University Politehnica of Bucharest

Faculty of Automatic Control and Computers

Computer Science and Engineering Department

Big Data
- An Overview -

Ciprian-Octavian Truică
ciprian.truica@cs.pub.ro

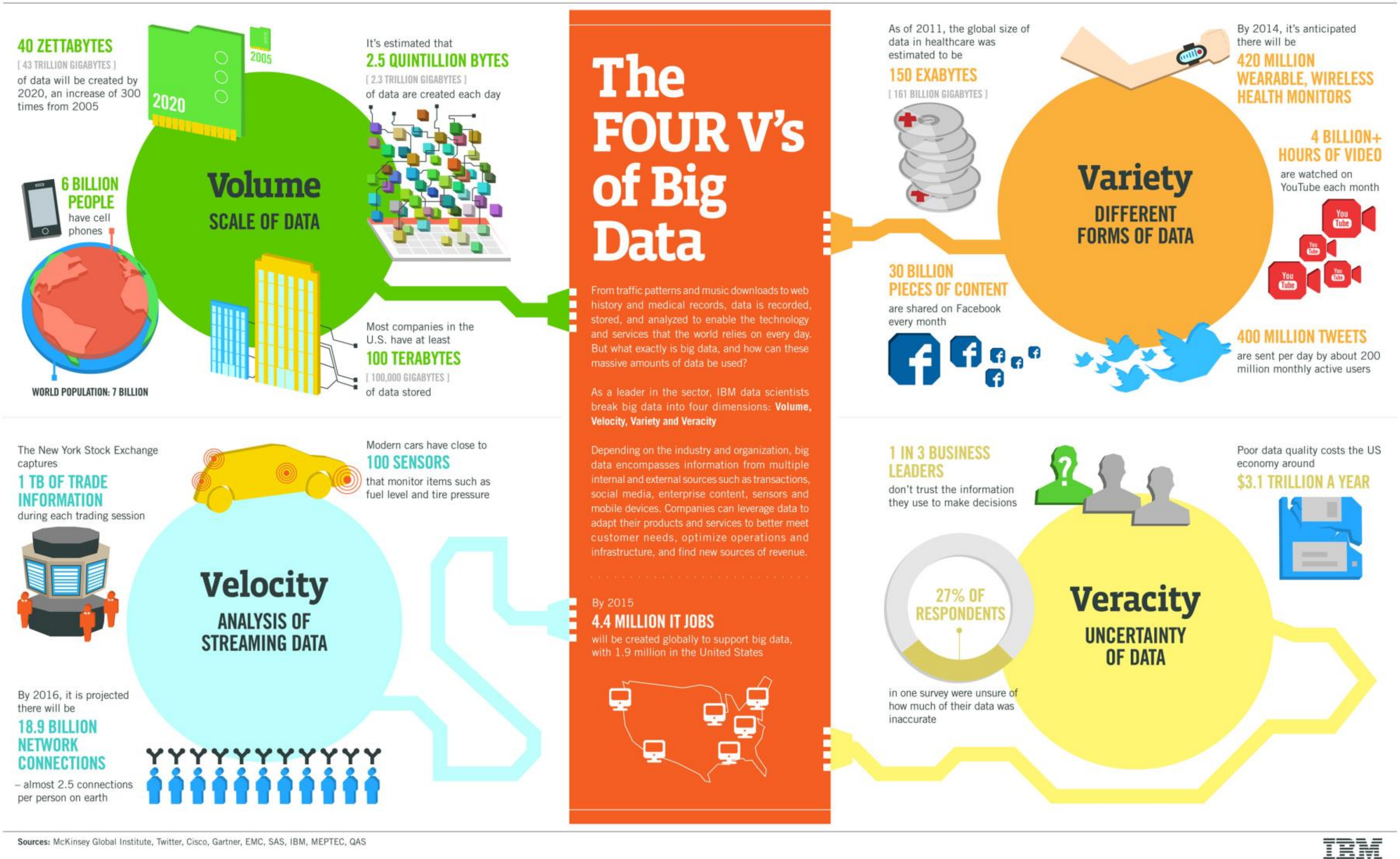
Overview

- What is Big Data?
- Why Big Data?
- Types of Big Data
- Techniques
- Distributed architecture
- Cloud Computing
- Storage and tools

What is Big Data?

- Big Data is high volume, high velocity and high variety of data that require new forms of processing to enable knowledge extraction, insight discovery, decision making, and process optimization

What is Big Data?



What is Big Data?

- The 4 V's of Big Data:

1. **Volume**
   - The main characteristic of Big Data is the volume
   - The volume of data impacts its analysis
   - Historical data is important especially for Business Intelligence
   - Data generated from different sources are stored together to create correlations and extract knowledge.

What is Big Data?

- The 4 V’s of Big Data:

2. Variety

- Variety refers to the types of data available for analysis
- Multiple types of data: numbers, dates, text, images, video, etc.
- Multiple sources for data: companies databases, social media, blogs, etc.
- Structured, unstructured and hybrid types of data.

What is Big Data?

- The 4 V's of Big Data:

3. **Veracity**
   - Veracity refers to the trustworthiness of the data.
   - The quality of the data can affect the analysis process
   - The data must be representative, relevant, consistent, accurate and current to discover patterns
   - The data must be preprocessed to extract relevant knowledge

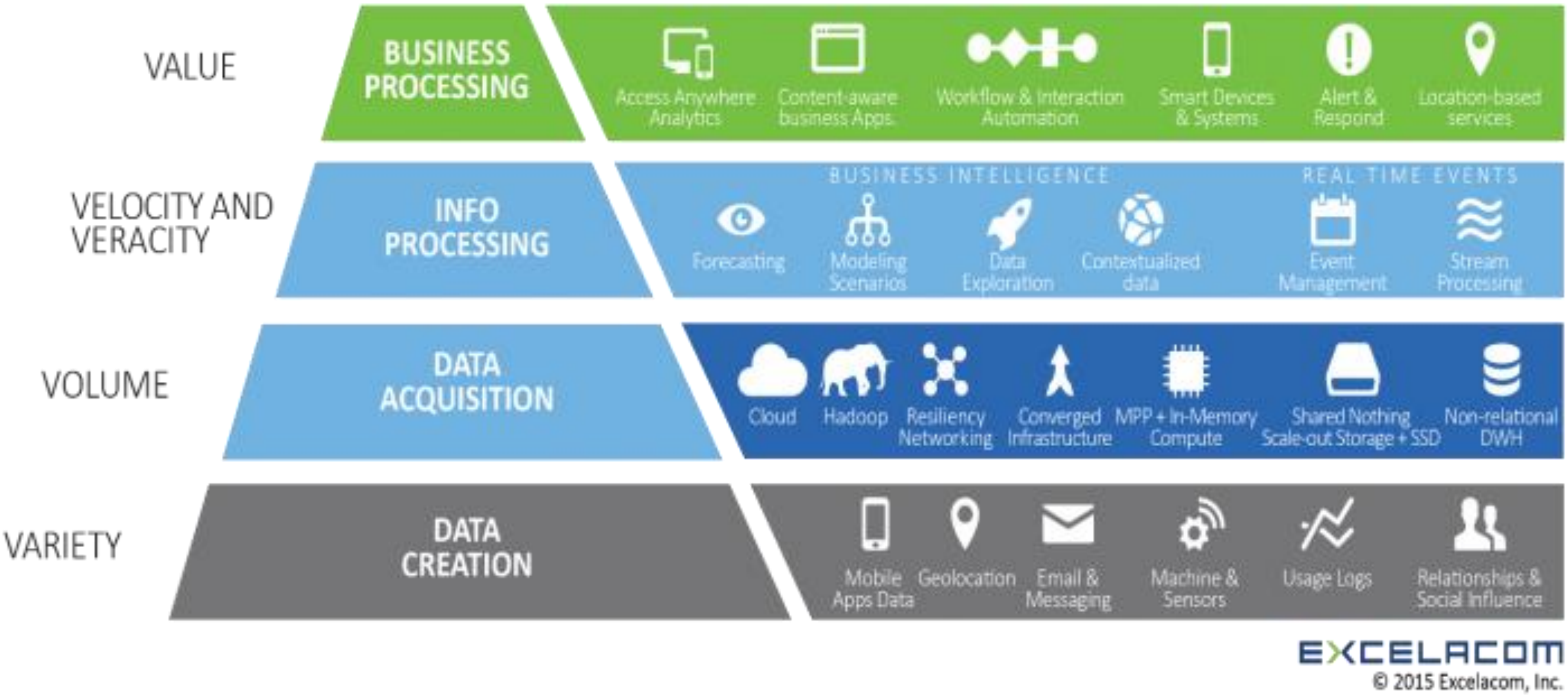
What is Big Data?

- The 4 V’s of Big Data:

4. Velocity

  – Velocity refers to how fast the data is processed
  – The faster the date is processed, the faster it can be queried and interpreted to extract knowledge
  – A rapid data analysis helps to correctly and rapidly make and take decision

What is Big Data?



What is Big Data?

- The 5 V's of Big Data:

5. **Value**:
   - A critical objective of Big Data is to use the 4 V to create value
   - Through data analysis, Value is achieved by making the data profitable and help to increase revenue and decrease costs

What is Big Data?

- The 8 V’s of Big Data:

6. **Variability** of the data must be also taken into account
   - The meaning of data is constantly changing
   - The context where the data appears must be understood

What is Big Data?

- The 8 V's of Big Data:

7. **Visualization** refers to presentation of data in a pictorial or graphical form
   - Analytics preserved visually
     - To understand difficult concept
     - To identify new patterns
   - Interactive visualization: charts, graphs, etc.
   - Develop new techniques for data visualization
   - Used for data exploration
   - Detect inconstancies in the data structure

What is Big Data?

- The 8 V's of Big Data:

8. **Vicissitude** refers to the challenges of scaling Big Data complex workflows.

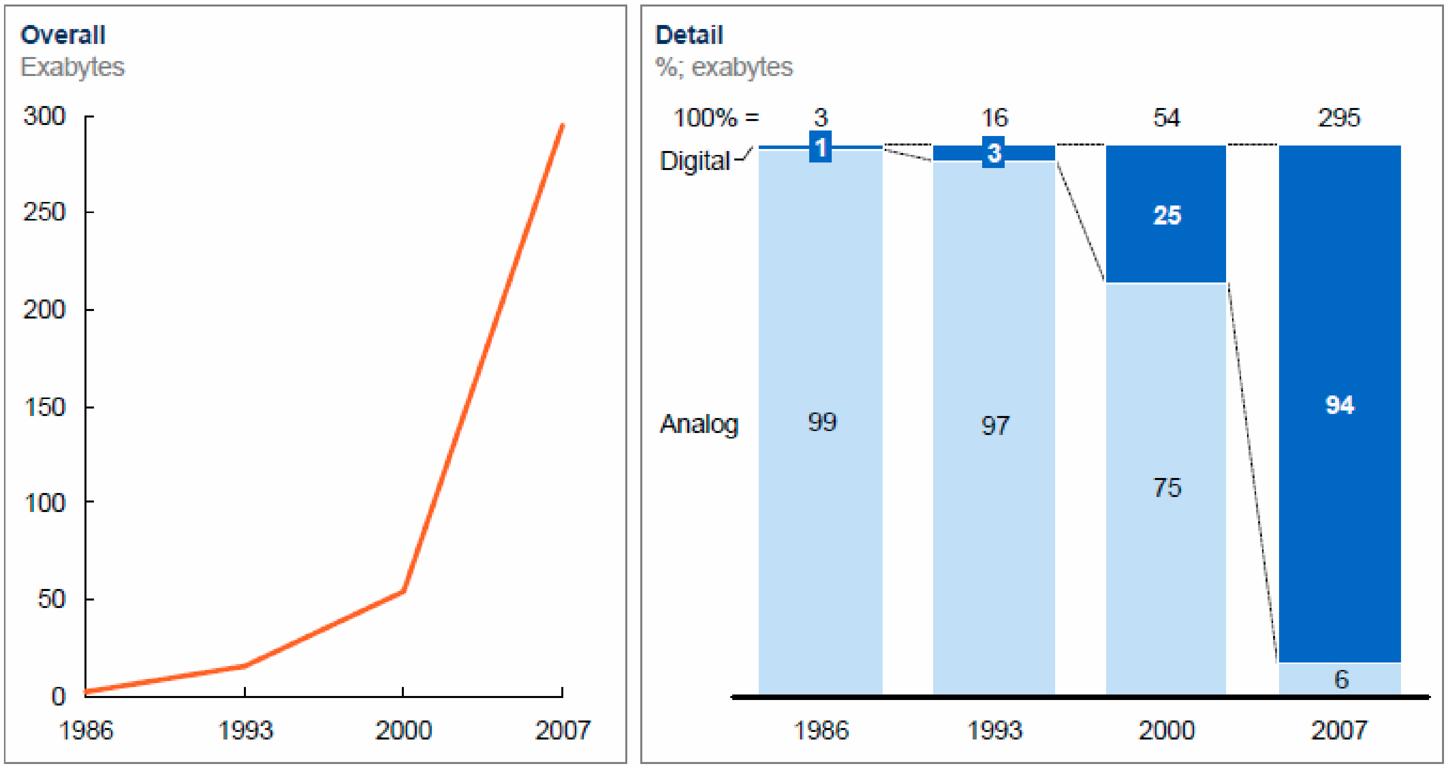
Why Big Data?

