LECTURE HANDOUTS

L01

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I- INTRODUCTION TO BIG DATA **Date of Lecture:** 22.02.2021

Topic of Lecture: Challenges of Conventional Systems

Introduction : (Maximum 5 sentences)

Three Challenges that big data face.

☆Data - The volume of data, especially machine-generated data, is exploding

☆Process - More than 80% of today's information is unstructured and it is typically too big to manage effectively.

☆Management - A lot of this data is unstructured, or has a complex structure that's hard to represent in rows and columns

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

Challenges of conventional system in big data

Three Challenges that big data face.

☆Data
☆Process
☆Management

Data or Volume

1. The volume of data, especially machine-generated data, is exploding,

2. How fast that data is growing every year, with new sources of data that are emerging.

3. For example, in the year 2000, 800,000petabytes (PB) of data were stored in the world, and it is expected to reach 35 zettabytes (ZB) by2020 (according to IBM).

Processing

More than 80% of today’s information is unstructured and it is typically too big to manage effectively.

Today, companies are looking to leverage a lot more.

Data from a wider variety of sources both inside and outside the organization.

Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

Management

A lot of this data is unstructured, or has a complex structure that’s hard to represent in rows and columns.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

https://snscourseware.org/snscenew/files/1563802060.pdf

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L02

MCA

I / II

Course Name with Code : 19CAB13 – Big Data Analytics

Course Faculty : Dr.M.Moorthy

Unit: I- INTRODUCTION TO BIG DATA Date of Lecture: 23.02.2021

Topic of Lecture: Intelligent data analysis

Introduction : (Maximum 5 sentences)

- It is one of the hot issues in the field of artificial intelligence and information.
- Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data.
- Intelligent data analysis is also a kind of decision support process.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

Intelligent Data Analysis (IDA)

- It is one of the hot issues in the field of artificial intelligence and information.
- Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data.
- Intelligent data analysis is also a kind of decision support process.
- Based on artificial intelligence, machine learning, pattern recognition, statistics, database and visualization technology mainly, IDA automatically extracts useful information, necessary knowledge and interesting models from a lot of online data in order to help decision makers make the right choices.

The process of IDA generally consists of the following three stages:

(1) data preparation;

(2) rule finding or data mining;

(3) result validation and explanation.

- Data preparation involves selecting the required data from the relevant data source and integrating this into a data set to be used for data mining.

- Rule finding is working out rules contained in the data set by means of certain methods or algorithms.
- Result validation requires examining these rules, and result explanation is giving intuitive, reasonable and understandable descriptions using logical reasoning.
- As the goal of intelligent data analysis is to extract useful knowledge, *the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on.*

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:1**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L03

MCA

I / II

Course Name with Code : 19CAB13 – Big Data Analytics

Course Faculty : Dr.M.Moorthy

Unit: I - INTRODUCTION TO BIG DATA Date of Lecture: 24.02.2021

Topic of Lecture: Nature of Data

Introduction : (Maximum 5 sentences)

Nature of Big Data

- *(i)* ***Volume –*** The name Big Data itself is related to a size which is enormous.
- *(ii)* ***Variety –*** The next aspect of Big Data is its **variety**.
- *(iii)* *(iii) **Velocity*** – The term **'velocity'** refers to the speed of generation of data.
- *(iv)* *(iv) **Variability*** – This refers to the inconsistency which can be shown by the data at times

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

Nature of Big Data

*(i) **Volume*** – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

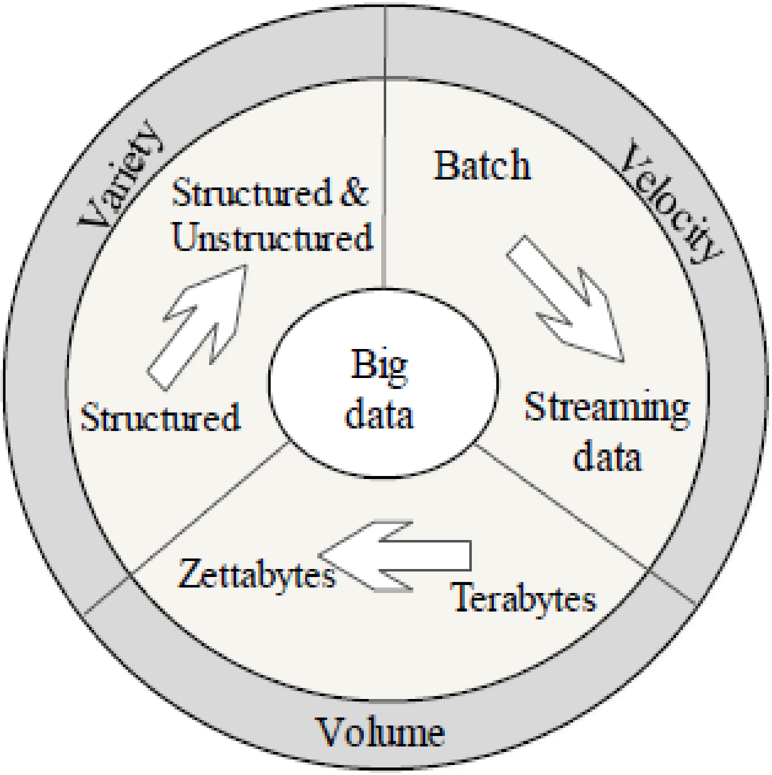
*(ii) **Variety*** – The next aspect of Big Data is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

*(iii) **Velocity*** – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

(iv) Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.



The Gartner's Vector model.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:8**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L04

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA **Date of Lecture:** 25.02.2021

Topic of Lecture: Analytic Processes and Tools

Introduction : (Maximum 5 sentences)

Best Analytic Processes and Big Data Tools

- Big data is the storage and analysis of large data sets.
- These are complex data sets which can be both structured and unstructured.
- They are so large that it is not possible to work on them with traditional analytical tools.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)
•Knowledge on Statistics
•Knowledge on DBMS

Detailed content of the Lecture:

Best Analytic Processes and Big Data Tools

- Big data is the storage and analysis of large data sets.
- These are complex data sets which can be both structured and unstructured.
- They are so large that it is not possible to work on them with traditional analytical tools.
- These days, organizations are realizing the value they get out of big data analytics and hence they are deploying big data tools and processes to bring more efficiency in their work environment.
- They are willing to hire good big data analytics professionals at a good salary.
- In order to be a big data analyst, you should get acquainted with big data first and get certification by enrolling yourself in analytics courses online.
- There are many big data tools and processes being utilised by companies these days.
- These are used in the processes of discovering insights and supporting decision making.
- The top big data tools used these days are open source data tools, data visualization tools, sentiment tools, data extraction tools and databases.

Some of the best used big data tools are mentioned below –

1. **R-Programming**
   - R is a free open source software programming language and a software environment for statistical computing and graphics.
   - It is used by data miners for developing statistical software and data analysis.
   - It has become a highly popular tool for big data in recent years.
2. **Data wrapper**
   - It is an online data visualization tool for making interactive charts.
   - You need to paste your data file in a csv, pdf or excel format or paste it directly in the field.
   - Datawrapper then generates any visualization in the form of bar, line, map etc.
   - It can be embedded into any other website as well. It is easy to use and produces visually effective charts.

3. **Tableau Public**

- Tableau is another popular big data tool.
- It is simple and very intuitive to use.
- It communicates the insights of the data through data visualisation.
- Through Tableau, an analyst can check a hypothesis and explore the data before starting to work on it extensively.

4. **Content Grabber**

- Content Grabber is a data extraction tool.
- It is suitable for people with advanced programming skills.
- It is a web crawling software.
- Businesses can use it to extract content and save it in a structured format.
- It offers editing and debugging facility among many others for analysis later.
- The market is full of big data tools these days. These tools help unlock the power that big data provides to business processes.
- By choosing the tools carefully, a company can increase its efficiency in its operations.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streamswith Advanced Analytics", John Wiley & sons, 2012 (**Page No:154**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L05

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA **Date of Lecture:** 26.02.2021

Topic of Lecture: Analysis vs. Reporting

Introduction : (Maximum 5 sentences)

Analysis Vs. Reporting

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.*

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:
Analysis vs. Reporting

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.*

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.

Through the process of performing **analysis** you may raise **additional questions**, but the goal is to

identify answers, or at least potential answers **that can be tested**.

In summary, reporting shows you ***what is happening*** while analysis focuses on explaining ***why it is happening*** and ***what you can do about it.***

Reporting provides data in a static and structured way, measuring and monitoring business performance. It **answers the "what" questions**.

In order to be a business enabler and stay on top of the competitive marketplace, the **"why," "how," and "when" are important parts of** the equation as well.

That's where **analysis** comes into play.

It doesn't just provide data, but provides answers behind the numbers.

It **takes your data and converts it into intelligence and insights.**

Reporting: Push Information to the Users

- Reporting is the information gathering part of a data-driven decision making process.
- Every organization has particular key performance indicators (KPIs) that they are looking to meet or a data range to stay between.
- Reporting provides the first alert system if something becomes wrong and raises the 'what is happening' flag, but it will not explain why these changes are relevant or not.
- With reporting, you take raw data, consolidate and format it, then push the summaries out to users in the form of static reports (that encompass fixed metrics and dimensions), dashboards (custom-made reports that provide a high-level view of business performance for particular audiences), or alerts (special reports sent if data falls outside a certain parameters).

Analysis: Pull Those Insights

- The ultimate goal of analysis is to bring deeper meaning to the reported data, ultimately using it in an actionable way in the organization.
- Reporting raises questions and analysis will answer them, with additional questions that may come to light.
- When analysis is being done, it's like conducting research — there are questions to be answered, data is examined, interpretations and comparisons are done, and then confirmations are distributed to answer **what has been happening**, **why it was taking place**, **when it took place**, and **the next steps toward resolution**.

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streamswith Advanced Analytics", John Wiley & sons, 2012 (**Page No:179**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L06

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA **Date of Lecture:** 01.03.2021

Topic of Lecture: Modern Data Analytic Tools

Introduction : (Maximum 5 sentences)
Modern Data Analytic Tools Available in Big Data

1. Data Storage and Management Tools
2. Data Cleaning Tools
3. Data Mining Tools
4. Data Analysis
5. Data Visualization

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

Modern Data Analytic Tools Available in Big Data

1. Data Storage and Management Tools

Storing a "Big Data" is really competitive task in major data centric organizations. Hence modern data storage and management tools can be used to handle many storage and management challenges. Such tools include Hadoop, Cloudera, and MongoDB.

1.1 Hadoop

Hadoop is open source distributed software for handling large datasets stored on computer clusters. Hence it provides wider storage, higher processing ability, and supports any number of virtual concurrent tasks. The implementation in Hadoop requires complete knowledge in java.

1.2 Mongo DB

Mongo DB serves the similar purpose as relational database.

It handles structured, semi-structured, unstructured and dynamic data.

With Mongo DB the operational overhead can be reduced up to 95%.

With the help of Wired Tiger storage engine new architectures for flexible storage are achieved.

Nearly seven to ten times better performance has been achieved.

Storage space utilization is reduced nearly 80% with compression of data.

1.3 Cloudera

Cloudera comes in hand with Hadoop.

It is also open source software which provides additional functionalities at an enterprise level.

It also provides certain level of security to the sensitive big data stored on enterprise computers.

1.4 Talend

Talend is an open source software company providing extensive data management to any kind of business firms using the concept of Master Data Management (MDM) System.

2. Data Cleaning Tools

Data analytics require complete cleaning of data before mining.

Data coming from web requires complete cleaning of data because the data may not be structured or it may be of different size.

Various data cleaning tools are Open Refine, Data Cleaner, etc.

2.1 Open Refine

The unstructured big data supported by wider community can be cleansed using Open Refine and Open Refine Wiki which make them user friendly.

2.2 Data Cleaner

Data visualization requires clean and structured data which is performed by Data Cleaner.

3. Data Mining Tools

Based on the data that we have data mining tools help us to predict and make decisions intelligently.

Various data mining tools are Rapid Miner, IBM SPSS Modeler, Oracle data mining etc.

3.1 Rapid Miner

Rapid Miner is mainly used for predictive analysis.

Even we can build our own algorithms for predictive analysis and can be fed into Rapid Miner using API's.

Working in Rapid Miner is easy because of graphical interface which does not require any prior knowledge in coding.

3.2 IBM SPSS Modeler

IBM SPSS Modeler mainly suits for performing mining in big companies.

It provides set of products for text analytics, decision support management, data optimization techniques, entity analytics, etc.

3.3 Oracle Data Mining

Oracle data mining is another big data analytics tool which provides greater insights into data mining.

It helps to build greater predictive systems, build customers model and identifies best customers from the past customer histories.

Data scientists as well as analyst, and Business analyst, build successful predictive models using GUI provided by Oracle Data Mining.

4. Data Analysis

Data Analysis is the task of finding answers for unanswered questions from the pattern mining. Big ML is identified as one of the best tools for Data Analysis.

4.1 Big ML

Big ML is a good machine learning tool.

It provides enhanced and efficient GUI to which we can load our data and get predictions out of that data.

It is also used for predictive analysis. It also provides a virtual private cloud environment to encage enterprise level problems.

5. Data Visualization

With the help of data visualization, one can view the complex and large set of numeric data in an easy diagrammatic and understandable way.

Such tools include Tableau, Silk, Carto DB, Chartio, Plot.ly, and Data wrapper.

Tableau, Silk, Carto DB, Chartio, Plot.ly

5.1 Tableau is mainly used in creating business intelligence.

One can create maps, scatter plots, bar charts without using any programming.

One can get live real time data by connecting to a database or API for real time visualization.

5.2 Silk is similar to Tableau one can create maps and charts with the click of mouse.

5.3 Carto DB is mainly used in making maps with the help of location data.

5.4 Chartio is mainly used in creating highly complex dashboards with a few clicks.

5.5 Plot.ly is used to create efficient 2D and 3D graphs without any programming.

Other Tools

There are many other tools like Block spring, Pentaho, etc. and also data languages like R programming, Python, RegEx, Xpath, etc. available for data integration.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:12**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L07

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA **Date of Lecture:** 02.03.2021

Topic of Lecture: Sampling Distributions & Re-Sampling

Introduction : (Maximum 5 sentences)

Sampling Distribution

- A random sample is a sampling method whereby each member of the population has an equal chance or probability of being selected.
- Statisticians have conducted detailed investigations of the behavior of statistics which are obtained on the basis of random sampling.
- The manner, in which a statistic such as the mean is distributed, when many random samples are drawn from a population, is referred to as the sampling distribution of the statistic.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

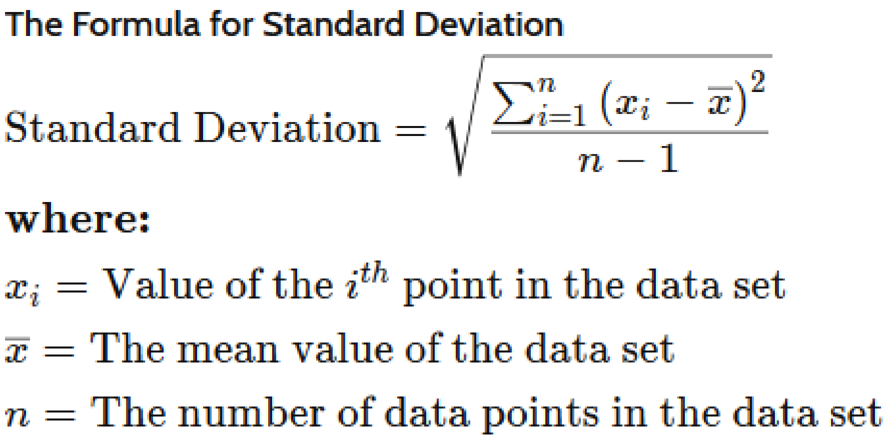
•Knowledge on Statistics

•Knowledge on DBMS

Detailed content of the Lecture:

Sampling Distribution

- A random sample is a sampling method whereby each member of the population has an equal chance or probability of being selected.
- Statisticians have conducted detailed investigations of the behavior of statistics which are obtained on the basis of random sampling.
- The manner, in which a statistic such as the mean is distributed, when many random samples are drawn from a population, is referred to as the sampling distribution of the statistic.
- The sampling distribution of a statistics is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population.
- It is a theoretical idea—we do not actually build it.
- The sampling distribution of a statistic is the probability distribution of that statistic.



The Formula for Standard Deviation

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

where:

x_i = Value of the i^{th} point in the data set

$\overline{x}$ = The mean value of the data set

n = The number of data points in the data set

Calculating the Standard Deviation

Standard deviation is calculated as follows:

1. The mean value is calculated by adding all the data points and dividing by the number of data points.
2. The variance for each data point is calculated by subtracting the value of the data point from the mean. Each of those resulting values is then squared and the results summed. The result is then divided by the number of data points less one.
3. The square root of the variance—results from no. 2—is then used to find the standard deviation.

Example of Standard Deviation

Say we have the data points 5, 7, 3, and 7, which total 22.

You would then divide 22 by the number of data points, in this case, four—resulting in a mean of 5.5.

This leads to the following determinations: $\bar{x} = 5.5$ and $N = 4$.

The variance is determined by subtracting the mean's value from each data point, resulting in -0.5, 1.5, -2.5, and 1.5.

Each of those values is then squared, resulting in 0.25, 2.25, 6.25, and 2.25.

The square values are then added together, giving a total of 11, which is then divided by the value of N minus 1, which is 3, resulting in a variance of approximately 3.67.

The square root of the variance is then calculated, which results in a standard deviation measure of approximately 1.915.

Resampling is the method that consists of drawing repeated samples from the original data samples.

The method of Resampling is a nonparametric method of statistical inference.

In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.

Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter.
Resampling methods, in fact, make use of a nested **resampling** method.

