Scalable and Emerging Information System Techniques

Chapter-8

Syllabus

- Scalable and Emerging information System techniques (8 hours)
  - Techniques for voluminous data
  - Cloud computing technologies and their types
  - MapReduce and Hadoop systems
  - Data management in the cloud
  - Information retrieval in the cloud
  - Link analysis in cloud setup
  - Case studies of voluminous data environment

Techniques for Voluminous Data

- Big Data
- Cloud Computing
- Data Mining

Big Data

- Big Data applies to information that can’t be processed or analyzed using traditional processes or tools.
- Big Data is also **data** but with a **huge size**.
- Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time.
- In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Example

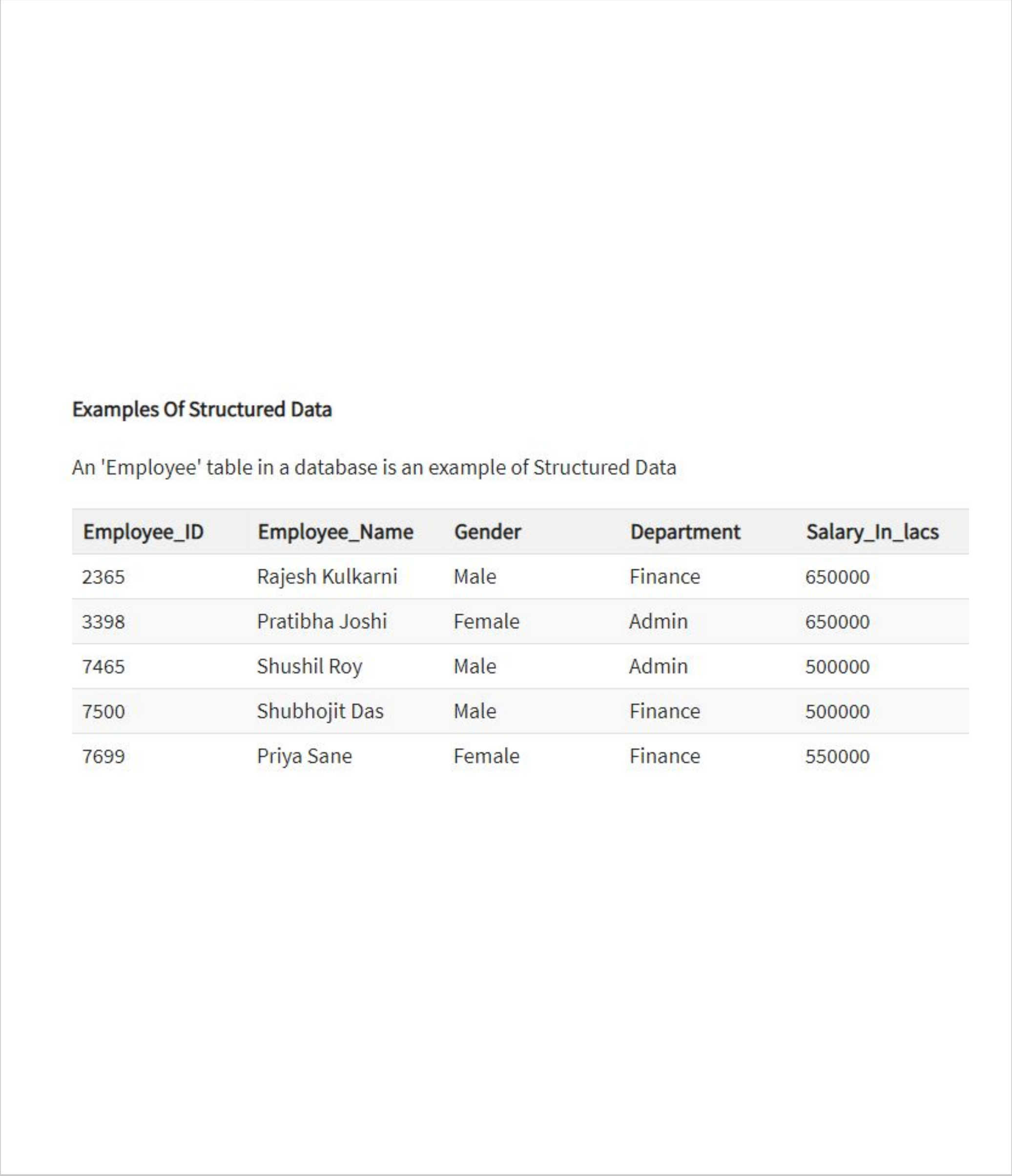
- **Social Media**
- The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
- A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Types of Big Data

