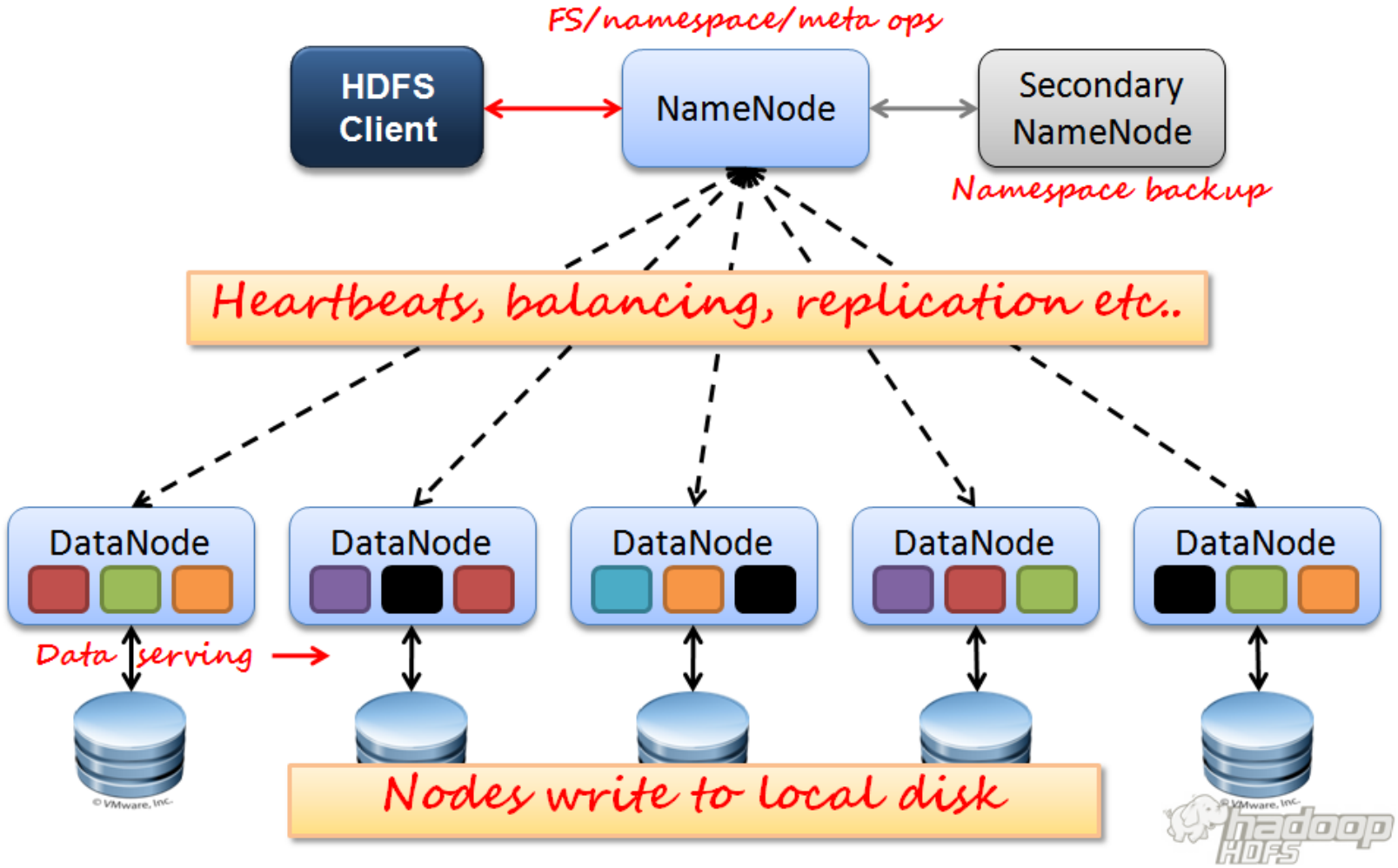
Terminology

- **HDFS**: Hadoop Distributed File System
- **Datanode**: A DataNode stores data in HDFS.
- **Namenode**: The centerpiece of an HDFS file system.
  - Keeps the directory tree of all files in the file system
  - Tracks where across the cluster the file data is kept.
    - Does not store the data of these files itself.
  - Active : Actively serving request
  - Standby: Becomes Active if the current Active node fails

Terminology

- **Secondary Namenode**:
  - helper node for namenode
  - Puts a checkpoint in filesystem which will help Namenode to function better