Some aspects to consider prior to collecting a data sample include:

- **Sample Goal**. The population property that you wish to estimate using the sample.
- **Population**. The scope or domain from which observations could theoretically be made.
- **Selection Criteria**. The methodology that will be used to accept or reject observations in your sample.
- **Sample Size**. The number of observations that will constitute the sample.
- **Sampling** is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population.
- **The Main Characteristics of Sampling**
- In sampling, we assume that samples are drawn from the population and sample means and population means are equal. A population can be defined as a whole that includes all items and characteristics of the research taken into study. However, gathering all this information is time consuming and costly. We therefore make inferences about the population with the help of samples.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No: 29 & 68**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L08

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA

Date of Lecture: 03.03.2021

Topic of Lecture: Statistical Inference

Introduction : (Maximum 5 sentences)

Statistical inference consists in the use of **statistics** to draw conclusions about some unknown aspect of a population based on a random sample from that population. **Statistical inference** can be divided into two areas: estimation and hypothesis testing.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

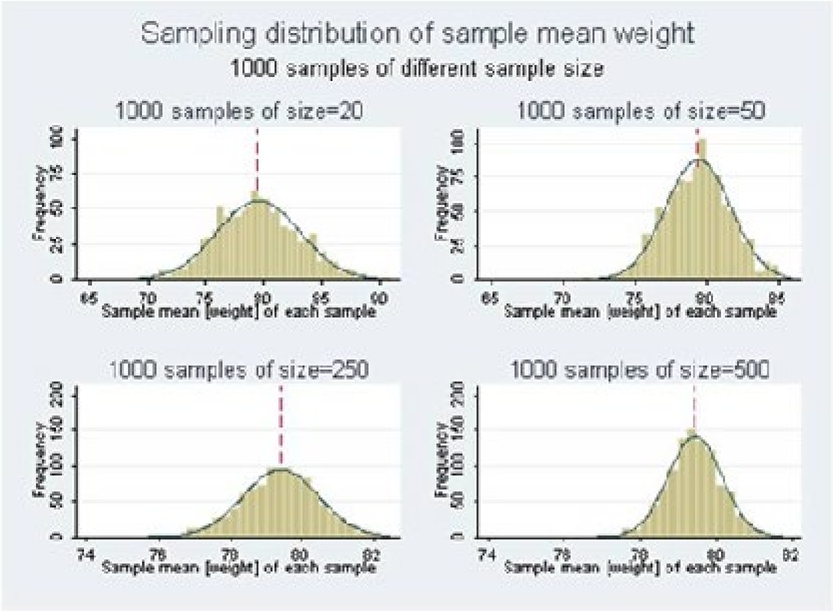
Statistical inference consists in the use of **statistics** to draw conclusions about some unknown aspect of a population based on a random sample from that population. **Statistical inference** can be divided into two areas: estimation and hypothesis testing.

The purpose of **statistical inference** is to estimate this sample to sample variation or uncertainty.
The **four pillars of statistical inference**

Significance : How strong is the evidence?
Generalisation : How broadly do the results apply?
Estimation : How large is the effect?
Causation : Can we say what caused the effect?

Basic concepts of statistical inference

Histograms of sample means



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:33**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L09

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: I - INTRODUCTION TO BIG DATA **Date of Lecture:** 04.03.2021

Topic of Lecture: Prediction Error

Introduction : (Maximum 5 sentences)

- A **prediction error** is the failure of some expected event to occur.
- **Errors** are an inescapable element of **predictive** analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its **predictions** are expected to be.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Statistics**
- **Knowledge on DBMS**

Detailed content of the Lecture:

- A **prediction error** is the failure of some expected event to occur.
- **Errors** are an inescapable element of **predictive** analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its **predictions** are expected to be.
- **Prediction error** quantifies one of two things:
- In regression analysis, it's a measure of how well the model predicts the response variable.
- In classification (machine learning), it's a measure of how well samples are classified to the correct category.
- When predictions fail, humans can use metacognitive functions, examining prior predictions and failures and deciding, for example, whether there are correlations and trends, such as consistently being unable to foresee outcomes accurately in particular situations.
- Applying that type of knowledge can inform decisions and improve the quality of future predictions.
- In artificial intelligence (AI), the analysis of prediction errors can help guide machine learning (ML), similarly to the way it does for human learning.
- In reinforcement learning, for example, an agent might use the goal of minimizing error feedback as a way to improve.

- Prediction errors, in that case, might be assigned a negative value and predicted outcomes a positive value, in which case the AI would be programmed to attempt to maximize its score.
- That approach to ML, sometimes known as error-driven learning, seeks to stimulate learning by approximating the human drive for mastery.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:46**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L10

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 05.03.2021

Topic of Lecture: Stream Data Model and Architecture

Introduction : (Maximum 5 sentences)
Streaming data refers to data that is **continuously generated**, usually in **high volumes** and at **high velocity**. A streaming data source would typically consist of a stream of logs that record events as they happen – such as a user clicking on a link in a web page, or a sensor reporting the current temperature.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

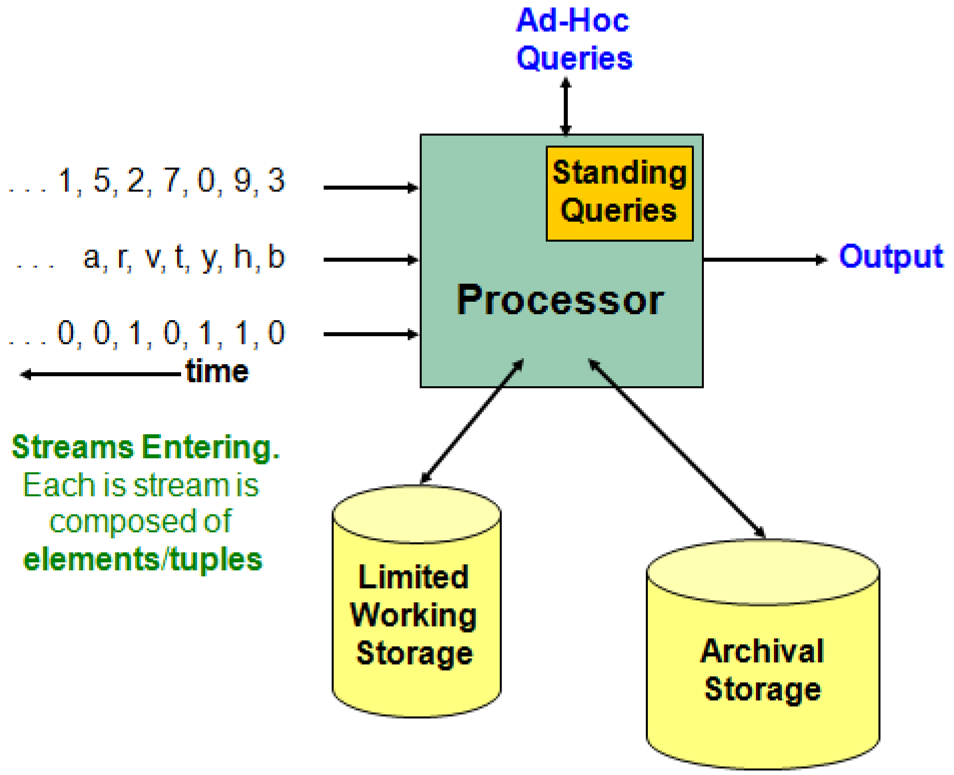
Detailed content of the Lecture:

Streaming data refers to data that is **continuously generated**, usually in **high volumes** and at **high velocity**. A streaming data source would typically consist of a stream of logs that record events as they happen – such as a user clicking on a link in a web page, or a sensor reporting the current temperature.

Common examples of streaming data include:

- IoT sensors
- Server and security logs
- Real-time advertising
- Click-stream data from apps and websites

General Stream Processing Model



Streams Entering. Each is stream is composed of **elements/tuples**

Modern Streaming Architecture

- In modern streaming data deployments, many organizations are adopting a full stack approach rather than relying on patching together open-source technologies.
- The modern data platform is built on business-centric value chains rather than IT-centric coding processes, wherein the complexity of traditional architecture is abstracted into a single self-service platform that turns event streams into analytics-ready data.
- The idea behind Upsolver is to act as the centralized data platform that automates the labor-intensive parts of working with streaming data: message ingestion, batch and streaming ETL, storage management and preparing data for analytics.

Benefits of a modern streaming architecture:

- Can eliminate the need for large data engineering projects
- Performance, high availability and fault tolerance built in
- Newer platforms are cloud-based and can be deployed very quickly with no upfront investment
- Flexibility and support for multiple use cases

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CambridgeUniversity Press, 2014 (**Page No:131**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L11

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 06.03.2021

Topic of Lecture: Stream Computing

Introduction : (Maximum 5 sentences)

Stream Computing

The stream processing computational paradigm consists of assimilating data readings from collections of software or hardware sensors in stream form (i.e., as an infinite series of tuples), analyzing the data, and producing actionable results, possibly in stream format as well.

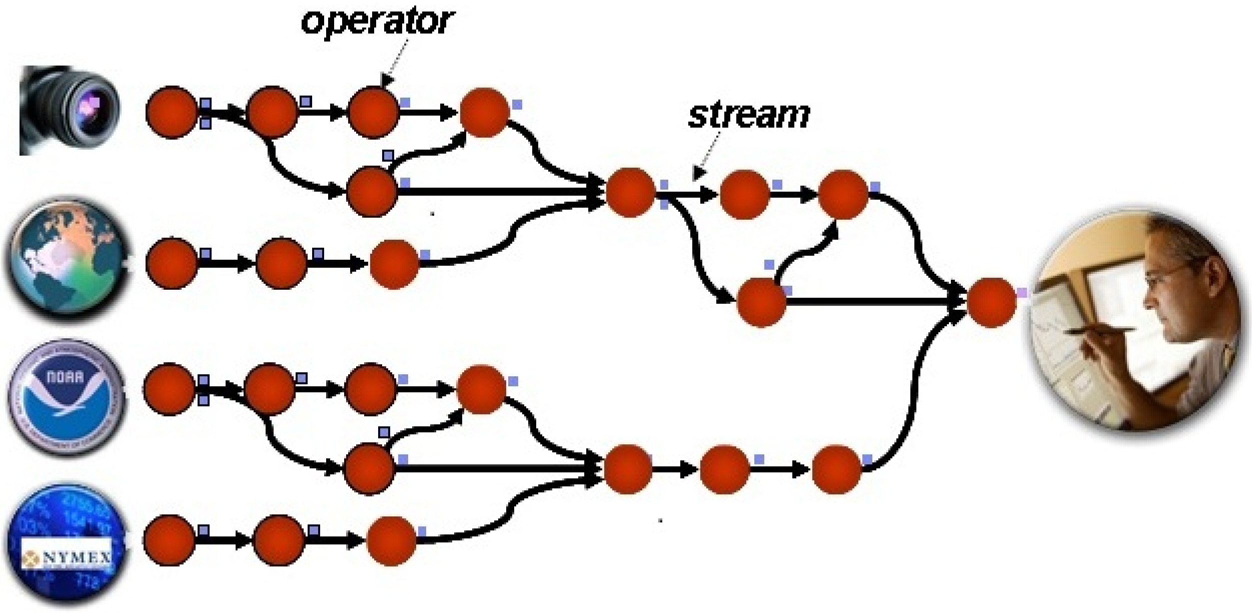
Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Stream Computing

- The stream processing computational paradigm consists of assimilating data readings from collections of software or hardware sensors in stream form (i.e., as an infinite series of tuples), analyzing the data, and producing actionable results, possibly in stream format as well.



- In a stream processing system, applications typically act as continuous queries, ingesting data continuously, analyzing and correlating the data, and generating a stream of results.
- Applications are represented as data-flow graphs composed of operators and interconnected by streams, as shown in the figure.
- The individual operators implement algorithms for data analysis, such as parsing, filtering, feature extraction, and classification.
- Such algorithms are typically single-pass because of the high data rates of external feeds (e.g., market information from stock exchanges, environmental sensors readings from sites in a forest, etc.).
- Stream processing applications are usually constructed to identify new information by incrementally building models and assessing whether new data deviates from model predictions and, thus, is interesting in some way.
- For example, in a financial engineering application, one might be constructing pricing models for options on securities, while at the same time detecting mispriced quotes, from a live stock market feed.
- Streams applications may consist of dozens to hundreds of analytic operators, deployed on production systems hosting many other potentially interconnected stream applications, distributed over a large number of processing nodes.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

https://researcher.watson.ibm.com/researcher/view_group.php?id=2531

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L12

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 08.03.2021

Topic of Lecture: Sampling Data in a Stream

Introduction : (Maximum 5 sentences)

Sampling from a Data Stream

- Since **we cannot store the entire stream,** one obvious approach is to store a **sample**

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Sampling from a Data Stream

- Since **we cannot store the entire stream,** one obvious approach is to store a **sample**
- **Two different problems:**
  - **(1)** Sample a **fixed proportion** of elements in the stream (say 1 in 10)
  - **(2)** Maintain a **random sample of fixed size** over a potentially infinite stream
    - At any "time" ***k*** we would like a random sample of ***s*** elements
      - **What is the property of the sample we want to maintain?** For all time steps ***k***, each of ***k*** elements seen so far has equal prob. of being sampled
- **Types of queries one wants on answer on a data stream:**

- **Filtering a data stream**
    - Select elements with property ***x*** from the stream
- **Counting distinct elements**
    - Number of distinct elements in the last ***k*** elements of the stream
- **Estimating moments**
    - Estimate avg./std. dev. of last ***k*** elements
- **Finding frequent elements**

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CambridgeUniversity Press, 2014 (**Page No:136**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L13

MCA

I / II

Course Name with Code : 19CAB13 – Big Data Analytics

Course Faculty : Dr.M.Moorthy

Unit: II - MINING DATA STREAMS Date of Lecture: 09.03.2021

Topic of Lecture: Filtering Streams

Introduction : (Maximum 5 sentences)

Filtering Data Streams

- **Each element of data stream is a tuple**
- Given a list of keys **S**
- **Determine which tuples of stream are in *S***
- **Obvious solution: Hash table**
  - But suppose we **do not have enough memory** to store all of ***S*** in a hash table
  - E.g., we might be processing millions of filters on the same stream

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Filtering Data Streams

- **Each element of data stream is a tuple**
- Given a list of keys **S**
- **Determine which tuples of stream are in *S***
- **Obvious solution: Hash table**
  - But suppose we **do not have enough memory** to store all of ***S*** in a hash table
  - E.g., we might be processing millions of filters on the same stream

Applications

- **Example: Email spam filtering**
    - We know 1 billion “good” email addresses
    - If an email comes from one of these, it is **NOT** spam
- **Publish-subscribe systems**
    - You are collecting lots of messages (news articles)
    - People express interest in certain sets of keywords
    - Determine whether each message matches user’s interest

Bloom Filters by Example

- A Bloom filter is a data structure designed to tell you, rapidly and memory-efficiently, whether an element is present in a set.
- The price paid for this efficiency is that a Bloom filter is a **probabilistic data structure**: it tells us that the element either *definitely is not* in the set or *may be* in the set.
- The base data structure of a Bloom filter is a **Bit Vector**. Here's a small one we'll use to demonstrate:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

- Each empty cell in that table represents a bit, and the number below it its index. To add an element to the Bloom filter, we simply hash it a few times and set the bits in the bit vector at the index of those hashes to 1.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:139**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L14

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 10.03.2021

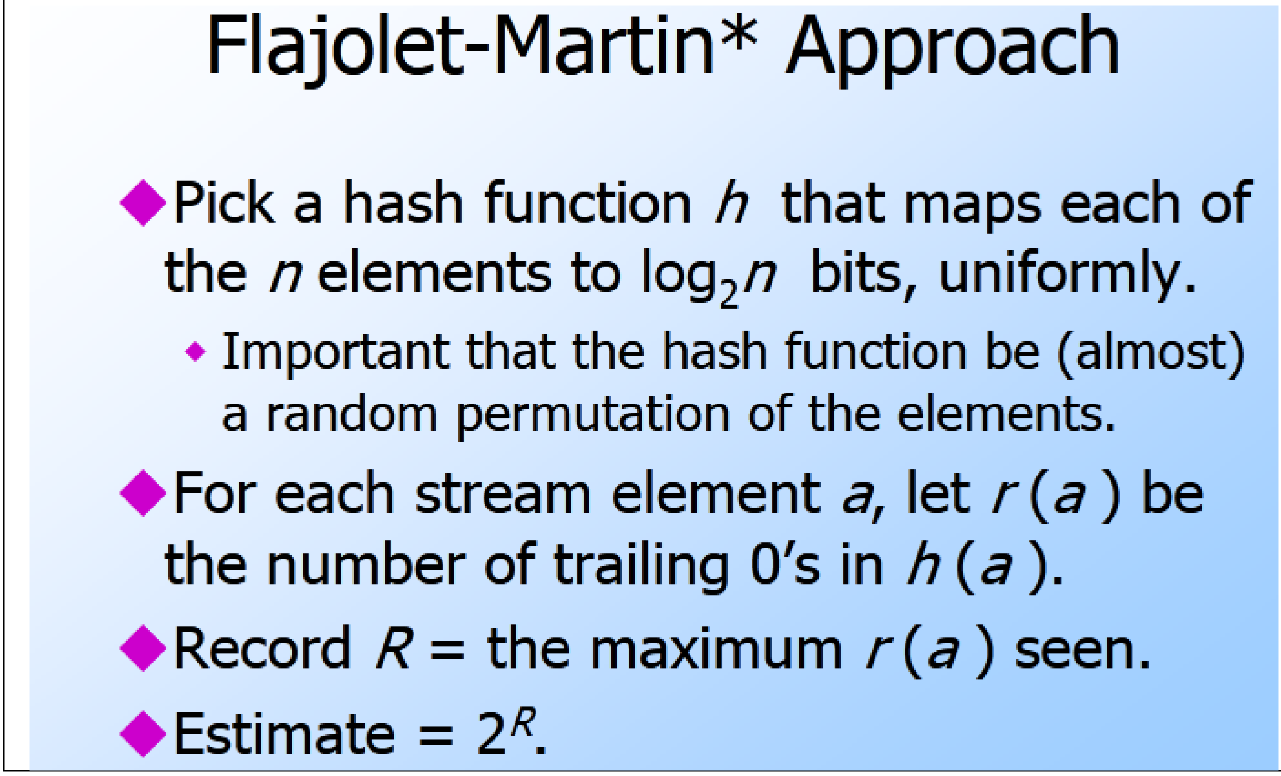
Topic of Lecture: Counting Distinct Elements in a Stream