- **Structured**
- **Unstructured**
- **Semi-structured**

Structured

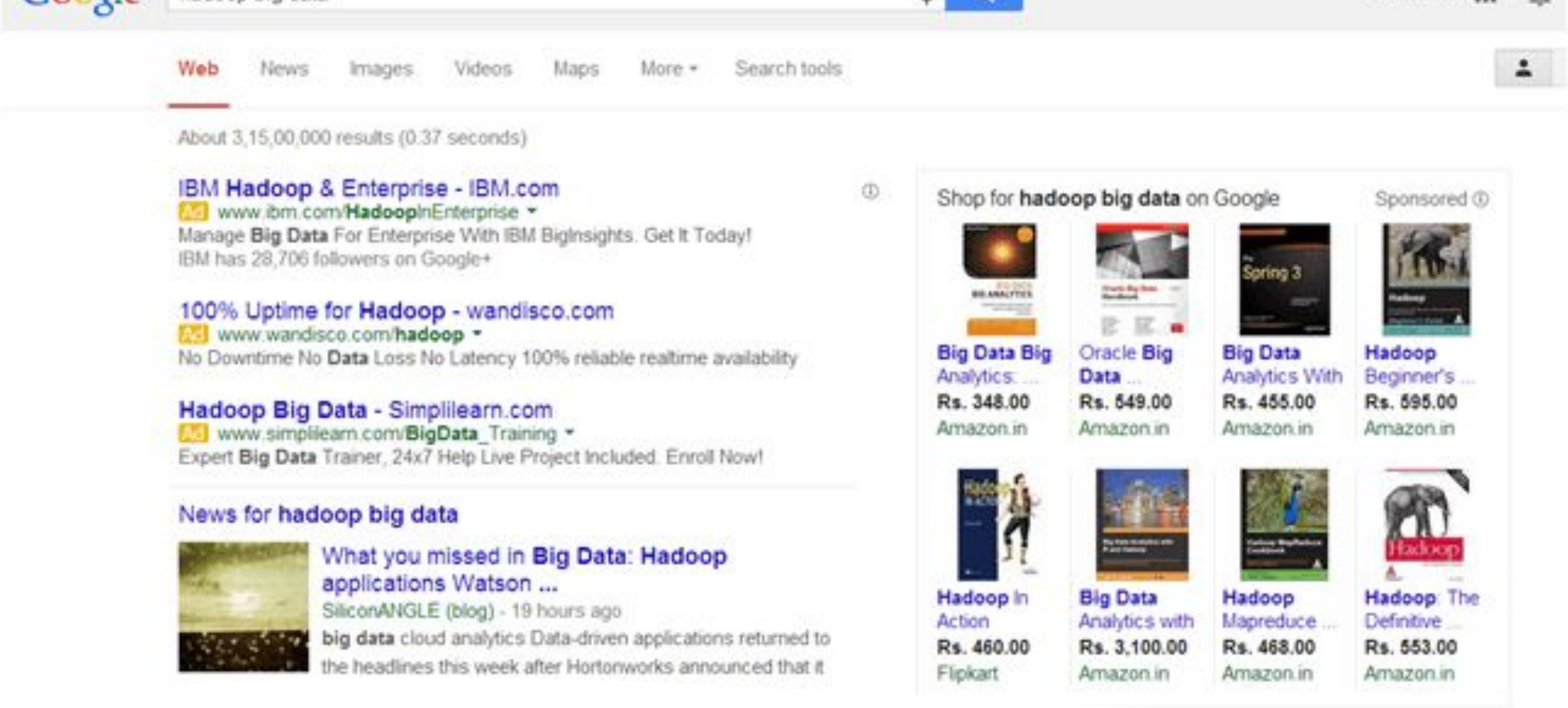
- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.



Examples Of Structured Data

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000



About 3,15,00,000 results (0.37 seconds)

IBM **Hadoop** & Enterprise - IBM.com
Ad www.ibm.com/**Hadoop**InEnterprise
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for **Hadoop** - wandisco.com
Ad www.wandisco.com/**hadoop**
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/**BigData**_Training
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for **hadoop big data**

What you missed in **Big Data**: **Hadoop** applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it

