

Data Analytics (IT-3006)

**Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024**

School of Computer Engineering



Strictly for internal circulation (within KIIT) and reference only. Not for outside circulation without permission

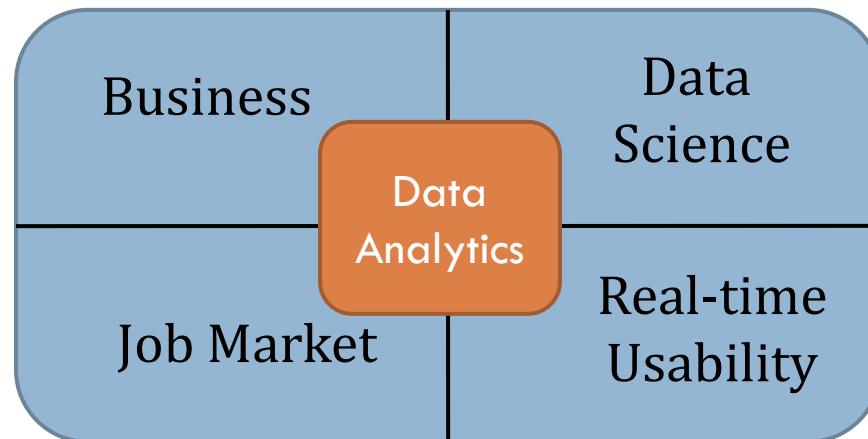
3 Credit

Lecture Note

Importance of the Course

2

- ❑ The data analytics is indeed a revolution in the field of computer.
- ❑ The use of data analytics by the companies is enhancing every year and the primary focus of the companies is on customers.
- ❑ Many organizations are actively looking out for the right talent to analyze vast amounts of data.
- ❑ Following four perspectives leads to importance of data analytics.



Why Learn Data Analytics?

3

- ❑ A priority for top organizations.
- ❑ Gain problem solving skills.
- ❑ High demand
 - ❑ Increasing job opportunities.
 - ❑ Increasing pay.
 - ❑ Various job titles from which to choose (Metrics and Analytics Specialist, Data Analyst, Big Data Engineer, Data Analytics Consultant)
- ❑ Analytics is everywhere.
- ❑ It's only becoming more important.
- ❑ It represents perfect freelancing opportunities.
- ❑ Develop new revenue streams

Course Contents



4

Sr #	Major and Detailed Coverage Area	Hrs
1	Introduction to Big Data Introduction to Data, Big Data Characteristics, Types of Big Data, Challenges of Traditional, Systems, Web Data, Evolution of Analytic Scalability, OLTP, MPP, Grid Computing, Cloud Computing, Fault Tolerance, Analytic Processes and Tools, Analysis Versus Reporting, Statistical Concepts, Types of Analytics.	9
2	Data Analysis Introduction to Data Analysis, Importance of Data Analysis, Data Analytics Applications, Regression Modelling Techniques: Linear Regression, Multiple Linear Regression, Non Linear Regression, Logistic Regression, Bayesian Modelling, Bayesian Networks, Support Vector Machines, Time Series Analysis, Rule Induction, Sequential Cover Algorithm.	12
3	Mining Data Streams Introduction to Mining Data Streams, Data Stream Management Systems, Data Stream Mining, Examples of Data Stream Applications, Stream Queries, Issues in Data Stream Query Processing, Sampling in Data Streams, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Querying on Windows – Counting Ones in a Window, Decaying Windows, Real-Time Analytics Platform (RTAP).	10

Course Contents continue...



5

Sr #	Major and Detailed Coverage Area	Hrs
4	Frequent Itemsets and Clustering Introduction to Frequent Itemsets, Market-Basket Model, Algorithm for Finding Frequent Itemsets, Association Rule Mining, Apriori Algorithm, Introduction to Clustering, Overview of Clustering Techniques, Hierarchical Clustering, Partitioning Methods, K-Means Algorithm, Clustering High-Dimensional Data.	10
5	Frameworks and Visualization Introduction to framework and Visualization, Introduction to Hadoop, Core Components of Hadoop, Hadoop Ecosystem, Physical Architecture, Hadoop Limitations, Hive, MapReduce and The New Software Stack, MapReduce, Algorithms Using MapReduce, NOSQL, NoSQL Business Drivers, NoSQL Case Studies, NoSQL Data Architectural Patterns, Variations of NoSQL, Architectural Patterns, Using NoSQL to Manage Big Data, Visualizations	8

Course Outcome

6

CO #	CO	Unit
CO1	Understand and classify the characteristics, concepts and principles of big data.	Introduction to Big Data
CO2	Apply the data analytics techniques and models.	Data Analysis
CO3	Implement and analyze the data analysis techniques for mining data streams.	Mining Data Streams
CO4	Examine the techniques of clustering and frequent item sets.	Frequent Itemsets and Clustering
CO5	Analyze and evaluate the framework and visualization for big data analytics.	Frameworks and Visualization
CO6	Formulate the concepts, principles and techniques focusing on the applications to industry and real world experience.	Applications of all units

Prerequisites

- NIL

Books

7

Textbook

- ❑ Data Analytics, Radha Shankarmani,M. Vijayalaxmi, Wiley India Private Limited, ISBN: 9788126560639.

Reference Books

- ❑ Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (Editor), Wiley, 2014
- ❑ Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with advanced analytics, John Wiley & sons, 2012.
- ❑ Glenn J. Myatt, Making Sense of Data, John Wiley & Sons, 2007 Pete Warden, Big
- ❑ Data Glossary,O'Reilly, 2011.
- ❑ Jiawei Han, MichelineKamber “Data Mining Concepts and Techniques”, Second Edition, Elsevier, Reprinted 2008.
- ❑ Stephan Kudyba, Thomas H. Davenport, Big Data, Mining, and Analytics, Components of Strategic Decision Making, CRC Press, Taylor & Francis Group. 2014
- ❑ Big Data, Black Book, DT Editorial Services, Dreamtech Press, 2015

Evaluation

8

Grading:

- Internal assessment – 30 marks
 - 2 quizzes = $2.5 \times 2 = 5$ marks
 - 5 group assignments = $2 \times 5 = 10$ marks
 - Class participation = 5 marks
 - Mini Project = 10 marks
- Mid-Term exam - 20 marks
- End-Term exam - 50 marks



Data

9

- ❑ A representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalized manner.
- ❑ Data is either intended to be processed, is being processed, or has been processed.
- ❑ It can be in any form stored internally in a computer system or computer network or in a person's mind.
- ❑ Since the mid-1900s, people have used the word **data** to mean computer information that is transmitted or stored.
- ❑ Data is the plural of datum (a Latin word meaning something given), a single piece of information. In practice, however, people use data as both the singular and plural form of the word.
- ❑ It must be interpreted, by a human or machine to derive meaning.
- ❑ It is present in homogeneous sources as well as heterogeneous sources.
- ❑ The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights.

Data → Information → Knowledge → Actionable Insights

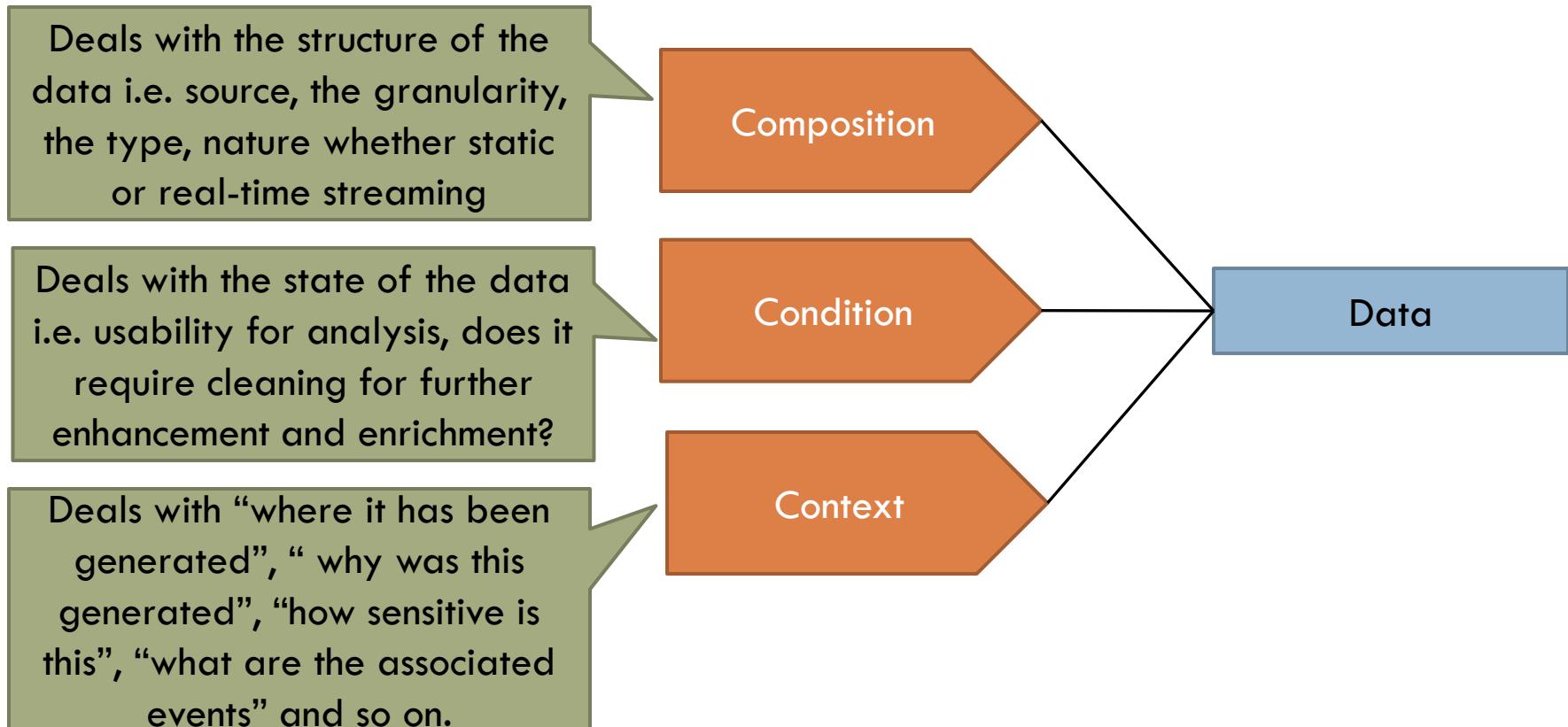
Importance of Data

10

- ❑ The ability to analyze and act on data is increasingly important to businesses. It might be part of a study helping to cure a disease, boost a company's revenue, understand and interpret market trends, study customer behavior and take financial decisions
- ❑ The pace of change requires companies to be able to react quickly to changing demands from customers and environmental conditions. Although prompt action may be required, decisions are increasingly complex as companies compete in a global marketplace
- ❑ Managers may need to understand high volumes of data before they can make the necessary decisions
- ❑ Relevant data creates strong strategies - Opinions can turn into great hypotheses, and those hypotheses are just the first step in creating a strong strategy. It can look something like this: "Based on X, I believe Y, which will result in Z"
- ❑ Relevant data strengthens internal teams
- ❑ Relevant data quantifies the purpose of the work

Characteristics of Data

11



Human vs. Machine Readable data



12

- ❑ Human-readable refers to information that only humans can interpret and study, such as an image or the meaning of a block of text. If it requires a person to interpret it, that information is human-readable.
- ❑ Machine-readable refers to information that computer programs can process. A program is a set of instructions for manipulating data. Such data can be automatically read and processed by a computer, such as CSV, JSON, XML, etc.

Non-digital material (for example printed or hand-written documents) is by its non-digital nature not machine-readable. But even digital material need not be machine-readable. For example, a PDF document containing tables of data. These are definitely digital but are not machine-readable because a computer would struggle to access the tabular information - even though they are very human readable. The equivalent tables in a format such as a spreadsheet would be machine readable.

Another example scans (photographs) of text are not machine-readable (but are human readable!) but the equivalent text in a format such as a simple ASCII text file can be machine readable and processable.

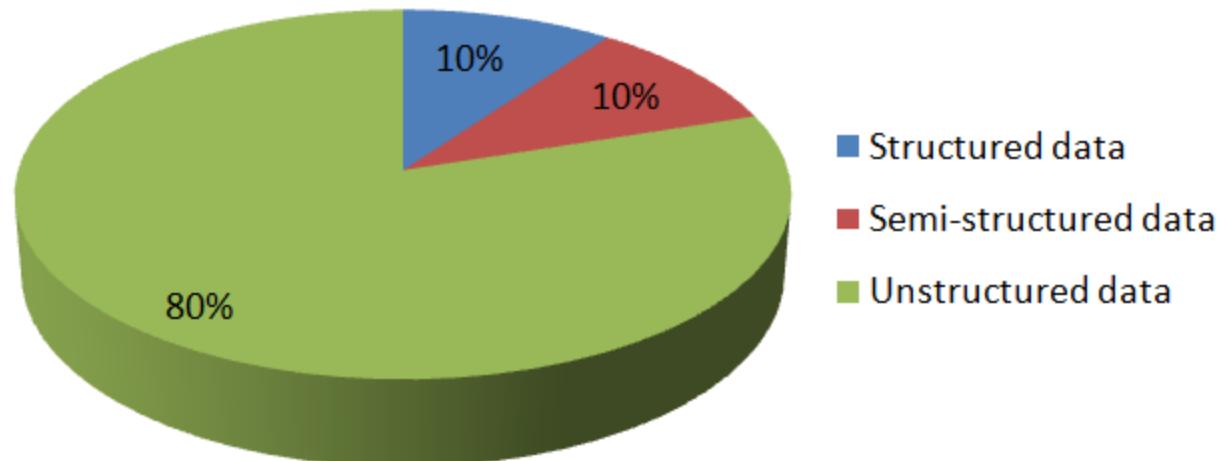
Classification of Digital Data

13

Digital data is classified into the following categories:

- Structured data
- Semi-structured data
- Unstructured data

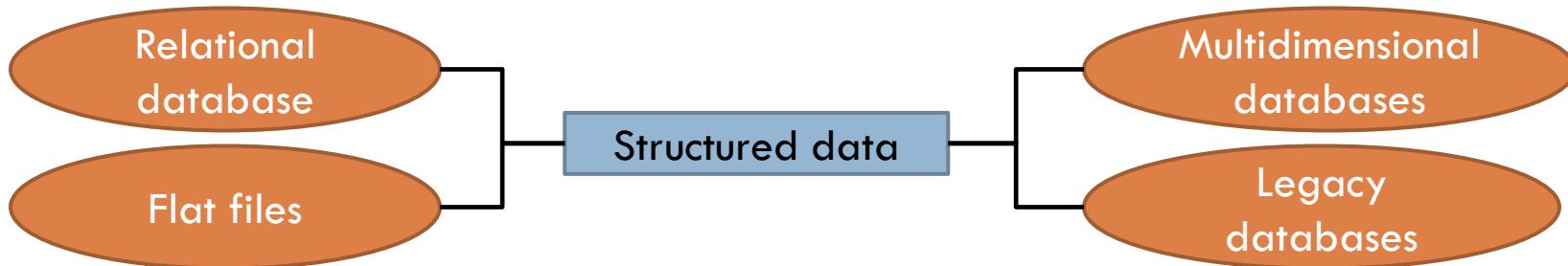
Approximate percentage distribution of digital data



Structured Data

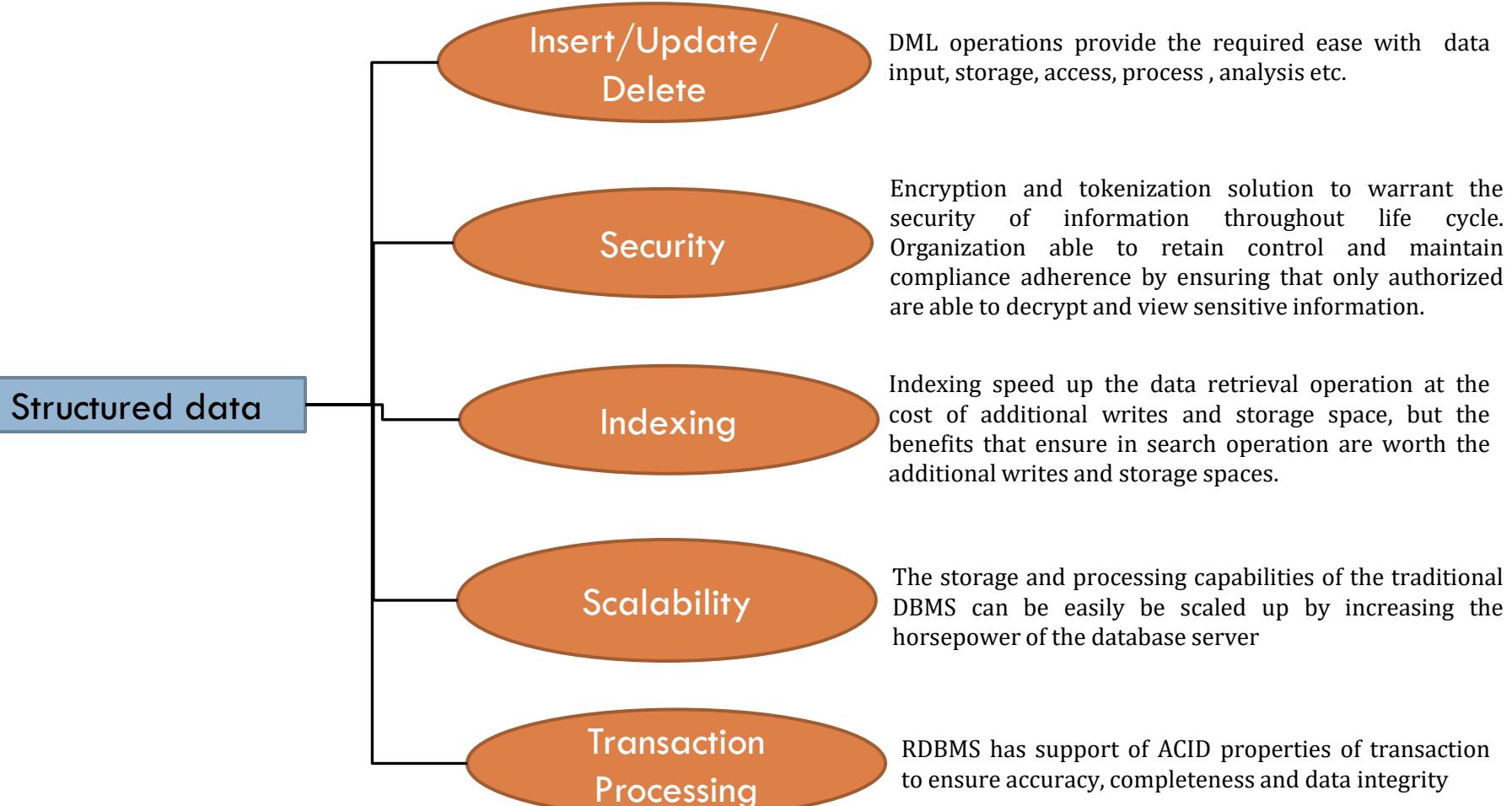
14

- ❑ It is defined as the data that has a defined repeating pattern and this pattern makes it easier for any program to sort, read, and process the data.
- ❑ This data is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- ❑ Relationships exist between entities of data.
- ❑ Structured data:
 - ❑ Organize data in a pre-defined format
 - ❑ Is stored in a tabular form
 - ❑ Is the data that resides in a fixed fields within a record of file
 - ❑ Is formatted data that has entities and their attributes mapped
 - ❑ Is used to query and report against predetermined data types
- ❑ Sources:



Ease with Structured Data

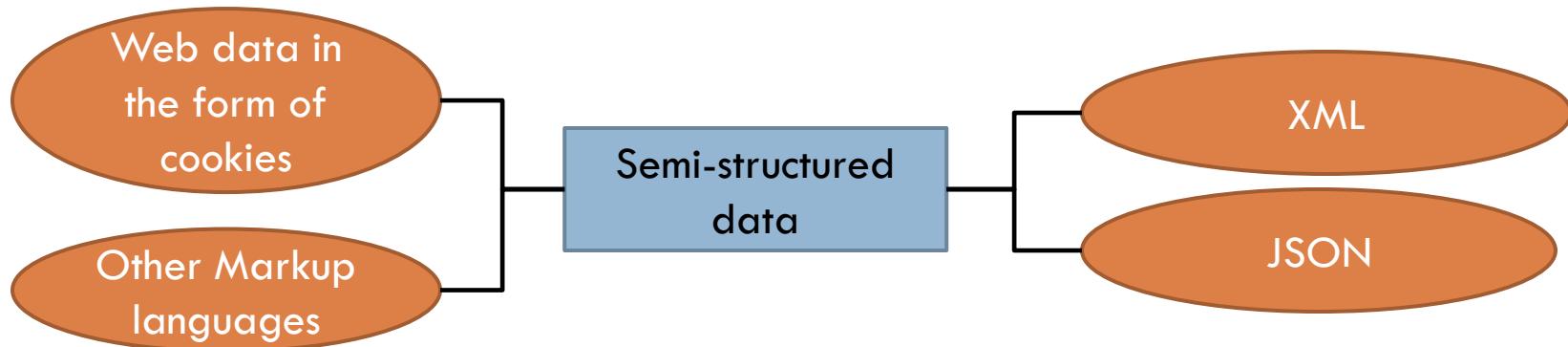
15



Semi-structured Data

16

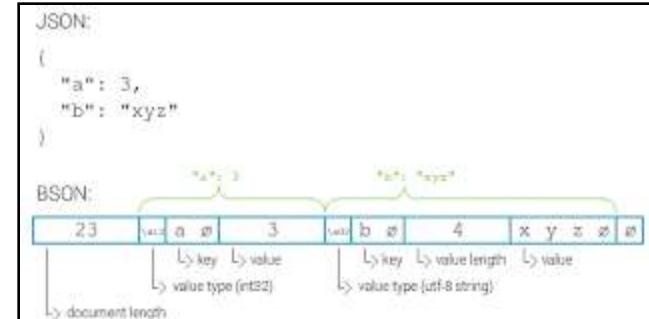
- ❑ Semi-structured data, also known as having a schema-less or self-describing structure, refers to a form which does not conform to a data model as in relational database but has some structure.
- ❑ In other words, data is stored inconsistently in rows and columns of a database.
- ❑ However, it is not in a form which can be used easily by a computer program.
- ❑ Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
- ❑ Sources:



XML, JSON, BSON format

17

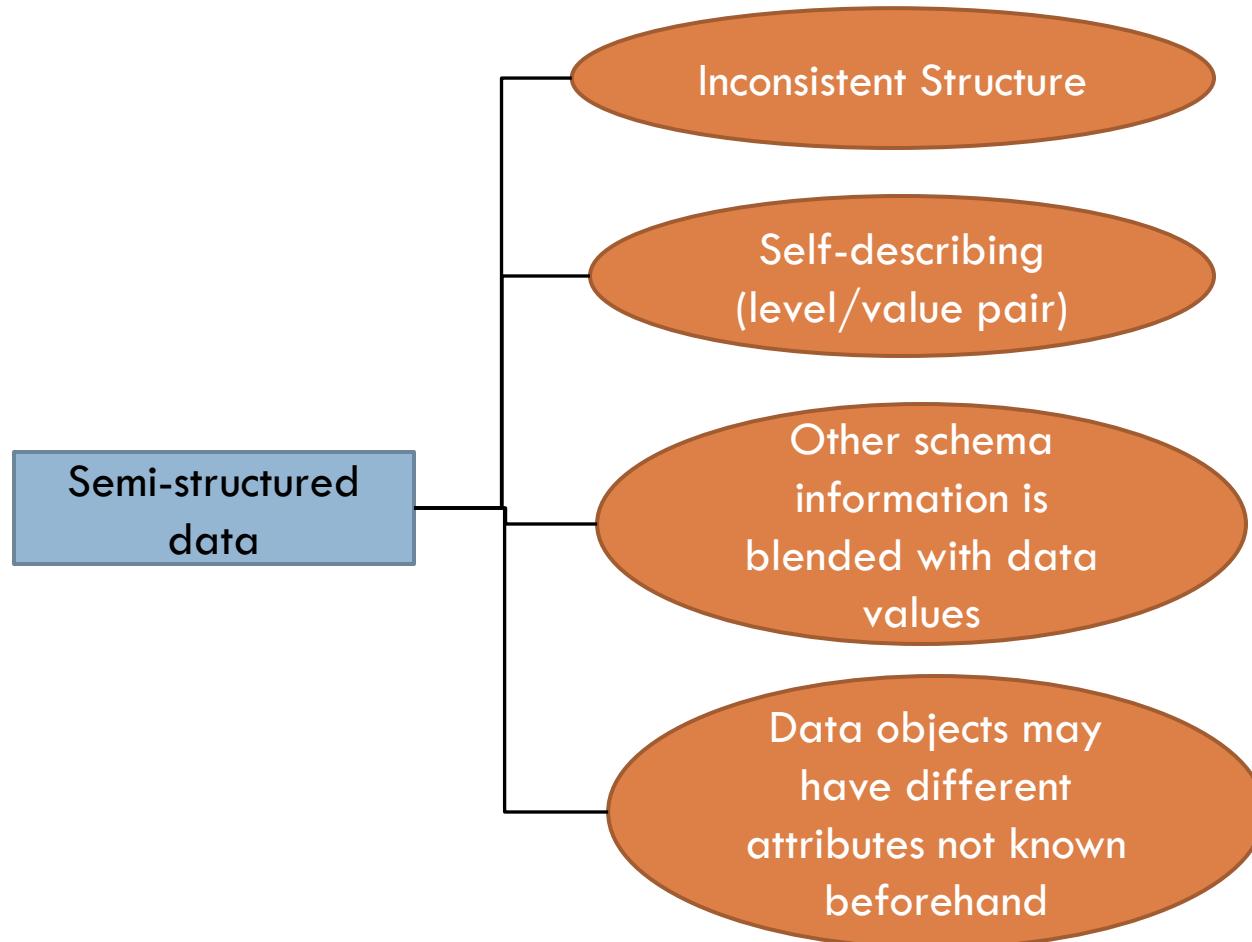
XML	JSON
<pre> <Node> <id>10002</id> <Name>john</Name> </Node> <Node> <id>10003</id> <Name>Scott</Name> </Node> <Node> <id>10004</id> <Name>Mohan</Name> </Node> <Node> <id>10001</id> <Name>Deepak </Name> </Node> </pre>	<pre> [{ "id":10002, "name":"john" }, { "id":10003, "name":"Scott" }, { "id":10004, "name":"Mohan" }, { "id":10001, "name":"Deepak" }] </pre>



Source (XML & JSON): <http://sqllearnergroups.blogspot.com/2014/03/how-to-get-json-format-through-sql.html>
 Source (JSON & BSON): <http://www.expert-php.fr/mongodb-bson/>

Characteristics of Semi-structured Data

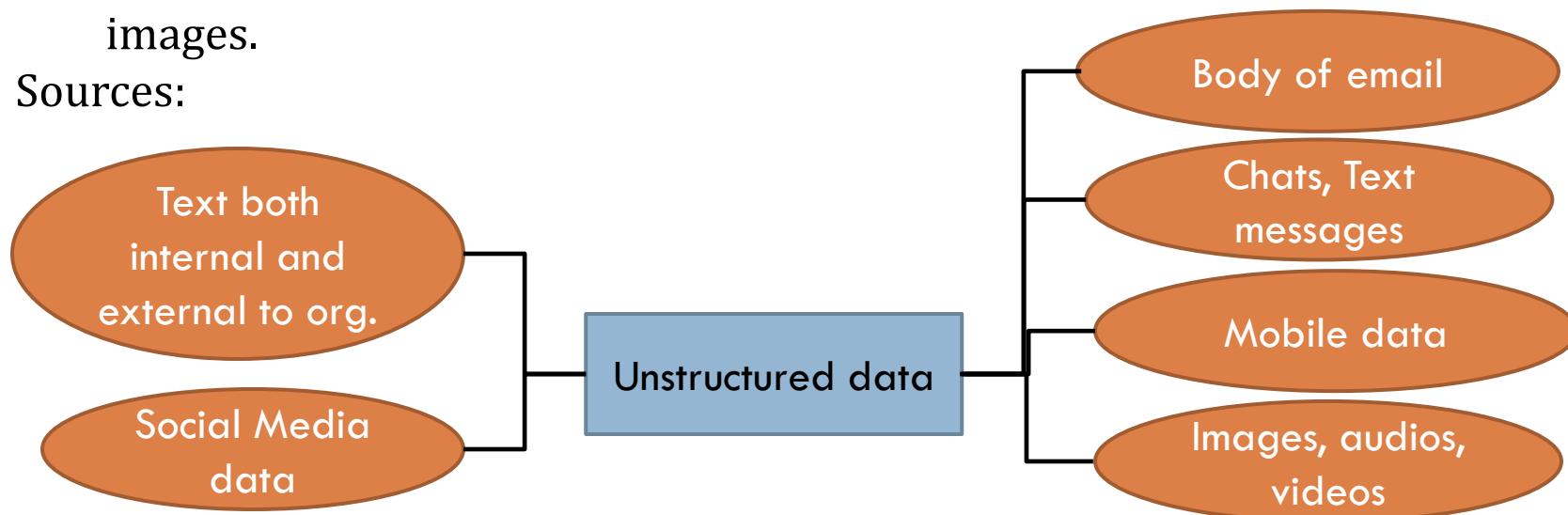
18



Unstructured Data

19

- ❑ Unstructured data is a set of data that might or might not have any logical or repeating patterns and is not recognized in a pre-defined manner.
- ❑ About 80 percent of enterprise data consists of unstructured content.
- ❑ Unstructured data:
 - ❑ Typically consists of metadata i.e. additional information related to data.
 - ❑ Comprises of inconsistent data such as data obtained from files, social media websites, satellites etc
 - ❑ Consists of data in different formats such as e-mails, text, audio, video, or images.
- ❑ Sources:



Challenges associated with Unstructured data

20

Working with unstructured data poses certain challenges, which are as follows:

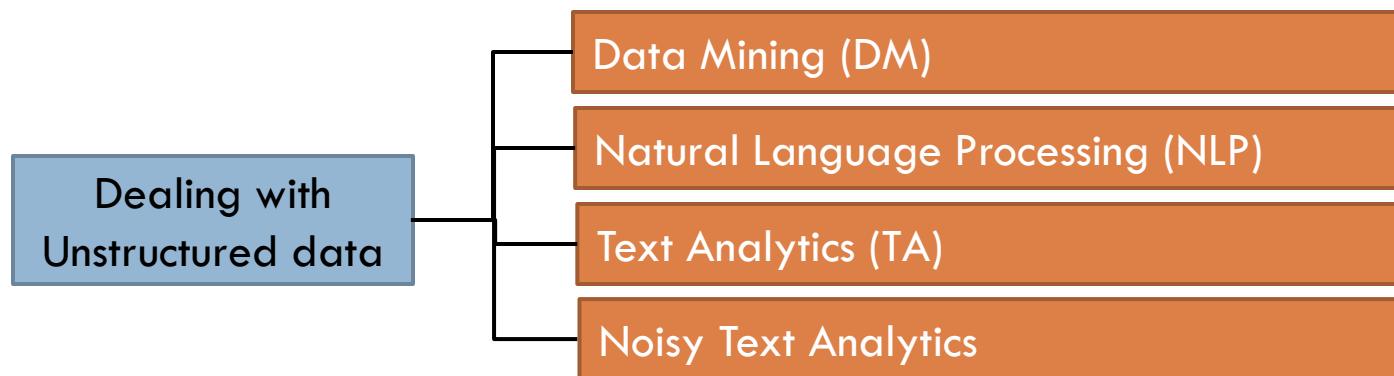
- ❑ Identifying the unstructured data that can be processed
- ❑ Sorting, organizing, and arranging unstructured data in different sets and formats
- ❑ Combining and linking unstructured data in a more structured format to derive any logical conclusions out of the available information
- ❑ Costing in terms of storage space and human resources need to deal with the exponential growth of unstructured data

Data Analysis of Unstructured Data

The complexity of unstructured data lies within the language that created it. Human language is quite different from the language used by machines, which prefer structured information. Unstructured data analysis is referred to the process of analyzing data objects that doesn't follow a predefine data model and/or is unorganized. It is the analysis of any data that is stored over time within an organizational data repository without any intent for its orchestration, pattern or categorization.

Dealing with Unstructured data

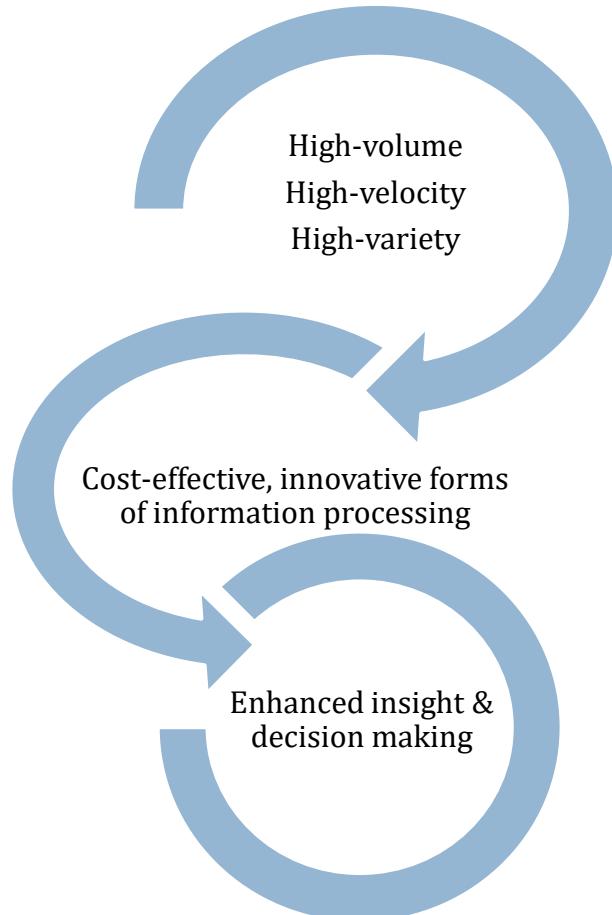
21



Note: Refer to Appendix for further details.

Definition of Big Data

22



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.

Source: Gartner IT Glossary

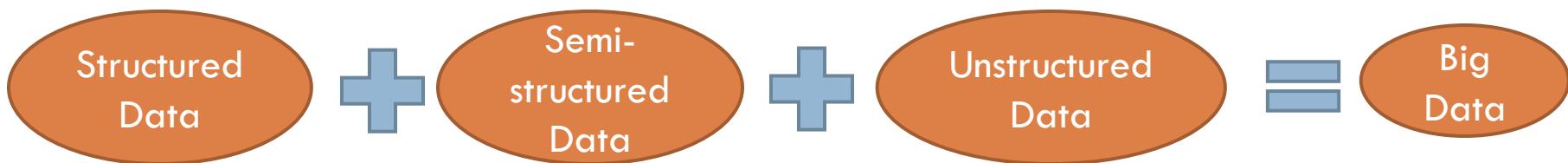
What is Big Data?

23

Think of following:

- Every second, there are around 822 tweets on Twitter
- Every minutes, nearly 510 comments are posted, 293 K statuses are updated, and 136K photos are uploaded in Facebook
- Every hour, Walmart, a global discount departmental store chain, handles more than 1 million customer transactions.
- Everyday, consumers make around 11.5 million payments by using PayPal.

In the digital world, data is increasing rapidly because of the ever increasing use of the internet, sensors, and heavy machines at a very high rate. The sheer volume, variety, velocity, and veracity of such data is signified the term '**Big Data**'.



Elements of Big Data

24

In most big data circles, these are called the four V's: **volume**, **variety**, **velocity**, and **veracity**. (One might consider a fifth V, **value**.)

Volume - refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc. The vast amounts of data have become so large in fact it can no longer store and perform data analysis using traditional database technology. So using distributed systems, where parts of the data is stored in different locations and brought together by software.

Variety - defined as the different types of data the digital system now use. Data today looks very different than data from the past. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

Velocity - refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every second of every day data is increasing. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access. Big data technology allows to analyze the data while it is being generated, without ever putting it into databases.

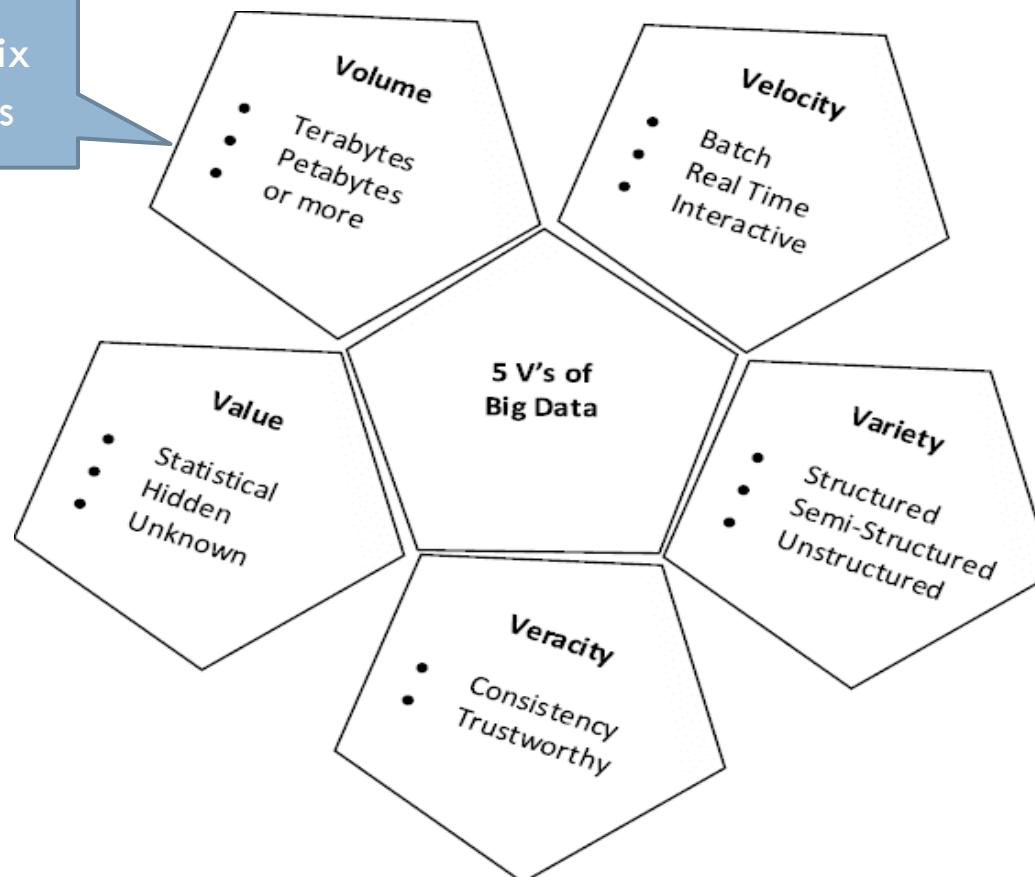
Veracity - is the quality or trustworthiness of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content.

Elements of Big Data cont'd

25

Value - refers to the ability to transform a tsunami of data into business. Having endless amounts of data is one thing, but unless it can be turned into value it is useless.

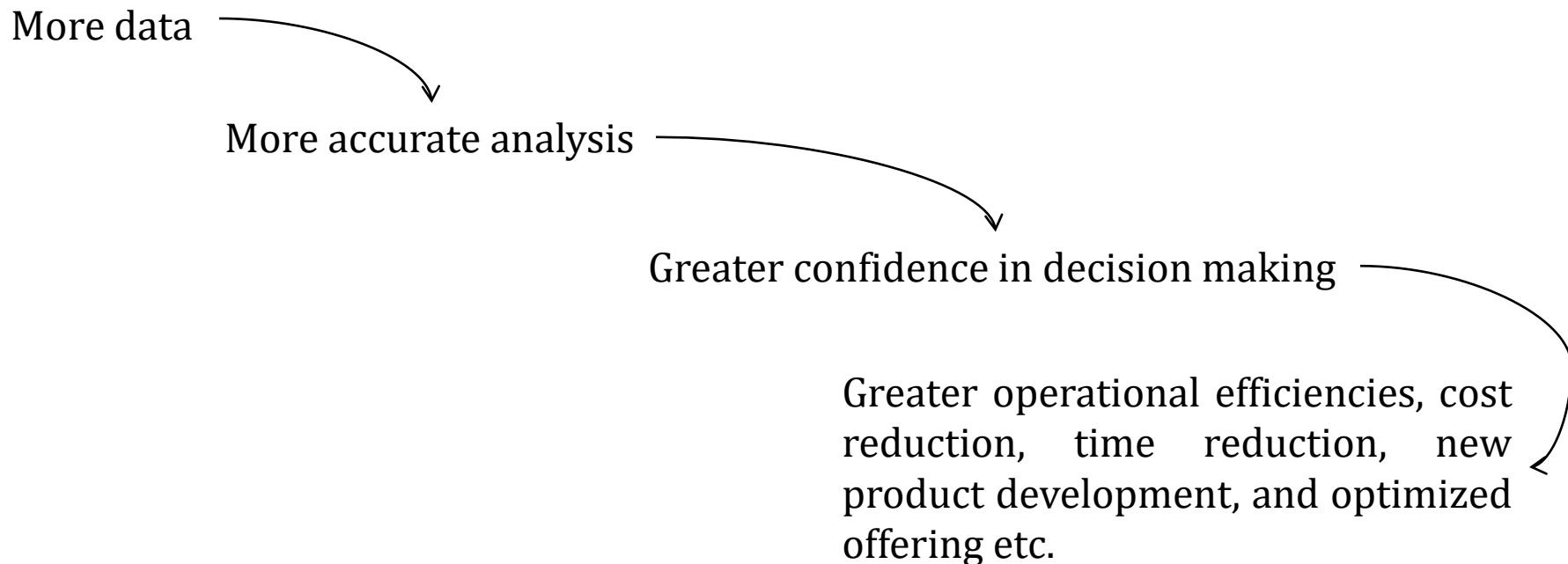
Refer to Appendix
for data volumes



Why Big Data?

26

More data for analysis will result into **greater analytical accuracy** and greater **confidence in the decisions** based on the analytical findings. This would entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services and optimizing existing services.



Challenges of Traditional Systems

27

The main challenge in the traditional approach for computing systems to manage 'Big Data' because of immense speed and volume at which it is generated. Some of the challenges are:

- ❑ Traditional approach cannot work on unstructured data efficiently
- ❑ Traditional approach is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. This approach will not adequate for big data
- ❑ Traditional approach is batch oriented and need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained
- ❑ Traditional data management, warehousing, and analysis systems fizzle to analyze this type of data. Due to its complexity, big data is processed with parallelism. Parallelism in a traditional system is achieved through costly hardware like MPP (Massively Parallel Processing) systems
- ❑ Inadequate support of aggregated summaries of data

Challenges of Traditional Systems cont'd

28

Other challenges can be categorized as:

- ❑ Data Challenges:
 - ❑ Volume, velocity, veracity, variety
 - ❑ Data discovery and comprehensiveness
 - ❑ Scalability
- ❑ Process challenges
 - ❑ Capturing Data
 - ❑ Aligning data from different sources
 - ❑ Transforming data into suitable form for data analysis
 - ❑ Modeling data(Mathematically, simulation)
- ❑ Management Challenges:
 - ❑ Security
 - ❑ Privacy
 - ❑ Governance
 - ❑ Ethical issues

Web Data

29

- ❑ It refers to the data that is publicly available on the web sites.
- ❑ The web data has documents in pdf, doc, docx, plain text as well as images, music, and videos.
- ❑ The most widely used and best-known source of big data today is the detailed data collected from web sites.
- ❑ The data is unstructured and inappropriate for access by software application, and hence is converted to either semi-structured or structured format that is well suited for both humans and machines.

Distributed vs. Parallel Computing

30

Parallel Computing	Distributed Computing
Shared memory system	Distributed memory system
Multiple processors share a single bus and memory unit	Autonomous computer nodes connected via network
Processor is order of Tbps	Processor is order of Gbps
Limited Scalability	Better scalability and cheaper
	Distributed computing in local network (called cluster computing). Distributed computing in wide-area network (grid computing)

EDW, OLTP, MPP

31

- ❑ **Enterprise Data Warehouse:** An enterprise data warehouse (EDW) is a database, or collection of databases, that centralizes a business's information from multiple sources and applications, and makes it available for analytics and use across the organization. EDWs can be housed in an on-premise server or in the cloud. The data stored in this type of digital warehouse can be one of a business's most valuable assets, as it represents much of what is known about the business, its employees, its customers, and more.
- ❑ **Online Transactional Processing (OLTP):** It is a category of data processing that is focused on transaction-oriented tasks. OLTP typically involves inserting, updating, and/or deleting small amounts of data in a database. OLTP mainly deals with large numbers of transactions by a large number of users.
- ❑ **Massively Parallel Processing (MPP):** It is a storage structure designed to handle the coordinated processing of program operations by multiple processors. This coordinated processing can work on different parts of a program, with each processor using its own operating system and memory. This allows MPP databases to handle massive amounts of data and provide much faster analytics based on large datasets.

Hadoop

32

- ❑ Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.
- ❑ It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
- ❑ It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
- ❑ Importance:
 - ❑ Ability to store and process huge amounts of any kind of data, quickly.
 - ❑ **Computing power:** It's distributed computing model processes big data fast.
 - ❑ **Fault tolerance:** Data and application processing are protected against hardware failure.
 - ❑ **Flexibility:** Unlike traditional relational databases, preprocess of data does not require before storing it.
 - ❑ **Low cost:** The open-source framework is free and uses commodity hardware to store large quantities of data.
 - ❑ Scalability: System can easily grow to handle more data simply by adding nodes. Little administration is required.

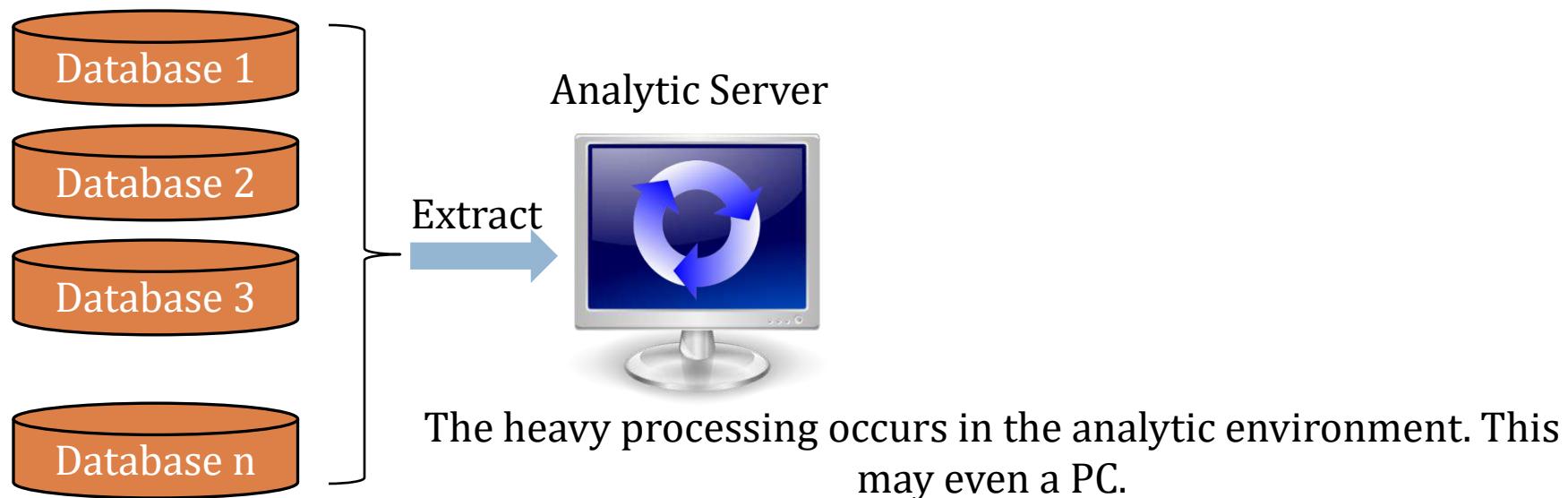
Evolution of Analytics Scalability

33

- ❑ As the amount of data organizations process continue to increase, the world of big data requires new levels of scalability. Organizations need to update the technology to provide a higher level of scalability.
- ❑ Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.
- ❑ The technologies are:
 - ❑ MPP (massively parallel processing)
 - ❑ Cloud computing
 - ❑ Grid computing
 - ❑ MapReduce

Traditional Analytics Architecture

34



Modern In-Database Analytics Architecture



35



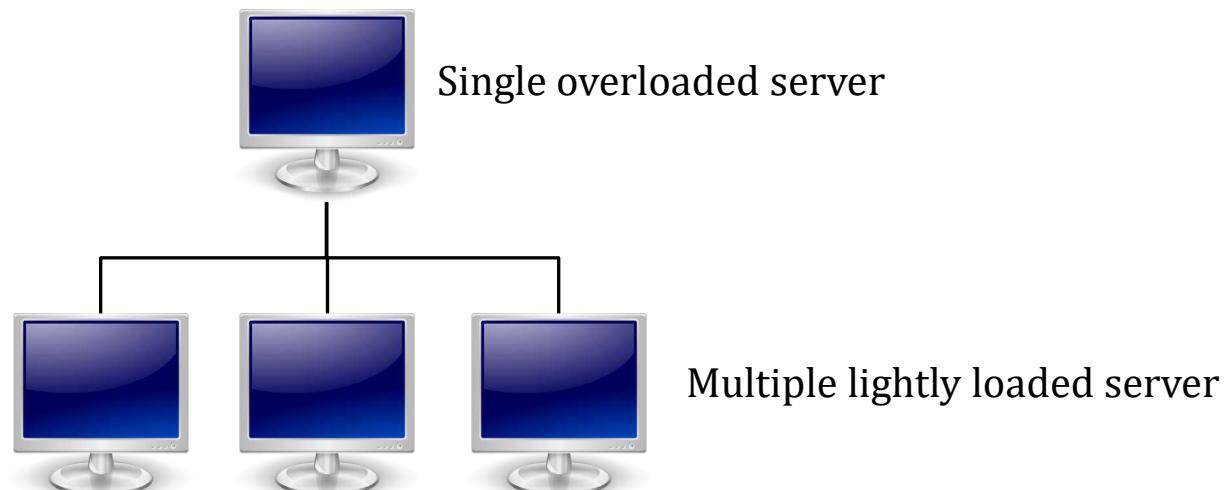
In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

MPP Analytics Architecture

36

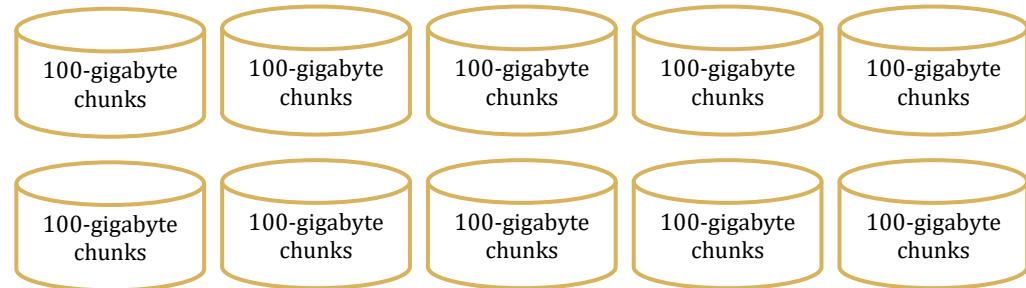
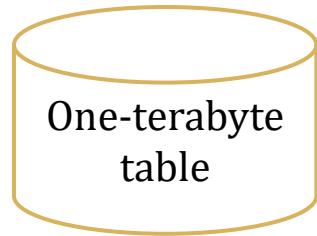
Massively parallel processing (MPP) database systems is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data. An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources. Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house. The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.

In stead of single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU.



MPP Database Example

37



A Traditional database will query a one-terabyte table one row at time

10 simultaneous 100-gigabyte queries

MPP database is based on the principle of **SHARE THE WORK!**

A MPP database spreads data out across multiple sets of CPU and disk space. Think logically about dozens or hundreds of personal computers each holding a small piece of a large set of data. This allows much faster query execution, since many independent smaller queries are running simultaneously instead of just one big query

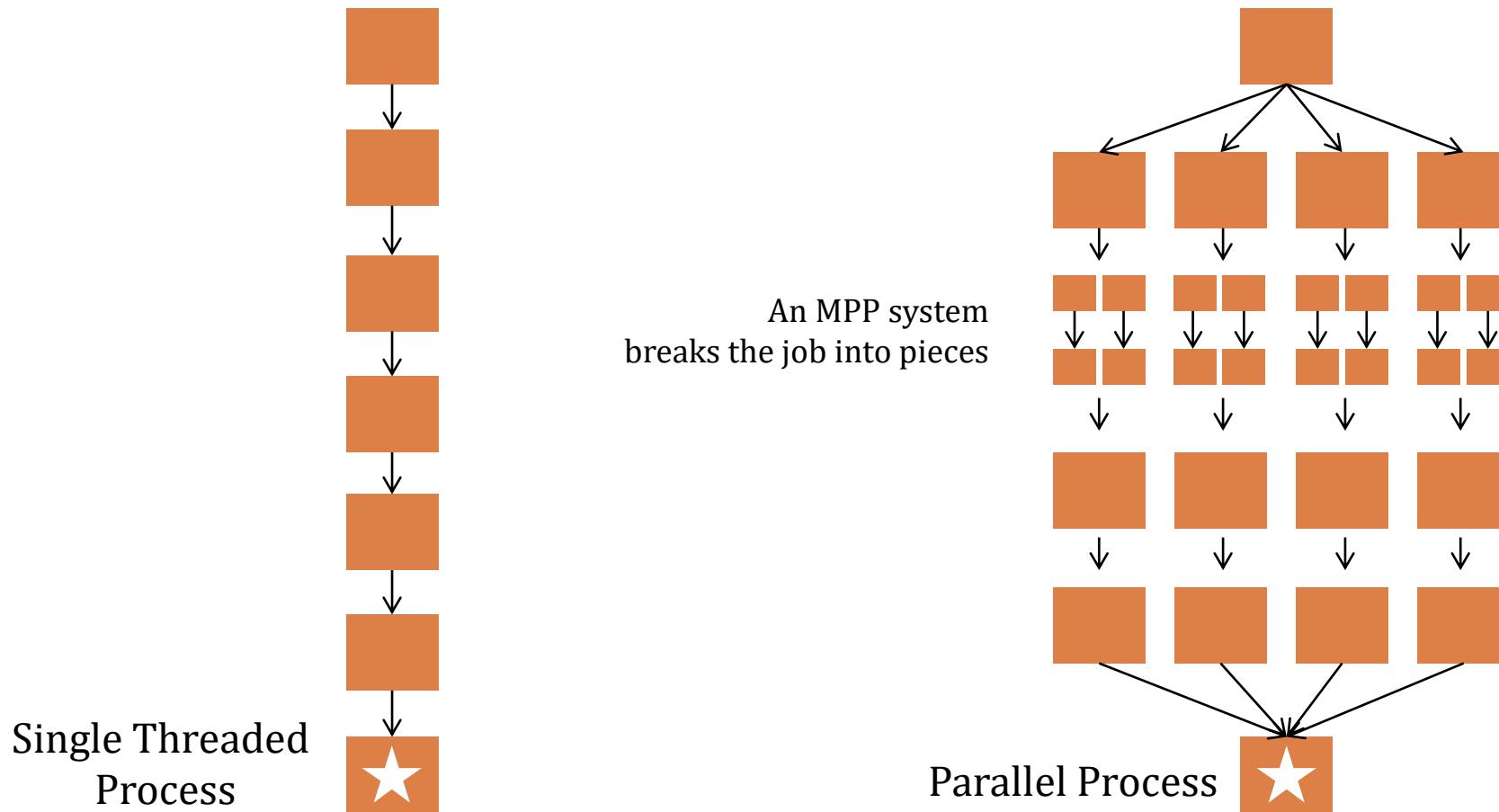
If more processing power and more speed are required, just bolt on additional capacity in the form of additional processing units

MPP systems build in redundancy to make recovery easy and have resource management tools to manage the CPU and disk space

MPP Database Example cont'd

38

An MPP system allows the different sets of CPU and disk to run the process concurrently



OLTP vs. MPP vs. Hadoop

39

OLTP	MPP
Examples: Oracle, DB2, SQL Server etc.	Examples: Netezza, Teradata, Vertica etc.
It needs to read data from disk to memory before start processing, so very fast in memory calculation.	Takes the processing as close possible to the data, so less data movement
It is good for smaller OLTP (transaction) operations. It also maintains very high level of data integrity.	It is good for batch processing. Some of the MPP (Netezza, Vertica) overlooks integrity like enforcing unique key for the sake of batch performance.

MPP	Hadoop
Stores data in a matured internal structure. So data loading and data processing is efficient.	There are no such structured architecture for data stored on Hadoop. So, accessing and loading data is not as efficient as conventional MPP systems.
It support only relational models.	Support virtually any kind of data.
However the main objective of MPP and Hadoop is same, process data parallelly near storage.	

How to choose what?

40

- ❑ OLTP Databases (Oracle, DB2, MySQL, MS SQL, Exadata):
 - ❑ Transaction based application
 - ❑ Smaller DWH
- ❑ MPP (Netezza, Teradata, Vertica):
 - ❑ Bigger Data warehouse (may be having tables with size more than 4-5 TB)
 - ❑ Needs no or little pre-processing
 - ❑ Needs faster batch processing speed
 - ❑ In database analytics
- ❑ Only Hadoop:
 - ❑ All data as heavily unstructured (documents, audio, video etc)
 - ❑ Need to process in batch

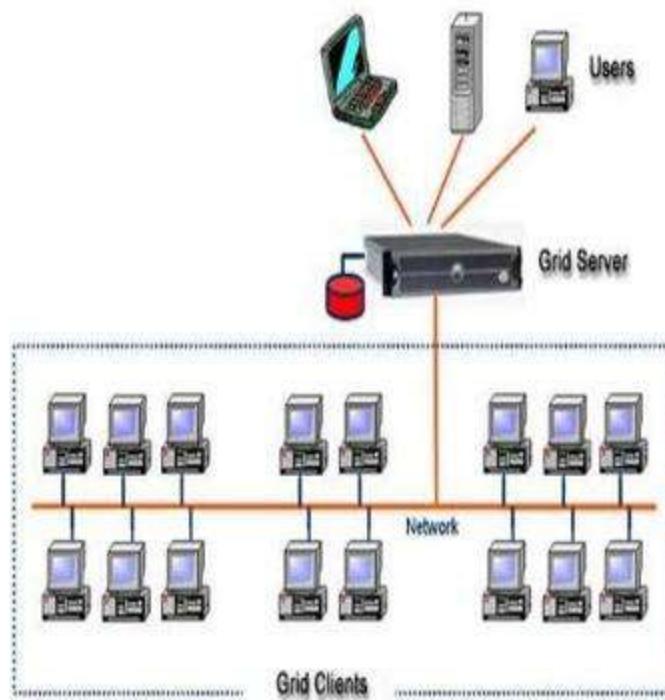
Grid Computing

41

- ❑ Grid Computing can be defined as a network of computers working together to perform a task that would rather be difficult for a single machine.
- ❑ The task that they work on may include analysing huge datasets or simulating situations which require high computing power.
- ❑ Computers on the network contribute resources like processing power and storage capacity to the network.
- ❑ Grid Computing is a subset of distributed computing, where a virtual super computer comprises of machines on a network connected by some bus, mostly Ethernet or sometimes the Internet.
- ❑ It can also be seen as a form of parallel computing where instead of many CPU cores on a single machine, it contains multiple cores spread across various locations.

How Grid Computing works?

42



In general, a grid computing system requires:

- At least one computer, usually a server, which handles all the administrative duties for the System
- A network of computers running special grid computing network software.
- A collection of computer software called middleware

Cloud Computing

43

- ❑ It is a internet-based computing and relies on sharing computing resources on-demand rather than having local PCs and other devices.
- ❑ It is the delivery of on-demand computing services - from applications to storage and processing power over the internet and on a pay-as-you-go basis.
- ❑ It uses high-capacity networks, low-cost computers, and storage devices and adopts hardware virtualization, service-oriented architecture, and utility computing.
- ❑ Rather than owning their own computing infrastructure or data centers, companies can rent access to anything from applications to storage from a cloud service provider and can scale up and scale down as per their computing demands.
- ❑ There are 3 types of cloud environment named public cloud, private cloud and hybrid cloud.

Public Cloud

44

- ❑ It is the most common type of cloud computing deployment.
- ❑ The cloud resources (like servers and storage) are owned and operated by a third-party cloud service provider and delivered over the internet.
- ❑ With a public cloud, all hardware, software and other supporting infrastructure are owned and managed by the cloud provider.
- ❑ In a public cloud, the same hardware, storage and network devices are shared with other organizations or cloud “tenants,” and the adopter access services and manage account using a web browser.
- ❑ Public cloud deployments are frequently used to provide web-based email, online office applications, storage and testing and development environments.
- ❑ Advantages of public clouds are lower costs, no maintenance, high reliability etc.

Private Cloud

45

- ❑ A private cloud consists of cloud computing resources used exclusively by one business or organization.
- ❑ The private cloud can be physically located at your organization's on-site datacenter or it can be hosted by a third-party service provider.
- ❑ The services and infrastructure are always maintained on a private network and the hardware and software are dedicated solely to the organisation.
- ❑ It is often used by government agencies, financial institutions, any other mid- to large-size organizations with business-critical operations seeking enhanced control over their environment.
- ❑ Advantages of private clouds are more flexibility, more control, and more scalability etc.

Hybrid Cloud

46

- ❑ A hybrid cloud combines on-premises infrastructure or a private cloud with a public cloud.
- ❑ It allows data and apps to move between the two environments.
- ❑ Many organizations choose a hybrid cloud approach due to business imperatives such as meeting regulatory and data sovereignty requirements, taking full advantage of on-premises technology investment or addressing low latency issues.
- ❑ A hybrid cloud platform gives organizations many advantages—such as greater flexibility, more deployment options, security, compliance and getting more value from their existing infrastructure.
- ❑ When computing and processing demand fluctuates, hybrid cloud computing gives businesses the ability to seamlessly scale up their on-premises infrastructure to the public cloud to handle any overflow—without giving third-party datacenters access to the entirety of their data.

Fault Tolerance

47

- ❑ Fault tolerance refers to the ability of a system (computer, network, cloud cluster, etc.) to continue operating without interruption when one or more of its components fail.
- ❑ The objective of creating a fault-tolerant system is to prevent disruptions arising from a single point of failure, ensuring the high availability and business continuity of mission-critical applications or systems.
- ❑ Fault-tolerant systems use backup components that automatically take the place of failed components, ensuring no loss of service. These include:
 - ❑ **Hardware systems** that are backed up by identical or equivalent systems. For example, a server can be made fault tolerant by using an identical server running in parallel, with all operations mirrored to the backup server.
 - ❑ **Software systems** that are backed up by other software instances. For example, a database with customer information can be continuously replicated to another machine. If the primary database goes down, operations can be automatically redirected to the second database.
 - ❑ **Power sources** that are made fault tolerant using alternative sources. For example, many organizations have power generators that can take over in case main line electricity fails.

Analytic Processes and Tools

48

Self-Study from the book

Points to cover

- Spreadsheets and Analytics Tool
- Analytics Engine
- CRM and Online Marketing Solutions

Analysis vs. Reporting

49

Reporting: The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

Analysis: The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

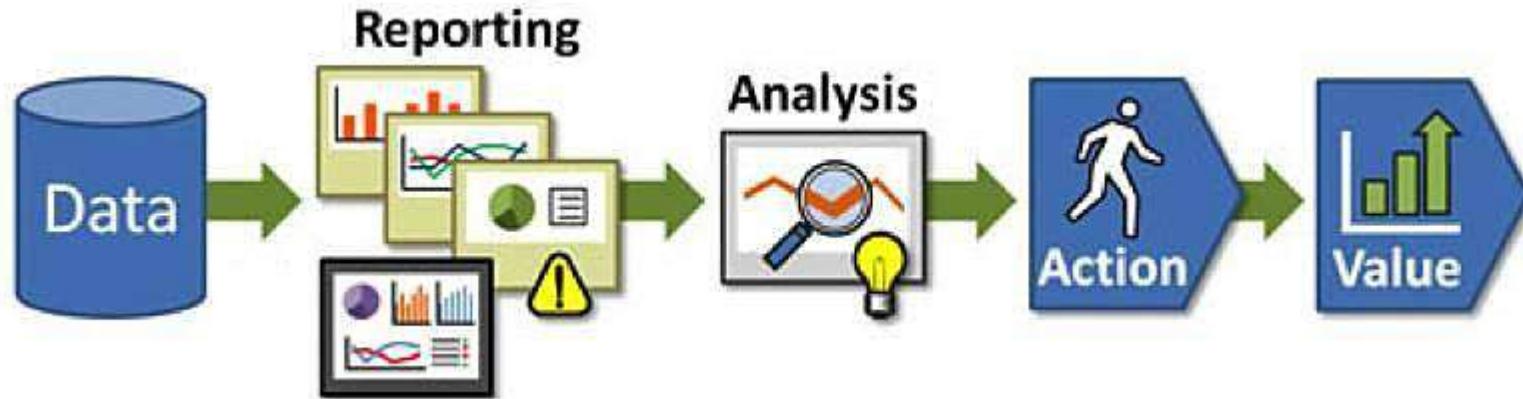
Difference b/w Reporting and Analysis:

- ❑ Reporting translates raw data into information. Analysis transforms data and information into insights.
- ❑ Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges. Good reporting should raise questions about the business from its end users. The goal of analysis is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.
- ❑ In summary, **reporting shows you what is happening while analysis focuses on explaining why it is happening and what you can do about it.**

Goal of Analysis and Reporting



50



Reporting uses data to track the performance of your business, while an analysis uses data to answer strategic questions about your business. Though they are distinct, reporting and analysis rely on each other. Reporting sheds light on what questions to ask, and an analysis attempts to answer those questions.

Simply put,

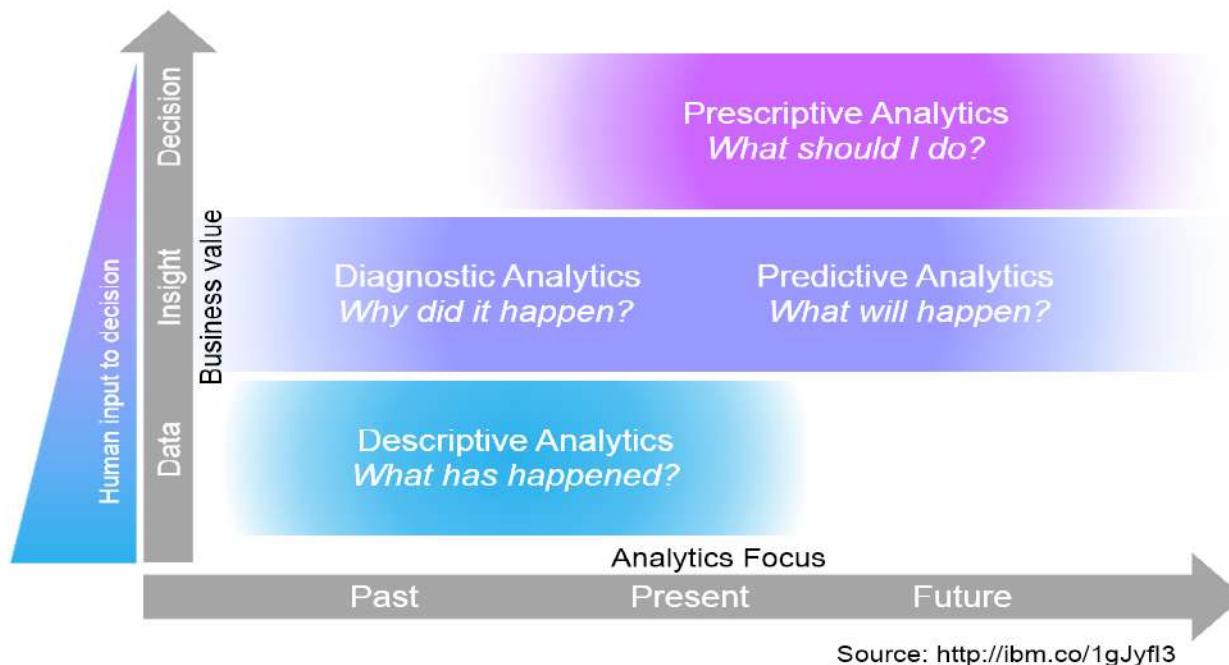
- Data Reporting Reveals The Right Questions.
- Data Analysis Helps Find Answers.

Data Analytics

51

Data analytics is the process of extracting useful information by analysing different types of data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful information for the benefit of faster decision making.

There are 4 types of analytics:



Types of Analytics

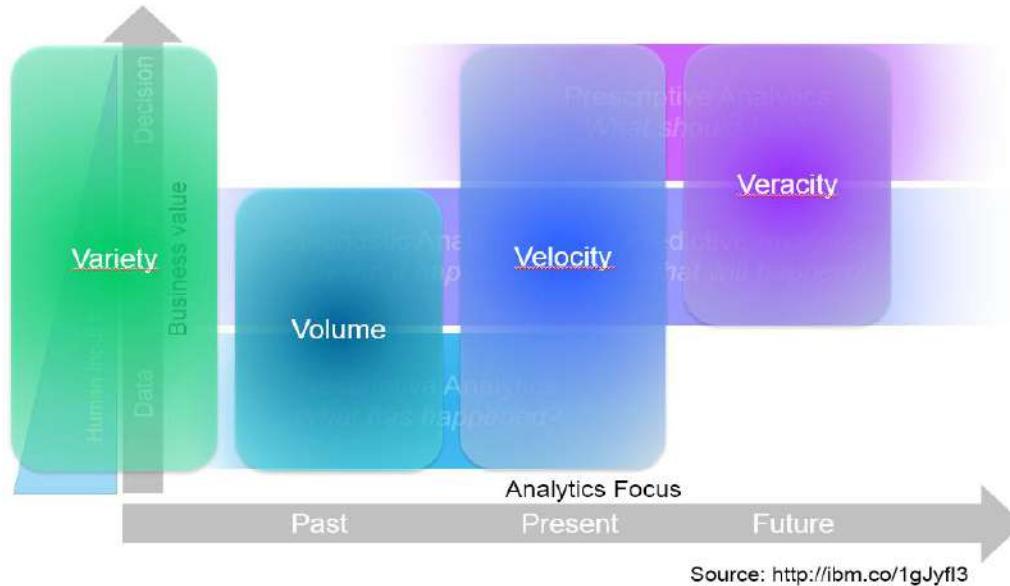
52

Approach	Explanation
Descriptive	<p>What's happening in my business?</p> <ul style="list-style-type: none">• Comprehensive, accurate and historical data• Effective Visualisation
Diagnostic	<p>Why is it happening?</p> <ul style="list-style-type: none">• Ability to drill-down to the root-cause• Ability to isolate all confounding information
Predictive	<p>What's likely to happen?</p> <ul style="list-style-type: none">• Decisions are automated using algorithms and technology• Historical patterns are being used to predict specific outcomes using algorithms
Prescriptive	<p>What do I need to do?</p> <ul style="list-style-type: none">• Recommended actions and strategies based on champion/challenger strategy outcomes• Applying advanced analytical algorithm to make specific recommendations

Mapping of Big Data's Vs to Analytics Focus



53



History data can be quite large. There might be a need to process huge amount of data many times a day as it gets updated continuously. Therefore volume is mapped to history. Variety is pervasive. Input data, insights, and decisions can span a variety of forms, hence it is mapped to all three. High velocity data might have to be processed to help real time decision making and plays across descriptive, predictive, and prescriptive analytics when they deal with present data. Predictive and prescriptive analytics create data about the future. That data is uncertain, by nature and its veracity is in doubt. Therefore veracity is mapped to prescriptive and predictive analytics when it deal with future.

Big Data Analytics

54

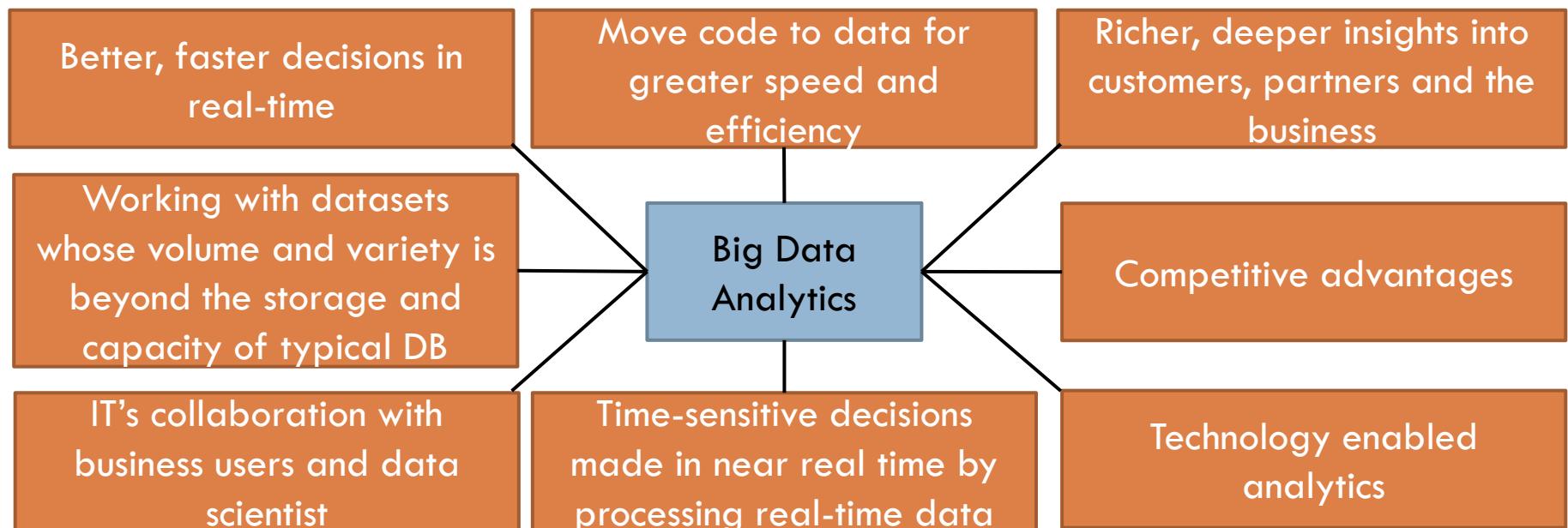
Big data analytics is the process of extracting useful information by analysing different types of big data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful info for the benefit of faster decision making.

Big Data Application in different Industries

<p>Retail/Consumer</p> <ul style="list-style-type: none"> ❖ Merchandizing and market basket analysis ❖ Campaign management and customer loyalty programs ❖ Supply-chain management and analytics ❖ Event- and behavior-based targeting ❖ Market and consumer segmentations 	<p>Finances & Frauds Services</p> <ul style="list-style-type: none"> ❖ Compliance and regulatory reporting ❖ Risk analysis and management ❖ Fraud detection and security analytics ❖ Credit risk, scoring and analysis ❖ High speed arbitrage trading ❖ Trade surveillance ❖ Abnormal trading pattern analysis 	<p>Web and Digital media</p> <ul style="list-style-type: none"> ❖ Large-scale clickstream analytics ❖ Ad targeting, analysis, forecasting and optimization ❖ Abuse and click-fraud prevention ❖ Social graph analysis and profile segmentation ❖ Campaign management and loyalty programs
<p>Health & Life Sciences</p> <ul style="list-style-type: none"> ❖ Clinical trials data analysis ❖ Disease pattern analysis ❖ Campaign and sales program optimization ❖ Patient care quality and program analysis ❖ Medical device and pharmacy supply-chain management ❖ Drug discovery and development analysis 	<p>Telecommunications</p> <ul style="list-style-type: none"> ❖ Revenue assurance and price optimization ❖ Customer churn prevention ❖ Campaign management and customer loyalty ❖ Call detail record (CDR) analysis ❖ Network performance and optimization ❖ Mobile user location analysis 	<p>Ecommerce & customer service</p> <ul style="list-style-type: none"> ❖ Cross-channel analytics ❖ Event analytics ❖ Recommendation engines using predictive analytics ❖ Right offer at the right time ❖ Next best offer or next best action

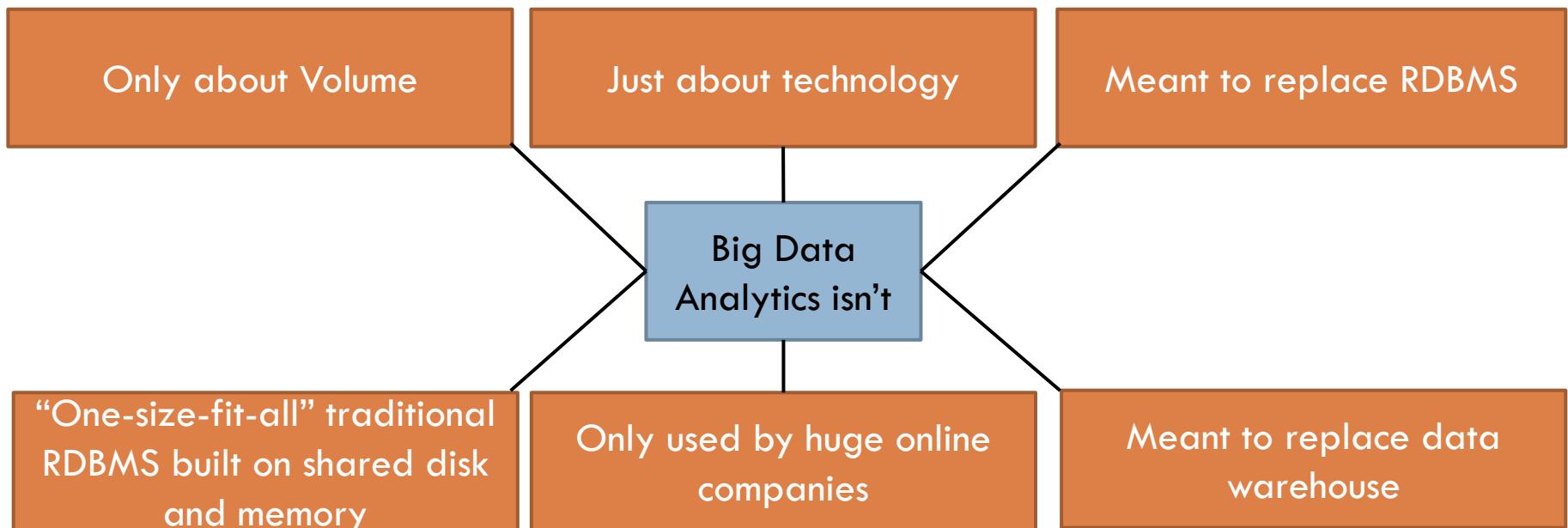
What is Big Data Analytics ?

55



What is Big Data Analytics isn't?

56



Top challenges facing Big Data

57

- 1. Scale:** Storage is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should scale vertically or horizontally?
- 2. Security:** Most of the NoSQL (Not only SQL) big data platforms have poor security mechanism (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data.
- 3. Schema:** Rigid schema have no place. The need of the hour is dynamic schema and static (pre-defined) schemas are passed.
- 4. Data Quality:** How to maintain data quality – data accuracy, completeness, timeliness etc. Is the appropriate metadata in place?
- 5. Partition Tolerant:** How to build partition tolerant systems that can take care of both hardware and software failures?
- 6. Continuous availability:** The question is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.

Kind of Technologies to help meet the challenges posed by Big Data

58

1. Cheap and abundant storage
2. Faster processors to help with quicker processing of big data
3. Affordable open-source, distributed big data platforms
4. Parallel processing, clustering, visualisation, large grid environments, high connectivity, and high throughputs rather than low latency
5. Cloud computing and other flexible resource allocation agreements

Summary

59

Detailed Lessons

Introduction to Data, Big Data Characteristics, Types of Big Data, Challenges of Traditional Systems, Web Data, Evolution of Analytic Scalability, OLTP, MPP, Grid Computing, Cloud Computing, Fault Tolerance, Analytic Processes and Tools, Analysis Versus Reporting, Statistical Concepts, Types of Analytics.

How was the journey?



THANK YOU!

Appendix

61

- ❑ **Data Mining:** Data mining is the process of looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/Previously unknown relationships amongst the data. It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.
- ❑ **Natural Language Processing (NLP):** NLP gives the machines the ability to read, understand and derive meaning from human languages.
- ❑ **Text Analytics (TA):** TA is the process of extracting meaning out of text. For example, this can be analyzing text written by customers in a customer survey, with the focus on finding common themes and trends. The idea is to be able to examine the customer feedback to inform the business on taking strategic action, in order to improve customer experience.
- ❑ **Noisy text analytics:** It is a process of information extraction whose goal is to automatically extract structured or semi-structured information from noisy unstructured text data.

Appendix cont...

62

Example of Data Volumes

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

Data Analytics (IT-3006)

**Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024**

School of Computer Engineering



Strictly for internal circulation (within KIIT) and reference only. Not for outside circulation without permission

3 Credit

Lecture Note – Unit 2

Course Contents



2

Sr #	Major and Detailed Coverage Area	Hrs
2	Data Analysis Introduction to Data Analysis, Importance of Data Analysis, Data Analytics Applications, Regression Modelling Techniques: Linear Regression, Multiple Linear Regression, Non Linear Regression, Logistic Regression, Bayesian Modelling, Basian Networks, Support Vector Machines, Time Series Analysis, Rule Induction, Sequential Cover Algorithm.	12

Introduction

3

- ❑ Rapid advances in computing, data storage, networks etc have dramatically increased the ability to access, store, and process huge amount of data.
- ❑ The field of scientific research and business are challenged with the need to extract relevant information from the huge amounts of data from heterogeneous data sources such as sensors, text achieves, images, videos, audio etc.
- ❑ In such voluminous data, general patterns, structures, regularities go undetected. In many cases, such patterns can be exploited to increase the productivity of an enterprise.

Introduction cont...

4

- ❑ In the data-rich age, understanding how to analyze and extract true meaning from the insights is one of the primary drivers of success.
- ❑ Despite the colossal volume of data created every day, a mere 0.5% is actually analyzed and used for data discovery, improvement, and intelligence.
- ❑ With so much data and so little time, knowing how to collect, curate, organize, and make sense of all of this potentially business-boosting information can be a minefield – but data analysis is the solution.
- ❑ To help the business leaders to understand the potential of analysis, the meaning, and how it can use it to enhance business practices, one has to answer a host of important analytical questions.

Analytical questions

5

- ❑ **Retailers** use it to understand their customer needs and buying habits to predict trends, recommend new products, and boost their business.
- ❑ **Healthcare** industries analyze patient data to provide lifesaving diagnoses and treatment options. They also deal with healthcare plans, insurance information to derive key insights.
- ❑ **Manufacturing** industries can discover new cost-saving and revenue opportunities. They can solve complex supply chain issues, labor constraints, and equipment breakdowns.
- ❑ **Banking** institutions gather and access large volumes of data to derive analytical insights and make sound financial decisions. They find out probable loan defaulters, customer churn out rate, and detect frauds in transactions.
- ❑ **Logistics** companies use data analytics to develop new business models, optimize routes, improve productivity, and order processing capabilities as well as performance management.

Data Analysis

6

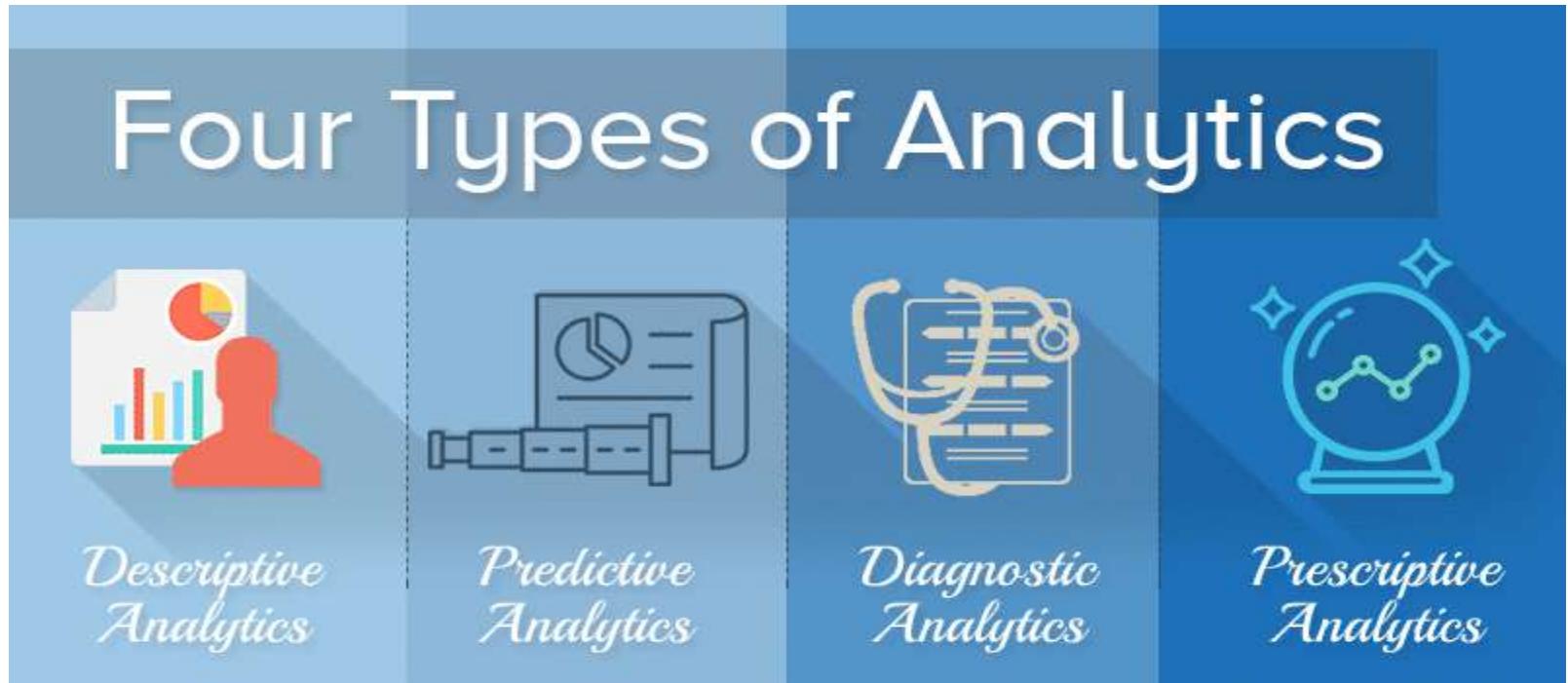
- ❑ Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.
- ❑ Intelligent data analysis (IDA) uses the concept from artificial intelligence (AI), information retrieval (IR), machine learning (ML), pattern reorganization, visualization, distributed programming and a host of other computer science concepts to automate the task of extracting unknown, valuable information /knowledge from the large amount of data.
- ❑ IDA process demands a combination of processes like extraction, analysis, conversion, classification, organization, and reasoning.
- ❑ The IDA process consists of 3 stages namely:
 - ❑ Data preparation
 - ❑ Data mining and rule finding
 - ❑ Result validation and interpretation.

Data Analytics



7

- ❑ Data analytics refers to the process of examining datasets to draw conclusions about the information they contain.
- ❑ Data analytic techniques enable to take raw data and uncover patterns to extract valuable insights from it.



Data Analysis vs. Data Analytics



8

Basis for Comparison	Data Analytics	Data Analysis
Form	Data analytics is 'general' form of analytics which is used in businesses to make decisions from data which are data-driven.	Data analysis is a specialized form of data analytics used in businesses to analyze data and take some insights of it.
Structure	Data analytics consist of data collection and inspect in general and has one or more users.	Data analysis consisted of defining a data, investigation, cleaning, transforming the data to give a meaningful outcome.
Tools	R, Tableau Public, Python, SAS, Apache Spark, Excel are used.	OpenRefine, KNIME, RapidMiner, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha are used.

Data Analysis vs. Data Analytics cont...



9

Basis for Comparison	Data Analytics	Data Analysis
Sequence	The life cycle consist of Business Case Evaluation, Data Identification, Data Acquisition & Filtering, Data Extraction, Data Validation & Cleansing, Data Aggregation & Representation, Data Analysis, Data Visualization, Utilization of Analysis Results.	The sequence followed are data gathering, data scrubbing, analysis of data and interpret the data precisely so that you can understand what data want to convey.
Usage	Find masked patterns, anonymous correlations, customer preferences, market trends and other necessary information that can help to make more notify decisions for business purpose.	Descriptive analysis, exploratory analysis, inferential analysis, predictive analysis and take useful insights from the data.

Data Analysis vs. Data Analytics cont...



10

Basis for Comparison	Data Analytics	Data Analysis
Example	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find that what the customers next possible purchases.	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find what happened so far.

Summary

- ❑ Both data analytics and data analysis are used to uncover patterns, trends, and anomalies lying within data, and thereby deliver the insights businesses need to enable evidence-based decision making.
- ❑ Where they differ, data analysis looks at the past, while data analytics tries to predict the future.
- ❑ Analysis is the detailed examination of the elements or structure of something. Analytics is the systematic computational analysis of data.

Importance of Data Analysis

11

- ❑ Data analysis is important to businesses will be an understatement. In fact, no business can survive without analyzing available data. Visualize the following situations:
 - ❑ A pharmacy company is performing trials on number of patients to test its new drug to fight cancer. The number of patients under the trial is well over 500.
 - ❑ A company wants to launch new variant of its existing line of fruit juice. It wants to carry out the survey analysis and arrive at some meaningful conclusion.
 - ❑ Sales director of a company knows that there is something wrong with one of its successful products, however hasn't yet carried out any market research data analysis. How and what does he conclude?
- ❑ These situations are indicative enough to conclude that data analysis is the lifeline of any business. Whether one wants to arrive at some marketing decisions or fine-tune new product launch strategy, data analysis is the key to all the problems.

Importance of Data Analysis cont...

12

- ❑ Merely analyzing data isn't sufficient from the point of view of making a decision. How does one interpret from the analyzed data is more important. Thus, data analysis is not a decision making system, but decision supporting system.
- ❑ Data analysis can offer the following benefits:
 - ❑ Structuring the findings from survey research or other means of data collection.
 - ❑ Break a macro picture into a micro one.
 - ❑ Acquiring meaningful insights from the dataset.
 - ❑ Basing critical decisions from the findings.
 - ❑ Ruling out human bias through proper statistical treatment.

Data Analytics Applications

13

1. **Understanding and targeting customers:** Data analytics is extremely useful to understand and predict customer behaviour. The trend is towards getting a 360 degree view of each customer which includes data from traditional customer purchase data as well as the non-traditional unstructured data sets like social media, web logs, customer clicks on e-retail sites, etc. This picture helps businesses predict which customers may move to a rival (called customer churn), predict what products will sell, predict living patterns that can help insurance companies charge a differential premium and so on. The list is endless with potential business benefits.
2. **Understanding and optimizing business processes:** Information is collected from social media resulting in sentiment analysis. This along with company profiles data are analyzed using data analytics tools to effectively predict demand for products and thus help to retain stock in warehouses to an optimal level. For example, Apple excessively uses sentiment analysis information from social media to gauge the potential sales of their new iPhone 6s offering. By providing a holistic view of assets and business processes, enterprises are now able to gain unparalleled insight into optimizing those assets and processes. For example, a Fortune 50 company was able to optimise accounts receivable collections without increasing collector headcount. This was possible because data analytics tools crawled and mined historical data, identified the factors that affect late payments, provided insights in the collection system, and provided on-going recommendations that helped improved Accounts Receivable collections by 65% over the prior year.

Data Analytics Applications cont...



14

3. **Personal quantification and performance optimization:** Personal quantification is an emerging trend in big data science. Self-quantification of personal health and wellness data are contributing heavily towards more self-managed health care. Advances in network and wearable sensor technologies easily help to capture and share significant health-related information on a daily basis. New functionalities in wearable devices and the associated apps enable individuals to measure vital signs, access analytical tools and quantify data about themselves faster and more ubiquitously than ever before. For example, keeping diaries of food intake, converting these collected data into numbers, analysing them and using them to make better decisions regarding personal health are part of personal quantification.
4. **Improving healthcare and public health:** The healthcare industry historically has massive amounts of data in its archives. This may be due to record keeping, compliance and regulatory requirements, continual patient care, etc. This voluminous data totally renders itself to data analytics applications. This data includes clinical data from hospitals and clinical decision support systems (physician's written notes and prescriptions, medical imaging, laboratory test data, pharmacy prescriptions and sales, insurance data, patient data in electronic patient records (EPRs)) machine generated/sensor data, such as from monitoring vital signs; social media data, news feeds, and articles in medical journals, etc. This data can be analyzed to effectively provide customised medical care to patients, detect epidemics, curtail infections and several such applications. Interesting current application include holistic cancer treatment, genomics, identifying and stopping hospital fraud and the like.

Data Analytics Applications cont...



15

5. **Improving sports performance:** Any sports, be it football, car racing or sailing, gets affected by advances in the capture, storage and analysis of data. Data analytics allows athletes to train better and more effectively and it allows teams to alter their in-game decision-making based on what they are seeing. Like other businesses, sports teams strive to make better decisions faster. Coaching staff, scouts and players are leveraging analytics to better understand the performance of their own teams as well as that of the opposition.

It is also changing the way broadcasters produce sports entertainment and the way teams and broadcasters engage with fans. Big data is also playing an increasingly important role not just in broadcast production, but in broadcast and digital distribution. Multi-platform distribution of content and the use of social media and consumer-generated content in broadcasts uses metrics to determine popular content and are able to offer different pricing models.

Data Analytics Applications cont...



16

6. **Improving science and research:** The emergent field of data analytics is rapidly changing the direction and speed of scientific research by letting people fine-tune their inquiries by tapping into giant data sets. In the past, certain fields of science relied heavily on big data sets, such as high-energy particle physics or research on nuclear fusion. But as information becomes available from more sources, collecting and analysing large amounts of data is becoming common in other fields of research too. One such recent example is the research being conducted at CERN, the Swiss nuclear physics laboratory with its Large Hadron Collider. The CERN data centre has 65,000 processors to analyze its 30 petabytes of data. It uses the computing powers of thousands of computers distributed across 150 data centres worldwide to analyze the data.
7. **Optimizing machine and device performance:** Big data analytics helps machines and devices become smarter and more autonomous. Data analytics can be utilised to collect energy usage data from smart meters, analyze usage patterns and effectively provide smart grids that can optimise energy usage. Looking into usage patterns of machines at large manufacturing plants, analytics can effectively predict machine down time and that helps to perform preventive maintenance resulting in saving of huge amount of resources.

Data Analytics Applications cont...



17

8. **Improving security and law enforcement:** By taking advantage of big data, crime analysts identify trends and make recommendations based on their observations. Through analysis and computer mapping, crime analysts play a crucial role in helping law enforcement agencies quantify, evaluate, and respond to the changing landscape of criminal activity in their jurisdictions. Typical applications use big data techniques to detect and prevent cyberattacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data to detect fraudulent transactions.
9. **Improving and optimizing cities and countries:** Today's towns and cities generate about 5 terabytes per day per square kilometre of urbanized land area. This includes location data collected by smart phones to data generated by GPS instruments, payment cards, smart ID cards, loyalty and store cards, bank cards, toll payments, etc. Further sources of data include data created by traffic management systems, from traffic lights to the sensors on our roads; from the provision of utilities such as gas, electricity and drinking water, etc. All this data can be analyzed to improve many aspects of our cities and citizen's daily life. For example, we can have intelligent route planning systems that are based on real-time traffic information as well as social media and weather data. Smart cities are planned by integrating and analysing all subsystems in a city like energy, traffic, police, etc.

Data Analytics Applications cont...



18

10. **Financial trading:** High-frequency trading is an area where big data finds a lot of use today. Data analytics technologies have advanced sufficiently to provide millisecond latency on large data sets. Here, big data analysis algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that take into account signals from social media networks and news websites, weather predictions, etc. to make, buy and sell decisions in split seconds.

Regression Modelling Techniques



19

- ❑ One of the fundamental task in data analysis is to find how different variables are related to each other and one of the central tool for learning about such relationships is regression.
- ❑ Lets take a simple example: Suppose your manager asked you to predict annual sales. There can be factors (drivers) that affects sales such as competitive pricing, product quality, shipping time & cost, online reviews, easy return policy, loyalty rewards, word of mouth recommendations, ease of checkout etc. In this case, sales is your dependent variable. Factors affecting sales are independent variables.
- ❑ Regression analysis would help to solve this problem. In simple words, regression analysis is used to model the relationship between a dependent variable and one or more independent (predictors) variables and then use the relationships to make predictions about the future.
- ❑ Regression analysis helps to answer the following questions:
 - ❑ Which of the drivers have a significant impact on sales?
 - ❑ Which is the most important driver of sales?
 - ❑ How do the drivers interact with each other?

Regression Modelling Techniques cont...



20

- The regression analysis allows to model the dependent variable as a function of its predictors i.e. $Y = f(X_i, \beta) + e_i$ where Y is dependent variable, f is the function, X_i is the independent variable, β is the unknown parameters, e_i is the error term, and i varies from 1 to n.
- *Terminologies*
 - **Outliers:** Suppose there is an observation in the dataset which is having a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier. In simple words, it is extreme value. An outlier is a problem because many times it hampers the results we get.
 - **Multicollinearity:** When the predictors are highly correlated to each other then the variables are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance or it makes job difficult in selecting the most important independent variable (factor).
 - **Heteroscedasticity:** When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity. Example -As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.

Terminologies cont...

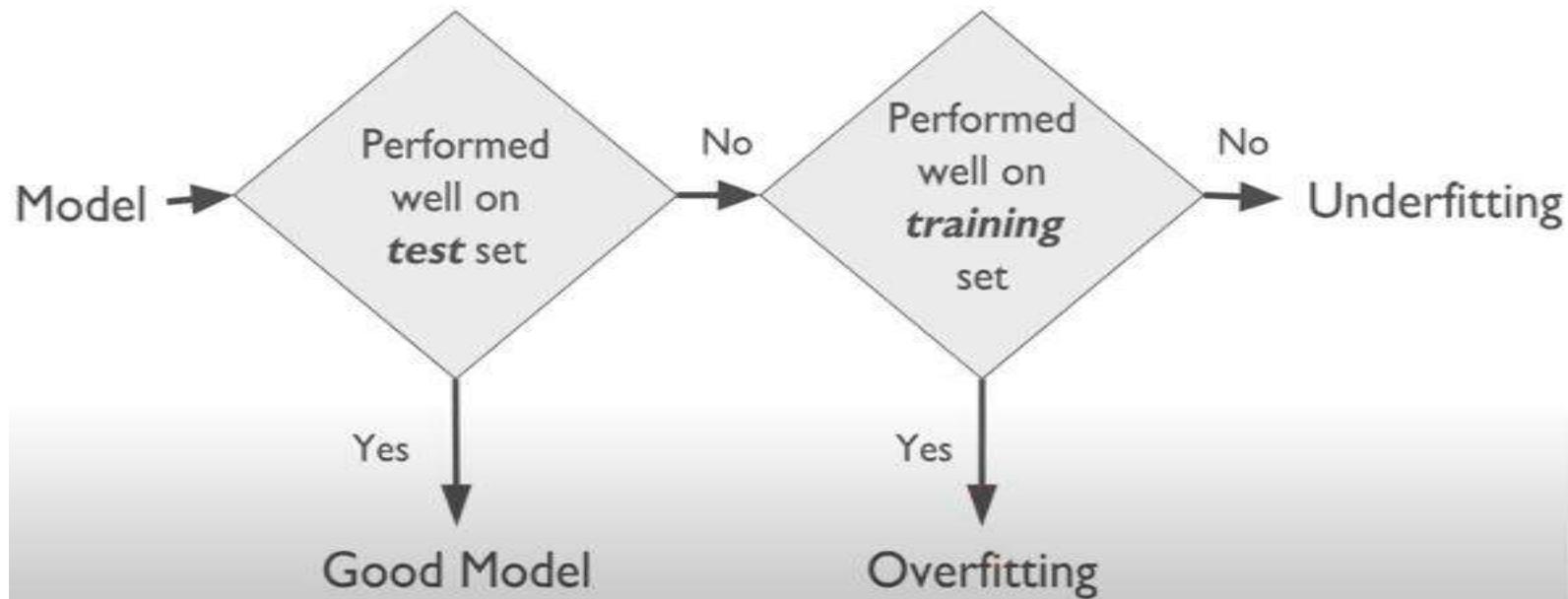
21

- ❑ **Training and Test dataset:** In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. So, training data is used to fit the model and testing data to test it.
- ❑ **Overfitting:** It means that model works well on the training dataset but is unable to perform better on the test datasets. It is also known as problem of **high variance**. Variance indicates how much the estimate of the model will alter if different training data were used.
- ❑ **Underfitting:** When the model works so poorly that it is unable to fit even training set well then it is said to be underfitting the data. It is also known as problem of **high bias**. A bias is the amount that a model's prediction differs from the target value.

Terminologies cont...

22

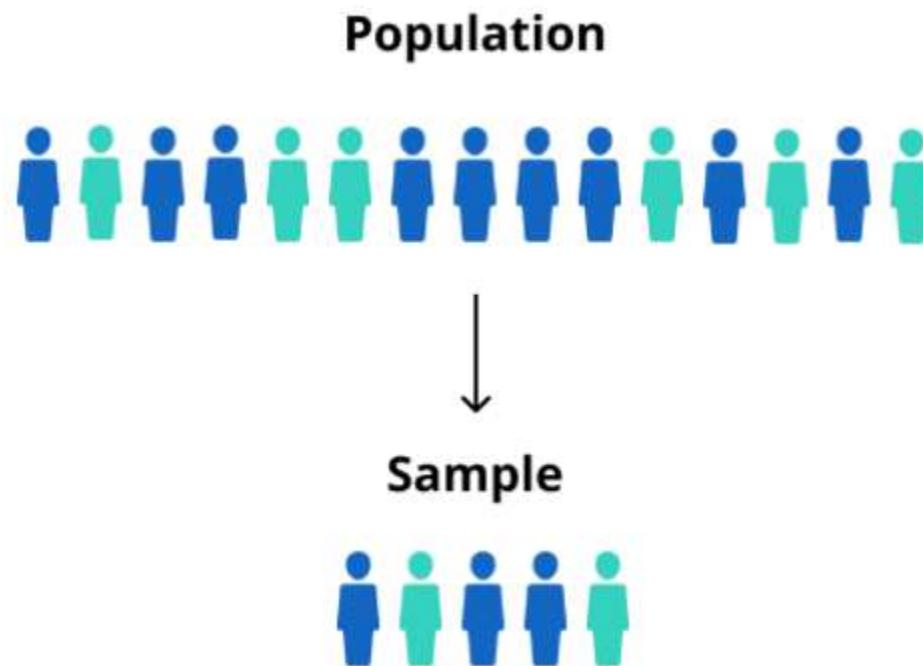
❑ Overfitting and Underfitting



Terminologies cont...

23

- **Sample and population:** A population is the entire group of elements meant to draw conclusions. A sample is a smaller part of the whole, i.e., a subset of the entire population. The size of the sample is always less than the total size of the population.



Terminologies cont...

24

Correlation: Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

- ❑ **A positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
- ❑ **A negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- ❑ **A zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

Terminologies cont... Scattergrams

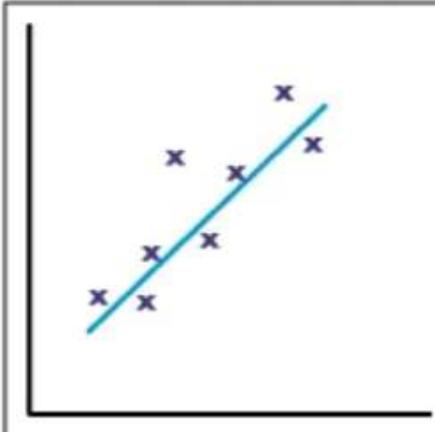
25

- ❑ A correlation can be expressed visually. This is done by drawing a scattergram (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram).
- ❑ A scattergram is a graphical display that shows the relationships or associations between two numerical variables (or co-variables), which are represented as points (or dots) for each pair of score.
- ❑ A scattergraph indicates the strength and direction of the correlation between the co-variables.
- ❑ When you draw a scattergram it doesn't matter which variable goes on the x-axis and which goes on the y-axis.
- ❑ Correlations always deals with paired scores, so the values of the 2 variables should be taken together and used to make the diagram.

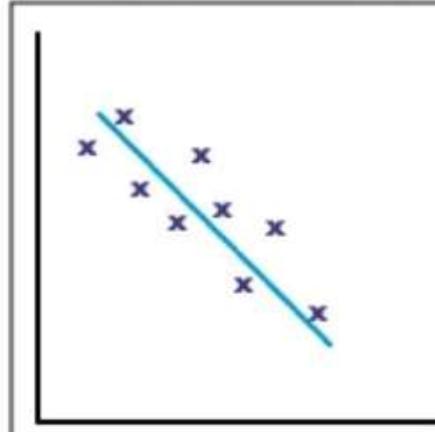
Terminologies cont... Scattergrams

26

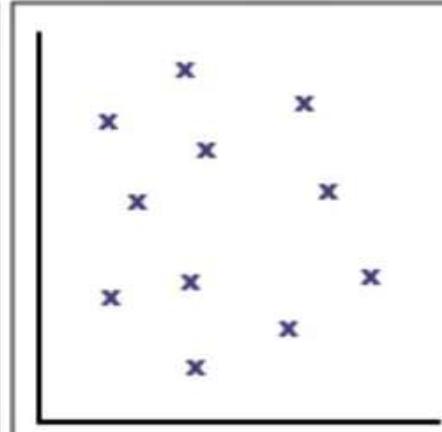
Positive correlation



Negative correlation



No correlation



The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

The points lie close to a straight line, which has a negative gradient.

This shows that as one variable **increases**, the other **decreases**.

There is no pattern to the points.

This shows that there is **no connection** between the two variables.

Terminologies cont... Correlation Coefficients: Determining Correlation Strength



27

- ❑ Instead of drawing a scattergram a correlation can be expressed numerically as a coefficient, ranging from -1 to +1.
- ❑ The correlation coefficient (r) indicates the extent to which the pairs of numbers for these two variables lie on a straight line. Values over zero indicate a positive correlation, while values under zero indicate a negative correlation.
- ❑ A correlation of -1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that as one variable goes up, the other goes up.
- ❑ There is no rule for determining what size of correlation is considered strong, moderate or weak. The interpretation of the coefficient depends on the topic of study. In my research area, I generally consider correlations above 0.4 to be relatively strong; correlations between 0.2 and 0.4 are moderate, and those below 0.2 are considered weak.

Terminologies cont.. Correlation Coefficients

28

- The following formula is normally used to estimate the correlation coefficients between two variables X and Y.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- x is the independent variable and y is the dependent variable.
- n is the number of observations
- r, the computed value is known as the correlation coefficients .

Correlation Coefficients Calculation

29

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

- $n = 10, \sum X = 80, \sum Y = 255, \sum XY = 2289$
- $\sum X^2 = 756, \sum Y^2 = 7097, (\sum X)^2 = 6400, (\sum Y)^2 = 65025, r = 0.95$

Class Exercise

30

Find the correlation coefficients of the below sample.

Subject	Age (x)	Glucose Level (y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Types of Regression

31

- ❑ Every regression technique has some assumptions attached to it which need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution. The types of regression algorithms are:
 - ❑ Linear Regression
 - ❑ Multiple Linear Regression
 - ❑ Non Linear Regression
 - ❑ Logistic Regression
 - ❑ Polynomial Regression
 - ❑ Quantile Regression
 - ❑ Ridge Regression
 - ❑ Lasso Regression
 - ❑ Elastic Net Regression
 - ❑ Principal Components Regression (PCR)
 - ❑ Partial Least Squares (PLS) Regression
 - ❑ Support Vector Regression
 - ❑ Ordinal Regression
 - ❑ Poisson Regression
 - ❑ Negative Binomial Regression
 - ❑ Quasi Poisson Regression
 - ❑ Cox Regression

Linear Regression

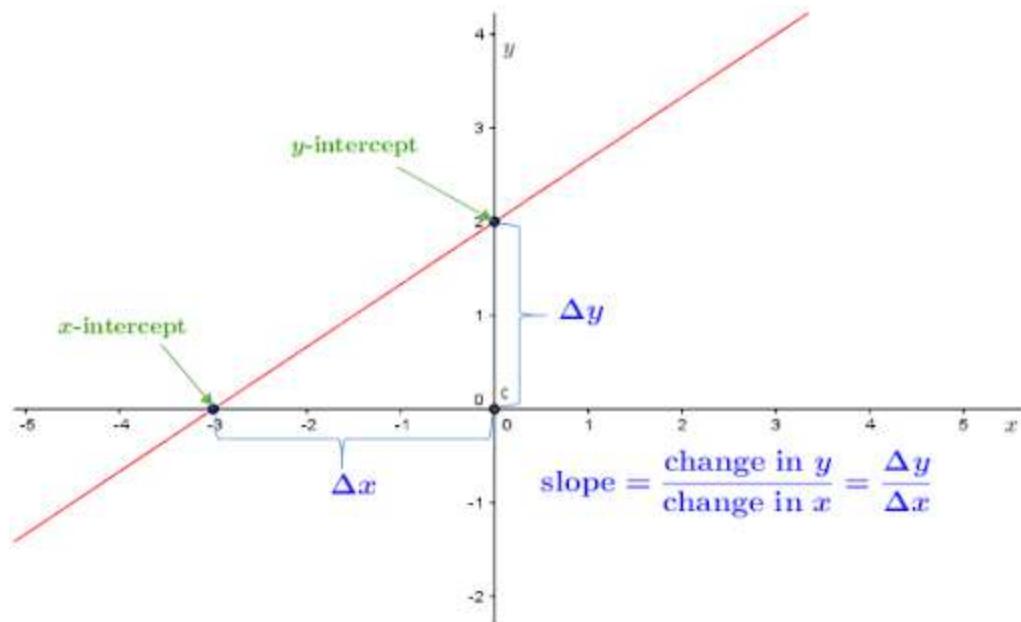
32

- ❑ Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory (independent) variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
- ❑ Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables.
- ❑ A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.
- ❑ If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

Linear Regression cont...

33

- ❑ A linear regression line has an equation of the form $\mathbf{Y} = \mathbf{a} + \mathbf{bX} + \mathbf{e}$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, a is the intercept (the value of y when x = 0), and e is the random error.
- ❑ The slope and intercept is as follows in the following linear equation of line:



- ❑ In the above snap, a is considered as the y-intercept.

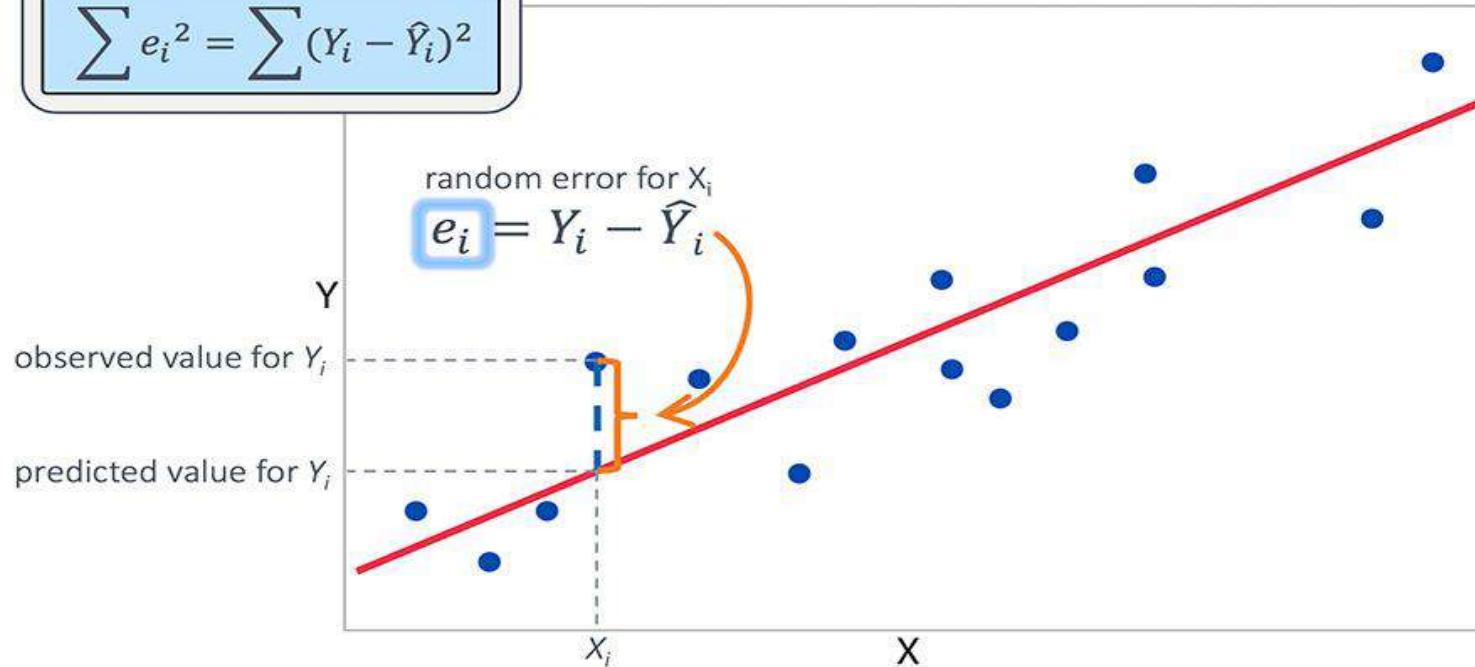
Linear Regression cont...

34

- The random error in the following linear equation of line:

Method of Least Squares

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

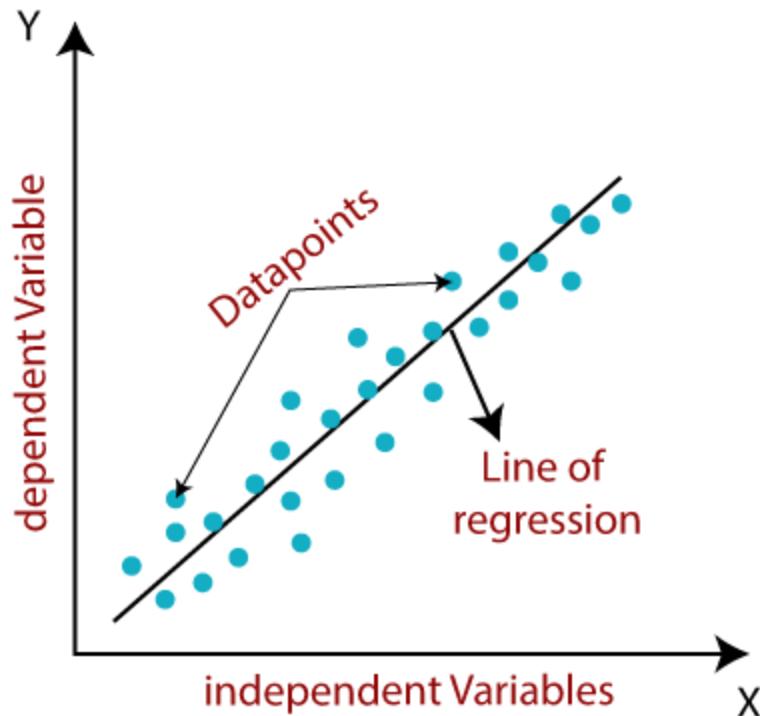


- To fit the regression line, a statistical approach known as least squares method.

Linear Regression cont...

35

- The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Linear Regression cont...

36

- The calculation of b and a is as follows:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y}{n} - b \cdot \frac{\sum X}{n}$$

- If $b > 0$, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.
- If $b < 0$, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.
- If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^n (\text{actual_output} - \text{predicted_output}) ** 2$$

Linear Regression cont...

37

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$$b = \frac{10 \times 2289 - (80 \times 255)}{[10 \times 756 - (80)^2]} = 2.1466; \quad a = \frac{255}{10} - 2.1466 \frac{80}{10} = 8.3272$$

Linear Regression cont...

38

- ❑ The linear regression will thus be Predicted (Y) = 8.3272 + 2.1466 X
- ❑ The above equation can be used to predict the volume of sales for an insurance company given its agent number. Thus if a company has 1000 agents (10 hundreds) the predicted value of sales will be around ?
- ❑ In summary, linear regression consists of the following steps:
 - ❑ Collection of sample of independent and dependent variable.
 - ❑ Compute b and a.
 - ❑ Use these values to formulate the linear regression equation.
 - ❑ Given the new values for X predict the value of Y.
- ❑ Larger and better the sample of data, more accurate would be the regression model and would lead to more accurate forecasts.

Multiple Linear Regression

39

- ❑ Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression.
- ❑ Example:
 - ❑ Do age and intelligence quotient (IQ) scores predict grade point average (GPA)?
 - ❑ Do weight, height, and age explain the variance in cholesterol levels?
 - ❑ Do height, weight, age, and hours of exercise per week predict blood pressure?
- ❑ The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \dots + \beta_nx_n + e$$

where, y = the predicted value of the dependent variable.

β_0 = the y -intercept (value of y when all other parameters are set to 0)

β_1x_1 = the regression coefficient (β_1) of the first independent variable (x_1)

β_nx_n = the regression coefficient (β_n) of the last independent variable (x_n)

e = model error

Multiple Linear Regression with Two Independent Variables

40

- The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

where, y = the predicted value of the dependent variable.

β_0 = the y -intercept (value of y when all other parameters are set to 0)

$\beta_1 x_1$ = the regression coefficient (β_1) of the first independent variable (x_1)

$\beta_2 x_2$ = the regression coefficient (β_2) of the second independent variable (x_2)

e = model error

- β_1 and β_2 is calculated as follows:

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

- β_0 is calculated as follows:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

where $\bar{y} = \frac{\sum Y}{n}$ and $\bar{x}_i = \frac{\sum X_i}{n}$

Non-Linear Regression

41

- ❑ In the case of linear and multiple linear regression, the dependent variable is linearly dependent on the independent variable(s). But, in several situations, the situation is no simple where the two variables might be related in a non-linear way.
- ❑ This may be the case where the results from the correlation analysis show no linear relationship but these variables might still be closely related.
- ❑ If the result of the data analysis show that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model.
- ❑ The non-linear data can be handled in 2 ways:
 - ❑ Use of polynomial rather than linear regression model
 - ❑ Transform the data and then use linear regression model.

Non-Linear Regression cont...

42

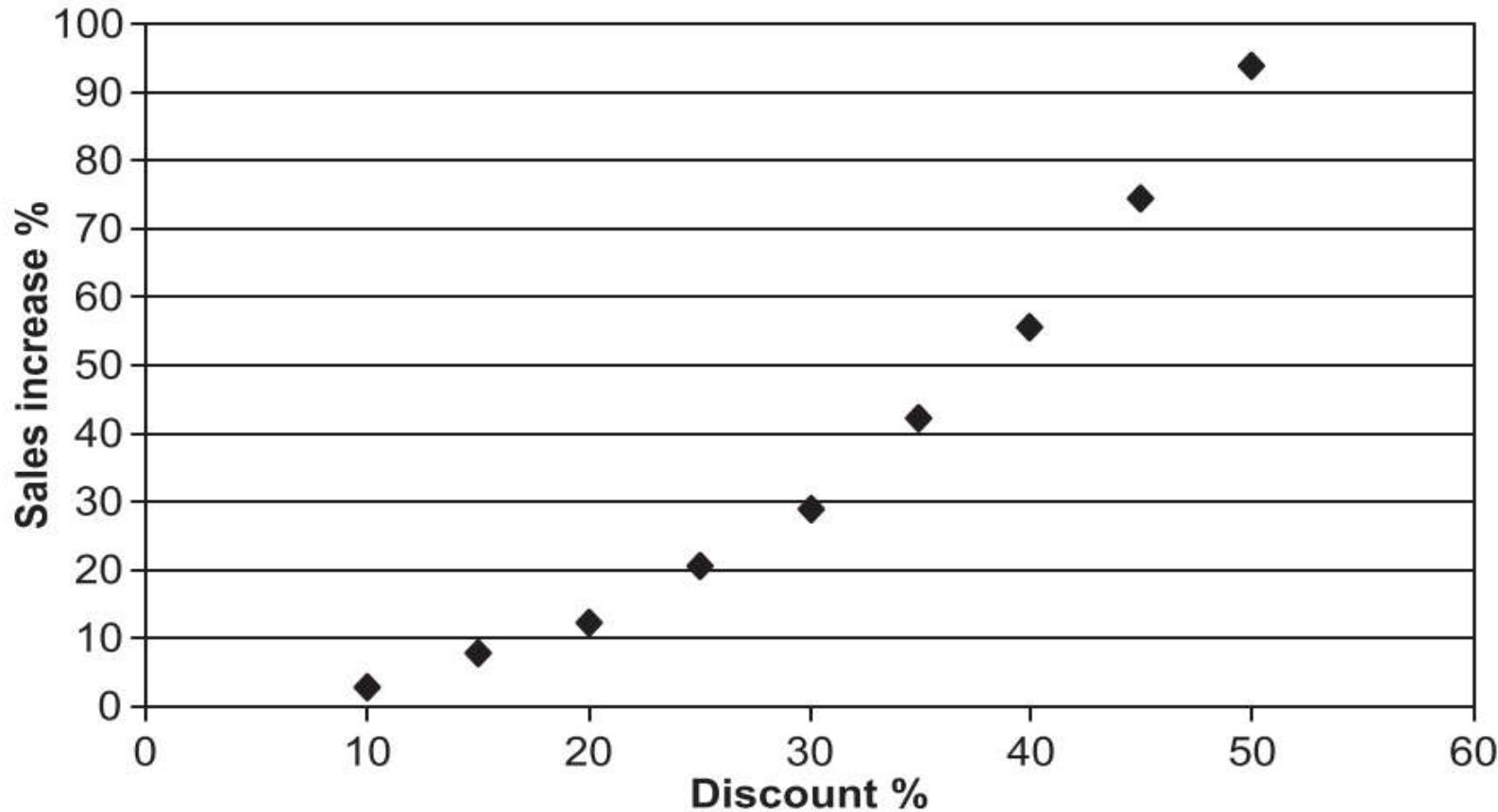
Product	Increase in sale in % (Y)	Discount in % (X)
A	3.05	10
B	7.62	15
C	12.19	20
D	20.42	25
E	28.65	30
F	42.06	35
G	55.47	40
H	74.68	45
I	93.88	50

Non-Linear Regression cont...



