RAMAIAH
Institute of Technology

Cloud Computing and Big Data

Subject Code: CS71 (Credits: 4:0:0)

Textbook:

1. Cloud Computing Theory and Practice – **DAN C**. Marinescu – Morgan Kaufmann Elsevier.
2. Cloud Computing A hands - on approach – Arshdeep Bahga & **Vijay madisetti** Universities press
3. Big Data Analytics, Seema Acharya and Subhashini Chellappan. 2nd edition, Wiley India Pvt. Ltd. 2019

NOTE: I declare that the PPT content is picked up from the prescribed course text books or reference material prescribed in the syllabus book and Online Portals.

Unit III

Cloud Security:

- Risks,
- privacy and privacy impacts assessments,
- Trust, OS, VM security,
- security of virtualization,
- risk posed by shared images, mgmt OS, Xoar,
- Trusted VMM.

Introduction to Big data:

- Types of digital data; Big data – definition,
- characteristics, evolution of Big data, Challenges;
- Comparison with BI; Cloud Computing and Big Data,
- Cloud Services for Big Data, In-Memory Computing Technology for Big Data.

Cloud Security: Risks

- Security has been a concern since the early days of computing. In an **interconnected world**, various embodiments of **malware can migrate easily** from one system to another, cross national borders and infect systems all over the globe.

  - **Threats and vulnerability are part of risks:**

    Risk = Threats x Vulnerabilities

  - Threats (effects) generally can NOT be controlled.
  - **Threats need to be identified**, but they often remain outside of your control. is a function of the enemy's capability and intent

    tThreat = capability x intent

  - Risk CAN be mitigated.
  - Risk can be managed to either lower vulnerability or the overall impact on the business.

    Risk = probability x harm

  - Vulnerability CAN be treated.
  - Weaknesses should be identified and proactive measures taken to correct identified vulnerabilities

Three broad classes of Risk

1. **Traditional Security Threats**
2. **Threats related to System Availability**
3. **Threats related to Third Party Data Control.**

Traditional threats → are those experienced for some time by **any system connected to the Internet**, but with some cloud specific twists.

- The impact is amplified due to the **vast amount of cloud resources** and the **large user population** that can be affected.
- The **fuzzy bounds of responsibility** between the providers of cloud services and users and the difficulties in accurately identifying the cause of a problem adds to cloud users concerns.
- The threat begins at users site. The **user must protect the infrastructure used to connect to cloud** and to interact with the application running on the cloud.
- The task is more difficult because some **components of this infrastructure are outside the firewall** protecting the user.

Authentication and authorization →

DDoS attacks: Distributed denial-of-service, which **prevents legitimate users** accessing cloud services.

Phishing: is an attack aiming to gain information from a site database by masquerading(pretend) as a trustworthy entity.

Such information could be names, **credit card numbers, SSN, or other personal information stored by online merchants** or other service providers.

SQL injection : is a form of attack typically **used against a website**. An SQL command entered in a web form causes the contents of a dbase used by the website to be dumped to the attacker or altered.

Cross–site scripting: A browser permits the attacker to **insert client scripts into the web pages** and thus by pass the access controls at the web site.

Attacks in a cloud computing environment

Three actors involved; six types of attacks possible.

The user can be attacked by:

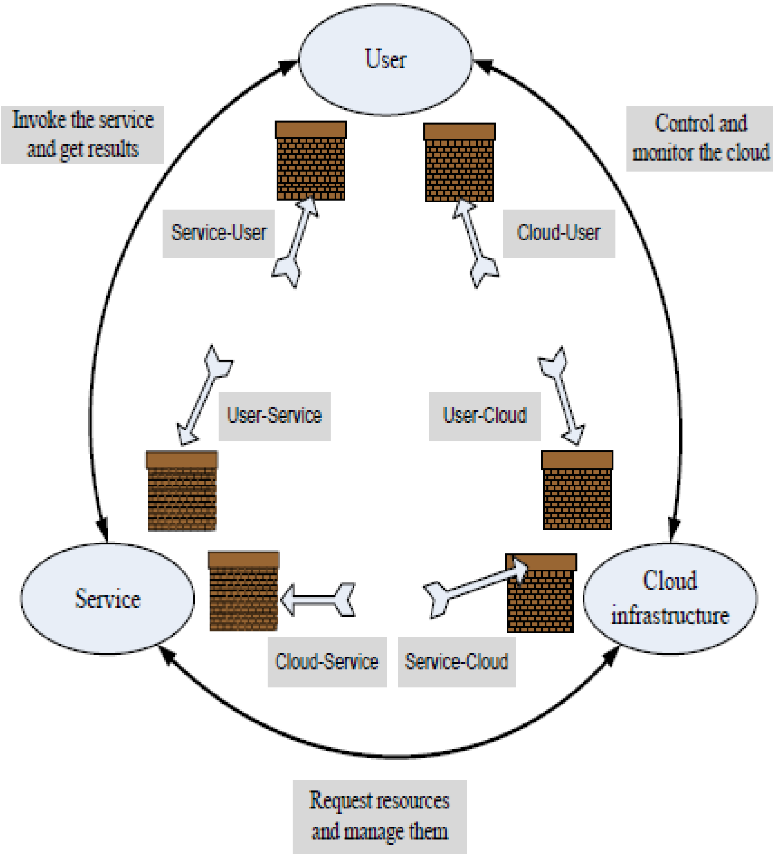
- **Service** → SSL certificate spoofing, attacks on browser caches, or phishing attacks.
- **The cloud infrastructure** → attacks that either originates at the cloud or spoofs to originate from the cloud infrastructure.

The service can be attacked by:

- **A user**→ buffer overflow, SQL injection, and privilege escalation are the common types of attacks.
- **The cloud infrastructure** → the most serious line of attack. Limiting access to resources, privilege-related attacks, data distortion, injecting additional operations.

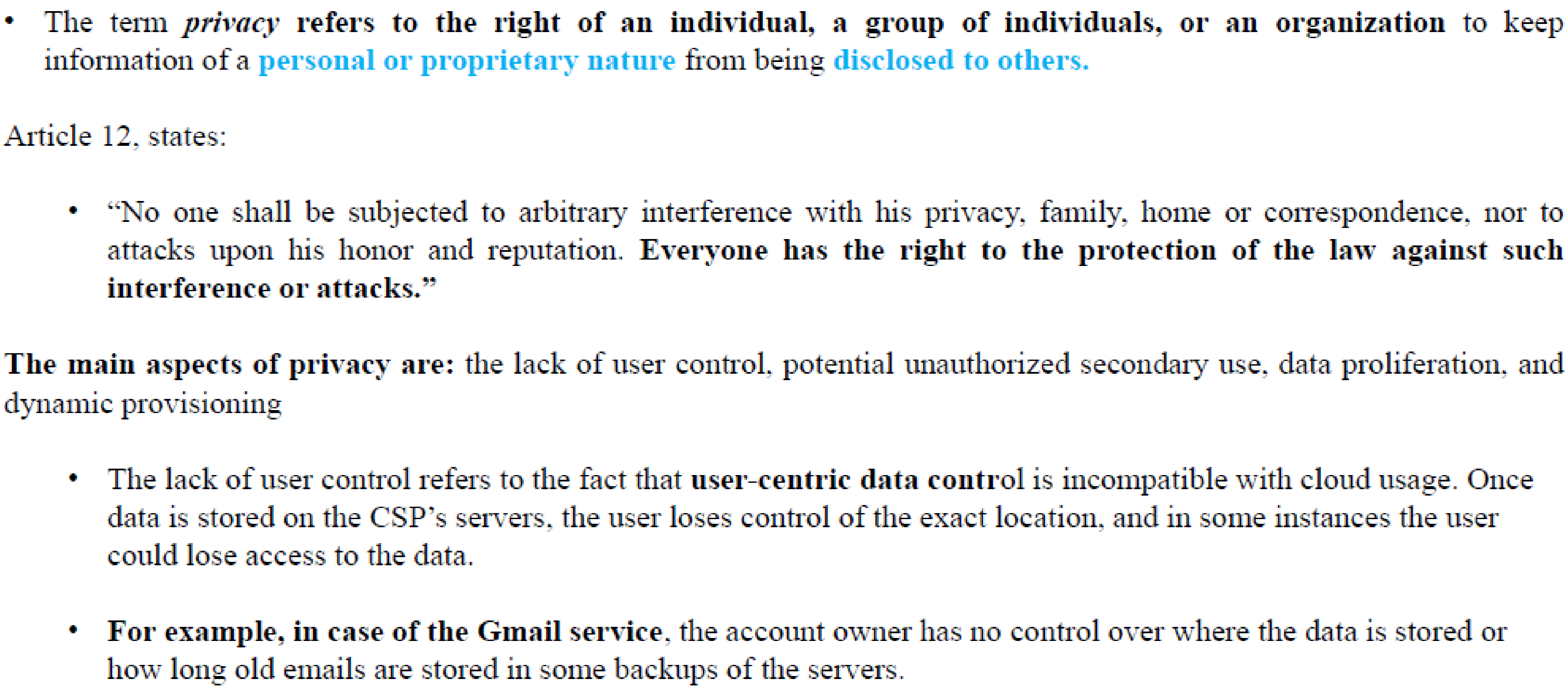
The cloud infrastructure can be attacked by:

- **A user** → targets the cloud control system.
- **A service** → requesting an excessive amount of resources and causing the exhaustion of the resources.



Surfaces of attacks in a cloud computing environment.

Privacy and privacy Impact Assessment



- The term ***privacy*** **refers to the right of an individual, a group of individuals, or an organization** to keep information of a **personal or proprietary nature** from being **disclosed to others.**

Article 12, states:

- "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. **Everyone has the right to the protection of the law against such interference or attacks."**

The main aspects of privacy are: the lack of user control, potential unauthorized secondary use, data proliferation, and dynamic provisioning

- The lack of user control refers to the fact that **user-centric data control** is incompatible with cloud usage. Once data is stored on the CSP's servers, the user loses control of the exact location, and in some instances the user could lose access to the data.

- **For example, in case of the Gmail service**, the account owner has no control over where the data is stored or how long old emails are stored in some backups of the servers.

There is a need for legislation addressing the **multiple aspects of privacy in the digital age**.

- "**Consumer-oriented commercial Web sites** that collect personal identifying information from or about consumers online would be required to comply with the four widely accepted fair information practices:
  - Notice. Web sites would be required to provide consumers **clear and conspicuous notice of their information practices**, including what information they collect, how they collect it, how they use it, how they provide Choice, Access, and Security to consumers
  - Choice. Web sites would be required to **offer consumers choices as to how their personal identifying information is used** beyond the use for which the information was provided (e.g., to consummate a transaction)
  - Access. Web sites would be required to offer **consumers reasonable access to the information a Web site has collected about them**, including a reasonable opportunity to review information and to correct inaccuracies or delete information.
  - Security. Web sites would be required to **take reasonable steps to protect the security of the information they collect from consume**rs. The Commission recognizes that the implementation of these practices may vary with the nature of the information collected and the uses to which it is put, as well as with technological developments.

Trust

- **Trust in the context of cloud computing** is intimately related to the general problem of **trust in online activities.**

- Two conditions must exist for trust to develop.
  - The first condition is ***risk*****, the perceived probability of loss;** indeed, trust would not be necessary if there were no risk involved, if there is a certainty that an action can succeed.
  - The second condition is ***interdependence*****,** the idea that the interests of **one entity cannot be achieved without reliance on other entities.**

- A trust relationship goes though three phases:

  (1) a building phase, when trust is formed;

  (2) a stability phase, when trust exists; and

  (3) a dissolution phase, when trust declines.

- **There are different reasons for and forms of trust**
  - Deterrence-based trust
  - Calculus-based trust
  - Relational trust
  - Persistent trust
  - Dynamic trust

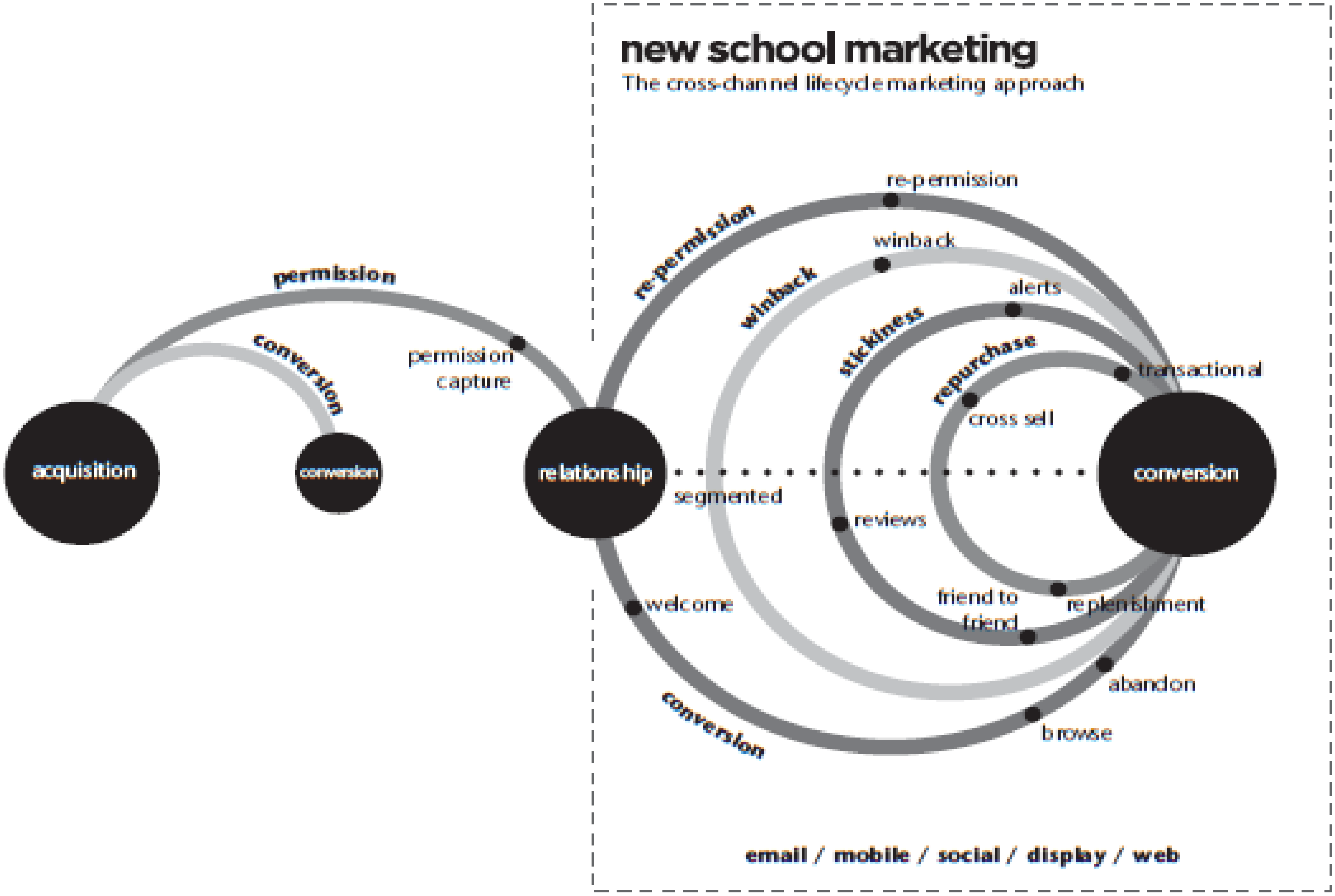


Figure 2.1 New School of Marketing
Source: Responsys Inc. 2012.