Shop for **hadoop big data** on Google

Sponsored

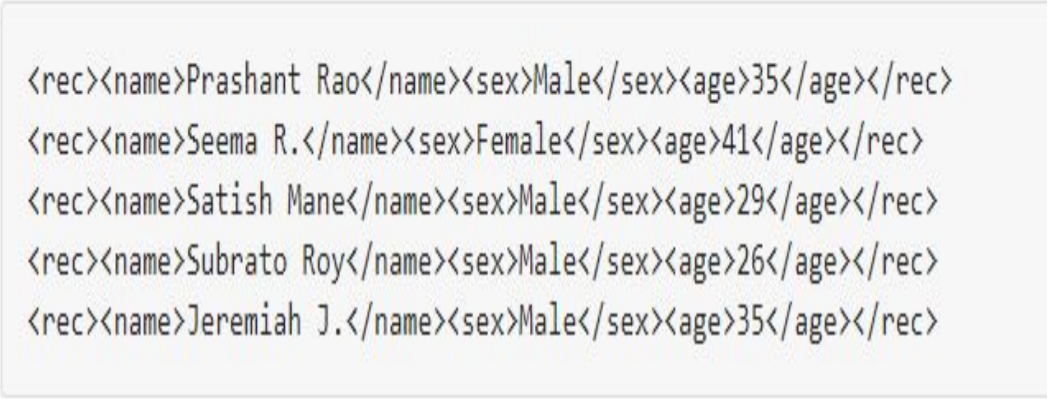
Big Data Big Analytics: ...	Oracle **Big Data** ...	**Big Data** Analytics With	**Hadoop** Beginner's ...
Rs. 348.00	**Rs. 549.00**	**Rs. 455.00**	**Rs. 595.00**
Amazon.in	Amazon.in	Amazon.in	Amazon.in
Hadoop In Action	**Big Data** Analytics with	**Hadoop** Mapreduce ...	**Hadoop**: The Definitive ...
Rs. 460.00	**Rs. 3,100.00**	**Rs. 468.00**	**Rs. 553.00**
Flipkart	Amazon.in	Amazon.in	Amazon.in

Unstructured

- Any data with unknown form or the structure is classified as unstructured data.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