43

The scatter diagram of sales increase for various discount percentage looks as follows:



The value of r is 0.97 which indicates a very strong, almost perfect, positive correlation, and the data value appears to form a slight curve.

Non-Linear Regression cont...

44

Polynomials are the equations that involve powers of the independent variables. A second degree (quadratic), third degree (cubic), and n degree polynomial functions:

- Second degree: $y = \beta_0 + \beta_1x + \beta_2x^2 + e$
- Third degree: $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + e$
- n degree: $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \dots + \beta_nx^n + e$

Where:

- β_0 is the intercept of the regression model
- $\beta_1, \beta_2, \beta_3$ are the coefficient of the predictors.

How to find the right degree of the equation?

As we increase the degree in the model, it tends to increase the performance of the model. However, increasing the degrees of the model also increases the risk of over-fitting and under-fitting the data. So, one of the approach can be adopted:

- Forward Selection:** This method increases the degree until it is significant enough to define the best possible model.
- Backward Elimination:** This method decreases the degree until it is significant enough to define the best possible model.

Non-Linear Regression cont...



45

- ❑ The techniques of fitting of the polynomial model in one variable can be extended to the fitting of polynomial models in two or more independent variables.
- ❑ A second-order polynomial is more used in practice, and its model with two independent variables is specified by: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + e$
- ❑ This is also termed as **response surface**. The methodology of response surface is used to fit such models and helps in designing an experiment. This type is generally covered in the topics in the design of experiment.

Class work

- ❑ Define the second-order polynomial model with two independent variables.
- ❑ Define the second-order polynomial model with three independent variables.
- ❑ Define the third-order polynomial model with two independent variables.

Non-Linear Regression cont...



46

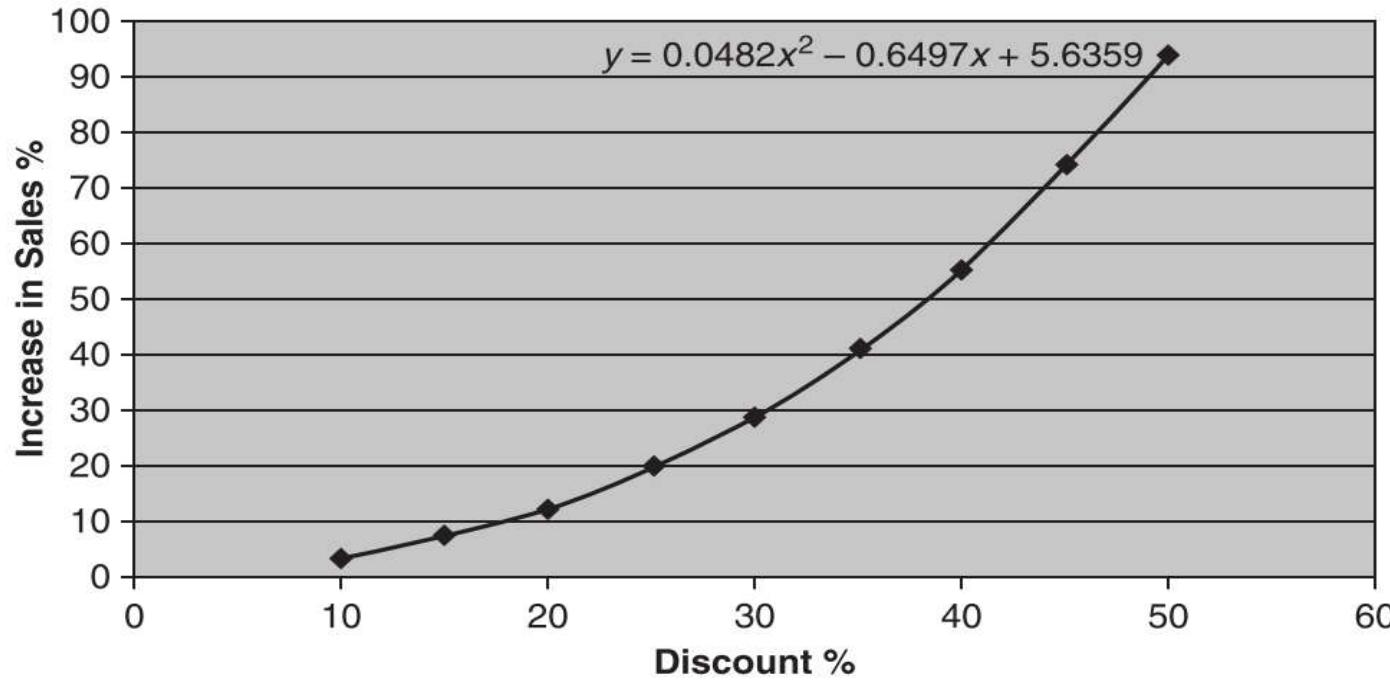
A polynomial regression is regression that involves multiple powers of predictor(s). So, regression tools and diagnostics can be applied to polynomial regression.

Non-Linear Regression cont...



47

- ❑ The tools exists in software such as SAS, Excel or the language such as Python, R can estimate the value of coefficients of predictor such as β_0, β_1 etc and to fit a curve in a non-linear fashion for the given data.
- ❑ Following figure depicts the graph of increase in sale vs. discount.

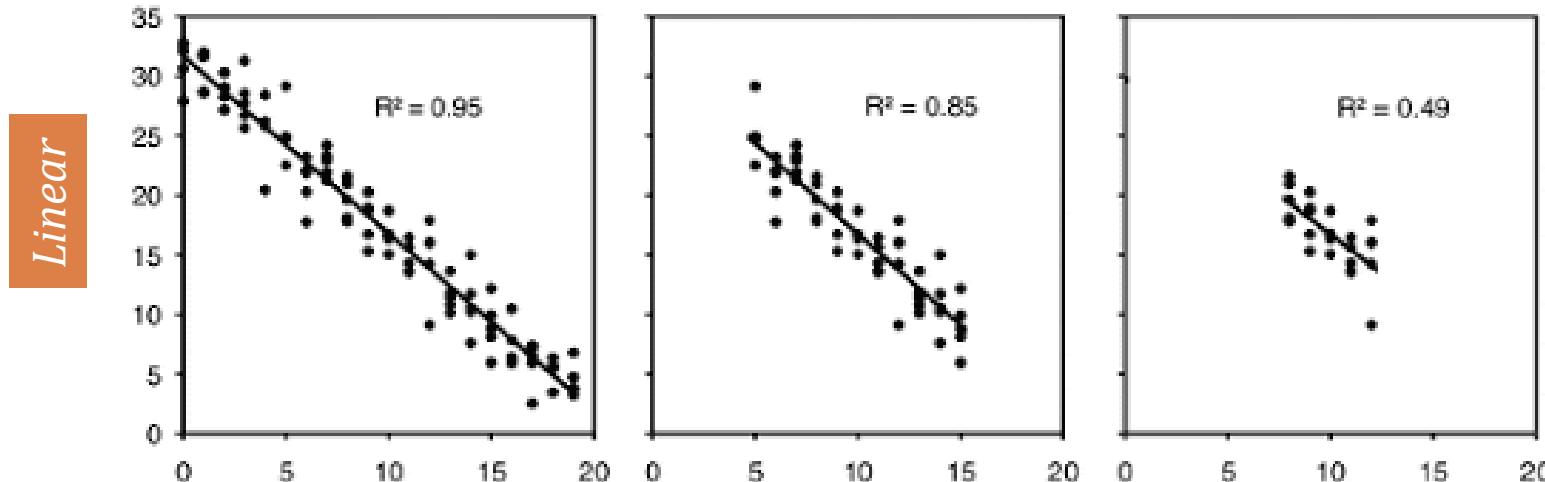


The predicted model is $Y = 5.6359 - 0.6497 x + 0.0482 x^2$

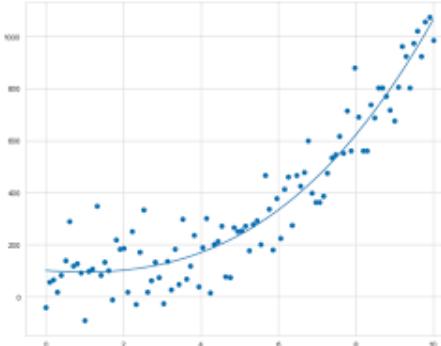
Non-Linear Regression cont...

48

R^2 is known as coefficient of determination and it's a number that indicates how well the data fits into the developed model i.e. a line or curve.



Curve



The figure shows a scatter plot with blue data points forming a parabolic shape. A red curve is drawn through the points, representing a quadratic regression model. The x-axis ranges from 0 to 10, and the y-axis ranges from 0 to 10. The data points are approximately at (0,1), (1,3), (2,1), (3,5), (4,3), (5,7), (6,5), (7,9), (8,7), (9,11), (10,9).

Non-Linear Regression cont...

49

- An R^2 of 1 indicates that the regression model perfectly fits the data while an R^2 of 0 indicate that model does not fit the data at all.
- An R^2 is calculated as follows:

$$\text{Sum of Squares Regression (SSR)} = \sum(\hat{Y}_i - \bar{Y}_i)^2$$

$$\text{Sum of Squares Error (SSE)} = \sum(\hat{Y}_i - Y_i)^2$$

$$\text{Sum of Squares Total (SST)} = \sum(Y_i - \bar{Y})^2$$

$$SST = SSR + SSE$$

$$R^2 = 1 - \frac{SSE}{SST}$$

where

\bar{Y} is the mean of the actual values of Y

\hat{Y}_i is predicted values of Y_i .

Non-Linear Regression cont...



50

- In the example, a value of 0.99 for R^2 indicates that a quadratic model is good fit for the data.
- Another preferable way to perform non-linear regression is to try to transform the data in order to make the relationship between the two variables more linear and then use a regression model rather than a polynomial one. Transformations aim to make a non-linear relationship between two variables more linear so that it can be described by a linear regression model.
- Three most popular transformations are the:
 - Square root (\sqrt{X})
 - Logarithm ($\log X$)
 - Negative reciprocal ($-1/X$)

Non-Linear Regression cont...



51

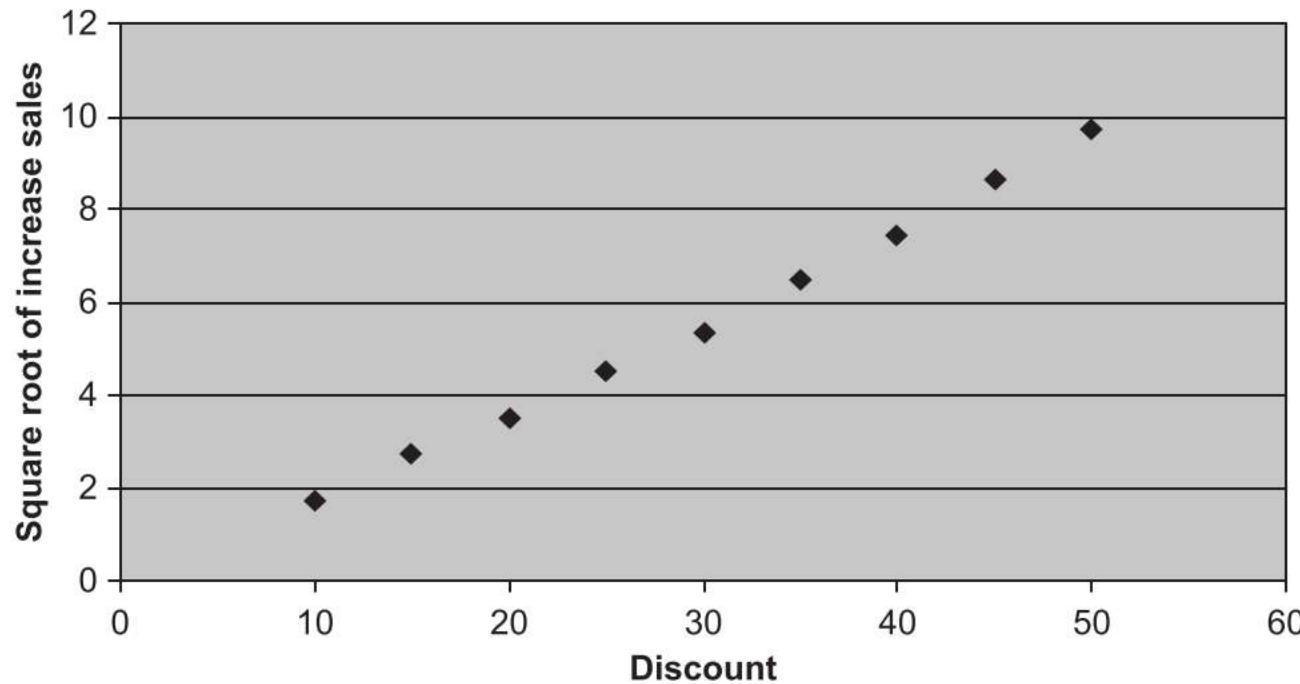
Application of Square root (\sqrt{Y})

Product	Discount in % (X)	Increase in sale in % (Y)	SQRT (Y)
A	10	3.05	$\sqrt{3.05} = 1.75$
B	15	7.62	$\sqrt{7.62} = 2.76$
C	20	12.19	$\sqrt{12.19} = 3.49$
D	25	20.42	$\sqrt{20.42} = 4.52$
E	30	28.65	$\sqrt{28.65} = 5.35$
F	35	42.06	$\sqrt{42.06} = 6.49$
G	40	55.47	$\sqrt{55.47} = 7.45$
H	45	74.68	$\sqrt{74.68} = 8.64$
I	50	93.88	$\sqrt{93.88} = 9.69$

Non-Linear Regression cont...

52

Square root of transformation



In the similar fashion, Logarithm and negative reciprocal techniques can be applied to the dependent variable followed up by the application of linear regression model.

Logistic Regression

53

- ❑ Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications.
- ❑ Logistic Regression is used when the dependent variable (target) is categorical. For example:
 - ❑ To predict whether an email is spam (1) or not (0). If the model infers a value of 0.932 on a particular email message, it implies a 93.2% probability that the email message is spam. The model predicts the email message is spam 93.2% of the time and the remaining 6.8% will not.
 - ❑ Whether the tumor is malignant (1) or not (0)
- ❑ There are 3 types of Logistic Regression
 - ❑ **Binary Logistic Regression:** The categorical response has only two possible outcomes. Example: Spam or Not.
 - ❑ **Multinomial Logistic Regression:** Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
 - ❑ **Ordinal Logistic Regression:** Three or more categories with ordering. Example: Movie rating from 1 to 5.

Logistic Regression cont...

54

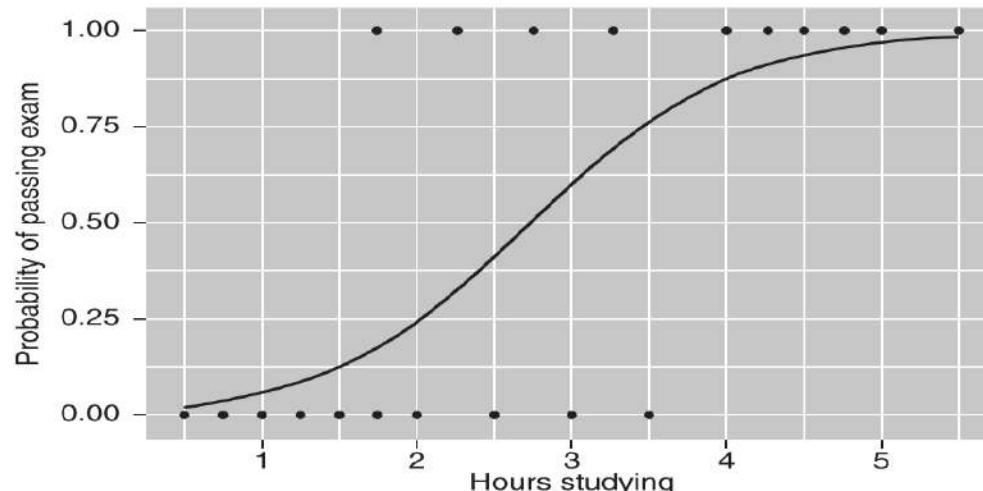
- ❑ It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function which is the cumulative logistic distribution.
- ❑ Since the predicted values are probabilities and therefore are restricted to (0, 1), a logistic regression model **only predicts the probability of particular outcome given the values of the existing data.**
- ❑ **Example:** A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam? The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used. The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Logistic Regression cont...

55

- ❑ In logistic regression, we don't directly fit a straight line to the data like in linear regression. Instead, we fit a S shaped curve, called **sigmoid** or **logistic regression curve**. A logistic regression curve showing probability of passing an exam versus hours studying is shown below.
- ❑ Y-axis goes from 0 to 1. This is because the sigmoid function always takes as maximum (i.e. 1) and minimum (i.e. 0), and this fits very well to the goal of classifying samples in two different categories (fail or pass).
- ❑ The sigmoid function is $\text{sigmoid}(x) = 1 / (1 + e^{-x})$ where x is the weighted sum of independent variable i.e. $x = \beta_0 + \beta_1 x_i$ where i is the individual independent variable instance.



Logistic Regression cont...

56

Consider a model with one predictor X_1 , and one binary response variable Y , which we denote $p = P(Y = 1 | X_1 = x)$, where p is the probability of success. p should meet criteria: (i) it must always be positive, (ii) it must always be less than or equal to 1.

We assume a linear relationship between the independent variable and the logit of the event i.e. $Y = 1$. In statistics, the logit is the logarithm of the **odds** i.e. $p / (1-p)$. This linear relationship can be written in the following mathematical form (where ℓ is the logit, b is the base of the logarithm, and β is the parameter of the model).

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

The odds can be recovered by exponentiation of the logit:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1} \rightarrow p = \frac{b^{\beta_0 + \beta_1 x_1}}{b^{\beta_0 + \beta_1 x_1} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1)}} = S_b(\beta_0 + \beta_1 x_1)$$

Where S_b is the sigmoid function with base b . However in some cases it can be easier to communicate results by working in base 2, base 10, or exponential constant e . In reference to the students example, solving the equation with software tool and considering base as e , the coefficient is $\beta_0 = -4.0777$ and $\beta_1 = 1.5046$

Logistic Regression cont...

57

- For example, for a student who studies 2 hours, entering the value Hours = 2 in the equation gives the estimated probability of passing the exam of 0.26.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 2 - 4.0777))} = 0.26$$

- Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 4 - 4.0777))} = 0.87$$

- Following table shows the probability of passing the exam for several values of hours studying.

Hours of study	Probability of passing the exam
1	0.07
2	0.26
3	0.61
5	0.97

Bayesian Modelling

58

- ❑ Bayesian data analysis deals with set of practical methods for making inferences from the available data. The methods used probability models to model the given data and also predict future values.
- ❑ Thus, essentially it is a statistical paradigm that answers research questions about unknown parameters using probability statements. For example, what is the probability that the average male height is between 70 and 80 inches or that the average female height is between 60 and 70 inches?
- ❑ Bayesian data analysis consists of 3 important steps:
 - ❑ **Setting up the prior distribution:** Using domain expertise or prior knowledge to develop a **joint probability distribution** for all independent variable of the data under consideration and also the dependent variable. This terms as prior distribution.
 - ❑ **Setting up the posterior distribution:** After taking into the account the observed data, calculate and interpret the appropriate posterior distribution. This is estimating the **conditional probability distribution** of the data parameters, given the observed data.
 - ❑ **Evaluating the fit of the model:** This is to seek answer for the questions: how well does the developed model fit the data? Are the conclusion reasonable? How sensitive are the results to the modelling assumptions (as per step 1)? the In response, the model can be altered or expanded with the three steps.

Joint Probability

59

- ❑ A statistical measure that calculates the likelihood of two events occurring together and at the same point in time is called Joint probability.
- ❑ Let A and B be the two events, joint probability is the probability of event B occurring at the same time that event A occurs.
- ❑ The formula $P(A \cap B)$ represents the joint probability of events with intersection, where, A and B are the two events. The symbol “ \cap ” in a joint probability is called an intersection. The probability of event A and event B happening is the same thing as the point where A and B intersect. Hence, the joint probability is also called the intersection of two or more events.
- ❑ **Example:** Find the probability that the number three will occur twice when two dice are rolled at the same time.

Solution: Number of possible outcomes when a dice is rolled = 6 i.e. {1, 2, 3, 4, 5, 6}. Let A be the event of occurring 3 on first dice and B be the event of occurring 3 on the second dice. Both the dice have six possible outcomes, the probability of a three occurring on each die is $1/6$. $P(A) = 1/6$, $P(B) = 1/6$ and $P(A \cap B) = 1/6 \times 1/6 = 1/36$.

Conditional Probability

60

- ❑ Conditional probability is the probability of one thing being true given that another thing is true. This is distinct from joint probability, which is the probability that both things are true without knowing that one of them must be true.
- ❑ For example, one joint probability is "the probability that your left and right socks are both black," whereas a conditional probability is "the probability that your left sock is black if you know that your right sock is black,"
- ❑ Event A is that it is raining outside, and it has a 0.3 (30%) chance of raining today. Event B is that you will need to go outside, and that has a probability of 0.5 (50%). A conditional probability would look at these two events in relationship with one another, such as the probability that it is both raining and you will need to go outside. The formula for conditional probability is: $P(B|A) = P(A \cap B) / P(A)$
- ❑ **Example:** In a group of 100 sports car buyers, 40 bought alarm systems, 30 purchased bucket seats, and 20 purchased an alarm system and bucket seats. If a car buyer chosen at random bought an alarm system, what is the probability they also bought bucket seats?

Bayesian Interface

61

- ❑ Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.
- ❑ It is the process of fitting a probability model to a set of data which results in a probability distribution on the parameters of the model.
- ❑ The model is then extended to new unobserved data and thus makes predictions for new observations.
- ❑ Axioms of probability:
 - ❑ **Sum Rule:** The sum rule is $P(A + B) = P(A) + P(B)$ where A and B are each events that could occur, but cannot occur at the same time. **Example:** The probability that the next person walking into class will be a student and the probability that the next person will be a teacher. If the probability of the person being a student is 0.8 and the probability of the person being a teacher is 0.1, then the probability of the person being either a teacher or student is $0.8 + 0.1 = 0.9$.
 - ❑ **Product Rule:** The product rule is $P(E \cdot F) = P(E) * P(F)$ where E and F are events that are independent. **Example:** When picking cards from a deck of 52 cards, the probability of getting an ace is $4/52 = 1/13$, because there are 4 aces among the 52 cards. The probability of picking a heart is $13/52 = 1/4$, because there are 13 hearts among the 52 cards.. The probability of picking the ace of hearts is $1/4 * 1/13 = 1/52$.
 - ❑ **Not Rule:** The not rule is $P(\bar{A}) = 1 - P(A)$ where A is an event.

Bayesian Modelling cont...

62

Bayesian methods are based on three important theories in probability:

- Bayes Theorem Law of total probability Normalization

Bayes Theorem

Mathematically Bayes' theorem is defined as:

$$P(A | B) = (P(B | A) * P(A)) / P(B)$$

where A and B are events, $P(A|B)$ is the conditional probability that event A occurs given that event B has already occurred. $P(B|A)$ has the same meaning but with the roles of A and B reversed, and $P(A)$ and $P(B)$ are the marginal probabilities of event A and event B occurring respectively.

Example: There are 52 cards in the pack, 26 of them are red and 26 are black. What is the probability of the card being a 4 given that we know the card is red? Event A is the event that the card picked is a 4 and event B is the card being red. Hence, $P(A|B)$ in the equation above is $P(4|\text{red})$.

$$\begin{aligned}P(A) &= P(4) = 4/52 = 1/13, \\P(B) &= P(\text{red}) = 26/52 = 1/2, \\P(B|A) &= P(\text{red}|4) = 1/2 \\P(4|\text{red}) &= (1/2 * 1/13) / (1/2) = 1/13\end{aligned}$$

Law of total probability

63

The rule states that if the probability of an event is unknown, it can be calculated using the known probabilities of several distinct events. Consider the image:



There are three events: A, B, and C. Events B and C are distinct from each other while event A intersects with both events. We do not know the probability of event A. However, we know the probability of event A under condition B and the probability of event A under condition C. The total probability rule states that by using the two conditional probabilities, we can find the probability of event A. Mathematically, the total probability rule can be written in the following equation where n is the number of events and B_n is the distinct event.

$$P(A) = \sum_n P(A \cap B_n) \quad \text{where, } P(A \cap B) = P(A|B) \times P(B)$$

Law of total probability cont...



64

As per the diagram, the total probability of event A from the situation can be found using the equation is : $P(A) = P(A \cap B) + P(A \cap C)$.

Example: You are a stock analyst following ABC Corp. You discovered that the company is planning to launch a new project that is likely to affect the company's stock price. You have identified the following probabilities:

- There is a 60% probability of launching a new project.
- If a company launches the project, there is a 75% probability that its stock price will increase.
- If a company does not launch the project, there is a 30% probability that its stock price will increase.

You want to find the probability that the company's stock price will increase.

Solution:

$$P(\text{Launch a project} \mid \text{Stock price increases}) = 0.6 \times 0.75 = 0.45$$

$$P(\text{Do not launch} \mid \text{Stock price increases}) = 0.4 \times 0.3 = 0.12$$

$$P(\text{Stock price increases}) = P(\text{Launch a project} \mid \text{Stock price increases}) + P(\text{Do not launch} \mid \text{Stock price increases}) = 0.45 + 0.12 = 0.57. \text{ Thus, there is a 57\% probability that the company's share price will increase.}$$

Law of total probability cont...



65

Class Exercise 1: A person has undertaken a mining job. The probabilities of completion of job on time with and without rain are 0.42 and 0.90 respectively. If the probability that it will rain is 0.45, then determine the probability that the mining job will be completed on time.

Solution: ?

Further study: The Total Probability Rule and Decision Tree @

<https://corporatefinanceinstitute.com/resources/knowledge/other/total-probability-rule/>

Class Exercise 2: An airport screens bags for forbidden items, and an alarm is supposed to be triggered when a forbidden item is detected. Suppose that 5 percent of bags contain forbidden items. If a bag contains a forbidden item, there is a 98 percent chance that it triggers the alarm. If a bag doesn't contain a forbidden item, there is an 8 percent chance that it triggers the alarm. Given a randomly chosen bag triggers the alarm, what is the probability that it contains a forbidden item? Draw the decision tree.

Solution: ?

Normalization

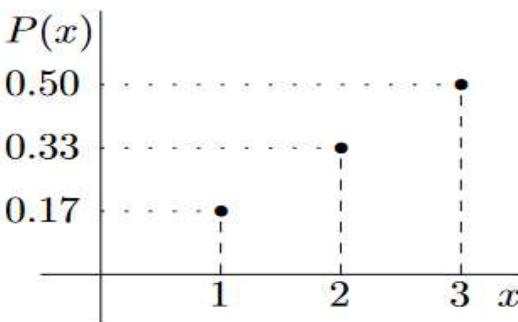
66

A probability distribution function is said to be “normalized” if the sum of all its possible results is equal to **one**.

Example: Let's assume that we have a dice with 6 sides. One side is marked 1, two sides are marked 2, and three sides are marked 3. Since there are six total sides, this means that the probability of rolling each number is as shown below:

Number Rolled	Probability
1	$1/6 \approx 0.17$
2	$2/6 = 1/3 \approx 0.33$
3	$3/6 = 1/2 = 0.50$

It can also be represented as a graph of values versus probabilities



Class Exercise

If you roll this dice 25 times, about how many times will you expect to get each value (1, 2, and 3)?

Bayesian Interface cont...

67

Bayesian interface computes the posterior probability according to Bayes' theorem:

$$P(H | E) = \frac{P(H) * P(E | H)}{P(E)}$$

where:

- ❑ H is the hypothesis whose probability is affected by data.
- ❑ E is the evidence i.e. the unseen data which was not used in computing the prior probability
- ❑ $P(H)$ is the prior probability i.e. it is the probability of H before E is observed
- ❑ $P(H | E)$ is the posterior probability i.e. the probability of H given E and after E is observed.
- ❑ $P(E | H)$ is the probability of observing E given H. It indicates the compatibility of the evidence with the given hypothesis.
- ❑ $P(E)$ is the marginal likelihood or model evidence.

Bayesian Interface cont...

68

Bayes' theorem also can be written as: $P(H | E) = (P(H) * P(E | H)) * \lambda$, where $\lambda = 1 / P(E)$ and is the normalizing constant ensuring that $P(H | E)$ sums to 1 for each state of E .

Class Exercise

Consider the use of online dating sites by age group:

	18-29	30-49	50-64	65+	Total
Used online dating site	60	86	58	21	225
Did not use online dating site	255	426	450	382	1513
Total	315	512	508	403	1738

Based on above table,

1. What is the probability that an 18-29 year old uses online dating sites?
2. What is the probability that 65+ year old do not uses online dating sites?
3. What is the probability that an 18-29 year and 30-49 year old uses online dating sites?
4. What is the probability that an 18-29 year and 30-49 and 50-64 year old uses online dating sites?

Bayesian Model – Naïve Bayes Classifier

69

- **Naive Bayes classifiers (NBC)** are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. *every pair of features being classified is independent of each other.*
- Consider the problem of playing golf, and the dataset is shown on right.
- The need is to classify whether the day is suitable for playing golf, given the features of the day. The columns represent these features and the rows represent individual entries.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naïve Bayes Classifier cont...



70

- If we take the first row of the dataset, we can observe that it is not suitable for playing golf if the outlook is rainy, temperature is hot, humidity is high and it is not windy.
- We make two assumptions. The first, these predictors are independent i.e. if the temperature is hot, it does not necessarily mean that the humidity is high. Another assumption is that all the predictors have an equal effect on the outcome i.e. the day being windy does not have more importance in deciding to play golf or not.
- The Bayes theorem can be written as $P(y|X) = (P(X|y) * P(y)) / P(X)$ where the variable y is the dependent variable(play golf), which represents if it is suitable to play golf or not given the conditions. X represent the independent variables (outlook, temperature, humidity and windy).
- X is given as $X = (x_1, x_2, \dots, x_n)$ where represent the feature and mapped to outlook, temperature, humidity and windy.
- By substituting for X and expanding using the chain rule we get:

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)}$$

- Now, the values for each can be obtained by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remains static. Therefore, the denominator can be removed and a proportionality can be introduced.

Naïve Bayes Classifier cont...



71

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y)$$

- In the example, the class variable(y) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, the need is to find the class y with maximum probability.

$$y = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i|y)$$

- Using the above function, we can obtain the class, given the predictors.
- $P(Y) = 9/14$ and $P(N) = 5/14$ where Y stands for Yes and N stands for No.
- The outlook probability is: $P(\text{sunny} | Y) = 2/9$, $P(\text{overcast} | Y) = 4/9$, $P(\text{rain} | Y) = 3/9$, $P(\text{sunny} | N) = 3/5$, $P(\text{overcast} | N) = 0$, $P(\text{rain} | N) = 2/5$
- The temperature probability is: $P(\text{hot} | Y) = 2/9$, $P(\text{mild} | Y) = 4/9$, $P(\text{cool} | Y) = 3/9$, $P(\text{hot} | N) = 2/5$, $P(\text{mild} | N) = 2/5$, $P(\text{cool} | N) = 1/5$

Naïve Bayes Classifier cont...



72

- ❑ The humidity probability is: $P(\text{high} \mid Y) = 3/9$, $P(\text{normal} \mid Y) = 6/9$, $P(\text{high} \mid N) = 4/5$, $P(\text{normal} \mid N) = 2/5$.
- ❑ The windy probability is: $P(\text{true} \mid Y) = 3/9$, $P(\text{false} \mid Y) = 6/9$, $P(\text{true} \mid N) = 3/5$, $P(\text{false} \mid N) = 2/5$
- ❑ Now we want to predict “Enjoy Sport” on a day with the conditions: <outlook = sunny; temperature = cool; humidity = high; windy = strong>
- ❑ $P(Y) P(\text{sunny} \mid Y) P(\text{cool} \mid Y) P(\text{high} \mid Y) P(\text{strong} \mid Y) = .005$ and $P(N) P(\text{sunny} \mid N) P(\text{cool} \mid N) P(\text{high} \mid N) P(\text{strong} \mid N) = .021$
- ❑ Since, the probability of No is the larger, we can predict “Enjoy Sport” to be No on that day.

Types of Naive Bayes Classifier

- ❑ **Multinomial Naive Bayes:** This is mostly used for document classification problem, i.e. whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
- ❑ **Bernoulli Naive Bayes:** This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
- ❑ **Gaussian Naive Bayes:** The predictors take up a continuous value and are not discrete.

Naïve Bayes Classifier cont...



73

Pros

- ❑ It is easy and fast to predict class of test data set. It also performs well in multi class prediction.
- ❑ When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression and you need less training data.
- ❑ It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons

- ❑ The assumption of independent predictors. In real life, it is almost impossible to get a set of predictors which are completely independent.
- ❑ If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

Time Series Analysis

74

- ❑ Whether to predict the trend in financial markets or electricity consumption, time is an important factor that must be considered in the model. For example, it would be interesting to forecast at what hour during the day there going to be a peak consumption in electricity, such as to adjust the price or the production of electricity.
- ❑ A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. As the name suggests, it involves working on time (years, weeks, days, hours, minutes) based data, to derive hidden insights to make informed decision making.
- ❑ **Example of Time Series Data:**

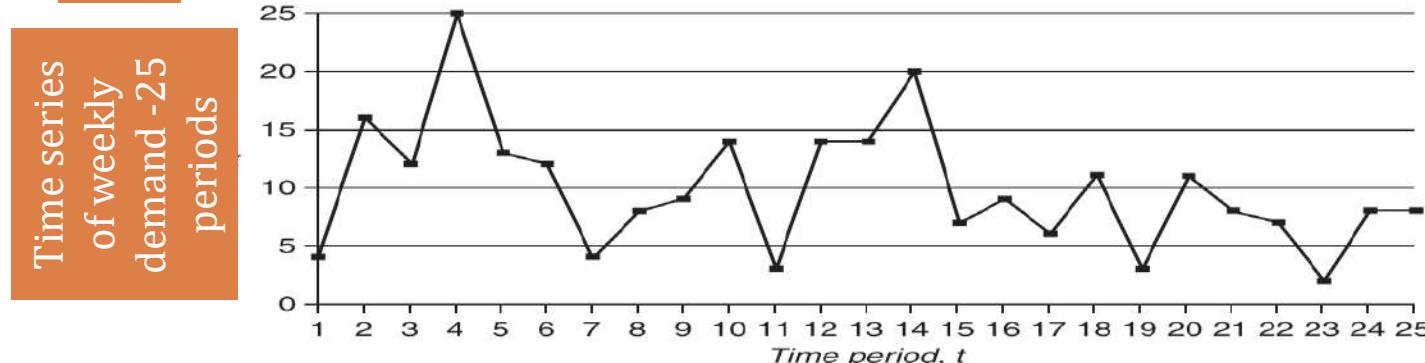
Field	Example Topics
Economics	Gross Domestic Product (GDP), Consumer Price Index (CPI), and unemployment rates
Medicine	Blood pressure tracking, weight tracking, cholesterol measurements, heart rate monitoring
Physical sciences	Global temperatures, monthly sunspot observations, pollution levels.
Social Sciences	Birth rates, population, migration data, political indicators
Epidemiology	Disease rates, mortality rates, mosquito populations

Time Series Model

75

- A time series is a sequential set of data points, measured typically at successive times. It is mathematically defined as a set of vectors $x(t)$ where $t = 0, 1, 2 \dots$ where t represents the time elapsed. The variable $x(t)$ is treated as random variable.
- A time series model generally reflect the fact that observations close together in time which are closely related than the observations further apart.
- The data shown below represent the weekly demand of some product. The model uses x to indicate an observation and t to represent the index of the time period. The data from 1 to t is: x_1, x_2, \dots, x_t . Time series of 25 periods is shown below.

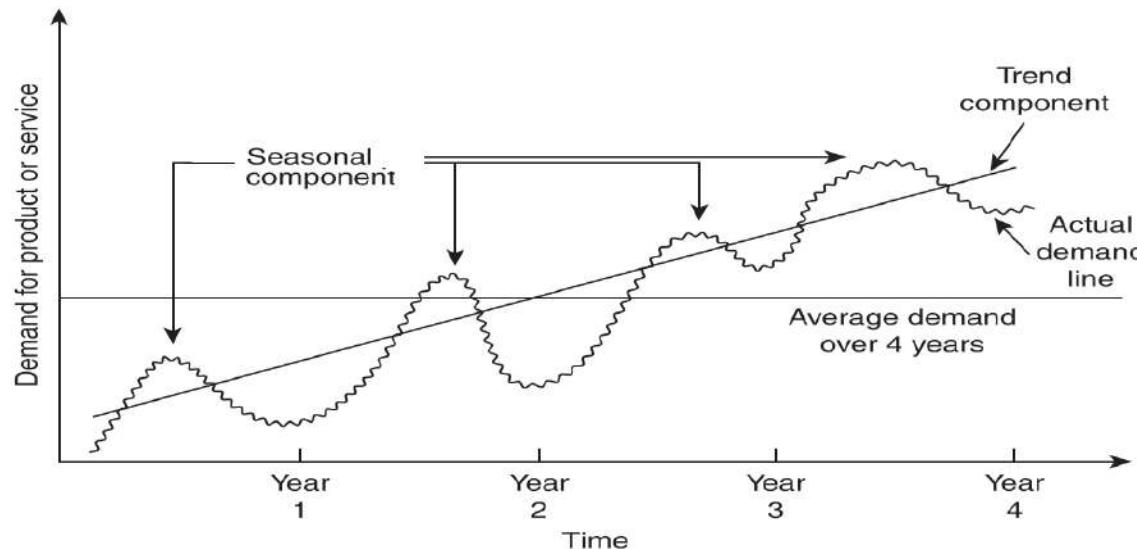
Weekly demand	Time Observations	
	1 – 10	11 – 20
	21 – 30	
4	16	12
3	14	20
8	7	2
25	7	9
13	6	11
12	11	3
4	16	11
8	9	4
9	3	14
14	11	11
1	14	8
15	7	9
16	10	9
17	7	7
18	11	11
19	16	3
20	10	8
21	11	9
22	8	7
23	2	4
24	8	8
25	8	8



Time Series Model cont...

76

- ❑ Any time series is composition of many individual component times series. Some of these components are predictable whereas other components may be almost random which can be difficult to predict.
- ❑ This calls for the decomposition methods that will generate individual component series from the original series. Decomposing a series into such components enable to analyze the behaviour of each component and thus improve the accuracy of the final forecast.
- ❑ **Example:** A typical sales time series.



Time Series Model Component



77

Time series models are characterized of four components:

- Trend component Seasonal component
- Cyclical component Irregular component

Trend component

- The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency.
- It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.
- It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable.
- The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Time Series Model Component cont...



78

Seasonal component

- ❑ These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.
- ❑ These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.
- ❑ The effect of person-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Time Series Model Component cont...



79

Cyclical component

- ❑ The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.
- ❑ It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular or not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Irregular component

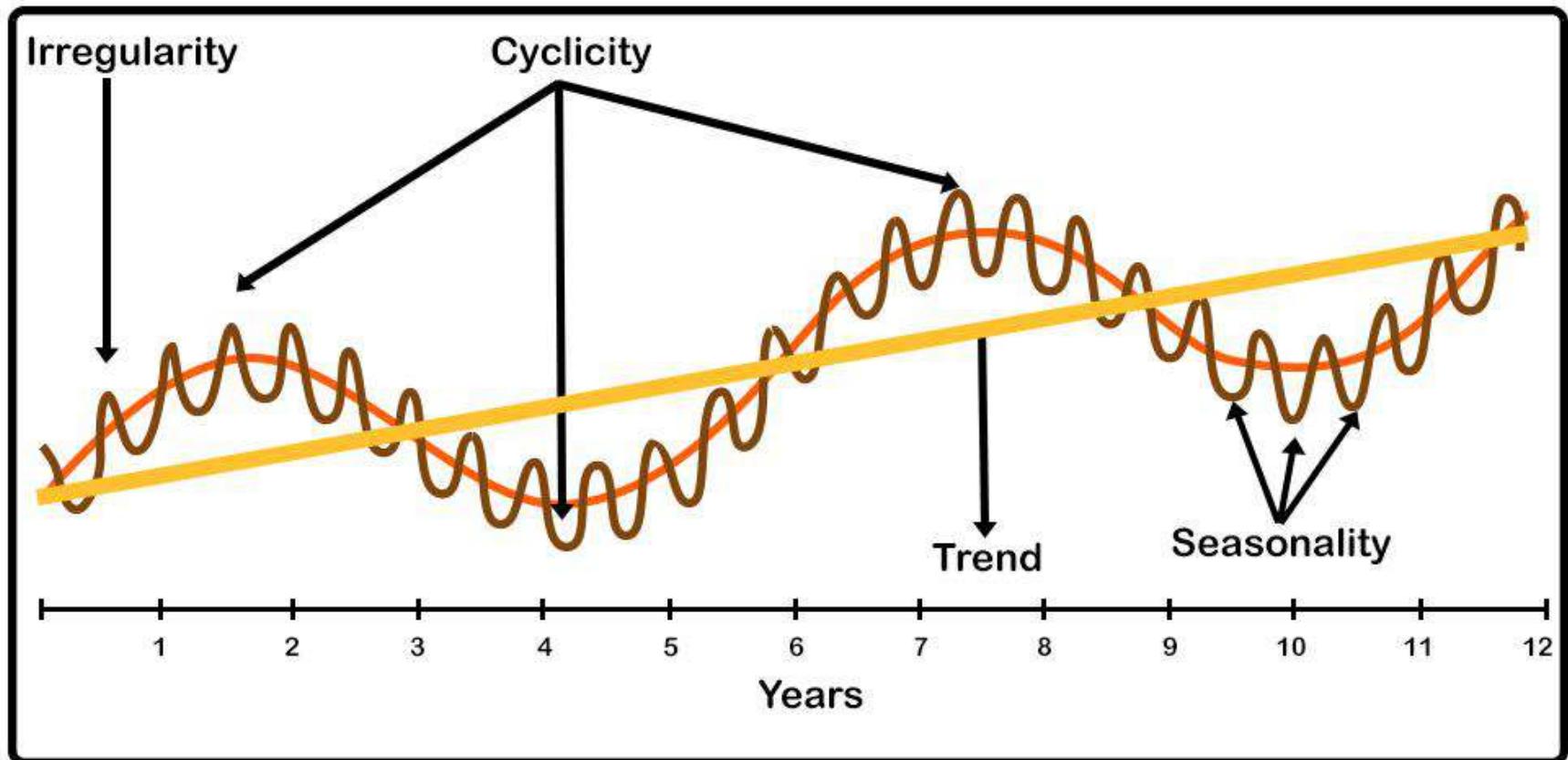
- ❑ They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

Time Series Model Component cont...



80

Pictorial depiction of different component



Decomposition Model

81

- ❑ Mathematical representation of the decomposition approach is $Y_t = f(T_t, S_t, C_t, I_t)$ where Y_t is the time series value at time t . T_t , S_t , C_t , and I_t are the trend, seasonal, cyclic and irregular component value at time t respectively.
- ❑ There are 3 types of decomposition model:
 - ❑ Additive model
 - ❑ Multiplicative model
 - ❑ Mixed model

Additive model

- ❑ According to this model, a time series is expressed as $Y_t = T_t + S_t + C_t + I_t$
- ❑ The model is appropriate when the amplitude of both the seasonal and irregular variations do not change as the level of trend rises or falls.
- ❑ This model assumes that all four components of the time series act independently of each other.

Multiplicative model

- ❑ According to this model, a time series is expressed as $Y_t = T_t * S_t * C_t * I_t$
- ❑ The model is appropriate when the amplitude of both the seasonal and irregular variations increase as the level of trend rises.
- ❑ The model assumes that the various components operate proportionately to each other.

Decomposition Model cont...

82

Mixed model

- Different assumptions lead to different combinations of additive and multiplicative models as $Y_t = T_t + S_t + C_t * I_t$
- The time series analysis can also be done using the model as:
 - $Y_t = T_t + S_t * C_t * I_t$
 - $Y_t = T_t * S_t + C_t * I_t$

Home Work

- How to determine if a time series has a trend component?
- How to determine if a time series has a seasonal component?
- How to determine if a time series has both a trend and seasonal component?

Time Series Forecasting Model



83

- Time series forecasting models can be classified into 2 categories.
- One group is called as **averaging methods** in which all observations (time series values) are equally weighted.
- The second group called **exponential smoothing methods** that applies unequal weights to past data, typically decaying in an exponential manner as one goes from recent to distinct past.

Averaging Model

- The simple average method uses the mean of all the past values to forecast the next value. This method is seen to be no use in a practical scenario.
- This method is used when the time series has attained some level of stability and no longer dependent on any external parameters.
- This would happen in sales forecasting, only when the product for which the forecast is needed is at a mature stage in its life cycle.
- The Averaging Model is represented as follows where F is the forecasted value at instance of time $t+1$, t is the current time and Y_i is the value of series at time instant i .

$$F_{t+1} = \frac{1}{t} \sum_{i=1}^t Y_i$$

Averaging Model

84

Supplier	Amount
1	9
2	8
3	9
4	12
5	9
6	12
7	11
8	7
9	13
10	9
11	11
12	10

A manager of a warehouse wants to know how much a typical supplier delivers in 10 dollar units. He/she has taken a sample of 12 suppliers at random, obtaining the result as shown in the table.

The computed mean of the amount is 10 and hence the manager decides to use this as the estimate for the expenditure of a typical supplier.

It is more reasonable to assume that the recent points in past are better predictors than the whole history. This is particularly true for sales forecasting. Every product has a life cycle, initial stage, middle volatile period and a more or less stable mature stage and an end stage. Hence, a better method of forecasting would be to use moving averages (MAs).

Moving Averages (MAs)

85

- The MA approach calculates an average of a finite number of past observations and then employs that average as the forecast for the next period.
- The number of sample observations to be included in the calculation of the average is specified at the start of the process. The term MA refer to the fact that as a new observation becomes available, a new average is calculated by dropping the oldest observation in order to include the newest one.
- An MA of order k, represented with MA(k) is calculated as:

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^t Y_i$$

- MA(3), MA(5) and MA(12) are commonly used for monthly data and MA(4) is normally used for quarterly data.
- MA(4), and MA(12) would average out the seasonality factors in quarterly and monthly data respectively.
- The advantage of MA method is that the data requirement is very small.
- The major disadvantage is that it assumes the data to be stationary.
- MA also called as **simple moving average**.

Moving Averages (MAs) cont...



86

Month	Demand
1	89
2	57
3	144
4	221
5	177
6	280
7	223
8	286
9	212
10	275
11	188
12	312

- $MA(3) = (275 + 188 + 312) / 3 = 258.33$
- $MA(6) = (223+286+212+275+188+312)/6 = 249.33$
- $MA(12) = (89+57+144+221+177+280+223+286+212+275+188+312)/ 3= 205.33$

Home Work

Calculate:

- $MA(5)$
- $MA(4)$
- $MA(10)$

Exponential Smoothing Model



87

- ❑ The extension to the MA method is to have a weighted MA, whereas in Single Moving Averages the past observations are weighted equally, Exponential Smoothing assigns exponentially decreasing weights as the observation get older. In other words, recent observations are given relatively more weight in forecasting than the older observations.
- ❑ In the case of moving averages, the weights assigned to the observations are the same and are equal to $1/N$. In exponential smoothing, however, there are one or more smoothing parameters to be determined (or estimated) and these choices determine the weights assigned to the observations.
- ❑ This class of techniques consists of a range of methods, starting from simple exponential smoothing (SES) used for the data with no trend or seasonality, to the sophisticated widely used Holt-Winter's method which is able to provide forecasts for data that exhibit both seasonality and trend.
- ❑ ***In these methods, the observations are weighted in an exponentially decreasing manner as they become older.***

Simple Exponential Smoothing



88

- For any time period t , the smoothed value S_t is found by computing $S_t = \alpha * y_{t-1} + (1-\alpha) * S_{t-1}$ where $0 < \alpha \leq 1$ and $t \geq 3$ and y_t represents the actual value at time t , S_t smoothed observation at time t , α is the smoothing constant.
- **Why is it called Exponential?**

Let us expand the basic equation by first substituting for S_{t-1} in the basic equation to obtain:

$$\begin{aligned}S_t &= \alpha * y_{t-1} + (1-\alpha) * [\alpha * y_{t-2} + (1-\alpha) * S_{t-2}] \\&= \alpha * y_{t-1} + \alpha * (1-\alpha) * y_{t-2} + (1-\alpha)^2 * S_{t-2}\end{aligned}$$

By substituting for S_{t-2} , then for S_{t-3} , and so forth, until we reach S_2 (which is just y_1), it can be shown that the expanding equation can be written as:

$$S_t = \alpha \sum_{i=1}^{t-2} (1 - \alpha)^{i-1} y_{t-i} + (1 - \alpha)^{t-2} S_2, \quad t \geq 2$$

This illustrates the exponential behavior. The weights, $\alpha * (1-\alpha)^t$ decrease geometrically.

Simple Exponential Smoothing cont...



89

□ What is the best value for α ?

The speed at which the older responses are dampened (smoothed) is a function of the value of α . When α is close to 1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below.

α	$(1-\alpha)$	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.0625
0.1	0.9	0.81	0.729	0.6561

Error calculation

- The error is calculated as $E_t = y_t - S_t$ (i.e. difference of actual and smooth at time t)
- Then error square is calculated i.e. $ES_t = E_t * E_t$
- Then, sum of the squared errors (SSE) is calculated i.e. $SSE = \sum ES_i$ for $i = 1$ to n where n is the number of observations.
- Then, the mean of the squared errors is calculated i.e. $MSE = SSE/(n-1)$
- ***The best value for α is choose so the value which results in the smallest MSE.***

Simple Exponential Smoothing cont...



90

Let us illustrate this principle with an example. Consider the following data set consisting of 12 observations taken over time with α as 0.1:

Time	y_t	s_t	e_t	es_t
1	71			
2	70	$0.1 * 70 - (1-0.1) * 71 = 71$	$70 - 71 = -1.0$	$(-1.0)^2 = 1.00$
3	69	$0.1 * 69 - (1-0.1) * 70 = 70.9$	$69 - 70.9 = -1.90$	$(-1.90)^2 = 3.61$
4	68	70.71	-2.71	7.34
5	64	70.44	-6.44	41.47
6	65	69.80	-4.80	23.04
7	72	69.32	2.68	7.18
8	78	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	75	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

Simple Exponential Smoothing cont...



91

- ❑ The sum of the squared errors (SSE) = 208.94. The mean of the squared errors (MSE) is the SSE /11 = 19.0.
- ❑ In the similar fashion, the MSE was again calculated for $\alpha=0.5$ and turned out to be 16.29, so in this case we would prefer an α of 0.5.
- ❑ **Can we do better?**
 - ❑ We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between 0.1 and 0.9.
 - ❑ We determine the best initial choice for α and then search between $\alpha-\Delta$ and $\alpha+\Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.

In general, most well designed statistical software programs should be able to find the value of α that minimizes the MSE.

Holt's Method

92

- ❑ Holt (1957) extended simple exponential smoothing to allow the forecasting of data with a trend. This method involves a forecast equation and two smoothing equations (i.e. one for the level and one for the trend).
- ❑ This method is used when a series has no seasonality but exhibits some form of trend.
- ❑ The k step ahead forecast function for a given time series X is
$$X_{t+k} = \ell_t + k * b_t$$
 where ℓ_t denotes an estimate of the level of the series at time t, b_t denotes an estimate of the trend (slope) of the time series at time t.
- ❑ The equation for level is
$$\ell_t = \alpha * y_t + (1 - \alpha) * (\ell_{t-1} + b_{t-1})$$
- ❑ The equation for trend is
$$b_t = \beta * (\ell_t - \ell_{t-1}) + (1 - \beta) * b_{t-1}$$
where,
 α is the smoothing parameter for the level, $0 \leq \alpha \leq 1$,
 β is the smoothing parameter for the trend, $0 \leq \beta \leq 1$.
- ❑ Reasonable starting values for level and slope are $\ell_1 = X_1$ and $b_1 = X_2 - X_1$

Evaluation of Forecasting Accuracy



93

- ❑ What makes a good forecast? Of course, a good forecast is an accurate forecast.
- ❑ A forecast “error” is the difference between an observed value and its forecast. The “error” does not mean a mistake, it means the unpredictable part of an observation.
- ❑ Error measure plays an important role in calibrating and refining forecasting model/method and helps the analyst to improve forecasting method.
- ❑ The choice of an error measure may vary according to the situation , number of time series available and on whether the task is to select the most accurate method or to calibrate a given model.
- ❑ The popular and highly recommended error measures are
 - ❑ Mean Square Error (MSE)
 - ❑ Root Mean Square Error (RMSE)
 - ❑ Mean Absolute Percentage Error (MAPE)

Mean Square Error (MSE)

94

MSE is defined as mean or average of the square of the difference between actual and estimated values. Mathematically it is represented as:

$$\text{MSE} = \frac{\sum_{j=1}^N (\text{observation } (j) - \text{prediction } (j))^2}{N}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56 and MSE = 56 / 12 = 4.6667

Root Mean Square Error (RMSE)

95

It is just the square root of the mean square error. Mathematically it is represented as:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^N (\text{observation } (j) - \text{prediction } (j))^2}{N}}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56, MSE = 56 / 12 = 4.6667, RMSE = SQRT(4.667) = 2.2

Mean Absolute Percentage Error (MAPE)

96

The formula to calculate MAPE is as follows:

$$\text{MAPE} = (100 / n) \times \sum_{i=1}^n \frac{|X'(t) - X(t)|}{X(t)}$$

Here, $X'(t)$ represents the forecasted data value of point t and $X(t)$ represents the actual data value of point t. Calculate MAPE for the below dataset.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38

- ❑ MAPE is commonly used because it's easy to interpret and easy to explain. For example, a MAPE value of 11.5% means that the average difference between the forecasted value and the actual value is 11.5%.
- ❑ The lower the value for MAPE, the better a model is able to forecast values e.g. a model with a MAPE of 2% is more accurate than a model with a MAPE of 10%.

THANK YOU!

Data Analytics (IT-3006)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering



Strictly for internal circulation (within KIIT) and reference only. Not for outside circulation without permission

3 Credit

Lecture Note – Unit 3

Course Contents



2

Sr #	Major and Detailed Coverage Area	Hrs
2	Mining Data Streams Introduction to Mining Data Streams, Data Stream Management Systems, Data Stream Mining, Examples of Data Stream Applications, Stream Queries, Issues in Data Stream Query, Sampling in Data Streams, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Querying on Windows – Counting Ones in a Window, Decaying Windows , Real-Time Analytics Platform (RTAP).	10

Data Stream

3

Data Stream – Large data volume, likely unstructured and structured arriving at a very high rate, which requires real time/near real time analysis for effective decision making.

- ❑ It is basically continuously generated data and arrives in a stream (sequence of data elements made available over time). It is generally time-stamped and geo-tagged (in the form of latitude and longitude).
- ❑ Stream is composed of synchronized sequence of elements or events.
- ❑ If it is not processed immediately, then it is lost forever.
- ❑ In general, such data is generated as part of application logs, events, or collected from a large pool of devices continuously generating events such as ATM or PoS.

Example:

Data Center: Large network deployment of a data center with hundreds of servers, switches, routers and other devices in the network. The event logs from all these devices at real time create a stream of data. This data can be used to prevent failures in the data center and automate triggers so that the complete data center is fault tolerant.

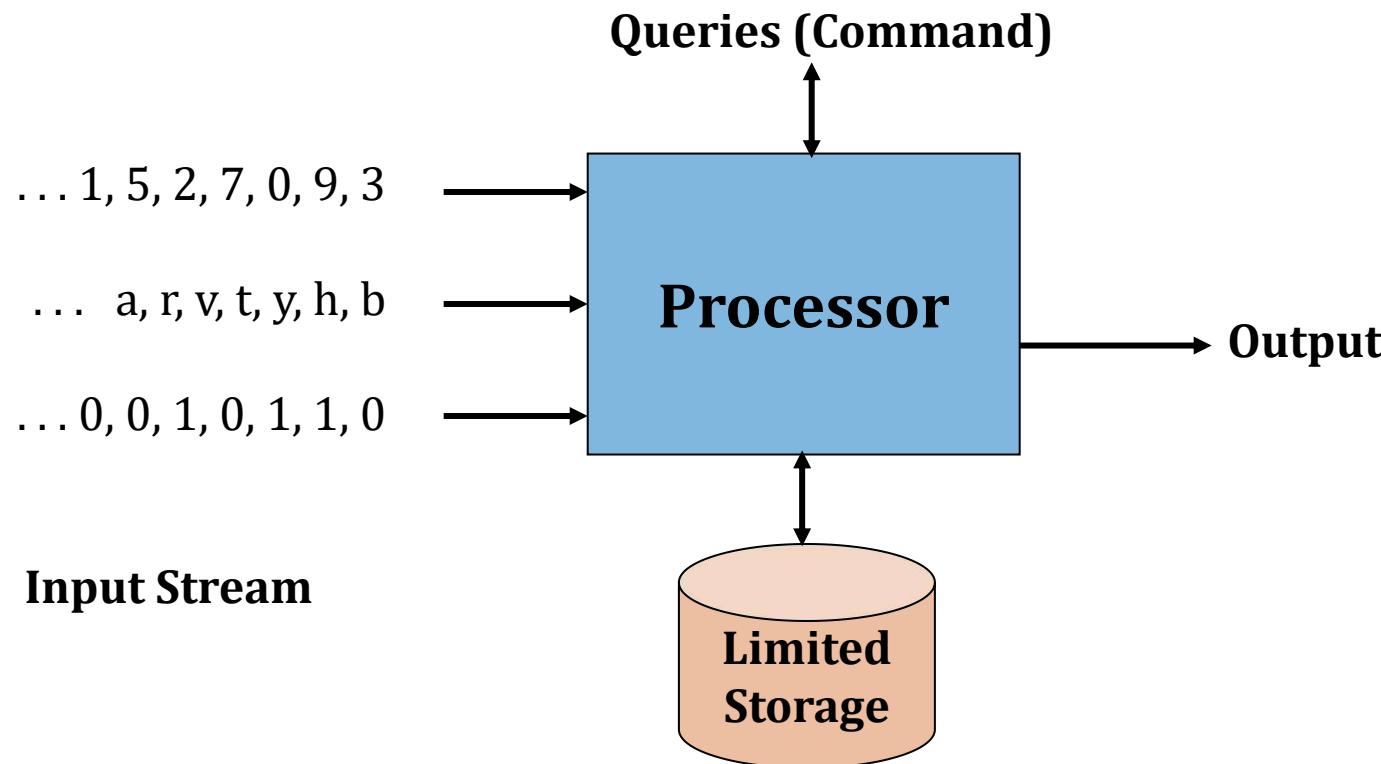
Stock Market: The data generated here is a stream of data where a lot of events are happening in real-time. The price of stock are continuously varying. These are large continuous data streams which needs analysis in real-time for better decisions on trading.

Basic Model of Stream data



4

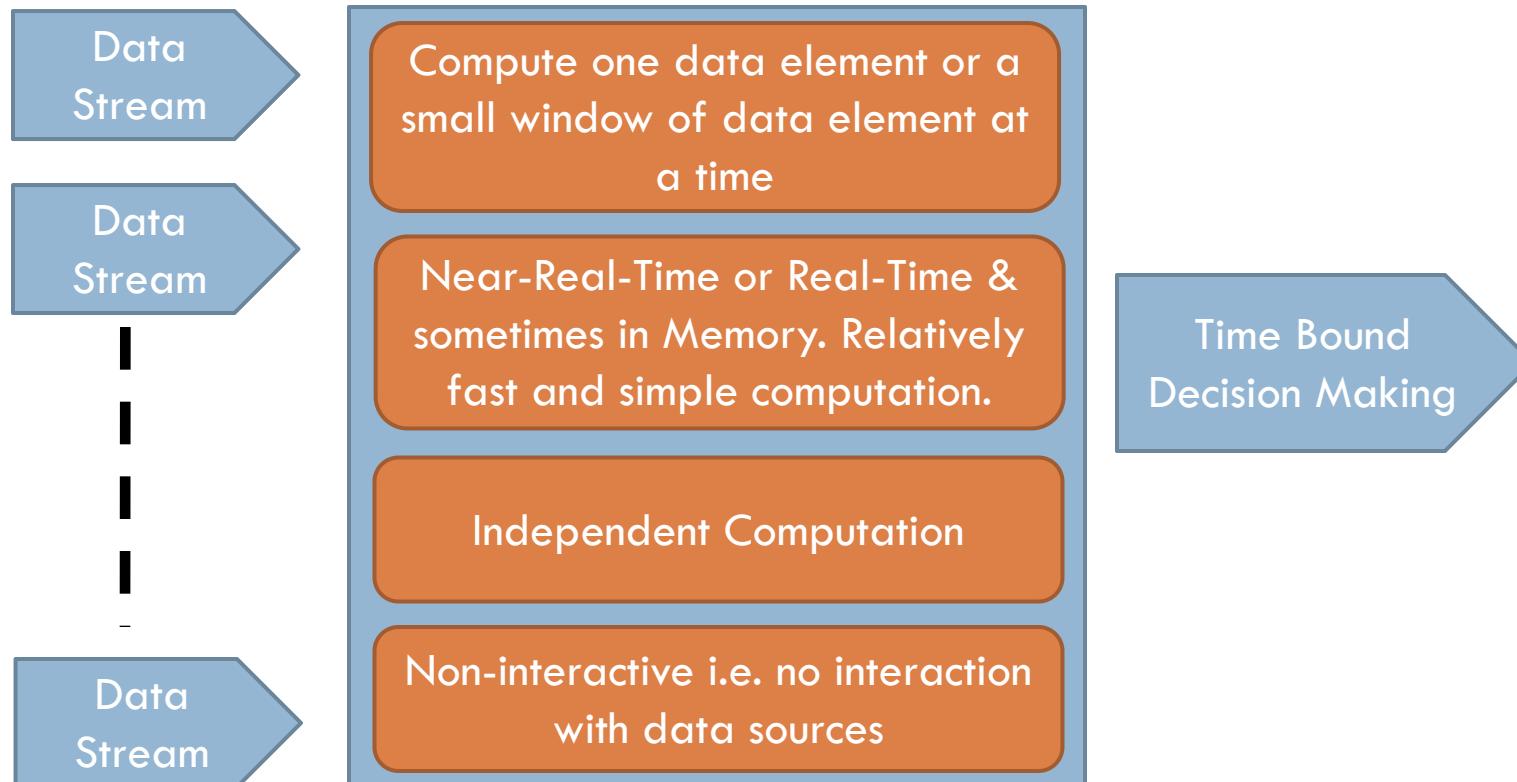
- ❑ Input data rapidly and streams needn't have the same data rates or data types.
- ❑ The system cannot store the data entirely.
- ❑ Queries tends to ask information about recent data.
- ❑ The scan never turn back.



Streaming Data System

5

Streaming Data System



Data-at-Rest vs. Data-in-Motion



6

- ❑ **Data-at-rest** - This refers to data that has been collected from various sources and is then analyzed after the event occurs. The point where the data is analyzed and the point where action is taken on it occur at two separate times. For example, a retailer analyzes a previous month's sales data and uses it to make strategic decisions about the present month's business activities. The action takes place after the data-creating event has occurred. For data at rest, a batch processing method would be most likely.
- ❑ **Data-in-motion** - The collection process for data in motion is similar to that of data at rest; however, the difference lies in the analytics. In this case, the analytics occur in real-time as the event happens. For example – sensor data from self-driving vehicles. For data in motion, you'd want to utilize a real-time processing method.

Data-at-Rest vs. Data-in-Motion Infrastructure Option



7

Data-at-rest



Public Cloud

Public cloud can be an ideal infrastructure choice in such scenario from a cost standpoint, since virtual machines can easily be spun up as needed to analyze the data and spun down when finished.

Data-in-motion



Bare-Metal Cloud

Bare-Metal cloud can be an preferable infrastructure choice. It involves the use of dedicated servers that offers cloud-like features without the use of virtualization.

Streaming Data Changes over Time

8

Change can be periodic or sporadic

**Periodic: evening,
weekends etc**

People post Facebook messages more in the evening in comparison to during day, working hours.

**Sporadic: major
events**

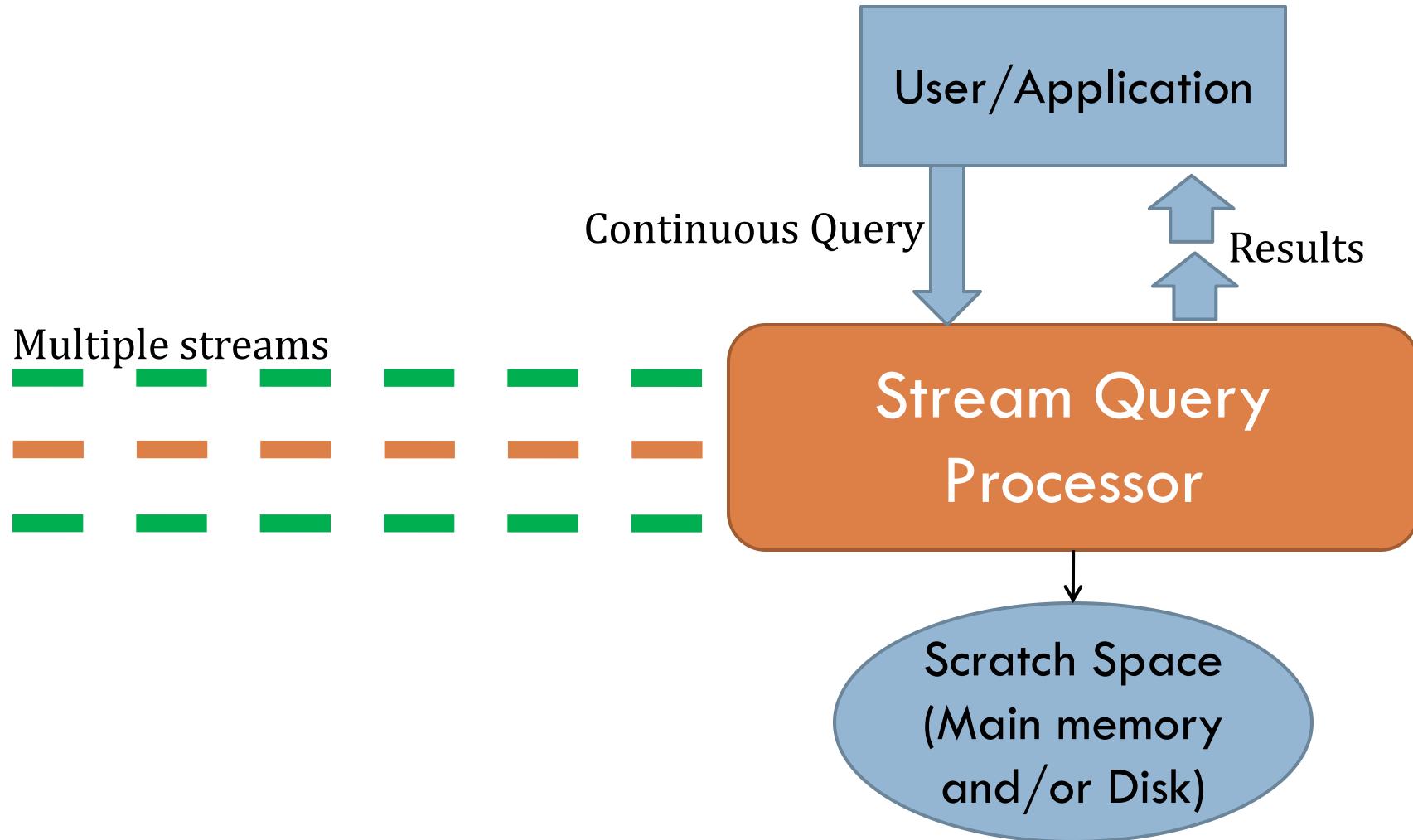
BREAKING NEWS

In summary, streaming data:

- ❑ Size is unbounded i.e. it continually generated and can't process all at once
- ❑ Size and Frequency is unpredictable due to human behavior
- ❑ Processing is must be relatively fast and simple

Architecture: Stream Query Processing

9

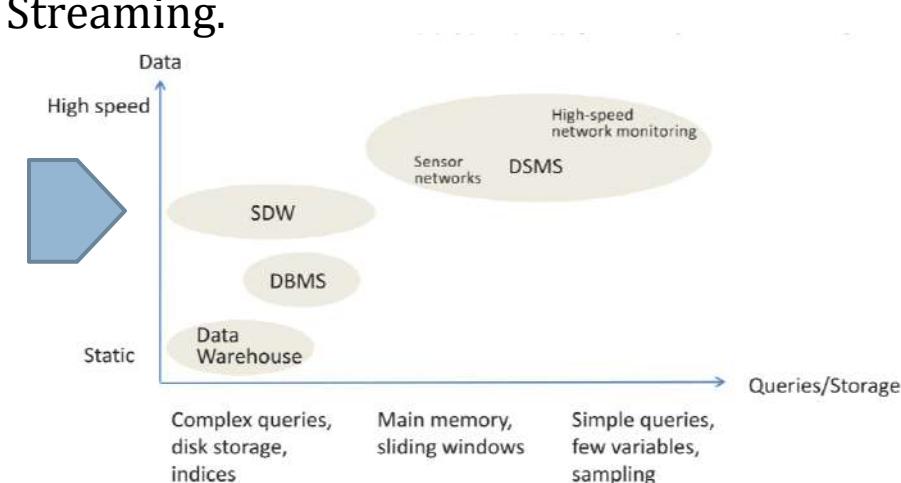


Data Stream Management Systems



10

- ❑ Traditional relational databases store and retrieve records of data that are static in nature and do not perceive a notation of time unless time is added as an attribute during the schema design.
- ❑ The model is adequate for legacy applications and older repositories of information, but many current and emerging application require support for online analysis of rapidly arriving and changing data streams.
- ❑ This has resulted in data stream management system (DSMS) with an emphasis on continuous query languages and query evaluation.
- ❑ There are two complementary techniques for end-to-end stream processing: Data Stream Management Systems (DSMSs) and Streaming.
- ❑ Comparison of DSMS and SDW with traditional database and warehouse systems, wherein data rates are on the y-axis, and query complexity and available storage on the x-axis.



Summary difference between DBMS and DSMS

11

	DBMS	DSMS
Data	Persistent relations	Streams, time windows
Data access	Random	Sequential, One-pass
Updates	Arbitrary	Append-only
Update Rates	Relatively Low	High, bursty
Processing model	Query driven (pull-based)	Data driven (push-based)
Queries	One-time	Continuous
Query Plans	Fixed	Adaptive
Query Optimization	One query	Multi-query
Query Answers	Exact	Exact or approximate
Latency	Relatively high	Low

Summary difference between Traditional data warehouse and SDW



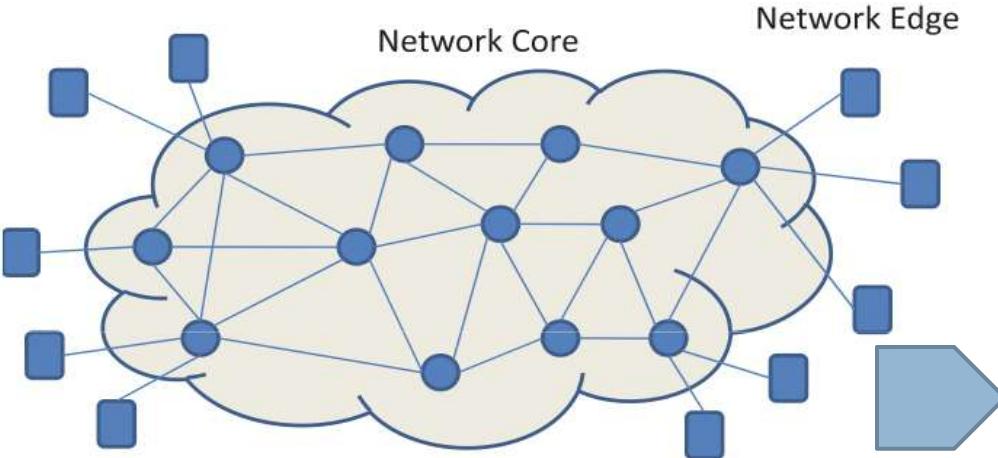
12

	Traditional data warehouse	SDW
Update frequency	Low	High
Update propagation	Synchronous	Asynchronous
Data	Historical	Recent and historical
ETL Process	Complex	Fast and light-weight

Network monitoring - A use case



13



The network monitoring illustrates a simple IP network with high-speed routers and links in the core, and hosts (clients and servers) at the edge. A large network contains thousands of routers and links, and its core links may carry many thousands of packets per second; in fact, optical links in the Internet backbone can reach speeds of over 100 million packets per second.

- ❑ The traffic flowing through the network is itself a high-speed data stream, with each data packet containing fields such as a timestamp, the source and destination IP addresses, and ports.
- ❑ Other network monitoring data streams include real-time system and alert logs produced by routers, routing and configuration updates, and periodic performance measurements.
- ❑ However, it is not feasible to perform complex operations on high-speed streams or to keep transmitting terabytes of raw data to a data management system.
- ❑ Instead, there is a need of scalable and flexible end-to-end data stream management solutions, ranging from real-time low-latency alerting and monitoring, ad-hoc analysis and early data reduction on raw streaming data, to long-term analysis of processed data.

Network monitoring – DBMS, DSMS, SDW

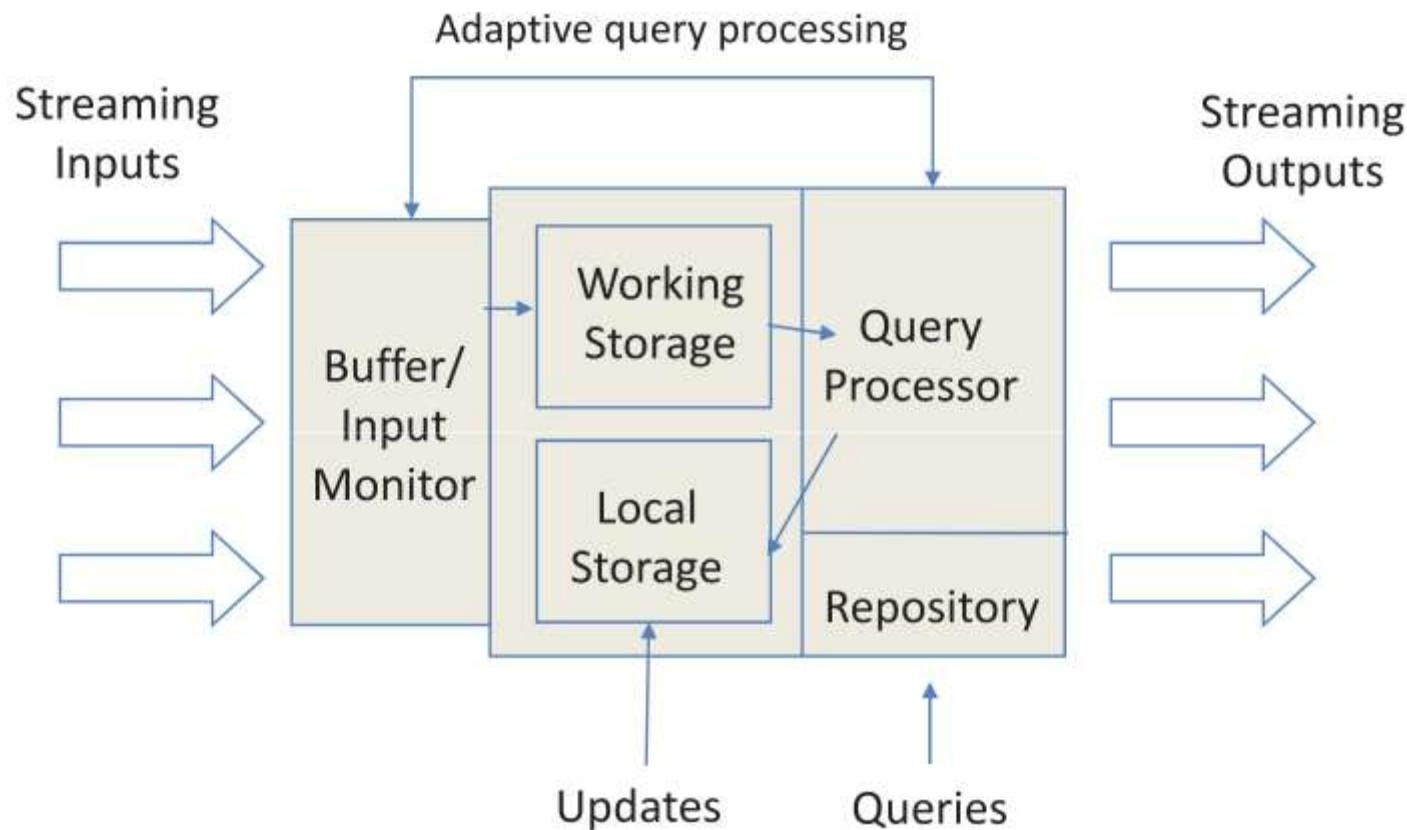


14

- ❑ Database Management Systems (DBMSs) handle somewhat more dynamic workloads, consisting of ad-hoc queries and data manipulation statements, i.e., insertions, updates and deletions of a single row or groups of rows. On the other hand, DSMSs lie in the top right corner as they evaluate continuous queries on data streams that accumulate over time.
- ❑ SDWs, also known as Active Data Warehouses, combine the real-time response of a DSMS (by attempting to load and propagate new data across materialized views as soon as they arrive) with a data warehouse's ability to manage Terabytes of historical data on secondary storage.
- ❑ In applications such as troubleshooting a live network, the data rates may be so high that only the simplest continuous queries that require very little working memory and per-tuple processing are feasible, such as simple filters and simple aggregates over non-overlapping windows.
- ❑ In network monitoring, an SDW may store traffic streams that have been pre-aggregated or otherwise pre-processed by a DSMS, as well as various network performance and configuration feeds that arrive with a wide range of inter-arrival times, e.g., once a minute to once a day.

Reference architecture of a DSMS

15



Reference architecture of a DSMS cont...

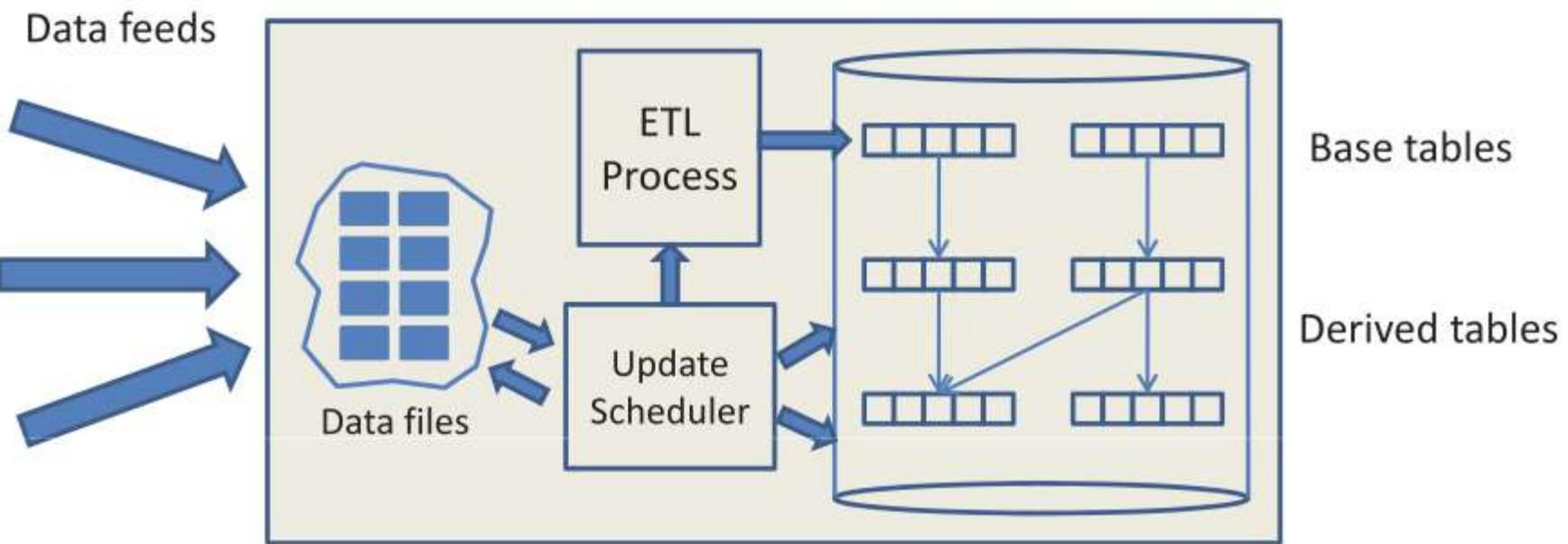


16

- ❑ The input buffer captures the streaming inputs. Optionally, an input monitor may collect various statistics such as inter-arrival times or drop some incoming data in a controlled fashion (e.g., via random sampling) if the system cannot keep up.
- ❑ The working storage component temporarily stores recent portions of the stream and/or various summary data structures needed by queries. Depending on the arrival rates, this ranges from a small number of counters in fast RAM to memory-resident sliding windows.
- ❑ Local storage may be used for metadata such as foreign key mappings, e.g., translation from numeric device IDs that accompany router performance data to more user-friendly router names. Users may directly update the metadata in the local storage, but the working storage is used only for query processing.
- ❑ Continuous queries are registered in the query repository and converted into execution plans; similar queries may be grouped for shared processing. While superficially similar to relational query plans, continuous query plans also require buffers, inter-operator queues and scheduling algorithms to handle continuously streaming data. Conceptually, each operator consumes a data stream and returns a modified stream for consumption by the next operator in the pipeline.
- ❑ The query processor may communicate with the input monitor and may change the query plans in response to changes in the workload and the input rates.
- ❑ Finally, results may be streamed to users, to alerting or event-processing applications, or to a SDW for permanent storage and further analysis.

Reference architecture of a SDW

17



Reference architecture of a SDW cont...



18

- ❑ Data streams or feeds arrive periodically from various sources, often in the form of text or zipped files.
- ❑ An update scheduler decides which file or batch of files to load next.
- ❑ The data then pass through an ETL process, as in traditional data warehouses. Examples of ETL tasks include unzipping compressed files, and simple data cleaning and standardization (e.g., converting strings to lower or upper case or converting timestamps to GMT).
- ❑ Base tables are sourced directly from the raw files, while derived tables correspond to materialized views (over base or other derived tables).
- ❑ Base and derived tables are usually partitioned by time so that arrivals of new data only affect the most recent partitions.

Mining Big Data Stream

19

- ❑ Mining big data streams faces three principal challenges: volume, velocity, and volatility.
- ❑ Volume and velocity require a high volume of data to be processed in limited time. Starting from the first arriving instance, the amount of available data constantly increases from zero to potentially infinity. This requires incremental approaches that incorporate information as it becomes available, and online processing if not all data can be kept.
- ❑ Volatility corresponds to a dynamic environment with ever-changing patterns. Here, old data is of limited use, even if it could be saved and processed again later.
- ❑ This can affect the data mining models in multiple ways:
 - ❑ Change of the target variable.
 - ❑ Change in the available feature information.
 - ❑ Drift.

Mining Big Data Stream cont...



20

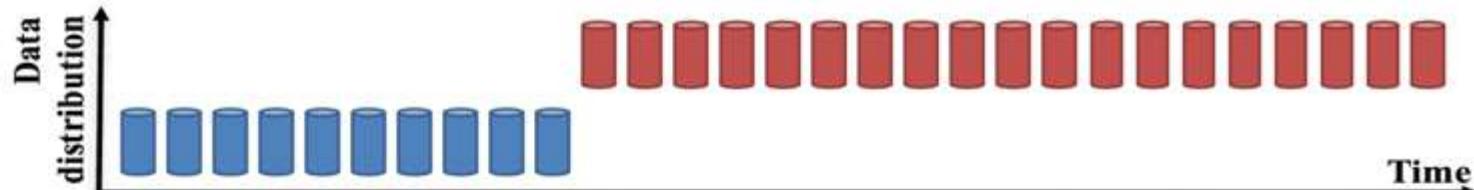
- ❑ Changes of the target variable occur for example in credit scoring, when the definition of the classification target “default” versus “non-default” changes due to business or regulatory requirements.
- ❑ Changes in the available feature information arise when new features become available, e.g. due to a new sensor or instrument. Similarly, existing features might need to be excluded due to regulatory requirements, or a feature might change in its scale, if data from a more precise instrument becomes available.
- ❑ Finally, drift is a phenomenon that occurs when the distributions of features x and target variables y change in time, e.g. sudden changes in the popularity of movie several days after its release due to good reviews from those who watched it and also due to changes in the price of the movie.

Types of drift

21

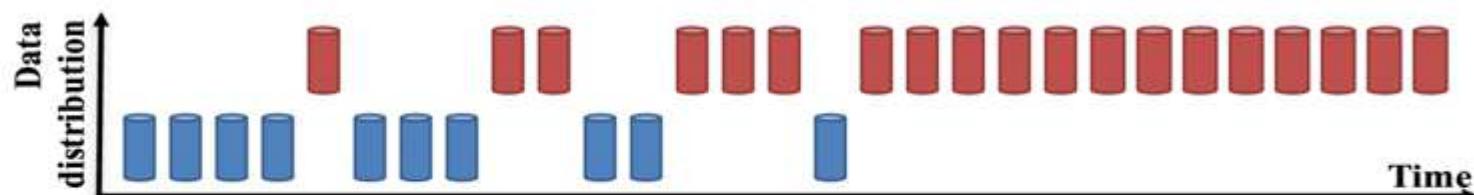
Sudden Drift:

A new concept occurs within a short time.



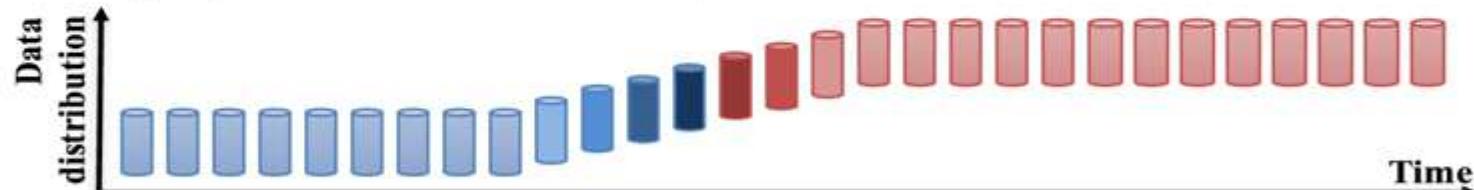
Gradual Drift:

A new concept gradually replaces an old one over a period of time.



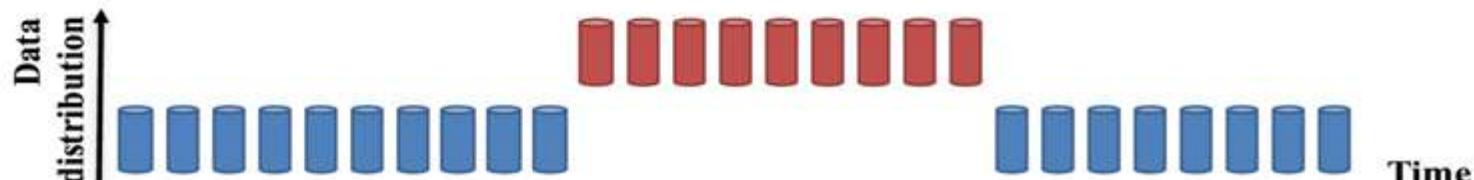
Incremental Drift:

An old concept incrementally changes to a new concept over a period of time.



Reoccurring Drift:

An old concept may reoccur after some time.



Mining Big Data Stream cont...



22

Mining big data stream is challenging in the two aspects.

- ❑ First, random access to fast and large data streams may be impossible. Thus multi-pass algorithm (ones that load data into main memory multiple times) are often infeasible.
- ❑ Second, the exact answers from data streams are often too expensive.

The most common data stream mining tasks are:

- ❑ Clustering.
- ❑ Classification.
- ❑ Frequent pattern mining.

Examples of Data Stream Applications



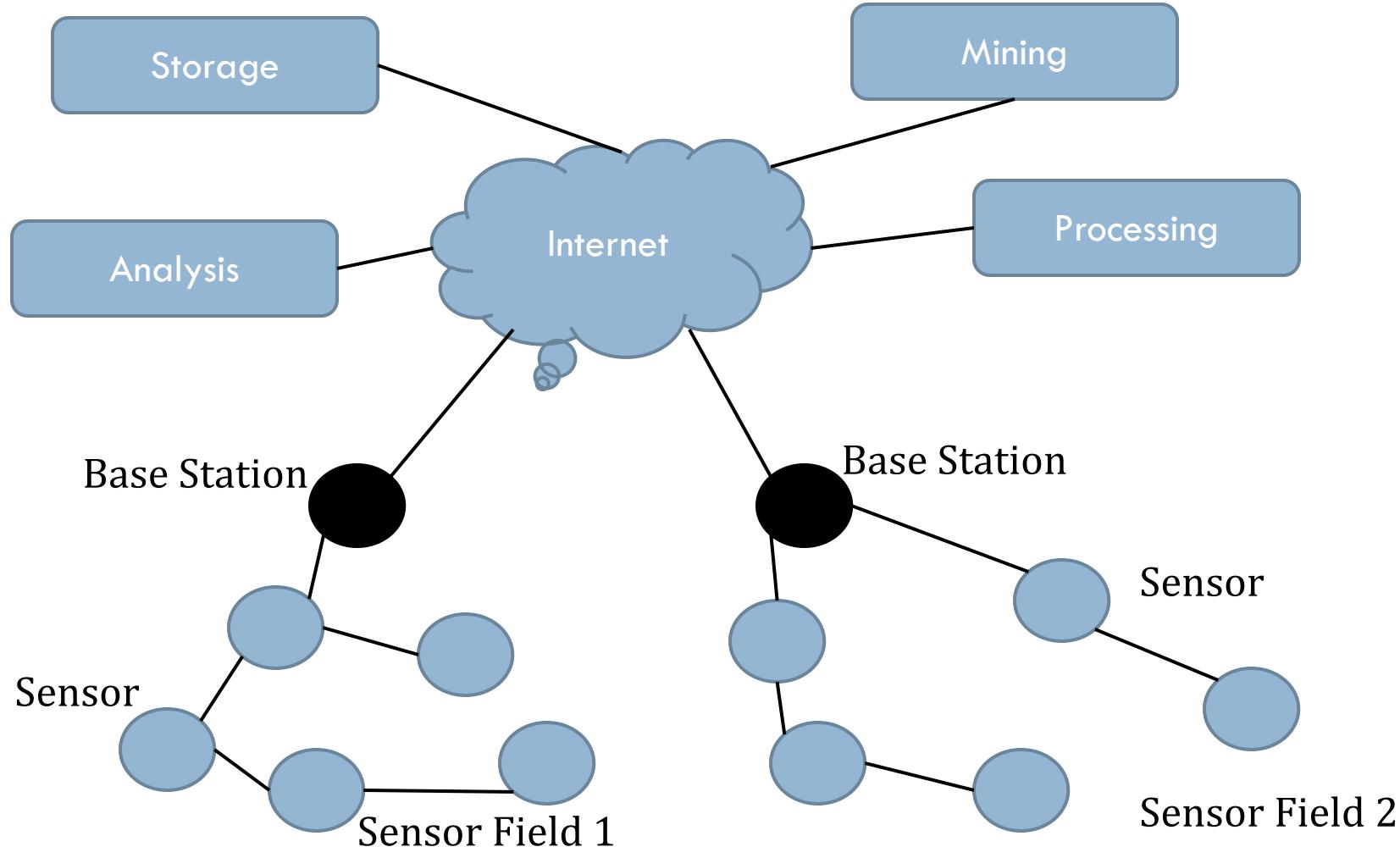
23

- Sensor Networks
- Network Traffic Analysis
- Financial Applications
- Transaction Log Analysis

Sensor Networks

- Sensor networks (SNs) consist of spatially distributed devices communicating through wireless radio and cooperatively sensing physical or environmental conditions to provide a high degree of visibility into the environmental physical processes. These are notably useful in emergency scenarios, such as floods, fires, volcanoes, battlefields, where human participation is too dangerous and infrastructure networks are either impossible or too expensive.
- SNs are a huge source of data occurring in streams. They require constant monitoring of several variables, based on which important decisions are made.
- In many cases, alerts and alarms may be generated as a response to the information received from a series of sensors. To perform such analysis, aggregations and joins over multiple data stream corresponding to various sensors are required.
- Examples of queries involves:
 - Perform join of several data stream like temperature streams, ocean streams etc at weather stations to give alerts or warning of disasters like cyclones and tsunami.
 - Constant monitoring of a stream of power usage statistics to a power station, group them by locations, user types etc to manage power distribution efficiently.

Sensor networks



Network traffic analysis



25

Network traffic analysis (NTA) is a method of monitoring network availability and activity to identify anomalies, including security and operational issues. Common use cases for NTA include:

- ❑ Collecting a real-time and historical record of what's happening on the network.
- ❑ Detecting malware such as ransom ware activity.
- ❑ Detecting the use of vulnerable protocols and ciphers.
- ❑ Troubleshooting a slow network.
- ❑ Improving internal visibility and eliminating blind spots.

Implementing a data stream solution can continuously monitor network traffic gives the insight to optimize network performance, minimize attack surface, enhance security, and improve the management of resources. Examples queries include:

- ❑ Check whether a current stream of actions over a time window are similar to previous identified intrusions on the network.
- ❑ Check if several routes over which traffic is moving has several common intermediate nodes which may potential indicate a congestion on that route.

Financial Applications



26

Online analysis of stock prices and making hold or sell decisions requires quickly identification of correlations and fast changing trends and to an extent forecasting future valuations as data is constantly arriving from several sources like news, current stock movement etc. Typical queries include:

- ❑ Find the stocks priced between \$1 and \$200, which is showing very large buying in the last one hour based on some federal bank news about tax rates for a particular industry.
- ❑ Find all the stocks trading above their 100 day moving average by more than 10% and also with volume exceeding a million shares.

Transactional Log Analysis



27

Online mining of web usage logs, telephone call records and ATM are the examples of data streams since they continuously output data and are potentially infinite. The goal is to find interesting customer behavior patterns, identifying suspicious spending behavior that could indicate fraud etc. Typical queries include:

- ❑ Examine current buying pattern of users at a website and potentially plan advertising campaigns and recommendations.
- ❑ Continuously monitor location, average spends etc of credit card customers and identify potential frauds.

Stream Queries

28

- ❑ Streams Queries are similar to SQL in that one can specify which data like to include in the stream, any conditions that the data has to match, etc.
- ❑ Streams queries are composed in the following format: **SELECT** <select criteria> **WHERE** <where criteria> **HAVING** <having criteria>
- ❑ Two types of queries can be identified as typical over data streams. The first distinction is between one-time queries and continuous queries.
 - ❑ **One-time queries:** These are the queries that are evaluated once over a point-in-time snapshot of the dataset, with the answers returned to the users. For example, a stock price checker may alert the user when a stock price crosses a particular price point.
 - ❑ **Continuous queries:** These are evaluated continuously as data streams continue to arrive. The answer to a continuous query is produced over time, always reflecting the stream data so far. It may be stored and updated as new data arrives, or they may produced as data streams themselves. Typically, aggregation queries such as finding maximum, average, count etc. are the continuous queries where values are stored.

Stream Queries cont...

29

The second distinction is between predefined queries and ad hoc queries:

- ❑ **Predefined query:** A predefined query is one that is supplied to the data stream management system before any relevant data has arrived. Predefined queries are generally continuous queries, although scheduled one-time queries can also be predefined.
- ❑ **Ad hoc queries:** Ad hoc queries, are issued online after the data streams have already begun. Ad hoc queries can be either one-time queries or continuous queries. Ad hoc queries complicate the design of a data stream management system, both because they are not known in advance for the purposes of query optimization, identification of common sub expressions across queries, etc., and more importantly because the correct answer to an ad hoc query may require referencing data elements that have already arrived on the data streams.

Issues in data stream queries



30

Query processing in the data stream model comes with its own unique challenges and are:

- Unbounded Memory Requirements
- Approximate Query Answering
- Sliding Windows
- Batch Processing, Sampling and Synopses
- Blocking Operators

Unbounded Memory Requirements

- Data streams are potentially unbounded (i.e. it continually generated and can't process all at once) in size, and the amount of storage required to compute an exact answer to a data stream query may also grow without bound.
- The continuous data stream model is most applicable to problems where timely query responses are important and there are large volumes of data that are being continually produced at a high rate over time.
- New data is constantly arriving even as the old data is being processed; the amount of computation time per data element must be low, or else the latency of the computation will be too high and the algorithm will not be able to keep pace with the data stream.
- For this reason, the interest is in algorithms that are able to confine themselves to main memory without accessing disk.

Approximate Query Answering



31

- ❑ When we are limited to a bounded amount of memory it is not always possible to produce exact answers for data stream queries; however, high-quality approximate answers are often acceptable in lieu of exact answers.
- ❑ Approximation algorithms for problems defined over data streams has been a fruitful research area in the algorithms community in recent years, which has led to some general techniques for data reduction and synopsis construction, including: sketches, random sampling, histograms, and wavelets.

Note:

Data reduction: It is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form.

Synopsis construction: Creating synopsis of data refers to the process of applying summarization techniques that are capable of summarizing the incoming stream for further analysis. Since synopsis of data does not represent all the characteristics of the dataset, approximate answers are produced when using data synopsis. Synopses structures can be used both for answering queries and applying data mining algorithms to streams.

Sliding Windows

32

- ❑ One technique for producing an approximate answer to a data stream query is to evaluate the query not over the entire past history of the data streams, but rather only over sliding windows of recent data from the streams. For example, only data from the last week could be considered in producing query answers, with data older than one week being discarded.
- ❑ Imposing sliding windows on data streams is a natural method for approximation that has several attractive properties. It is well defined and easily understood: the semantics of the approximation are clear, so that users of the system can be confident that they understand what is given up in producing the approximate answer.
- ❑ Most importantly, it emphasizes recent data, which in the majority of real-world applications is more important and relevant than old data: if one is trying in real-time to make sense of network traffic patterns, or phone call or transaction records, or scientific sensor data, then in general insights based on the recent past will be more informative and useful than insights based on stale data.
- ❑ In fact, for many such applications, sliding windows can be thought of not as an approximation technique reluctantly imposed due to the infeasibility of computing over all historical data, but rather as part of the desired query semantics explicitly expressed as part of the user's query.

Sliding Windows Example

33

Consider the demo stream **STREAM_001** with the schema as:

```
(TICKER_SYMBOL VARCHAR(4),  
 SECTOR varchar(16),  
 CHANGE REAL,  
 PRICE REAL)
```

- ❑ Suppose a application to compute aggregates using a sliding 1-minute window i.e. for each new record that appears on the stream, the application to emit an output by applying aggregates on records in the preceding 1-minute window.
- ❑ The following time-based windowed query can be used. The query uses the WINDOW clause to define the 1-minute range interval. The PARTITION BY in the WINDOW clause groups records by ticker values within the sliding window.

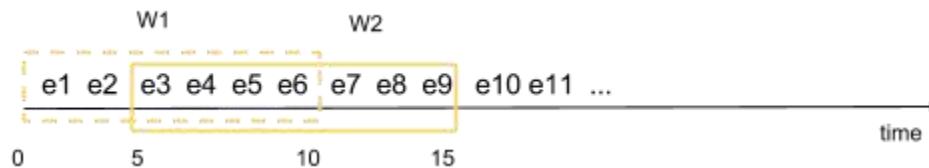
```
SELECT STREAM ticker_symbol,  
       MIN(Price) OVER W1 AS Min_Price,  
       MAX(Price) OVER W1 AS Max_Price,  
       AVG(Price) OVER W1 AS Avg_Price  
  FROM  "STREAM_001"  
WINDOW W1 AS (  
    PARTITION BY ticker_symbol  
    RANGE INTERVAL '1' MINUTE PRECEDING);
```

Sliding Windows and Tumbling Windows



34

- ❑ In a sliding window, tuples are grouped within a window that slides across the data stream according to a specified interval. A time-based sliding window with a length of ten seconds and a sliding interval of five seconds contains tuples that arrive within a ten-second window. The set of tuples within the window are evaluated every five seconds. Sliding windows can contain overlapping data; an event can belong to more than one sliding window.
- ❑ In the following image, the first window (w1, in the box with dashed lines) contains events that arrived between the zero th and ten th seconds. The second window (w2, in the box with solid lines) contains events that arrived between the fifth and fifteenth seconds. Note that events e3 through e6 are in both windows. When window w2 is evaluated at time $t = 15$ seconds, events e1 and e2 are dropped from the event queue.



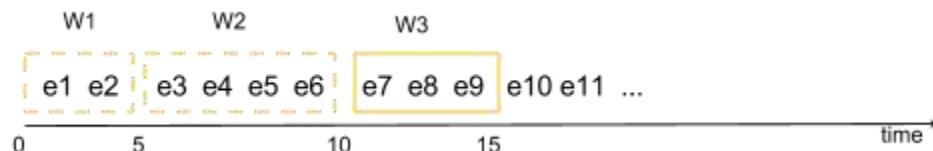
An example would be to compute the moving average of a stock price across the last five minutes, triggered every second.

Sliding Windows and Tumbling Windows cont...



35

- ❑ In a tumbling window, tuples are grouped in a single window based on time or count. A tuple belongs to only one window.
- ❑ For example, consider a time-based tumbling window with a length of five seconds. The first window (w1) contains events that arrived between the zero th and fifth seconds. The second window (w2) contains events that arrived between the fifth and tenth seconds, and the third window (w3) contains events that arrived between tenth and fifteenth seconds. The tumbling window is evaluated every five seconds, and none of the windows overlap; each segment represents a distinct time segment.



An example would be to compute the average price of a stock over the last five minutes, computed every five minutes.

Batch Processing, Sampling and Synopses



36

Another class of techniques for producing approximate answers is to give up on processing every data element as it arrives, resorting to some sort of sampling or batch processing technique to speed up query execution.

- ❑ In **batch processing**, rather than producing a continually up-to-date answer, the data elements are buffered as they arrive, and the answer to the query is computed periodically as time permits. The query answer may be considered approximate in the sense that it is not timely, i.e., it represents the exact answer at a point in the recent past rather than the exact answer at the present moment. This approach of approximation through batch processing is attractive because it does not cause any uncertainty about the accuracy of the answer, sacrificing timeliness instead. It is also a good approach when data streams are bursty.
- ❑ **Sampling** is based on the principle that it is futile to attempt to make use of all the data when computing an answer, because data arrives faster than it can be processed. Instead, some tuples must be skipped altogether, so that the query is evaluated over a sample of the data stream rather than over the entire data stream.
- ❑ For classes of data stream queries where no exact data structure with the desired properties exists, one can often design an approximate data structure that maintains a small **synopsis** or sketch of the data rather than an exact representation, and therefore is able to keep computation per data element to a minimum.

Blocking Operators

37

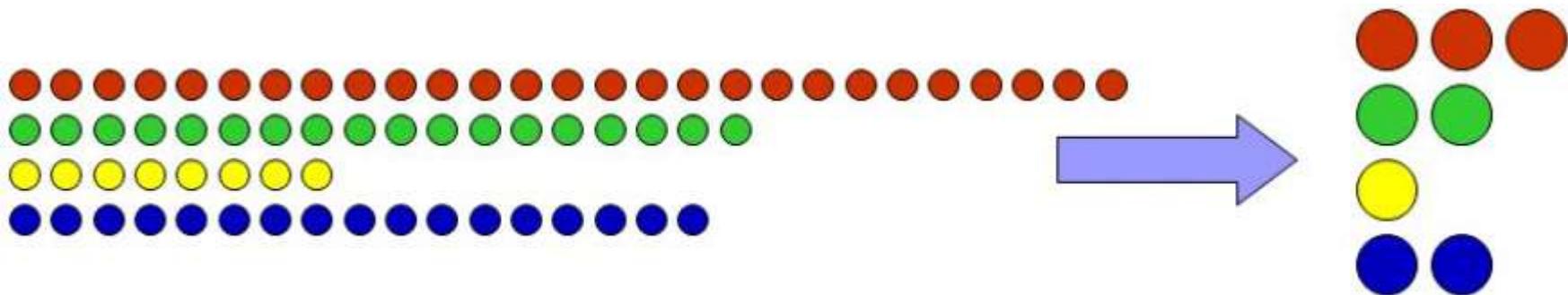
- ❑ A blocking query operator is a query operator that is unable to produce the first tuple of its output until it has seen its entire input. Sorting is an example of a blocking operator, as are aggregation operators such as SUM, COUNT, MIN, MAX, and AVG.
- ❑ If one thinks about evaluating continuous stream queries using a traditional tree of query operators, where data streams enter at the leaves and final query answers are produced at the root, then the incorporation of blocking operators into the query tree poses problems.
- ❑ Since continuous data streams may be infinite, a blocking operator that has a data stream as one of its inputs will never see its entire input, and therefore it will never be able to produce any output.
- ❑ Clearly, blocking operators are not very suitable to the data stream computation model, but aggregate queries are extremely common, and sorted data is easier to work with and can often be processed more efficiently than unsorted data.
- ❑ Doing away with blocking operators altogether would be problematic, but dealing with them effectively is one of the more challenging aspects of data stream computation.

Sampling in Data Streams



38

Sampling is a common practice for selecting a subset of data to be analyzed. Instead of dealing with an entire data stream, the instances at periodic intervals are selected. Sampling is used to compute statistics (expected values) of the stream. While sampling methods reduce the amount of data to process, and, by consequence, the computational costs, they can also be a source of errors. The main problem is to obtain a representative sample, a subset of data that has approximately the same properties of the original data.



Need & how of sampling - System cannot store the entire stream conveniently, so

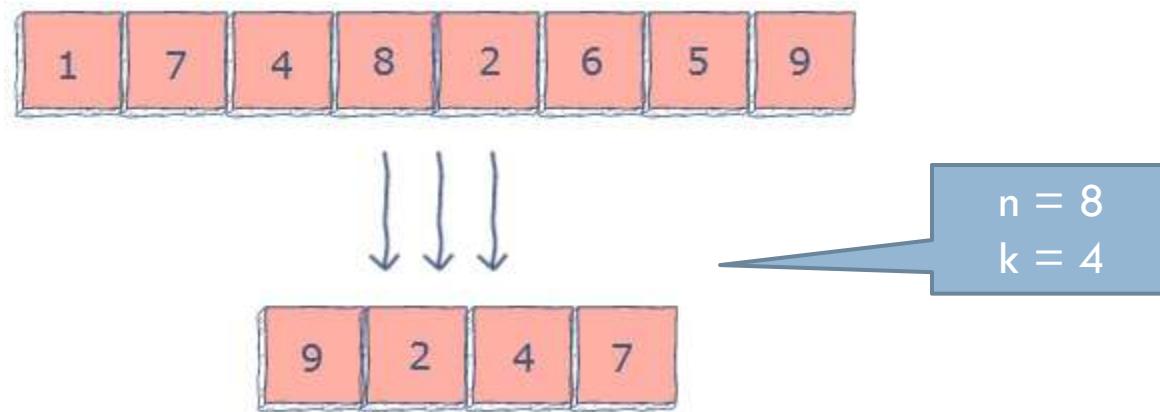
- ❑ how to make critical calculations about the stream using a limited amount of (primary or secondary) memory?
- ❑ Don't know how long the stream is, so when and how often to sample?

Three solutions namely (i) Reservoir , (ii) Biased Reservoir , and (iii) Concise sampling

Reservoir Sampling

39

Reservoir sampling is a randomized algorithm that is used to select k out of n samples; n is usually very large or unknown. For example, it can be used to obtain a sample of size k from a population of people with brown hair. This algorithm takes $O(n)$ to select k elements with uniform probability.



The key idea behind reservoir sampling is to create a 'reservoir' from a big ocean of data. Each element of the population has an equal probability of being present in the sample and that probability is (k/n) . With this key idea, a subsample to be created. It has to be noted, when a sample is created, the distributions should be identical not only row-wise but also column-wise, wherein columns are the features.

Reservoir Sampling Algorithm



40

0. Start

1. Create an array reservoir[0..k-1] and copy first k items of stream[] to it.
2. Iterate from k to n-1. In each iteration i:
 - 2.1. Generate a random number from 0 to i. Let the generated random number is j.
 - 2.2. If j is in range 0 to k-1, replace reservoir[j] with arr[i]
3. Stop

Illustration

Input:

The list of integer stream: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}, and the value of k = 6

Output:

k-selected items in the given array: 8 2 7 9 12 6

Biased Reservoir Sampling

41

In many cases, the stream data may evolve over time, and the corresponding data mining or query results may also change over time. Thus, the results of a query over a more recent window may be quite different from the results of a query over a more distant window. Similarly, the entire history of the data stream may not relevant for use in a repetitive data mining application such as classification. The simple reservoir sampling algorithm can be adapted to sample from a moving window over data streams. This is useful in many data stream applications where a small amount of recent history is more relevant than the entire previous stream. However, this can sometimes be an extreme solution, since for some applications we may need to sample from varying lengths of the stream history. While recent queries may be more frequent, it is also not possible to completely disregard queries over more distant horizons in the data stream. **Biased reservoir sampling** is a bias function to regulate the sampling from the stream. This bias gives a higher probability of selecting data points from recent parts of the stream as compared to distant past. This bias function is quite effective since it regulates the sampling in a smooth way so that the queries over recent horizons are more accurately resolved.

Concise sampling

42

Many a time, the size of the reservoir is sometimes restricted by the available main memory. It is desirable to increase the sample size within the available main memory restrictions. For this purpose, the technique of concise sampling is quite effective. Concise sampling exploits the fact that the number of distinct values of an attribute is often significantly smaller than the size of the data stream. In many applications, sampling is performed based on a single attribute in multi-dimensional data. For example, customer data in an e-commerce site sampling may be done based on only customer ids. The number of distinct customer ids is definitely much smaller than “n” the size of the entire stream.

The repeated occurrence of the same value can be exploited in order to increase the sample size beyond the relevant space restrictions. We note that when the number of distinct values in the stream is smaller than the main memory limitations, the entire stream can be maintained in main memory, and therefore, sampling may not even be necessary. For current systems in which the memory sizes maybe the order of several gigabytes, very large sample sizes can be main memory resident as long as the number of distinct values do not exceed the memory constraints.

Other Sampling Techniques

43

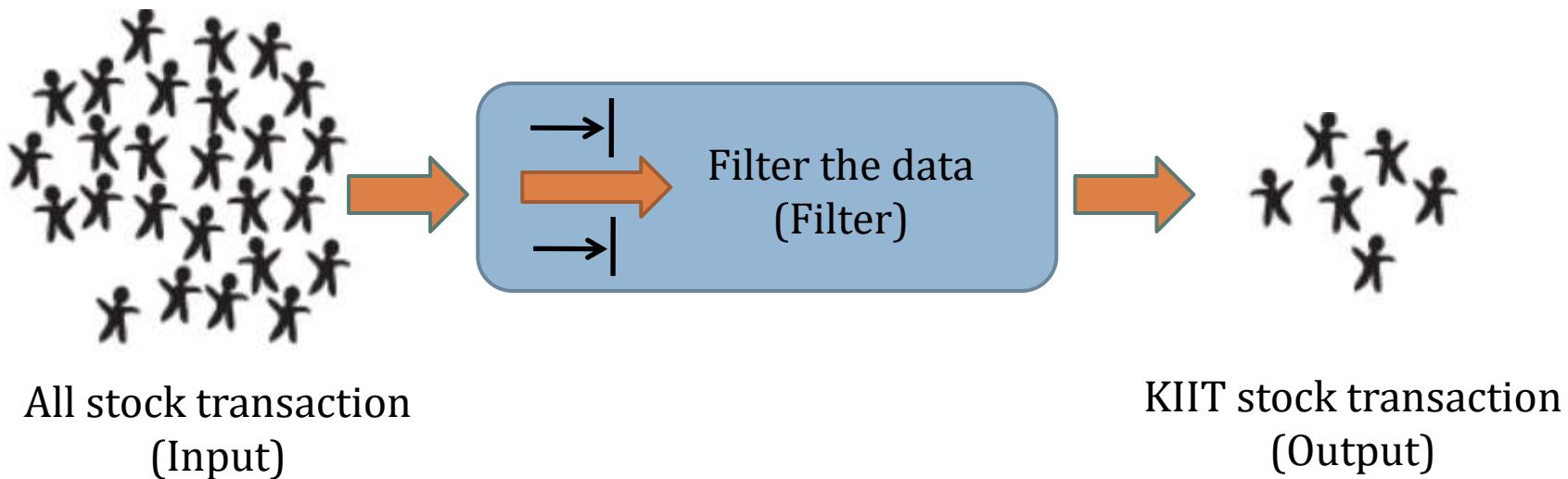
- Inverse Sampling
- Weighted Sampling
- Biased Sampling
- Priority Sampling
- Dynamic Sampling
- Chain Sampling



Filtering Stream

44

A filter is a program or section of code that is designed to examine each input or output stream for certain qualifying criteria and then process or forward it accordingly by producing another stream.



In this example, the streams processing application needs to filter the stock transaction data for KIIT transaction records.

Bloom Filter

45

Bloom filter is a space-efficient probabilistic data structure conceived by Burton Howard Bloom in 1970, that is used to test whether an element is a member of a set. **False positive** matches are possible, but **False Negatives are not** – in other words, a query returns either "**possibly in set**" or "**definitely not in set**"

False Positive = "**possibly in set**" or "**definitely not in set**"

False Negative = "**possibly not in set**" or "**definitely in set**"

Overview

x : An element

S: A set of elements

Input: x, S

Output:

- TRUE if x in S

- FALSE if x not in S

Bloom Filter cont...

46

Suppose you are creating an account on Facebook, you want to enter a cool username, you entered it and got a message, “Username is already taken”. You added your birth date along username, still no luck. Now you have added your university roll number also, still got “Username is already taken”. It’s really frustrating, isn’t it?

But have you ever thought how quickly Facebook check availability of username by searching millions of username registered with it. There are many ways to do this job –

Linear search : Bad idea!

Binary Search : There must be something better!!

Bloom Filter is a data structure that can do this job.

Bloom Filter cont...

47

For understanding bloom filters, one must know hashing. A hash function takes input and outputs a unique identifier of fixed length which is used for identification of input.

Properties of Bloom Filter

- ❑ Unlike a standard hash table, a Bloom filter of a fixed size can represent a set with an arbitrarily large number of elements.
- ❑ Adding an element never fails. However, the false positive rate increases steadily as elements are added until all bits in the filter are set to 1, at which point all queries yield a positive result.
- ❑ Bloom filters never generate false negative result, i.e., telling you that a username doesn't exist when it actually exists.
- ❑ Deleting elements from filter is not possible because, if we delete a single element by clearing bits at indices generated by k hash functions, it might cause deletion of few other elements.

Working of Bloom Filter

48

A empty bloom filter is a bit array of n bits, all set to zero, like below:

0	0	0	0	0	0	0	0	0	0
0	1	2	3	4	5	6	7	8	9

We need k number of hash functions to calculate the hashes for a given input. When we want to add an item in the filter, the bits at k indices $h_1(x), h_2(x), \dots, h_k(x)$ are set, where indices are calculated using hash functions.

Example – Suppose we want to enter “geeks” in the filter, we are using 3 hash functions and a bit array of length 10, all set to 0 initially. First we’ll calculate the hashes as following :

$$h_1(\text{"geeks"}) \% 10 = 1, h_2(\text{"geeks"}) \% 10 = 4, \text{ and } h_3(\text{"geeks"}) \% 10 = 7$$

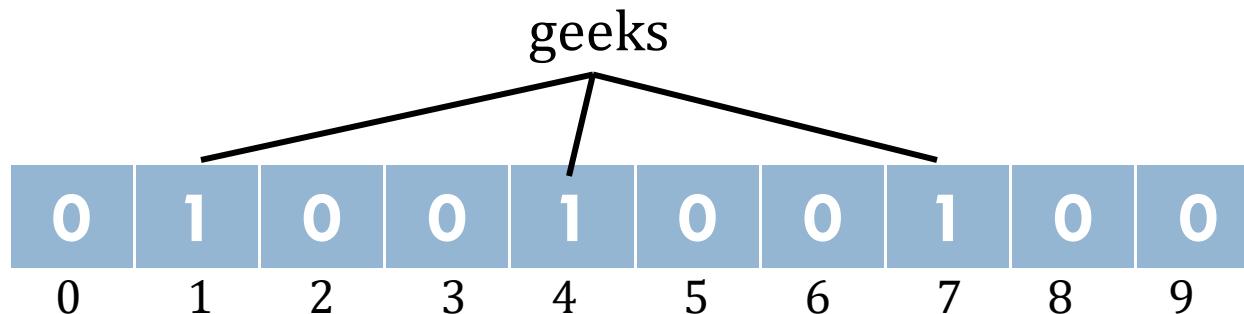
Note: *These outputs are random for explanation only.*

Working of Bloom Filter cont...



49

Now we will set the bits at indices 1, 4 and 7 to 1



Again we want to enter “nerd”, similarly we’ll calculate hashes

$$h_1(\text{"nerd"}) \% 10 = 3$$

$$h_2(\text{"nerd"}) \% 10 = 5$$

$$h_3(\text{"nerd"}) \% 10 = 4$$

Note: *These outputs are random for explanation only.*

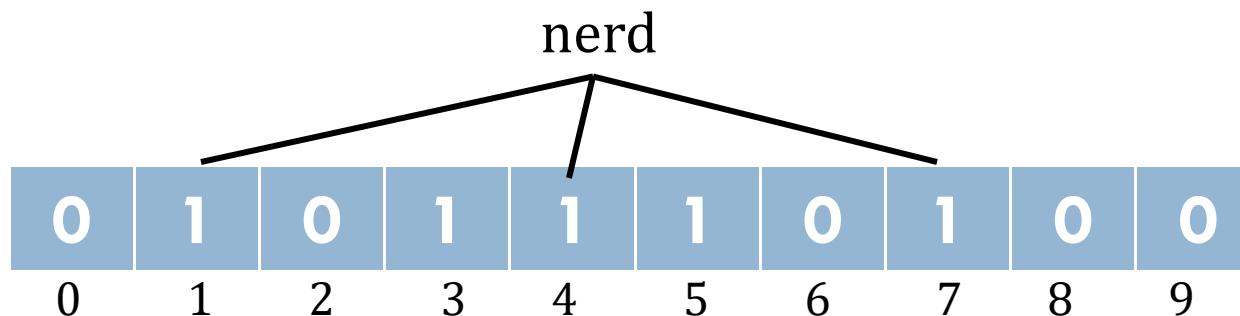
Set the bits at indices 3, 5 and 4 to 1

Working of Bloom Filter cont...



50

Now we will set the bits at indices 3, 5 and 4 to 1



Now if we want to check “geeks” is present in filter or not. We’ll do the same process but this time in reverse order. We calculate respective hashes using h_1 , h_2 and h_3 and check if all these indices are set to 1 in the bit array. If all the bits are set then we can say that “geeks” is probably present. If any of the bit at these indices are 0 then “geeks” is definitely not present.