The Right Approach: Cross-Channel Lifecycle Marketing:

- Cross-Channel Lifecycle Marketing really starts with the **capture of customer permission, contact information, and preferences for multiple channels**. It also requires marketers to have the right integrated marketing and customer information systems,

(1) They can have **complete understanding of customers** through **stated preferences** and **observed behavior at any given time**; and

(2) They can **automate and optimize** their **programs and processes** throughout the **customer lifecycle**. Once marketers have that, they need a practical framework for planning marketing activities.

- Let 's take a look at the various loops that guide marketing strategies and tactics in the **Cross-Channel Lifecycle Marketing approach**: **conversion, repurchase, stickiness, win-back, and re-permission** (see Figure 2.1).

Operating System Security

- An operating system (OS) allows multiple applications to share the hardware resources of a physical system, subject to a set of policies.
    - A critical function of an **OS is to protect applications against a wide range of malicious attacks** such as **unauthorized access to privileged information, tempering with executable code, and spoofing.**

- **Access control, authentication usage, and cryptographic usage policies** are all elements of mandatory OS security.

    - The first policy specifies how the **OS controls the access to different system objects**,
    - The second defines the authentication mechanisms the OS uses to authenticate a principal, and the last specifies the **cryptographic mechanisms used to protect the data.**

- **Applications with special privileges** that perform security-related functions are called ***trusted applications***. Such applications should only be allowed the lowest level of privileges required to perform their functions.
    - For example, type enforcement is a mandatory security mechanism that can be used to **restrict a trusted application to the lowest level of privileges.**

- A **trusted-path mechanism** is required to prevent malicious software invoked by an authorized application to tamper with the attributes of the object and/or with the policy rules.

Virtual Machine Security

Virtual Machine Security

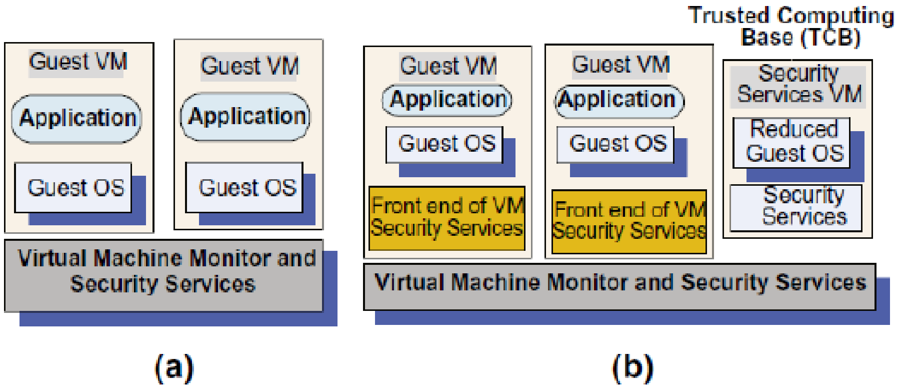

Guest VM
Application
Guest OS
Guest VM
Application
Guest OS
Virtual Machine Monitor and Security Services
(a)
Trusted Computing Base (TCB)
Guest VM
Application
Guest OS
Front end of VM Security Services
Guest VM
Application
Guest OS
Front end of VM Security Services
Security Services VM
Reduced Guest OS
Security Services
Virtual Machine Monitor and Security Services
(b)

FIGURE 9.2

(a) Virtual security services provided by the VMM. (b) A dedicated security VM.

Virtual security services are typically provided by the VMM, as shown in Figure 9.2 (a).
Another alternative is to have **a dedicated security services VM**, as shown in Figure 9.2(b).

A secure trusted **computing base (TCB)** is a necessary condition for security in a virtual machine environment; **if the TCB is compromised, the security of the entire system is affected.**

- VM technology provides **a stricter isolation of virtual machines** from one another than the isolation of processes in a traditional operating system.

- Indeed, a **VMM controls the execution of privileged operations** and can thus enforce memory isolation as well as disk and network access.

- The VMMs are **considerably less complex and better structured** than traditional operating systems; Thus, they are in a better position to respond to security attacks.

- A guest OS runs on simulated hardware, and the **VMM has access to the state of all virtual machines operating** on the same hardware.
- The state of a guest virtual machine can be saved, restored, cloned, and encrypted by the VMM.

The security group involved with the NIST project has identified the following VMM-and VM-based threats:

- **VMM-based threats:**

    1. Starvation of resources and denial of service for some VMs. Probable causes:

        (a) badly configured resource limits for some VMs;

        (b) a rogue VM with the capability to **bypass resource** limits set in the VMM.

    2. VM side-channel attacks. Malicious attacks on one or more VMs by a rogue VM under the same VMM. Probable causes:

        (a) lack of proper isolation of inter-VM traffic due to **misconfiguration of the virtual network** residing in the VMM;

        (b) limitation of packet inspection devices to **handle high-speed traffic**, e.g., video traffic;

        (c) presence of VM instances built from insecure VM images, e.g., a VM image having **a guest OS without the latest patches.**

    3. Buffer overflow attacks.

VM-based threats:

1. Deployment of rogue or insecure VM. **Unauthorized users** may create **insecure instances** from images or may perform unauthorized administrative actions on existing VMs. Probable cause:

    improper configuration of access controls on VM administrative tasks such as instance creation, launching, suspension, reactivation, and so on.

2. Presence of **insecure and tampered VM images** in the VM image repository. Probable causes:

    (a) lack of access control to the VM image repository; (b) lack of mechanisms to verify the integrity of the images, e.g., digitally signed image

Security of Virtualization

- Important virtues of virtualization is that the complete state of an operating system running under a virtual machine is captured by the VM. *This state can be saved in a file and then the file can be copied and shared*. There are several useful implications regarding this fact

1. **Ability to support the IaaS delivery model.** In this model a **user selects an image matching the local en**vironment used by the application
2. **Increased reliability.** An operating system with all the applications running under it can be replicated.
3. **Straightforward mechanisms to implement resource management policies**:
   - **To balance the load of a system**, an OS and the applications running under it can be moved to another server when the load on the current server exceeds a high-water mark.
   - **To reduce power consumption**, the load of lightly loaded servers can be moved to other servers and then these servers can be turned off or set on standby mode.
4. **Improved intrusion prevention and detection**. In a virtual environment a clone can look for known patterns in system activity and detect intrusion. The operator can switch to a hot standby when suspicious events are detected.
5. **Secure logging and intrusion protection**. Intrusion detection can be disabled and logging can be modified by an intruder when implemented at the OS level. When these services are implemented at the VMM/hypervisor layer, the services cannot be disabled or modified.
6. **More efficient and flexible software testing**. Instead of a very large number of dedicated systems running under different operating systems, different versions of each operating system, and different patches for each version, virtualization allows the multitude of OS instances to share a small number of physical systems.

Security risks posed by shared images

- Even when we assume that **a cloud service provider is trustworthy**, many **users either ignore or underestimate** the danger posed by other sources of concern. One of them, especially critical to the ***IaaS* cloud delivery model**, is image sharing.

- For example, a user of AWS has the option to choose between **Amazon Machine Images (AMIs),** accessible through the Quick Start or the **Community AMI** menus of the **EC2 service**.

  - The option of using one of these AMIs is especially tempting for a first-time or less sophisticated user.

  - First, let's review the process to create an AMI. We can start from a running system, from another AMI, or from the image of a VM and **copy the contents of the file system to the S3**, the so-called **bundling.**

  - The first of the **three steps in bundling** is to **create an image**, the second step is to **compress and encrypt the image**, and the last step is to **split the image into several segments** and then upload the segments to the S3.

- Two procedures for the creation of an image are available:
  - **ec2-bundle-image and**
  - **ec2-bundle-volume**

Security risks posed by a management OS

- We often hear that virtualization enhances security because a virtual machine monitor or hypervisor is considerably smaller than an operating system.

- A hypervisor supports stronger isolation between the VMs running under it than the isolation between processes supported by a traditional operating system. Yet the **hypervisor must rely on a management OS** to create VMs and to transfer data in and out from a guest VM to storage devices and network interfaces.

- The **trusted computer base (TCB)** of a cloud computing environment includes not only the **hypervisor** but also the management OS. The **management OS** supports administrative tools, live migration, device drivers, and device emulators.