- Increased storage capacities
  - Storage capacities have grown
  - Storage (now) is cheap – cost of storage/GB decreased

Why Big Data?

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Why Big Data?

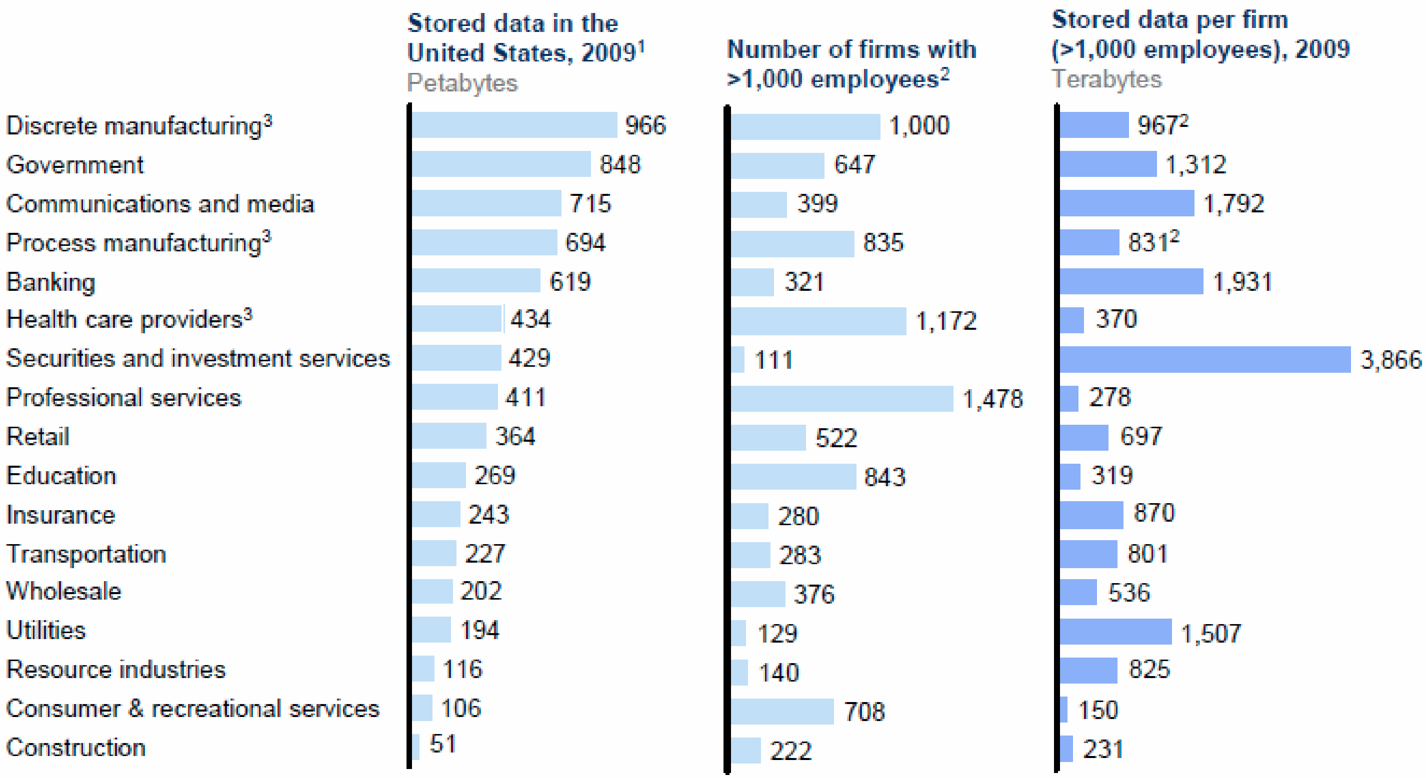
- Increase of processing power
  - Computation capacity has also risen sharply
  - New means of scaling computations: GPU, Hadoop, Spark, etc.

Why Big Data?

- Availability of data
  - More data is available
  - Different types of data are available:
    - Private sector (financier corporations, banks, etc.)
    - Social media
    - Internet of Things

Why Big Data?

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

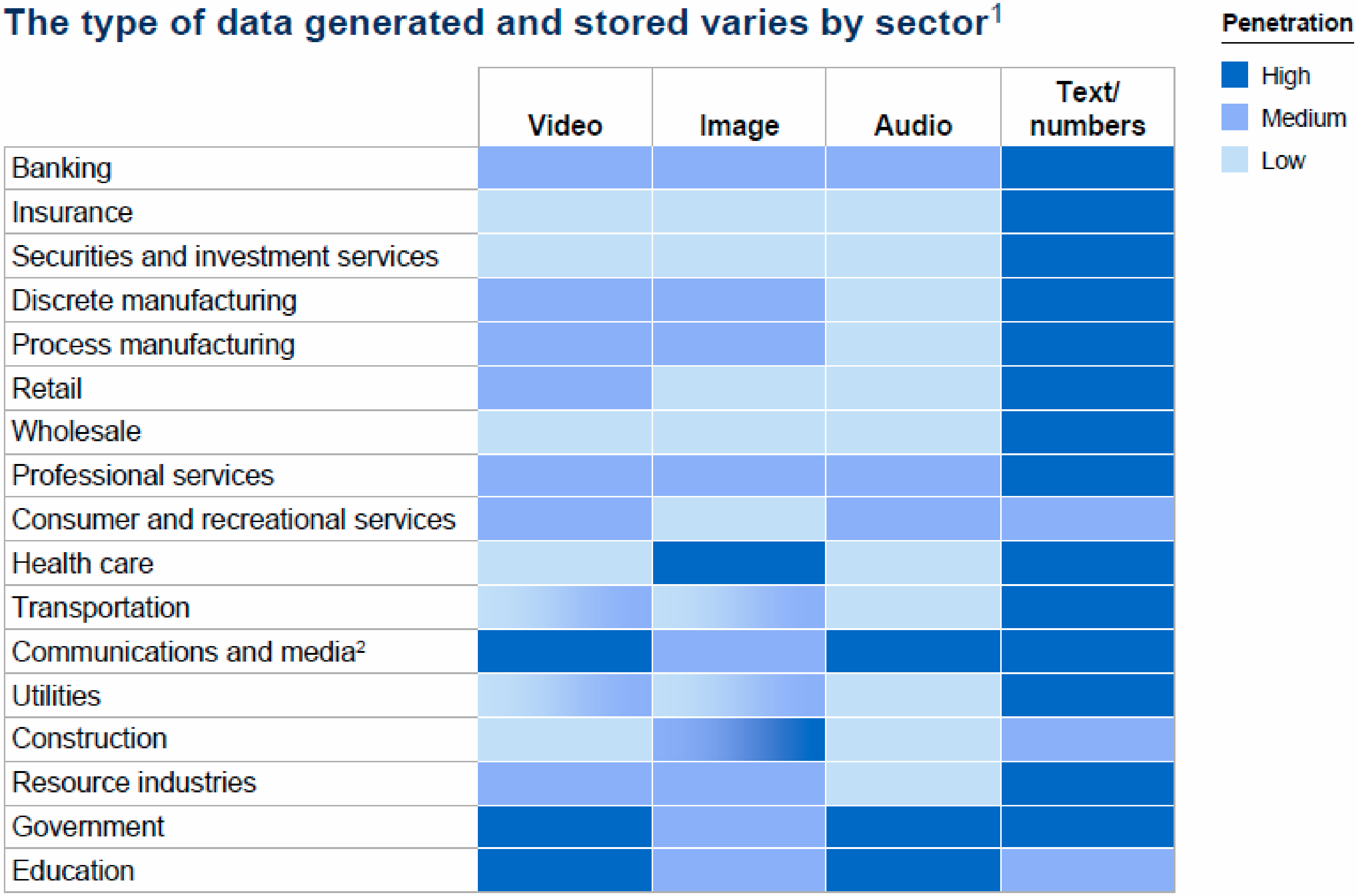


1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

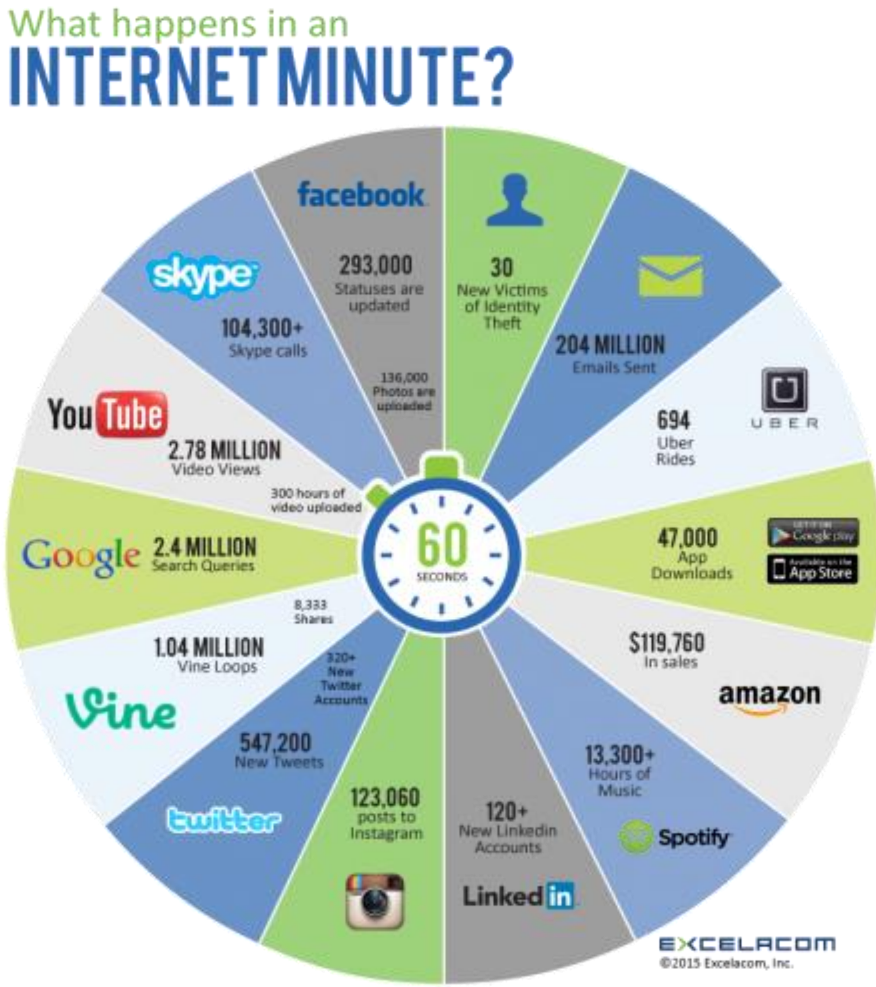
Why Big Data?

