

DIGITAL NOTES ON BIG DATA ANALYTICS

**B.TECH IV YEAR - I SEM
(2019-20)**



DEPARTMENT OF INFORMATION TECHNOLOGY

**MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)**

(Affiliated to JNTUH, Hyderabad, Approved by AICTE - Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)
Maisammaguda, Dhulapally (Post Via. Hakimpet), Secunderabad – 500100, Telangana State, INDIA.



MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

SYLLABUS

(R15A0530) BIG DATA ANALYTICS (ASSOCIATE ANALYTICS — II)
(Elective III)

Unit I:

Data Management (NOS 2101):

Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/signal/GPS etc. Data Management, Data Quality (noise, outliers, missing values, duplicate data) and Data Pre-processing.

Export all the data onto Cloud ex. AWS/Rackspace etc.

Maintain Healthy, Safe & Secure Working Environment (NOS 9003):

Introduction, workplace safety, Report Accidents & Emergencies, Protect health & safety as your work, course conclusion, assessment

Unit II

Big Data Tools (NOS 2101):

Introduction to Big Data tools like Hadoop, Spark, Impala etc., Data ETL process, Identify gaps in the data and follow-up for decision making.

Provide Data/Information in Standard Formats (NOS 9004):

Introduction, Knowledge Management, Standardized reporting & compliances, Decision Models, course conclusion. Assessment.

Unit III

Big Data Analytics:

Run descriptives to understand the nature of the available data, collate all the data sources to suffice business requirement, Run descriptive statistics for all the variables and observe the data ranges, Outlier detection and elimination.

Unit IV

Machine Learning Algorithms (NOS 9003):

Hypothesis testing and determining the multiple analytical methodologies, Train Model on 2/3 sample data using various Statistical/Machine learning algorithms, Test model on 1/3 sample for prediction etc.

Unit V

(NOS 9004)

Data Visualization (NOS 2101):

Prepare the data for Visualization, Use tools like Tableau, QlickView and D3, Draw insights out of Visualization tool. Product Implementation

TEXT BOOK

1. Student's Handbook for Associate Analytics.

REFERENCE BOOKS:

1. Introduction to Data Mining, Tan, Steinbach and Kumar, Addison Wesley, 2006
2. Data Mining Analysis and Concepts, M. Zaki and W. Meira (the authors have kindly made an online version available): <http://www.dataminingbook.info/uoloads/book.pdf>
3. Mining of Massive Datasets Jure Leskovec Stanford Univ. Anand RajaramanMilliway Labs Jeffrey D. Ullman Stanford Univ.
4. (<http://www.vistrails.org/index.php/Course: Big Data Analysis>)



MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

INDEX

S. No	Unit	Topic	Page no
1	I	Design Data Architecture and manage the data for analysis	6
2	I	understand various sources of Data like Sensors/signal/GPS etc.	7
3	I	Data Management, Data Quality (noise, outliers, missing values, duplicate data)	8
4	I	Data Pre-processing	9
5	I	Export all the data onto Cloud ex. AWS/Rackspace etc.	11
7	I	Introduction, workplace safety, Report Accidents & Emergencies, Protect health & safety as your work, course conclusion, assessment	13
8	II	Introduction to Big Data tools like Hadoop, Spark, Impala etc, Data ETL process, Identify gaps in the data and follow-up for decision making.	19
9	II	Provide Data/Information in Standard Formats	20
10	II	Knowledge Management	22
11	II	Standardized reporting & compliances	24
12	II	Decision Models, Course conclusion. Assessment	25
13	III	Run descriptives to understand the nature of the available data	28
14	III	collate all the data sources to suffice business requirement	32
15	III	Run descriptive statistics for all the variables and observer the data ranges	33

16	III	Outlier detection and elimination	36
17	IV	Hypothesis testing and determining the multiple analytical methodologies	37
18	IV	Train Model on 2/3 sample data using various Statistical/Machine learning algorithms,	39
19	IV	Test model on 1/3 sample for prediction etc.	40
20	V	Prepare the data for Visualization	41
21	V	Use tools like Tableau, QlickView and D3	42
22	V	Draw insights out of Visualization tool. Product Implementation	43

UNIT I Data Management (NOS 2101)

Design Data Architecture and manage the Data for analysis

Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing needs.

Enterprise requirements

These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management. In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes. One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a data warehouse).

Technology drivers

These are usually suggested by the completed data architecture and database architecture designs. In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing).

Economics

These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost. External factors such as the business cycle, interest rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

Business policies

Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency. These policies and rules will help describe the manner in which enterprise wishes to process their data.

Data processing needs

These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational initiatives as required (i.e. annual budgets, new product development).

The General Approach is based on designing the Architecture at three Levels of Specification :-

- The Logical Level
- The Physical Level
- The Implementation Level

Understand various sources of the Data

Data can be generated from two types of sources namely Primary and Secondary

Sources of Primary Data

The sources of generating primary data are -

- Observation Method
- Survey Method
- Experimental Method
- Experimental Method

There are number of experimental designs that are used in carrying out and experiment. However, Market researchers have used 4 experimental designs most frequently. These are -

CRD - Completely Randomized Design

RBD - Randomized Block Design - The term Randomized Block Design has originated from agricultural research. In this design several treatments of variables are applied to different blocks of land to ascertain their effect on the yield of the crop. Blocks are formed in such a manner that each block contains as many plots as a number of treatments so that one plot from each is selected at random for each treatment. The production of each plot is measured after the treatment is given. These data are then interpreted and inferences are drawn by using the analysis of Variance Technique so as to know the effect of various treatments like different dozes of fertilizers, different types of irrigation etc.

LSD - Latin Square Design - A Latin square is one of the experimental designs which has a balanced two way classification scheme say for example - 4 X 4 arrangement. In this scheme each letter from A to D occurs only once in each row and also only once in each column. The balance arrangement, it may be noted that, will not get disturbed if any row gets changed with the other.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

The balance arrangement achieved in a Latin Square is its main strength. In this design, the comparisons among treatments, will be free from both differences between rows and columns. Thus the magnitude of error will be smaller than any other design.

FD - Factorial Designs - This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyzes the impacts of each of the variables.

In a true experiment, randomization is essential so that the experimenter can infer cause and effect without any bias.

Sources of Secondary Data

While primary data can be collected through questionnaires, depth interview, focus group interviews, case studies, experimentation and observation; The secondary data can be obtained through

- Internal Sources - These are within the organization
- External Sources - These are outside the organization
- Internal Sources of Data

If available, internal secondary data may be obtained with less time, effort and money than the external secondary data. In addition, they may also be more pertinent to the situation at hand since they are from within the organization. The internal sources include

Accounting resources- This gives so much information which can be used by the marketing researcher. They give information about internal factors.

Sales Force Report- It gives information about the sale of a product. The information provided is of outside the organization.

Internal Experts- These are people who are heading the various departments. They can give an idea of how a particular thing is working

Miscellaneous Reports- These are what information you are getting from operational reports. If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources.

