Chapter -1
Introduction to Big Data and Big Data Analytics

Evolution of Big Data

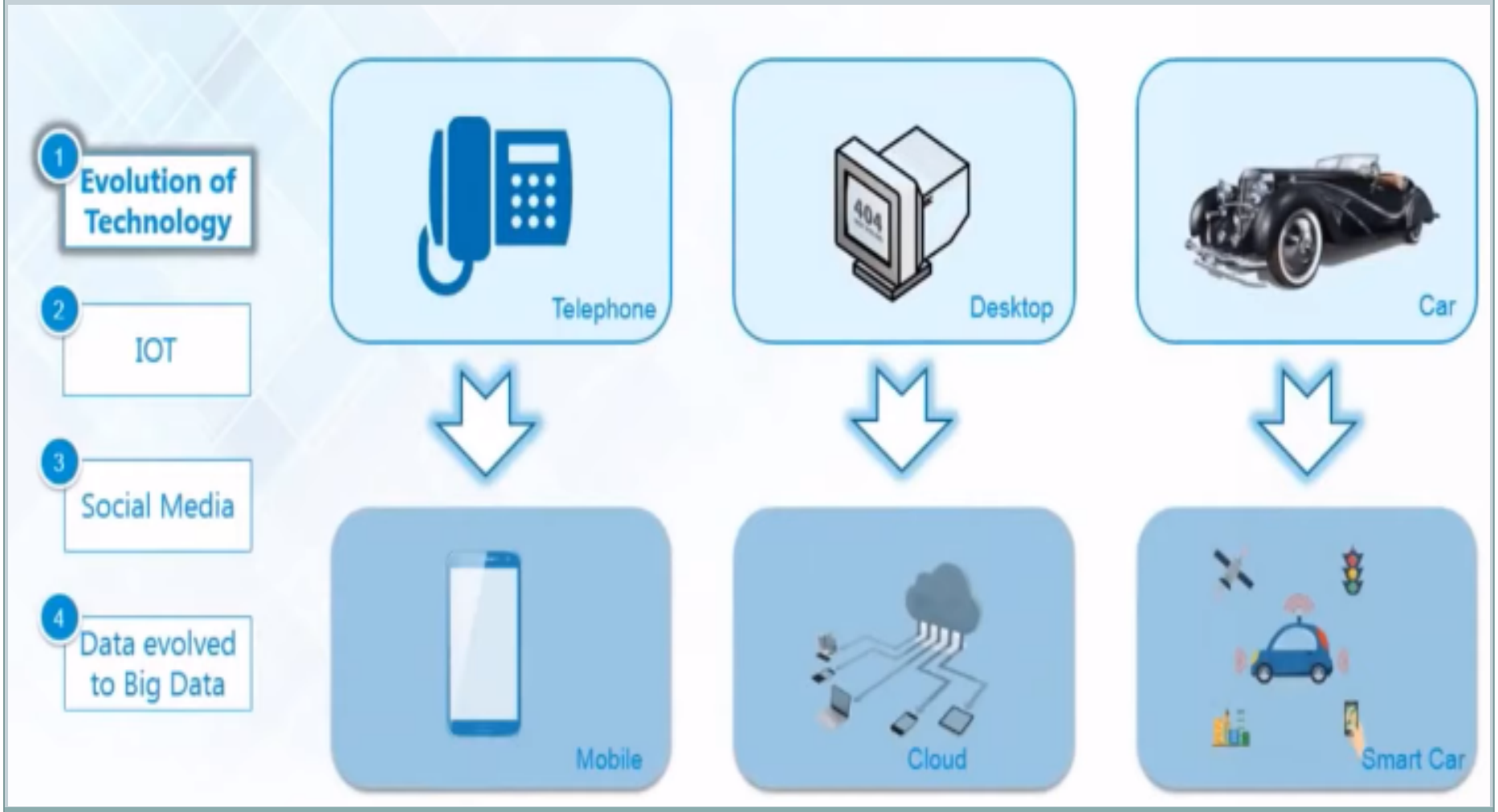
- **The Model of Generating/Consuming Data has Changed**

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

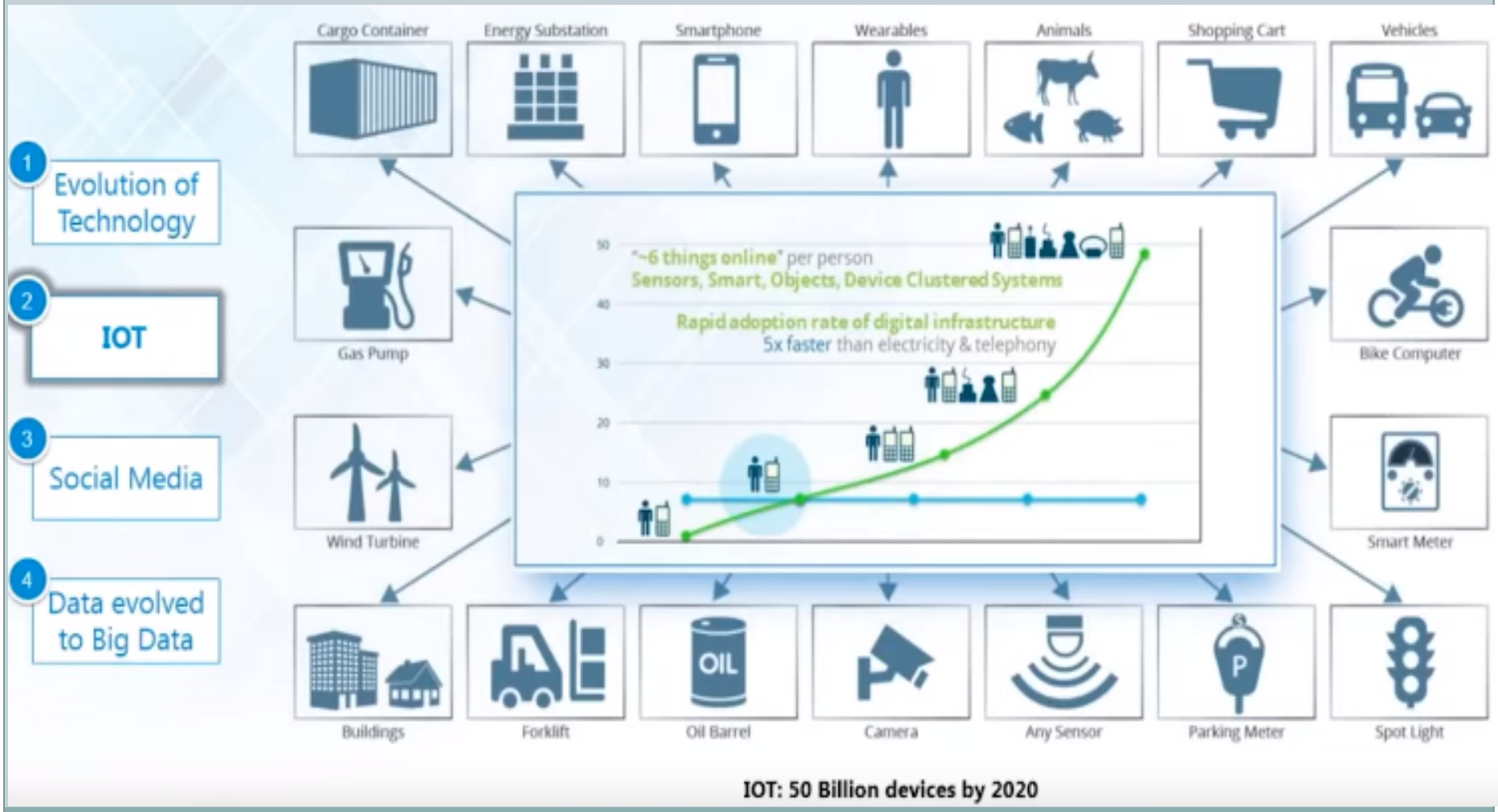
New Model: all of us are generating data, and all of us are consuming data

Unit of Data size	Exact size	Approximate Size	Examples	
KB (kilobyte)	2^{10} or 1024 bytes	(10^3 or one thousand) bytes	A typical joke =1KB	
MB(megabyte)	2^{20} bytes	(10^6 or one million) bytes	Complete work of Shakespeare =5MB	
GB (gigabyte)	2^{30} bytes	(10^9 or one billion) bytes	Ten yards of books on a shelf = 1GB	
TB (terabyte)	2^{40} bytes	(10^{12} or one trillion) bytes	All the X-rays for a large hospital =1TB Tweets; created daily =121TB;	
PB (peta byte)	2^{50} bytes	(10^{15} or one quadrillion) bytes	All U.S. academic research libraries = 2PB Data processed in a day by Google =24PB	BIG DATA
EB (exa byte)	2^{60} bytes	(10^{18} or one Quintillion) bytes	Total global data created in 2006 = 161EB	
ZB (zetta byte)	2^{70} bytes	(10^{21} or one Sextillion) bytes	Total amount of global data created in 2012 = 2.7 ZB and expected 44 ZB by 2020	
YB (yotta byte)	2^{80} bytes	(10^{24} or one Septillion) bytes		

Evolution of Big Data by technology

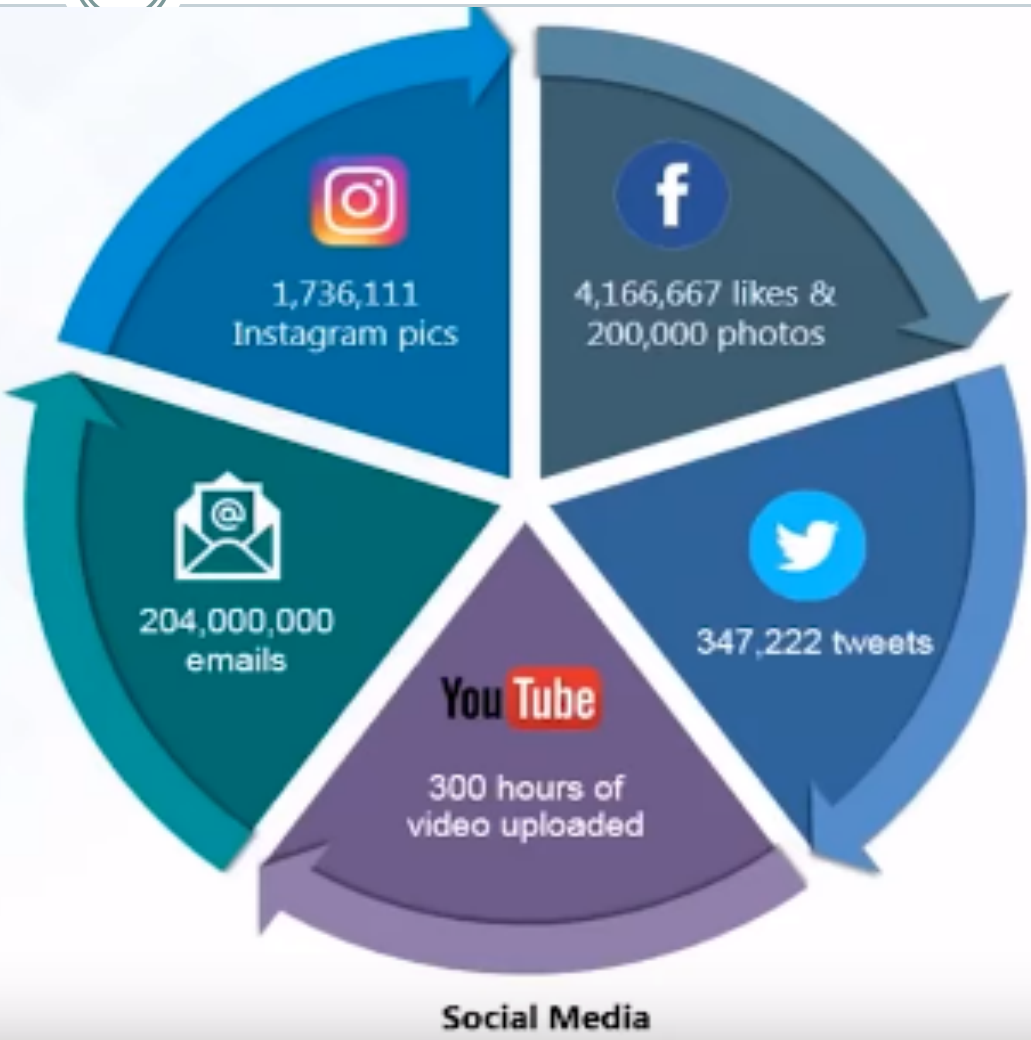


Evolution of Big Data by Internet Of Things



Evolution of Big Data by Social Media

1. Evolution of Technology
2. IOT
3. **Social Media**
4. Data evolved to Big Data



Evolution of Big Data by other factors

1. Evolution of Technology
2. IOT
3. Social Media
4. Other Factors

Big Data sources

- Human Generated Data
  - is emails, documents, photos and tweets. We are generating this data faster than ever. Just imagine the number of videos uploaded to You Tube and tweets swirling around. This data can be Big Data too.
- Machine Generated Data
  - is a new breed of data. This category consists of sensor data, and logs generated by 'machines'
  - such as email logs, click stream logs, etc. Machine generated data is orders of magnitude larger than Human Generated Data.