Semi structured

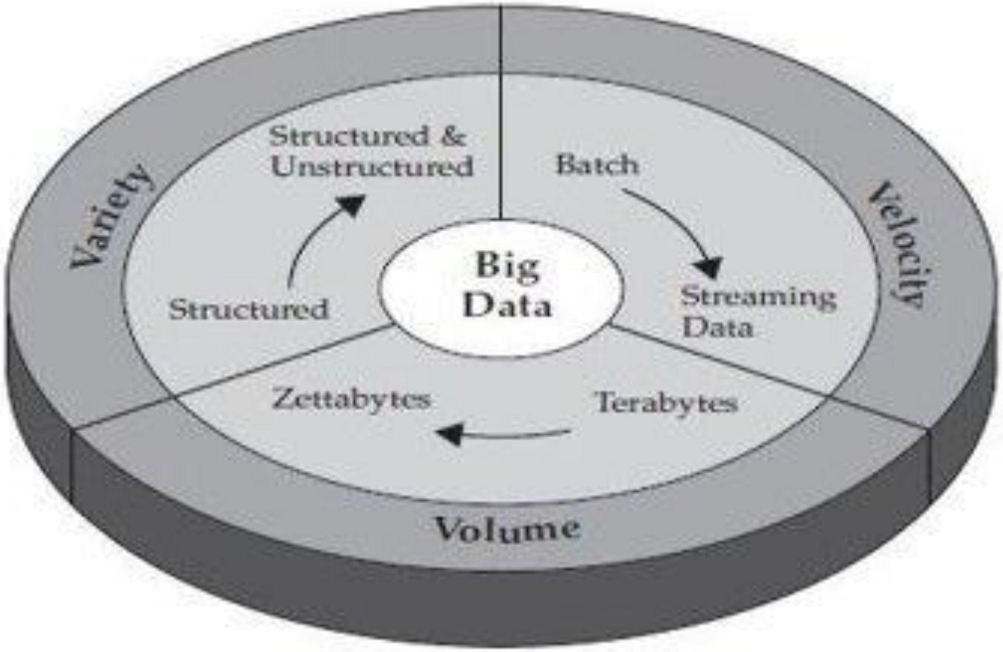
- Semi-structured data can contain both the forms of data.
- We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
- Example of semi-structured data is a data represented in an XML file.



```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Characteristics Of Big Data

- ***(i) Volume –*** The name Big Data itself is related to a size which is enormous.
- ***(ii) Variety –***
- Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- ***(iii) Velocity –*** The term **'velocity'** refers to the speed of generation of data.
- ***(iv) Variability –*** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.



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

Benefits

- Businesses can utilize outside intelligence while taking decisions
- Improved customer service
- Early identification of risk to the product/services, if any

Data Mining

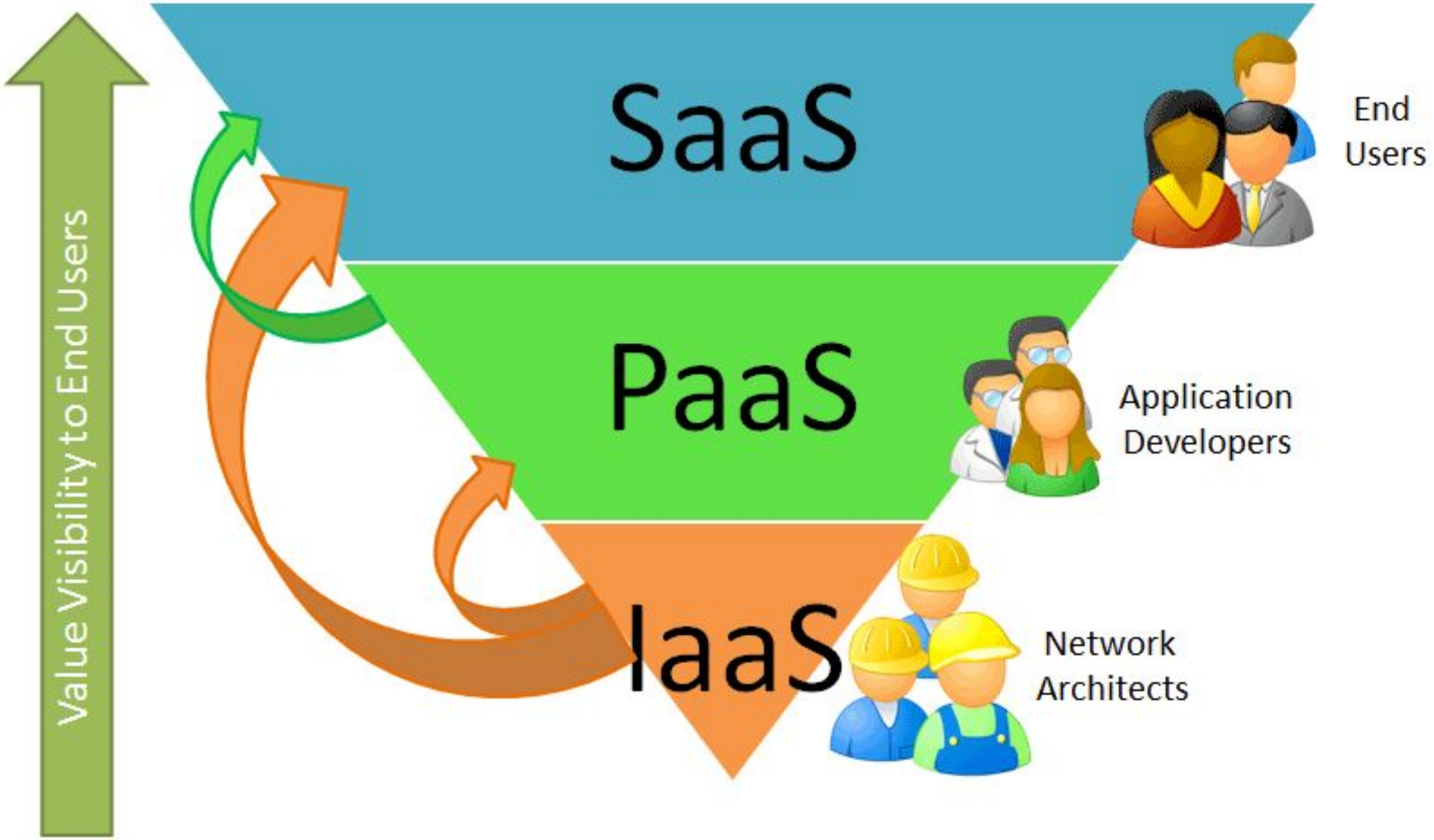
- Studied in chapter 4

Cloud Computing

Cloud computing consists of hardware and software resources made available on the internet as managed third-party services.

These services typically provide access to advanced software applications and high-end networks of server computers.

Cloud Computing Layers



Types of Cloud Computing

SaaS (Software as a Service):