- For example, the **TCB of an environment based on Xen** includes not only the **hardware** and the **hypervisor** but also the **management operating system** running in the so-called **Dom0** (see Figure 9.3)

- Dom0 manages the building of all **user domains (DomU),** a process consisting of several steps:

  1. Allocate memory in the Dom0 address space and load the kernel of the guest operating system from secondary storage.
  2. Allocate memory for the new VM and use foreign mapping17 to load the kernel to the new VM.
  3. Set up the initial page tables for the new VM.
  4. Release the foreign mapping on the new VM memory, set up the virtual CPU registers, and launch the new VM.

Security risks posed by a management OS

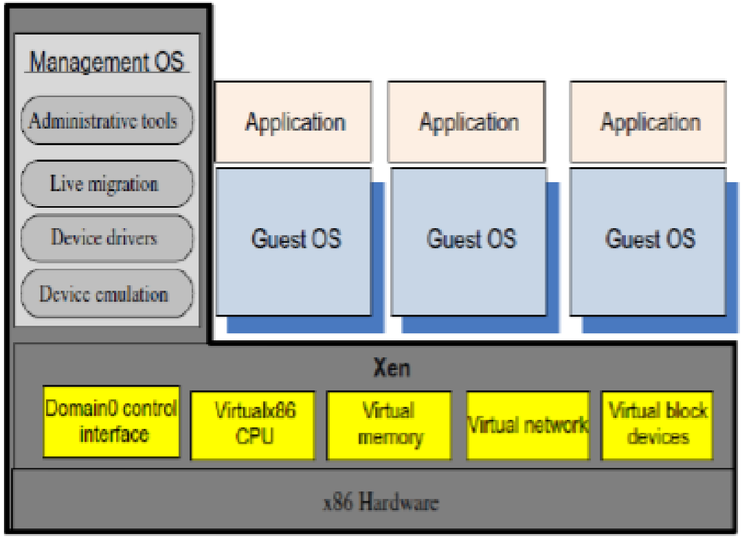


FIGURE 9.3

The trusted computing base of a *Xen*-based environment includes the hardware, *Xen*, and the management operating system running in *Dom0*. The management OS supports administrative tools, live migration, device drivers, and device emulators. A guest operating system and applications running under it reside in a *DomU*.

A malicious Dom0 can play several nasty tricks at the time when it creates a DomU :

- Refuse to carry out the steps necessary to start the new VM, an action that can be considered a **denial-of-service attack.**

- **Modify the kernel of the guest operating system** in ways that will allow a third party to monitor and control the execution of applications running under the new VM.

- **Undermine the integrity of the new VM** by setting the wrong page tables and/or setting up incorrect virtual CPU registers.

- **Refuse to release the foreign mapping and access the memory** while the new VM is running

Xoar : Breaking the monolithic design of the TCB

***Xoar* is a modified version of *Xen* that is designed to boost system security** The security model of *Xoar* assumes that the system is professionally managed and that privileged access to the system is granted only to system administrators. The design goals of Xoar are:

1. Maintain the functionality provided by Xen.
2. Ensure transparency with existing management and VM interfaces.
3. Maintain tight control of privileges; each component should only have the privileges required by its function.
4. Minimize the interfaces of all components to reduce the possibility that a component can be used by an attacker.
5. Eliminate sharing and make sharing explicit whenever it cannot be eliminated to allow meaningful logging and auditing.
6. Reduce the opportunity of an attack targeting a system component by limiting the time window when the component runs.

The Xoar system has four types of components: permanent, self-destructing, restarted upon request, and restarted on timer

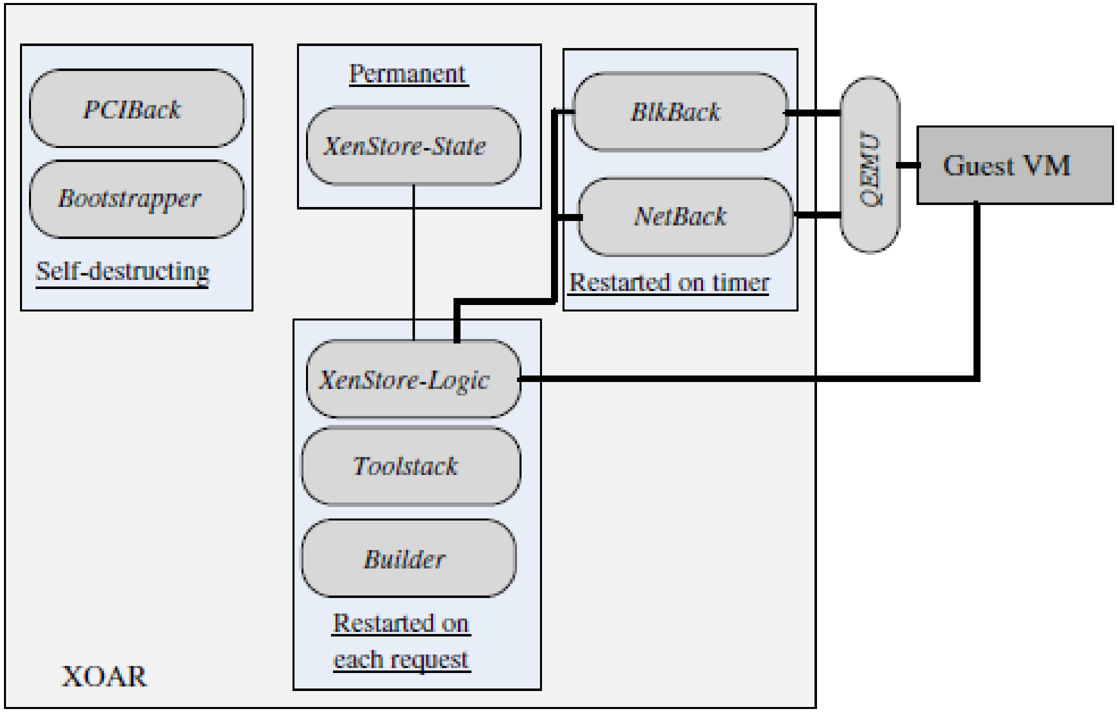


FIGURE 9.4

Xoar has nine classes of components of four types: permanent, self-destructing, restarted upon request, and restarted on timer. A guest VM is started using the *Toolstack* by the *Builder*, and it is controlled by the *XenStore-Logic*. The devices used by the guest VM are emulated by the *QEMU* component.

1. **Permanent components.** XenStore-State maintains all information regarding the state of the system.

2. **Components used to boot the system.** These components self-destruct before any userVMis started. Two components discover the hardware configuration of the server, including the PCI drivers, and then boot the system:

   - **PCIBack.** Virtualizes access to PCI bus configuration.
   - **Bootstrapper**. Coordinates booting of the system.