Introduction : (Maximum 5 sentences)
In computer science, the **count-distinct** problem (also known in applied mathematics as the cardinality estimation problem) is the problem of finding the number of **distinct elements** in a **data stream** with repeated **elements**. This is a well-known problem with numerous applications.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:



Flajolet-Martin* Approach

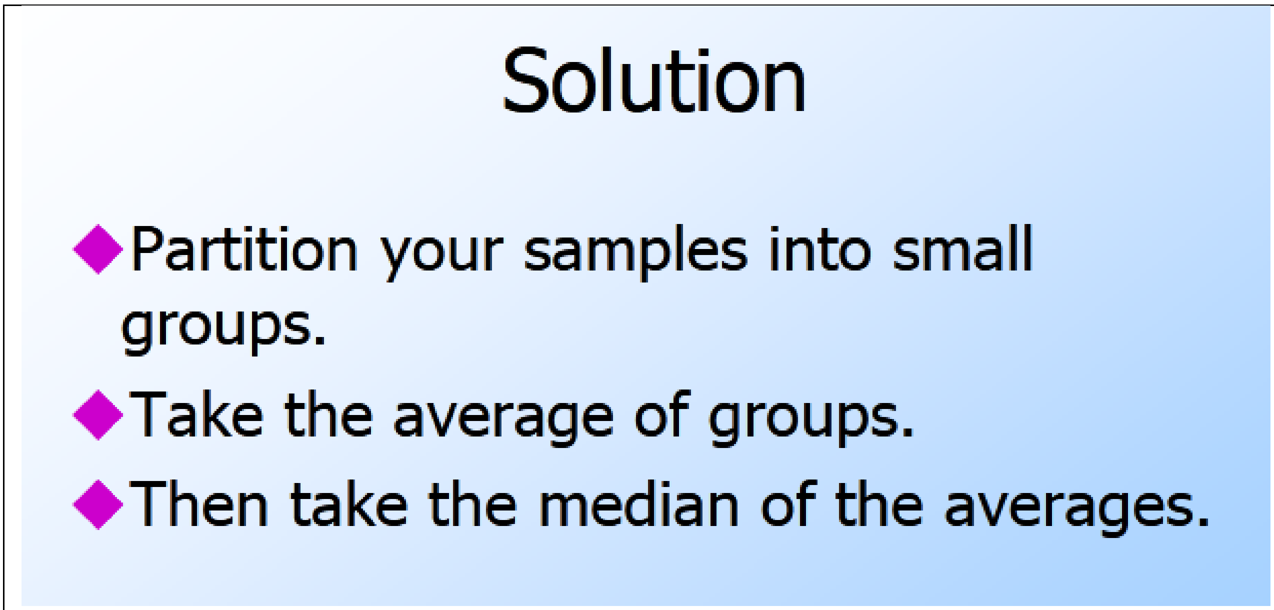
- Pick a hash function *h* that maps each of the *n* elements to $\log_2 n$ bits, uniformly.
  - Important that the hash function be (almost) a random permutation of the elements.
- For each stream element *a*, let *r* (*a*) be the number of trailing 0's in *h* (*a*).
- Record *R* = the maximum *r* (*a*) seen.
- Estimate = 2^R.

Why It Works

- The probability that a given element a has $h(a) \geq r$ is 2^{-r}.
- If there are m elements in the stream, the probability that $R \geq r$ is $1 – (1 - 2^{-r})^m$.
- If $2^r >> m$, prob $\approx m / 2^r$ (small).
- If $2^r << m$, prob ≈ 1.
- Thus, 2^R will almost always be around m.

Why It Doesn't Work

- $E(2^R)$ is actually infinite.
  - Probability halves when R -> $R+1$, but value doubles.
- That means using many hash functions and getting many samples.
- How are samples combined?
  - Average? What if one very large value?
  - Median? All values are a power of 2.



Solution

- Partition your samples into small groups.
- Take the average of groups.
- Then take the median of the averages.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:142**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L15

MCA

I / II

Course Name with Code : 19CAB13 – Big Data Analytics

Course Faculty : Dr.M.Moorthy

Unit: II - MINING DATA STREAMS **Date of Lecture:** 11.03.2021

Topic of Lecture: Estimating Moments

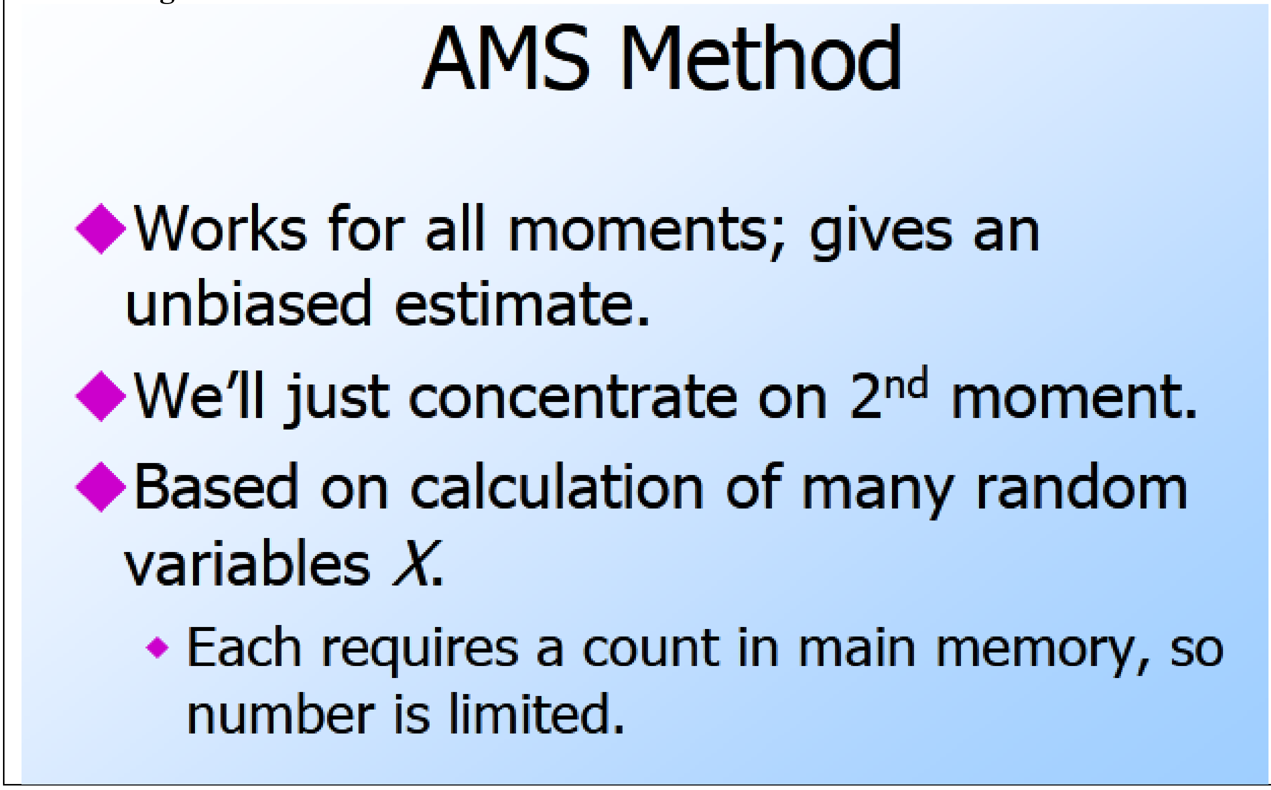
Introduction : (Maximum 5 sentences)
Estimating moment is a process that is carried out in a data stream using different algorithms. AMS method is one among them.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- Knowledge on Mathematics
- Knowledge on Algorithms
- Knowledge on GUI programming

Detailed content of the Lecture:

Estimating Moments



AMS Method

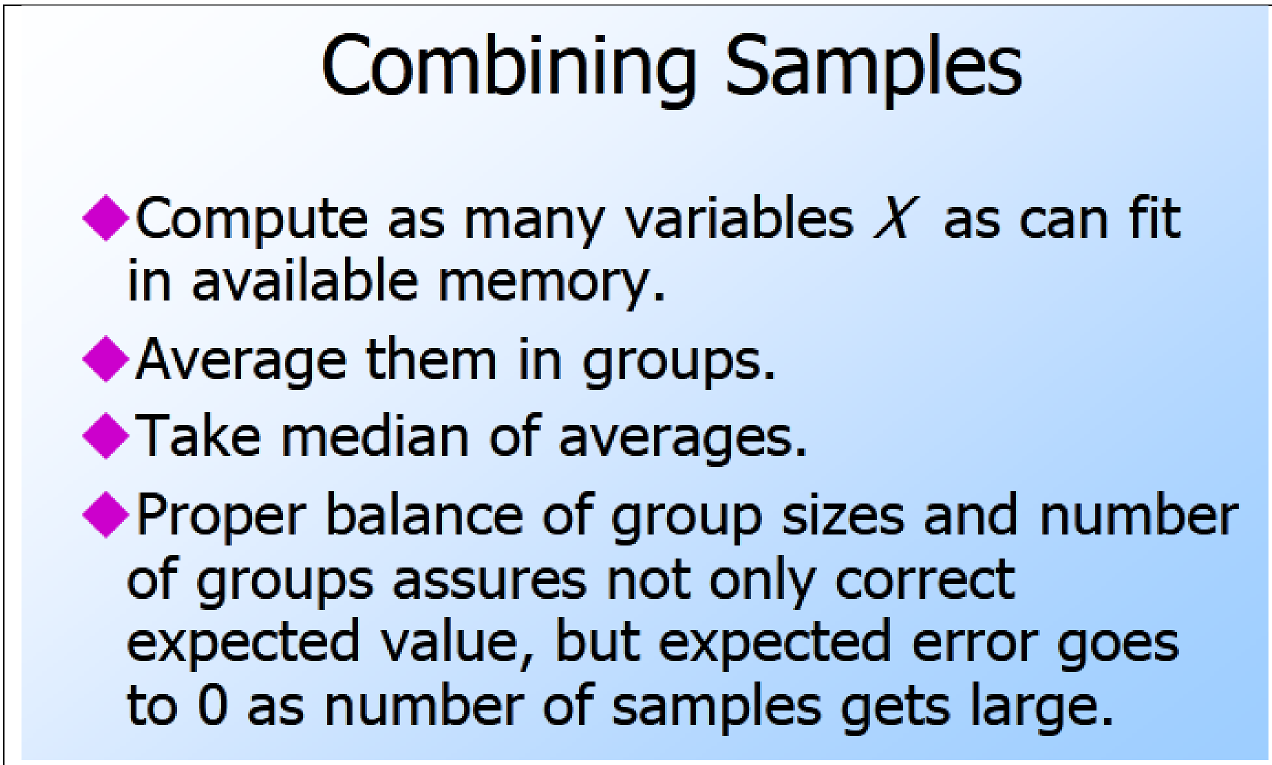
- Works for all moments; gives an unbiased estimate.
- We'll just concentrate on 2nd moment.
- Based on calculation of many random variables *X*.
  - Each requires a count in main memory, so number is limited.

One Random Variable

- Assume stream has length n.
- Pick a random time to start, so that any time is equally likely.
- Let the chosen time have element a in the stream.
- $X = n *$ ((twice the number of a's in the stream starting at the chosen time) – 1).

Expected Value of X

- 2^{nd} moment is $\Sigma_a (m_a)^2$.
- $E(X) = (1/n)\big(\Sigma_{\text{all times } t}$ of $n *$ (twice the number of times the stream element at time t appears from that time on) – 1$\big)$.
- $= \Sigma_a (1/n)(n)(1+3+5+\ldots+2m_a-1)$.
- $= \Sigma_a (m_a)^2$.



Combining Samples

- Compute as many variables X as can fit in available memory.
- Average them in groups.
- Take median of averages.
- Proper balance of group sizes and number of groups assures not only correct expected value, but expected error goes to 0 as number of samples gets large.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:145**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L16

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 12.03.2021

Topic of Lecture: Counting Oneness in a Window

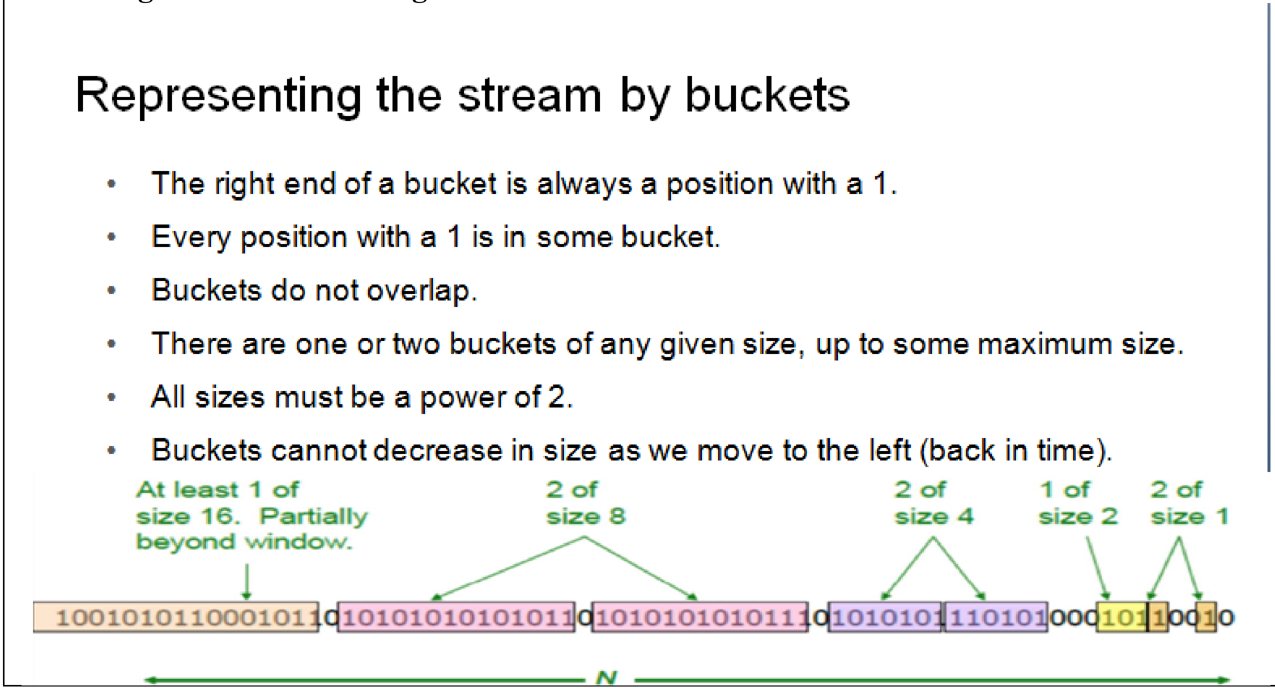
Introduction : (Maximum 5 sentences)
It is a process that is carried out in a data stream using DGIM algorithm. It represents the stream by buckets.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Counting bits with DGIM algorithm



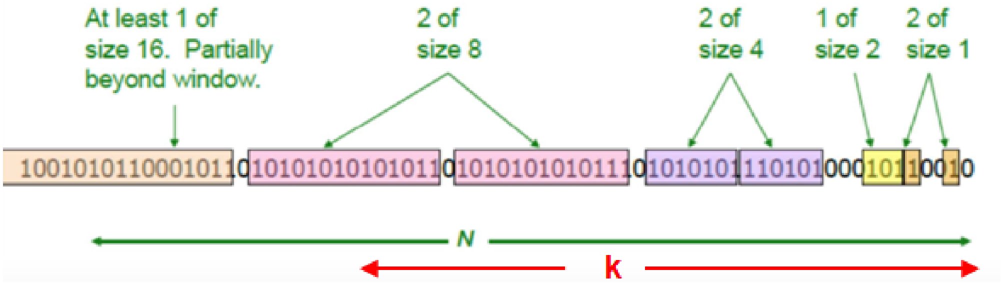
Representing the stream by buckets

- The right end of a bucket is always a position with a 1.
- Every position with a 1 is in some bucket.
- Buckets do not overlap.
- There are one or two buckets of any given size, up to some maximum size.
- All sizes must be a power of 2.
- Buckets cannot decrease in size as we move to the left (back in time).

How many ones are in the most recent k bits?

- Find all buckets **overlapping** with last k bits
- Sum the sizes of all but the oldest one
- Add the half of the size of the oldest one

Ans = 1 + 1 + 2 + 4 + 4 + 8 + 8/2 = 24



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CambridgeUniversity Press, 2014 (**Page No:150**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L17

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 13.03.2021

Topic of Lecture: Real time Analytics Platform(RTAP)Application

Introduction : (Maximum 5 sentences)

Real time Analytics Platform (RTAP) Applications

The power of streaming analytics is such that it allows for the streaming of millions of events in a second and thus allows enterprises to build mission-critical applications that require the performance to be quick and efficient.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Real time Analytics Platform (RTAP) Applications

- The power of streaming analytics is such that it allows for the streaming of millions of events in a second and thus allows enterprises to build mission-critical applications that require the performance to be quick and efficient.
- Real-time streaming analytics can, for example, present to you the statistics if your latest online ad campaign is working as expected, or if it needs some tweaking to work better. Such applications want to always stay upgraded for performance benefits.

Here are the top platforms being used all over the world for streaming analytics solutions:

Apache Flink

- Flink is an open-source platform that handles distributed stream and batch data processing. At its core is a streaming data engine that provides for data distribution, fault tolerance, and communication, for undertaking distributed computations over the data streams.
- In the last year, the Apache Flink community saw three major version releases for the platform and the community event Flink Forward in San Francisco.
- Apache Flink's open-source contributors on Github have increased from 258 in December 2016 to 352 in December 2017.
- Apache Flink contains several APIs to enable creating applications that use the Flink engine.

