Chapter -8

Scalable and Emerging Information System Techniques

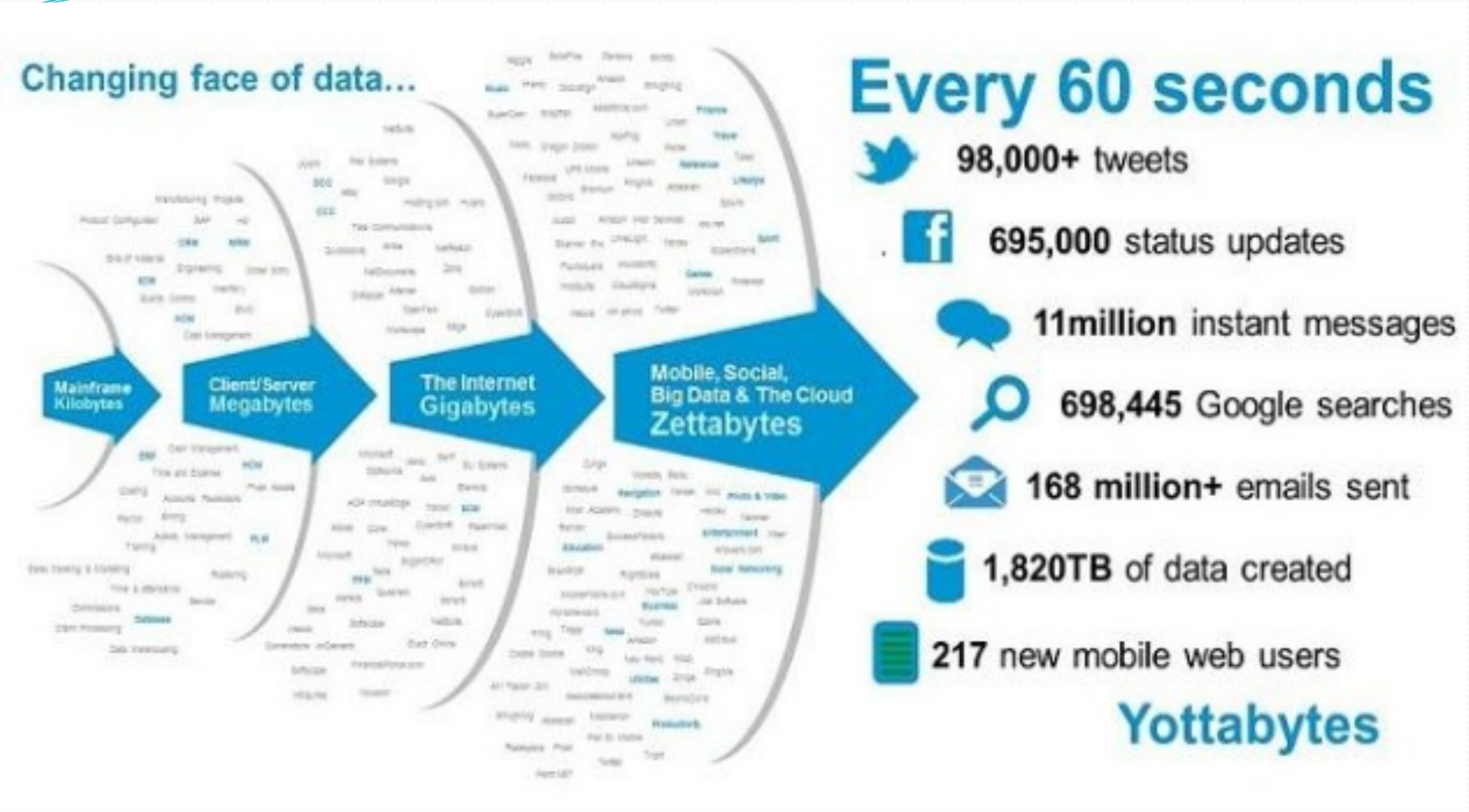
Information System *(CT 751)*

BCT IV/II

By: Shayak Raj Giri

Outline

- **Scalable and Emerging information System techniques**
  - Techniques for voluminous data
  - Cloud computing technologies and their types
  - Map Reduce and Hadoop systems
  - Data management in the cloud
  - Information retrieval in the cloud
  - Link analysis in cloud setup
  - Case studies of voluminous data environment

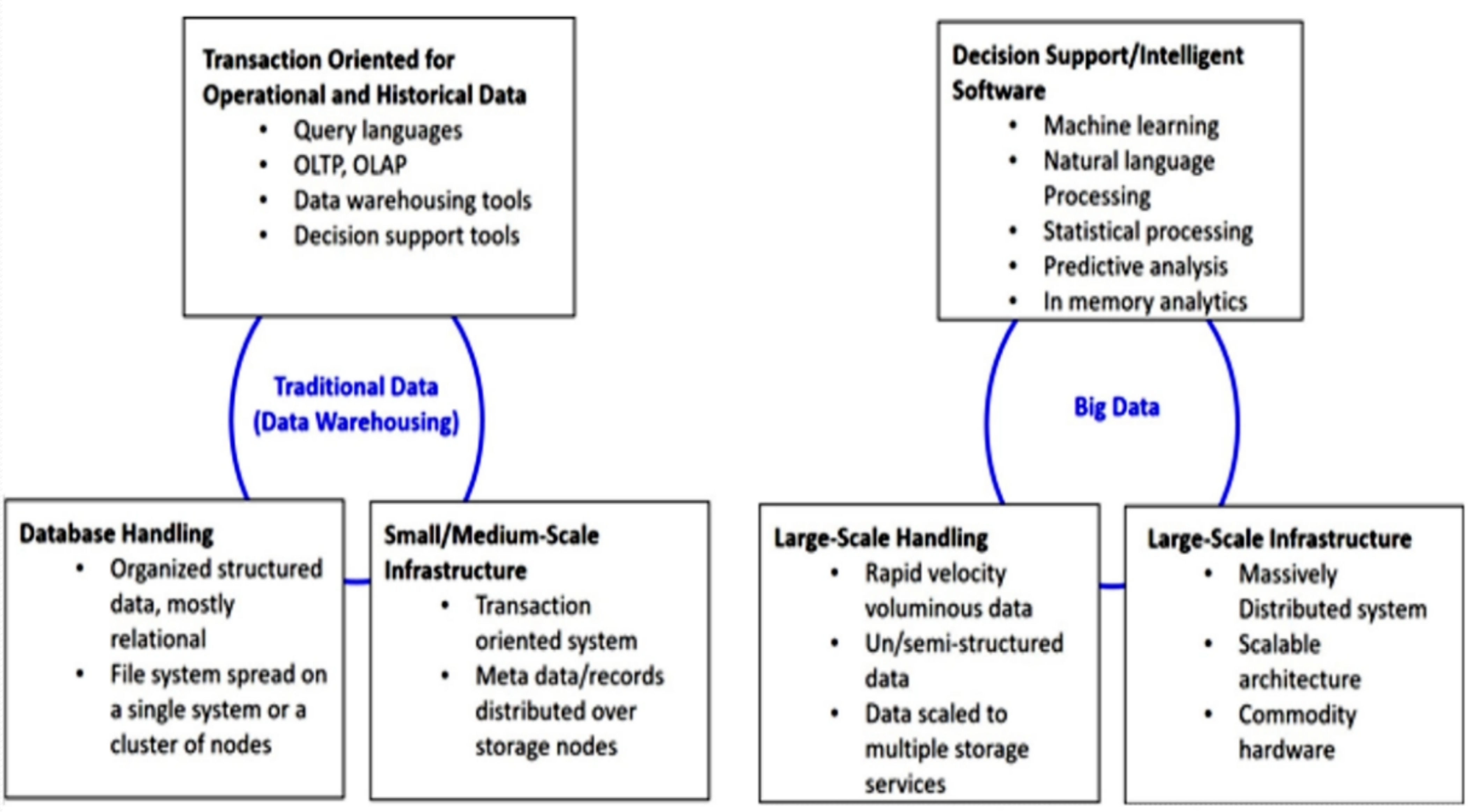


Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users

Yottabytes

Traditional Data vs. Big Data



Transaction Oriented for Operational and Historical Data
- Query languages
- OLTP, OLAP
- Data warehousing tools
- Decision support tools

Traditional Data (Data Warehousing)

Database Handling
- Organized structured data, mostly relational
- File system spread on a single system or a cluster of nodes

Small/Medium-Scale Infrastructure
- Transaction oriented system
- Meta data/records distributed over storage nodes

Decision Support/Intelligent Software
- Machine learning
- Natural language Processing
- Statistical processing
- Predictive analysis
- In memory analytics