External Sources of Data

External Sources are sources which are outside the company in a larger environment. Collection of external data is more difficult because the data have much greater variety and the sources are much more numerous.

External data can be divided into following classes.

Government Publications- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are:

Registrar General of India- It is an office which generates demographic data. It includes details of gender, age, occupation etc.

Central Statistical Organization- This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

Director General of Commercial Intelligence- This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Ministry of Commerce and Industries- This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc. It also generates All India Consumer Price Index numbers for industrial workers, urban, non manual employees and cultural labourers.

Planning Commission- It provides the basic statistics of Indian Economy.

Reserve Bank of India- This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

Labour Bureau- It provides information on skilled, unskilled, white collared jobs etc.

National Sample Survey- This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.

Department of Economic Affairs- It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

State Statistical Abstract- This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

Non Government Publications- These includes publications of various industrial and trade associations, such as

The Indian Cotton Mill Association

Various chambers of commerce

The Bombay Stock Exchange (it publishes a directory containing financial accounts, key profitability and other relevant matter)

Various Associations of Press Media.

Export Promotion Council.

Confederation of Indian Industries (CII)

Small Industries Development Board of India

Different Mills like - Woolen mills, Textile mills etc

The only disadvantage of the above sources is that the data may be biased. They are likely to colour their negative points.

Syndicate Services- These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services. So the services are designed in such a way that the information suits the subscriber. These services are useful in television viewing, movement of consumer goods etc. These syndicate services provide information data from both household as well as institution.

In collecting data from household they use three approaches

Survey- They conduct surveys regarding - lifestyle, sociographic, general topics.

Mail Diary Panel- It may be related to 2 fields - Purchase and Media.

Electronic Scanner Services- These are used to generate data on volume.

They collect data for Institutions from

Whole sellers

Retailers, and

Industrial Firms

Various syndicate services are Operations Research Group (ORG) and The Indian Marketing Research Bureau (IMRB).

Importance of Syndicate Services

Syndicate services are becoming popular since the constraints of decision making are changing and we need more of specific decision-making in the light of changing environment. Also Syndicate services are able to provide information to the industries at a low unit cost.

Disadvantages of Syndicate Services

The information provided is not exclusive. A number of research agencies provide customized services which suits the requirement of each individual organization.

International Organization- These includes

The International Labour Organization (ILO)- It publishes data on the total and active population, employment, unemployment, wages and consumer prices

The Organization for Economic Co-operation and development (OECD) - It publishes data on foreign trade, industry, food, transport, and science and technology.

The International Monetary Fund (IMA) - It publishes reports on national and international foreign exchange regulations.

Export all the Data onto the cloud like Amazon web services S3

We usually export our data to cloud for purposes like safety, multiple access and real time simultaneous analysis.

There are various vendors which provide cloud storage services. We are discussing Amazon S3.

An Amazon S3 export transfers individual objects from Amazon S3 buckets to your device, creating one file for each object. You can export from more than one bucket and you can specify which files to export using manifest file options.

Export Job Proces

You create an export manifest file that specifies how to load data onto your device, including an encryption PIN code or password and details such as the name of the bucket that contains the data to export. For more information, see [The Export Manifest File](#). If you are going to mail us multiple storage devices, you must create a manifest file for each storage device.

You initiate an export job by sending a CreateJob request that includes the manifest file. You must submit a separate job request for each device. Your job expires after 30 days. If you do not send a device, there is no charge.

You can send a CreateJob request using the AWS Import/Export Tool, the AWS Command Line Interface (CLI), the AWS SDK for Java, or the AWS REST API. The easiest method is the AWS Import/Export Tool. For details, see

[Sending a CreateJob Request Using the AWS Import/Export Web Service Tool](#)

[Sending a CreateJob Request Using the AWS SDK for Java](#)

[Sending a CreateJob Request Using the REST API](#)

AWS Import/Export sends a response that includes a job ID, a signature value, and information on how to print your pre-paid shipping label. The response also saves a SIGNATURE file to your computer.

You will need this information in subsequent steps.

You copy the SIGNATURE file to the root directory of your storage device. You can use the file AWS sent or copy the signature value from the response into a new text file named SIGNATURE. The file name must be SIGNATURE and it must be in the device's root directory.

Each device you send must include the unique SIGNATURE file for that device and that JOBID. AWS Import/Export validates the SIGNATURE file on your storage device before starting the data load. If the SIGNATURE file is missing invalid (if, for instance, it is associated with a different job request), AWS Import/Export will not perform the data load and we will return your storage device.

Generate, print, and attach the pre-paid shipping label to the exterior of your package. See [Shipping Your Storage Device](#) for information on how to get your pre-paid shipping label.

You ship the device and cables to AWS through UPS. Make sure to include your job ID on the shipping label and on the device you are shipping. Otherwise, your job might be delayed. Your job expires after 30 days. If we receive your package after your job expires, we will return your device. You will only be charged for the shipping fees, if any. You must submit a separate job request for each device.

Note

You can send multiple devices in the same shipment. If you do, however, there are specific guidelines and limitations that govern what devices you can ship and how your devices must be packaged. If your shipment is not prepared and packed correctly, AWS Import/Export cannot process your jobs. Regardless of how many devices you ship at one time, you must submit a separate job request for each device. For complete details about packaging requirements when shipping multiple devices, see [Shipping Multiple Devices](#).

AWS Import/Export validates the signature on the root drive of your storage device. If the signature doesn't match the signature from the CreateJob response, AWS Import/Export can't load your data.

Once your storage device arrives at AWS, your data transfer typically begins by the end of the next business day. The time line for exporting your data depends on a number of factors, including the availability of an export station, the amount of data to export, and the data transfer rate of your device.

AWS reformats your device and encrypts your data using the PIN code or password you provided in your manifest.

We repack your storage device and ship it to the return shipping address listed in your manifest file. We do not ship to post office boxes.

You use your PIN code or TrueCrypt password to decrypt your device. For more information, see [Encrypting Your Data](#)

Health, Safety and Security

Why Workplace Safety

Ask the question to the participants and gather responses.

Discuss the responses with the group to understand the significance of workplace safety.

Basic Workplace Safety Guidelines

Prompt participants to come up with basic safety rules that they follow at their workplace.

Fire Safety

Employees should be aware of all emergency exits, including fire escape routes, of the office building and also the locations of fire extinguishers and alarms.

Falls and Slips

To avoid falls and slips, all things must be arranged properly. Any spilt liquid, food or other items such as paints must be immediately cleaned to avoid any accidents. Make sure there is proper lighting and all damaged equipment, stairways and light fixtures are repaired immediately.

First Aid

Employees should know about the location of first-aid kits in the office. First-aid kits should be kept in places that can be reached quickly. These kits should contain all the important items for first aid, for example, all the things required to deal with common problems such as cuts, burns, headaches, muscle cramps, etc.

Security

Employees should make sure that they keep their personal things in a safe place.

Electrical Safety

Employees must be provided basic knowledge of using electrical equipment and common problems. Employees must also be provided instructions about electrical safety such as keeping water and food items away from electrical equipment. Electrical staff and engineers should carry out routine inspections of all wiring to make sure there are no damaged or broken wires.