- Some of the most popular APIs on the platform are- DataStream API for unbounded streams, DataSet API for static data embedded in Python, Java, and Scala, and the Table API with a SQL-like language.

Spark Streaming

- Apache Spark is used to build scalable and fault-tolerant streaming applications.
- With Spark Streaming, you get to use Apache Spark's language-integrated API which lets you write streaming jobs in the similar way as you write batch jobs.
- Spark Streaming supports the three languages- Java, Scala, Python.
- Apache Spark is being used in various leading industries today, such as- Healthcare, Finance, e-commerce, Media and Entertainment, Travel industry, etc.
- The popularity of Apache Spark adds the glitter to the platform Spark Streaming.

IBM Streams

- This streaming analytics platform from IBM enables the applications developed by users to gather, analyze, and correlate information that comes to them from a variety of sources.
- The solution is known to handle high throughput rates and up to millions of events and messages per second, making it a leading proprietary streaming analytics solution for real-time applications.
- IBM Stream computing helps analyze large streams of data in the form of unstructured texts, audio, video, and geospatial, and allows for organizations to spot risks and opportunities and make efficient decisions.

Software AG's Apama Streaming Analytics

- Apama Streaming analytics platform is built for streaming analytics and automated action on fast-moving data on the basis of intelligent decisions.
- The software bundles up other aspects like messaging, event processing, in-memory data management and visualization and is ideal for fast-moving Big Data Analytics Solutions.
- Sensors that bring in loads of data from different sources can be churned using this solution in real-time.
- With Apama, you can act on high-volume business operations in real-time.

Azure Stream Analytics

- Azure Stream Analytics facilitates the development and deployment of low-cost solutions that can gain real-time insights from devices, applications, and sensors.
- It is recommended to be used for IoT scenarios like real-time remote management and monitoring, connected cars, etc.

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:
https://www.dauniv.ac.in/public/frontassets/coursematerial/
BigDataAnallyticsPPTs/BDACh07L06RealTimeAnalyticsPlatform.pdf

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L18

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: II - MINING DATA STREAMS **Date of Lecture:** 15.03.2021

Topic of Lecture: Case Studies

Introduction : (Maximum 5 sentences)

Case Studies

Streams may be archived in a large archival store, but we assume it is not possible to answer queries from the archival store. It could be examined only under special circumstances using time-consuming retrieval processes.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Mathematics**
- **Knowledge on Algorithms**
- **Knowledge on GUI programming**

Detailed content of the Lecture:

Case Studies

- Streams may be archived in a large archival store, but we assume it is not possible to answer queries from the archival store. It could be examined only under special circumstances using time-consuming retrieval processes.
- There is also a working store, into which summaries or parts of streams may be placed, and which can be used for answering queries. The working store might be disk, or it might be main memory, depending on how fast we need to process queries. But either way, it is of sufficiently limited capacity that it cannot store all the data from all the streams.

Sensor Data

- Imagine a temperature sensor bobbing about in the ocean, sending back to a base station a reading of the surface temperature each hour. The data produced by this sensor is a stream of real numbers. It is not a very interesting stream, since the data rate is so low. It would not stress modern technology, and the entire stream could be kept in main memory, essentially forever. Now, give the sensor a GPS unit, and let it report surface height instead of temperature.
- The surface height varies quite rapidly compared with temperature, so we might have the sensor send back a reading every tenth of a second. If it sends a 4-byte real number each time, then it produces 3.5 megabytes per day. It will still take some time to fill up main memory, let alone a single disk. But one sensor might not be that interesting.
- To learn something about ocean behavior, we might want to deploy a million sensors, each sending back a stream, at the rate of ten per second. A million sensors isn't very many; there would be one for every 150 square miles of ocean. Now we have 3.5 terabytes arriving every day, and we definitely need to think about what can be kept in working storage and what can only be archived.

Image Data

Satellites often send down to earth streams consisting of many terabytes of images per day. Surveillance cameras produce images with lower resolution than satellites, but there can be many of them, each producing a stream of images at intervals like one second. London is said to have six million such cameras, each producing a stream.

Internet and Web Traffic

- A switching node in the middle of the Internet receives streams of IP packets from many inputs and routes them to its outputs. Normally, the job of the switch is to transmit data and not to retain it or query it. But there is a tendency to put more capability into the switch, e.g., the ability to detect denial-of-service attacks or the ability to reroute packets based on information about congestion in the network.
- Web sites receive streams of various types. For example, Google receives several hundred million search queries per day. Yahoo! Accepts billions of “clicks” per day on its various sites. Many interesting things can be learned from these streams. For example, an increase in queries like “sore throat” enables us to track the spread of viruses. A sudden increase in the click rate for a link could indicate some news connected to that page, or it could and that the link is broken and needs to be repair.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:
https://www.semanticscholar.org/paper/A-Case-Study%3A-Stream-Data-Mining-Classification-Desale-Ade/8ef869c91fcb3bb042677c2a470f40dc7fd413f7

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L19

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 16.03.2021

Topic of Lecture: History of Hadoop

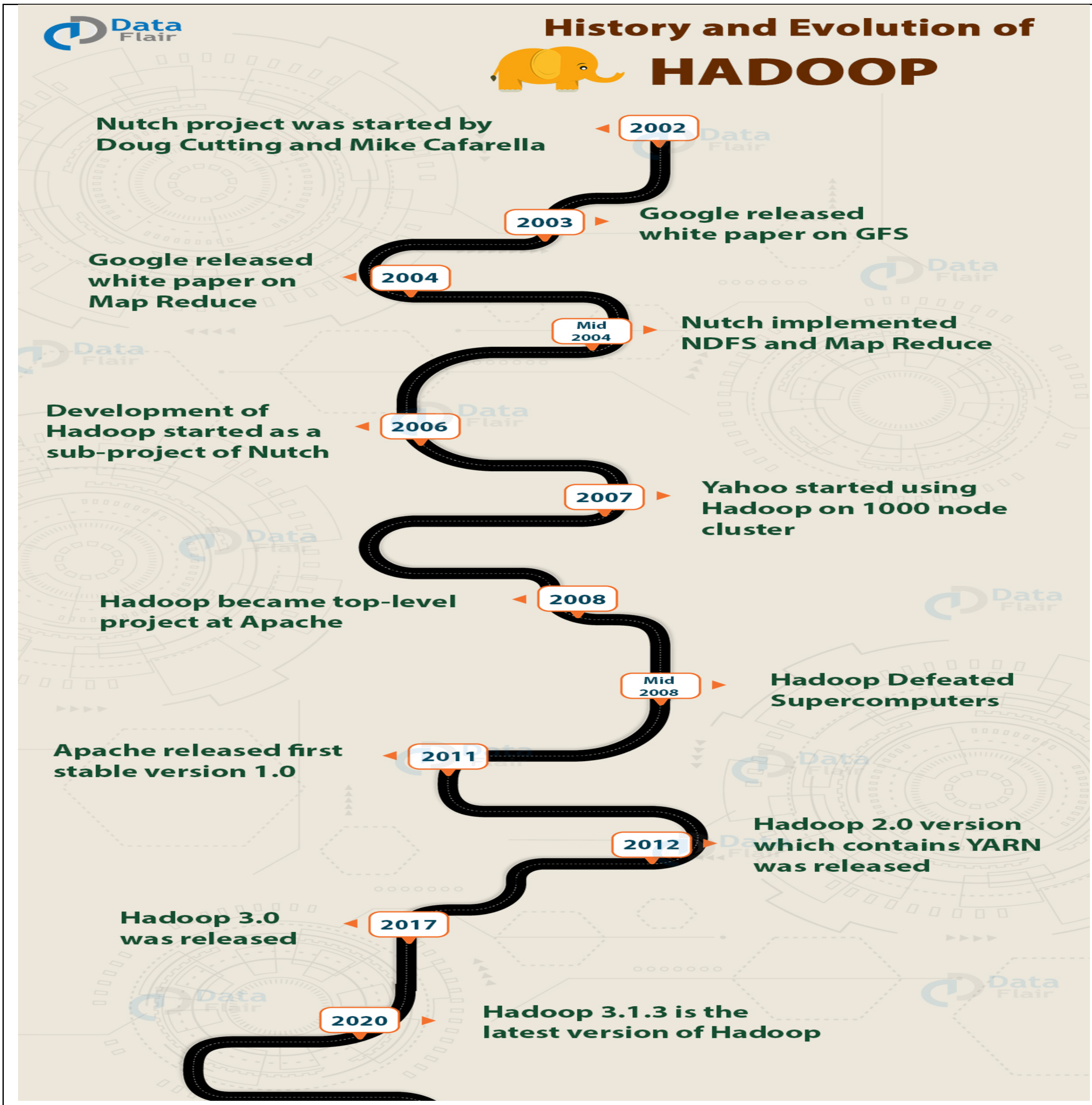
Introduction : (Maximum 5 sentences)
Hadoop is an **open-source software** framework for storing and processing large datasets ranging in size from **gigabytes** to **petabytes**. Hadoop was developed at the Apache Software Foundation.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Hadoop is an **open-source software** framework for storing and processing large datasets ranging in size from **gigabytes** to **petabytes**. Hadoop was developed at the Apache Software Foundation.



History and Evolution of HADOOP

2002 – Nutch project was started by Doug Cutting and Mike Cafarella

2003 – Google released white paper on GFS

2004 – Google released white paper on Map Reduce

Mid 2004 – Nutch implemented NDFS and Map Reduce

2006 – Development of Hadoop started as a sub-project of Nutch

2007 – Yahoo started using Hadoop on 1000 node cluster

2008 – Hadoop became top-level project at Apache

Mid 2008 – Hadoop Defeated Supercomputers

2011 – Apache released first stable version 1.0

2012 – Hadoop 2.0 version which contains YARN was released

2017 – Hadoop 3.0 was released

2020 – Hadoop 3.1.3 is the latest version of Hadoop

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Publishing, 2012 (**Page No:54**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L20

MCA

I / II

Course Name with Code : 19CAB13 – Big Data Analytics

Course Faculty : Dr.M.Moorthy

Unit: III - HADOOP ENVIRONMENT Date of Lecture: 17.03.2021

Topic of Lecture: The Hadoop Distributed File System

Introduction : (Maximum 5 sentences)

HDFS

- HDFS stands for **Hadoop Distributed File System**.
- It provides for data storage of Hadoop.
- HDFS splits the data unit into smaller units called blocks and stores them in a distributed manner.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

HDFS

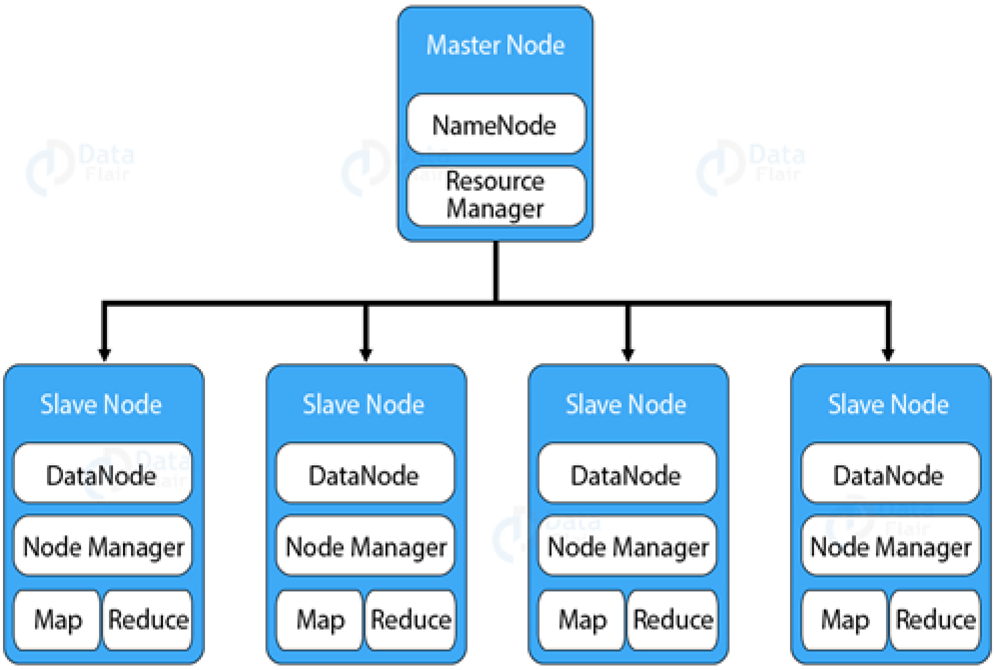
- HDFS stands for **Hadoop Distributed File System**.
- It provides for data storage of Hadoop.
- HDFS splits the data unit into smaller units called blocks and stores them in a distributed manner.
- It has got two daemons running.
- One for master node – NameNode
- And other for slave nodes – DataNode.

NameNode and DataNode

- HDFS has a **Master-slave architecture**.
- The daemon called NameNode runs on the master server. It is responsible for Namespace management and regulates file access by the client.
- DataNode daemon runs on slave nodes. It is responsible for storing actual business data.
- Internally, a file gets split into a number of data blocks and stored on a group of slave

machines.

- Namenode manages modifications to file system namespace. These are actions like the opening, closing and renaming files or directories.
- NameNode also keeps track of mapping of blocks to DataNodes. This DataNodes serves read/write request from the file system's client.
- DataNode also creates, deletes and replicates blocks on demand from NameNode.



- **Java is the native language** of HDFS.
- Hence one can deploy DataNode and NameNode on machines having Java installed.
- In a typical deployment, there is one dedicated machine running NameNode.
- And all the other nodes in the cluster run DataNode.
- The NameNode contains metadata like the location of blocks on the DataNodes.
- And arbitrates resources among various competing DataNodes.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data", McGrawHill Publishing, 2012 (**Page No:56**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L21

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 18.03.2021

Topic of Lecture: Components of Hadoop

Introduction : (Maximum 5 sentences)

What is Hadoop Architecture?

Hadoop has a master-slave topology. In this topology, we have *one master node and multiple slave nodes*. Master node's function is to assign a task to various slave nodes and manage resources. The slave nodes do the actual computing. Slave nodes store the real data whereas on master we have metadata. This means it stores data about data.

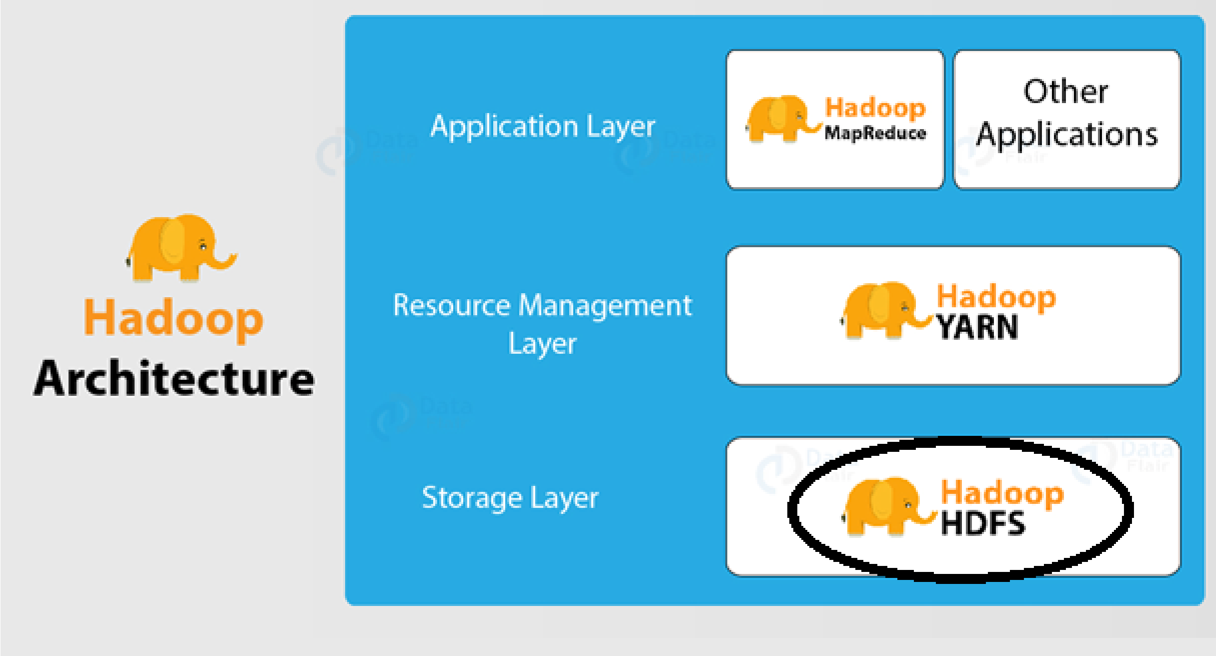
Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

What is Hadoop Architecture?

Hadoop has a master-slave topology. In this topology, we have *one master node and multiple slave nodes*. Master node's function is to assign a task to various slave nodes and manage resources. The slave nodes do the actual computing. Slave nodes store the real data whereas on master we have metadata. This means it stores data about data.



MapReduce

MapReduce is the data processing component of Hadoop. It applies the computation on sets of data in parallel thereby improving the performance. MapReduce works in two phases –

Map Phase – This phase takes input as key-value pairs and produces output as key-value pairs. It can write custom business logic in this phase. Map phase processes the data and gives it to the next phase.

Reduce Phase – The MapReduce framework sorts the key-value pair before giving the data to this phase. This phase applies the summary type of calculations to the key-value pairs.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data", McGrawHill Publishing, 2012 (**Page No:55**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L22

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 19.03.2021

Topic of Lecture: Analyzing the Data with Hadoop

Introduction : (Maximum 5 sentences)

Hadoop Data Analysis Technologies

The existing open source Hadoop data analysis technologies to analyze the huge stock data being generated very frequently.