False Positive in Bloom Filters



51

The question is why we said “probably present”, why this uncertainty. Let’s understand this with an example. Suppose we want to check whether “cat” is present or not. We’ll calculate hashes using h_1 , h_2 and h_3

$$h_1(\text{"cat"}) \% 10 = 1$$

$$h_2(\text{"cat"}) \% 10 = 3$$

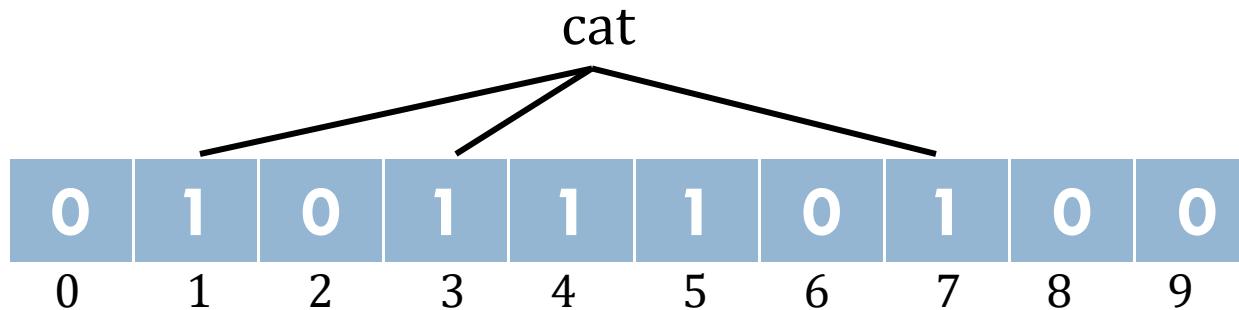
$$h_3(\text{"cat"}) \% 10 = 7$$

If we check the bit array, bits at these indices are set to 1 but we know that “cat” was never added to the filter. Bit at index 1 and 7 was set when we added “geeks” and bit 3 was set when we added “nerd”.

False Positive in Bloom Filters cont...



52



So, because bits at calculated indices are already set by some other item, bloom filter erroneously claim that “cat” is present and generating a false positive result. Depending on the application, it could be huge downside or relatively okay.

We can control the probability of getting a false positive by controlling the size of the Bloom filter. More space means fewer false positives. If we want decrease probability of false positive result, we have to use more number of hash functions and larger bit array. This would add latency in addition of item and checking membership.

Bloom Filter Algorithm

53

Insertion

Data: e is the element to insert into the Bloom filter.

```
insert(e)
begin
/* Loop all hash functions k */
for j : 1 ... k do
    m ← hj(e) //apply the hash function on e
    Bm ← bf[m] //retrieve val at mth pos from Bloom filter bf
    if Bm == 0 then
        /* Bloom filter had zero bit at index m */
        Bm ← 1;
    end if
end for
end
```

Lookup

Data: x is the element for which membership is tested.

```
bool isMember(x) /* returns true or false to the membership test */
begin
    t ← 1
    j ← 1
    while t == 1 and j ≤ k do
        m ← hj(x)
        Bm ← bf[m]
        if Bm == 0 then
            t ← 0
        end if
        j ← j + 1;
    end while
    return (bool) t
end
```

Bloom Filter Performance



54

A Bloom filter requires space $O(n)$ and can answer membership queries in $O(1)$ time where n is number item inserted in the filter. Although the asymptotic space complexity of a Bloom filter is the same as a hash map, $O(n)$, a Bloom filter is more space efficient.

Class Exercise



- A empty bloom filter is of size 11 with 4 hash functions namely
 - $h_1(x) = (3x+ 3) \bmod 6$
 - $h_2(x) = (2x+ 9) \bmod 2$
 - $h_3(x) = (3x+ 7) \bmod 8$
 - $h_4(x) = (2x+ 3) \bmod 5$

Illustrate bloom filter insertion with 7 and then 8.

Perform bloom filter lookup/membership test with 10 and 48

False Positive in Bloom Filters cont...



55

0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	2	3	4	5	6	7	8	9	10	11	

K=4 (# of Hash Function)

Note: *These outputs are random for explanation only.*

INSERT (x_1), $x_1 = 7$

$h_1(x_1) = 0$

$h_2(x_1) = 1$

$h_3(x_1) = 4$

$h_4(x_1) = 2$

INSERT (x_2), $x_2 = 8$

$h_1(x_2) = 3$

$h_2(x_2) = 1$

$h_3(x_2) = 7$

$h_4(x_2) = 4$

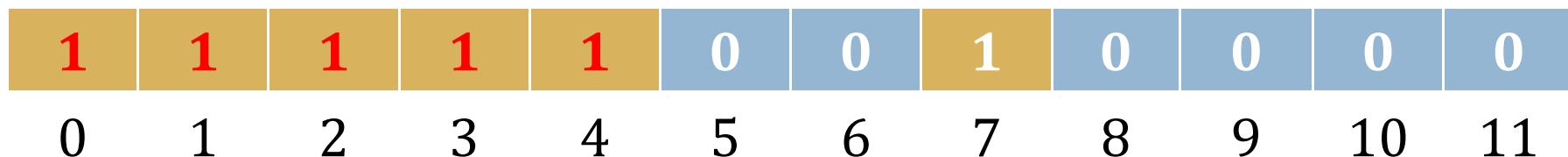
State of bloom filter post to the insertion of x_1 and x_2

1	1	1	1	1	0	0	1	0	0	0	0
0	1	2	3	4	5	6	7	8	9	10	11

False Positive in Bloom Filters cont'd



56



LOOKUP(x_3), $x_3 = 10$

$$h_1(x_3) = 3$$

$$h_2(x_3) = 1$$

$$h_3(x_3) = 5$$

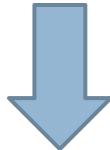
LOOKUP(x_4), $x_4 = 48$

$$h_1(x_4) = 3$$

$$h_2(x_4) = 1$$

$$h_3(x_4) = 7$$

$$h_4(x_4) = 4$$



**X_3 doesn't
exist**



**X_4 - Case of
FALSE POSITIVE**

Optimum number of hash functions



57

The number of hash functions k must be a positive integer. If n is size of bit array and m is number of elements to be inserted, then k can be calculated as :

$$k = \frac{n}{m} \ln(2)$$

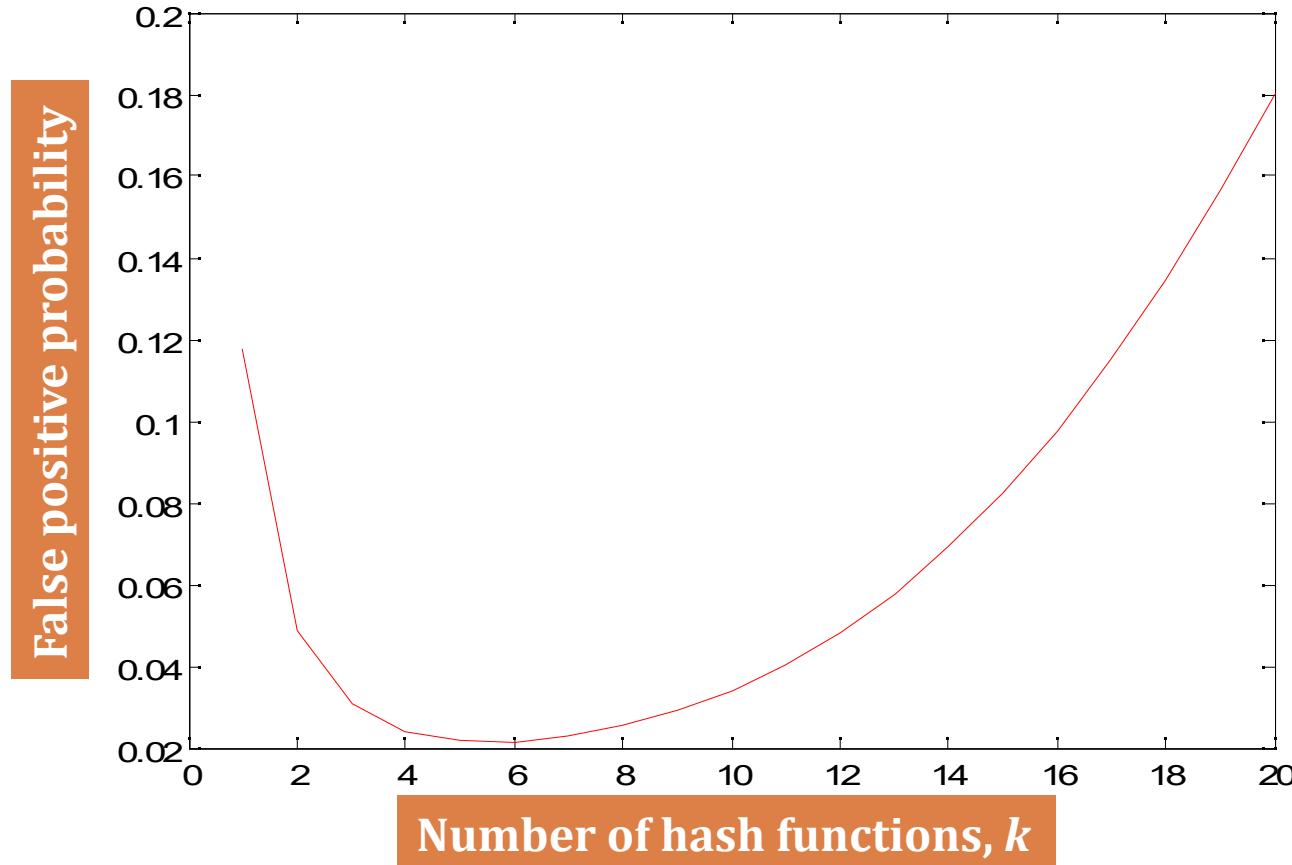


Class Exercise

Calculate the optimal number of hash functions for 10 bit length bloom filter having 3 numbers of input elements.

What happens to increasing k ?

58



Calculating probability of False Positives



59

- ❑ Probability that a slot is hashed = $1/n$
- ❑ Probability that a slot is not hashed = $1 - (1/n)$
- ❑ Probability that a slot is not hashed after insertion of an element for all the k hash function is:

$$\left(1 - \frac{1}{n}\right)^k$$

- ❑ Probability that a slot is not set to 1 after insertion of m element is:

$$\left(1 - \frac{1}{n}\right)^{km}$$

- ❑ Probability that a slot is set to 1 after insertion of m element is:

$$1 - \left(1 - \frac{1}{n}\right)^{km}$$

Calculating probability of False Positives cont'd



60

Let n be the size of bit array, k be the number of hash functions and m be the number of expected elements to be inserted in the filter, then the probability of false positive p can be calculated as:

$$\left(1 - \left(\frac{1}{e}\right)^{\frac{km}{n}}\right)^k$$

Class Exercise

Calculate the probability of false positives with table size 10 and no. of items to be inserted are 3.

Counting distinct elements in a stream – Approach 1



61

(1, 2, 2, 1, 3, 1, 5, 1, 3, 3, 3, 2, 2)

Number of distinct elements = 4

How to calculate?

1. Initialize the hashtable (large binary array) of size n with all zeros.
2. Choose the hash function $h_i : i \in \{1, \dots, k\}$
3. For each flow label $f \in \{1, \dots, m\}$, compute $h(f)$ and mark that position in the hashtable with 1
4. Count the number of positions in the hashtable with 1 and call it c.
5. The number of distinct items is $m * \ln(m / (m - c))$

Class Exercise



Count the distinct elements in a data stream of elements {1, 2, 2, 1, 3, 1, 5, 1, 3, 3, 3, 2, 2} with the hash function $h(x) = (5x+1) \bmod 6$ of size 11.

Counting distinct elements in a stream using Flajolet-Martin algorithm – Approach 2



62

Count the distinct elements in a data stream of elements {6,8,4,6,3,4} with hash function $h(x) = (5x+1) \bmod 6$ of size 11.

How to calculate?

Step 1:

Apply Hash function(s) to the data stream and compute the slots.

$h(6)=1$, $h(8)=5$, $h(4)=3$, $h(6)=1$, $h(3)=4$, $h(4)=3$.

The slot numbers obtained are:{1,5,3,1,4,3}

Step 2: Convert the numbers to binary

$h(6)=1=001$, $h(8)=5=101$, $h(4)=3=001$, $h(6)=1=001$, $h(3)=4=100$,
 $h(4)=3=011$

Step 3: Calculate the maximum trailing zeros

$TZ = \{0, 0, 0, 0, 2, 0\}$ /* TZ stands for Trailing Zeros */

$R = \text{MAX}(TZ) = \text{MAX}(0, 0, 0, 0, 2, 0) = 2$

Step 4: Estimate the distinct elements with the formula 2^R

Number of distinct elements = $2^R = 2^2 = 4$

Bloom Filter Use Cases



63

- ❑ Bitcoin uses Bloom filters to speed up wallet synchronization and also to improve Bitcoin wallet security
- ❑ Google Chrome uses the Bloom filter to identify malicious URLs - it keeps a local Bloom filter as the first check for Spam URL
- ❑ Google BigTable and Apache Cassandra use Bloom filters to reduce disk lookups for non-existent rows or columns
- ❑ The Squid Web Proxy Cache uses Bloom filters for cache digests - proxies periodically exchange Bloom filters for avoiding ICP messages
- ❑ Genomics community uses Bloom filter for classification of DNA sequences and efficient counting of k-mers in DNA sequences
- ❑ Used for preventing weak password choices using a dictionary of easily guessable passwords as bloom filter
- ❑ Used to implement spell checker using a predefined dictionary with large number of words

Other Types of Bloom Filter



64

- ❑ **Compressed Bloom Filter** - Using a larger but sparser Bloom Filter can yield the same false positive rate with a smaller number of transmitted bits.
- ❑ **Scalable Bloom Filter** - A Scalable Bloom Filters consist of two or more Standard Bloom Filters, allowing arbitrary growth of the set being represented.
- ❑ **Generalized Bloom Filter** - Generalized Bloom Filter uses hash functions that can set as well as reset bits.
- ❑ **Stable Bloom Filter** - This variant of Bloom Filter is particularly useful in data streaming applications.

Estimating Moments

65

- ❑ Assume a stream A of length N which is composed of m different types of items a_1, \dots, a_m each of which repeats itself n_1, \dots, n_m times (in arbitrary order). The frequency moments of order k, f_k is defined as

$$f_k = \sum_{i=1}^m n_i^k$$

- ❑ The 0th order moment i.e. f_0 is the number of distinct elements in the stream.
- ❑ The 1st order moment i.e. f_1 is the length of the stream.
- ❑ The 2nd order moment f_2 is an important quantity which represent show “skewed” the distribution of the elements in stream is.

Example

Consider the stream a, b, c, b, d, a, c, d, a, b, d, c, a, a, b wherein $n_a = 5$, $n_b = 4$, $n_c = 3$ and $n_d = 3$. In this case:

- ❑ $f_0 = \text{number of distinct elements} = 4$
- ❑ $f_1 = \text{length of the stream} = n_a + n_b + n_c + n_d = 5 + 4 + 3 + 3 = 15$
- ❑ $f_2 = n_a^2 + n_b^2 + n_c^2 + n_d^2 = 5^2 + 4^2 + 3^2 + 3^2 = 59$

Real-Time Analytics Platform (RTAP)



66

- ❑ Real-time analytics makes use of all available data and resources when they are needed, and it consists of dynamic analysis and reporting, based on data entered into a system before the actual time of use.
- ❑ Real-time denotes the ability to process data as it arrives, rather than storing the data and retrieving it at some point in the future.
- ❑ For example, consider an e-merchant like Flipkart or Snapdeal; real time means the time elapsed from the time a customer enters the website to the time the customer logs out. Any analytics procedure, like providing the customer with recommendations or offering a discount based on current value in the shopping cart, etc., will have to be done within this timeframe which may be about 15 minutes to an hour.
- ❑ But from the point of view of a military application where there is constant monitoring say of the air space, time needed to analyze a potential threat pattern and make decision maybe a few milliseconds.

RTAP cont...

67

Real-Time Analytics is thus discovering meaningful patterns in data for something urgent. There are two specific and useful types of real-time analytics i.e. On-Demand and Continuous.

- ❑ **On-Demand Real-Time Analytics** is reactive because it waits for users to request a query and then delivers the analytics. This is used when someone within a company needs to take a pulse on what is happening right this minute. For instance, a movie producer may want to monitor the tweets and identify sentiments about his movie on the first day first show and be prepared for the outcome.
- ❑ **Continuous Real-Time Analytics** is more proactive and alerts users with continuous updates in real time. The best example could be monitoring the stock market trends and provide analytics to help users make a decision to buy or sell all in real time.

THANK YOU!

Data Analytics (IT-3006)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering



Strictly for internal circulation (within KIIT) and reference only. Not for outside circulation without permission

3 Credit

Lecture Note – Unit 4

Course Contents



2

Sr #	Major and Detailed Coverage Area	Hrs
4	Frequent Itemsets and Clustering Introduction to Frequent Itemsets, Market-Basket Model, Algorithm for Finding Frequent Itemsets, Association Rule Mining, Apriori Algorithm, Introduction to Clustering, Overview of Clustering Techniques, Hierarchical Clustering, Partitioning Methods , K- Means Algorithm, Clustering High-Dimensional Data .	10

Introduction

3

- ❑ Data mining is the process of analyzing massive volumes of data to discover business intelligence that helps business to solve problems, mitigate risks, and seize new opportunities.
- ❑ Frequent itemsets play an essential role in many data mining tasks where the need is to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and many more.
- ❑ One of the most popular applications of frequent itemset mining is discovery of association rules. The identification of sets of items, products, symptoms, characteristics and so forth that often occur together in the given database can be seen as one of the most basic tasks in data mining.
- ❑ For instance, customers of an on-line bookstore could be considered examples, each represented by the set of books he or she has purchased. A set of books, such as {"Machine Learning," "The Elements of Statistical Learning," "Pattern Classification,"} is a frequent itemset if it has been bought by sufficiently many customers.
- ❑ Many techniques have been invented to mine databases for frequent events. These techniques work well in practice on smaller datasets, but are not suitable for truly big data. Applying frequent itemset mining to large databases is a challenge as very large databases do not fit into main memory.

Frequent Itemsets

4

Let $I = \{i_1, \dots, i_k\}$ be a set of items. Let D , be a set of transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$.

Transaction	TID	Items bought
T1	10	Beer, Nuts, Diaper
T2	20	Beer, Coffee, Diaper
T3	30	Beer, Diaper, Eggs
T4	40	Nuts, Eggs, Milk
T5	50	Nuts, Coffee, Diaper, Eggs, Milk

What is k-itemset?

- When $k=1$, then k-Itemset is itemset 1.
- When $k=2$, then k-Itemset is itemset 2.
- When $k=3$, then k-Itemset is itemset 3.
- When $k=4$, then k-Itemset is itemset 4.
- When $k=5$, then k-Itemset is itemset 5.

What is a frequent itemset?

An itemset is frequent if its support is no less than “minimum support”. Minimum support is always according to the choice. One can select any minimum support to decide that the itemset is frequent or not. Support of an item is the number of transactions in which the item appears. E.g. support of Beer is 3, support of Eggs is 3, support for Coffee is 2 and so on. If the minimum support is 3, then the frequent itemset are {Beer}, {Nuts}, {Diaper}, {Eggs}, {Beer, Diaper}.

Class Exercise

5

Items = {milk (m), coke (c), pepsi (p), beer (b), juice (j)}

Minimum support $s = 3$

Transactions

1. $T_1 = \{m, c, b\}$
2. $T_2 = \{m, p, j\}$
3. $T_3 = \{m, b\}$
4. $T_4 = \{c, j\}$
5. $T_5 = \{m, p, b\}$
6. $T_6 = \{m, c, b, j\}$
7. $T_7 = \{c, b, j\}$
8. $T_8 = \{b, c\}$



Determine the frequent Itemsets.

Market-Basket Model



6

What is Market Basket Analysis?

- It is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.
- It creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased. The rules are probabilistic in nature or, in other words, they are derived from the frequencies of co-occurrence in the observations. Frequency is the proportion of baskets that contain the items of interest. The rules can be used in pricing strategies, product placement, and various types of cross-selling strategies.
- Market Basket Analysis takes data at transaction level, which lists all items bought by a customer in a single purchase. The technique determines relationships of what products were purchased with which other product(s). These relationships are then used to build profiles containing If-Then rules of the items purchased.
- The rules could be written as: If {A} Then {B}. The If part of the rule (the {A}) is known as the **antecedent** and the THEN part of the rule is known as the **consequent** (the {B}). The antecedent is the condition and the consequent is the result.

What is Market Basket Analysis cont...



7

- ❑ For example, you are in a supermarket to buy milk. Referring to the below example, there are nine baskets containing varying combinations of milk, cheese, apples, and bananas.
- ❑ **Question** - are you more likely to buy apples or cheese in the same transaction than somebody who did not buy milk?

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

- ❑ The next step is to determine the relationships and the rules. So, association rule mining is applied in this context. It is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

Market-Basket Model cont...



8

- ❑ The association rule has three measures that express the degree of confidence in the rule, i.e. Support, Confidence, and Lift. Since the market-basket has its origin in retail application, it is sometimes called **transaction**.
- ❑ **Support:** The number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
Example: Referring to the earlier dataset, $\text{Support}(\text{milk}) = 6/9$, $\text{Support}(\text{cheese}) = 7/9$, $\text{Support}(\text{milk} \& \text{cheese}) = 6/9$. This is often expressed as $\text{milk} \Rightarrow \text{cheese}$ i.e. bought milk and cheese together.
- ❑ **Confidence:** It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
Example: Referring to the earlier dataset, $\text{Confidence}(\text{milk} \Rightarrow \text{cheese}) = (\text{milk} \& \text{cheese}) / (\text{milk}) = 6/6$.
- ❑ **Lift:** The lift of the rule $A \Rightarrow B$ is the confidence of the rule divided by the expected confidence, assuming that the itemsets A and B are independent of each other.
Example: Referring to the earlier dataset, $\text{Lift}(\text{milk} \Rightarrow \text{cheese}) = [(\text{milk} \& \text{cheese}) / (\text{milk})] / [\text{cheese} / \text{Total}] = [6/6] / [7/9] = 1 / 0.777$.

Class Exercise

9

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

Basket	Item(s)	How many items?	Support	Confidence	Lift
	Milk	?	?		
	Cheese	?	?		
1	Milk => Cheese	?	?	?	?
	Apple, Milk	?	?		
2	(Apple, Milk) => Cheese	?	?	?	?
	Apple, Cheese	?	?		
3	(Apple, Cheese) => Milk	?	?	?	?

Class Exercise cont...

10

$$\text{Support} = \frac{(A + B)}{\text{Total}}$$

$$\text{Support for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Total}} = \frac{6}{9} = .6666667$$

$$\text{Confidence} = \frac{(A + B)}{A}$$

$$\text{Confidence for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Milk}} = \frac{6}{6} = 1.000$$

$$\text{Lift} = \left(\frac{\left(\frac{(A + B)}{A} \right)}{\left(\frac{B}{\text{Total}} \right)} \right)$$

$$\text{Lift for Basket 1} = \left(\frac{\left(\frac{(\text{Milk} + \text{Cheese})}{\text{Milk}} \right)}{\left(\frac{\text{Cheese}}{\text{Total}} \right)} \right) = \left(\frac{\left(\frac{6}{6} \right)}{\left(\frac{7}{9} \right)} \right) = \left(\frac{1}{.7777778} \right) = 1.2857$$

Class Exercise

11

Transaction ID	Grapes	Apple	Mango	Orange
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

- Support(Grapes) ?
- Confidence({Grapes, Apple} => {Mango}) ?
- Lift ({Grapes, Apple} => {Mango}) ?

Market-Basket Model cont...



12

The likelihood of a customer buying both A and B together is ‘lift-value’ times more than the chance if purchasing alone.

- Lift ($A \Rightarrow B$) = 1 means that there is no correlation within the itemset.
- Lift ($A \Rightarrow B$) > 1 means that there is a positive correlation within the itemset, i.e., products in the itemset, A, and B, are more likely to be bought together.
- Lift ($A \Rightarrow B$) < 1 means that there is a negative correlation within the itemset, i.e., products in itemset, A, and B, are unlikely to be bought together.

The Apriori Algorithm

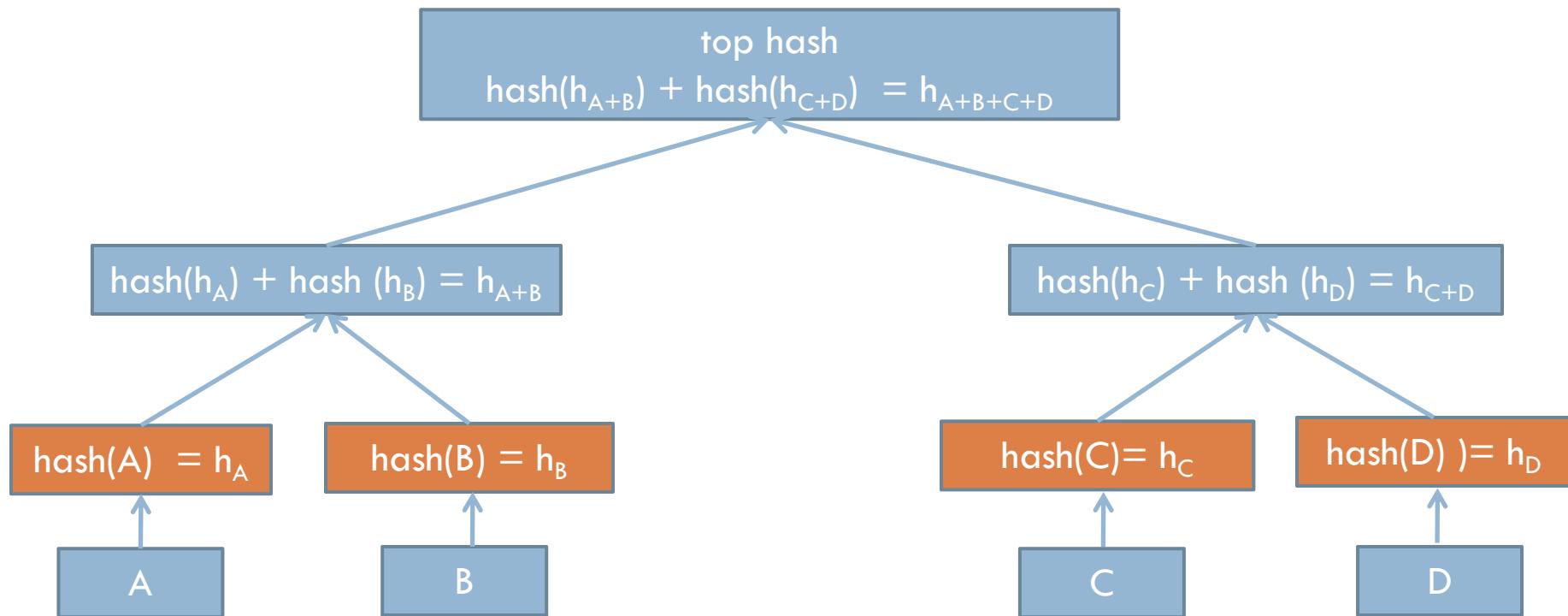
13

- ❑ Imagine you're at the supermarket, and in your mind, you have the items you wanted to buy. But you end up buying a lot more than you were supposed to. This is called **impulsive buying** and brands use the **Apriori algorithm** to leverage this phenomenon.
- ❑ The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions.
- ❑ With the help of these association rule, it determines how strongly or how weakly two objects are connected.
- ❑ It is the iterative process for finding the frequent itemsets from the large dataset.
- ❑ This algorithm uses a **breadth-first search** and **Hash Tree** to calculate the itemset associations efficiently.
- ❑ It is mainly used for market basket analysis and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Hash Tree

14

- ❑ Hash tree is a data structure used for data verification and synchronization.
- ❑ It is a tree data structure where each non-leaf node is a hash of its child nodes. All the leaf nodes are at the same depth and are as far left as possible.
- ❑ It is also known as Merkle Tree.



The Apriori Algorithm cont...

15

Apriori says the probability that item I is not frequent if:

- $P(I) <$ minimum support threshold, then I is not frequent.
- $P(I+A) <$ minimum support threshold, then $I+A$ is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below minimum support, and thus can be ignored. This property is called the Antimonotone property.

The steps followed in the Apriori algorithm of data mining are:

- Join Step:** This step generates $(k+1)$ itemset from k-itemsets by joining each item with itself.
- Prune Step:** This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Steps for Apriori Algorithm

16

Below are the main steps for the Apriori algorithm:

1. Calculate the support of itemsets (of size $k = 1$) in the transactional database. This is called generating the candidate set.
2. Prune the candidate set by eliminating items with a support less than the given threshold.
3. Join the frequent itemsets to form sets of size $k + 1$, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support.

Note:

Support refers to items' frequency of occurrence and confidence is a conditional probability.

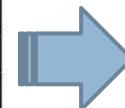
Apriori Algorithm Example

17

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given: Minimum Support= 2, Minimum Confidence= 50%

Now, we will take out all the itemsets that have the greater support count than the minimum support (2). It will give us the table for the frequent itemset **L1**. Since all the itemsets have greater or equal support count than the minimum support, except the E, so E itemset will be removed.



Step-1: Calculating C1 and L1:

In the first step, a table is created that contains support count (The frequency of each itemset individually in the dataset) of each itemset in the given dataset. This table is called the candidate set or C1.



C1

Itemset	Support Count
A	6
B	7
C	6
D	2
E	1



Apriori Algorithm Example cont...



18

L1

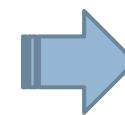
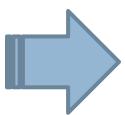
Itemset	Support Count
A	6
B	7
C	6
D	2

C2

Itemset	Support Count
{A, B}	4
{A,C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

Step-2: Candidate Generation C2, and L2:

In this step, we will generate C2 with the help of L1. In C2, we will create the pair of the itemsets of L1 in the form of subsets. After creating the subsets, we will again find the support count from the main transaction table of datasets, i.e., how many times these pairs have occurred together in the given dataset. So, we will get the below table for C2:



Again, we need to compare the C2 Support count with the minimum support count, and after comparing, the itemset with less support count will be eliminated from the table C2. It will give us the below table for L2. In this case, {A,D}, {C,D} itemset will be removed.

Apriori Algorithm Example cont...



19

L2

Itemset	Support Count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2



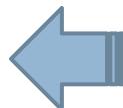
Step-3: Candidate Generation C3, and L3:

For C3, we will repeat the same two processes, but now we will form the C3 table with subsets of three itemsets together, and will calculate the support count from the dataset. It will give the below table:



C3

Now we will create the L3 table. As we can see from the above C3 table, there is only one combination of itemset that has support count equal to the minimum support count. So, the L3 will have only one combination, i.e., {A, B, C}.



L3

Itemset	Support Count
{A, B, C}	2

Itemset	Support Count
{A, B, C}	2
{B, C, D}	0
{A, C, D}	0
{A, B, D}	0

Apriori Algorithm Example cont...



20

Step-4: Finding the association rules for the subsets:

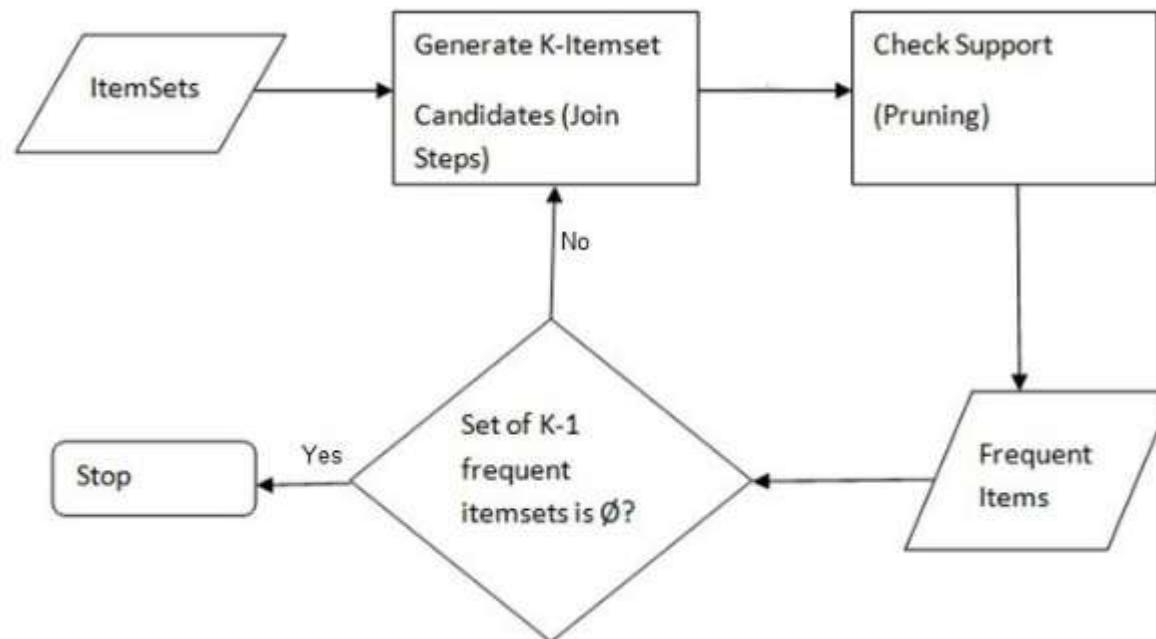
To generate the association rules, first, we will create a new table (AR) with the possible rules from the occurred combination. For all the rules, we will calculate the Confidence using formula $\text{sup}(A \wedge B)/A$. After calculating the confidence value for all rules, we will exclude the rules that have less confidence than the minimum threshold (50%).

Rules	Support	Confidence	AR
$A \wedge B \rightarrow C$	2	$\text{sup}\{(A \wedge B) \wedge C\}/\text{sup}(A \wedge B) = 2/4 = 0.5 = 50\%$	
$B \wedge C \rightarrow A$	2	$\text{sup}\{(B \wedge C) \wedge A\}/\text{sup}(B \wedge C) = 2/4 = 0.5 = 50\%$	
$A \wedge C \rightarrow B$	2	$\text{sup}\{(A \wedge C) \wedge B\}/\text{sup}(A \wedge C) = 2/4 = 0.5 = 50\%$	
$C \rightarrow A \wedge B$	2	$\text{sup}\{(C \wedge (A \wedge B))\}/\text{sup}(C) = 2/5 = 0.4 = 40\%$	
$A \rightarrow B \wedge C$	2	$\text{sup}\{(A \wedge (B \wedge C))\}/\text{sup}(A) = 2/6 = 0.33 = 33.33\%$	
$B \rightarrow B \wedge C$	2	$\text{sup}\{(B \wedge (B \wedge C))\}/\text{sup}(B) = 2/7 = 0.28 = 28\%$	

As the given threshold or minimum confidence is 50%, so the first three rules $A \wedge B \rightarrow C$, $B \wedge C \rightarrow A$, and $A \wedge C \rightarrow B$ can be considered as the strong association rules for the given problem.

Apriori Algorithm Flow

21



Class Exercise



22

Transaction	List of items
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

Minimum support = 2
Minimum confidence = 50%



Find frequent itemsets and generate association rules for them.
Illustrate it with step-by-step process.

Class Exercise

23

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

Choose minimum support and minimum confidence to your choice.

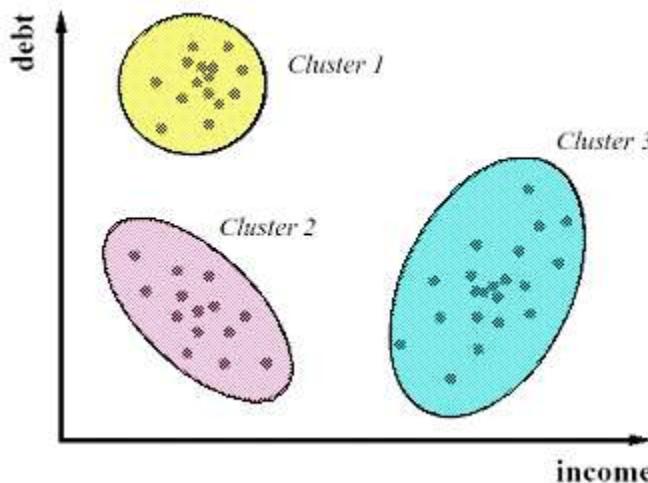


Find frequent itemsets.

Clustering

24

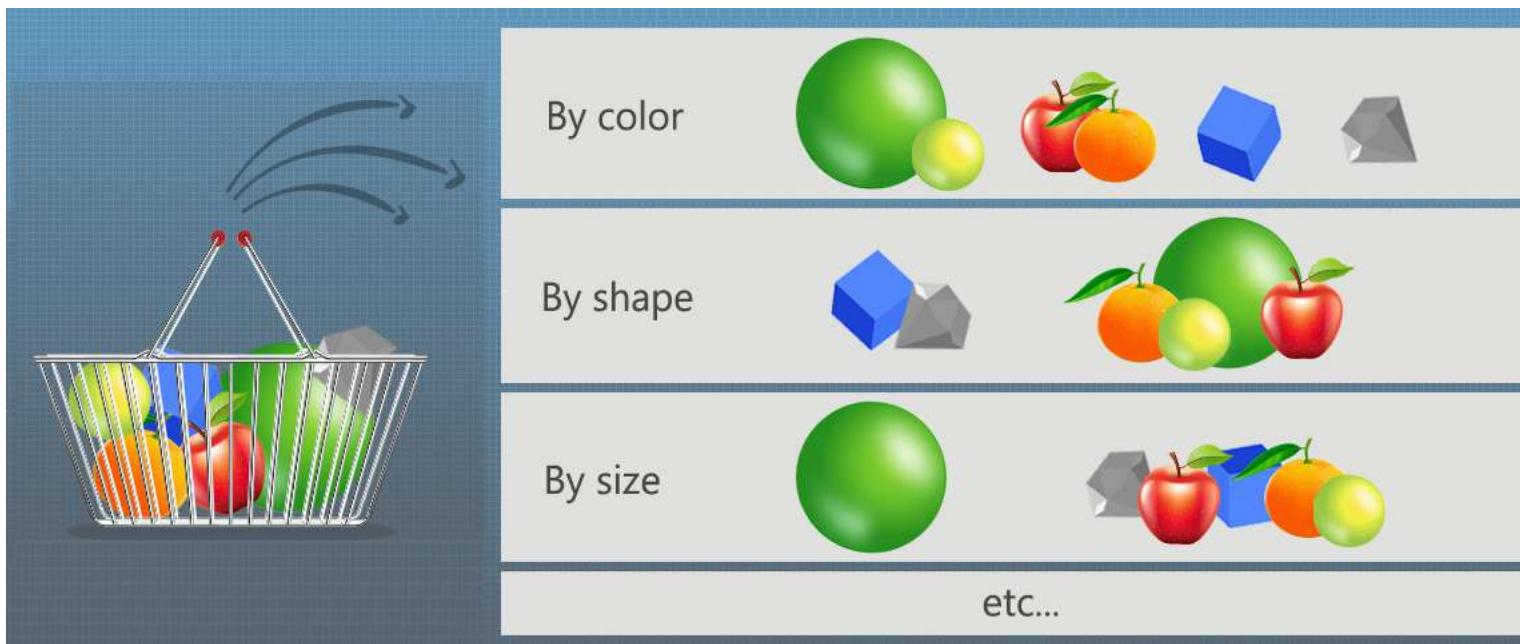
- ❑ Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- ❑ It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- ❑ Following is an example of finding clusters of population based on their income and debt.



Clustering Example

25

Imagine of a number of objects in a basket. Each item has a distinct set of features (size, shape, color, etc.). Now the task is to group each of the objects in the basket. A natural first question to ask is, 'on what basis these objects should be grouped?' Perhaps size, shape, or color.

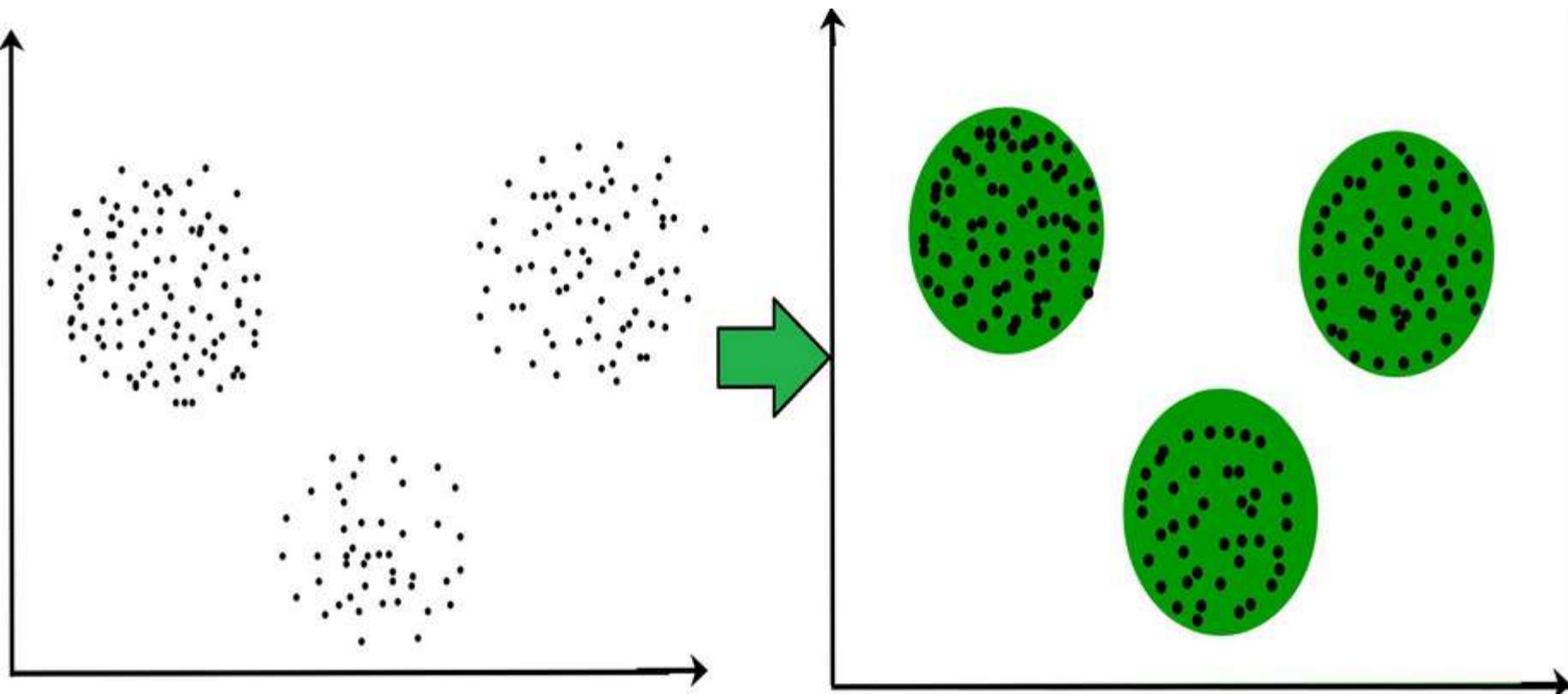


Clustering Example cont...



26

The data points clustered together can be classified into one single group. The clusters can be distinguished, and can identify that there are 3 clusters.

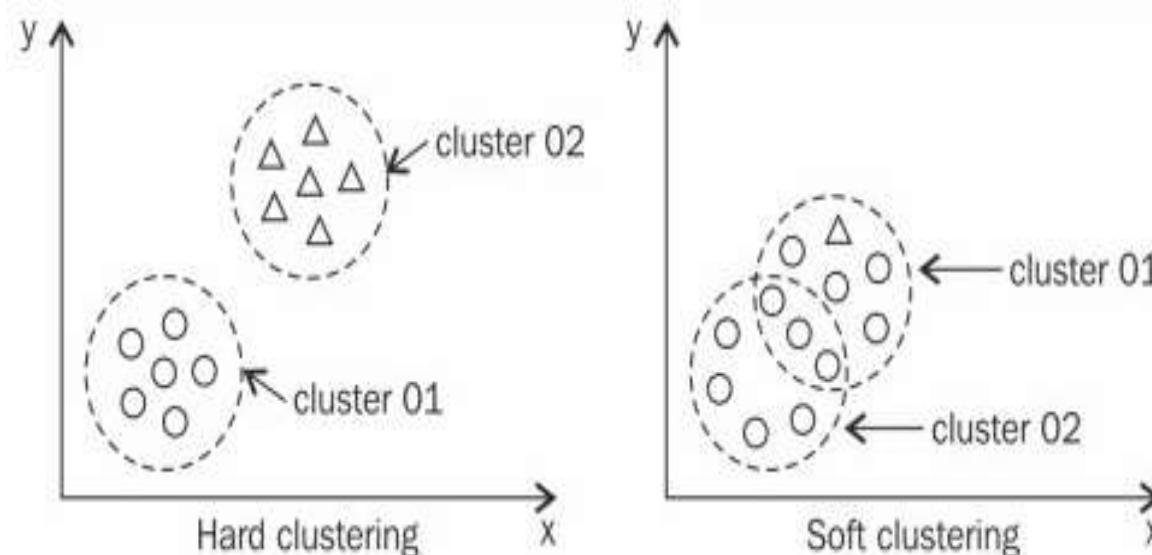


Clustering Types

27

Clustering can be divided into two subgroups:

- ❑ **Hard Clustering:** each data point either belongs to a cluster completely or not.
- ❑ **Soft Clustering:** Instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.



Clustering Application

28

- ❑ **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
- ❑ **Biology:** It can be used for classification among different species of plants and animals.
- ❑ **Libraries:** It is used in clustering different books on the basis of topics and information.
- ❑ **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.
- ❑ **City Planning:** It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
- ❑ **Earthquake studies:** By learning the earthquake affected areas, the dangerous zones can be determined.
- ❑ **Healthcare:** It can be used in identifying and classifying the cancerous gene.
- ❑ **Search Engine:** It is the backbone behind the search engines. Search engines try to group similar objects in one cluster and the dissimilar objects far from each other.
- ❑ **Education:** It can be used to monitor the students' academic performance. Based on the students' score they are grouped into different-different clusters, where each clusters denoting the different level of performance.

Requirements of Clustering in Data Mining



29

- Scalability** – Highly scalable clustering algorithms are needed to deal with large databases.
- Ability to deal with different kinds of attributes** - Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- Discovery of clusters with attribute shape** - The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- High dimensionality** - The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data** – Dataset contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretability** – The clustering results should be interpretable, comprehensible, and usable.

Types of clustering algorithms

30

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarity' among data points. The popular types are:

- Connectivity models Distribution models
- Centroid models Density models

Connectivity models: These models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

Centroid models: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset.

Types of clustering algorithms cont...

31

Distribution models: These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

Density Models: These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

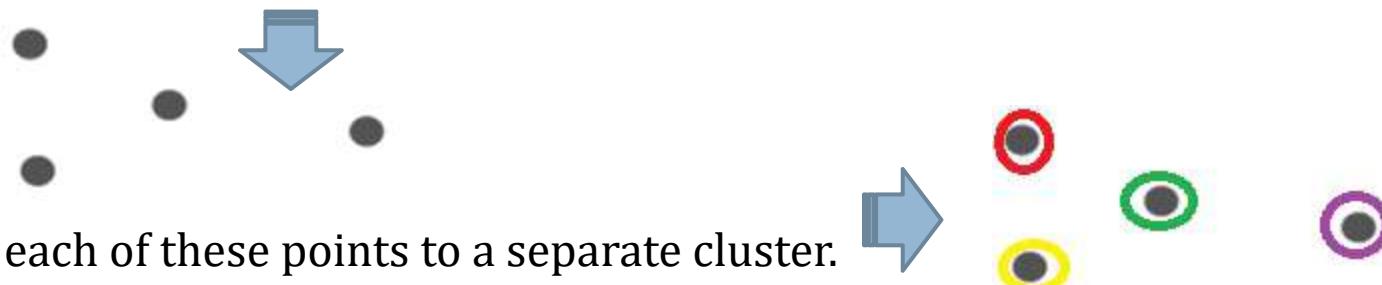
Summary

- ❑ **Connectivity based:** Create a hierarchical decomposition of the set of data using some criterion.
- ❑ **Centroid based:** Construct various partitions and then evaluate them by some criterion.
- ❑ **Distribution based:** A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.
- ❑ **Density based:** Based on connectivity and density functions.

Hierarchical Clustering

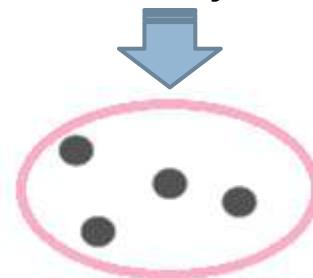
32

Let's say we have the below points and we want to cluster them into groups.



We can assign each of these points to a separate cluster.

Now, based on the similarity of these clusters, the most similar clusters combined together and this process is repeated until only a single cluster is left.



We are essentially building a hierarchy of clusters. That's why this algorithm is called hierarchical clustering.

Hierarchical Clustering Example



33

Suppose a faculty wants to divide the students into different groups. The faculty has the marks scored by each student in an assignment and based on these marks, he/she wants to segment them into groups. There's no fixed target as to how many groups to have. Since the faculty does not know what type of students should be assigned to which group, it cannot be solved as a supervised learning problem. So, hierarchical clustering is applied to segment the students into different groups. Let's take a sample of 5 students.

Roll No	Mark
1	10
2	7
3	28
4	20
5	35

Creating a Proximity Matrix

First, a proximity matrix to be created which tell the distance between each of these points (marks). Since the distance is calculated of each point from each of the other points, a square matrix of shape $n \times n$ (where n is the number of observations) is obtained. Let's make the 5×5 proximity matrix for the example.

Hierarchical Clustering Example cont...

34

Proximity Matrix		1	2	3	4	5
Roll No		1	2	3	4	5
1	0	3	18	10	25	
2	3	0	21	13	28	
3	18	21	0	8	7	
4	10	13	8	0	15	
5	25	28	7	15	0	

The diagonal elements of this matrix is always 0 as the distance of a point with itself is always 0. The Euclidean distance formula is used to calculate the rest of the distances. So, to calculate the distance between

$$\text{Point 1 and 2: } \sqrt{(10-7)^2} = \sqrt{9} = 3$$

$$\text{Point 1 and 3: } \sqrt{(10-28)^2} = \sqrt{324} = 18 \text{ and so on...}$$

Similarly, all the distances are calculated and the proximity matrix is filled.

Steps to Perform Hierarchical Clustering



35

Step 1: First, all the points to an individual cluster is assigned. Different colors here represent different clusters. Hence, 5 different clusters for the 5 points in the data.



Step 2: Next, look at the smallest distance in the proximity matrix and merge the points with the smallest distance. Then the proximity matrix is updated.

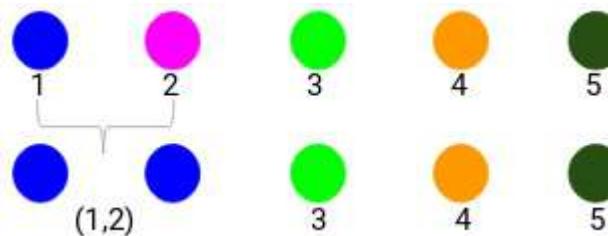
Roll No	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Here, the smallest distance is 3 and hence point 1 and 2 is merged.

Steps to Perform Hierarchical Clustering cont...



36



Let's look at the updated clusters and accordingly update the proximity matrix. Here, we have taken the maximum of the two marks (7, 10) to replace the marks for this cluster. Instead of the maximum, the minimum value or the average values can also be considered.

Roll No	Mark
(1, 2)	10
3	28
4	20
5	35

Now, the proximity matrix for these clusters is calculated again.

Steps to Perform Hierarchical Clustering cont...



37

Revised Proximity Matrix

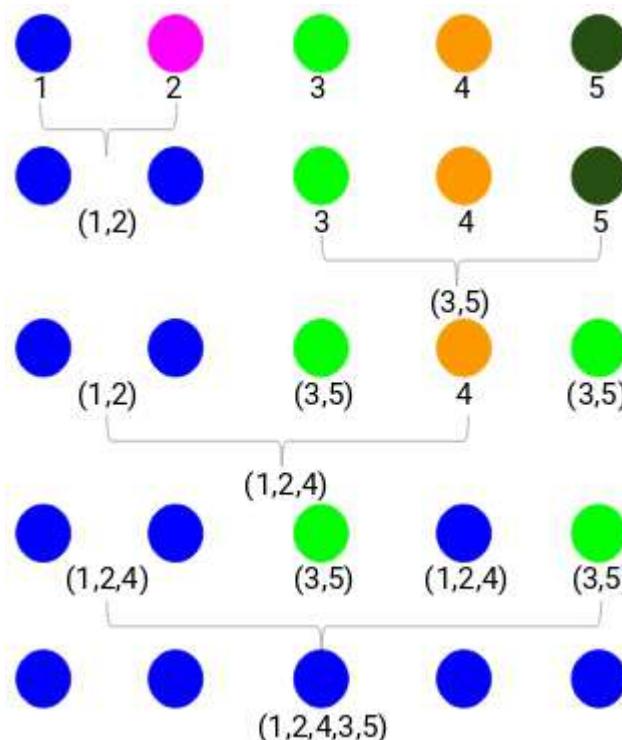
Roll No	1, 2	3	4	5
1, 2	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Step 3: Step 2 is repeated until only a single cluster is left. So, look at the minimum distance in the proximity matrix and then merge the closest pair of clusters. We will get the merged clusters after repeating these steps:

Steps to Perform Hierarchical Clustering cont...



38



We started with 5 clusters and finally have a single cluster.

Types of Hierarchical Clustering

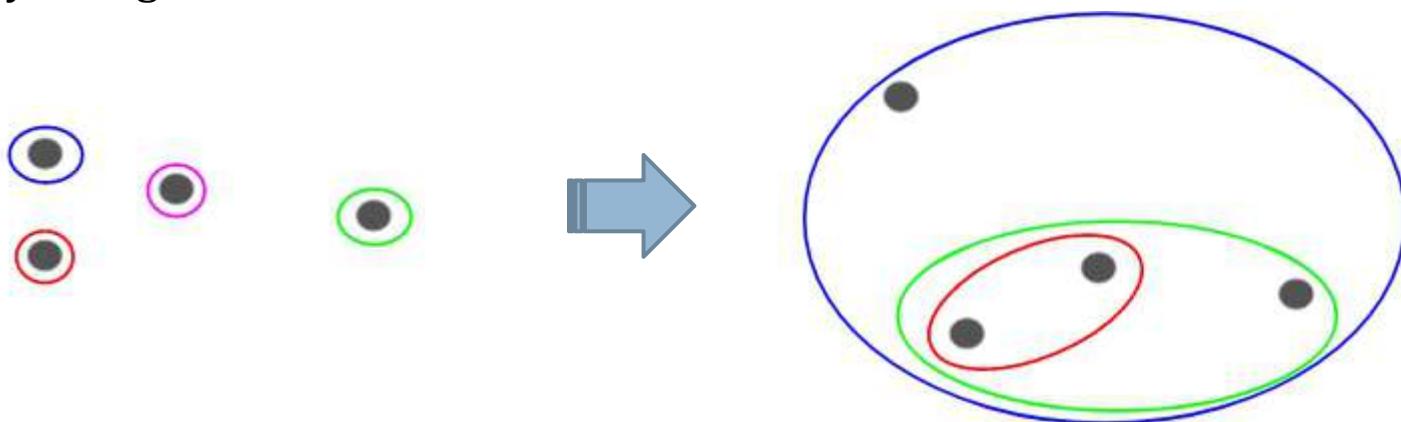
39

There are mainly two types of hierarchical clustering:

- Agglomerative hierarchical clustering
- Divisive Hierarchical clustering

Agglomerative Hierarchical Clustering

- Each point is assigned to an individual cluster in this technique. Suppose there are 4 data points, so each of these points would be assigned to a cluster and hence there would be 4 clusters in the beginning.
- Then, at each iteration, closest pair of clusters are merged and this step is repeated until only a single cluster is left.

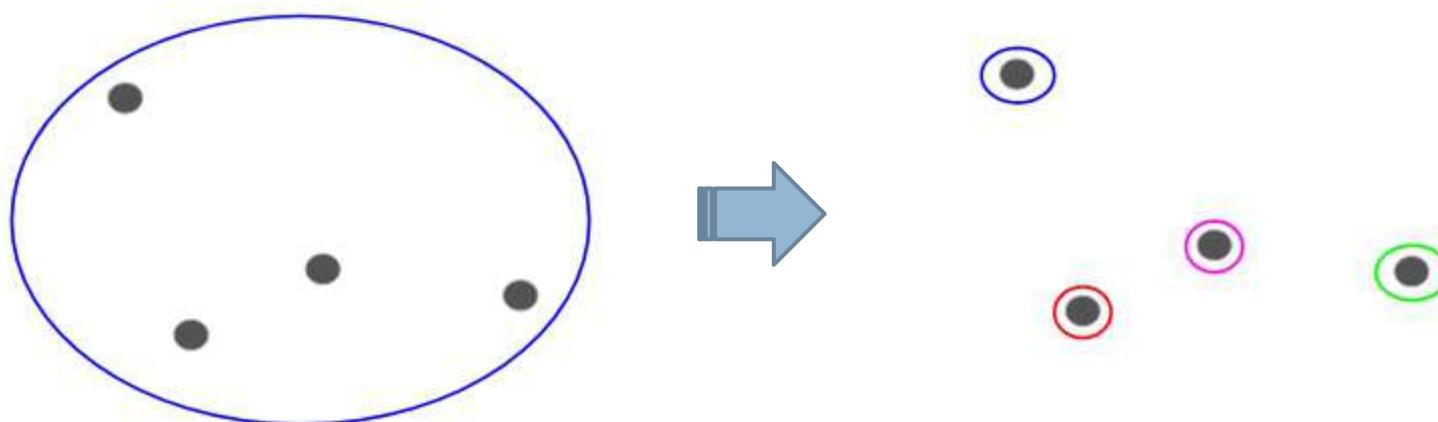


Types of Hierarchical Clustering cont...

40

Divisive Hierarchical clustering

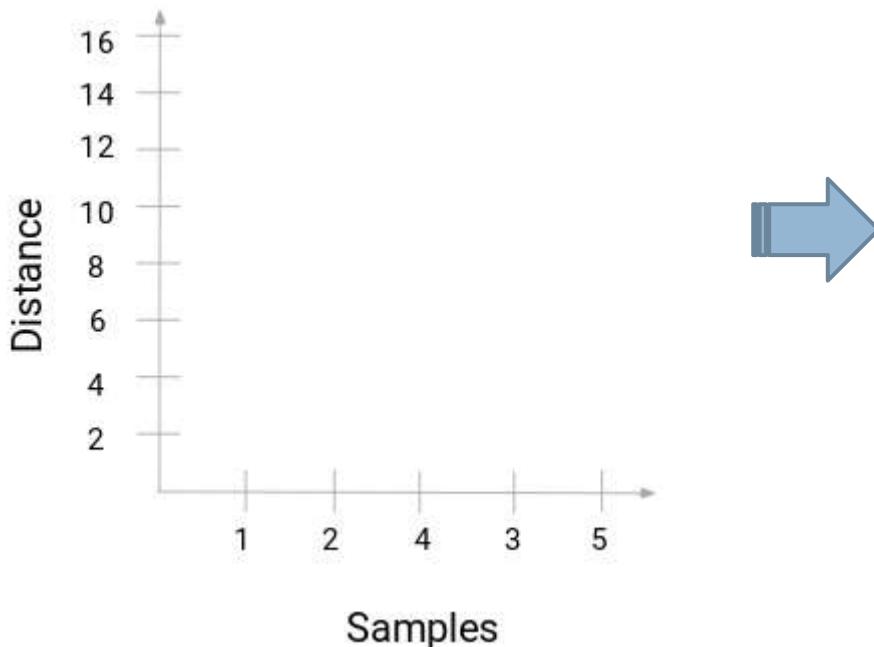
- ❑ Divisive hierarchical clustering works in the opposite way. Instead of starting with n clusters (in case of n observations), we start with a single cluster and assign all the points to that cluster. So, it doesn't matter if we have 10 or 1000 data points. All these points will belong to the same cluster at the beginning.
- ❑ Now, at each iteration, farthest point in the cluster is split and this process is repeated until each cluster only contains a single point.



Dendrogram

41

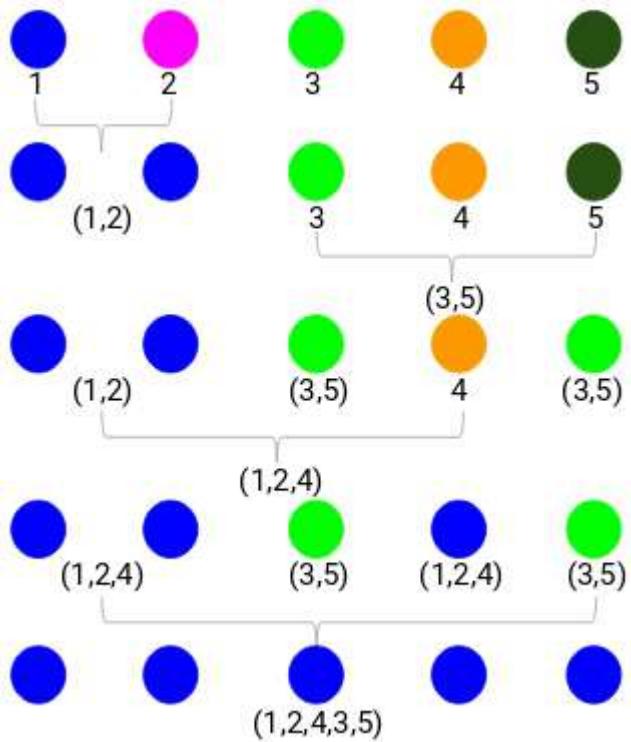
- ❑ To get the number of clusters for hierarchical clustering, we make use of the concept called a **Dendrogram**.
- ❑ A dendrogram is a tree-like diagram that records the sequences of merges or splits.
- ❑ Let's get back to faculty-student example. Whenever we merge two clusters, a dendrogram record the distance between these clusters and represent it in graph form.



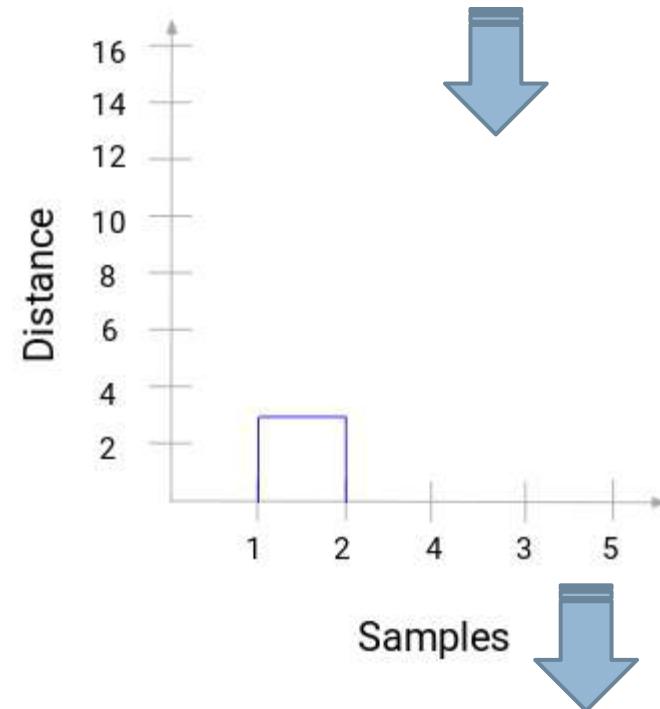
We have the samples of the dataset on the x-axis and the distance on the y-axis. Whenever two clusters are merged, we will join them in this dendrogram and the height of the join will be the distance between these points.

Dendrogram cont...

42



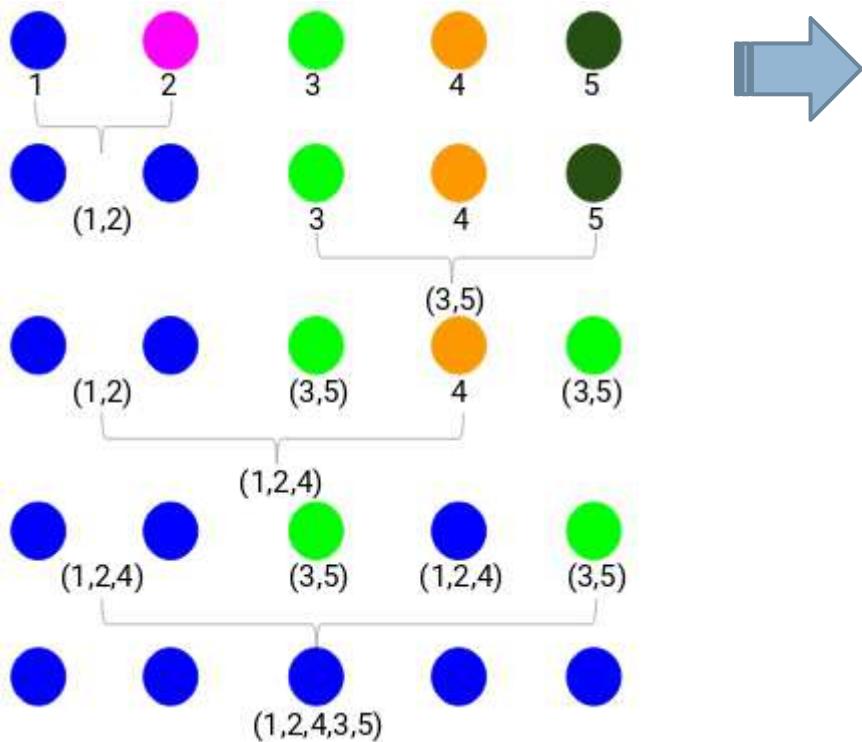
We started by merging sample 1 and 2 and the distance between these two samples was 3. Let's plot this in the dendrogram.



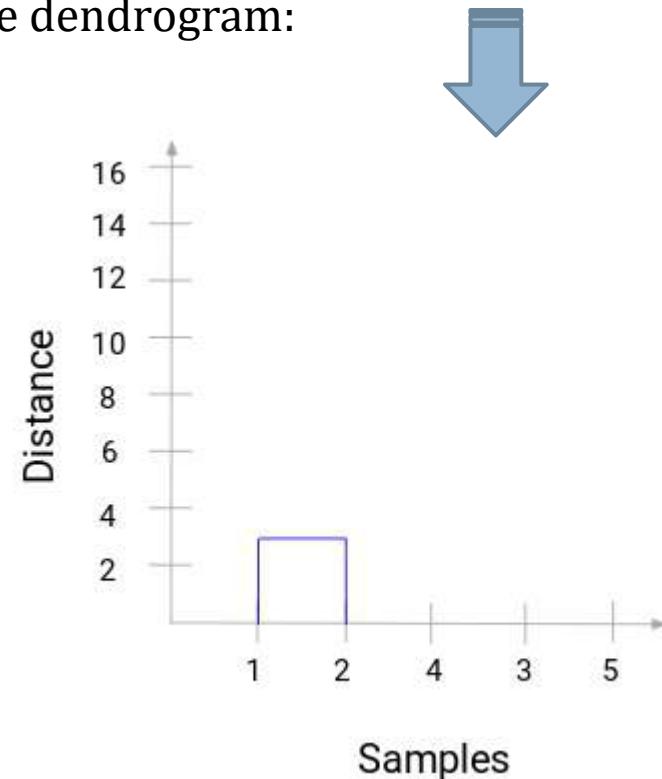
Here, we can see that we have merged sample 1 and 2. The vertical line represents the distance between these samples.

Dendrogram cont...

43



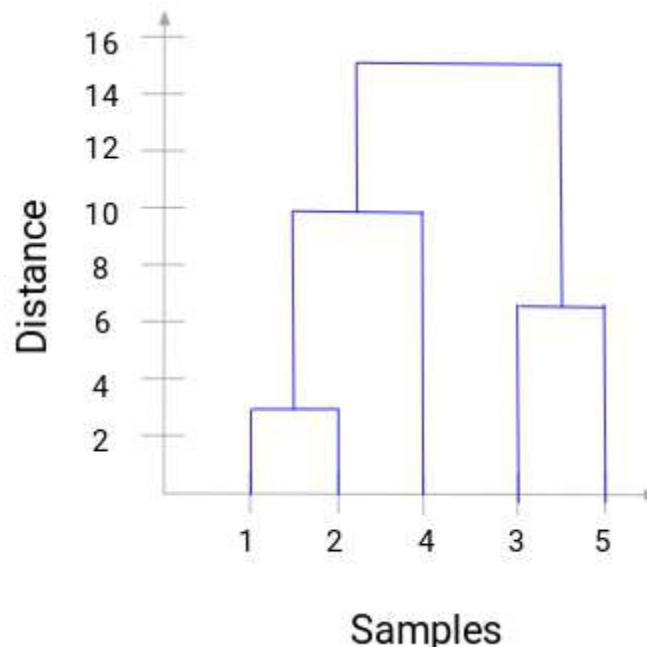
We started by merging sample 1 and 2 and the distance between these two samples was 3. Let's plot this in the dendrogram:



Dendrogram cont...

44

Similarly, we plot all the steps where we merged the clusters and finally, we get a dendrogram like this:

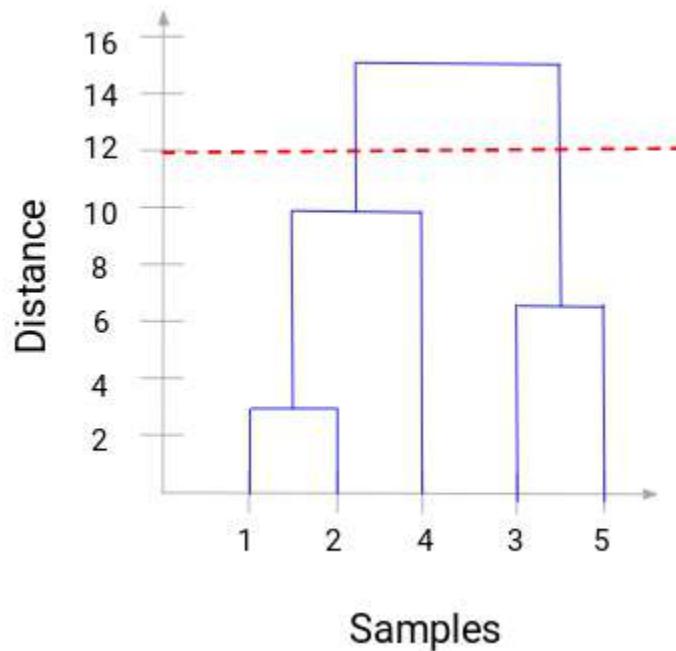


We can clearly visualize the steps of hierarchical clustering. **More the distance of the vertical lines in the dendrogram, more the distance between those clusters.**

Dendrogram cont...

45

Now, we can set a threshold distance and draw a horizontal line (Generally, the threshold is set in such a way that it cuts the tallest vertical line). Let's set this threshold as 12 and draw a horizontal line:



The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. In the above example, since the red line intersects 2 vertical lines, we will have 2 clusters. One cluster will have a sample (1,2,4) and the other will have a sample (3,5).

Hierarchical Clustering closeness of two clusters



46

The decision of merging two clusters is taken on the basis of closeness of these clusters. There are multiple metrics for deciding the closeness of two clusters and primarily are:

- Euclidean distance
- Squared Euclidean distance
- Manhattan distance
- Maximum distance
- Mahalanobis distance

Names	Formula
Euclidean distance	$\ a - b\ _2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\ a - b\ _2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\ a - b\ _1 = \sum_i a_i - b_i $
maximum distance	$\ a - b\ _\infty = \max_i a_i - b_i $
Mahalanobis distance	$\sqrt{(a - b)^\top S^{-1} (a - b)}$ where S is the Covariance matrix

K- Means Clustering Algorithm



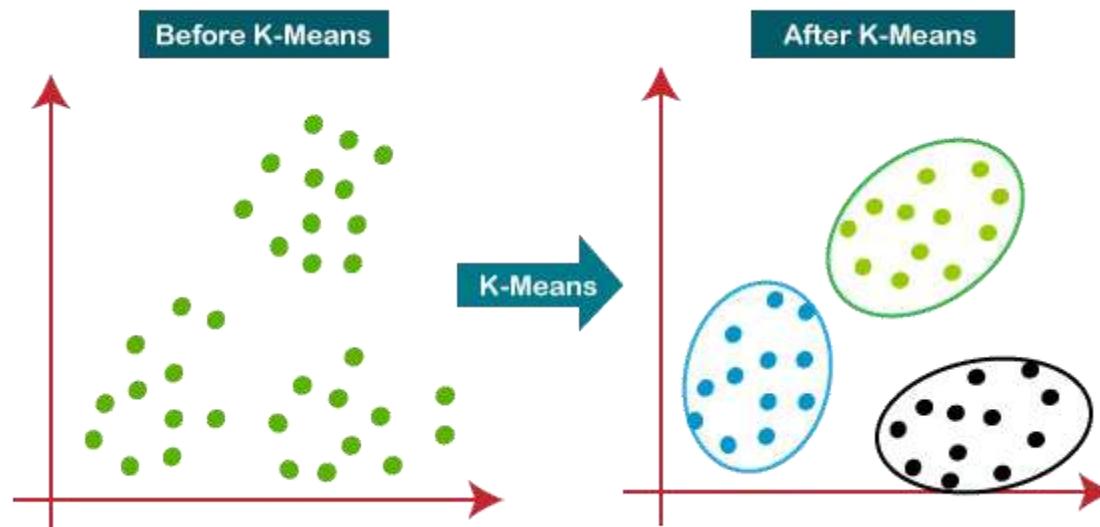
47

- ❑ It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.
- ❑ It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. ***A centroid is a data point (imaginary or real) at the center of a cluster.***
- ❑ The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
- ❑ The k-means clustering algorithm mainly performs two tasks:
 - ❑ Determines the best value for K center points or centroids by an iterative process.
 - ❑ Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.
- ❑ Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

K-Means Algorithm cont...

48

The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?



49

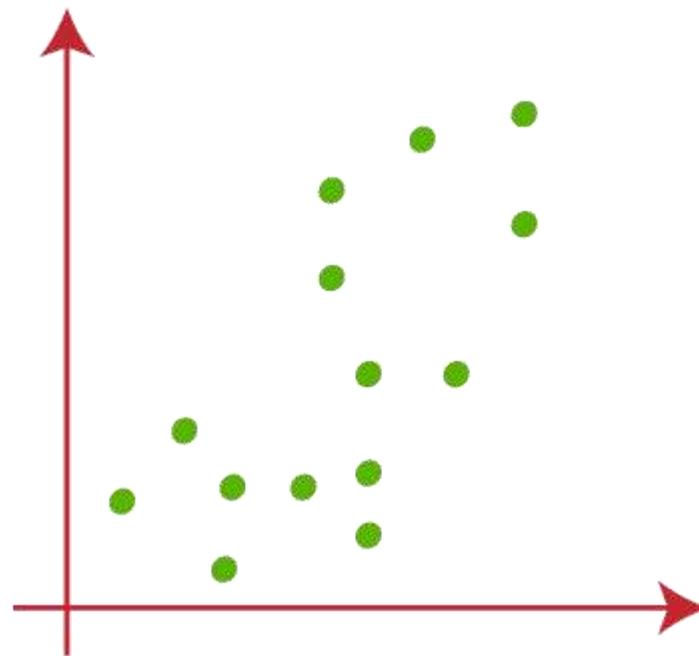
1. Begin
2. Step-1: Select the number K to decide the number of clusters.
3. Step-2: Select random K points or centroids. (It can be other from the input dataset).
4. Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
5. Step-4: Calculate the variance and place a new centroid of each cluster.
6. Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
7. Step-6: If any reassignment occurs, then go to step-4 else go to step-7.
8. Step-7: The model is ready.
9. End

Working of K-Means Algorithm



50

Suppose we have two variables x and y. The x-y axis scatter plot of these two variables is given below:

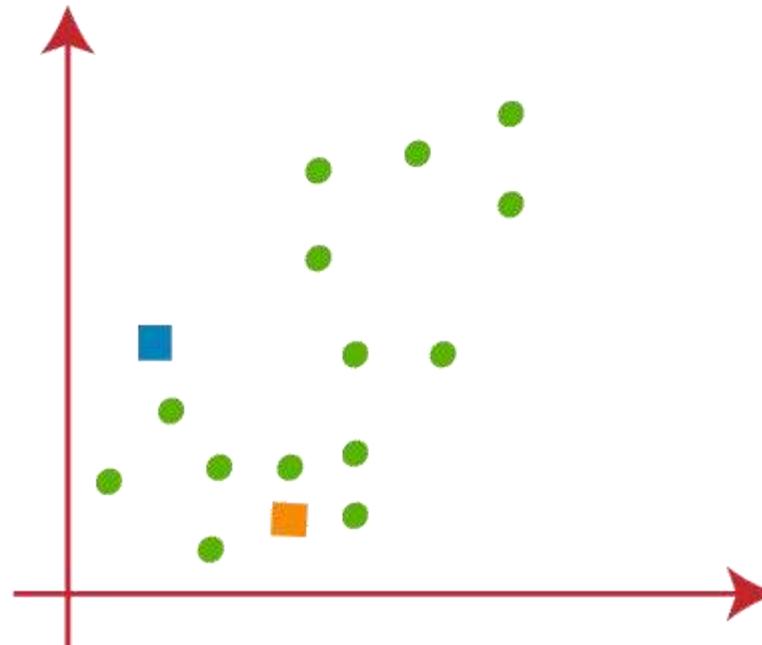


Working of K-Means Algorithm cont...



51

- ❑ Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
- ❑ We need to choose some random K points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as K points, which are not the part of dataset. Consider the below image:

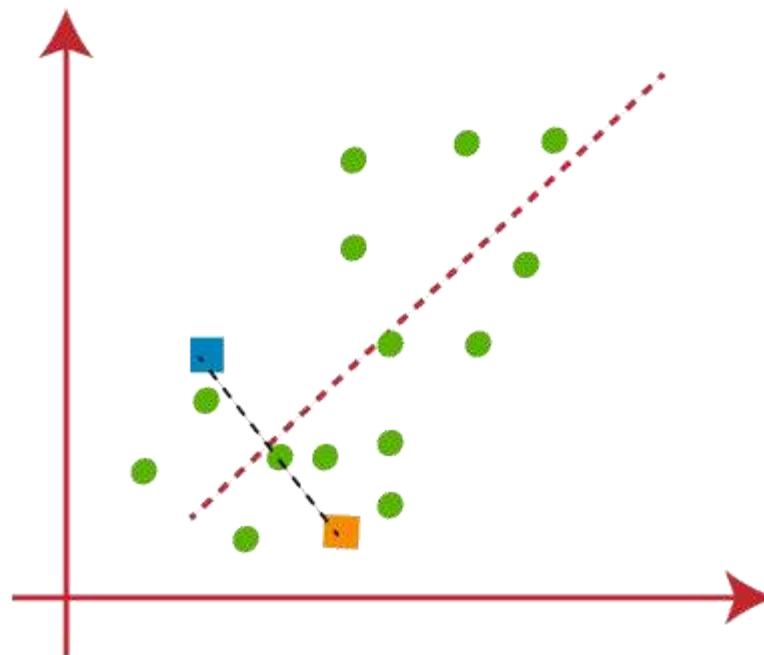


Working of K-Means Algorithm cont...



52

- Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by calculating the distance between two points. So, we will draw a median between both the centroids. Consider the below image:

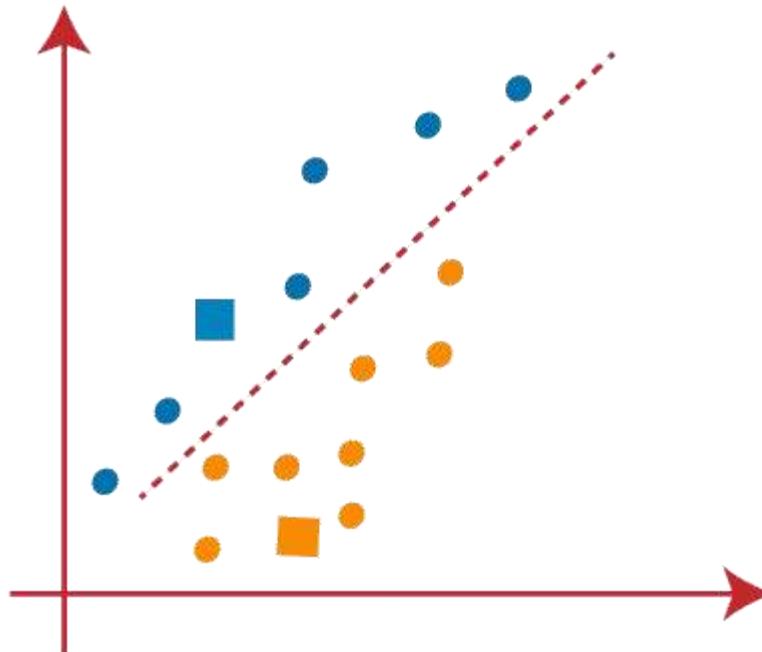


Working of K-Means Algorithm cont...



53

- From the image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid i.e. K2. Let's color them as blue and yellow for clear visualization.

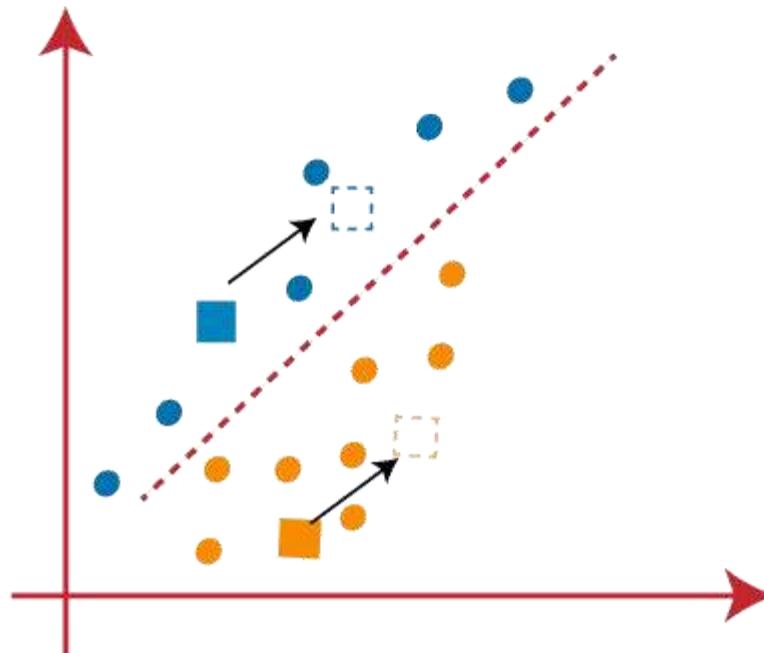


Working of K-Means Algorithm cont...



54

- ❑ As we need to find the closest cluster, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:

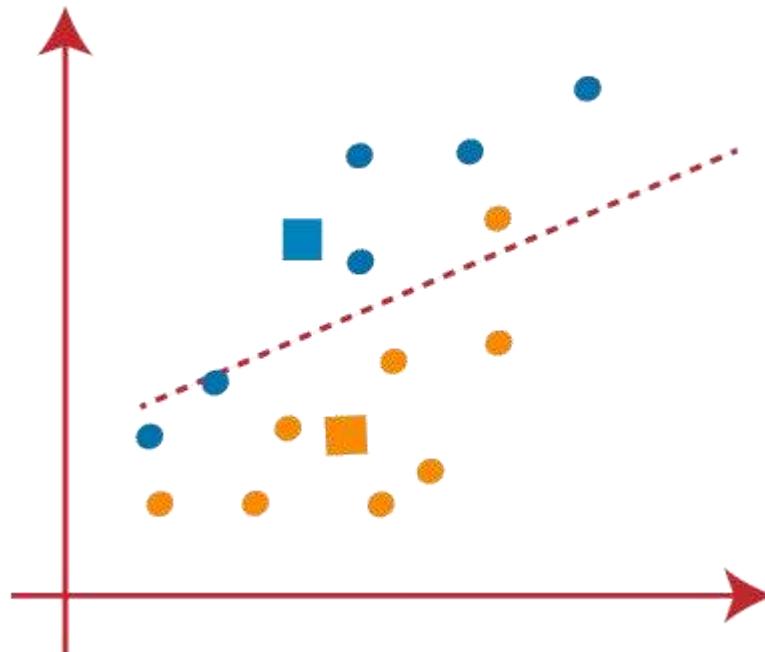


Working of K-Means Algorithm cont...



55

- ❑ Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be like below image:

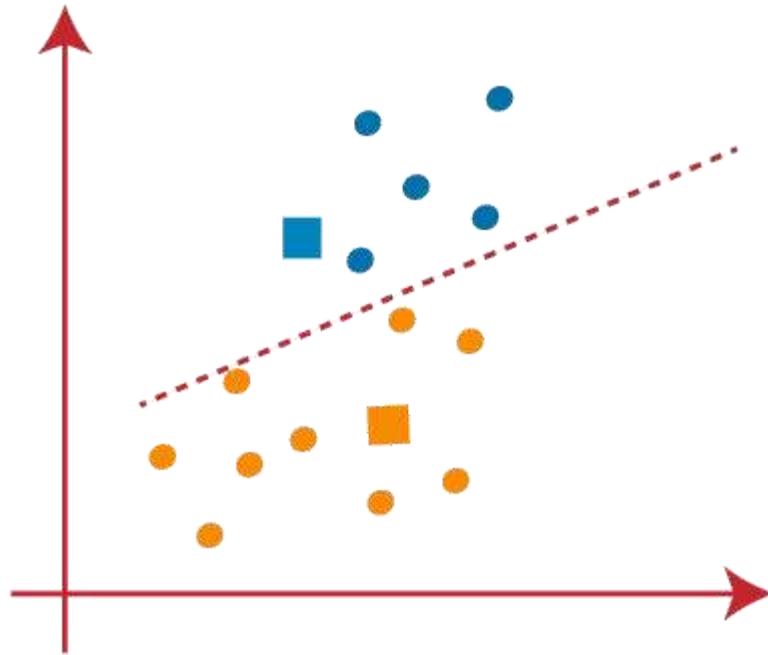


From the above image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids.

Working of K-Means Algorithm cont...



56



- ❑ As reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.

Working of K-Means Algorithm cont...



57

- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:

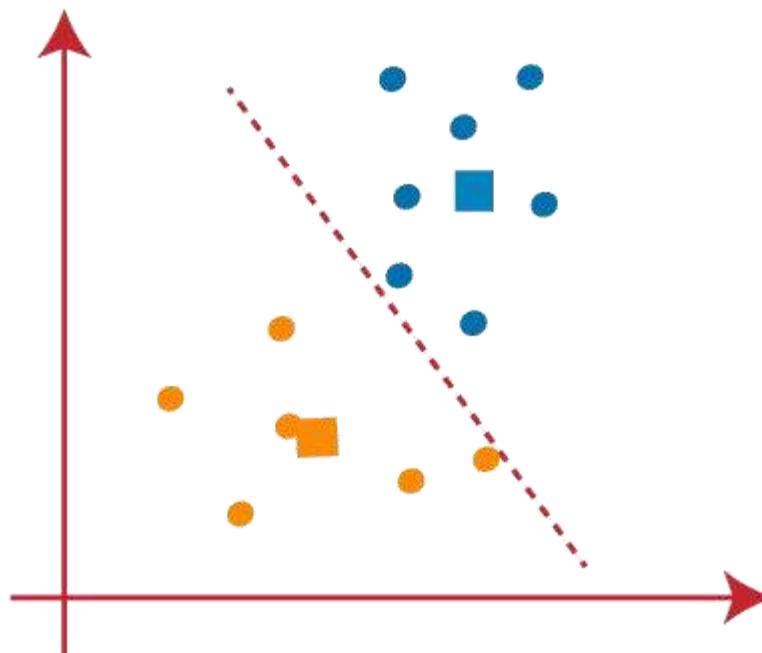


Working of K-Means Algorithm cont...



58

- ❑ As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:

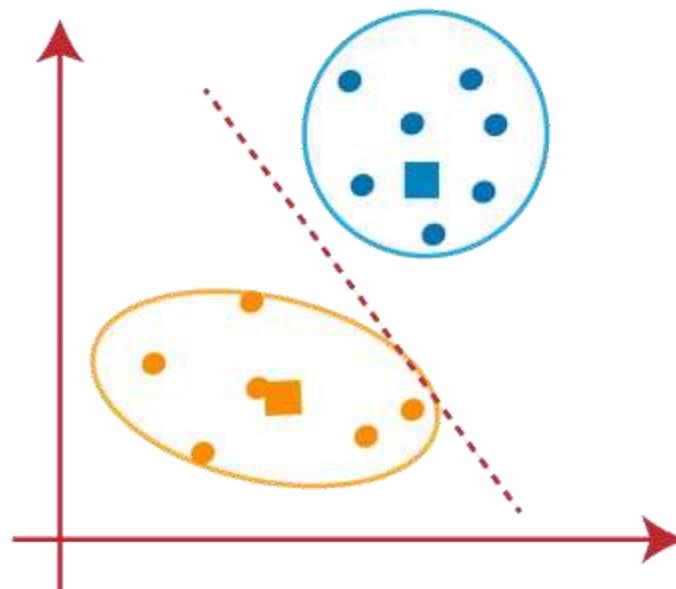


Working of K-Means Algorithm cont...



59

- We can see in the previous image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:

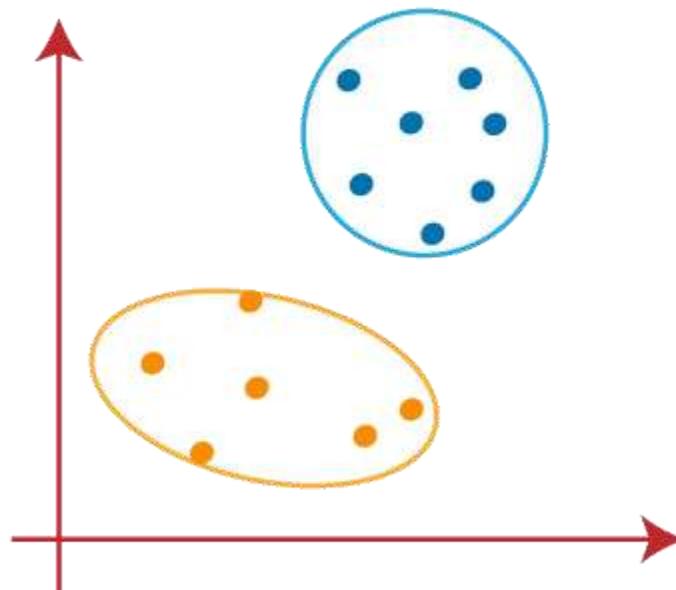


Working of K-Means Algorithm cont...



60

- ❑ As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



How to Determine the optimal K for K-Means?



61

How to determine the optimal value of K in K-Means clustering? There are 2 methods:

- The Elbow Method
- The Silhouette Method

Self-Study

THANK YOU!

Data Analytics (IT-3006)

**Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024**

School of Computer Engineering



Strictly for internal circulation (within KIIT) and reference only. Not for outside circulation without permission

3 Credit

Lecture Note – Unit 5

Course Contents



2

Sr #	Major and Detailed Coverage Area	Hrs
5	Frameworks and Visualization Introduction to framework and Visualization, Introduction to Hadoop, Core Components of Hadoop, Hadoop Ecosystem, Physical Architecture, Hadoop Limitations, Hive, MapReduce and The New Software Stack , MapReduce, Algorithms Using MapReduce, NOSQL, NoSQL Business Drivers, NoSQL Case Studies, NoSQL Data Architectural Patterns, Variations of NoSQL Architectural Patterns, Using NoSQL to Manage Big Data, Visualizations	8

Introduction

3

- ❑ Huge volume of unstructured data are produced by heterogeneous scenarios in various applications from scientific computing to social networks, from e-government applications to medical information systems and so on, have to be stored in big data stores and analyzed in order to derive intelligence and extract useful knowledge from them.
- ❑ There is a need for good visualization of the results produced from analytic engines.
- ❑ Visualization issues are a big problem in data warehousing and OLAP research.
- ❑ These issues get multifold in the context of big data analytics, as visualization is expected to give a stronger decision-support value.
- ❑ Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers.

Hadoop

4

Hadoop is an open-source project of the Apache Foundation. Apache Hadoop is written in Java and a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data and uses Google's MapReduce and Google File System as its foundation.

Hadoop
Apache open-source software framework
Inspired by:
- Google MapReduce
- Google File System

Hadoop provides various tools and technologies, collectively termed as Hadoop ecosystem, to enable development and deployment of Big Data solutions. It accomplishes two tasks namely i) Massive data storage, and ii) Faster data processing.

Flood of data

5

Few statistics to get an idea of data gets generated every day, every minute, and every second.

- ❑ Every day
 - ❑ NYSE generates 1.5 billion shares and trade data
 - ❑ Facebook stores 2.7 billion comments and likes
 - ❑ Google processes about 24 petabytes of data
- ❑ Every minutes
 - ❑ Facebook users share nearly 2.5 million pieces of content.
 - ❑ Amazon generates over \$ 80,000 in online sale
 - ❑ Twitter users tweet nearly 300,000 times.
 - ❑ Instagram users post nearly 220,000 new photos
 - ❑ Apple users download nearly 50,000 apps.
 - ❑ Email users send over 2000 million messages
 - ❑ YouTube users upload 72 hrs of new video content
- ❑ Every second
 - ❑ Banking applications process more than 10,000 credit card transactions.

Data Challenges

6

To process, analyze and make sense of these different kinds of data, a system is needed that scales and address the challenges as shown:



“I am flooded with data”. How to store terabytes of mounting data?

“I have data in various sources. I have data that rich in variety – structured, semi-structured and unstructured”. How to work with data that is so very different?



“I need this data to be proceed quickly. My decision is pending”. How to access the information quickly?

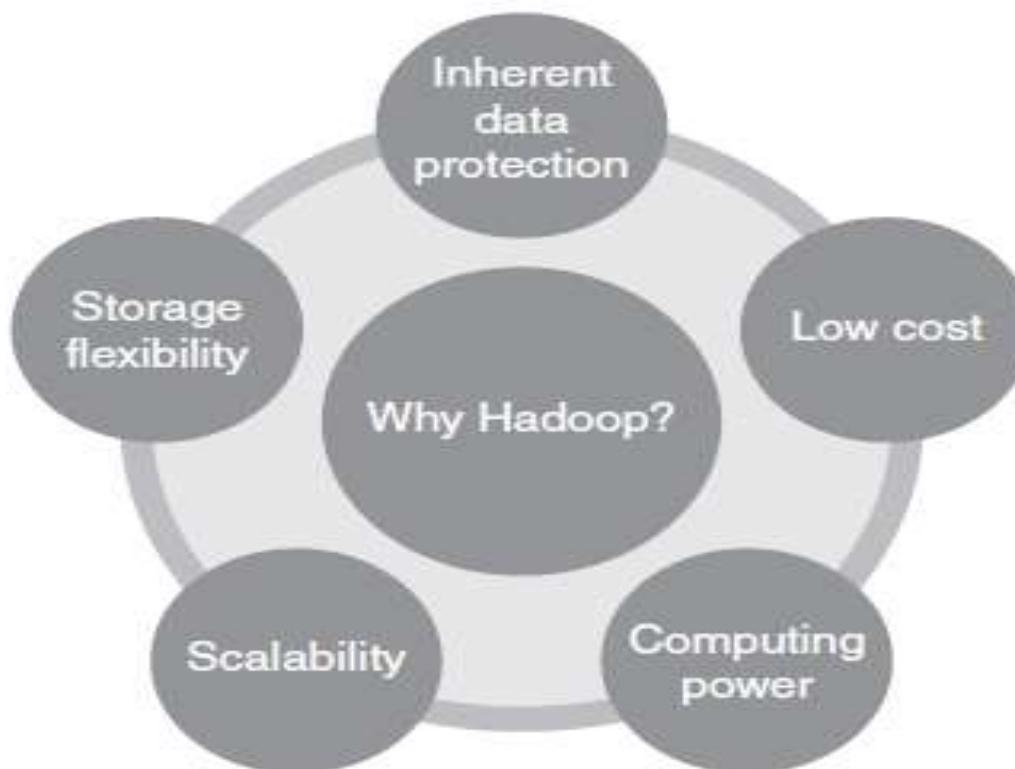


Why Hadoop

7

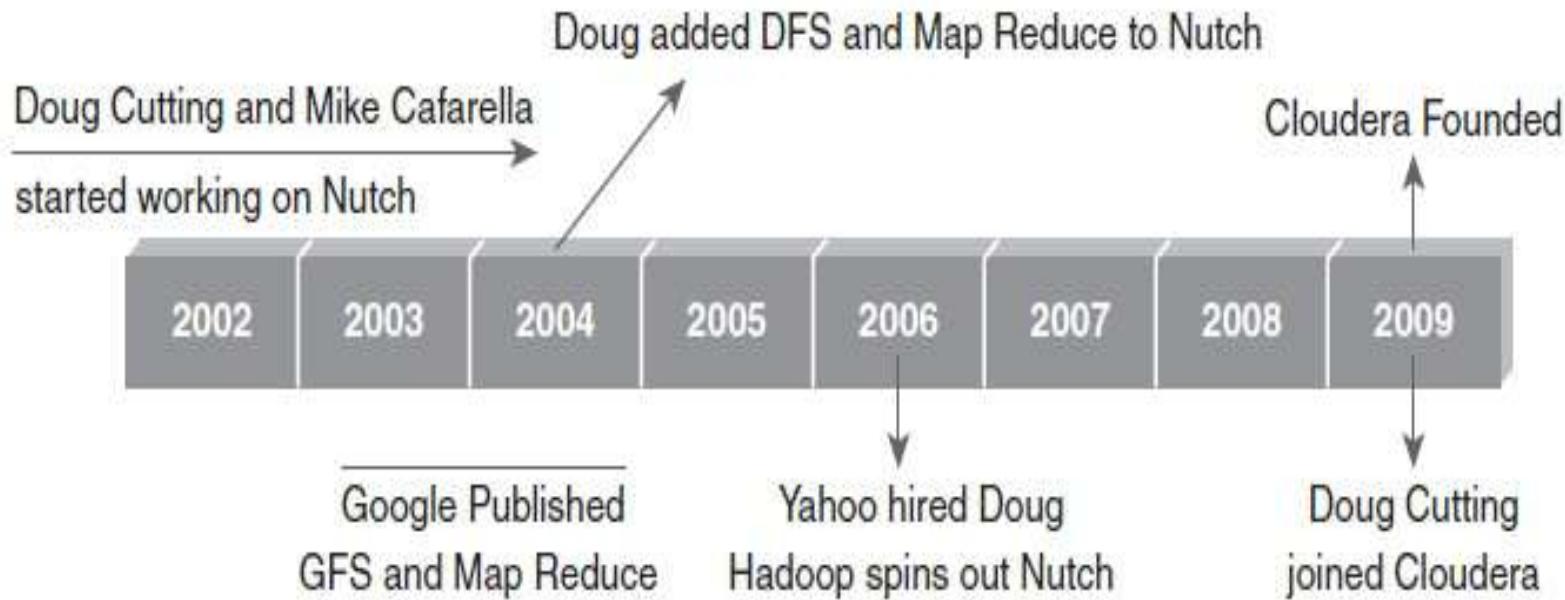
Its capability to handle massive amounts of data, different categories of data – fairly quickly.

Considerations



Hadoop History

8

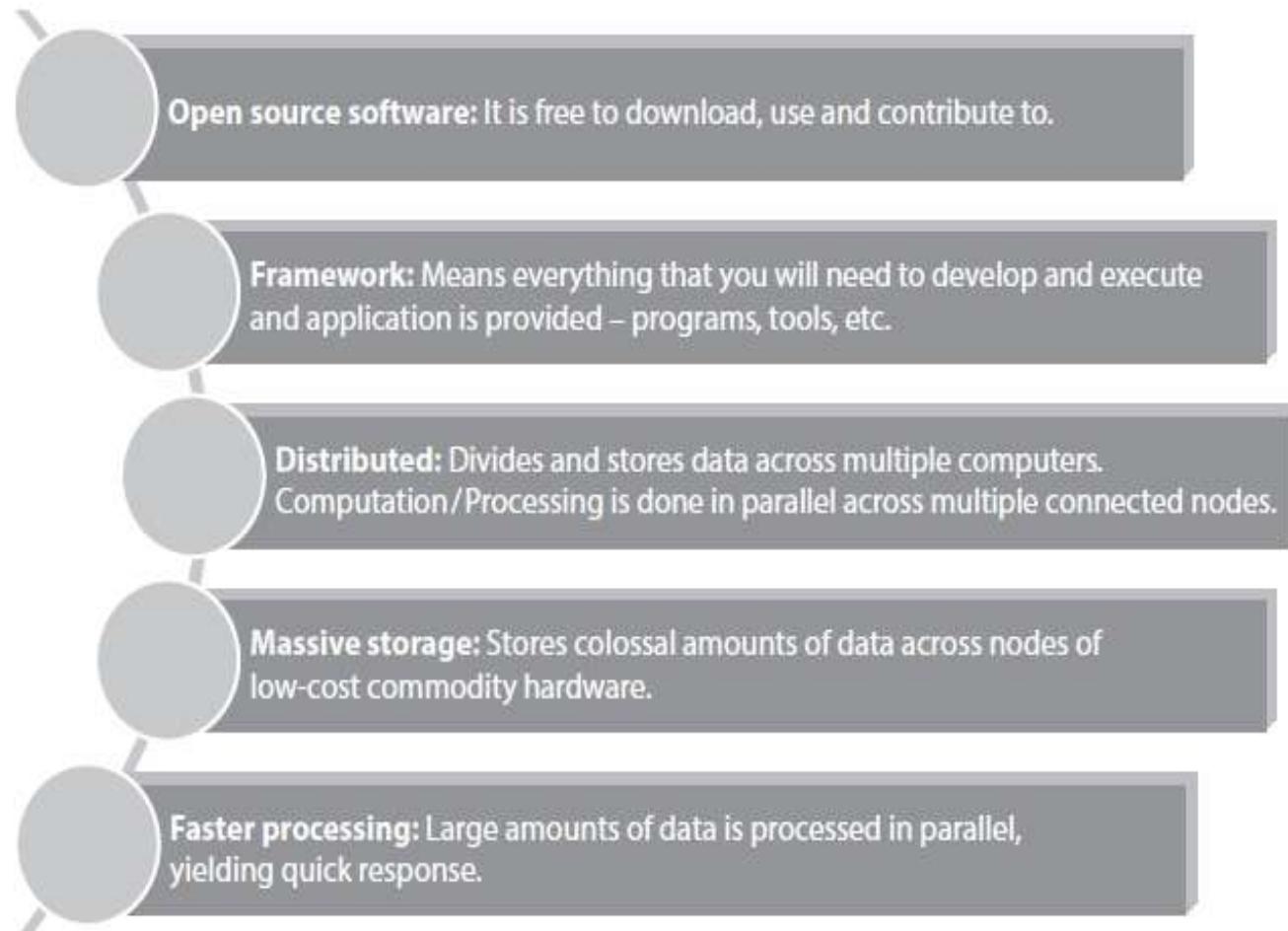


Hadoop was created by Doug Cutting, the creator of Apache Lucene (text search library). Hadoop was part of Apace Nutch (open-source web search engine of Yahoo project) and also part of Lucene project. The name Hadoop is not an acronym; it's a made-up name.

Key Aspects of Hadoop

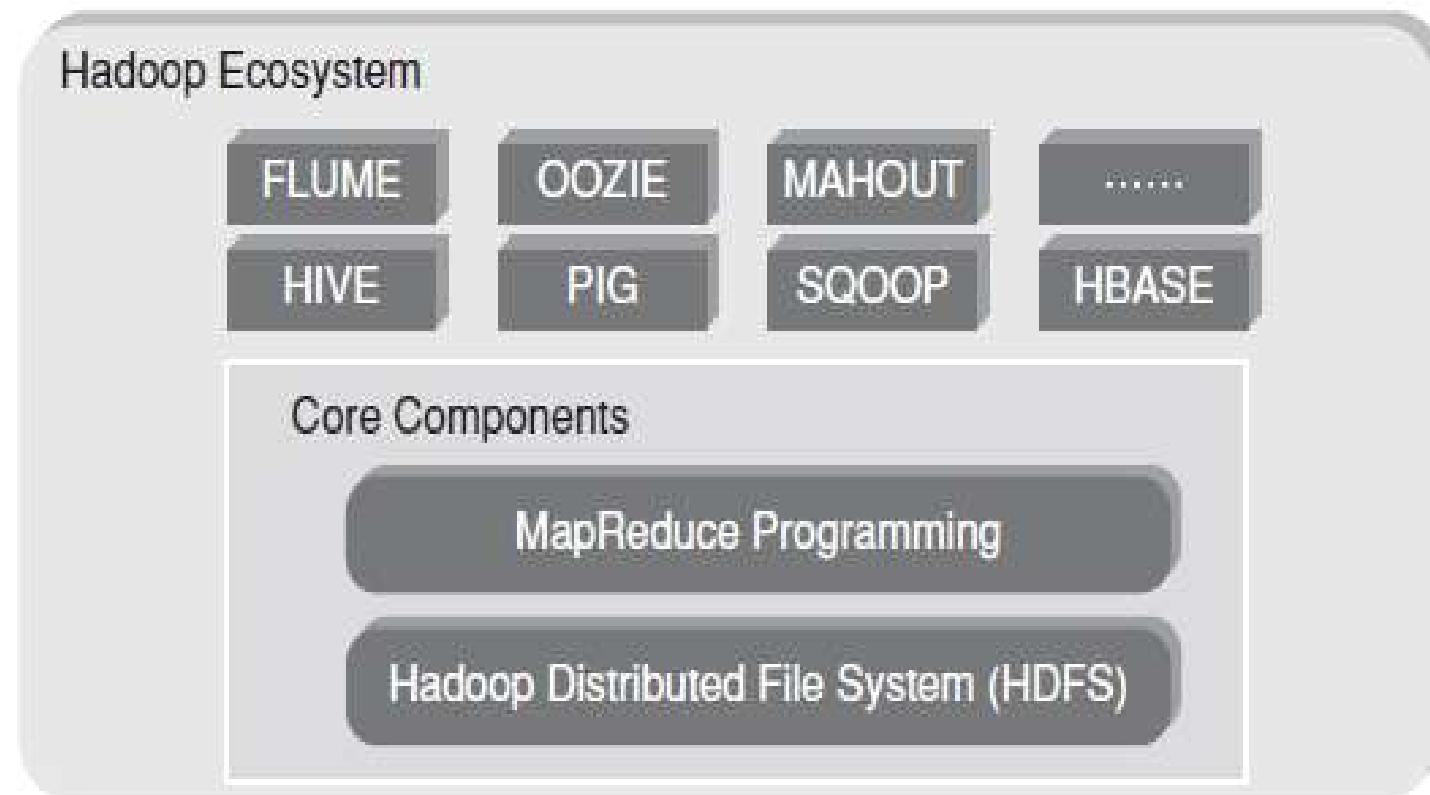


9



Hadoop Components

10



Hadoop Components cont'd



11

Hadoop Core Components:

- ❑ HDFS
 - ❑ Storage component
 - ❑ Distributed data across several nodes
 - ❑ Natively redundant
- ❑ MapReduce
 - ❑ Computational Framework
 - ❑ Splits a task across multiple nodes
 - ❑ Process data in parallel

Hadoop Ecosystems: These are support projects to enhance the functionality of Hadoop Core components. The projects are as follows:

- | | | |
|---------|----------|---------|
| ❑ Hive | ❑ Flume | ❑ HBase |
| ❑ Pig | ❑ Oozie | |
| ❑ Sqoop | ❑ Mahout | |

The Hadoop Ecosystem

Hadoop Common

- Contains Libraries and other modules

HDFS

- Hadoop Distributed File System

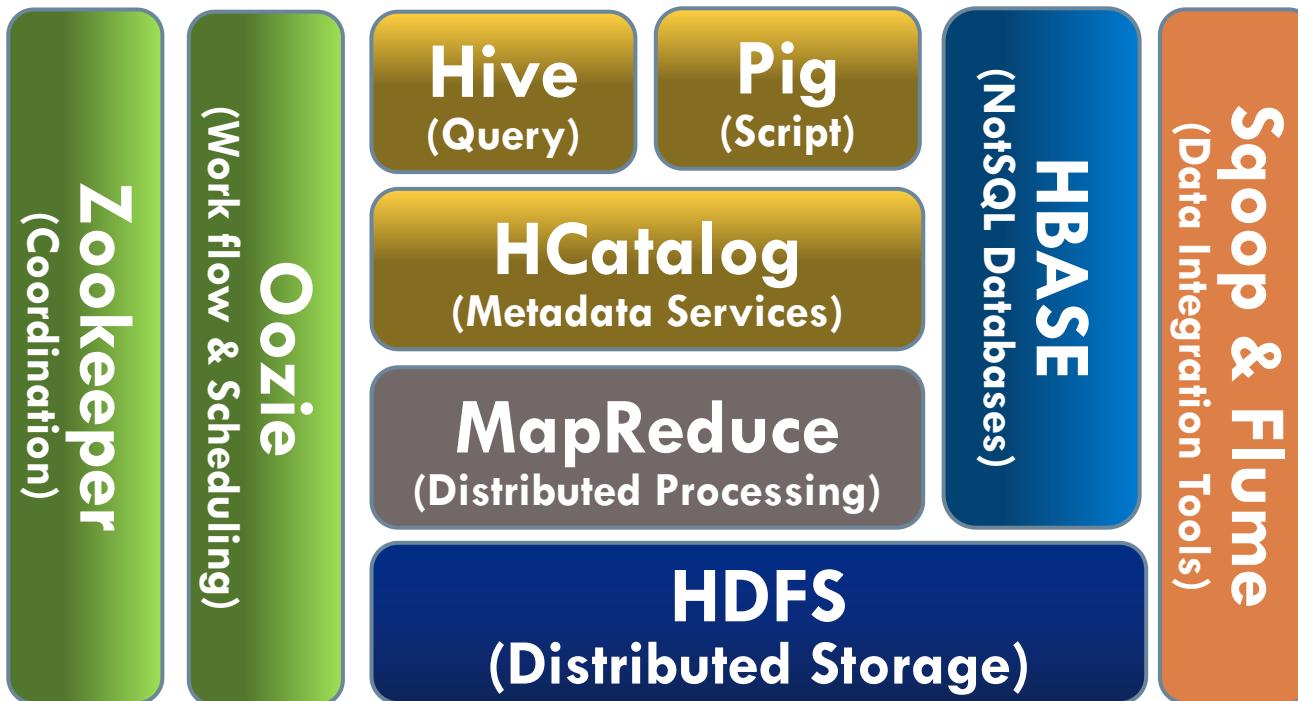
Hadoop YARN

- Yet Another Resource Negotiator

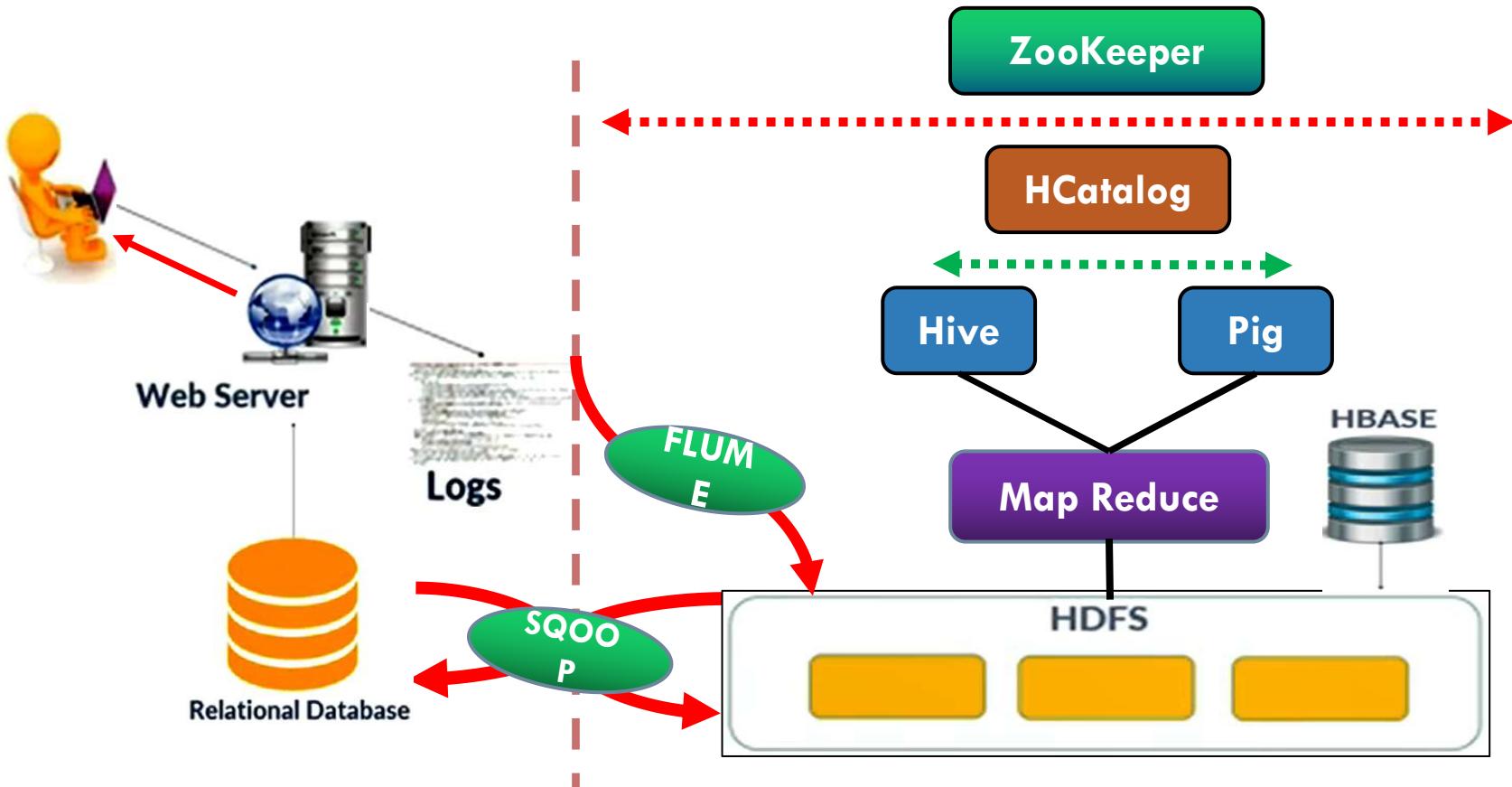
Hadoop MapReduce

- A programming model for large scale data processing

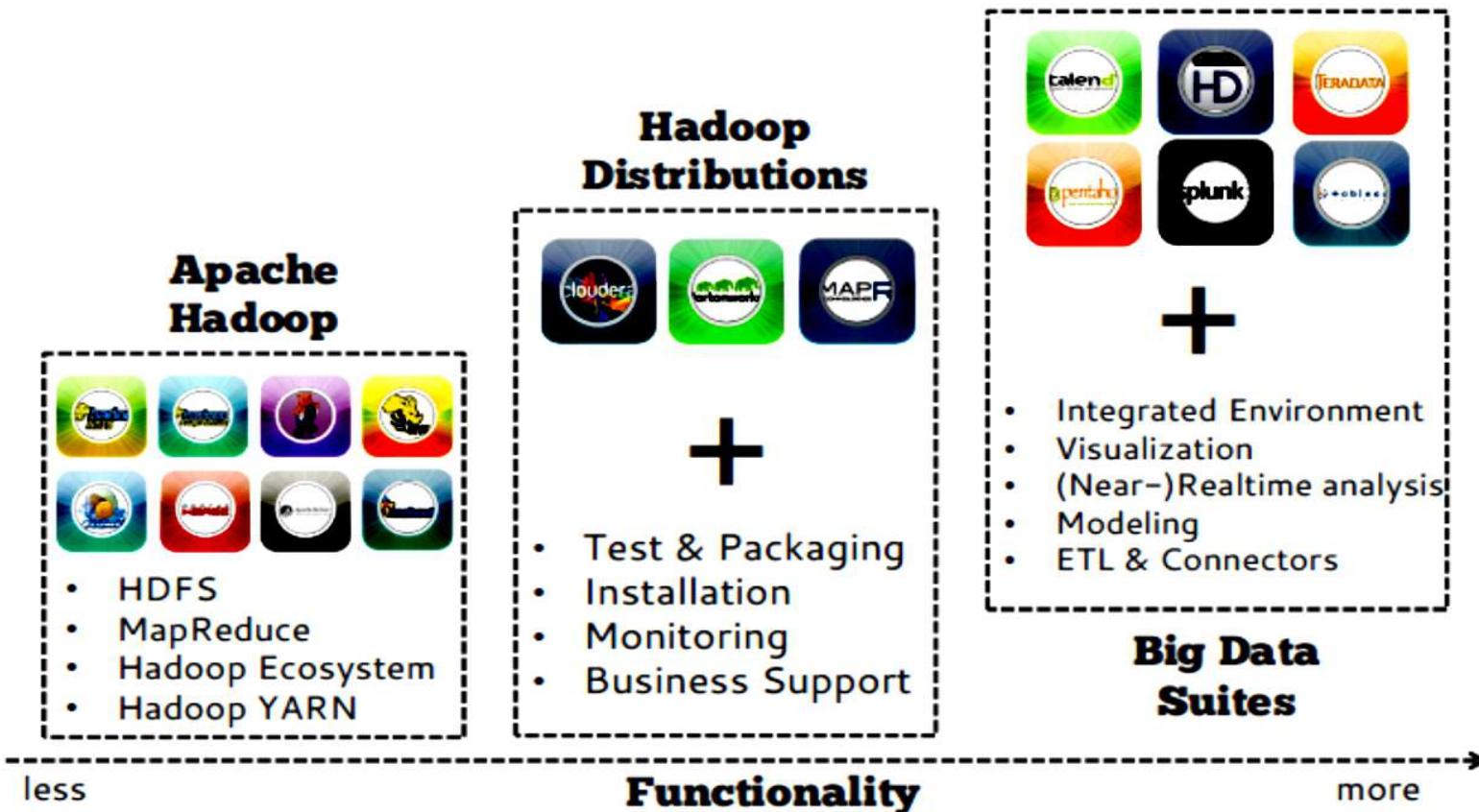
The Hadoop Ecosystem



Hadoop Use case

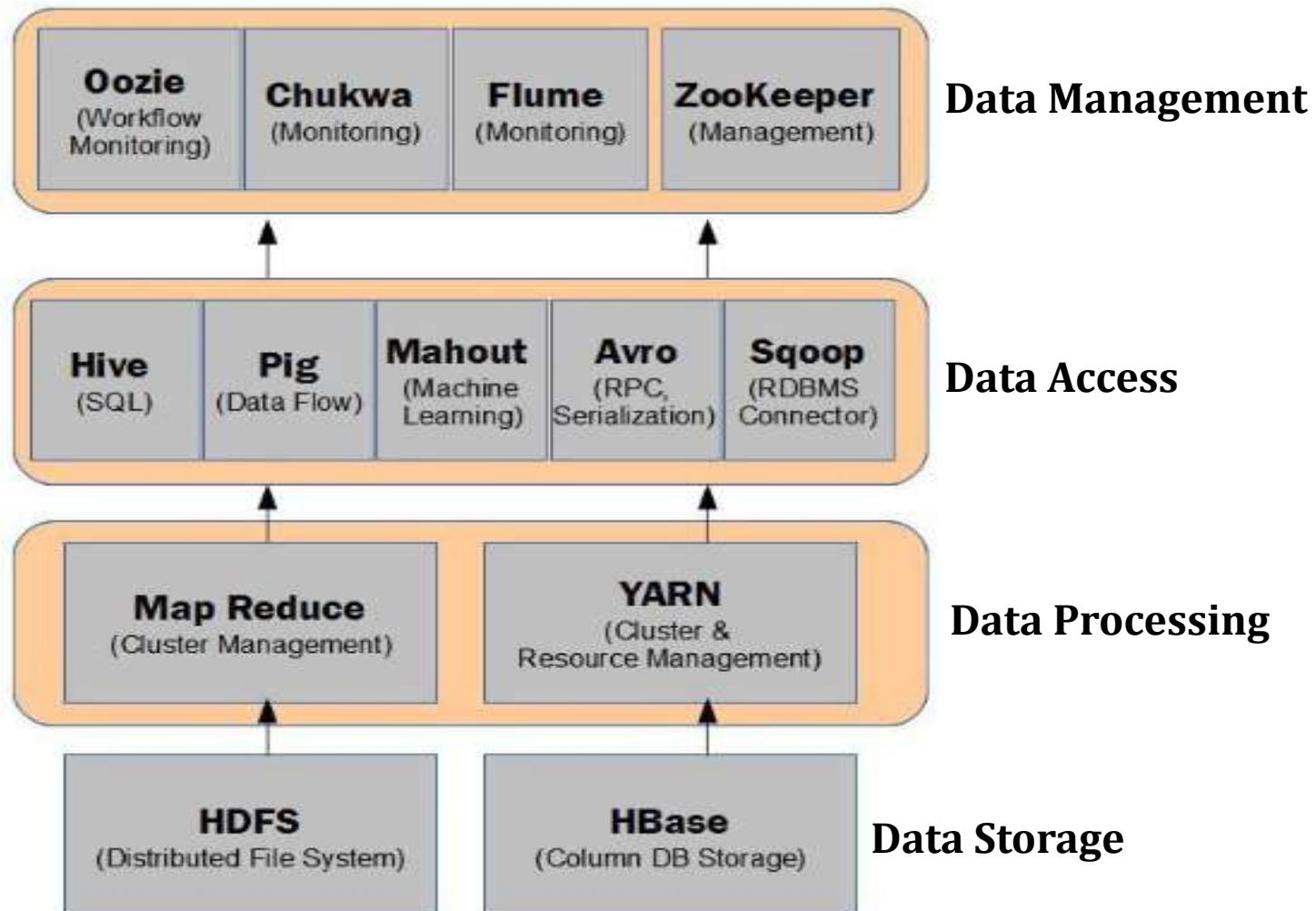


The Hadoop App Store



Hadoop Ecosystem

16



Version of Hadoop

17

There are 3 versions of Hadoop available:

- Hadoop 1.x Hadoop 3.x
- Hadoop 2.x

Hadoop 1.x vs. Hadoop 2.x

Hadoop 1.x

MapReduce
Data Processing & Resource Management

HDFS
Distributed File Storage
(redundant, reliable storage)

Hadoop 2.x

MapReduce

Other Data Processing Framework

YARN
Resource Management

HDFS2
Distributed File Storage
(redundant, highly-available, reliable storage)

Hadoop 2.x vs. Hadoop 3.x



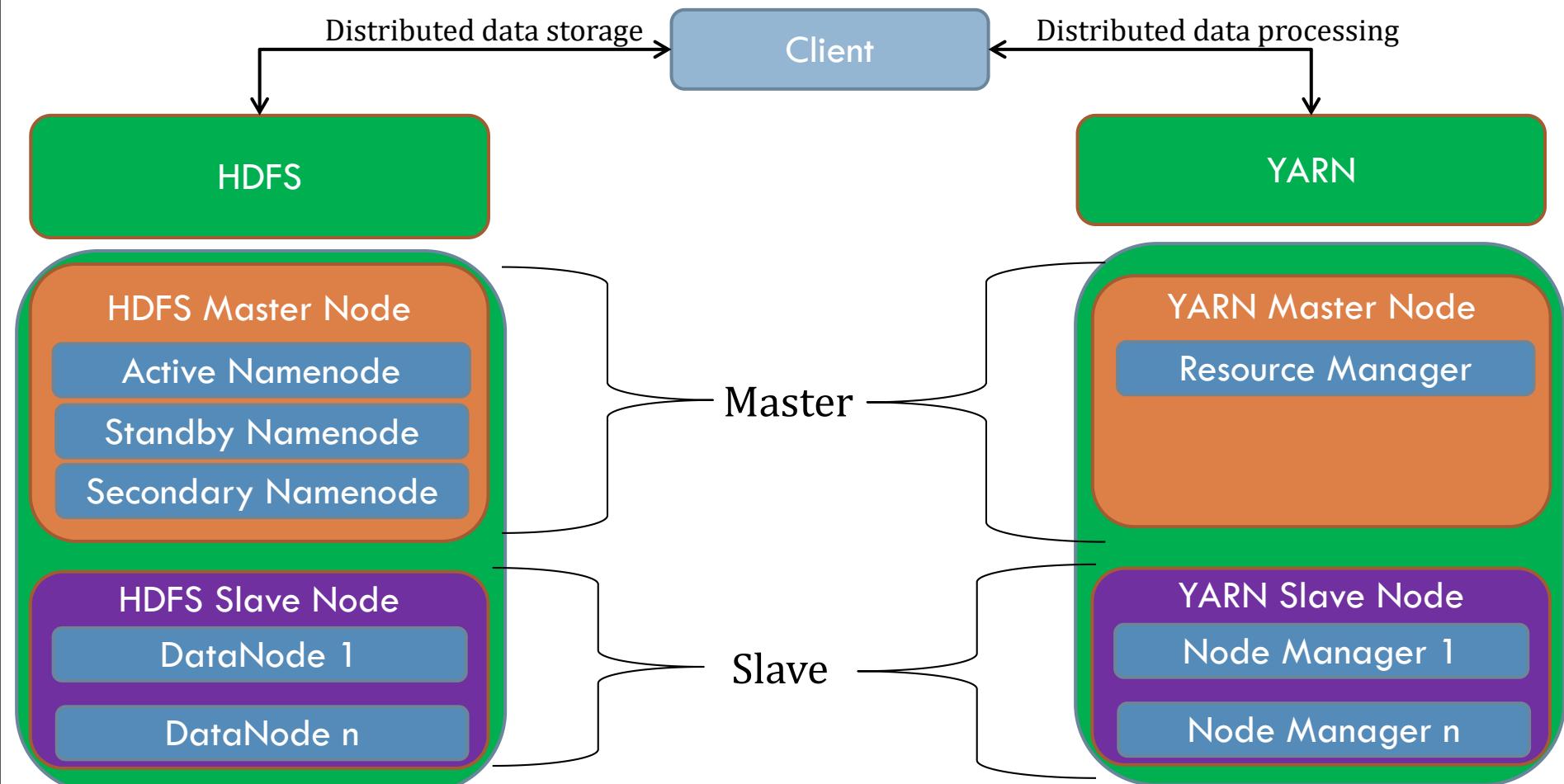
18

Characteristics	Hadoop 2.x	Hadoop 3.x
Minimum supported version of java	Java 7	Java 8
Fault tolerance	Handled by replication (which is wastage of space).	Handled by erasure coding
Data Balancing	Uses HDFS balancer	Uses Intra-data node balancer, which is invoked via the HDFS disk balancer CLI.
Storage Scheme	Uses 3X replication scheme. E.g. If there is 6 block so there will be 18 blocks occupied the space because of the replication scheme.	Support for erasure encoding in HDFS. E.g. If there is 6 block so there will be 9 blocks occupied the space 6 block and 3 for parity.
Scalability	Scale up to 10,000 nodes per cluster.	Scale more than 10,000 nodes per cluster.