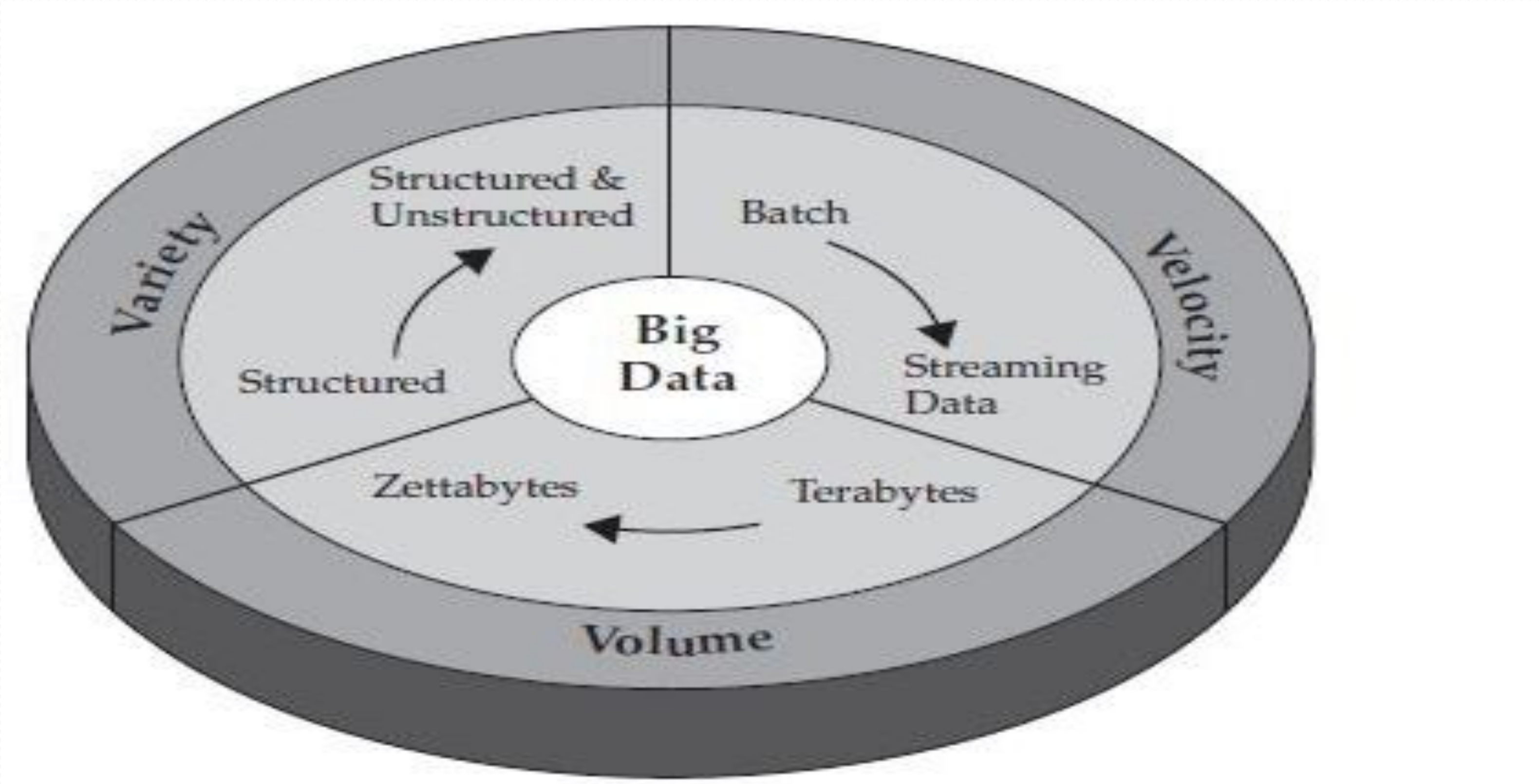
Big Data

Large-Scale Handling
- Rapid velocity voluminous data
- Un/semi-structured data
- Data scaled to multiple storage services

Large-Scale Infrastructure
- Massively Distributed system
- Scalable architecture
- Commodity hardware

Big Data

- Big Data applies to information that can't be processed or analyzed using traditional processes or tools.



IBM characterizes Big Data by its volume, velocity, and variety—or simply,

5 V's: Characteristics of Big Data

- **Volume** refers to the vast amounts of data generated every second. Just think of all the emails, twitter messages, photos, video clips, sensor data etc. we produce and share every second. We are not talking Terabytes , but Zettabytes or Brontobytes.
- **Velocity** refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds.
- **Variety** refers to the different types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases.
- With big data technology we can now harness differed types of data (**structured and unstructured**) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.

5 V’s: Characteristics of Big Data

- **Veracity** refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable but big data and analytics technology now allows us to work with these type of data.
- **Value** is the most important aspect in the big data. Though, the potential value of the big data is huge. It is all well and good having access to big data but unless we can turn it into value , becomes useless.

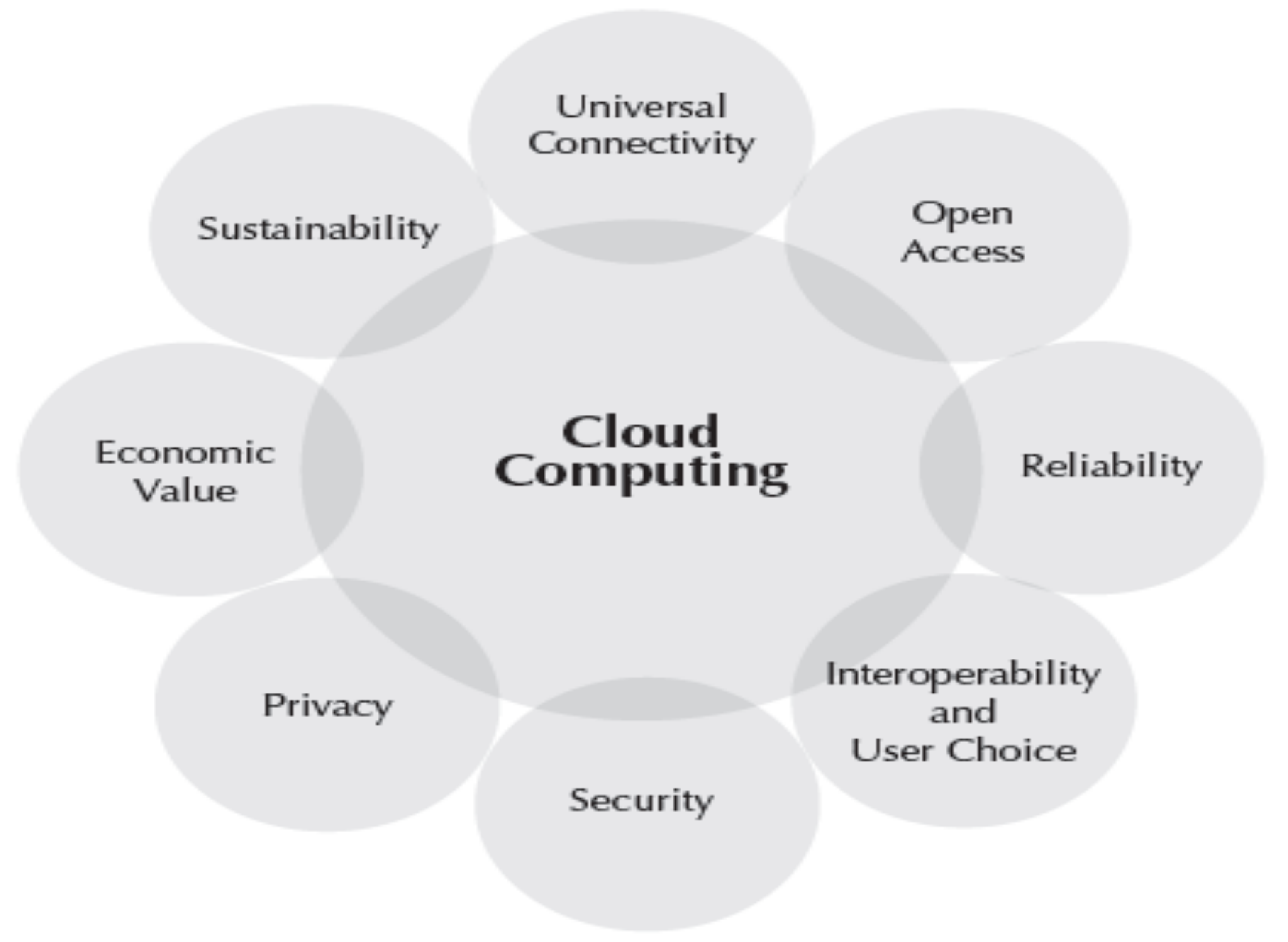
Techniques for Voluminous Data

- Cloud Computing is an efficient method to balance between dealing with voluminous data and keeping costs competitive, is designed to deliver IT services consumable on demand, is scalable as per user need and uses a pay-per-use model.
- Business houses are progressively turn towards retaining core competencies, and shedding the non-core competencies for on-demand technology, business innovation and savings.