Big Data sources

- **Web Data**
  - **Social media data** : Sites like Facebook, Twitter, LinkedIn generate a large amount of data
  - **Click stream data** : when users navigate a website, the clicks are logged for further analysis (like navigation patterns). Click stream data is important in on line advertising and E-Commerce

12+ TBs of tweet data every day

25+ TBs of log data every day

? TBs of data every day

Big Data sources

sensor data : sensors embedded in roads to monitor traffic and misc.

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

What is Big Data?

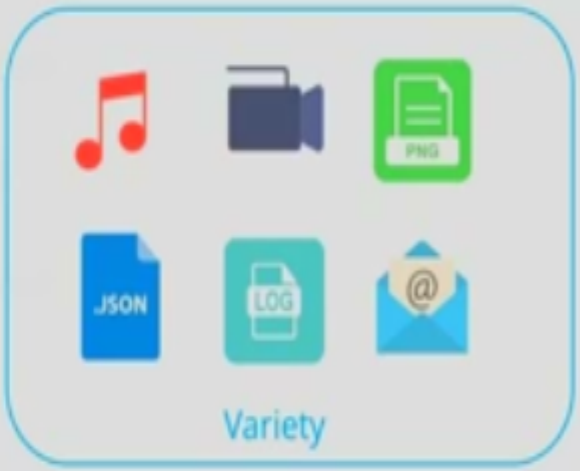
Big data

is the term for a collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Real world examples of Big Data

- Facebook : has 40 PB of data and captures 100 TB / day
- Yahoo : 60 PB of data
- Twitter : 8 TB / day
- EBay : 40 PB of data, captures 50TB/ day

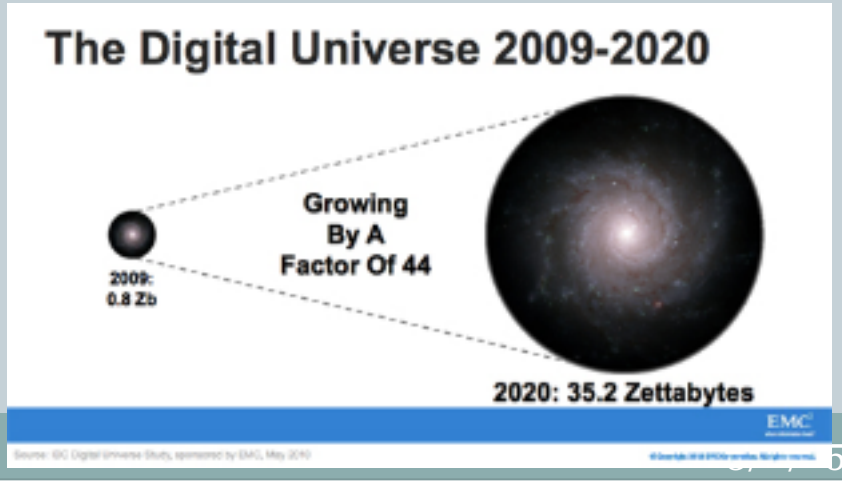
Characteristics of Big Data(5 Vs of Big data)



Characteristics of Big Data(5 Vs of Big data)

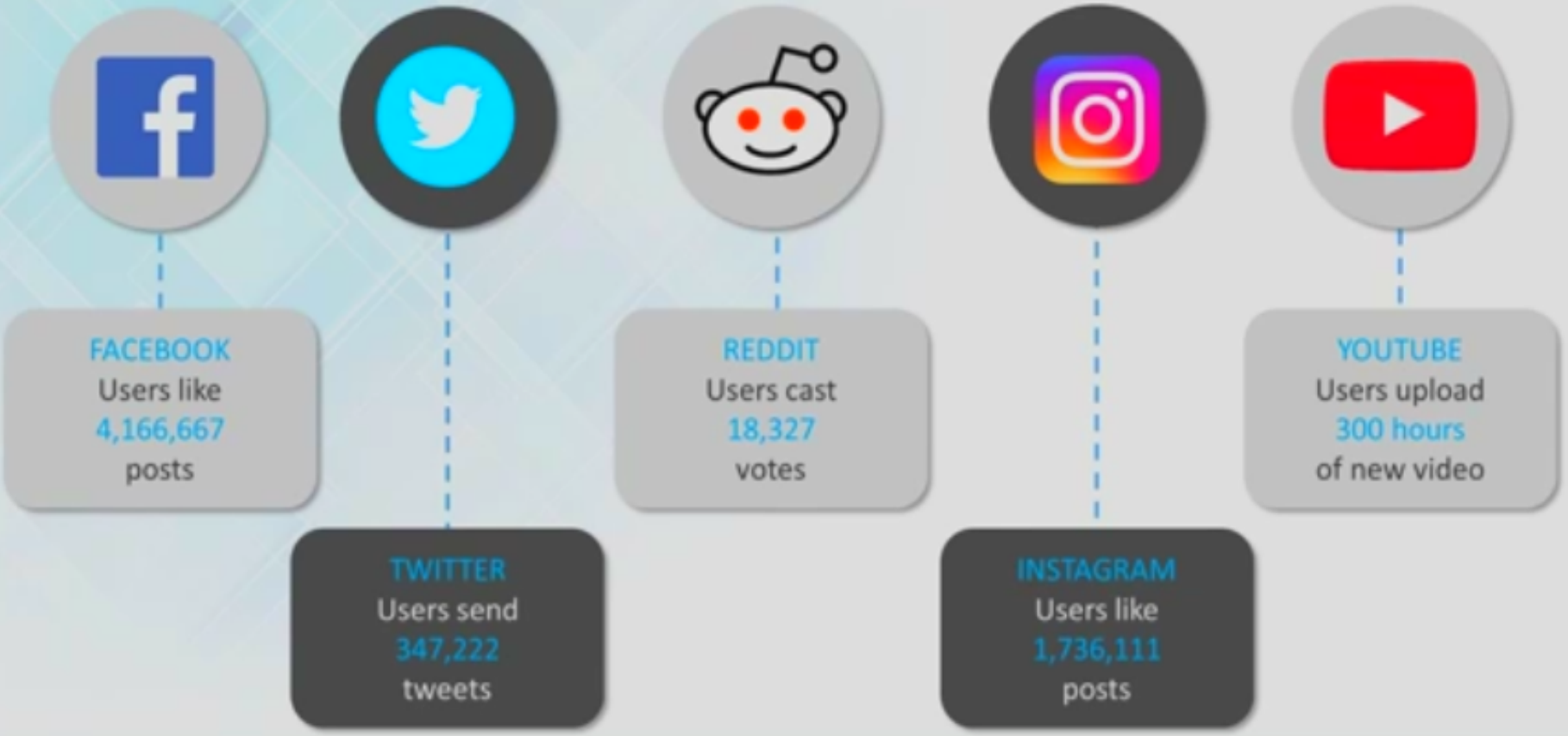
- 1st V-volume

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



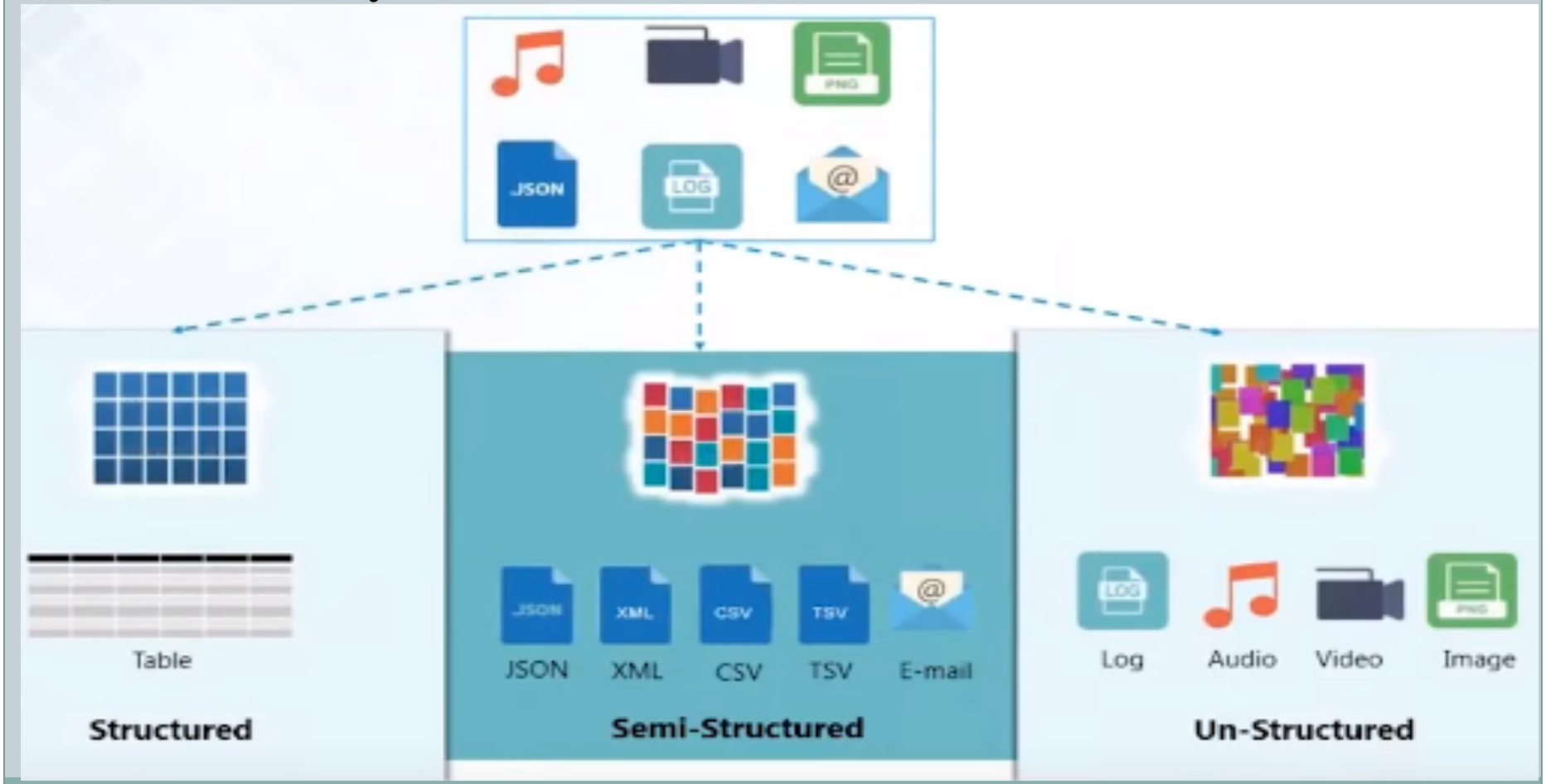
Characteristics of Big Data(5 Vs of Big data)