High Level Hadoop 2.0 Architecture

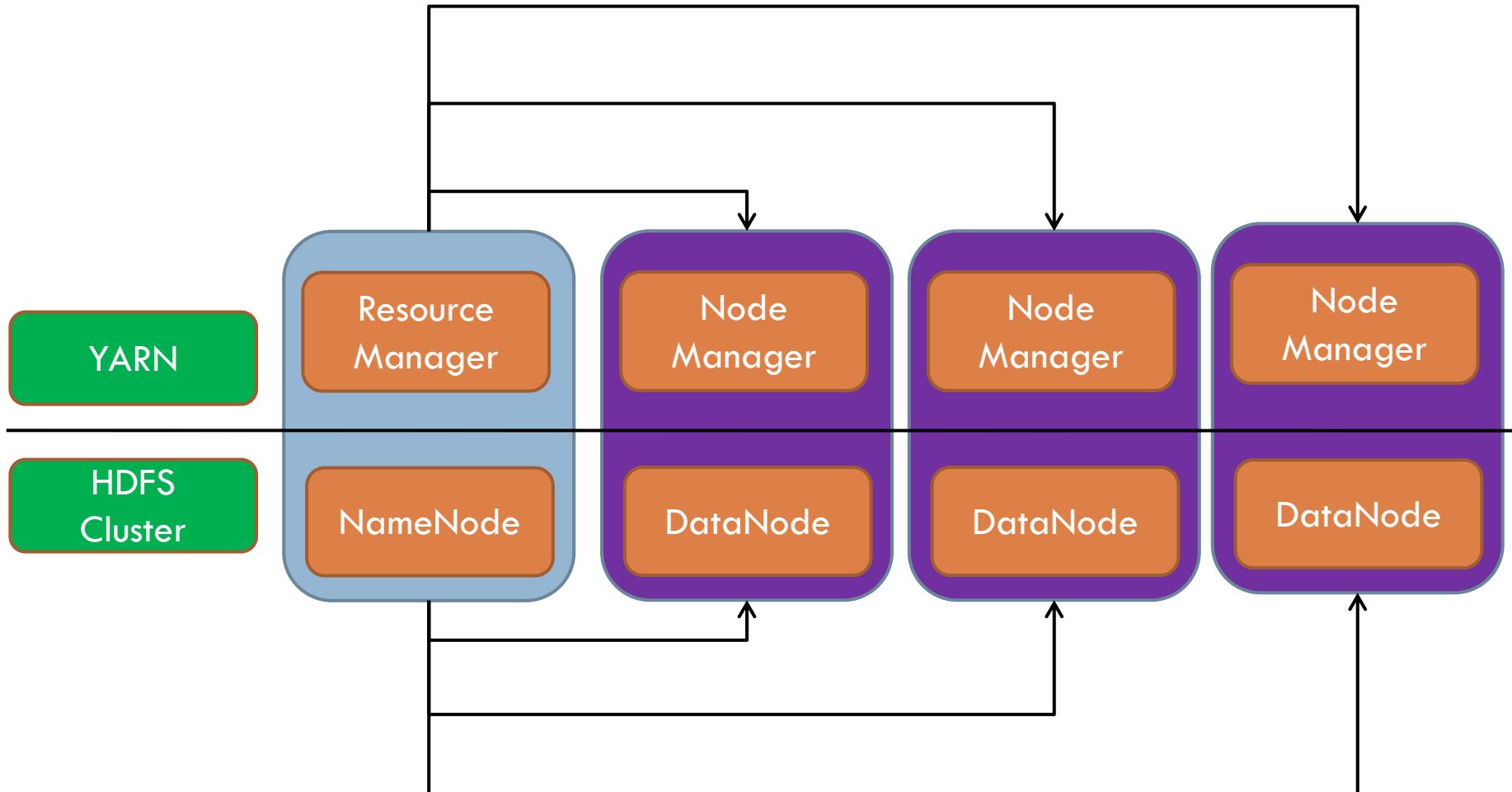
19

Hadoop is distributed Master-Slave architecture.



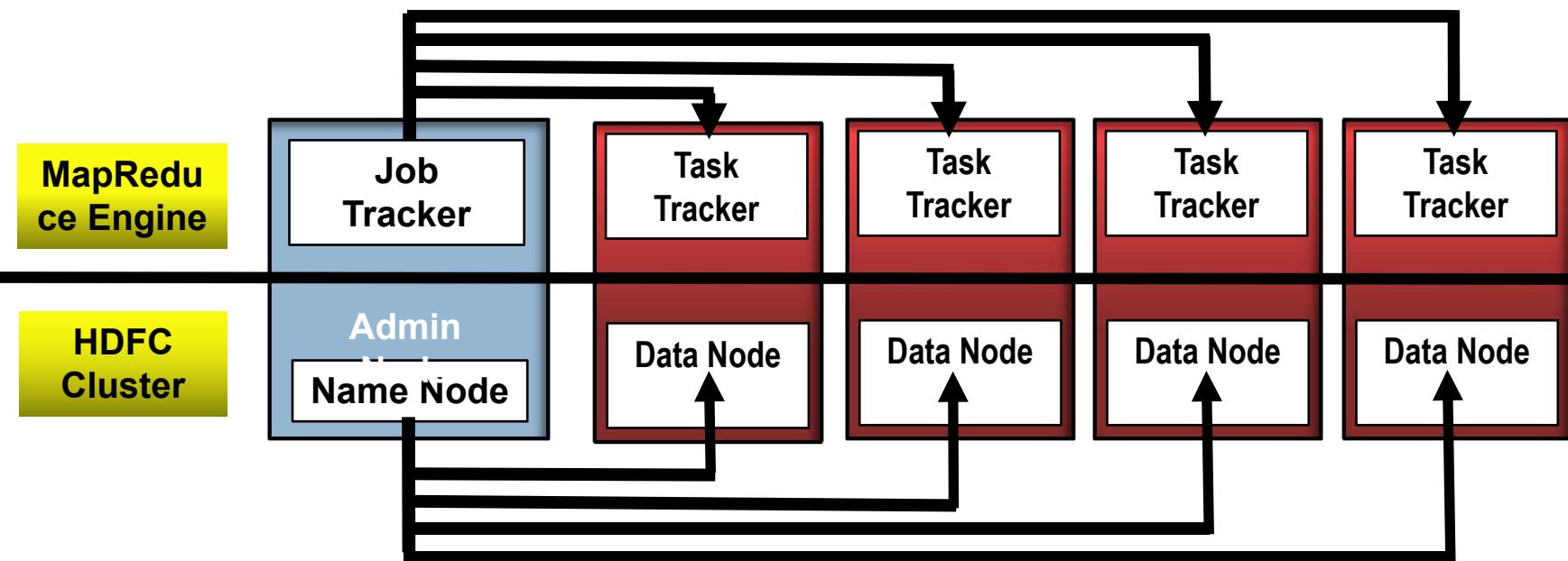
High Level Hadoop 2.0 Architecture cont'd

20



Hadoop Core Components

- HDFC - Hadoop Distributed File System (storage)
- MapReduce (processing)



Hadoop HDFS

22

- ❑ The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications.
- ❑ HDFS holds very large amount of data and employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.
- ❑ To store such huge data, the files are stored across multiple machines.
- ❑ These files are stored in redundant fashion to rescue the system from possible data losses in case of failure.
- ❑ It's run on commodity hardware.
- ❑ Unlike other distributed systems, HDFS is highly fault-tolerant and designed using low-cost hardware.

Hadoop HDFS Key points

23

Some key points of HDFS are as follows:

1. Storage component of Hadoop.
2. Distributed File System.
3. Modeled after Google File System.
4. Optimized for high throughput (HDFS leverages large block size and moves computation where data is stored).
5. One can replicate a file for a configured number of times, which is tolerant in terms of both software and hardware.
6. Re-replicates data blocks automatically on nodes that have failed.
7. Sits on top of native file system

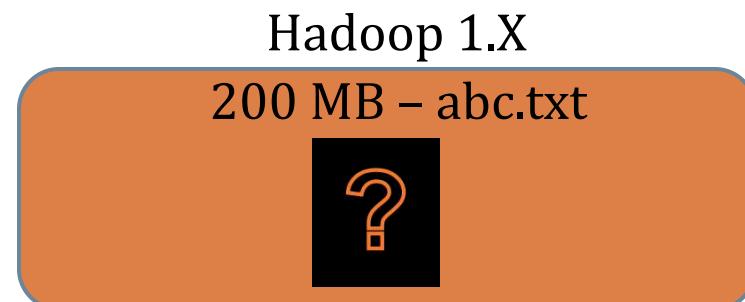
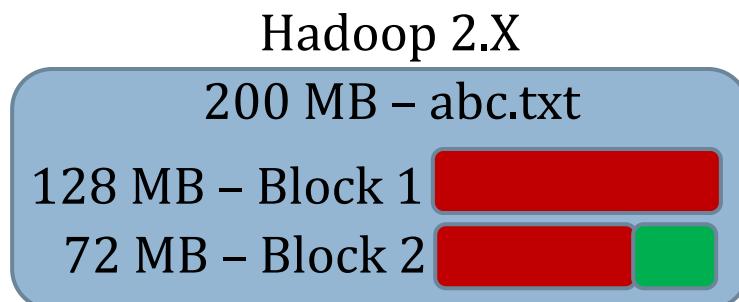
HDFS Physical Architecture

24

Key components of HDFS are as follows:

1. NameNode
2. DataNodes
3. Secondary NameNode
4. Standby NameNode

Blocks: Generally the user data is stored in the files of HDFS. HDFS breaks a large file into smaller pieces called **blocks**. In other words, the minimum amount of data that HDFS can read or write is called a block. By default the block size is 128 MB in Hadoop 2.x and 64 MB in Hadoop 1.x. But it can be increased as per the need to change in HDFS configuration.



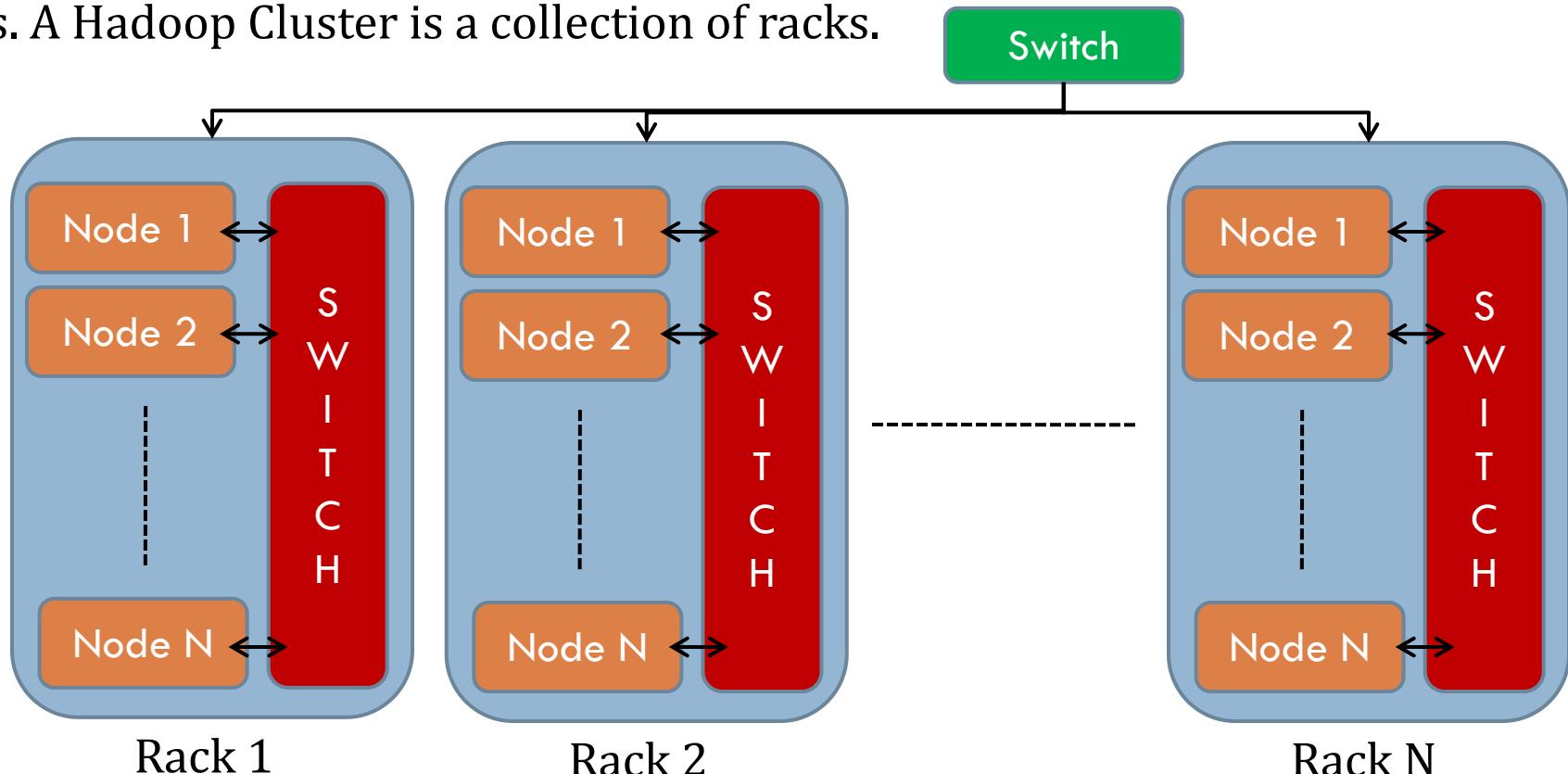
Why block size is large?

1. Reduce the cost of seek time and
2. Proper usage of storage space

Rack

25

A rack is a collection of 30 or 40 nodes that are physically stored close together and are all connected to the same network switch. Network bandwidth between any two nodes in rack is greater than bandwidth between two nodes on different racks. A Hadoop Cluster is a collection of racks.



NameNode

26

1. NameNode is the centerpiece of HDFS.
2. NameNode is also known as the Master.
3. NameNode only stores the metadata of HDFS – the directory tree of all files in the file system, and tracks the files across the cluster.
4. NameNode does not store the actual data or the dataset. The data itself is actually stored in the DataNodes
5. NameNode knows the list of the blocks and its location for any given file in HDFS. With this information NameNode knows how to construct the file from blocks.
6. NameNode is usually configured with a lot of memory (RAM).
7. NameNode is so critical to HDFS and when the NameNode is down, HDFS/Hadoop cluster is inaccessible and considered down.
8. NameNode is a single point of failure in Hadoop cluster.

Configuration

Processors: 2 Quad Core CPUs running @ 2 GHz

RAM: 128 GB

Disk: 6 x 1TB SATA

Network: 10 Gigabit Ethernet

NameNode Metadata

27

1. Metadata stored about the file consists of file name, file path, number of blocks, block Ids, replication level.
2. This metadata information is stored on the local disk. Namenode uses two files for storing this metadata information.
 - FsImage
 - EditLog
3. NameNode in HDFS also keeps in it's memory, location of the DataNodes that store the blocks for any given file. Using that information Namenode can reconstruct the whole file by getting the location of all the blocks of a given file.

Example

(File Name, numReplicas, rack-ids, machine-ids, block-ids, ...)
/user/in4072/data/part-0, 3, r:3, M3, {1, 3}, ...
/user/in4072/data/part-1, 3, r:2, M1, {2, 4, 5}, ...
/user/in4072/data/part-2, 3, r:1, M2, {6, 9, 8}, ...

DataNode

28

1. DataNode is responsible for storing the actual data in HDFS.
2. DataNode is also known as the Slave
3. NameNode and DataNode are in constant communication.
4. When a DataNode starts up it announce itself to the NameNode along with the list of blocks it is responsible for.
5. When a DataNode is down, it does not affect the availability of data or the cluster. NameNode will arrange for replication for the blocks managed by the DataNode that is not available.
6. DataNode is usually configured with a lot of hard disk space. Because the actual data is stored in the DataNode.

Configuration

Processors: 2 Quad Core CPUs running @ 2 GHz

RAM: 64 GB

Disk: 12-24 x 1TB SATA

Network: 10 Gigabit Ethernet

Secondary NameNode

29

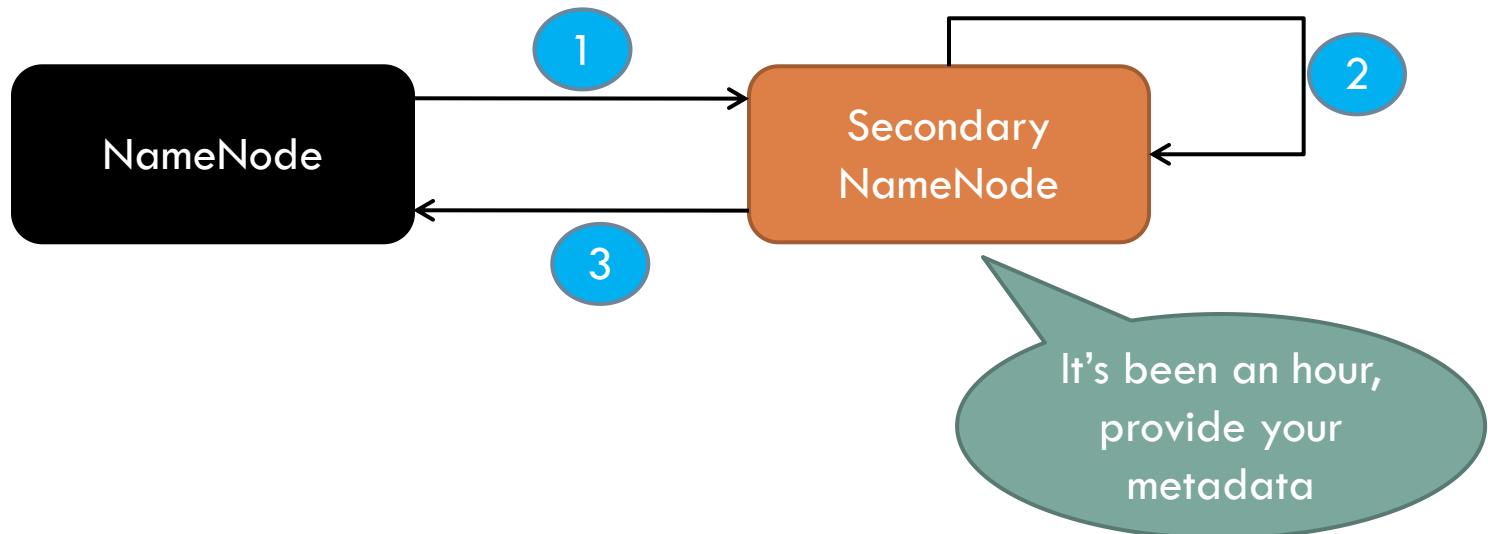
1. Secondary NameNode in Hadoop is more of a helper to NameNode, it is not a backup NameNode server which can quickly take over in case of NameNode failure.
2. EditLog– All the file write operations done by client applications are first recorded in the EditLog.
3. FsImage– This file has the complete information about the file system metadata when the NameNode starts. All the operations after that are recorded in EditLog.
4. When the NameNode is restarted it first takes metadata information from the FsImage and then apply all the transactions recorded in EditLog. NameNode restart doesn't happen that frequently so EditLog grows quite large. That means merging of EditLog to FsImage at the time of startup takes a lot of time keeping the whole file system offline during that process.
5. Secondary NameNode take over this job of merging FsImage and EditLog and keep the FsImage current to save a lot of time. Its main function is to check point the file system metadata stored on NameNode.

Secondary NameNode cont'd

30

The process followed by Secondary NameNode to periodically merge the fsimage and the edits log files is as follows:

1. Secondary NameNode pulls the latest FsImage and EditLog files from the primary NameNode.
2. Secondary NameNode applies each transaction from EditLog file to FsImage to create a new merged FsImage file.
3. Merged FsImage file is transferred back to primary NameNode.



Standby NameNode

31

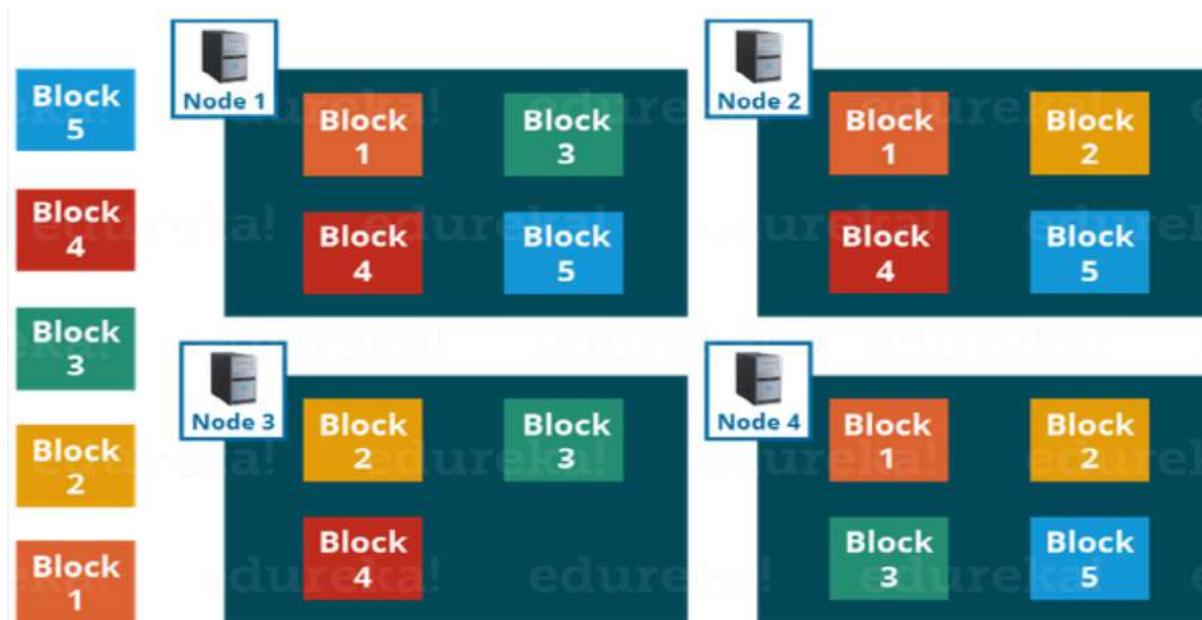
With Hadoop 2.0, built into the platform, HDFS now has **automated failover** with a **hot standby**, with full stack resiliency.

- 1. Automated Failover:** Hadoop pro-actively detects NameNode host and process failures and will automatically switch to the standby NameNode to maintain availability for the HDFS service. There is no need for human intervention in the process – System Administrators can sleep in peace!
- 2. Hot Standby:** Both Active and Standby NameNodes have up to date HDFS metadata, ensuring seamless failover even for large clusters – which means no downtime for your HDP cluster!
- 3. Full Stack Resiliency:** The entire Hadoop stack (MapReduce, Hive, Pig, HBase, Oozie etc.) has been certified to handle a NameNode failure scenario without losing data or the job progress. This is vital to ensure long running jobs that are critical to complete on schedule will not be adversely affected during a NameNode failure scenario.

Replication

32

HDFS provides a reliable way to store huge data in a distributed environment as data blocks. The blocks are also replicated to provide fault tolerance. The default replication factor is 3 which is configurable. Therefore, if a file to be stored of 128 MB in HDFS using the default configuration, it would occupy a space of 384 MB (3×128 MB) as the blocks will be replicated three times and each replica will be residing on a different DataNode.



Rack Awareness

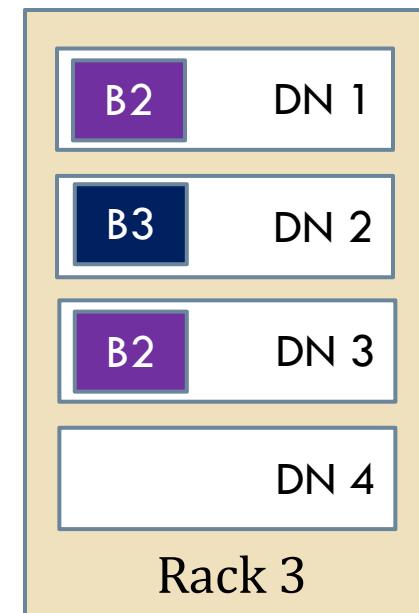
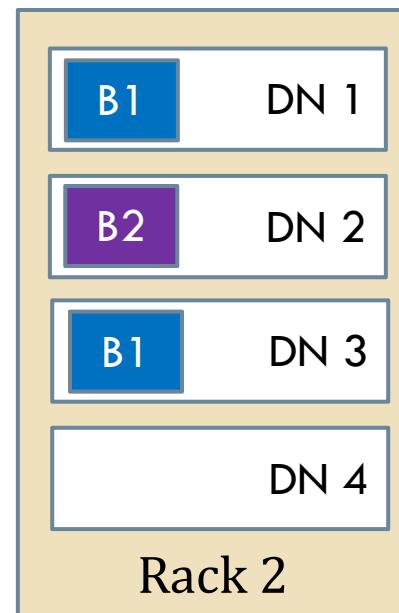
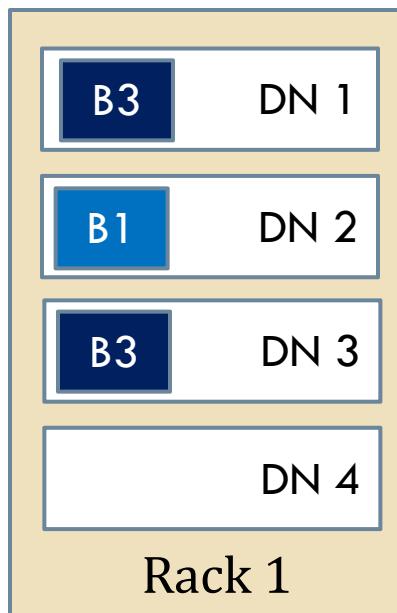
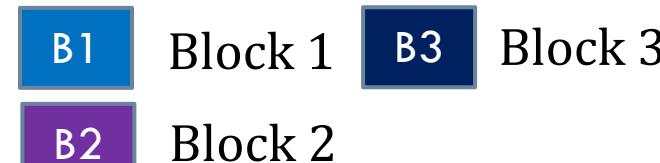
33

All machines in rack are connected using the same network switch and if that network goes down then all machines in that rack will be out of service. Thus the rack is down. Rack Awareness was introduced by Apache Hadoop to overcome this issue. In Rack Awareness, NameNode chooses the DataNode which is closer to the same rack or nearby rack. NameNode maintains Rack ids of each DataNode to achieve rack information. Thus, this concept chooses DataNodes based on the rack information. NameNode in Hadoop makes ensures that all the replicas should not stored on the same rack or single rack. Default replication factor is 3. Therefore according to Rack Awareness Algorithm:

- ❑ When a Hadoop framework creates new block, it places first replica on the local node, and place a second one in a different rack, and the third one is on different node on same remote node.
- ❑ When re-replicating a block, if the number of existing replicas is one, place the second on a different rack.
- ❑ When number of existing replicas are two, if the two replicas are in the same rack, place the third one on a different rack.

Rack Awareness & Replication

34



High-Level Overview

- When data is loaded onto the system it is divided into blocks
 - Typically 64MB or 128MB
- Tasks are divided into two phases
 - Map tasks which are done on small portions of data where the data is stored
 - Reduce tasks which combine data to produce the final output
- A master program allocates work to individual nodes

Fault Tolerance

- Failures are detected by the master program which reassigns the work to a different node
- Restarting a task does not affect the nodes working on other portions of the data
- If a failed node restarts, it is added back to the system and assigned new tasks
- The master can redundantly execute the same task to avoid slow running nodes

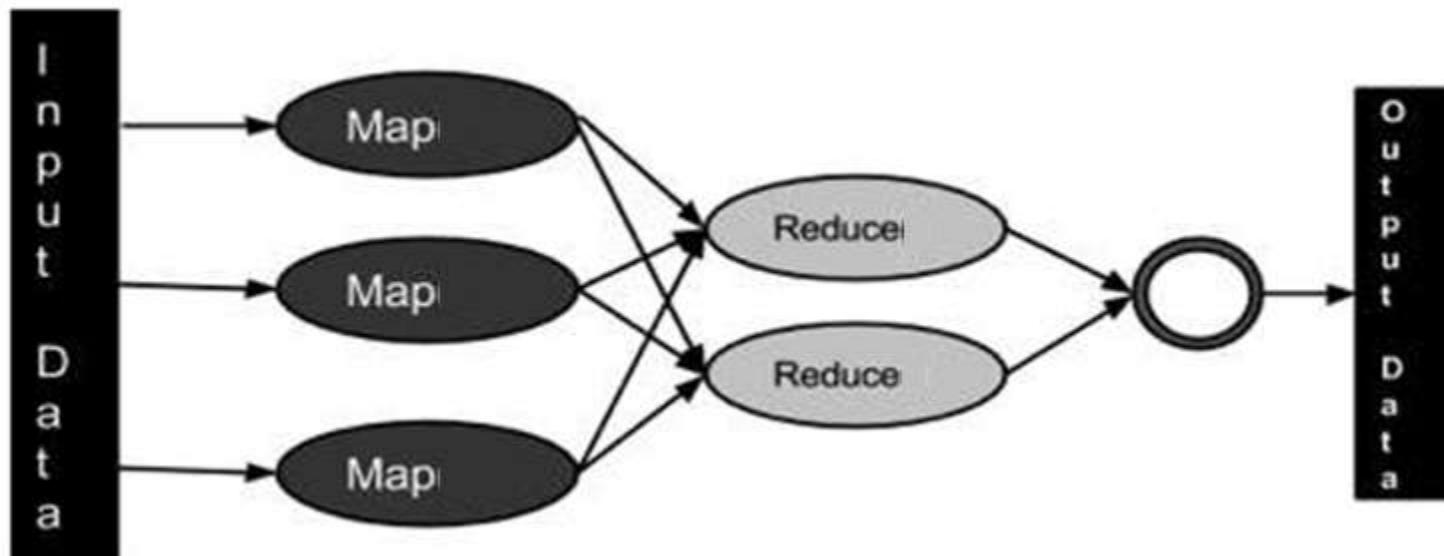
MapReduce

37

1. MapReduce is a processing technique and a program model for distributed computing based on java. It is built on divide and conquer algorithm.
2. In MapReduce Programming, the input dataset is split into independent chunks.
3. It contains two important tasks, namely Map and Reduce.
4. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The processing primitive is called mapper. The processing is done in parallel manner. The output produced by the map tasks serves as intermediate data and is stored on the local disk of that server.
5. Reduce task takes the output from a map as an input and combines those data tuples into a smaller set of tuples. The processing primitive is called reducer. The input and output are stored in a file system.
6. Reduce task is always performed after the map job.
7. The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes and takes care of other tasks such as scheduling, monitoring, re-executing failed tasks etc.

MapReduce cont'd

38



MapReduce cont'd

39

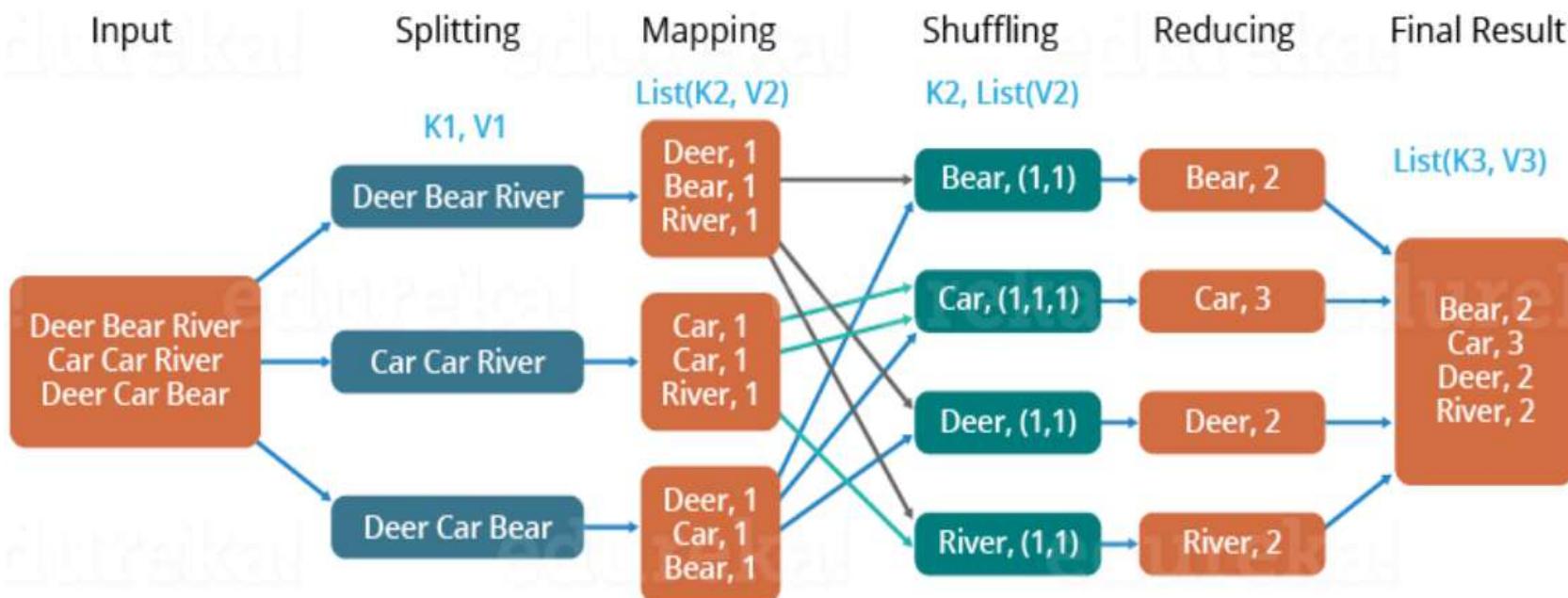
- ❑ The main advantages is that we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster with a configuration change.
- ❑ MapReduce program executes in three stages: map stage, shuffle & sorting stage, and reduce stage.
- ❑ **Map Stage:** The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
- ❑ **Shuffle & Sorting Stage:** Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. Sort phase in MapReduce covers the merging and sorting of map outputs.
- ❑ **Reducer Stage:** The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

How MapReduce Work?

40

At the crux of MapReduce are two functions: Map and Reduce. They are sequenced one after the other.

- ❑ The Map function takes input from the disk as <key,value> pairs, processes them, and produces another set of intermediate <key,value> pairs as output.
- ❑ The Reduce function also takes inputs as <key,value> pairs, and produces <key,value> pairs as output.



Working of MapReduce

41

The types of keys and values differ based on the use case. All inputs and outputs are stored in the HDFS. While the map is a mandatory step to filter and sort the initial data, the reduce function is optional.

$$\begin{aligned} <\text{k1}, \text{v1}\text{>} \rightarrow \text{Map}() \rightarrow \text{list}(<\text{k2}, \text{v2}\text{>}) \\ <\text{k2}, \text{list}(\text{v2})\text{>} \rightarrow \text{Reduce}() \rightarrow \text{list}(<\text{k3}, \text{v3}\text{>}) \end{aligned}$$

Mappers and Reducers are the Hadoop servers that run the Map and Reduce functions respectively. It doesn't matter if these are the same or different servers.

- ❑ **Map:** The input data is first split into smaller blocks. Each block is then assigned to a mapper for processing. For example, if a file has 100 records to be processed, 100 mappers can run together to process one record each. Or maybe 50 mappers can run together to process two records each. The Hadoop framework decides how many mappers to use, based on the size of the data to be processed and the memory block available on each mapper server.

Working of MapReduce cont'd

42

- **Reduce:** After all the mappers complete processing, the framework shuffles and sorts the results before passing them on to the reducers. A reducer cannot start while a mapper is still in progress. All the map output values that have the same key are assigned to a single reducer, which then aggregates the values for that key.

Class Exercise 1

Draw the MapReduce process to count the number of words for the input:

Dog Cat Rat
Car Car Rat
Dog car Rat
Rat Rat Rat

Class Exercise 2

Draw the MapReduce process to find the maximum electrical consumption for each year:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

Hadoop Ecosystem

43

Following are the components that collectively form a Hadoop ecosystem:

- ❑ HDFS: Hadoop Distributed File System
 - ❑ YARN: Yet Another Resource Negotiator
 - ❑ MapReduce: Programming based Data Processing
-
- ❑ Spark: In-memory data processing
 - ❑ PIG, HIVE: query based processing of data services
 - ❑ HBase: NoSQL Database
 - ❑ Mahout, Spark MLLib: Machine Learning algorithm libraries
 - ❑ Solar, Lucene: Searching and Indexing
 - ❑ Zookeeper: Managing cluster
 - ❑ Oozie: Job Scheduling
 - ❑ Sqoop: Data transfer between Hadoop and RDBMS or mainframes
 - ❑ HCatalog: Metadata services

Hadoop Ecosystem cont...

44

PIG

- ❑ It was developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.
- ❑ It is a platform for structuring the data flow, processing and analyzing huge data sets.
- ❑ Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.
- ❑ Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM.
- ❑ Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

Hbase

- ❑ It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.
- ❑ At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data.

Hadoop Ecosystem cont...

45

HIVE

- ❑ With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).
- ❑ It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.
- ❑ Similar to the Query Processing frameworks, HIVE too comes with two components: JDBC Drivers and HIVE Command Line. JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

Oozie

- ❑ It simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit.
- ❑ There are two kinds of jobs i.e Oozie workflow and Oozie coordinator jobs. Oozie workflow is the jobs that need to be executed in a sequentially ordered manner whereas Oozie Coordinator jobs are those that are triggered when some data or external stimulus is given to it.

Hadoop Ecosystem cont...

46

Zookeeper

- ❑ There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency, often.
- ❑ Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.

Mahout

- ❑ It allows machine learning ability to a system or application. Machine Learning helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
- ❑ It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.

Spark

- ❑ It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.
- ❑ It consumes in memory resources hence, thus being faster than the prior in terms of optimization.
- ❑ It is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.

Hadoop Ecosystem cont...

47

Sqoop

- ❑ It is a tool designed to transfer data between Hadoop and relational database.
- ❑ It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.

HCatlog

- ❑ It is a table storage management tool for Hadoop that exposes the tabular data of HIVE metastore to other Hadoop applications.
- ❑ It enables users with different data processing tools (Pig, MapReduce) to easily write data onto a grid.
- ❑ It ensures that users don't have to worry about where or in what format their data is stored.

Solr, Lucene

- ❑ These are the services that perform the task of searching and indexing built on top of Lucene (full text search engine).
- ❑ As Hadoop handles a large amount of data, Solr & Lucene helps in finding the required information from such a large source.
- ❑ It is a scalable, ready to deploy, search/storage engine optimized to search large volumes of text-centric data.

Hadoop Limitations

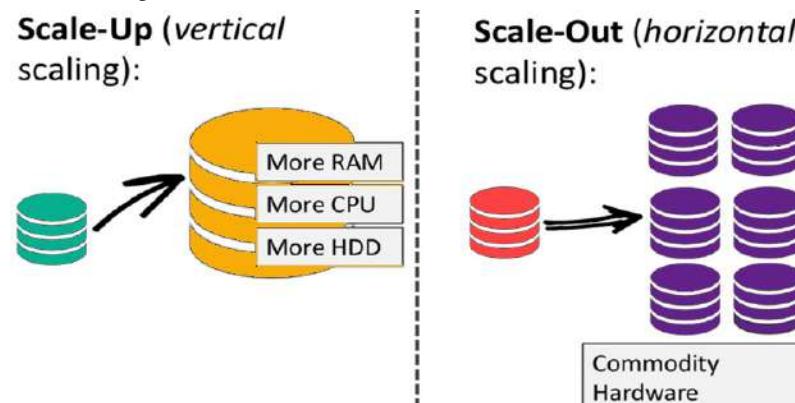
48

- ❑ **Not fit for small data:** Hadoop does not suit for small data. HDFS lacks the ability to efficiently support the random reading of small files because of its high capacity design. The solution to this drawback of Hadoop to deal with small file issue is simple. Just merge the small files to create bigger files and then copy bigger files to HDFS.
- ❑ **Security concerns:** Hadoop is challenging in managing the complex application. If the user doesn't know how to enable a platform who is managing the platform, data can be a huge risk. At storage and network levels, Hadoop is missing encryption, which is a major point of concern. Hadoop supports Kerberos authentication, which is hard to manage. Spark provides a security bonus to overcome the limitations of Hadoop.
- ❑ **Vulnerable by nature:** Hadoop is entirely written in Java, a language most widely used, hence java been most heavily exploited by cyber criminals and as a result, implicated in numerous security breaches.
- ❑ **No caching:** Hadoop is not efficient for caching. In Hadoop, MapReduce cannot cache the intermediate data in memory for a further requirement which diminishes the performance of Hadoop. Spark can overcome this limitation.

NoSQL

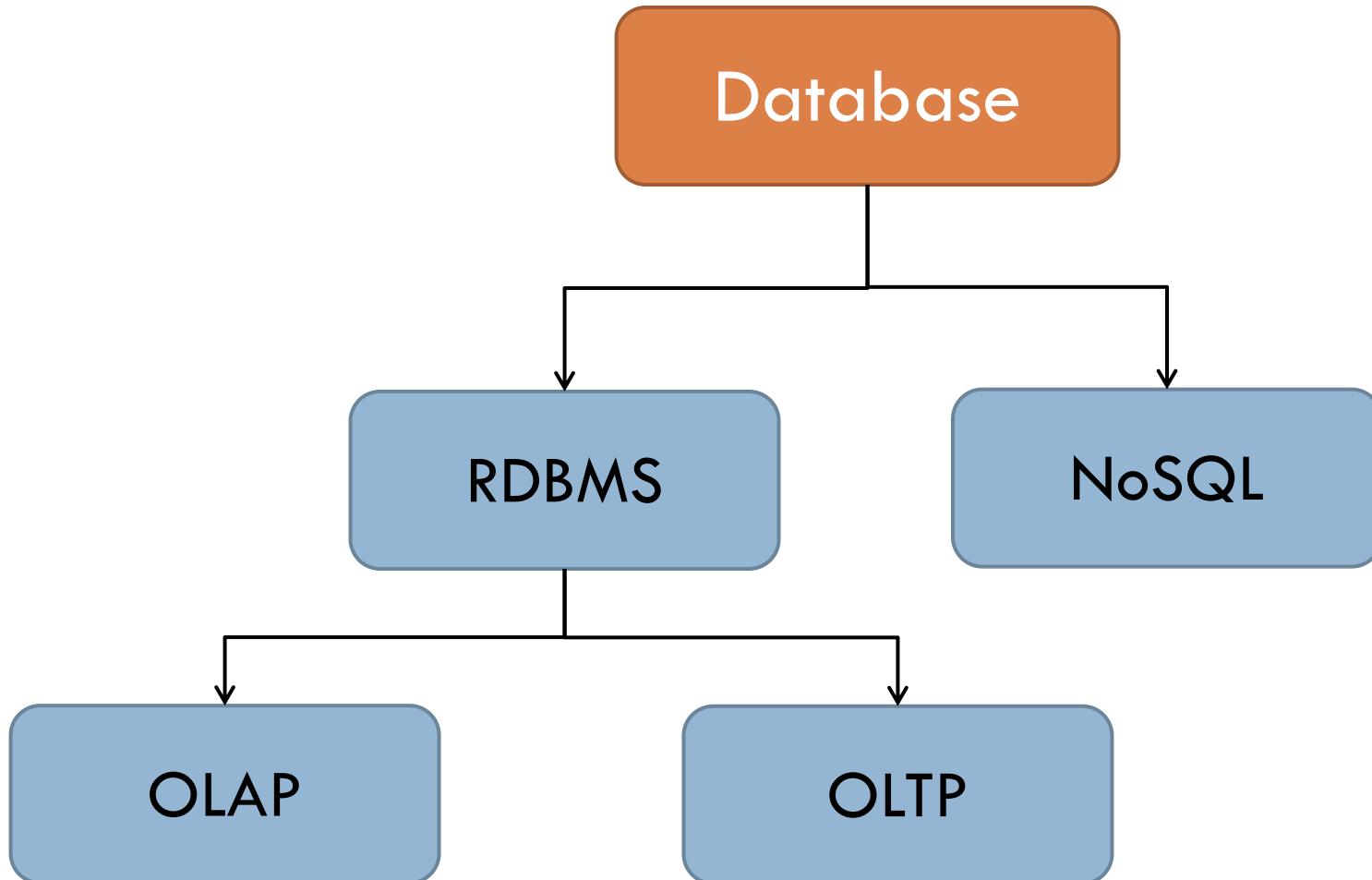
49

- ❑ NoSQL database stands for "Not Only SQL" or "Not SQL."
- ❑ It is a non-relational database, that does not require a fixed schema, and avoids joins.
- ❑ It is used for distributed data stores and specifically targeted for big data, for example Google or Facebook which collects terabytes of data every day for their users.
- ❑ Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, and unstructured data.
- ❑ **It adhere to Brewer's CAP theorem.**
- ❑ The tables are stored as ASCII files and each field is separated by tabs
- ❑ The data scale horizontally.



NoSQL cont...

50



RDBMS vs. NoSQL



51

RDBMS	NoSQL
Relational database	Non-relational, distributed database
Relational model	Model-less approach
Pre-defined schema	Dynamic schema for unstructured data
Table based databases	Document-based or graph-based or wide column store or key-value pairs databases
Vertically scalable (by increasing system resources)	Horizontally scalable (by creating a cluster of commodity machines)
Uses SQL	Uses UnQL (Unstructured Query Language)
Not preferred for large datasets	Largely preferred for large datasets
Not a best fit for hierarchical data	Best fit for hierarchical storage as it follows the key-value pair of storing data similar to JSON
Emphasis on ACID properties	Follows Brewer's CAP theorem

RDBMS vs. NoSQL cont'd

52

RDBMS	NoSQL
Excellent support from vendors	Relies heavily on community support
Supports complex querying and data keeping needs	Does not have good support for complex querying
Can be configured for strong consistency	Few support strong consistency (e.g., MongoDB), few others can be configured for eventual consistency (e.g., Cassandra)
Examples: Oracle, DB2, MySQL, MS SQL, PostgreSQL, etc.	Examples: MongoDB, HBase, Cassandra, Redis, Neo4j, CouchDB, Couchbase, Riak, etc.

OLTP vs. OLAP

53

OLTP	OLAP
Many short transactions	Long transactions, complex queries
Example: <ul style="list-style-type: none">- Update account balance- Add book to shopping cart- Enroll in course	Example: <ul style="list-style-type: none">- Count the classes with fewer than 10 classes- Report total sales for each dept in each month
Queries touch small amounts of data (few records)	Queries touch large amounts of data
Updates are frequent	Updates are infrequent
Concurrency is biggest performance problem	Individual queries can require lots of resources

CAP Theorem

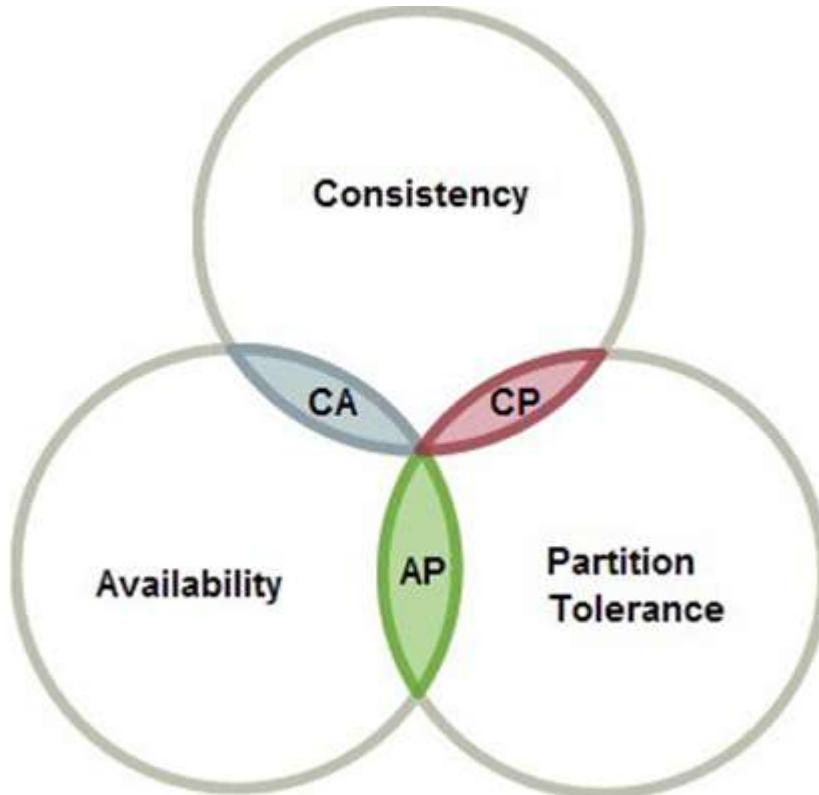
54

CAP Theorem: In the past, when there was a need to store more data or increase processing power, the common option was to scale vertically (get more powerful machines) or further optimize the existing code base. However, with the advances in parallel processing and distributed systems, it is more common to expand horizontally, or have more machines to do the same task in parallel. However, in order to effectively pick the tool of choice like Spark, Hadoop, Kafka, Zookeeper and Storm in Apache project, a basic idea of CAP Theorem is necessary. The CAP theorem is called the **Brewer's Theorem**. It states that a distributed computing environment can only have 2 of the 3: **Consistency**, **Availability** and **Partition Tolerance** – one must be sacrificed.

- ❑ **Consistency** implies that every read fetches the last write
- ❑ **Availability** implies that reads and writes always succeed. In other words, each non-failing node will return a response in a reasonable amount of time
- ❑ **Partition Tolerance** implies that the system will continue to function when network partition occurs

CAP Theorem cont'd

55



Source: Towards Data Science

The CAP theorem categorizes systems into three categories:

CP (Consistent and Partition Tolerant) - a system that is consistent and partition tolerant but never available. CP is referring to a category of systems where availability is sacrificed only in the case of a network partition.

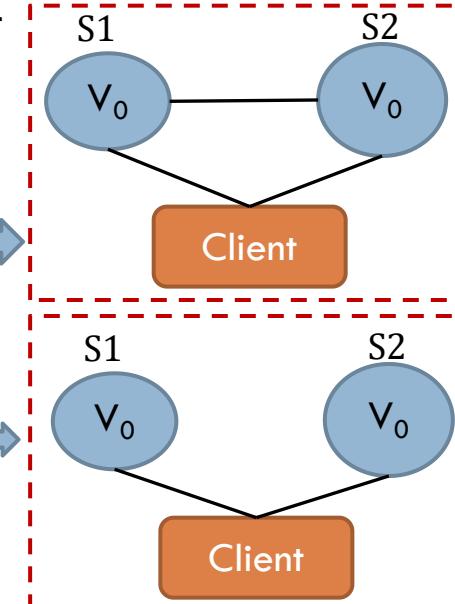
CA (Consistent and Available) - CA systems are consistent and available systems in the absence of any network partition. Often a single node's DB servers are categorized as CA systems. Single node DB servers do not need to deal with partition tolerance and are thus considered CA systems.

AP (Available and Partition Tolerant) - These are systems that are available and partition tolerant but cannot guarantee consistency.

CAP Theorem Proof

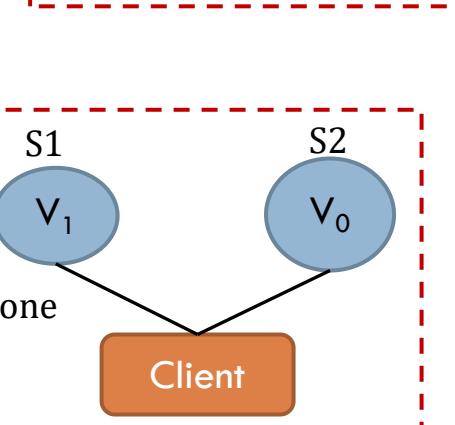
56

Let's consider a very simple distributed system. Our system is composed of two servers, S1 and S2. Both of these servers are keeping track of the same variable, v, whose value is initially v_0 . S1 and S2 can communicate with each other and can also communicate with external client. Here's what the system looks like.

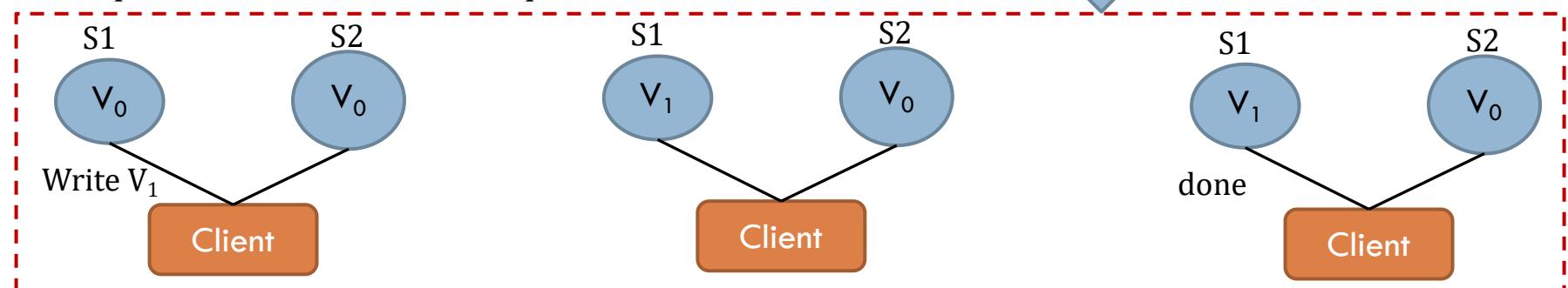


Assume for contradiction that the system is consistent, available, and partition tolerant.

The first thing we do is partition our system. It looks like this.

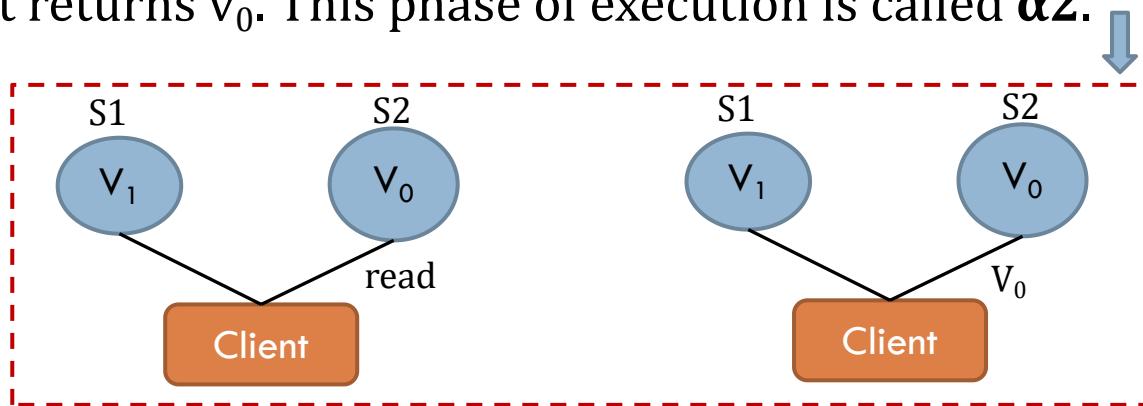


Next, the client request that v_1 be written to S1. Since the system is available, S1 must respond. Since the network is partitioned, however, S1 cannot replicate its data to S2. This phase of execution is called **$\alpha 1$** .



CAP Theorem Proof cont'd

Next, the client issue a read request to S2. Again, since the system is available, S2 must respond and since the network is partitioned, S2 cannot update its value from S1. It returns v_0 . This phase of execution is called $\alpha 2$.



S2 returns v_0 to the client after the client had already written v_1 to S1. This is inconsistent.

We assumed a consistent, available, partition tolerant system existed, but we just showed that there exists an execution for any such system in which the system acts inconsistently. Thus, no such system exists.

Summary - what is NoSQL?

58

- ❑ **It's more than rows in tables** —NoSQL systems store and retrieve data from many formats: key-value stores, graph databases, column-family (Bigtable) stores, document stores, and even rows in tables.
- ❑ **It's free of joins** —NoSQL systems allow you to extract your data using simple interfaces without joins.
- ❑ **It's schema-free** — NoSQL systems allow to drag-and-drop data into a folder and then query it without creating an entity-relational model.
- ❑ **It works on many processors** — NoSQL systems allow you to store your database on multiple processors and maintain high-speed performance.
- ❑ **It uses shared-nothing commodity computers** — Most (but not all) NoSQL systems leverage low-cost commodity processors that have separate RAM and disk.
- ❑ **It supports linear scalability** — When you add more processors, you get a consistent increase in performance.
- ❑ **It's innovative** — NoSQL offers options to a single way of storing, retrieving, and manipulating data. NoSQL supporters (also known as NoSQLers) have an inclusive attitude about NoSQL and recognize SQL solutions as viable options. To the NoSQL community, NoSQL means “Not only SQL.”

Summary - what NoSQL is not?

59

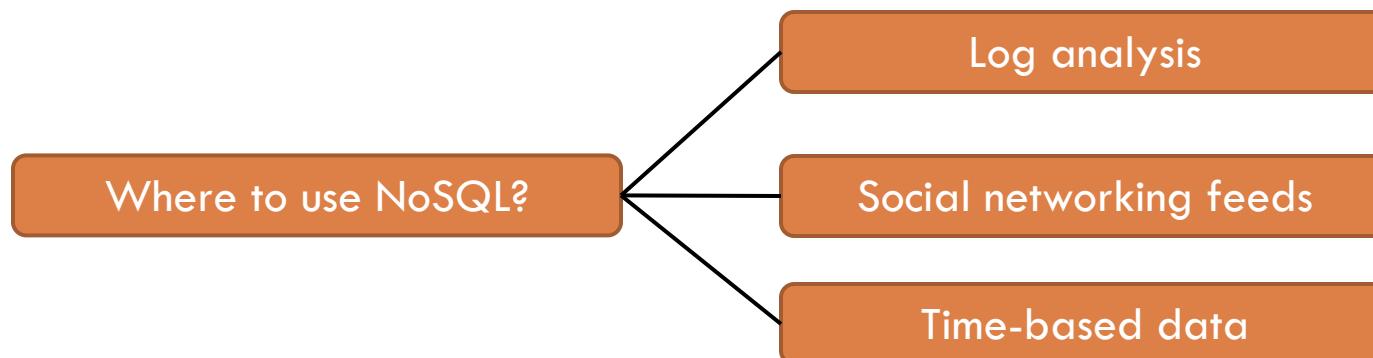
- ❑ **It's not about the SQL language** — The definition of NoSQL isn't an application that uses a language other than SQL. SQL as well as other query languages are used with NoSQL databases.
- ❑ **It's not only open source** — Although many NoSQL systems have an open source model, commercial products use NOSQL concepts as well as open source initiatives. You can still have an innovative approach to problem solving with a commercial product.
- ❑ **It's not only big data** — Many, but not all, NoSQL applications are driven by the inability of a current application to efficiently scale when big data is an issue. Though volume and velocity are important, NoSQL also focuses on variability and agility.
- ❑ **It's not about cloud computing** — Many NoSQL systems reside in the cloud to take advantage of its ability to rapidly scale when the situation dictates. NoSQL systems can run in the cloud as well as in your corporate data center.
- ❑ **It's not about a clever use of RAM and SSD** — Many NoSQL systems focus on the efficient use of RAM or solid state disks to increase performance. Though this is important, NoSQL systems can run on standard hardware.
- ❑ **It's not an elite group of products** — NoSQL isn't an exclusive club with a few products. There are no membership dues or tests required to join. To be considered a NoSQLer, you only need to convince others that you have innovative solutions to their business problems.

Why and Uses of NoSQL

60

Why: In today's time data is becoming easier to access and capture through third parties such as Facebook, Google+ and others. Personal user information, social graphs, geo location data, user-generated content and machine logging data are just a few examples where the data has been increasing exponentially. To avail the above service properly, it is required to process huge amount of data which SQL databases were never designed. The evolution of NoSql databases is to handle these huge data properly.

Uses:

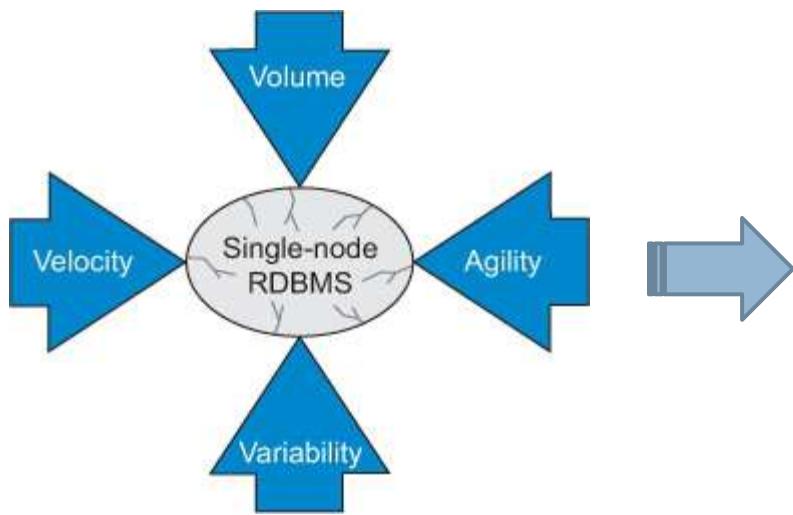


NoSQL Business Drivers



61

Enterprises today need highly reliable, scalable and available data storage across a configurable set of systems that act as storage nodes. The needs of organizations are changing rapidly. Many organizations operating with single CPU and relational database management systems (RDBMS) were not able to cope up with the speed in which information needs to be extracted. Businesses have to capture and analyze a large amount of variable data, and make immediate changes in their business based on their findings.



The figure shows how the demands of volume, velocity, variability, and agility play a key role in the emergence of NoSQL solutions. As each of these drivers applies pressure to the single-processor relational model, its foundation becomes less stable and in time no longer meets the organization's needs.

NoSQL Business Drivers cont...



62

Volume

- ❑ Without a doubt, the key factor pushing organizations to look at alternatives to their current RDBMSs is a need to query big data using clusters of commodity processors.
- ❑ Until around 2005, performance concerns were resolved by purchasing faster processors. In time, the ability to increase processing speed was no longer an option. As chip density increased, heat could no longer dissipate fast enough without chip overheating. This phenomenon, known as the **power wall**, forced systems designers to shift their focus from increasing speed on a single chip to using more processors working together.
- ❑ The need to scale out (also known as horizontal scaling), rather than scale up (faster processors), moved organizations from serial to parallel processing where data problems are split into separate paths and sent to separate processors to divide and conquer the work.

NoSQL Business Drivers cont...



63

Velocity

- ❑ Though big data problems are a consideration for many organizations moving away from RDBMSs, the ability of a single processor system to rapidly read and write data is also key.
- ❑ Many single-processor RDBMSs are unable to keep up with the demands of real-time inserts and online queries to the database made by public-facing websites.
- ❑ RDBMSs frequently index many columns of every new row, a process which decreases system performance.
- ❑ When single-processor RDBMSs are used as a back end to a web store front, the random bursts in web traffic slow down response for everyone, and tuning these systems can be costly when both high read and write throughput is desired.

NoSQL Business Drivers cont...



64

Variability

- ❑ Companies that want to capture and report on exception data struggle when attempting to use rigid database schema structures imposed by RDBMSs. For example, if a business unit wants to capture a few custom fields for a particular customer, all customer rows within the database need to store this information even though it doesn't apply.
- ❑ Adding new columns to an RDBMS requires the system be shut down and ALTER TABLE commands to be run.
- ❑ When a database is large, this process can impact system availability, costing time and money.

NoSQL Business Drivers cont...



65

Agility

- ❑ The most complex part of building applications using RDBMSs is the process of putting data into and getting data out of the database.
- ❑ If the data has nested and repeated subgroups of data structures, one needs to include an object-relational mapping layer. The responsibility of this layer is to generate the correct combination of INSERT, UPDATE, DELETE, and SELECT SQL statements to move object data to and from the RDBMS persistence layer.
- ❑ This process isn't simple and is associated with the largest barrier to rapid change when developing new or modifying existing applications.

NoSQL Business Drivers cont...



66

The desirable features of NoSQL that drive business are:

- ❑ **24x 7 Data availability:** In the highly competitive world today, downtime is equated to real dollars lost and is deadly to a company's reputation. Hardware failures are bound to occur. Care has to be taken that there is no single point of failure and system needs to show fault tolerance. For this, both function and data are to be replicated so that if database servers or "nodes" fail, the other nodes in the system are able to continue with operations without data loss. NoSQL database environments are able to provide this facility. System updates can be made dynamically without having to take the database offline.
- ❑ **Location transparency:** The ability to read and write to a storage node regardless of where that I/O operation physically occurs is termed as "Location Transparency or Location Independence". Customers in many different geographies need to keep data local at those sites for fast access. Any write functionality that updates a node in one location, is propagated out from that location so that it is available to users and systems at other locations.

NoSQL Business Drivers cont...



67

- ❑ **Schema-less data model:** Most of the business data is unstructured and unpredictable which a RDBMS cannot cater to. NoSQL database system is a schema-free flexible data model that can easily accept all types of structured, semi-structured and unstructured data. Also relational model has scalability and performance problems when it has to manage large data volumes. NoSQL data model can handle this easily to deliver very fast performance for both read and write operations.
- ❑ **Modern day transaction analysis:** Most of the transaction details relate to customer profile, reviews on products, branding, reputation, building business strategy, trading decisions, etc. that do not require ACID transactions. The data consistency denoted by “C” in ACID property in RDBMSs is enforced via foreign keys/referential integrity constraints. This type of consistency is not required to be used in progressive data management systems such as NoSQL databases since there is no JOIN operation. Here, the “Consistency” is stated in the CAP theorem that signifies the immediate or eventual consistency of data across all nodes that participate in a distributed database.

NoSQL Business Drivers cont...



68

- ❑ **Architecture that suits big data:** NoSQL solutions provide modern architectures for applications that require high degrees of scale, data distribution and continuous availability. For this multiple data center support with which a NoSQL environment complies is one of the requirements. The solution should not only look into today's big data needs but also suit greater time horizons.
- ❑ **Analytics and business intelligence:** A key business strategic driver that suggests the implementation of a NoSQL database environment is the need to mine the data that is being collected in order to derive insights to gain competitive advantage. Traditional relational database system poses great difficulty in extracting meaningful business intelligence from very high volumes of data. NoSQL database systems not only provide storage and management of big data but also deliver integrated data analytics that provides instant understanding of complex datasets and facilitate various options for easy decision-making.

NoSQL Case Studies



69

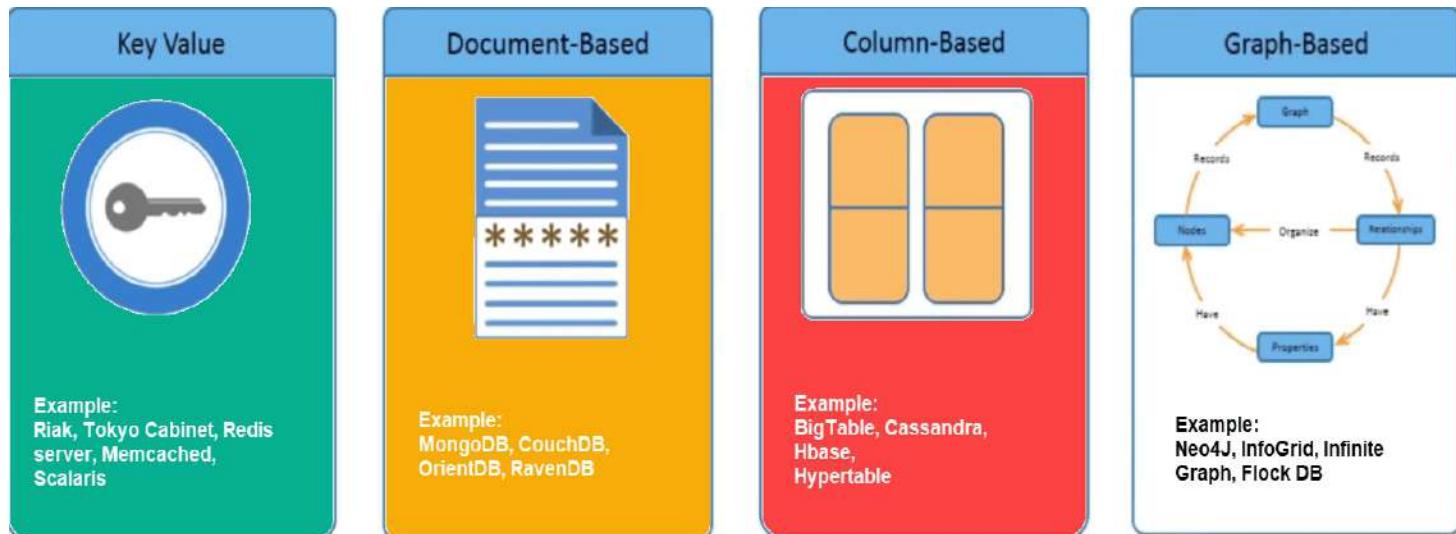
Case study	Driver	Finding
Amazon's DynamoDB	Need to accept a web order 24 hours a day, 7 days a week.	A key-value store with a simple interface can be replicated even when there are large volumes of data to be processed.
Google's Bigtable	Need to flexibly store tabular data in a distributed system.	By using a sparse matrix approach, users can think of all data as being stored in a single table with billions of rows and millions of columns without the need for up-front data modeling.
MarkLogic	Need to query large collections of XML documents stored on commodity hardware using standard query languages.	By distributing queries to commodity servers that contain indexes of XML documents, each server can be responsible for processing data in its own local disk and returning the results to a query server.

NoSQL Data Architectural Patterns



70

There are mainly four categories of NoSQL data stores. Each of these categories has its unique attributes and limitations.



Performance	High	High	High	Variable
Scalability	High	High	Moderate	Minimal
Flexibility	High	Variable (high)	High	Variable (low)
Functionality	Variable	Variable	Variable	Graph Theory

Key Value

71

Data is stored in key/value pairs. It is designed in such a way to handle lots of data and heavy load. Key-value pair storage databases store data as a hash table where each key is unique, and the value can be a JSON, BLOB, string, etc. It is one of the most basic types of NoSQL databases. This kind of NoSQL database is used as a collection, dictionaries, associative arrays, etc. Key value stores help the developer to store schema-less data. They work best for shopping cart contents. Redis, Dynamo, Riak are some examples of key-value store.

SQL

ID = 1	Name John	Age 27	State California
ID = 2	Name Daniel	Age 32	State Montana
ID = 3	Name Mary	Age 31	State Washington



ID	Name	Age	State
1	John	27	California



NoSQL – Key Value	
Key (i.e. ID)	Values
1	Name: John Age:27 State: California

Document-Based

72

Document-Oriented NoSQL DB stores and retrieves data as a key value pair but the value part is stored as a document. The document is stored in JSON or XML formats. The document type is mostly used for CMS (Content Management Systems), blogging platforms, real-time analytics & e-commerce applications. It should not use for complex transactions which require multiple operations or queries against varying aggregate structures.

SQL				→	NoSQL – Document-Based	
ID	Name	Age	State		Key (ID)	Value (JSON)
1	John	27	California		1	{ "Name": John, "Age": 27, "State": California }

JSON vs. XML format

73

JSON

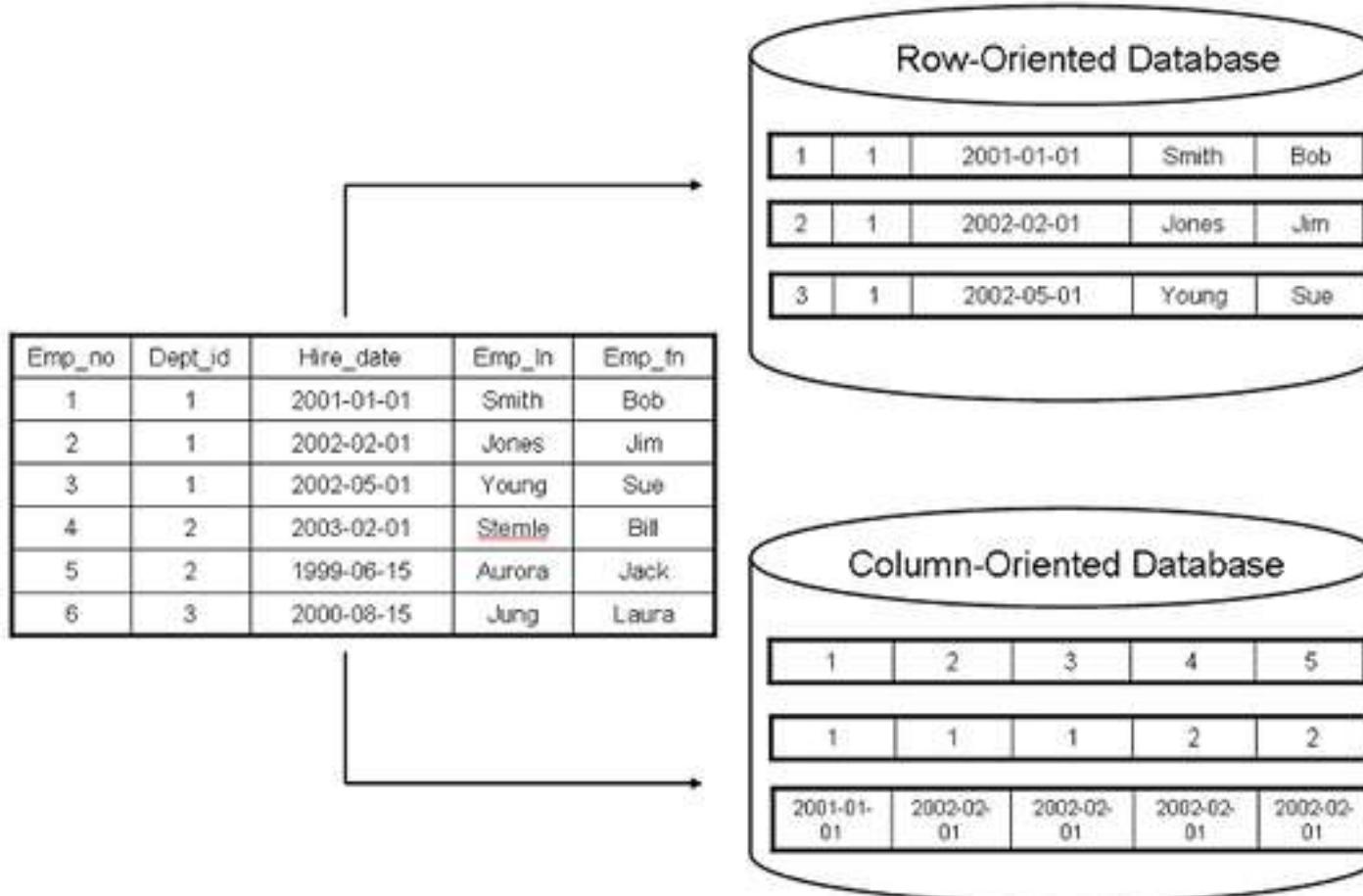
```
{  
    "firstName": "John",  
    "lastName": "Smith",  
    "age": 25,  
    "address": {  
        "streetAddress": "21 2nd Street",  
        "city": "New York",  
        "state": "NY",  
        "postalCode": "10021-3100"  
    },  
    "phoneNumbers": [  
        {  
            "type": "home",  
            "number": "212 555-1234"  
        },  
        {  
            "type": "office",  
            "number": "646 555-4567"  
        }  
    ]  
}
```

XML

```
<person>  
    <firstName>John</firstName>  
    <lastName>Smith</lastName>  
    <age>25</age>  
    <address>  
        <streetAddress>21 2nd Street</streetAddress>  
        <city>New York</city>  
        <state>NY</state>  
        <postalCode>10021</postalCode>  
    </address>  
    <phoneNumbers>  
        <phoneNumber>  
            <type>home</type>  
            <number>212 555-1234</number>  
        </phoneNumber>  
        <phoneNumber>  
            <type>fax</type>  
            <number>646 555-4567</number>  
        </phoneNumber>  
    </phoneNumbers>  
</person>
```

Column-Oriented vs. Row-Oriented Database

74



Column-Oriented vs. Row-Oriented Database cont'd

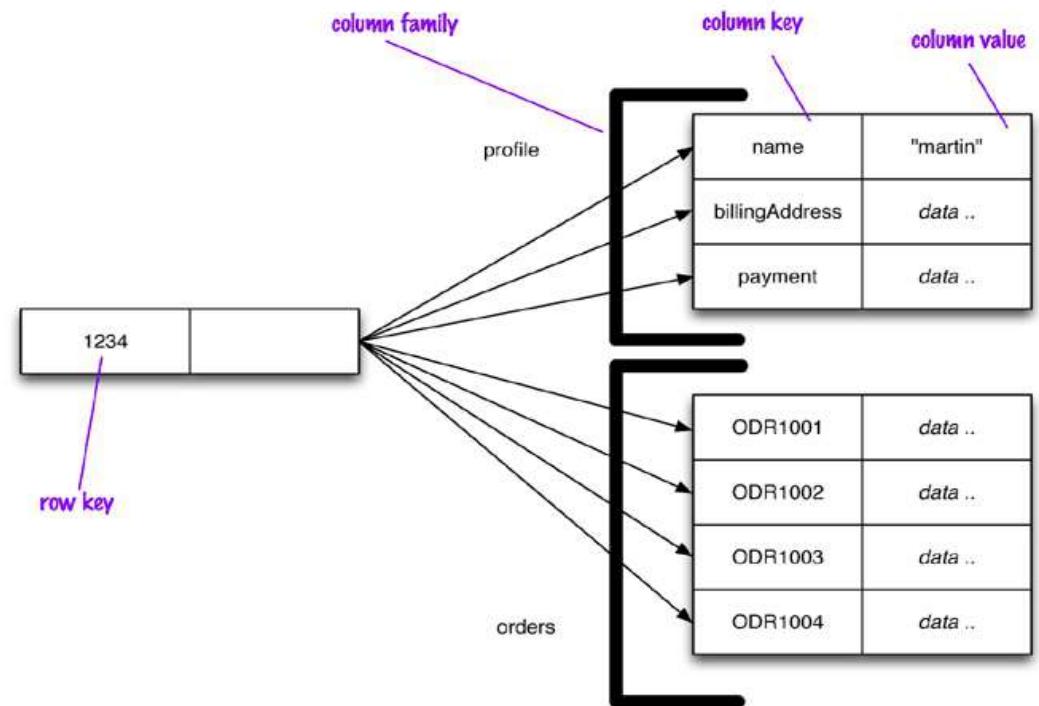
Row Oriented	Column Oriented
Data is stored and retrieved one row at a time and hence could read unnecessary data if some of the data in a row are required.	In this type of data stores, data are stored and retrieve in columns and hence it can only able to read only the relevant data if required.
Records in Row Oriented Data stores are easy to read and write.	In this type of data stores, read and write operations are slower as compared to row-oriented.
Row-oriented data stores are best suited for online transaction system.	Column-oriented stores are best suited for online analytical processing.
These are not efficient in performing operations applicable to the entire datasets and hence aggregation in row-oriented is an expensive job or operations.	These are efficient in performing operations applicable to the entire dataset and hence enables aggregation over many rows and columns.

Column-Based Database



76

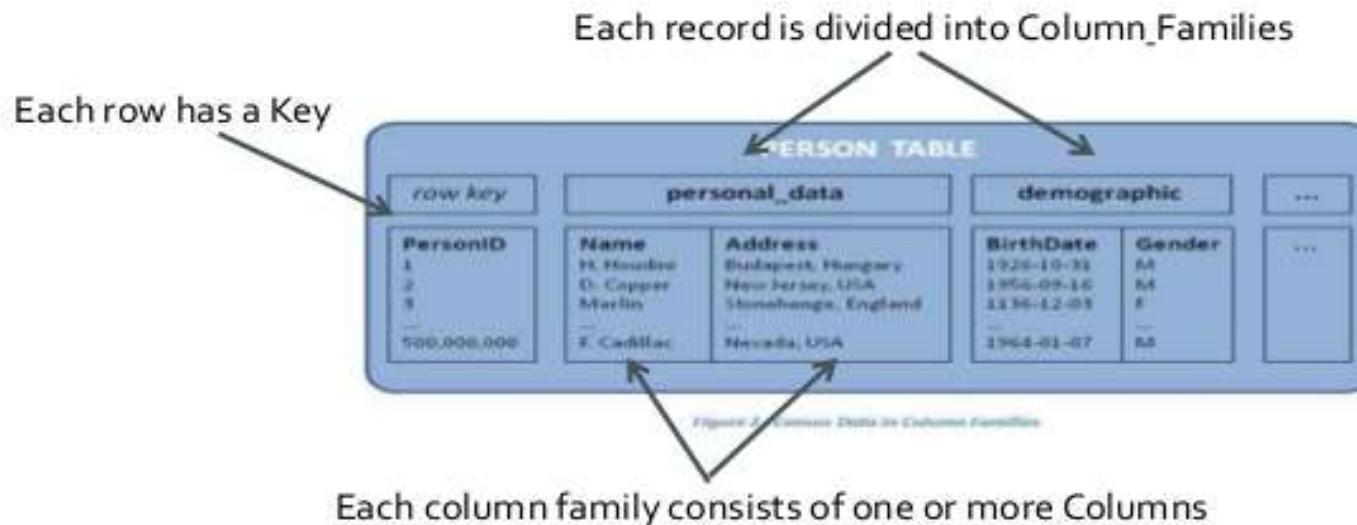
Column-oriented databases work on column family and based on BigTable paper by Google. Every column is treated separately. Values of single column databases are stored contiguously. They deliver high performance on aggregation queries like SUM, COUNT, AVG, MIN etc. as the data is readily available in a column. Such NoSQL databases are widely used to manage data warehouses, business intelligence, CRM, Library card catalogs etc.



Column-Based cont'd

77

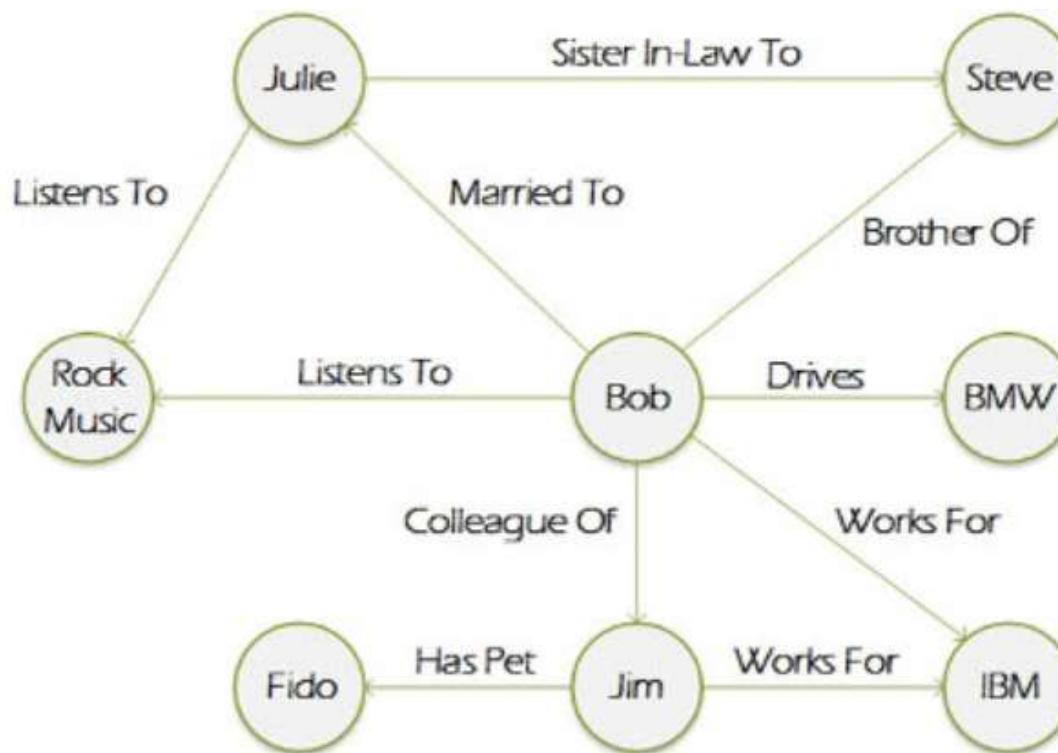
Column Families



Graph-Based

78

A graph type database stores entities as well the relations amongst those entities. The entity is stored as a node with the relationship as edges. An edge gives a relationship between nodes. Every node and edge has a unique identifier. Graph base database mostly used for social networks, logistics, spatial data.



Variations of NoSQL Architectural Patterns

79

- ❑ The key-value store, column family store, document store and graph store patterns can be modified based on different aspects of the system and its implementation. Database architecture could be distributed (manages single database distributed in multiple servers located at various sites) or federated (manages independent and heterogeneous databases at multiple sites).
- ❑ The variations in architecture are based on system requirements like agility, availability (anywhere, anytime), intelligence, scalability, collaboration and low latency. Various technologies support the architectural strategies to satisfy the above requirement. For example, agility is given as a service using virtualization or cloud computing; availability is the service given by internet and mobility; intelligence is given by machine learning and predictive analytics; scalability (flexibility of using commodity machines) is given by Big Data Technologies/cloud platforms; collaboration is given by (enterprise-wide) social network application; and low latency (event driven) is provided by in-memory databases.

Using NoSQL to Manage Big Data

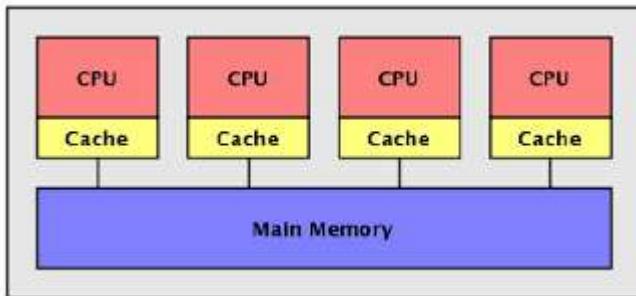
80

- ❑ NoSQL solution is used to handle and manage big data. NoSQL with their inherently horizontal scale out architectures solves big data problems by moving data to queries, uses hash rings to distribute the load, replicates the scale reads, and allows the database to distribute queries evenly in order to make systems run fast.
- ❑ In the distributed computing architecture, there are two ways of resource sharing possible or share nothing. The memory can be shared or disk can be shared (by CPUs); or no resources shared. The three of them can be considered as shared memory, shared disk, and shared-nothing. Each of these architectures works with different types of data to solve big data problems. In shared memory, many CPUs access a single shared memory over a high-speed bus. This system is ideal for large computation and also for graph stores. For graph traversals to be fast, the entire graph should be in main memory. The shared disk system, processors have independent memory but shares disk space using a storage area network (SAN). **Big data uses commodity machines which shares nothing (shares no resources).**

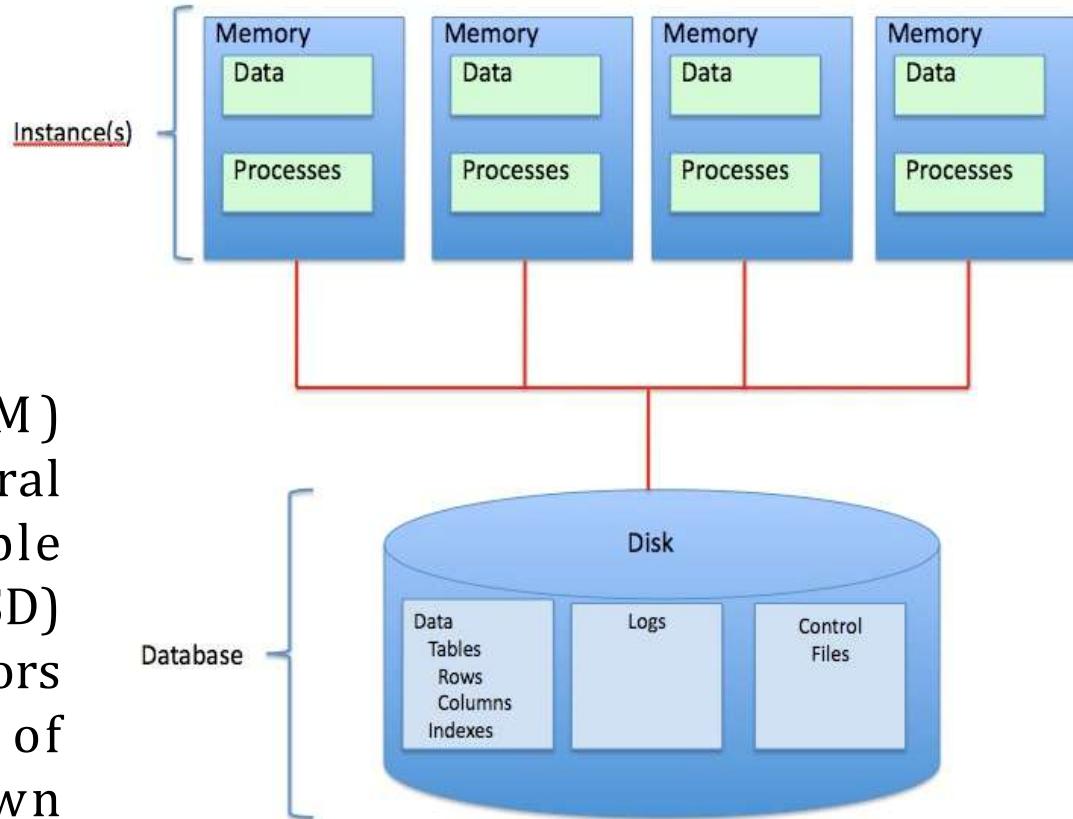
Using NoSQL to Manage Big Data cont...

81

SM



SD



In a shared memory (SM) architecture, a common central memory is shared by multiple processors. In a shared disk (SD) architecture, multiple processors share a common collection of disks while having their own private memory.

Using NoSQL to Manage Big Data cont...

82

In a shared nothing (SN) architecture, neither memory nor disk is shared among multiple processors.

Advantages:

- ❑ **Fault Isolation:** provides the benefit of isolating fault. A fault in a single machine or node is contained and confined to that node exclusively and exposed only through messages.
- ❑ **Scalability:** If the disk is a shared resource, synchronization will have to maintain a consistent shared state and it means that different nodes will have to take turns to access the critical data. This imposes a limit on how many nodes can be added to the distributed shared disk system, this compromising on scalability.

Advantages of NoSQL

83

- ❑ Can be used as Primary or Analytic Data Source
- ❑ Big Data Capability
- ❑ No Single Point of Failure
- ❑ Easy Replication
- ❑ No Need for Separate Caching Layer
- ❑ Provides fast performance and horizontal scalability.
- ❑ Can handle structured, semi-structured, and unstructured data with equal effect
- ❑ NoSQL databases don't need a dedicated high-performance server
- ❑ Support Key Developer Languages and Platforms
- ❑ Simple to implement than using RDBMS
- ❑ It can serve as the primary data source for online applications.
- ❑ Handles big data which manages data velocity, variety, volume, and complexity
- ❑ Excels at distributed database and multi-data center operations
- ❑ Eliminates the need for a specific caching layer to store data
- ❑ Offers a flexible schema design which can easily be altered without downtime or service disruption

Disadvantages of NoSQL

84

- ❑ No standardization rules
- ❑ Limited query capabilities
- ❑ RDBMS databases and tools are comparatively mature
- ❑ It does not offer any traditional database capabilities, like consistency when multiple transactions are performed simultaneously.
- ❑ When the volume of data increases it is difficult to maintain unique values as keys become difficult
- ❑ Doesn't work as well with relational data
- ❑ The learning curve is stiff for new developers
- ❑ Open source options so not so popular for enterprises.

Visualization

85

Visualization is a pictorial representation technique. Anything which is represented in pictorial or graphical form, with the help of diagrams, charts, pictures, flowcharts etc. is known as visualization. Data visualization is a pictorial or visual representation of data with the help of visual aids such as graphs, bar, histograms, tables, pie charts, mind maps etc.

Ways of Representing Visual Data

The data is first analyzed and then the result is visualized. There are 2 ways to visualize a data, namely, Infographics and data visualization.

Infographics – It is the visual representation of information or data rapidly.

Data Visualization – It is the study of representing data or information in a visual form.

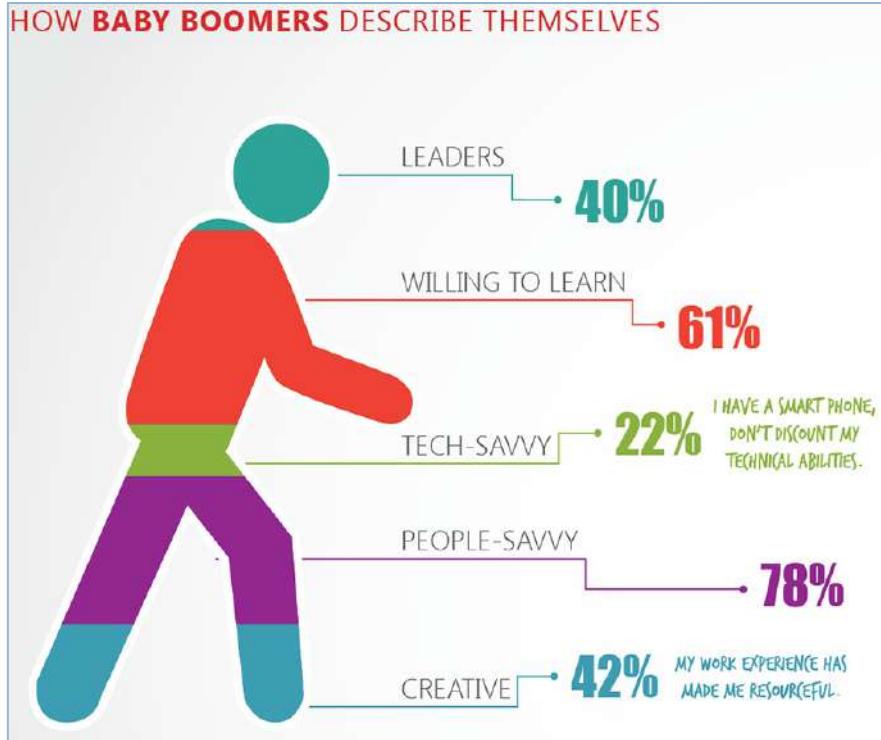
Difference: **Infographics** tell a premeditated story to guide the audience to conclusions (subjective). **Data visualizations** let the audience draw their own conclusions (objective).

An infographic can contain data visualizations but not the other way around.

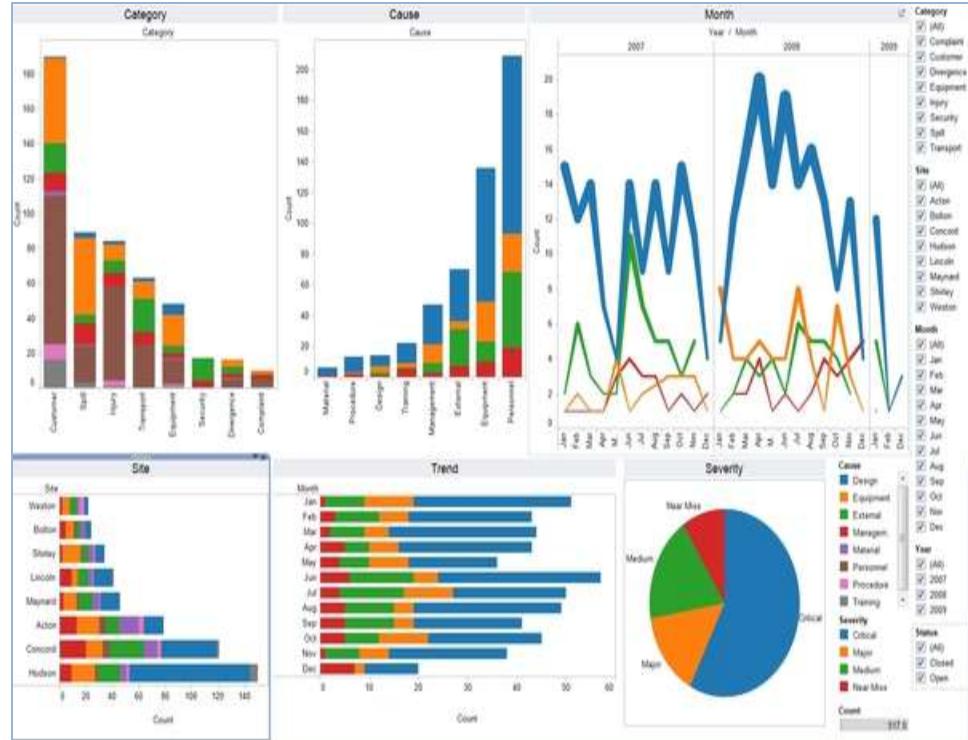
Infographics vs. Data Visualization

86

Infographics



Data Visualization



Infographics vs. Data Visualization cont'd

87

Infographics are:

- ❑ Best for telling a premeditated story and offer subjectivity.
- ❑ Best for guiding the audience to conclusions and point out relationships.
- ❑ Created manually for one specific dataset.

It is used for Marketing content, Resumes, Blog posts, and Case studies etc.

Data visualizations are:

- ❑ Best for allowing the audience to draw their own conclusions, and offer objectivity
- ❑ Ideal for understanding data at a glance
- ❑ Automatically generated for arbitrary datasets

It is used for Dashboards, Scorecards, Newsletters, Reports, and Editorials etc.

Data Visualization Purpose



88

- Data presented in the form of graphics can be **analyzed better** than the data presented in words.
- Patterns, trends, outliers and correlations that **might go undetected in text-based data** can be **exposed and recognized easier** with data visualization software.
- Data scientists can use data visualizations to make their information **more actionable**. Illustrations, graphs, charts and spreadsheets can turn dull reports into something illuminating, where it's **easier to gather insight** and **actionable results**.
- Data Visualization help to **transmit a huge amount of information** to the human brain **at a glance**.
- Data Visualization **point out key or interesting breakthrough in a large dataset**.

Techniques Used for Data Representation



89

Data can be presented in various forms, which include simple line diagrams, bar graphs tables, metrics etc. Techniques used for a visual representation of the data are as follows:

- Map
- Parallel Coordinate Plot
- Venn Diagram
- Timeline
- Euler Diagram
- Hyperbolic Trees
- Cluster Diagram
- Ordinogram
- Isoline
- Isosurface
- Streamline
- Direct Volume Rendering (DVR)

Map

90

It is generally used to represent the location of different areas of a country and is generally drawn on a plain surface. Google maps is generally widely used for data visualization. Now-a-days it is widely used for finding the location in different domains of country.

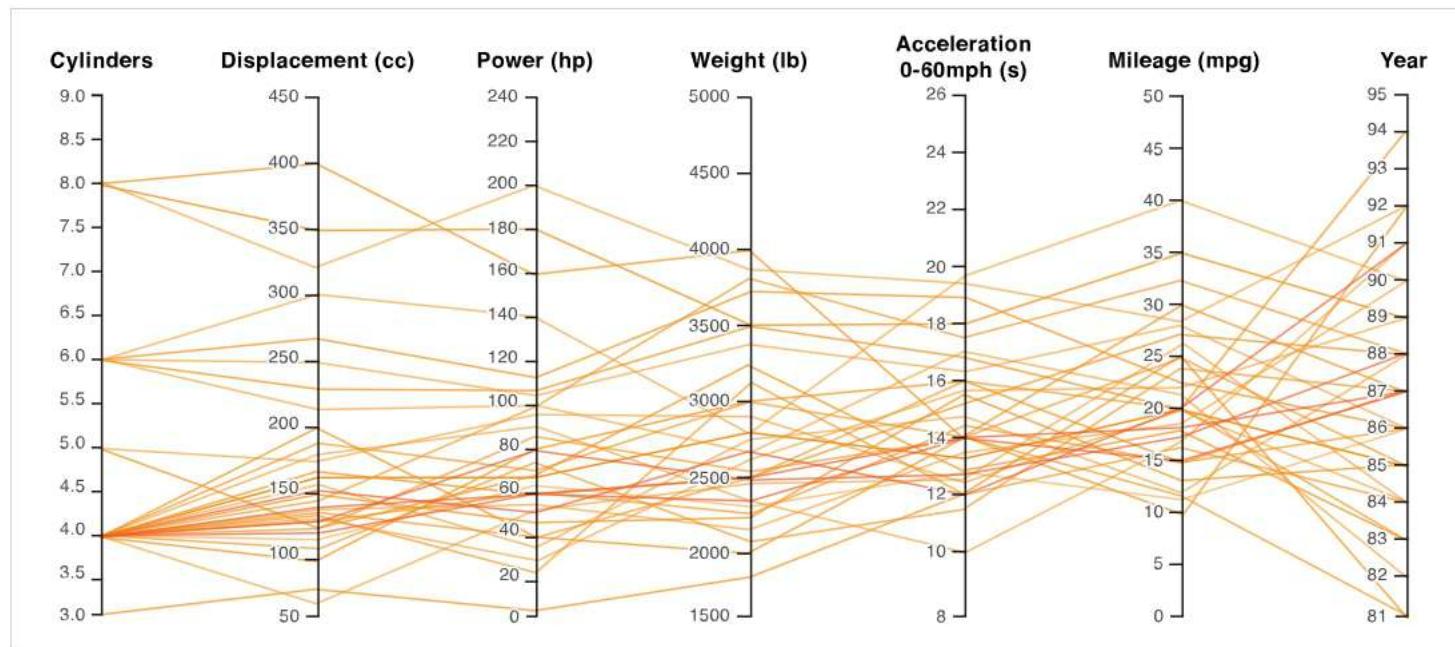


Parallel Coordinate Plot



91

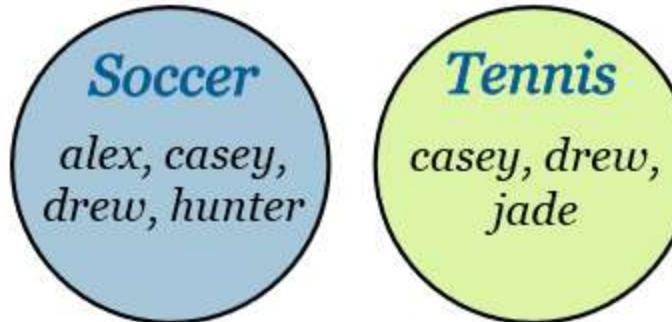
It is a visualization technique of representing multidimensional data. This type of visualisation is used for plotting multivariate, numerical data. Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them.



Venn Diagram

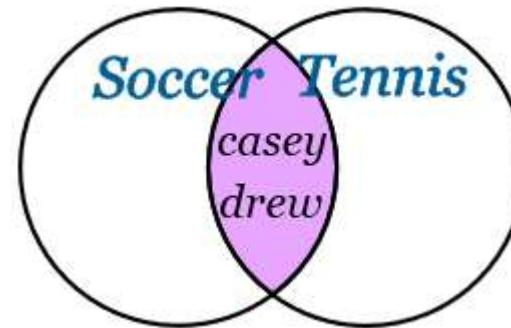
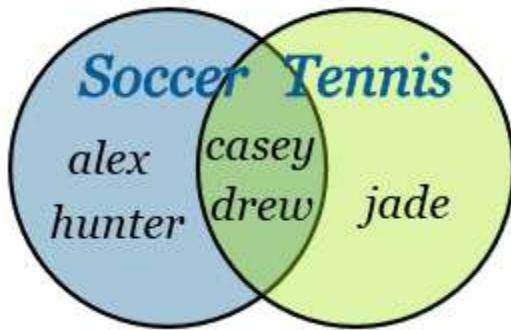
92

It is used to represent the logical relations between finite collection of sets.



List your friends that play Soccer OR Tennis

List your friends that play Soccer AND Tennis



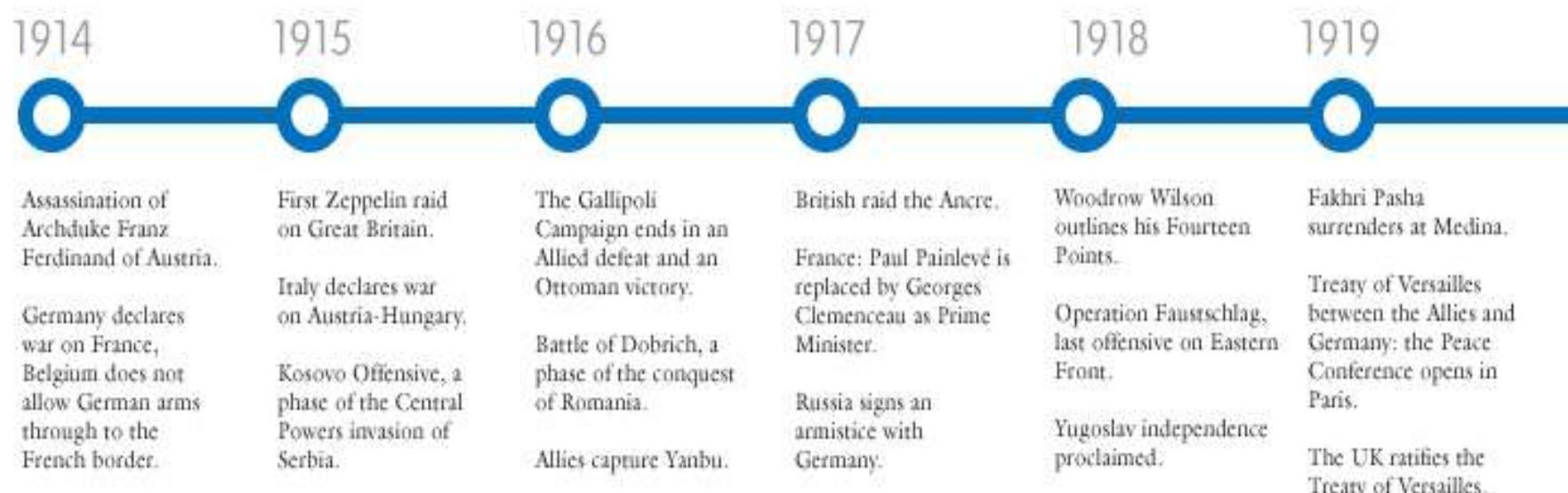
- Draw the Venn Diagram to show people that play Soccer but NOT Tennis
- Draw the Venn Diagram to show people that play Soccer or play Tennis, but not the both.

Timeline

93

It is used to represent a chronological display of events.

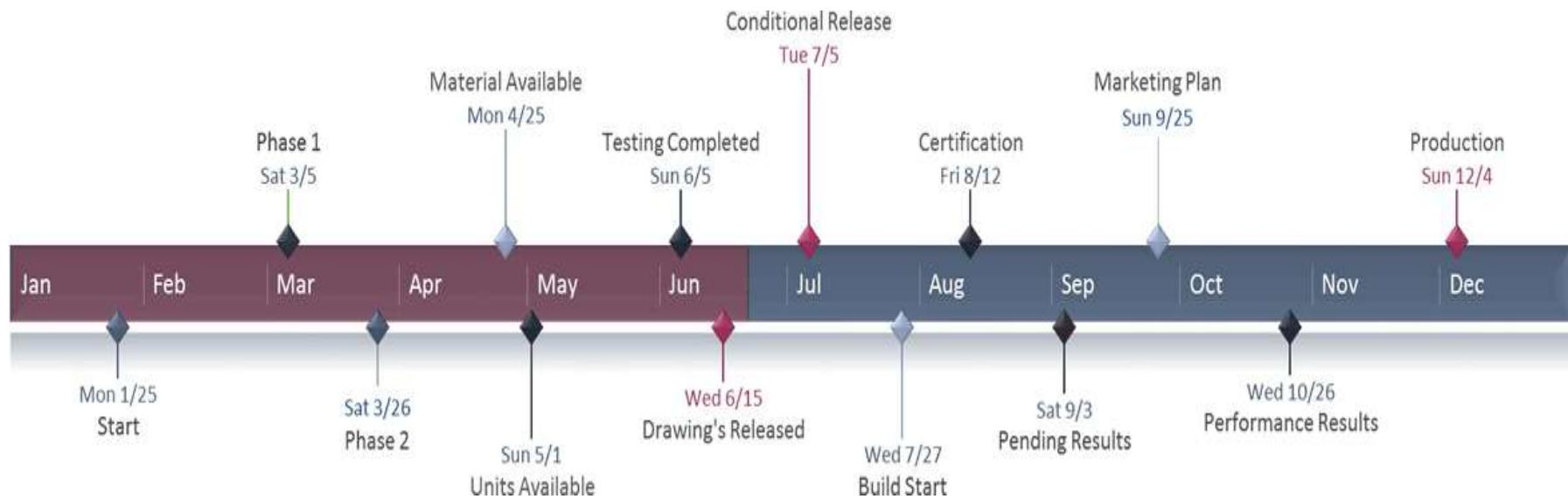
Timeline of World War I



Source: datavizcatalogue.com

Timeline cont'd

94



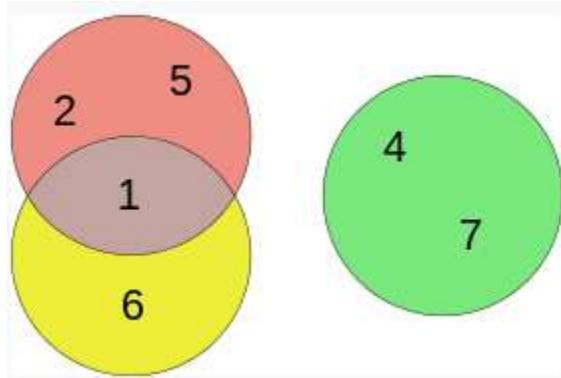
Source: officetimeline.com

Euler Diagram

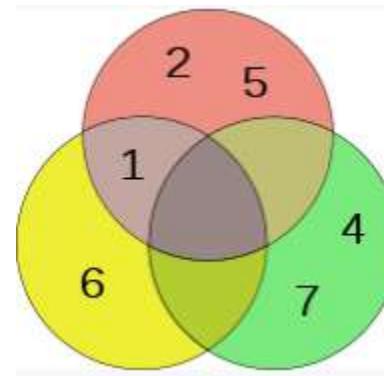
95

It is a representation of the relationships between sets.

Example: Let's take 3 sets namely $A = \{1, 2, 5\}$, $B = \{1, 6\}$ and $C = \{4, 7\}$. The Euler diagram of the sets looks like:



Draw the Equivalent Venn Diagram



Source: wikipedia

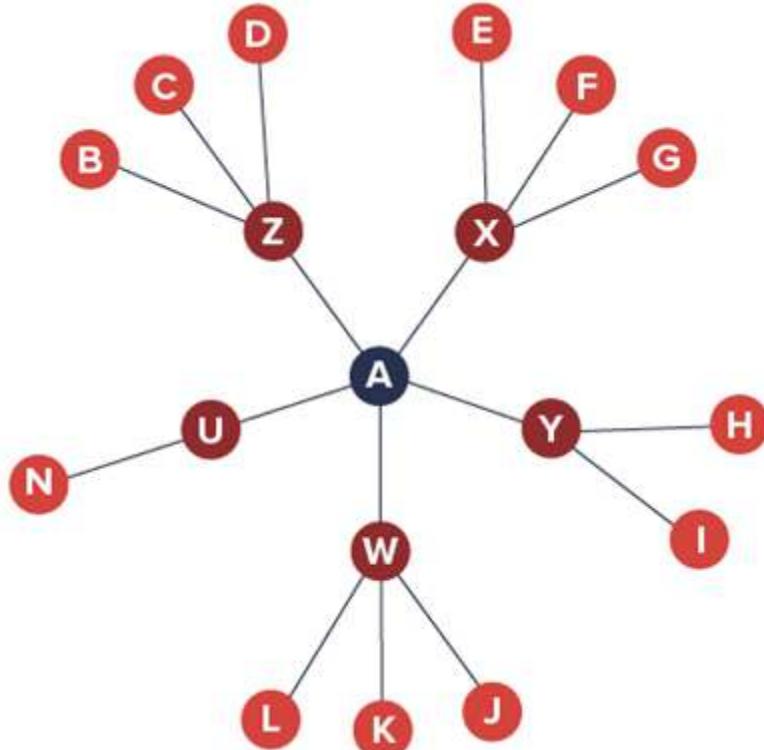
Class Exercise

Draw the Euler diagram of the sets, $X = \{1, 2, 5, 8\}$, $Y = \{1, 6, 9\}$ and $Z = \{4, 7, 8, 9\}$. Then draw the equivalent Venn Diagram.

Hyperbolic Trees

96

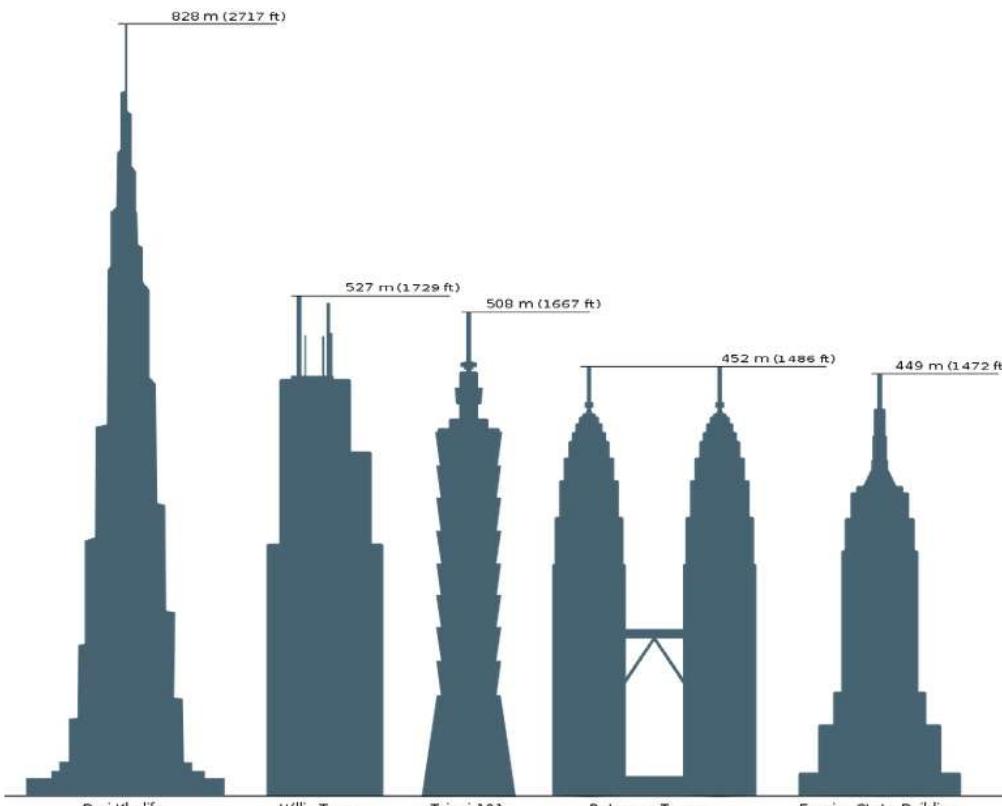
A hyperbolic tree (often shortened as hypertree) is an information visualization and graph drawing method inspired by hyperbolic geometry.