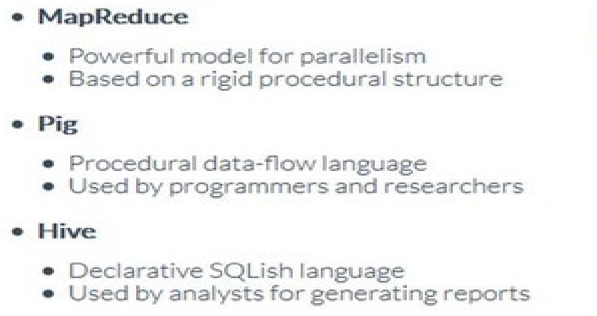
Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

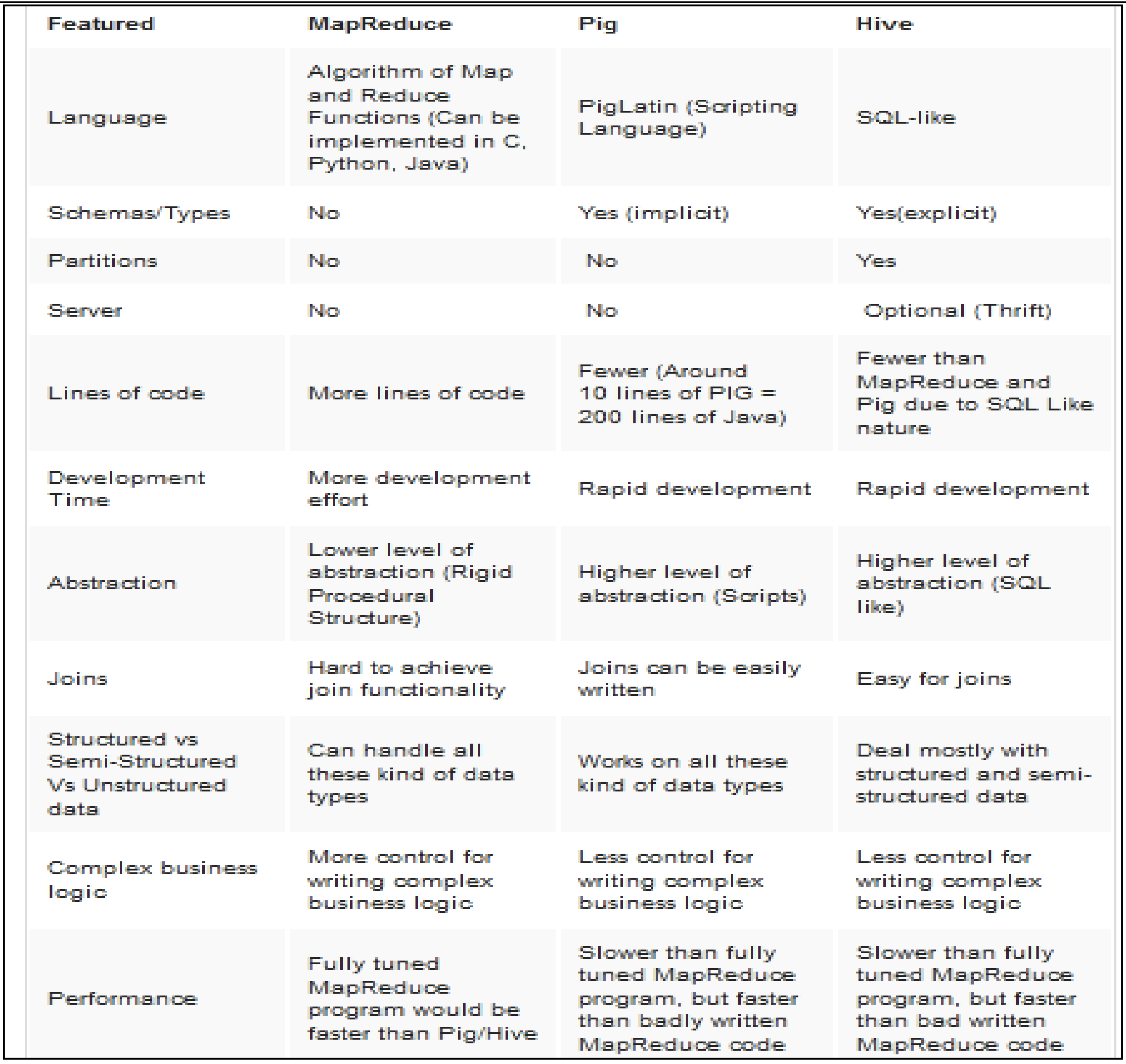
Detailed content of the Lecture:

Hadoop Data Analysis Technologies

The existing open source Hadoop data analysis technologies to analyze the huge stock data being generated very frequently.



- MapReduce
  - Powerful model for parallelism
  - Based on a rigid procedural structure
- Pig
  - Procedural data-flow language
  - Used by programmers and researchers
- Hive
  - Declarative SQLish language
  - Used by analysts for generating reports



Featured	MapReduce	Pig	Hive
Language	Algorithm of Map and Reduce Functions (Can be implemented in C, Python, Java)	PigLatin (Scripting Language)	SQL-like
Schemas/Types	No	Yes (implicit)	Yes(explicit)
Partitions	No	No	Yes
Server	No	No	Optional (Thrift)
Lines of code	More lines of code	Fewer (Around 10 lines of PIG = 200 lines of Java)	Fewer than MapReduce and Pig due to SQL Like nature
Development Time	More development effort	Rapid development	Rapid development
Abstraction	Lower level of abstraction (Rigid Procedural Structure)	Higher level of abstraction (Scripts)	Higher level of abstraction (SQL like)
Joins	Hard to achieve join functionality	Joins can be easily written	Easy for joins
Structured vs Semi-Structured Vs Unstructured data	Can handle all these kind of data types	Works on all these kind of data types	Deal mostly with structured and semi-structured data
Complex business logic	More control for writing complex business logic	Less control for writing complex business logic	Less control for writing complex business logic
Performance	Fully tuned MapReduce program would be faster than Pig/Hive	Slower than fully tuned MapReduce program, but faster than badly written MapReduce code	Slower than fully tuned MapReduce program, but faster than bad written MapReduce code

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Tom White " Hadoop: The Definitive Guide" Fourth Edition, O‟reilly Media, 2015 (**Page No:18**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L23

MCA **I / II**

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 20.03.2021

Topic of Lecture: Hadoop filesystems

Introduction : (Maximum 5 sentences)

- HDFS (Hadoop Distributed File System) is a unique design that provides storage for *extremely large files* with streaming data access pattern and it runs on *commodity hardware*.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Hadoop Distributed File System (HDFS)

- With growing data velocity the data size easily outgrows the storage limit of a machine. A solution would be to store the data across a network of machines. Such filesystems are called *distributed filesystems*.
- Since data is stored across a network all the complications of a network come in. This is where Hadoop comes in.
- It provides one of the most reliable filesystems. HDFS (Hadoop Distributed File System) is a unique design that provides storage for *extremely large files* with streaming data access pattern and it runs on *commodity hardware*.

  Let’s elaborate the terms:

- ***Extremely large files***: Here we are talking about the data in range of petabytes(1000 TB).
- ***Streaming Data Access Pattern***: HDFS is designed on principle of *write-once and read-many-times*. Once data is written large portions of dataset can be processed any number times.
- ***Commodity hardware:*** Hardware that is inexpensive and easily available in the market. This is one of feature which specially distinguishes HDFS from other file system.

Nodes: Master-slave nodes typically form the HDFS cluster.

1. **MasterNode:**
    - Manages all the slave nodes and assign work to them.
    - It executes filesystem namespace operations like opening, closing, renaming files and directories.
    - It should be deployed on reliable hardware which has the high config. not on commodity hardware.
2. **NameNode:**
    - Actual worker nodes, who do the actual work like reading, writing, processing etc.
    - They also perform creation, deletion, and replication upon instruction from the master.
    - They can be deployed on commodity hardware.

HDFS deamons: Deamons are the processes running in background.

Namenodes:

- Run on the master node.
- Store metadata (data about data) like file path, the number of blocks, block Ids. etc.
- Require high amount of RAM.
- Store meta-data in RAM for fast retrieval i.e to reduce seek time. Though a persistent copy of it is kept on disk.

DataNodes:

- Run on slave nodes.
- Require high memory as data is actually stored here.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

IQAC

LECTURE HANDOUTS

L24

MCA | **I / II**

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 22.03.2021

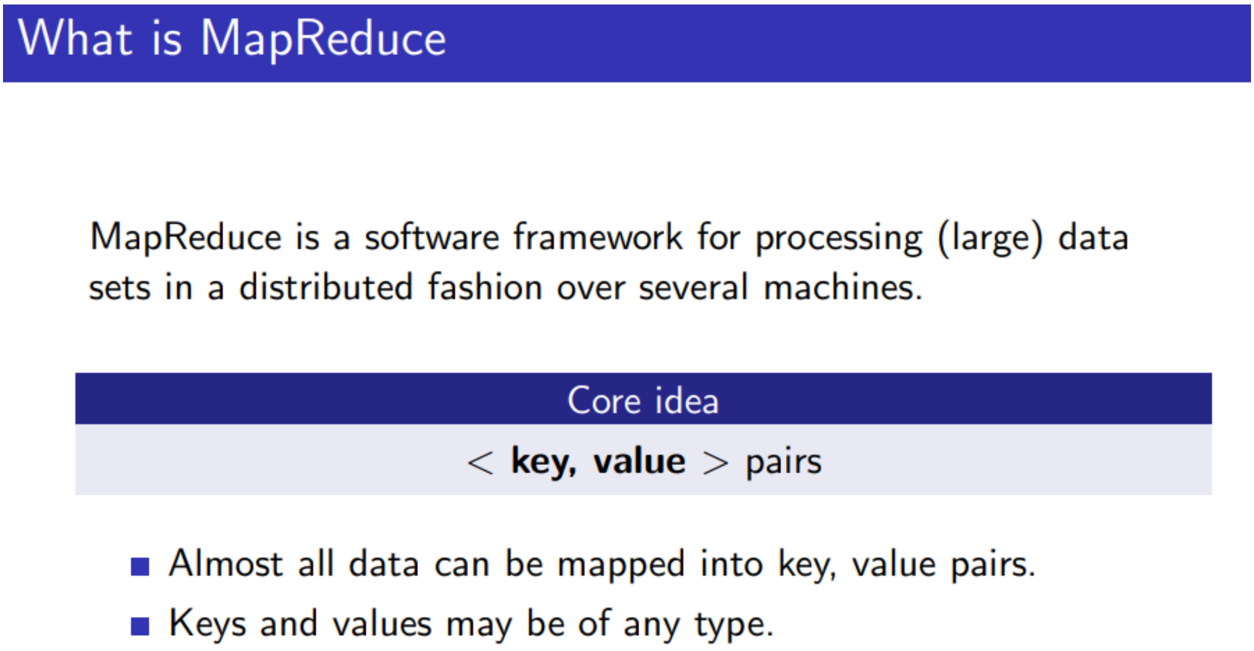
Topic of Lecture: Developing a Map Reduce Application

Introduction : (Maximum 5 sentences)
Map Reduce is software framework for processing large datasets. It uses two phases viz., Mapping phase and Reduce phase.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:



What is MapReduce

MapReduce is a software framework for processing (large) data sets in a distributed fashion over several machines.

Core idea

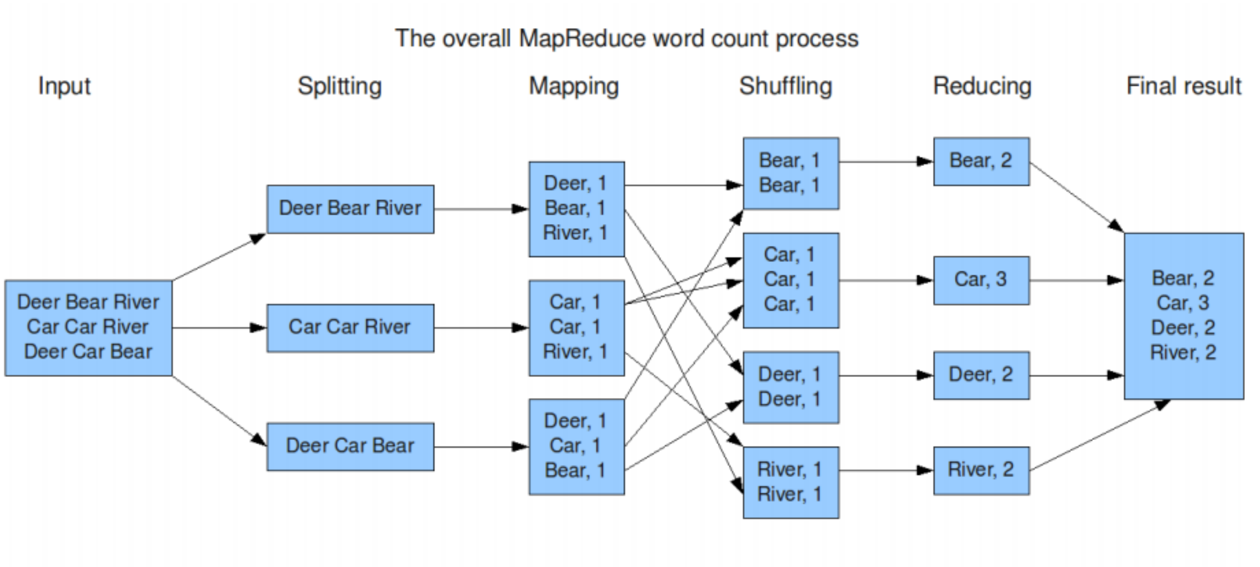
< **key, value** > pairs

- Almost all data can be mapped into key, value pairs.
- Keys and values may be of any type.

Task: Counting the word occurances (frequencies) in a text file (or set of files).
< word, count > as < key, value > pair

Mapper: Emits < word, 1 > for each word (no counting at this part).

Shuffle in between: pairs with same keys grouped together and passed to a single machine.
Reducer: Sums up the values (1s) with the same key value.



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Publishing, 2012 (**Page No:64**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L25

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 23.03.2021

Topic of Lecture: How Map Reduce Works

Introduction : (Maximum 5 sentences)
Map Reduce is software framework for processing large datasets. It uses two phases viz., Mapping phase and Reduce phase. It works with key-value pair.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Phases of the MapReduce model

MapReduce model has three major and one optional phase:

1. Mapper

- It is the first phase of MapReduce programming and contains the coding logic of the mapper function.
- The conditional logic is applied to 'n' number of data blocks spread across various data nodes.
- Mapper function accepts key-value pairs as input as (k, v), where the key represents the offset address of each record and value represents the entire record content.
- The output of the Mapper phase will also be in the key-value format as (k', v').

2. Shuffle and Sort

- The output of various mappers (k', v'), then goes into Shuffle and Sort phase.
- All the duplicate values are removed, and different values are grouped together based on similar keys.
- The output of the Shuffle and Sort phase will be key-value pairs again as key and array of values (k, v[]).

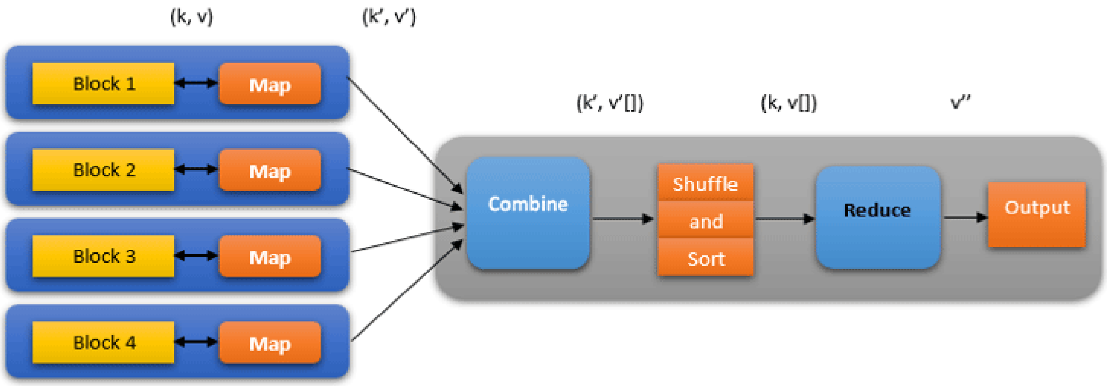
3. Reducer

- The output of the Shuffle and Sort phase (k, v[]) will be the input of the Reducer phase.
- In this phase reducer function's logic is executed and all the values are aggregated against their corresponding keys.
- Reducer consolidates outputs of various mappers and computes the final job output.
- The final output is then written into a single file in an output directory of HDFS.

4. Combiner

- It is an optional phase in the MapReduce model.
- The combiner phase is used to optimize the performance of MapReduce jobs.

- In this phase, various outputs of the mappers are locally reduced at the node level.
- For example, if different mapper outputs (k, v) coming from a single node contains duplicates, then they get combined i.e. locally reduced as a single (k, v[]) output.
- This phase makes the Shuffle and Sort phase work even quicker thereby enabling additional performance in MapReduce jobs.