- 2nd V-velocity: **Data is being generated at every minute**



Characteristics of Big Data(5 Vs of Big data)

- 3rd V-Variety: **different kinds of data generated from various sources**



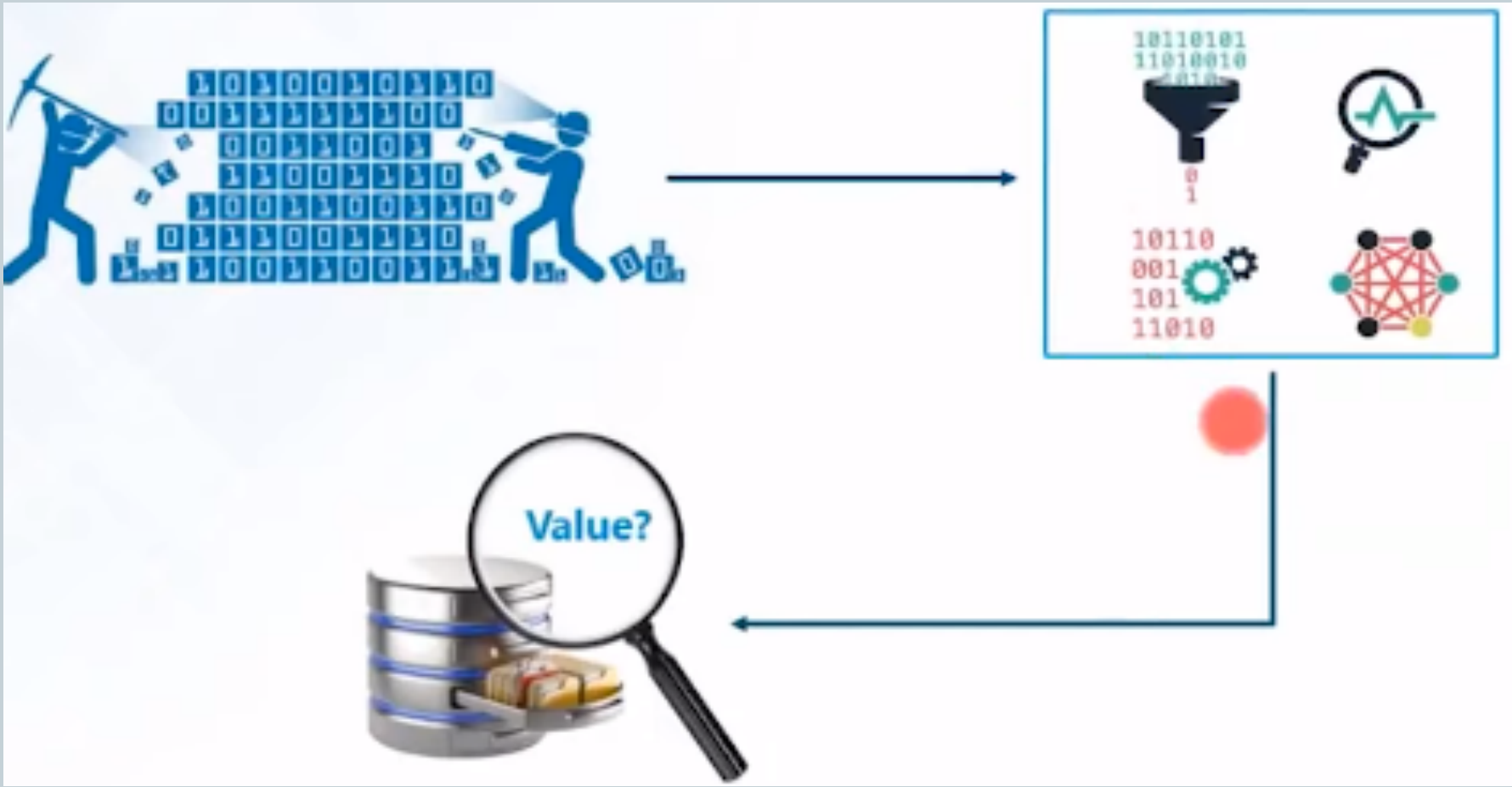
Characteristics of Big Data(5 Vs of Big data)

- 4th V - Veracity: **uncertainties and inconsistencies in big data**

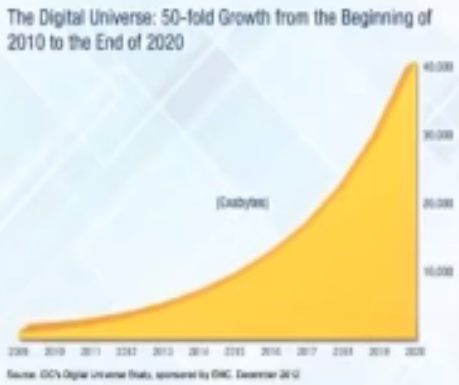
Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Characteristics of Big Data(5 Vs of Big data)

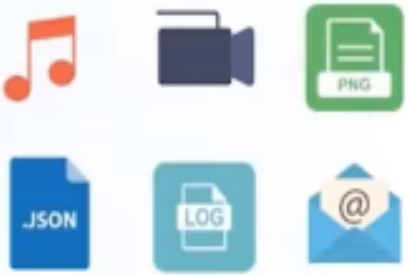
- 5th V - Value: **Mechanism to bring correct meaning out of the data**



Characteristics of Big Data(5 Vs of Big data)



Volume



Different kinds of data is being generated from various sources

Variety

Data is being generated at an alarming rate

Velocity

Mechanism to bring the correct meaning out of the data

Value

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

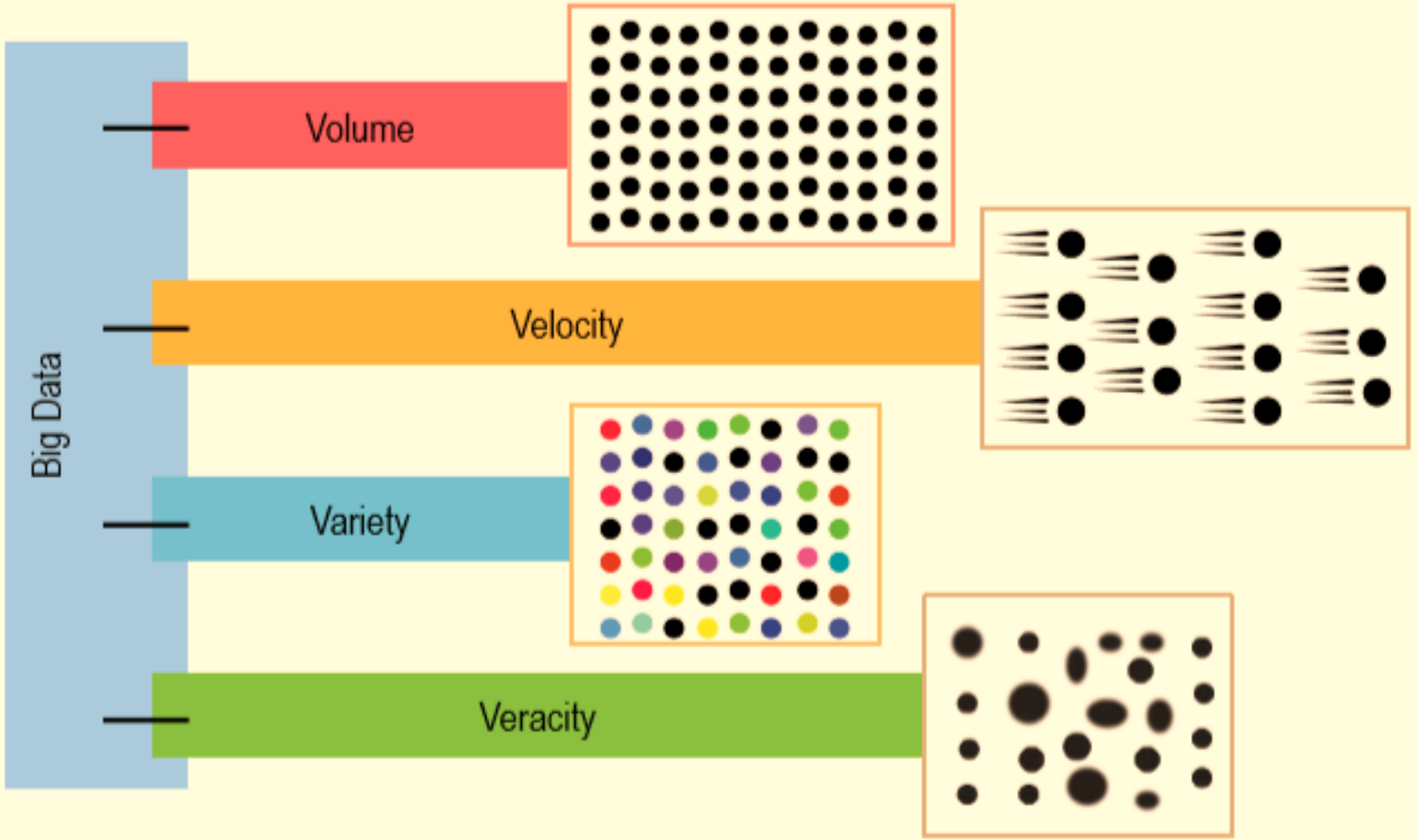
Uncertainty and inconsistencies in the data

Veracity

... ...