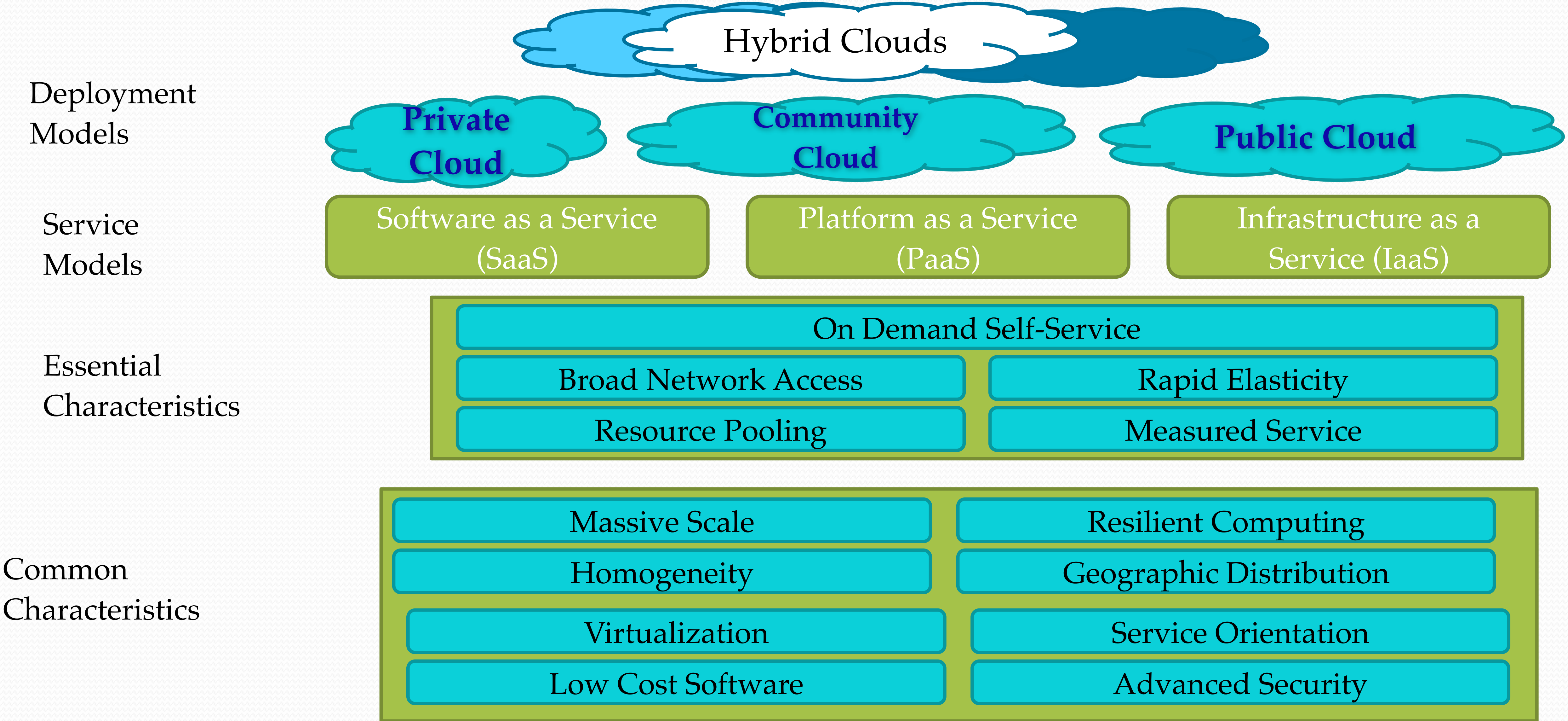
Cloud Computing

- “Cloud computing is a model for enabling **ubiquitous, convenient, on–demand network access** to a **shared pool** of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”
- It provides **high level abstraction** of computation and storage model.
- This cloud model promotes availability and is composed of five essential **characteristics,** three **service models**, and four **deployment models**.

Fundamental Elements of cloud Computing



The NIST (National Institute of Standards and Technology) Cloud Definition Framework



Hybrid Clouds

Deployment Models

- Private Cloud
- Community Cloud
- Public Cloud

Service Models

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

Essential Characteristics

- On Demand Self-Service
- Broad Network Access
- Rapid Elasticity
- Resource Pooling
- Measured Service

Common Characteristics

- Massive Scale
- Resilient Computing
- Homogeneity
- Geographic Distribution
- Virtualization
- Service Orientation
- Low Cost Software
- Advanced Security

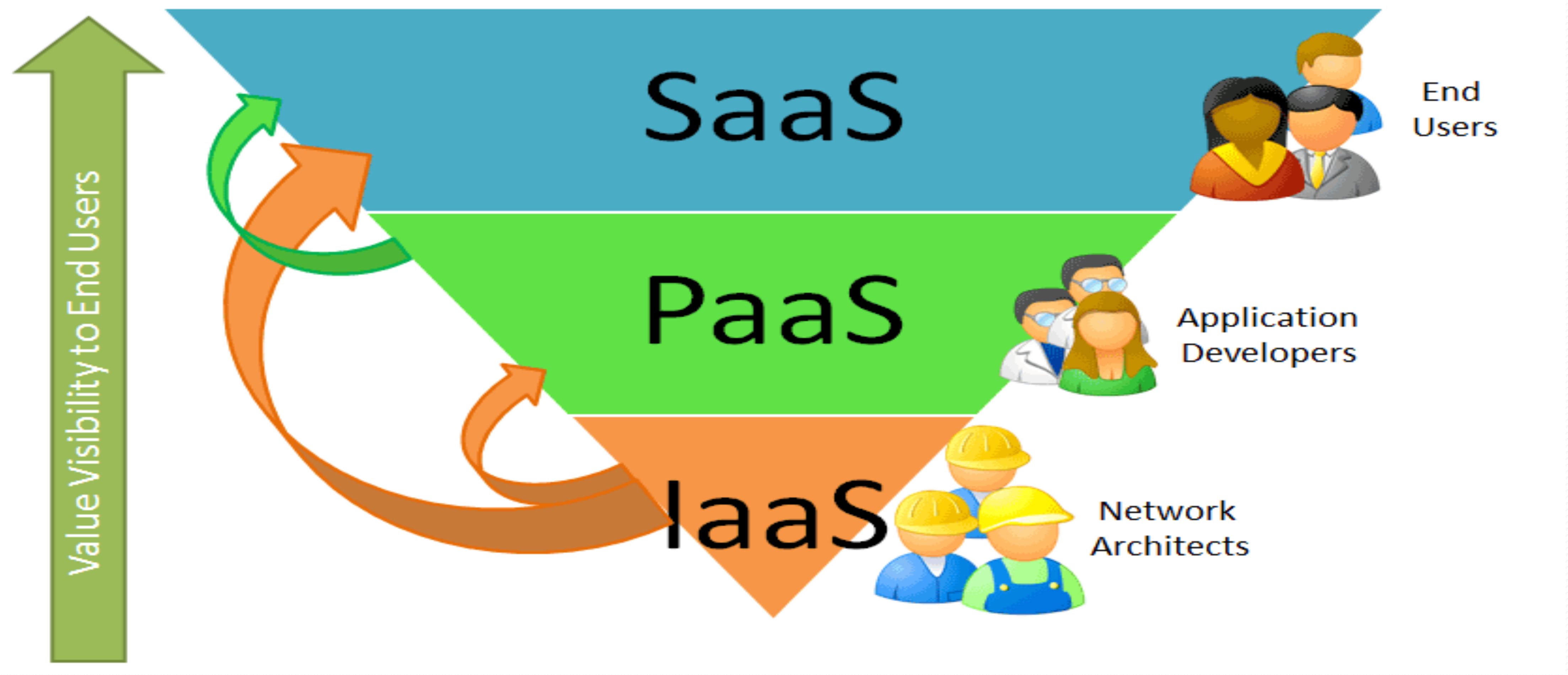
Essential Cloud Characteristics

- On-demand self-service
  - Get computing capabilities as needed automatically
- Broad network access
  - Services available over the net using desktop, laptop, PDA, mobile phone
- Resource pooling
  - Location independence
  - Provide resources pool to serve multiple clients
- Rapid elasticity
  - Ability to quickly scale in/out service
- Measured service
  - control, optimize services based on metering