This is the idea of providing a given application to multiple tenants, typically using the browser which supports business applications of host and delivery type as a service. End Customers for instance Google Doc, Myspace.com

Common features of SaaS:

a) User applications run on cloud infrastructure

b) Accessible by users through web browser

c) Suitable for CRM (Customer Resource Management) applications

d) Supports multi-tenant environment

PaaS (Platform as a Service): This is a variant of SaaS. You run your own applications but you do it on the cloud provider’s infrastructure. Provides a comprehensive stack for developers to create Cloud-ready business applications.

Developers for instance Force.com, Google App Engine, Azure and Salesforce.com etc

Features of PaaS are:

a) Supports web-service standards

b) Dynamically scalable as per demand

c) Supports multi-tenant environment

IaaS (Infrastructure as a Service): These are virtual storage and server options that organizations can access on demand, even allowing the creation of a virtual data center. Delivers computing hardware like Servers, Network, Storage, etc. For instance Rackspace.com, GoGrid.com etc.

Typical features are:

a)Users use resources but have no control of underlying cloud infrastructure

b)Users pay for what they use

c)Flexible scalable infrastructure without extensive pre-planning

Benefit of Cloud Computing

- 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.

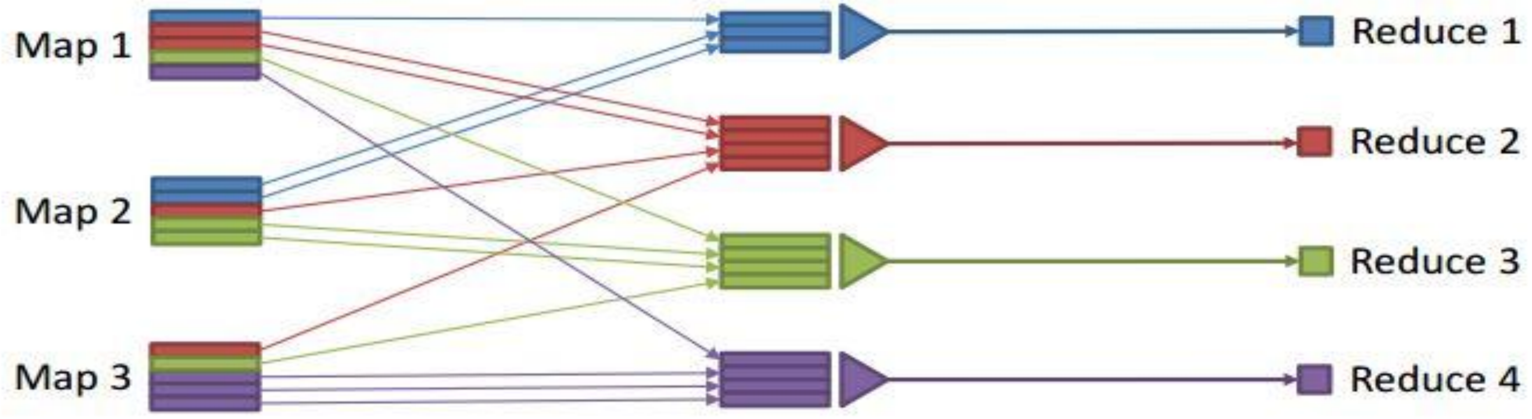
MapReduce and Hadoop Systems

- 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.

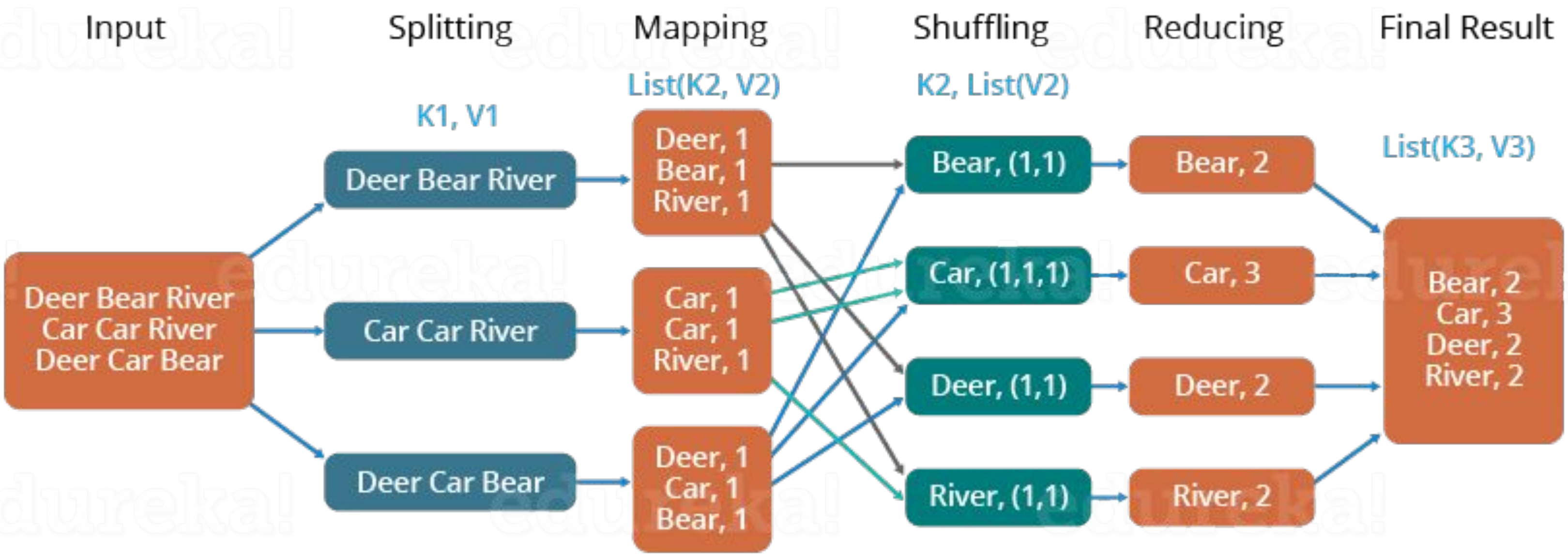
MapReduce Interaction



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

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

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

- 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

- Many “Cloud Killer Apps” are in fact data-intensive

– Batch Processing as with MapReduce

– Online Transaction Processing (OLTP) as in automated business applications

– Offline Analytical Processing (OLAP) as in data mining or machine learning

Data Management in the cloud

- A database system must implement for it to run well in the cloud, in potential database applications to consider for cloud deployment.
- Data management applications are best suited for deployment on top of cloud computing infrastructure.
- Data management is the proper management of a data resource for an organization.
- Data management consists of a set of theories, concepts, principles, and techniques for properly managing data.
- The primary objective is to support the business information as needs of the organization.

Data Management in Cloud

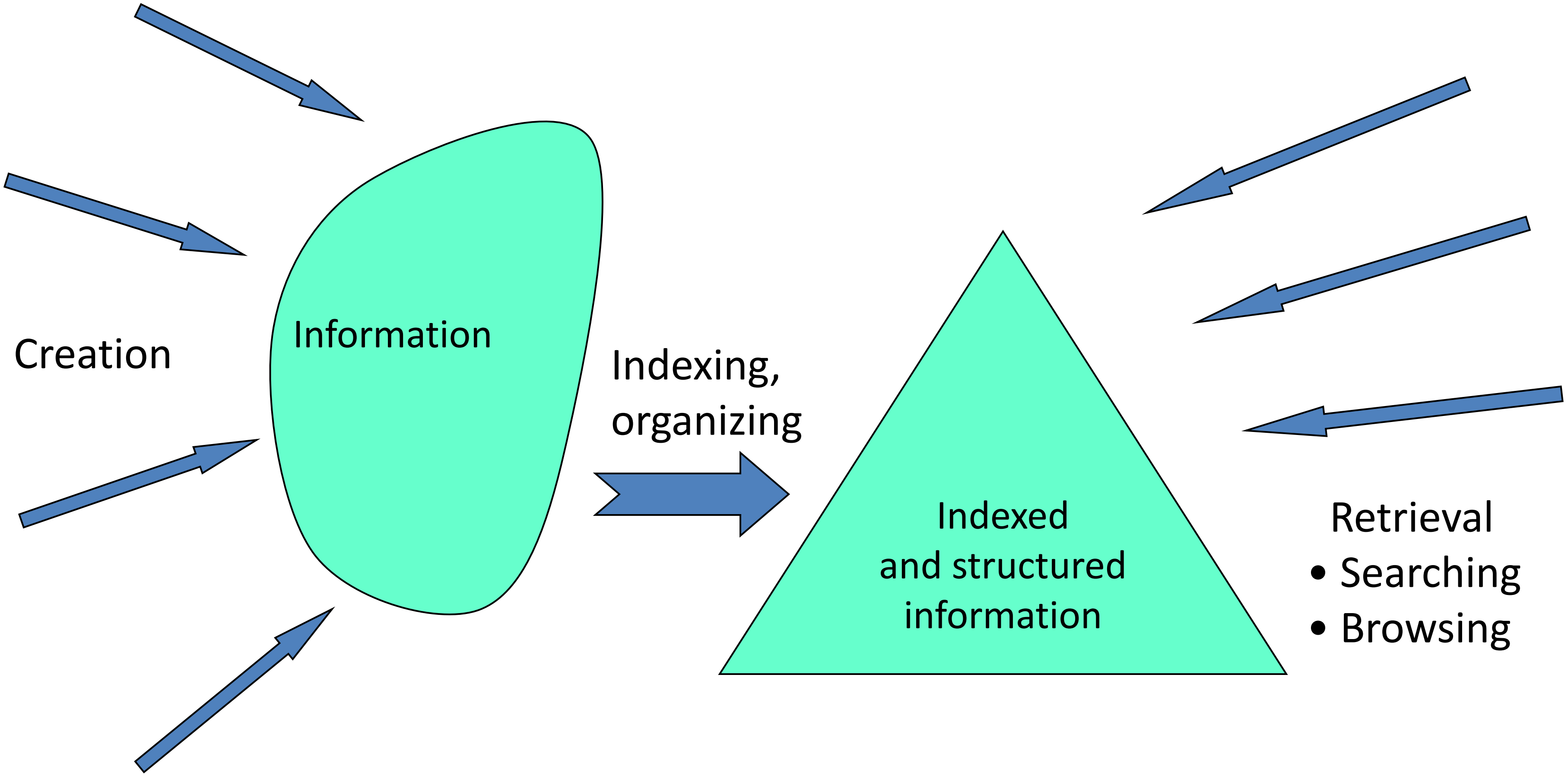
There are three characteristics of a cloud computing environment.

- Compute power is elastic, but only if workload is parallelizable.