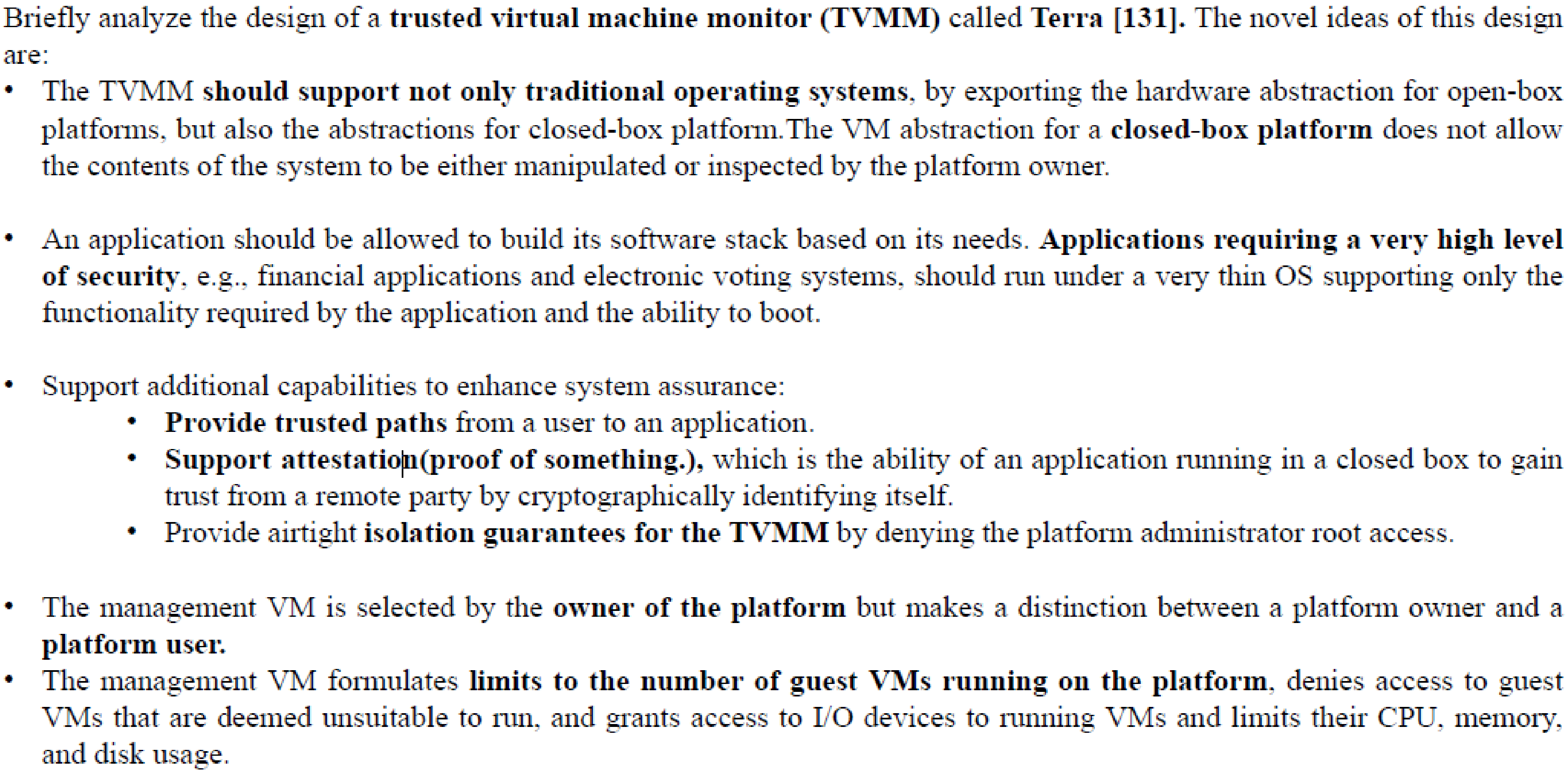
3. **Components restarted on each request:**

   - **XenStore**-Logic.
   - **Toolstack.** Handles VM management requests, e.g., it requests the Builder to create a new guest VM in response to a user request.
   - **Builder.** Initiates user VMs.

4. **Components restarted on a timer.** Two components export physical storage device drivers and the physical network driver to a guest VM:

   - **Blk-Back.** Exports physical storage device drivers using udev21 rules.
   - **NetBack.** Exports the physical network driver.

A trusted virtual machine monitor



Briefly analyze the design of a **trusted virtual machine monitor (TVMM)** called **Terra [131].** The novel ideas of this design are:

- The TVMM **should support not only traditional operating systems**, by exporting the hardware abstraction for open-box platforms, but also the abstractions for closed-box platform.The VM abstraction for a **closed-box platform** does not allow the contents of the system to be either manipulated or inspected by the platform owner.

- An application should be allowed to build its software stack based on its needs. **Applications requiring a very high level of security**, e.g., financial applications and electronic voting systems, should run under a very thin OS supporting only the functionality required by the application and the ability to boot.

- Support additional capabilities to enhance system assurance:
    - **Provide trusted paths** from a user to an application.
    - **Support attestation(proof of something.),** which is the ability of an application running in a closed box to gain trust from a remote party by cryptographically identifying itself.
    - Provide airtight **isolation guarantees for the TVMM** by denying the platform administrator root access.

- The management VM is selected by the **owner of the platform** but makes a distinction between a platform owner and a **platform user.**
- The management VM formulates **limits to the number of guest VMs running on the platform**, denies access to guest VMs that are deemed unsuitable to run, and grants access to I/O devices to running VMs and limits their CPU, memory, and disk usage.

Introduction to Big data:

- Types of digital data; Big data – definition,
- characteristics, evolution of Big data, Challenges;
- Comparison with BI; Cloud Computing and Big Data,
- Cloud Services for Big Data,
- In-Memory Computing Technology for Big Data.

Introduction to Big Data

- The **"Internet of Things"** and its widely ultra-connected nature are leading to a burgeoning(increase rapidly) rise in big data. There is no dearth(scarcity) of data for today's enterprise.

- That brings us to the following questions:

  1. Why is it that we cannot forego big data?
  2. How has it come to assume such magnanimous importance in running business?
  3. How does it compare with the **traditional Business Intelligence (BI)** environment?
  4. Is it here to replace the traditional, relational database management system and data warehouse environment or is it likely to complement their existence?"

- **Big data** is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of traditional database architectures

- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.

Where Is This “Big Data” Coming From ?

12+ TBs of tweet data every day

7 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

76 million smart meters in 2009… 200M by 2014

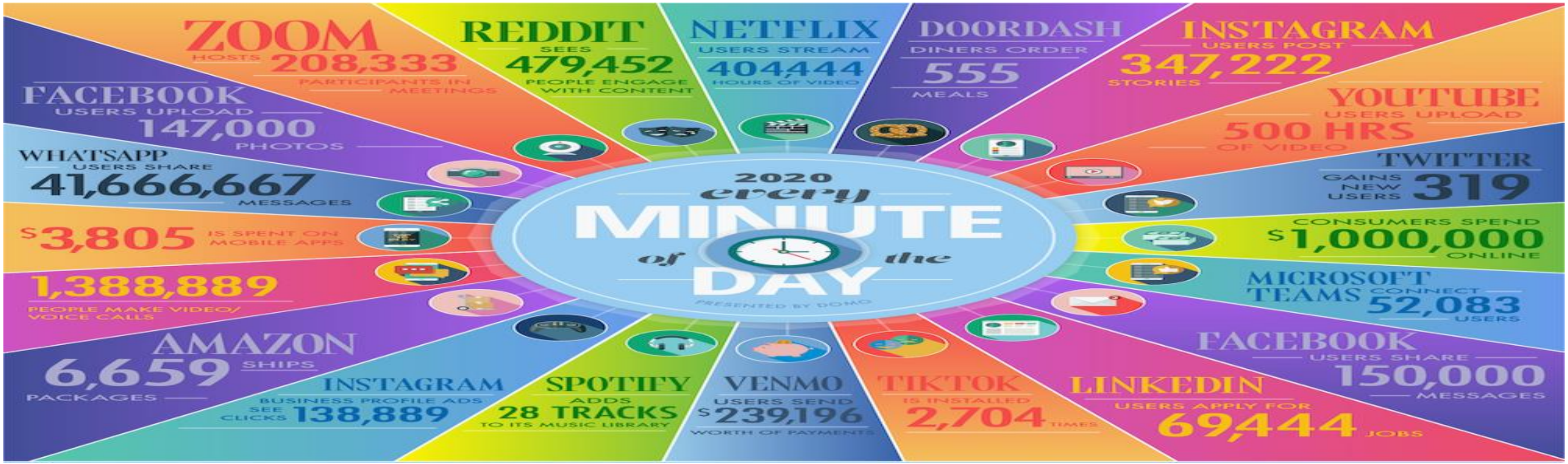
2+ billion people on the Web by end 2011

- **Retail:** Big Data presents many opportunities to improve **sales and marketing analytics**. An example of this is the U.S. retailer Target. After **analyzing consumer purchasing behavior,** Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.

- **Twitter and Facebook generate massive amounts of unstructured data** and use Hadoop and its ecosystem of tools to manage this high volume.

