BIG DATA ANALYTICS (ASSOCIATE ANALYTICS — II) [R15A0531]

LECTURE NOTES

B.TECH IV YEAR – I SEM(R15) (2019-20)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous Institution – UGC, Govt. of India)
Recognized under 2(f) and 12 (B) of UGC ACT 1956
(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

(R15A0531) BIG DATA ANALYTICS (ASSOCIATE ANALYTICS – II)

(Core Elective-III)

L T/P/D C
4 -/-/- 3

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 Preprocessing. 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, and Standardized reporting & compliances, Decision Models, course conclusion. Assessment.

Unit – III

Big Data Analytics: Run descriptive 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 observer 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 03, Draw insights out of Visualization tool. Product Implementation

TEXT BOOK

Student's Handbook for Associate Analytics.

REFERENCE BOOKS

- Introduction to Data Mining, Tan, Steinbach and Kumar, Addison Wesley, 2006

- Data Mining Analysis and Concepts, M. Zaki and W. Meira (the authors have kindly made an online version available):
- http:/iwww.dataminingbook. infoluploads/book.pdf
- Mining of Massive Datasets Jure Leskovec Stanford Univ. Anand RajaramanMilliway Labs Jeffrey D. Ullman Stanford Univ.
- (hftp:llwww.vistrai ls.orglindex.phplCourse:_Big_Data_Analysis)

INDEX

(R15A0531) BIG DATA ANALYTICS (ASSOCIATE ANALYTICS — II)
(Core 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 Preprocessing. 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.

Data Management:

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 Process

1 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.

2 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

3 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.

4 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.

5 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.

6 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.

7 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.

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

9 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

Maintain Healthy, Safe & Secure Working Environment

<u>Basic Workplace Safety Guidelines</u>

□ 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.

Report Accidents & Emergencies

Accidents and Emergencies

Discuss the definition of 'accidents and emergencies' and the events that fall in the category of accidents.

An accident is an unplanned, uncontrolled, or unforeseen event resulting in injury or harm to people and damages to goods. For example, a person falling down and getting injured or a glassware item that broke upon being knocked over. Emergency is a serious or crisis situation that needs immediate attention and action. For example, a customer having a heart attack or sudden outbreak of fire in your organization needs immediate attention.

Each organization or chain of organizations has procedures and practices to handle and report accidents and take care of emergencies. Although you will find most of these procedures and practices common across the industry, some procedures might be modified to fit a particular type of business within the industry. For example, procedure to handle accidents caused by slipping or falling will be similar across the industry. You need to be aware of the general procedures and practices as well as the ones specific to your organization.

The following are some of the guidelines for identifying and reporting an accident or emergency:

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.

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.

Slip and fall: People may lose foothold on the floor and stairs resulting in injuries. Slips are mainly due to wet floors. Other causes: spilling of liquids or throwing of other slip-causing material on floors, such fruit peels. Tripping and slipping is generally caused by negligence, which can be either from the side of organization employees or from the side of customers. It can also be due to broken or uneven walking surface, such as broken or loose floor tile. However, you should prevent any such negligence. In addition, people should be properly cautioned against tripping and slipping. For example, a "wet floor" sign will warn people to walk carefully on freshly

mopped floors. Similarly, "watch your steps" signs can prevent accidents on a staircase with a sharp bent or warn against a loose floor tile.

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.

Types of Emergencies

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.

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.

Handling General Emergencies

It is important to have policies and procedures to tackle the given categories of emergencies. You should be aware of at least the basic procedures to handle emergencies. The basic procedures that you should be aware of depend on the business of your organization. Typically, you should seek answers to the following questions to understand what basic emergency procedures that you should be aware of:

- What is the evacuation plan and procedure to follow in case of an emergency?

- Who all should you notify within the organization?

- Which external agencies, such as police or ambulance, you should notify in which emergency?

What all services and equipment should you shut down during which emergency? Here are some general emergency handling procedures that you can follow:

- Keep a list of numbers to call during emergency, such as those of police, fire brigade, security, ambulance etc. Ensure that these numbers are fed into the organizations telephone program and hard copies of the numbers are placed at strategic locations in the organization.

- Regularly check that all emergency handling equipment's are in working condition, such as the fire extinguisher and fire alarm system.

Ensure that emergency exits are not obstructed and keys to such exists are easily accessible. Never place any objects near the emergency doors or windows.

Summary

Identify and report accidents and emergencies:

- Notice and correctly identify accidents and emergencies.
- Get help promptly and in the most suitable way.
- Follow company policy and procedures for preventing further injury while waiting for help to arrive.
- Act within the limits of your responsibility and authority when accidents and emergencies arise.
- Promptly follow the instructions given by senior staff and the emergency services personnel.

Handling accidents:

- Attend the injured person immediately.
- Inform your supervisor about the accident giving details.
- Assist your supervisor in investigating and finding out the actual cause of the accident.

General emergency handling procedures:

- Keep a list of numbers to call during emergencies.
- Regularly check that all emergency handling equipment is in working condition.
- Ensure that emergency exits are not obstructed.

Protect Health & Safety

Hazards

What are hazards?

In relation to workplace safety and health, hazard can be defined as any source of potential harm or danger to someone or any adverse health effect produced under certain condition.

A hazard can harm an individual or an organization. For example, hazard to an organization include loss of property or equipment while hazard to an individual involve harm to health or body.

A variety of sources can be potential source of hazard at workplace. These hazards include practices or substances that may cause harm. Here are a few examples of potential hazards:

- Material: Knife or sharp edged nails can cause cuts.
- Substance: Chemicals such as Benzene can cause fume suffocation.
  Inflammable substances like petrol can cause fire.
- Electrical energy: Naked wires or electrodes can result in electric shocks.
- Condition: Wet floor can cause slippage. Working conditions in mines can cause health hazards.
- Gravitational energy: Objects falling on you can cause injury.
- Rotating or moving objects: Clothes entangled into ratting objects can cause serious harm.

Potential Sources of Hazards in an Organization

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 fight against all the given hazards related to working environment.

General Evacuation Procedures

Each organization will has 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.
- Guide your customers to the emergency exits.
- 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.
- 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.
- 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.
- 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.
- Do not go back to the building you have evacuated till you are informed by authorized personnel that it is safe to go inside.

Summary

- Hazards can be defined as any source of potential harm or danger to someone or any adverse health effect produced under certain condition.

- Some potential sources of hazards in an organization are as follows:
- Using computers
- Handling office equipment
- Handling objects
- Stress at work
- Working environment
- Every employee should be aware of evacuation procedures and follow them properly during an emergency evacuation.
- Follow all safety rules and warning to keep your workplace free from accidents.
- Recognize all safety signs in offices.
- Report any incidence of non-compliance to safety rules and anything that is a safety hazard.

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, and Standardized reporting & compliances, Decision Models, course conclusion. Assessment.

Big Data Tools

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

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 contra variance, 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. The name Scala is a portmanteau of "scalable" and "language", signifying that it is designed to grow with the demands of its users.

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:

The expectation-maximization algorithm

full information maximum likelihood estimation Interpolation

In the mathematical field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points.

Model-Based Techniques

Model based techniques, often using graphs, offer additional tools for testing missing data types (MCAR, MAR, MNAR) and for estimating parameters under missing data conditions. For example, a test for refuting MAR/MCAR reads as follows:

For any three variables X,Y, and Z where Z is fully observed and X and Y partially observed, the data should satisfy: .

In words, the observed portion of X should be independent on the missingness status of Y, conditional on every value of Z. Failure to satisfy this condition indicates that the problem belongs to the MNAR category.

When data falls into MNAR category techniques are available for consistently estimating parameters when certain conditions hold in the model.For example, if Y explains the reason for missingness in X and Y itself has missing values, the joint probability distribution of X and Y can still be estimated if the missingness of Y is random. The estimand in this case will be:

$$\begin{aligned} P(X,Y) &= P(X|Y)P(Y) \\ &= P(X|Y,R_x=0,R_y=0)P(Y|R_y=0) \end{aligned}$$

Where Rx=0 and Ry=0 denote the observed portions of their respective variables.

Different model structures may yield different estimand and different procedures of estimation whenever consistent estimation is possible. The preceding estimand calls for first estimating P(X/Y) from complete data and multiplying it by P(Y) estimated from cases in which Y is observed regardless of the status of X. Moreover, in order to obtain a consistent estimate it is crucial that the first term be P(X/Y) as opposed to P(Y/X).
In many cases model based techniques permit the model structure to undergo refutation tests. Any model which implies the independence between a partially observed variable X and the missingness indicator of another variable Y(i.e. Ry), conditional on can be submitted to the following refutation

R_x

test: $X \perp\!\!\!\perp R_y | R_x = 0$

Finally, the estimand that emerge from these techniques are derived in closed form and do not require iterative procedures such as Expectation Maximization that are susceptible to local optima.

Provide data/information in standard formats

Knowledge Management

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

Knowledge Management- Industrializing Collective Brain Power

- What is knowledge management?

  Knowledge management (KM) is the process of capturing, developing, sharing, and effectively using organizational **knowledge**. It refers to a multi-disciplinary approach to achieving organizational objectives by making the best use of **knowledge**.
- KM to support an organization’s growth engine:
  - 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
- Decrease time to innovate/deliver through faster knowledge distribution

Fig : Knowledge management needs to deal with a lot of knowledge items

Fig : Knowledge Management Approach

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.