- Data is stored at unstructured host.
- Data is replicated, often across large geographic distances

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”.

The stages of IR


Creation

Information

Indexing, organizing

Indexed and structured information

Retrieval

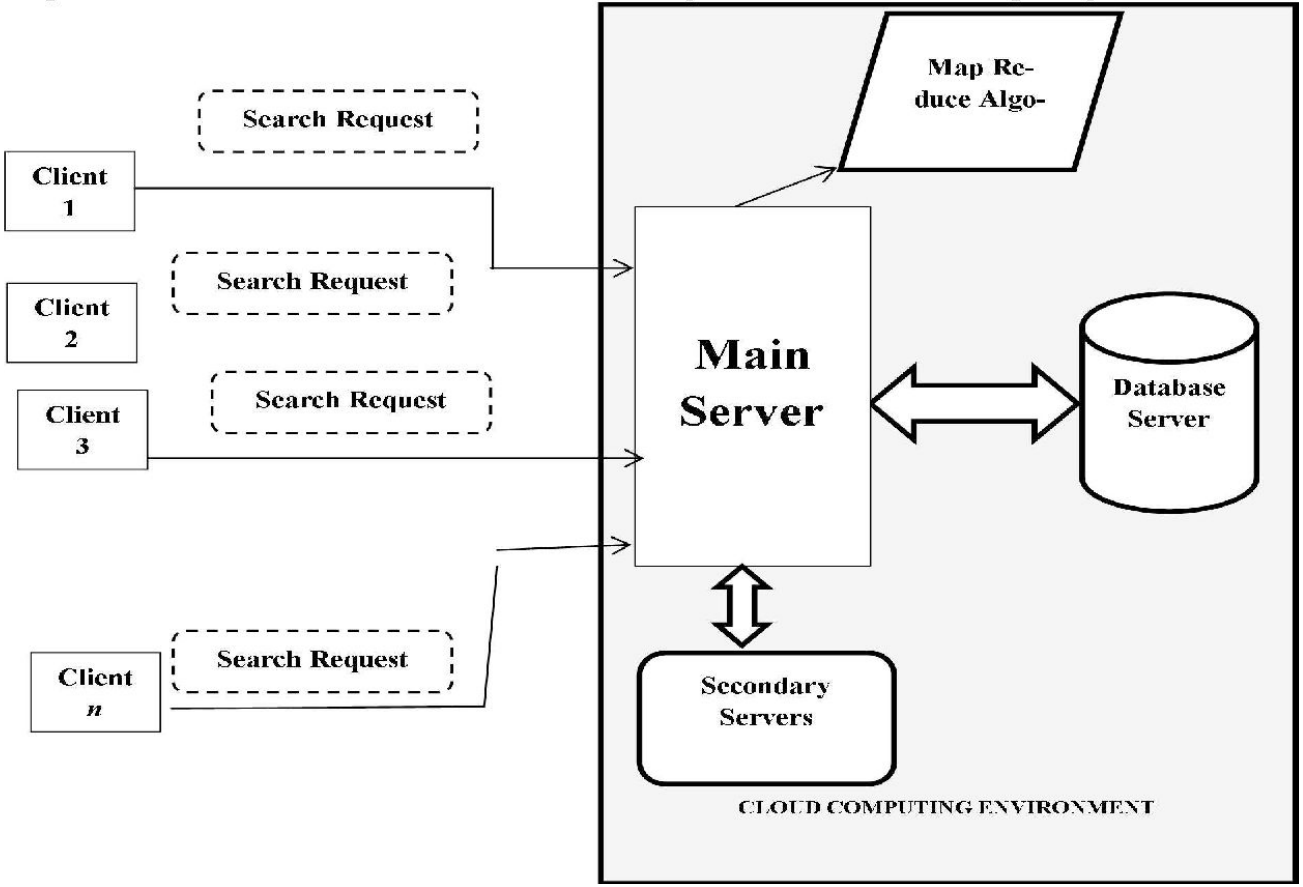
- Searching
- Browsing

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.

FIGURE 2

Implementation of Information Retrieval (IR) Algorithm in a Cloud computing Environment

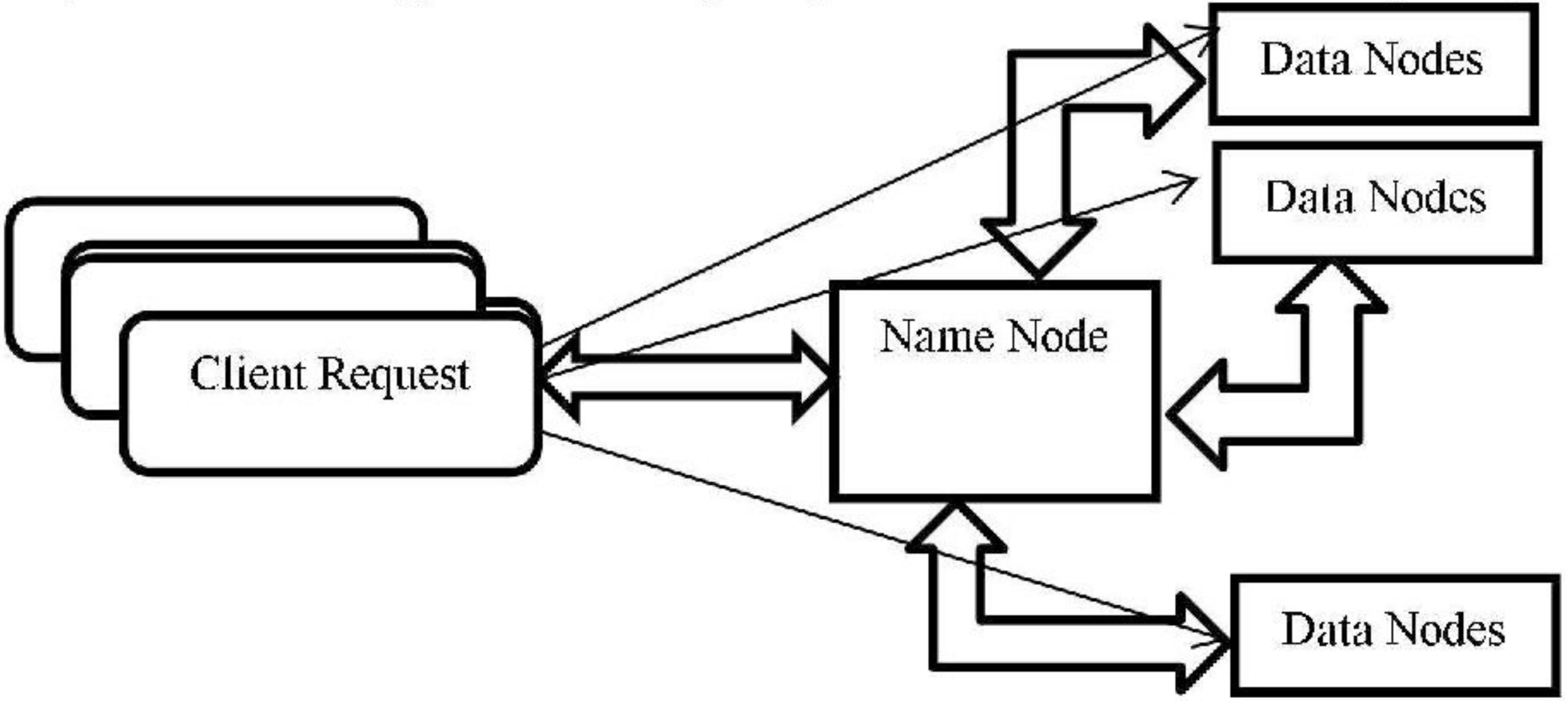


- 1. Main Server
- 2. Secondary Server

- 3. Database Serve The cloud architecture has both master nodes and slave nodes.
- In this enactment, a main server is one that gets client requests and handles them.
- The master node is present in main server and the slave nodes in secondary server.
- Search requests are forwarded to the MapReduce algorithm present in main server.
- MapReduce takes care of the searching and indexing procedure by instigating a large number of Map and Reduce processes.
- Once the MapReduce process for a particular search key is completed, it returns the output value to the main server and in turn to the client.

- The information required by the client is send directly to the Main Server.
- For simplicity, the Main server is termed as Name node and stores the Meta data about the information.
- The Meta data includes the size of the file, exact location of the file, block locations amongst others.
- Each of the information (file) is replicated in number of Secondary Servers, named as Data nodes.
- Data nodes are actually responsible to track the data from the data centers.

l Steps of the IR Algorithm using MapReduce in a Cloud Computing En



- The complete functionality of the MapReduce algorithm operates as follows:
- 1. The client requests arrive at the Main Node.
- 2. The Main node has the MapReduce algorithm in place and does the task of mapping. In nutshell, Name node keeps trajectory of complete file directory structure and the placement of chunks.
- Thus Name node is the essential control point for the complete system. To read a file, the client API will calculate the chunk index based on the offset of the file pointer and make a request to the Name node.
- The Name node will reply which Data nodes has a copy of that chunk. From this point, the client contacts the Data node directly without going through the Name node.

3. The client pushes its changes to all Data nodes, and the change is stored in a buffer of each Data node. After changes are buffered at all Data nodes, the client send a “commit” request, and client gets the response about the success.

Click to add text

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.

- See More on Chapter 7 for Link Analysis

Thank you