V's associated with Big Data may grow with time

Characteristics of Big Data(5 Vs of Big data)



Characteristics of Big Data Systems

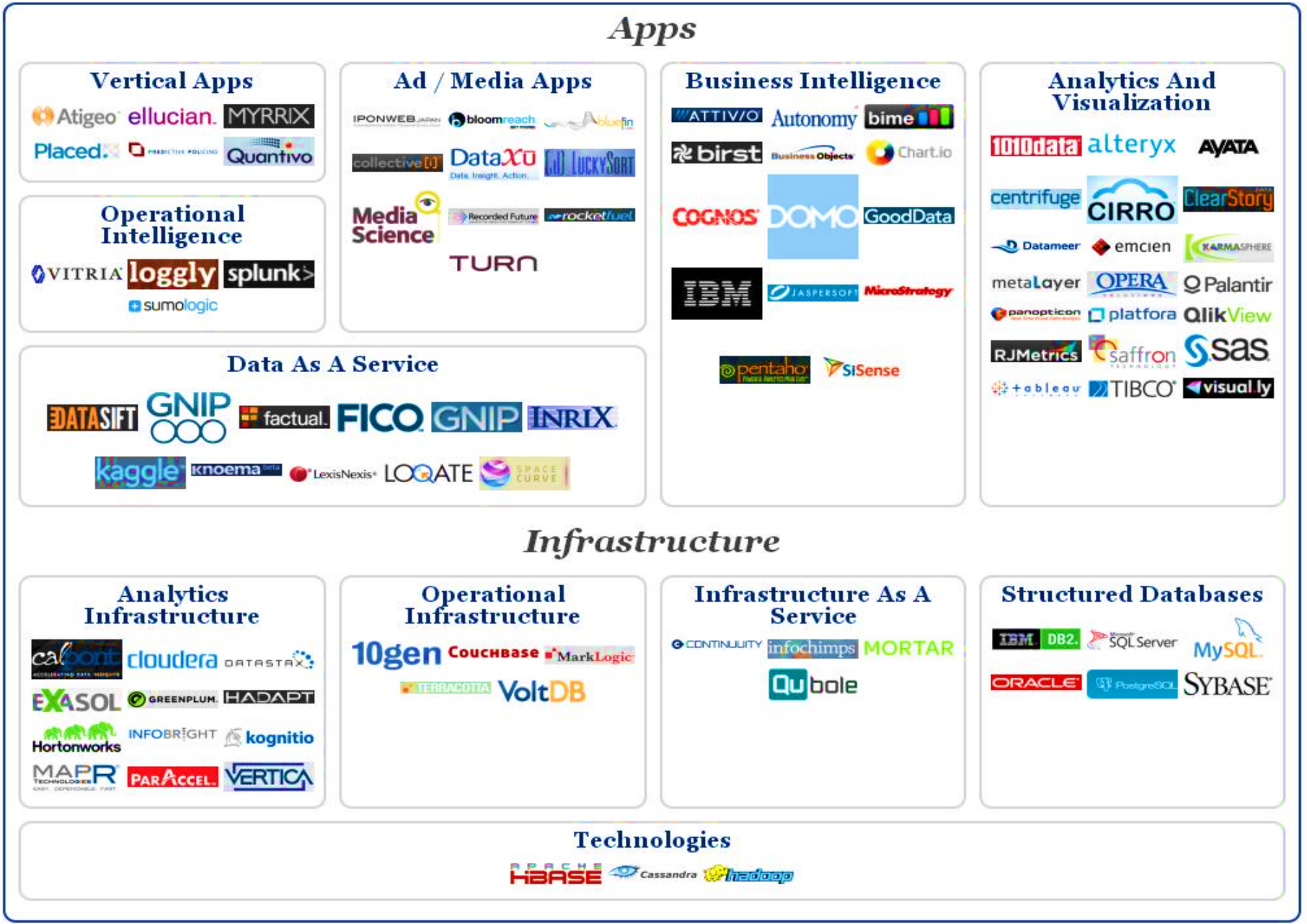
- **Distributed Computing: Horizontal Scaling instead of Vertical Scaling**
  - Computations are done ***closer*** to where data is stored
  - Instead of centrally located parallel computing architecture with super-computing capabilities (Giga/Teraflops), low capacity distributed storage/computing solution is used
- **Use of Low Cost Commodity hardware**
  - Big Data solutions use large number of low cost, commodity hardware, organized in clusters to carry out storage/computing tasks
- **Reliability, Fault Tolerance and Recovery**
  - Individual nodes can fail anytime, so to ensure reliability, data is replicated across multiple nodes
- **Scaling with Demand**
  - The solutions are scalable and allow cluster sizes to grow as per requirement
- **Storage of unstructured Data**
  - Traditional RDBMS systems require well defined schema to be created, before data can be stored (schema on write)
  - New data storage paradigm – 'NoSQL' has evolved to cater to need to store *any type* of data. This provides for schema on read i.e. schema is applied when data is read.
- **No Archiving**
  - Data is always online, so no archiving. The big data solutions ***do not assume*** what data queries will be using, so rule is to *store all data* in raw form.

Traditional DB vs Big Data

Traditional data base/ data warehouse	Big Data

• Data
  ○ TB to PB
  ○ Only structured
• Hardware
  ○ big central servers
  ○ Expensive
  ○ Hardware reliability
  ○ Limited scalability
• Software
  ○ Centralized
  ○ Schema based
  ○ Oracle/mysql/sql server
	• Data
	  ○ PB to ZB
	  ○ structured and unstructured
	• Hardware
	  ○ computer clusters
	  ○ Cost effective
	  ○ Unreliable HW
	  ○ Scales further
	• Software
	  ○ Distributed
	  ○ Not schema based
	  ○ Hadoop

Big data tools



What is Big data analytics

"Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights"

Stages in Big data analytics

Big data analytics goals

Big data analytics goals

1.Making organizations more smarter and efficient

Big data analytics goals

2 **Optimize Business Operations by analysing customer behaviour**

- Analysing all the clicks of every visitor on a website
- Studying the paths leading them to buy products
- Customer Satisfaction

Amazon uses customer click-stream data and historical purchase data of more than 300 million customers and each user is shown customized results on customized web pages.

Big data analytics goals

3 Cost Reduction

Parkland Hospital uses analytics and predictive modelling to identify high-risk patients and predict likely outcomes once patients are sent home. As a result, Parkland reduced 30-day readmissions for patients with heart failure, by 31 percent, saving $500,000 annually.

Big data analytics goals

4 Next Generation Products

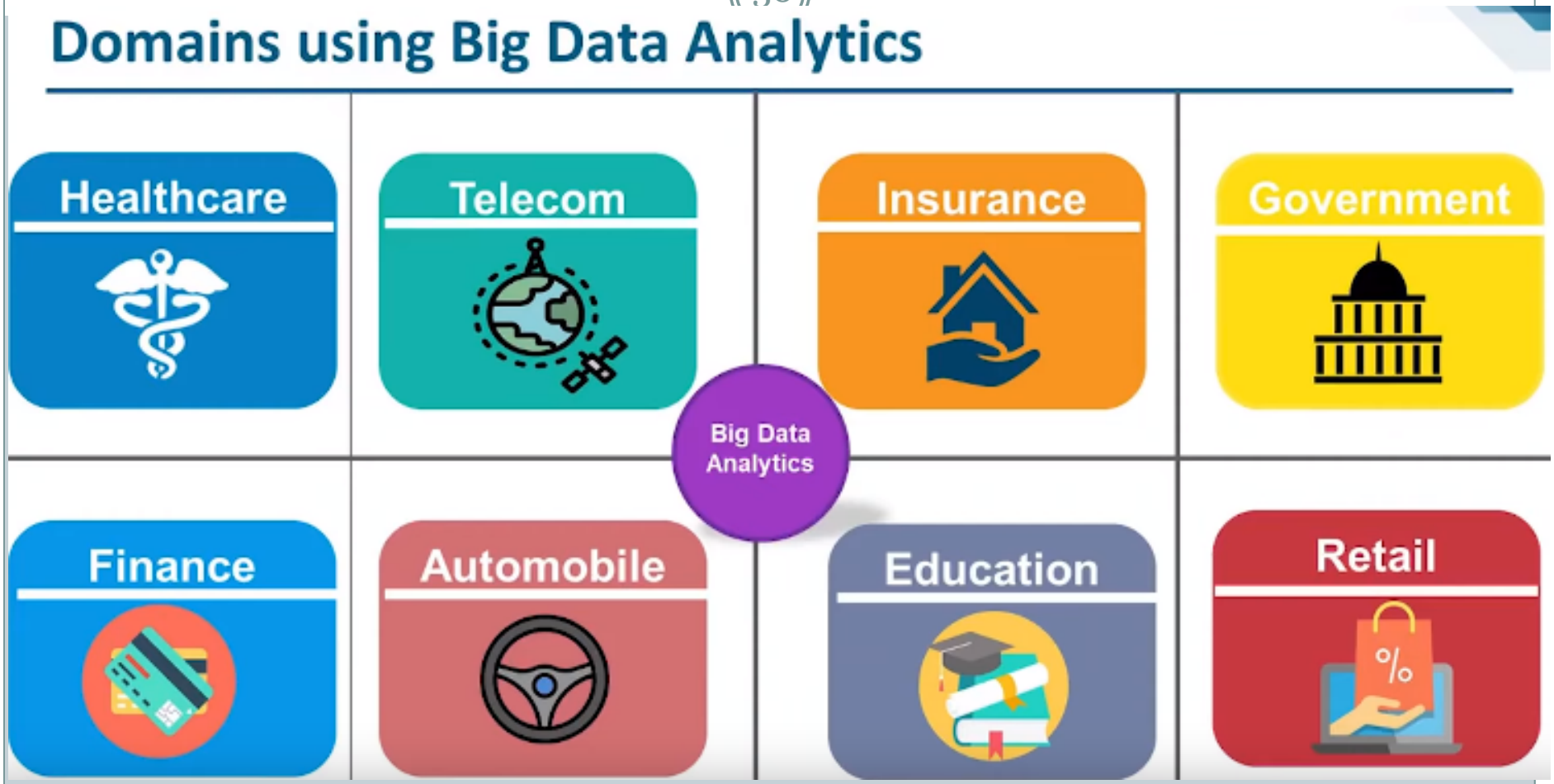
Big Data tools are used to operate Google's Self Driving Cars. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

Netflix launched the seasons of its TV show House of Cards based on the user reviews, ratings and viewership.

NETFLIX

A smart yoga mat has sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice.

Big data analytics application domains



Big data analytics use cases

Use Case 1 - Starbucks

Starbucks uses behavioural analytics to cater to its customers

Starbucks gather a lot of info about their customers' coffee-buying habits from their preferred drinks to what time of day they're usually ordering

The company directs exciting offers and coupons to their customers and ensures to maintain their interest

Big data analytics use cases

Use Case 2 – Procter & Gamble

P&G uses Market Basket Analysis and price optimization to optimize their products

Procter&Gamble

Market Basket Analysis, analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets"

P&G

The company uses simulation models and predictive analysis in order to create the best design for its products.