The type of data generated and stored varies by sector[1]

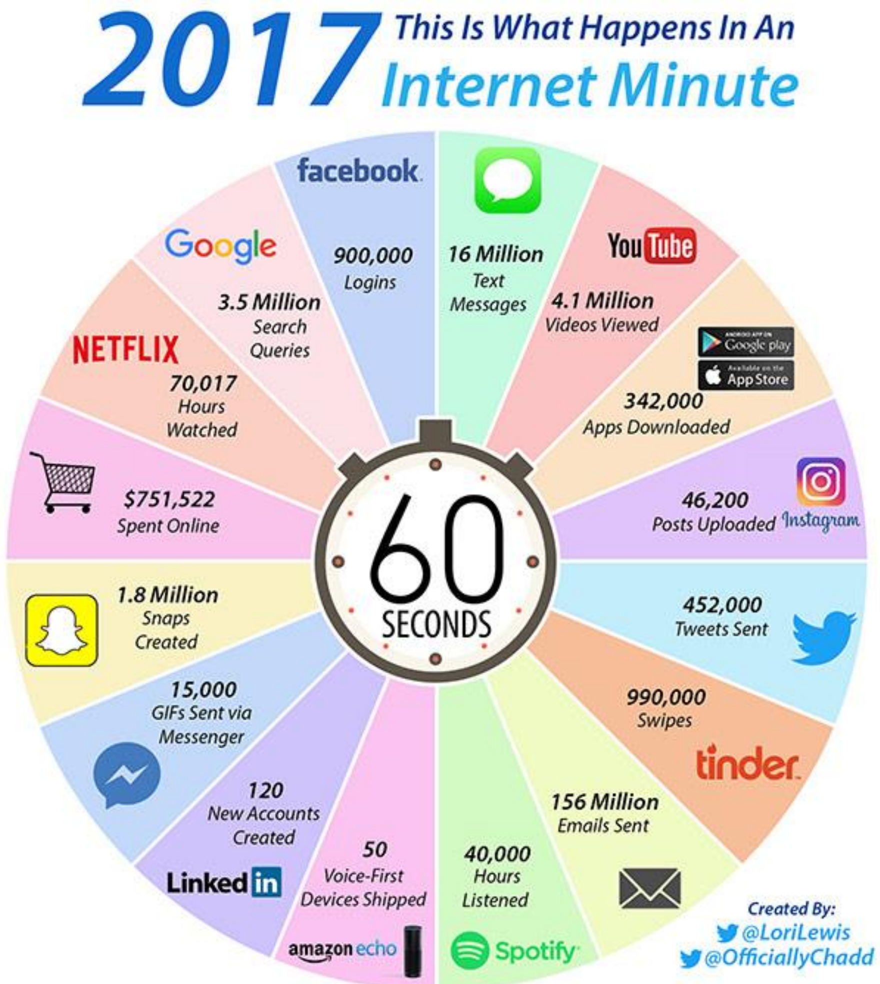


1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

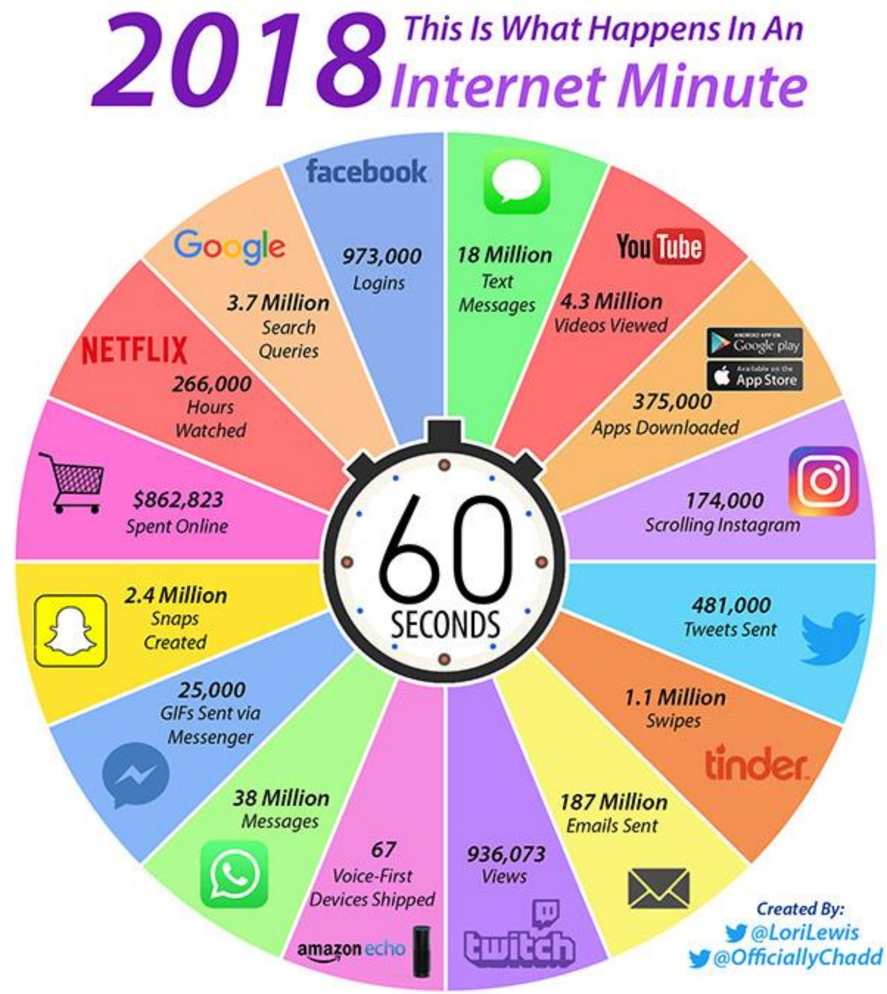
Why Big Data?


What happens in an
INTERNET MINUTE?
facebook
293,000
Statuses are updated
136,000
Photos are uploaded
30
New Victims of Identity Theft
204 MILLION
Emails Sent
694
Uber Rides
UBER
47,000
App Downloads
App Store
$119,760
In sales
amazon
13,300+
Hours of Music
Spotify
120+
New Linkedin Accounts
LinkedIn
123,060
posts to Instagram
547,200
New Tweets
320+
New Twitter Accounts
twitter
1.04 MILLION
Vine Loops
Vine
8,333
Shares
Google
2.4 MILLION
Search Queries
300 hours of video uploaded
YouTube
2.78 MILLION
Video Views
skype
104,300+
Skype calls
60
SECONDS
EXCELACOM
©2015 Excelacom, Inc.

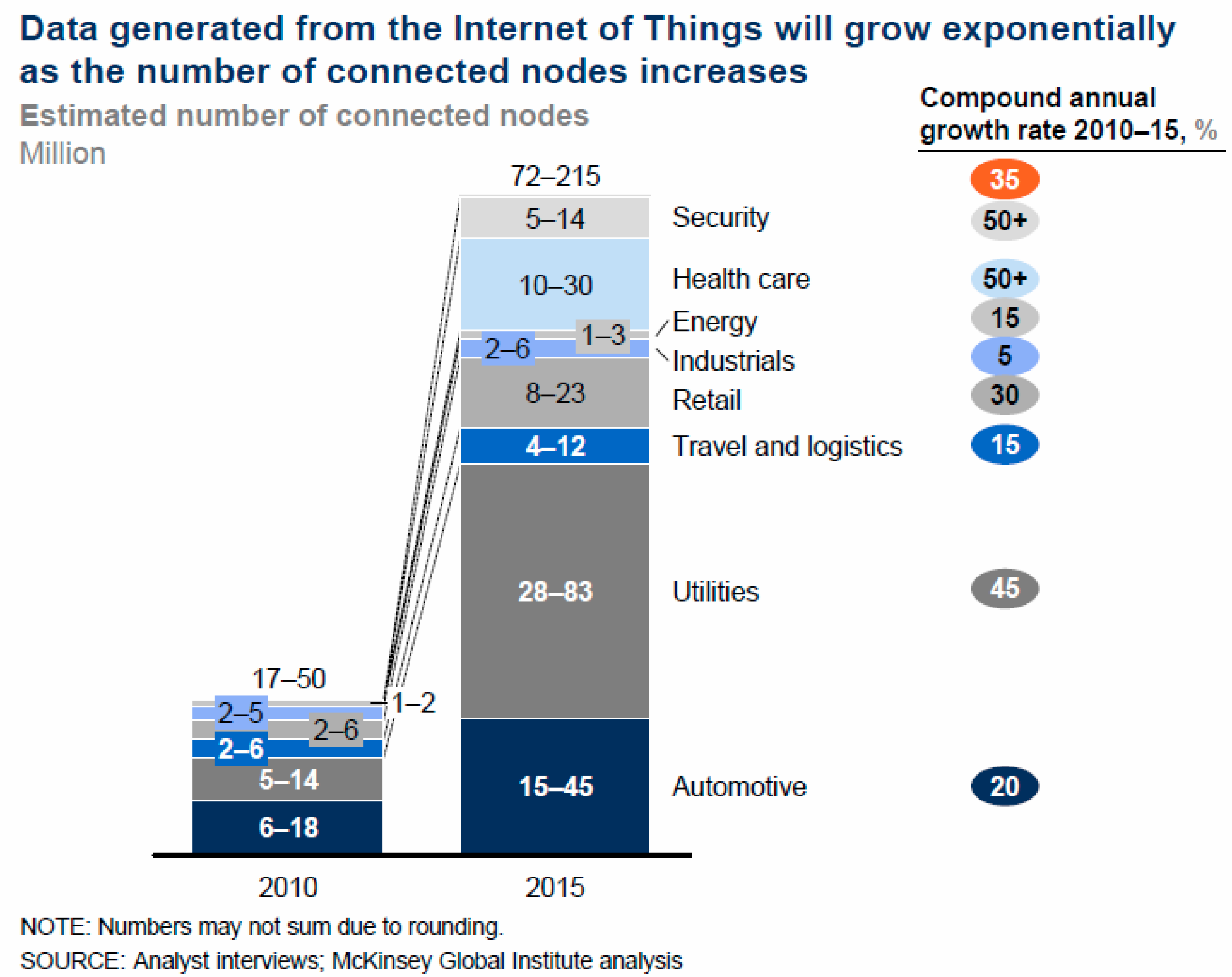
Why Big Data?



2017 This Is What Happens In An Internet Minute
facebook
900,000 Logins
16 Million Text Messages
YouTube
4.1 Million Videos Viewed
Google
3.5 Million Search Queries
NETFLIX
70,017 Hours Watched
Google play
App Store
342,000 Apps Downloaded
$751,522 Spent Online
46,200 Posts Uploaded
Instagram
60 SECONDS
1.8 Million Snaps Created
452,000 Tweets Sent
15,000 GIFs Sent via Messenger
990,000 Swipes
tinder
120 New Accounts Created
LinkedIn
50 Voice-First Devices Shipped
amazon echo
40,000 Hours Listened
Spotify
156 Million Emails Sent
Created By:
@LoriLewis
@OfficiallyChadd



2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
Google
3.7 Million Search Queries
NETFLIX
266,000 Hours Watched
Google play
App Store
375,000 Apps Downloaded
$862,823 Spent Online
174,000 Scrolling Instagram
60 SECONDS
2.4 Million Snaps Created
481,000 Tweets Sent
25,000 GIFs Sent via Messenger
1.1 Million Swipes
tinder
38 Million Messages
67 Voice-First Devices Shipped
amazon echo
936,073 Views
twitch
187 Million Emails Sent
Created By:
@LoriLewis
@OfficiallyChadd


Data generated from the Internet of Things will grow exponentially as the number of connected nodes increases
Estimated number of connected nodes
Million
Compound annual growth rate 2010–15, %
72–215
5–14
Security
50+
10–30
Health care
50+
Energy
15
2–6
1–3
Industrials
5
8–23
Retail
30
4–12
Travel and logistics
15
28–83
Utilities
45
15–45
Automotive
20
35
17–50
2–5
2–6
2–6
1–2
5–14
6–18
2010
2015
NOTE: Numbers may not sum due to rounding.
SOURCE: Analyst interviews; McKinsey Global Institute analysis

Types of Big Data

- Structured Big Data
- Unstructured Big Data
- Hybrid (Multi-Structured) Big data

Types of Big Data

- Structured Big Data
  - Refers to data that have a predefined data model
  - Has a strict schema which generally adheres to the Relational Algebra Theory
  - The data set's attributes are predefined and have a strict data type: number, datetime, string, etc.

Types of Big Data

- Structured Big Data sources:
  - Computer or machine generated data is data that generally refers to data that is created by a machine without human intervention.
  - Some examples:
    - Sensor data: generated by sensors, medical devices, GPS, etc.
    - Web log data: data captured by machine's activates, it is generated by servers, applications, networks, etc.
    - Financial data: financial systems use predefined rules that automate their processes, e.g. stock-trading

Types of Big Data

- Structured Big Data sources:
  - Human generated data is data that is generated by the interaction of human with computers
  - Some examples:
    - Input Data: information and data that humans input in a computer, e.g. when buying an airplane ticket a human will input his name, age, gender, passport ID, etc.
    - Click-stream data: data that is generated when humans click a link on a website. This data is used to analyze and determine customer behavior and buying patterns
    - Gaming related data: used to record the player's moves in a game to understand and develop new wining strategies

Types of Big Data

- Structured Big Data technologies:
  - Relational databases: Oracle, PostgreSQL, MySQL, Microsoft SQL server, Teradata
  - Data Warehouses and Data Marts
  - Enterprise Resource Planning software (ERP)
  - Customer Relationship Management software (CRM)
  - Mainframes

Types of Big Data

- Unstructured Big Data
  - Refers to data that does not have a predefined data model
  - Does not have a strict schema, generally it uses a flexible schema-free model
  - The data set’s attributes are not predefined and they do not have a strict data type
  - Employs the use of different data structures: sets, lists, maps, etc.