Case studies of hazardous events

Case 1: On Friday, June 13, 1997 a fire broke out at Uphaar Cinema, Green Park, Delhi, while the film Border was being shown. The fire happened because of a blast in a transformer in an underground parking lot in the five-organization building which housed the cinema hall and several offices. 59 people died and 103 were seriously hurt when people rushed to move out of the exit doors. Many people were trapped on the balcony and died because the exit doors were locked.

Case 2: 43 people died when fire broke out on the fifth and sixth floors of the Stephen Court building in Kolkata.

Case 3: 9 people were killed and 68 hurt when a fire accident took place in a commercial complex in Bangalore.

Case 4: In Kolkata, more than 90 people were killed when a fire broke out at the Advanced Medicare and Research Institute (AMRI) Hospitals at Dhakuria.

Accidents and Emergencies

Notice and correctly identify accidents and emergencies: You need to be aware of what constitutes an emergency and what constitutes an accident in an organization. The organization's policies and guidelines will be the best guide in this matter. You should be able to accurately identify such incidents in your organization. You should also be aware of the procedures to tackle each form of accident and emergency.

Follow company policies and procedures for preventing further injury while waiting for help to arrive: If someone is injured, do not act as per your impulse or gut feeling. Go as per the procedures laid down by your organization's policy for tackling injuries. You need to stay calm and follow the prescribed procedures. If you panic or act outside the prescribed guidelines, you may end up further aggravating the emergency situation or putting the injured person into further danger. You may even end up injuring yourself.

Act within the limits of your responsibility and authority when accidents and emergencies arise: Provide help and support within your authorized limit. Provide medical help to the injured only if you are certified to provide the necessary aid. Otherwise, wait for the professionals to arrive and give necessary help. In case of emergencies also, act within your authorized limits and let the professionals do the task allocated to them. Do not attempt to handle any emergency situation for which you do not have formal training or authority. You may end up harming yourself and the people around you.

Promptly follow instructions given by senior staff and the emergency services: Provide necessary services as described by the organization's policy for your role. Also, follow the instructions of senior staff that are trained to handle particular situations. Work under their supervision when handling accidents and emergencies.

Types of Accidents

The following are some of commonly occurring accidents in organizations:

Trip and fall: Customers or employees can trip on carelessly left loose material and fall down, such as tripping on loose wires, goods left on aisles, elevated threshold. This type of accident may result in simple bruises to serious fractures.

Injuries caused due to escalators or elevators (or lifts): Although such injuries are uncommon, they mainly happen to children, ladies, and elderly. Injuries can be caused by falling on escalators and getting hurt. People may be injured in elevators by falling down due to sudden, jerking movement of elevators or by tripping on elevators' threshold. They may also get stuck in elevators resulting in panic and trauma. Escalators and elevators should be checked regularly for proper and safe functioning by the right person or department. If you notice any sign of malfunctioning of escalators or elevators, immediately inform the right people. If organization's procedures are not being followed properly for checking and maintaining these, escalate to appropriate authorities in the organization.

Accidents due to falling of goods: Goods can fall on people from shelves or wall hangings and injure them. This typically happens if pieces of goods have been piled improperly or kept in an inappropriate manner. Always check that pieces of goods are placed properly and securely.

Accidents due to moving objects: Moving objects, such as trolleys, can also injure people in the organization. In addition, improperly kept props and lighting fixtures can result in accidents. For example, nails coming out dangerously from props can cause cuts. Loosely plugged in lighting fixtures can result in electric shocks.

Handling Accidents

Try to avoid accidents in your organization by finding out all potential hazards and eliminating them. If a colleague or customer in the organization is not following safety practices and precautions, inform your supervisor or any other authorized personnel. Always remember that one person's careless action can harm the safety of many others in the organization. In case of an injury to a colleague or a customer due to an accident in your organization, you should do the following:

Attend to the injured person immediately. Depending on the level and seriousness of the injury, see that the injured person receives first aid or medical help at the earliest. You can give medical treatment or first aid to the injured person only if you are qualified to give such treatments. Let trained authorized people give first aid or medical treatment.

Inform your supervisor about the accident giving details about the probable cause of accident and a description of the injury.

Assist your supervisor in investigating and finding out the actual cause of the accident. After identifying the cause of the accident, help your supervisor to take appropriate actions to prevent occurrences of similar accidents in future.

Each organization also has policies and procedures to tackle emergency situations. The purpose of these policies and procedures is to ensure safety and well-being of customers and staff during emergencies. Categories of emergencies may include the following:

Medical emergencies, such as heart attack or an expectant mother in labor: It is a medical condition that poses an immediate risk to a person's life or a long-term threat to the person's health if no actions are taken promptly.

Substance emergencies, such as fire, chemical spills, and explosions:
Substance emergency is an unfavourable situation caused by a toxic, hazardous, or inflammable substance that has the capability of doing mass scale damage to properties and people.

Structural emergencies, such as loss of power or collapsing of walls: Structural emergency is an unfavourable situation caused by development of some faults in the building in which the organization is located. Such an emergency can also be caused by the failure of an essential function or service in the building, such as electricity or water supply failure. Such emergencies result in a long-term or permanent disruption of the organization's functions.

Key Points

Security emergencies, such as armed robberies, intruders, and mob attacks or civil disorder:
Security emergency is an unfavourable situation caused by a breach in security posing a significant danger to life and property.

Natural disaster emergencies, such as floods and earthquakes: It is an emergency situation caused by some natural calamity leading to injuries or deaths, as well as a large-scale destruction of properties and essential service infrastructures.

– Protect Health & Safety as You Work

Here are some potential sources of hazards in an organization:

Using computers: Hazards include poor sitting postures or excessive duration of sitting in one position. These hazards may result in pain and strain. Making same movement repetitively can also cause muscle fatigue. In addition, glare from the computer screen can be harmful to eyes. Stretching up at regular intervals or doing some simple yoga in your seat only can mitigate such hazards.

Handling office equipment: Improper handling of office equipment can result in injuries. For example, sharp-edged equipment if not handled properly can cause cuts. Staff members should be trained to handle equipment properly. Relevant manual should be made available by administration on handling equipment.

Handling objects: Lifting or moving heavy items without proper procedure or techniques can be a source of potential hazard. Always follow approved procedure and proper posture for lifting or moving objects.

Stress at work: In today's organization, you may encounter various stress causing hazards. Long working hours can be stressful and so can be aggressive conflicts or arguments with colleagues. Always look for ways for conflict resolution with colleagues. Have some relaxing hobbies for stress against long working hours.

Working environment: Potential hazards may include poor ventilation, inappropriate height chairs and tables, stiffness of furniture, poor lighting, staff unaware of emergency procedures, or poor housekeeping. Hazards may also include physical or emotional intimidation, such as bullying or ganging up against someone. Staff should be made aware of organization's policies to

General Evacuation Procedures

Each organization will have its own evacuation procedures as listed in its policies. An alert employee, who is well-informed about evacuation procedures, can not only save him or herself, but also helps others in case of emergencies. Therefore, you should be aware of these procedures and follow them properly during an emergency evacuation. Read your organization's policies to know about the procedures endorsed by it. In addition, here are a few general evacuation steps that will always be useful in such situations:

Leave the premises immediately and start moving towards the nearest emergency exit.