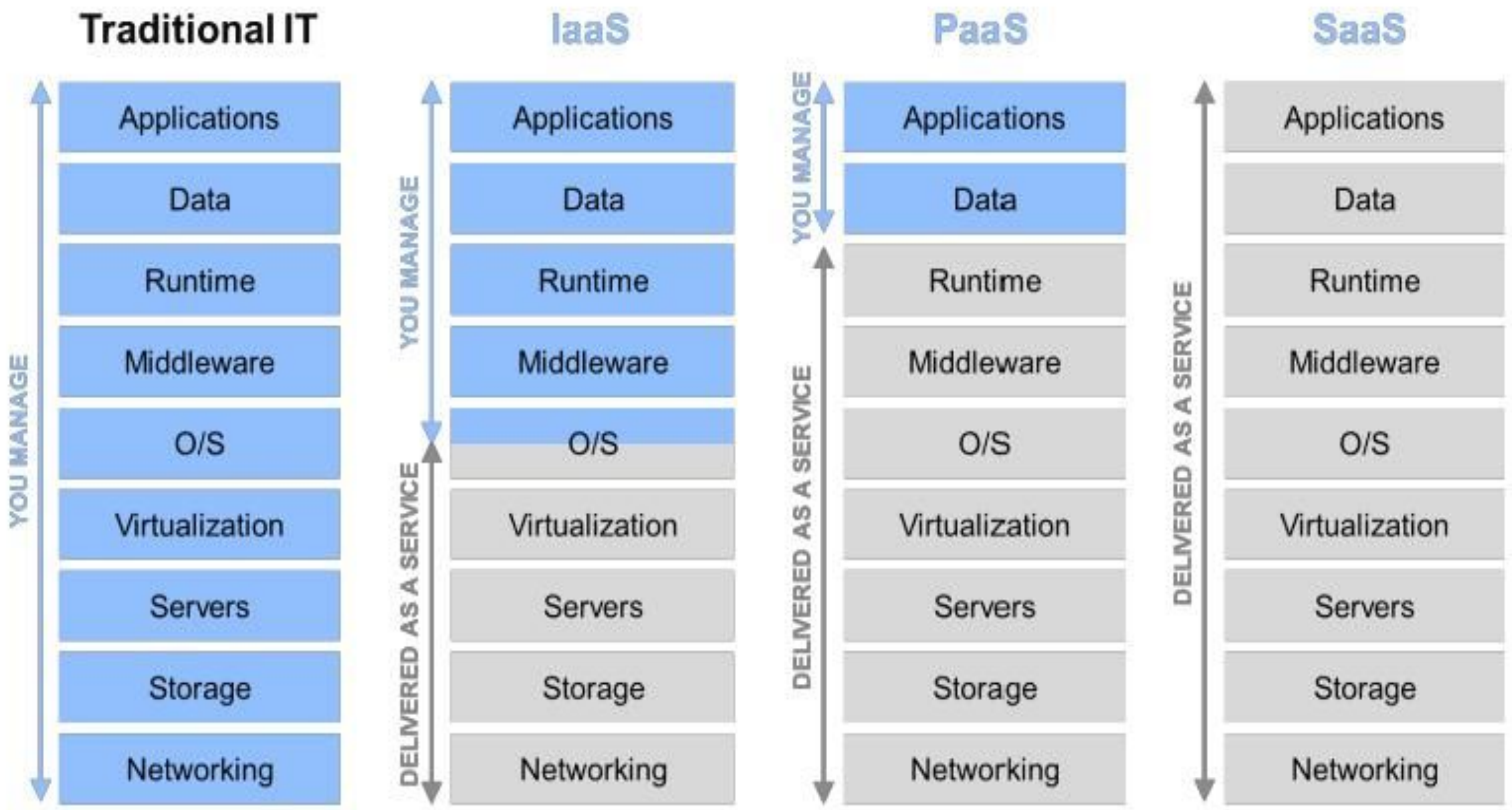
Service Models Overview

- What if you want to have an IT department ?
  - Similar to ***build a new house***
    - You can rent some virtualized infrastructure and build up your own IT system among those resources, which may be fully controlled.
    - Technical speaking, use the ***Infrastructure as a Service (IaaS)*** solution.
  - Similar to ***buy an empty house***
    - You can directly develop your IT system through one cloud platform, and do not care about any lower level resource management.
    - Technical speaking, use the ***Platform as a Service (PaaS)*** solution.
  - Similar to ***live in a hotel***
    - You can directly use some existed IT system solutions, which were provided by some cloud application service provider, without knowing any detail technique about how these service was achieved.
    - Technical speaking, use the ***Software as a Service (SaaS)*** solution.