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Tom White " Hadoop: The Definitive Guide" Fourth Edition, O"reilly Media, 2015 (**Page No:167**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L26

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 24.03.2021

Topic of Lecture: Setting up a Hadoop Cluster

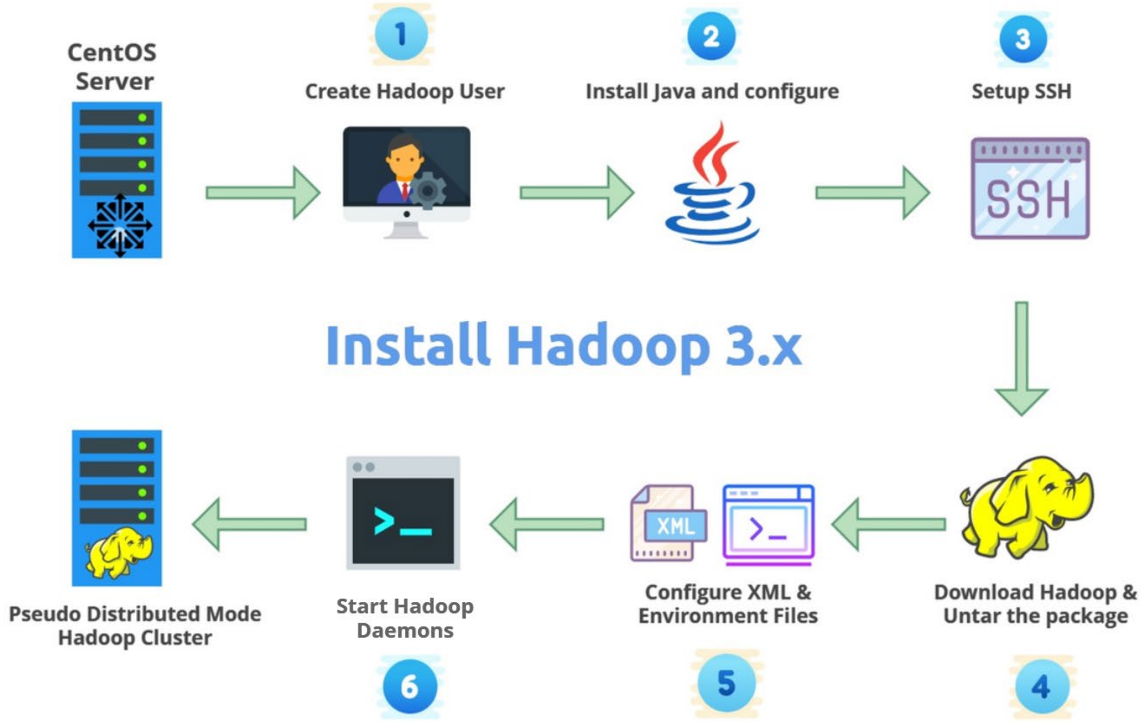
Introduction : (Maximum 5 sentences)

Hadoop uses distributed systems to store and process the data. Hence there is a necessity to create cluster.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:


CentOS Server
1
Create Hadoop User
2
Install Java and configure
3
Setup SSH
Install Hadoop 3.x
4
Download Hadoop & Untar the package
5
Configure XML & Environment Files
XML
6
Start Hadoop Daemons
Pseudo Distributed Mode Hadoop Cluster

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L27

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: III - HADOOP ENVIRONMENT **Date of Lecture:** 25.03.2021

Topic of Lecture: Hadoop Configuration

Introduction : (Maximum 5 sentences)

Hadoop uses xml files to store configuration. They are

core-site.xml
mapred-site.xml
hdfs-site.xml
yarn-site.xml

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Java Programming**
- **Knowledge on Algorithms**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

core-site.xml

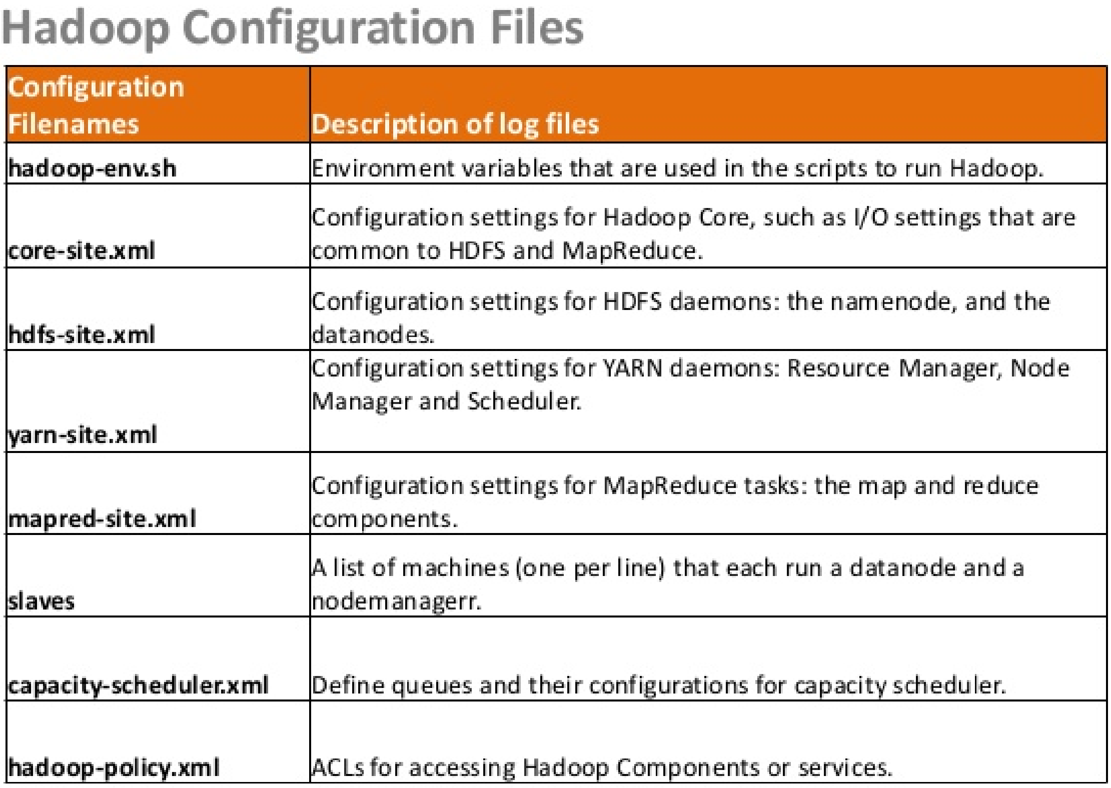
```
  <property>
      <name>fs.defaultFS</name>
      <value>hdfs://localhost:9000</value>
   </property>
```

mapred-site.xml

```
<property>
     <name>mapreduce.framework.name</name>
     <value>yarn</value>
  </property>
```

yarn-site.xml

```
<property>
          <name>yarn.nodemanager.aux-services</name>
          <value>mapreduce_shuffle</value>
  </property>
  <property>
          <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
          <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
```



Hadoop Configuration Files

Configuration Filenames	Description of log files
hadoop-env.sh	Environment variables that are used in the scripts to run Hadoop.
core-site.xml	Configuration settings for Hadoop Core, such as I/O settings that are common to HDFS and MapReduce.
hdfs-site.xml	Configuration settings for HDFS daemons: the namenode, and the datanodes.
yarn-site.xml	Configuration settings for YARN daemons: Resource Manager, Node Manager and Scheduler.
mapred-site.xml	Configuration settings for MapReduce tasks: the map and reduce components.
slaves	A list of machines (one per line) that each run a datanode and a nodemanagerr.
capacity-scheduler.xml	Define queues and their configurations for capacity scheduler.
hadoop-policy.xml	ACLs for accessing Hadoop Components or services.

hdfs-site.xml

```
<property>
     <name>dfs.replication</name>
     <value>1</value>
  </property>
  <property>
     <name>dfs.namenode.name.dir</name>
     <value>C:\data\namenode</value>
  </property>
  <property>
     <name>dfs.datanode.data.dir</name>
     <value>C:\data\datanode</value>
  </property>
```

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

http://hadoop.apache.org/docs/r2.6.4/api/org/apache/hadoop/conf/Configuration.html?is-external=true

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L28

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 26.03.2021

Topic of Lecture: Link Analysis

Introduction : (Maximum 5 sentences)

Link analysis is a process of finding connections between different entities, such as connecting customers to other customers or customer to products.

These relationships can be between various types of objects (nodes), including people, organizations and even transactions.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Link Analysis

- Link analysis is a process of finding connections between different entities, such as connecting customers to other customers or customer to products.
- These relationships can be between various types of objects (nodes), including people, organizations and even transactions.
- Link analysis is essentially a kind of knowledge discovery that can be used to visualize data to allow for better analysis, especially in the context of links, whether Web links or relationship links between people or between different entities.
- Link analysis is often used in search engine optimization as well as in intelligence, in security analysis and in market and medical research.

Link analysis has three primary purposes:

- Find matches for known patterns of interests between linked objects.
- Find anomalies by detecting violated known patterns.
- Find new patterns of interest (for example, in social networking and marketing and business intelligence).

- Some examples include analyzing telephone call detail records to examine links established when a connection is initiated at one telephone number to a different telephone number, determining whether two individuals are connected via a social network, or the degree to which similar travelers select travel on specific flights.
- For telephone connectivity, some examples include the frequency of the calls, the duration of the calls, or the times at which those calls are made.
- Link analysis is useful for analytical applications that rely on graph theory for drawing conclusions. One example is looking for closely connected groups of people.
- Another analytical area for which link analysis is useful is process optimization. An example might be evaluating the allocation of airplanes and pilots (who are trained for flying specific kinds of airplanes) to the many routes that an airline travels.
- A third use is in assessing viral influence of individuals within a social networking environment. One participant might not necessarily account for a significant number of product purchases directly, but her recommendation may be followed by many individuals within her "sphere of influence."
- Link analysis is critical for mapping and then understanding spheres of influence as looking at various kinds of relationships and questions such as:

    - Closely connected groups of customers.
    - Collections of individuals linked by certain attributes such as location, purchased products, or other demographic variables.
    - The speed at which communication flows across a social network.
    - Which customers are known as having a particular expertise in using products.
    - Which customers exercise significant influence over the broadest collection of individuals.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CambridgeUniversity Press, 2014 (**Page No:163**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L29

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 27.03.2021

Topic of Lecture: Efficient Computation of PageRank

Introduction : (Maximum 5 sentences)

An algorithm used by Google Search to rank web pages in their search engine results.

Named after Larry Page, one of the founders of Google.

A way of measuring the importance of website pages.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

PageRank (PR)

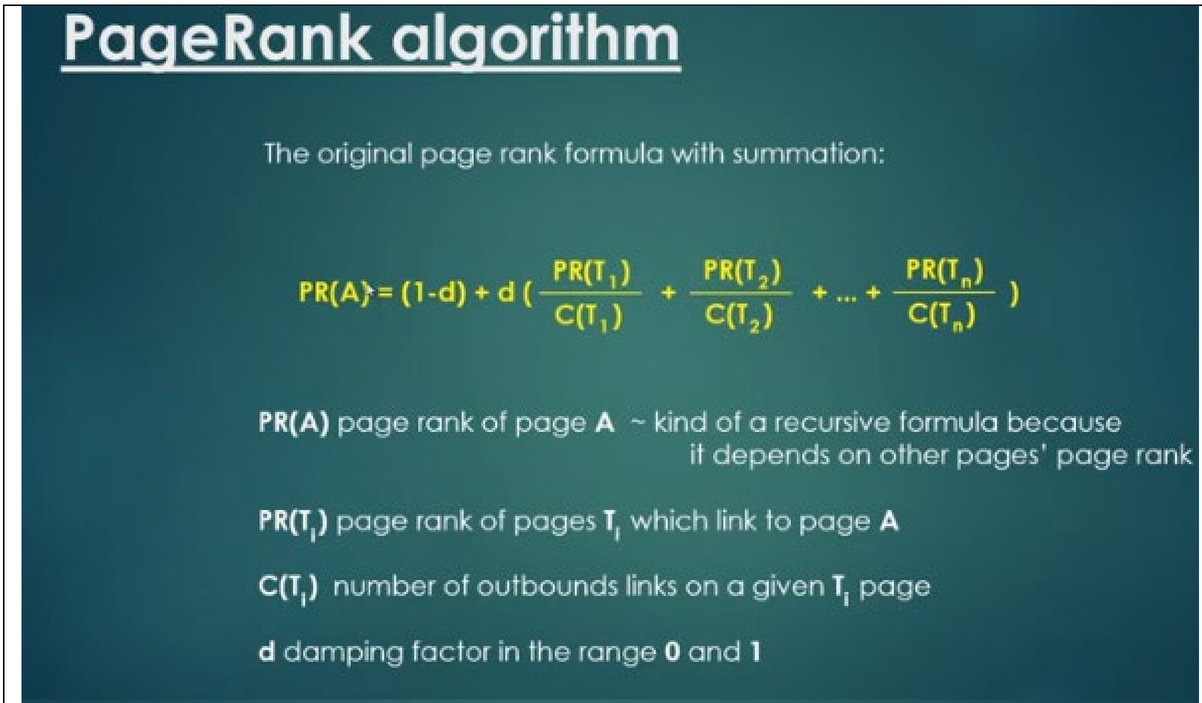
An algorithm used by Google Search to rank web pages in their search engine results.

Named after Larry Page, one of the founders of Google.

A way of measuring the importance of website pages.

According to Google:

- PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.
- The underlying assumption is that more important websites are likely to receive more links from other websites.



PageRank algorithm

The original page rank formula with summation:

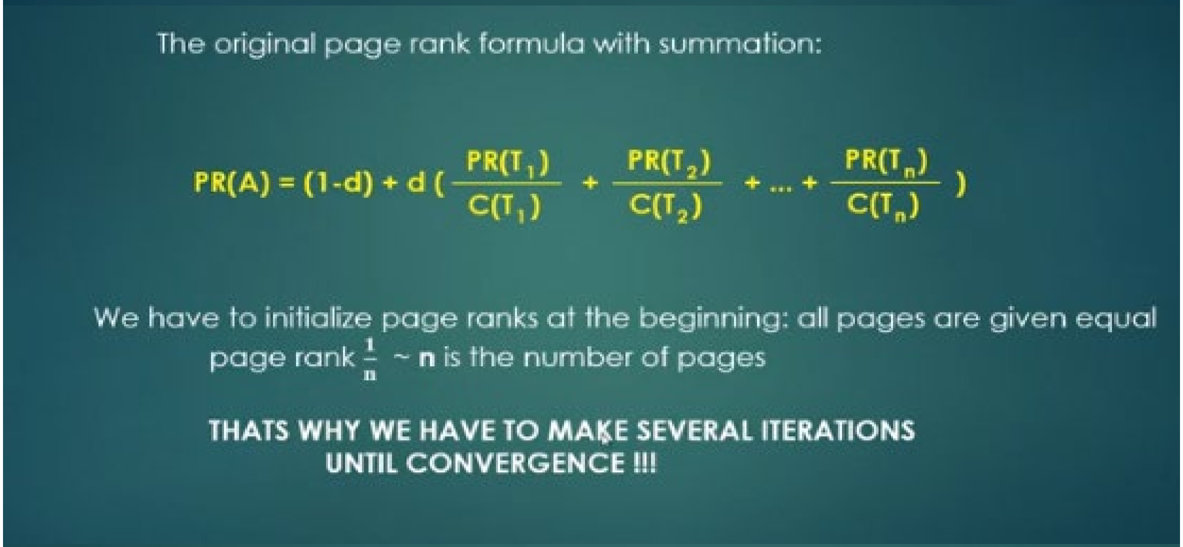
$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right)$$

PR(A) page rank of page **A** ~ kind of a recursive formula because it depends on other pages' page rank

PR(T$_i$) page rank of pages **T$_i$** which link to page **A**

C(T$_i$) number of outbounds links on a given **T$_i$** page

d damping factor in the range **0** and **1**

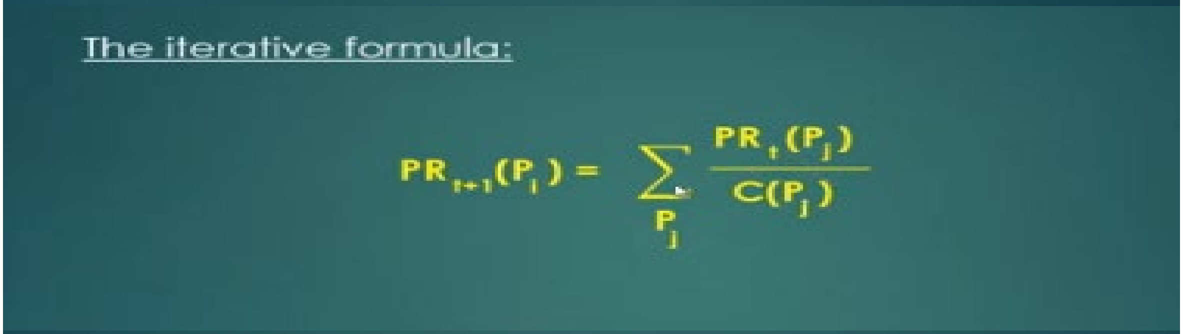


The original page rank formula with summation:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right)$$

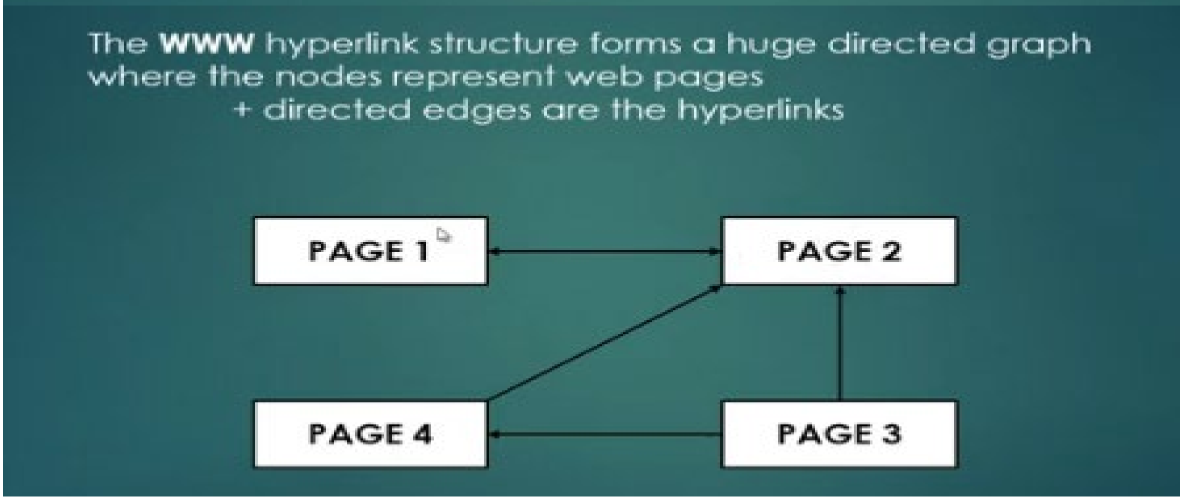
We have to initialize page ranks at the beginning: all pages are given equal page rank $\frac{1}{n}$ ~ **n** is the number of pages

THATS WHY WE HAVE TO MAKE SEVERAL ITERATIONS UNTIL CONVERGENCE !!!