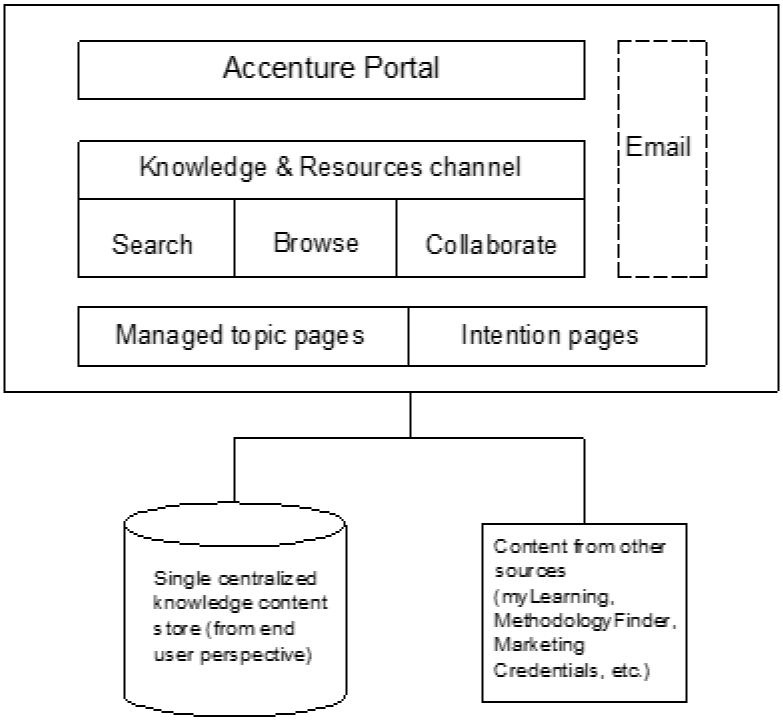

Accenture Portal
Email
Knowledge & Resources channel
Search
Browse
Collaborate
Managed topic pages
Intention pages
Single centralized knowledge content store (from end user perspective)
Content from other sources (myLearning, MethodologyFinder, Marketing Credentials, etc.)

Fig : Accenture Knowledge Management solution structure

Exhibit 3

USER REQUEST FLOW

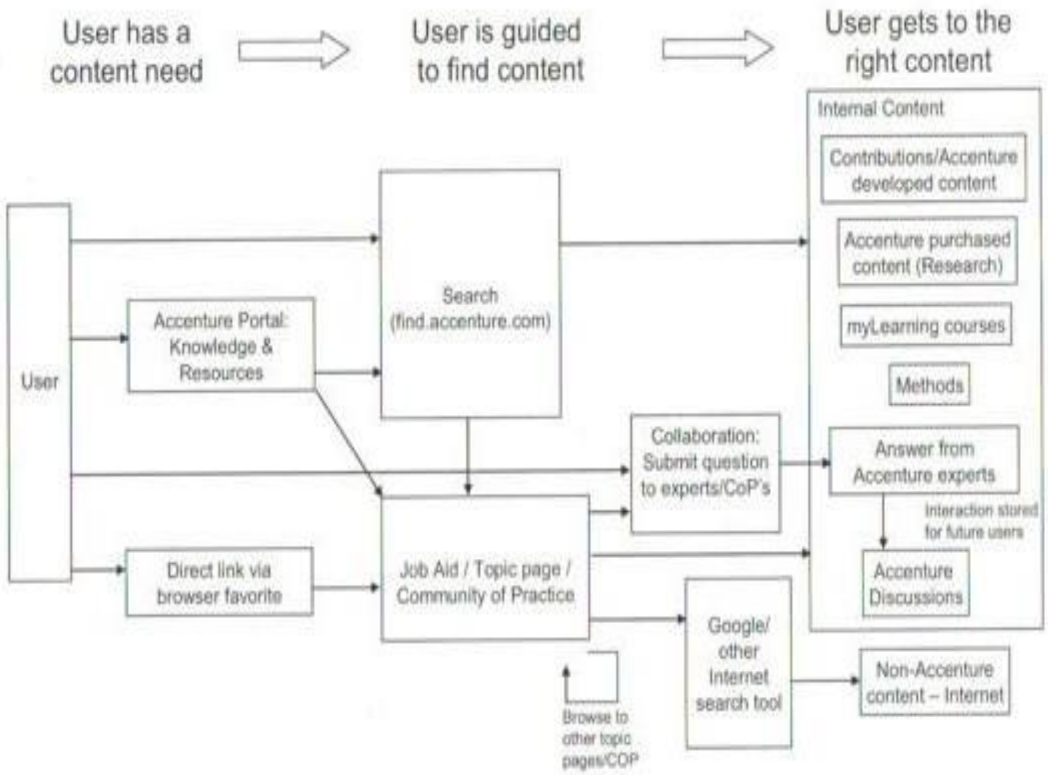

User has a content need
User is guided to find content
User gets to the right content
User
Accenture Portal: Knowledge & Resources
Direct link via browser favorite
Search (find.accenture.com)
Job Aid / Topic page / Community of Practice
Browse to other topic pages/COP
Collaboration: Submit question to experts/CoP's
Google/ other Internet search tool
Internal Content
Contributions/Accenture developed content
Accenture purchased content (Research)
myLearning courses
Methods
Answer from Accenture experts
Interaction stored for future users
Accenture Discussions
Non-Accenture content – Internet

Fig : Accenture Knowledge Management report extraction

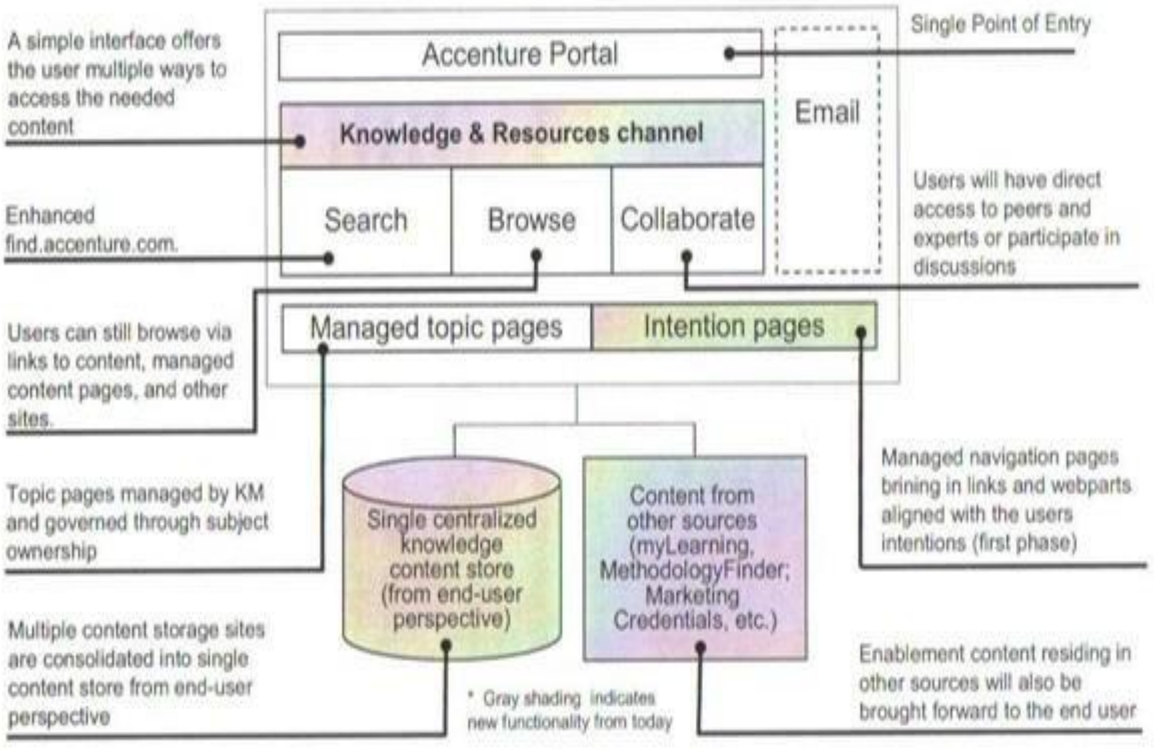

A simple interface offers the user multiple ways to access the needed content
Accenture Portal
Email
Single Point of Entry
Knowledge & Resources channel
Enhanced find.accenture.com.
Search
Browse
Collaborate
Users will have direct access to peers and experts or participate in discussions
Users can still browse via links to content, managed content pages, and other sites.
Managed topic pages
Intention pages
Topic pages managed by KM and governed through subject ownership
Single centralized knowledge content store (from end-user perspective)
Content from other sources (myLearning, MethodologyFinder; Marketing Credentials, etc.)
Managed navigation pages brining in links and webparts aligned with the users intentions (first phase)
Multiple content storage sites are consolidated into single content store from end-user perspective
* Gray shading indicates new functionality from today
Enablement content residing in other sources will also be brought forward to the end user