Types of Big Data

- Unstructured Big Data sources:
  - As in the case of Structured Big Data, they are machine or human generated
  - Some examples of machine generated unstructured data:
    - Satellite images includes weather data or the data the governments captures in their surveillance satellites
    - Scientific data includes seismic imagery, atmospheric data, etc.
    - Photographs and videos includes security, surveillance and traffic videos
    - Radar and sonar data includes meteorological and oceanographic seismic profiles.

Types of Big Data

- Unstructured Big Data sources:
  - Some examples of human generated unstructured data:
    - Text internal to a company: word documents, excel spreadsheets, emails, etc.
    - Social media data: Facebook, Twitter, YouTube, LinkedIn, Instagram, etc.
    - Mobile data: text messages, location information, etc.
    - Website content: News sites, YouTube, Instagram, etc.

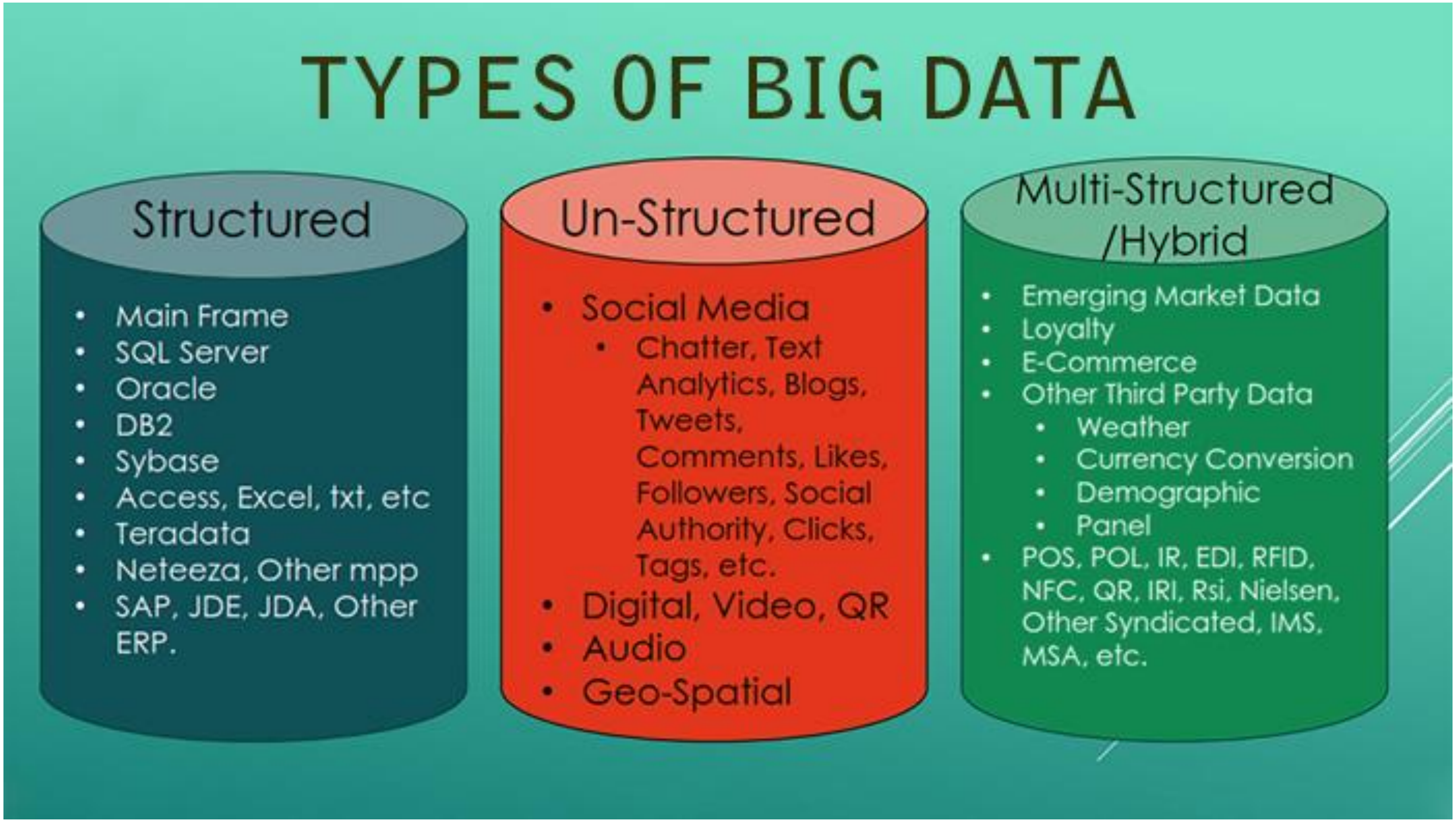
Types of Big Data

- Unstructured Big Data technologies:
  - NoSQL databases: MongoDB, CouchDB, Neo4J, Cassandra, etc.
  - Distributed storage frameworks: Hadoop
  - Data Lakes (a storage repository that stores vast amounts of raw data in its native format).
  - Enterprise Content Management Systems (CMS) that manage the complete life cycle of contend

Types of Big Data

- Hybrid Big Data
  - Combines Structured and Unstructured Big Data
  - Combines the data sources for Structured and Unstructured Big Data and correlates the information
  - Combines the technologies used for structured and unstructured Big Data

Types of Big Data



Techniques

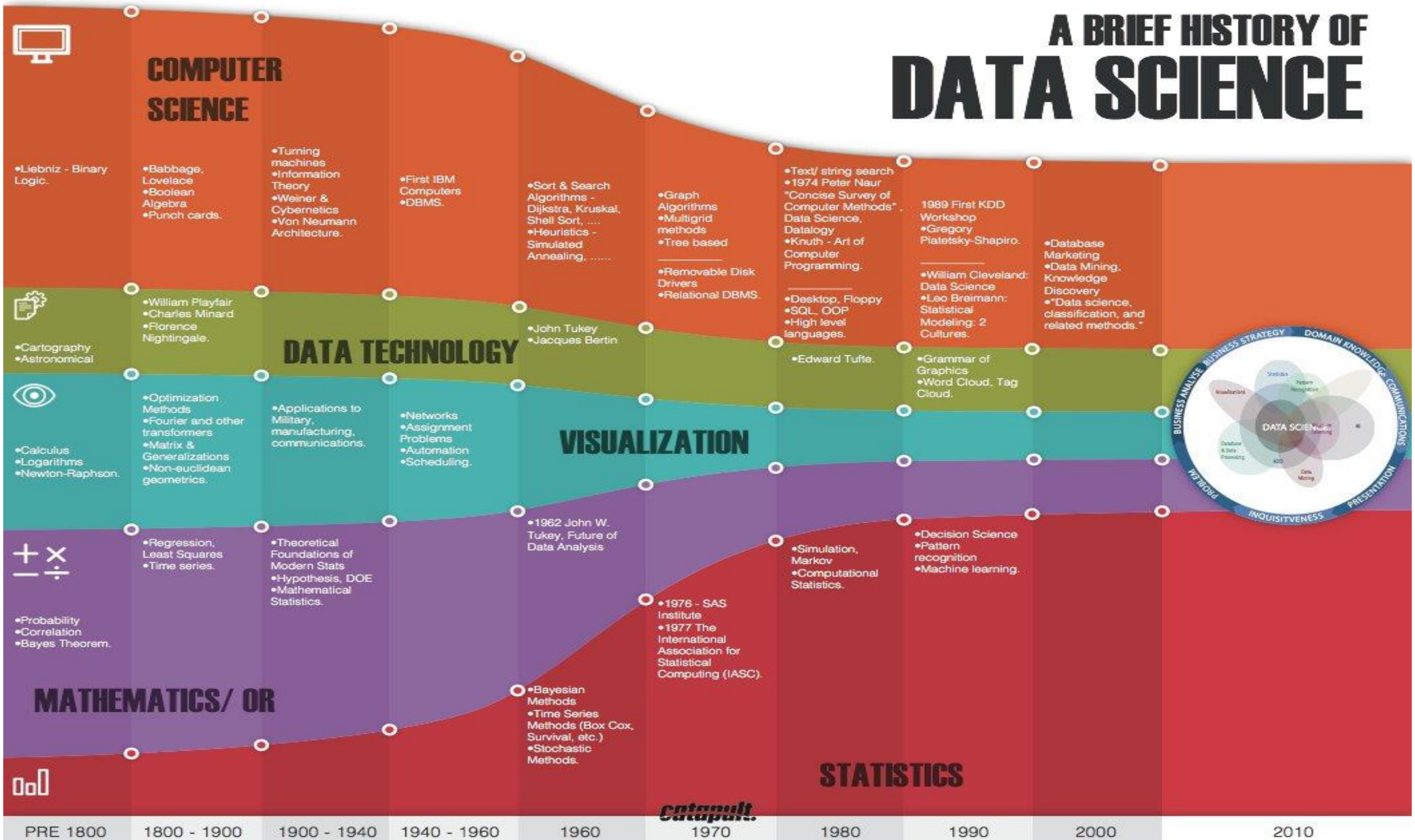
- When Big data is a problem?

1. When the operations on the data are complex:
   - Modeling and reasoning
   - Knowledge extraction
   - Pattern extraction
2. When usual algorithms cannot handle big amounts of data:
   - Performance and query execution time
   - Parallelism and distribution
   - Resource usage (CPU, RAM)

Techniques

- Big Data combines multiple research fields:
  - Databases
  - Knowledge Discovery in Databases (KDD)
  - Data Mining and Text Mining
  - Parallel and distributed algorithms
  - High Performance Computing (HPC)
  - Information Retrieval (IR) and subdomains: Multi-Lingual Information Retrieval (MLIR)
  - Natural Language Processing (NLP)
  - Semantic Web
  - Probabilities and Statistics
  - Machine Learning
  - Etc.

Techniques



A BRIEF HISTORY OF
DATA SCIENCE
COMPUTER SCIENCE
•Liebniz - Binary Logic.
•Babbage, Lovelace
•Boolean Algebra
•Punch cards.
•Turning machines
•Information Theory
•Weiner & Cybernetics
•Von Neumann Architecture.
•First IBM Computers
•DBMS.
•Sort & Search Algorithms - Dijkstra, Kruskal, Shell Sort, ...
•Heuristics - Simulated Annealing, ...
•Graph Algorithms
•Multigrid methods
•Tree based
•Removable Disk Drivers
•Relational DBMS.
•Text/ string search
•1974 Peter Naur "Concise Survey of Computer Methods" - Data Science, Datalogy
•Knuth - Art of Computer Programming.
•Desktop, Floppy
•SQL, OOP
•High level languages.
1989 First KDD Workshop
•Gregory Piatetsky-Shapiro.
•William Cleveland: Data Science
•Leo Breimann: Statistical Modeling: 2 Cultures.
•Database Marketing
•Data Mining, Knowledge Discovery
•"Data science, classification, and related methods."
•Cartography
•Astronomical
•William Playfair
•Charles Minard
•Florence Nightingale.
DATA TECHNOLOGY
•John Tukey
•Jacques Bertin
•Edward Tufte.
•Grammar of Graphics
•Word Cloud, Tag Cloud.
•Calculus
•Logarithms
•Newton-Raphson.
•Optimization Methods
•Fourier and other transformers
•Matrix & Generalizations
•Non-euclidean geometrics.
•Applications to Military, manufacturing, communications.
•Networks
•Assignment Problems
•Automation
•Scheduling.
VISUALIZATION
•Probability
•Correlation
•Bayes Theorem.
•Regression, Least Squares
•Time series.
•Theoretical Foundations of Modern Stats
•Hypothesis, DOE
•Mathematical Statistics.
•1962 John W. Tukey, Future of Data Analysis
•1976 - SAS Institute
•1977 The International Association for Statistical Computing (IASC).
•Simulation, Markov
•Computational Statistics.
•Decision Science
•Pattern recognition
•Machine learning.
MATHEMATICS/ OR
•Bayesian Methods
•Time Series Methods (Box Cox, Survival, etc.)
•Stochastic Methods.
STATISTICS
catapult.
BUSINESS STRATEGY
DOMAIN KNOWLEDGE
COMMUNICATIONS
PRESENTATION
INQUISITIVENESS
PROBLEM
BUSINESS ANALYSE
DATA SCIENCE
PRE 1800
1800 - 1900
1900 - 1940
1940 - 1960
1960
1970
1980
1990
2000
2010