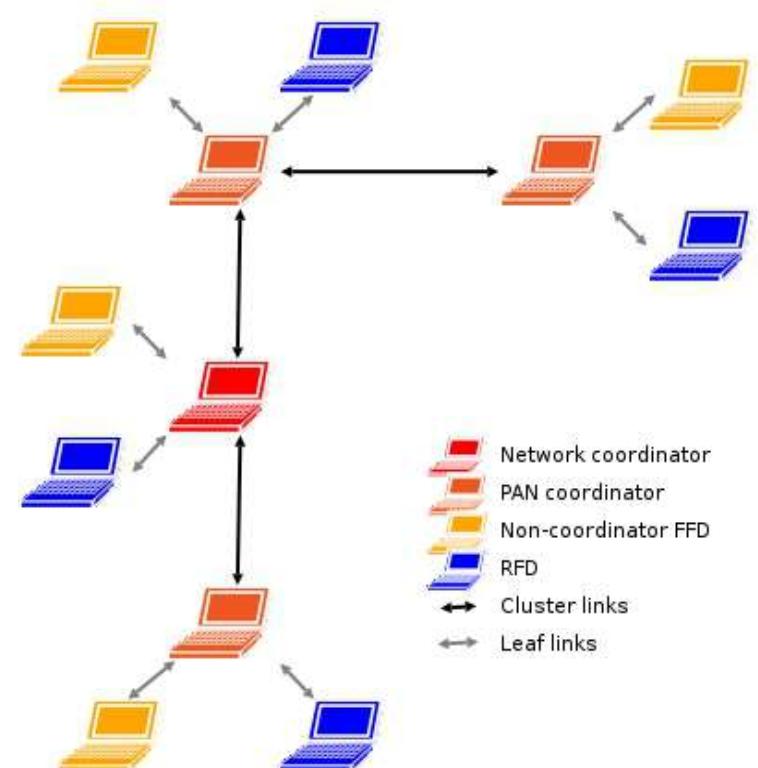
Cluster Diagram

97

A cluster diagram or clustering diagram is a general type of diagram, which represents some kind of cluster. A cluster in general is a group or bunch of several discrete items that are close to each other.



Comparison diagram of sky scraper



Computer network diagram

Ordinogram

98

It is generally used to perform the analysis operation of various sets of **multivariate objects** which are generally used in different domain. Simple two-dimensional graph is an example of ordinogram.

Univariate data – This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Height (cm)	164	167	170	170.4	176.5	180	179.2	165	175
-------------	-----	-----	-----	-------	-------	-----	-------	-----	-----

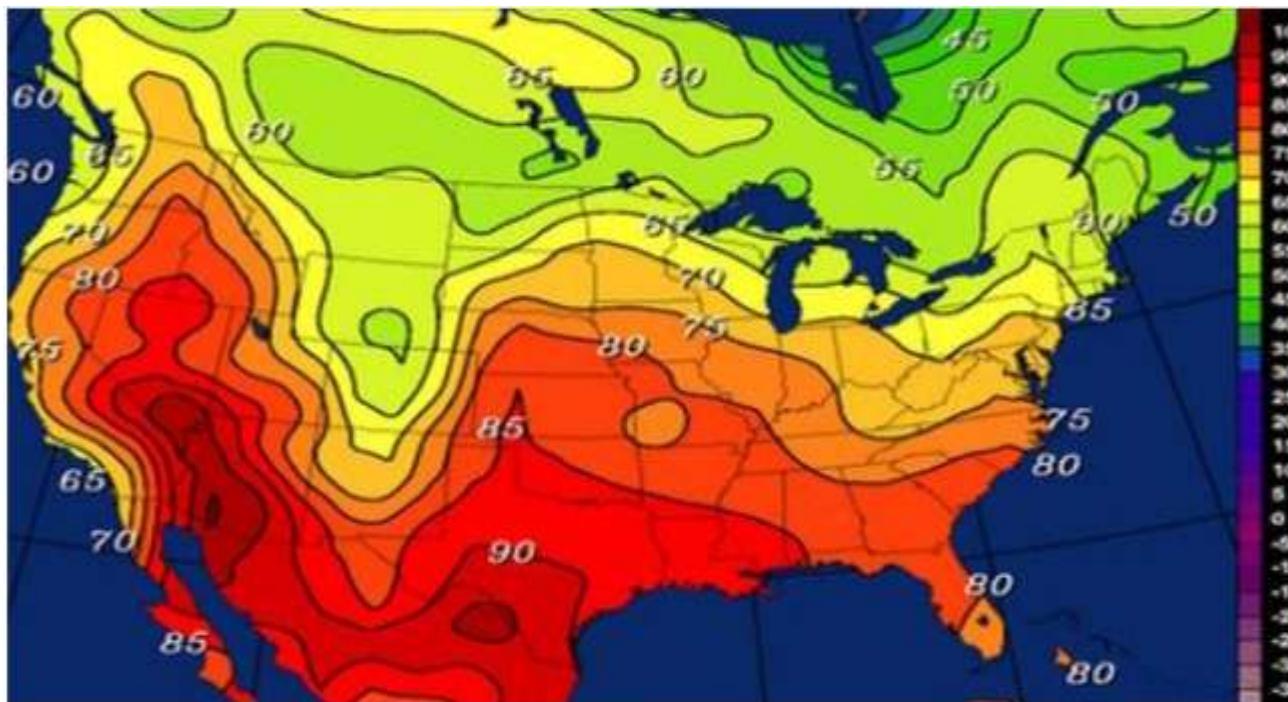
Multivariate data – This type of data involves two or more than two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the variables.

Temperature in Celsius	Ice Cream Sales
20	2000
35	5000

Isoline

99

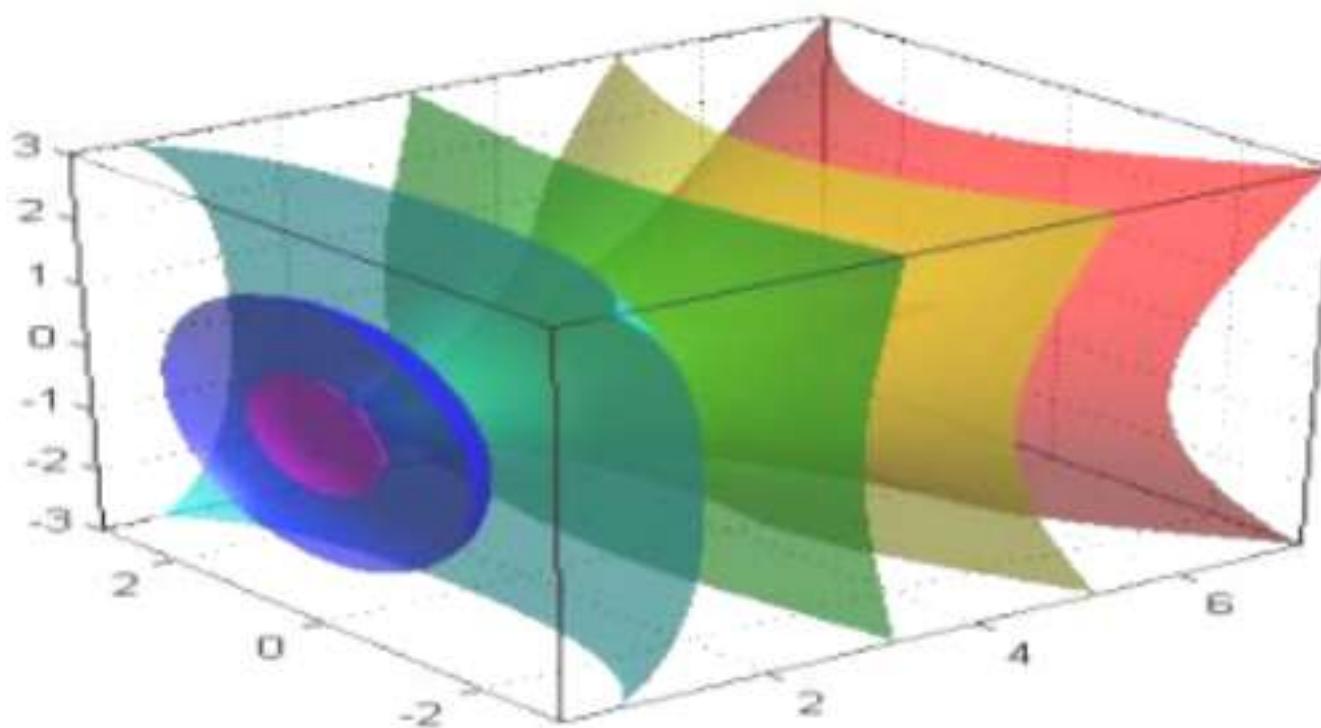
It is basically a 2D data representation of a curved line that generally transfers constantly on the surface of the graph, the plotting of line generally drawn on the basis of data arrangement instead of data visualization.



Isosurface

100

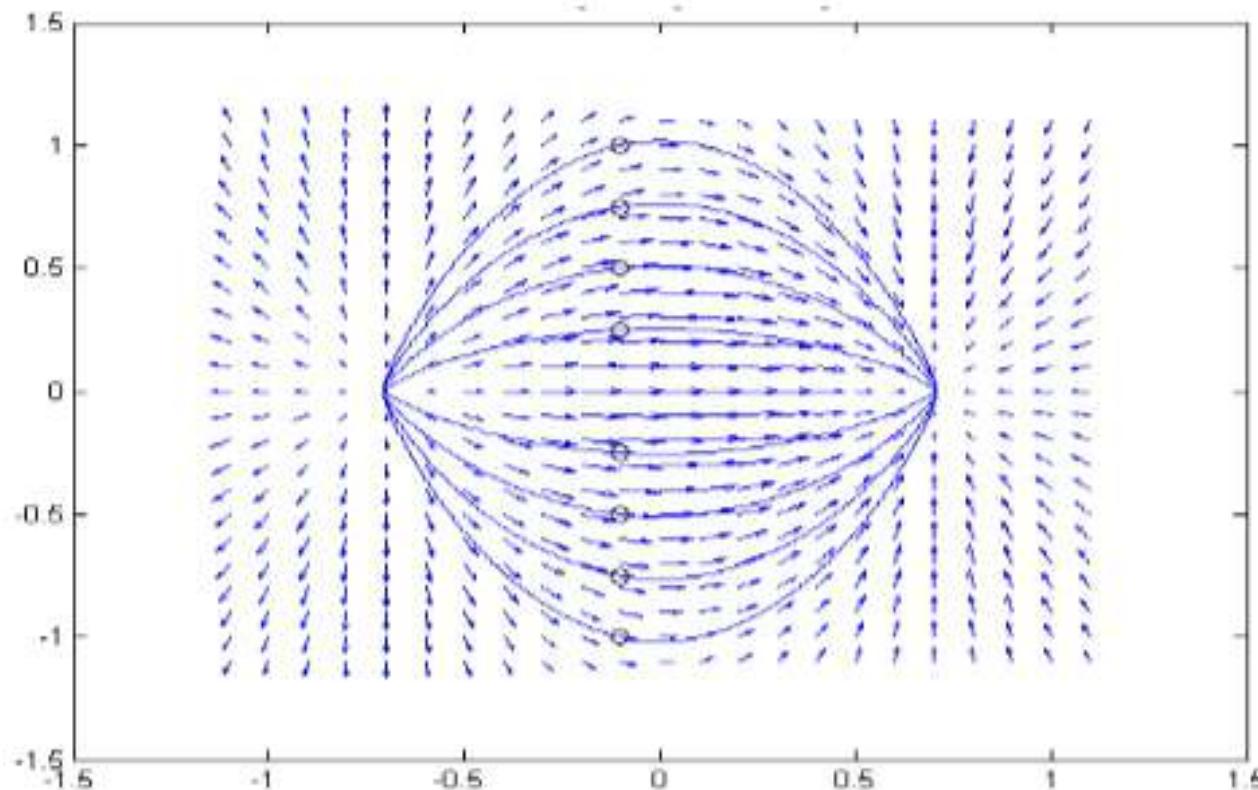
It is a 3D representation of an Isoline. Isosurfaces are designed to present points that are bound by a constant value in a volume of space i.e. in a domain that covers 3D space.



Streamline

101

It is a field that is generated from the description of velocity vector field of the data flow.



Application of Data Visualization

102

There are 3 ways to use data visualization in a company.

- 1. Internal Communication:** Any key data that influences decision-making is prime for data visualization. This is specifically true for the information delivered to higher-ups such as boss or other key stakeholders. Examples are presentation, reports or financial statements.
- 2. Client Reporting:** With data visualization, results reporting to clients or customers is more impactful.
- 3. Marketing Content:** Public-facing content for thought leadership or promotion is more credible with data. Content such as blogs, whitepapers, infographics etc. can be beneficial.

THANK YOU!

Holt's Method

- ⇒ y_t is also called Holt's exponential method.
- ⇒ Assume the time series has trend but no seasonality.
- ⇒ $F_{t+K} = L_t + K T_t$

$K=1$ forecasting is done for 1 step ahead
 step = depends upon your time series. If dataset contains monthly data then 1 step ahead means next month ahead.

- ⇒ Updating the Level

$$L_t = \alpha Y_t + (1-\alpha) (L_{t-1} + T_{t-1})$$

α - smoothing coefficient for level

- ⇒ Updating the Trend

$$T_t = \beta (L_t - L_{t-1}) + (1-\beta) T_{t-1}$$

β - smoothing coefficient for trend

- ⇒ choosing α and β values depend upon minimize either (MSE, RMSE or MAPE)

$$MSE = \left(\sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \right) \quad RMSE = \sqrt{MSE}, \quad MAPE = \frac{\sum_{i=1}^n |\hat{Y}_i - Y_i|}{n} \times 100$$

Y_i = actual value, \hat{Y}_i = Predicted value, n = total sample

Example Using Holt's exponential smoothing solve the forecasting for 1 time stamp ahead. Assume $\alpha = 0.2$, $\beta = 0.2$, $K = 1$ time stamp

<u>Date</u>	<u>y_t Sales</u>	<u>L_t Level</u>	<u>T_t Trend</u>	<u>F_t Forecast</u>	<u>Error ($E = Y - F$)</u>
0 Mar - 2019	155	-	-	-	
1 June - 2019	180	180	25	-	
2 Sept - 2019	260	216	27.2	205	55
3 Dec - 2019	560	306.56	39.87	243.2	316.8
4 Mar - 2020	160	309.15	32.41	346.432	-186.43
5 June - 2020	185	310.25	26.15	341.56	-156.56
6 Sept - 2020	270	323.12	23.496	336.40	-66.40
7 Dec - 2020	600	397.29	33.63	346.61	253.3833

Here, RMSE is considered to set α & β values.

$$\Rightarrow \text{Initially } L_1 = y_1 = 180$$

$$T_1 = y_1 - y_0 = 180 - 155 = 25$$

$$F_{t+K} = L_t + K \cdot T_t \Rightarrow F_{1+1} = L_1 + 1 \cdot T_1$$

$$t=1 \quad \Rightarrow F_2 = 180 + 25 = 205$$

Update Level and Trend

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$\Rightarrow L_2 = 0.2 * 260 + (1-0.2)(180 + 25) = 216$$

$$T_t = \beta (L_t - L_{t-1}) + (1-\beta) T_{t-1}$$

$$\Rightarrow T_2 = 0.2 (L_2 - L_1) + (1-0.2) T_1$$

$$= 0.2 (216 - 180) + (1-0.2) 25$$

$$= 27.2$$

$$E_t = Y_t - F_t$$

$$\Rightarrow E_2 = Y_2 - F_2 = 260 - 205 = 55$$

$$MSE = \left(\sum_{i=2}^{n=6} (Y_i - F_i)^2 \right) / n$$

$$RMSE = \sqrt{MSE}$$

\Rightarrow By choosing different values of α & β in next epoch, RMSE value will be compared. Least RMSE value will be considered and respective α & β will be set.

Introduction to Hadoop

Topics

The goal of this presentation is to give you a basic understanding of Hadoop's core components (HDFS and MapReduce)

In this presentation, I will try and answer the following few questions:

- What is Hadoop and what are its significant features
- How it offers reliable storage for massive amounts of data with HDFS
- How it supports large scale data processing through MapReduce
- How 'Hadoop Ecosystem' tools can boost an analyst's productivity

The Traditional Workaround

How has industry typically dealt with storage problem?

- Perform an ETL (Extract, Transform, Load) on the data to summarize and denormalize the data, before archiving the result in a data warehouse
 - Discarding the details
 - Run queries against the summary data

Unfortunately, this process results in loss of detail

- But there could be real nuggets in the lost detail
- Value lost with data summarization or deletion
 - Think of the opportunities lost when we summarize 10000 rows of data into a single record
 - We may not learn much from a single tweet, but we can learn a lot from 10000 tweets
- More data == Deeper understanding
 - “There’s no data like more data” (Moore 2001)
 - “It’s not who has the best algorithms that wins. It’s who has the most data”

Early Large Scale Computing

- Historically computation was processor-bound
 - Data volume has been relatively small
 - Complicated computations are performed on that data
- Advances in computer technology has historically centered around improving the power of a single machine

Distributed Systems

- Allows developers to use multiple machines for a single task
- Programming on a distributed system is much more complex
- “You know you have a distributed system when the crash of a computer you’ve never heard of stops you from getting any work done.” –*Leslie Lamport*
- Distributed systems must be designed with the expectation of failure

Distributed System: Data Storage

- Typically divided into Data Nodes and Compute Nodes
- At compute time, data is copied to the Compute Nodes
- Fine for relatively small amounts of data
- Modern systems deal with far more data than was gathering in the past.
 - Facebook - 500 TB per day
 - Yahoo - Over 170 PB
 - eBay - Over 6 PB
- Getting the data to the processors becomes the bottleneck.

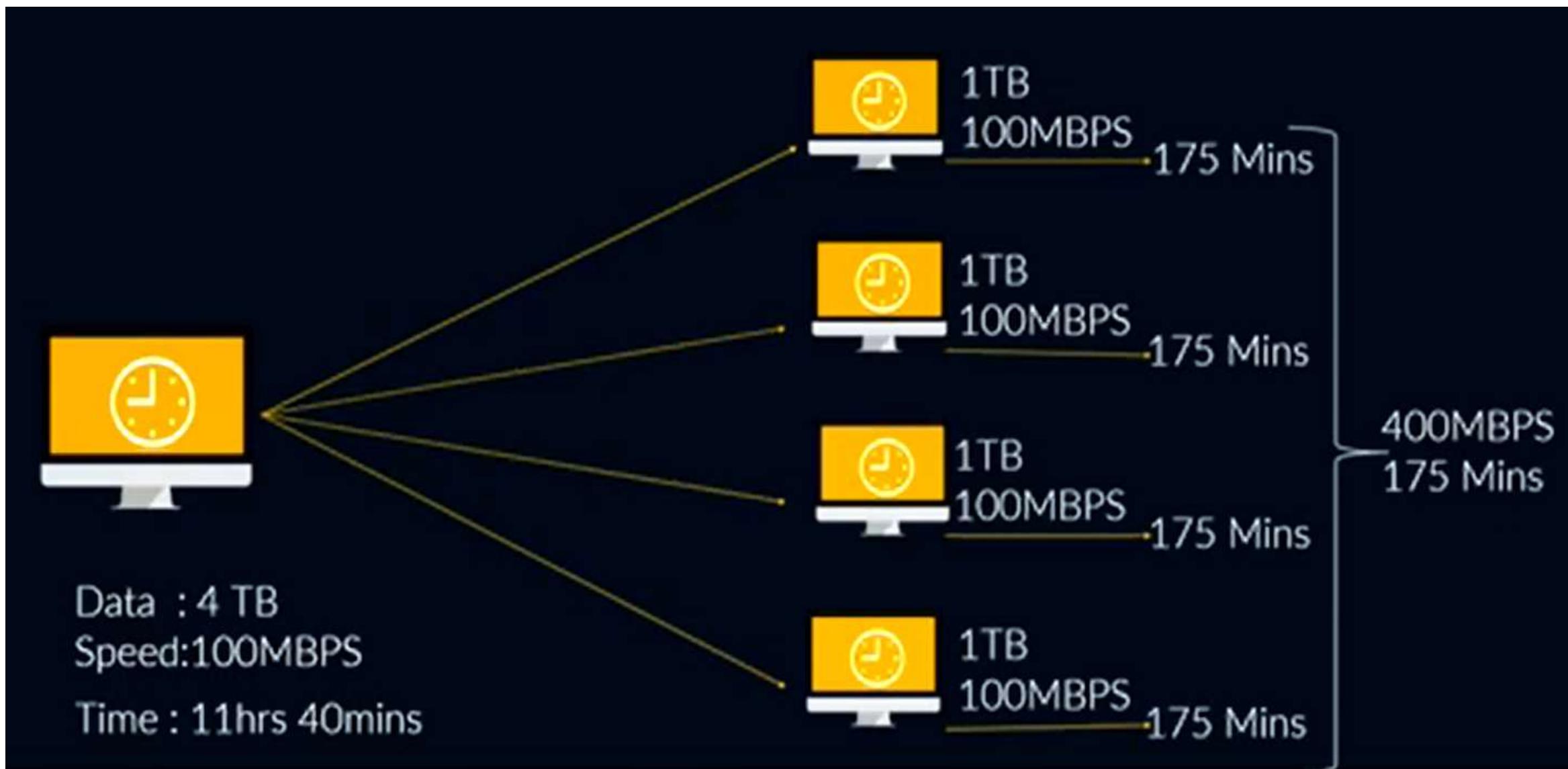
Reading Data with Single Machine



DATA : 4 TB
SPEED: 100MBPS

Calculation: $\frac{4 * 1024 * 1024}{100 * 60}$
=700 Mins(approx.)

Parallel Processing



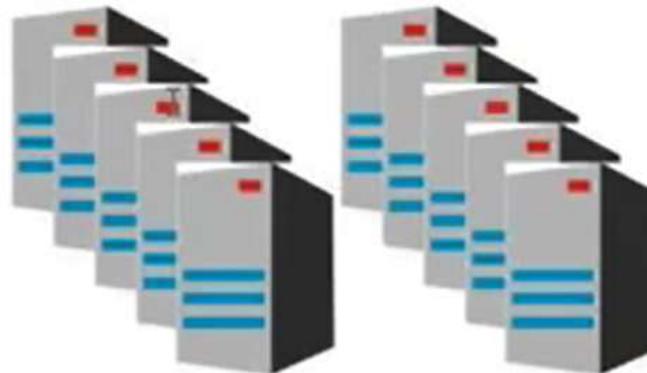
Why Distributed File Systems?

Read 1 TB Data



1 Machine

- 4 I/O Channels
- Each Channel – 100 MB/s



10 Machines

- 4 I/O Channels
- Each Channel – 100 MB/s



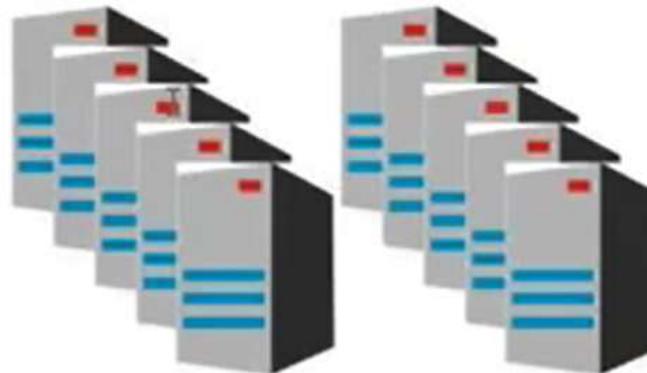
Why Distributed File Systems?

Read 1 TB Data



1 Machine

- 4 I/O Channels
- Each Channel – 100 MB/s



10 Machines

- 4 I/O Channels
- Each Channel – 100 MB/s

43-45 minutes

4.3-4.5 minutes

What is Distributed File Systems?

Before DFS consolidation



\Chicago_server\Homedirs



\Chicago_maxi\Projects



\Houston_server\Reports



\Denver_server\Software

After DFS consolidation



\maxi-pedia.com\Public\Homedirs
\maxi-pedia.com\Public\Projects
\maxi-pedia.com\Public\Reports
\maxi-pedia.com\Public\Software

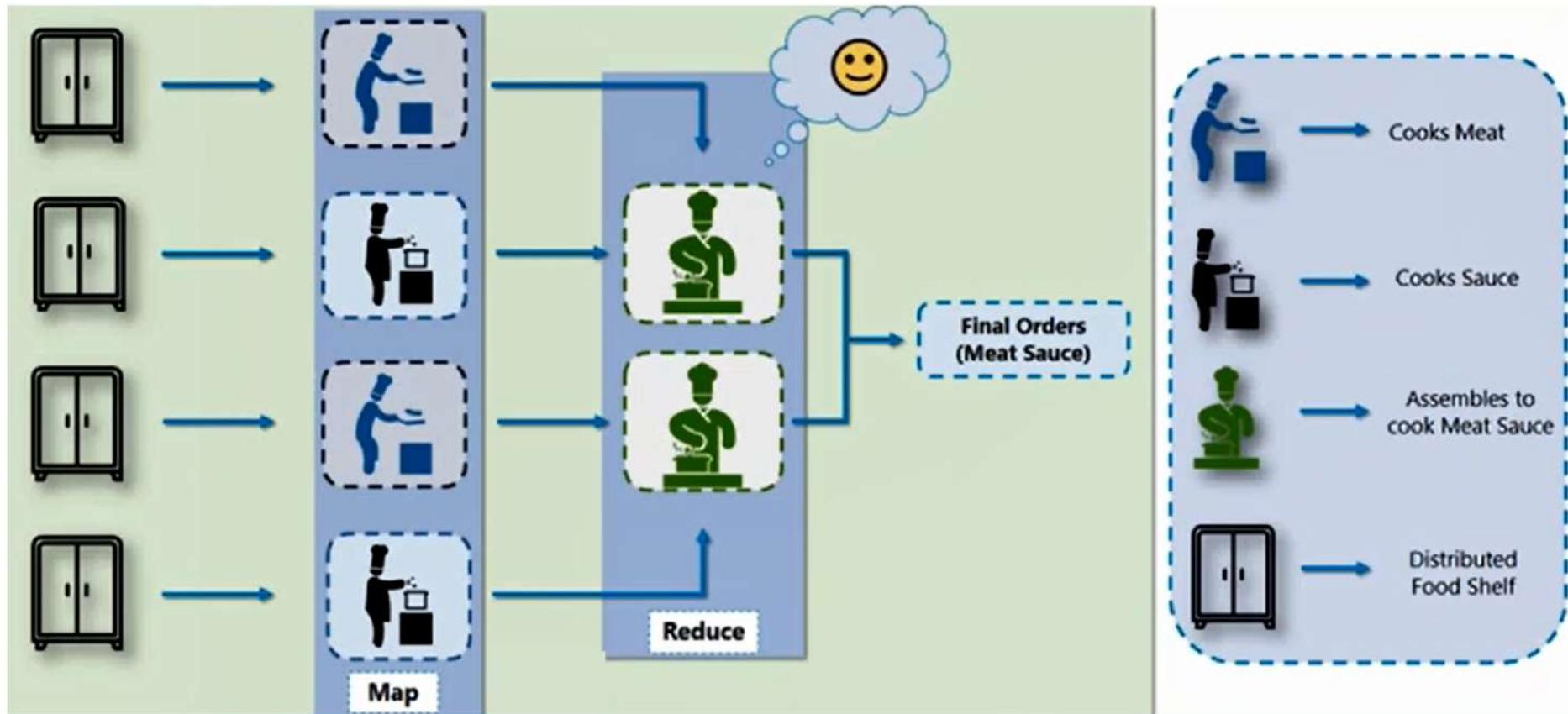
Distributed and Parallel Approach

Case study : Restrurant Service

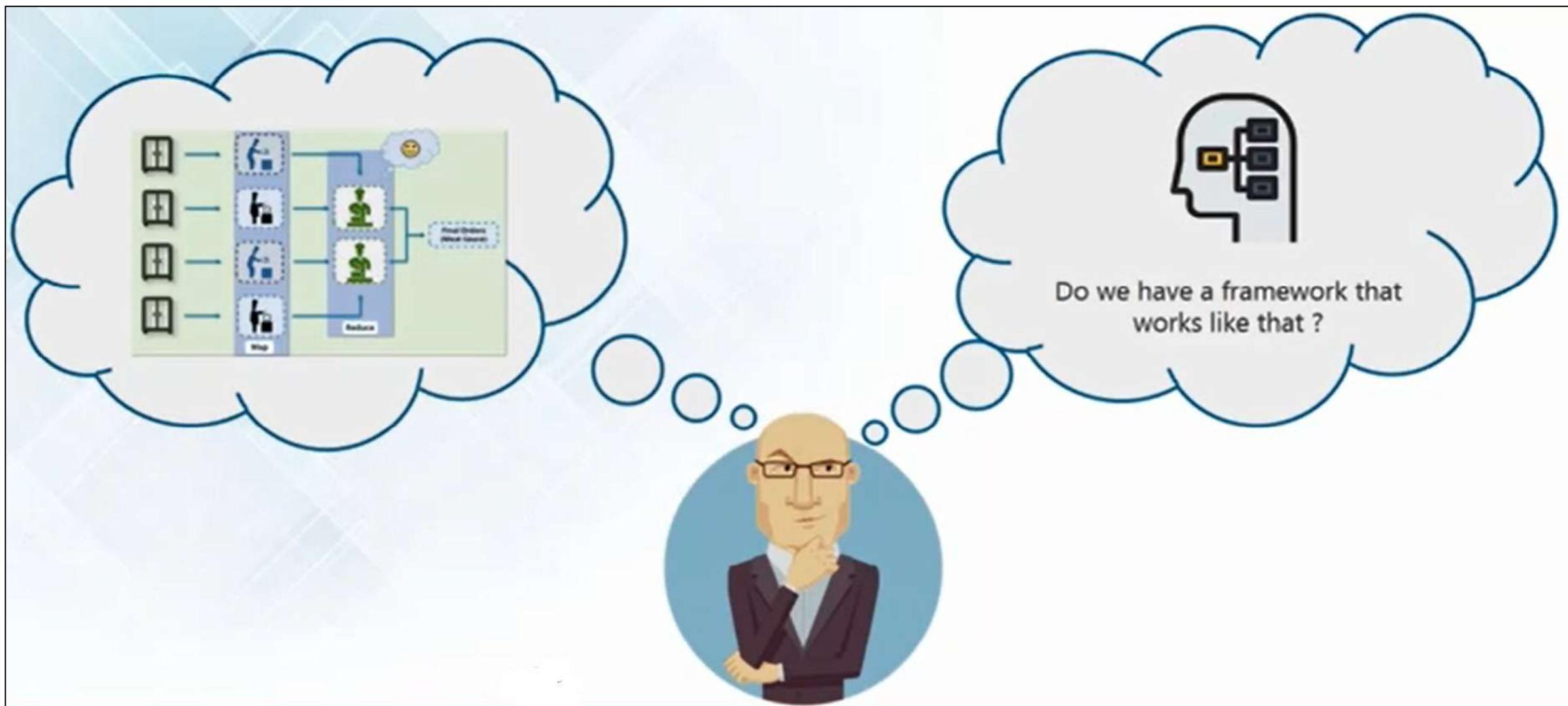
Issue 1: Too many orders per hour - Hiring multiple cook

Issue 2: Foods shelf becomes the bottleneck - Distributed & Parallel Approach

Effective Solution is,

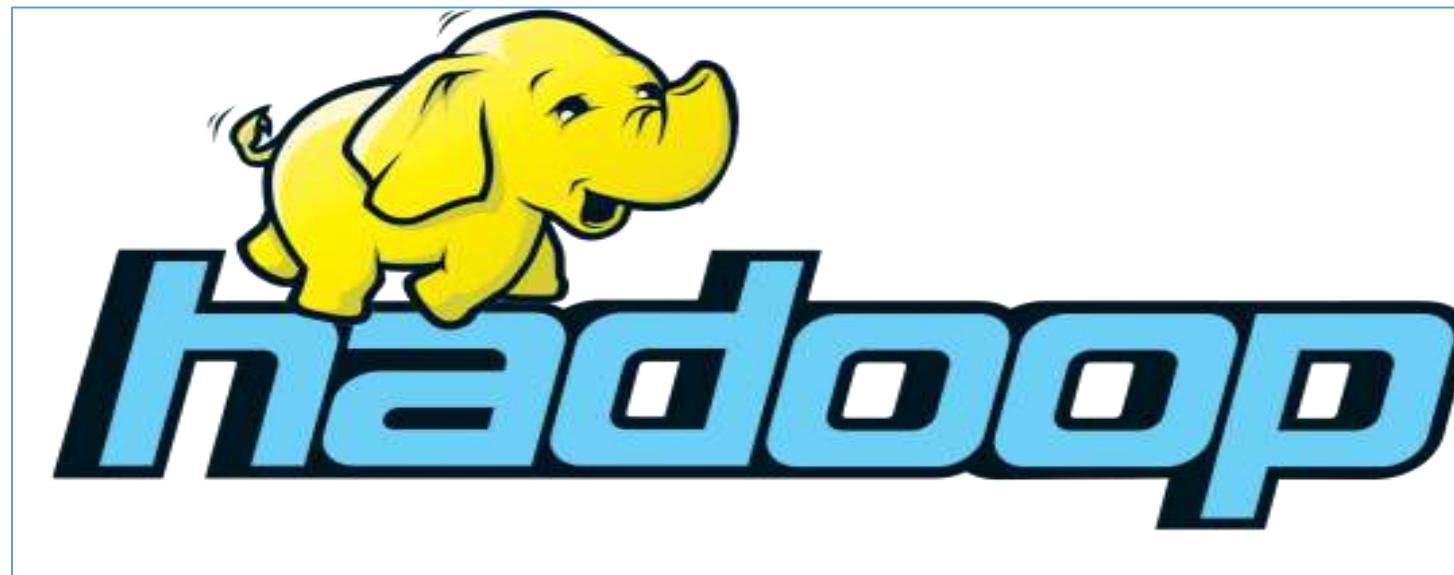


Need of a Framework...



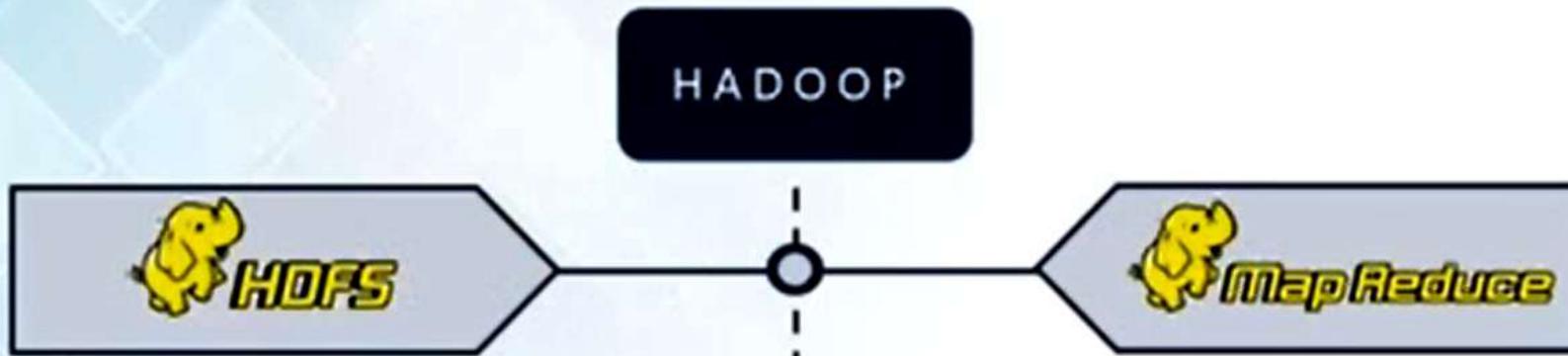
We Need A System That Scales

- We're generating too much data to process with traditional tools
- Two key problems to address
 - How can we reliably store large amounts of data at a reasonable cost?
 - How can we analyze all the data we have stored?

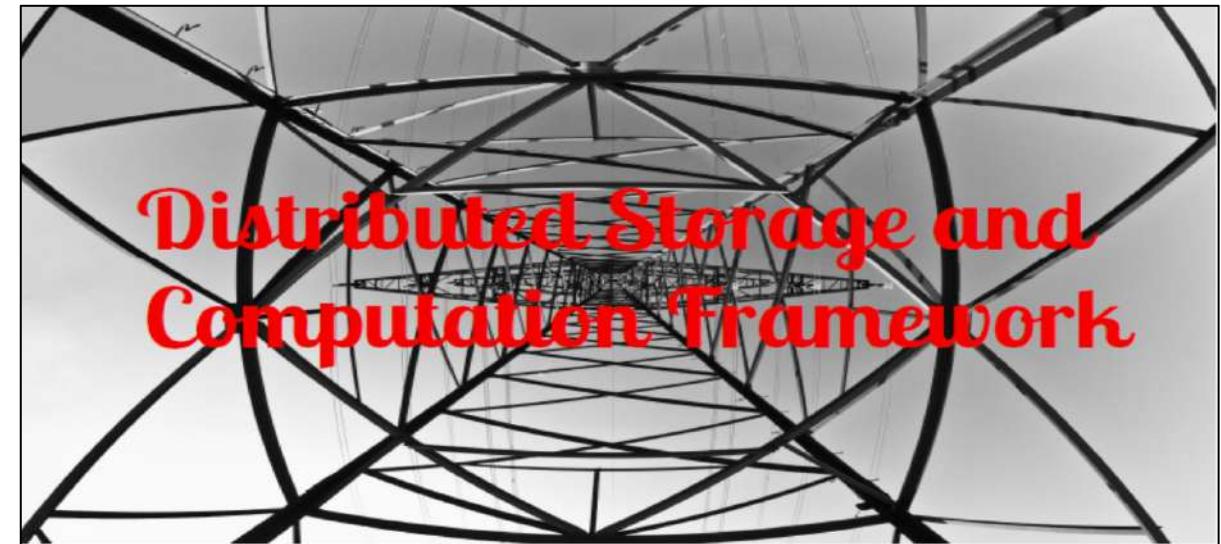
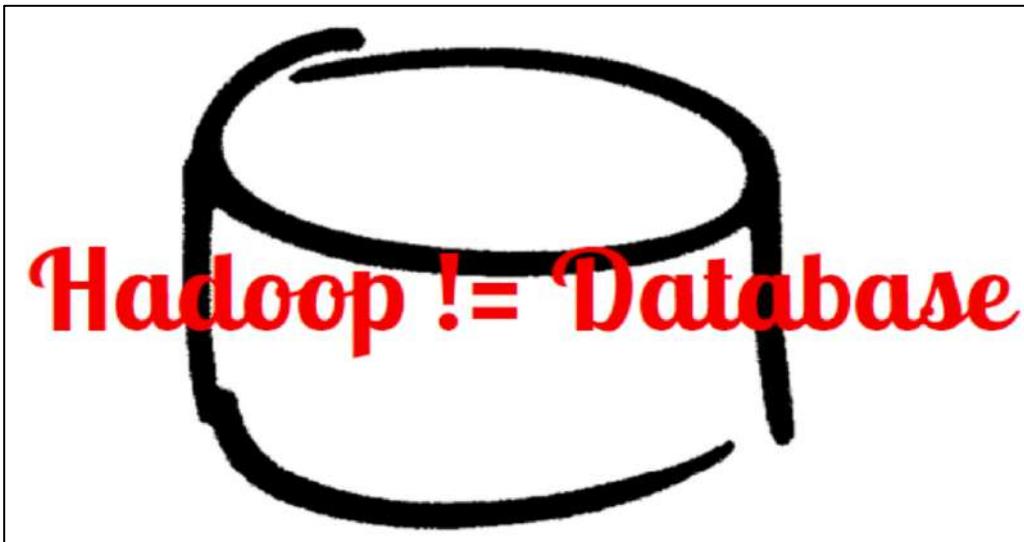


Apache Hadoop: Framework to Process Big Data

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



What is Hadoop?



What is Hadoop?

- Hadoop is a free/open source software package that provides tools that enable data discovery, question answering, and rich analytics based on very large volumes of information.
- It is a framework designed for storage and processing of large scale data on clusters of commodity hardware
- All this in fractions of the time required by traditional databases.

Requirements for Hadoop

- Must support **partial failure**
- Must be scalable

Partial failure

- Failure of a single component must not cause the failure of the entire system only a degradation of the application performance
- Failure should not result in the loss of any data



Component Recovery

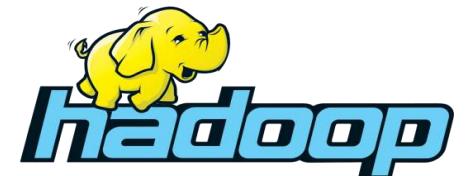
- If a component fails, it should be able to recover without restarting the entire system
- Component failure or recovery during a job must not affect the final output

Scalability

- Increasing resources should increase load capacity
- Increasing the load on the system should result in a graceful decline in performance for all jobs
 - Not system failure

Where Did Hadoop Come From?

- Based on work done by Google in the early 2000s
- Google's objective was to index the entire World Wide Web
- Google had reached the limits of scalability of RDBMS technology
 - “The Google File System” in 2003
 - “MapReduce: Simplified Data Processing on Large Clusters” in 2004
- A developer by the name of Doug Cutting (at Yahoo!) was wrestling with many of the same problems in the implementation of his own open-source search engine,
- He started an open-source project based on Google’s research and created Hadoop in 2005.
- ***Hadoop was named after his son’s toy elephant.***
- The core idea was to distribute the data as it is initially stored
 - Each node can then perform computation on the data it stores without moving the data for the initial processing



Uses for Hadoop

- Data-intensive text processing
- Assembly of large genomes
- Graph mining
- Machine learning and data mining
- Large scale social network analysis
- many ore....

The Hadoop Ecosystem

Hadoop Common

- Contains Libraries and other modules

HDFS

- Hadoop Distributed File System

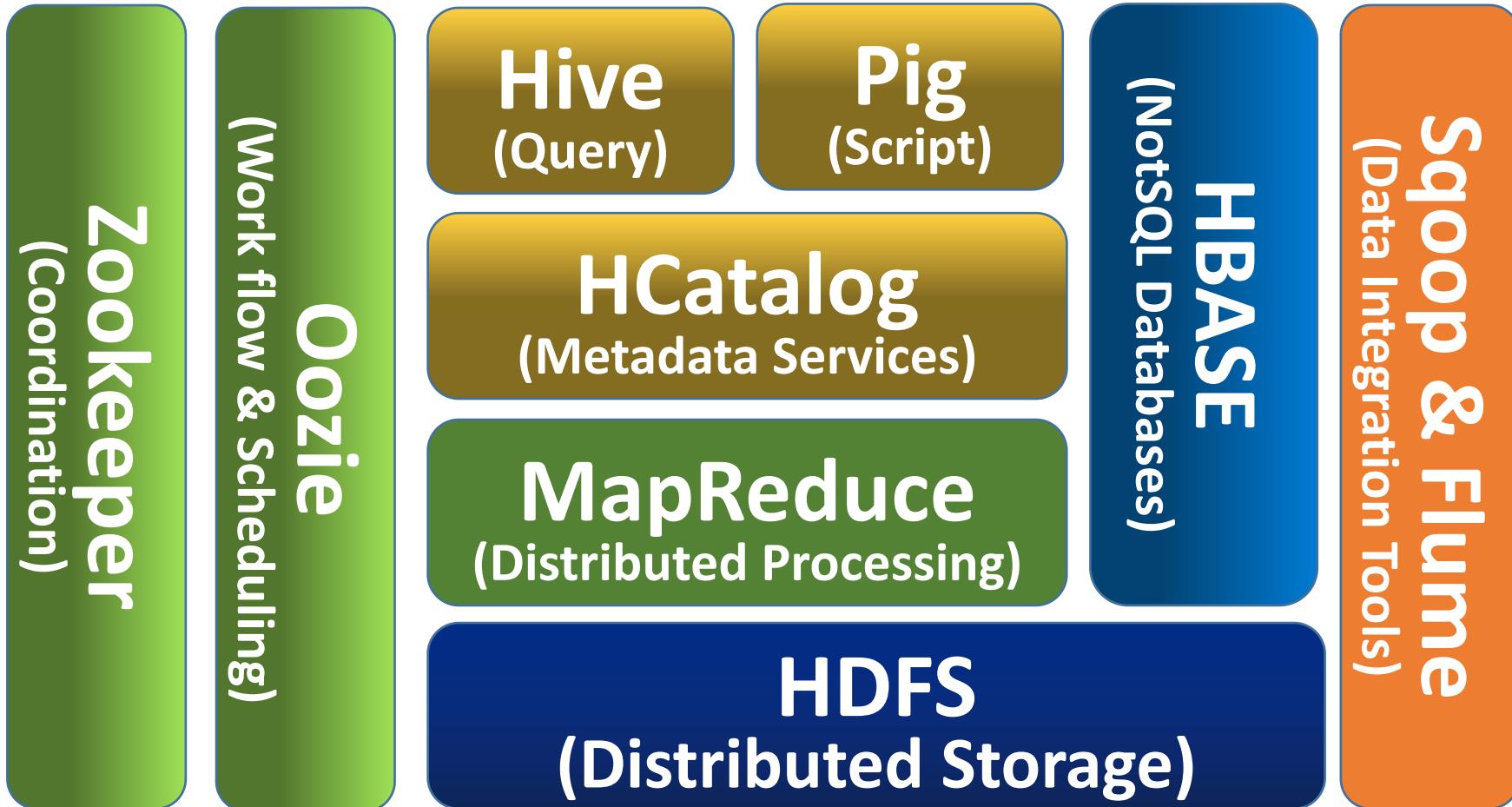
Hadoop YARN

- Yet Another Resource Negotiator

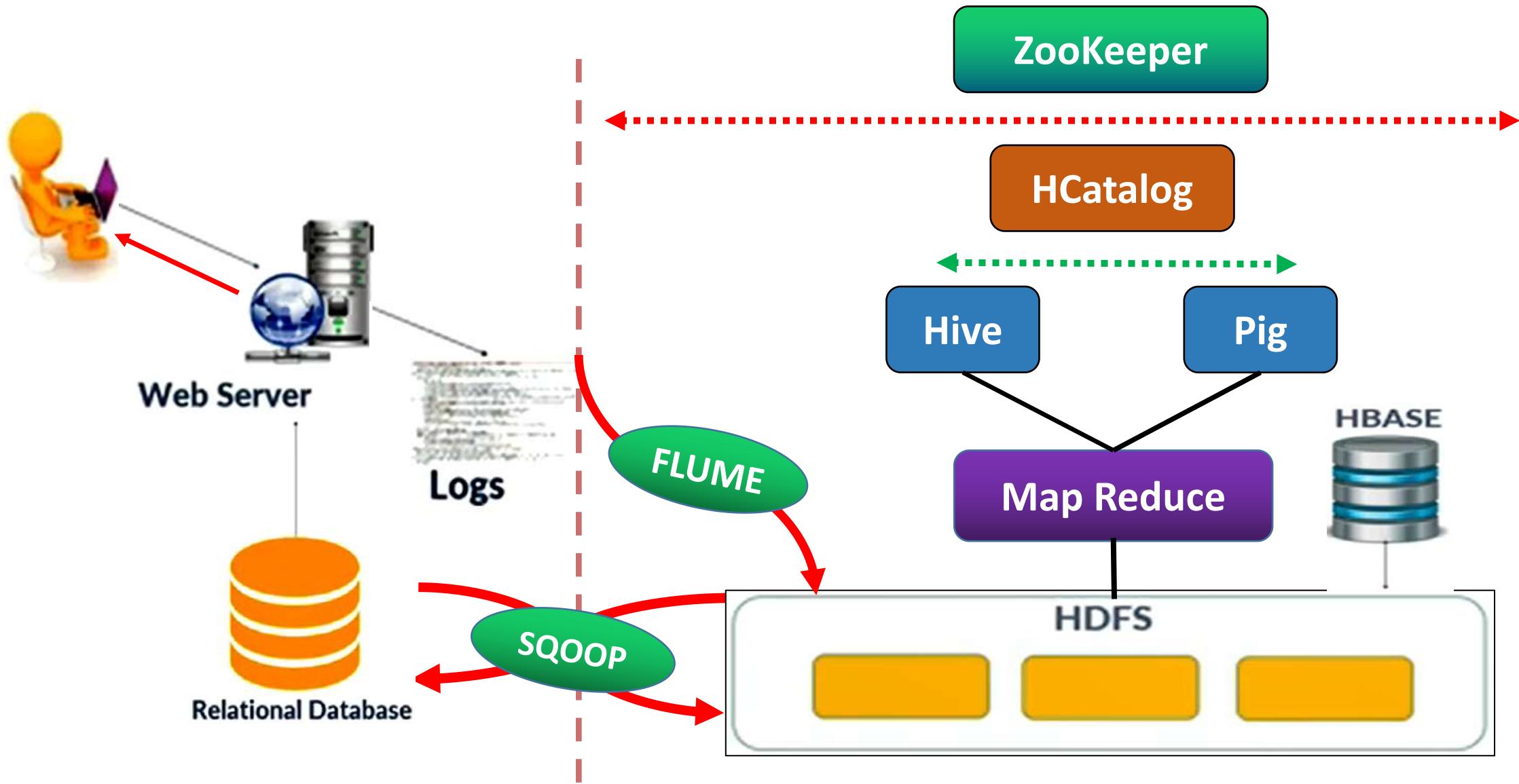
Hadoop MapReduce

- A programming model for large scale data processing

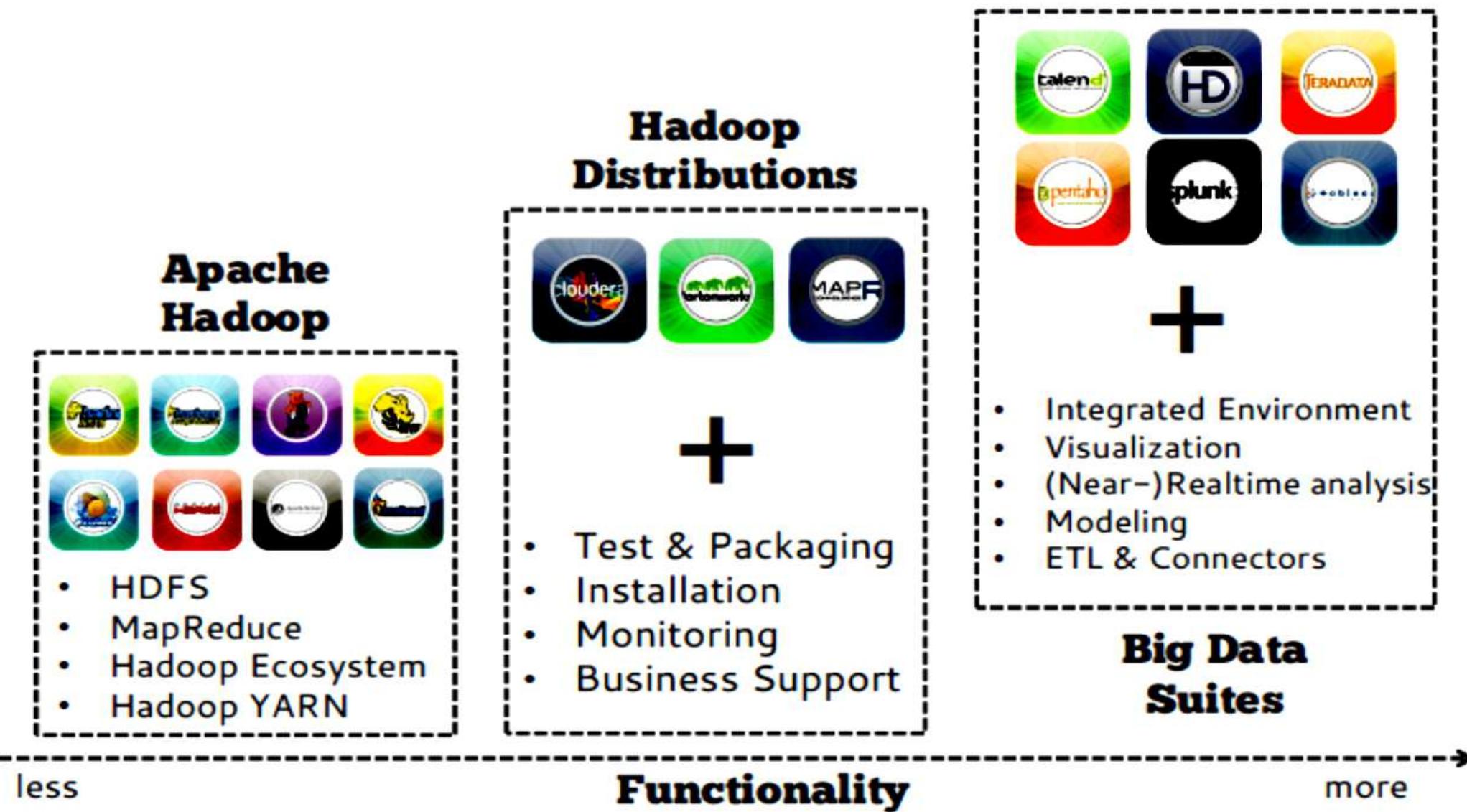
The Hadoop Ecosystem



Hadoop Use case



The Hadoop App Store

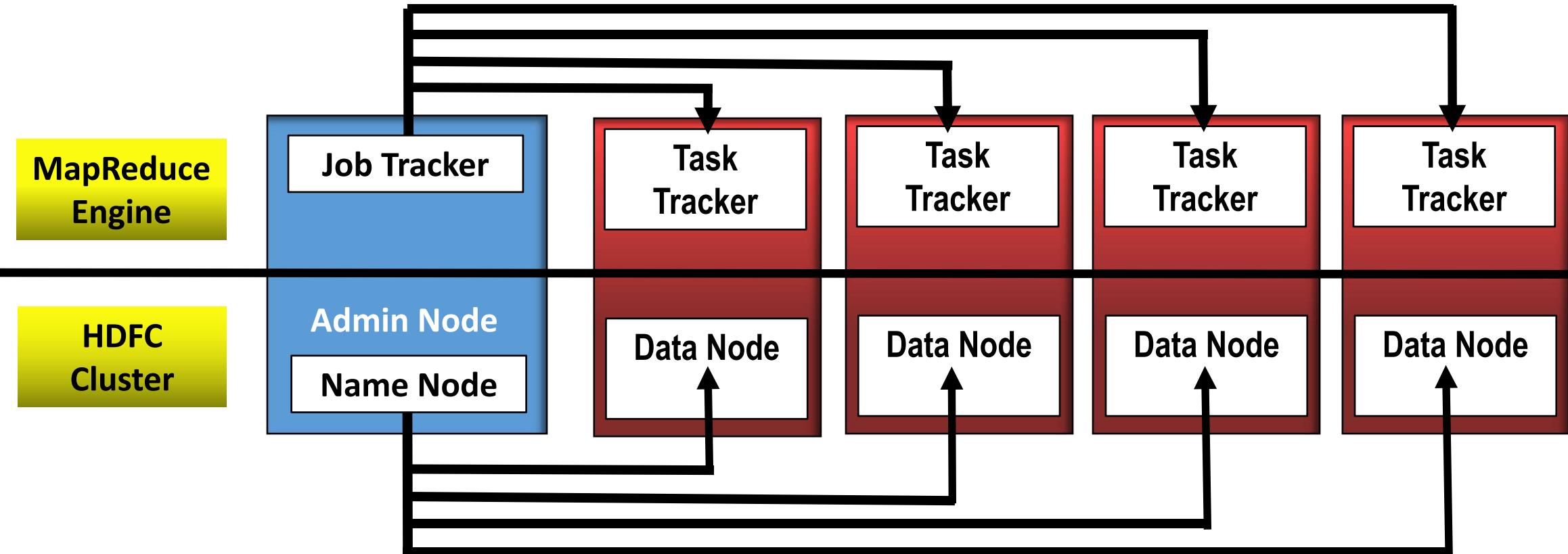


Core Hadoop Concepts

- Applications are written in a high-level programming language
 - No network programming or temporal dependency.
- Nodes should communicate as little as possible
 - A “shared nothing” architecture.
- Data is spread among the machines in advance
 - Perform computation where the data is already stored as often as possible.

Hadoop Core Components

- HDFC - Hadoop Distributed File System (storage)
- MapReduce (processing)



Scalability

Hadoop is a distributed system

- A collection of servers running Hadoop software is called a **cluster**

Individual servers within a cluster are called **nodes**

- Typically standard rack-mount servers running Linux
- Each node both stores and processes data

Add more nodes to the cluster to increase scalability

- A cluster may contain up to several thousand nodes
 - Facebook and Yahoo are each running clusters in excess of 4400 nodes
- Scalability is linear
- And its accomplished on standard hardware, greatly reducing the storage costs

High-Level Overview

- When data is loaded onto the system it is divided into blocks
 - Typically 64MB or 128MB
- Tasks are divided into two phases
 - Map tasks which are done on small portions of data where the data is stored
 - Reduce tasks which combine data to produce the final output
- A master program allocates work to individual nodes

Fault Tolerance

- Failures are detected by the master program which reassigns the work to a different node
- Restarting a task does not affect the nodes working on other portions of the data
- If a failed node restarts, it is added back to the system and assigned new tasks
- The master can redundantly execute the same task to avoid slow running nodes

Reliability / Availability

Reliability and availability are traditionally expensive to provide

- Requiring expensive hardware, redundancy, auxiliary resources, and training

With Hadoop, providing reliability and availability is much less expensive

- If a node fails, just replace it
 - At Google, servers are not bolted into racks, they are attached to the racks using velcro, in anticipation of the need to replace them quickly
- If a task fails, it is automatically run somewhere else
 - No data is lost
 - No human intervention is required

In other words, reliability / availability are provided by the framework itself

How are reliability, availability, and fault tolerance provided?

A file in Hadoop is broken down into **blocks** which are:

- typically 64MB each (a typical windows block size is 4KB)
- Distributed across the cluster
- Replicated across multiple nodes of the cluster (Default: 3)

Consequences of this architecture:

- If a machine fails (even an entire rack) no data is lost
- Tasks can run elsewhere, where the data resides
- If a task fails, it can be dispatched to one of the nodes containing redundant copies of the same data block
- When a failed node comes back online, it can automatically be reintegrated with the cluster

HDFS: Hadoop Distributed File System

HDFS: Hadoop Distributed File System

- Based on Google's GFS (Google File System)
- Provides inexpensive and reliable storage for massive amounts of data
 - Optimized for a relatively small number of large files
 - Each file likely to exceed 100 MB, multi-gigabyte files are common
 - Store file in hierarchical directory structure
 - e.g. , /sales/reports/asia.txt
- Cannot modify files once written
 - Need to make changes? remove and recreate
- Data is distributed across all nodes at load time
 - Provides for efficient Map Reduce processing
- Use Hadoop specific utilities to access HDFS

HDFS Design

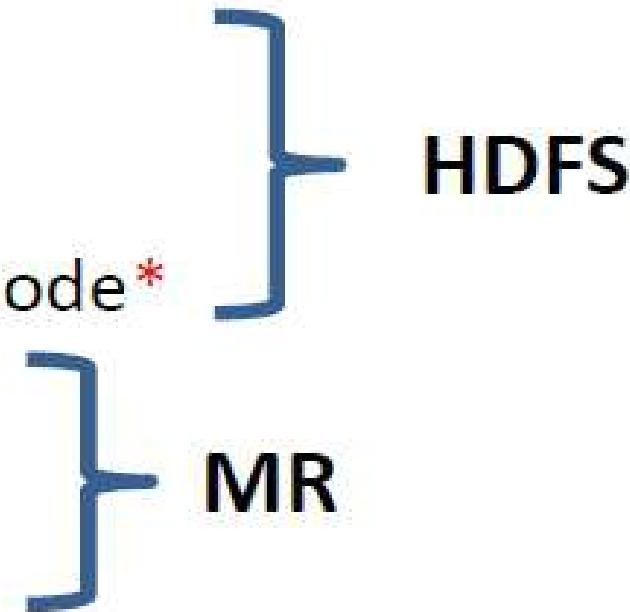
- Runs on commodity hardware
 - Assumes high failure rates of the components
- Works well with lots of large files
 - Hundred of Gigabytes or Terabytes in size
- Built around the idea of “write-once, read many-times”
- Large streaming reads
 - Not random access
- Responsible for storing data on the cluster
- Data files are split into blocks and distributed across the nodes in the cluster (Each block is replicated multiple times)
- High throughput is more important than low latency

HDFS and Unix File System

- In some ways, HDFS is similar to a UNIX filesystem
 - Hierarchical, with UNIX/style paths (e.g. /sales/reports/asia.txt)
 - UNIX/style file ownership and permissions
- There are also some major deviations from UNIX
 - No concept of a current directory
 - Cannot modify files once written
 - You can delete them and recreate them, but you can't modify them
 - Must use Hadoop specific utilities or custom code to access HDFS

Daemons in Hadoop Core

- NameNode
- DataNode
- Secondary NameNode*
- JobTracker*
- TaskTracker*



- Daemon Process
 - process which runs in background and has no controlling terminal.

HDFS Architecture

Hadoop has a **master/slave** architecture

HDFS master daemon: **Name Node**

- It stores metadata and manages access
- Manages namespace (file to block mappings) and metadata (block to machine mappings)
- Monitors slave nodes

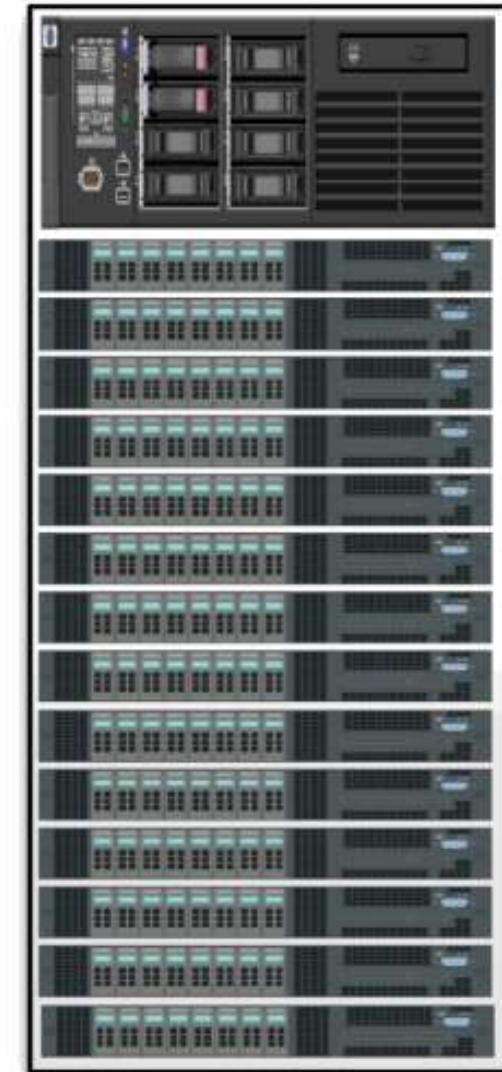
HDFS slave daemon: **Data Node**

- Reads and writes the actual data

Provides reliability through replication

- Each Block is replicated across several Data Nodes

A Small Hadoop Cluster



How are Files Stored

- Generally the user data is stored in the files of HDFS.
- Files are split into blocks
- Blocks are split across many machines at load time
 - Different blocks from the same file will be stored on different machines
- Blocks are replicated across multiple machines
- The NameNode keeps track of which blocks make up a file and where they are stored
- In other words, the minimum amount of data that HDFS can read or write is called a Block.
- The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

HDFS Architecture

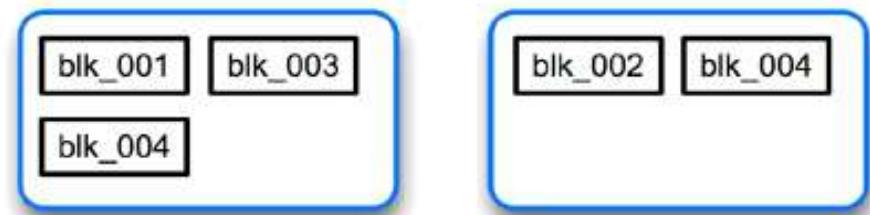
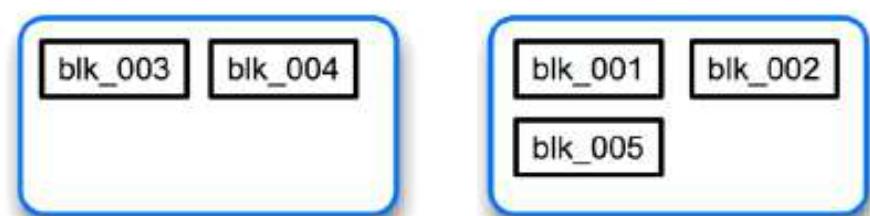
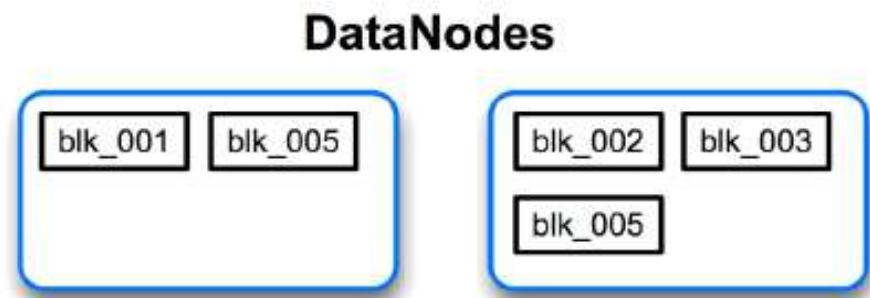
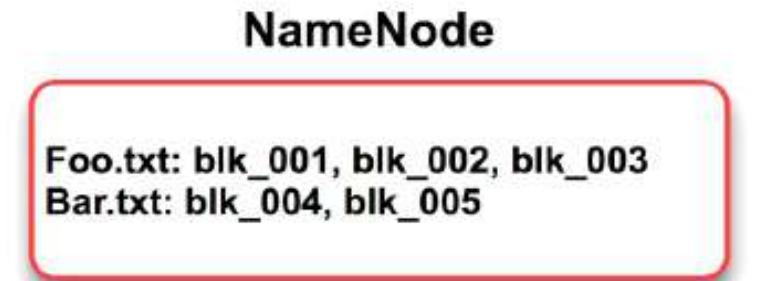
Example:

The NameNode holds metadata for the two files

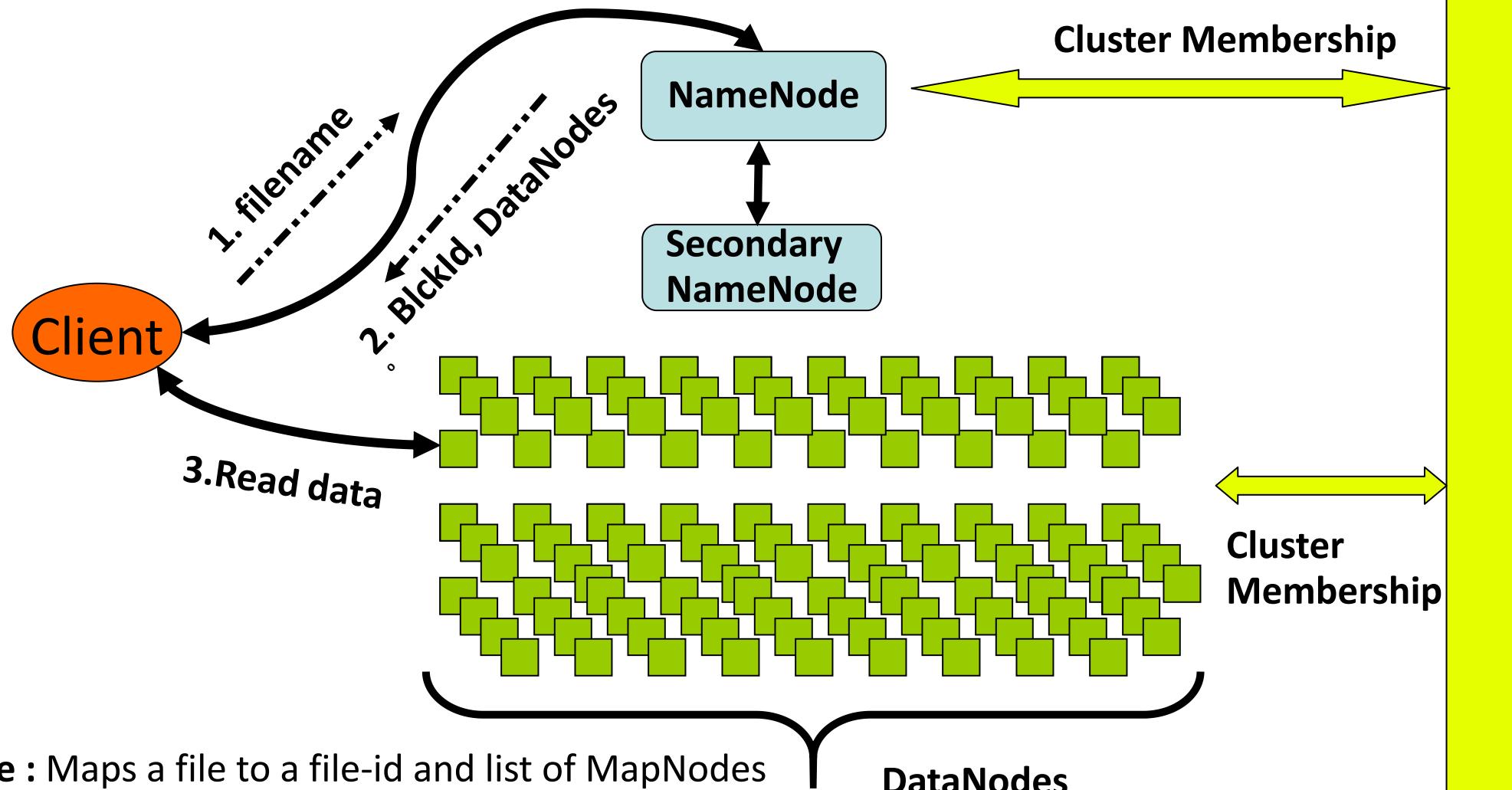
- Foo.txt (300MB) and Bar.txt (200MB)
- Assume HDFS is configured for 128MB blocks

The DataNodes hold the actual blocks

- Each block is 128MB in size
- Each block is replicated three times on the cluster
- Block reports are periodically sent to the NameNode



HDFS Architecture

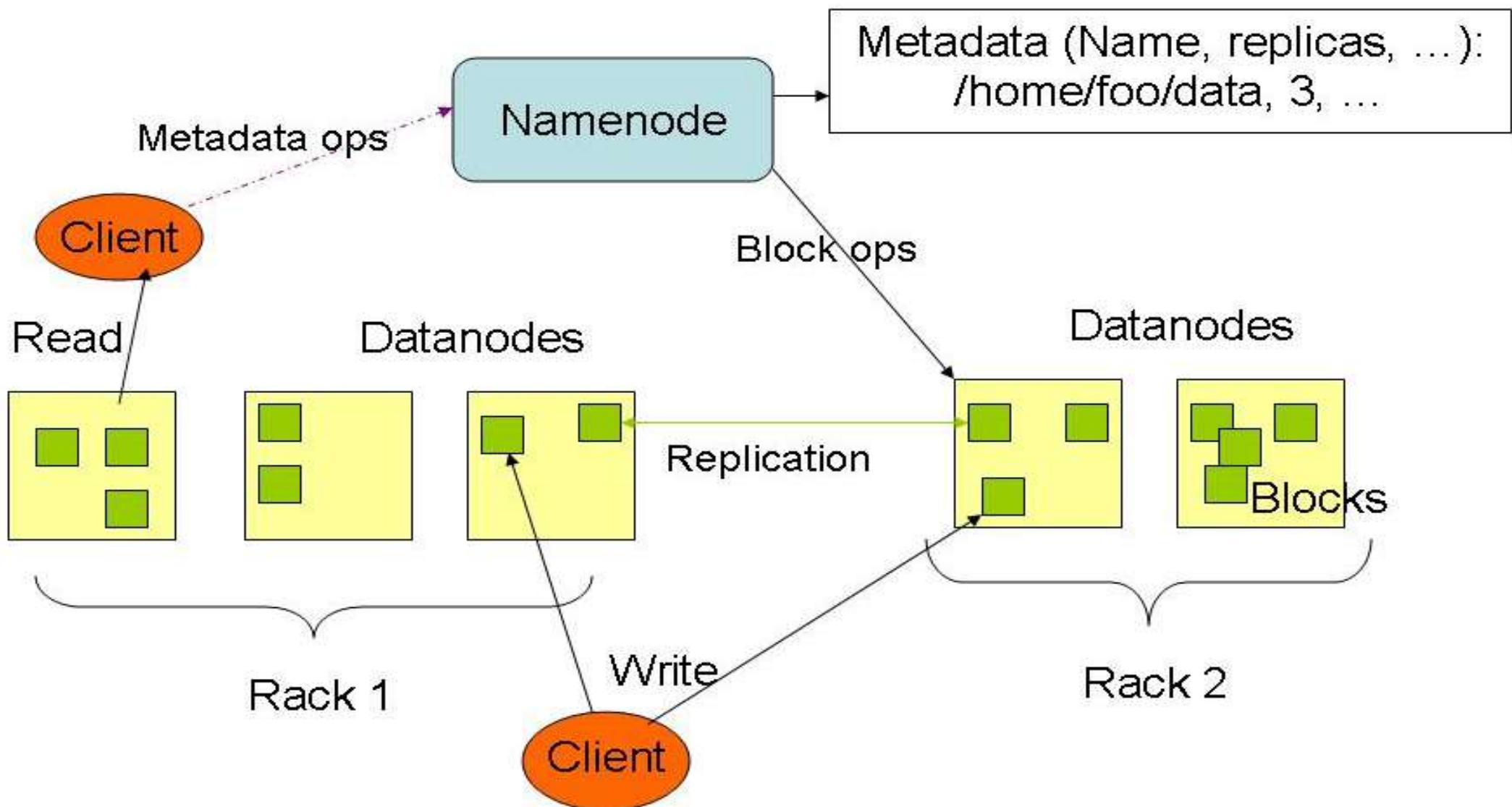


NameNode : Maps a file to a file-id and list of MapNodes

DataNode : Maps a block-id to a physical location on disk

SecondaryNameNode: Periodic merge of Transaction log

HDFS Architecture



Role of NameNode

- The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software.
- It is a software that can be run on commodity hardware.
- The system having the namenode acts as the master server.
 - It stores all metadata: filenames, locations of each block on Data Nodes, file attributes, etc...
 - Block and Replica management
 - Health of Data Nodes through block reports
 - Keeps metadata in RAM for fast lookup
 - Regulates client's access to files.
 - It also executes file system operations such as renaming, closing, and opening files and directories

Functionalities of NameNode

- Running on a single machine, the **NameNode daemon determines and tracks where the various blocks of a data file are stored.**
- **If a client application wants to access a particular file stored in HDFS, the application contacts the NameNode.**
- **NameNode provides the application with the locations of the various blocks for that file.**

Functionalities of NameNode

- For performance reasons, the NameNode resides in a machine's memory.
- Because the NameNode is critical to the operation of HDFS, any unavailability or corruption of the NameNode results in a data unavailability event on the cluster.
- Thus, the NameNode is viewed as a single point of failure in the Hadoop environment.
- To minimize the chance of a NameNode failure and to improve performance, the NameNode is typically run on a dedicated machine.

Role of DataNode

- The datanode is a commodity hardware having the GNU/Linux operating system and datanode software.
- For every node (Commodity hardware/System) in a cluster, there will be a datanode.
 - The **DataNode daemon manages the data stored on each machine.**
 - It stores file contents as blocks.
 - Different blocks of the same file are stored on different Datanodes
 - Same block is replicated across several Datanodes for redundancy

Functionalities of DataNode

- Datanodes perform read-write operations on the file systems, as per client request.
- They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.
- Each DataNode periodically builds a report about the blocks stored on the DataNode and sends the report to the NameNode.

Role of Secondary NameNode

- A helper node for Namenode
- Performs memory-intensive administrative functions for the Namenode
- Have a check point for the file system (HDFS)
- ***Not a Backup Node***
- Recommended to run on a separate machine
 - It requires as much RAM as the primary NameNode

Functionalities of Secondary NameNode

- A third daemon, the **Secondary NameNode**.
- It **provides the capability to perform some of the NameNode tasks** to reduce the load on the NameNode.
- Such tasks include updating the file system image with the contents of the file system edit logs.
- In the event of a NameNode outage, the NameNode must be restarted and initialized with the last file system image file and the contents of the edits logs.
- Periodically combines a prior filesystem snapshot and editing into a new snapshot. New snapshot is sent back to the NameNode.

NameNode Failure

- Loosing a NameNode is equivalent to losing all the files on the filesystem
- Hadoop provides two options:
 - Back up files that make up the persistent state of the file system (local or NFS mount)
 - Run a Secondary NameNode

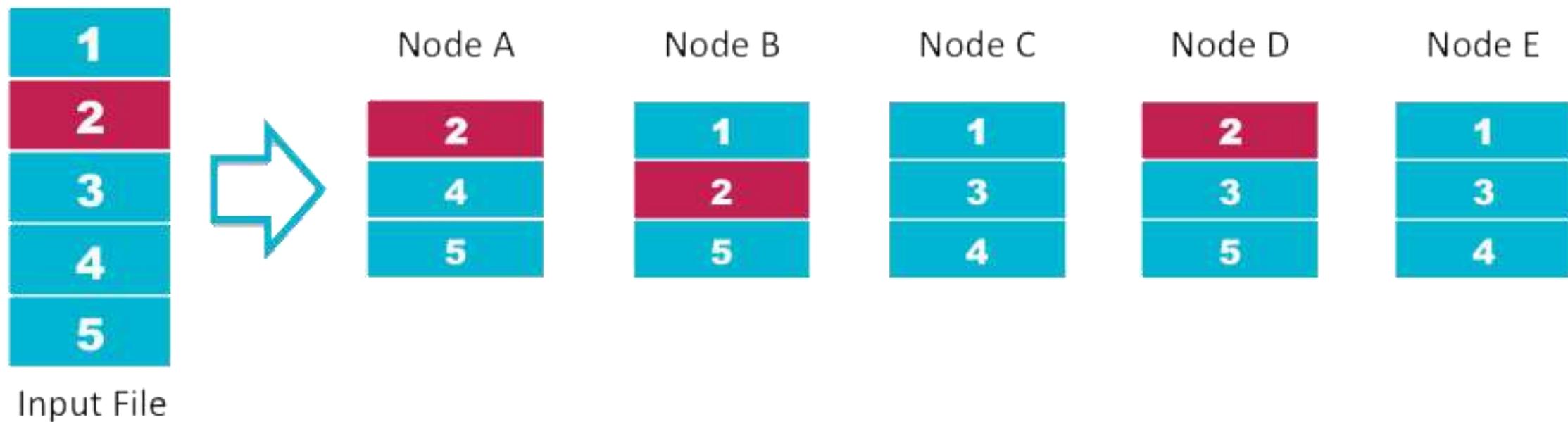
DataNode Failure and Recovery

- DataNodes exchange **heartbeats** with NameNode
- If no heartbeat received within a certain time period DataNode is assumed to be lost.
 - NameNode determines which blocks were on the lost node
 - NameNode finds other copies of these 'lost' blocks and replicates them to other nodes.
 - Block replication is actively maintained.

Data Replication

- Default replication is 3-fold

HDFS Data Distribution



Block Concept

TestFile1.txt -> 1GB

Block Size -> 64 MB

Files are splitted into number of chunks(Blocks) of pre-defined size

No of Blocks = 1GB / 64MB = 16 blocks

Blocks are B1,B2,.....B16

DataNode

DataNode

DataNode

DataNode

B1

B8

B10

B13

B3

B7

B12

B16

B4

B5

B11

B14

B2

B6

B9

B15

Block Concept

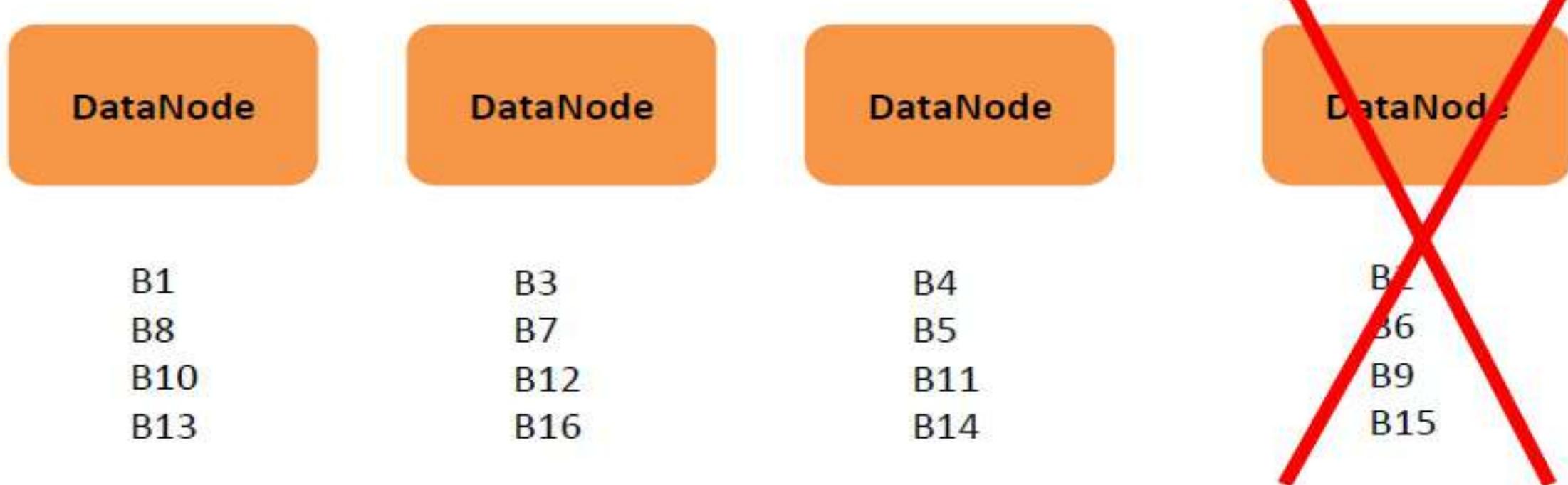
TestFile1.txt -> 1GB

Block Size -> 64 MB

No of Blocks = 1GB / 64MB = 16 blocks

Blocks are B1,B2,...,B16

What happens to my data if
node 4 goes down??



Fault Tolerant in HDFS

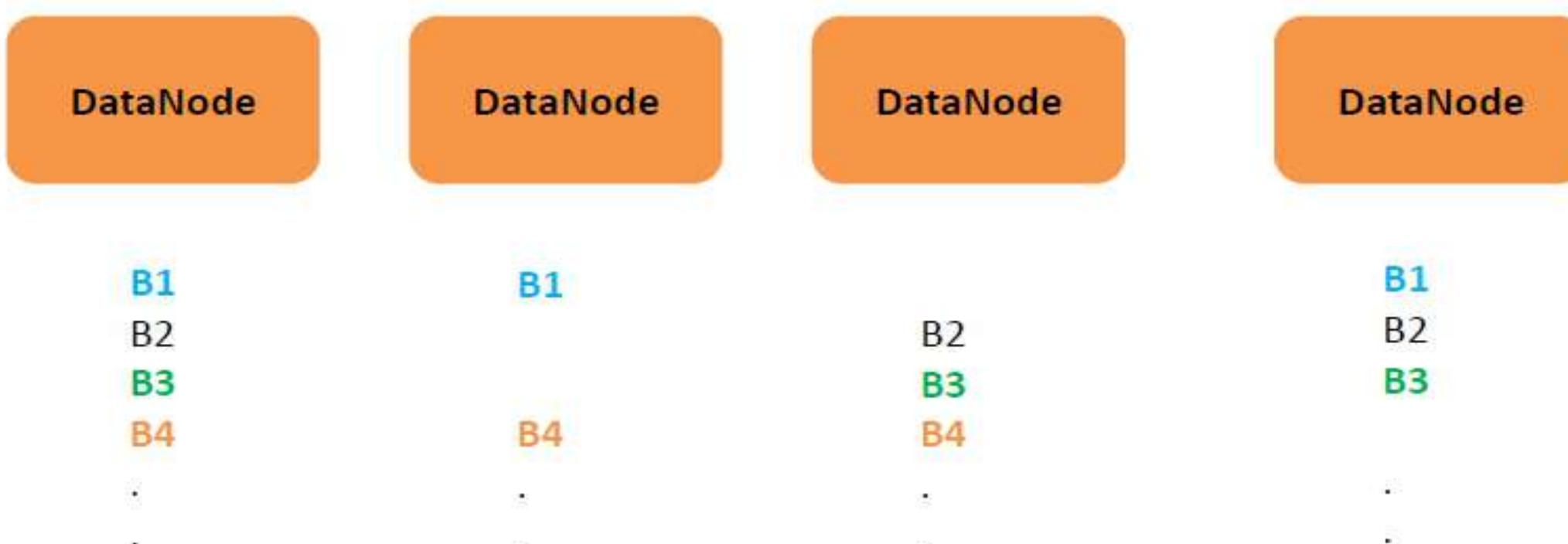
TestFile1.txt → 1GB

Block Size → 64 MB

HDFS provides fault tolerant by replication of each block by 3

No of Blocks = 1GB / 64MB = 16 blocks

Blocks are B1,B2,...,B16



Block Placement

- Default strategy:
 - One replica on local node
 - Second replica on a node in the remote rack
 - Third replica on the some node in the remote rack
 - Additional replicas are random
- Clients always read from nearest node
- Client retrieves a list of DataNodes on which to place replicas of a block.
- Client writes block to the first DataNode

Block Placement

- The first DataNode forwards the data to the next DataNode in the pipeline.
- When all replicas are written, the client moves on the write the next block in file
- Important is, that multiple machines are involved in writing one files and they could be different machines.

Balancing Hadoop Cluster

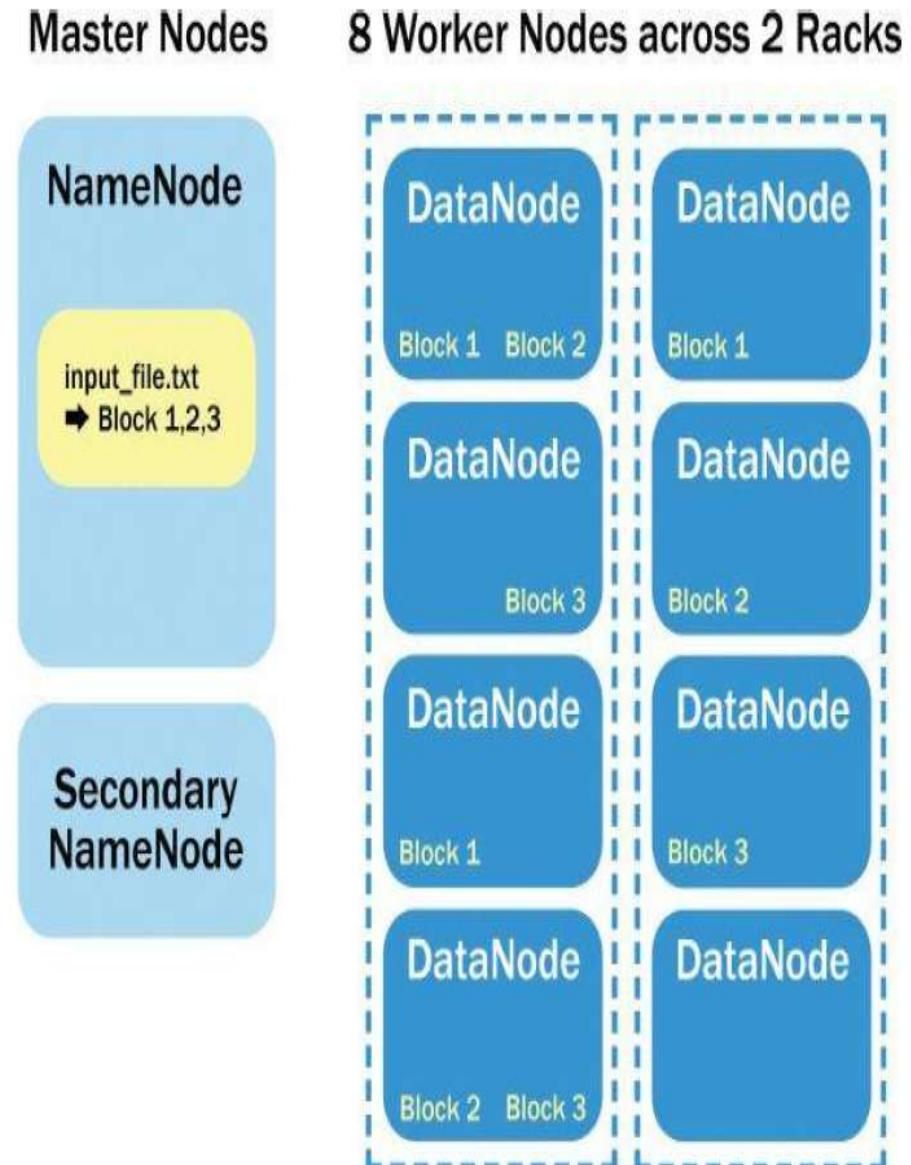
- Hadoop works best when blocks are evenly spread out
- Goal: Have % Disk Full on DataNodes at the same level.
- Balancer - Hadoop daemon
 - **% *start-balancer.sh***
 - Re-distributes blocks from over-utilised to under-utilised DataNodes
 - Run it when new DataNodes are added.
 - Runs in the background and can be throttled to avoid network congestion/negative cluster impact.

Data Retrieval

- When a client wants to retrieve data
 - Communicates with the NameNode to determine which blocks make up a file and on which data nodes those blocks are stored
 - Then communicated directly with the data nodes to read the data

Data Retrieval

- The latest versions of Hadoop provide an HDFS High Availability (HA) feature.
- This feature enables the use of **two NameNodes: one in an active state, and the other in a standby state**.
- If an active NameNode fails, the standby NameNode takes over.
- When using the HDFS HA feature, a Secondary NameNode is unnecessary.
- **Figure illustrates a Hadoop cluster with ten machines and the storage of one large file requiring three HDFS data blocks.**
- Furthermore, this file is stored using triple replication.
- The machines running the NameNode and the Secondary NameNode are considered master nodes.



Introducing MapReduce

Introducing MapReduce

- Now that we have described how Hadoop stores data, lets turn our attention to how it processes data
- We typically process data in Hadoop using MapReduce
- MapReduce is not a language, it's a programming model
- MapReduce is a method for distributing a task across multiple nodes. Each node processes data stored on that node.
- MapReduce consists of two functions:
 - **map (K1, V1) -> (K2, V2)**
 - **reduce (K2, list(V2)) -> list(K3, V3)**

Why Map Reduce is So Popular?

- Automatic parallelization and distribution (The biggest advantage).
- Fault-tolerance (individual tasks can be retried)
- Hadoop comes with standard status and monitoring tools.
- A clean abstraction for developers.
- MapReduce programs are usually written in Java (possibly in other languages using streaming)

Understanding Map and Reduce

The **map** function always runs first

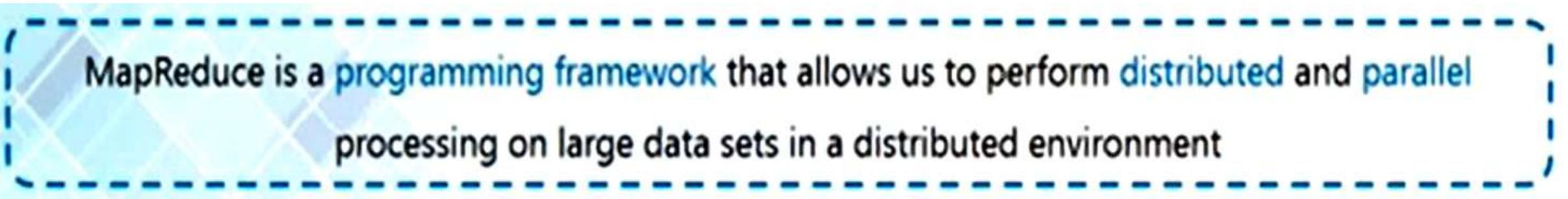
- Typically used to “break down”
 - Filter, transform, or parse data, e.g. Parse the stock symbol, price and time from a data feed
- The output from the map function (eventually) becomes the input to the reduce function

The **reduce** function

- Typically used to aggregate data from the map function
 - e.g. Compute the average hourly price of the stock
- Not always needed and therefore optional
 - You can run something called a “map-only” job

MapReduce paradigm

- **Divide & conquer:** partition a large problem into smaller sub-problems
 - **Independent sub-problems** can be executed in parallel by workers (anything from threads to clusters)
 - Intermediate results from each worker are **combined** to get the final result



MapReduce is a **programming framework** that allows us to perform **distributed** and **parallel** processing on large data sets in a distributed environment

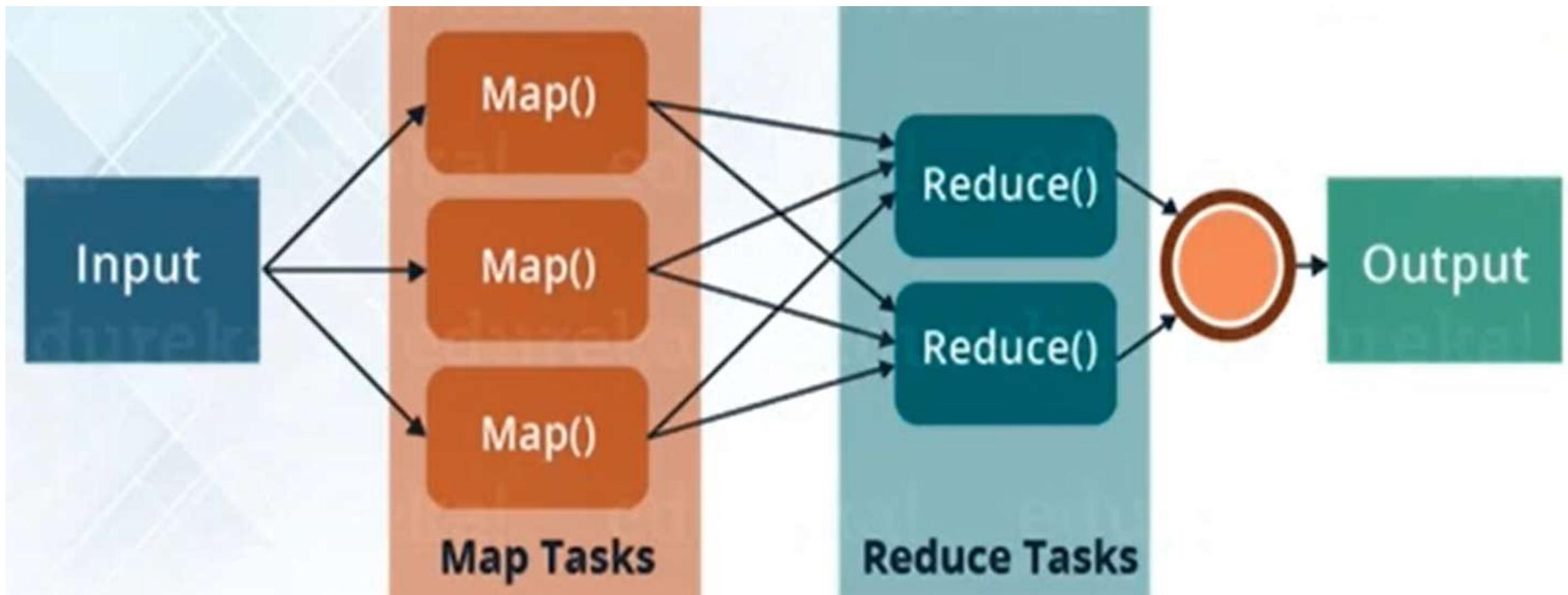
Understanding Map and Reduce

Between these two tasks there is typically a hidden phase known as the
“Shuffle and Sort”

- Which organizes map output for delivery to the reducer

Each individual piece is simple, but collectively are quite powerful

- Analogous to a pipe / filter in Unix



Typical Large Data Problem

- **Iterate over a large number of records**
- **Extract something of interest from each**
- **Shuffle and sort intermediate results**
- **Aggregate intermediate results**
- **Generate final output**

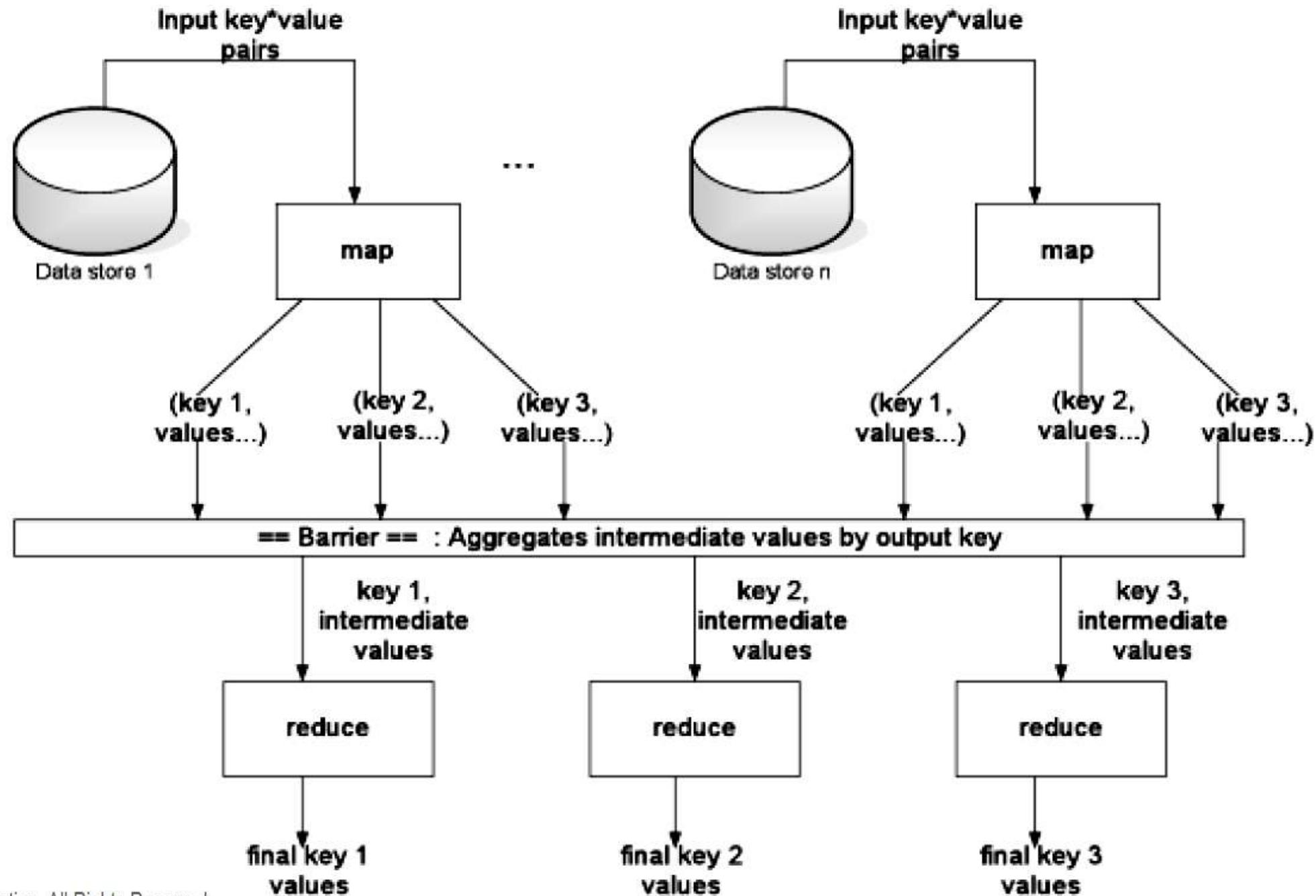


JobTracker and TaskTracker

- ▶ JobTracker
 - Determines the execution plan for the job
 - Assigns individual tasks

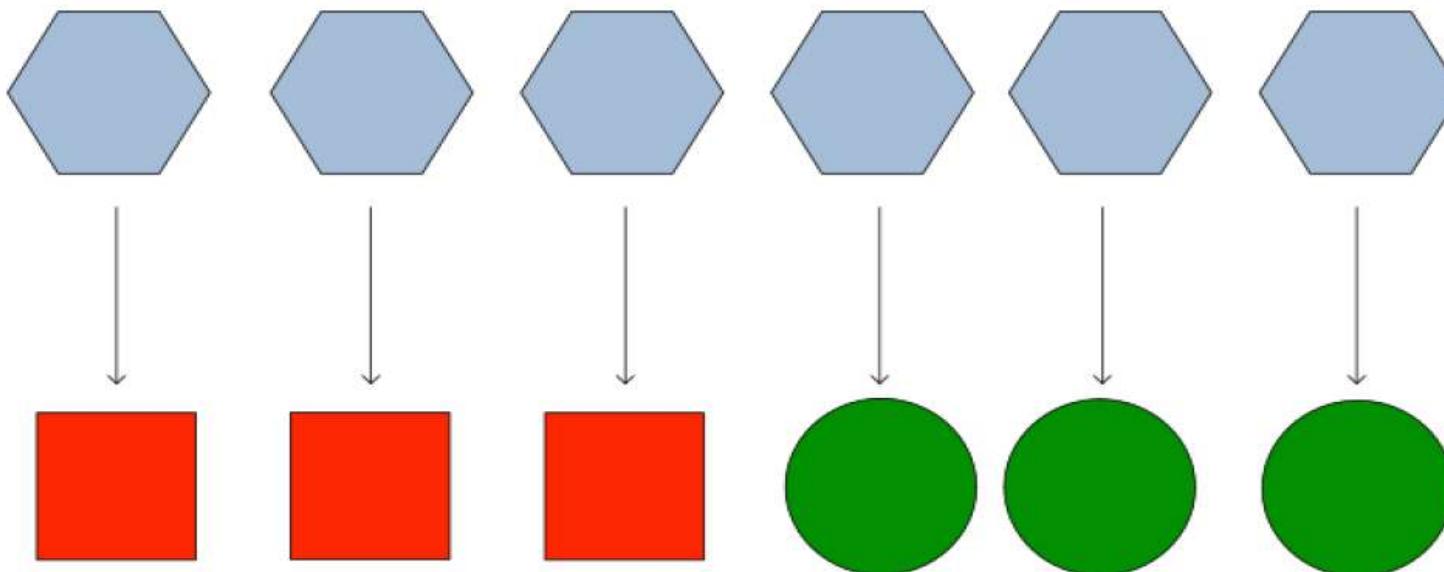
- ▶ TaskTracker
 - Keeps track of the performance of an individual mapper or reducer

MapReduce: The Big Picture



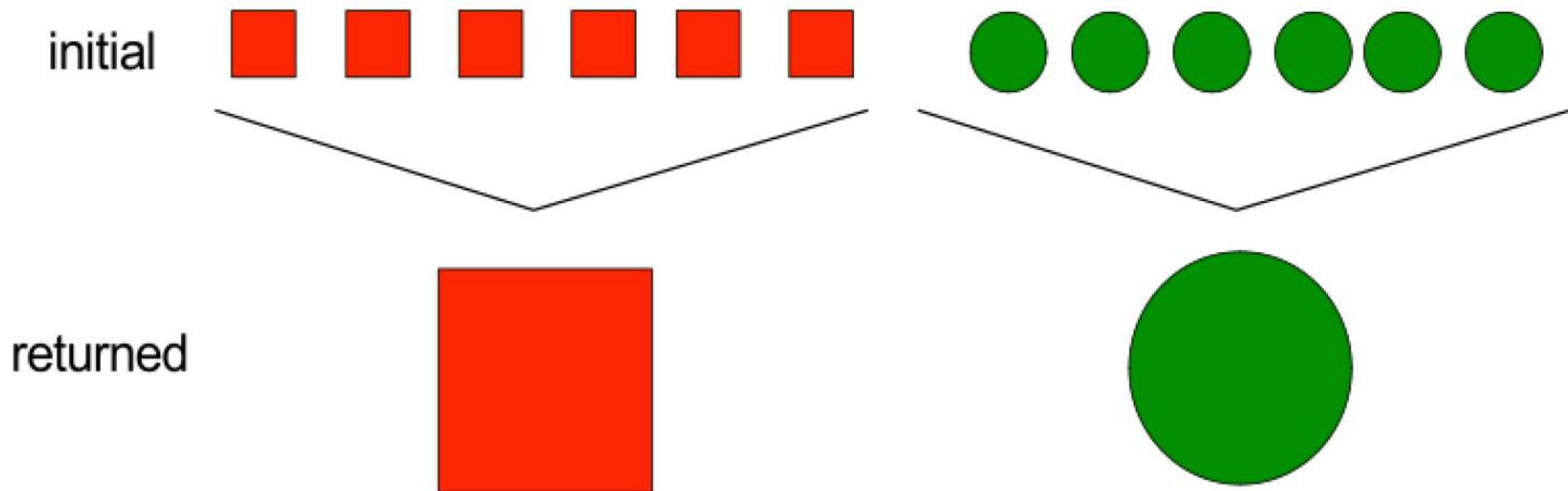
Map Process

- **map (in_key, in_value) → (out_key, out_value)**



Reduce Process

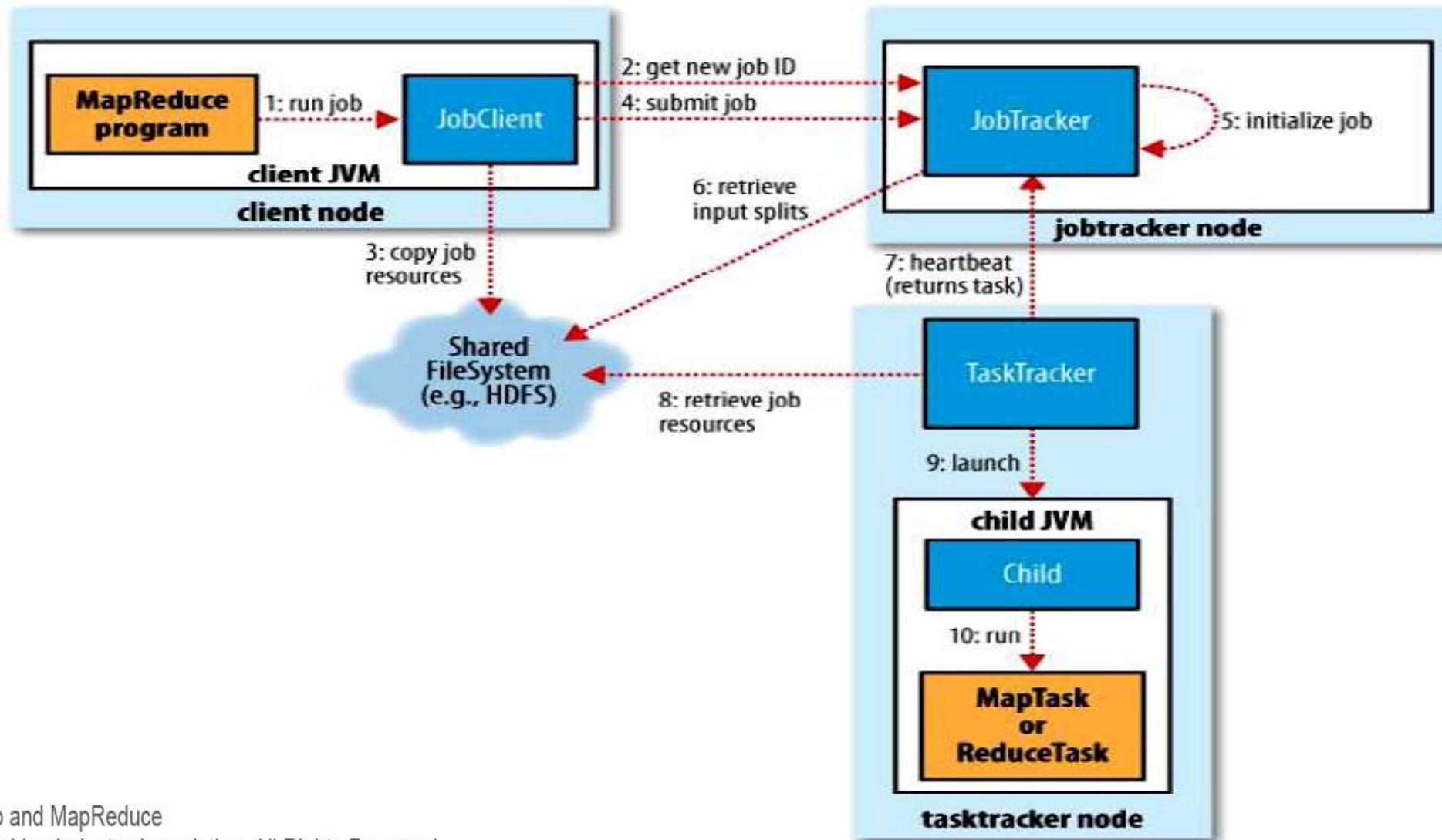
- **reduce (out_key, out_value list) → (final_key, final_value list)**



Terminology

- The client program submits a job to Hadoop.
 - The job consists of a mapper, a reducer, and a list of inputs.
- The job is sent to the **JobTracker** process on the Master Node.
- Each Slave Node runs a process called the **TaskTracker**.
- The JobTracker instructs TaskTrackers to run and monitor tasks.
- A Map or Reduce over a piece of data is a single task.
- A task attempt is an instance of a task running on a slave node.

MapReduce : High Level



MapReduce Failure Recovery

- Task processes send heartbeats to the TaskTracker.
- TaskTrackers send heartbeats to the JobTracker.
- Any task that fails to report in 10 minutes is assumed to have failed- its JVM is killed by the TaskTracker.
- Any task that throws an exception is said to have failed.
- Failed tasks are reported to the JobTracker by the TaskTracker.
- The JobTracker reschedules any failed tasks - it tries to avoid rescheduling the task on the same TaskTracker where it previously failed.
- If a task fails more than 4 times, the whole job fails.

TaskTracker Recovery

- Any TaskTracker that fails to report in 10 minutes is assumed to have crashed.
 - All tasks on the node are restarted elsewhere
 - Any TaskTracker reporting a high number of failed tasks is blacklisted, to prevent the node from blocking the entire job.
 - There is also a “global blacklist”, for TaskTrackers which fail on multiple jobs.
- The JobTracker manages the state of each job and partial results of failed tasks are ignored.

Example: Word Count

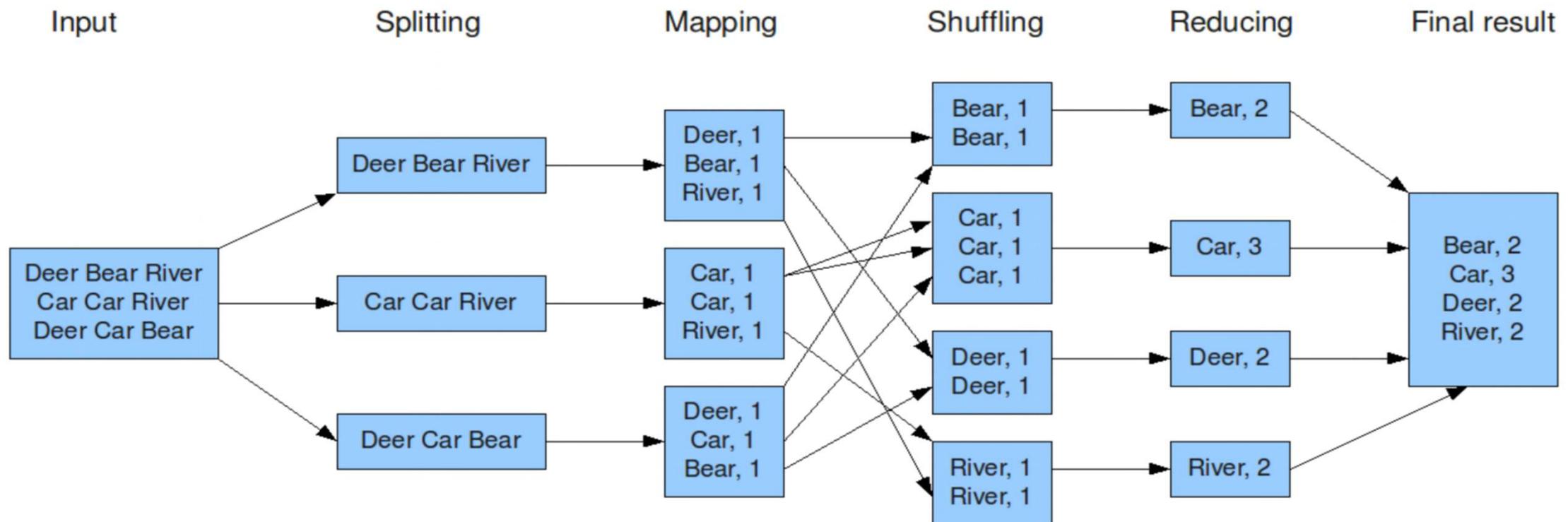
- We have a large file of words, one word to a line
- Count the number of times each distinct word appears in the file
- *Sample application:* analyze web server logs to find popular URLs

MapReduce

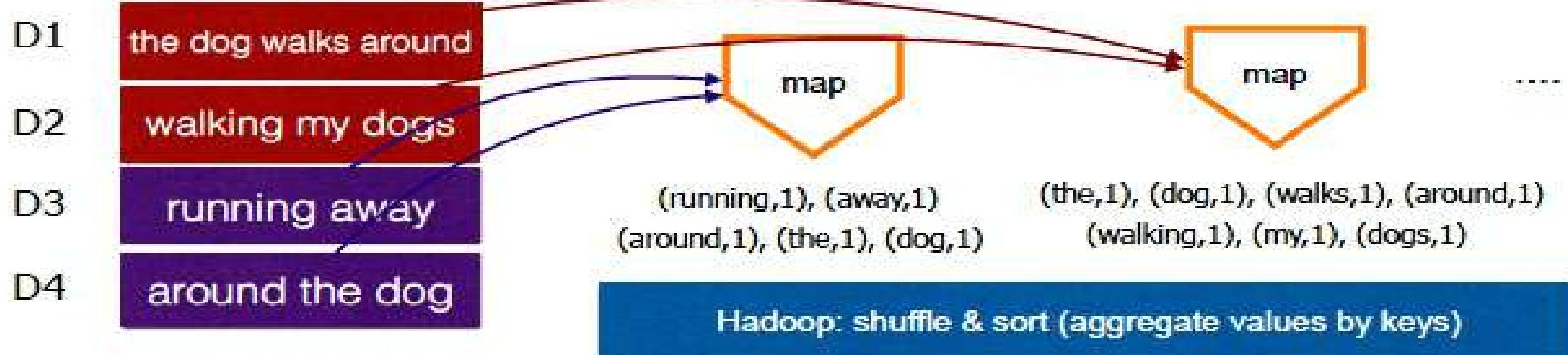
- Input: a set of key/value pairs
- User supplies two functions:
 - $\text{map}(k,v) \rightarrow \text{list}(k_1, v_1)$
 - $\text{reduce}(k_1, \text{list}(v_1)) \rightarrow v_2$
- (k_1, v_1) is an intermediate key/value pair
- Output is the set of (k_1, v_2) pairs

MapReduce: Word Count

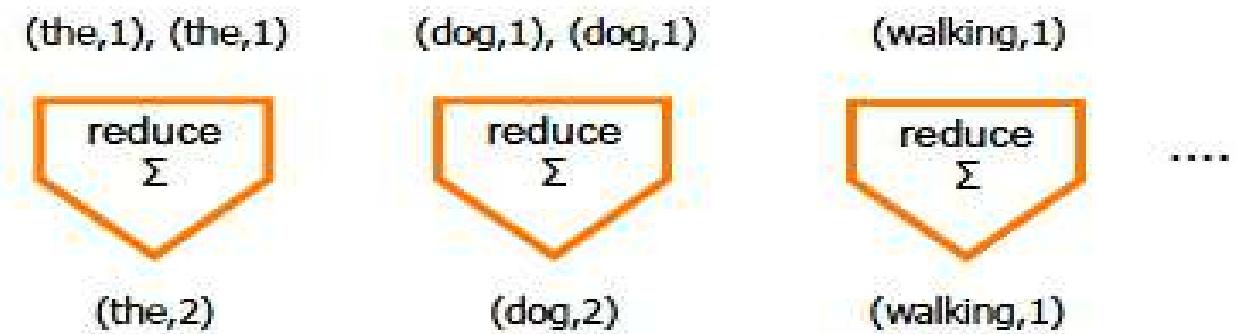
The overall MapReduce word count process



Example: word count



Term	#tf
the	2
dog	2
walks	1
around	2
walking	1
my	1
....	...