Big data analytics use cases

Walmart boosted its sales by leveraging the power of Big Data

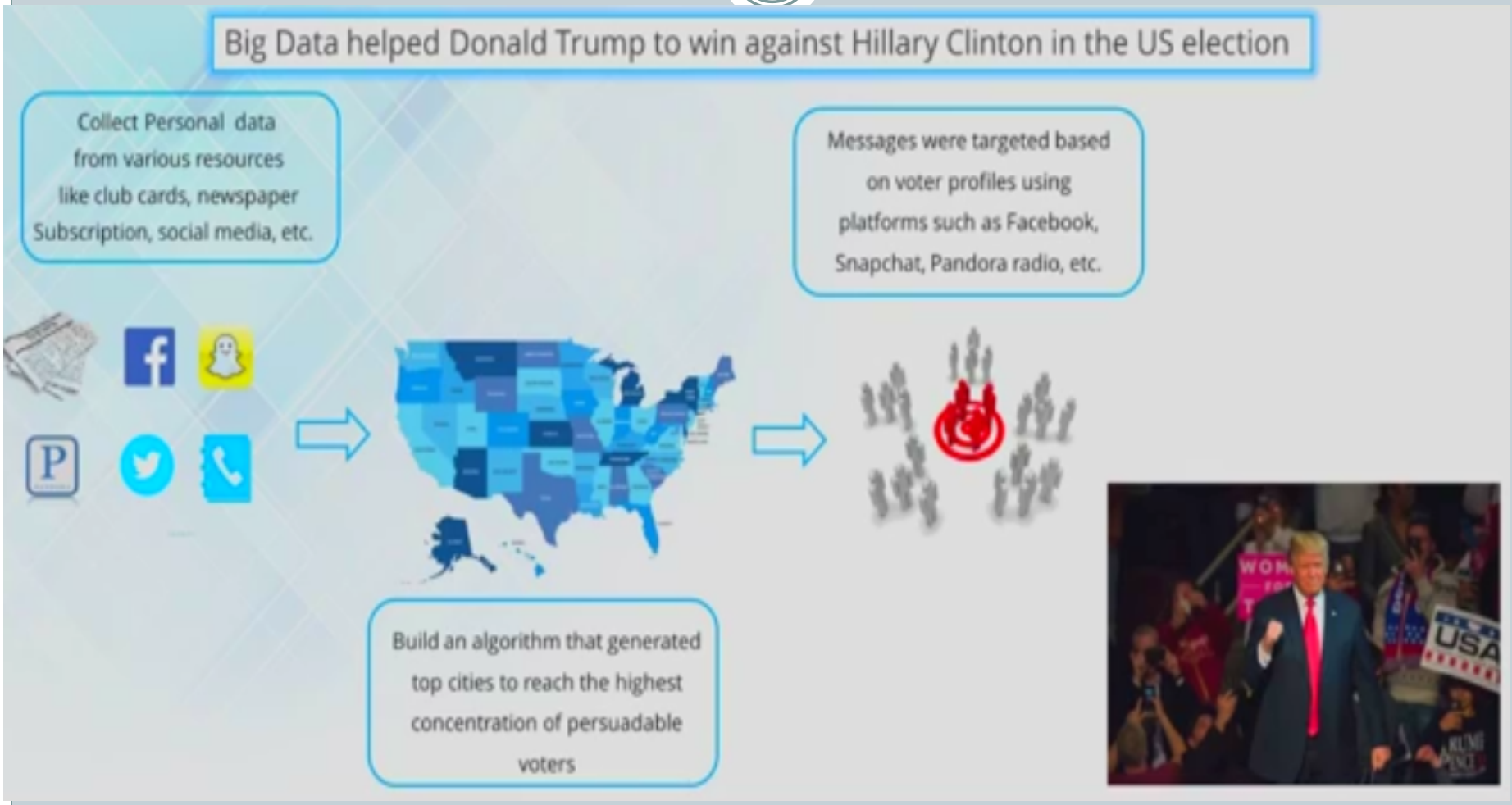
While forecasting the demand for emergency supplies for approaching Hurricane Sandy, they gain some amazing insights:

Walmart

Extra supplies of Strawberry Pop Tarts were dispatched to stores in Hurricane Sandy's path in 2012, and sold extremely well

Along with flashlights and emergency equipment, they found an upsurge in sales of strawberry Pop Tarts

Big data analytics use cases



Big data analytics use cases

Apixio uses big data analytics to improve healthcare decision

APIXIO

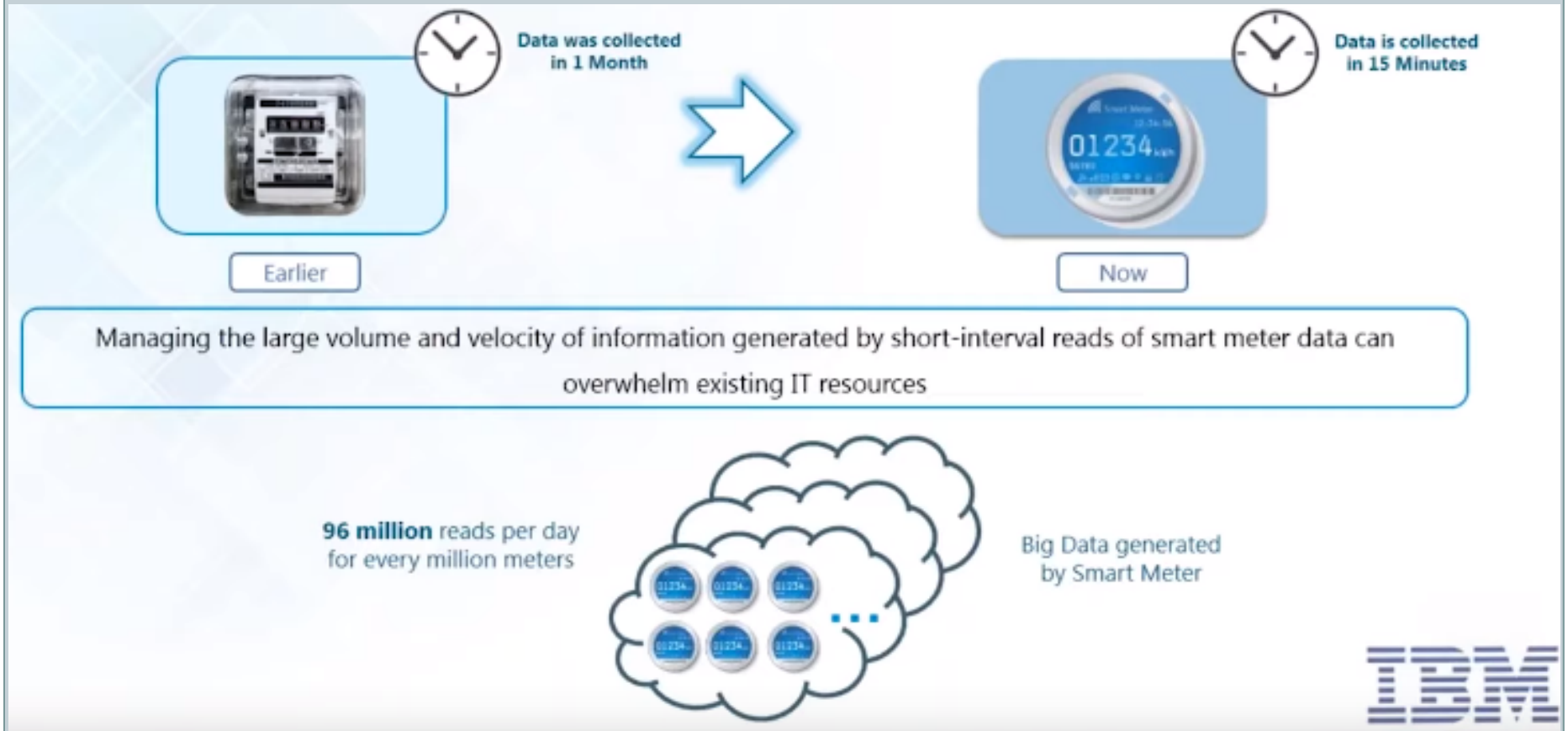
80% of medical and clinical information about patients is in unstructured format, such as written physician notes

Analysis of medical data using variety of different methodologies & algorithms that are machine learning based and have NLP capabilities

The patient data model generated is aggregated across population to derive larger insights like disease prevalence, treatment patterns, etc.

Big data analytics use cases

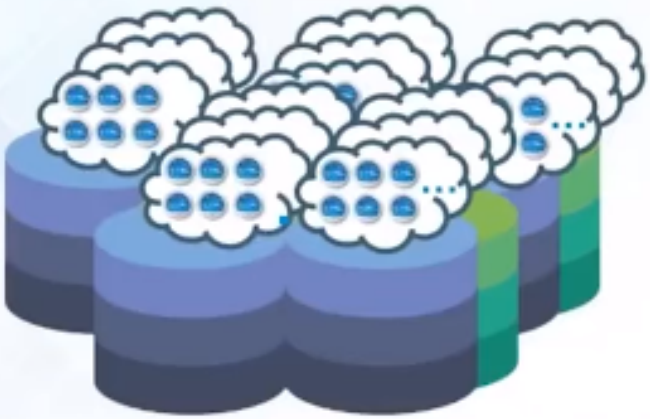
- IBM Big data analytics – Big data collected by smart meters



Big data analytics use cases

- IBM Big data analytics – problem with smart meter big data

To manage and use this information to gain insight, utility companies must be capable of high-volume data management and advanced analytics designed to transform data into actionable insights.



Store

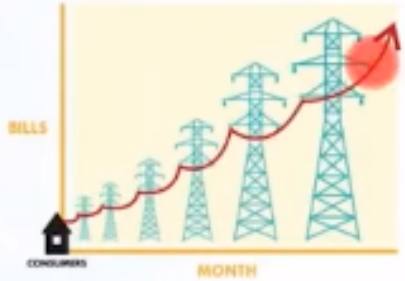
Analyze

IBM

Big data analytics use cases

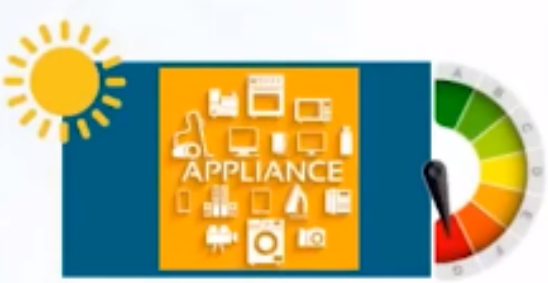
- IBM Big data analytics – how smart meter big data analysed

Before analyzing Big Data

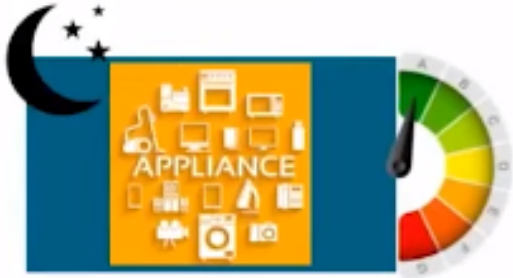


Energy utilization and billing has increased

After analyzing Big Data



During peak-load the users require more energy



During off-peak times the users required less energy

Time-of-use pricing encourages cost-savvy retail like industrial heavy machines to be used at off-peak times

Big data analytics use cases

- IBM Big data analytics – IBM smart meter solution

IBM offers an integrated suite of products designed to enable IT to leverage big data in a variety of ways that can contribute to the success of energy companies

IBM Solution

1. Managing smart meter data
2. Monitoring the distribution grid
3. Optimizing unit commitment
4. Optimizing energy trading
5. Forecasting and scheduling loads

Types of Big data analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analytics

Types of Big Data Analytics

What is happening now based on incoming data.

Google Analytics Tool is the best example for descriptive analysis. A business gets result from the web server through the tool which help understand what actually happened in the past and validate if a promotional campaign was successful or not based on basic parameters like page views.

Types of Big data analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analytics

Types of Big Data Analytics

What might happen in the future

For example, Southwest Airlines analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction, thus allowing the airlines to the necessary repairs before its schedule.

Types of Big data analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analytics

Types of Big Data Analytics

What action should be taken.

Google's self-driving car is a perfect example of prescriptive analytics. It analyses the environment and decides the direction to take based on data.

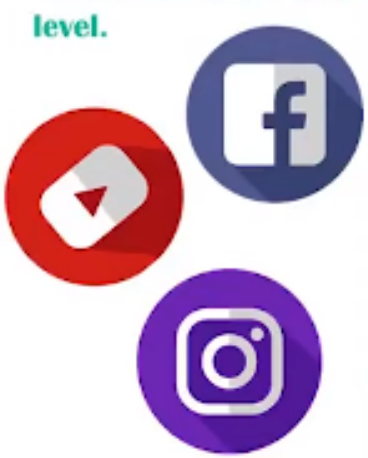
Types of Big data analytics

1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analytics

Types of Big Data Analytics

Why did it happen

For a Social Media marketing campaign, you can use diagnostic analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc. and analyse the failure and success rate of the campaign at a fundamental level.