- o Guide your customers to the emergency exits.
- o If possible, assist any person with disability to move towards the emergency exit. However, do not try to carry anyone unless you are trained to do so.
- o Keep yourself light when evacuating the premises. You may carry your hand-held belongings, such as bags or briefcase as you move towards the emergency exit. However, do not come back into the building to pick up your belongings unless the area is declared safe.
- o Do not use the escalators or elevators (lifts) to avoid overcrowding and getting trapped, in case there is a power failure. Use the stairs instead.
- o Go to the emergency assembly area. Check if any of your colleagues are missing and immediately inform the personnel in charge of emergency evacuation or your supervisor.
- o Do not go back to the building you have evacuated till you are informed by authorized personnel that it is safe to go inside.

Safety Signs

Review: Safety Guidelines Checklist

1. Store all cleaning chemicals in tightly closed containers in separate cupboards.
2. Keep the kitchen clean and dry all the time.
3. Throw away rubbish daily.
4. Make sure all areas have proper lighting.
5. In case of any injury or fracture, do not move the person until he or she has received medical attention.
6. Do not wear loose clothing or jewelry when working with machines. It may catch on moving equipment and cause a serious injury.
7. Never distract the attention of people who are working near fire or with some machinery, tools or equipment.
8. Where required, wear protective items, such as goggles, safety glasses, masks, gloves, hair nets, etc.
9. Shut down all machines before leaving for the day.
10. Do not play with electrical controls or switches.
11. Do not operate machines or equipment until you have been properly trained and allowed to do so by your supervisor.
12. Do not adjust, clean or oil moving machinery.
13. Stack all shelves in an orderly way.
14. Stack all boxes and crates properly.
15. Never leave dishrags, aprons and other clothing near any hot surface.
16. Repair torn wires or broken plugs before using any electrical equipment.
17. Do not use equipment if it smokes, sparks or looks unsafe.
18. Cover all food with a lid, plastic wrap or aluminium foil.
19. Do not smoke in “No Smoking” areas.
20. Report any unsafe condition or acts to your supervisor. These could include:

UNIT II Big Data Tools (NOS 2101)

Introduction to the Big Data tools like Spark, Scala, Impala

The tools used for Big Data handling and analysis and further reporting are called Big Data Tools.

1. Apache Spark :-

Apache Spark is an open source cluster computing framework originally developed in the AMPLab at University of California, Berkeley but was later donated to the Apache Software Foundation where it remains today. In contrast to Hadoop's two-stage disk-based Map Reduce paradigm, Spark's multi-stage in-memory primitives provide performance up to 100 times faster for certain applications. By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well-suited to machine learning algorithms.

Spark requires a cluster manager and a distributed storage system. For cluster management, Spark supports standalone (native Spark cluster), Hadoop YARN, or Apache Mesos. For distributed storage, Spark can interface with a wide variety, including Hadoop Distributed File System (HDFS), Cassandra, OpenStack Swift, Amazon S3, or a custom solution can be implemented. Spark also supports a pseudo-distributed local mode, usually used only for development or testing purposes, where distributed storage is not required and the local file system can be used instead; in such a scenario, Spark is run on a single machine with one executor per CPU core.

2. Scala:-

Scala is a programming language for general software applications. Scala has full support for functional programming and a very strong static type system. This allows programs written in Scala to be very concise and thus smaller in size than other general-purpose programming languages. Many of Scala's design decisions were inspired by criticism of the shortcomings of Java.

Scala source code is intended to be compiled to Java byte code, so that the resulting executable code runs on a Java virtual machine. Java libraries may be used directly in Scala code and vice versa (language interoperability). Like Java, Scala is object-oriented, and uses curly-brace syntax reminiscent of the C programming language. Unlike Java, Scala has many features of functional programming languages like Scheme, Standard ML and Haskell, including currying, type inference, immutability, lazy evaluation, and pattern matching. It also has an advanced type system supporting algebraic data types, covariance and contravariance, higher-order types (but not higher-rank types), and anonymous types. Other features of Scala not present in Java include operator overloading, optional parameters, named parameters, raw strings, and no checked exceptions.

3. Cloudera Impala:-

Cloudera Impala is Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop.

Cloudera Impala is a query engine that runs on Apache Hadoop. The project was announced in October 2012 with a public beta test distribution and became generally available in May 2013.

Impala brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation. Impala is integrated with Hadoop to use the same file and data formats, metadata, security and resource management frameworks used by MapReduce, Apache Hive, Apache Pig and other Hadoop software.

Impala is promoted for analysts and data scientists to perform analytics on data stored in Hadoop via SQL or business intelligence tools. The result is that large-scale data processing (via MapReduce) and interactive queries can be done on the same system using the same data and metadata – removing the need to migrate data sets into specialized systems and/or proprietary formats simply to perform analysis.

Features include:

- Supports HDFS and Apache HBase storage,
- Reads Hadoop file formats, including text, LZO, SequenceFile, Avro, RCFile, and Parquet,
- Supports Hadoop security (Kerberos authentication),
- Fine-grained, role-based authorization with Apache Sentry,
- Uses metadata, ODBC driver, and SQL syntax from Apache Hive.
- In early 2013, a column-oriented file format called Parquet was announced for architectures including Impala. In December 2013, Amazon Web Services announced support for Impala. In early 2014, MapR added support for Impala.

Identify gaps in the data and follow-up for decision making

There can be two types of gap in Data:-

1. Missing Data Imputation
2. Model based Techniques

For missing values we have got several treatments like replacement with Average value or Removal. While for analysis to be proper we select the variables for modeling based on correlation test results.

Techniques of dealing with missing data

Missing data reduce the representativeness of the sample and can therefore distort inferences about the population. If it is possible try to think about how to prevent data from missingness before the actual data gathering takes place. For example, in computer questionnaires it is often not possible to skip a question. A question has to be answered, otherwise one cannot continue to the next. So missing values due to the participant are eliminated by this type of questionnaire, though this method may not be permitted by an ethics board overseeing the research. And in survey research, it is common to make multiple efforts to contact each individual in the sample, often sending letters to attempt to persuade those who have decided not to participate to change their minds. However, such techniques can either help or hurt in terms of reducing the negative inferential effects of missing data, because the kind of people who are willing to be persuaded to participate after initially refusing or not being home are likely to be significantly different from the kinds of people who will still refuse or remain unreachable after additional effort.

In situations where missing data are likely to occur, the researcher is often advised to plan to use methods of data analysis methods that are robust to missingness. An analysis is robust when we are confident that mild to moderate violations of the technique's key assumptions will produce little or no bias, or distortion in the conclusions drawn about the population.