Fig : Accenture Knowledge Management Portal

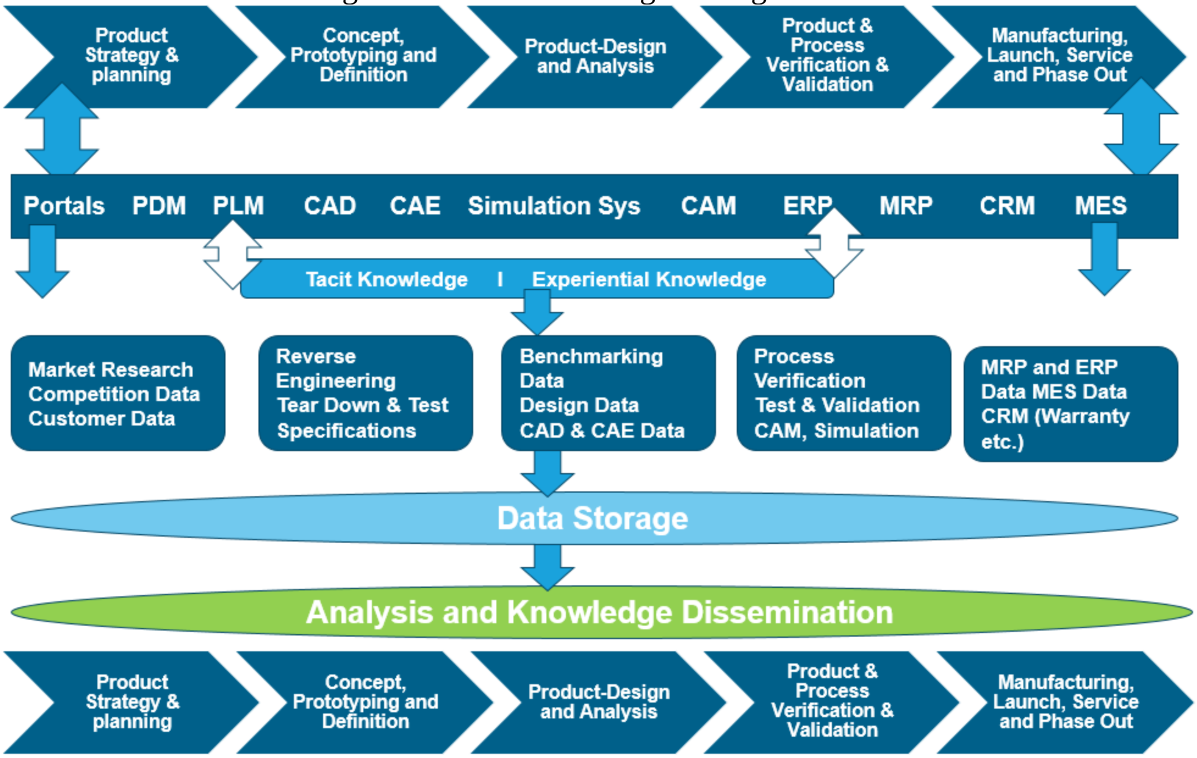

Product Strategy & planning
Concept, Prototyping and Definition
Product-Design and Analysis
Product & Process Verification & Validation
Manufacturing, Launch, Service and Phase Out
Portals
PDM
PLM
CAD
CAE
Simulation Sys
CAM
ERP
MRP
CRM
MES
Tacit Knowledge I Experiential Knowledge
Market Research Competition Data Customer Data
Reverse Engineering Tear Down & Test Specifications
Benchmarking Data Design Data CAD & CAE Data
Process Verification Test & Validation CAM, Simulation
MRP and ERP Data MES Data CRM (Warranty etc.)
Data Storage
Analysis and Knowledge Dissemination
Product Strategy & planning
Concept, Prototyping and Definition
Product-Design and Analysis
Product & Process Verification & Validation
Manufacturing, Launch, Service and Phase Out

Fig : TCS Enterprise View of Big Data Sources & Knowledge Management

Summary

- Every organization has its own knowledge management framework
- It is very important to adhere to a given KM framework and understand the same in as much details as possible.
- The KM framework of an organization is a very dynamic and evolving structure
- The KM framework is designed keeping in mind the important core functions of the organization

Standardized Reporting and compliance

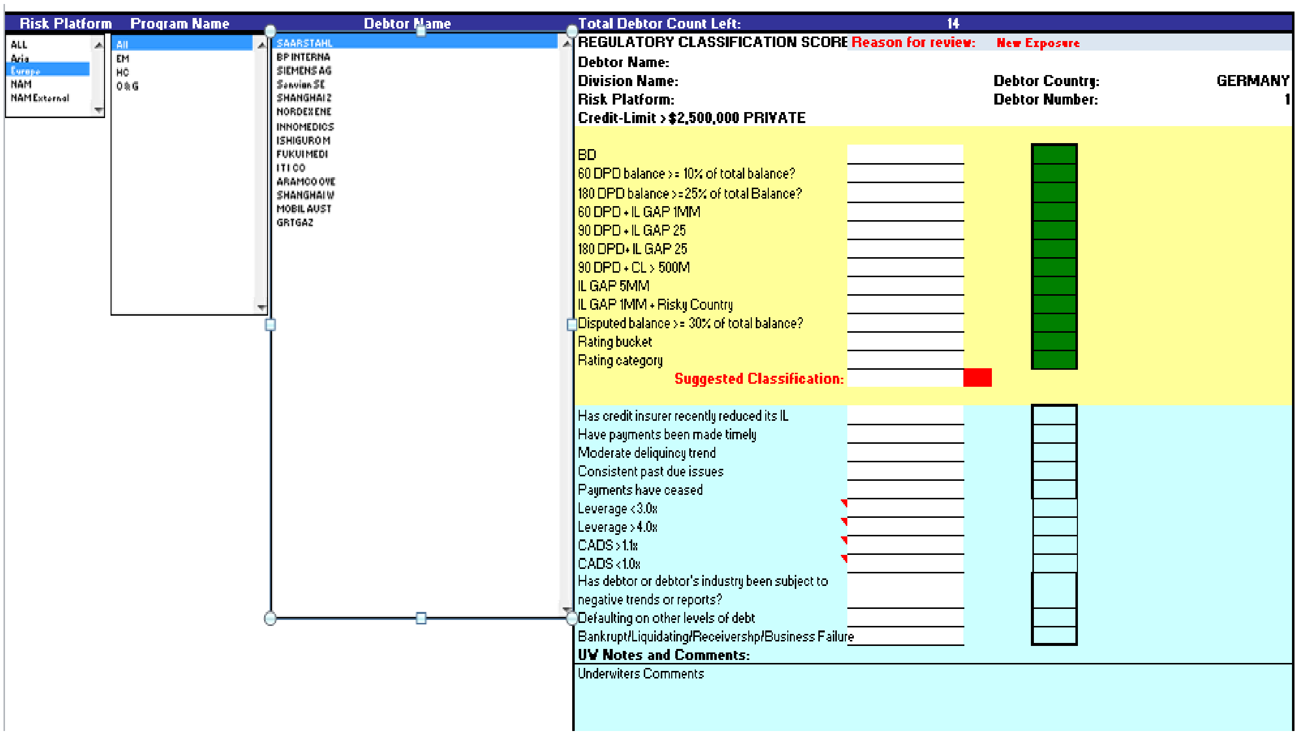
What are standard reporting templates

What are templates?

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



Total Debtor Count Left: 14

REGULATORY CLASSIFICATION SCORE Reason for review: New Exposure

Debtor Name:

Division Name: Debtor Country: GERMANY

Risk Platform: Debtor Number: 1

Credit-Limit > $2,500,000 PRIVATE

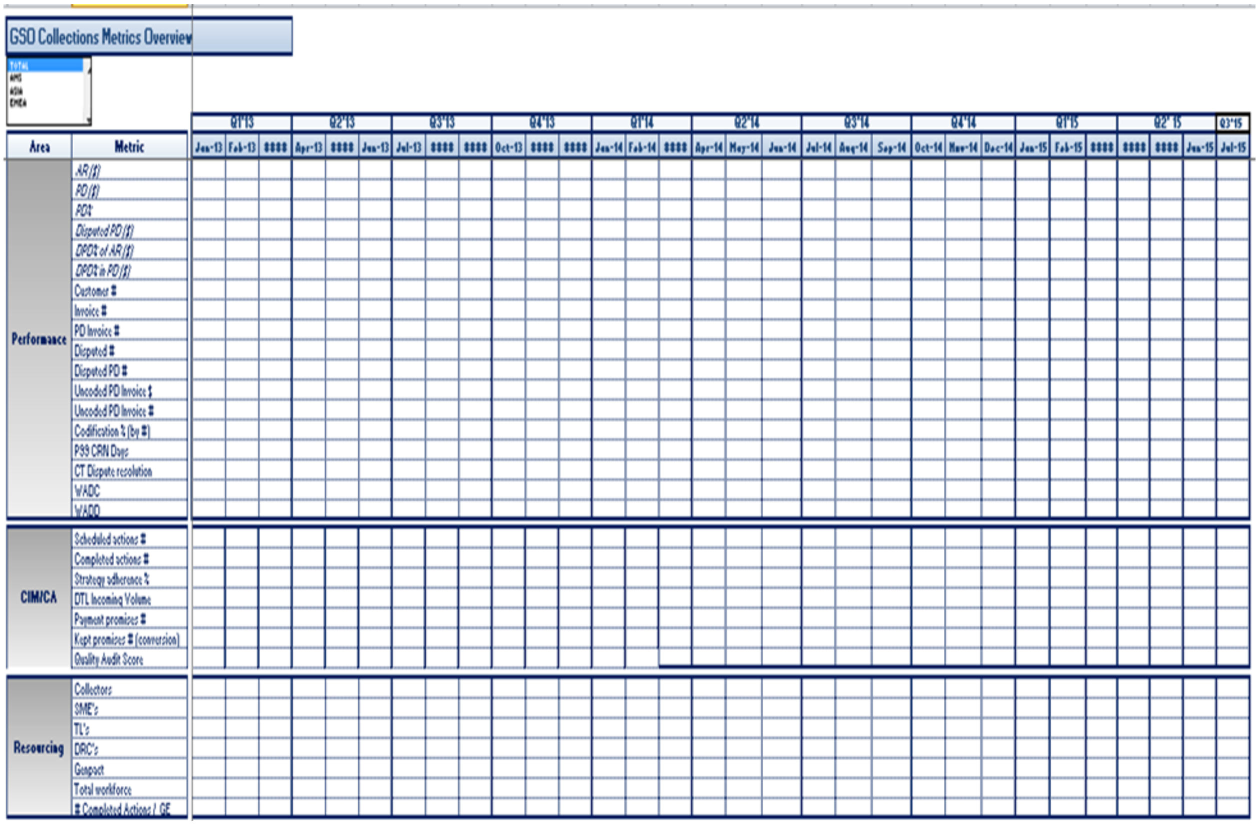
BD
60 DPD balance >= 10% of total balance?
180 DPD balance >=25% of total Balance?
60 DPD + IL GAP 1MM
90 DPD + IL GAP 25
180 DPD+ IL GAP 25
90 DPD + CL > 500M
IL GAP 5MM
IL GAP 1MM + Risky Country
Disputed balance >= 30% of total balance?
Rating bucket
Rating category
Suggested Classification:

Has credit insurer recently reduced its IL
Have payments been made timely
Moderate deliquincy trend
Consistent past due issues
Payments have ceased
Leverage <3.0x
Leverage >4.0x
CADS >1.1x
CADS <1.0x
Has debtor or debtor's industry been subject to negative trends or reports?
Defaulting on other levels of debt
Bankrupt/Liquidating/Receivershp/Business Failure

UW Notes and Comments:

Underwiters Comments

Fig : Reporting Templates Sample– Debtor Information & Score Card



GSO Collections Metrics Overview

TOTAL
AMS
ASIA
EMEA

		Q1'13			Q2'13			Q3'13			Q4'13			Q1'14			Q2'14			Q3'14			Q4'14			Q1'15			Q2' 15			Q3'15
Area	Metric	Jan-13	Feb-13	####	Apr-13	####	Jun-13	Jul-13	####	####	Oct-13	####	####	Jan-14	Feb-14	####	Apr-14	May-14	Jun-14	Jul-14	Aug-14	Sep-14	Oct-14	Nov-14	Dec-14	Jan-15	Feb-15	####	####	####	Jun-15	Jul-15
Performance	AR ($)																															
	PD ($)																															
	PD%																															
	Disputed PD ($)																															
	DPD% of AR ($)																															
	DPD% in PD ($)																															
	Customer #																															
	Invoice #																															
	PD Invoice #																															
	Disputed #																															
	Disputed PD #																															
	Uncoded PD Invoice $																															
	Uncoded PD Invoice #																															
	Codification % (by #)																															
	P99 CRN Days																															
	CT Dispute resolution																															
	WADC																															
	WADD																															
CIM/CA	Scheduled actions #																															
	Completed actions #																															
	Strategy adherence %																															
	DTL Incoming Volume																															
	Payment promises #																															
	Kept promises # (conversion)																															
	Quality Audit Score																															
Resourcing	Collectors																															
	SME's																															
	TL's																															
	DRC's																															
	Genpact																															
	Total workforce																															
	# Completed Actions / GE																															

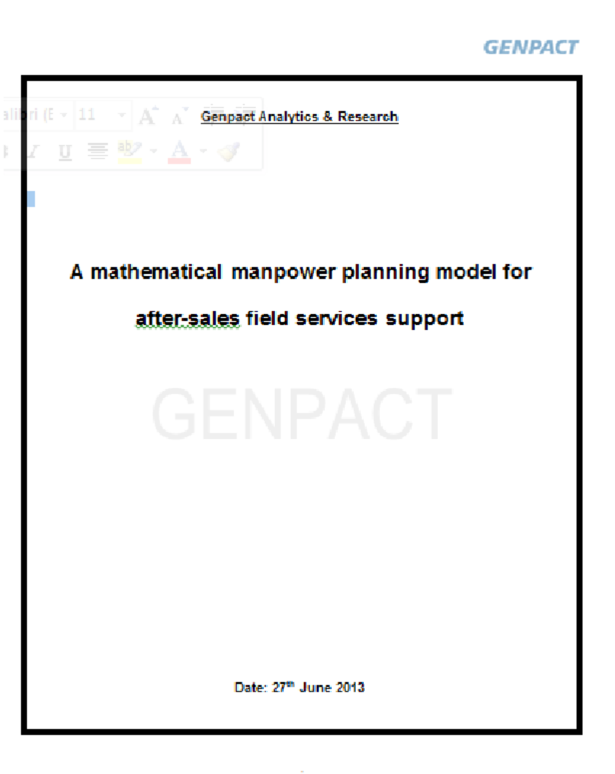
Fig : Reporting Templates Sample– Collection Metrics Template

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

Structure of a sample whitepaper


GENPACT

Genpact Analytics & Research

A mathematical manpower planning model for

after-sales field services support

GENPACT

Date: 27th June 2013

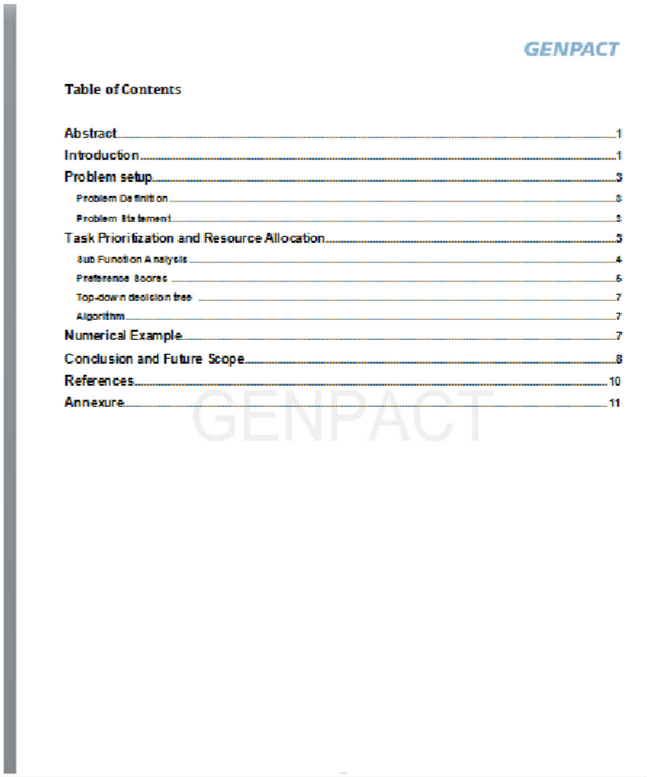
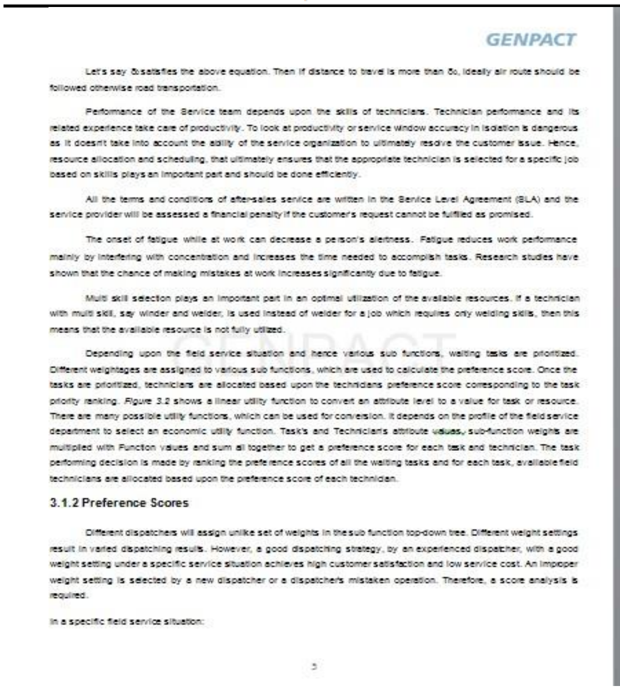

GENPACT

Table of Contents

GENPACT


GENPACT

Let's say δ satisfies the above equation. Then if distance to travel is more than δ₀, ideally air route should be followed otherwise road transportation.

Performance of the Service team depends upon the skills of technicians. Technician performance and its related experience take care of productivity. To look at productivity or service window accuracy in isolation is dangerous as it doesn't take into account the ability of the service organization to ultimately resolve the customer issue. Hence, resource allocation and scheduling, that ultimately ensures that the appropriate technician is selected for a specific job based on skills plays an important part and should be done efficiently.

All the terms and conditions of after-sales service are written in the Service Level Agreement (SLA) and the service provider will be assessed a financial penalty if the customer's request cannot be fulfilled as promised.

The onset of fatigue while at work can decrease a person's alertness. Fatigue reduces work performance mainly by interfering with concentration and increases the time needed to accomplish tasks. Research studies have shown that the chance of making mistakes at work increases significantly due to fatigue.

Multi skill selection plays an important part in an optimal utilization of the available resources. If a technician with multi skill, say winder and welder, is used instead of welder for a job which requires only welding skills, then this means that the available resource is not fully utilized.