Challenges/problems with Big data

Problem 1: Storing exponentially growing huge datasets

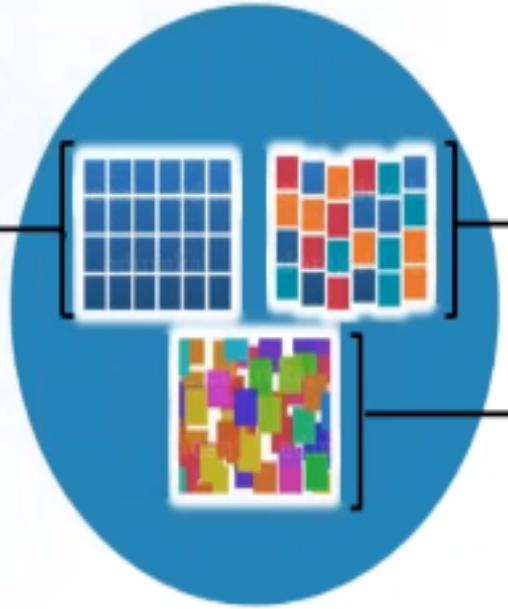
- Data generated in past **2 years** is more than the previous history in total
- By 2020, total digital data will grow to **44 Zettabytes** approximately
- By 2020, about **1.7 MB** of new info will be created every second for every person

Challenges/problems with Big data

Problem 2: Processing data having complex structure

Structured

- Organized data format
- Data schema is fixed
- Ex: RDBMS data, etc.



Semi – Structured

- Partial organized data
- Lacks formal structure of a data model
- Ex: XML & JSON files, etc.

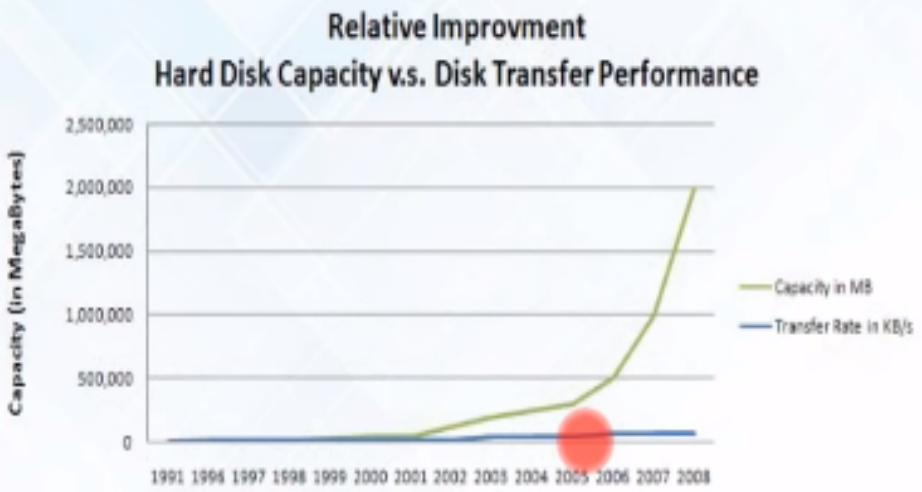
Unstructured

- Un-organized data
- Unknown schema
- Ex: multi-media files, etc.

Challenges/problems with Big data

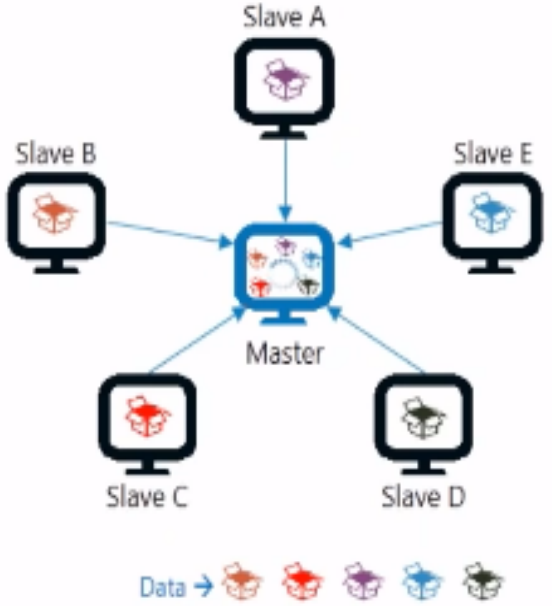
Problem 3: Processing data faster

The data is growing at much faster rate than that of disk read/write speed



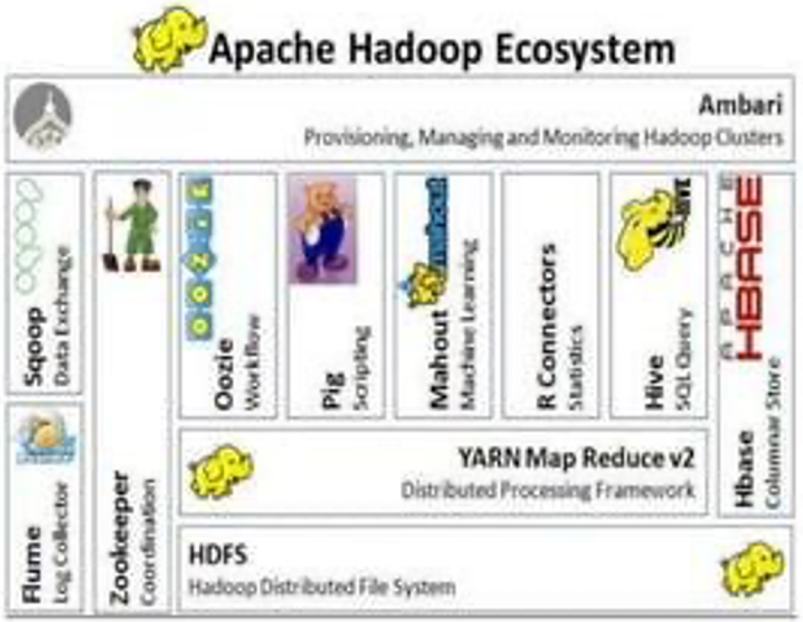
Source: *Tom's Hardware*

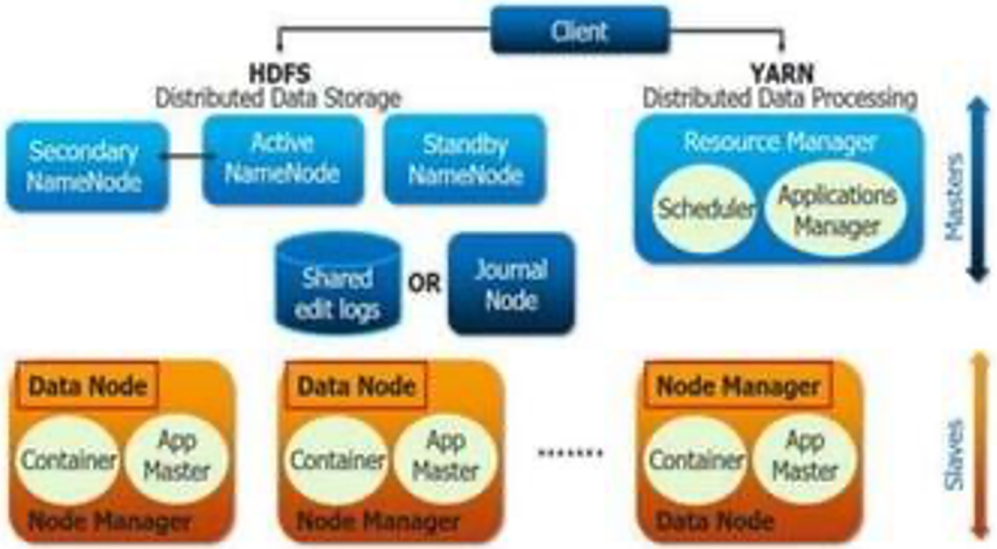
Bringing huge amount of data to computation unit becomes a bottleneck



Apache hadoop [6]

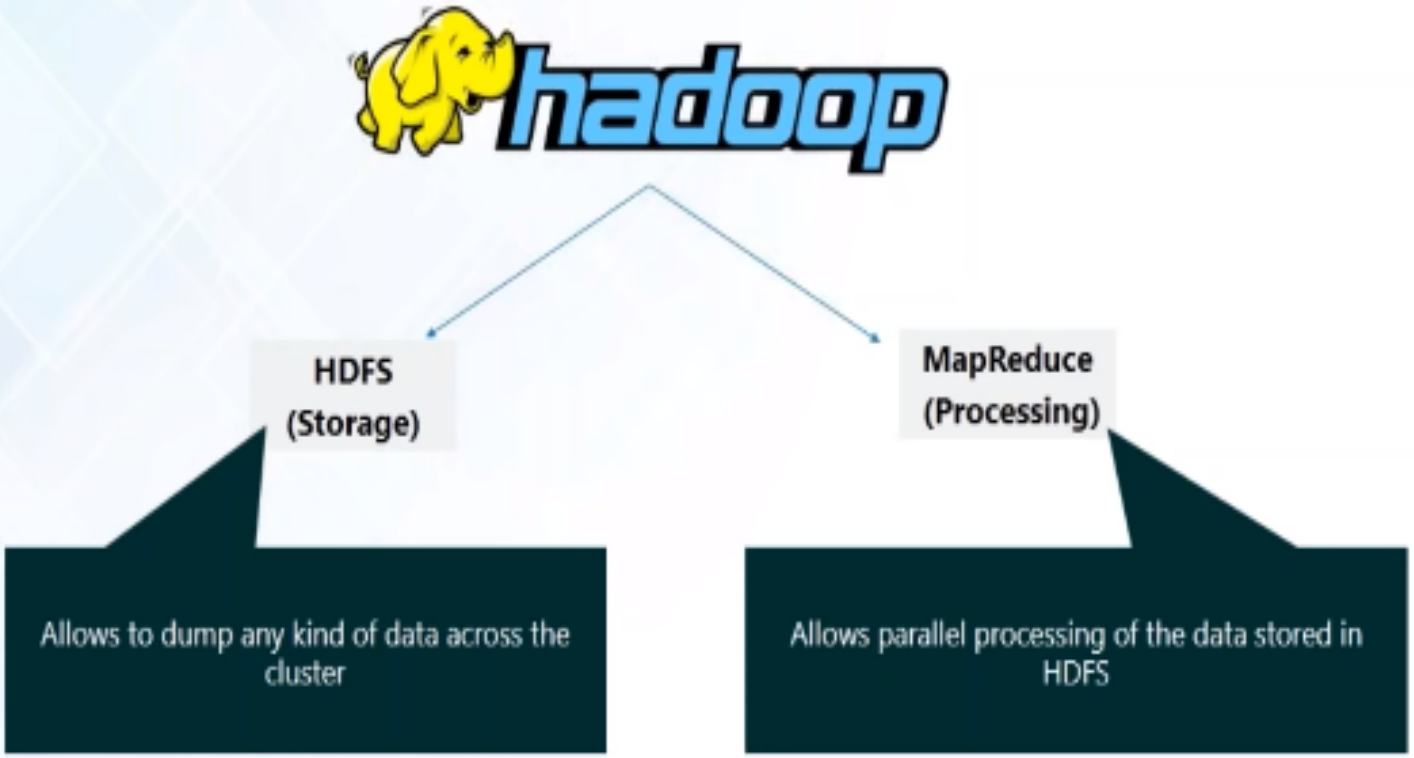
- **Apache Hadoop is** widely used, open-source software for reliable, scalable, distributed computing. Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage





HADOOP is solution to Big data problems

Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion

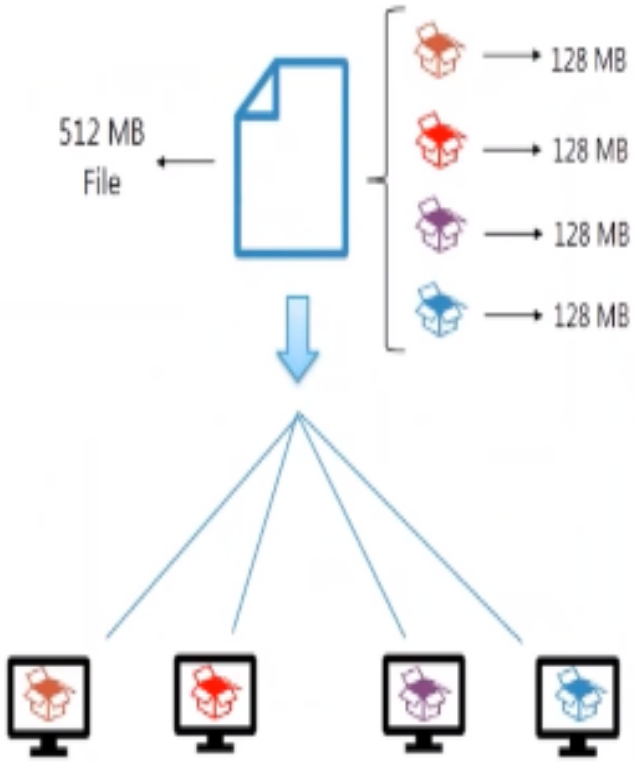


HADOOP is solution to Big data problems

Problem 1: Storing exponentially growing huge datasets

Solution: HDFS

- Storage unit of Hadoop
- It is a Distributed File System
- Divide files (input data) into smaller chunks and stores it across the cluster
- Scalable as per requirement

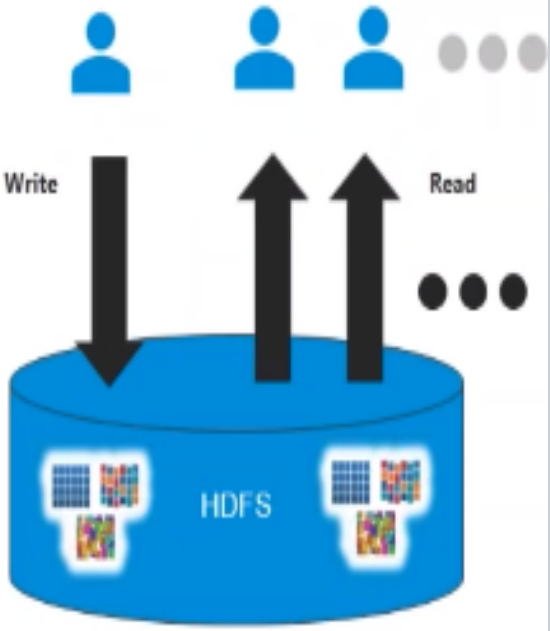


HADOOP is solution to Big data problems

Problem 2: Storing unstructured data

Solution: HDFS

- Allows to store any kind of data, be it structured, semi-structured or unstructured
- Follows WORM (Write Once Read Many)
- No schema validation is done while dumping data

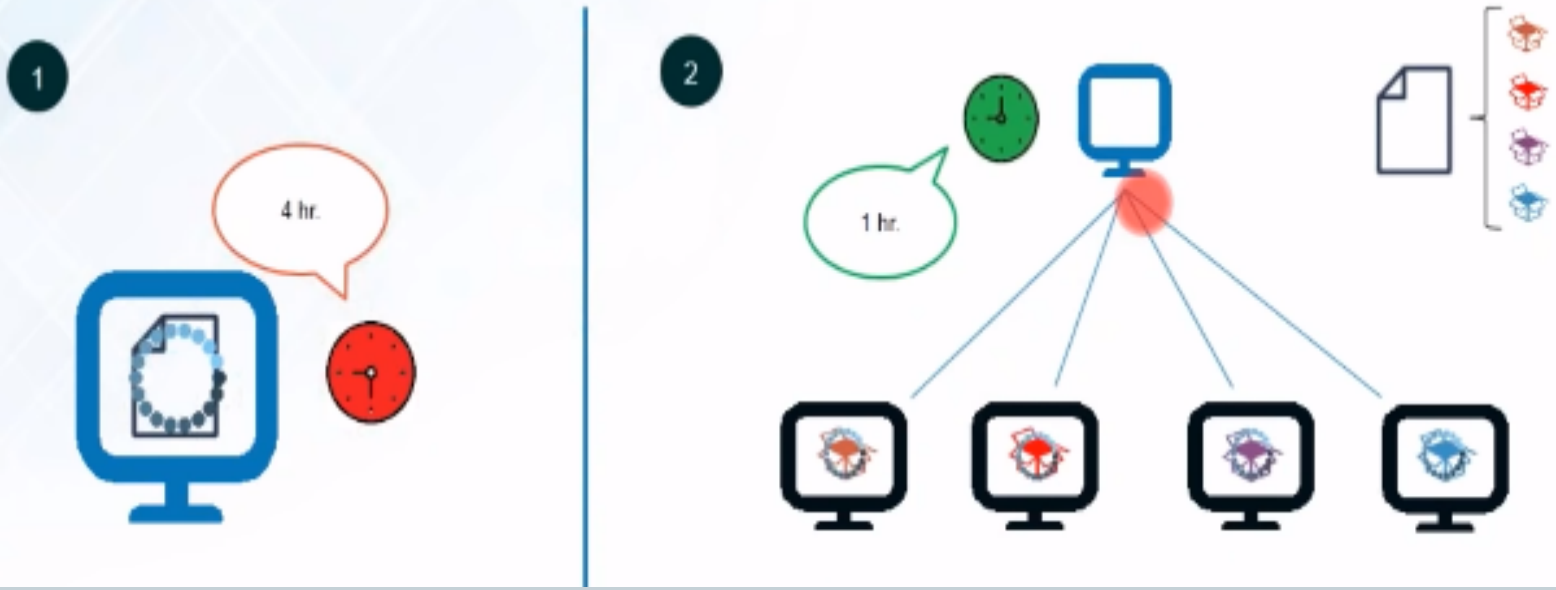


HADOOP is solution to Big data problems

Problem 3: Processing data faster

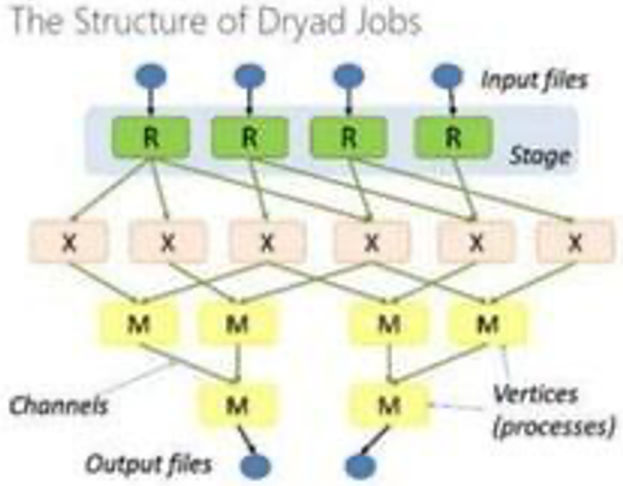
Solution: Hadoop MapReduce

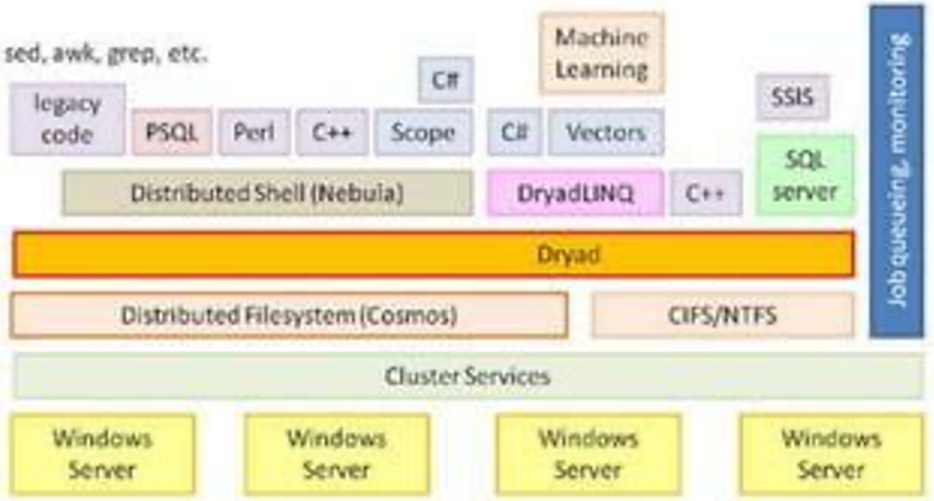
- Provides parallel processing of data present in HDFS
- Allows to process data locally i.e. each node works with a part of data which is stored on it



Big Data Solutions: Dryad[12]

- Microsoft Dryad is a R&D project, which provides an infrastructure to allow a programmer to use the resources of a computer cluster or a data center for running data-parallel programs
- A Dryad programmer can use thousands of machines, each of them with multiple processors or cores
- A Dryad job is a graph generator which can synthesize any directed acyclic graph
  - These graphs can even change during execution, in response to important events in the computation.

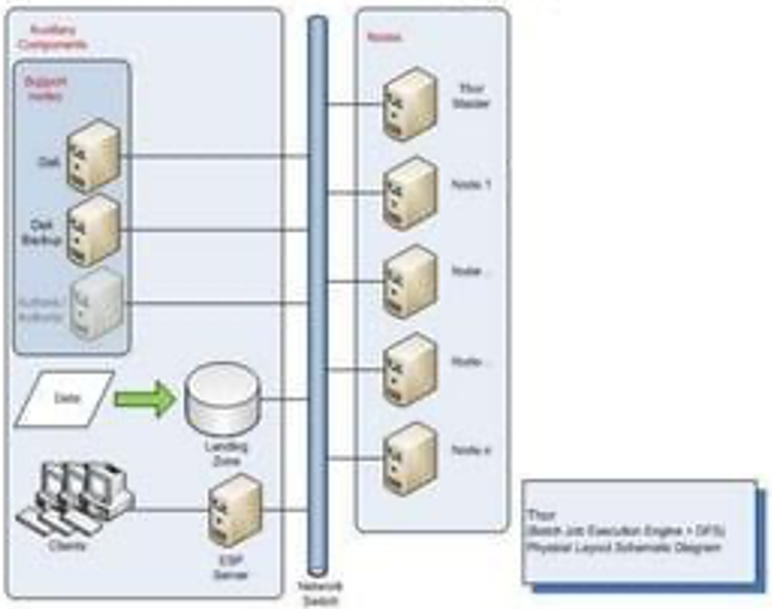




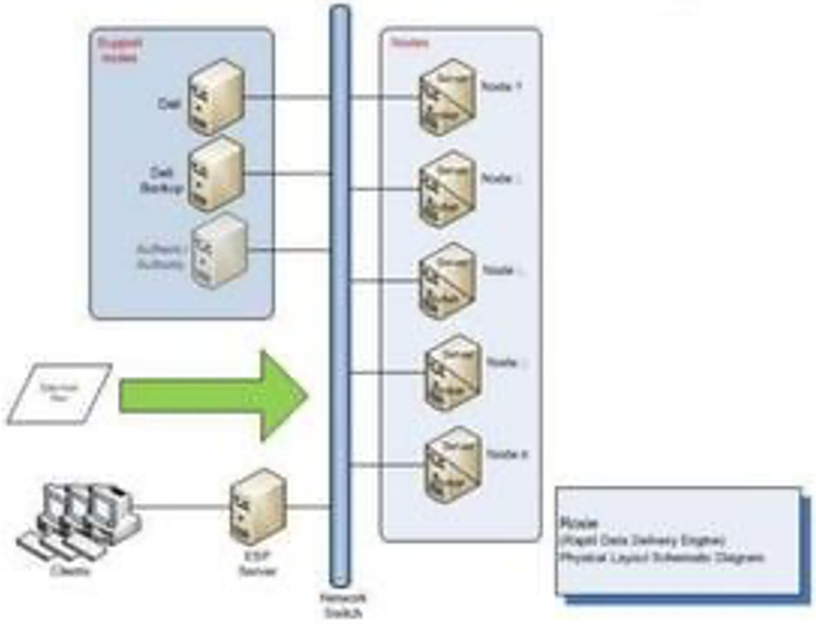
LexisNexis – HPCC[22]

* **HPCC** (High-Performance Computing Cluster), also known as **DAS** (Data Analytics Supercomputer), is an open source, data-intensive computing system platform developed by LexisNexis Risk Solutions. The HPCC platform incorporates a software architecture implemented on commodity computing clusters to provide high-performance, data-parallel processing for applications utilizing big data.