Distributed architectures

- A distributed architecture is a model in which resources located on network computers are put together in a resource pool to achieve a common goal.
- A cluster is a set of tightly connected computers that work together so that, in many respect they can be seen as a single system

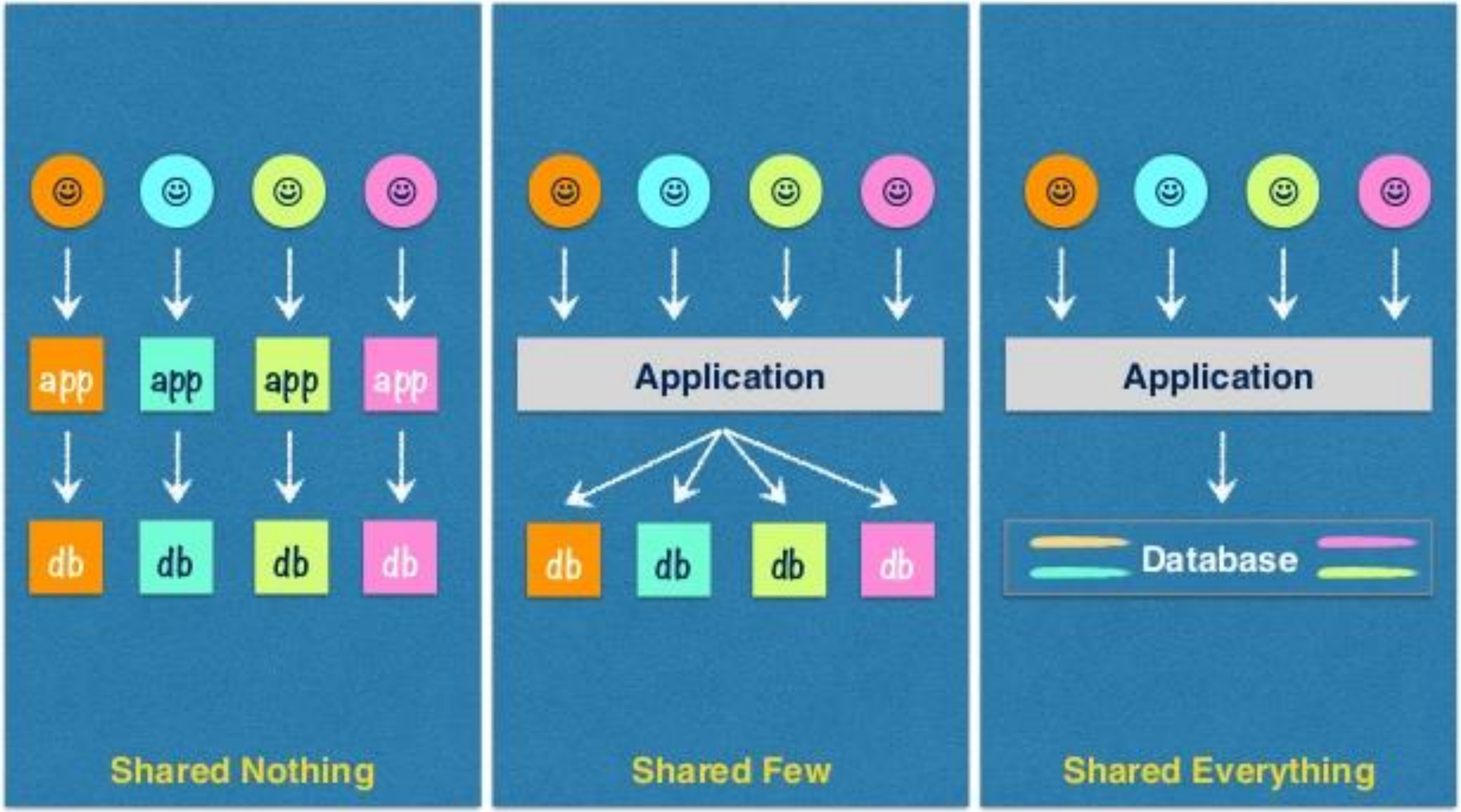
Distributed architectures

- Scalability is the capability of a system to handle a growing amount of work
- Types of scaling
  - Vertical scaling (scale up/down)
    - Means that you can scale by adding more resources to an existing machine, e.g. CPU, RAM, HDD, etc.
  - Horizontal scaling (scale out/in)
    - Means that you can scale by adding more machines into the resource pool, e.g. add nodes to the cluster.

Distributed architectures

- Resource sharing in a cluster
  - Shared everything architecture
    - Is a distributed computer architecture in which each node is added its resources to the resource pool
    - More specifically, the user sees the total amount of memory, CPU or disk storage and not individual amounts
  - Shared nothing architecture
    - Is a distributed computer architecture in which each node is independent and self sufficient
    - More specifically, none of the nodes share memory, CPU or disk storage

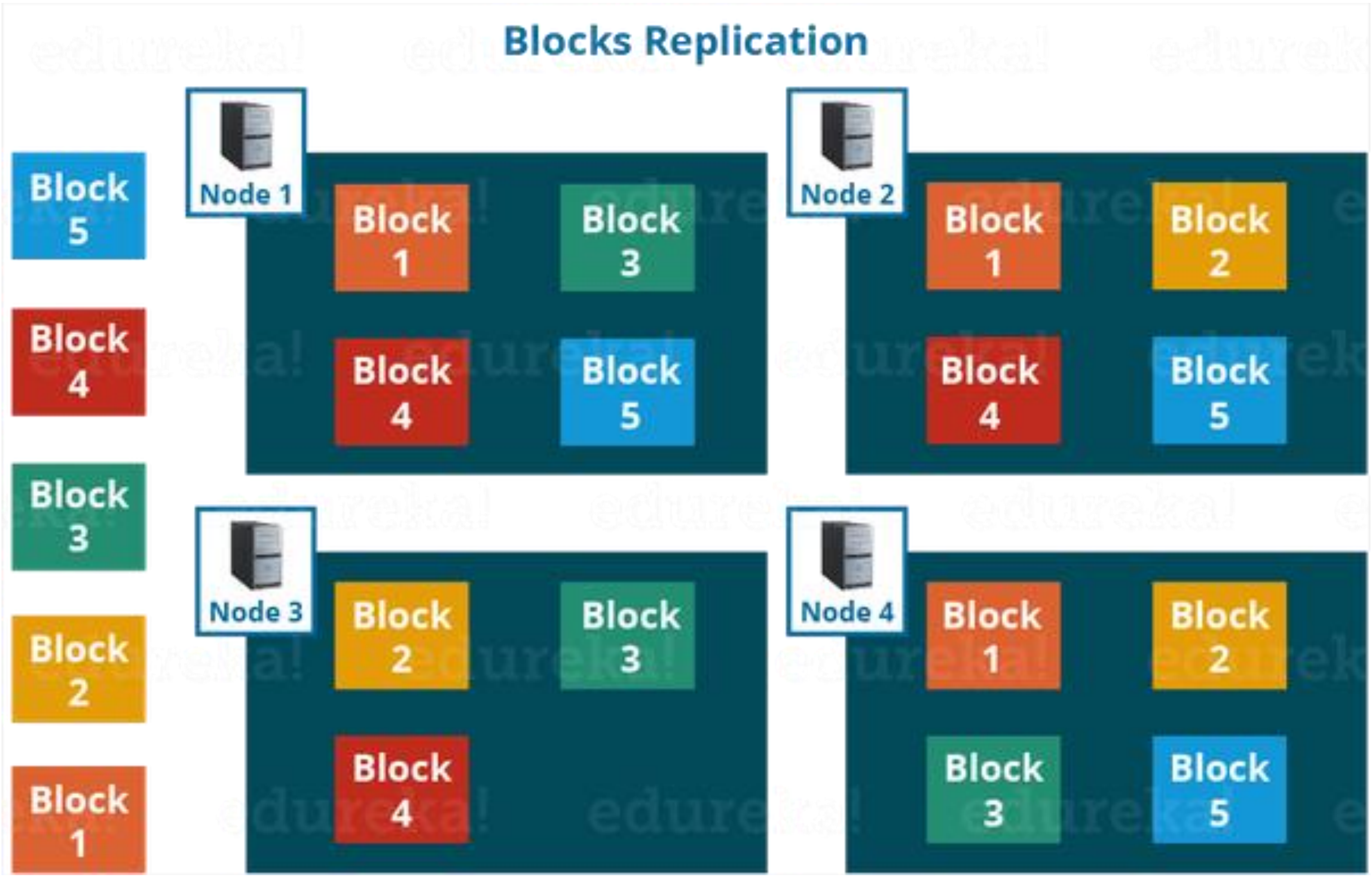
Distributed architectures



Distributed architectures

- Replication involves sharing information so as to ensure consistency between redundant resources and to improve reliability, fault-tolerance, or accessibility.
- Replication is archived when the same data is stored on multiple storage devices

Distributed architectures



Distributed architectures

- Replication from the architecture perspective
  - Master-Slave
    - The Master node supports all the operations, i.e. create, read, update, delete (CRUD)
    - Only the Master node can start a transaction
    - Only the read operation is supported on the Slave node
  - Multi-Master
    - All the nodes support all the CRUD operations
    - All the nodes can start a transaction

Distributed architectures

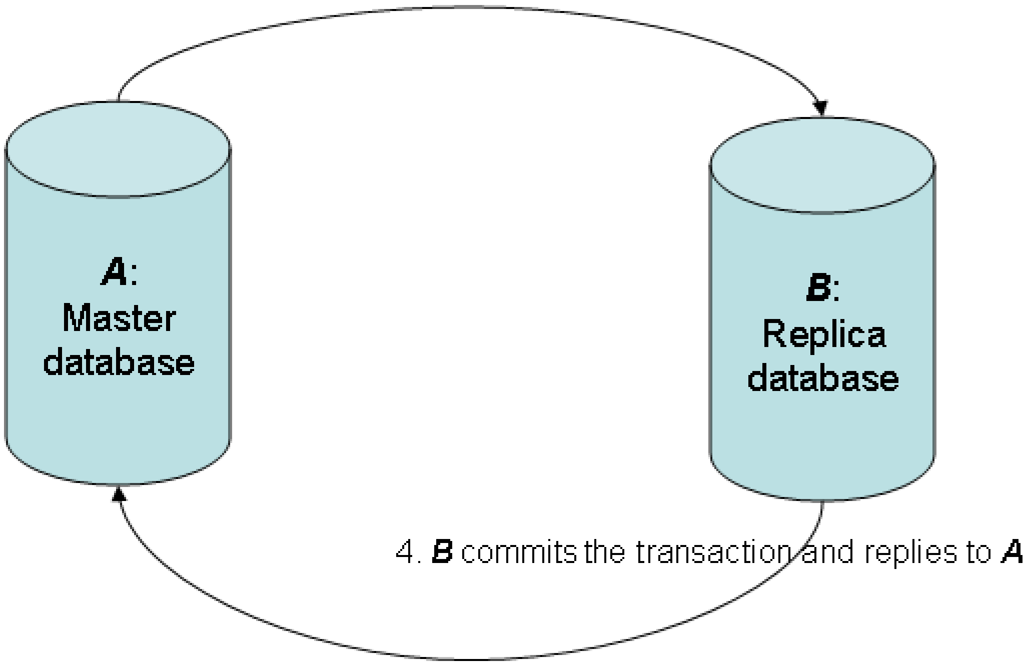
- Types of replication from the transaction perspective
  - Synchronous replication
    - Guarantees 0 data loss, i.e. a transaction either is completed on all the nodes or not at all
    - A transaction is not considered complete without acknowledgment from all the nodes
    - Applications wait for a write transaction to complete before proceeding with further work, hence overall performance decreases considerably.

Distributed architectures

- Types of replication from the transaction perspective
  - Asynchronous replication
    - Transaction is considered complete as soon as the local storage acknowledges it
    - Remote nodes are updated but with (a small) lag
    - Performance is greatly increased, but in case of losing a local storage, the remote storage is not guaranteed to have the current copy of data and most recent data may be lost.