Imputation

If it is known that the data analysis technique which is to be used isn't content robust, it is good to consider imputing the missing data. This can be done in several ways. Recommended is to use multiple imputations. Rubin (1987) argued that even a small number (5 or fewer) of repeated imputations enormously improves the quality of estimation.

For many practical purposes, 2 or 3 imputations capture most of the relative efficiency that could be captured with a larger number of imputations. However, a too-small number of imputations can lead to a substantial loss of statistical power, and some scholars now recommend 20 to 100 or more.[8] Any multiply-imputed data analysis must be repeated for each of the imputed data sets and, in some cases, the relevant statistics must be combined in a relatively complicated way.

Examples of imputations are listed below.

Partial imputation

The expectation-maximization algorithm is an approach in which values of the statistics which would be computed if a complete dataset were available are estimated (imputed), taking into account the pattern of missing data. In this approach, values for individual missing data-items are not usually imputed.

Partial deletion

Methods which involve reducing the data available to a dataset having no missing values include:

Listwise deletion/casewise deletion

Pairwise deletion

Full analysis

Methods which take full account of all information available, without the distortion resulting from using imputed values as if they were actually observed:

Knowledge Mangement

Introductions

I am <Facilitator's Name> and I am your facilitator today.”

Briefly review the roles of the Lead Facilitator and Support Facilitator, if any.

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]

“Why are you here today?”

After reviewing and arranging responses, summarize the responses and map the responses to the suggested course benefits below.

“Regardless of why you're here today, we're all going to walk away with some key benefits – let's discuss those briefly.”

Debrief the following:

Why is knowledge management so important?

- It is important to put data into information
- Retention of information is one of the most important challenges an organization has
- Information needs to be presented as reports which should be standardized to as much extent possible
- When publishing reports, it is important to collaborate with everyone
- We also need to look at some decision models which help us in taking the right decisions

Importance of knowledge management

Provide a brief overview of the session. Discuss the importance of knowledge management from an organization's standpoint.

Open up the discussion for the session and ask participants to share their thoughts on "knowledge management"?

The first part of this session discusses that the following are the functions of Knowledge Management:

- Capture uniqueness of each project in new growth and improvise existing work
- Decouple the Art with Process and make complex work scalable
- Reduce people dependencies

Knowledge Management- Industrializing Collective Brain Power

Effectively using organizational **knowledge**. It refers to a multi-disciplinary approach

KM to support an organization's growth engine :

Capture uniqueness of each project in new growth and improvise existing work

- o Decouple the Art with Process and make complex work scalable
 - o Reduce people dependencies

KM processes of a few organizations

Now we are going to look at the Knowledge Management Processes of a few organizations. It is important to note that each organization will have some set standards, methods and approaches towards knowledge management.

As we go through the various approaches of various organizations, let's try to evaluate and find out the commonalities between them.

Standardise Reporting and Compliance

Reporting templates are pre-created structures based on which reports are to be created. These templates can be of any of the following types:

- Financial reporting templates
- Marketing and sales reporting templates
- Data entry templates
- Research templates
- Pricing and product costing templates
- Any other reporting or data presentation requirements

Some Examples of standardized templates as used in Organizations

Whitepapers

A white paper is an authoritative report or guide informing readers in a concise manner about a complex issue and presenting the issuing body's philosophy on the matter. It is meant to help readers understand an issue, solve a problem, or make a decision.

Organizations create Whitepapers all the time to document standard processes

Organizing data/information

Industry Experts states that several key success factors or mechanisms can lead to high quality knowledge content. These mechanisms assure knowledge base used by the analysts remain up to date, relevant, and valid. The five major mechanisms are:

- Standardized content formats
- A clearly specified knowledge content production process
- Informal or format peer review assuring that the document knowledge is valid & relevant
- Information quality criteria
- Guidelines – specifying minimal requirements in terms of document content, style, size & ownership and format

Decision Model

The business logic behind a business **decision**. An informal definition of business logic is

Decision Models are used to model a decision being made once as well as to model a repeatable decision-making approach that will be used over and over again.

The Vroom-Yetton-Jago Decision Model

Origin : This model was originally described by Victor Vroom and Philip Yetton in their 1973 book titled Leadership and Decision Making. Later in 1988, Vroom and Arthur Jago, replaced the decision tree system of the original model with an expert system based on mathematics. Hence you will see the model called Vroom-Yetton, Vroom-Jago, and Vroom-Yetton-Jago. The model here is based on the Vroom-Jago version of the model.

Understanding the Model

When you sit down to make a decision, your style, and the degree of participation you need to get from your team, are affected by three main factors:

- Decision Quality** – how important is it to come up with the "right" solution? The higher the quality of the decision needed, the more you should involve other people in the decision.
- Time Constraints** – How much time do
- you have to make the decision? The more time you have, the more you have the luxury of including others, and of using the decision as an opportunity for teambuilding.

Specific Leadership Styles

The way that these factors impact on you helps you determine the best leadership and decision-making style to use. Vroom-Jago distinguishes three styles of leadership, and five different processes of decision-making that you can consider using:

OODA LOOPS

It can be fun to read books like *The Art of War*, written in 6th Century China by Sun Tzu, and to think about how these can be applied to business strategy. So when former US Air Force Colonel John Boyd developed his model for decision-making in air combat, its potential applications in business soon became apparent.

Boyd developed his model after analyzing the success of the American F-86 fighter plane compared with that of the Soviet MIG-15. Although the MIG was faster and could turn better, the American plane won more battles because, according to Boyd, the pilot's field of vision was far superior.

This improved field of vision gave the pilot a clear competitive advantage, as it meant he could assess the situation better and faster than his opponent. As a result, he could out-maneuver the enemy pilot, who would be put off-balance, wouldn't know what to expect, and would start making mistakes.

Success in business often comes from being one step ahead of the competition and, at the same time, being prepared to react to what they do. With global, real-time communication, ongoing rapid improvements in information technology, and economic turbulence, we all need to keep updating and revising our strategies to keep pace with a changing environment.

See the similarities with Boyd's observations? Brought together in his model, they can hold a useful lesson for modern business.

Proactive decision-making. The four stages are:

1. **Observe** – collect current information from as many sources as practically possible.
2. **Orient** – analyze this information, and use it to update your current reality.
3. **Decide** – determine a course of action.
4. **Act** – follow through on your decision.

You continue to cycle through the OODA Loop by observing the results of your actions, seeing whether you've achieved the results you intended, reviewing and revising your initial decision, and moving to your next action.

Observing and orienting correctly are key to a successful decision. If these steps are flawed, they'll lead you to a flawed decision, and a flawed subsequent action. So while speed is important, so too is improving your analytical skills and being able to see what's really happening.

The OODA Loop model is closely related to **Plan Do Check Act**. Both highlight the importance of analyzing a situation accurately, checking that your actions are having the intended results, and making changes as needed.

Stage 1. Observe

At this initial point in the loop, you should be on the look-out for new information, and need to be aware of unfolding circumstances. The more information you can take in here, the more accurate your perception will be. Like an F-86 pilot with a wide field of vision, you want to capture as much

Stage 2. Orient

One of the main problems with decision-making comes at the Orient stage: we all view events in a way that's filtered through our own experiences and perceptions. Boyd identified five main influences:

Cultural traditions.