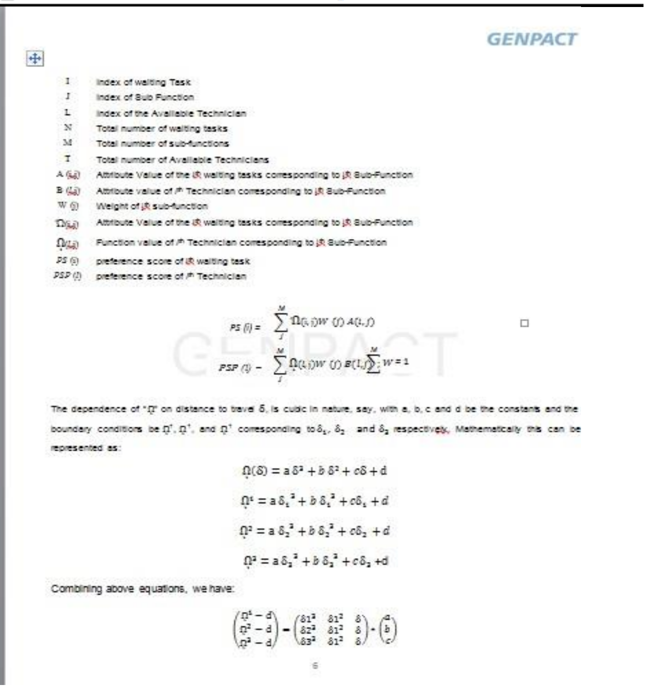
Depending upon the field service situation and hence various sub functions, waiting tasks are prioritized. Different weightages are assigned to various sub functions, which are used to calculate the preference score. Once the tasks are prioritized, technicians are allocated based upon the technicians preference score corresponding to the task priority ranking. *Figure 3.2* shows a linear utility function to convert an attribute level to a value for task or resource. There are many possible utility functions, which can be used for conversion. It depends on the profile of the field service department to select an economic utility function. Task's and Technician's attribute values, sub-function weights are multiplied with Function values and sum all together to get a preference score for each task and technician. The task performing decision is made by ranking the preference scores of all the waiting tasks and for each task, available field technicians are allocated based upon the preference score of each technician.

3.1.2 Preference Scores

Different dispatchers will assign unlike set of weights in the sub function top-down tree. Different weight settings result in varied dispatching results. However, a good dispatching strategy, by an experienced dispatcher, with a good weight setting under a specific service situation achieves high customer satisfaction and low service cost. An improper weight setting is selected by a new dispatcher or a dispatcher's mistaken operation. Therefore, a score analysis is required.

In a specific field service situation:

5


GENPACT

I	Index of waiting Task
J	Index of Sub Function
L	Index of the Available Technician
N	Total number of waiting tasks
M	Total number of sub-functions
T	Total number of Available Technicians
A (i,j)	Attribute Value of the ith waiting tasks corresponding to jth Sub-Function
B (l,j)	Attribute value of lth Technician corresponding to jth Sub-Function
W (j)	Weight of jth sub-function
Ω(i,j)	Attribute Value of the ith waiting tasks corresponding to jth Sub-Function
Ω(l,j)	Function value of lth Technician corresponding to jth Sub-Function
PS (i)	preference score of ith waiting task
PSP (l)	preference score of lth Technician

$$PS(i) = \sum_{j}^{M} \Omega(i,j) W(j) A(i,j)$$ □

$$PSP(l) = \sum_{j}^{M} \Omega(l,j) W(j) B(l,j) \sum_{j}^{M} W = 1$$

The dependence of "Ω" on distance to travel δ, is cubic in nature, say, with a, b, c and d be the constants and the boundary conditions be Ω¹, Ω¹, and Ω¹ corresponding to δ_1, δ_2 and δ_3 respectively. Mathematically this can be represented as:

$$\Omega(\delta) = a\delta^3 + b\delta^2 + c\delta + d$$

$$\Omega^1 = a\delta_1^3 + b\delta_1^2 + c\delta_1 + d$$

$$\Omega^2 = a\delta_2^3 + b\delta_2^2 + c\delta_2 + d$$

$$\Omega^3 = a\delta_3^3 + b\delta_3^2 + c\delta_3 + d$$

Combining above equations, we have:

$$\begin{pmatrix} \Omega^1 - d \\ \Omega^2 - d \\ \Omega^3 - d \end{pmatrix} = \begin{pmatrix} \delta 1^3 & \delta 1^2 & \delta \\ \delta 2^3 & \delta 1^2 & \delta \\ \delta 3^3 & \delta 1^2 & \delta \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

6

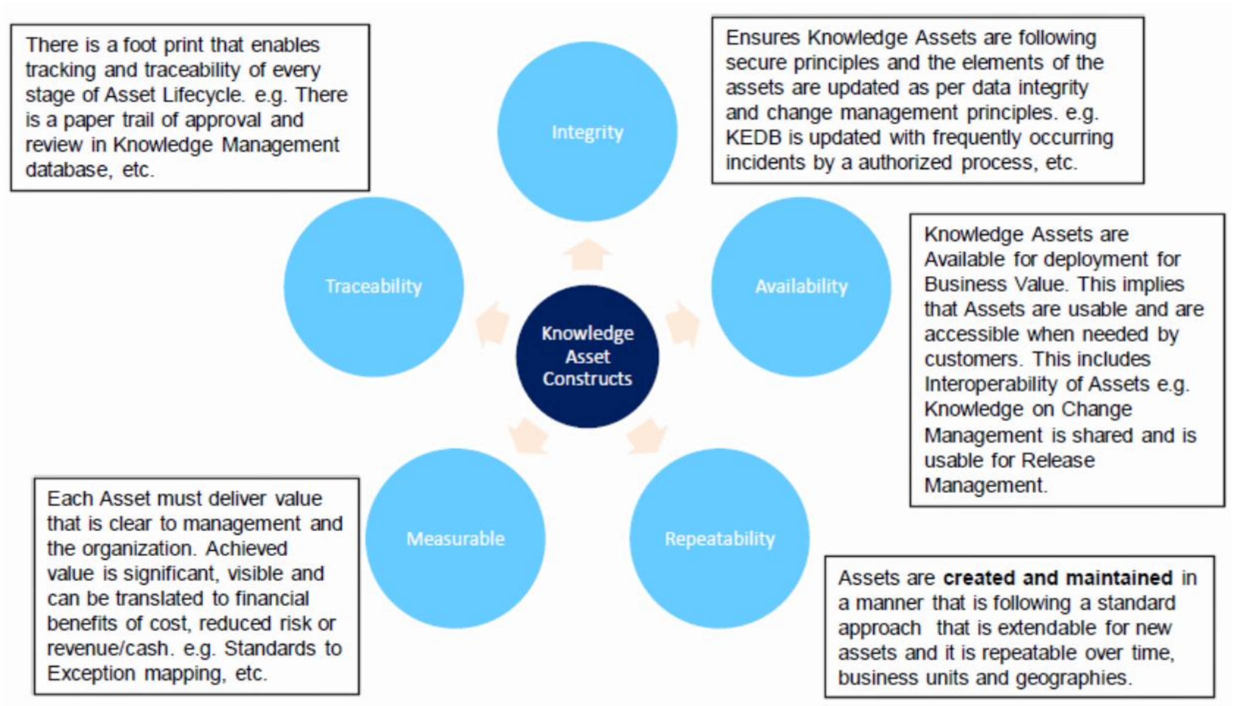
Organizing data/information

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

Policies and procedures for recording and sharing information



There is a foot print that enables tracking and traceability of every stage of Asset Lifecycle. e.g. There is a paper trail of approval and review in Knowledge Management database, etc.

Ensures Knowledge Assets are following secure principles and the elements of the assets are updated as per data integrity and change management principles. e.g. KEDB is updated with frequently occurring incidents by a authorized process, etc.

Knowledge Assets are Available for deployment for Business Value. This implies that Assets are usable and are accessible when needed by customers. This includes Interoperability of Assets e.g. Knowledge on Change Management is shared and is usable for Release Management.

Each Asset must deliver value that is clear to management and the organization. Achieved value is significant, visible and can be translated to financial benefits of cost, reduced risk or revenue/cash. e.g. Standards to Exception mapping, etc.

Assets are **created and maintained** in a manner that is following a standard approach that is extendable for new assets and it is repeatable over time, business units and geographies.

Importance of Compliance

What is Compliance?

In general, compliance means conforming to a rule, such as a specification, policy, standard or law. Regulatory compliance describes the goal that organizations aspire to achieve in their efforts to ensure that they are aware of and take steps to comply with relevant laws and regulations.

Why is compliance important?

- An effective compliance program can reduce many of the company's greatest risks, reduce the severity of claims and penalties when violations of law occur despite the program, and enhance company performance and profitability.
- When it comes to information technology and security, regulatory compliance for IT can impose added costs on
company operations depending upon the industry.

- At the same token, the cost of not complying with regulations both internally and externally can be significantly higher in terms of fines and time invested following up on a security breach.

• One of the primary issues with compliance is information security and the
potential for data leaks. Although there may be policies in place, it is necessary to ensure that employees follow t
e policies as well as the entire staff within a company.

• This is an ongoing process and one that can lead to a high profile data breach if companies become too lax on policy enforcement.

Summary

- Standard reporting templates help in saving a lot of time and efforts in getting standardized reports out.
- Reporting templates are pre-created structures based on which reports are to be created.
- 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 very important to comply with organizational norms for reporting and documentation to avoid any internal or external risk scenarios.

Decision Models

What is a decision model?

The Decision Model is an intellectual template for perceiving, organizing, and managing the business logic behind a business decision. An informal definition of business logic is it is a set of business rules represented as atomic elements of conditions leading to conclusions.

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.

There are many different decision models, and we are going to study a few of them here

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.
Subordinate Commitment – how important is it that your team and others buy into the decision? When teammates need to embrace the decision you should increase the participation levels.
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:

Style:
Autocratic – you make the decision and inform others of it.
There are two separate processes for decision making in an autocratic style:
Processes:
Autocratic 1(A1) – you use the information you already have and make the decision
Autocratic 2 (A2) – you ask team members for specific information and once you have it, you make the decision. Here you don't necessarily tell them what the information is needed for.
Style:
Consultative – you gather information from the team and other and then make the decision.
Processes:
Consultative 1 (C1) – you inform team members of what you're doing and may individually ask opinions, however, the group is not brought together for discussion. You make the decision.
Consultative 2 (C2) – you are responsible for making the decision, however, you get together as a group to discuss the situation, hear other perspectives, and solicit suggestions.