The iterative formula:

$$PR_{t+1}(P_i) = \sum_{P_j} \frac{PR_t(P_j)}{C(P_j)}$$



The **WWW** hyperlink structure forms a huge directed graph where the nodes represent web pages + directed edges are the hyperlinks

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:177**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L30

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 29.03.2021

Topic of Lecture: Topic-Sensitive PageRank

Introduction : (Maximum 5 sentences)
A context-**sensitive** ranking algorithm for Web search. For searches done in context (e.g., when the search query is performed by highlighting words in a Web page), we compute the topic-sensitive PageRank scores using the topic of the context in which the query appeared.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search. For searches done in context (e.g., when the search query is performed by highlighting words in a Web page), we compute the topic-sensitive PageRank scores using the topic of the context in which the query appeared.

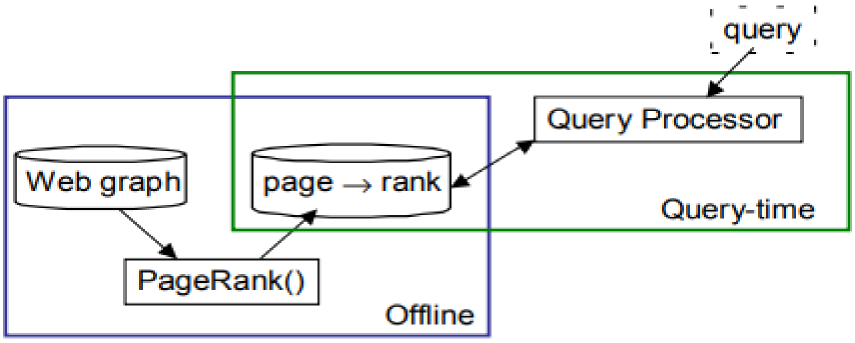
Motivation

- Improve search results
- Current engines work well for us “computer types”, but not for novice users
- Exploit search context in a tractable and effective way

Search Context

Query context
    Highlighted word on page
    Previous queries issued
User context
    Bookmarks
    Browsing history

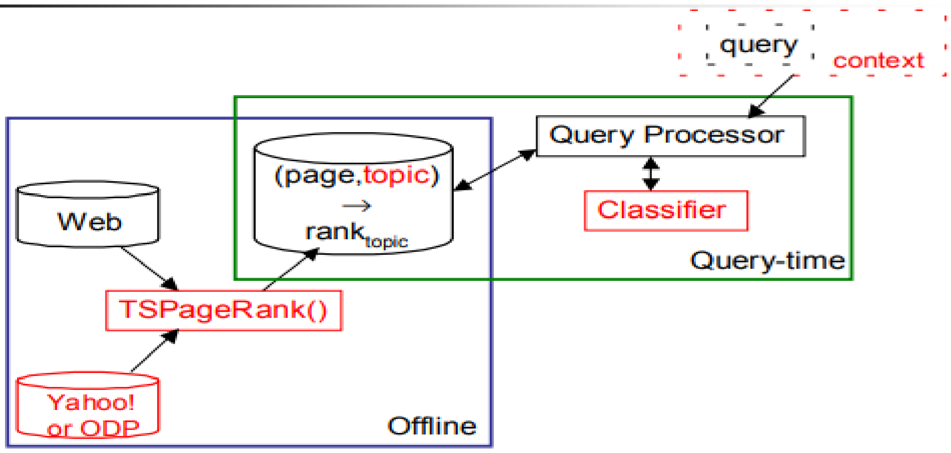
Original PageRank



Topic-Sensitive PageRank

- Assigns *multiple* a-priori “importance” estimates to pages
- One PageRank score per *basis topic*
  - + Query specific rank score
  - + Make use of context
  - + Inexpensive at runtime

Topic-Sensitive PageRank



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:183**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L31

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 30.03.2021

Topic of Lecture: Link Spam

Introduction : (Maximum 5 sentences)
Link spam is the posting of out-of-context links on websites, discussion forums, blog comments, guest books or any other online venue that displays user comments. Link spam is also known as comment spam, blog spam or wikispam.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

- Link spam is the posting of out-of-context links on websites, discussion forums, blog comments, guest books or any other online venue that displays user comments.
- Link spam is also known as comment spam, blog spam or wiki spam.
- Link spammers usually don't leave comments of any value along with their links.
- Link spam is defined as links between pages that are present for reasons other than merit.
- Link spam takes advantage of link-based ranking algorithms, which gives websites higher rankings the more other highly ranked websites link to it.

How do you identify spam links?

- The first thing you should do is to look at the anchor text of the link.
- If it sounds suspicious, incoherent or doesn't coincide with your niche, it can be a strong signal of a spam link. 5.
- You can also click on a suspicious link to check if the content on that page is low quality.

Video Content / Details of website for further learning (if any): **https://www.tutorialspoint.com/big_data_tutorials.htm**
Important Books/Journals for further learning including the page nos.: AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:187**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L32

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 31.03.2021

Topic of Lecture: A Model for Recommendation Systems

Introduction : (Maximum 5 sentences)

Within recommendation systems, there is a group of models called collaborative-filtering, which tries to find similarities between users or between items based on recorded user-item preferences or ratings. A subgroup of collaborative systems called memory-based models. They are called memory-based because the algorithm is not complicated, but requires a lot of memory to keep track of the results.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Model-Based Recommendation Systems

- Within recommendation systems, there is a group of models called collaborative-filtering, which tries to find similarities between users or between items based on recorded user-item preferences or ratings. A subgroup of collaborative systems called memory-based models. They are called memory-based because the algorithm is not complicated, but requires a lot of memory to keep track of the results.
- Another subgroup of collaborative-filtering models: model-based models (which is a rather silly name). As opposed to the memory-based approaches, this uses some sort of machine learning algorithm. There are many different variations within this group, what we are going to concentrate on is the singular value decomposition methods.

How do recommender systems work?

Understanding relationships

- Relationships provide recommender systems with tremendous insight, as well as an understanding of customers. There are three main types that occur:

User-Product Relationship

- The user-product relationship occurs when some users have an affinity or preference towards specific products that they need. For example, a cricket player might have a preference for cricket-related items, thus the e-commerce website will build a user-product relation of player->cricket.

Product-Product Relationship

Product-product relationships occur when items are similar in nature, either by appearance or description. Some examples include books or music of the same genre, dishes from the same cuisine, or news articles from a particular event.

User-User Relationship

User-user relationships occur when some customers have similar taste with respect to a particular product or service. Examples include mutual friends, similar backgrounds, similar age, etc.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:307**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L33

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 01.04.2021

Topic of Lecture: Content-Based Recommendations

Introduction : (Maximum 5 sentences)

Recommender systems are active information filtering systems which **personalize the information** coming to a user based on his interests, relevance of the information etc. Recommender systems are used widely for recommending movies, articles, restaurants, places to visit, items to buy etc.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

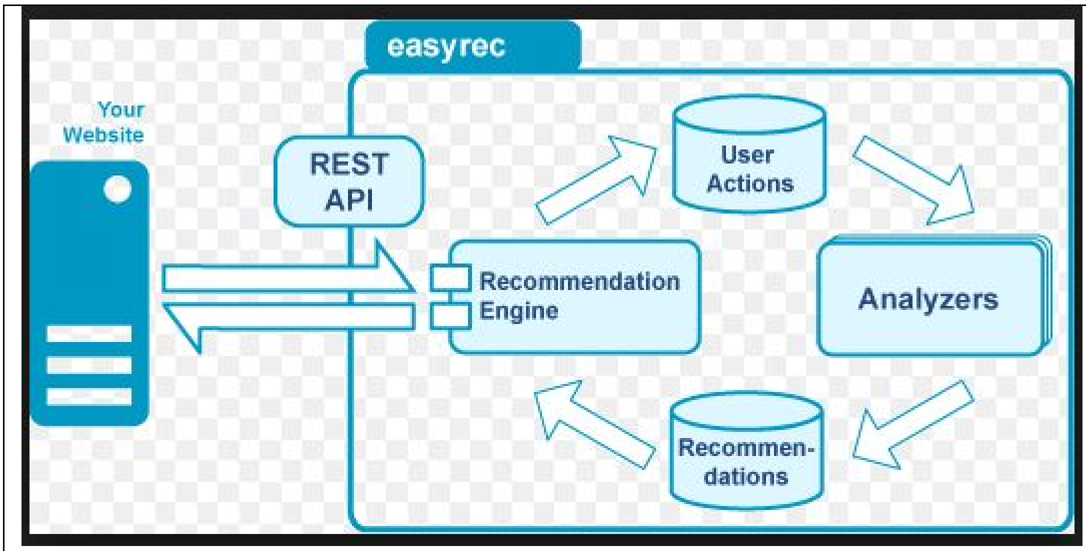
- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

- Recommender systems are active information filtering systems which **personalize the information** coming to a user based on his interests, relevance of the information etc. Recommender systems are used widely for recommending movies, articles, restaurants, places to visit, items to buy etc.

How do Content Based Recommender Systems work?

- A content based recommender works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user. As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate.



What are the concepts used in Content Based Recommenders?

- The concepts of Term Frequency (***TF***) and Inverse Document Frequency (***IDF***) are used in information retrieval systems and also content based filtering mechanisms (such as a content based recommender). They are used to determine the relative importance of a document / article / news item / movie etc.

Term Frequency (TF) and Inverse Document Frequency (IDF)

- TF is simply the frequency of a word in a document. IDF is the inverse of the document frequency among the whole corpus of documents. TF-IDF is used mainly because of two reasons: Suppose we search for ***"****the rise of analytics****"*** on Google**.** It is certain that "*the****"*** will occur more frequently than "*analytics****"*** but the relative importance of analytics is higher than the search query point of view. In such cases, TF-IDF weighting negates the effect of high frequency words in determining the importance of an item (document).

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CambridgeUniversity Press, 2014 (**Page No:312**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L34

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 03.04.2021

Topic of Lecture: Dimensionality Reduction

Introduction : (Maximum 5 sentences)
Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

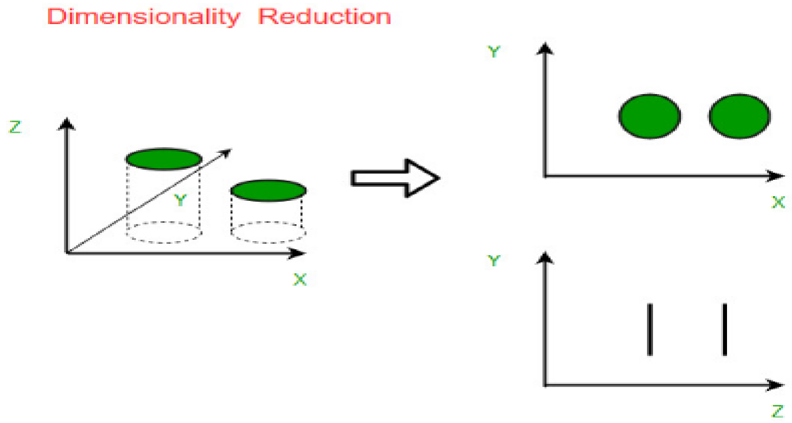
Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.

Components of Dimensionality Reduction

There are two components of dimensionality reduction:

- **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  1. Filter
  2. Wrapper
  3. Embedded
- **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

The below figure illustrates this concept, where a 3-D feature space is split into two 1-D feature spaces, and later, if found to be correlated, the number of features can be reduced even further.



Methods of Dimensionality Reduction

The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be both linear and non-linear, depending upon the method used.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CambridgeUniversity Press, 2014 (**Page No:328**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L35

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 05.04.2021

Topic of Lecture: Visual data analysis techniques

Introduction : (Maximum 5 sentences)
Data visualization is used in many areas to model complex events and visualize phenomena that cannot be observed directly, such as weather patterns, medical conditions or mathematical relationships.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Visual data analysis techniques

- Data visualization is used in many areas to model complex events and visualize phenomena that cannot be observed directly, such as weather patterns, medical conditions or mathematical relationships. Here we review basic data visualization tools and techniques.
- Data visualization is applied in practically every field of knowledge. Scientists in various disciplines use computer techniques to model complex events and visualize phenomena that cannot be observed directly, such as weather patterns, medical conditions or mathematical relationships.
- Data visualization provides an important suite of tools and techniques for gaining a qualitative understanding. The basic techniques are the following plots:

Line Plot

- The simplest technique, a line plot is used to plot the relationship or dependence of one variable on another. To plot the relationship between the two variables, we can simply call the plot function.

Bar Chart

- Bar charts are used for comparing the quantities of different categories or groups. Values of a category are represented with the help of bars and they can be configured with vertical or horizontal bars, with the length or height of each bar representing the value.

Pie and Donut Charts

- There is much debate around the value of pie and donut charts. As a rule, they are used to compare the parts of a whole and are most effective when there are limited components and when text and percentages are included to describe the content. However, they can be difficult to interpret because the human eye has a hard time estimating areas and comparing visual angles.

Histogram Plot

- A histogram, representing the distribution of a continuous variable over a given interval or period of time, is one of the most frequently used data visualization techniques in machine learning. It plots the data by chunking it into intervals called 'bins'. It is used to inspect the underlying frequency distribution, outliers, skewness, and so on.

Scatter Plot

- Another common visualization technique is a scatter plot that is a two-dimensional plot representing the joint variation of two data items. Each marker (symbols such as dots, squares and plus signs) represents an observation. The marker position indicates the value for each observation. When you assign more than two measures, a scatter plot matrix is produced that is a series of scatter plots displaying every possible pairing of the measures that are assigned to the visualization. Scatter plots are used for examining the relationship, or correlations, between X and Y variables.

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:418**)

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L36

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: IV - DATA ANALYSIS SYTEMS AND VISUALIZATION **Date of Lecture:** 06.04.2021

Topic of Lecture: Interaction techniques

Introduction : (Maximum 5 sentences)

Interaction visualization is the conversation between visualization and audience, a.k.a. users.
- Users can create changes to the visualization
- When a user enacts a cause the visualization responds with effect

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

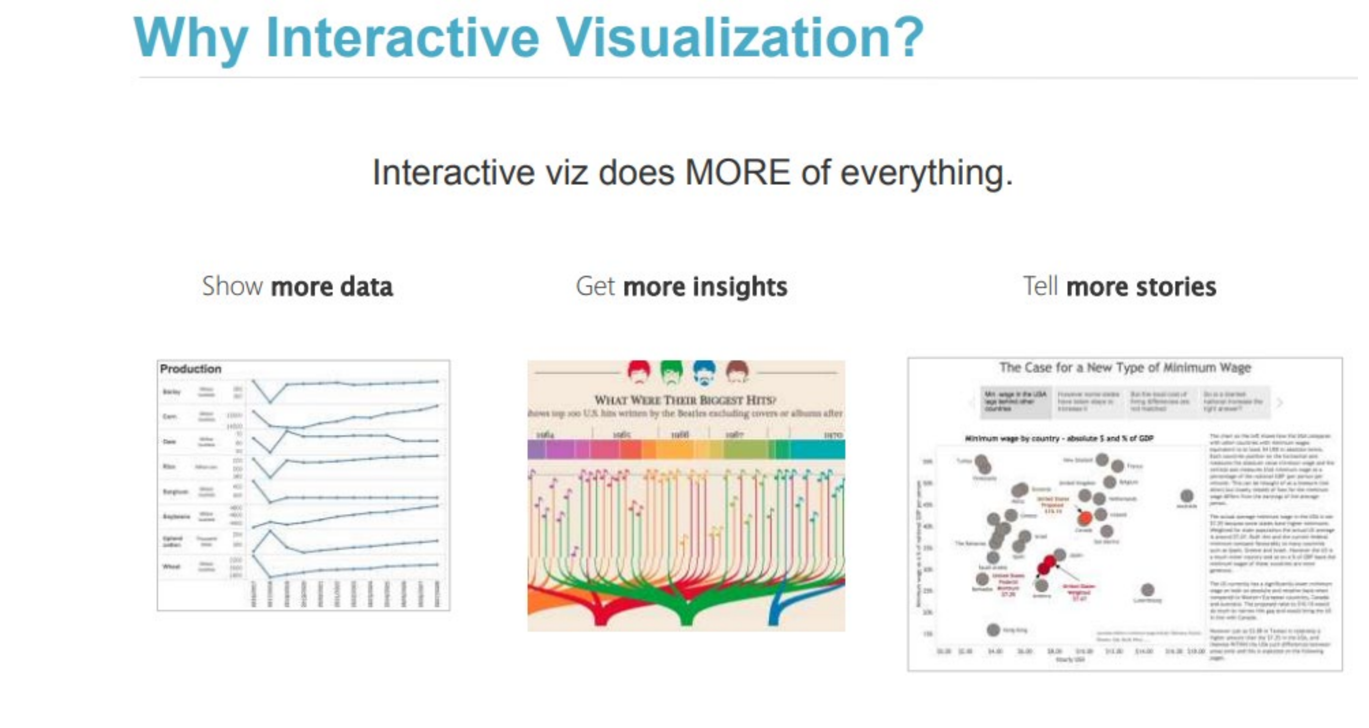
- **Knowledge on Internet**
- **Knowledge on Graph**
- **Knowledge on GUI Tools**

Detailed content of the Lecture:

Interaction visualization is the conversation between visualization and audience, a.k.a. users.
- Users can create changes to the visualization
- When a user enacts a cause the visualization responds with effect

Digital interactions:

- Zoom
- Highlight
- Tooltips
- Filters



Why Interactive Visualization?