MapReduce Example

MapReduce code for Hadoop is typically written in Java

- But it is possible to use nearly any language with Hadoop Streaming

The following slides will explain an entire MapReduce job

- **Input:** Text file containing order ID, employee name, and sale amount
- **Output:** Sum of all sales per employee

Job Input

0	Alice	3625
1	Bob	5174
2	Alice	893
3	Alice	2139
4	Diana	3581
5	Carlos	1039
6	Bob	4823
7	Alice	5834
8	Carlos	392
9	Diana	1804

Job Output

Alice	12491
Bob	1431
Carlos	9997
Diana	5385

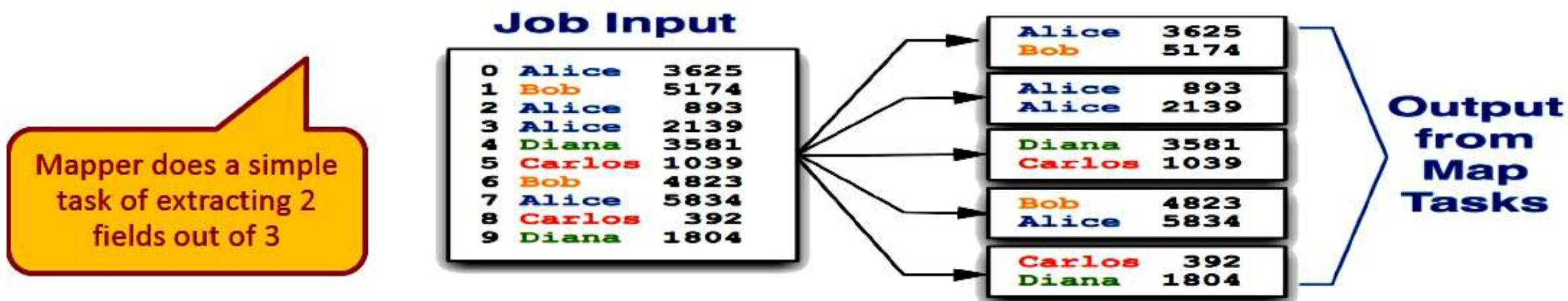
Explanation Of The Map Function

Hadoop splits a job into many individual map tasks

- The number of map tasks is determined by the amount of input data
- Each map task receives a portion of the overall job input to process in parallel
- **Mappers** process one input record at a time
- For each input record, they emit zero or more records as output

In this case, 5 map tasks each parse a subset of the input records

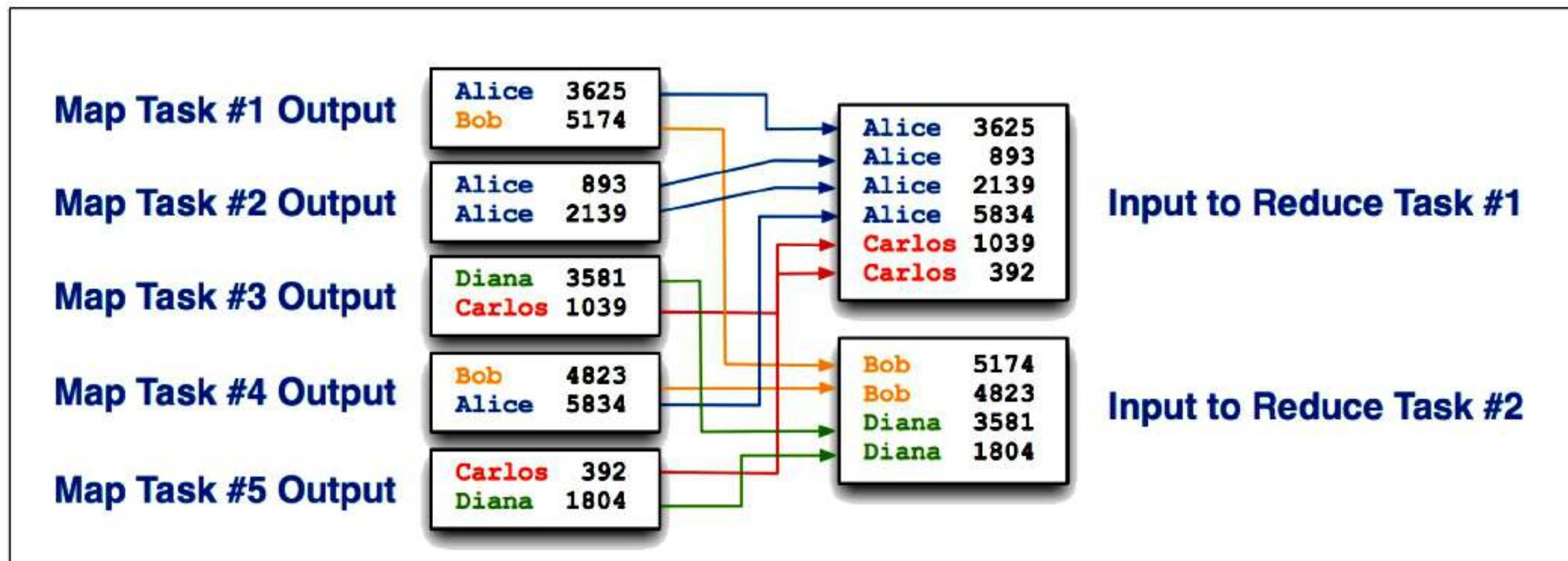
- And then output the name and sales fields for each as output



Shuffle and Sort

Hadoop automatically sorts and merges output from all map tasks

- Sorting records by key (name, in this example)
- This (transparent) intermediate process is known as the “Shuffle and Sort”
- The result is supplied to reduce tasks



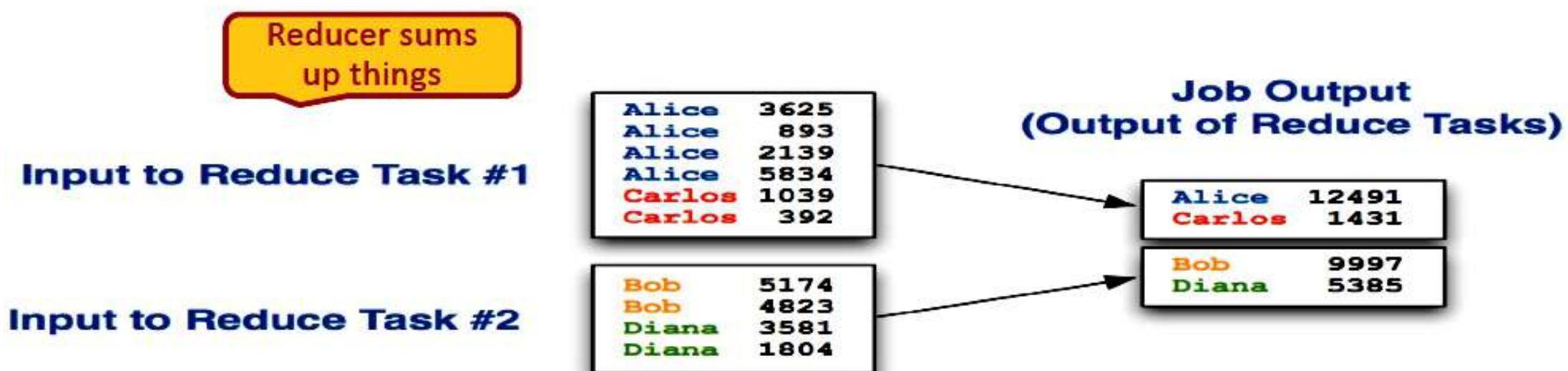
Explanation of Reduce Function

Reducer input comes from the shuffle and sort process

- As with map, the reduce function receives one record at a time
- A given reducer receives all records for a given key
- For each input record, reduce can emit zero or more output records

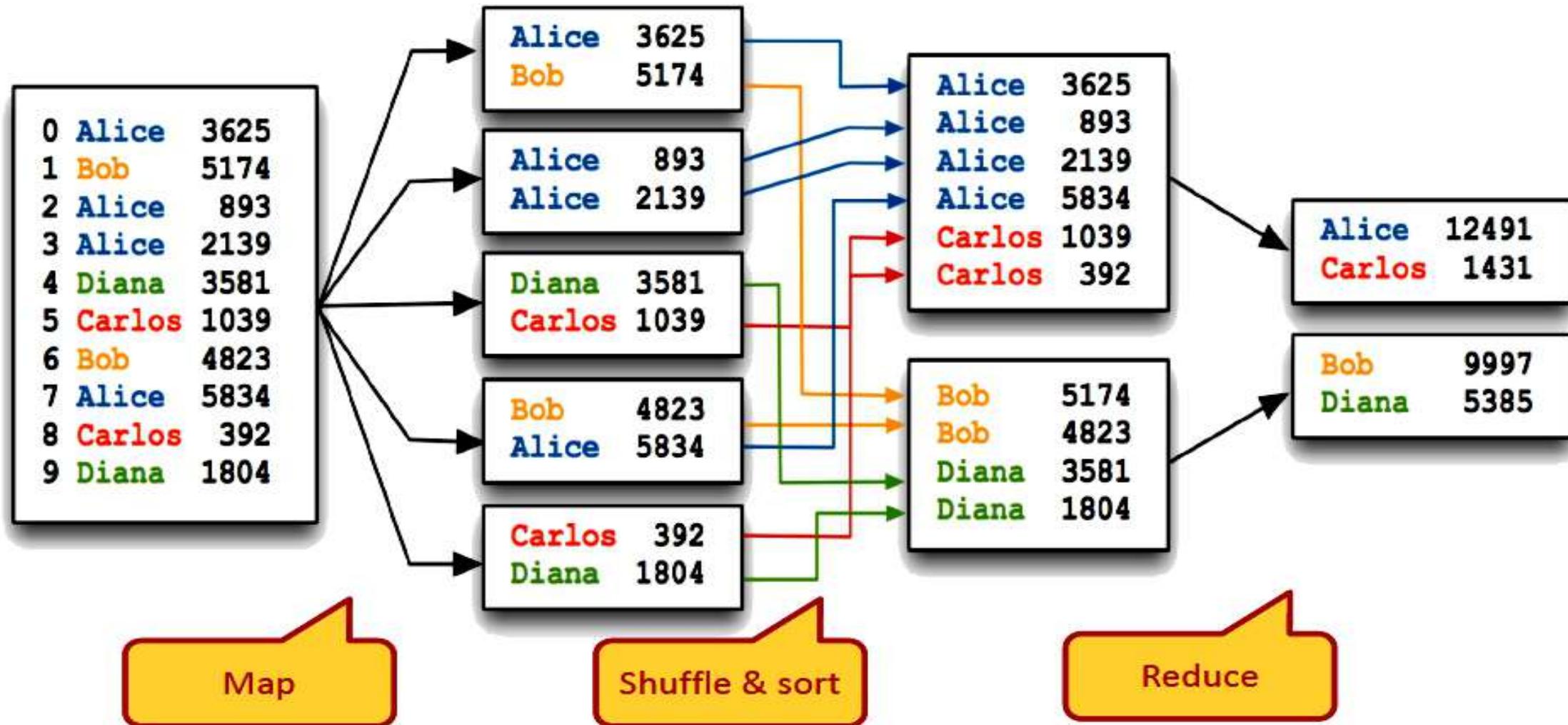
This reduce function sums the total sales per person

- And emits the employee name (key) and total (sales) as output to a file
- Each reducer output file is written to a job specific directory in HDFS



Putting It All Together

Here's the data flow for the entire MapReduce job



Benefits of MapReduce

- Simplicity (via fault tolerance)
 - Particularly when compared with other distributed programming models
- Flexibility
 - Offers more analytic capabilities and works with more data types than platforms like SQL
- Scalability
 - Because it works with
 - Small quantities of data at a time
 - Running in parallel across a cluster
 - Sharing nothing among the participating nodes

Hadoop Resource Management: YARN

(Yet Another Resources Negotiator)

What is YARN ?

It is a cluster management technology which is an open source platform distributed for processing framework.

YARN is the Hadoop processing layer that contains

- A **resource manager**
- A **job scheduler**

YARN allows multiple data processing engines to run on a single Hadoop cluster.

- Batch programs (e.g Spark, Mapreduce)
- Interactive SQL (e.g. Impala)
- Advanced analytics (e.g. Spark, Impala)
- Streaming (e.g. Spark Streaming)

Objectives of YARN

- The main objective of YARN is to construct a framework on Hadoop that allows the cluster resources to be allocated to the specified applications and consider MapReduce has one of these applications.
 - It separates each tasks of the job tracker into separate entities.
 - The job tracker maintains track of both job scheduling which matches the tasks with task tracker.
- Another one is task progress monitoring that take care of tasks and starts again the failed or slower tasks and doing the task bookkeeping like as maintaining counter totals.
- **YARN is a completely rewritten architecture of Hadoop cluster. It seems to be a game-changer for the way distributed applications are implemented and executed on a cluster of commodity machines.**

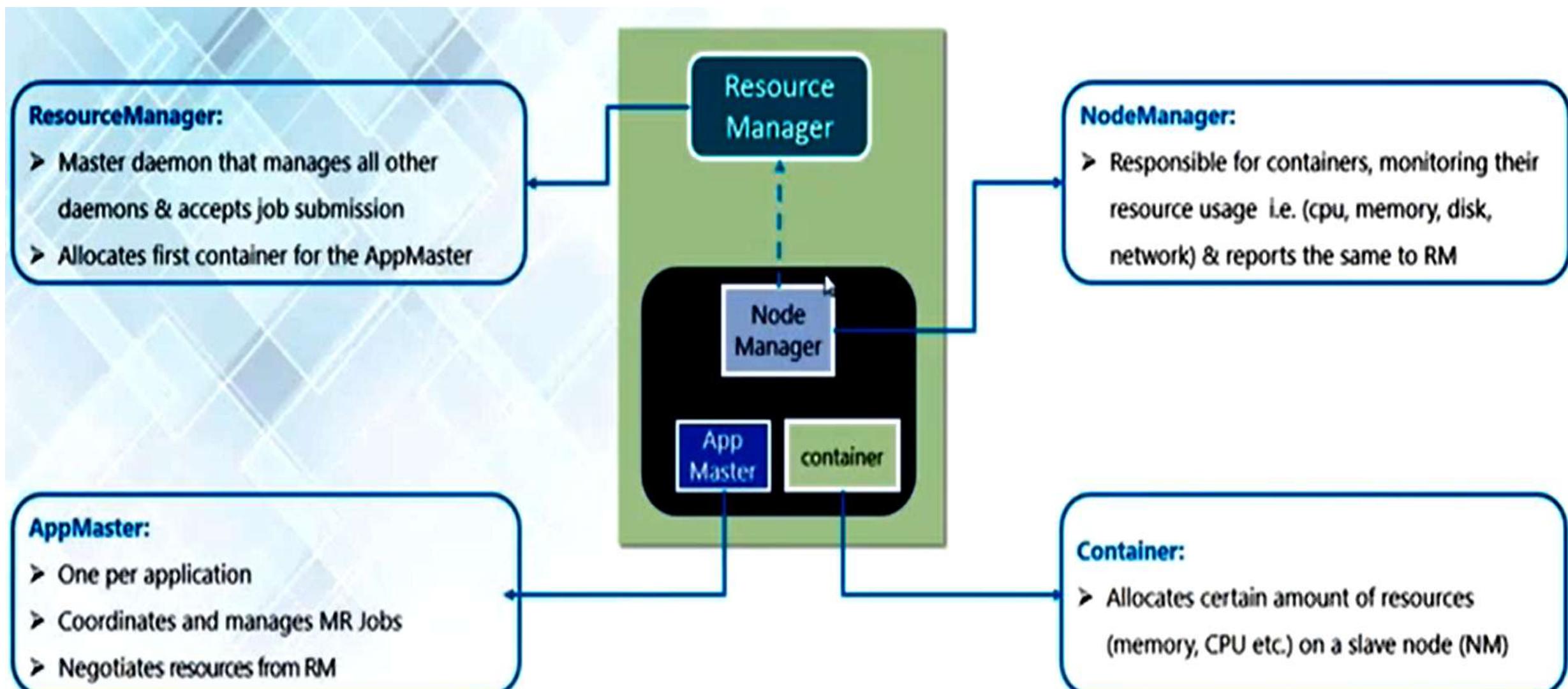
YARN: The next generation of Hadoop's compute platform

YARN came into existence because there was an urgent need to separate the two distinct tasks that go on in a Hadoop ecosystem and which are known as TaskTracker and the JobTracker entities.

Let's slightly change the terminology now. The following name changes give a bit of insight into the design of YARN:

- **ResourceManager** instead of a cluster manager
- **ApplicationMaster** instead of a dedicated and short-lived JobTracker
- **NodeManager** instead of TaskTracker
- A **distributed application** instead of a MapReduce job

Key components of YARN



Key components of YARN (cont..)

- The **Node Manager** and the **Resource Manager** became the main reason on which the new distributed application works.
- The various resources manager are allocated to the system applications using the power of the Resource Manager.
- Application Master works along with the Node Manager and also works on specific framework to get resources from the Resource Manager to manage the various task components.

Key components of YARN (cont..)

- A **scheduler** works with the RM(Resource Manager) framework for the right allocation of resources and ensuring all the constraints of the user limit and queue capacities are adhered are provided at all times.
- As per the requirements of each application the scheduler will provide the right resource.
- The **Application Master** works on basis of coordination with the scheduler in order to get the right resource containers keep an eye on the status and also keep tracking the progress of the process.

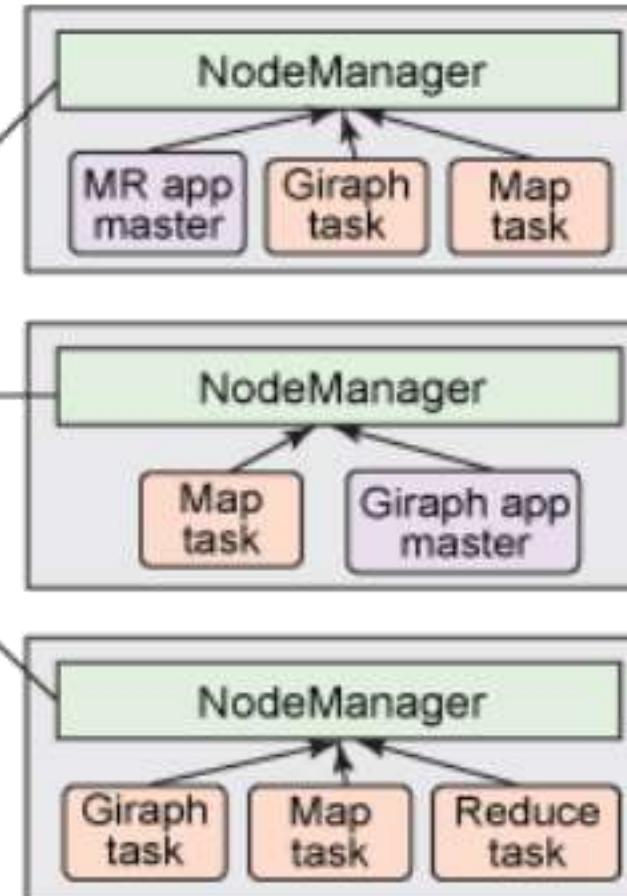
Key components of YARN (cont..)

- The **Node Manager** manages the application containers and launches it when it is required, tracks down the uses of the resources like the memory, processor, network and the disk utilization and gives the entire detailed report to the Resource Manager.

Architecture of YARN

ResourceManager (RM)

- Keeps track of live NodeManagers and available resources
- Allocates available resources to appropriate applications and tasks
- Monitors application masters



Client

- Can submit any type of application supported by YARN

NodeManager (NM)

- Provides computational resources in form of containers
- Manages processes running in containers

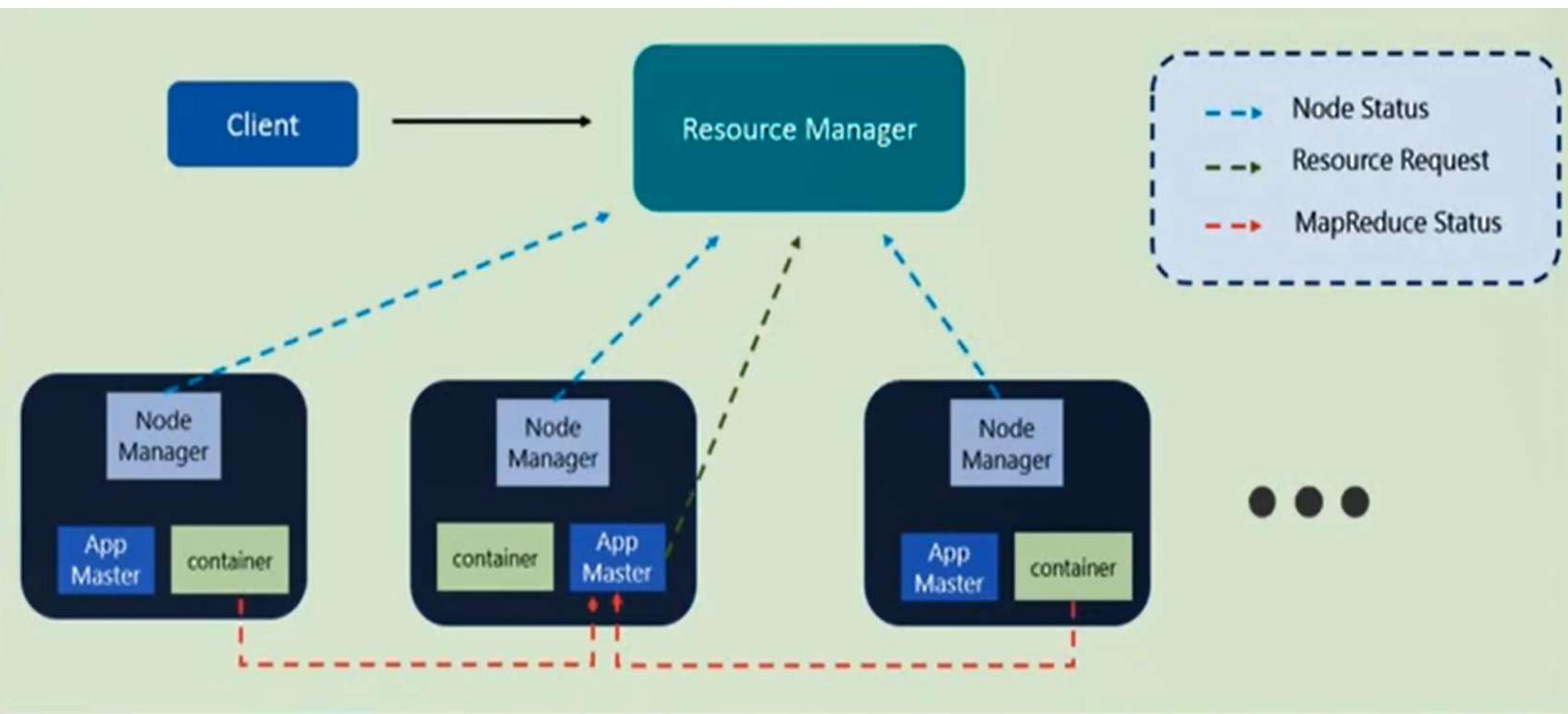
ApplicationMaster (AM)

- Coordinates the execution of all tasks within its application
- Asks for appropriate resource containers to run tasks

Containers

- Can run different types of tasks (also Application Masters)
- Has different sizes e.g. RAM, CPU

Architecture of YARN



Application submission in YARN

➤ *Suppose that users submit applications to the ResourceManager*

- The **ResourceManager** maintains the list of applications running on the cluster and the list of available resources on each live **NodeManager**.
- The ResourceManager needs to determine which application should get a portion of cluster resources next.
- The ResourceManager uses a pluggable **Scheduler**.
- The **Scheduler** focuses only on scheduling; it manages who gets cluster resources (in the form of containers) and when, but it does not perform any monitoring of the tasks within an application so it does not attempt to restart failed tasks.

Application submission in YARN (cont..)

- When the ResourceManager accepts a new application submission, one of the first decisions the Scheduler makes is selecting a container in which ApplicationMaster will run.
- After the **ApplicationMaster** is started, it will be responsible for a whole life cycle of this application.
- First and foremost, it will be sending resource requests to the ResourceManager to ask for containers needed to run an application's tasks.
- A resource request is simply a request for a number of containers that satisfies some resource requirements, such as:
 - *An amount of resources, today expressed as megabytes of memory and CPU shares*
 - *A preferred location, specified by hostname, rackname, or * to indicate no preference*
 - *A priority within this application, and not across multiple applications*

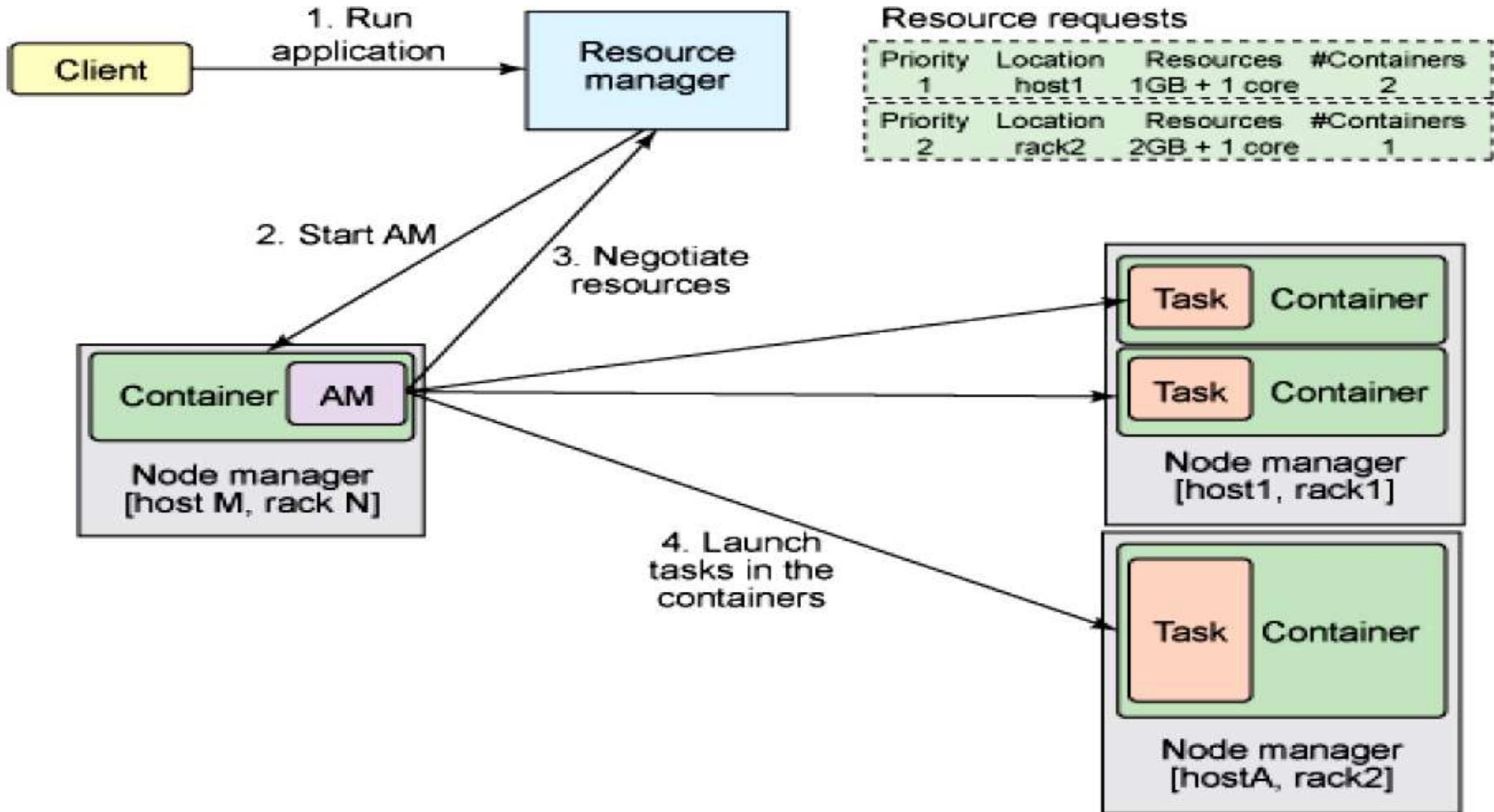
Application submission in YARN (cont..)

- If and when it is possible, the ResourceManager grants a container (expressed as container ID and hostname) that satisfies the requirements requested by the ApplicationMaster in the resource request.
- A container allows an application to use a given amount of resources on a specific host.
- After a container is granted, the ApplicationMaster will ask the NodeManager (that manages the host on which the container was allocated) to use these resources to launch an application-specific task.
- This task can be any process written in any framework.
- The NodeManager does not monitor tasks; it only monitors the resource usage in the containers and, for example, it kills a container if it consumes more memory than initially allocated.

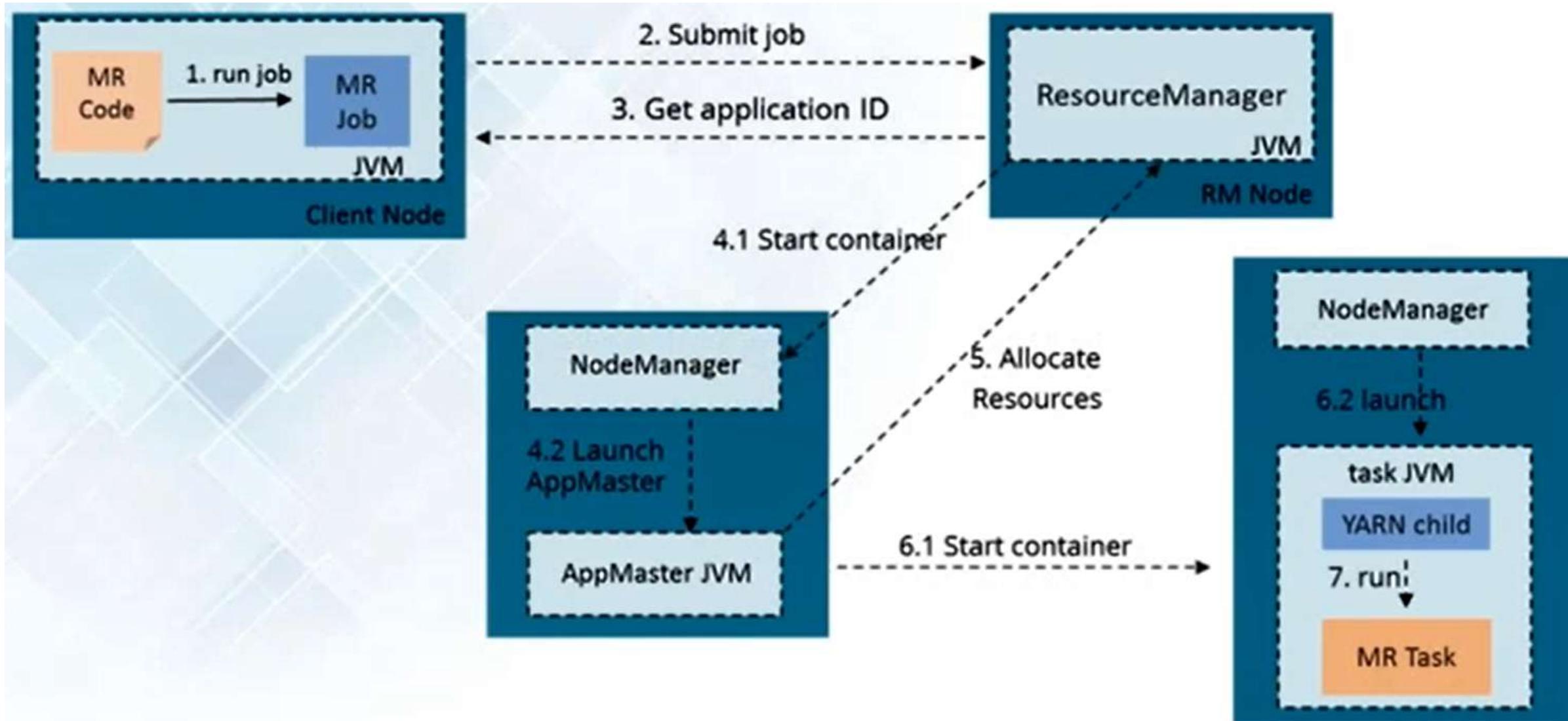
Application submission in YARN (cont..)

- The ApplicationMaster spends its whole life negotiating containers to launch all of the tasks needed to complete its application.
- It also monitors the progress of an application and its tasks, restarts failed tasks in newly requested containers, and reports progress back to the client that submitted the application.
- After the application is complete, the ApplicationMaster shuts itself down and releases its own container.
- Though the ResourceManager does not perform any monitoring of the tasks within an application, it checks the health of the ApplicationMasters.
- If the ApplicationMaster fails, it can be restarted by the ResourceManager in a new container.
- You can say that the ResourceManager takes care of the ApplicationMasters, while the ApplicationMasters takes care of tasks.

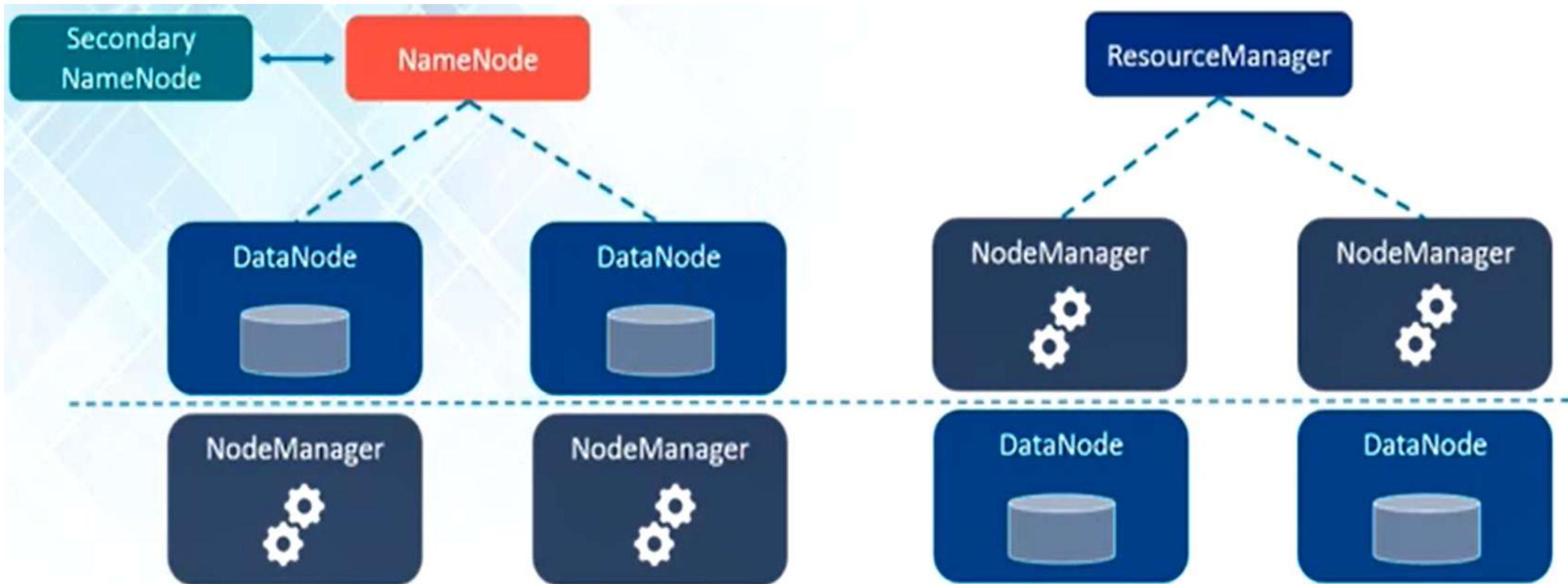
Application submission in YARN (cont..)



MapReduce Job Workflow



HDFC & YARN Architecture



Features of YARN

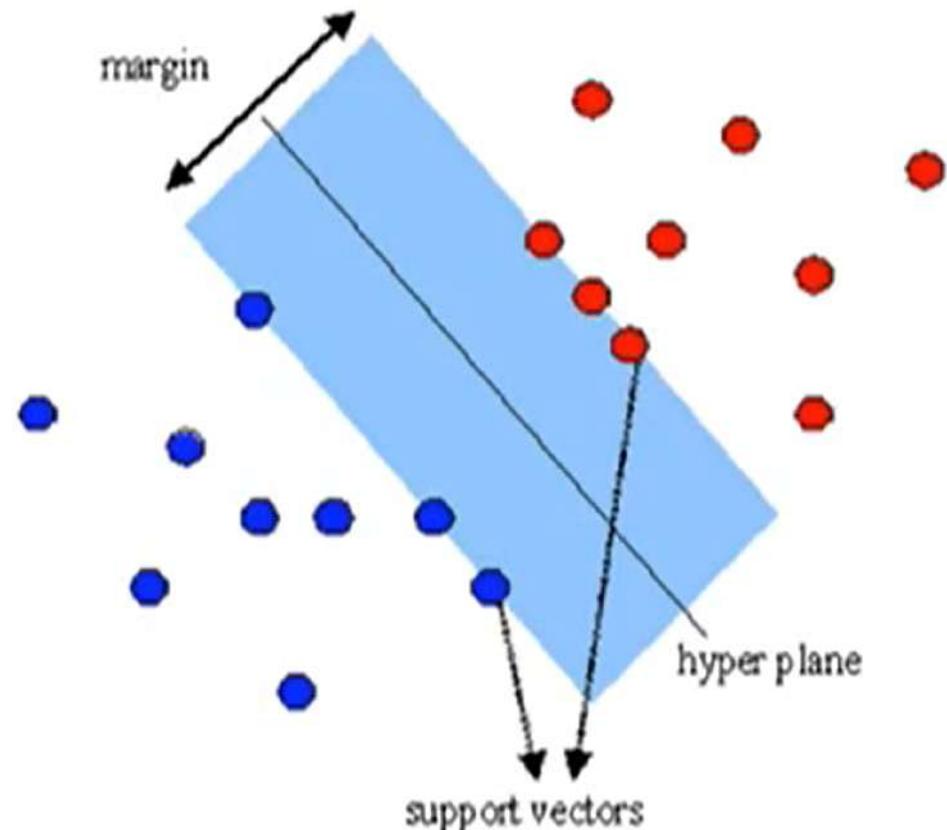
- **High degree of compatibility:** The applications created are using the Map Reduce framework which can easily run on YARN.
- **Better cluster utilization:** YARN allocates all the cluster resources in an efficient and dynamic manner and which leads to utilizes it in much better way compared to previous version of Hadoop.
- **Utmost scalability:** As and when the required number of nodes in the Hadoop cluster expands, the YARN Resource Manager ensures that it meets the user requirements and processing power of the data center does not face any problems in solving.
- **Multi-tenancy:** Various engines that access data on the Hadoop cluster can efficiently works all thanks goes to YARN being a highly versatile technology.

- YARN offers clear advantages in scalability, efficiency, and flexibility compared to the classicalMapReduce engine in the first version of Hadoop. Both small and large Hadoop clusters greatly benefit from YARN.

Support Vector Machine (SVM)

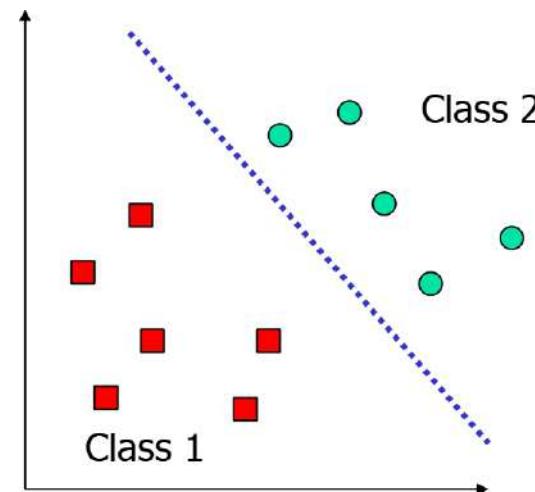
Support Vector Machine (SVM)

- **Supervised** machine learning algorithm
- A SVM performs ***classification*** by finding the ***hyperplane*** that ***maximizes*** the margin between two classes.
- (Note: **Linearly separable** and **binary**)
- It draws the widest channel, or street between two classes
- The two class labels are **+1** (positive examples) and **-1** (negative examples)

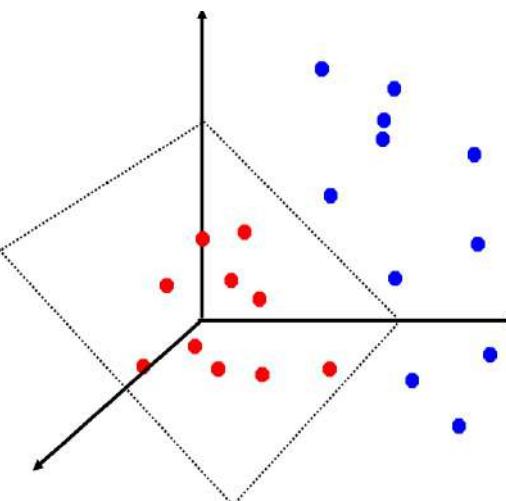
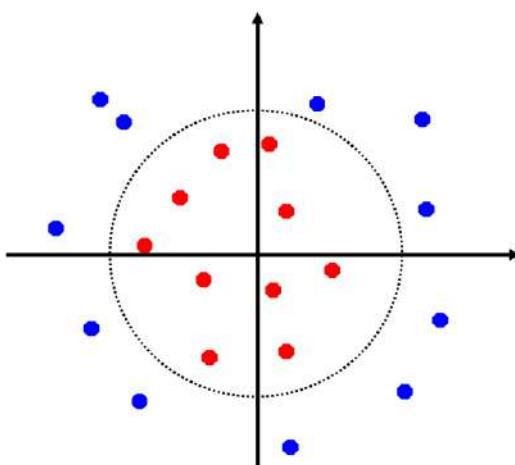
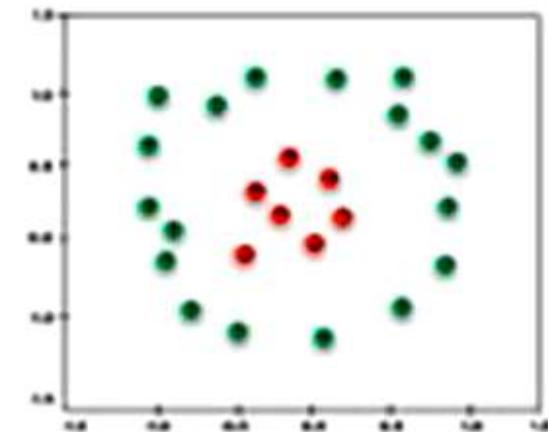


Basic knowledge

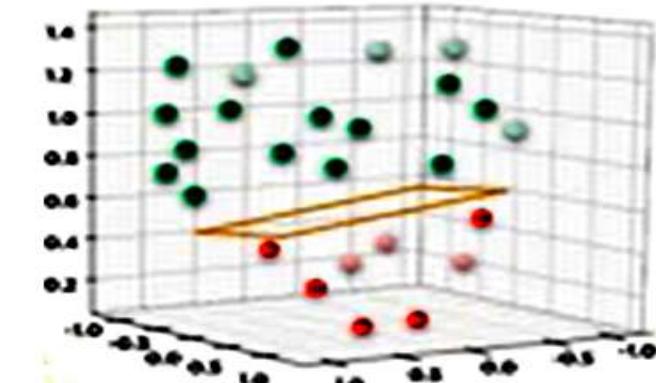
- In 2D we are seeking a line
- In 3D we are seeking a plane (surface)
- In $>3D$ we are seeking a hyperplane
- We will focus on 2D and whatever we can do in 2D, we should be able to do it in higher dimensions



Linear Separable Case
in 2D

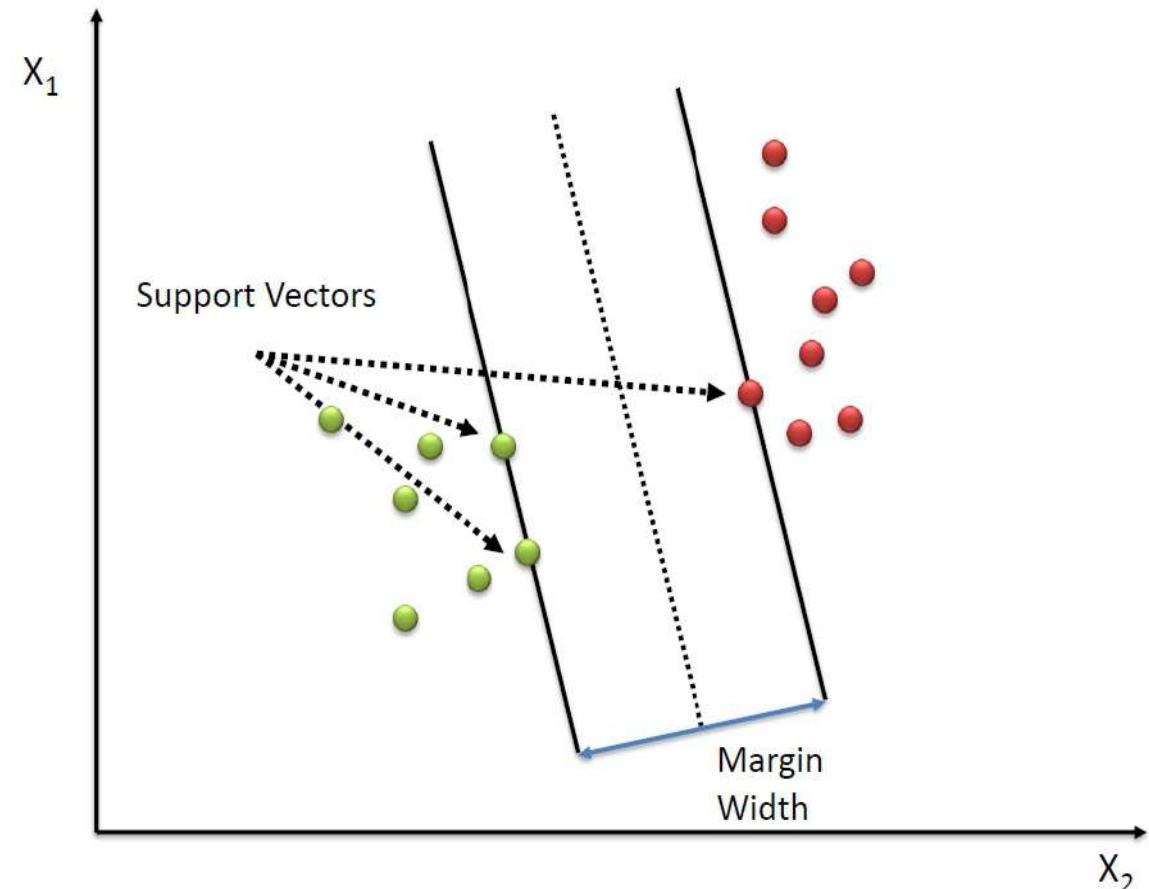


Non-linear SVMs case in
3D or $>3D$



Intuition behind SVM

- Points (instances) are like vectors, $p=(x_1, x_2, \dots, x_n)$
- SVM finds the **closest two points** from the two classes that **support (define)** the best separating line/plane.
- The SVM draws a line connecting them (support vectors from both the classes)
- After that, SVM decides that the best separating line is the line that **bisects**, and is **perpendicular** to the connecting line.



Vectors

- We mentioned that each point will be a vector of the form $\mathbf{p} = (x_1, x_2, \dots, x_n)$
- So, our dataset can be represented as:

$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where n is the number of instances

- x_i is the i^{th} point (example or vector) and y_i is the class associated with that point.
- y_i can either be **-1** or **+1**

- A key concept required here is the dot product between two vectors (inner, or scalar, product)

$$A = [a_1, a_2, \dots, a_n], B = [b_1, b_2, \dots, b_n]$$

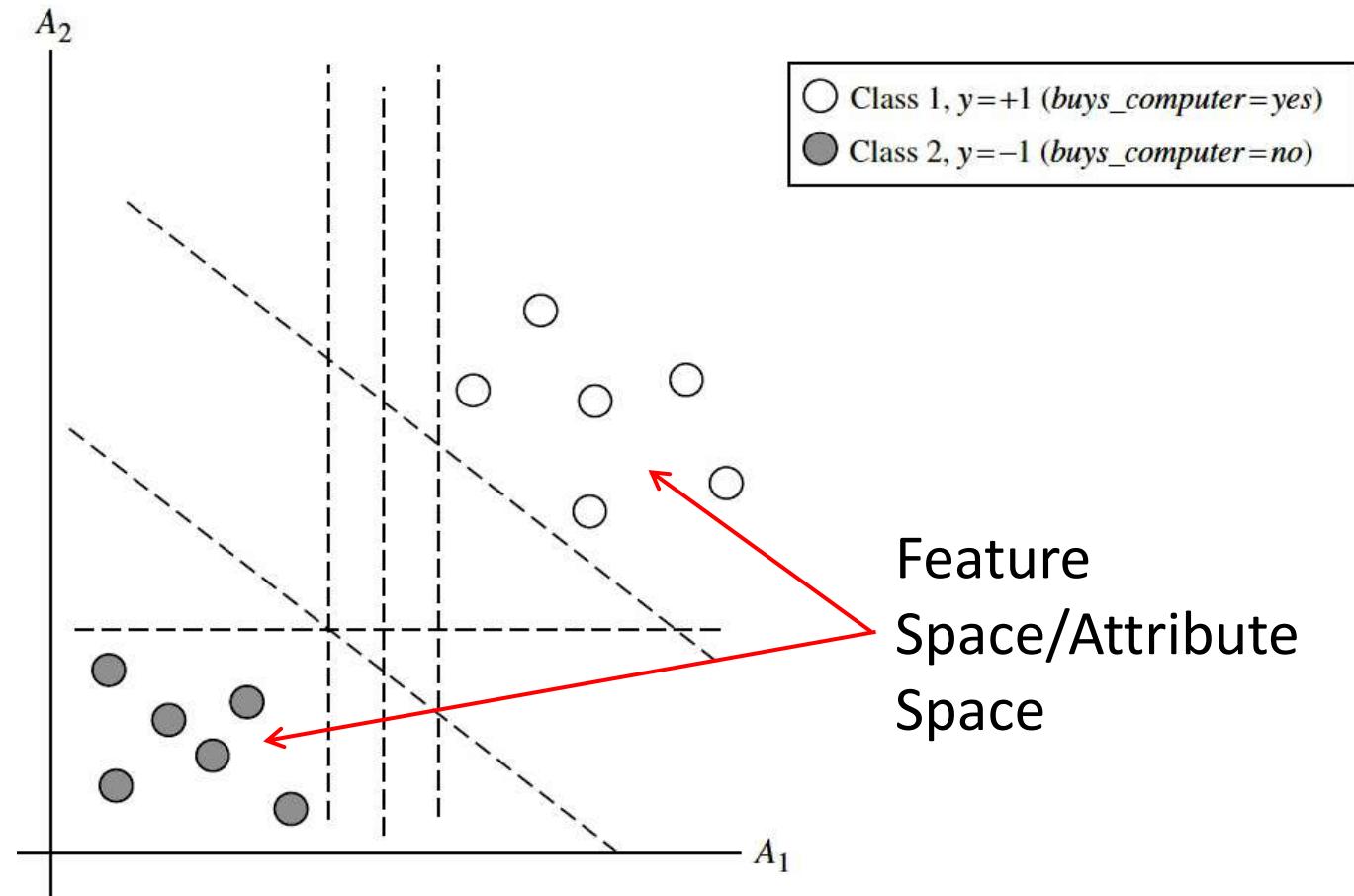
$$A \cdot B = a_1 * b_1 + a_2 * b_2 + \dots + a_n * b_n$$

When the Data Are Linearly Separable? Which Hyperplane?

Let the data set D be given as $(X_1, y_1), (X_2, y_2), \dots, (X_{|D|}, y_{|D|})$, where X_i is the set of training tuples with associated class labels, y_i .

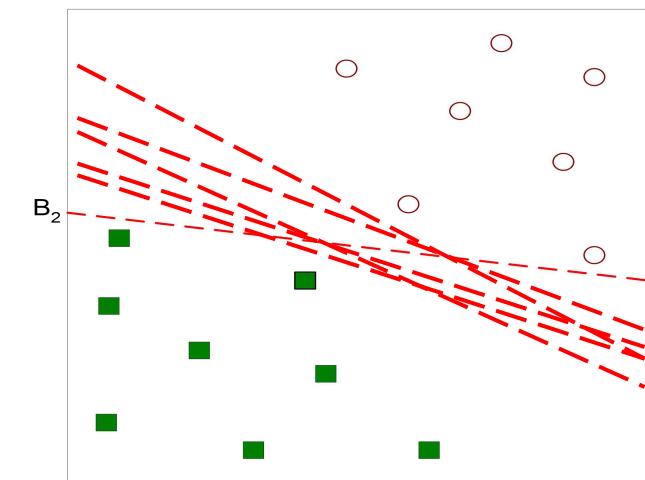
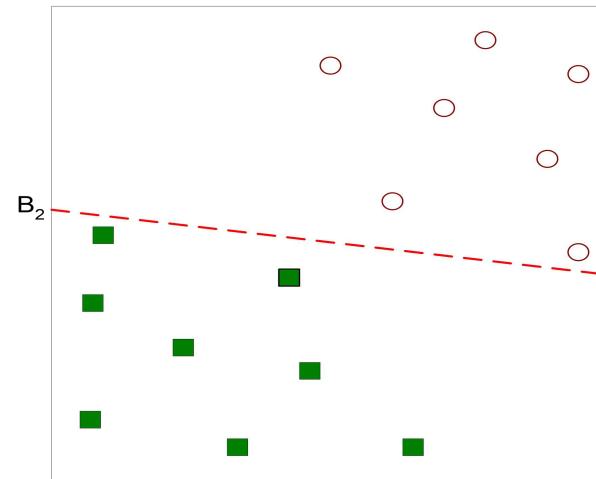
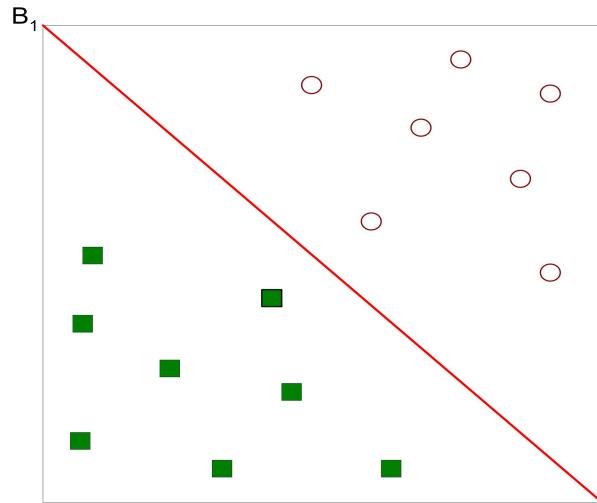
- Each y_i can take one of two values, either +1 or -1 (i.e., $y_i \in \{+1, -1\}$), corresponding to the classes ***buys computer = yes*** and ***buys computer = no***, respectively.

We see that the 2-D data are linearly separable (or “linear,” for short), because a straight line can be drawn to separate all the tuples of class +1 from all the tuples of class -1.



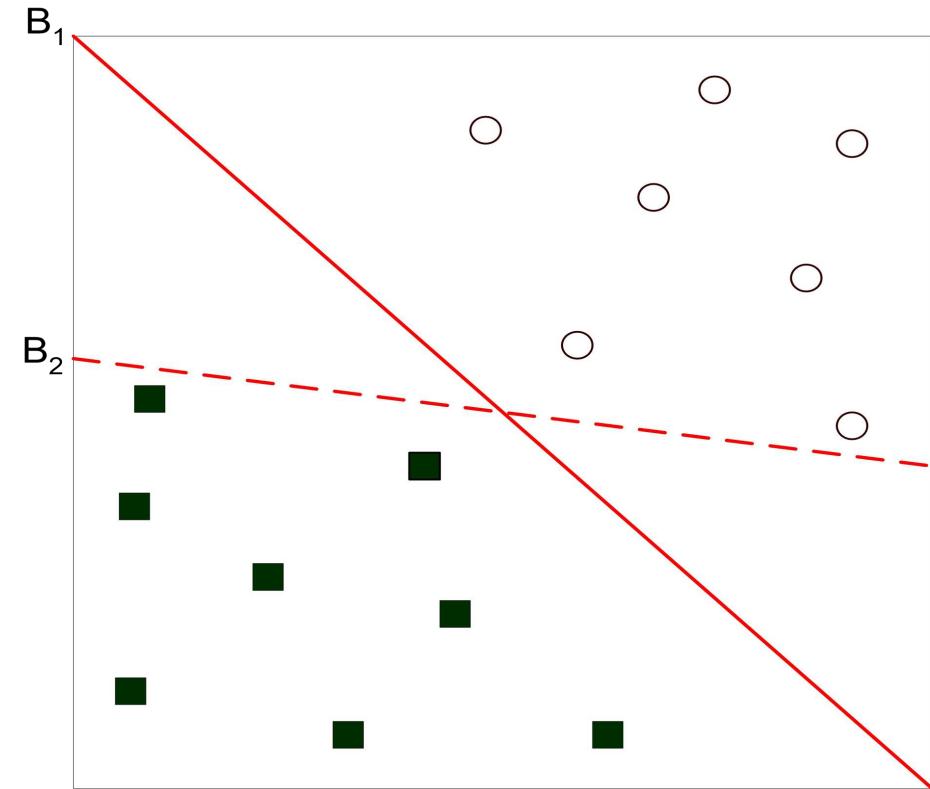
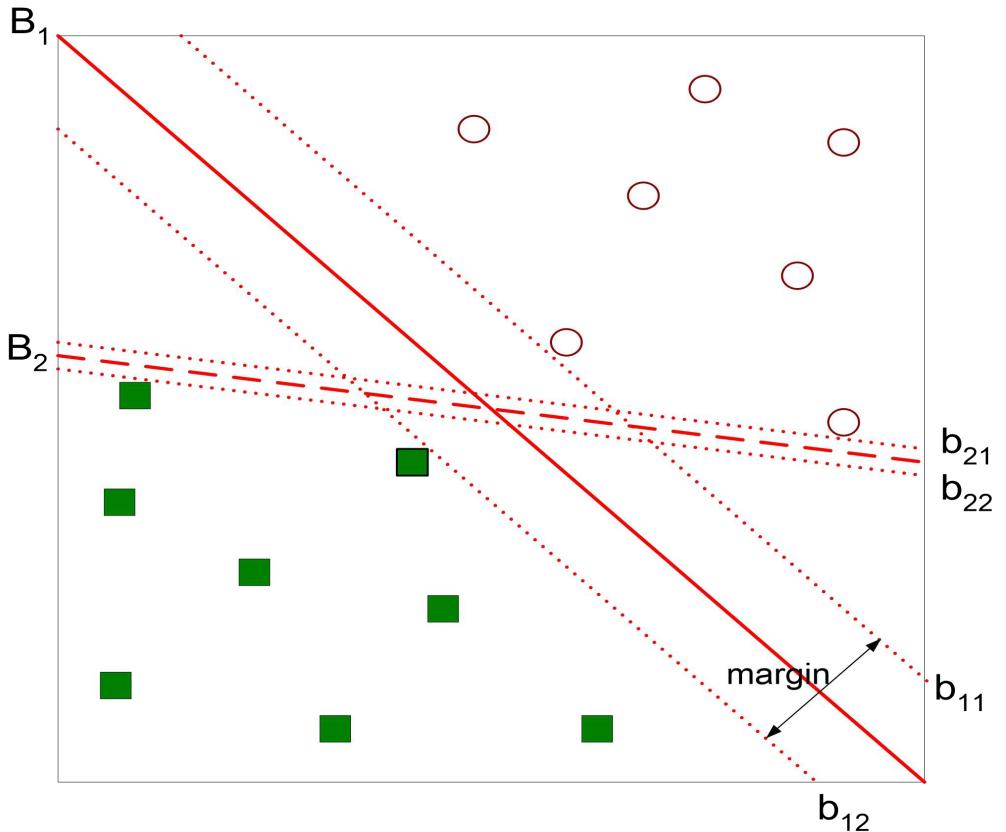
How can we find this best line? or how can we find the best hyperplane?

- There are an infinite number of separating lines that could be drawn. We want to find the “**best**” one, that is, one that will have the minimum classification error.



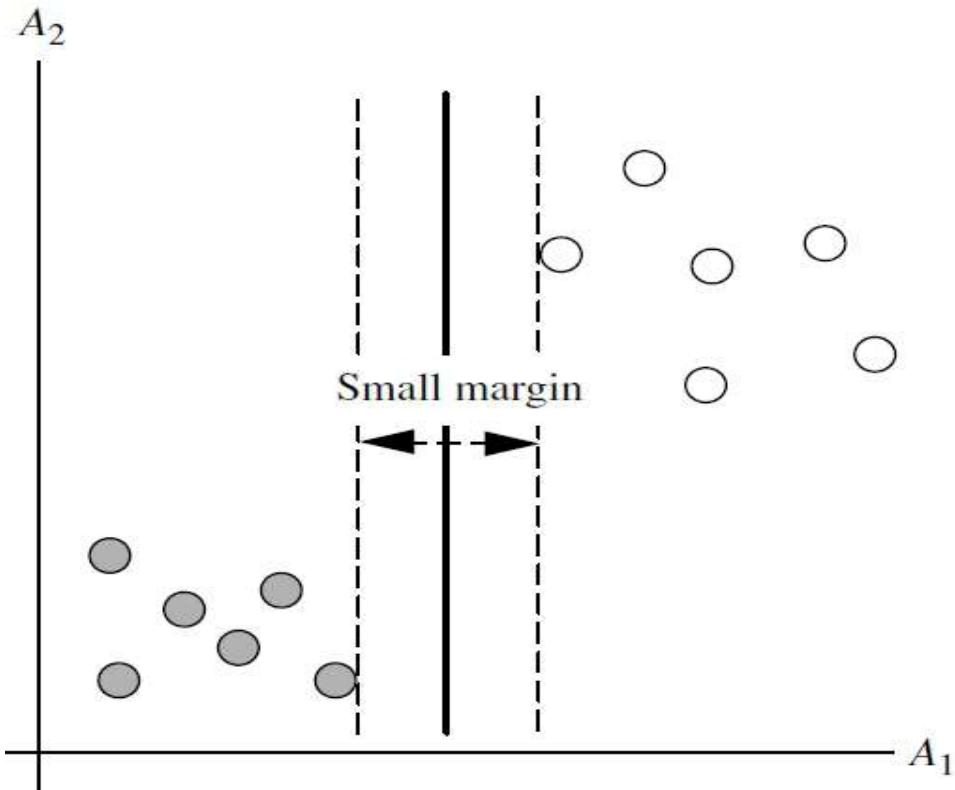
How can we find this best line?

- Which one is better? B1 or B2?
- How do you define better?



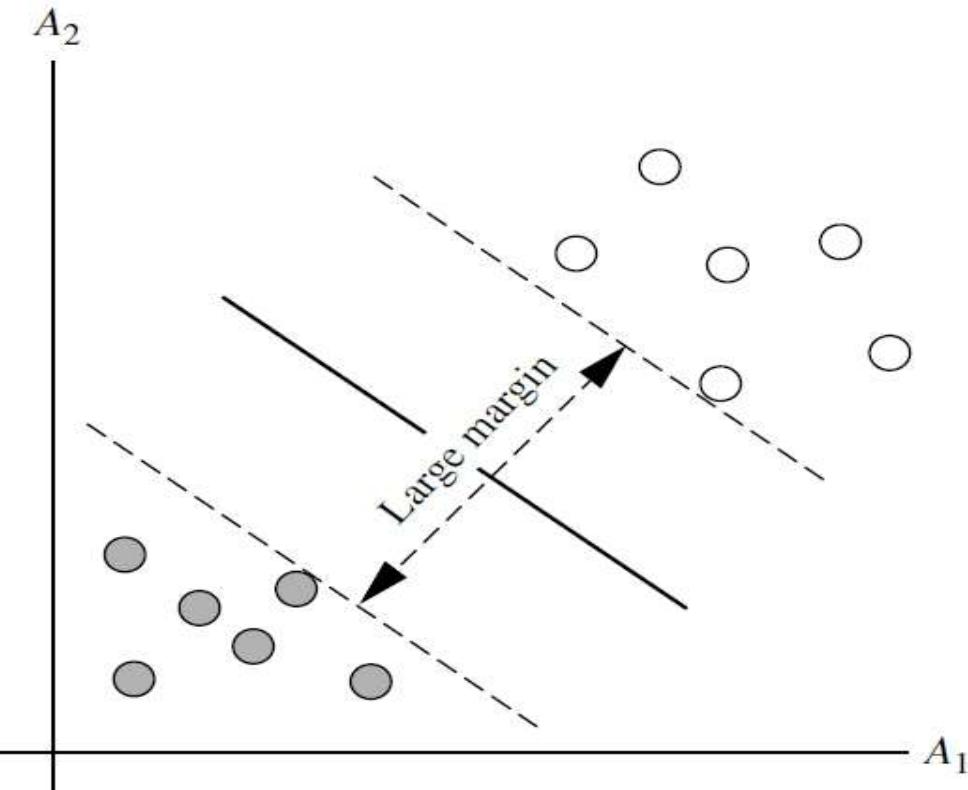
- Find hyperplane **maximizes** the margin => B1 is better than B2

Here we see just two possible separating hyperplanes and their associated margins. Which one is better? The one with the larger margin (b) should have greater generalization accuracy.



- Class 1, $y=+1$ (*buys_computer=yes*)
- Class 2, $y=-1$ (*buys_computer=no*)

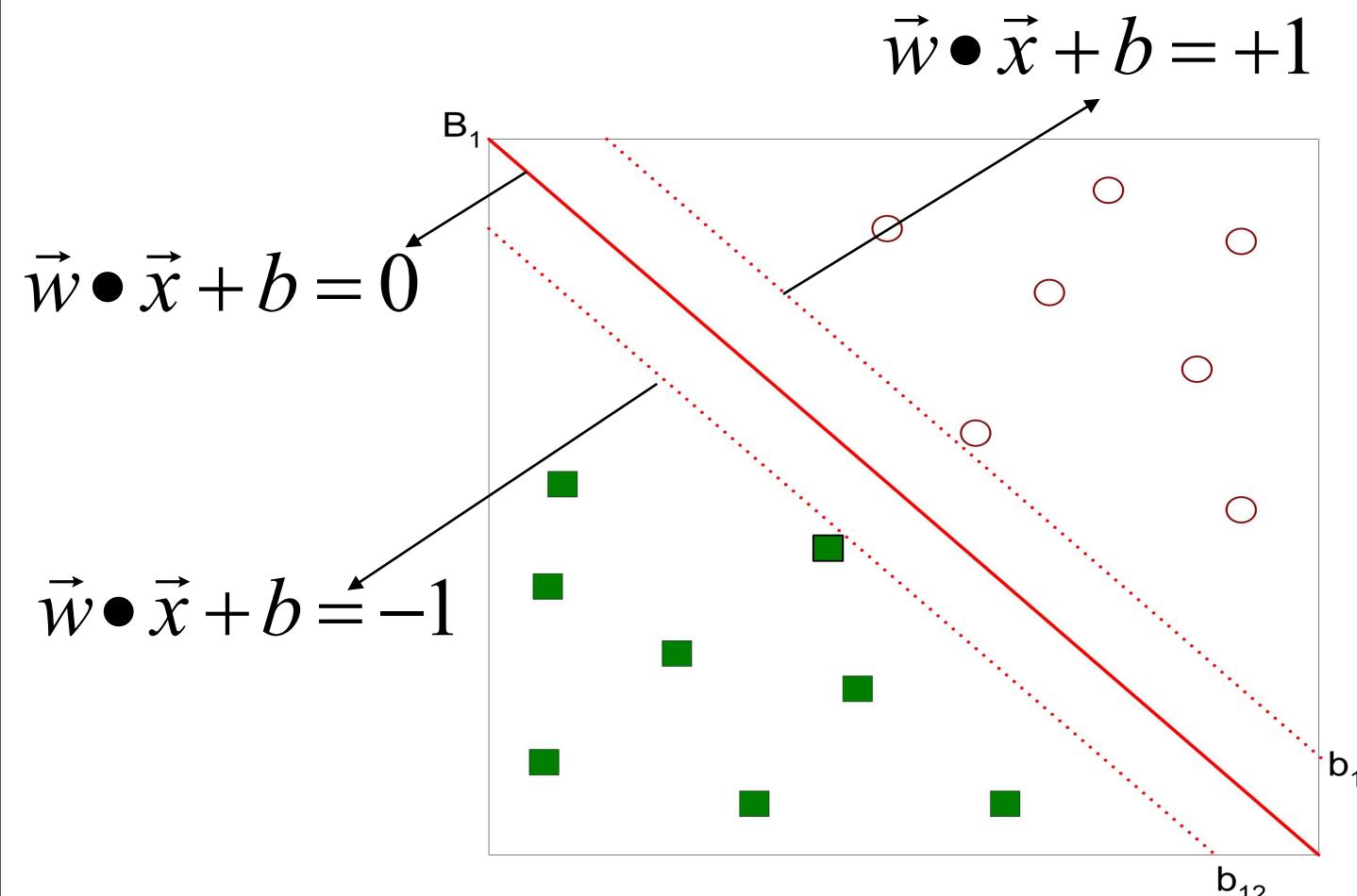
(a)



- Class 1, $y=+1$ (*buys_computer=yes*)
- Class 2, $y=-1$ (*buys_computer=no*)

(b)

Support Vector Machines



A separating hyperplane can be written as

$$\vec{w} \cdot \vec{x} + b = 0$$

Given: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

Find a 'good' Linear Discriminant which will work well for classifying future points.

We can write down a linear discriminant as an equation of a straight line.

$$\vec{W} = \{w_1, w_2, \dots, w_n\} \text{ & } \vec{X} = \{x_1, x_2, \dots, x_n\}$$

So, $w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0$. n is the number of attributes; and b is a scalar, often referred to as a bias.

So, x_1 is the x coordinate, x_2 is the y coordinate, w_1 & w_2 are the slopes.

We can give values of w_1 , w_2 , w_b and get different-different straight lines.

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \cdot \vec{x} + b \leq -1 \end{cases}$$

Support Vector Machines

- Then another way of writing this equation, $f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$

This is nothing but the transpose of the W vector into the X vector.

Find w, b such that:

$$w^T x + b > 0, \text{ if } y_i = +1$$

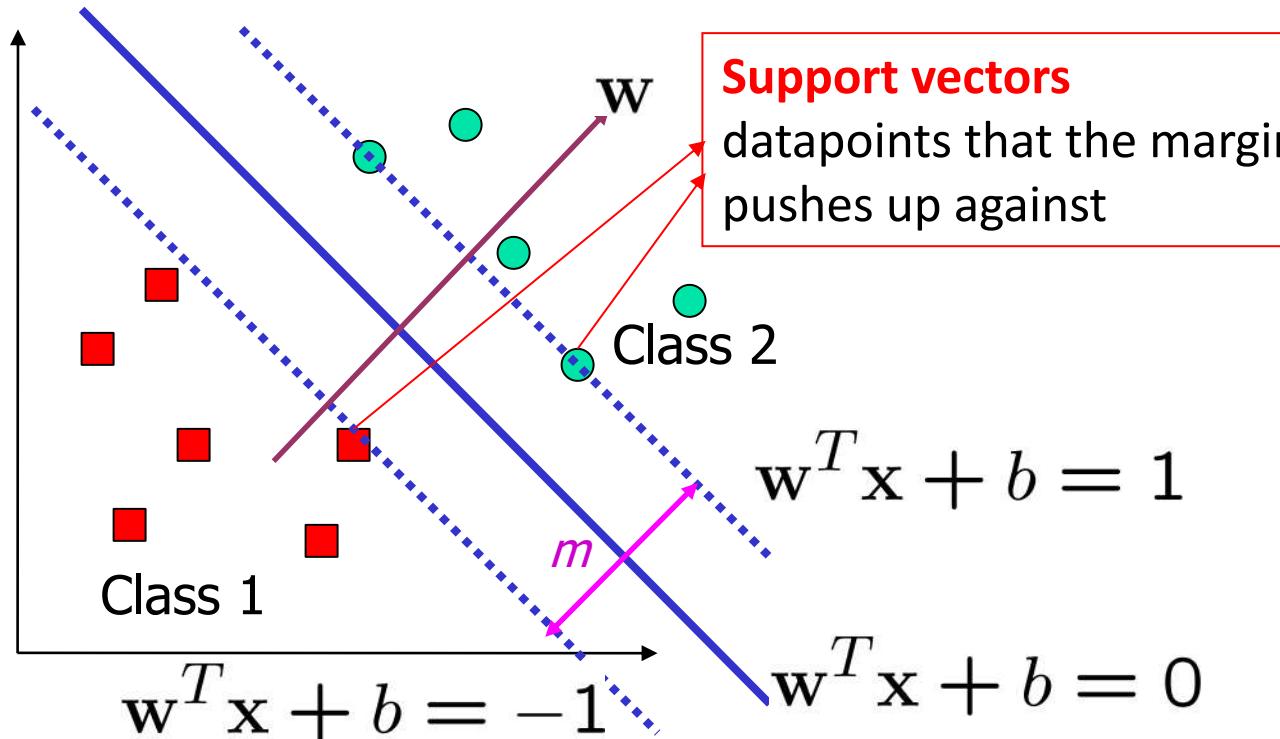
$$w^T x + b < 0, \text{ if } y_i = -1$$

So,

$$w^T x + b = 0$$

$$w^T x + b = 1$$

$$w^T x + b = -1$$



$$f(x_i) = \text{sign}(w^T x_i + b)$$

- So, the first condition will basically tell that all plus y are in non origin side and minus y are in origin side.

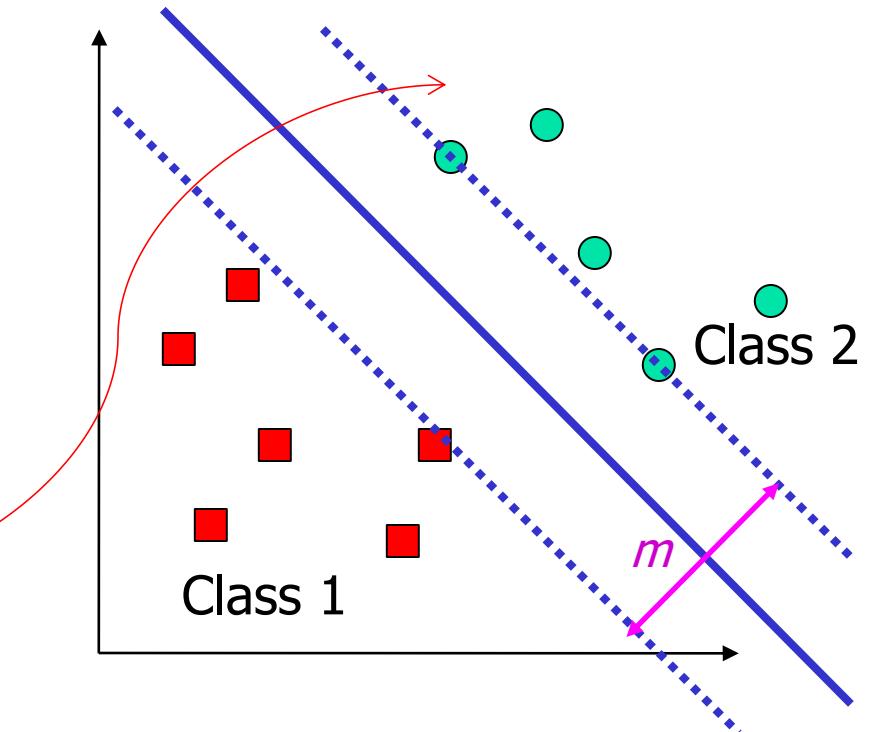
Find w,b such that:

$$w^T x_i + b > 0, \text{ if } y_i = +1 \quad (1)$$

$$w^T x_i + b < 0, \text{ if } y_i = -1 \quad (2)$$

So, Combining the two inequalities of Equations (1) & (2) we get,

$y_i(w^T x_i + b) \geq 1, \forall x_i. \text{ (Condition-1)}$

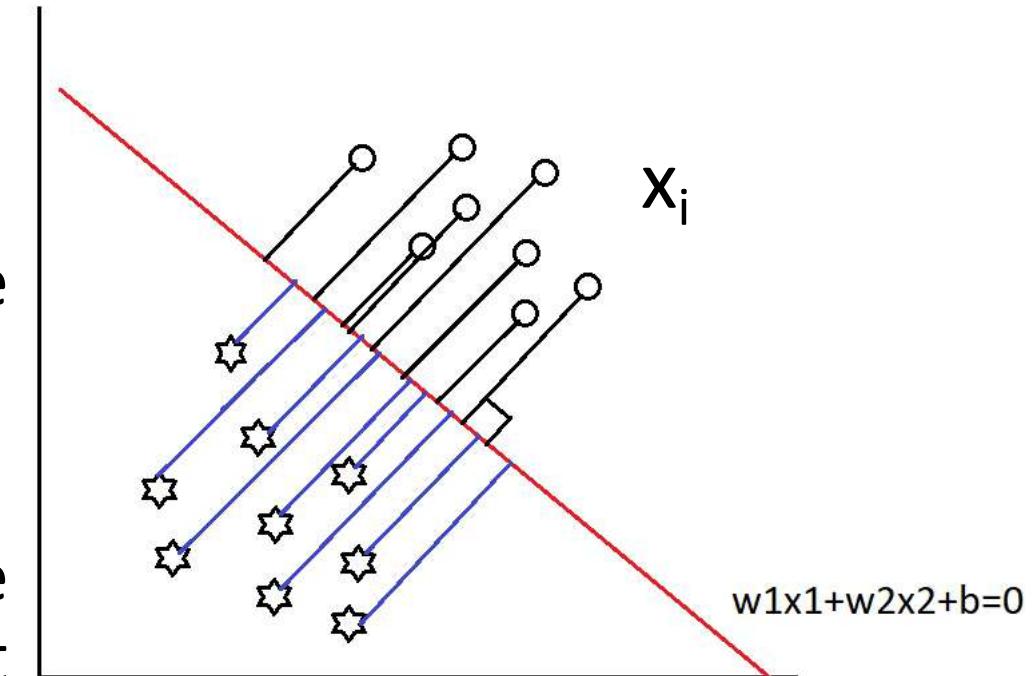


There are many lines which satisfy condition 1.

So, what we want to do is to sort of pick up one of this infinite line, which we consider is good. Let us see how do I pick up one among this infinite line.

Margin of Line

- Define a quantity called Margin of a line
- What is the margin of a line? You take the closest point to that line.
- How do you find out the closest point?
- Draw perpendicular from every point to the line. See which point has the closest smallest perpendicular distance, let us say this point has the perpendicular distance & is the margin of a line.



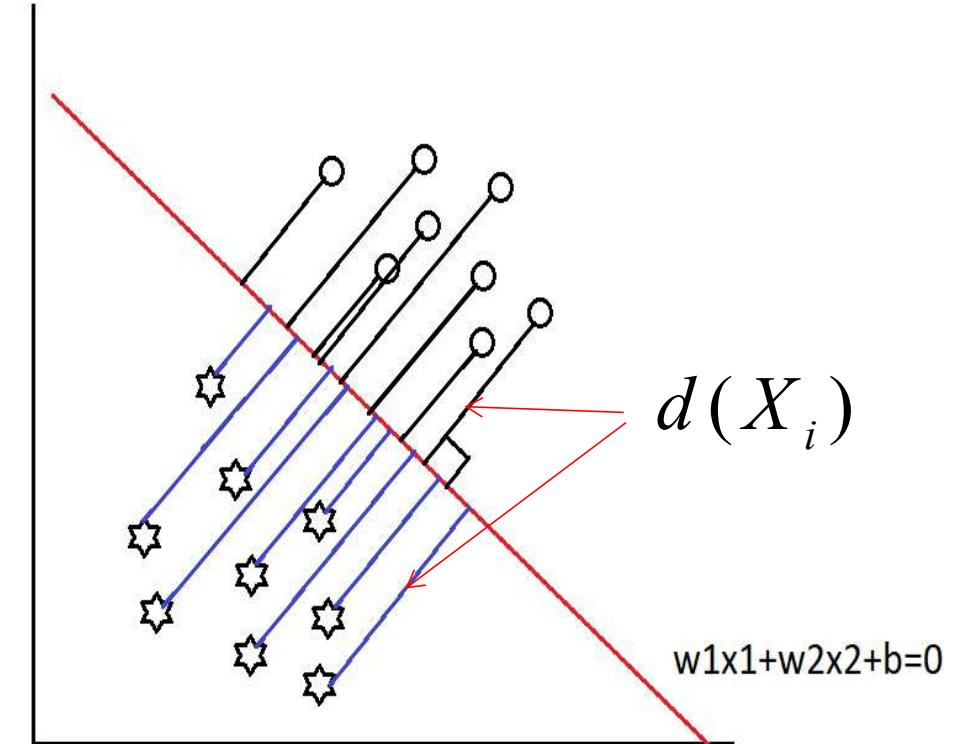
Perpendicular distance from a point to a line (*from coordinate geometry*)

$$d(X_i) = \frac{|w_1 \cdot x_{1i} + w_2 \cdot x_{2i} + b|}{\sqrt{w_1^2 + w_2^2}}$$

Margin of Line

$$d(X_i) = \frac{|w_1 \cdot x_{1i} + w_2 \cdot x_{2i} + b|}{\sqrt{w_1^2 + w_2^2}}$$

Where $W=[w_1, w_2]$ & $X_i=[x_{1i}, x_{2i}]$ in 2-Dimensional.
But in higher dimensional it will be w_3, w_4, \dots &
 x_{3i}, x_{4i}, \dots



Let, introduce again the vector notation for this that will be

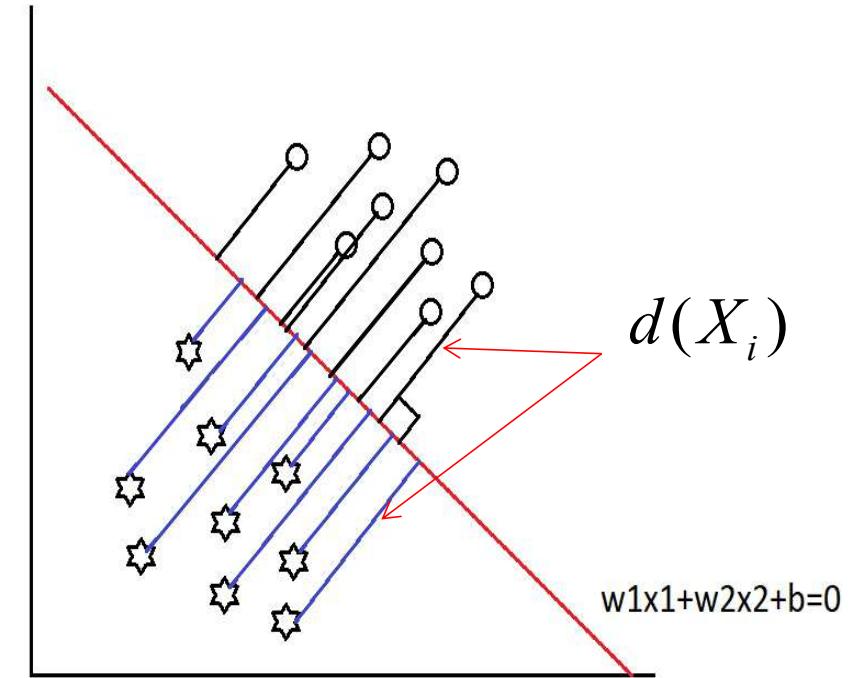
$$d(X_i) = \frac{|W^T X_i + b|}{\|W\|}$$

Norm of W

As per vector representation, $W \cdot W^T = \|W\|^2$

What is the margin?

- The margin now the smallest of $d(X_i)$ values the closest point the perpendicular the closest point you will have the smallest perpendicular distance.
- So, margin is nothing, but the minimum of these distances the smallest of these distances over all the X_i 's.



$$\text{Margin} = \min_{X_i} d(X_i) = \min_{X_i} \frac{|W^T X_i + b|}{\|W\|}$$

Choosing the best margin

- If you examine this 3 lines they have different values of w_1, w_2, b different, but they are really different lines. If you plot them they will turn out to be the same line they will have the same margin. NO
- So, which one should we take which value of w and b should we take.
- So, we will scale w_1, w_2 and b by multiplying by some factor. We will multiply w and b by a constant such that this smallest value turns out to be 1. So, that the smallest this value overall X_i 's becomes 1

$$\begin{aligned}2x_1 + 3x_2 + 4 &= 0 \\4x_1 + 6x_2 + 8 &= 0 \\6x_1 + 9x_2 + 12 &= 0\end{aligned}$$

$$\text{Margin} = \min_{X_i} d(X_i) = \min_{X_i} \frac{|W^T X_i + b|}{\|W\|} = 1$$

So, it can be rewrite as, Margin = $\frac{1}{\|W\|}$

Choosing the best margin

- Find w, b , such that, Maximum Margin = $\frac{1}{\|W\|}$ **(Objective)**
- It has to satisfy 2 conditions maximize margin and linearly separate.

So, $y_i(w^T x_i + b) > 0$ for all x_i

min=1

Such that, $y_i(w^T x_i + b) \geq 1$ for all $i = 1..N$ **(Constraint)**

So, this is called as optimization problem to get w and b , as maximise margin subject to separability.

- Maximise Margin = $\frac{1}{\|W\|}$

In fact, since norm of W is positive, we can actually minimize the square of norm of w.

can write down Minimise $\|W\|^2 = \left(\sqrt{W_1^2 + W_2^2} \right)^2 = W^T \cdot W = [w_1, w_2] \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

Such that, $y_i (w^T x_i + b) \geq 1$ for all $i = 1..N$

Primal Optimization Problem

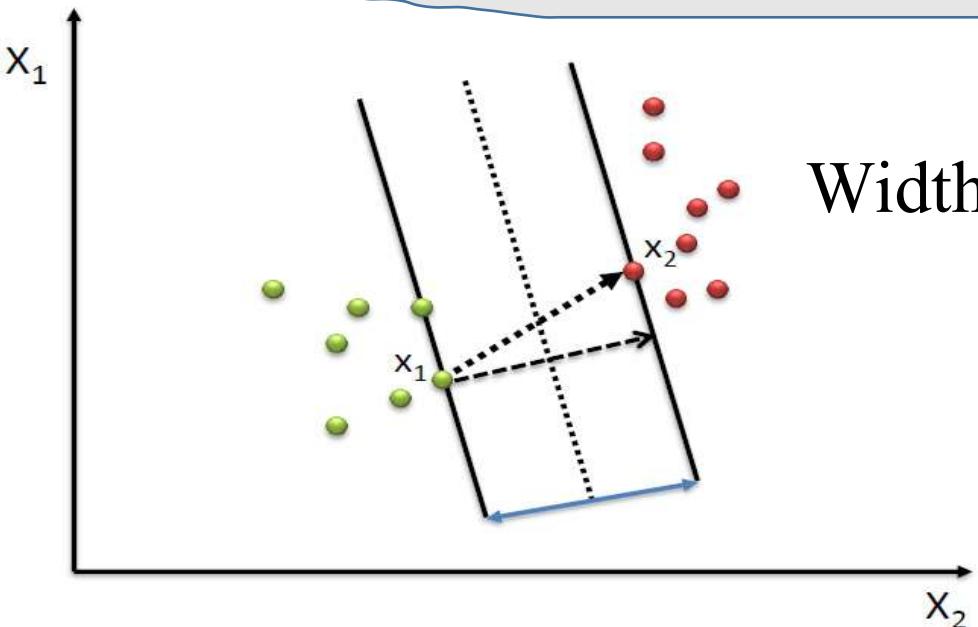
$$w \cdot x_2 + b = 1$$

$$w \cdot x_1 + b = -1$$

$$w \cdot x_2 + b - w \cdot x_1 - b = 1 - (-1)$$

$$w \cdot x_2 - w \cdot x_1 = 2$$

$$\frac{w}{\|w\|} (x_2 - x_1) = \frac{2}{\|w\|}$$



$$\text{Width (maximize)} = \frac{2}{\|W\|}$$

SVM – Optimization

- Learning the SVM can be formulated as an optimization:

$$\max_{\mathbf{w}} \frac{2}{\|\mathbf{w}\|} \text{ subject to } \mathbf{w}^\top \mathbf{x}_i + b \begin{cases} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{cases} \text{ for } i = 1 \dots N$$

- Or equivalently

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 \text{ subject to } y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \text{ for } i = 1 \dots N$$

- This is a quadratic optimization problem subject to linear constraints and there is a unique minimum

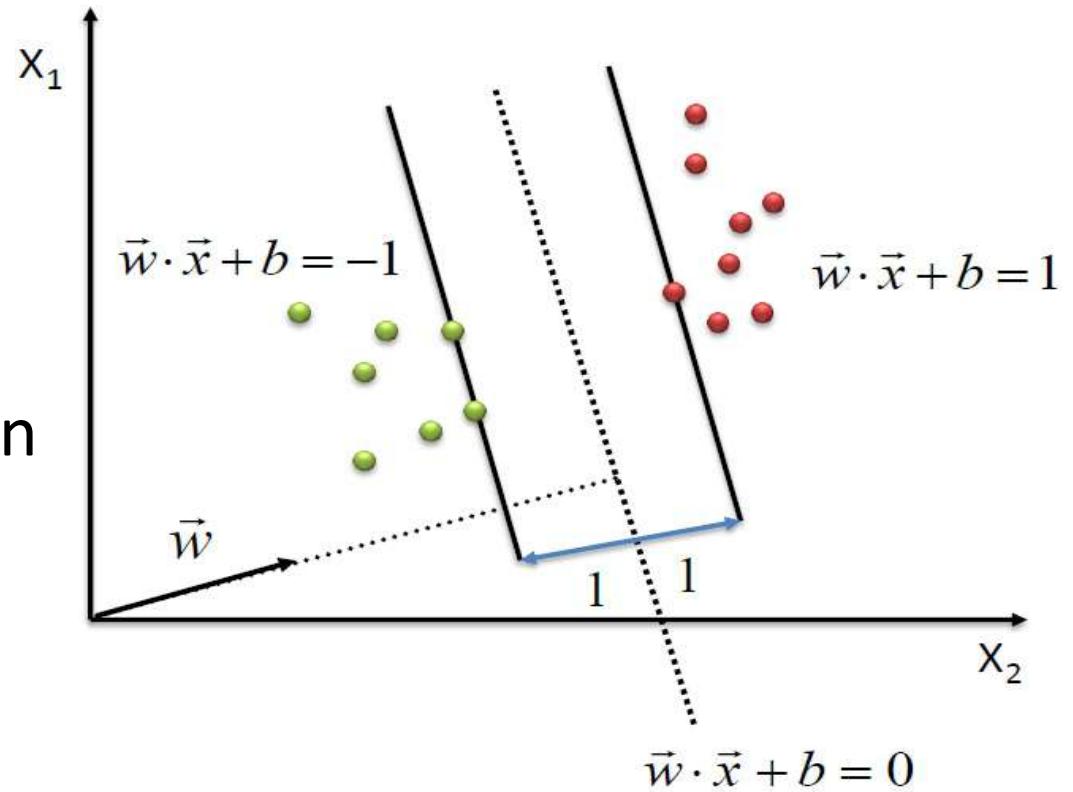
SVM as a Minimization Problem

- Hence SVM becomes a minimization problem

$$\min \frac{1}{2} \|\mathbf{W}\|^2$$

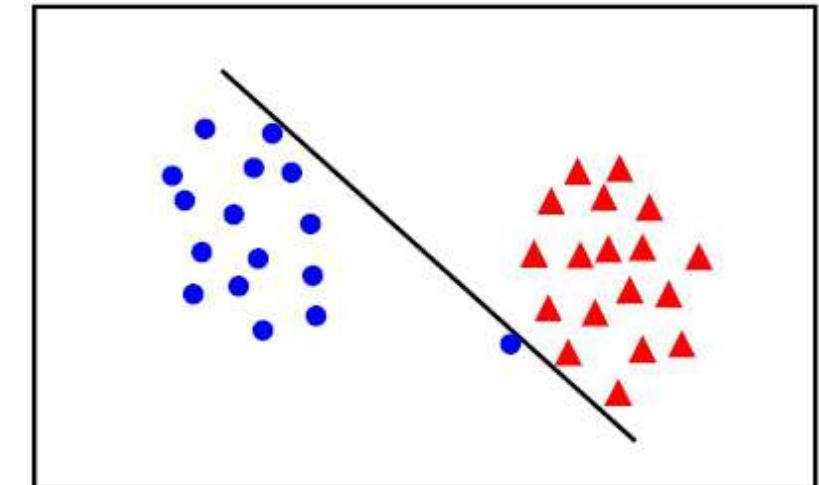
such that $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall x_i$

- We are now optimizing a quadratic function subject to linear constraints.

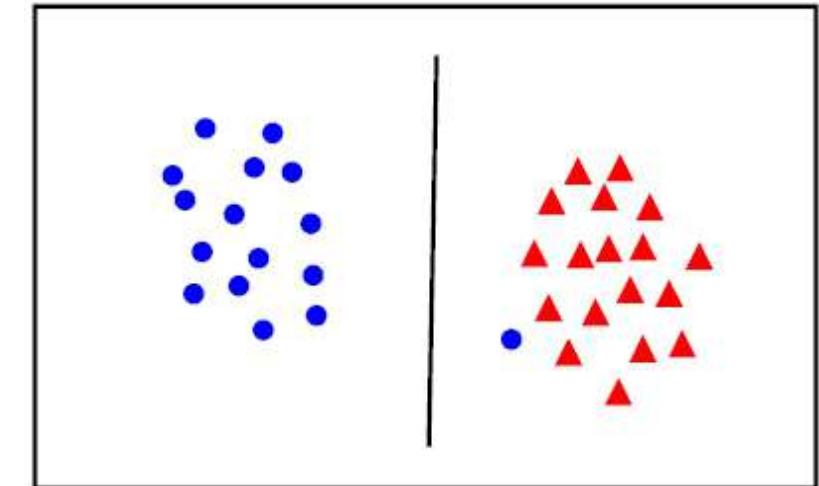


Linear separability again: What is the best w?

- the points can be linearly separated but there is a very narrow margin



- but possibly the large margin solution is better, even though one constraint is violated

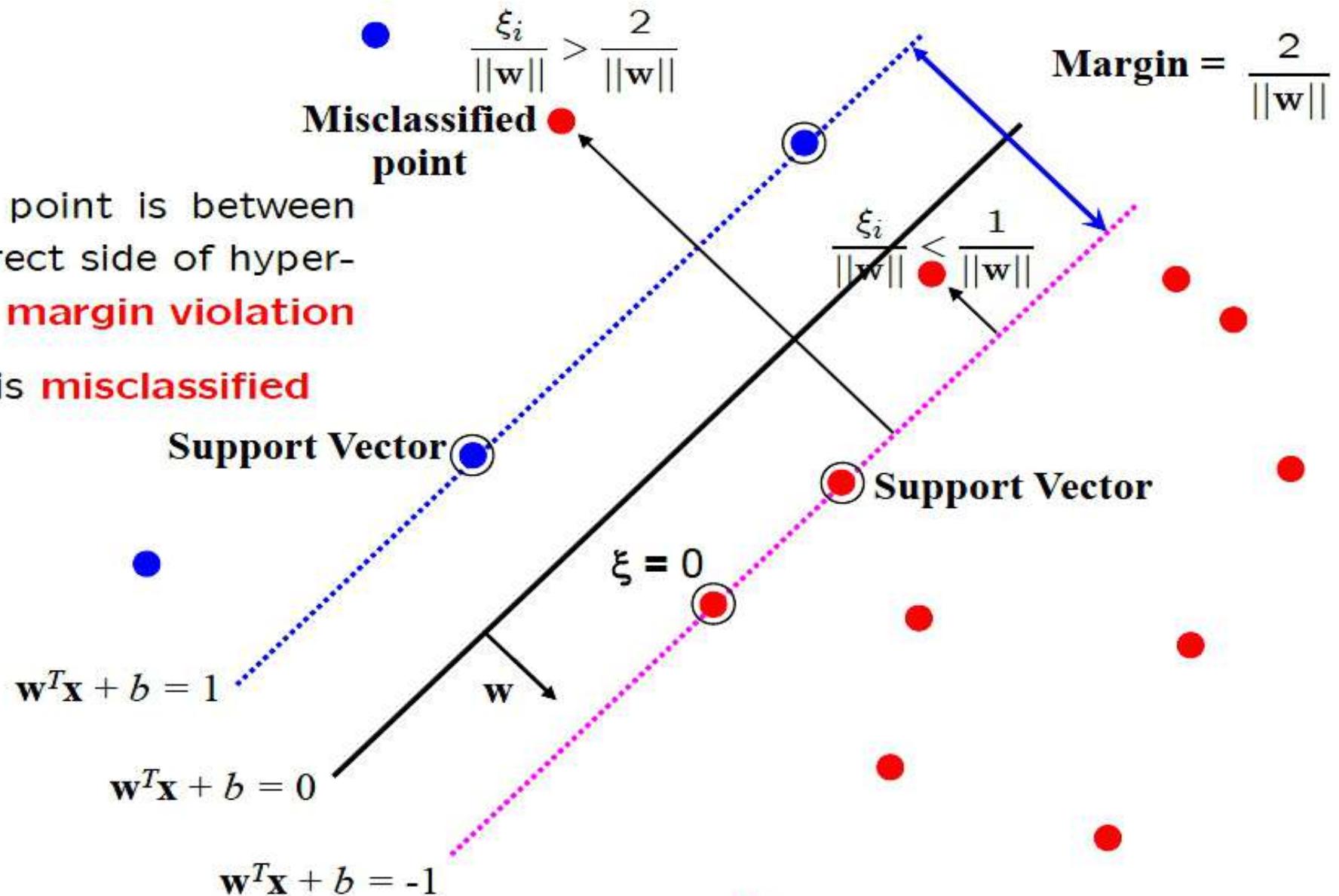


In general there is a trade off between the margin and the number of mistakes on the training data

Introduce “slack” variables

$$\xi_i \geq 0$$

- for $0 < \xi \leq 1$ point is between margin and correct side of hyperplane. This is a **margin violation**
- for $\xi > 1$ point is **misclassified**



Discriminating Hyper Plane

The hyper plane that discriminates the positive class from the negative class is given by: $\tilde{W} = \sum_i \alpha_i \tilde{X}_i$ the “supportiveness” parameters α_i

When solving SVM problems, there are some useful equations to keep in mind:

1. $\vec{w} \cdot \vec{x} + b = 0$ defines the boundary, and in particular $\vec{w} \cdot \vec{x} + b \geq 0$ defines the positive side of the boundary.
2. $\vec{w} \cdot \vec{x} + b = \pm 1$ defines the positive and negative gutters.
3. The distance between the gutters (the width of the margin) is
margin-width = $\frac{2}{\|\vec{w}\|} = \frac{2}{\sqrt{\vec{w} \cdot \vec{w}}}$
4. $\sum_{\text{training data}} y_i \alpha_i = 0$
5. $\sum_{\text{training data}} y_i \alpha_i \vec{x}_i = \vec{w}$

Support vector guideline



Support vector guideline 1: To see if a point \vec{x}_j is a support vector, imagine deleting it and see if you would draw a different SVM boundary. If you would draw a different SVM boundary, the point is a support vector ($\alpha_j > 0$). If you would draw the same boundary, even if the point were deleted, the point isn't a support vector ($\alpha_j = 0$).



Support vector guideline 2:

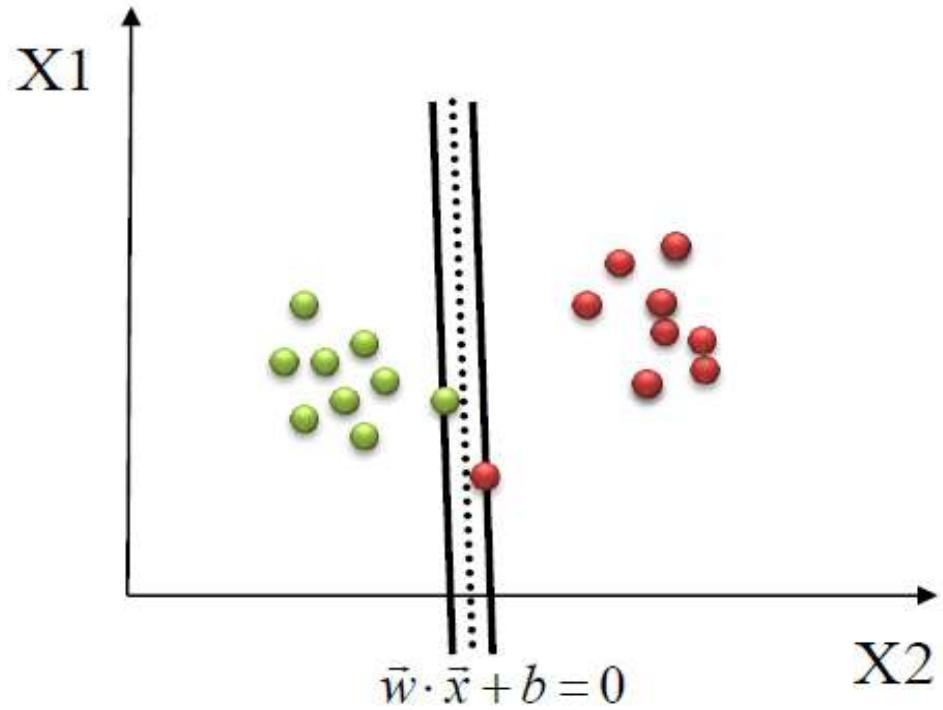
- Only points on the gutter can be support vectors; if a point isn't on the gutter, it's not a support vector.
- If a point is on the gutter, it *might or might not* be a support vector—you can determine whether it is using Support vector guideline 1.



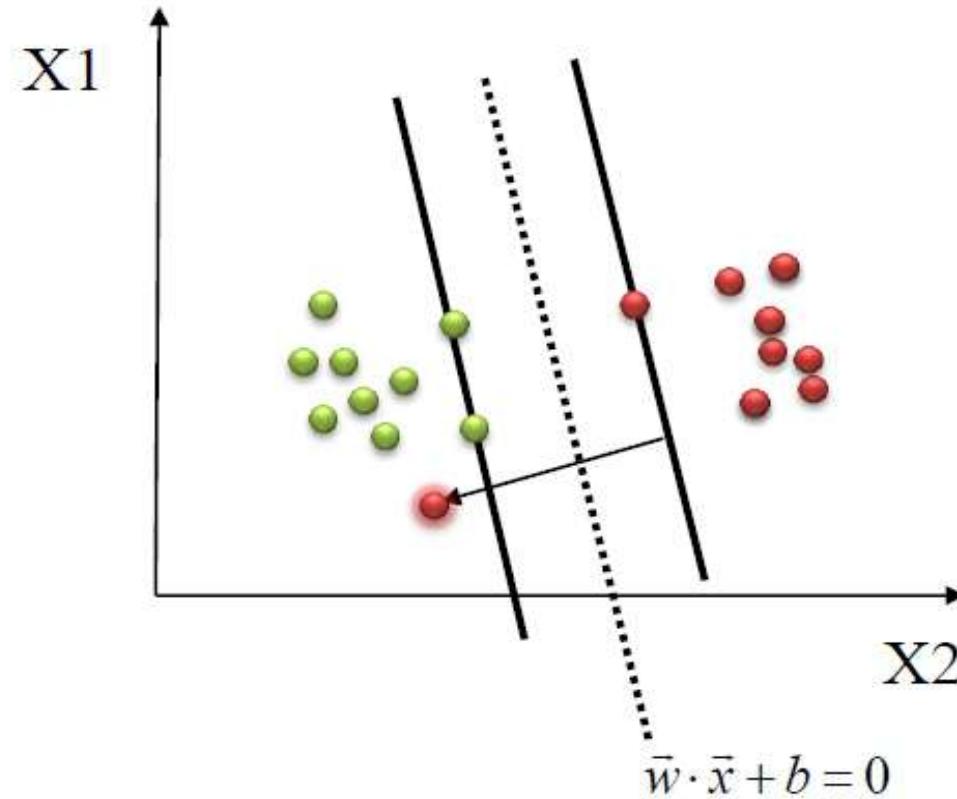
Support vector guideline 3: You may find it useful to think of the α_i as measuring “supportiveness”. This means, for example, that:

- α_i is zero for non-support vectors, i.e. training points that do not determine the boundary, and which would not affect the placement of the boundary if deleted.
- When you compare two separate SVM problems, where the first has support vectors that are far from the boundary, and the second has support vectors very close to the boundary, the latter support vectors have comparatively higher α_i values.

Soft vs Hard Margin SVMs



Hard Margin SVM



Soft Margin SVM

Formulating the Optimization Problem

Constraint becomes :

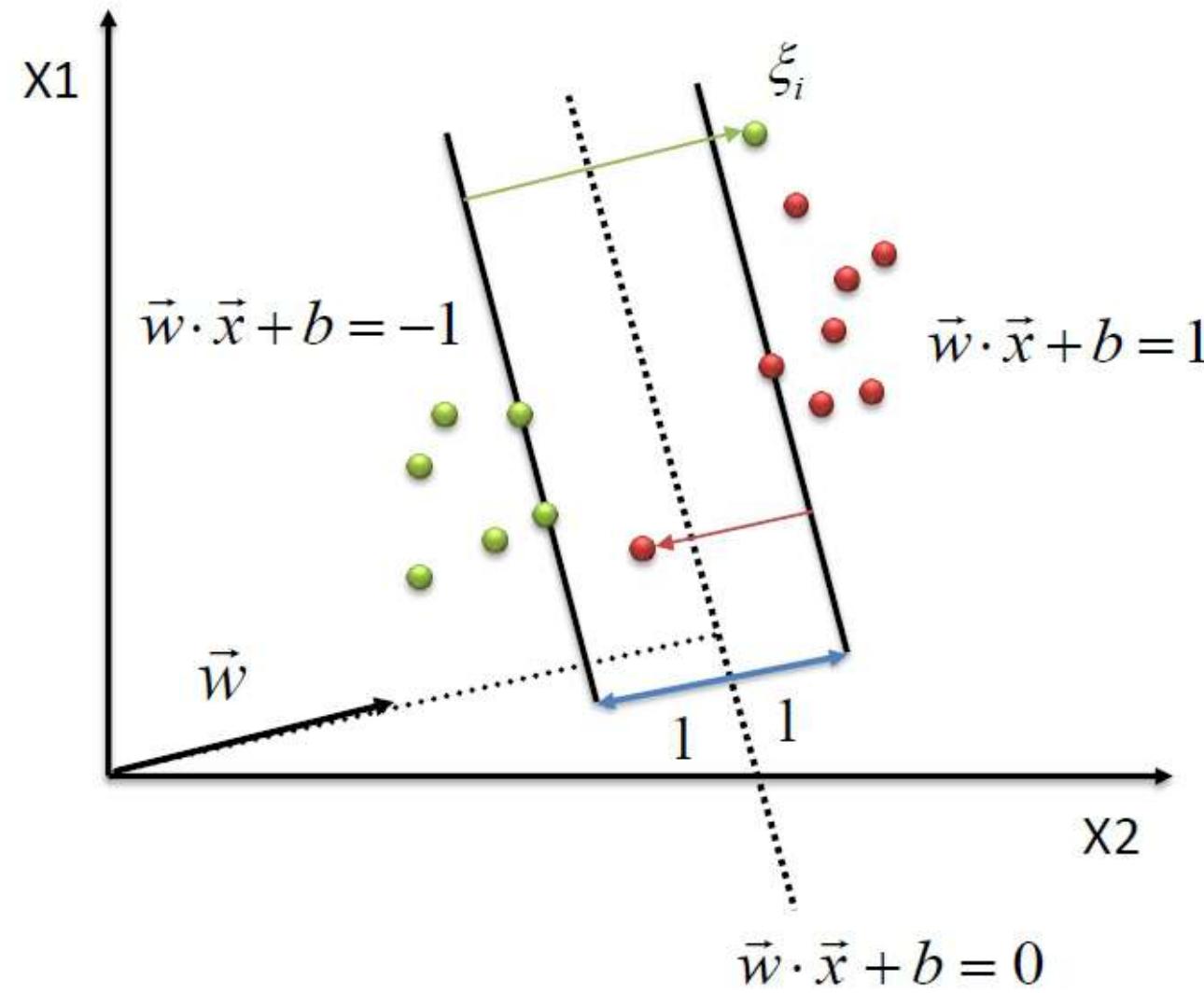
$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad \forall x_i \\ \xi_i \geq 0$$

Objective function

penalizes for misclassified instances and those within the margin

$$\min \frac{1}{2} \|\vec{w}\|^2 + C \sum_i \xi_i$$

C trades-off margin width and misclassifications



“Soft” margin solution

The optimization problem becomes

$$\min_{\mathbf{w} \in \mathbb{R}^d, \xi_i \in \mathbb{R}^+} \|\mathbf{w}\|^2 + C \sum_i^N \xi_i$$

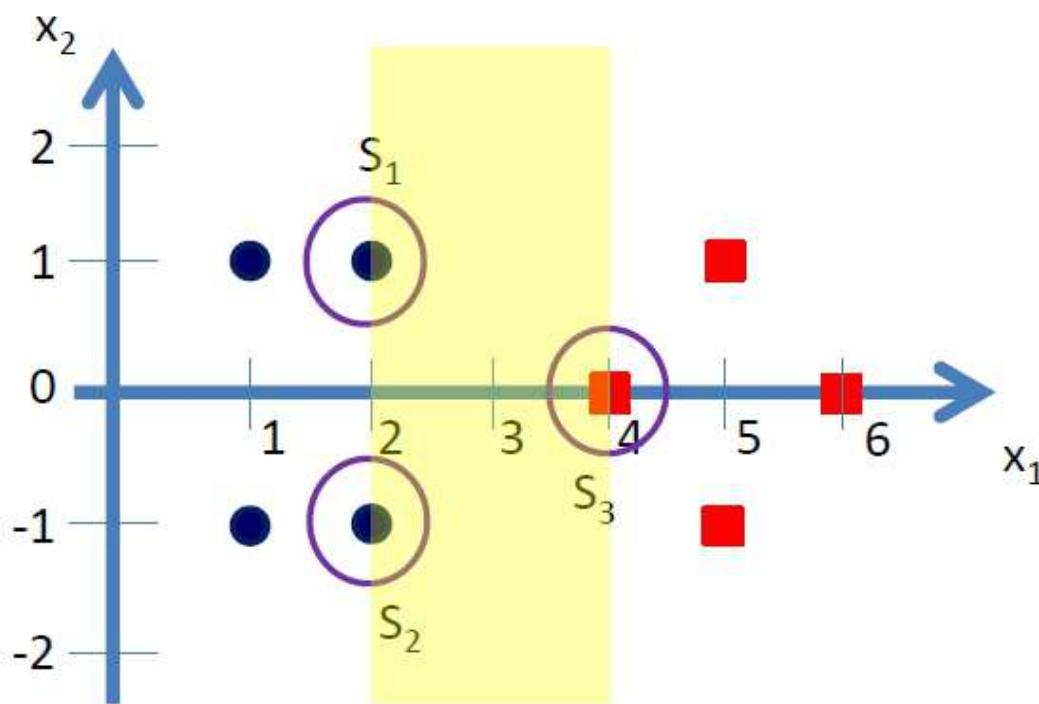
subject to

$$y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \text{ for } i = 1 \dots N$$

- Every constraint can be satisfied if ξ_i is sufficiently large
- C is a *regularization* parameter:
 - small C allows constraints to be easily ignored \rightarrow large margin
 - large C makes constraints hard to ignore \rightarrow narrow margin
 - $C = \infty$ enforces all constraints: hard margin
- This is still a quadratic optimization problem and there is a unique minimum. Note, there is only one parameter, C .

- Example

- Here we select 3 Support Vectors to start with.
- They are S_1, S_2 and S_3 .



$$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

- Here we will use vectors augmented with a 1 as a bias input, and for clarity we will differentiate these with an over-tilde.

That is:

$$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

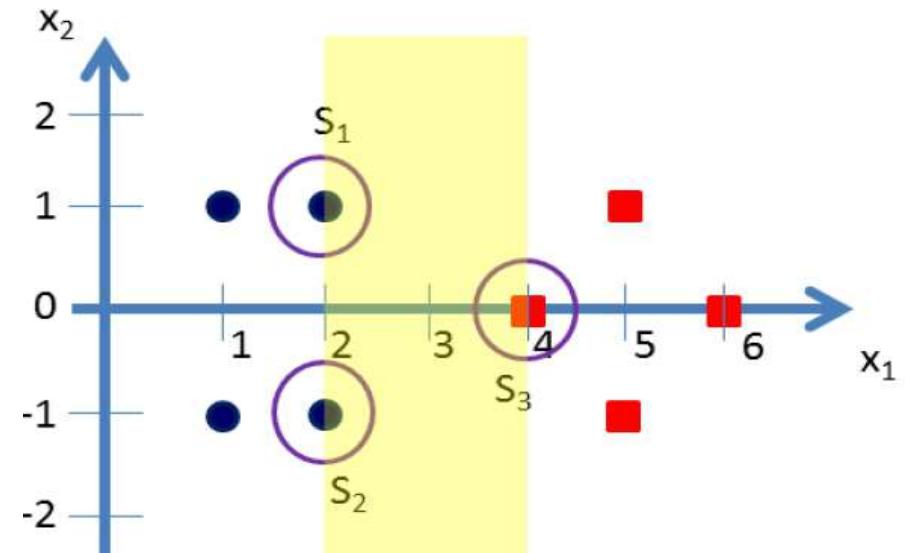
$$S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

$$\tilde{S}_1 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

$$\tilde{S}_2 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

$$\tilde{S}_3 = \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$



- Now we need to find 3 parameters α_1, α_2 , and α_3 based on the following 3 linear equations:

$$\alpha_1 \tilde{S}_1 \cdot \tilde{S}_1 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_1 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_1 = -1 \quad (-ve\ class)$$

$$\alpha_1 \tilde{S}_1 \cdot \tilde{S}_2 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_2 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_2 = -1 \quad (-ve\ class)$$

$$\alpha_1 \tilde{S}_1 \cdot \tilde{S}_3 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_3 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_3 = +1 \quad (+ve\ class)$$

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_1 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_1 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_1 = -1 \text{ } (-ve \text{ class})$$

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_2 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_2 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_2 = -1 \text{ } (-ve \text{ class})$$

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_3 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_3 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_3 = +1 \text{ } (+ve \text{ class})$$

- Let's substitute the values for \widetilde{S}_1 , \widetilde{S}_2 and \widetilde{S}_3 in the above equations.

$$\widetilde{S}_1 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \quad \widetilde{S}_2 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \quad \widetilde{S}_3 = \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = +1$$

- After simplification we get:

$$6\alpha_1 + 4\alpha_2 + 9\alpha_3 = -1$$

$$4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1$$

$$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = +1$$

- Simplifying the above 3 simultaneous equations we get: $\alpha_1 = \alpha_2 = -3.25$ and $\alpha_3 = 3.5$.

$\alpha_1 = \alpha_2 = -3.25$ and $\alpha_3 = 3.5$

$$\widetilde{S}_1 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

$$\widetilde{S}_2 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

$$\widetilde{S}_3 = \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

- The hyper plane that discriminates the positive class from the negative class is given by:

$$\tilde{w} = \sum_i \alpha_i \tilde{S}_i$$

- Substituting the values we get:

$$\tilde{w} = \alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

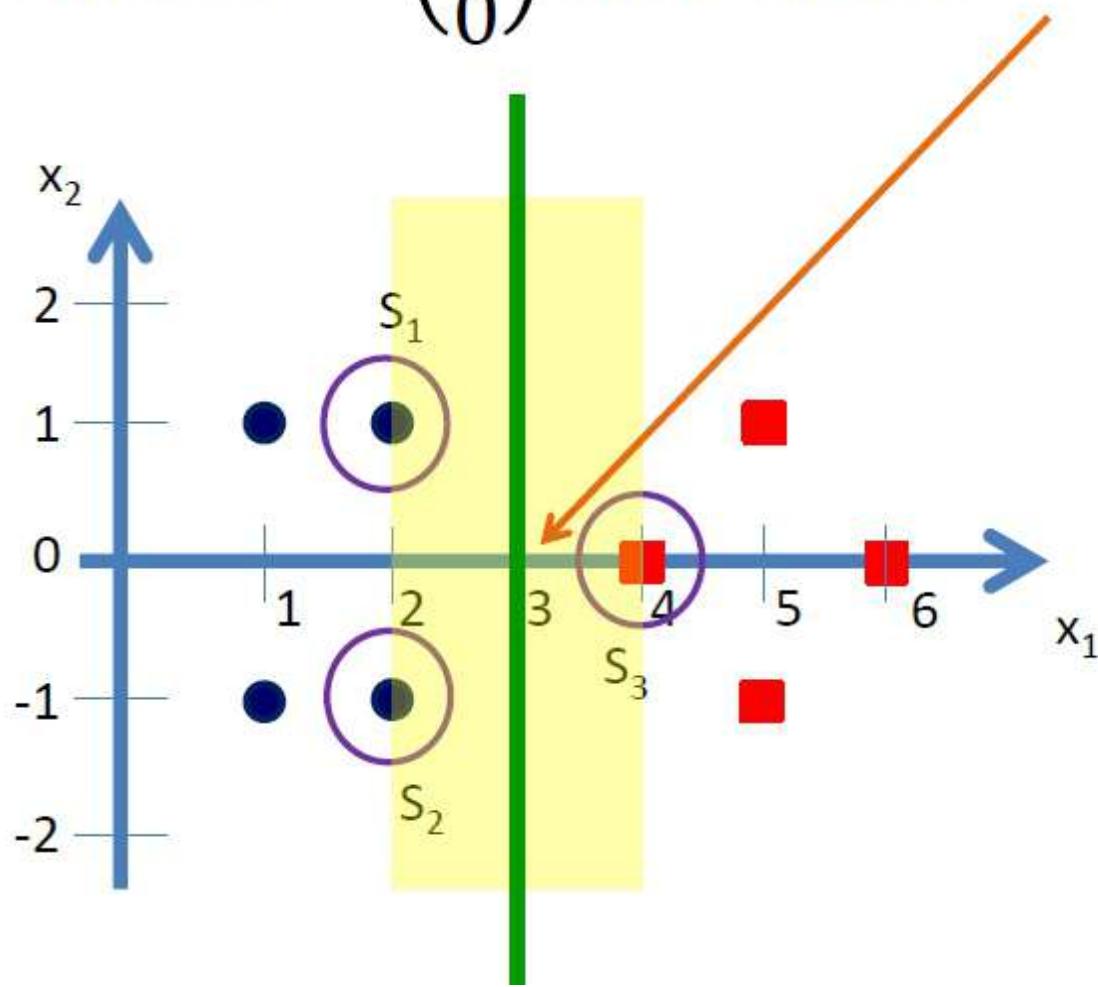
$$\tilde{w} = (-3.25) \cdot \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + (-3.25) \cdot \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + (3.5) \cdot \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$$

$$\tilde{w} = (-3.25) \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + (-3.25) \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + (3.5) \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$$

- Our vectors are augmented with a bias.
- Hence we can equate the entry in \tilde{w} as the hyper plane with an offset b .
- Therefore the separating hyper plane equation

$$y = wx + b \text{ with } w = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and offset } b = -3.$$

- $y = wx + b$ with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and offset $b = -3$.

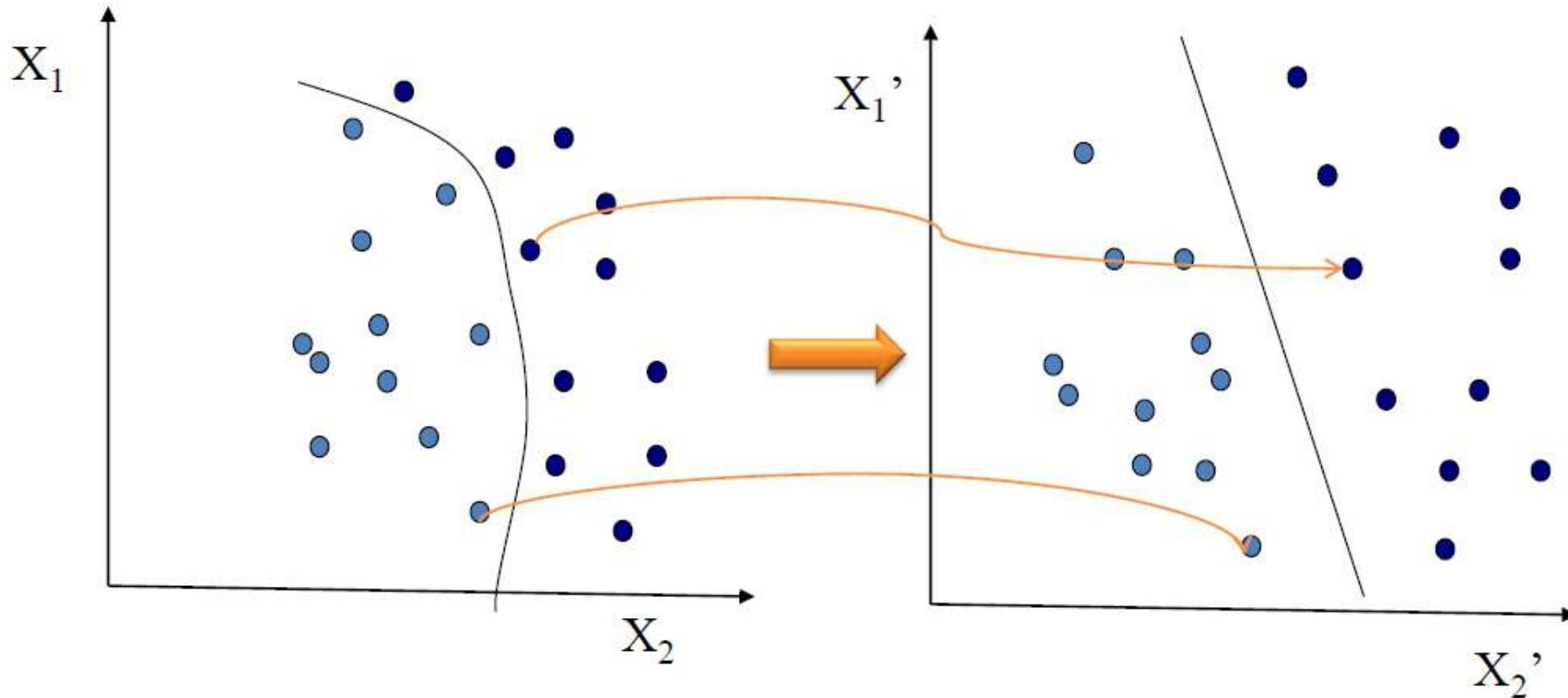


- This is the expected decision surface of the LSVM.

Non-Liner classification SVM

- The original optimal hyperplane algorithm proposed by Vladimir Vapnikin 1963 was a linear classifier.
- However, in 1992, Bernhard Boser, Isabelle Guyon and Vapnik suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes.
- The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function.
- This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space.
- The transformation may be non-linear and the transformed space high dimensional though the classifier is a hyperplane in the high-dimensional feature space, it may be non-linear in the original input space.

Linear Classifiers in High-Dimensional Spaces



Find function $\Phi(x)$ to map to a different space

Mapping Data to a High-Dimensional Space

- Find function $\Phi(x)$ to map to a different space, then SVM formulation becomes:

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \quad \begin{aligned} & s.t. \quad y_i(w \cdot \Phi(x) + b) \geq 1 - \xi_i, \forall x_i \\ & \xi_i \geq 0 \end{aligned}$$

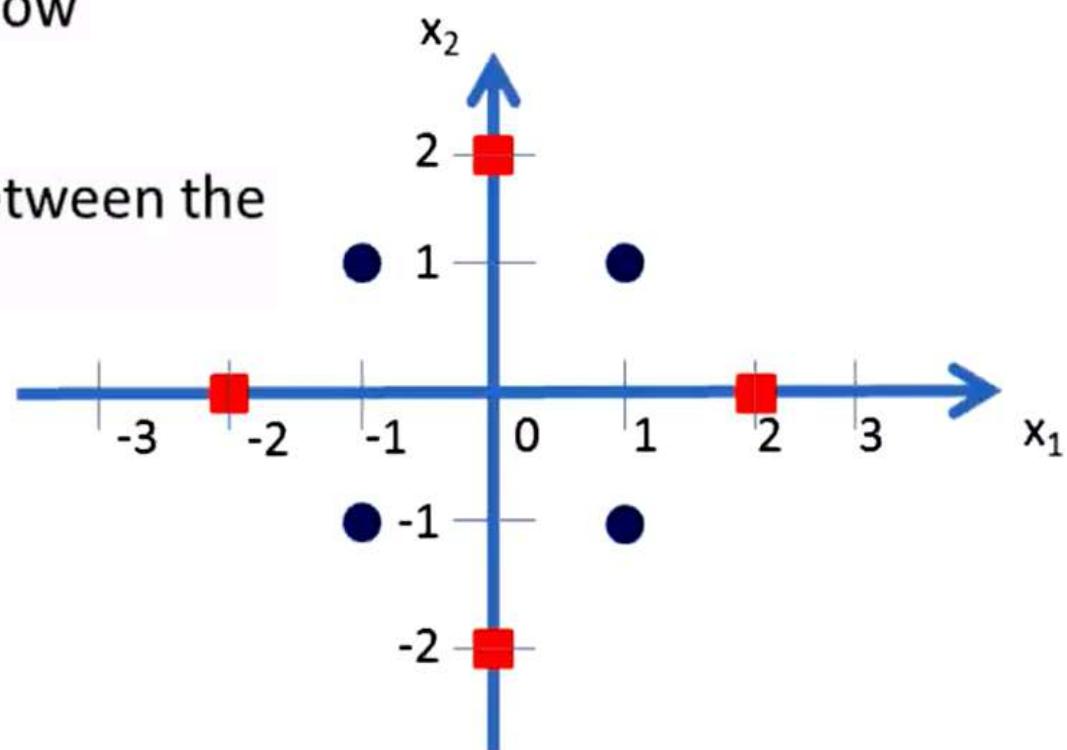
- Data appear as $\Phi(x)$, weights w are now weights in the new space.
- Explicit mapping expensive if $\Phi(x)$ is very high dimensional.
- Solving the problem without explicitly mapping the data is desirable.