Architecture



Storing file on HDFS

Motivation: Reliability, Availability , Network Bandwidth

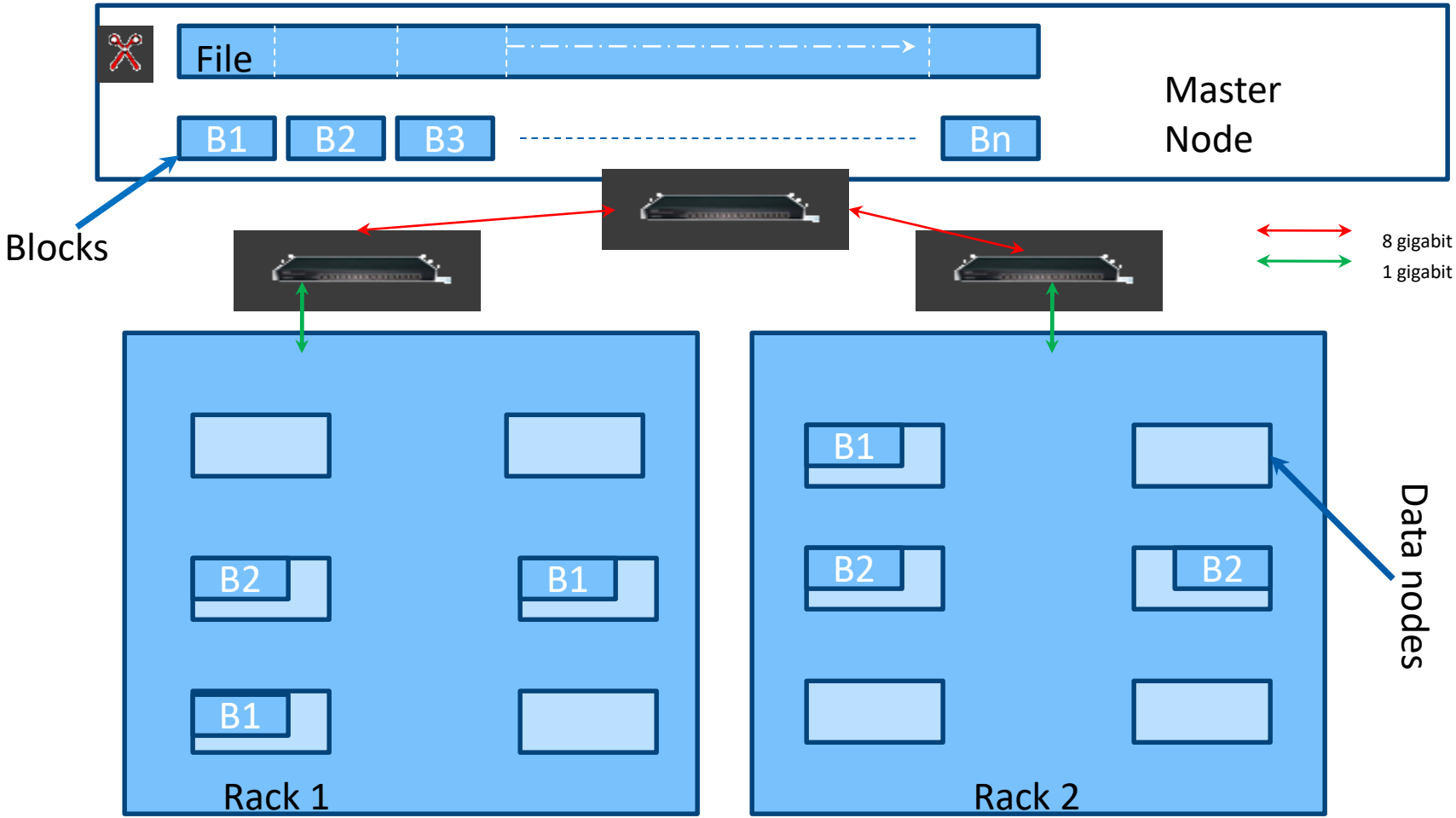
- The input file (say 1 TB) is split into smaller chunks/blocks of 128 MB
- The chunks are stored on multiple nodes as independent files on data nodes

Storing file on HDFS

- To ensure that data is not lost, data can typically be replicated on:
  - local rack
  - remote rack (in case local rack fails)
  - remote node (in case local node fails)
  - randomly
- Default replication factor is 3

Storing file on HDFS

- Default replication factor is 3
  - first replica of a block will be stored on a local rack
  - the next replica will be stored on a remote rack
  - the third replica will be stored on the same remote rack but on a different Datanode
  - Why?
- More replicas?
  - the rest will be placed on random Datanodes
  - As far as possible, no more than two replicas are kept on the same rack


File
B1
B2
B3
Bn
Master Node
Blocks
8 gigabit
1 gigabit
B2
B1
B1
Rack 1
B1
B2
B2
Rack 2
Data nodes

Tasks of NameNode

- Manages File System
  - mapping files to blocks and blocks to data nodes
- Maintaining status of data nodes
  - Heartbeat
    - Datanode sends heartbeat at regular intervals
    - If heartbeat is not received, datanode is declared dead
  - Blockreport
    - DataNode sends list of blocks on it
    - Used to check health of HDFS

NameNode Functions

- Replication
  - On Datanode failure
  - On Disk failure
  - On Block corruption
- Data integrity
  - Checksum for each block
  - Stored in hidden file
- Rebalancing - balancer tool
  - Addition of new nodes
  - Decommissioning
  - Deletion of some files

HDFS Robustness

- Safemode
  - At startup: No replication possible
  - Receives Heartbeats and Blockreports from Datanodes
  - Only a percentage of blocks are checked for defined replication factor
- Replicate blocks wherever necessary

All is well ☺ ➔ Exit Safemode

HDFS Summary

- Fault tolerant
- Scalable
- Reliable
- File are distributed in large blocks for
  - Efficient reads
  - Parallel access

Questions?