Synchronous replication

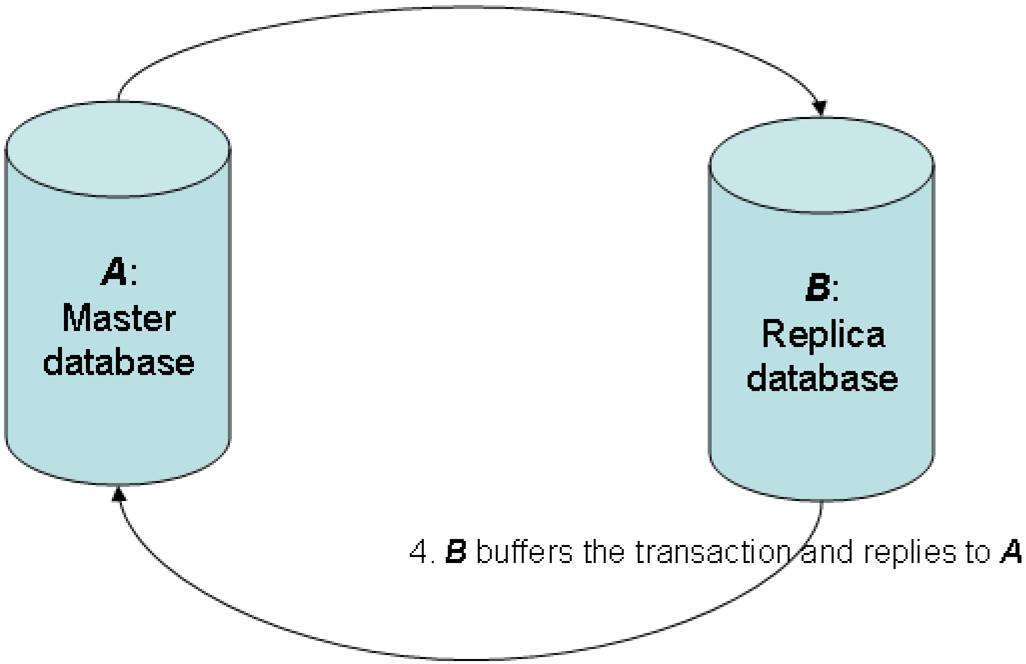
1. Application send a transaction to ***A***
2. ***A*** commits transaction
3. ***A*** sends transaction to ***B*** and wait for reply



5. ***A*** receives the reply from ***B*** and can return from the transaction
6. Application returns from the transaction and can continue

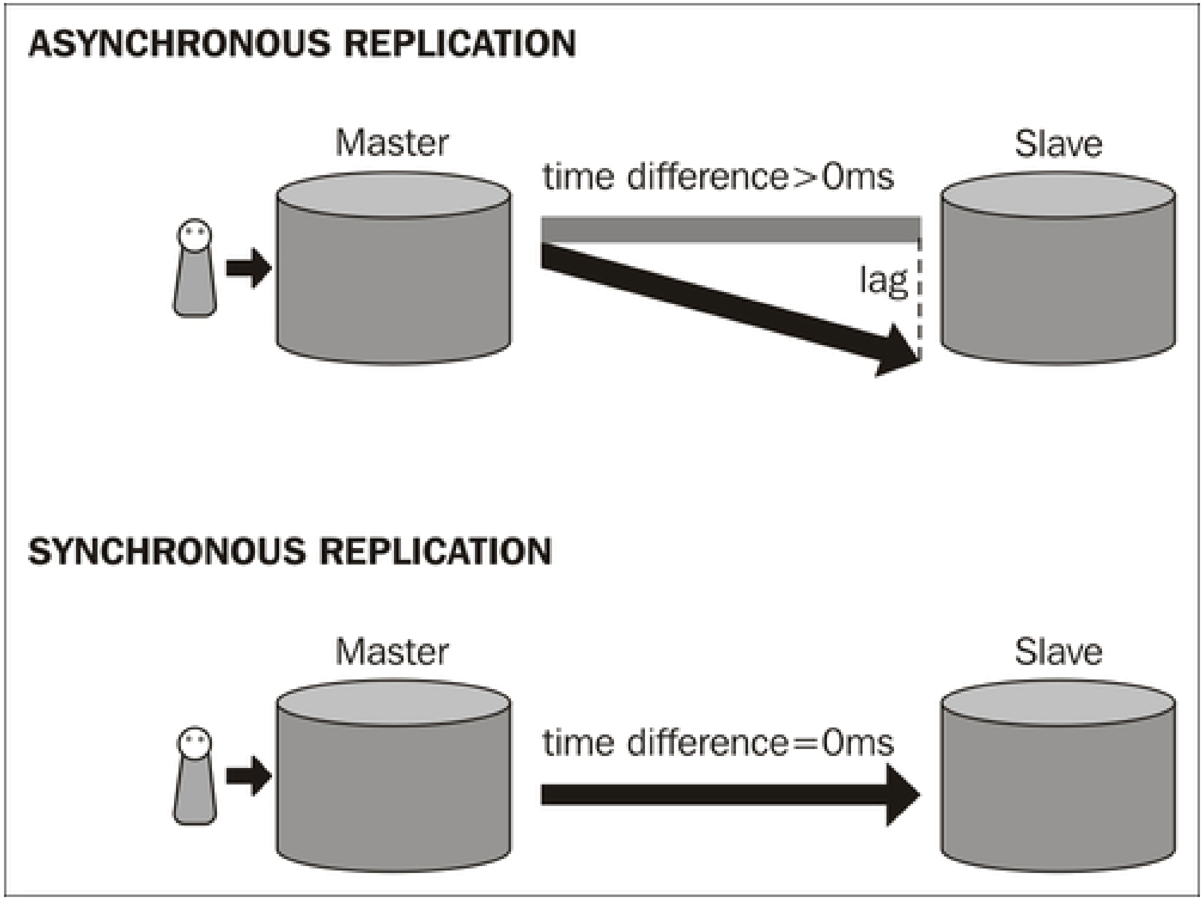
Asynchronous replication

1. Application send a transaction to ***A***
2. ***A*** commits transaction
3. ***A*** sends transaction to ***B*** and wait for reply



5. ***A*** receives the reply from ***B*** and can return from the transaction
6. Application returns from the transaction and can continue

Distributed architectures



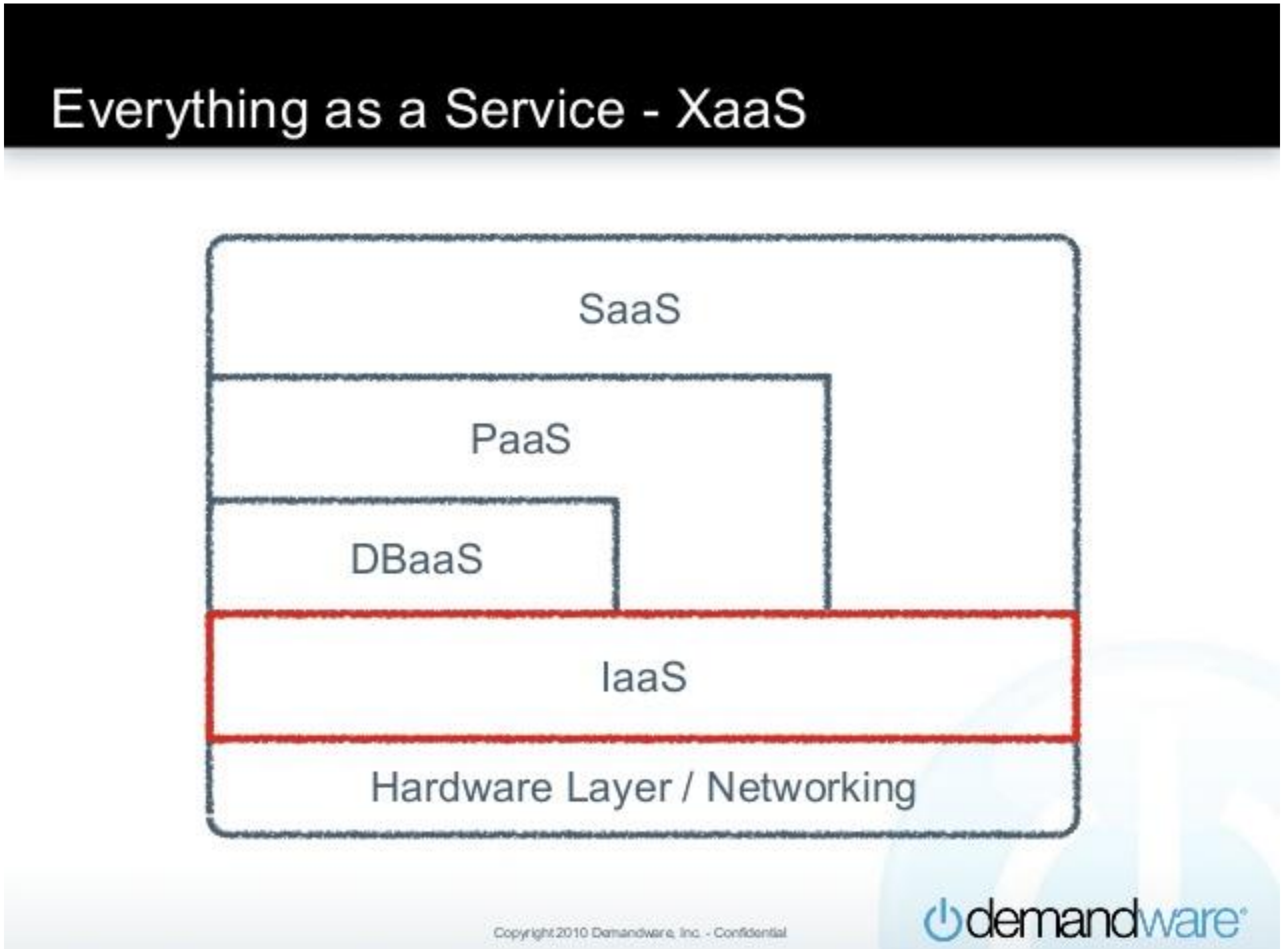
Cloud Computing

- Cloud Computing
  - Is the delivery of on-demand computing resources over the internet on a pay-for-use basis
  - The resources available include everything from data centers to applications
  - Often referred as simply “the cloud”

Cloud Computing

- Cloud Computing
  - Infrastructure as a Service (IaaS)
  - Database as a Service (DBaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)

Cloud Computing



Cloud Computing

- IaaS (Cloud Computing)
  - Is an instant computing infrastructure provisioned and managed over the internet.
  - It provides virtual computing resources over the internet

Cloud Computing

- IaaS features:
  - Provision fundamental computing resources in the form of Virtual Machines (VM)
  - Resources are distributed and support dynamic scaling
  - Deploy and run arbitrary software, meaning middleware, and operating systems;
  - Control over operating systems, storage, and deployed applications
  - User does not manage or control the underlying Cloud infrastructure

Cloud Computing

- DBaaS (Cloud Computing)
  - Is an architectural and operational approach enabling DBAs to deliver database functionality as a service to internal and/or external users
  - Some IT specialists consider it as a part of PaaS, others don’t

- DBaaS features:
  - Pre-configured, automatically installed database services for users
  - Management of database instances using on demand, self-service mechanisms to the users
  - Maximum availability

Cloud Computing

- PaaS (Cloud Computing):
  - Provides a platform allowing users to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an application.
  - Is a complete development and deployment environment in the cloud, with resources that enable you to deliver everything from simple cloud-based apps to sophisticated, cloud-enabled enterprise applications

Cloud Computing

- PaaS features:
  - Deploy applications onto the Cloud infrastructure
  - Integrated development framework: programming languages, tools and libraries are supported by the provider
  - Analytics and or business intelligence tools are provided as a service with PaaS to allow users to analyze and mine their data
  - Control over the deployed applications

Cloud Computing

- SaaS (Cloud Computing):
  - Allows users to connect to and use cloud-based applications over the Internet
  - It is sometimes referred to as “on-demand software”

Cloud Computing

- SaaS features:
  - Applications are supplied by the cloud provider
  - Applications are accessible from various client devices (e.g. web browsers, APIs);
  - Examples: GMAIL, Dropbox, Office 365, Amazon Web Services

Cloud Computing

- Deployment models for Cloud Computing:
  - Private cloud
  - Public cloud
  - Hybrid cloud

Cloud Computing

- Deployment models for Cloud Computing :
  - Private cloud
    - Single-tenant architecture
    - On-premise hardware
    - Direct control of infrastructure
    - E.g. OpenStrack, VMware ESXi

Cloud Computing

- Deployment models for Cloud Computing :
  - Public cloud
    - Multi-tenant architecture
    - Pay-as-you-go pricing model
    - E.g. Amazon AWS, Microsoft Azure, Google Cloud Platform

Cloud Computing

- Deployment models for Cloud Computing :
  - Hybrid cloud
    - Cloud bursting capabilities
    - Benefits both public and private environments
    - Combination of both public and private providers

Storage and tools

- Relational and NoSQL databases
- MapReduce
- Hadoop
- HDSF
- YARN
- Spark

Storage and tools

- Database (DB) is an organized collection of data. The collection of data is organized in schemas, tables, reports, views and other object.
- Database Management Systems (DBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data.
- A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases.