Interactive viz does MORE of everything.

Show **more data**

Get **more insights**

Tell **more stories**

Production
Barley
Corn
Oats
Rice
Sorghum
Soybeans
Upland cotton
Wheat
What Were Their Biggest Hits?
1964
1965
1966
1967
1970
The Case for a New Type of Minimum Wage
Minimum wage by country - absolute $ and % of GDP

Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/big_data_tutorials.htm

Important Books/Journals for further learning including the page nos.:

Da Ruan,Guoquing Chen, Etienne E.Kerre, Geert Wets, Intelligent Data Mining, Springer,2007 (**Page No:418**)

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L37

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V – DATA ANALYTICS USING PYTHON **Date of Lecture:** 07.04.2021

Topic of Lecture: Pandas - Introduction

Introduction: **(Maximum 5 sentences)**

Pandas is a Python library used for working with data sets.

It has functions for analyzing, cleaning, exploring, and manipulating data.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

What is Pandas?

Pandas is a Python library used for working with data sets.

It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

Relevant data is very important in data science.

Data Science: is a branch of computer science where we study how to store, use and analyze data for deriving information from it.

What Can Pandas Do?

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called *cleaning* the data.

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L38

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 08.04.2021

Topic of Lecture: Working with Series, DataFrame

Introduction : (Maximum 5 sentences)

A Pandas Series is like a column in a table.

It is a one-dimensional array holding data of any type.

A Pandas DataFrame is a 2 dimensional data structure, like a 2 dimensional array, or a table with rows and columns.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

What is a Series?

A Pandas Series is like a column in a table.

It is a one-dimensional array holding data of any type.

Example

Create a simple Pandas Series from a list:

```
import pandas as pd

a = [1, 7, 2]

myvar = pd.Series(a)

print(myvar)
```

What is a DataFrame?

A Pandas DataFrame is a 2 dimensional data structure, like a 2 dimensional array, or a table with rows and columns.

Example

Create a simple Pandas DataFrame:

```
import pandas as pd

data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}

#load data into a DataFrame object:
df = pd.DataFrame(data)

print(df)
```

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L39

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 09.04.2021

Topic of Lecture: Series basic functionality

Introduction : (Maximum 5 sentences)
There are eight series basic functionalities such as axes, dtype, empty, ndim, size, values, head() and tail.

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Series Basic Functionality

Sr.No.	Attribute or Method & Description	
1	**Axes**	Returns a list of the row axis labels
2	**Dtype**	Returns the dtype of the object.
3	**Empty**	Returns True if series is empty.
4	**Ndim**	Returns the number of dimensions of the underlying data, by definition 1.
5	**Size**	Returns the number of elements in the underlying data.
6	**Values**	Returns the Series as ndarray.
7	**head()**	Returns the first n rows.
8	**tail()**	Returns the last n rows.

Example

```
import pandas as pd
import numpy as np

#Create a series with 100 random numbers
s = pd.Series(np.random.randn(4))
print s
```

Its **output** is as follows −

```
0   0.967853
1  -0.148368
2  -1.395906
3  -1.758394
dtype: float64
```

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L40

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 10.04.2021

Topic of Lecture: Descriptive Statistics

Introduction : (Maximum 5 sentences)
These are functions under Descriptive Statistics in Python Pandas.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Functions & Description

Let us now understand the functions under Descriptive Statistics in Python Pandas. The following table list down the important functions −

Sr.No.	Function	Description
1	count()	Number of non-null observations
2	sum()	Sum of values
3	mean()	Mean of Values
4	median()	Median of Values
5	mode()	Mode of values
6	std()	Standard Deviation of the Values
7	min()	Minimum Value
8	max()	Maximum Value
9	abs()	Absolute Value
10	prod()	Product of Values

The **describe()** function computes a summary of statistics pertaining to the DataFrame columns.

```
import pandas as pd
import numpy as np

#Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack',
   'Lee','David','Gasper','Betina','Andres']),
   'Age':pd.Series([25,26,25,23,30,29,23,34,40,30,51,46]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8,3.78,2.98,4.80,4.10,3.65])}

#Create a DataFrame
df = pd.DataFrame(d)
print df.describe()
```

Its **output** is as follows −

```
            Age         Rating
count    12.000000      12.000000
mean     31.833333       3.743333
std      9.232682       0.661628
min      23.000000       2.560000
25%      25.000000       3.230000
50%      29.500000       3.790000
75%      35.500000       4.132500
```

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L41

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 12.04.2021

Topic of Lecture: Working with Columns

Introduction : (Maximum 5 sentences)
It refers the various actions that can take place on columns like addition, deletion, etc…

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

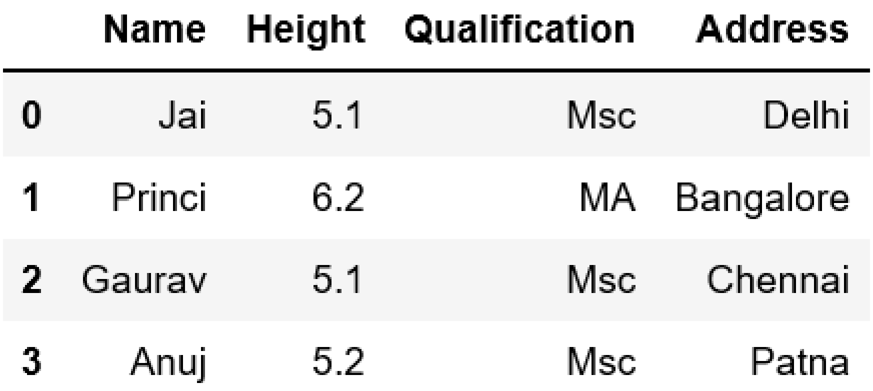
- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Adding new column to existing DataFrame in Pandas

```
# Import pandas package
import pandas as pd

# Define a dictionary containing Students data
data = {'Name': ['Jai', 'Princi', 'Gaurav', 'Anuj'],
        'Height': [5.1, 6.2, 5.1, 5.2],
        'Qualification': ['Msc', 'MA', 'Msc', 'Msc']}

# Convert the dictionary into DataFrame
df = pd.DataFrame(data)

# Declare a list that is to be converted into a column
address = ['Delhi', 'Bangalore', 'Chennai', 'Patna']

# Using 'Address' as the column name
# and equating it to the list
df['Address'] = address

# Observe the result
df
```

Output:



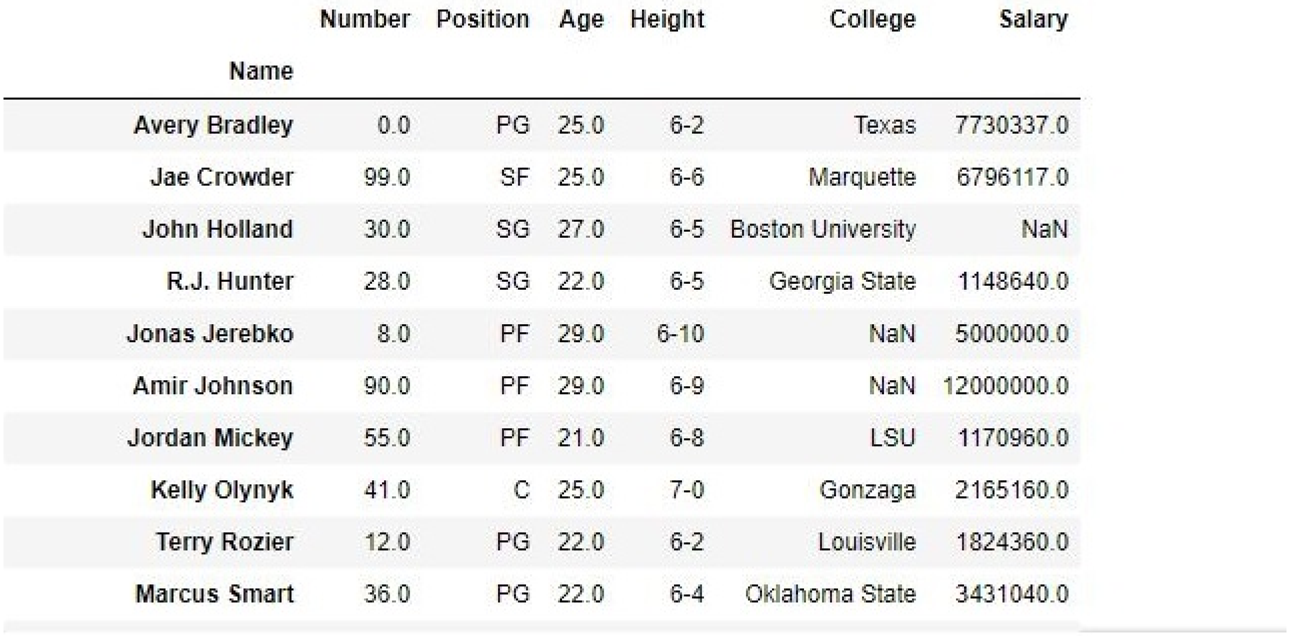
	Name	Height	Qualification	Address
0	Jai	5.1	Msc	Delhi
1	Princi	6.2	MA	Bangalore
2	Gaurav	5.1	Msc	Chennai
3	Anuj	5.2	Msc	Patna

Dropping columns with column name

In his code, Passed columns are dropped using column names. axis parameter is kept 1 since 1 refers to columns.

```
# importing pandas module
import pandas as pd

# making data frame from csv file
data = pd.read_csv("nba.csv", index_col ="Name" )

# dropping passed columns
data.drop(["Team", "Weight"], axis = 1, inplace = True)

# display
data
```

Data Frame after Dropping Columns-



	Number	Position	Age	Height	College	Salary
Name						
Avery Bradley	0.0	PG	25.0	6-2	Texas	7730337.0
Jae Crowder	99.0	SF	25.0	6-6	Marquette	6796117.0
John Holland	30.0	SG	27.0	6-5	Boston University	NaN
R.J. Hunter	28.0	SG	22.0	6-5	Georgia State	1148640.0
Jonas Jerebko	8.0	PF	29.0	6-10	NaN	5000000.0
Amir Johnson	90.0	PF	29.0	6-9	NaN	12000000.0
Jordan Mickey	55.0	PF	21.0	6-8	LSU	1170960.0
Kelly Olynyk	41.0	C	25.0	7-0	Gonzaga	2165160.0
Terry Rozier	12.0	PG	22.0	6-2	Louisville	1824360.0
Marcus Smart	36.0	PG	22.0	6-4	Oklahoma State	3431040.0

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L42

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 15.04.2021

Topic of Lecture: Working with rows

Introduction : (Maximum 5 sentences)
It refers the actions invoked on rows like insert, delete, etc..

Prerequisite knowledge for Complete understanding and learning of Topic:
(Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Pandas provide data analysts a way to delete and filter data frame using `.drop()` method. Rows or columns can be removed using index label or column name using this method.

Syntax:
DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')

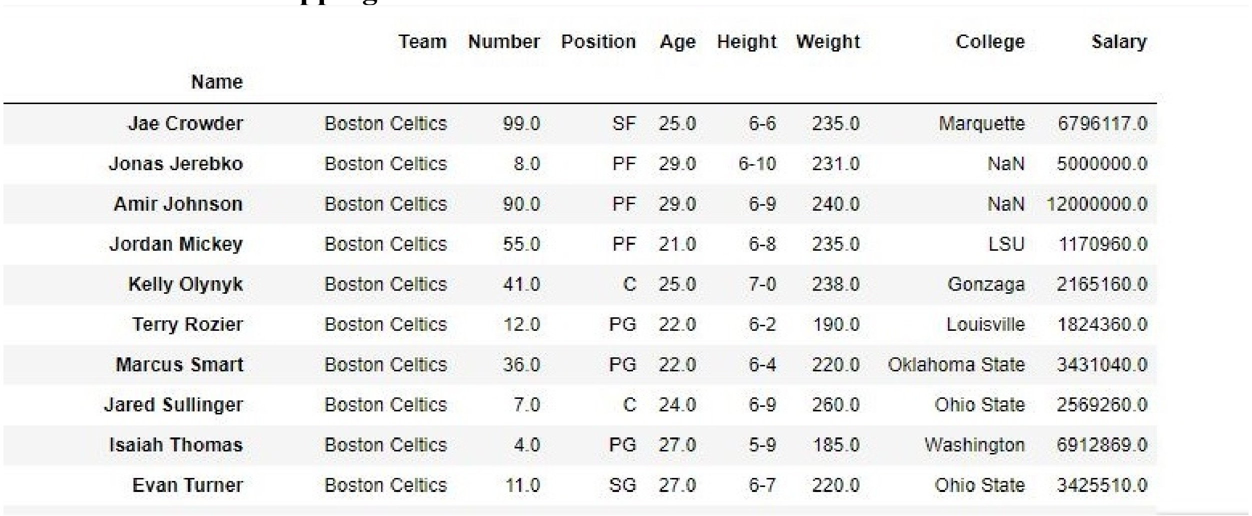
Parameters:

labels: String or list of strings referring row or column name.
axis: int or string value, 0 'index' for Rows and 1 'columns' for Columns.
index or columns: Single label or list. index or columns are an alternative to axis and cannot be used together.
level: Used to specify level in case data frame is having multiple level index.
inplace: Makes changes in original Data Frame if True.
errors: Ignores error if any value from the list doesn't exists and drops rest of the values when errors = 'ignore'
Return type: Dataframe with dropped values

```
# importing pandas module
import pandas as pd

# making data frame from csv file
data = pd.read_csv("nba.csv", index_col ="Name" )
```

```
# dropping passed values
data.drop(["Avery Bradley", "John Holland", "R.J. Hunter",
                        "R.J. Hunter"], inplace = True)

# display
data
```

Data Frame after Dropping values-



	Team	Number	Position	Age	Height	Weight	College	Salary
Name								
Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0
Jared Sullinger	Boston Celtics	7.0	C	24.0	6-9	260.0	Ohio State	2569260.0
Isaiah Thomas	Boston Celtics	4.0	PG	27.0	5-9	185.0	Washington	6912869.0
Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L43

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 17.04.2021

Topic of Lecture: Working with CSV file

Introduction : (Maximum 5 sentences)
It refers the Comma Separated Values file for processing.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

read_csv is an important pandas function to read csv files and do operations on it.

```
# Import pandas
import pandas as pd

# reading csv file
pd.read_csv("filename.csv")
```

Opening a CSV file through this is easy. But there are many others thing one can do through this function only to change the returned object completely. For instance, one can read a csv file not only locally, but from a URL through read_csv or one can choose what columns needed to export so that we don’t have to edit the array later.

Here is the list of parameters it takes with their **Default values**.

Video Content / Details of website for further learning (if any): https://www.geeksforgeeks.org/pandas-tutorial/
Important Books/Journals for further learning including the page nos.: https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L44

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 17.04.2021

Topic of Lecture: Handling missing data

Introduction : (Maximum 5 sentences)
There are different actions to be carried out on missing data in the data frame.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Cleaning / Filling Missing Data

Pandas treat None and NaN as essentially interchangeable for indicating missing or null values. To facilitate this convention, there are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

- isnull()
- notnull()
- dropna()
- fillna()
- replace()
- interpolate()

Replace NaN with a Scalar Value

The following program shows how you can replace "NaN" with "0".

Live Demo

```
import pandas as pd
import numpy as np

df = pd.DataFrame(np.random.randn(3, 3), index=['a', 'c', 'e'],columns=['one',
'two', 'three'])

df = df.reindex(['a', 'b', 'c'])
```

```
print df
print ("NaN replaced with '0':")
print df.fillna(0)
```

Its **output** is as follows −

```
        one        two     three
a  -0.576991  -0.741695  0.553172
b        NaN        NaN       NaN
c   0.744328  -1.735166  1.749580

NaN replaced with '0':
        one        two     three
a  -0.576991  -0.741695  0.553172
b   0.000000   0.000000  0.000000
c   0.744328  -1.735166  1.749580
```

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L45

MCA

I / II

Course Name with Code : **19CAB13 – Big Data Analytics**

Course Faculty : **Dr.M.Moorthy**

Unit: V - DATA ANALYTICS USING PYTHON **Date of Lecture:** 19.04.2021

Topic of Lecture: Querying, Sorting and Grouping

Introduction : (Maximum 5 sentences)

Query to extract values, Sorting for ordering data and Grouping is done for finding details based on group.

Prerequisite knowledge for Complete understanding and learning of Topic: (Max. Four important topics)

- **Knowledge on Internet**
- **Knowledge on SQL**
- **Knowledge on online transactions**

Detailed content of the Lecture:

Pandas now supports three types of Multi-axes indexing; the three types are mentioned in the following table −

Sr.No	Indexing & Description	
1	**.loc()**	Label based
2	**.iloc()**	Integer based
3	**.ix()**	Both Label and Integer based

Example

```
#import the pandas library and aliasing as pd
import pandas as pd
import numpy as np

df = pd.DataFrame(np.random.randn(8, 4),
index = ['a','b','c','d','e','f','g','h'], columns = ['A', 'B', 'C', 'D'])

#select all rows for a specific column
print df.loc[:,'A']
```

Its **output** is as follows −

```
a   0.391548
```

```
b  -0.070649
c  -0.317212
d  -2.162406
e   2.202797
f   0.613709
g   1.050559
h   1.122680
Name: A, dtype: float64
# importing pandas package
```

Sorting

```
import pandas as pd
# making data frame from csv file
data = pd.read_csv("nba.csv")

# sorting data frame by name
data.sort_values("Name", axis = 0, ascending = True,
                                    inplace = True, na_position ='last')
# display
data
```

Grouping

```
# using groupby function
# with one key

df.groupby('Name')
print(df.groupby('Name').groups)

# Using multiple keys in
# groupby() function
df.groupby(['Name', 'Qualification'])

print(df.groupby(['Name', 'Qualification']).groups)
```

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/pandas-tutorial/

Important Books/Journals for further learning including the page nos.:

https://www.geeksforgeeks.org/pandas-tutorial/

Course Faculty

Verified by HOD