Thor: Batch Processing Engine



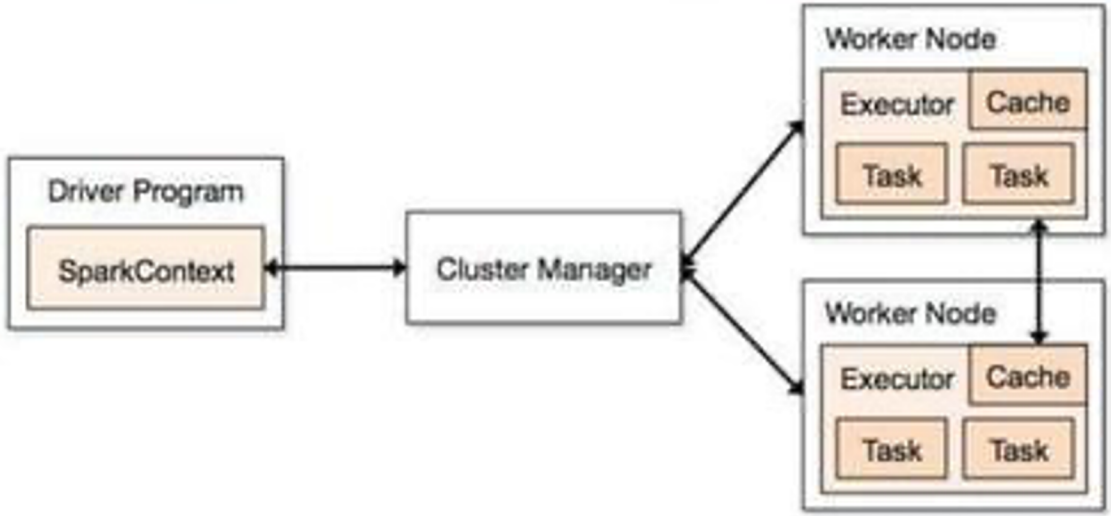
Roxie: High Perf. Query Engine



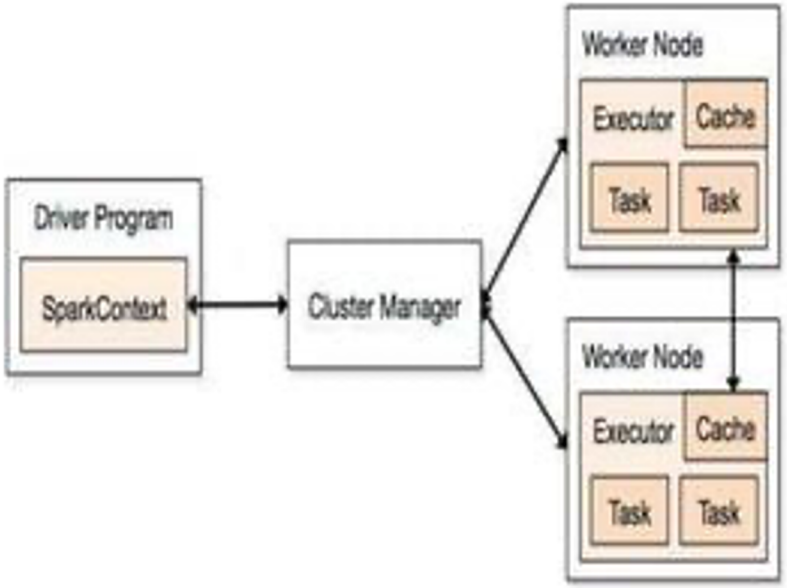
Apache Spark [28]
Lightning-fast cluster computing

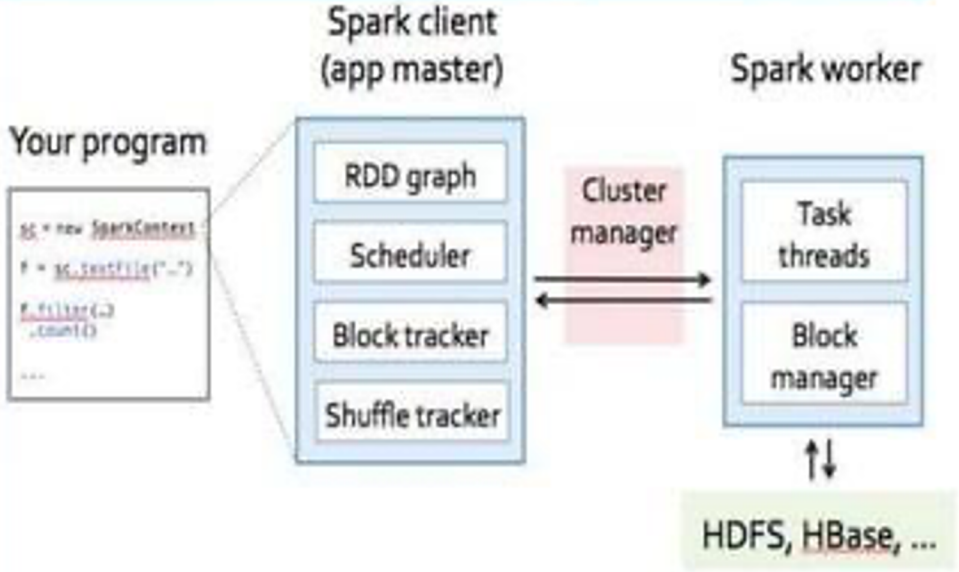
* **Apache Spark** is an open-source cluster computing framework originally developed in the AMPLab at UC Berkeley. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's in-memory primitives provide performance up to 100 times faster for certain applications

* Spark applications run as independent sets of processes on a cluster, coordinated by the SparkContext object(aka *driver program*).



Apache Spark [28]
Lightning-fast cluster computing





* The *SparkContext* can connect to several types of *cluster managers* (either Spark's own standalone cluster manager or Mesos or YARN), which allocate resources across applications
* Once connected, first it acquires *executors* (processes that run computations and store data) on nodes. Next, it sends application code (defined by JAR or Python files passed to *SparkContext*) to the *executors*.
* Finally, *SparkContext* sends *tasks* for the executors to run.

Big Data Applications

- **Government Operation:** National Archives and Records Administration, Census Bureau
- **Commercial:** Finance in Cloud, Cloud Backup, Mendeley (Citations), Netflix, Web Search, Digital Materials, Cargo shipping (as in UPS)
- **Defense:** Sensors, Image surveillance, Situation Assessment
- **Healthcare and Life Sciences:** Medical records, Graph and Probabilistic analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity models, Biodiversity
- **Deep Learning and Social Media:** Driving Car, Geolocate images/cameras, Twitter, Crowd Sourcing, Network Science, NIST benchmark datasets
- **The Ecosystem for Research:** Metadata, Collaboration, Language Translation, Light source experiments
- **Astronomy and Physics:** Sky Surveys compared to simulation, Large Hadron Collider at CERN, Belle Accelerator II in Japan
- **Earth, Environmental and Polar Science:** Radar Scattering in Atmosphere, Earthquake, Ocean, Earth Observation, Ice sheet Radar scattering, Earth radar mapping, Climate simulation datasets, Atmospheric turbulence identification, Subsurface Biogeochemistry (microbes to watersheds), AmeriFlux and FLUXNET gas sensors
- **Energy:** Smart grid

Emerging Trends in Big Data

* **Real Time Analytics:** Banking and Finance, Disaster detection and recovery, even monitoring etc. applications need vast data, coming at very fast pace to be processed within strict time limits

* **Artificial Intelligence/Business Intelligence:**
  * Intelligent Maintenance Systems: is a system that utilizes the collected data from the machinery in order to predict and ***prevent the potential failures*** in them

* **IoT/M2M:** These applications are generating data at a very fast rate (high velocity, from huge number of sources (high volume) and require big data solutions to process and derive meaningful information.

* **Transreality gaming**, sometimes written as **trans-reality gaming**, describes a type or a mode of gameplay that combines playing a game in a virtual environment with game-related, physical experiences in the real world and vice versa.

Emerging Trends in Cloud Computing – Complementary Technologies

- Cloud computing advances have helped Big Data emerge as a mass scale solution
  - Leased/Rented data storage, computing clusters, enable even startups to have global scale Big Data capability, without major capital investment

Terms[3] – 1/

- **Massively parallel processing** refers to a multitude of individual processors working in parallel to execute a particular program
- **The Big Data paradigm** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.
- **Big Data Engineering**: Advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.
- **NoSQL:** Non-relational models, also known as NoSQL, refer to logical data models that do not follow relational algebra for the storage and manipulation of data.
- **Federated database** system is a type of meta2-database management system (DBMS), which transparently maps multiple autonomous database systems into a single federated database.

Terms[3] – 2/

- **The data science paradigm** is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing.
- **The data lifecycle** is the set of processes that transform raw data into actionable knowledge.
- **Analytics** is the extraction of knowledge from information.
- **Data science** is the construction of actionable knowledge from raw data through the complete data lifecycle process.
- **A data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle.
- **Schema-on-read** is the application of a data schema through preparation steps such as transformations, cleansing, and integration at the time the data is read from the database.
- **Computational portability** is the movement of the computation to the location of the data.

Terms[3] – 3/

- **Transaction processing** is a style of computing that divides work into individual, indivisible operations, called transactions.
- **Relational databases** have traditionally supported the ACID transaction model. ACID transactions are:
  - **Atomic** Either all of the actions in a transaction are completed (i.e., transaction is committed) or none of them are completed (i.e., transaction is rolled back).
  - **Consistent** The transaction must begin and end with the database in a consistent state and must comply with all protocols (i.e., rules) of the database.
  - **Isolated** The transaction will behave as if it is the only operation being performed upon the database.
  - **Durable** The results of a committed transaction can survive system malfunctions.
- **The BASE acronym** is often used to describe the types of transactions typically supported by nonrelational databases. A BASE System is described in contrast to an ACID-compliant systems as:
  - Basically Available, Soft state, and Eventually Consistent
  - BASE transactions allow a database to be in a temporarily inconsistent state that will eventually be resolved.

Terms[3] – 4/

* **CAP Theorem** states that a distributed system can support only two of the following three characteristics:
  * **C**onsistency The client perceives that a set of operations has occurred all at once.
  * **A**vailability Every operation must terminate in an intended response.
  * **P**artition tolerance Operations will complete, even if individual components are unavailable.