Cloud Service Models

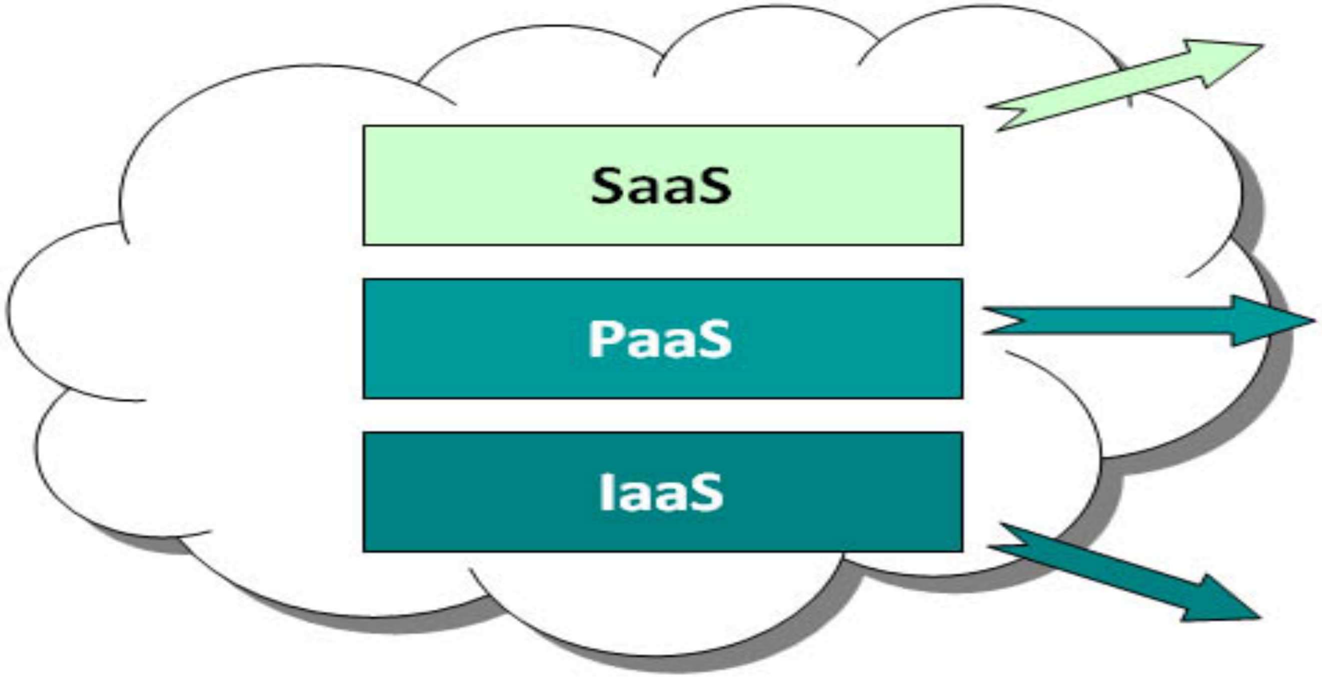


Cloud Service Models



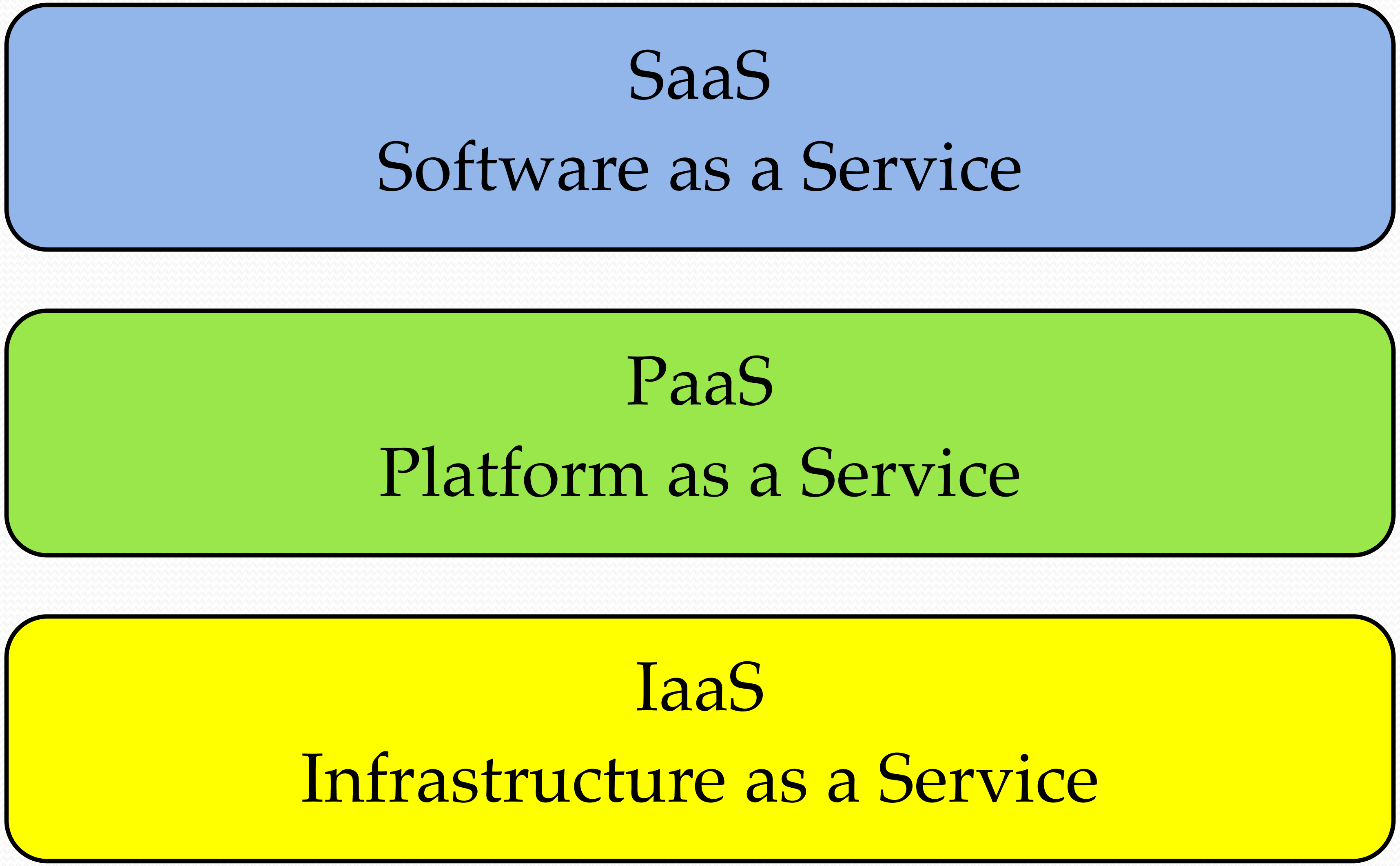
Traditional IT	IaaS	PaaS	SaaS
Applications	Applications	Applications	Applications
Data	Data	Data	Data
Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware
O/S	O/S	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

Source: Microsoft.



Who Uses It	What Services are available	Why use it?
Business Users	EMail, Office Automation, CRM, Website Testing, Wiki, Blog, Virtual Desktop ...	To complete business tasks
Developers and Deployers	Service and application test, development, integration and deployment	Create or deploy applications and services for users
System Managers	Virtual machines, operating systems, message queues, networks, storage, CPU, memory, backup services	Create platforms for service and application test, development, integration and deployment

	Amazon	Google	Salesforce	Customer Implications
Software as Service		Google Apps	salesforce	+ Application logic, platform and infrastructure abstracted + Significant reduction in effort to deploy, run and manage - Apps can be configured but may not meet highly customized requirements
Platform as Service			force.com	+ Platform & infrastructure abstracted + Custom apps can be built order of magnitude more quickly and cheaply - Custom apps still need to be supported and managed
Infrastructure as Service	amazon web services			+ Physical infrastructure abstracted + Can be scaled up and down as needed - Needs to be provisioned/managed - Higher levels of stack still need to be managed, maintained and supported



SaaS
Software as a Service

PaaS
Platform as a Service

IaaS
Infrastructure as a Service

SaaS
Software as a Service

SaaS

Software delivery model

- No hardware or software to manage
- Service delivered through a browser

SaaS

Advantages

- Pay per use
- Instant Scalability
- Security
- Reliability
- APIs (application programming Interface)

SaaS

Examples

- CRM
- Financial Planning
- Human Resources
- Word processing

Commercial Services:

- Salesforce.com
- emailcloud

PaaS
Platform as a Service

Platform delivery model

PaaS

- Platforms are built upon Infrastructure, which is expensive
- Estimating demand is not a science!
- Platform management is not fun!

Popular services

PaaS

- Storage
- Database
- Scalability

Advantages

PaaS

- Pay per use
- Instant Scalability
- Security
- Reliability
- APIs

Examples

PaaS

- Google App Engine
- Mosso
- AWS: S3

IaaS
Infrastructure as a Service

Computer infrastructure delivery model

Access to infrastructure stack:

- Full OS access
- Firewalls
- Routers
- Load balancing

IaaS

Advantages

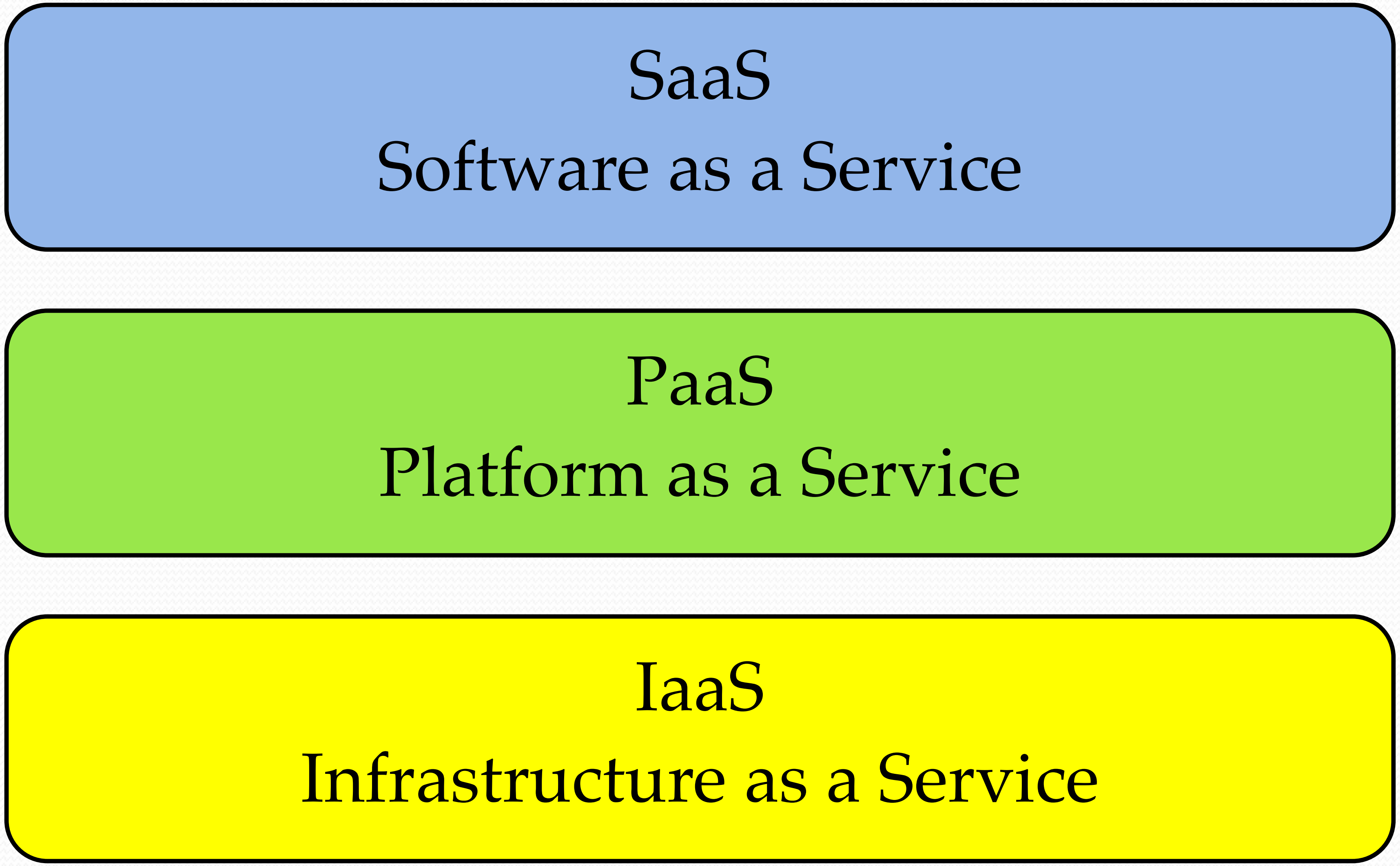
- Pay per use
- Instant Scalability
- Security
- Reliability
- APIs