- **social media:** It represents a tremendous opportunity to leverage social and professional interactions to derive new insights.

- **LinkedIn** represents a company in which data itself is the product. Early on, Linkedln founder Reid Hoffman saw the opportunity to create a social network for working professionals



Why Big Data?

Understanding and Targeting Customers

- This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences.
- Using big data, **Telecom companies** can now better predict customer churn; **Wal-Mart** can predict what products will sell, and **car insurance companies** understand how well their customers actually drive. Even **government election campaigns** can be optimized using big data analytics.

2.Understanding and Optimizing Business Processes

- Big data is also increasingly used to optimize business processes.
- **Retailers** are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts.
- One particular business process that is seeing a lot of big data analytics is **supply chain or delivery route optimization**.

Characteristics of Data:

- **Composition:** The composition of data **deals with the structure of data**, that is,
  - the sources of data,
  - the granularity,
  - the types, and
  - the nature of data as to whether it is **static or real-time streaming**.

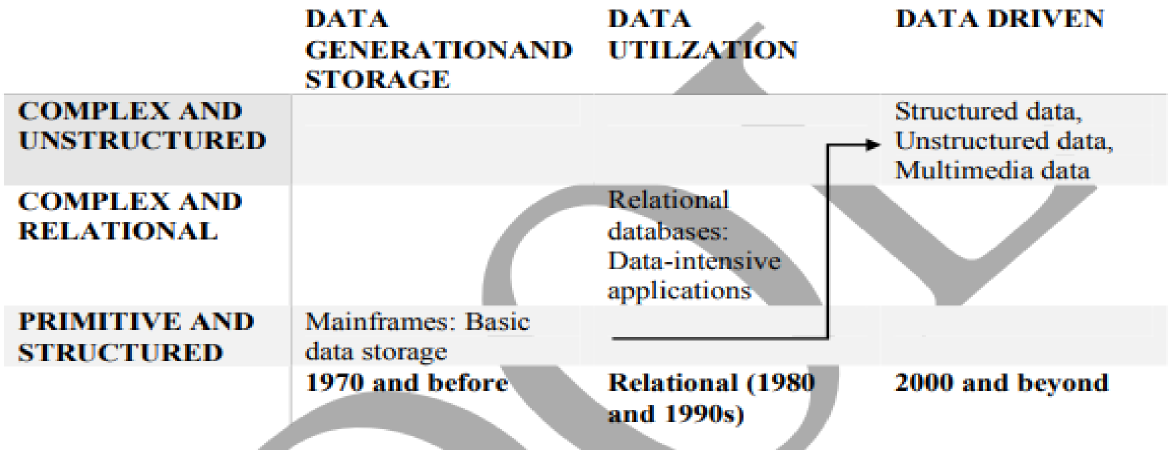
- **Condition:** The condition of data **deals with the state of data**, that is,
  - "Can one use this data as is foranalysis?" or
  - "Does it require cleansing for further enhancement and enrichment?"

- **Context:** The context of data deals with

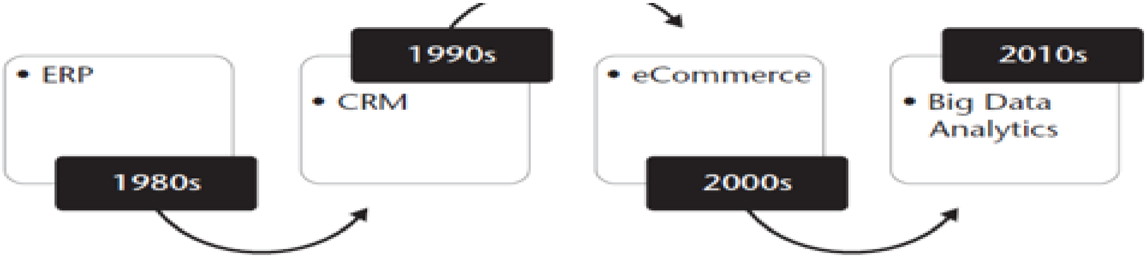
  - **Where has this data been generated?**
  - **"Why was this datagenerated?**
  - **How sensitive is this data?**
  - **What are the events associated with this data?** and so on.

Evolution of Big Data:

- 1970s and before was the era of mainframes. The data was essentially primitive and structured.
- Relational databases evolved in 1980s and 1990s. The era was of data intensive applications.
- The World Wide Web (WWW) and the Internet of Things (IOT) have led to an onslaught of structured, unstructured, and multimedia data



	DATA GENERATIONAND STORAGE	DATA UTILZATION	DATA DRIVEN
COMPLEX AND UNSTRUCTURED			Structured data, Unstructured data, Multimedia data
COMPLEX AND RELATIONAL		Relational databases: Data-intensive applications	
PRIMITIVE AND STRUCTURED	Mainframes: Basic data storage		
	1970 and before	Relational (1980 and 1990s)	2000 and beyond

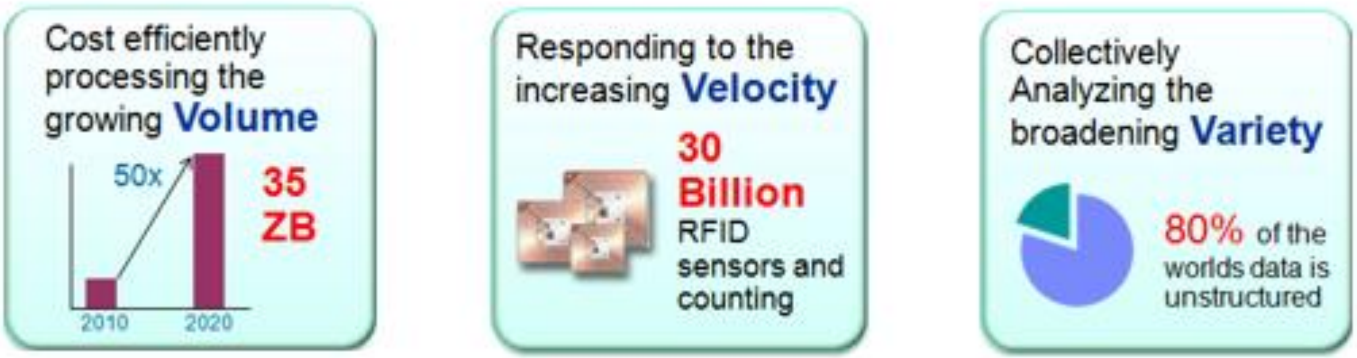


Definition of Big Data:

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

- Big data refers to **datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.**

- Big data in many sectors today will range from a **few dozen terabytes to multiple petabytes (thousands of terabytes).**


Cost efficiently processing the growing Volume
50x
35 ZB
2010
2020
Responding to the increasing Velocity
30 Billion RFID sensors and counting
Collectively Analyzing the broadening Variety
80% of the worlds data is unstructured

- **Part I of the definition:**

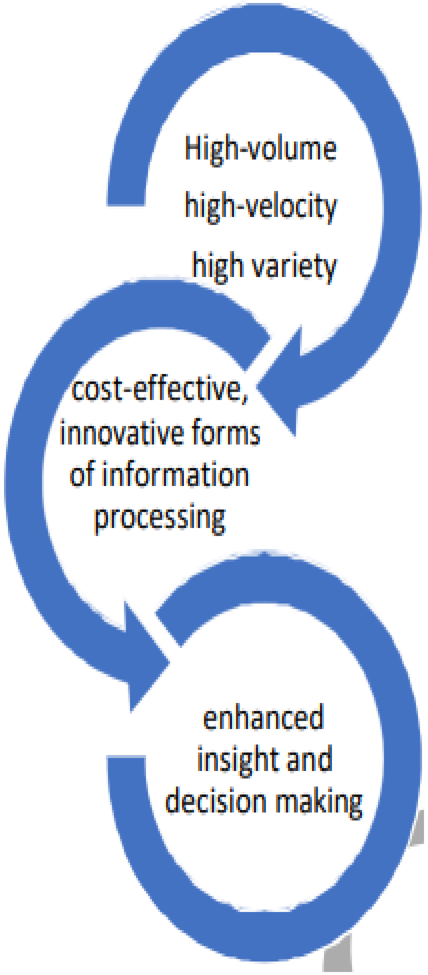
"Big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of structured, semi-structured. and unstructured data) and will require a good speed/pace for storage, preparation, processing and analysis.