Storage and tools

- There are multiple types of DBMS, usually classified by the way they store data:
  - Relational database management systems (RDBMSs): Oracle, Microsoft SQL Server, MySQL, PostgreSQL, IBM DB2, **Hive**
  - XML Databases: BaseX
  - Object-oriented DBMSs: ObjectDB, Caché
  - NoSQL DBMSs:
    - Key-Value DBMSs: **Riak**, Redis
    - Document-oriented DBMSs: **MongoDB**, Apache CouchDB
    - Column-oriented DBMSs: Apache **HBase**, **Cassandra**
    - Graph DBMSs: Neo4J

Storage and tools

- XML Databases
  - Stores the information in XML format
  - This data can be queried, transformed, exported and returned to a calling system
  - To manipulate the data, different technologies are used:
    - XQuery (XML Query) is a language used to query and transform collations of structured and unstructured data in XML format
    - XPath (XML Path Language) is a query language used for selecting nodes from a XML document
    - XSLT (Extensible Stylesheet Language Transformations) is a language for transforming XML documents into other XML documents or formats (e.g. HTML)

Storage and tools

- Object Oriented Databases:
  - Stores the information in the form of objects
  - Supports Object Oriented Programming principles: Polymorphism, Inheritance and Encapsulation
  - Are integrated to work with different frameworks, e.g. ObjectDB is integrated in Java EE (J2EE) and Spring web application and can be deployed on servlet containers (Tomcat, Jetty) as well as J2EE application servers (GlassFish, JBoss).
  - Support of different query languages, e.g. JDOQL (Java Data Objects Query language – based on the Java syntax), JPSQL (JPA Query Language), etc.

Storage and tools

- NoSQL Databases

Type	Example	
Key-Value Store	redis	riak
Wide Column Store	H·BASE	cassandra
Document Store	mongoDB	CouchDB relax
Graph Store	Neo4j	InfiniteGraph The Distributed Graph Database

Storage and tools

- NoSQL Databases Motives
  - Avoidance of unneeded complexity
    - Avoid strict data consistency
    - ACID properties for transactions are to restrictive
  - High throughput
    - NoSQL database provide a significantly higher data throughput than traditional RDBMSs
  - Easy horizontal scaling
  - Avoidance of expensive object-relational mapping
  - Decrease the complexity and cost of setting up database clusters
  - Compromising reliability for better performance
  - Moving to the cloud

Storage and tools

- NoSQL: Key-Value Databases
  - Stores the information as a map/dictionary
  - Allows clients to put and request values per key
  - Favor high scalability over consistency
  - Omit rich ad-hoc querying and analytics features (join and aggregate operations are set aside)

Storage and tools

- NoSQL: Key-Value Databases

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Storage and tools

- NoSQL: Document-Oriented Databases
  - The next logical step from Key-Value Databases
  - Are schema-less semi-structured databases
  - Support datatypes and collections type (integers, floats, strings, datetime, array, objects, etc.)
  - Supports nested documents
  - MongoDB supports atomic update.

Storage and tools

- NoSQL: Document-Oriented Databases

```
{
  first_name: 'Paul',
  surname: 'Miller',
  cell: 447557505611,
  city: 'London',
  location: [45.123,47.232],
  Profession: ['banking', 'finance', 'trader'],
  cars: [
    { model: 'Bentley',
      year: 1973,
      value: 100000, … },
    { model: 'Rolls Royce',
      year: 1965,
      value: 330000, … }
  ]
}
```

Fields

Typed field values

String

Number

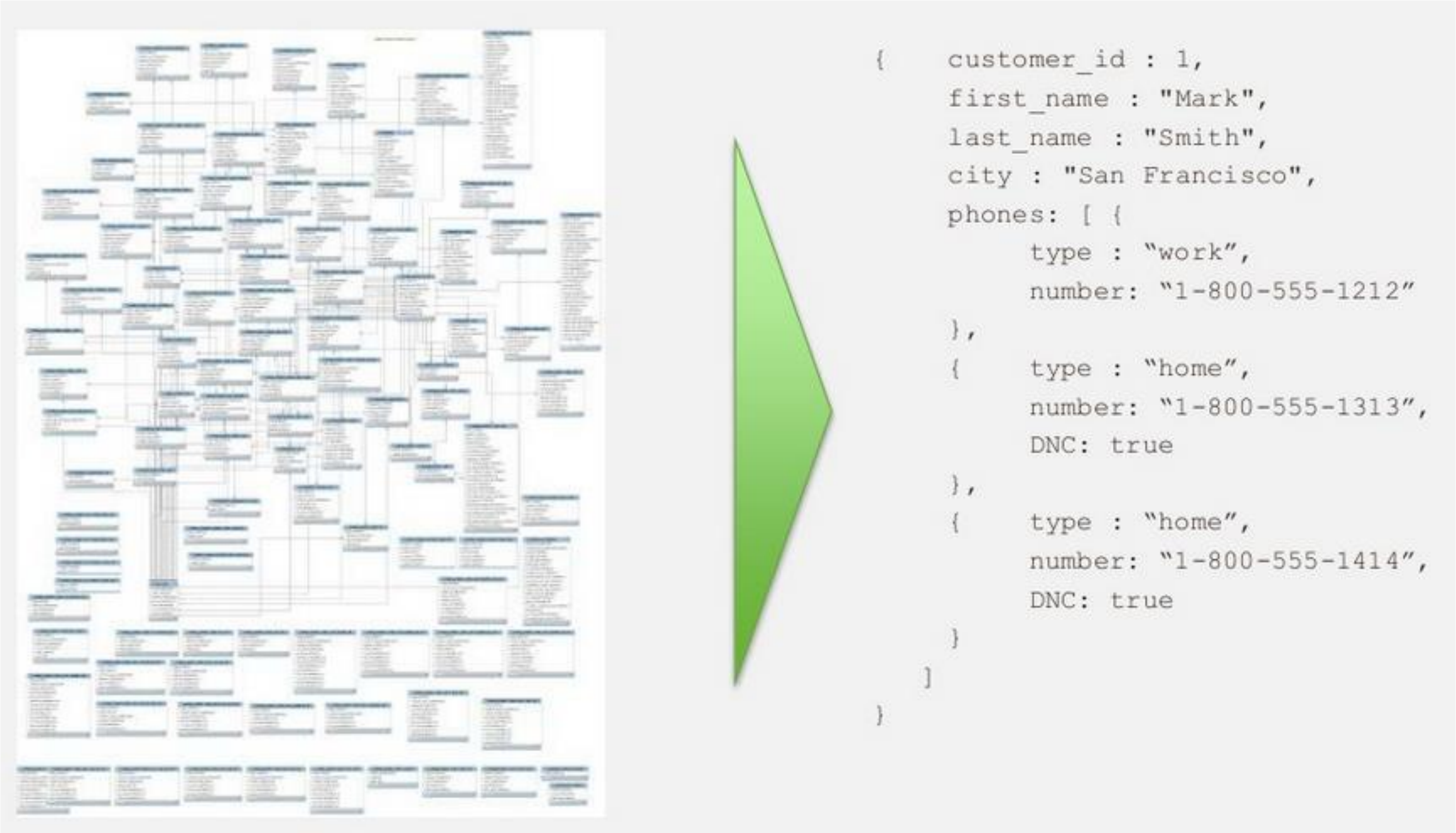
Geo-Coordinates

Fields can contain arrays

Fields can contain an array of sub-documents

Storage and tools

- NoSQL: Document-Oriented Databases



Storage and tools

- NoSQL: Column-Oriented Databases
  - Are used to store and process data by column instead of row
  - Its origin is in analytics and business intelligence
  - Column-stores operate in a shared-nothing massively parallel processing architecture
  - Can be used to build high-performance applications

Storage and tools

- NoSQL: Column-Oriented Databases

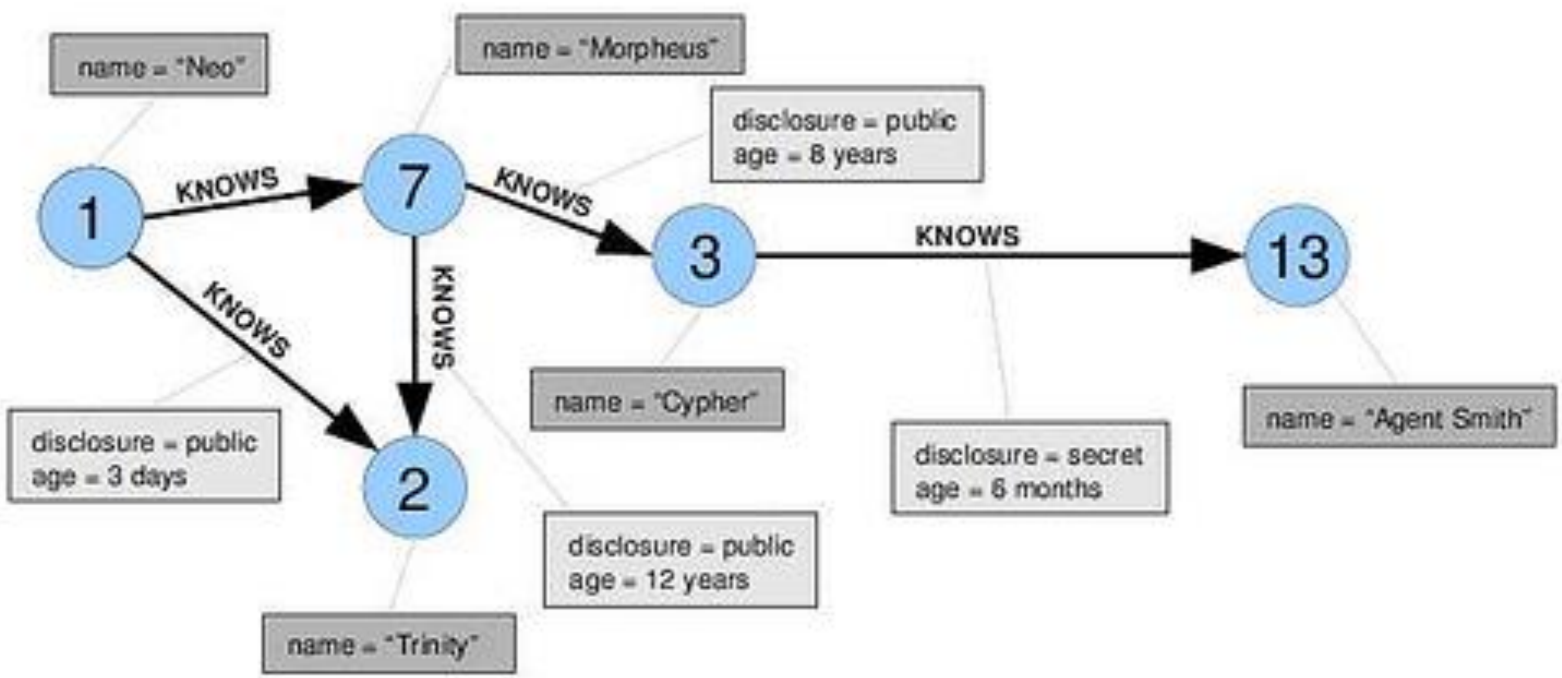
KEY	COLUMN FAMILIES	
ID	CUSTOMERINFO	ADDRESSINFO
1001	FirstName: Tom MiddleName: T LastName: Tester	Address1: 2001 Bayfront Dr. Address2: Suite#813 City: Tampa State: FL Zip: 34637 Country: US
1002	FirstName: Bob MiddleName: B LastName: Builder	Address1: 1234 Sunny Circle City: Beverly Hills State: CA Zip: 90210

Storage and tools

- NoSQL: Graph Databases:
  - Uses graph structures for semantic queries with nodes, edges and properties to represent and store data
  - A key concept is the graph (vertices and edges) which directly relates data items in the store
  - The vertices are the items (data, things)
  - The edges represent the relationship between the data