Style:
Collaborative – you and your team work together to reach a consensus.
Process:
Group (G2) – The team makes a decision together. Your role is mostly facilitative and you help the team come to a final decision that everyone agrees on.

The Kepner-Tregoe Matrix

Origin : The Kepner-Tregoe Matrix provides an efficient, systematic framework for gathering, organizing and evaluating decision making information. The approach was developed by Charles H. Kepner and Benjamin B. Tregoe in the 1960's and they first wrote about it in the business classic, The Rational Manager (1965). The approach is well-respected and used by many of the world's top organizations including NASA and General Motors.
The Kepner-Tregoe Approach
The Kepner-Tregoe approach is based on the premise that the end goal of any decision is to

make the "best possible" choice. This is a critical distinction: The goal is not to make the perfect choice, or the choice that has no defects. So the decision maker must accept some risk. And an important feature of the Kepner- Tregoe Matrix is to help evaluate and mitigate the risks of your decision.

The Kepner-Tregoe Matrix approach guides you through the process of setting objectives, exploring and prioritizing alternatives, exploring the strengths and weaknesses of the top alternatives, and of choosing the final "best" alternative. It then prompts you to generate ways to control the potential problems that will crop up as a consequence of your decision.

This type of detailed problem and risk analysis helps you to make an unbiased decision. By skipping this analysis and relying on gut instinct, your evaluation will be influenced by your preconceived beliefs and prior experience – it's simply human nature. The structure of the Kepner-Tregoe approach limits these conscious and unconscious biases as much as possible.

The Kepner-Tregoe Matrix comprises four basic steps:

1. Situation Appraisal – identify concerns and outline the priorities.

2. Problem Analysis – describe the exact problem or issue by identifying and evaluating the causes.
3. Decision Analysis – identify and evaluate alternatives by performing a risk analysis for each and then make a final decision.
4. Potential Problem Analysis – evaluate the final decision for risk and identify the contingencies and preventive actions necessary to minimize that risk.

Going through each stage of this process will help you come to the "best possible choice", given your knowledge and understanding of the issues that bear on the decision.

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.

Understanding the Tool

Called the OODA Loop, the model outlines a four-point decision loop that supports quick, effective and 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

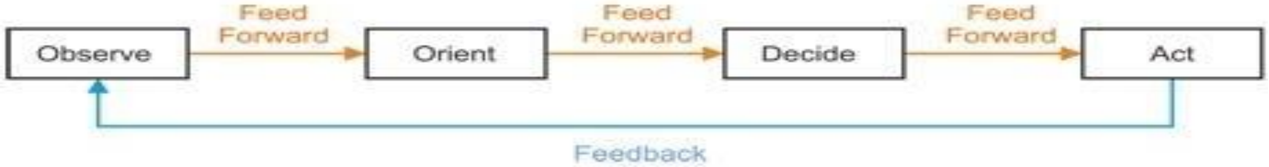


Figure 1 below shows the OODA Loop sequence:

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 . 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 incoming data as possible. The kind of questions you need to be asking are:

- What's happening in the environment that directly affects me?

- What's happening that indirectly affects me?

- What's happening that may have residual affects later on?

- Were my predictions accurate?

- Are there any areas where prediction and reality differ significantly?

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 and subsequent actions – essentially, you're learning as you continue to cycle through the steps. The results of your learning are brought in during the Orient phase, which in turn influences the rest of the decision making process.

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.

Summary

- Making good decisions is one of the main leadership tasks. Part of doing this is determining the most efficient and effective means of reaching the decision.
- Various decision models exist to aid in taking these decisions.
- The Vroom-Yetton-Jago decision model is a useful model, but it's quite complex and long-winded. Use it in new situations, or in ones which have unusual characteristics: Using it, you'll quickly get an feel for the right approach to use in more usual circumstances.
- The Kepner-Tregoe Matrix will help you come to the "best possible choice", given your knowledge and understanding of the issues that bear on the decision

Unit – III

Big Data Analytics: Run descriptive 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 observer the data ranges, Outlier detection and elimination.

Introduction to Big Data Analytics

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…

However, important limitations have appeared over the last 10 years, and providers of online services were the first who had to address these limitations.

From relational databases to Big Data

In particular they had to face five major weaknesses of relational databases:

the scaling of treatment

the scaling of data,

the redundancy

the **velocity**

the **variety and complexity**

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.

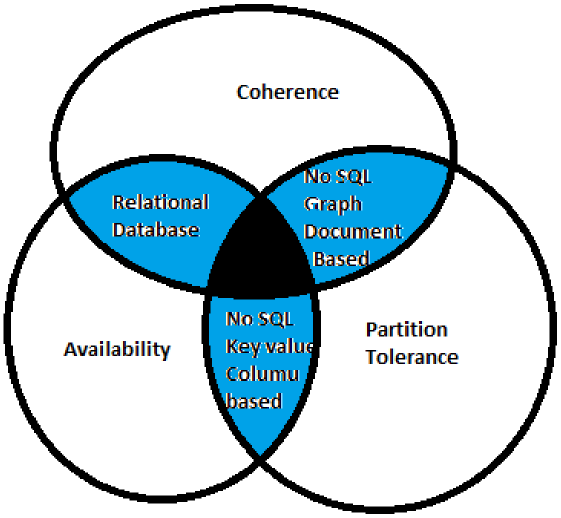
As a transactional database, they must respect the ACID constraints, i.e. the **Atomicity** of updates, the **Consistency** of the database, the **Isolation** and the **Durability** of queries. These constraints are perfectly applicable in a centralized architecture, but much more complex to insure in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

Partition Tolerance: each subnet-works must be autonomous



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:

- Key Value Store
- Column oriented Database
- Document Store
- Graph Database

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. However, some of them such as HBase give the possibility to index the values of the columns so that is also possible to query the database from the content of the columns.

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

Data preprocessing 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 (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.

Hypothesis testing and Determining the multiple analytical methodologies

What is Machine Learning :-

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.
- Between supervised and unsupervised learning is semi-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.

Summary

- 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.
- Three types of Machine Learning – Supervised, Unsupervised and Reinforced
- Machine learning focuses on prediction, based on known properties learned from the training data.

Unit – V

(NOS 9004) Data Visualization (NOS 2101): Prepare the data for Visualization, Use tools like Tableau, QlickView and 03, Draw insights out of Visualization tool. Product Implementation

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 Tablue

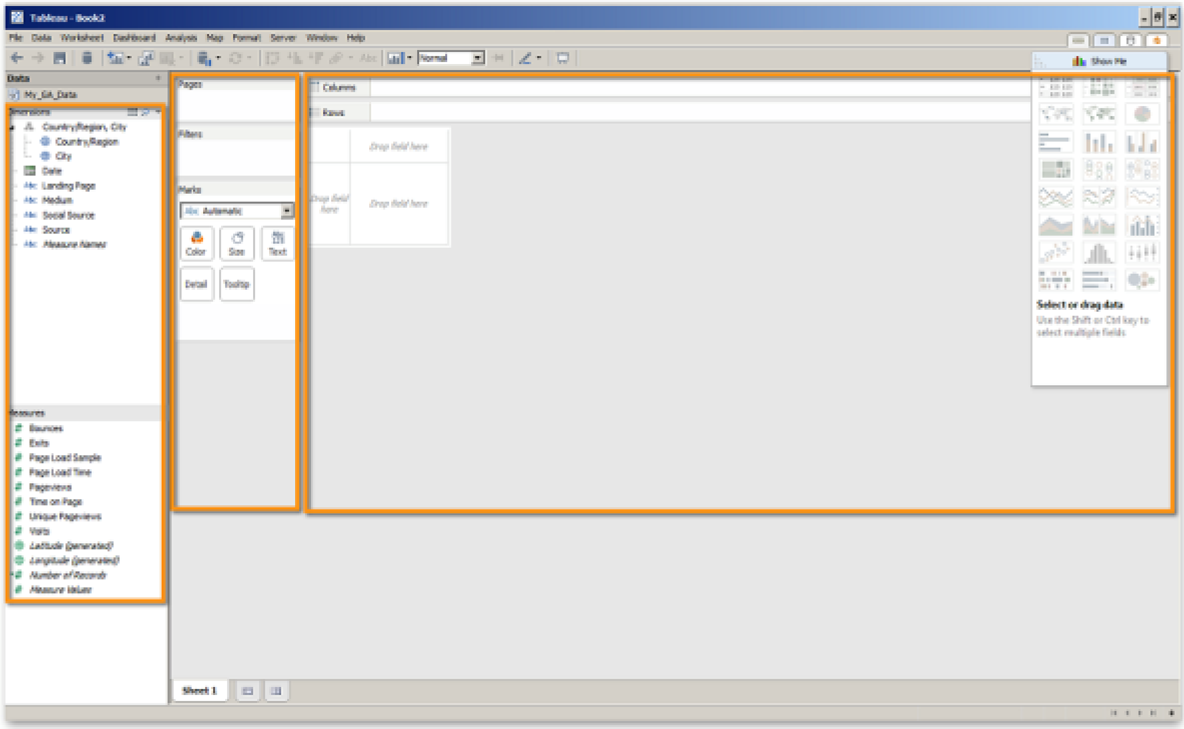
Extract The Data.

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.