- **Part II of the definition:**

"cost effective, innovative forms of information processing" talks about embracing new techniques and technologies to capture (ingest), store, process, persist, integrate and visualize the high-volume, high-velocity, and high-variety data.

- **Part III of the definition:**

"enhanced insight and decision making" talks about deriving deeper, richer and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge


High-volume
high-velocity
high variety
cost-effective, innovative forms of information processing
enhanced insight and decision making

Challenges with big data

Data volume: Data today is growing at an exponential rate. This high tide of data will continue to rise continuously. The key questions are

- “will all this data be useful for analysis?”,
- “Do we work with all this data or subset of it?”,
- “How will we separate the knowledge from the noise?” etc

Storage: Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity and easy upgrading / downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.

Data retention: How long should one retain this data? Some data may require for log-term decision, but some data may quickly become irrelevant and obsolete.

Skilled professionals: In order to develop, manage and run those applications that generate insights, organizations need professionals who possess a high-level proficiency in data sciences.

Other challenges: Other challenges of big data are with respect to capture, storage, search, analysis, transfer and security of big data.

Visualization: Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools.

- There is no explicit definition of how big the data set should be for it to be considered bigdata.
- Data visualization(computer graphics) is becoming popular as a separate discipline. There are very few data visualization experts.

- It’s important to remember that big companies have been **collecting and storing large amounts of data for a long time**

- Today, you have technologies like Hadoop, for example, that make it functionally practical to **access a tremendous amount of data, and then extract value from it**. **The availability of lower-cost hardware makes it easier and more feasible to retrieve and process information, quickly and at lower costs than ever before.**

- The industry has an evolving definition around Big Data that is currently defined by three dimensions:
  1. Volume
  2. Variety
  3. Velocity
- These are reasonable dimensions to quantify Big Data and take into account the typical measures around volume and variety plus introduce the velocity dimension, which is a key compounding factor.

- **Data volume** can be measured by the sheer **quantity of transactions, events, or amount of history that creates the data volume**, but the volume is often further exacerbated by the attributes, dimensions, or predictive variables.

- **Data variety** is the assortment of data. Traditionally data, especially operational data, is "structured" as it is put into a database based on the type of data **(i.e., character, numeric, floating point, etc.**). Over the past couple of decades, data has increasingly become "unstructured" as the sources of data have proliferated beyond operational applications Oftentimes, **text, audio, video, image, geospatial, and Internet data (including click streams and log files)** are considered unstructured data.

- **Data velocity** is about **the speed at which data is created, accumulated, ingested, and processed.** The increasing pace of the world has put demands on businesses to process information in real-time or with near real-time responses.

A Wider Variety of Data

- Internet data (i.e., clickstream, social media, social networking links)
- Primary research (i.e., surveys, experiments, observations)
- Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
- Location data (i.e., mobile device data, geospatial data)
- Image data (i.e., video, satellite image, surveillance)
- Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
- Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)
- The wide variety of data leads to complexities in ingesting the data into data storage. The variety of data also complicates the transformation (or the changing of data into a form that can be used in analytics processing) and analytic computation of the processing of the data.

Traditional Business Intelligence Vs Big Data

- In traditional BI environment, all the enterprise's data is housed in a **central server** whereas in a big data environment data resides in a **distributed file system.**

    - The distributed file system scales by scaling in(decrease) or out(increase) horizontally as compared to typical database server that scales vertically.

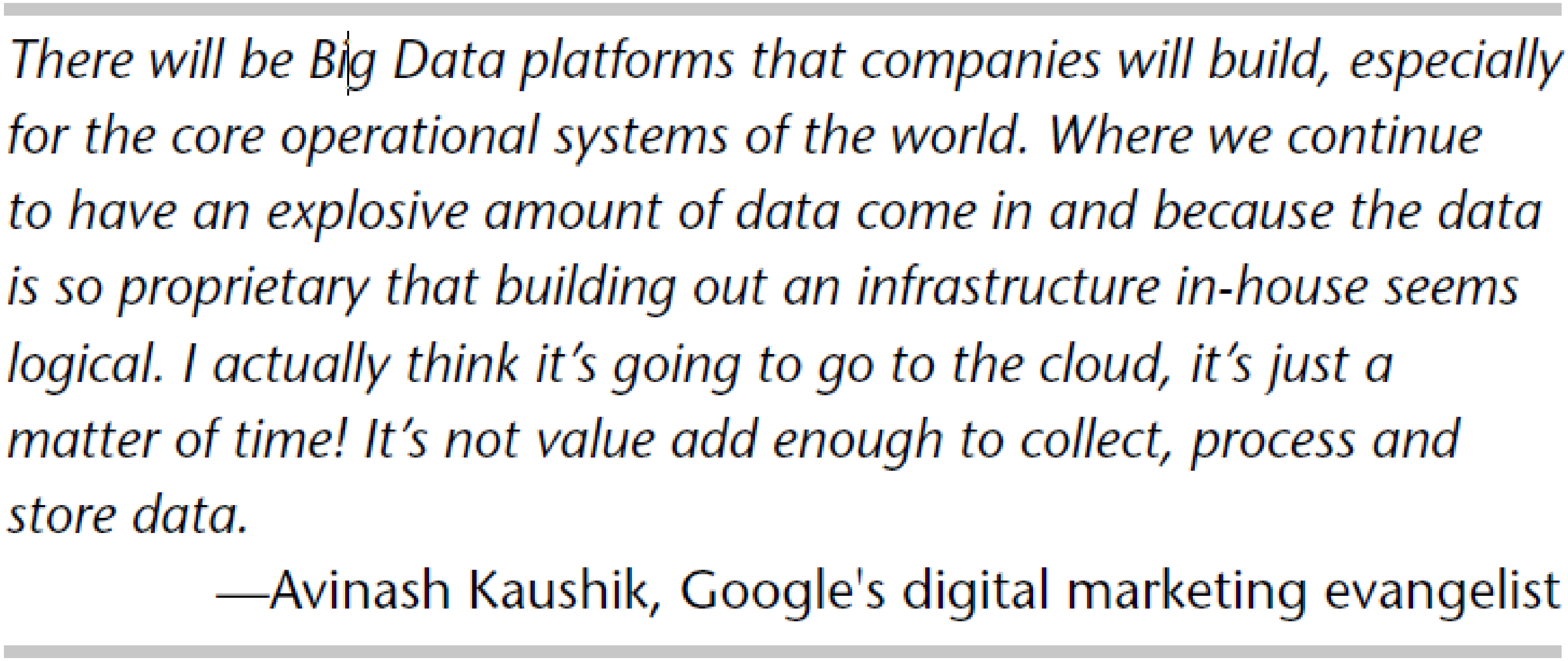
- In traditional BI, data is generally analysed in an **offline mode**

    - whereas in big data, it is analysed in both **real-time streaming** as well as in offline mode.

- Traditional BI is about **structured data** and it is here that data is taken to processing functions (move data to code)

    - whereas big data is about **variety: Structured, semi structured, and unstructured data** and here the processing functions are taken to the data (move code to data).

Cloud Computing and Big Data



> *There will be Big Data platforms that companies will build, especially for the core operational systems of the world. Where we continue to have an explosive amount of data come in and because the data is so proprietary that building out an infrastructure in-house seems logical. I actually think it's going to go to the cloud, it's just a matter of time! It's not value add enough to collect, process and store data.*
>
> —Avinash Kaushik, Google's digital marketing evangelist

With a cloud model, **you pay on a subscription basis** with no upfront capital expense. You don't incur the typical 30 percent maintenance fees—and all the updates on the platform are automatically available

The traditional cost of value chains is being completely disintermediated by **platforms—massively scalable platforms** where the marginal cost to deliver an incremental product or service is zero.