IaaS

Examples

- Flexiscale
- AWS: EC2

IaaS



SaaS

Software as a Service

PaaS

Platform as a Service

IaaS

Infrastructure as a Service

Common Factors

SaaS

PaaS

IaaS

- Pay per use
- Instant Scalability
- Security
- Reliability
- APIs

SaaS

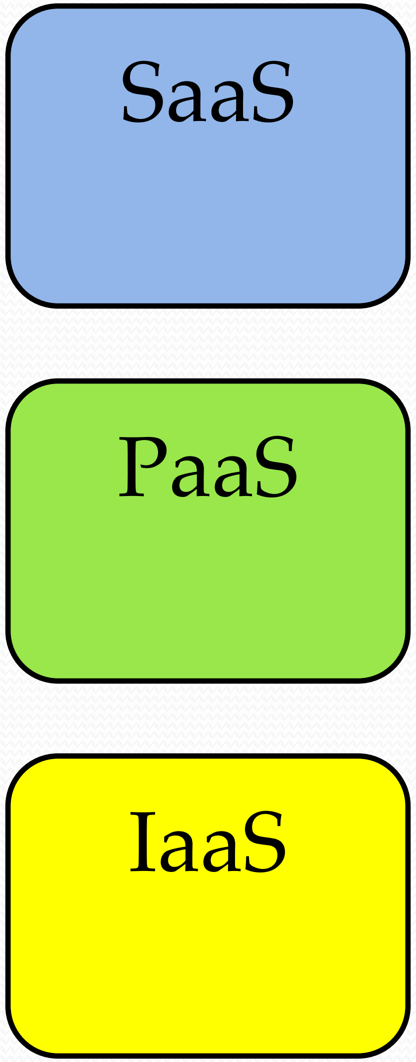
PaaS

IaaS

Advantages

- Lower cost of ownership
- Reduce infrastructure management responsibility
- Allow for unexpected resource loads
- Faster application rollout

Cloud Economics


SaaS

PaaS

IaaS

- Virtualisation lowers costs by increasing utilisation
- Economies of scale afforded by technology
- Automated update policy

SaaS

PaaS

IaaS

Risks

- Security
- Downtime
- Access
- Dependency
- Interoperability

Benefit of Cloud Computing on IS

- Reduced Cost : Cloud technology is paid incrementally, saving organizations money.
- Increased Storage: Organizations can store more data than on private computer systems.
- Highly Automated: No longer do IT personnel need to worry about keeping software up to date.
- Flexibility: Cloud computing offers much more flexibility than past computing methods.
- More Mobility: Employees can access information wherever they are, rather than having to remain at their desks.
- Allows IT to Shift Focus: No longer having to worry about constant server updates and other computing issues, government organizations will be free to concentrate on innovation.

MapReduce

- MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
- It was developed at Google for indexing Web pages and replaced their original indexing algorithms and heuristics in 2004.

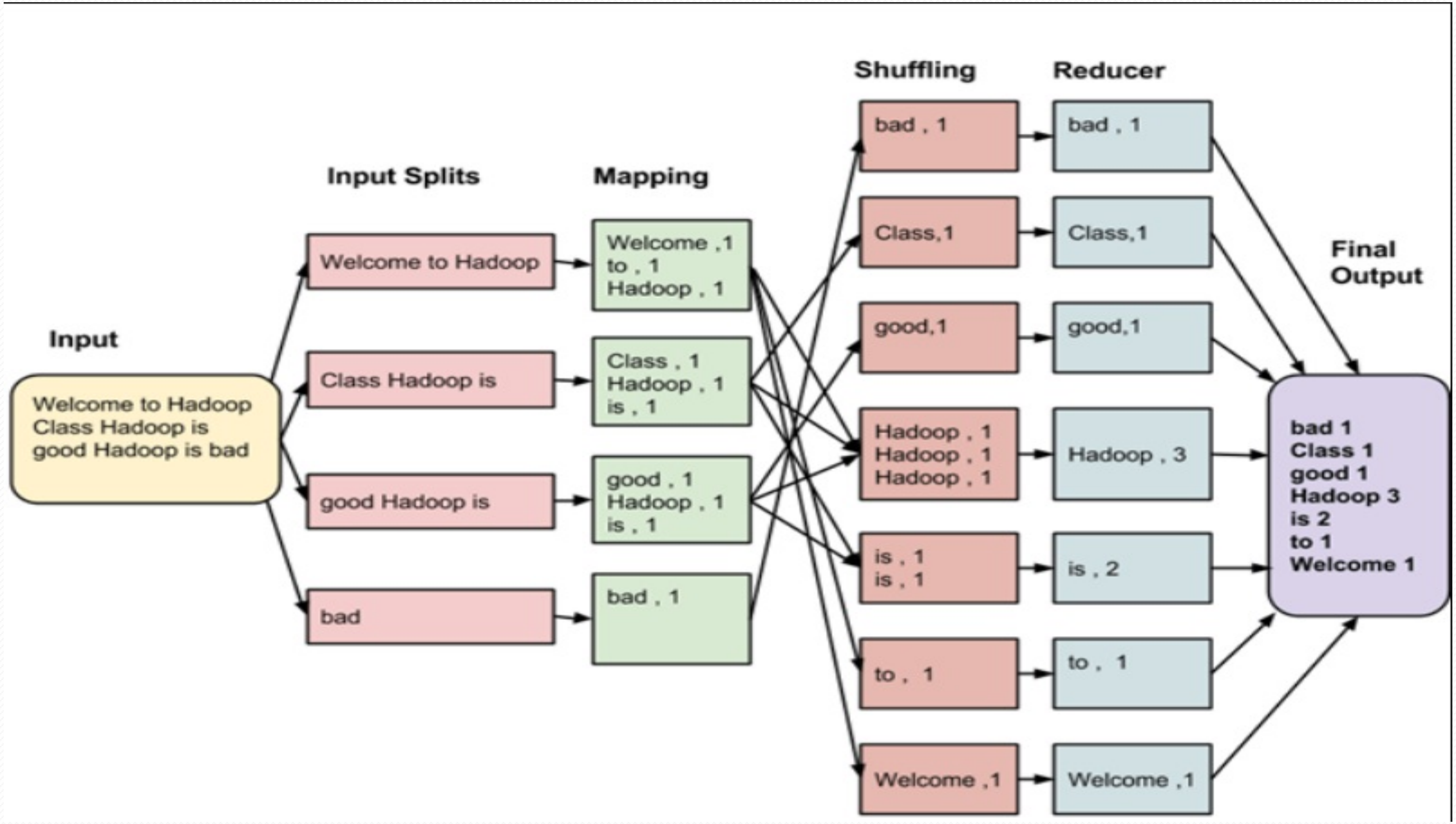
MapReduce and Hadoop Systems

- MapReduce is the heart of Hadoop.
- MapReduce allows data to be distributed across a large cluster, and can distribute out tasks across the data set to work on pieces of it independently, and in parallel.
- This allows big data to be processed in relatively little time.
- Apache has produced an open source MapReduce platform called **Hadoop**.