- Genetic heritage.
- The ability to analyze and synthesize.
- Previous experience.
- New information coming in.

Orientation is essentially how you interpret a situation. This then leads directly to your decision. The argument here is that by becoming more aware of your perceptions, and by speeding up your ability to orient to reality, you can move through the decision loop quickly and effectively. The quicker you understand what's going on, the better. And if you can make sense of the situation and the environment around you faster than your competition, you'll have an advantage.

And it's important to remember that you're constantly re-orienting. As new information comes in at the Observe stage, you need to process it quickly and revise your orientation accordingly.

Stage 3. Decide

Decisions are really your best guesses, based on the observations you've made and the orientation you're using. As such, they should be considered to be fluid works-in-progress. As you keep on cycling through the OODA Loop, and new suggestions keep arriving, these can trigger changes to your decisions.

Stage 4. Act

The Act stage is where you implement your decision. You then cycle back to the Observe stage, as you judge the effects of your action. This is where actions influence the rest of the cycle, and it's important to keep learning from what you, and your opponents, are doing.

UNIT III Big Data Analytics

Introduction to Big Data Analytics

Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence(BI) programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

Introduction

Relational and transactional databases based on SQL language have clearly dominated the market of data storage and data manipulation over the past 20 years. Several factors can explain this position of technological leadership. First of all, SQL is a standardized language, even if each vendor has implemented slight adaptation on it. This aspect is a key factor of cost reduction for enterprises in term of training in comparison of specific and proprietary technologies. Secondly, SQL is embedding most of commonly used functionalities to manage transactions and insure the integrity of data. Finally, this technology is very mature and over time a lot of powerful tools have been implemented in term of backup, monitoring, analytics...

From relational databases to Big Data

If the term “NoSQL” figures out that the SQL language is not adapted to distributed databases, in fact it is more the principle on which it is build that are difficult to apply: the relational and transactional data model, implemented in third normal form.

As a relational database, it provides a set of functionalities to access data across several entities (tables) by complex queries. It provides also integrity referential to insure the constant validity of the links between entities. Such mechanisms are extremely costly and complex to implement in distributed architecture, considering that it is necessary to insure that all data that are linked together have to be hosted on the same node. Moreover, it implies the definition a static data-model or schema, not applicable to the velocity of web data.

As a transactional database, they must respect the ACID constraints, i.e. the **Atomicity** of updates, the **Consistency** of the database, the **Isolation** and the **Durability** of queries. These constraints are perfectly applicable in a centralized architecture, but much more complex to insure in a distributed architecture. NoSQL and Big Data

In one word, both the **3rd normal form** and the **ACID constraints** make relational databases intolerant to the partitioning of data. However, three major criteria can be considered as a triptych in the implementation of a distributed architecture:

Coherence: All the nodes of the system have to see exactly the same data at the same time

Availability: The system must stay up and running even if one of its node is failing down

As established in the so called “CAP theorem”, the implementation of these three characteristics **at the same time** is not possible in a distributed architecture and a trade-off is necessary. On a practical point of view, a relational database insures the availability and coherence of data, but it shows many limitations regarding the tolerance to partitioning.

As a consequence, major players of the market of online services had to implement specific solutions in term of data storage, proprietary in a first hand, and then transmitted to open sources communities that have insured the convergence of these heterogeneous solutions in four major categories of NoSQL databases:

–value store,

Each of these four categories has its own area of applicability. *Key-Value* and *columns* databases address the volume of data and the scalability. They are implementing the *availability* of the database. *Document* and *graph* databases are more focused on the complexity of data, and thus on the *coherence* of the database. NoSQL and Big Data

Key-value store concept

This technology can address a large volume of data due to the simplicity of its data model. Each object is identified by a unique key and the access to this data is only possible through this key. The structure of the object is free. This model only provides the four basic operations to Create, Read, Update and Delete an object from its key. Generally, these databases are providing in façade a HTTP REST API so that they can interoperate with any language.

This simple approach has the benefit to provide exceptional performance in read and write access, and a large scalability of data. However, it provides only limited querying facilities, considering that data can only be retrieved from their key, and not their content.

Columns based databases concept

Columns based databases are storing data in grids, in which the column is the basic entity that represents a data field. Columns can be grouped together through the concept of columns NoSQL and Big Data families.

Rows of the grids are assimilated to records and identified by a unique Key such as in the *Key-value* model previously described. Additionally, some providers are also including in their model the concept of *version* as a third dimension of the grid.

The organization of the database in grids can appear similar to the *tables* of relational databases. However, the approach is completely different. While the columns of a relational table are static and present for each record, this is not the case in Columns Oriented Database so that it is possible to dynamically add a column to a table with no cost in term of storage space.

These databases are designed to store up to several millions of columns that can be fields of an entity or one-to-many relationships. Originally, their associated querying engine was designed to retrieve ranges of rows from the value of the keys, and columns from their names.

Document based databases concept

Document based databases are similar to Key-value stores except that the value associated to the key can be a structured and complex objects rather than a simple types. These complex objects are generally structured in XML or JSON formalism. This approach allows the implementation of queries on the content of the documents and not only through the key of the record. NoSQL and Big Data

Even if the documents are structured, these databases are *schemaless*, meaning that it is not necessary to previously determine the structure of the document. The simplicity and flexibility of this data model makes it particularly applicable to *Content Management Systems* (CMS).

Graph databases concept

The graph paradigm is a data model in which entities are *nodes* and associations between entities are *arcs* or *relationships*. Both nodes and relationships are characterized by a set of properties. This category of databases is typically designed to address the complexity of databases more than their volumetric. They are particularly relevant to use as soon as the number of relationships between business objects are increasing. In particular, they are applied in cartography, social networks, and more generally in network modelling.

MapReduce

MapReduce is a programming technique used to divide a database treatment in multiple sub-treatments that can be executed in parallel across the distributed architecture of the database. The term MapReduce actually refers to two separate and distinct tasks. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into key/value pairs. The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples.

As an example let's assume that we want to count the number of occurrences of each words of a book. The *Map* treatment would consist to launch one process on each node of the distributed architecture, taking in charge a range of page. The output of these processes would be an alphabetically sorted Map of key-values where keys are the words and values are the number of occurrences of that word. Then, the *Reduce* process would consist to concatenate and re-sort the output of the nodes alphabetically, and consolidate (by sum) the number of occurrences returned for each word by the sub processes.

Descriptive Statistics

Called the "simplest class of analytics", descriptive analytics allows you to condense big data into smaller, more useful bits of information or a summary of what happened.

It has been estimated that more than 80% of business analytics (e.g. social analytics) are descriptive. Some social data could include the number of posts, fans, followers, page views, check-ins,pins, etc. It would appear to be an endless list if we tried to list them all.

Outlier detection and elimination

- Data that don't conform to the normal and expected patterns are Outliers.
- Wide range of application in various domains including finance, security, intrusion detection in cyber security.
- Criteria for what constitutes an outlier depend the problem domain.
- Typically involve large amount of data which may be unstructured.
- Outliers elimination is already discussed in this book before.