The ability to build massively scalable platforms—platforms where you have the option to keep **adding new products and services for zero additional cost**—is giving rise to business models that weren't possible before

Mehta calls it **"the next industrial revolution**, **where the raw material is data and data factories replace manufacturing factories."**

He pointed out a few **guiding principles** that his firm stands by:

- **Stop saying "cloud."**
- **Acknowledge the business issues**
- **Fix some core technical gaps**

Stop saying “cloud.” :

- It ’s not about the fact that it is virtual, but the **true value lies in delivering software, data, and/or analytics in an “as a service” model.**
- Whether that is in a private hosted model or a publicly shared one does not matter. **The delivery, pricing, and consumption model matters**.

Acknowledge the business issues.

- There is no point to make light of matters around information **privacy, security, access, and delivery.**
- These issues are real, more often than not heavily regulated by multiple government agencies, and unless dealt with in a solution, will kill any platform sell.

Fix some core technical gaps.

- Everything from the ability to run analytics at scale in a virtual environment to **ensuring information processing and analytics authenticity are issues that need solutions and have to be fixed.**

Cloud Services for Big Data

Predictive Analytics Moves into the Limelight

- Enterprises will move from being in reactive positions **(business intelligence) to forward leaning positions (predictive analytics).**

- Using all the data available—traditional **internal data sources combined with new rich external data sources**—will make the **predictions more accurate and meaningful**.

- **Algorithmic trading** and **supply chain** optimization are just two typical examples where predictive analytics have greatly reduced the friction in business.

- Look for predictive analytics to proliferate in every facet of our lives, both personal and business. **Here are some leading trends** that are making their way to the forefront of businesses today:

Recommendation engines similar to those used in **Netflix and Amazon** that use past purchases and buying behavior to recommend new purchases.

Risk engines for a wide variety of business areas, including **market and credit risk, catastrophic risk, and portfolio risk**.

Innovation engines for new product innovation, **drug discovery,** and **consumer and fashion trends** to **predict potential new product formulations and discoveries**.

Customer insight engines that integrate a wide variety of customer related info, including **sentiment, behavior, and even emotions**.

- **Customer insight engines will be the backbone** in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to **maximize their spend**.

Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how **to seek natural resources to maximize output while reducing operational costs— or what potential competitive strategies should be used in a global business** that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

In-Memory Computing Technology for Big Data

The Elephant in the Room: Hadoop 's Parallel World

- At one-tenth the cost of traditional solutions, **Hadoop excels at supporting complex analyses**— including detailed, special-purpose computation—across large collections of data.

- Hadoop **handles a variety of workloads**, **including search, log processing, recommendation systems, data warehousing, and video/image analysis.**

- **Apache Hadoop is an open-source project** administered by the Apache Software Foundation. The software was originally **developed by the world 's largest Internet companies to capture and analyze the data that they generate.**

- Hadoop **stores terabytes, and even petabytes, of data inexpensively**. It is **robust and reliable and handles hardware and system failures automatically**, without losing data or interrupting data analyses.

- Hadoop runs on **clusters of commodity servers** and each of those servers has local CPUs and disk storage that can be leveraged by the system.

The Two critical components of Hadoop are:

The Hadoop Distributed File System (HDFS) .

- HDFS is the **storage system** for a **Hadoop cluster**.
- When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the **different servers participating in the cluster**.
- Each server stores just a small fragment of the complete data set, and each piece of **data is replicated on more than one server**.

MapReduce.

- Because Hadoop stores the entire dataset in small pieces across a collection of servers, **analytical jobs can be distributed, in parallel**, to each of the servers storing part of the data.
- Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a **comprehensive answer.**
- MapReduce is the agent that **distributes the work** and **collects the results**

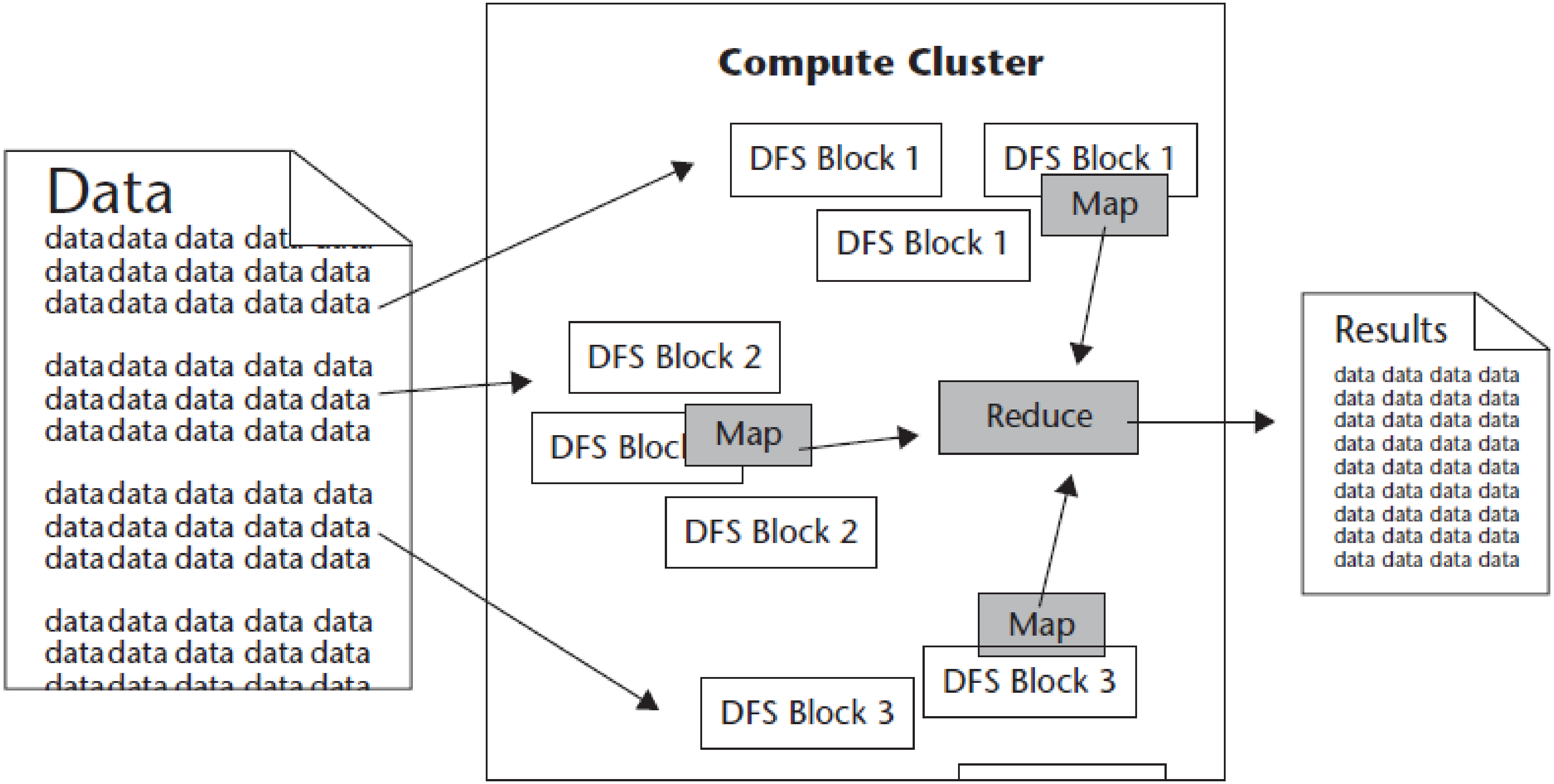
HDFS continually monitors the data stored on the cluster.

- If a **server becomes unavailable, a disk drive fails, or data is damaged**, whether due to hardware or software problems,
  - HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster.

MapReduce monitors progress of each of the servers participating in the job.

- If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data.

Because of the way that HDFS and MapReduce work, Hadoop provides **scalable, reliable, and fault-tolerant services** for **data storage and analysis** at very low cost.



Source: Apache Software Foundation.