Non-Liner classification SVM

We looked at the linear type SVM

Here we will look at an example like the one given below and find out how to carry out the classification.

Obviously there is no clear separating hyperplane between the red class and the blue class.



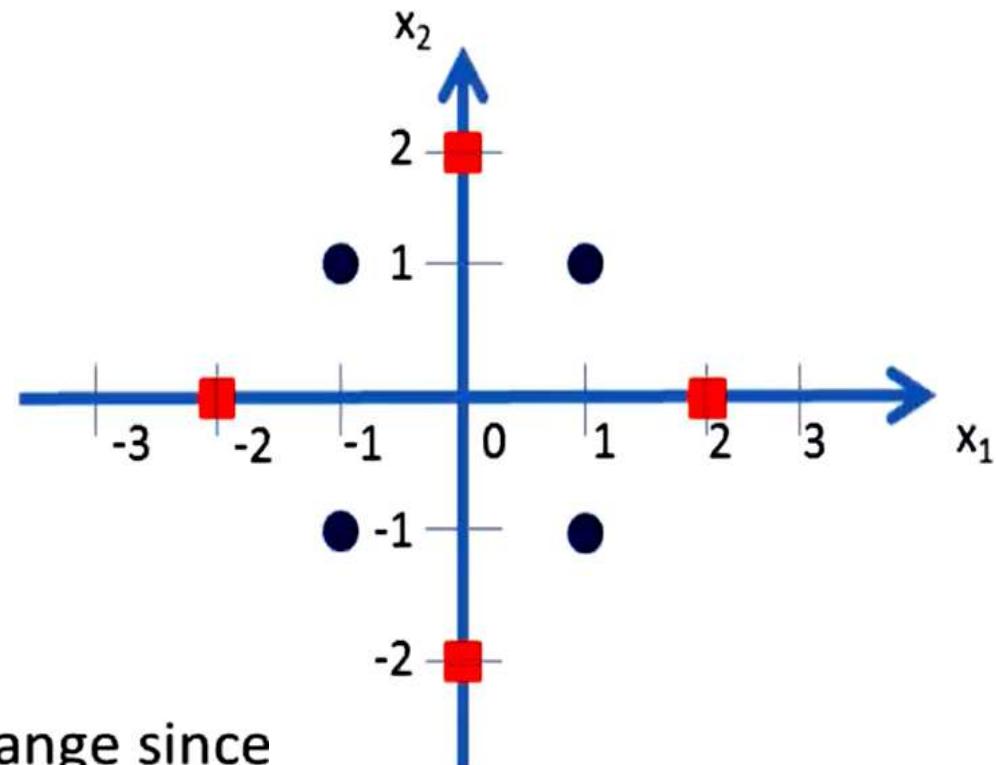
- Blue class vectors are: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- Red class vectors are: $\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -2 \end{pmatrix}$

Non-Liner classification SVM

- Here we need to find a non-linear mapping function Φ which can transform these data in to a new feature space where a separating hyperplane can be found.
- Let us consider the following mapping function.

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 6 - x_1 + (x_1 - x_2)^2 \\ 6 - x_2 + (x_1 - x_2)^2 \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} \geq 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

Blue class vectors are: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ no change since $\sqrt{x_1^2 + x_2^2} < 2$ for all the vectors



Non-Liner classification SVM

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 6 - x_1 + (x_1 - x_2)^2 \\ 6 - x_2 + (x_1 - x_2)^2 \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} \geq 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

Let us take Red class vectors : $\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -2 \end{pmatrix}$

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 6 - 2 + (2 - 0)^2 \\ 6 - 0 + (2 - 0)^2 \end{pmatrix} = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$$

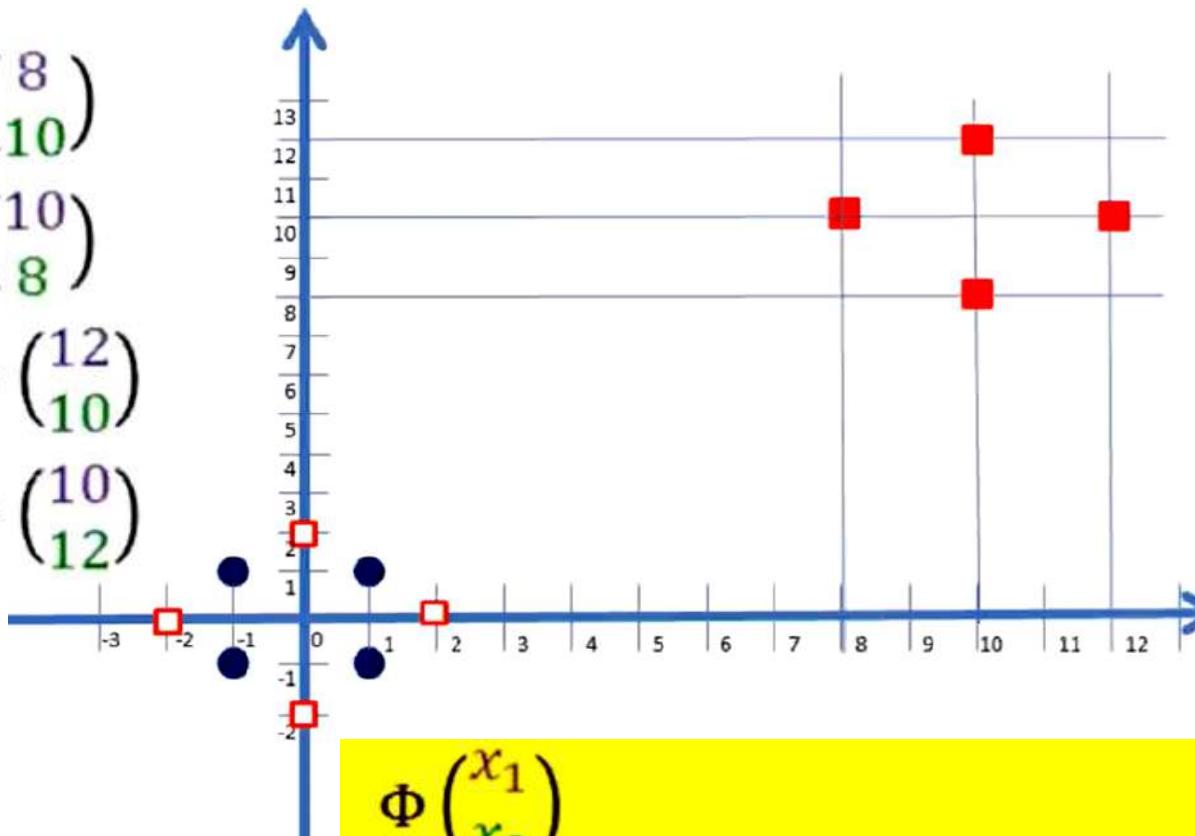
$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 - 0 + (0 - 2)^2 \\ 6 - 2 + (0 - 2)^2 \end{pmatrix} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$$

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} -2 \\ 0 \end{pmatrix} = \begin{pmatrix} 6 + 2 + (-2 - 0)^2 \\ 6 - 0 + (-2 - 0)^2 \end{pmatrix} = \begin{pmatrix} 12 \\ 10 \end{pmatrix}$$

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 6 - 0 + (0 + 2)^2 \\ 6 + 2 + (0 + 2)^2 \end{pmatrix} = \begin{pmatrix} 10 \\ 12 \end{pmatrix}$$

Non-Liner classification SVM

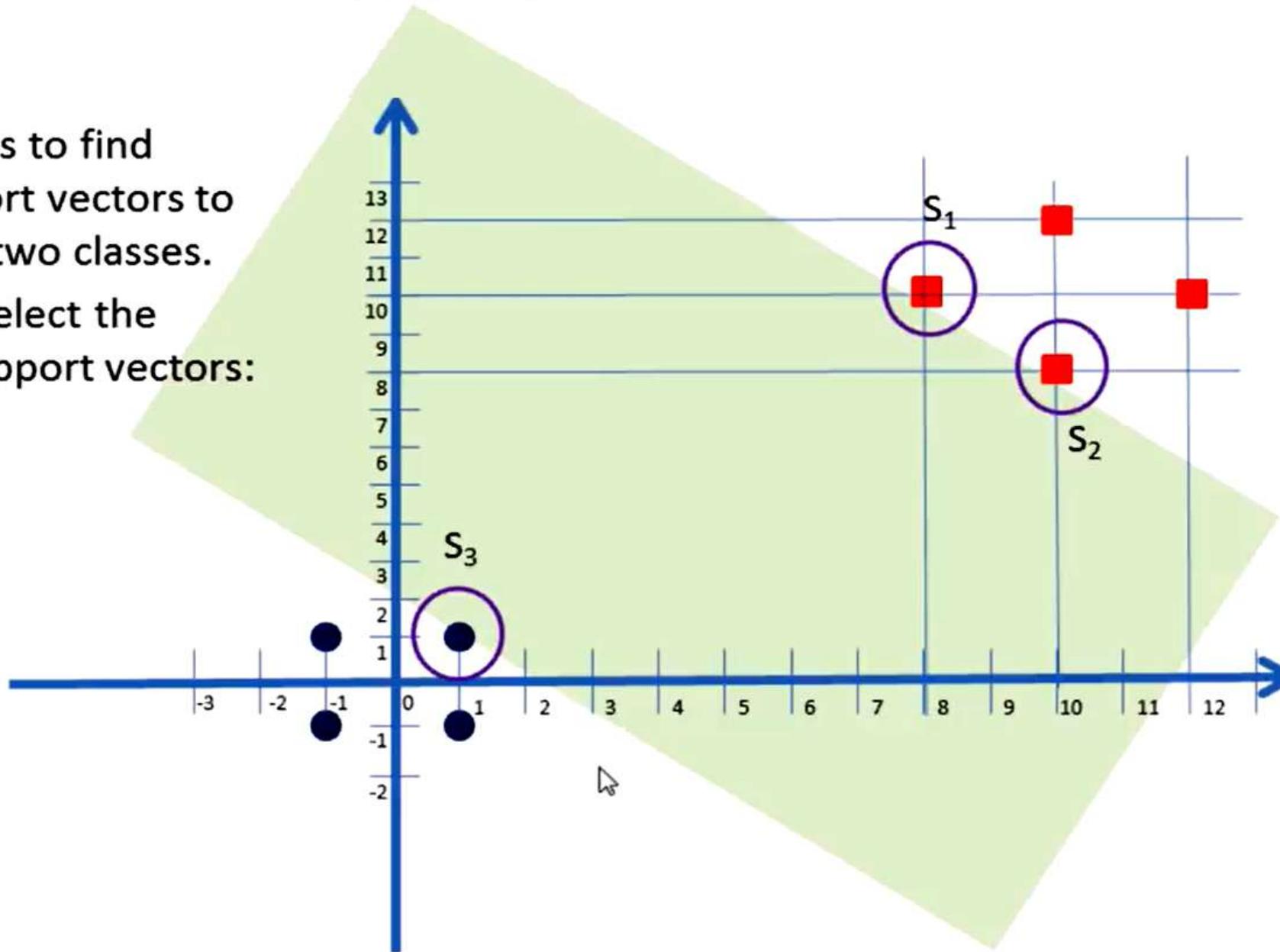
- $\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$
- $\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$
- $\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} -2 \\ 0 \end{pmatrix} = \begin{pmatrix} 12 \\ 10 \end{pmatrix}$
- $\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Phi \begin{pmatrix} 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 10 \\ 12 \end{pmatrix}$



$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 6 - x_1 + (x_1 - x_2)^2 \\ 6 - x_2 + (x_1 - x_2)^2 \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} \geq 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

Non-Liner classification SVM

- Now our task is to find suitable support vectors to classify these two classes.
- Here we will select the following 3 support vectors:
 - $s_1 = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$,
 - $s_2 = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$,
 - and $s_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$



Augmented Support Vector

- Here we will use vectors augmented with a 1 as a bias input, and for clarity we will differentiate these with an over-tilde.

That is:

$$S_1 = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\widetilde{S}_1 = \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix}$$

$$\widetilde{S}_2 = \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix}$$

$$\widetilde{S}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Finding Parameters

- Now we need to find 3 parameters α_1, α_2 , and α_3 based on the following 3 linear equations:

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_1 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_1 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_1 = +1 \quad (+ve\ class)$$

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_2 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_2 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_2 = +1 \quad (+ve\ class)$$

$$\alpha_1 \widetilde{S}_1 \cdot \widetilde{S}_3 + \alpha_2 \widetilde{S}_2 \cdot \widetilde{S}_3 + \alpha_3 \widetilde{S}_3 \cdot \widetilde{S}_3 = -1 \quad (-ve\ class)$$

- Let's substitute the values for \tilde{S}_1 , \tilde{S}_2 and \tilde{S}_3 in the above equations.

$$\tilde{S}_1 = \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \quad \tilde{S}_2 = \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \quad \tilde{S}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -1$$

Simultaneous Equations

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -1$$

- After multiplication we get:

$$165 \alpha_1 + 161 \alpha_2 + 19 \alpha_3 = +1$$

$$161 \alpha_1 + 165 \alpha_2 + 19 \alpha_3 = +1$$

$$19 \alpha_1 + 19 \alpha_2 + 3 \alpha_3 = -1$$

- Simplifying the above 3 simultaneous equations we get: $\alpha_1 = \alpha_2 = 0.859$ and $\alpha_3 = -1.4219$.

Discriminating Hyper Plane

- The hyper plane that discriminates the positive class from the negative class is given by:

$$\tilde{w} = \sum_i \alpha_i \tilde{s}_i$$

- Substituting the values we get:

$$\tilde{w} = \alpha_1 \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\tilde{w} = (0.0859) \cdot \begin{pmatrix} 8 \\ 10 \\ 1 \end{pmatrix} + (0.0859) \cdot \begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix} + (-1.4219) \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.1243 \\ 0.1243 \\ -1.2501 \end{pmatrix}$$

Equation for Discriminating Hyper Plane

- Our vectors are augmented with a bias.
- Hence we can equate the entry in \tilde{w} as the hyper plane with an offset b .
- Therefore the separating hyper plane equation

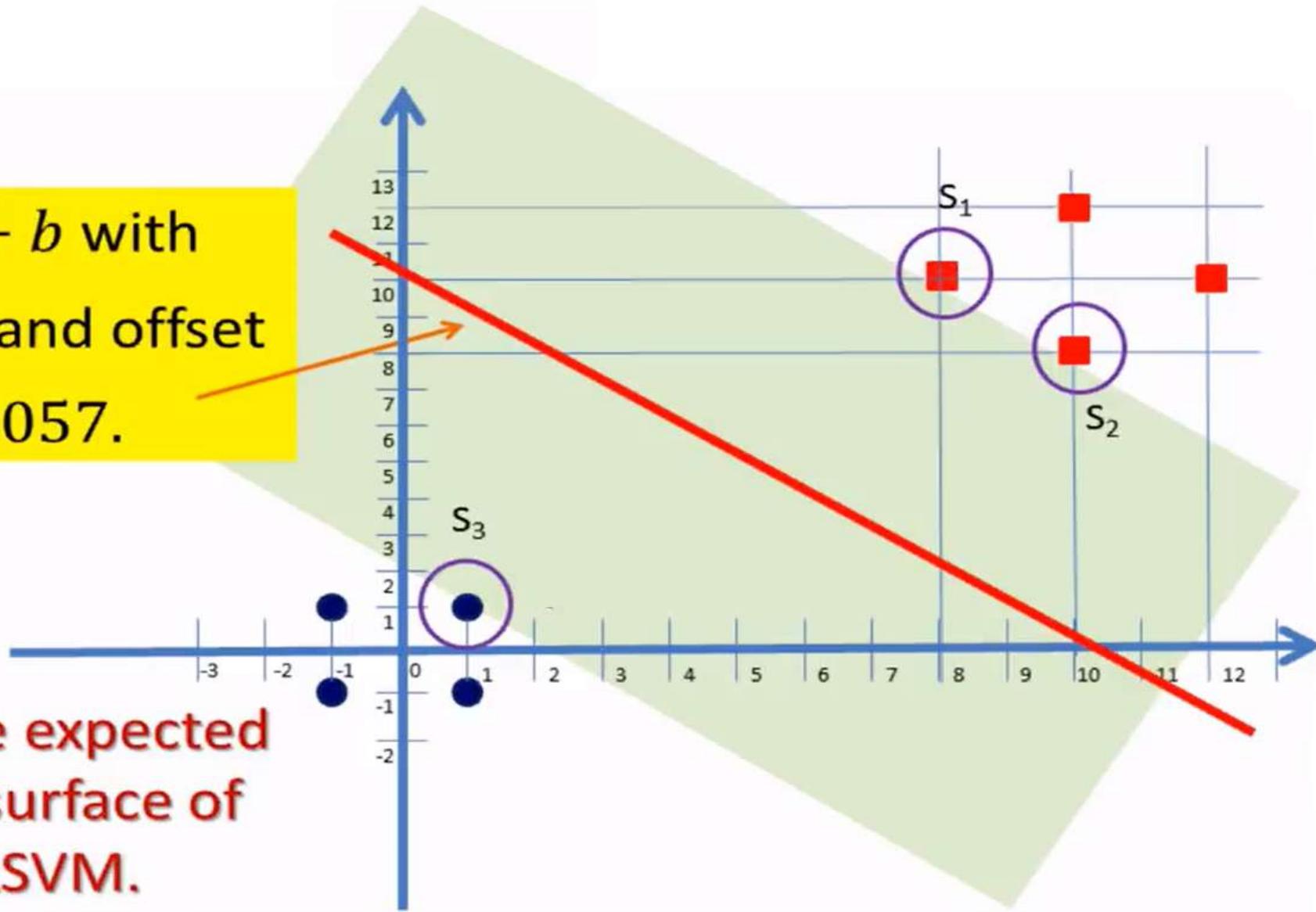
$$y = wx + b \text{ with } w = \begin{pmatrix} 0.1243/0.1243 \\ 0.1243/0.1243 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and an offset $b = -\frac{1.2501}{0.1243} = -10.057$.

Hyper Plane Separating the Two classes

- $y = wx + b$ with
 $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and offset
 $b = -10.057.$

- This is the expected decision surface of the Non LSVM.



Association Rule & Apriori Algorithm

Dr. H.K.Tripathy

Association rule mining

- Proposed by **Agrawal et al in 1993.**
- It is an important data mining model studied extensively by the database and data mining community.
- Assume all data are categorical.
- No good algorithm for numeric data.
- Initially used for **Market Basket Analysis** to find how items purchased by customers are related.

Bread → Milk [sup = 5%, conf = 100%]

What Is Association Mining?

- Motivation: finding regularities in data
 - What products were often purchased together? — Beer and diapers
 - What are the subsequent purchases after buying a PC?
 - What kinds of DNA are sensitive to this new drug?
 - Can we automatically classify web documents?

Association rule mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\}$,
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\}$,
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\}$,

Implication means co-occurrence, not causality!

Basket Data

Retail organizations, e.g., supermarkets, collect and store massive amounts sales data, called **basket data**.

A record consist of

transaction date

items bought

Or, basket data may consist of items bought by a customer over a period.

- Items frequently purchased together:

Bread \Rightarrow PeanutButter

Example Association Rule

90% of transactions that purchase bread and butter also purchase milk

“IF” part = **antecedent**

“THEN” part = **consequent**

“Item set” = the items (e.g., products) comprising the antecedent or consequent

- Antecedent and consequent are *disjoint* (i.e., have no items in common)

Antecedent: bread and butter

Consequent: milk

Confidence factor: 90%

Transaction data: supermarket data

- Market basket transactions:

t1: {bread, cheese, milk}

t2: {apple, eggs, salt, yogurt}

...

...

tn: {biscuit, eggs, milk}

- Concepts:

- An *item*: an item/article in a basket

- I : the set of all items sold in the store

- A *transaction*: items purchased in a basket; it may have TID (transaction ID)

- A *transactional dataset*: A set of transactions

Transaction data: a set of documents

- A text document data set. Each document is treated as a “bag” of keywords

doc1: Student, Teach, School

doc2: Student, School

doc3: Teach, School, City, Game

doc4: Baseball, Basketball

doc5: Basketball, Player, Spectator

doc6: Baseball, Coach, Game, Team

doc7: Basketball, Team, City, Game

Definition: Frequent Itemset

- **Itemset**
 - A collection of one or more items
 - Example: {Milk, Bread, Diaper}
 - k-itemset
 - An itemset that contains k items
- **Support count (σ)**
 - Frequency of occurrence of an itemset
 - E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- **Support**
 - Fraction of transactions that contain an itemset
 - E.g. $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- **Frequent Itemset**
 - An itemset whose support is greater than or equal to a *minsup* threshold

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

The model: data

- $I = \{i_1, i_2, \dots, i_m\}$: a set of *items*.
- **Transaction t** :
 - t a set of items, and $t \subseteq I$.
- **Transaction Database T** : a set of transactions $T = \{t_1, t_2, \dots, t_n\}$.
- **I : itemset**

{cucumber, parsley, onion, tomato, salt, bread, olives, cheese, butter}

- **T : set of transactions**

- 1 {{cucumber, parsley, onion, tomato, salt, bread},
- 2 {tomato, cucumber, parsley},
- 3 {tomato, cucumber, olives, onion, parsley},
- 4 {tomato, cucumber, onion, bread},
- 5 {tomato, salt, onion},
- 6 {bread, cheese}
- 7 {tomato, cheese, cucumber}
- 8 {bread, butter}}

The model: Association rules

- A transaction t contains X , a set of items (**itemset**) in I , if $X \subseteq t$.
- An **association rule** is an implication of the form:

$$X \rightarrow Y, \text{ where } X, Y \subset I, \text{ and } X \cap Y = \emptyset$$

- An **itemset** is a set of items.
 - E.g., $X = \{\text{milk, bread, cereal}\}$ is an itemset.
- A **k -itemset** is an itemset with k items.
 - E.g., $\{\text{milk, bread, cereal}\}$ is a 3-itemset

Rule strength measures

- **Support:** The rule holds with **support sup** in T (the transaction data set) if $sup\%$ of transactions contain $X \cup Y$.
 - sup = probability that a transaction contains $\Pr(X \cup Y)$
(Percentage of transactions that contain $X \cup Y$)
- **Confidence:** The rule holds in T with **confidence $conf$** if $conf\%$ of transactions that contain X also contain Y .
 - $conf$ = conditional probability that a transaction having X also contains Y
 $\Pr(Y | X)$
(Ratio of number of transactions that contain $X \cup Y$ to the number that contain X)
- An association rule is a pattern that states when X occurs, Y occurs with certain probability.

Support and Confidence

- **Support count:** The support count of an itemset X , denoted by $X.count$, in a data set T is the number of transactions in T that contain X . Assume T has n transactions.

- Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Goal: Find all rules that satisfy the user-specified *minimum support* (minsup) and *minimum confidence* (minconf).

Definition: Association Rule

- **Association Rule**

- An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
- Example:
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**

- Support (s)
 - ◆ Fraction of transactions that contain both X and Y
- Confidence (c)
 - ◆ Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk , Diaper } \} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk , Diaper, Beer })}{| T |} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer })}{\sigma(\text{Milk , Diaper })} = \frac{2}{3} = 0.67$$

Is minimum support and minimum confidence can be automatically determined in mining association rules?

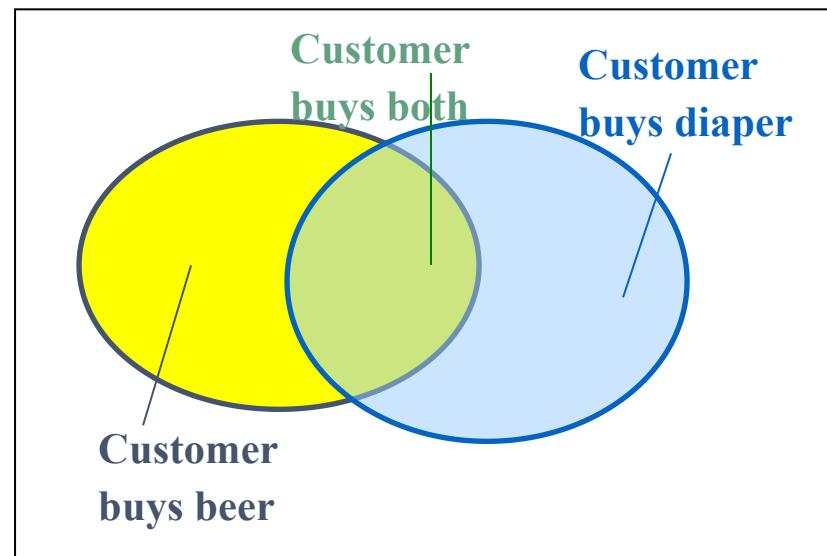
- For the **minimum support**, it all depends on the dataset. Usually, may start with a high value, and then decrease the values until to find a value that will generate enough patterns.
- For the **minimum confidence**, it is a little bit easier because it represents the confidence that you want in the rules. So usually, use something like 60 %. But it also depends on the data.
- In terms of performance, when ***minsup*** is higher you will find **less pattern** and the algorithm is faster. For ***minconf***, when it is set higher, there will be less pattern but it may not be faster because many algorithms don't use minconf to prune the search space. So obviously, setting these parameters also depends on how many rules you want.

An example

- Transaction data
 - Assume:
 - minsup = 30%
 - minconf = 80%
 - An example **frequent itemset**:
 $\{\text{Chicken, Clothes, Milk}\}$ [sup = 3/7]
 - **Association rules** from the itemset:
 $\text{Clothes} \rightarrow \text{Milk, Chicken}$ [sup = 3/7, conf = 3/3]
...
 $\text{Clothes, Chicken} \rightarrow \text{Milk, }$ [sup = 3/7, conf = 3/3]
- t1: Bread, Chicken, Milk
t2: Bread, Cheese
t3: Cheese, Boots
t4: Bread, Chicken, Cheese
t5: Bread, Chicken, Clothes, Cheese, Milk
t6: Chicken, Clothes, Milk
t7: Chicken, Milk, Clothes

Basic Concept: Association Rules

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F



- Let $\text{min_support} = 50\%$, $\text{min_conf} = 50\%:$
 - $A \rightarrow C$ (50%, 66.7%)
 - $C \rightarrow A$ (50%, 100%)

Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

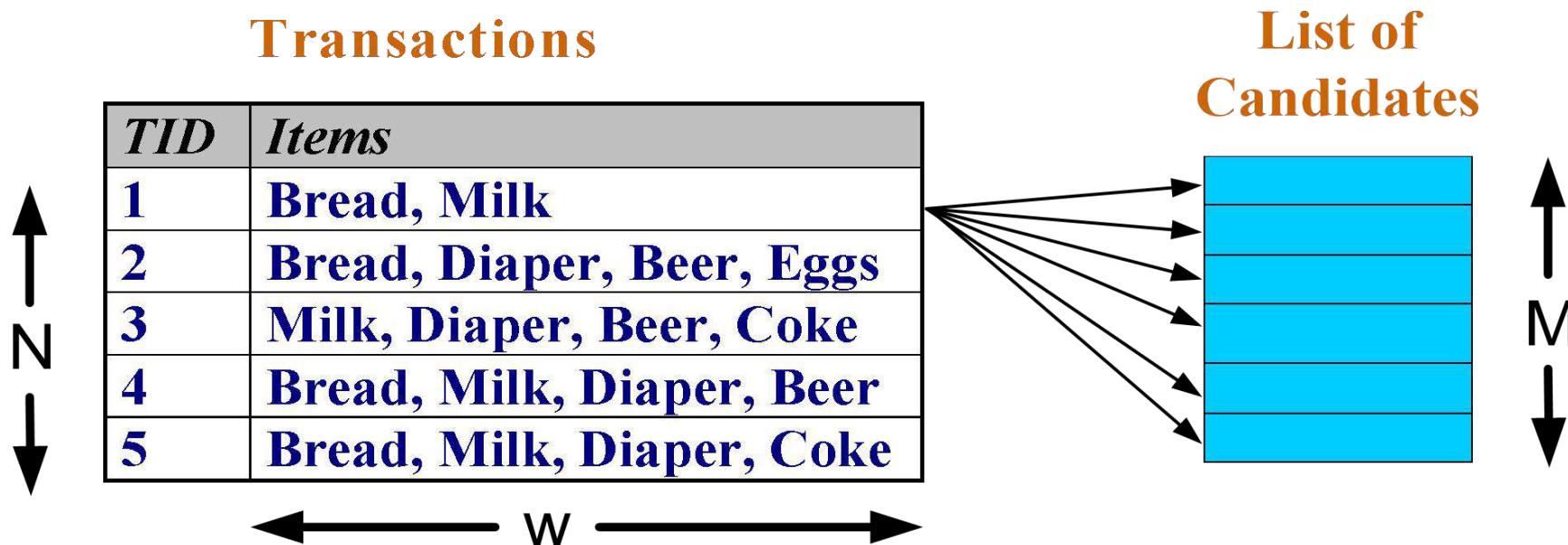
Association Rule Mining Task

- Given a set of transactions T, the goal of association rule mining is to find all rules having
 - **support $\geq \text{minsup}$ threshold**
 - **confidence $\geq \text{minconf}$ threshold**
- ***Brute-force approach:***
 - List all possible association rules
 - Compute the support and confidence for each rule
 - Prune rules that fail the minsup and minconf thresholds

⇒ Computationally prohibitive!

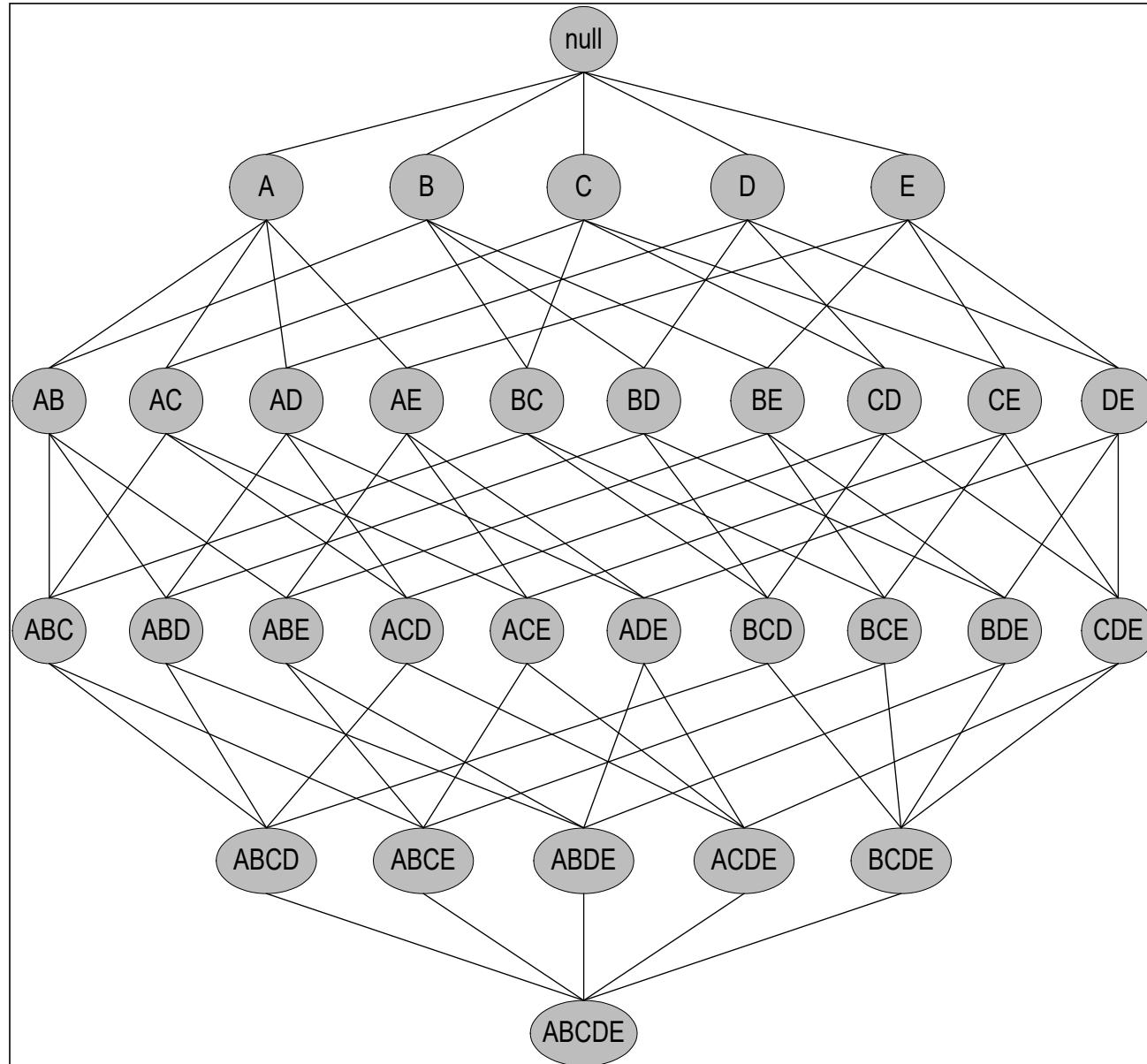
Frequent Itemset Generation

- Brute-force approach:
 - Each itemset in the lattice is a **candidate** frequent itemset
 - Count the support of each candidate by scanning the database



- Match each transaction against every candidate
- Complexity $\sim O(NMw)$ => **Expensive since $M = 2^d$!!!**

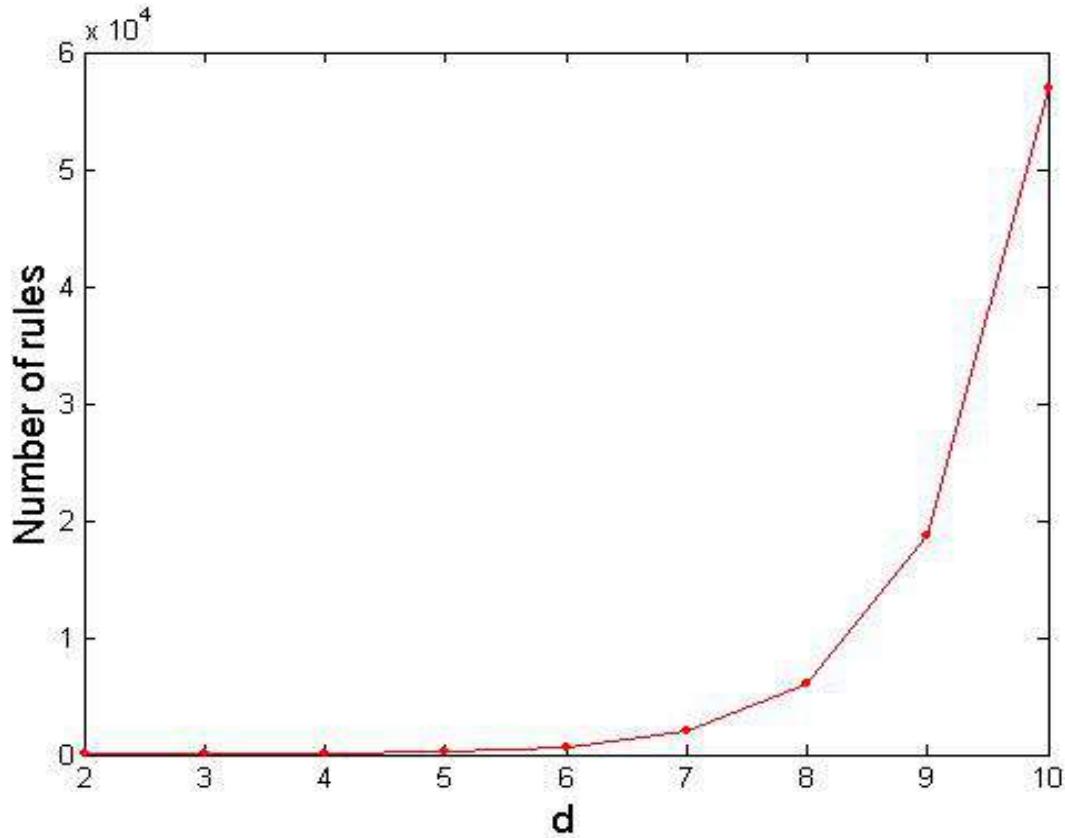
Brute-force approach:



Given d items, there are 2^d possible candidate itemsets

Computational Complexity

- Given d unique items:
 - Total number of itemsets = 2^d
 - Total number of possible association rules:



$$\begin{aligned}R &= \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right] \\&= 3^d - 2^{d+1} + 1\end{aligned}$$

If $d=6$, $R = 602$ rules

Mining Association Rules

- Two-step approach:

1. Frequent Itemset Generation

- Generate all itemsets whose support $\geq \text{minsup}$

if an itemset is frequent, each of its subsets is frequent as well.

- This property belongs to a special category of properties called **antimonotonicity** in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well.

1. Rule Generation

- Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive

Frequent Itemset Generation

- An itemset X is ***closed*** in a data set D if there exists no proper super-itemset Y^* such that Y has the same support count as X in D .
****(Y is a proper super-itemset of X if X is a proper sub-itemset of Y, that is, if $X \subset Y$. In other words, every item of X is contained in Y but there is at least one item of Y that is not in X.)***
- An itemset X is a ***closed frequent itemset*** in set D if X is both closed and frequent in D .
- An itemset X is a ***maximal frequent itemset (or max-itemset)*** in a data set D if X is frequent, and there exists no super-itemset Y such that $X \subset Y$ and Y is frequent in D .

Frequent Itemset Generation Strategies

- Reduce the **number of candidates** (M)
 - Complete search: $M=2^d$
 - Use pruning techniques to reduce M
- Reduce the **number of transactions** (N)
 - Reduce size of N as the size of itemset increases
 - Used by DHP (Direct Hashing & Purning) and vertical-based mining algorithms
- Reduce the **number of comparisons** (NM)
 - Use efficient data structures to store the candidates or transactions
 - No need to match every candidate against every transaction

Many mining algorithms

- There are a large number of them!!
- They use different strategies and data structures.
- Their resulting sets of rules are all the same.
 - Given a transaction data set T , and a minimum support and a minimum confident, the set of association rules existing in T is uniquely determined.
- Any algorithm should find the same set of rules although their computational efficiencies and memory requirements may be different.
- We study only one: the Apriori Algorithm

The Apriori algorithm

- The algorithm uses a level-wise search, where **k-itemsets** are used to explore **(k+1)-itemsets**
- In this algorithm, frequent subsets are extended one item at a time (this step is known as ***candidate generation process***)
- Then groups of candidates are tested against the data.
- It identifies the frequent individual items in the database and extends them to larger and larger item sets as long as those itemsets appear sufficiently often in the database.
- Apriori algorithm determines frequent itemsets that can be used to determine association rules which highlight general trends in the database.

The Apriori algorithm

- The Apriori algorithm takes advantage of the fact that any subset of a frequent itemset is also a frequent itemset.
 - *i.e., if $\{I_1, I_2\}$ is a frequent itemset, then $\{I_1\}$ and $\{I_2\}$ should be frequent itemsets.*
- The algorithm can therefore, reduce the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count.
- All infrequent itemsets can be pruned if it has an infrequent subset.

How do we do that?

- So we build a *Candidate list* of k-itemsets and then extract a *Frequent list* of k-itemsets using the support count
- After that, we use the *Frequent list* of k-itemsets in determining the *Candidate* and *Frequent list* of k+1-itemsets.
- We use *Pruning* to do that
- We repeat until we have an empty *Candidate* or *Frequent* of k-itemsets
 - Then we return the list of *k-1-itemsets*.

KEY CONCEPTS

- Frequent Itemsets: All the sets which contain the item with the minimum support (denoted by L_i for i^{th} itemset).
- Apriori Property: Any subset of frequent itemset must be frequent.
- Join Operation: To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself.

APRIORI ALGORITHM EXAMPLE

Database D

Minsup = 0.5

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Scan D

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3



L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_2

Scan D

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C_3

Scan D

itemset
{2 3 5}

L_3

itemset	sup
{2 3 5}	2

The Apriori Algorithm : Pseudo Code

- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code :
 C_k : Candidate itemset of size k
 L_k : frequent itemset of size k

```
 $L_1 = \{\text{frequent items}\};$ 
for ( $k = 1; L_k \neq \emptyset; k++$ ) do begin
     $C_{k+1} = \text{candidates generated from } L_k;$ 
    for each transaction  $t$  in database do
        increment the count of all candidates in  $C_{k+1}$ 
        that are contained in  $t$ 
     $L_{k+1} = \text{candidates in } C_{k+1} \text{ with min\_support}$ 
end
return  $\cup_k L_k;$ 
```

Apriori's Candidate Generation

- For $k=1$, $C_1 = \text{all 1-itemsets}$.
- For $k>1$, generate C_k from L_{k-1} as follows:

– *The join step*

$C_k = k-2$ way join of L_{k-1} with itself

If both $\{a_1, \dots, a_{k-2}, a_{k-1}\}$ & $\{a_1, \dots, a_{k-2}, a_k\}$ are in L_{k-1} ,
then add $\{a_1, \dots, a_{k-2}, a_{k-1}, a_k\}$ to C_k

(We keep items **sorted**).

– *The prune step*

Remove $\{a_1, \dots, a_{k-2}, a_{k-1}, a_k\}$ if it contains a non-frequent $(k-1)$ subset

Example – Finding frequent itemsets

Dataset D

TID	Items
T100	a1 a3 a4
T200	a2 a3 a5
T300	a1 a2 a3 a5
T400	a2 a5

minSup=0.5

1. scan D \rightarrow C_1 : a1:2, a2:3, a3:3, a4:1, a5:3

$\rightarrow L_1$: a1:2, a2:3, a3:3, a5:3

$\rightarrow C_2$: a1a2, a1a3, a1a5, a2a3, a2a5, a3a5

2. scan D \rightarrow C_2 : a1a2:1, a1a3:2, a1a5:1, a2a3:2, a2a5:3, a3a5:2

$\rightarrow L_2$: a1a3:2, a2a3:2, a2a5:3, a3a5:2

$\rightarrow C_3$: a1a2a3, a2a3a5

\rightarrow Pruned C_3 : a1a2a3

3. scan D \rightarrow L_3 : a2a3a5:2

Order of items can make difference in process

Dataset D

TID	Items
T100	1 3 4
T200	2 3 5
T300	1 2 3 5
T400	2 5

1. scan D $\rightarrow C_1: 1:2, 2:3, 3:3, 4:1, 5:3$

$\rightarrow L_1: 1:2, 2:3, 3:3, 5:3$

$\rightarrow C_2: 12, 13, 15, 23, 25, 35$

2. scan D $\rightarrow C_2: 12:1, 13:2, 15:1, 23:2, 25:3, 35:2$

Suppose the order of items is: 5,4,3,2,1

$\rightarrow L_2: 31:2, 32:2, 52:3, 53:2$

$\rightarrow C_3: 321, 532$

\rightarrow Pruned $C_3: 532$

3. scan D $\rightarrow L_3: 532:2$

minSup=0.5

Generating Association Rules

From frequent itemsets

- Procedure 1:
- Let we have the list of frequent itemsets
 - Generate all nonempty subsets for each frequent itemset I
 - For $I = \{1,3,5\}$, all nonempty subsets are $\{1,3\}, \{1,5\}, \{3,5\}, \{1\}, \{3\}, \{5\}$
 - For $I = \{2,3,5\}$, all nonempty subsets are $\{2,3\}, \{2,5\}, \{3,5\}, \{2\}, \{3\}, \{5\}$

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5
500	1 3 5

Itemset	Support
{1,3,5}	2
{2,3,5}	2

Generating Association Rules

From frequent itemsets

- Procedure 2:
- For every nonempty subset S of I, output the rule:

$$S \rightarrow (I - S)$$

- If $\text{support_count}(I)/\text{support_count}(s) \geq \text{min_conf}$
where min_conf is minimum confidence threshold

- Let us assume:
- minimum confidence threshold is 60%

Association Rules with confidence

- R1 : 1,3 -> 5
 - Confidence = $sc\{1,3,5\}/sc\{1,3\} = 2/3 = 66.66\%$ (R1 is selected)
- R2 : 1,5 -> 3
 - Confidence = $sc\{1,5,3\}/sc\{1,5\} = 2/2 = 100\%$ (R2 is selected)
- R3 : 3,5 -> 1
 - Confidence = $sc\{3,5,1\}/sc\{3,5\} = 2/3 = 66.66\%$ (R3 is selected)
- R4 : 1 -> 3,5
 - Confidence = $sc\{1,3,5\}/sc\{1\} = 2/3 = 66.66\%$ (R4 is selected)
- R5 : 3 -> 1,5
 - Confidence = $sc\{3,1,5\}/sc\{3\} = 2/4 = 50\%$ (R5 is REJECTED)
- R6 : 5 -> 1,3
 - Confidence = $sc\{5,1,3\}/sc\{5\} = 2/4 = 50\%$ (R6 is REJECTED)

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5
500	1 3 5

How to efficiently generate rules?

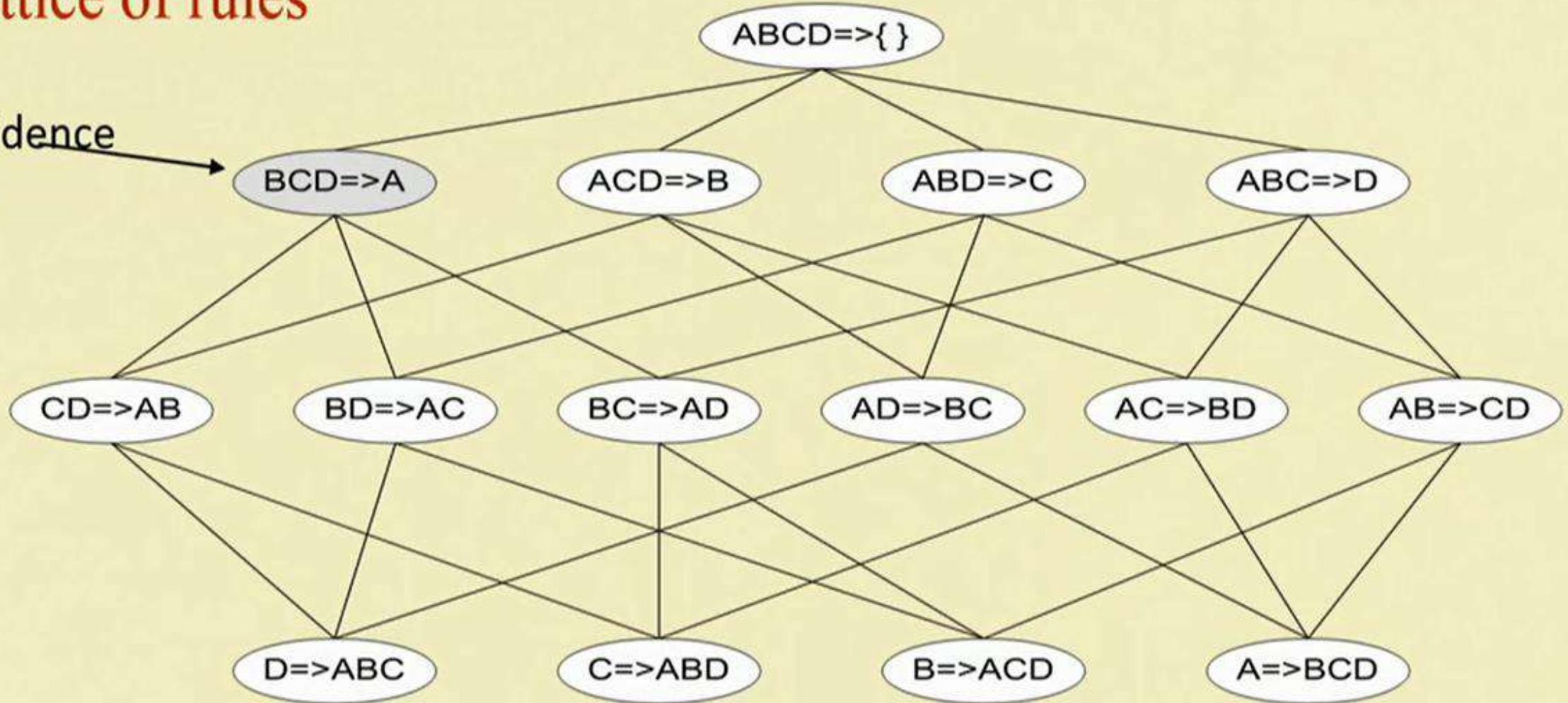
- In general, confidence does not have an anti-monotone property
 $c(ABC \rightarrow D)$ can be larger or smaller than $c(AB \rightarrow D)$
- But confidence of rules generated from the same itemset has an anti-monotone property
 - e.g., $L = \{A, B, C, D\}$
 $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

Confidence is anti-monotone w.r.t number of items on the RHS of the rule.

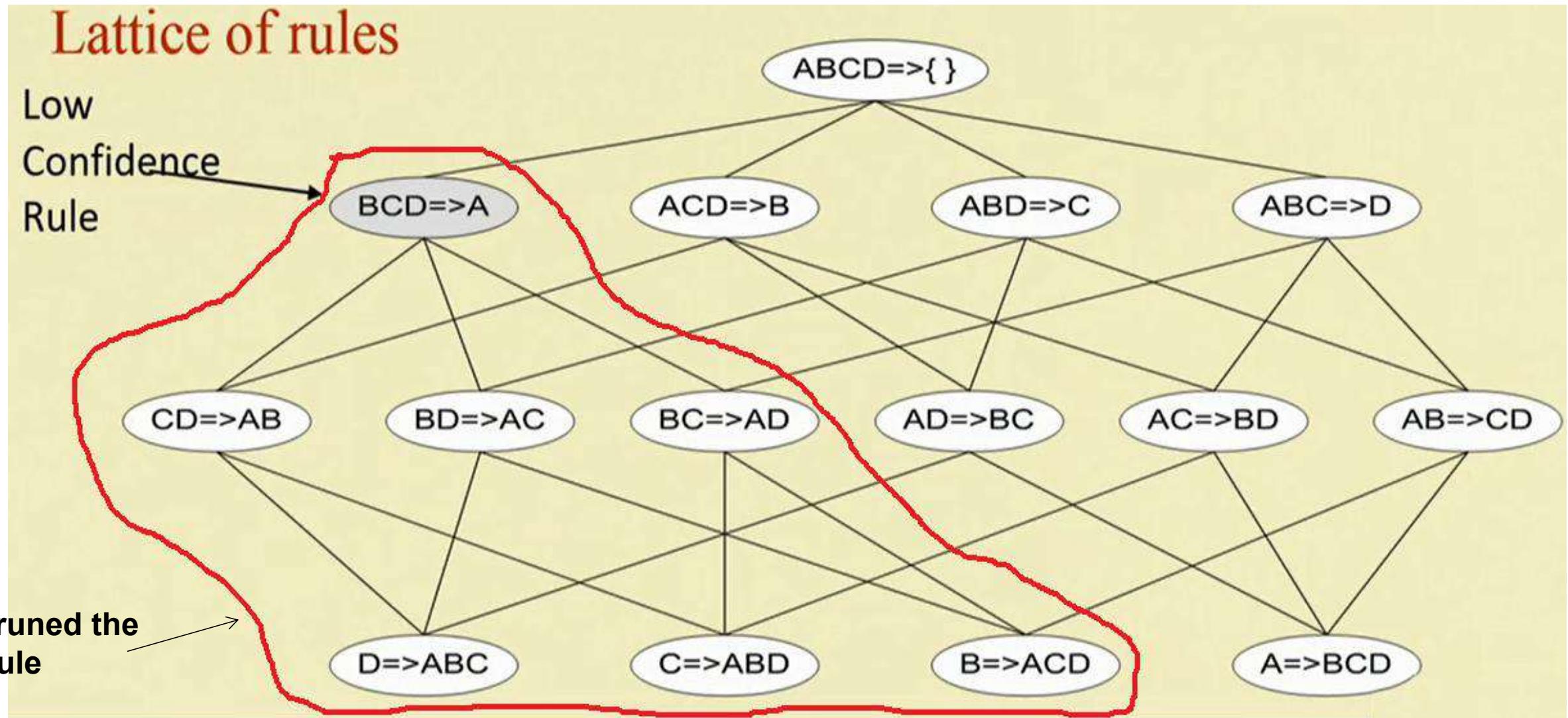
Rule generation for Apriori Algorithm

Lattice of rules

Low
Confidence
Rule

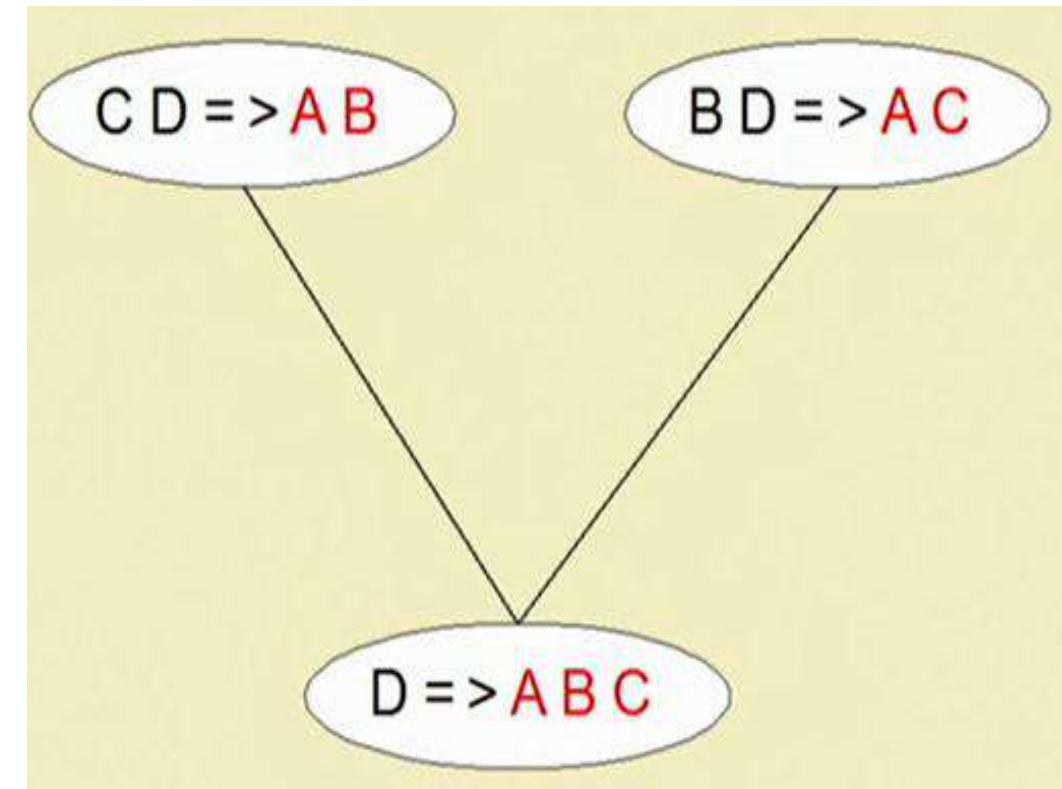


Rule generation for Apriori Algorithm



Rule generation for Apriori Algorithm

- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent
- join ($CD \Rightarrow AB$, $BD \Rightarrow AC$)
would produce the candidate rule,
 $D \Rightarrow ABC$
- Prune rule $D \Rightarrow ABC$ if its subset
 $AD \Rightarrow BC$ does not have high confidence



Association Analysis to Correlation Analysis

Pattern Evaluation

- Association rule algorithms tend to produce too many rules
 - many of them are uninteresting or redundant
 - Redundant if $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$ have same support & confidence
- Interestingness measures can be used to prune/rank the derived patterns
- In the original formulation of association rules, support & confidence are the only measures used

Which Patterns Are Interesting? – Pattern Evaluation Methods

- Strong Rules Are Not Necessarily Interesting.
- The support–confidence framework can be supplemented with additional interestingness measures based on correlation analysis.
- *“How can we tell which strong association rules are really interesting?”*

Computing Interestingness Measure

Contingency table for $X \rightarrow Y$

	Y	\bar{Y}	
X	f_{11}	f_{10}	f_{1+}
\bar{X}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$ T $

f_{11} : support of X and Y

f_{10} : support of X and \bar{Y}

f_{01} : support of \bar{X} and Y

f_{00} : support of \bar{X} and \bar{Y}

Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: $\text{Tea} \rightarrow \text{Coffee}$

$$\text{Confidence} = P(\text{Coffee}|\text{Tea}) = 0.75$$

$$\text{but } P(\text{Coffee}) = 0.9$$

⇒ Although confidence is high, rule is misleading

$$\Rightarrow P(\text{Coffee}|\overline{\text{Tea}}) = 0.9375$$

Association rules analysis

(From Association Analysis to Correlation Analysis)

- Association rules analysis is a technique to uncover how items are associated to each other. Traditionally, association rule mining is performed by using two interestingness measures named the ***support*** and ***confidence*** to evaluate rules.
- There are three common ways to measure association.
- **Support:** This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. You may then identify itemsets with support values above the threshold (***min_supp***) as significant itemsets.
- **Confidence.** This says how likely item Y is purchased when item X is purchased, expressed as $\{X \rightarrow Y\}$. This is measured by the proportion of transactions with item X, in which item Y also appears.
- ***One drawback of the confidence measure is that it might misrepresent the importance of an association. As discussed in last example.***
- ***To account for the base popularity of both constituent items, we use a third measure called lift or interest.***

- **Prior probability:** The probability of outcome $P(Y)$ is called prior probability, which can be calculated from the training dataset. Prior probability shows the likelihood of an outcome in a given dataset.
- **Conditional probability:** $P(Y|X)$ is called the conditional probability, which provides the probability of an outcome given the evidence, that is, when the value of X is known.
- **Posterior probability (Adjusted Probability):** $P(Y|X)$ is also called posterior probability. Calculating posterior probability is the objective of data science using Bayes' theorem. This is the likelihood of an outcome as the conditions are learnt.

Lift Measure

- The lift, also referred to as the interestingness measure, takes this into account by incorporating the **prior probability**, Lift of the rule consequent as follows:

$$\text{Lift}(X \rightarrow Y) = \left(\frac{\text{sup}(X \cup Y)}{N} \right) / \left(\frac{\text{sup}(X)}{N} * \frac{\text{sup}(Y)}{N} \right), \text{ where}$$

- N is the number of transactions in the transaction database,
- sup(X ∪ Y) is the number of transactions containing X, Y items together,
- sup(X) is the number of transactions containing X item(s) only.
- sup(Y) is the number of transactions containing Y item(s) only.
- The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule.
- This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

Correlation Concepts

- Lift is easier to understand when written in terms of probabilities.

$$\text{Lift} = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \cdot \text{Support}(Y)} = P(X, Y) / P(X) \cdot P(Y)$$

Probability = Focus to Events
Support = Focus to item togetherness

- The Lift measures the probability of X and Y **occurring together** divided by the probability of X and Y occurring if they were independent events.
- If the lift is equal to 1**, it means that two item sets **A and B are independent** (the occurrence of A is independent of the occurrence of item set B)
in this case, **support(A ∪ B) = support(A) · support(B)**
- If the lift is higher than 1**, it means that A and B are **positively correlated**.
- If the lift is lower than 1**, it means that A and B are **negatively correlated**.

Correlation Concepts [Cont.]

- $\text{corr}(A,B) > 1$ means that A and B are **positively correlated**
i.e. the occurrence of one implies the occurrence of the other.
- $\text{corr}(A,B) < 1$ means that the occurrence of A is **negatively correlated**
with (or discourages) the occurrence of B.
- $\text{corr}(A,B) = 0$ means that A and B are **independent** and there is **no correlation** between them.

Statistical Independence (Probabilistic)

- Population of 1000 students
 - 600 students know how to swim (S)
 - 700 students know how to bike (B)
 - 420 students know how to swim and bike (S,B)
- $P(S \wedge B) = 420/1000 = 0.42$
- $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
- $P(S \wedge B) = P(S) \times P(B) \Rightarrow$ Statistical independence
- $P(S \wedge B) > P(S) \times P(B) \Rightarrow$ Positively correlated
- $P(S \wedge B) < P(S) \times P(B) \Rightarrow$ Negatively correlated

Interestingness Measure: Correlations (Lift)

- $play\ basketball \Rightarrow eat\ cereal$ [40%, 66.7%] is misleading
 - The overall % of students eating cereal is 75% > 66.7%.
- $play\ basketball \Rightarrow not\ eat\ cereal$ [20%, 33.3%] is more accurate, although with lower support and confidence
- Measure of dependent/correlated events: **lift**

$$lift = \frac{Support(A \cup B)}{Support(A).Support(B)}$$

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

$$lift(B, C) = \frac{2000/5000}{3000/5000 * 3750/5000} = 0.89 \quad lift(B, \neg C) = \frac{1000/5000}{3000/5000 * 1250/5000} = 1.33$$

Example: Lift/Interest

	Coffee	$\bar{\text{Coffee}}$	
Tea	15	5	20
$\bar{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

$$\text{Confidence} = P(\text{Coffee}|\text{Tea}) = 0.75$$

$$\text{but } P(\text{Coffee}) = 0.9$$

$$\Rightarrow \text{Lift} = 0.75/0.9 = 0.8333 (< 1, \text{ therefore is negatively associated})$$

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

#	Measure	Formula
1	ϕ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(AB) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(AB) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(AB)} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(AB)} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$ $\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$
7	Mutual Information (M)	$\max \left(P(A,B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}B) \log \left(\frac{P(\bar{B} A)}{P(\bar{B})} \right), P(A,B) \log \left(\frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$
8	J-Measure (J)	$\max \left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2, P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] - P(A)^2 - P(\bar{A})^2 \right)$
9	Gini index (G)	$P(A,B)$ $\max(P(B A), P(A B))$ $\max \left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$ $\max \left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(B\bar{A})} \right)$ $\frac{\frac{P(A,B)}{P(A)P(\bar{B})}}{\sqrt{P(A)P(B)}}$ $P(A,B) - P(A)P(B)$ $\max \left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$ $\max(P(B A) - P(B), P(A B) - P(A))$ $\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$ $\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$ $\sqrt{P(A,B)} \max(P(B A) - P(B), P(A B) - P(A))$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B A), P(A B))$
12	Laplace (L)	$\max \left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$
13	Conviction (V)	$\max \left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(B\bar{A})} \right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(\bar{B})}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max \left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$
18	Added Value (AV)	$\max(P(B A) - P(B), P(A B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)} \max(P(B A) - P(B), P(A B) - P(A))$

Summary

The Frequent Pattern Mining Model

- **Definition (Support):** *The support of an itemset I is defined as the fraction of the transactions in the database $T = \{T_1 \dots T_n\}$ that contain I as a subset.*

$$\text{support } (A \Rightarrow B) = P(A \cup B)$$

Relative support: The itemset support defined in above equation is sometimes referred to as relative support.

Absolute support: whereas the occurrence frequency is called the absolute support.

(If the relative support of an itemset I satisfies a prespecified minimum support threshold (i.e., the absolute support of I satisfies the corresponding minimum support count threshold), then I is a frequent itemset)

- **Definition (Frequent Itemset Mining):** *Given a set of transactions $T = \{T_1 \dots T_n\}$, where each transaction T_i is a subset of items from U , determine all itemsets I that occur as a subset of at least a predefined fraction minsup of the transactions in T .*
- **Definition (Maximal Frequent Itemsets):** *A frequent itemset is maximal at a given minimum support level minsup , if it is frequent, and no superset of it is frequent.*

The Frequent Pattern Mining Model

- **Property (Support Monotonicity Property):** *The support of every subset J of I is at least equal to that of the support of itemset I .*

$$sup(J) \geq sup(I) \quad \forall J \subseteq I$$

- **Property (Downward Closure Property):** *Every subset of a frequent itemset is also frequent.*

Summary

Association Rule Generation Framework

- **Definition (Confidence):** Let X and Y be two sets of items. The confidence $\text{conf}(X \cup Y)$ of the rule $X \cup Y$ is the conditional probability of $X \cup Y$ occurring in a transaction, given that the transaction contains X . Therefore, the confidence $\text{conf}(X \Rightarrow Y)$ is defined as follows:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

- **Definition (Association Rules)** Let X and Y be two sets of items. Then, the rule $X \Rightarrow Y$ is said to be an association rule at a minimum support of minsup and minimum confidence of minconf , if it satisfies both the following criteria:

1. The support of the itemset $X \cup Y$ is at least minsup .
2. The confidence of the rule $X \Rightarrow Y$ is at least minconf .

- **Property 4.3.1 (Confidence Monotonicity)** Let X_1 , X_2 , and I be itemsets such that $X_1 \subset X_2 \subset I$. Then the confidence of $X_2 \Rightarrow I - X_2$ is at least that of $X_1 \Rightarrow I - X_1$.

$$\text{conf}(X_2 \Rightarrow I - X_2) \geq \text{conf}(X_1 \Rightarrow I - X_1)$$

Algorithm 6.1 Frequent itemset generation of the *Apriori* algorithm.

- 1: $k = 1$.
- 2: $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$. {Find all frequent 1-itemsets}
- 3: **repeat**
- 4: $k = k + 1$.
- 5: $C_k = \text{apriori-gen}(F_{k-1})$. {Generate candidate itemsets}
- 6: **for** each transaction $t \in T$ **do**
- 7: $C_t = \text{subset}(C_k, t)$. {Identify all candidates that belong to t }
- 8: **for** each candidate itemset $c \in C_t$ **do**
- 9: $\sigma(c) = \sigma(c) + 1$. {Increment support count}
- 10: **end for**
- 11: **end for**
- 12: $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$. {Extract the frequent k -itemsets}
- 13: **until** $F_k = \emptyset$
- 14: Result = $\bigcup F_k$.

Algorithm 6.2 Rule generation of the *Apriori* algorithm.

```
1: for each frequent  $k$ -itemset  $f_k$ ,  $k \geq 2$  do
2:    $H_1 = \{i \mid i \in f_k\}$            {1-item consequents of the rule.}
3:   call ap-genrules( $f_k, H_1.$ )
4: end for
```

Algorithm 6.3 Procedure ap-genrules(f_k, H_m).

```
1:  $k = |f_k|$     {size of frequent itemset.}
2:  $m = |H_m|$     {size of rule consequent.}
3: if  $k > m + 1$  then
4:    $H_{m+1} = \text{apriori-gen}(H_m).$ 
5:   for each  $h_{m+1} \in H_{m+1}$  do
6:      $conf = \sigma(f_k)/\sigma(f_k - h_{m+1}).$ 
7:     if  $conf \geq minconf$  then
8:       output the rule  $(f_k - h_{m+1}) \longrightarrow h_{m+1}.$ 
9:     else
10:      delete  $h_{m+1}$  from  $H_{m+1}.$ 
11:    end if
12:  end for
13:  call ap-genrules( $f_k, H_{m+1}.$ )
14: end if
```

Data Mining: Clustering

Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis

What is Cluster Analysis?

- Cluster: a collection of data objects
 - Similar to one another within the same cluster
 - Dissimilar to the objects in other clusters
- Cluster analysis
 - Grouping a set of data objects into clusters
- Clustering is **unsupervised classification**: no predefined classes
- Typical applications
 - As a **stand-alone tool** to get insight into data distribution
 - As a **preprocessing step** for other algorithms

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
 - create thematic maps in GIS by clustering feature spaces
 - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
 - Document classification
 - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earthquake studies:** Observed earth quake epicenters should be clustered along continent faults

What is not Cluster Analysis?

- Supervised classification
 - Have class label information
- Simple segmentation
 - Dividing students into different registration groups alphabetically, by last name
- Results of a query
 - Groupings are a result of an external specification
- Graph partitioning
 - Some mutual relevance and synergy, but areas are not identical

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
 - **high intra-class** similarity
 - **low inter-class** similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Data Structures

- Data matrix
 - (two modes)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
 - (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: $d(i, j)$
- There is a separate “quality” function that measures the “goodness” of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define “similar enough” or “good enough”
 - the answer is typically highly subjective.

Type of data in clustering analysis

- Interval-scaled variables:
- Binary variables:
- Nominal, ordinal, and ratio variables:
- Variables of mixed types:

Interval-valued variables

- Standardize data

- Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

where $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$

- Calculate the standardized measurement (z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$

- If $q = 2$, d is Euclidean distance: $d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$

Binary Variables

- A contingency table for binary data

		Object <i>j</i>		<i>sum</i>
		1	0	
<i>Object i</i>	1	<i>a</i>	<i>b</i>	<i>a+b</i>
	0	<i>c</i>	<i>d</i>	<i>c+d</i>
<i>sum</i>	<i>a+c</i>	<i>b+d</i>	<i>p</i>	

- Simple matching coefficient (invariant, if the binary variable is symmetric):

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

- Jaccard coefficient (noninvariant if the binary variable is asymmetric):

$$d(i, j) = \frac{b + c}{a + b + c}$$

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be set to 1, and the value N be set to 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
 - m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: use a large number of binary variables
 - creating a new binary variable for each of the M nominal states

Major Clustering Approaches

- **Partitioning algorithms**: Construct various partitions and then evaluate them by some criterion
- **Hierarchy algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- **Density-based**: based on connectivity and density functions
- **Grid-based**: based on a multiple-level granularity structure
- **Model-based**: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Major Clustering Approaches

- Important distinction between **partitional** and **hierarchical** sets of clusters
- **Partitional Clustering**
 - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering**
 - A set of nested clusters organized as a hierarchical tree

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database D of n objects into a set of k clusters
- Given a k , find a partition of k *clusters* that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic methods: *k-means* and *k-medoids* algorithms
 - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
 - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

K-Means Clustering

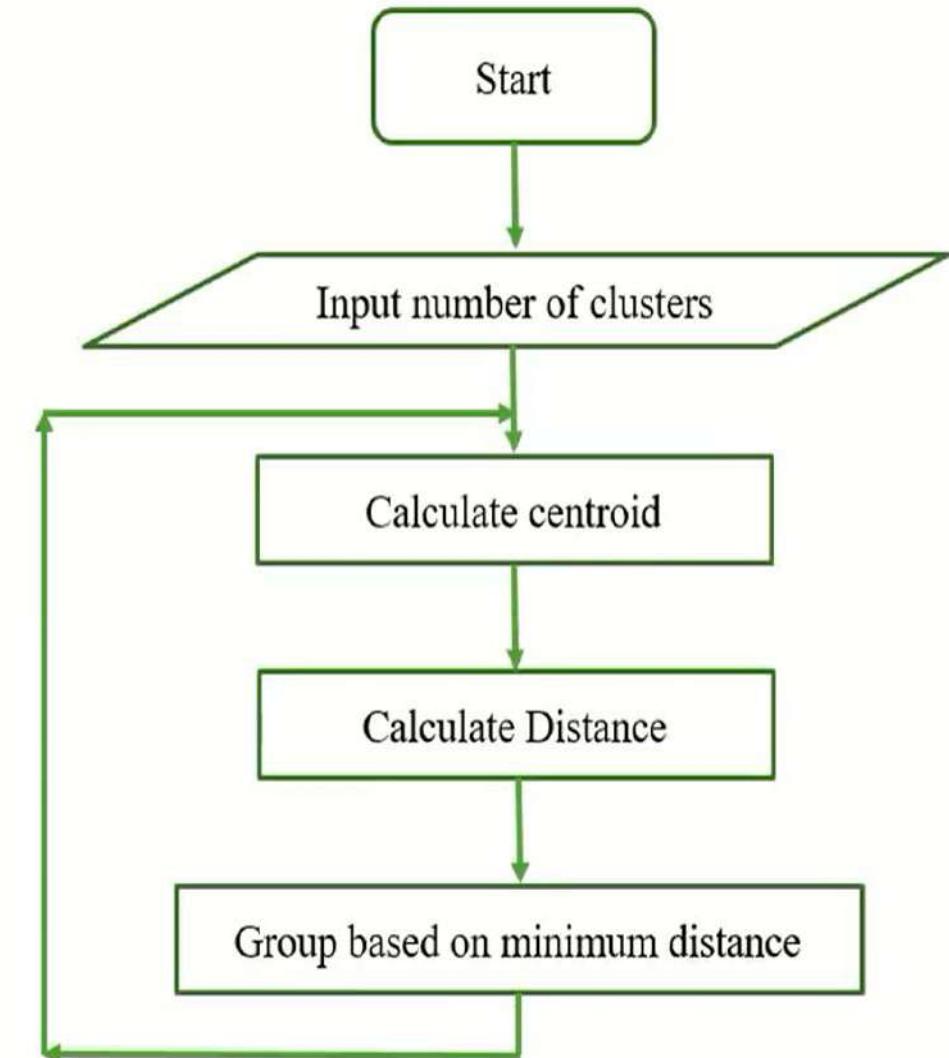
- Simple Clustering: K-means
- Given k , the *k-means* algorithm consists of four steps:

(Basic version works with numeric data only)

- 1) Select initial centroids at random - Pick a number (K) of cluster centers - centroids (at random)
- 2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
- 3) Move each cluster center to the mean of its assigned items
- 4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

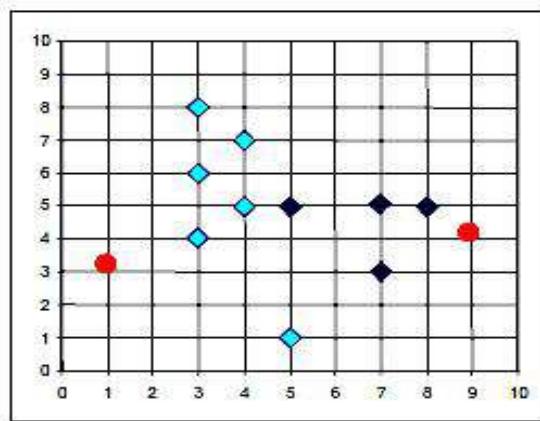
K-means Algorithms

- Initialization
 - Arbitrarily choose k objects as the initial cluster centers (centroids)
- Iteration until no change
 - For each object O_i ,
 - Calculate the distances between O_i and the k centroids
 - (Re)assign O_i to the cluster whose centroid is the closest to O_i
 - Update the cluster centroids based on current assignment



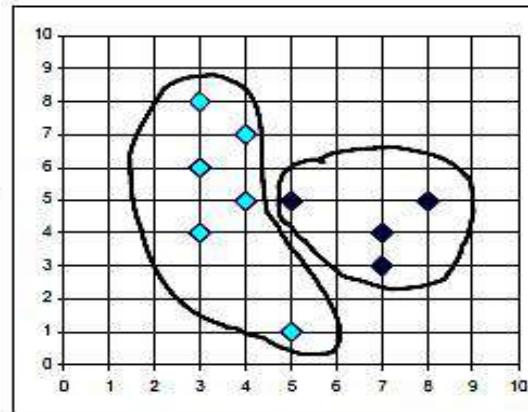
Illustrating K-Means

- Example

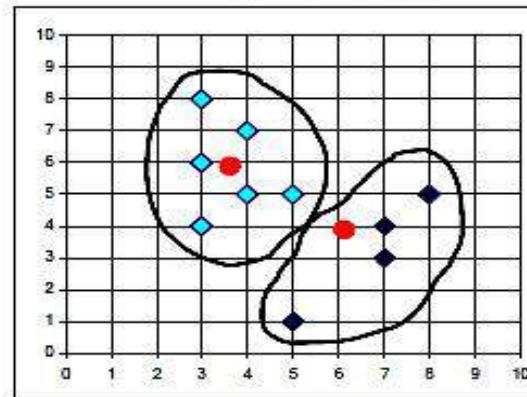


K=2
Arbitrarily choose K
object as initial
cluster center

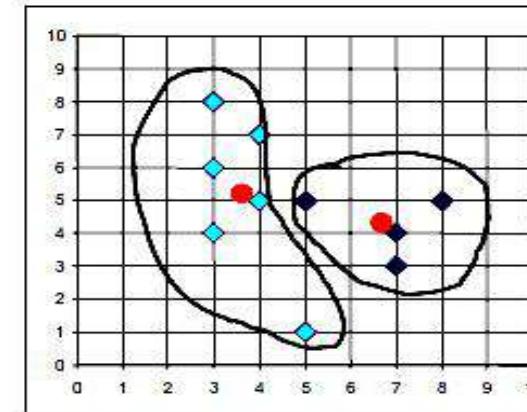
Assign
each
objects
to most
similar center



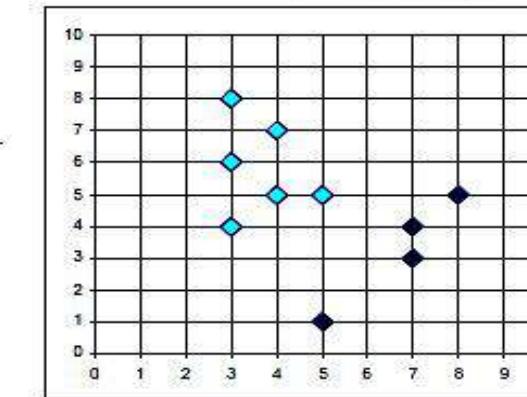
Update
the
cluster
means



Update
the
cluster
means

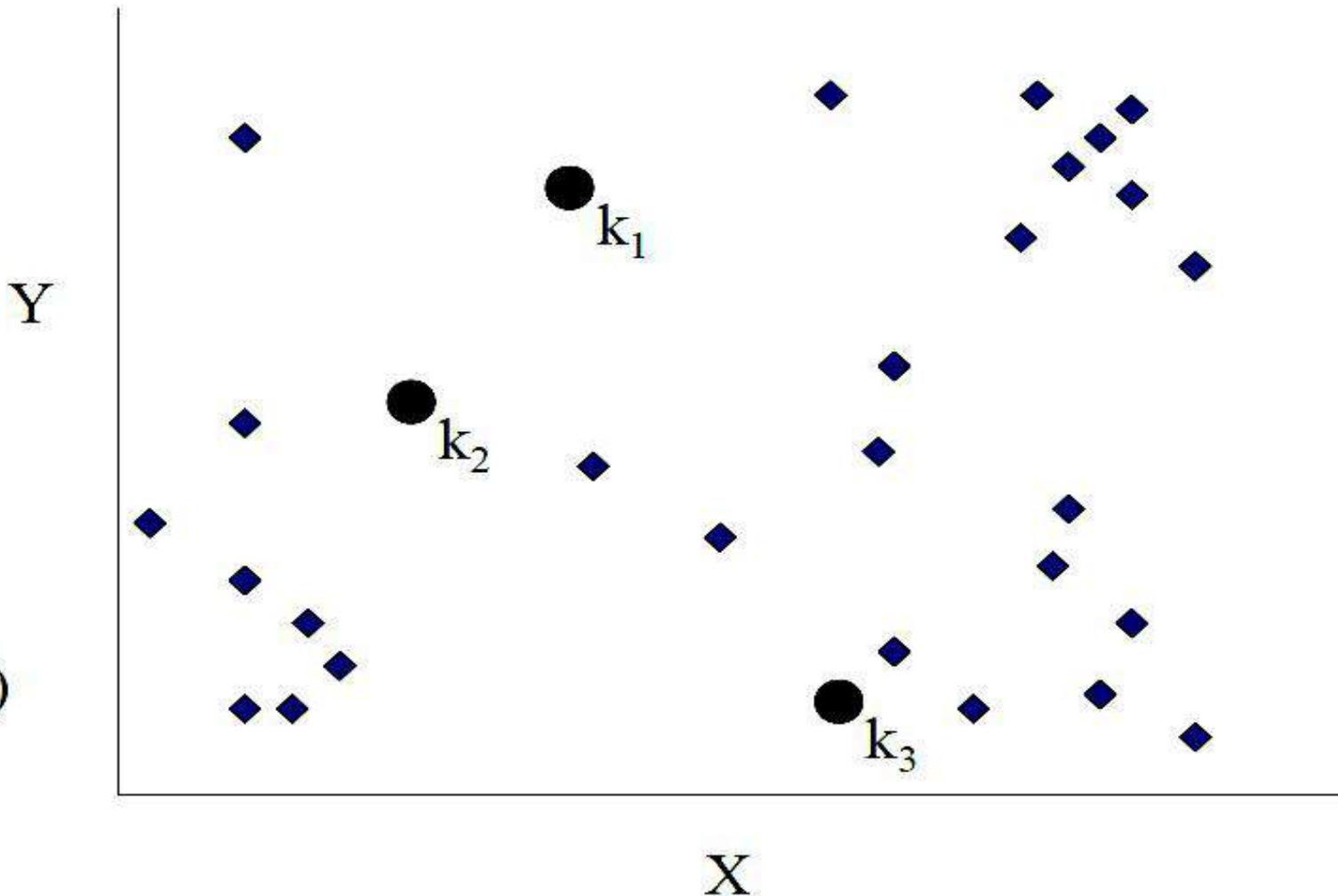


reassign

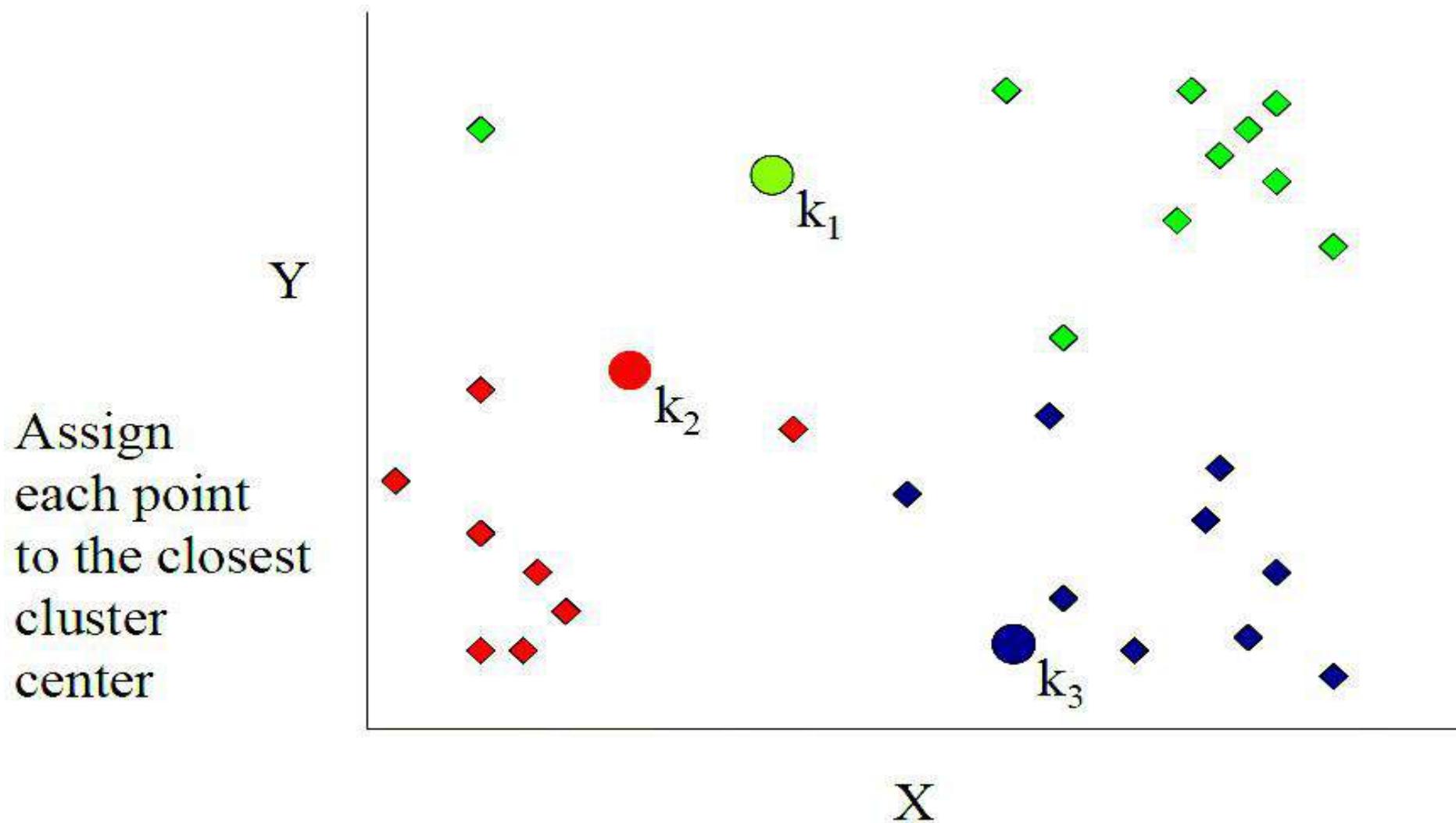


K-means example, step 1

Pick 3
initial
cluster
centers
(randomly)

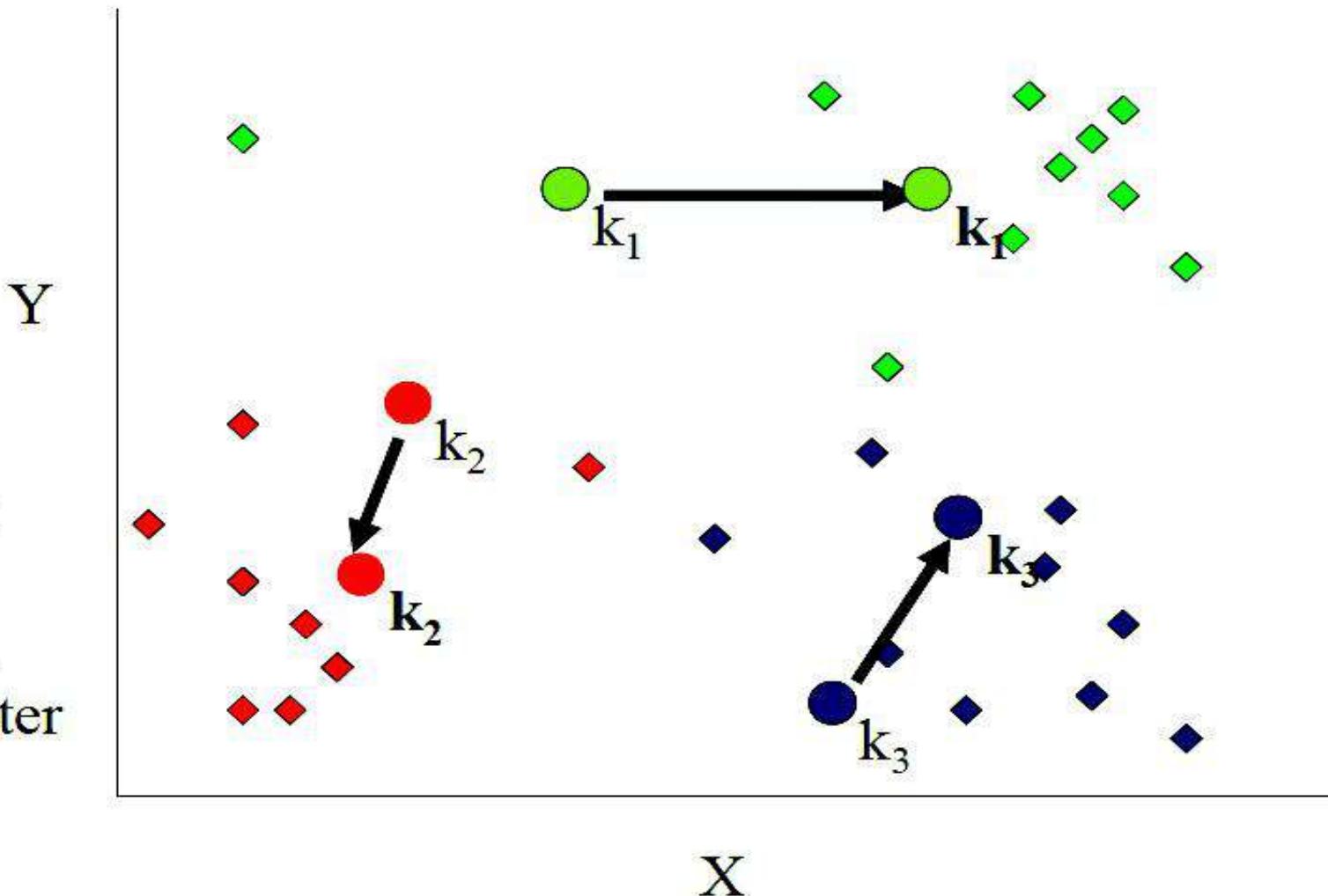


K-means example, step 2



K-means example, step 3

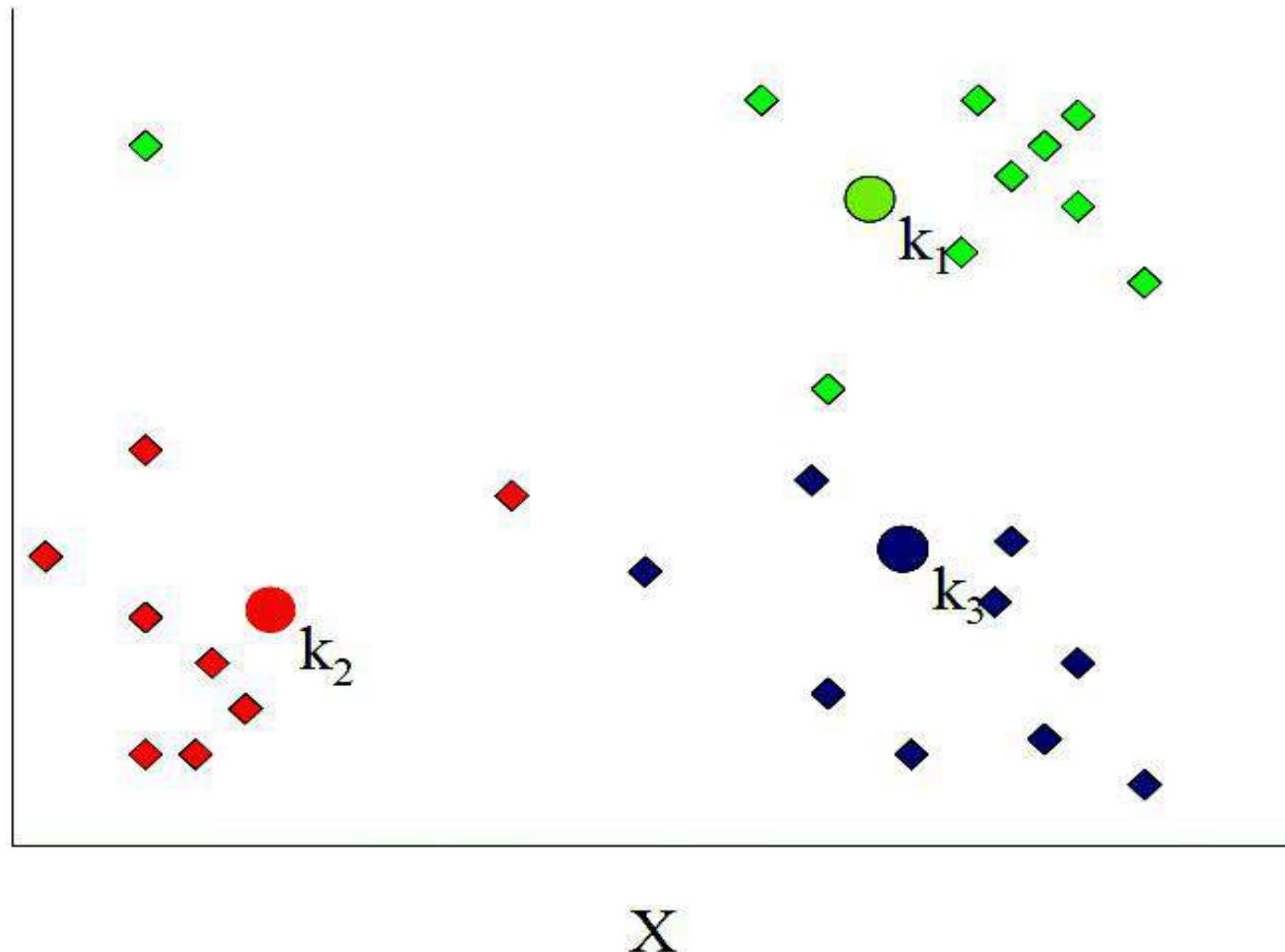
Move
each cluster
center
to the mean
of each cluster



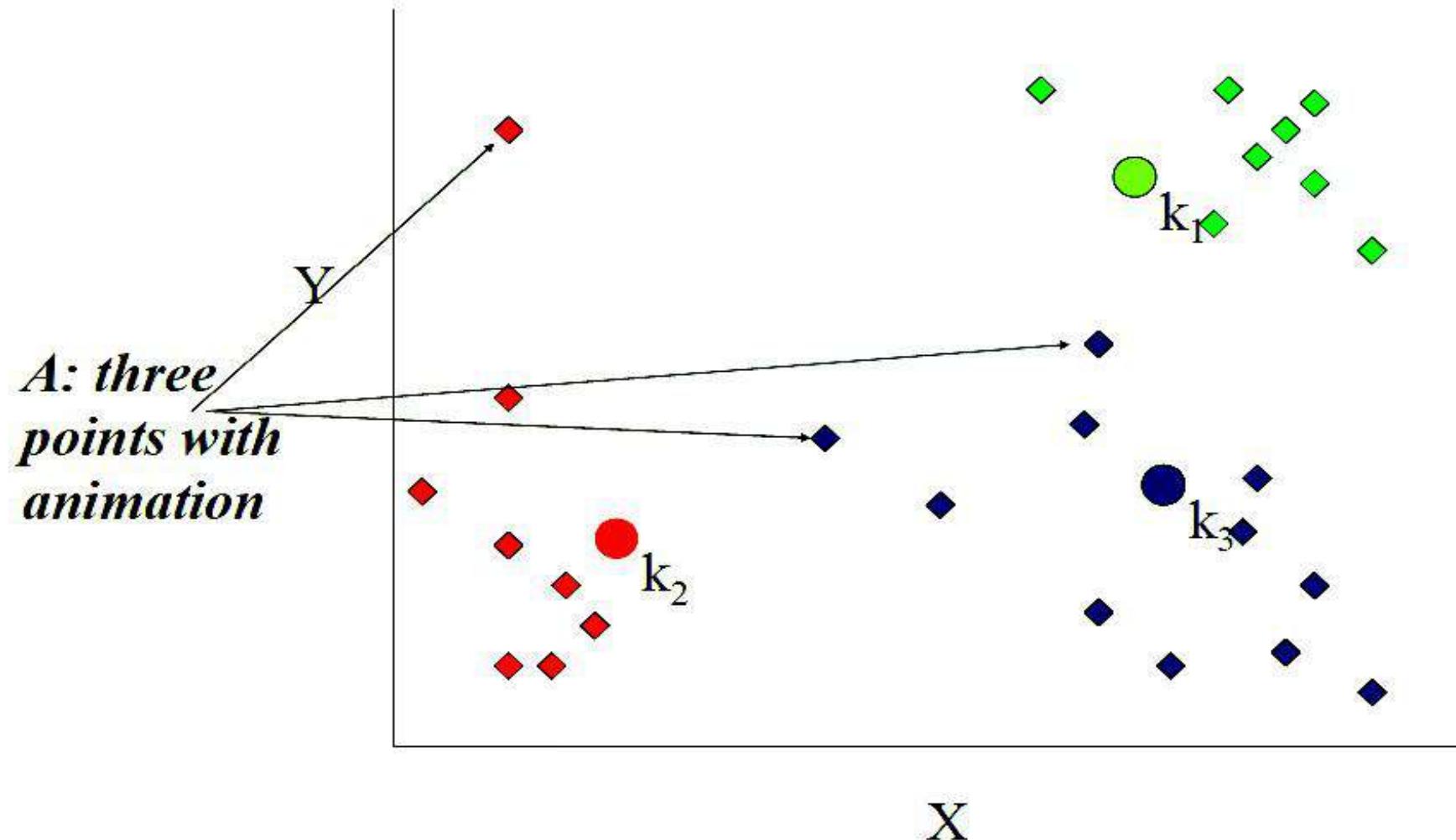
K-means example, step 4

Reassign
points
closest to a
different new
cluster center

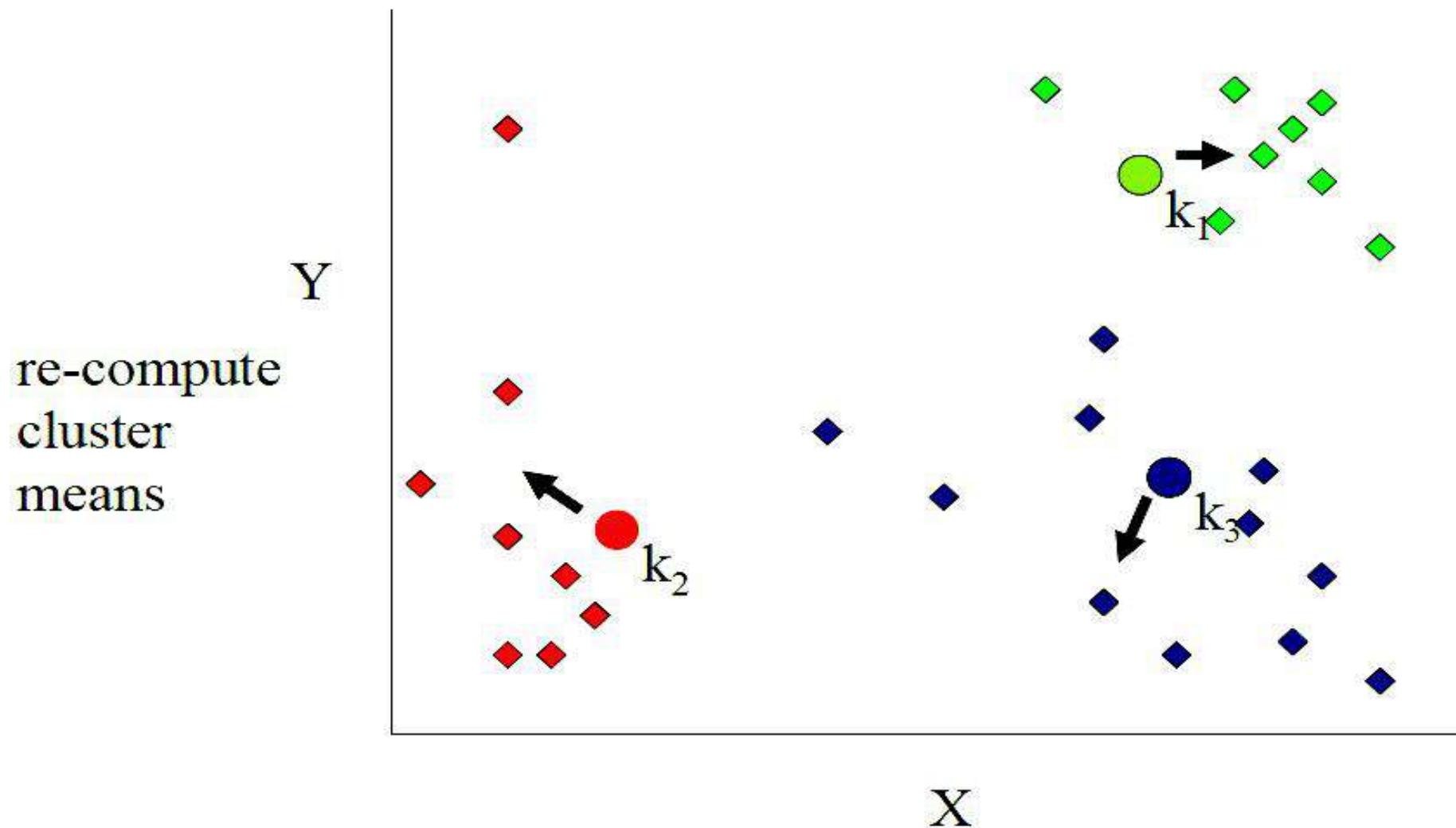
*Q: Which
points are
reassigned?*



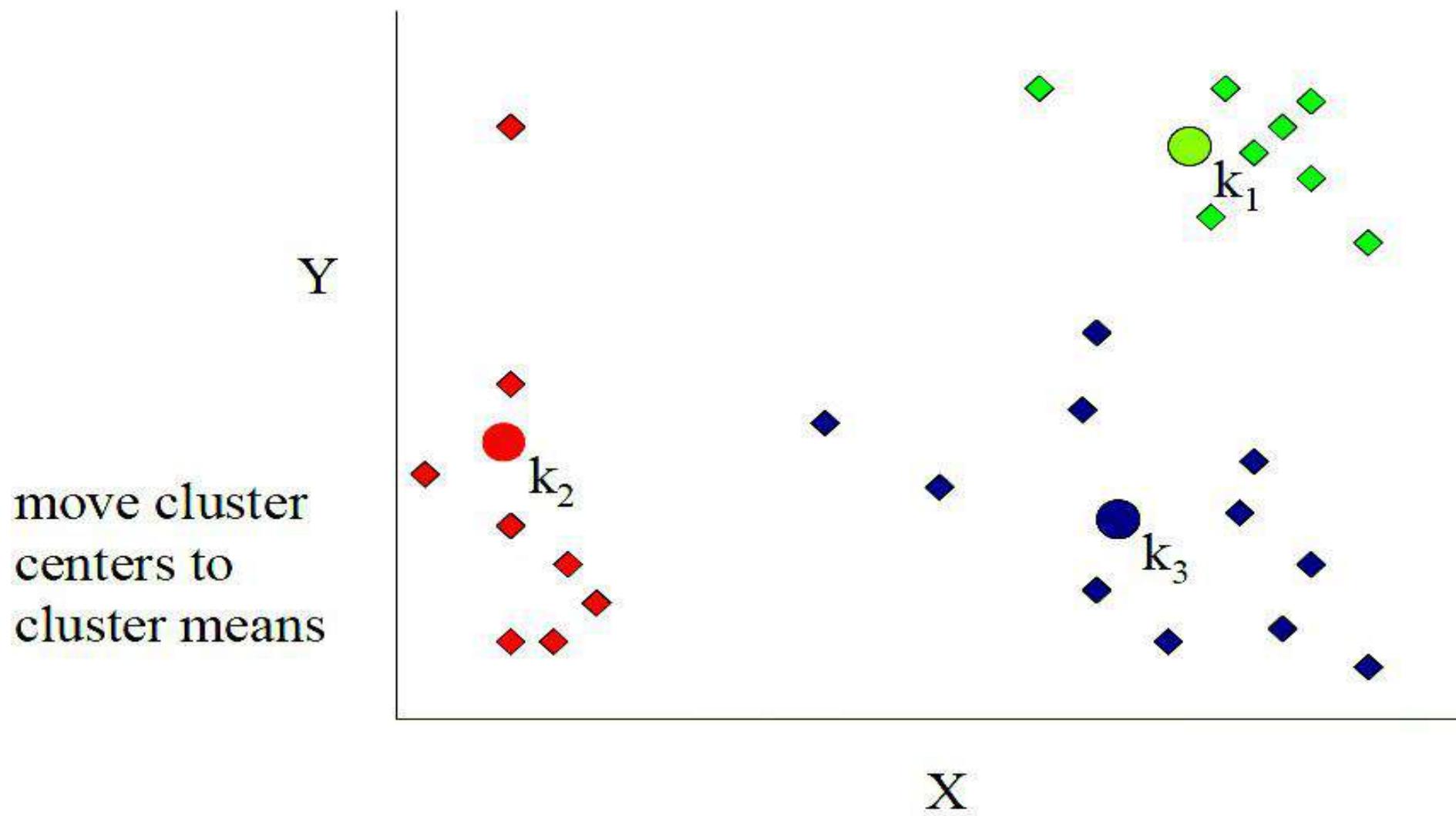
K-means example, step 4a



K-means example, step 4b



K-means example, step 5



Example:

- Apply K-mean clustering for the following data sets for two clusters. Tabulate all the assignments.

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Step-1: Given k=2, Initial Centroid

Cluster	X	Y
k1	185	72
k2	170	56

Calculate Euclidean distance using the given equation.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x - a)^2 + (y - b)^2}$$

$$\text{Cluster 1 } (185, 72) = \sqrt{(185 - 185)^2 + (72 - 72)^2} = 0$$

$$\text{Distance from Cluster 2} = \sqrt{(170 - 185)^2 + (56 - 72)^2}$$

$$\begin{aligned}(170, 56) &= \sqrt{(-15)^2 + (-16)^2} \\&= \sqrt{255 + 256} \\&= \sqrt{481} \\&= 21.93\end{aligned}$$

$$\text{Cluster 2 } (170, 56) = \sqrt{(170 - 170)^2 + (56 - 56)^2} = 0$$

Step-2: New Centroid

Cluster	Centroid		
	X	Y	ASSIGNMENT
k1	0	21.93	1
k2	21.93	0	2

Step-2: Distance calculation

Calculate Euclidean distance for the next dataset (168,60)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x - a)^2 + (y - b)^2}$$

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

$$\begin{aligned} \text{Distance from Cluster 1} &= \sqrt{(168 - 185)^2 + (60 - 72)^2} & \text{Distance from Cluster 2} &= \sqrt{(168 - 170)^2 + (60 - 56)^2} \\ (185, 72) &= \sqrt{(-17)^2 + (-12)^2} & (170, 56) &= \sqrt{(-2)^2 + (-4)^2} \\ &= \sqrt{283 + 144} & &= \sqrt{4 + 16} \\ &= \sqrt{433} & &= \sqrt{20} \\ &= 20.808 & &= 4.472 \end{aligned}$$

Dataset	Euclidean Distance			ASSIGNMENT
	Cluster 1	Cluster 2		
(168, 60)	20.808	4.472		2

Step-3: Update the cluster centroid

Cluster	X	Y
k1	185	72
k2	$= (170 + 168)/ 2$ $= 169$	$= (60+56)/ 2$ $= 58$

Step-4: Similarly process for next data set

Calculate Euclidean distance for the next dataset (179,68)

$$\text{Distance from Cluster 1} = \sqrt{(179 - 185)^2 + (68 - 72)^2}$$

$$\begin{aligned}(185,72) &= \sqrt{(-6)^2 + (-4)^2} \\ &= \sqrt{36 + 16} \\ &= \sqrt{52} \\ &= 7.211103\end{aligned}$$

$$\text{Distance from Cluster 2} = \sqrt{(179 - 169)^2 + (68 - 58)^2}$$

$$\begin{aligned}(169,58) &= \sqrt{(10)^2 + (10)^2} \\ &= \sqrt{100 + 100} \\ &= \sqrt{200} \\ &= 14.14214\end{aligned}$$

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Dataset	Euclidean Distance		ASSIGNMENT
	Cluster 1	Cluster 2	
(179,68)	7.211103	14.14214	1

Step-5: Update the cluster centroid

Cluster	X	Y
k1	= $185+179/2$ = 182	= $72+68/2$ = 70
k2	169	58

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (182,72)

$$\text{Distance from Cluster 1} = \sqrt{(182 - 182)^2 + (72 - 70)^2}$$

$$(182,70) = \sqrt{(0)^2 + (2)^2}$$

$$= \sqrt{0 + 4}$$

$$= \sqrt{4}$$

$$= 2$$

$$\text{Distance from Cluster 2} = \sqrt{(182 - 169)^2 + (72 - 58)^2}$$

$$(169,58) = \sqrt{(13)^2 + (14)^2}$$

$$= \sqrt{169 + 196}$$

$$= \sqrt{365}$$

$$= 19.10$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(182,72)	2	19.10	1

Step-6: Update the cluster centroid

Cluster	X	Y
k1	$= 182+182/2$ $= 182$	$= 70+72/2$ $= 71$
k2	169	58

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (188,77)

$$\begin{aligned} \text{Distance from Cluster 1} &= \sqrt{(188 - 182)^2 + (77 - 71)^2} \\ (182, 71) &= \sqrt{(6)^2 + (6)^2} \\ &= \sqrt{36 + 36} \\ &= \sqrt{72} \\ &= 8.4852 \end{aligned}$$

$$\begin{aligned} \text{Distance from Cluster 2} &= \sqrt{(188 - 169)^2 + (77 - 58)^2} \\ (169, 58) &= \sqrt{(19)^2 + (19)^2} \\ &= \sqrt{361 + 361} \\ &= \sqrt{722} \\ &= 26.87 \end{aligned}$$

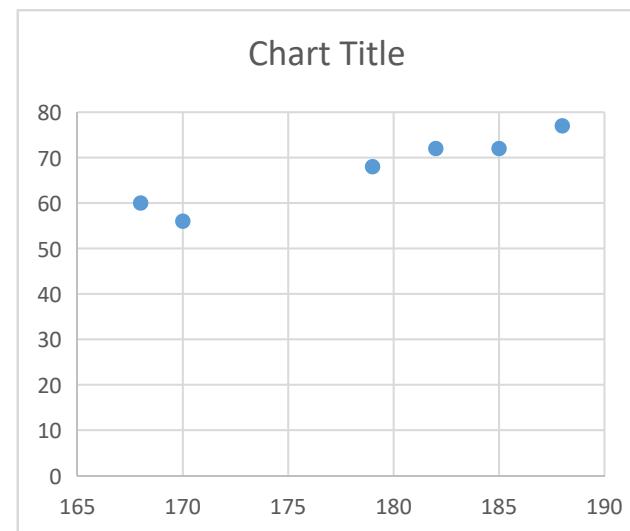
Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(188,77)	8.4852	26.87	1

Step-7: Update the cluster centroid

Cluster	X	Y
k1	$= 182+188/2$ $= 185$	$= 71+77/2$ $= 74$
k2	169	58

Final Assignment

Dataset No	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	77	1



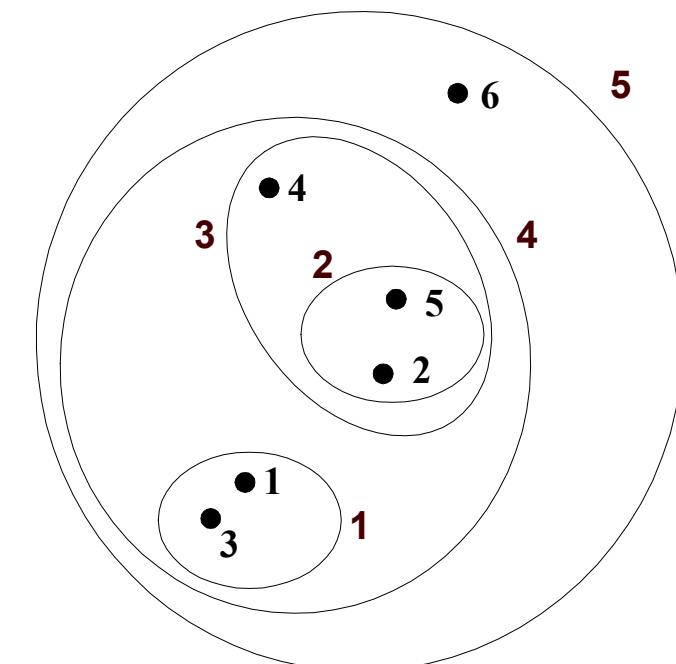
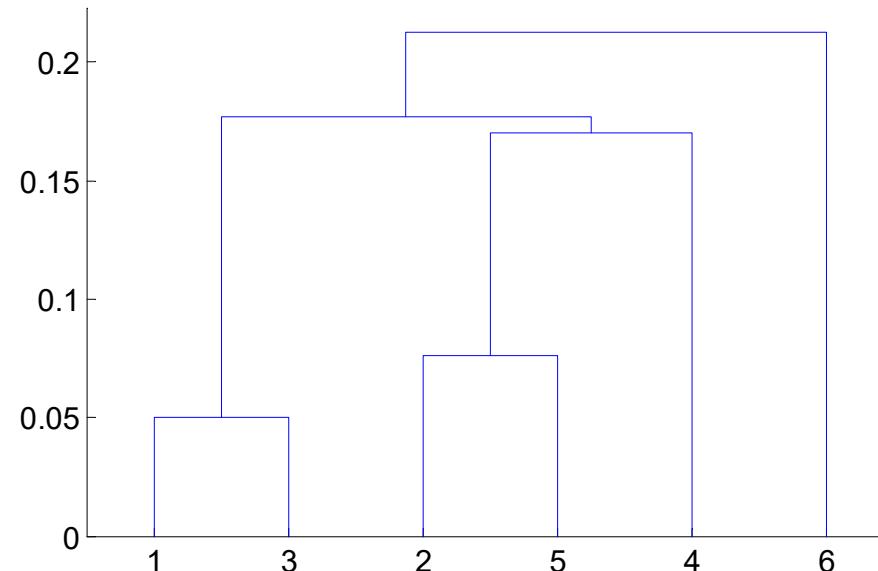
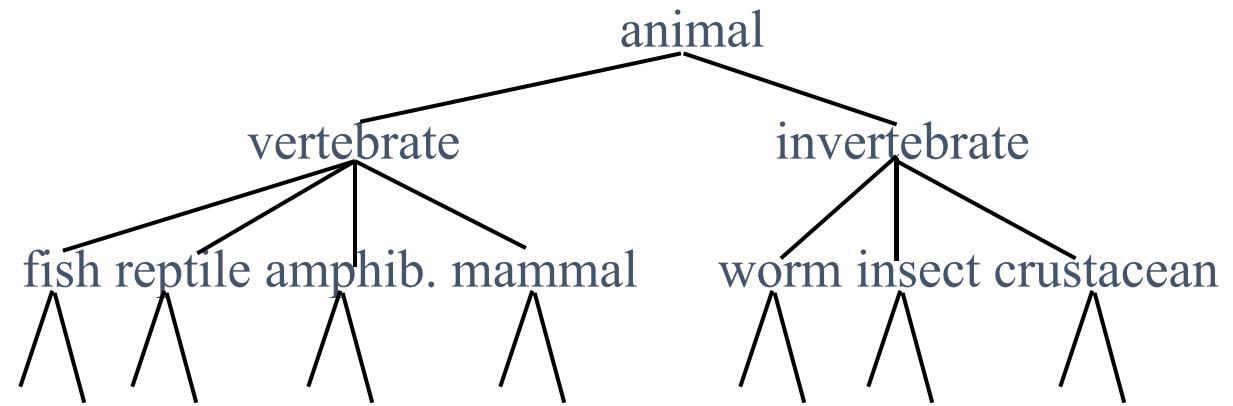
Hierarchical Clustering

Typical Alternatives to Calculate the Distance between Clusters

- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- **Average:** avg distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- **Centroid:** distance between the centroids of two clusters, i.e., $\text{dis}(K_i, K_j) = \text{dis}(C_i, C_j)$
- **Medoid:** distance between the medoids of two clusters, i.e., $\text{dis}(K_i, K_j) = \text{dis}(M_i, M_j)$
 - Medoid: one chosen, centrally located object in the cluster

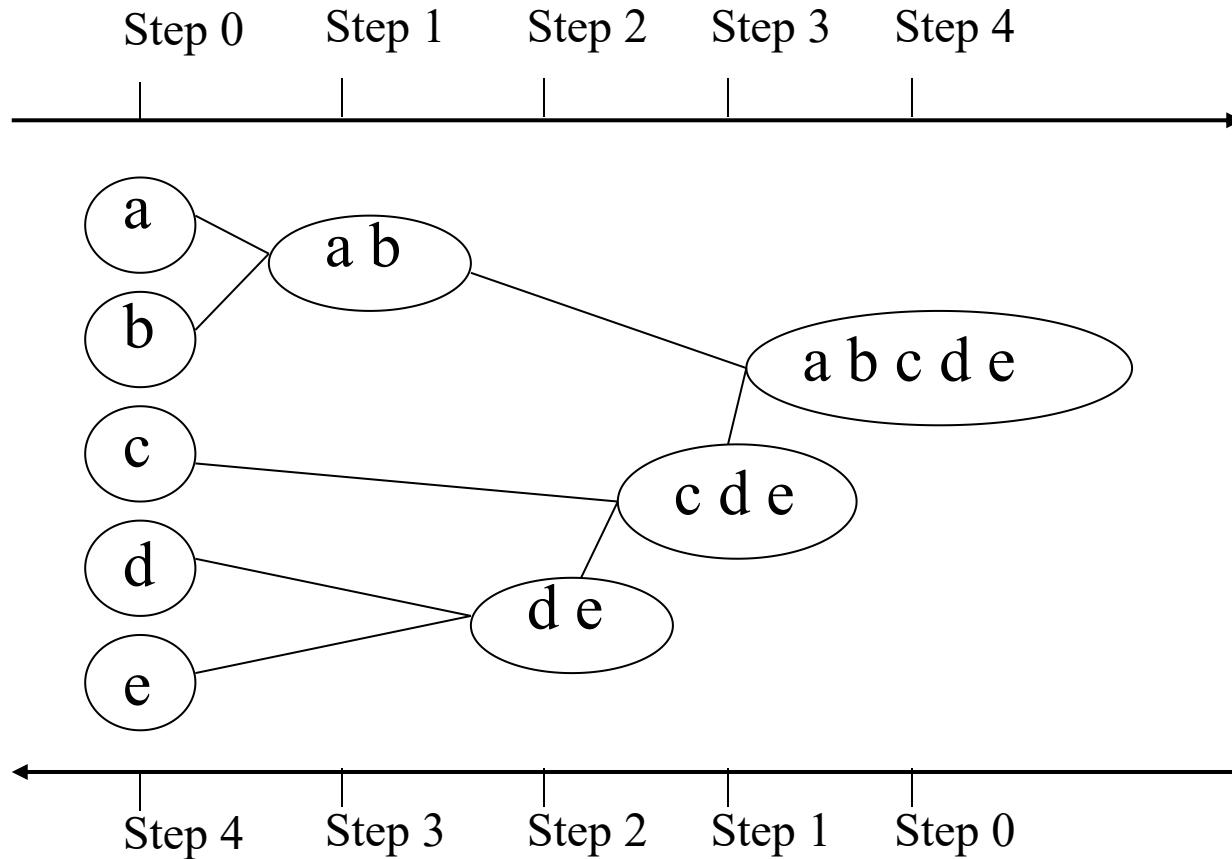
Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree from a set of unlabeled examples
- Can be visualized as a dendrogram
 - A tree like diagram that records the sequences of merges or splits



Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition



Agglomerative (bottom-up): (AGNES)

Start with each document being a single cluster.

Eventually all documents belong to the same cluster.

Divisive (top-down): (DIANA)

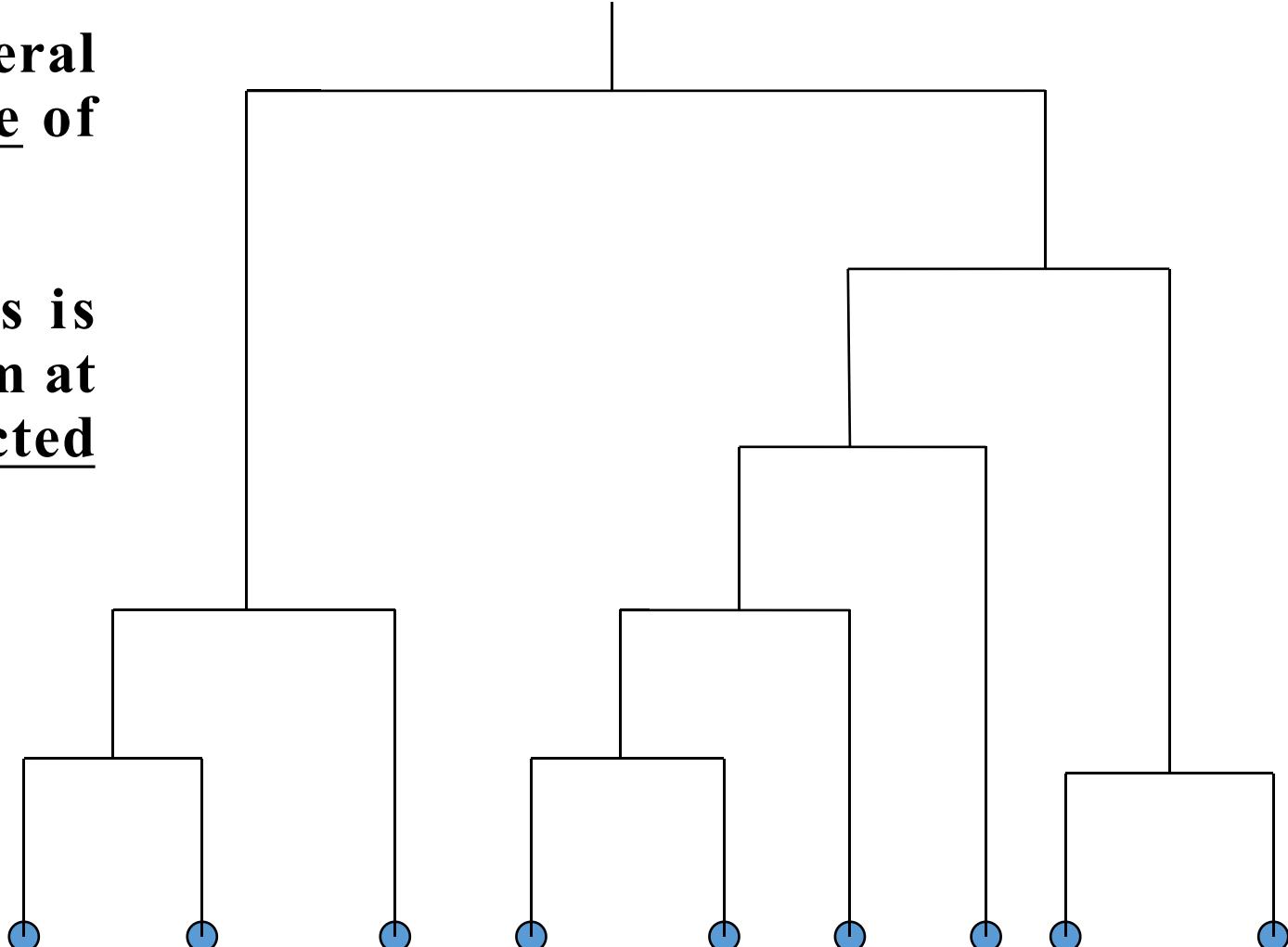
Start with all documents belong to the same cluster.