Data pre-processing for the analysis

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

UNIT IV Machine Learning Algorithms

Hypothesis testing and determining the multiple analytical methodologies

Machine learning usually refers to changes in systems that perform tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc.

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are:

- ❑ **Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- ❑ **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
- ❑ **Reinforcement learning:** A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing. Transduction is a special case of this principle where the entire set of problem instances is known at learning time, except that part of the targets is missing.

A support vector machine is a classifier that divides its input space into two regions, separated by a linear boundary. Here, it has learned to distinguish black and white circles.

Among other categories of machine learning problems, learning to learn learns its own inductive bias based on previous experience. Developmental learning, elaborated for robot learning, generates its own sequences (also called curriculum) of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers, and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation.

Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system.

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
- Density estimation finds the distribution of inputs in some space.
- Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space.

Machine learning and data mining often employ the same methods and overlap significantly. They can be roughly distinguished as follows:

- Machine learning focuses on prediction, based on known properties learned from the training data.
- Data mining focuses on the discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases.
-

The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind. On the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these two research communities (which do often have separate conferences and separate journals, ECML PKDD being a major exception) comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in Knowledge Discovery and Data Mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

Machine learning also has intimate ties to optimization: many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances. For example, in classification, one wants to assign a label to instances, and models are trained to

correctly predict the pre-assigned labels of a set examples.

The difference between the two fields arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples.

Train model using statistical/machine learning algorithms, Test model

To train the algorithm we feed it quality data known as a training set. A training set is the set of training examples we'll use to train our machine learning algorithms.

Train the algorithm: - This is where the machine learning takes place. This step and the next step are where the “core” algorithms lie, depending on the algorithm. You feed the algorithm good clean data from the first two steps and

extract knowledge or information. This knowledge you often store in a format that's readily useable by a machine for the next two steps. In the case of unsupervised learning, there's no training step because you don't have a target value. Everything is used in the next step.

Test the algorithm:- This is where the information learned in the previous step is put to use. When you're evaluating an algorithm, you'll test it to see how well it does. In the case of supervised learning, you have some known values you can use to evaluate the algorithm. In unsupervised learning, you may have to use some other metrics to evaluate the success.

Sample for prediction

For prediction various types of algorithms are used.

- **Collect data.** You could collect the samples by scraping a website and extracting data, or you could get information from an RSS feed or an API. You could have a device collect wind speed measurements and send them to you, or blood glucose levels, or anything you can measure. The number of options is endless. To save some time and effort, you could use publicly available data.

- **Prepare the input data.** Once you have this data, you need to make sure it's in a useable format. The format we'll be using in this book is the Python list. We'll talk about Python more in a little bit, and lists are reviewed in appendix A. The benefit of having this standard format is that you can mix and match algorithms and data sources. You may need to do some algorithm-specific formatting here. Some algorithms need features in a special format, some algorithms can deal with target variables and features as strings, and some

need them to be integers. We'll get to this later, but the algorithm-specific formatting is usually trivial compared to collecting data. One idea that naturally arises is combining multiple classifiers. Methods that do this are known as ensemble methods or meta-algorithms. Ensemble methods can take the form of using different algorithms, using the same algorithm with different settings, or assigning different parts of the dataset to different classifiers.

Explore the chosen algorithms for more accuracy

Analyze the input data. This is looking at the data from the previous task. This could be as simple as looking at the data you've parsed in a text editor to make sure that data is collected and prepared in proper way and are actually working and you don't have a bunch of empty values. You can also look at the data to see if you can recognize any patterns or if there's anything obvious, such as a few data points that are vastly different from the rest of the set. Plotting data in one, two, or three dimensions can also help. But most of the time you'll have more than three features and you can't easily plot the data across all features at one time. You could, however, use some advanced methods we'll talk about later to distill multiple dimensions down to two or three so you can visualize the data.

If you're working with a production system and you know what the data should look like, or you trust its source, you can skip this step. This step takes human involvement, and for an automated system you don't want human involvement. The value of this step is that it makes you understand you don't have garbage coming in.

UNIT V Data Visualization

Prepare the data for visualization

Data presentation architecture (DPA) is a skill-set that seeks to identify, locate, manipulate, format and present data in such a way as to optimally communicate meaning and proffer knowledge.

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It is not owned by any one field, but rather finds interpretation across many (e.g. it is viewed as a modern branch of descriptive statistics by some, but also as a grounded theory development tool by others). It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".

A primary goal of data visualization is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables, and charts selected. Effective visualization helps users in analyzing and reasoning about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look-up a specific measure of a variable, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science. The rate at which data is generated has increased, driven by an increasingly information-based economy. Data created by internet activity and an expanding number of sensors in the environment, such as satellites and traffic cameras, are referred to as "Big Data". Processing, analyzing and communicating this data present a variety of ethical and analytical challenges for data visualization. The field of data science and practitioners called data scientists has emerged to help address this challenge.

Draw insights out of the visualization tool

Graphical displays should:

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.
- Graphics reveal data. Indeed graphics can be more precise and revealing than conventional statistical computations.

Data Visualization in Tableau

Extract The Data.

We need to choose the *dimensions* and *measures* of the data you want to analyze. Dimensions are the category type data points such as landing page, source medium, etc. The measures are the number entries such as visits, bounces, etc.

Keep in mind that the more dimensions you add, the larger the data set will get. For example, adding a device type will give you one row of measures for each device. You can think of it this way: if your default data has 10,000 rows, and you add the hour dimension, you would have $10,000 * 24$ (hours). So, if you add hours and mobile device type, you would have $10,000 * 24 * 250 = 60,000,000$. So, make sure you only pull the dimensions that actually matter.

The Workspace

Now that we have loaded our data for this exercise, you should get familiar with the tool's workspace. You'll note that it is divided into three main sections: data, settings, and visualizations. In addition, you can see two sets of data on the left side of the screen — your dimensions are on the top, and your measures are on the bottom. Lastly, note the columns and rows sections near the top of the screen — they are a fundamental concept of Tableau.

Planning and Estimation

Planning and estimation are procedures that anticipate future demand based on current usage or growth patterns. The expectation of a value in the future is necessary to make appropriate policies, actions or strategies. For example, if the country is witnessing a rise in temperature in the current summer, planning and estimation techniques allow agricultural statisticians to gauge the temperature impact on the future winter crop and the impact of a potential downfall in yield. Estimates of the yield provides insights into food stocking and crop rotation schemes. Probability and statistics, hypothesis testing in particular allows professionals to test the sample size and relate to the how the population would behave, with a certain level of confidence. Analyzing crop yield from a particular unbiased farm allows to estimate the population (country) mean decrease in farm output. Further, if the data is recorded as a time series, then forecasting methods allow to project, again with certain prediction bounds, what the expected level of yield would be in the next season. Thus trend and seasonality effects are accounted for.