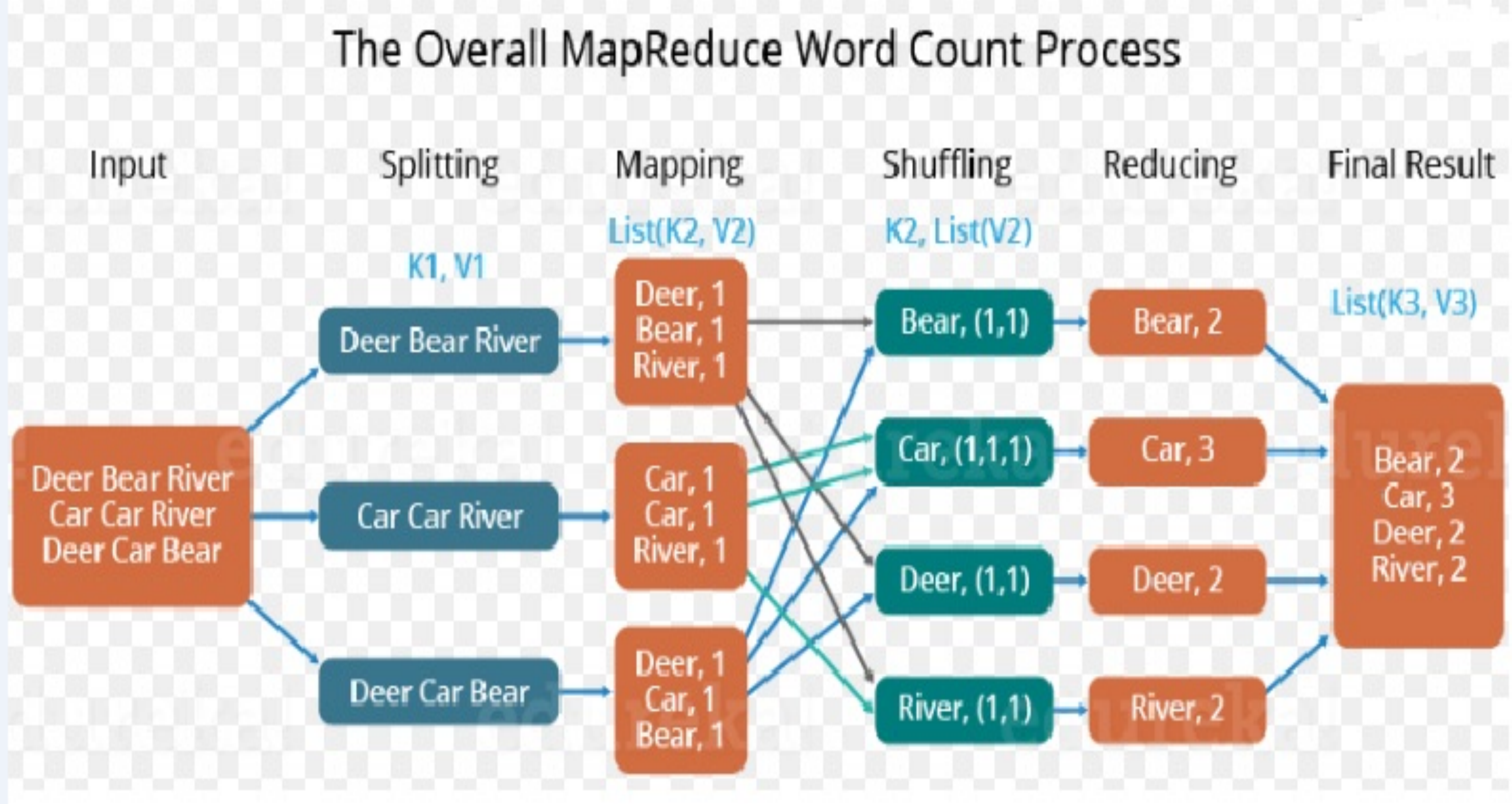
Framework is divided into two parts

- **Map**, a function that parcels out work to different nodes in the distributed cluster.
- **Reduce**, another function that collates the work and resolves the results into a single value.
- The MapReduce framework is fault-tolerant because each node in the cluster is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

Example

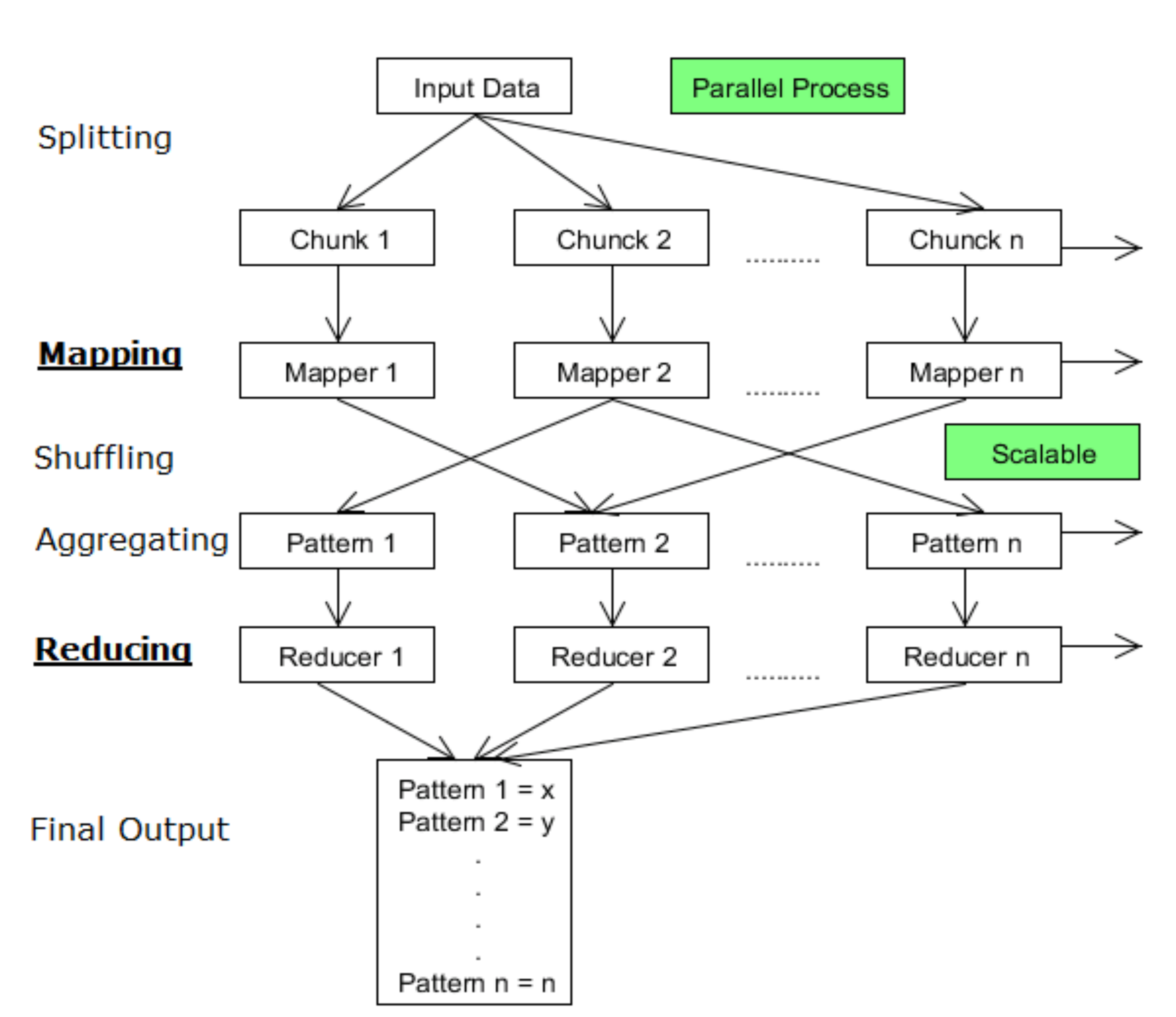

Input
Welcome to Hadoop
Class Hadoop is
good Hadoop is bad
Input Splits
Welcome to Hadoop
Class Hadoop is
good Hadoop is
bad
Mapping
Welcome ,1
to , 1
Hadoop , 1
Class , 1
Hadoop , 1
is , 1
good , 1
Hadoop , 1
is , 1
bad , 1
Shuffling
bad , 1
Class,1
good,1
Hadoop , 1
Hadoop , 1
Hadoop , 1
is , 1
is , 1
to , 1
Welcome ,1
Reducer
bad , 1
Class,1
good,1
Hadoop , 3
is , 2
to , 1
Welcome ,1
Final Output
bad 1
Class 1
good 1
Hadoop 3
is 2
to 1
Welcome 1

Example



The Overall MapReduce Word Count Process
Input
Splitting
Mapping
Shuffling
Reducing
Final Result
K1, V1
List(K2, V2)
K2, List(V2)
List(K3, V3)
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

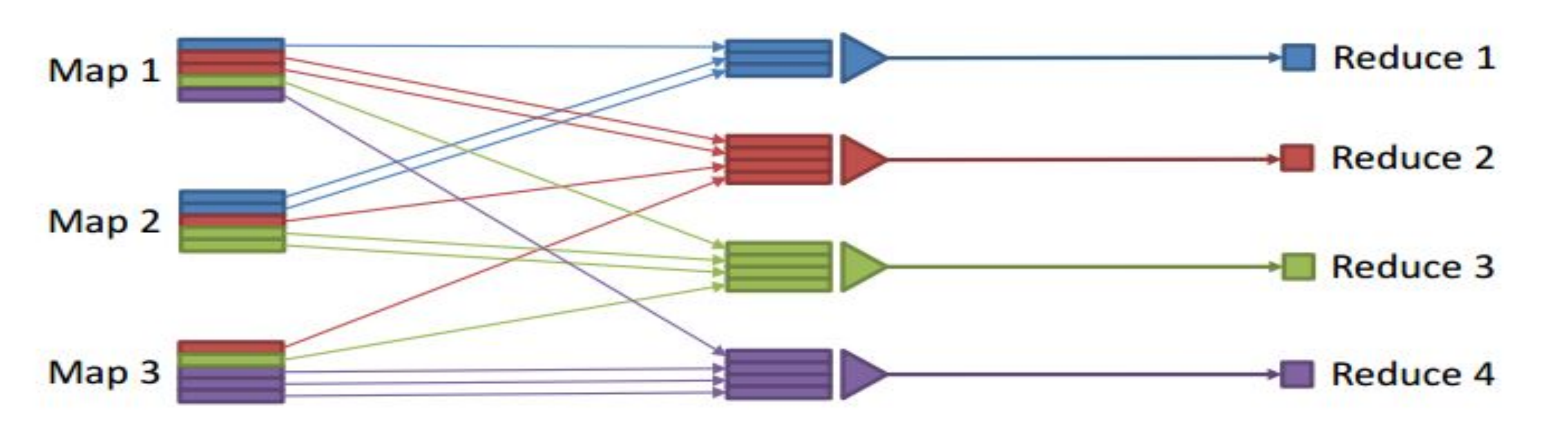
Flow of Map Reduce Algorithm



Flow of Map Reduce Algorithm

- The input data can be divided into n number of chunks depending upon the amount of data and processing capacity of individual unit.
- Next, it is passed to the mapper functions. Please note that all the chunks are processed simultaneously at the same time, which embraces the parallel processing of data.
- After that, shuffling happens which leads to aggregation of similar patterns.
- Finally, reducers combine them all to get a consolidated output as per the logic.
- This algorithm embraces scalability as depending on the size of the input data, we can keep increasing the number of the parallel processing units.