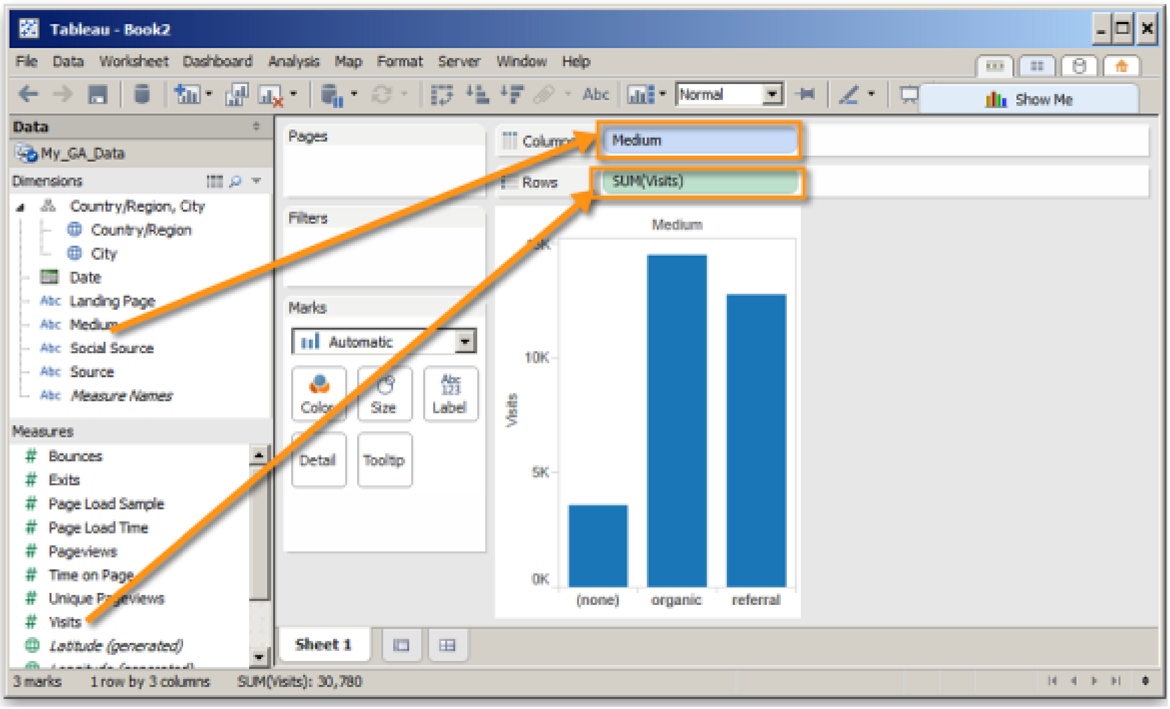
Keep in mind that the more dimensions you add, the larger the data set will get. For example, adding a device type will give you one row of measures for each device. You can think of it this way: if your default data has 10,000 rows, and you add the hour dimension, you would have 10,000*24 (hours). So,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

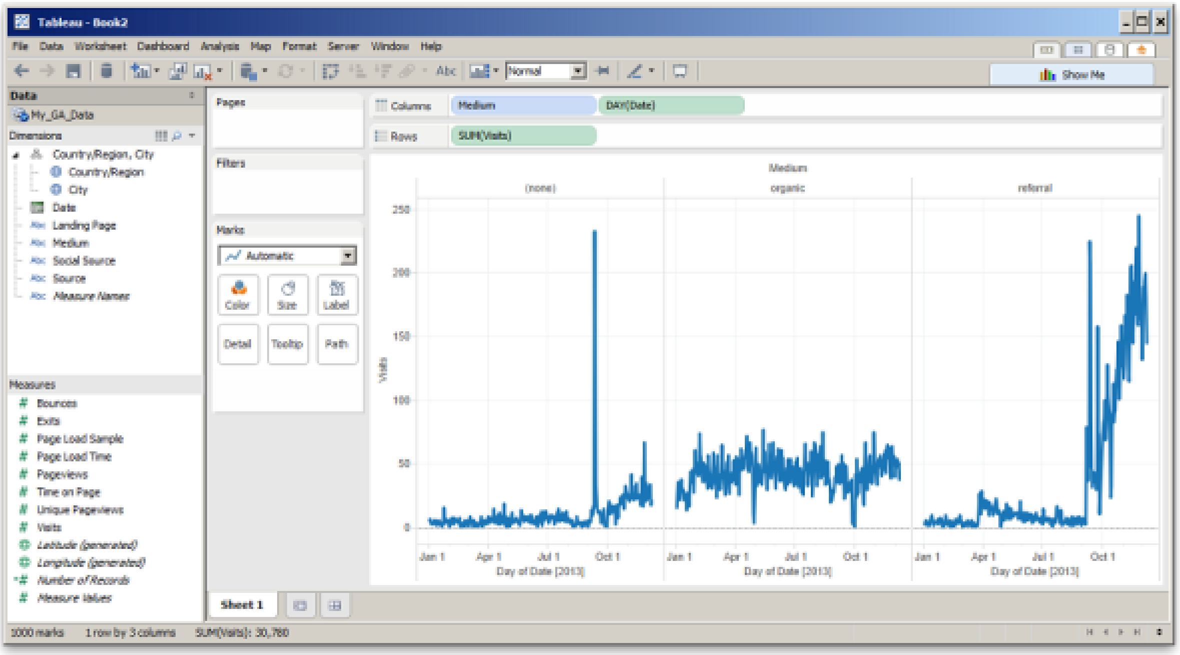


Your First Data Visualization. For our first effort, let's say we want to see what the traffic by medium looks like. To accomplish this, all you need to do is drag and drop icons from the dimensions and measures sections over to the columns and rows spots at the top. Specifically, drag the mediums icon to columns, and drag the visits icon to rows.



Now, how can we utilize this to take things to the next level? Let's look at historic performance. To do so, drag the date icon into the end of the column line. This will show us the performance by medium by year.

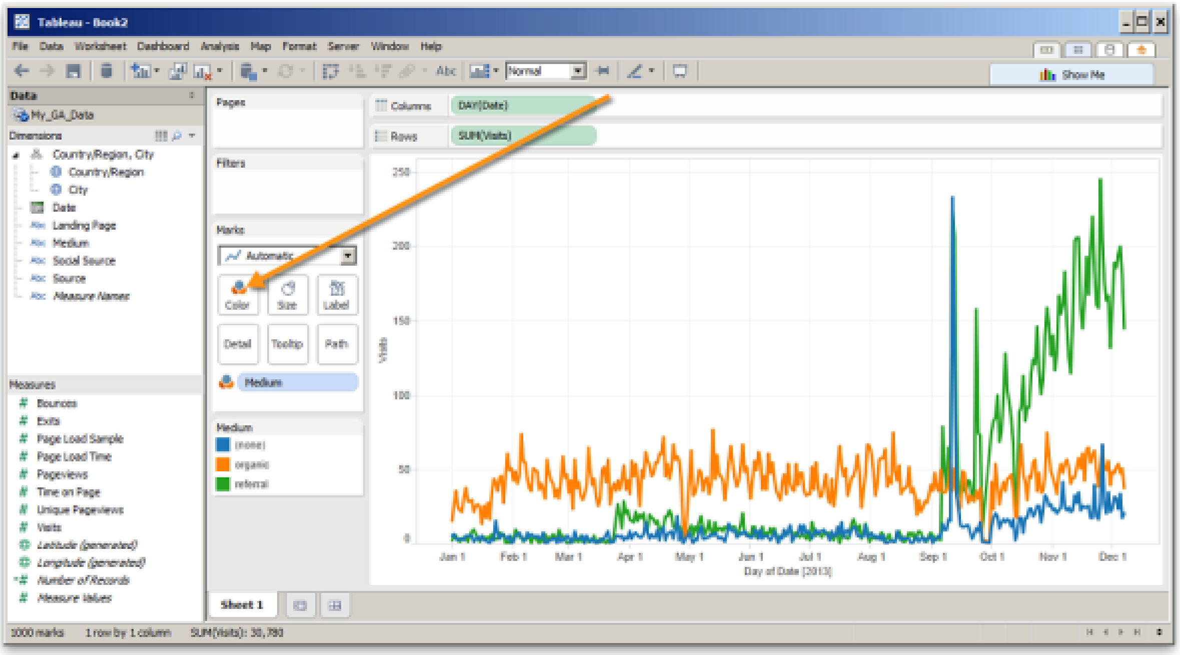
But since we only pulled 2013 data, the result is kind of boring. However, if you switch the dropdown in the date menu to month (or day) instead of year, you'll see that things get more interesting. You will have three line charts on the same axis comparing visits side by side.



Columns: Medium, DAY(Date)
Rows: SUM(Visits)
Marks: Automatic

1000 marks 1 row by 3 columns SUM(Visits): 30,780

But here is where the real power of DVTs comes into play. By simply dragging the medium from the columns area to the color area, you are instantly removing the columns for medium, combining it into one chart area, and then coloring by medium. This allows you to easily compare the data in a more visually interesting way.