Drivers of Asset/Engine risk categories analytics

Assets (vehicles, wind turbines, oil rigs etc.) are all prone to variability in operation due to temporal and environmental effects. It is important to note that not all components in a system fail at the same rate or in the same mode. Thus, it becomes imperative to understand the risk associated with each component (and system) by identifying the category or risk that it is attached to. High risk components are to have higher priority from either safety or regulatory stand point. To identify such risks, variables that correspond that the risk are first identified. For instance, temperature is a variable that is associated to a person's health (state of fever). Hence, every risk can have multiple such variables defining it. Consequently, to better manage the system, risk categories are developed to isolate particularly high risk bins that need immediate attention. However, the chance of dealing with multiple risks and multiple variables remains. Variable reduction methods are then employed to analyze the system using a reduced set of variables that is representative yet concise. For instance, if 10 variables are associated with a risk, then potentially 6 of them carry 90% of the information and are thus more important than the other 4. Model building is another tool using methods such as regression that allow to functionally characterize the risk. Newer methods such as clustering and classification allow to place the risk in its priority state so that the underlying causes can easily be identified.

Different approaches to Asset scoring models

Knowing the current health of an asset is important for 2 reasons: it allows to know how long it will last (provide service) and when to maintain/repair/replace it. To obtain these time stamps, reliability models are employed widely to isolate the survivability index of the asset. For instance, a tire used in a car is subjected to a variety of road and use conditions. Knowing how long it will last provides an estimate of replenishment time. In doing such studies, degradation models (hazard functions) are frequently employed to assess the state-of-health and the speed at which a certain asset is losing its operational health.

Key factors determining the Asset scoring

When assets are scored, it is important to know how or what metrics to be chosen. If cars are valued depending on their mileage, then the miles provided per gallon is an apt parameter. Systems are often studied using multiple parameters where ANOVA methods help to identify key contributing factors using p-values. Modern learning models such as Neural Nets and Random Forests can also be used to score and map the asset.

Validation and maintenance of Asset scoring models

Post scoring, assets are categorized based on the net value (score) that they possess. However, such scores need to be validated and periodically updated. In doing so, data is collected often and split into testing and training. The training data in conjunction with training algorithms allows to develop the asset score while the test data using methods such as k-fold cross validation ensures that the score developed is error minimized. Thus, the combination provides a validated asset score. Maintaining the score implies updating the model used to build the score. Data is collected and refreshed from time to time based on which the most current score is formulated.

Understanding the current Engineering / Manufacturing / Asset system

Prior to any analysis or analytics being employed, it is prudent to understand what the current state of a system/process is. This is usually referred to as baseline estimation. Also important is the comparison of the baseline with other similar systems/processes from contemporary industries. Benchmarking is a technique that allows such comparison. The outcome of this evaluation is the establishment of the industry standard and also the deviation from such standard as seen in the test system/process. Sampling tests allow the gauge the extent of such deviations statistically and also potential areas of improvement to match or exceed the current standards.

Creating the Business Understanding Document

In any analytics project, the most critical drivers are variables that are correlated directly to business goals. If profit margin increase is the desired business goal, then sales and revenues are correlated variables. In developing a business understanding document, it is thus imperative to not just identify the right variables but also discard extraneous causes. Business understanding thus necessarily means “What is to be done to achieve the objective?” If emissions from an automobile are to be reduced by say 10%, then the business goal is either cleaner combustion or better exhaust management, each of which has individual variables that characterize it. Sometimes, the business goal is convoluted, meaning that it can be multi-dimensional at which point trend analysis and covariance metrics are viable analysis options.

Understanding Data and creation of Data Dictionary

Data always tells a story, even when there is no discernible pattern. Descriptive statistics, simple explanations of mean, standard deviation, mode, number of observations, range etc. gives a fairly decent understanding of the characteristics of a sample (population). If a class has 10 students, collecting scores from 5 different tests and doing descriptive statistics provides an understanding of how the class is learning as a whole. Likewise, inferential statistics allows to gage specific improvement regions for the class as a whole based on inferences from the study.

Preparing the Data

Data makes sense if it is proper i.e., it is cleaned and sanitized. Clerical and entry errors are to be removed and adherence to existing reference documents allows data to be prepped for further analysis. Also, data understanding (structured vs. unstructured, numerical vs. text, continuous vs. categorical) provides some insight into how data is classified. Further, checks need to be done in terms of whether there are any missing values (discontinuous data) at which point imputation (filling in missing values) needs to be done. Therefore, preparation of data is critical for further analysis and final outcome. Additional checks such as collinearity effects (variables being dependent on each other) also need to be addressed.

Analysis and Modeling - which will include the Classing Report, Variable Reduction report, Model statistics

Analysis and modeling is the vital component of the analytics process. A method of assessment is selected based on the question to be answered. In defining the result, a robust metric carrying error information is also established. For instance, if one makes a choice of regression, then an evaluation of the cost of making an error must also be made. Thus analysis and modeling, which capture the inherent mathematical relationship between the desired output and the contributing

inputs provides the description on how a system behaves. Knowing this relationship then allows ways to better manage the system and also control the key influencing variables.

Exploring Employee Attrition data in R

Problem Statement

Employee churn is a major problem for many firms these days. Great talent is scarce, hard to keep and in high demand. Given the well-known direct relationship between happy employees and happy customers, it becomes of utmost importance to understand the drivers of employee dissatisfaction. In doing so, predictive analytics can be a core strategic tool to help facilitate employee engagement and set up well targeted employee retention campaigns.

Loading and understanding data

Download the data into local file system from [WA_Fn-UseC_-HR-Employee-Attrition](#). Using R, load the data into workspace. The data attributes are self-explanatory. Attrition is the dependent variable (outcome) which we want to determine as a function of several input variables.

Data exploration and visualization

- Identify the data types of each column and explore their statistical summaries.
- Code the categorical variables into numeric values
- Bin certain variables appropriately e.g., Age, MonthlyIncome
- Visualize and compare distributions of attrition across departments, performance rating and other categorical variables
 - Which departments are losing more employees than others?
 - Is the attrition higher amongst high performers or low performers?
 - Identify and plot other variables with respect to attrition
- Explore the correlation between numeric variables
 - Identify the variables which seem most correlated to employee attrition
 - Plot the correlations
 - Do some independent variables appear more correlated to each other than others

Fitting a logistic regression model

- Specify the null hypothesis for this model
- Run a logistic regression fit for Attrition against remaining variables
- Explain the following from fit summary

- AIC
- Intercept and coefficients
- Null and Residual deviance values
- Calculate the goodness of the model using log-likelihood function
- Try various combinations of encoding and scaling schemes to the data to improve the error score
- Finally, based on the results, should the null hypothesis be rejected

Applying different models

- Identify other models (Decision tree, Random Forest, etc) and run them with this data set
- Verify if the predictability improves or worsens with different models

Present your findings and conclusion

- Make a presentation based on your exploration, findings and research
- Use visualizations to bring out the story
- Explain your choice of the model used
- Share learnings and challenges

DATA SET ATTACHED. DATASET NAME → HR-Employee-Attrition.xls