MapReduce Interaction



Map functions create a user-defined “index” from source data

- Reduce functions compute grouped aggregates based on index
- Flexible framework
  - •users can cast raw original data in any model that they need
  - •wide range of tasks can be expressed in this simple framework

Hadoop System

- Developed by Apache as an open source distributed MapReduce platform, based off of Google's MapReduce.
- Runs on a Java architecture framework that supports the processing of large data sets in a distributed computing environment.
- Hadoop allows businesses to process large amounts of data quickly by distributing the work across several nodes.
- Good for Big data sets and on large cluster.

Hadoop - A Key Business Tool

Hadoop System is used by Large Content-Distribution Companies, such as…

Yahoo

- Hadoop is used for many of their tasks, and over 25,000 computers are running Hadoop.

Amazon

- Hadoop is good for Amazon, they have lots of product data, as well as user-generated content to index, and make searchable.

New York Times

- Hadoop is used to perform large-scale image conversions of public domain articles.

Hadoop - A Key Business Tool

Used by Non-Content-Distribution Companies, such as

- Facebook
- eHarmony

Other early adopters include anyone with big data:

- medical records
- tax records
- network traffic
- large quantities of data

Wherever there is a lot of data, a Hadoop cluster can generally process it relatively quickly.

Data Management in the Cloud

- Cloud data management is emerging as an alternative to data management using traditional on-premises software.
- Instead of buying on-premise storage resources and managing them, resources are bought on-demand in the cloud.
- **This service model allows organizations to receive dedicated cloud data management resources on an as-needed basis.**
- It can provide increased flexibility to meet changing business requirements. But challenge of: data security concerns .

Data Management in the Cloud

- Goals:
  - Availability
  - Scalability
  - Elasticity
  - Performance
  - Load balancing
  - Fault tolerance
  - Ability to run in a heterogeneous environment
  - Flexible query interface

- Challenges:
  - Availability of a Service
  - Data Confidentiality
  - Data lock-in
  - Data transfer bottlenecks
  - Application parallelization
  - Performance unpredictability
  - Application debugging in large-scale distributed systems

Data Management in the Cloud

- Data management applications are potential candidates for deployment in the cloud
  - **Industry:** enterprise database system have significant up-front cost that includes both hardware and software costs
  - **Academia:** manage, process and share mass-produced data in the cloud
- Two largest components of data management market:
  - Transactional Data Management --OLTP
  - Analytical Data Management -- OLAP

Transactional Data Management

- Banks, airline reservation, online e-commerce
- Not ready to move to the cloud for the following reasons:
  - Don’t use shared-nothing architecture
  - Hard to maintain ACID when data replication are all over the world
  - Enormous risks in storing transactional data on an untrusted host

Analytical Data Management

- Business planning, decision support
- Well-suited to run in a cloud environment:
  - Shared-nothing architecture is a good match
  - ACID guarantees are typically not needed
  - Particularly sensitive data can be left out of the cloud.

Information Retrieval

- **Information retrieval** is the activity of obtaining information resources relevant to an information need from a collection of information resources.
- Searches can be based on metadata or on full-text indexing.
- Automated information retrieval systems are used to reduce what has been called “information overload”.

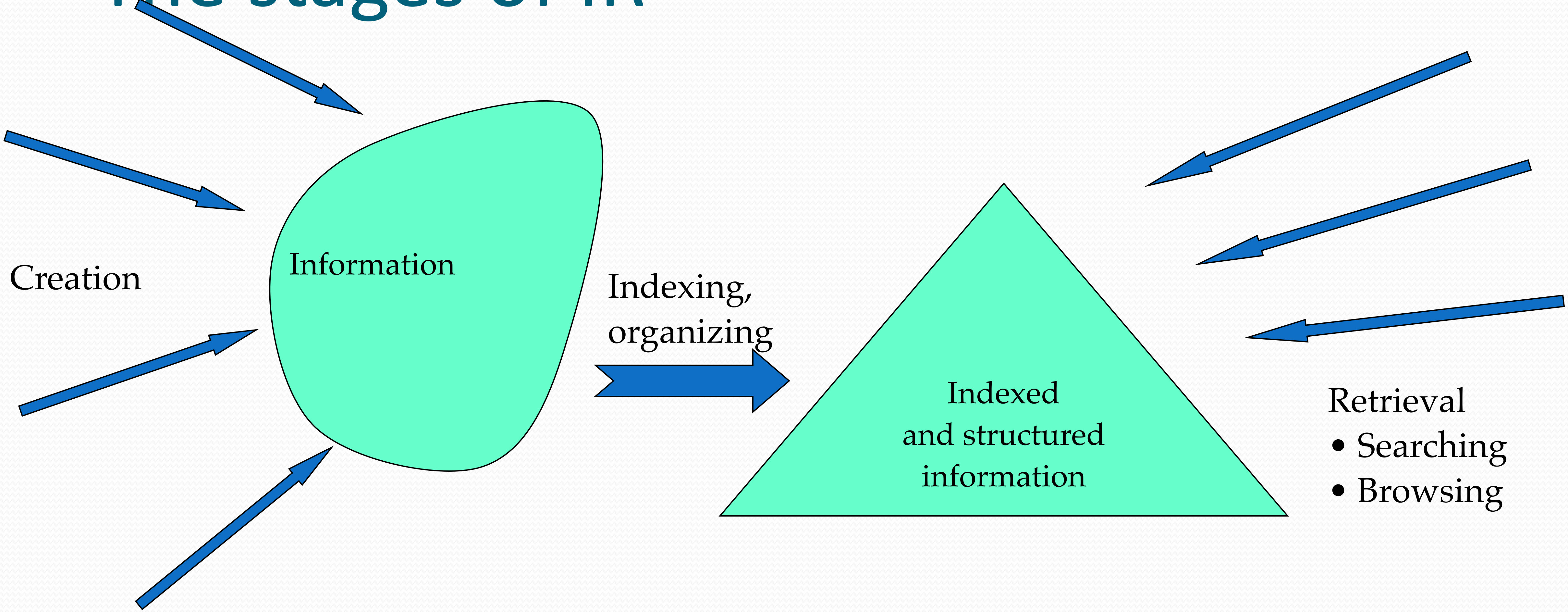
Information Retrieval in the Cloud

- IR user seeks actively information, pulling at it, by means of querying or browsing.
- In tag querying, user enters one or more tags in the search box to obtain an ordered list of resources which were in relation with these tags.
- When a user is scanning this list, the system also provide a list of related tags (i.e. tags with a high degree of co-occurrence with the original tag), allowing hypertext Browsing.

Information Retrieval System

  - Typically it refers to the automatic (rather than manual) retrieval of documents
    - Information Retrieval System (IRS)
- Information Retrieval is a research-driven theoretical and experimental discipline
  - The focus is on different aspects of the information–seeking process, depending on the researcher’s background or interest:
    - Computer scientist – fast and accurate search engine
    - Librarian – organization and indexing of information
    - Cognitive scientist – the process in the searcher’s mind
    - Philosopher – Is this really relevant ?

The stages of IR



Creation

Information

Indexing, organizing

Indexed and structured information

Retrieval
- Searching
- Browsing

Link Analysis in Cloud Setup

- The web is not just a collection of documents – its hyperlinks are important!
- A link from page *A* to page *B* may indicate:
  - *A* is related to *B*, or
  - *A* is recommending, citing, voting for or endorsing *B*
- Links are either
  - referential – *click here and get back home*, or
  - Informational – *click here to get more detail*
- Links effect the ranking of web pages and thus have commercial value.

Case studies of voluminous data environment

Case Studies Big data

Thank you

Good Luck

For your Examination

and

Bright Future Ahead