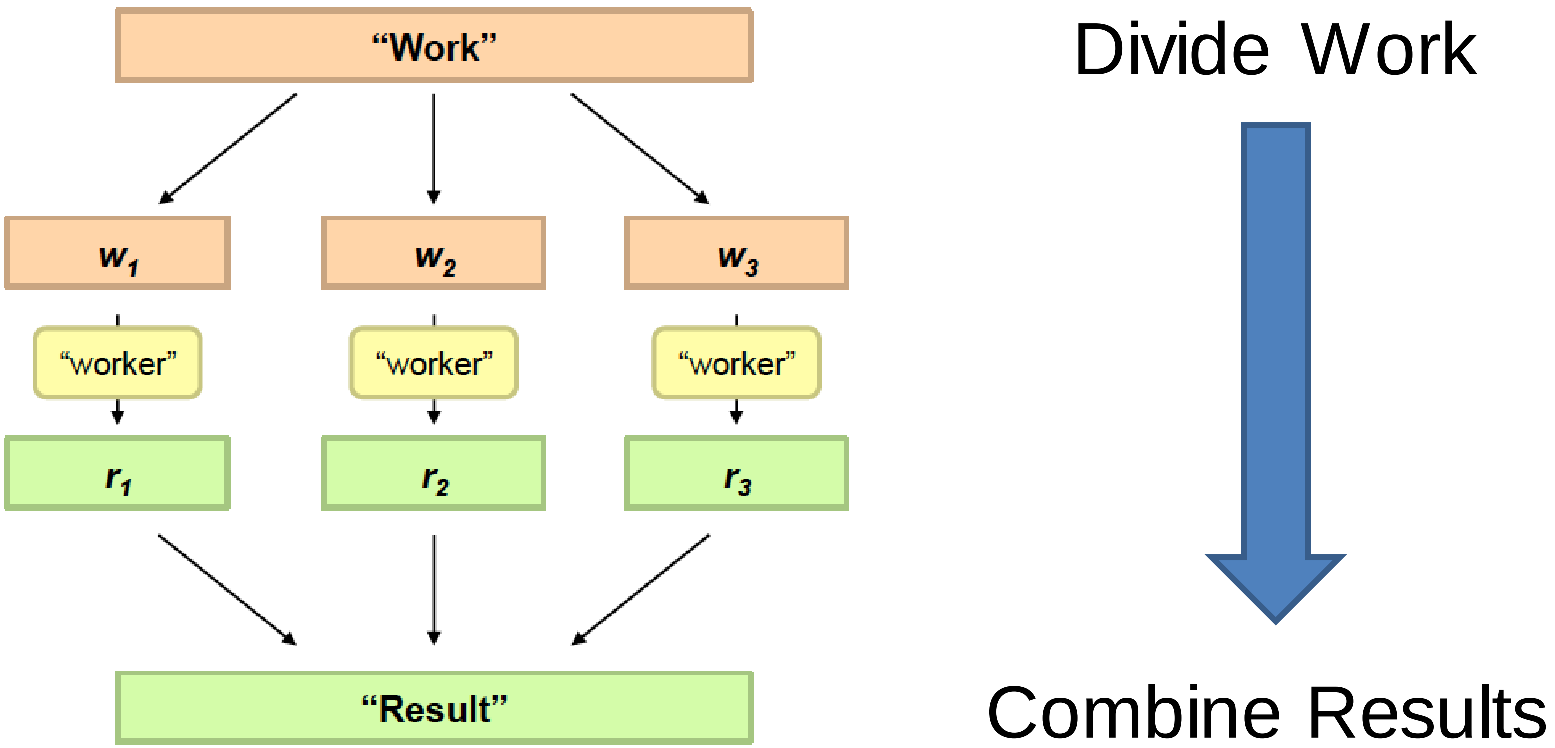
Storage and tools

- NoSQL: Graph Databases



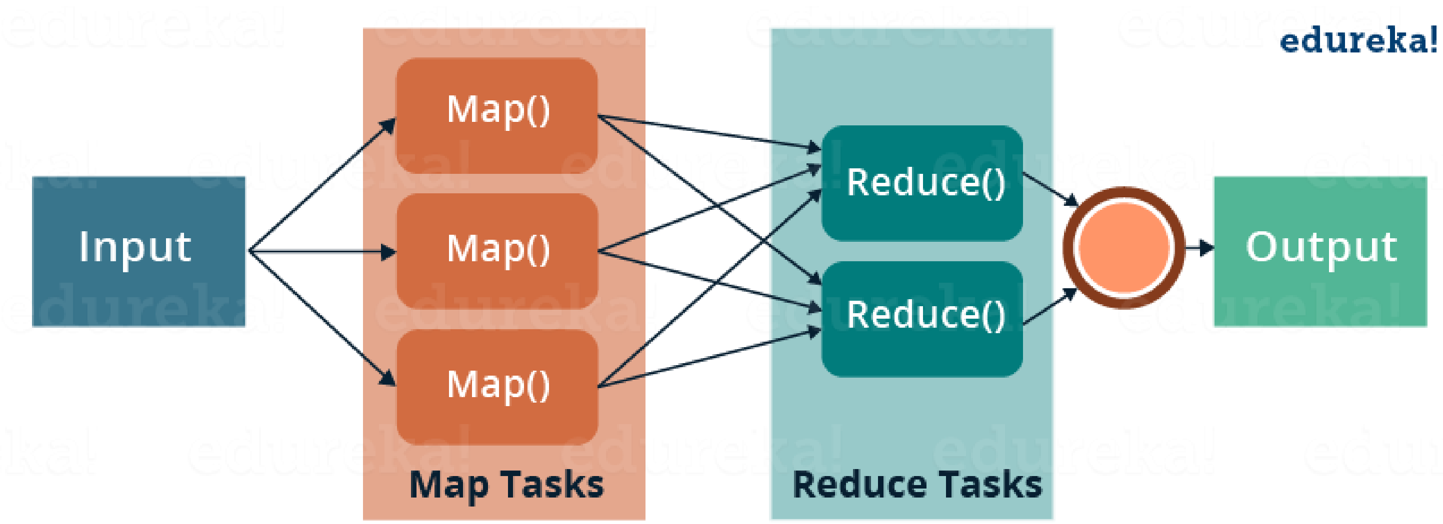
Storage and tools

- Philosophy to Scale for Big Data



Storage and tools

- MapReduce
  - Is a programming framework (paradigm or model) that allows us to perform distributed and parallel processing on large data sets in a distributed environment.

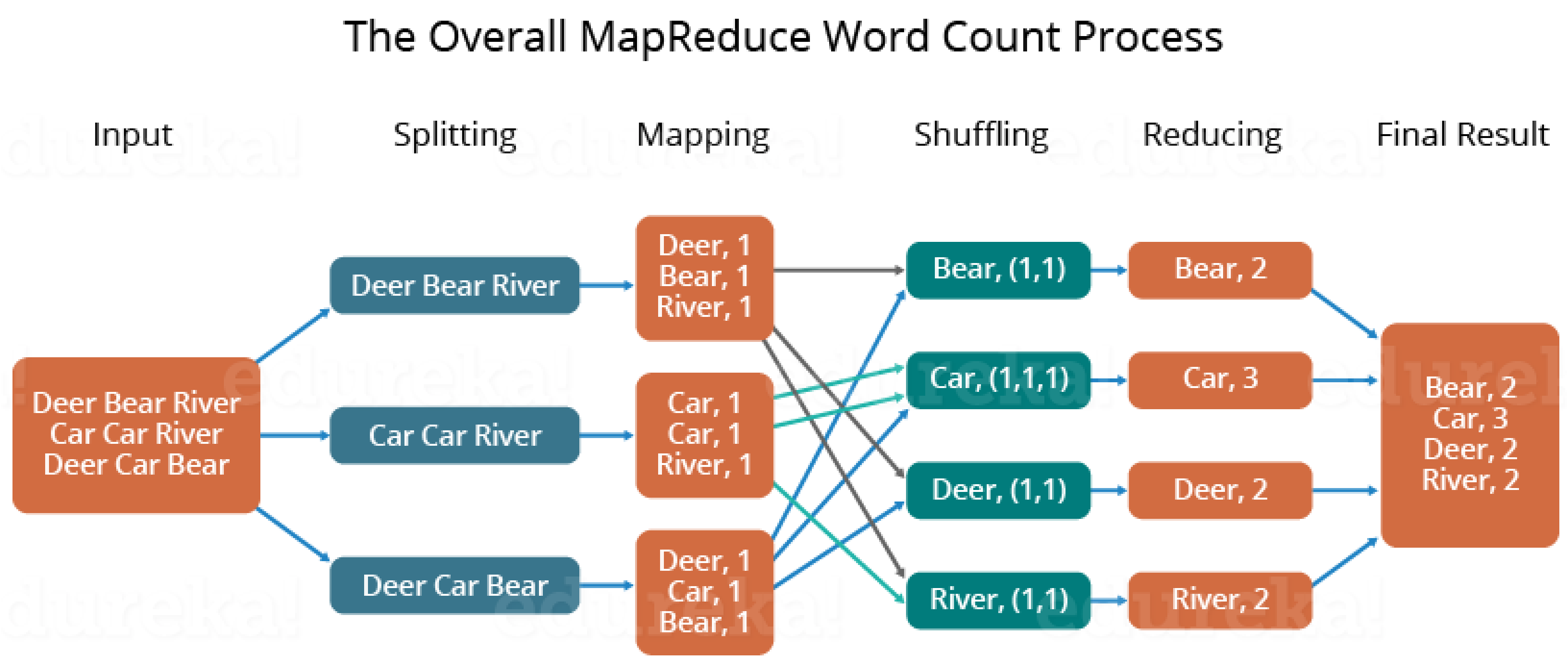


Storage and tools

- MapReduce
  - MapReduce consists of two distinct tasks – Map and Reduce.
  - As the name MapReduce suggests, reducer phase takes place after mapper phase has been completed.
  - The output of a Mapper or map job (key-value pairs) is input to the Reducer.
  - The reducer receives the key-value pair from multiple map jobs.
  - Then, the reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set of tuples or key-value pairs which is the final output.

Storage and tools

- MapReduce word count example



Storage and tools

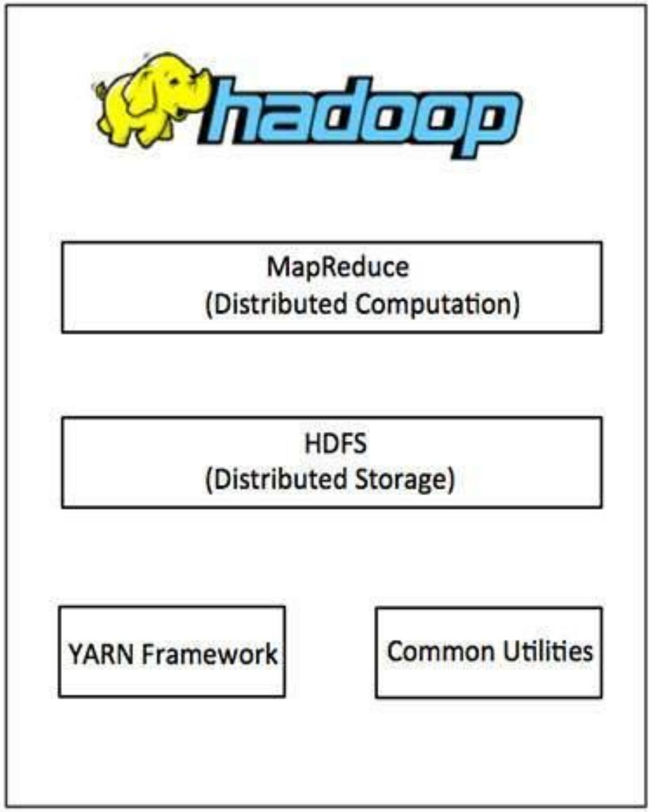
- Hadoop
  - Is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
  - Is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
  - Rather than rely on hardware to deliver high-availability, the framework itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Storage and tools

- Hadoop includes these modules
  - Hadoop Common: The common utilities that support the other Hadoop modules.
  - Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
  - Hadoop YARN: A framework for job scheduling and cluster resource management.
  - Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

Storage and tools

- Hadoop Ecosystem



Storage and tools

- HDFS
  - Hadoop distributed File System is based on Google File System (GFS)
  - Serves as the distributed file system for most tools in the Hadoop ecosystem
  - Scalability for large data sets
  - Reliability to cope with hardware failures

Storage and tools

- HDFS is good for:
  - Large files
  - Streaming data
- NDFS is not good for:
  - Lots of small files
  - Random access to files
  - Low latency access

Storage and tools

- Spark
  - An open-source and fast engine for large-scale data processing.
  - It supports data streaming and SQL, machine learning and graph processing.
  - Spark uses Hadoop’s client libraries for HDFS and YARN

Storage and tools

Storage and tools

- Hadoop Extended Ecosystem