Eventually each node forms a cluster on its own.

Dendrogram: Shows How the Clusters are Merged

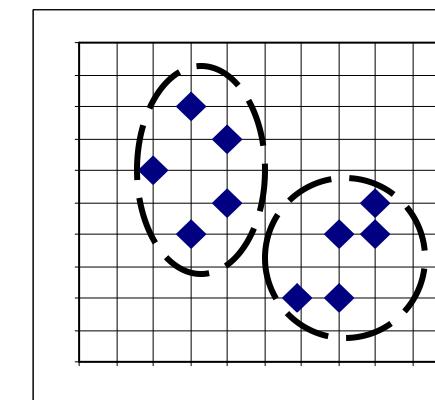
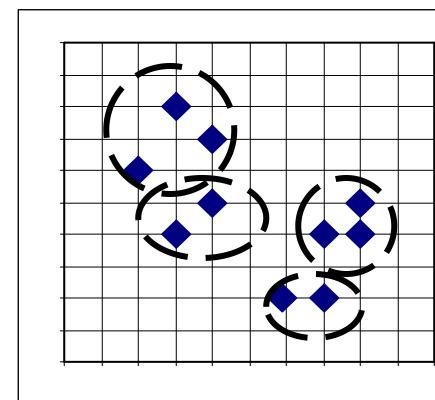
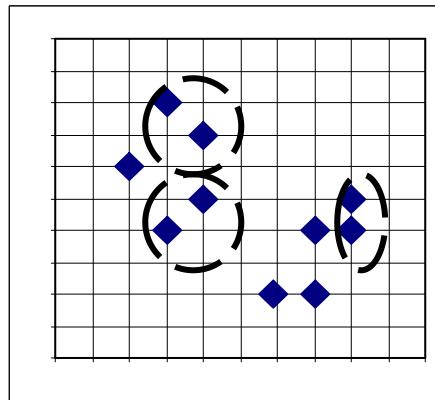
Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



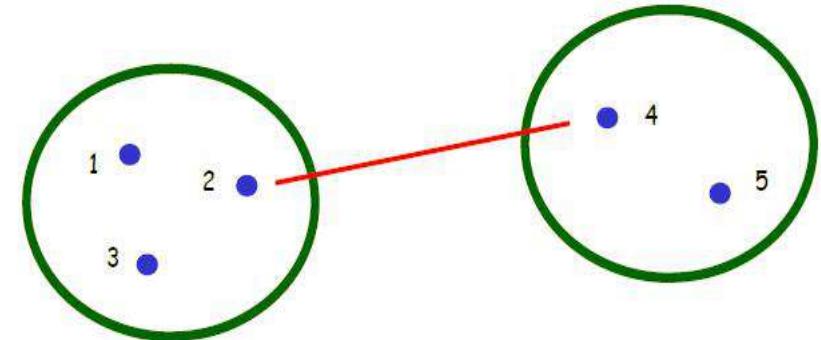
Agglomerative Algorithm

- Step1: Make each object as a cluster
- Step2: Calculate the Euclidean distance from every point to every other point. i.e., construct a Distance Matrix
- Step3: Identify two clusters with shortest distance.
 - Merge them
 - Go to Step 2
 - Repeat until all objects are in one cluster

Hierarchical Algorithms

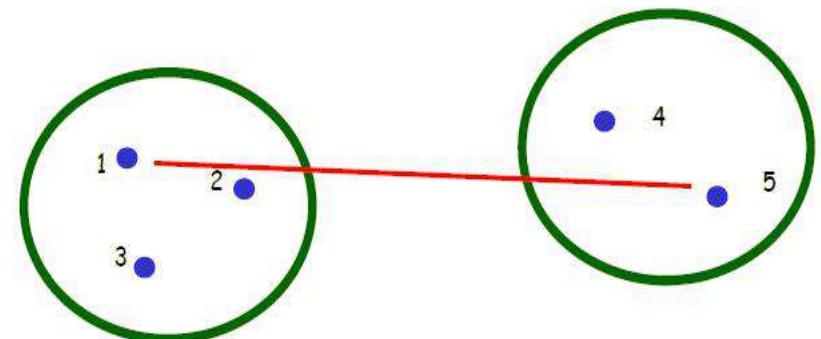
- **Single-link**

- Distance between two clusters set equal to the *minimum* of distances between all instances
- Single link (nearest neighbour). The distance between two clusters is determined by the distance of the **two closest objects** (nearest neighbours) in the different clusters.



- **Complete-link**

- Distance between two clusters set equal to *maximum* of all distances between instances in the clusters
- Complete link (furthest neighbour). The distances between clusters are determined by the **greatest distance** between any two objects in the different clusters (i.e., by the "furthest neighbours").
- Tightly bound, compact clusters



Hierarchical Algorithms cont..

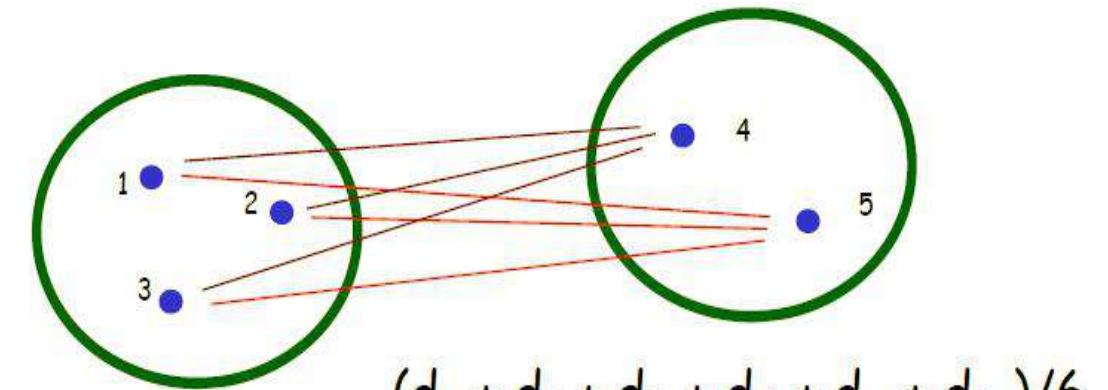
Pair-group average.

The distance between two clusters is calculated as the **average distance** between all pairs of objects in the two different clusters.

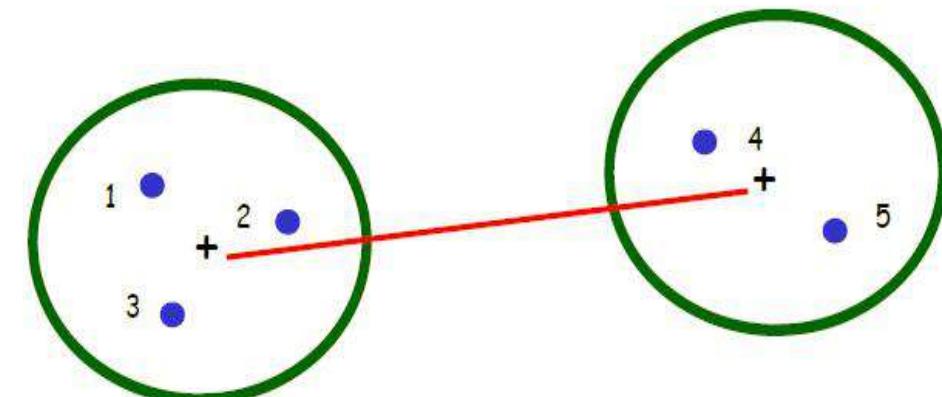
This method is also very efficient when the objects form natural distinct "clumps," however, it performs equally well with elongated, "chain" type clusters.

Pair-group centroid.

The distance between two clusters is determined as the distance **between centroids**.



$$(d_1 + d_2 + d_3 + d_4 + d_5 + d_6)/6$$



Single Link Agglomerative Clustering

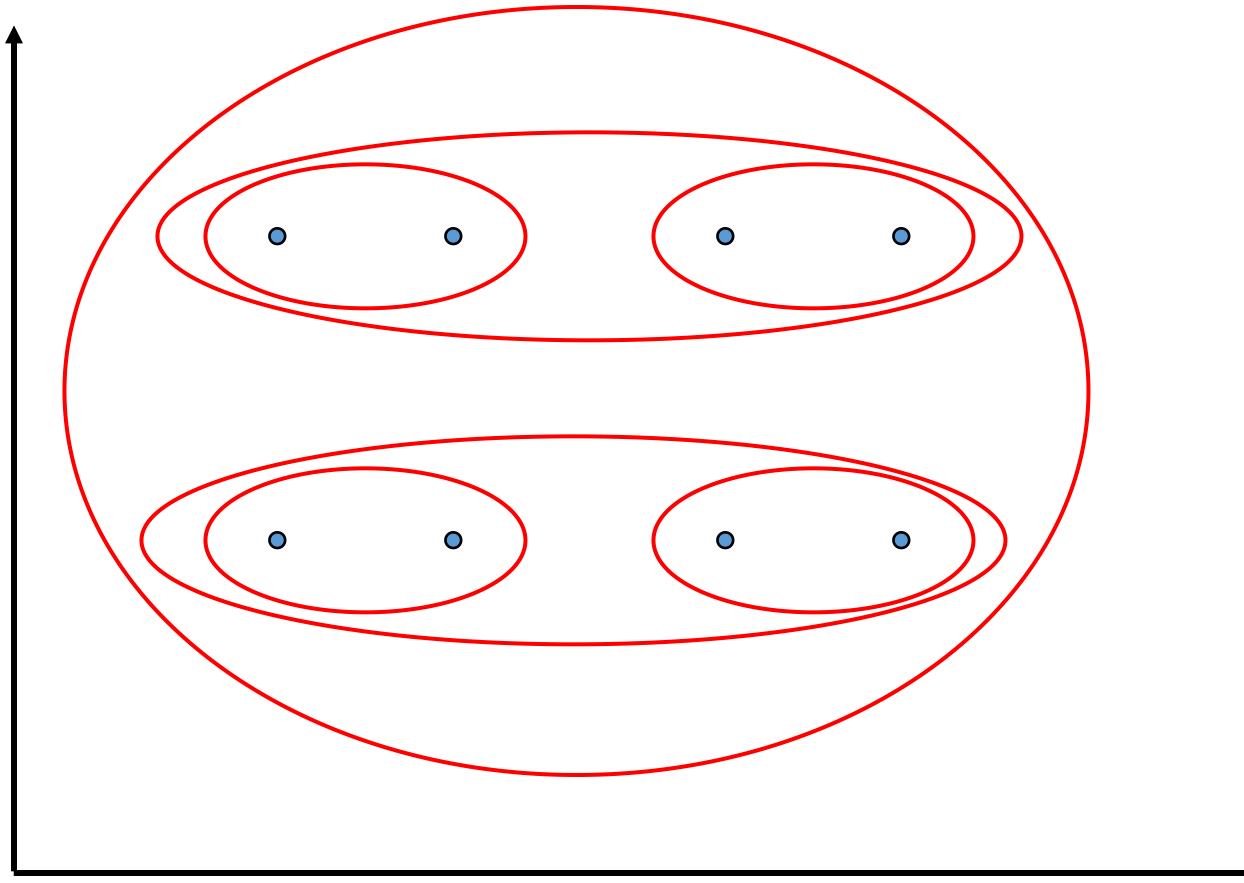
- Use maximum similarity of pairs:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Can result in “straggly” (long and thin) clusters due to chaining effect.
 - Appropriate in some domains, such as clustering islands: “Hawai’i clusters”
- After merging c_i and c_j , the similarity of the resulting cluster to another cluster, c_k , is:

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

Single Link Example



Complete Link Agglomerative Clustering

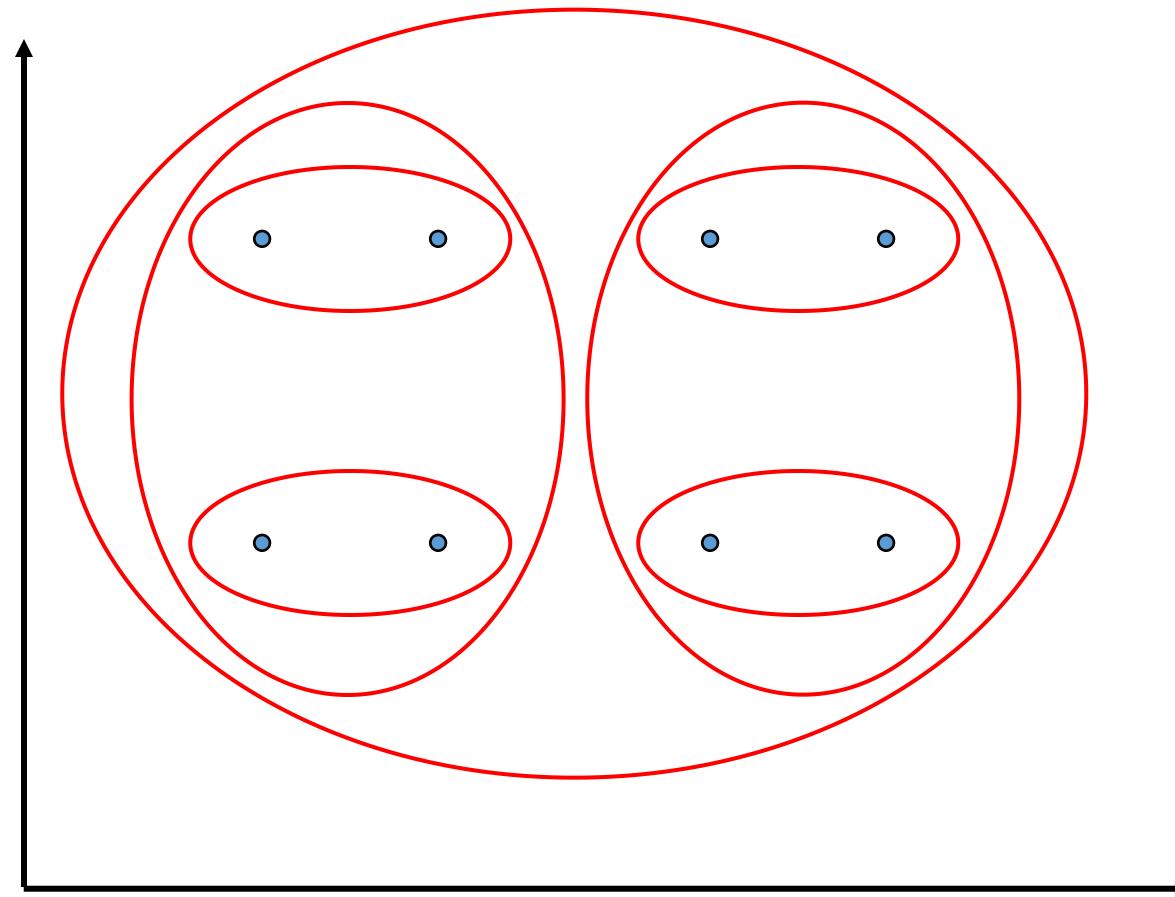
- Use minimum similarity of pairs:

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Makes “tighter,” spherical clusters that are typically preferable.
- After merging c_i and c_j , the similarity of the resulting cluster to another cluster, c_k , is:

$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$

Complete Link Example



Simple Example

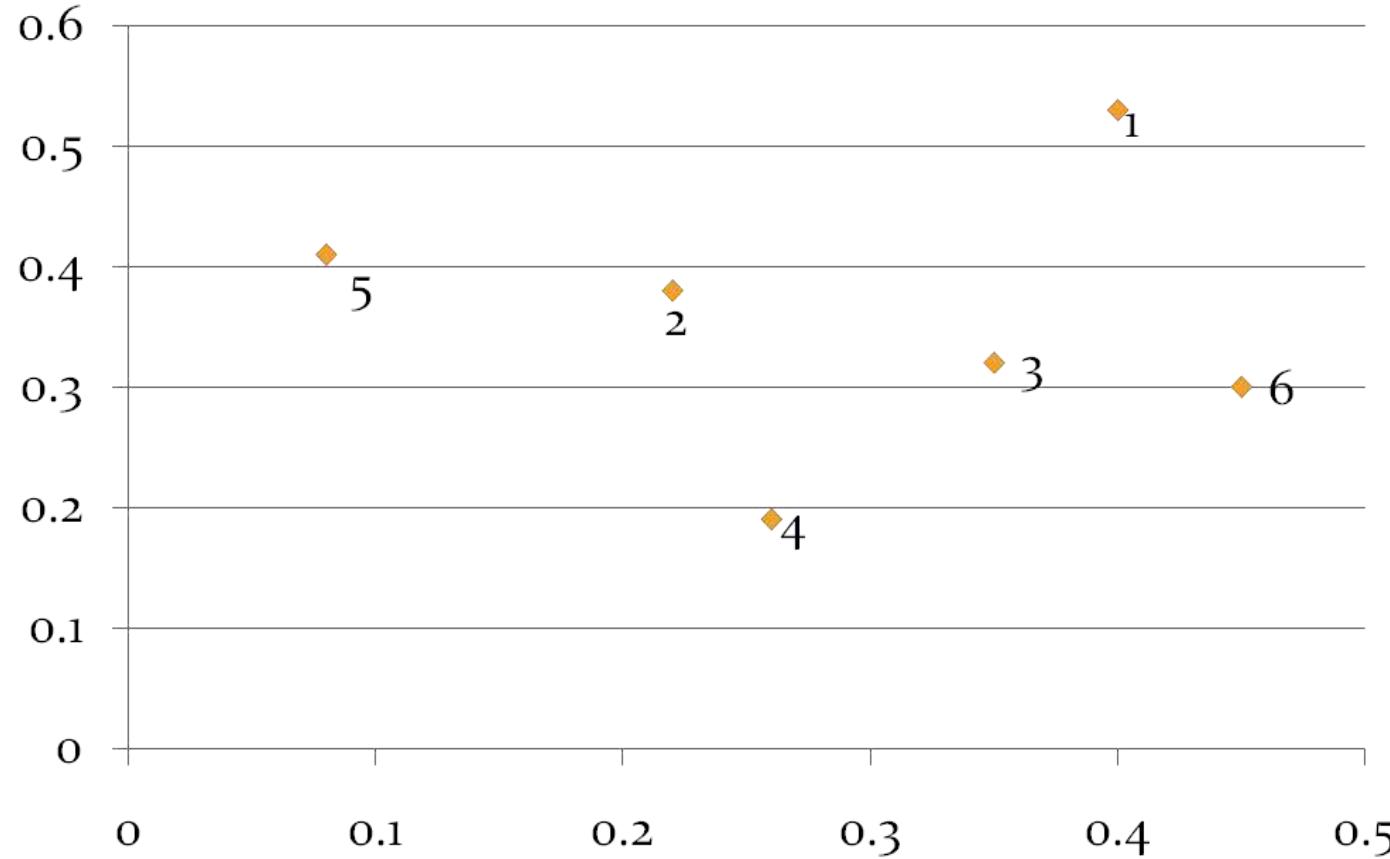
Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Another Example

- Find single link technique to find clusters in the given database.

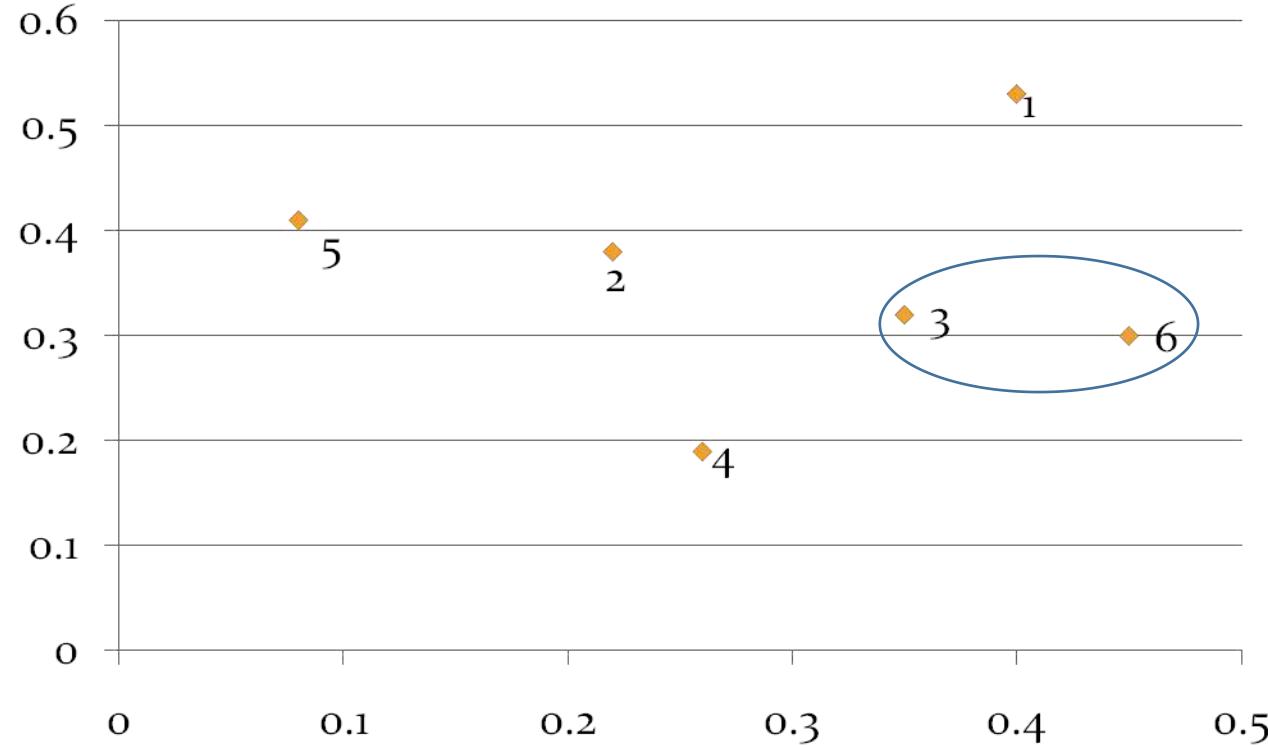
	x	y
1	0.4	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.3

Plot given data



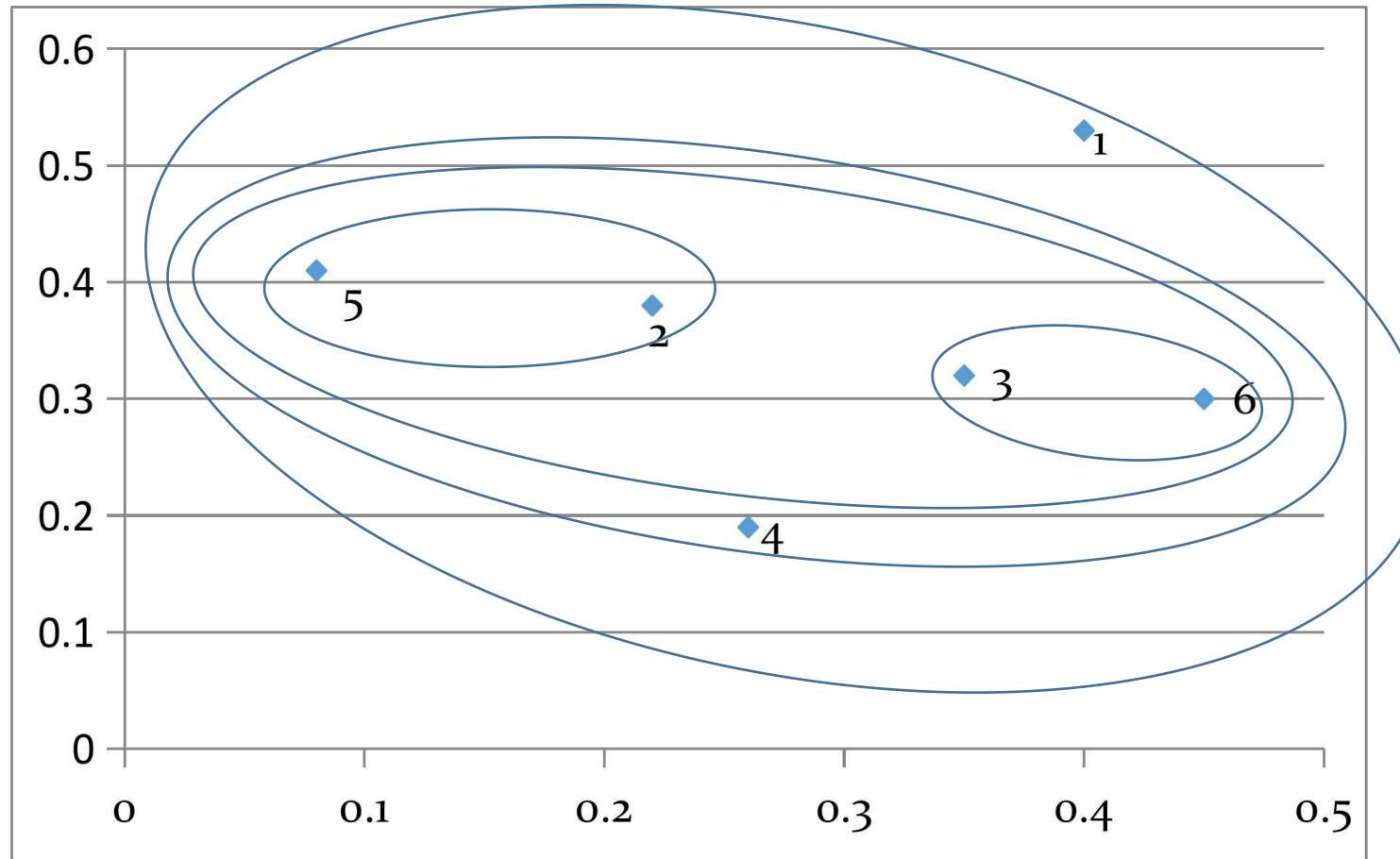
	X	Y
1	0.4	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.30

Identify two nearest clusters



	X	Y
1	0.4	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.3

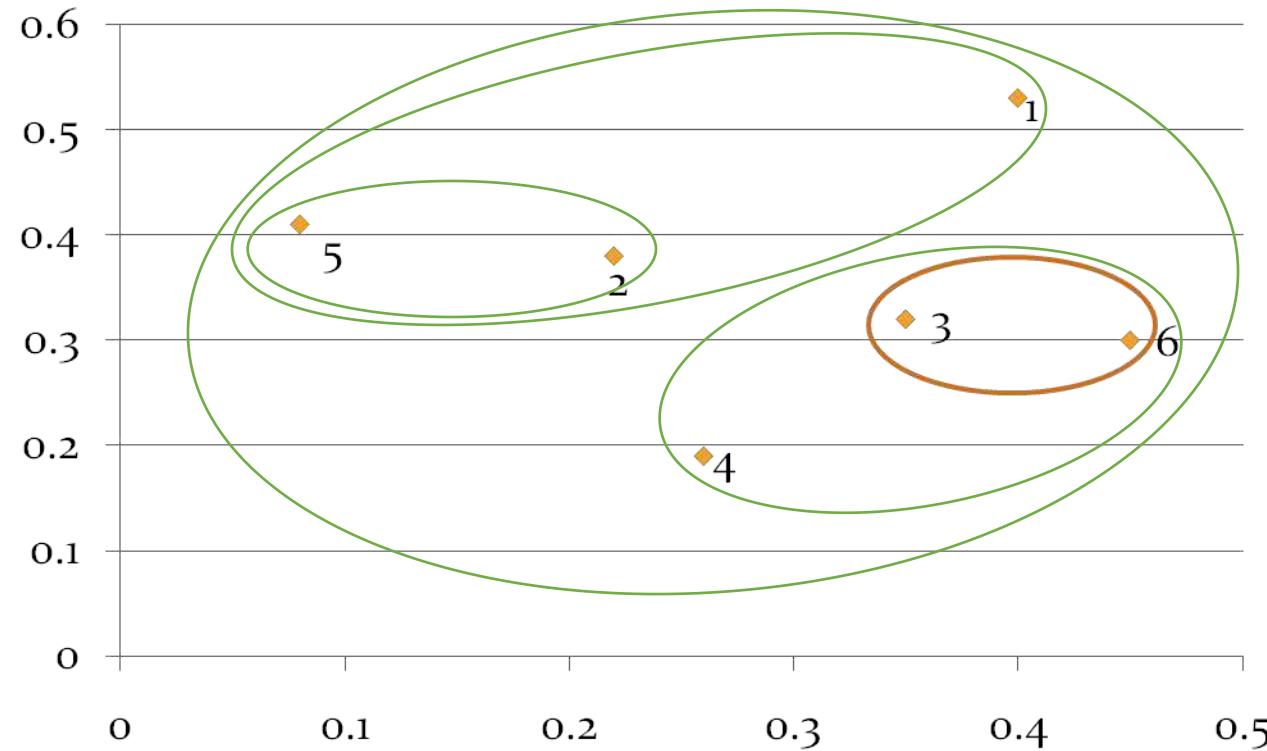
Repeat process until all objects in same cluster



	X	Y
1	0.4	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.3

Average link

- Average distance matrix



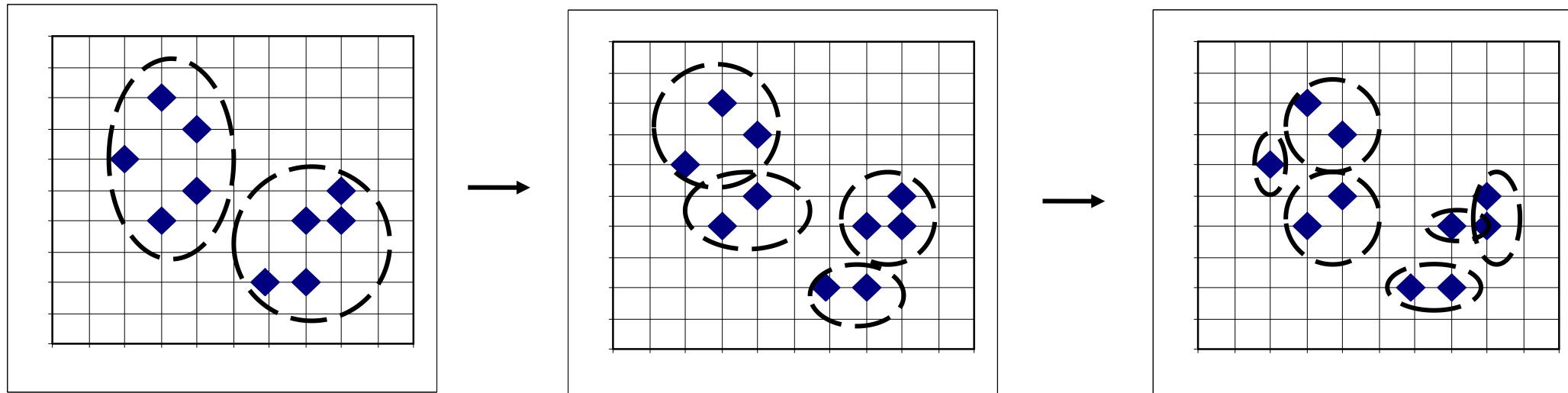
	X	Y
1	0.4	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.3

Construct a distance matrix

	1	2	3	4	5	6
1	0					
2	0.24	0				
3	0.22	0.15	0			
4	0.37	0.2	0.15	0		
5	0.34	0.14	0.28	0.29	0	
6	0.23	0.25	0.11	0.22	0.39	0

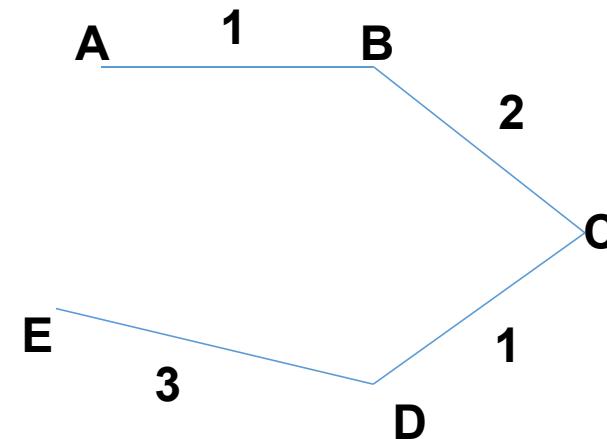
DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



Divisive Clustering

- All items are initially placed in one cluster
- The clusters are repeatedly split in two until all items are in their own cluster



Difficulties in Hierarchical Clustering

- Difficulties regarding the selection of merge or split points
- This decision is critical because the further merge or split decisions are based on the newly formed clusters
- Method does not scale well
- So hierarchical methods are integrated with other clustering techniques to form multiple-phase clustering

DATA ANALYTICS

Lecture-2

Dr. H.K.Tripathy

What is Data Analytics?

- The increase in size of the data has lead to a rise in need for carrying out inspecting, cleaning, transforming and modelling data with the goal of discovering useful information to gain insights from the data, suggesting conclusions and supporting decision-making.
- Intelligent data analysis (IDA) uses concepts from artificial intelligence, information retrieval, machine learning, pattern recognition, visualization, distributed programming.
- The process of IDA typically consists of the following three stages:
 - Data preparation
 - Data mining and rule finding
 - Result validation and interpretation
- It has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science and social science domains.

What is Data Analytics?

In Statistical applications, business analytics can be divided into 2 types

Exploratory Data Analysis (EDA)

Confirmatory Data Analysis (CDA)

- *It focuses on discovering new features in the data*
- *Exploratory Data Analysis involves things like: establishing the data's underlying structure, identifying mistakes and missing data, establishing the key variables, spotting anomalies, checking assumptions and testing hypotheses in relation to a specific model, estimating parameters.*

- *It focuses on confirming or falsifying existing hypotheses using traditional statistical tools such as significance, inference, and confidence.*
- *CDA involves processes like testing hypotheses, producing estimates, regression analysis (estimating the relationship between variables) and variance analysis (evaluating the difference between the planned and actual outcome).*

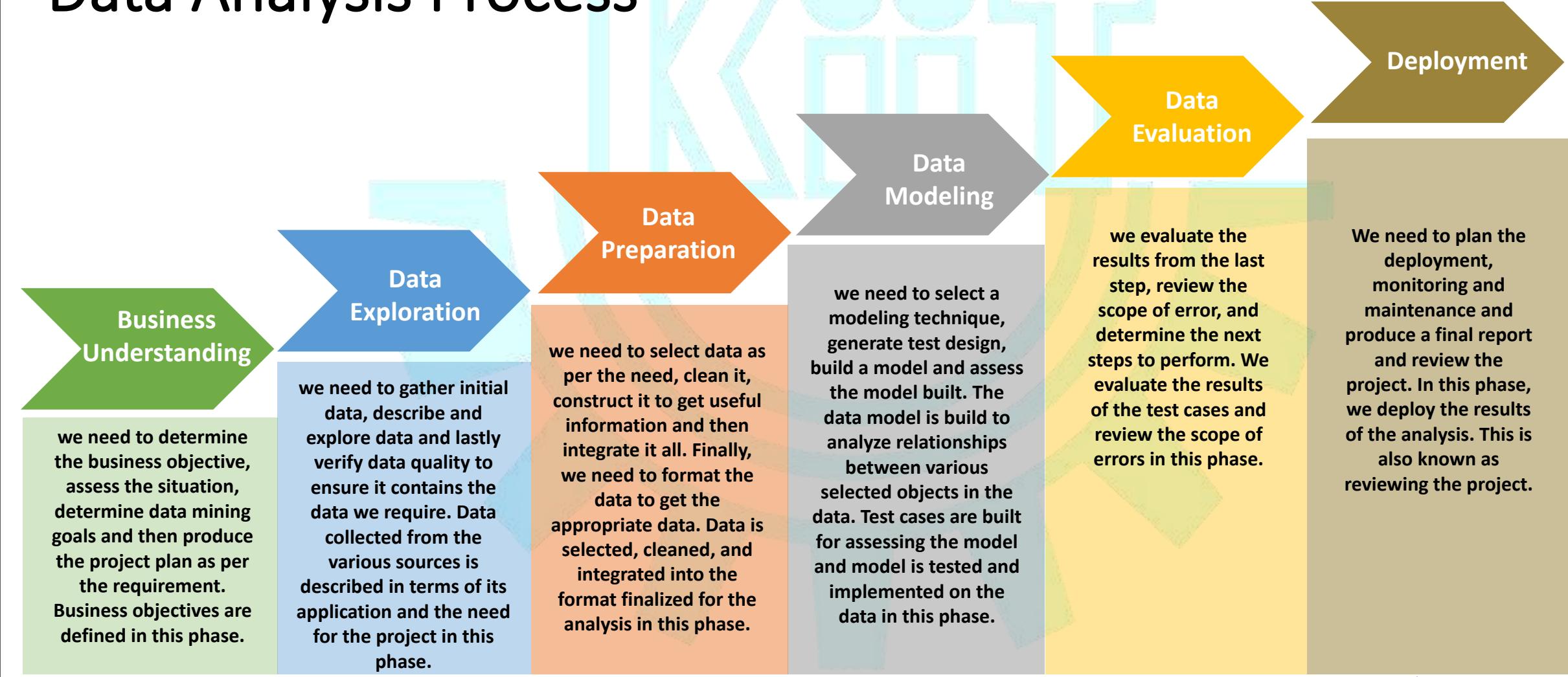
Importance of Data Analysis

- Data analysis offers the following benefits:
 - Structuring the findings from survey research or other means of data collection
 - Provides a picture of data at several levels of granularity from a macro picture into a micro one
 - Acquiring meaningful insights from the data set which can be effectively exploited to take some critical decisions to improve productivity
 - Helps to remove human bias in decision making, through proper statistical treatment
 - With the advent of big data, it is even more vital to find a way to analyze the ever (faster) growing disparate data coursing through their environments and give it meaning

Data Analytics Applications

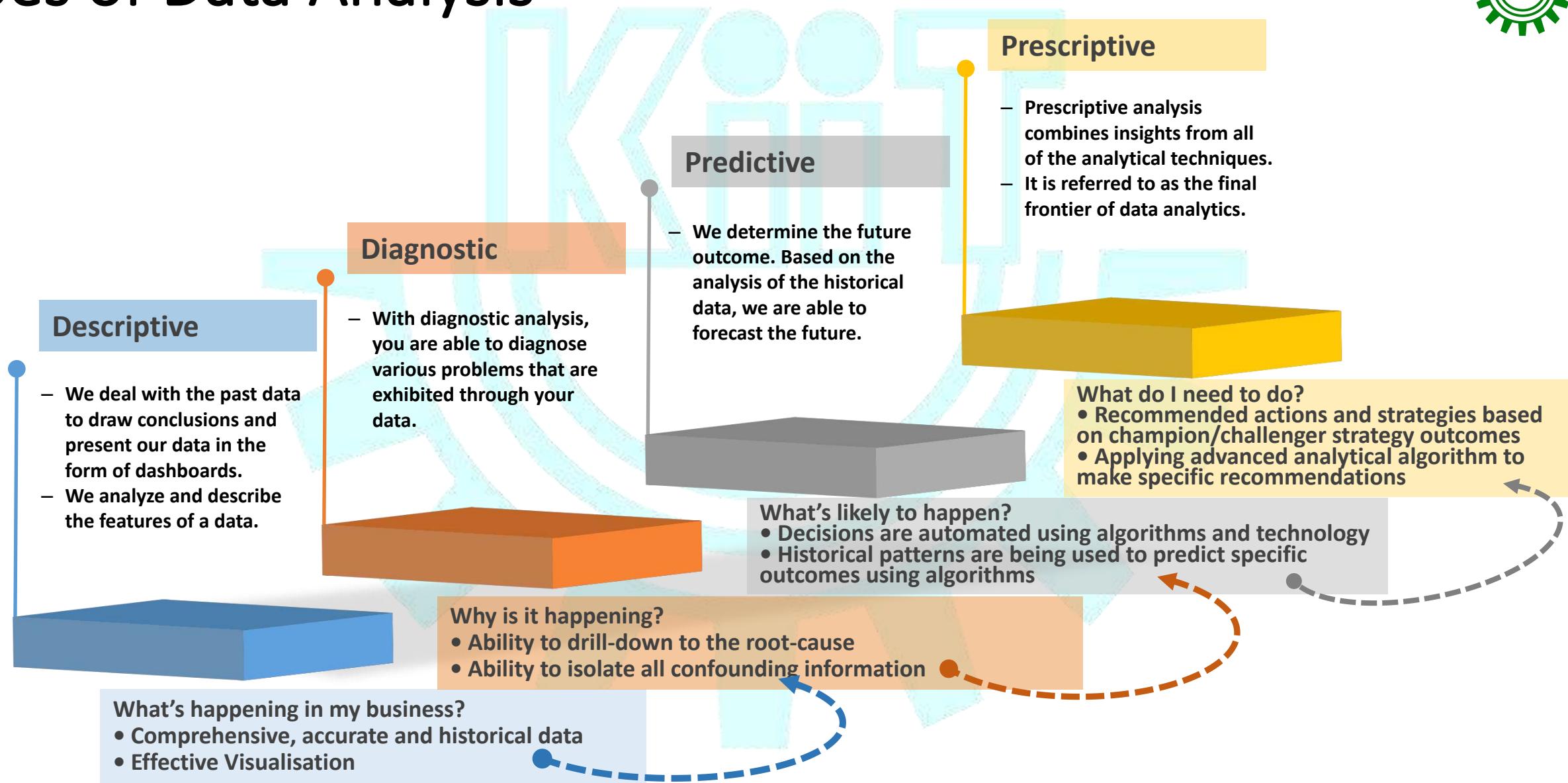
- Understanding and targeting customers
- Understanding and optimizing business processes
- Personal quantification and performance optimization
- Improving healthcare and public health
- Improving sports performance
- Improving science and research
- Optimizing machine and device performance
- Improving security and law enforcement
- Improving and optimizing cities and countries
- Financial trading

Data Analysis Process



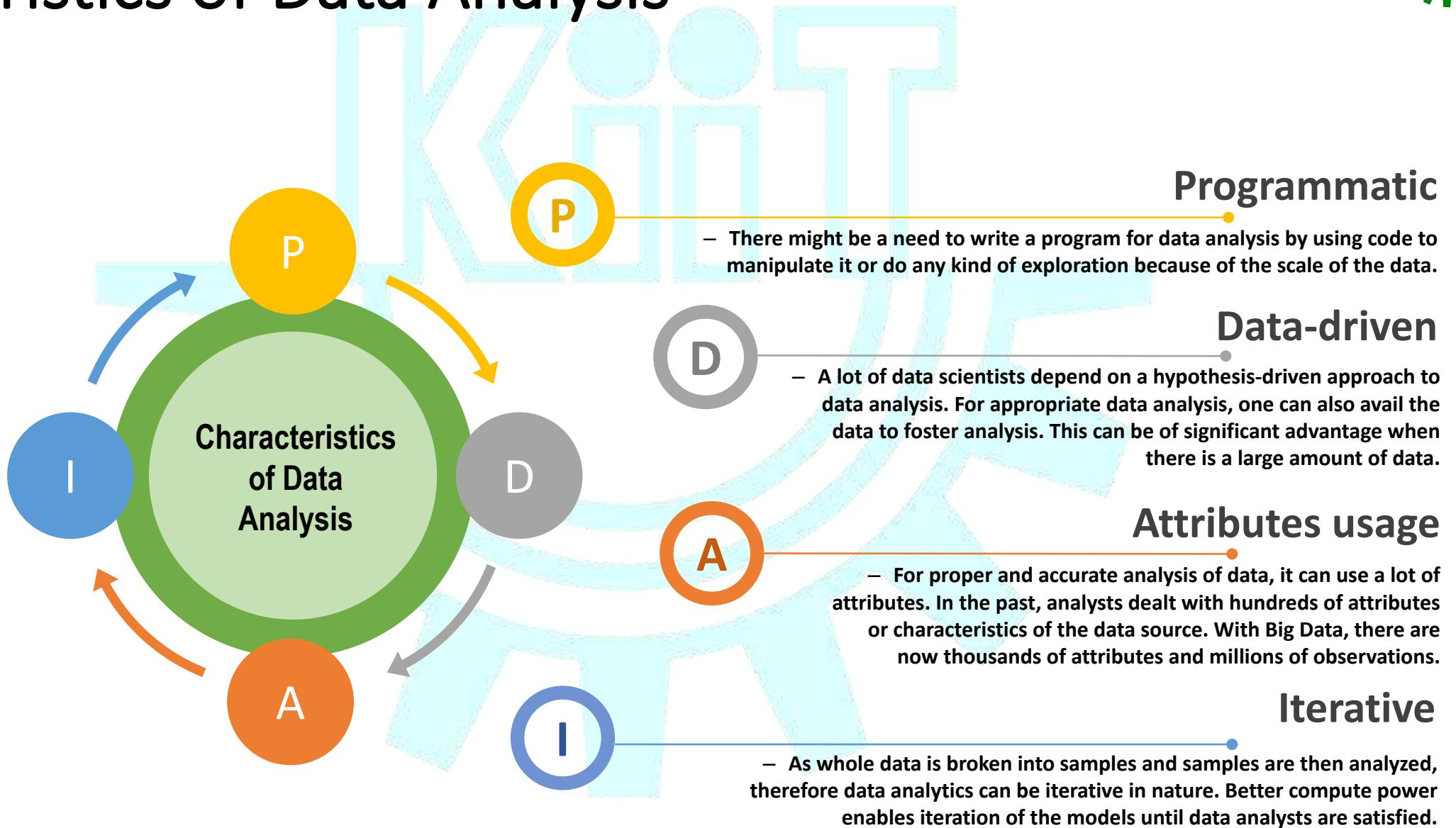
The complete process is known as Business Analytics Process

Types of Data Analysis



Characteristics of Data Analysis

The characteristics of the data analysis depend on different aspects such as volume, velocity, and variety.



How to Get a Better Analysis?

In order to have a great analysis, it is necessary to ask the right question, gather the right data to address it, and design the right analysis to answer the question. Only after careful analysis, we can define it as correct.

Statistical Significance

It states how is the problem statistically important for decision making. Statistical significance testing takes some assumptions and determines the probability of happening of results if the assumptions are correct.

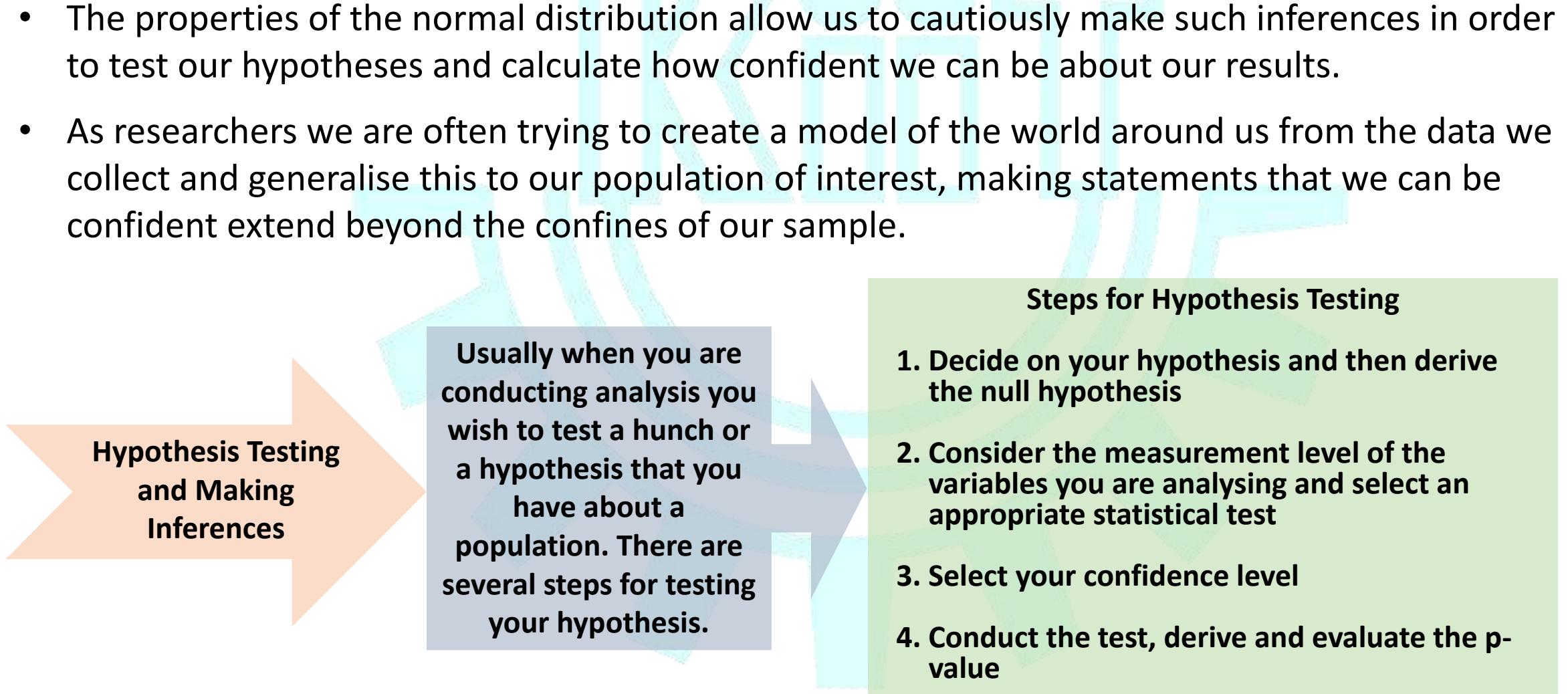
How to Get a Better Analysis?

Business Importance

It means how the problem is related to business and its importance. We will assign the results in the business context as part of the final process of validation.

Probability and Inferential Statistics

- The properties of the normal distribution allow us to cautiously make such inferences in order to test our hypotheses and calculate how confident we can be about our results.
- As researchers we are often trying to create a model of the world around us from the data we collect and generalise this to our population of interest, making statements that we can be confident extend beyond the confines of our sample.



Hypothesis Testing
and Making
Inferences

Usually when you are conducting analysis you wish to test a hunch or a hypothesis that you have about a population. There are several steps for testing your hypothesis.

Steps for Hypothesis Testing

1. Decide on your hypothesis and then derive the null hypothesis
2. Consider the measurement level of the variables you are analysing and select an appropriate statistical test
3. Select your confidence level
4. Conduct the test, derive and evaluate the p-value

Probability and Inferential Statistics

Let's start by talking about hypotheses.

- You probably noticed that there are two types of hypothesis mentioned in these steps; your initial hypothesis (often called the **alternate hypothesis**) and something called the **null hypothesis**.
- In order to explain these, let us take an example of a specific research question:

Do girls have higher educational achievement than boys at age 14?

Alternate hypothesis:

There is a relationship between gender and age 14 test score.

Null hypothesis:

There is no relationship between gender and age 14 test score. This is the default assumption (even if you do not think it is true!).

Statistical Significance - What is a P-value?

- A p-value is a probability. It is usually expressed as a proportion which can also be easily interpreted as a percentage.
- P-values become important when we are looking to ascertain how confident we can be in accepting or rejecting our hypotheses.
- Because we only have data from a sample of individual cases and not the entire population we can never be absolutely (100%) sure that the alternative hypothesis is true.
- However, by using the properties of the normal distribution we can compute the probability that the result we observed in our sample could have occurred by chance.
- The way that the p-value is calculated varies subtly between different statistical tests, which each generate a test statistic (called, for example, t, F or X^2 depending on the particular test).

Sampling Distribution

More precisely, sampling distributions are probability distributions and used **to describe the variability** of sample **statistics**.

Definition 5.1: Sampling distribution

The sampling distribution of a statistics is the probability distribution of that statistics.

The probability distribution of sample mean (hereafter, will be denoted as \bar{X}) is called the **sampling distribution of the mean** (also, referred to as the distribution of sample mean).

Like \bar{X} , we call **sampling distribution of variance** (denoted as S^2).

Using the values of \bar{X} and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution

Example 5.1:

Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls.

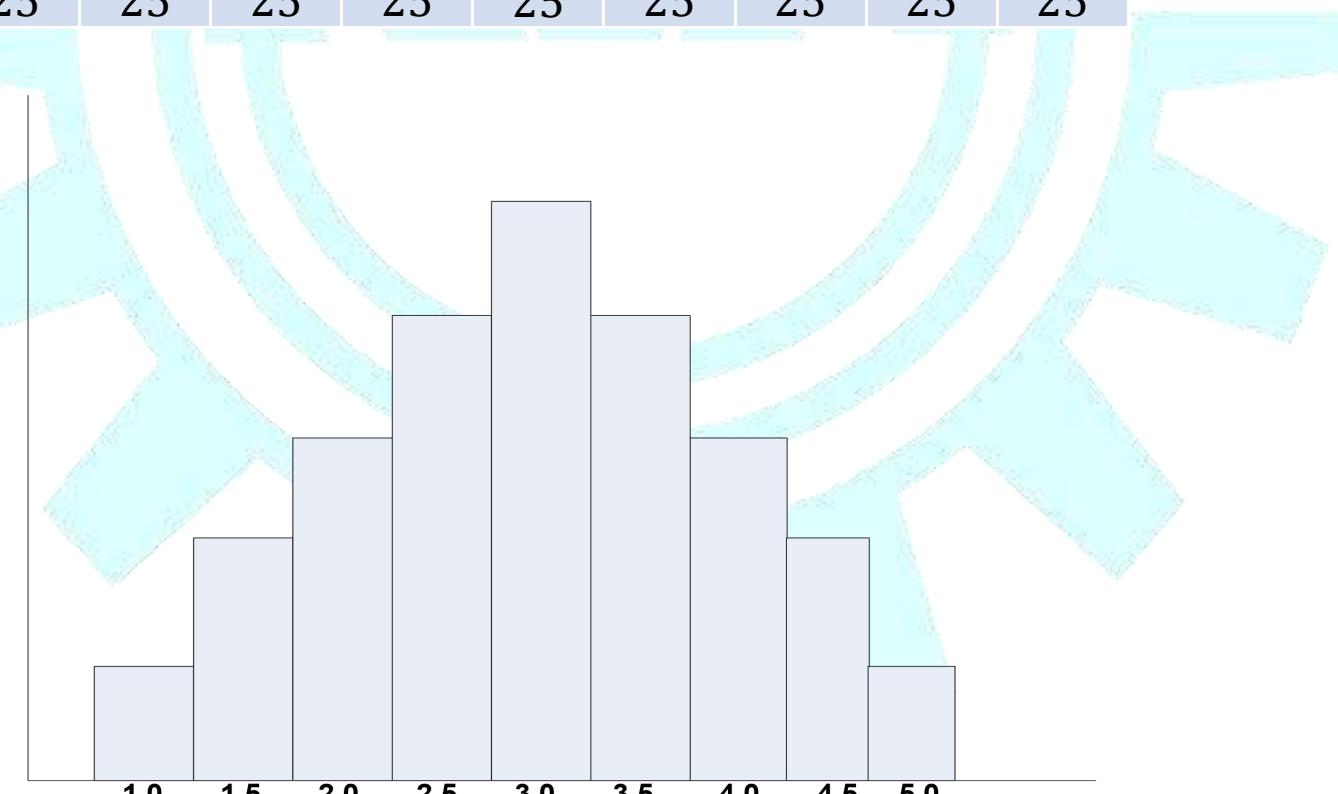
Following table lists all possible samples and their mean.

Sample (X)	Mean (\bar{X})	Sample (X)	Mean (\bar{X})	Sample (X)	Mean (\bar{X})
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

\bar{X}	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence probability distribution of **sample statistics**.
2. The sampling distribution of a statistics depends on
 - the size of the population
 - the size of the samples and
 - the method of choosing the samples.

The Idea of Statistical Significance

- Because sampling is imperfect
 - Samples may not ideally match the population
- Because hypothesis cannot be directly tested
 - Inference is subject to error

So, **Statistical Significance :**

The degree of risk that you are willing to take that you will reject a null hypothesis when it is actually true.

Degrees of Freedom

- Degrees of freedom (df) are the way in which the scientific tradition accounts for variation due to error
 - it specifies how many values vary within a statistical test
 - scientists recognize that collecting data can never be error-free
 - each piece of data collected can vary, or carry error that we cannot account for
 - by including df in statistical computations, scientists help account for this error
 - there are clear rules for how to calculate df for each statistical test

**Use statistical test to derive some calculated value
(e.g., t value or F value)**

Alternative and Null Hypotheses

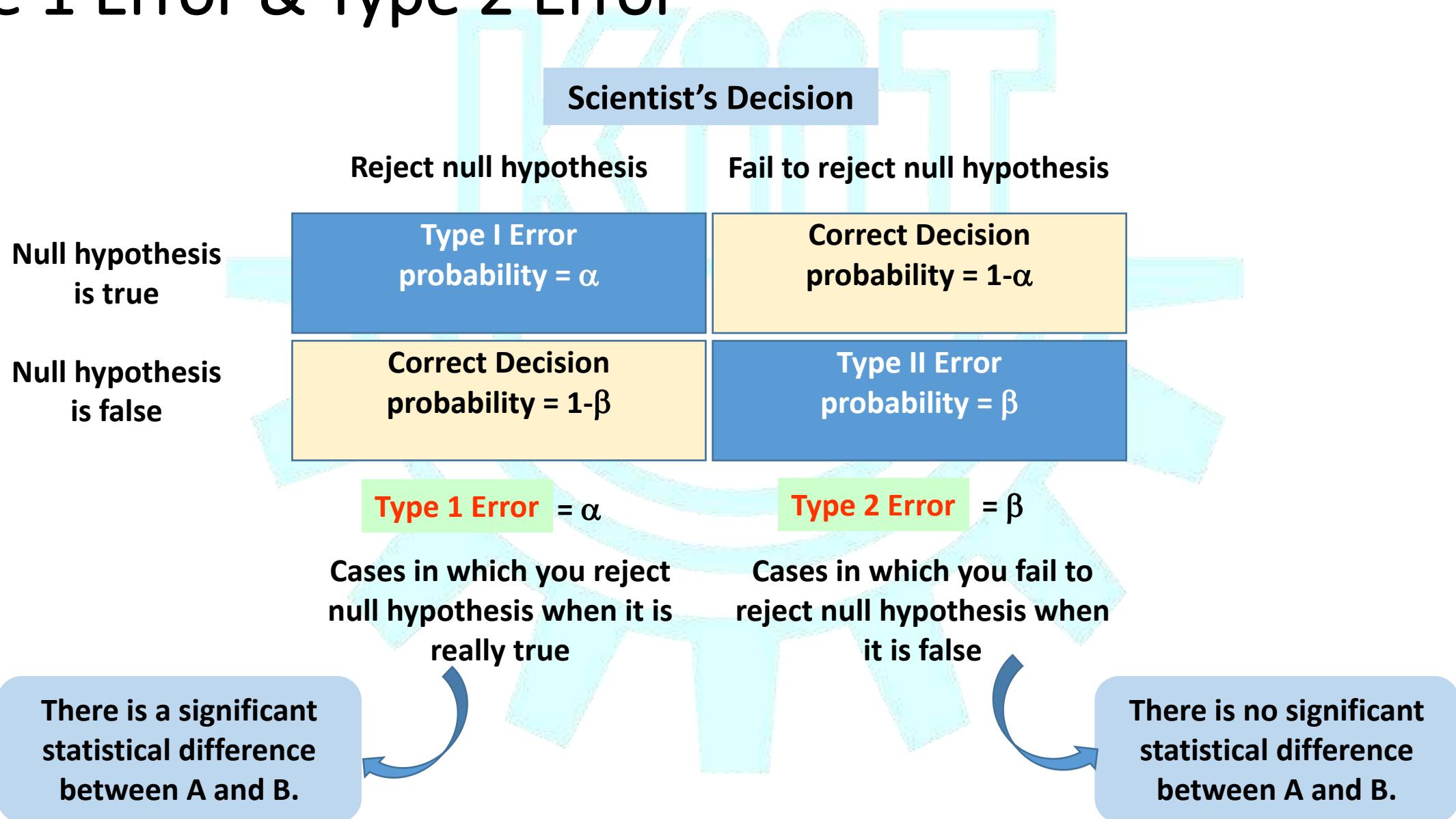
Inferential statistics test the likelihood that the **alternative (research) hypothesis (H_1)** is true and the **null hypothesis (H_0)** is not

- in testing differences, the H_1 would predict that differences would be found, while the H_0 would predict no differences
- by setting the significance level (generally at .05), the researcher has a criterion for making this decision
- If the .05 level is achieved (p is equal to or less than .05), then a researcher rejects the H_0 and accepts the H_1
- If the the .05 significance level is not achieved, then the H_0 is retained

Testing Hypothesis

- If reject H_0 and conclude groups are really different, it doesn't mean they're different for the reason you hypothesized
 - may be other reason
- Since H_0 testing is based on sample means, not population means, there is a possibility of making an error or wrong decision in rejecting or failing to reject H_0
 - **Type I error:** When we conclude that there is a relationship or effect but in fact there is not one (*false positive*).
 - **Type II error:** When we conclude that there is no relationship or effect when in fact there is one (*false negative*).

Type 1 Error & Type 2 Error



Testing for Statistical Significance

Significance Level (alpha = α)

The level in which we are allowed to reject the null hypothesis.

Probability value (p value)

The likelihood of an observed statistic occurring on the basis of the sampling distribution.

If P value is less than alpha ($p < .05$)

If P value is greater than alpha ($p > .05$)

Who decides the level?

The researcher
By convention, alpha is normally set to $\alpha = .05$

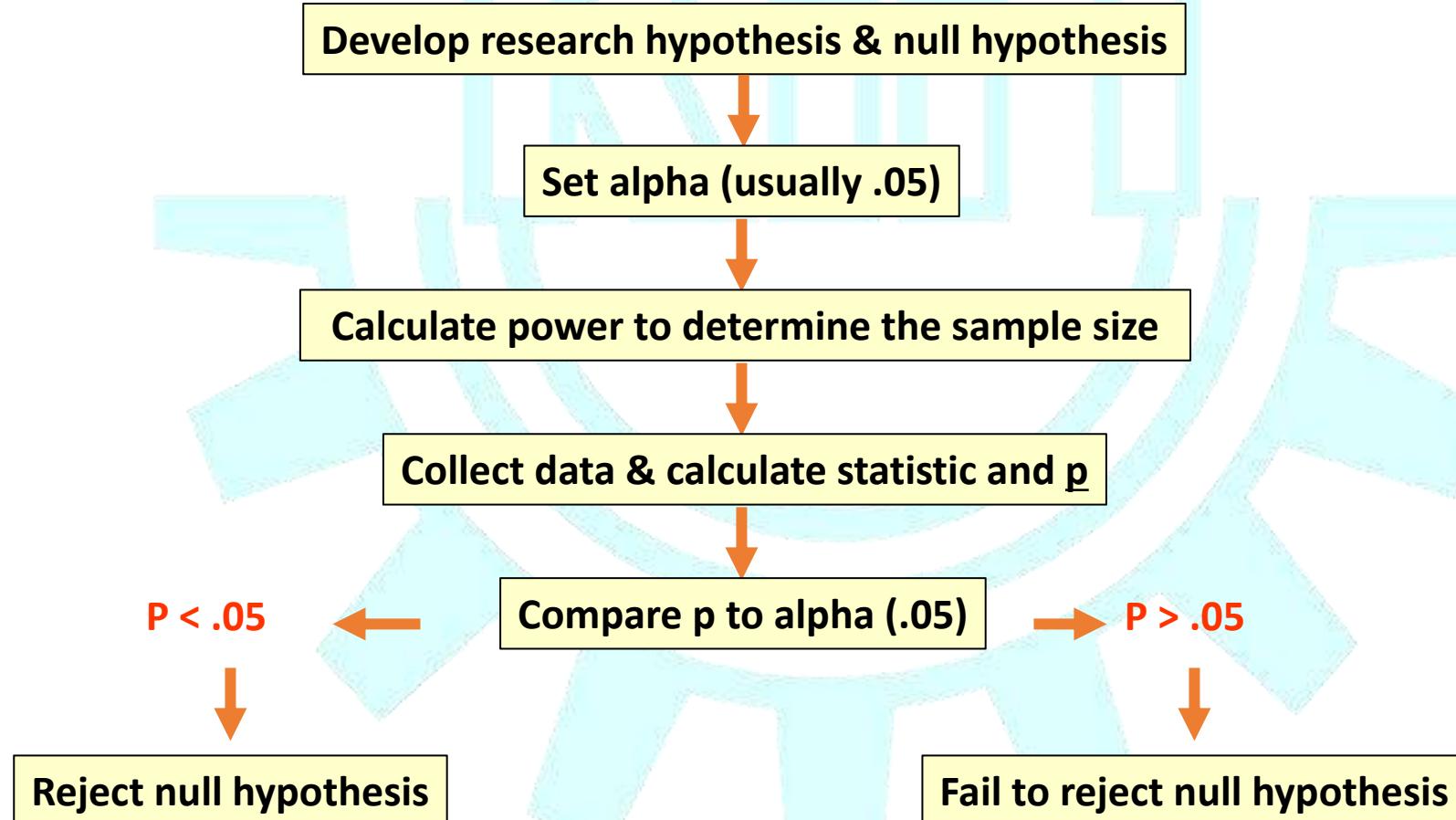
Reject null hypothesis

Fail to reject null hypothesis

Statistically Significant

Statistically nonsignificant

Hypothesis Testing Flow Chart



Regression

Lecture-3

Dr. H.K.Tripathy

What is Regression?

A way of predicting the value of one variable from another.

- It is a hypothetical model of the relationship between two variables.
- The model is used in a linear one.
- **Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.
 - *Any straight line can be represented by an equation of the form $Y = bX + a$, where b and a are constants.*
 - *The value of b is called the slope constant and determines the direction and degree to which the line is tilted.*
 - *The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.*

Main Objectives

Two main objectives:

- Establish if there is a **relationship** between two variables
 - Specifically, establish if there is a statistically significant relationship between the two.
 - Example: Income and expenditure, wage and gender, etc.
- Forecast new observations.
 - Can we use what we know about the relationship to forecast unobserved values?
 - Example: What will sales be over the next quarter?

Variable's Roles



Dependent

Independent

- This is the variable whose values we want to explain or forecast.
- Its values depend on something else.
- We denote it as Y.

- This is the variable that explains the other one.
- Its values are independent.
- We denote it as X.

$$Y = mX + c$$

A Linear Equation

You may remember one of these.

- $y = a + bx$
- $y = mx + b$
- In this regression discussion, we just use a different notation:
 - $y = \beta_0 + \beta_1 x$,
 - where, β_0 is called as intercept and β_1 is called as coefficient or slope
 - The values of the regression parameters β_0 , and β_1 are not known.
 - We estimate them from data.
 - β_1 indicates the change in the mean response per unit increase in X .
 - We call it “linear” because the equation represents a straight line in a bi-dimensional plot.

Regression Analysis

In regression analysis we use the independent variable (X) to estimate the dependent variable (Y)

- *Both the variables must be at least interval scale.*
- *The least squares criterion is used to determine the equation.*

Regression Equation:

An equation that express the linear relationship between two variables.

Least Squares Principle:

It determining a regression equation by minimising the sum of the squares of the vertical distances between the actual Y values and the predicted values of Y

Regression Analysis- Least Squares Principle

- The least squares principle
- Dots are actual values of Y
- Asterisks are the predicted values of Y for a given value of X



Chart 1



Chart 2

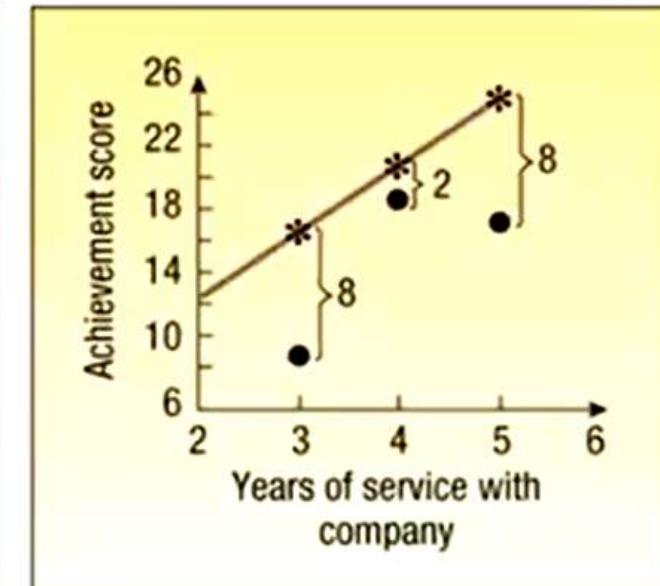


Chart 3

Regression Line

We will write an estimated regression line based on sample data as

$$\hat{y} = b_0 + b_1 x$$

The method of least squares chooses the values for b_0 , and b_1 to minimize the sum of squared errors

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Using calculus, we obtain estimating formulas:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

OR

$$b_1 = r \frac{S_y}{S_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Estimation of Mean Response

Fitted regression line can be used to estimate the mean value of y for a given value of x .

Example :

- The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table.

y	x
1250	41
1380	54
1425	63
1425	54
1450	48
1300	46
1400	62
1510	61
1575	64
1650	71

- From the data table we have:

$$\begin{aligned} n &= 10 & \sum x &= 564 & \sum x^2 &= 32604 \\ \sum y &= 14365 & \sum xy &= 818755 \end{aligned}$$

- The least squares estimates of the regression coefficients are:

$$b_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{10(818755) - (564)(14365)}{10(32604) - (564)^2} = 10.8$$

$$b_0 = 1436.5 - 10.8(56.4) = 828$$

Point Estimation of Mean Response

- The estimated regression function is:

$$\hat{y} = 828 + 10.8x$$

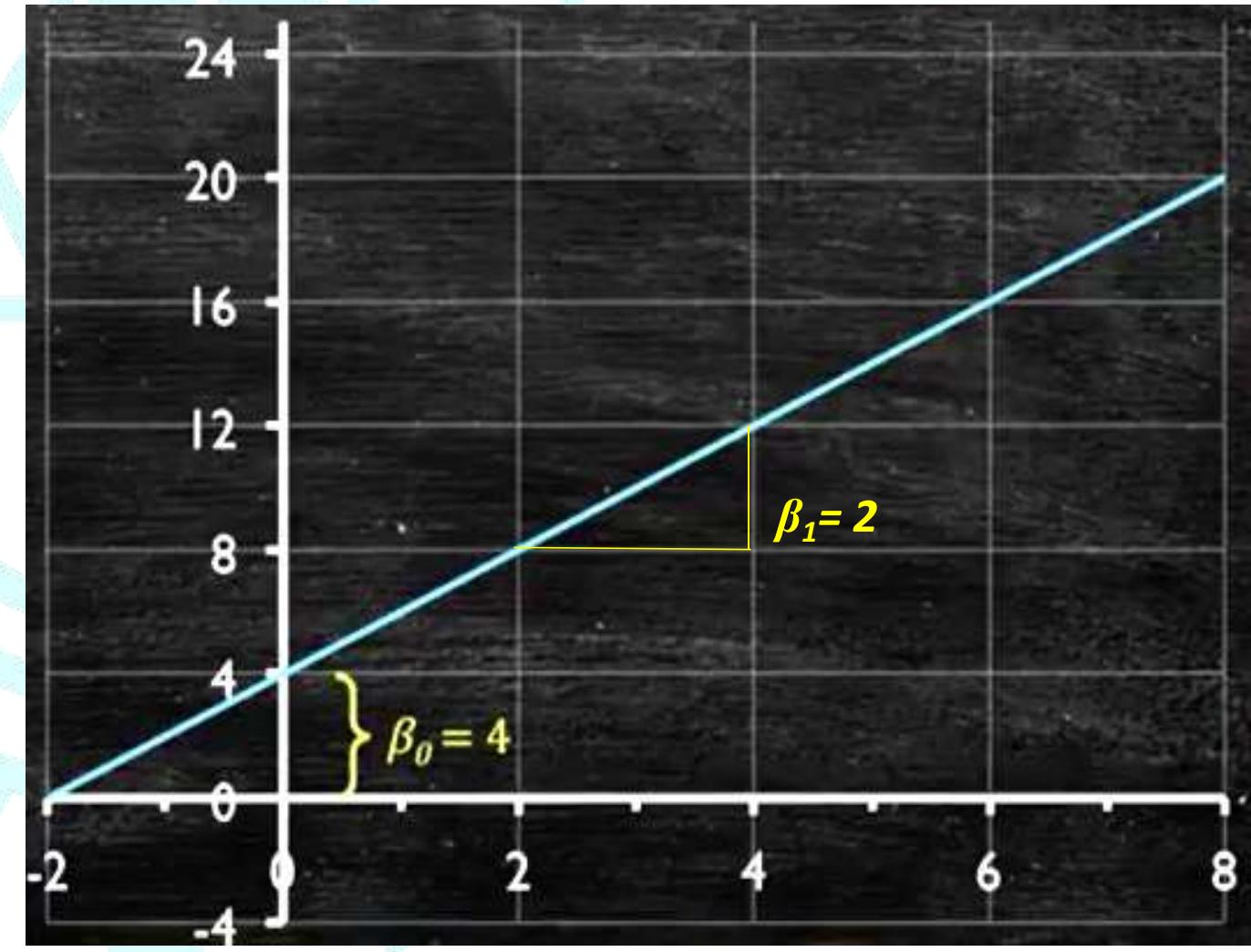
Sales = 828 + 10.8 Expenditure

- This means that if the weekly advertising expenditure is increased by \$1 we would expect the weekly sales to increase by \$10.8.

Linear Equation Example

$$y = \beta_0 + \beta_1 x$$

$$y = 4 + 2x$$

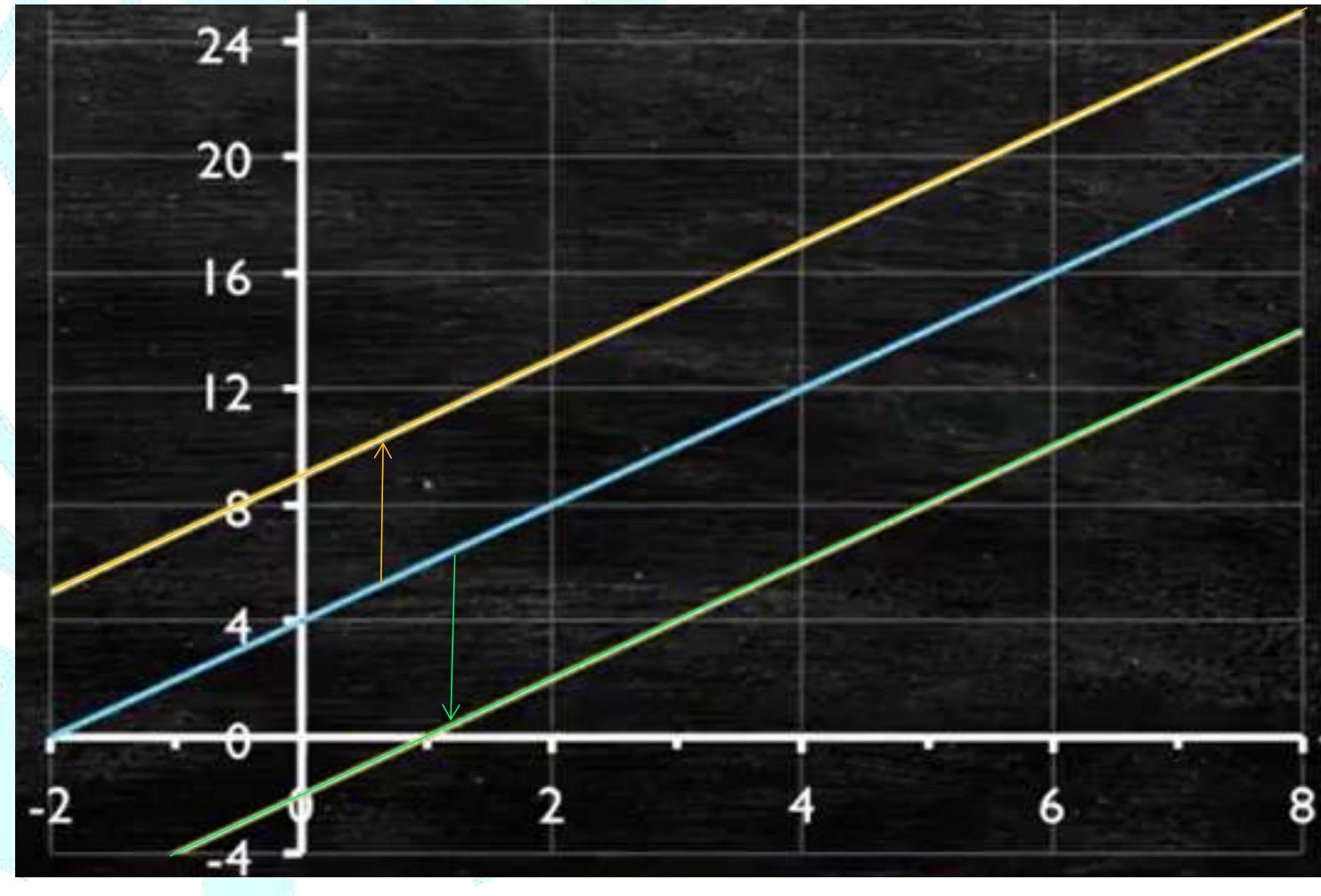
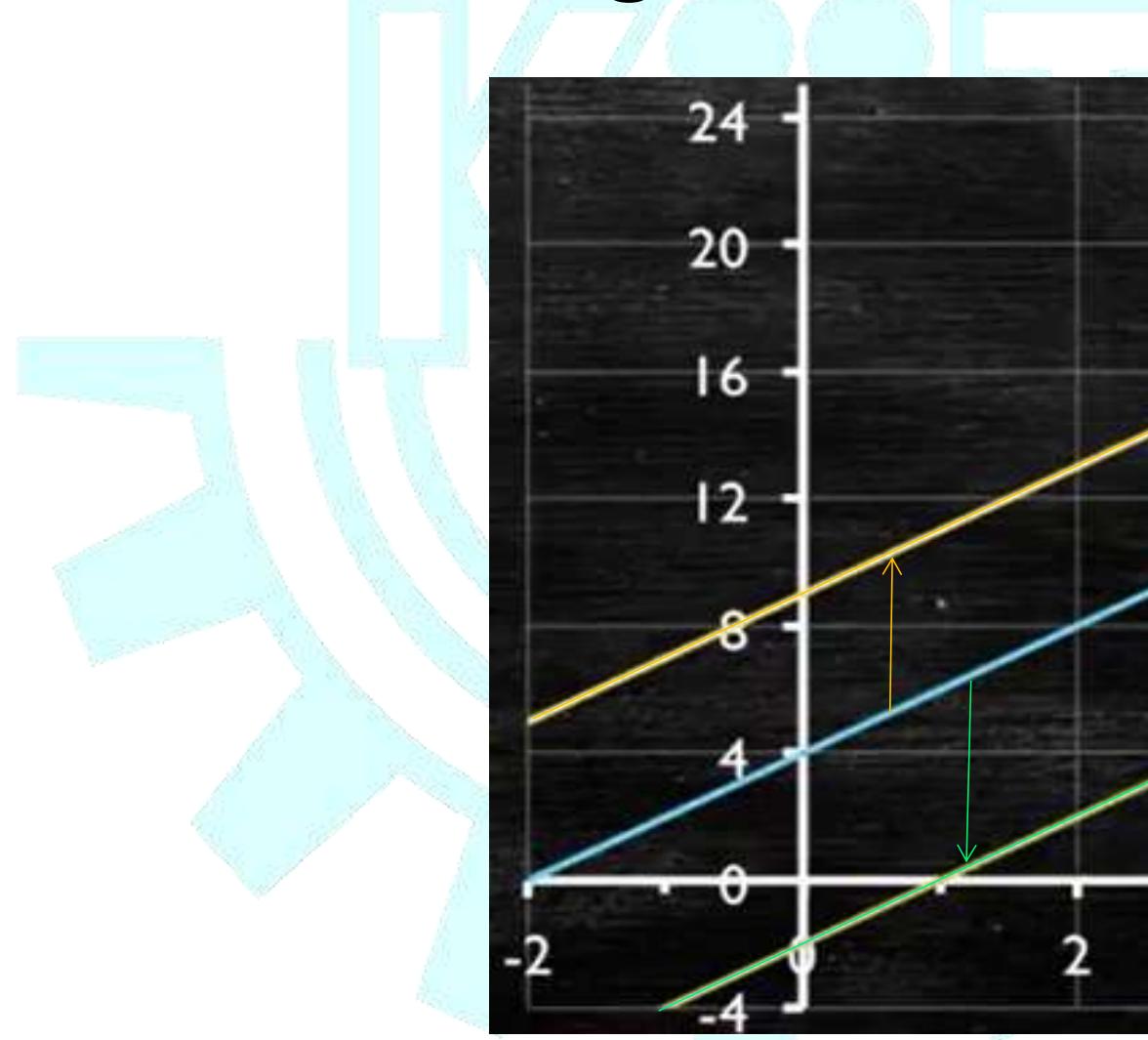


What happens if we change the intercept?

$$y = 4 + 2x$$

$$y = 9 + 2x$$

$$y = -2 + 2x$$



What happens if we change the slope?

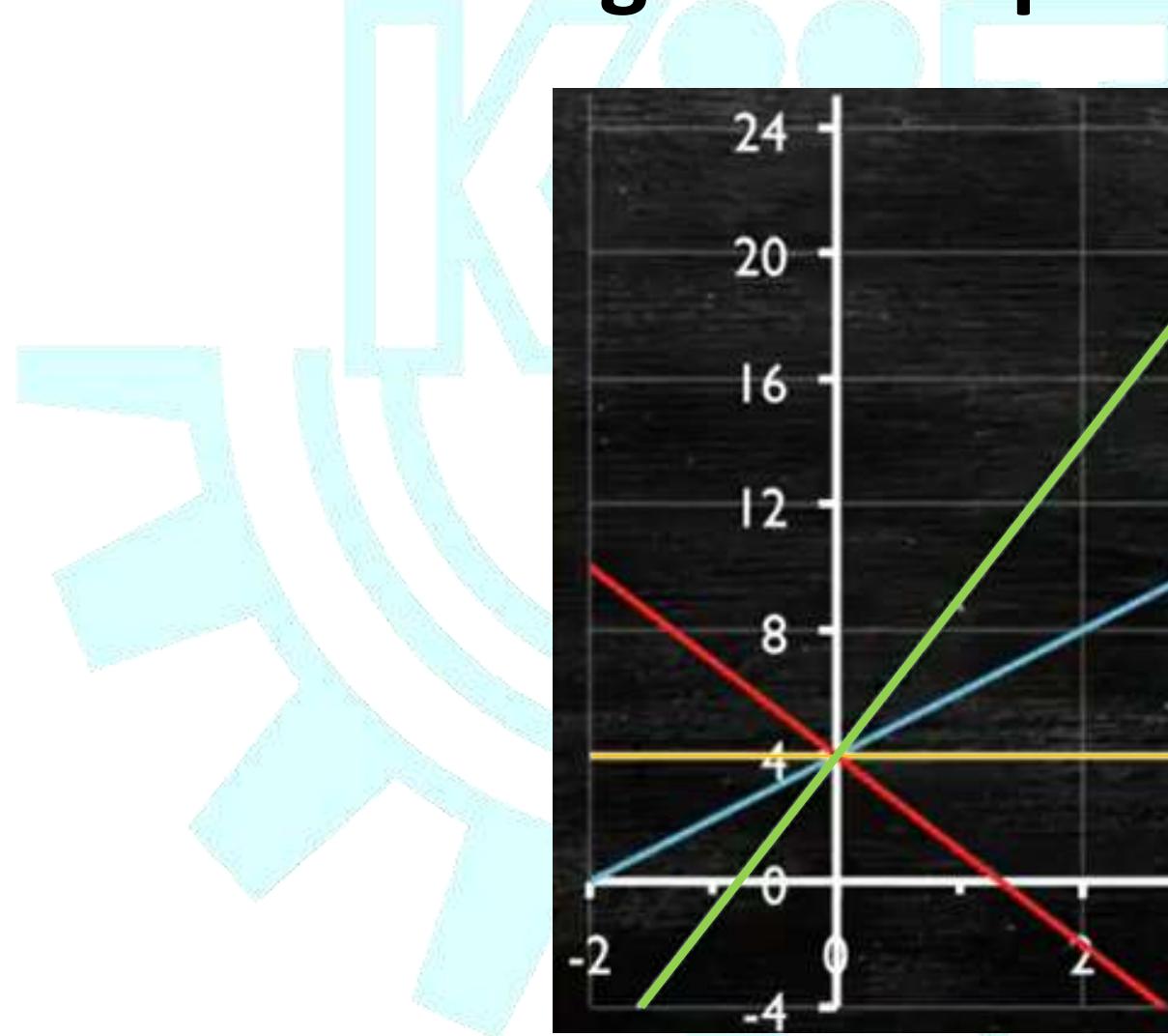
$$y = 4 + 2x$$

$$y = 4 + 5x$$

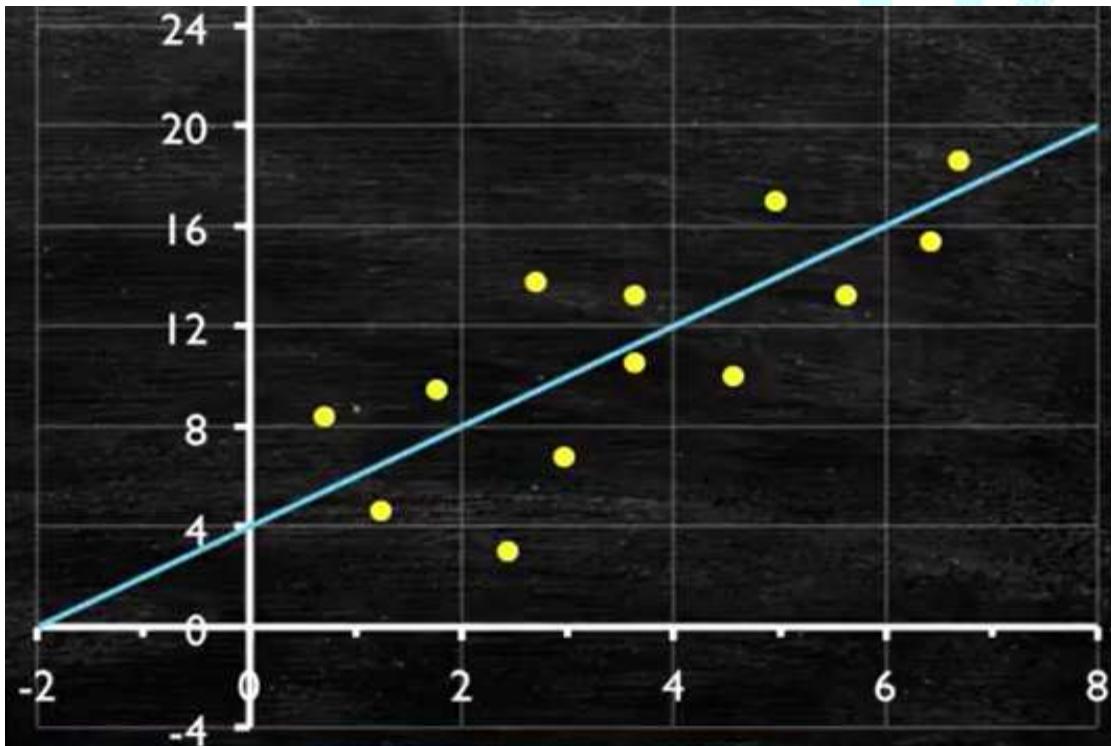
$$y = 4 + 0x$$

$$= 4$$

$$y = 4 - 3x$$



But, the world is not linear !

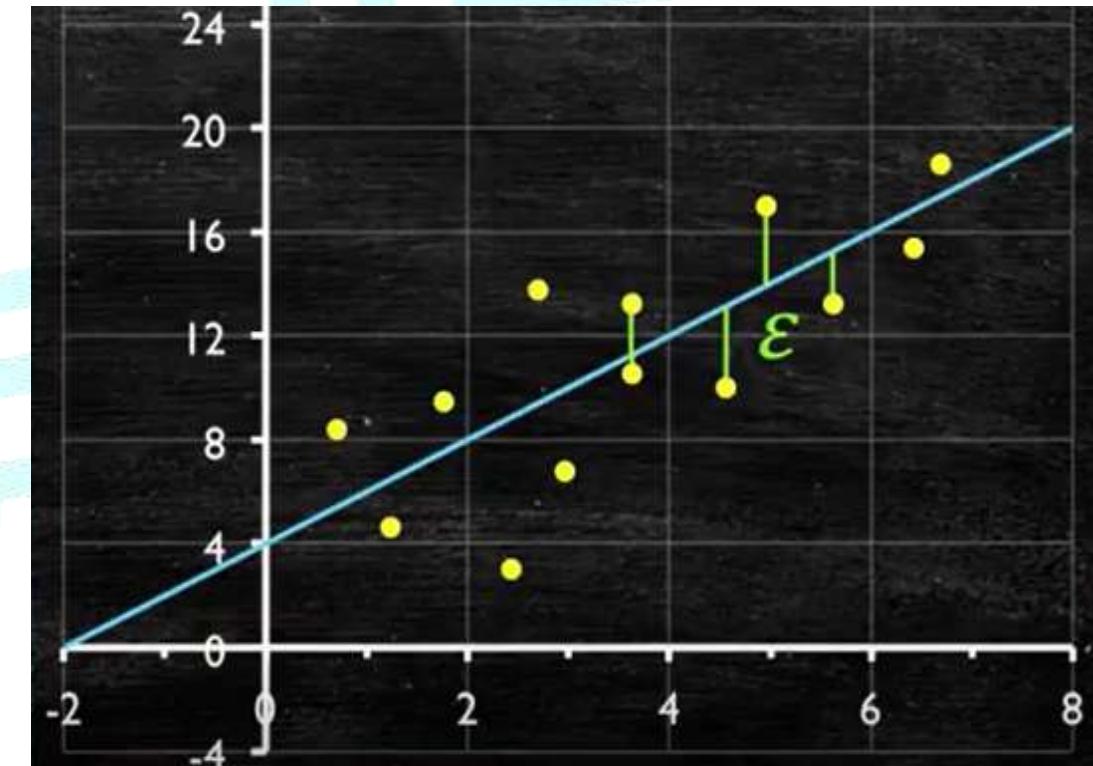


$$y = 4 + 2x$$

True Value

$y = \beta_0 + \beta_1 x + \varepsilon$

Simple Linear Regression Model



Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y is the dependent variable
- x is the independent variable
- β_0 is the constant or intercept
- β_1 is x 's slope or coefficient
- ε is the error term

$$\text{Error} = \sum_{i=1}^n (\underline{\text{actual_output}} - \underline{\text{predicted_output}})^2$$

Data for Linear Regression Example

ID	Income	Consumption
1	119	154
2	85	123
3	97	125
4	95	130
5	120	151
6	92	131
7	105	141
8	110	141
9	98	130
10	98	134
11	81	115
12	81	117
13	91	123
14	105	144
15	100	137
16	107	140
17	82	123
18	84	115
19	100	134
20	108	147

ID	Income	Consumption
21	116	144
22	115	144
23	93	126
24	105	141
25	89	124
26	104	144
27	108	144
28	88	129
29	109	137
30	112	144
31	96	132
32	89	125
33	93	126
34	114	140
35	81	120
36	84	118
37	88	119
38	96	131
39	82	127
40	114	150

Our assumption, which we will test, is that income explains consumption

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$\text{Consumption} = \beta_0 + \beta_1 \text{Income} + \varepsilon$$

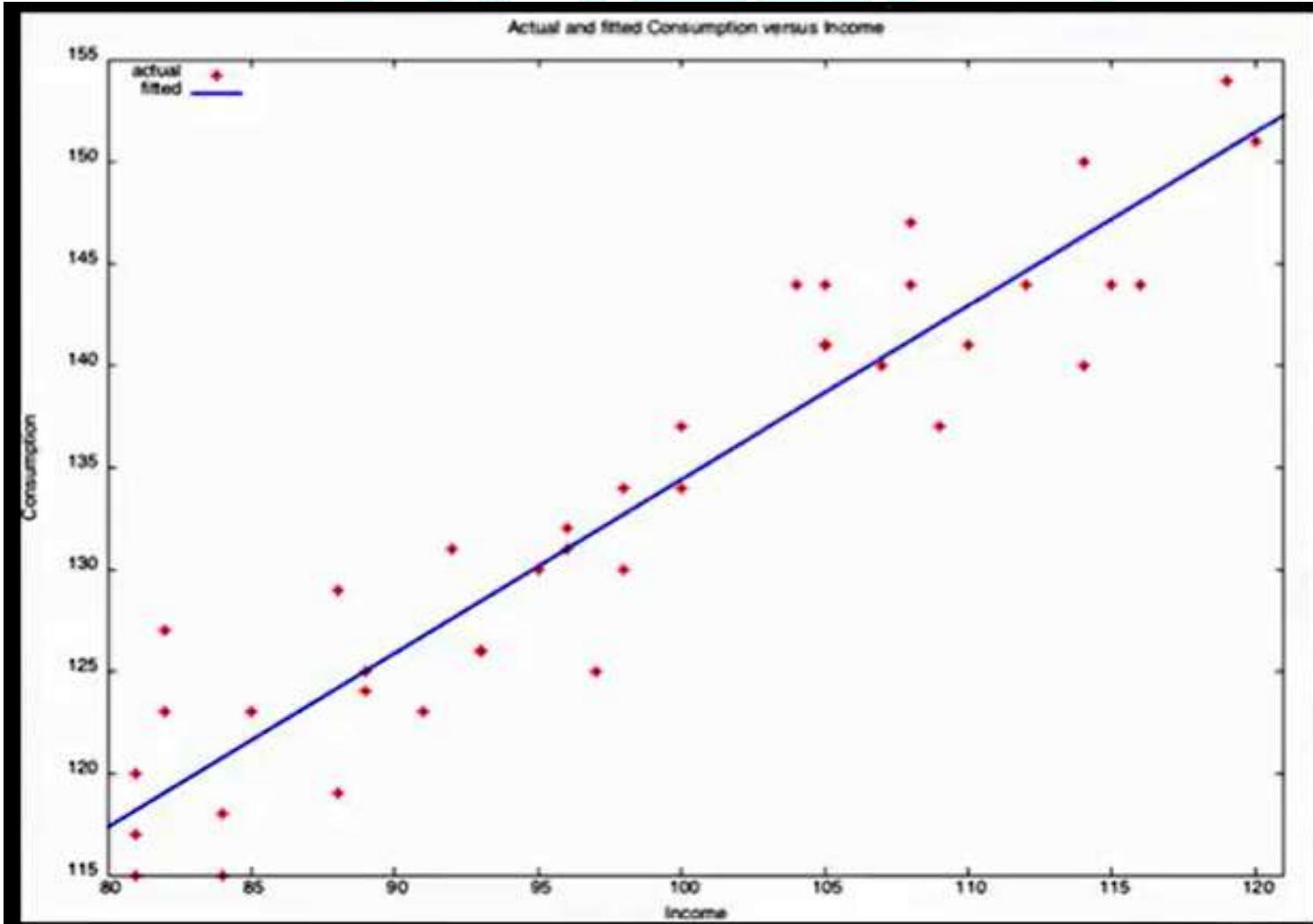
Interpreting the Coefficients

The estimate dmodel is

$$\textit{Consumption} = 49.13 + 0.85 \textit{Income} + \varepsilon$$

- 49.13 could be interpreted as the consumption level of a family with 0 income.
- 0.85 is the marginal effect of one unit of income on consumption: for every unit more of income a family has, we estimate its consumption grows by 0.85 units.

Estimated vs. Actual values



Logistic Regression

In linear regression the Y variable is always a continuous variable.

If suppose, the Y variable was categorical, you cannot use linear regression model it.

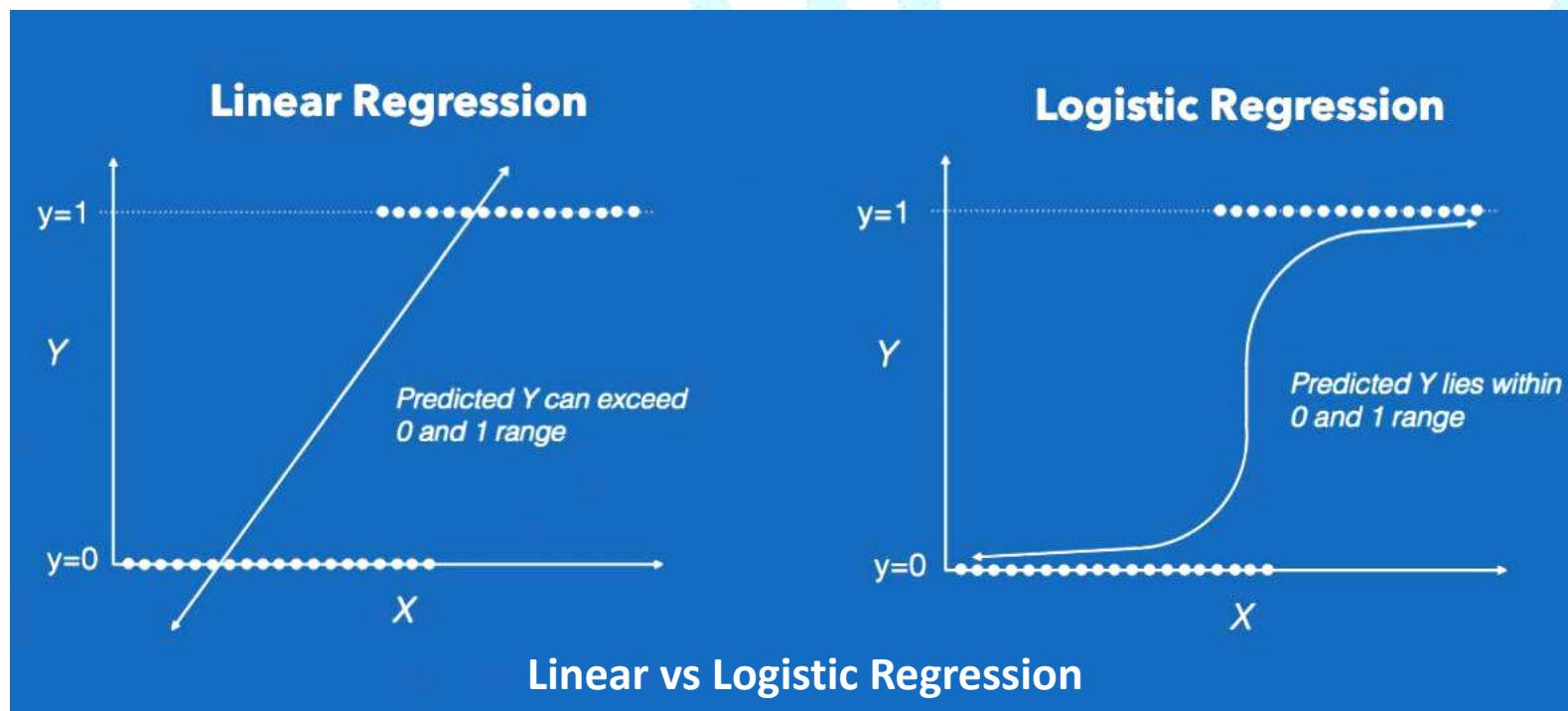
So what would you do when the Y is a categorical variable with 2 classes?

- Logistic regression can be used to model and solve such problems, also called as binary classification problems.
- Logistic Regression is one of the most commonly used Machine Learning algorithms that is used to model a binary variable that takes only 2 values – 0 and 1.
- The objective of Logistic Regression is to develop a mathematical equation that can give us a score in the range of 0 to 1.
- This score gives us the probability of the variable taking the value 1.

Why not linear regression?

When the response variable has only 2 possible values, it is desirable to have a model that predicts the value either as 0 or 1 or as a probability score that ranges between 0 and 1.

Linear regression does not have this capability. Because, If you use linear regression to model a binary response variable, the resulting model may not restrict the predicted Y values within 0 and 1.



This is where logistic regression comes into play. In logistic regression, you get a probability score that reflects the probability of the occurrence of the event.

Logistic Regression

- A key point to note here is that Y can have 2 classes only and not more than that.
- If Y has more than 2 classes, it would become a multi class classification and you can no longer use the logistic regression for that.
- Another advantage of logistic regression is that it computes a prediction probability score of an event. More on that when you actually start building the models.
- ***Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous***

Models relationship between set of variables X_i ,

- dichotomous (yes/no, smoker/nonsmoker,...)
- categorical (social class, race, ...)
- continuous (age, weight, gestational age, ...)

and

Dichotomous categorical response variable Y

e.g. Success/**Failure**, Remission/**No Remission**, Survived/**Died**, CHD/**No CHD**, Low Birth Weight/**Normal Birth Weight**, etc...

Logistic Regression



Logistic regression seeks to:

- **model the probability of an event occurring depending on the values of the independent variables, which can be categorical or numerical**
- **estimate the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur**
- **predict the effect of a series of variables on a binary response variable**
- **classify observations by estimating the probability that an observation is in a particular category (such as approved or not approved in our problem)**



Logistic Regression Example

Spam Detection: Spam detection is a binary classification problem where we are given an email and we need to classify whether or not it is spam. If the email is spam, we label it 1; if it is not spam, we label it 0.

Tumour Prediction: A Logistic Regression classifier may be used to identify whether a tumour is malignant or if it is benign. Several medical imaging techniques are used to extract various features of tumours. For instance, the size of the tumour, the affected body area, etc. These features are then fed to a Logistic Regression classifier to identify if the tumour is malignant or if it is benign.

Health : Predicting if a given mass of tissue is benign or malignant

Marketing : Predicting if a given user will buy an insurance product or not

Banking : Predicting if a customer will default on a loan.

Credit Card Fraud, etc.

Logistic Regression

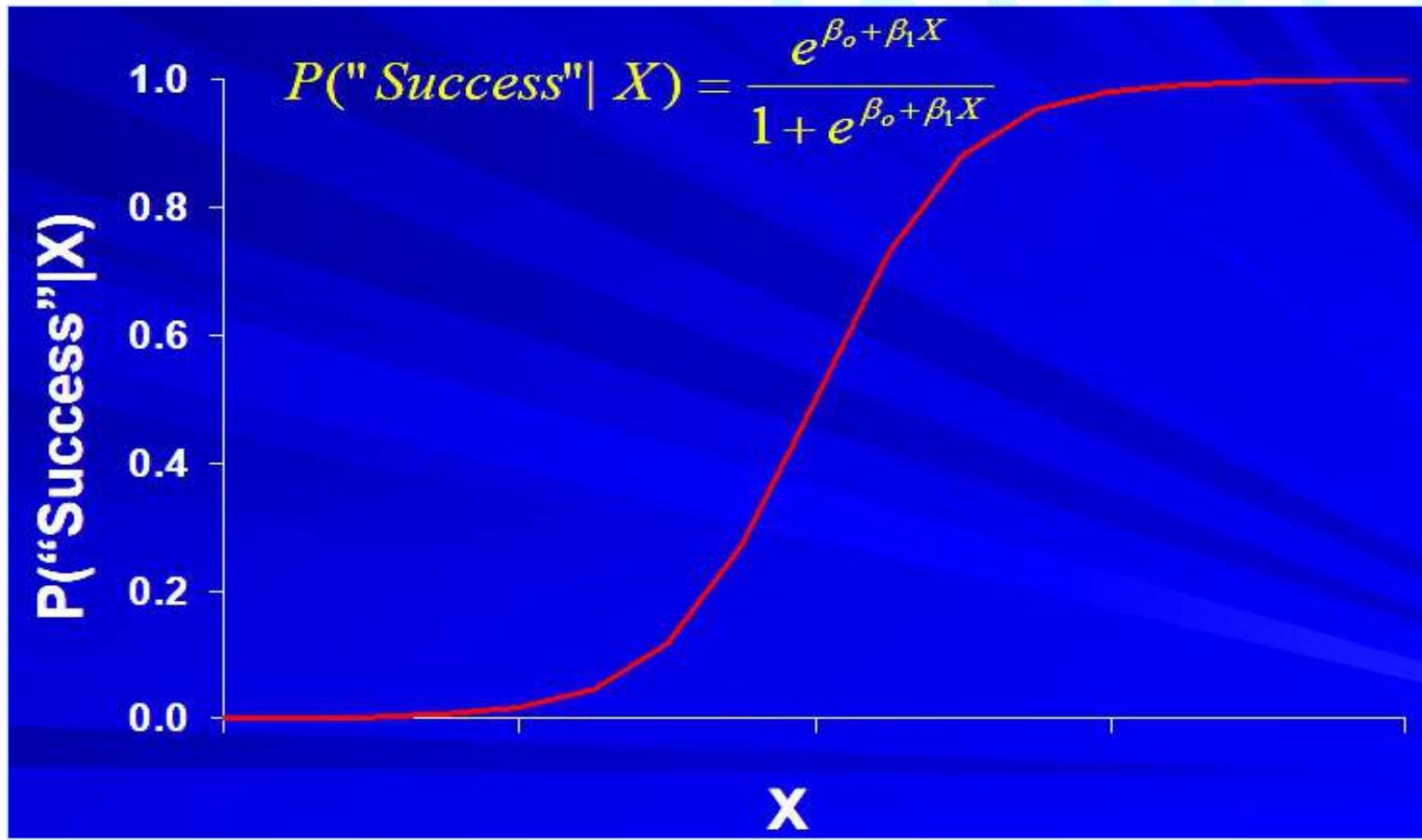
There are 3 types of Logistic Regression

- **Binary Logistic Regression (Binomial):** The categorical response has only two possible outcomes. “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
- **Multinomial Logistic Regression:** Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan), like “disease A” vs “disease B” vs “disease C”.
- **Ordinal Logistic Regression:** Three or more categories with ordering.

Example: Test score can be categorized as: “very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3., Movie rating from 1 to 5.

The Logistic Equation

Logistic regression achieves this by taking the log odds of the event $\ln(P/1-P)$, where, P is the probability of event. So P always lies between 0 and 1.



The values in the regression equation b_0 and b_1 take on slightly different meanings.

$b_0 \leftarrow$ The regression constant

log odds in unexposed
(moves curve left and right)

$b_1 \leftarrow$ The regression slope

**log odds ratio associated
with being exposed**
(steepness of curve)

$-\frac{b_0}{b_1} \leftarrow$ The threshold, where
probability of success = .50

Logit Transformation

The logistic regression model is given by

$$P(Y | X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

which is equivalent to

$$\ln\left(\frac{P(Y | X)}{1 - P(Y | X)}\right) = \ln\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 X$$

This is called the Logit Transformation

This implies that the odds for success can be expressed as

$$\frac{P}{1 - P} = e^{\beta_0 + \beta_1 X}$$

$$\text{odds} = p / (1-p)$$

Multiple Logistic Regression Model

$$\ln\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Logistic Regression uses Odds Ratios

- Does not model the outcome directly, which leads to effect estimates quantified by means (i.e., differences in means)
- Estimates of effect are instead quantified by “Odds Ratios”

Relationship between Odds & Probability

$$\text{Odds}(\text{event}) = \frac{\text{Probability}(\text{event})}{1 - \text{Probability}(\text{event})}$$

$$\text{Probability}(\text{event}) = \frac{\text{Odds}(\text{event})}{1 + \text{Odds}(\text{event})}$$

Fair coin flip

$$\text{odds}(\text{heads}) = \frac{0.5}{0.5} = 1 \text{ or } 1:1$$

Fair die roll

$$\text{odds}(1 \text{ or } 2) = \frac{0.333}{0.666} = \frac{1}{2} = 0.5 \text{ or } 1:2$$

Deck of playing cards

$$\text{odds}(\text{diamond card}) = \frac{0.25}{0.75} = \frac{1}{3} = 0.333 \text{ or } 1:3$$

Odds Ratios

Fair coin flip

$$P(\text{heads}) = \frac{1}{2} = 0.5$$

$$\text{odds}(\text{heads}) = \frac{0.5}{0.5} = 1 \text{ or } 1:1$$

Loaded coin flip

$$P(\text{heads}) = \frac{7}{10} = 0.7$$

$$\text{odds}(\text{heads}) = \frac{0.7}{0.3} = 2.333$$

$$\text{Odds ratio} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

$$\text{Odds ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

$$\text{Odds ratio} = \frac{\frac{.7}{.3}}{\frac{.5}{.5}} = \frac{.7}{.3} \times \frac{.5}{.5} = \frac{.35}{.15} = 2.333$$

The odds of getting “heads” on the loaded coin are 2.333x greater than the fair coin.

- The odds ratio for a variable in logistic regression represents how the odds change with a 1 unit increase in that variable holding all other variables constant
- For (fictitious) example:
 - Body weight and sleep apnea (two categories: apnea / no apnea)
 - Weight variable had an odds ratio of 1.07
 - This means a one pound increase in weight increases the odds of having sleep apnea by 1.07 (not very high b/c we are looking at 1lb increments)

Dichotomous Predictor

Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

Disease (Y)	Risk Factor (X)	
	Present ($X = 1$)	Absent ($X = 0$)
Yes ($Y = 1$)	$P(Y = 1 X = 1)$	$P(Y = 1 X = 0)$
No ($Y = 0$)	$1 - P(Y = 1 X = 1)$	$1 - P(Y = 1 X = 0)$

$$\frac{P}{1-P} = e^{\beta_0 + \beta_1 X} \left\{ \begin{array}{l} \text{Odds for Disease with Risk Present} = \frac{P(Y = 1 | X = 1)}{1 - P(Y = 1 | X = 1)} = e^{\beta_0 + \beta_1} \\ \text{Odds for Disease with Risk Absent} = \frac{P(Y = 1 | X = 0)}{1 - P(Y = 1 | X = 0)} = e^{\beta_0} \end{array} \right.$$

Therefore the odds ratio (OR) = $\frac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$

Odds Ratios

Definition of Odds Ratio: Ratio of two odds estimates.

So, if $\Pr(\text{response} | \text{trt}) = 0.40$ and $\Pr(\text{response} | \text{placebo}) = 0.20$

Then:

$$\hat{\text{Odds}}(\text{response} | \text{trt group}) = \frac{0.40}{1 - 0.40} \doteq 0.667$$

$$\hat{\text{Odds}}(\text{response} | \text{placebo group}) = \frac{0.20}{1 - 0.20} = 0.25$$

$$\Rightarrow \hat{\text{OR}}(\text{Trt vs. Placebo}) \doteq \frac{0.667}{0.25} \doteq 2.67$$

Logistic Regression Example

Consider an example dataset which maps the number of hours of study with the result of an exam. The result can take only two values, namely passed(1) or failed(0):

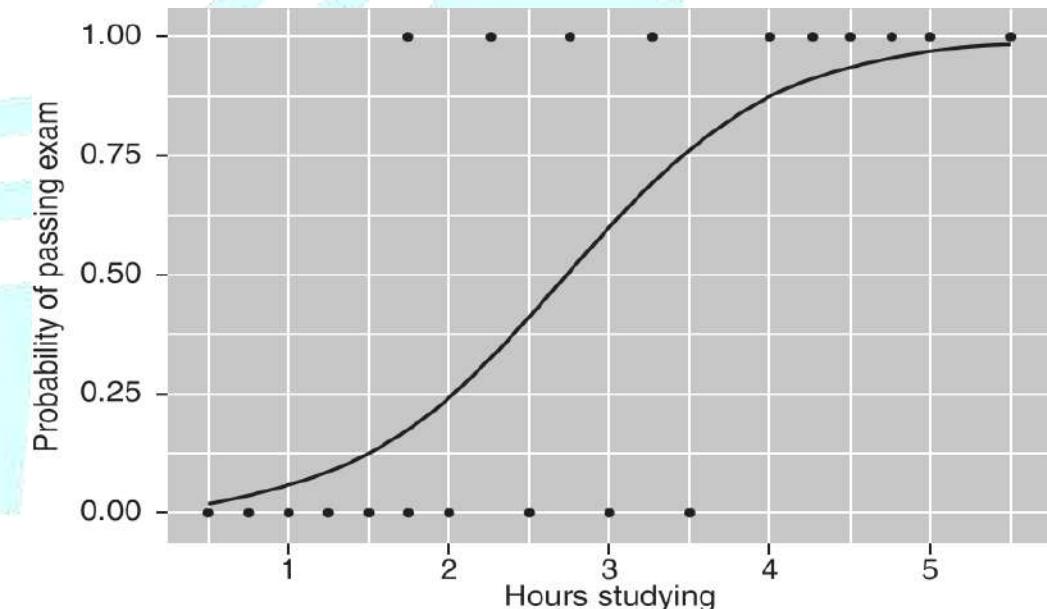
Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

So, we have

$$y = \begin{cases} 0, & \text{if fail} \\ 1, & \text{if pass} \end{cases}$$

i.e. y is a categorical target variable which can take only two possible type: "0" or "1". In order to generalize our model, we assume that:

The dataset has 'p' feature variables and 'n' observations. The feature matrix is represented as:



Logistic Regression Example

Consider a model with one predictor X_1 , and one binary response variable Y , which we denote $p = P(Y = 1 \mid X_1 = x)$. We assume a linear relationship between the independent variable and the logit of the event i.e. $Y = 1$. In statistics, the logit is the logarithm of the odds i.e. $p / (1-p)$. This linear relationship can be written in the following mathematical form (where ℓ is the logit, b is the base of the logarithm, and β is the parameter of the model).

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

$$P_i = E(y = 1|x_i) = \frac{e^z}{1 + e^z} = \frac{e^{\alpha + \beta_i x_i}}{1 + e^{\alpha + \beta_i x_i}}$$

The odds can be recovered by exponentiation of the logit:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1} \rightarrow p = \frac{b^{\beta_0 + \beta_1 x_1}}{b^{\beta_0 + \beta_1 x_1} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1)}} = S_b(\beta_0 + \beta_1 x_1)$$

Where S_b is the sigmoid function with base b . However in some cases it can be easier to communicate results by working in base 2, base 10, or exponential constant e . In reference to the students example, solving the equation with software tool and considering base as e , the coefficient is $\beta_0 = -4.0777$ and $\beta_1 = 1.5046$

Logistic Regression Example

- For example, for a student who studies 2 hours, entering the value Hours = 2 in the equation gives the estimated probability of passing the exam of 0.26.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 2 - 4.0777))} = 0.26$$

- Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 4 - 4.0777))} = 0.87$$

- Following table shows the probability of passing the exam for several values of hours studying.

Hours of study	Probability of passing the exam
1	0.07
2	0.26
3	0.61
5	0.97

Counting Distinct Elements

Definition

- Data stream consists of a universe of elements chosen from a set of N
- Maintain a count of number of distinct items seen so far.

Let us consider a stream :

32, 12, 14, 32, 7, 12, 32, 7, 6, 12, 4

Elements occur multiple times, we want to count the number of distinct elements.

Number of distinct element is n (=6 in this example)

Number of elements in this example is 11

Why do we count distinct elements?

- Number of distinct queries issued
- Unique IP addresses passing packages through a router
- Number of unique users accessing a website per month
- Number of different people passing through a traffic hub (airport, etc.)
- How many unique products we sold tonight?
- How many unique requests on a website came in today?
- How many different words did we find on a website?
 - Unusually large number of words could be indicative of spam

Question: how would you do it?

Now, let's constrain ourselves with limited storage...

- How to estimate (in an unbiased manner) the number of distinct elements seen?
 - **Flajolet-Martin (FM) Approach**
- FM algorithms approximates the number of unique objects in a stream or a database in one pass.
- If the stream contains n elements with m of them unique, this algorithm runs in $O(n)$ time and needs $O(\log(m))$ memory.

FM-sketch (Flajolet-Martin)

Task: given a data stream, estimate the number of distinct elements occurring in it.

- **Approach:** hash data stream elements uniformly to N bit values, i.e.:
$$h : a_i \rightarrow \{0, 1\}^N$$
- **Assumption:** the larger the number of distinct elements in the stream, the more distinct the occurring hash values, and the more likely one with an “unusual” property appears

FM-sketch (Flajolet-Martin)

- One possibility of interpreting “unusual” is the **hash tail**: the number of 0’s a binary hash value ends in

100110101110

100110101100

100110000000

Algorithm:

for all $a_i \in S$,

$h(a_i) \rightarrow \{0,1\}^N$

Maximum hash
tail seen so far

$R = \max_{a_i \in S} h(a_i)$
return $|S| = 2^R$

Important:

N must be long enough:
there must be more
possible results of the
hash function than
elements in the
universal set.

FM-sketch (Flajolet-Martin)

Pick a hash function that maps each of the N elements to at least $\log_2 N$ bits

For each stream element a , let $r(a)$ be the number of trailing 0s in $h(a)$

$r(a)$ = position of first 1 counting from the right

E.g., say $h(a) = 12$, then 12 is 1100 in binary, so $r(a) = 2$

Record R = the maximum $r(a)$ seen

$R = \max_a r(a)$, over all the items a seen so far

Estimated number of distinct elements = 2^R

Median of means for a stable result

Example

Determine the distinct element in the stream using FM.

- Input stream of integers $x = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1$.
- Hash Function, $h(x) = 6x + 1 \bmod 5$

$$h(1) = 6(1) + 1 \bmod 5$$

$$= 6 + 1 \bmod 5$$

$$= 7 \bmod 5$$

$$= 2$$

Therefore $h(1) = 2$

Similarly, calculate hash function for the complete stream.

Calculate hash functions $h(x)$

- For the given input stream 1,3,2,1,2,3,4,3,1,2,3,1.

$$h(x) = 6x + 1 \bmod 5$$

$$h(1) = 2$$

$$h(4) = 0$$

$$h(3) = 4$$

$$h(3) = 4$$

$$h(2) = 3$$

$$h(1) = 2$$

$$h(1) = 2$$

$$h(2) = 3$$

$$h(2) = 3$$

$$h(3) = 3$$

$$h(3) = 4$$

$$h(1) = 2$$

The numbers obtained are:

$$\{2, 4, 3, 2, 3, 4, 0, 4, 2, 3, 3, 2\}$$

Binnary Bits

- For the given input stream 1,3,2,1,2,3,4,3,1,2,3,1.
- For every hash function calculated, write the binary equivalent for the same.

$$h(1) = 2 = 010$$

$$h(4) = 0 = 000$$

$$h(3) = 4 = 100$$

$$h(3) = 4 = 100$$

$$h(2) = 3 = 011$$

$$h(1) = 2 = 010$$

$$h(1) = 2 = 010$$

$$h(2) = 3 = 011$$

$$h(2) = 3 = 011$$

$$h(3) = 4 = 100$$

$$h(3) = 4 = 100$$

$$h(1) = 2 = 010$$

Convesion to binanry:

{010,100,011,010,011,100,000,100,010,011,100,010}

Trailing Zeros

- For the given input stream 1,3,2,1,2,3,4,3,1,2,3,1.
- For every hash function calculated and the binary equivalent for the same is written.
- Now write the count of trailing zeros in each hash function bit.

$$h(1) = 2 = 010 = 1$$

$$h(3) = 4 = 100 = 2$$

$$h(2) = 3 = 011 = 0$$

$$h(1) = 2 = 010 = 1$$

$$h(2) = 3 = 011 = 0$$

$$h(3) = 4 = 100 = 2$$

$$h(4) = 0 = 000 = 0$$

$$h(3) = 4 = 100 = 2$$

$$h(1) = 2 = 010 = 1$$

$$h(2) = 3 = 011 = 0$$

$$h(3) = 4 = 100 = 2$$

$$h(1) = 2 = 010 = 1$$

Computing $r(a)$:

{1,2,0,1,0,2,0,2,1,0,2,1}

Distinct Elements

- From the binary equivalent trailing zero values, write the value of maximum number of trailing zeros.

So, $r(a)$: {1,2,0,1,0,2,0,2,1,0,2,1}

$$R = \max r(a) = 2$$

$$\text{Estimate} = 2^R = 2^2 = 2*2 = 4$$

➤ Hence , there are 4 distinct elements as 1,3,2,4