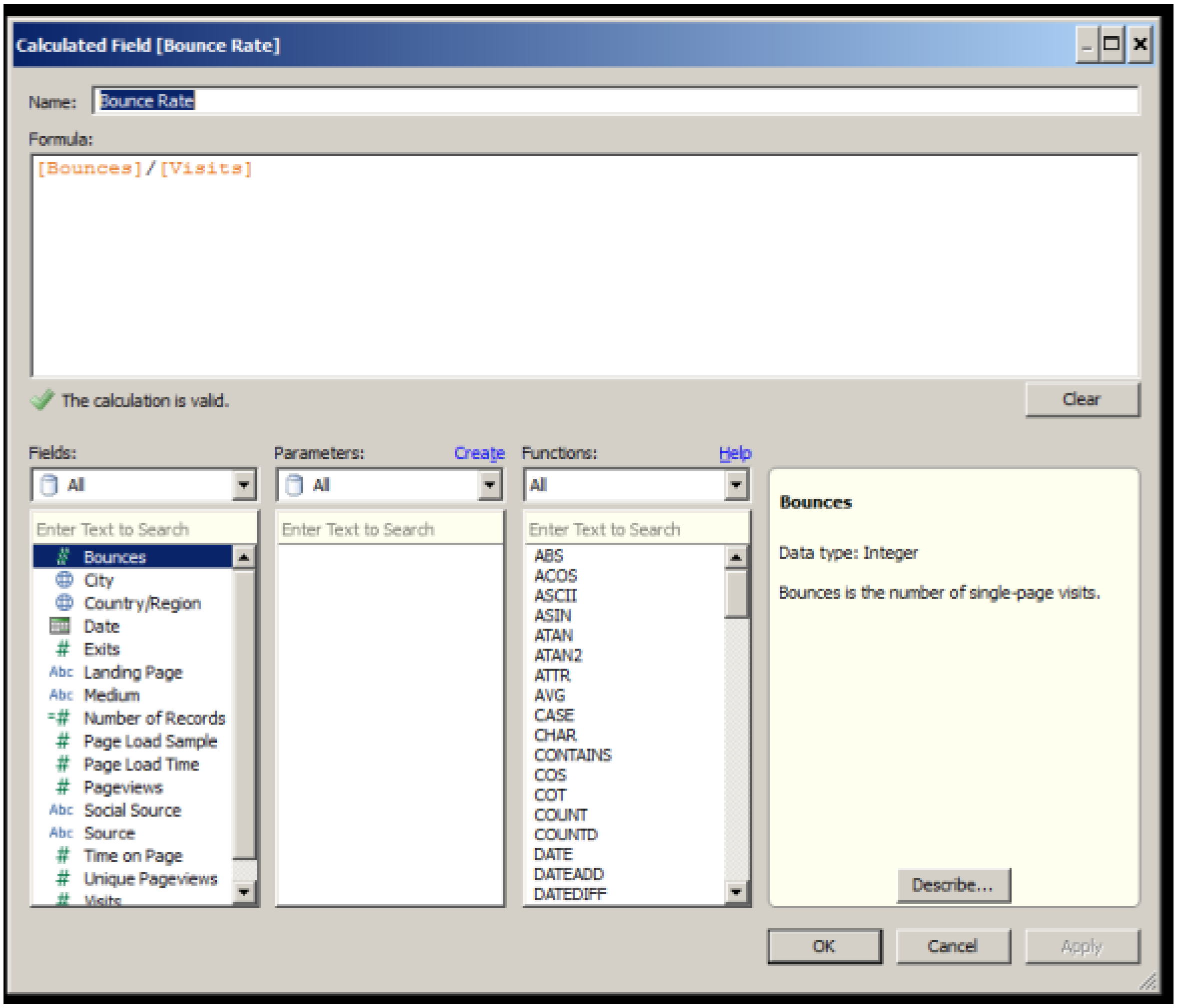


Columns: DAY(Date)
Rows: SUM(Visits)
Marks: Automatic
Color: Medium

1000 marks 1 row by 1 column SUM(Visits): 30,780

6. Enhancing Your Data. One of the differences between working in GA and working with raw data is that we still need to do some aggregation. For example, the raw data from GA includes the number of bounces and number of visits; however, it does not provide a bounce rate. Fortunately, that's not a problem for Tableau. This tool has a very powerful "calculated field" functionality that can be leveraged for either measures or dimensions.

For instance, let's say we want to calculate the bounce rate. Simply right-click in the measures area and select calculated fields. Then we would enter [Bounces]/[Visits].



Calculated Field [Bounce Rate]

Name: Bounce Rate

Formula:

```
[Bounces]/[Visits]
```

The calculation is valid.

Bounces

Data type: Integer

Bounces is the number of single-page visits.

The same approach can be used to do a variety of calculations. For instance, the code below could be used to distinguish weekend vs. weekday traffic:

```
if(DATEPart('weekday',[Date])=1) then 'Weekend' //Sunday
elseif(DATEPart('weekday',[Date])=2) then 'Weekday' //Monday
```

elseif(DATEPart('weekday',[Date])=3) then 'Weekday' //Tuesday

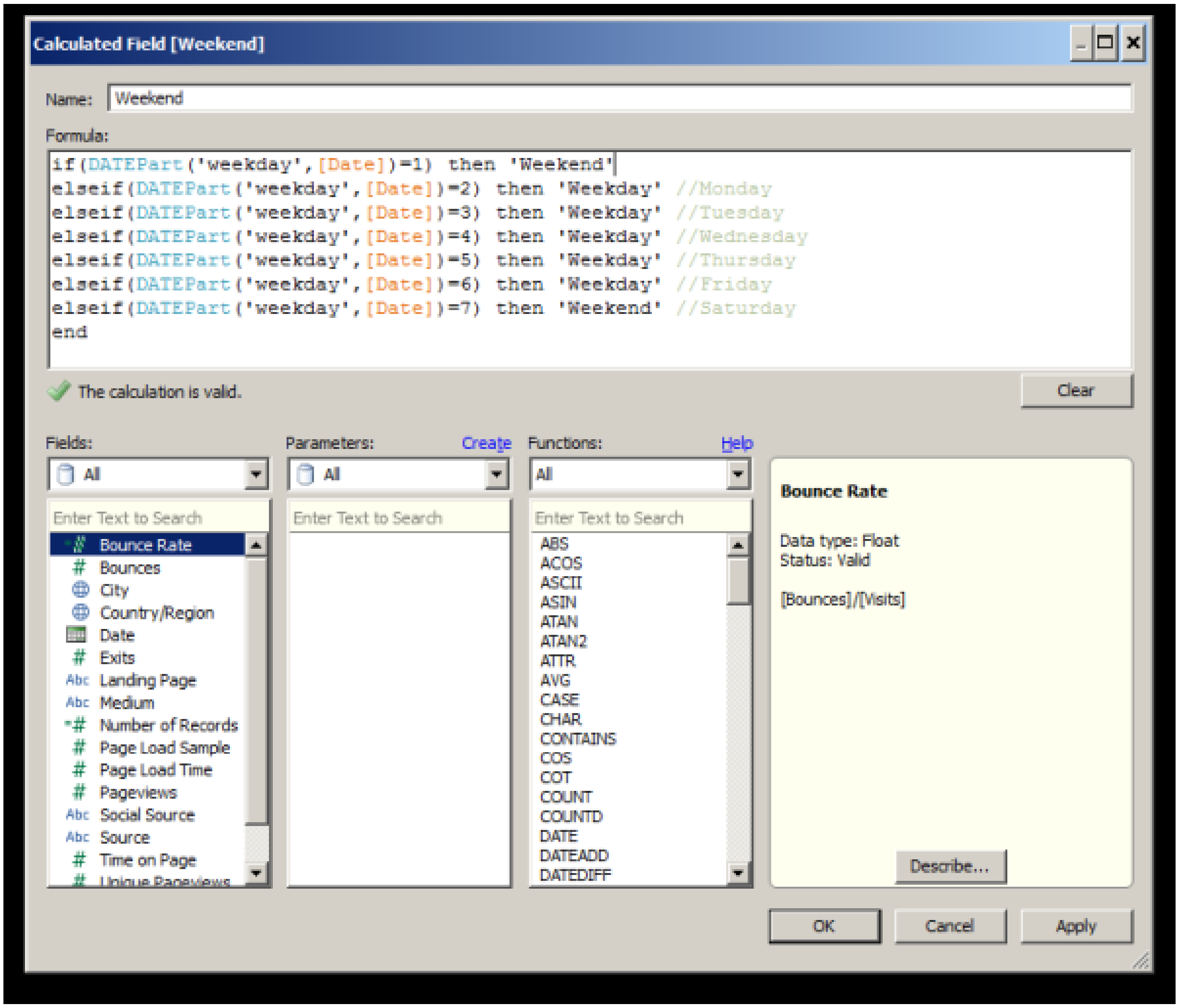
elseif(DATEPart('weekday',[Date])=4) then 'Weekday' //Wednesday

elseif(DATEPart('weekday',[Date])=5) then 'Weekday' //Thursday

elseif(DATEPart('weekday',[Date])=6) then 'Weekday' //Friday

elseif(DATEPart('weekday',[Date])=7) then 'Weekend' //Saturday

end



Calculated Field [Weekend]

Name: Weekend

Formula:

```
if(DATEPart('weekday',[Date])=1) then 'Weekend'
elseif(DATEPart('weekday',[Date])=2) then 'Weekday' //Monday
elseif(DATEPart('weekday',[Date])=3) then 'Weekday' //Tuesday
elseif(DATEPart('weekday',[Date])=4) then 'Weekday' //Wednesday
elseif(DATEPart('weekday',[Date])=5) then 'Weekday' //Thursday
elseif(DATEPart('weekday',[Date])=6) then 'Weekday' //Friday
elseif(DATEPart('weekday',[Date])=7) then 'Weekend' //Saturday
end
```

The calculation is valid.

Bounce Rate

Data type: Float
Status: Valid

[Bounces]/[Visits]

The above will give you a new dimension that allows you to separate your traffic by weekend and weekday. Overall, you can find some pretty interesting stories from similar behavioral segmentation. We have seen a lot of beauty brands show very distinctive behavior in terms of daytime parts.

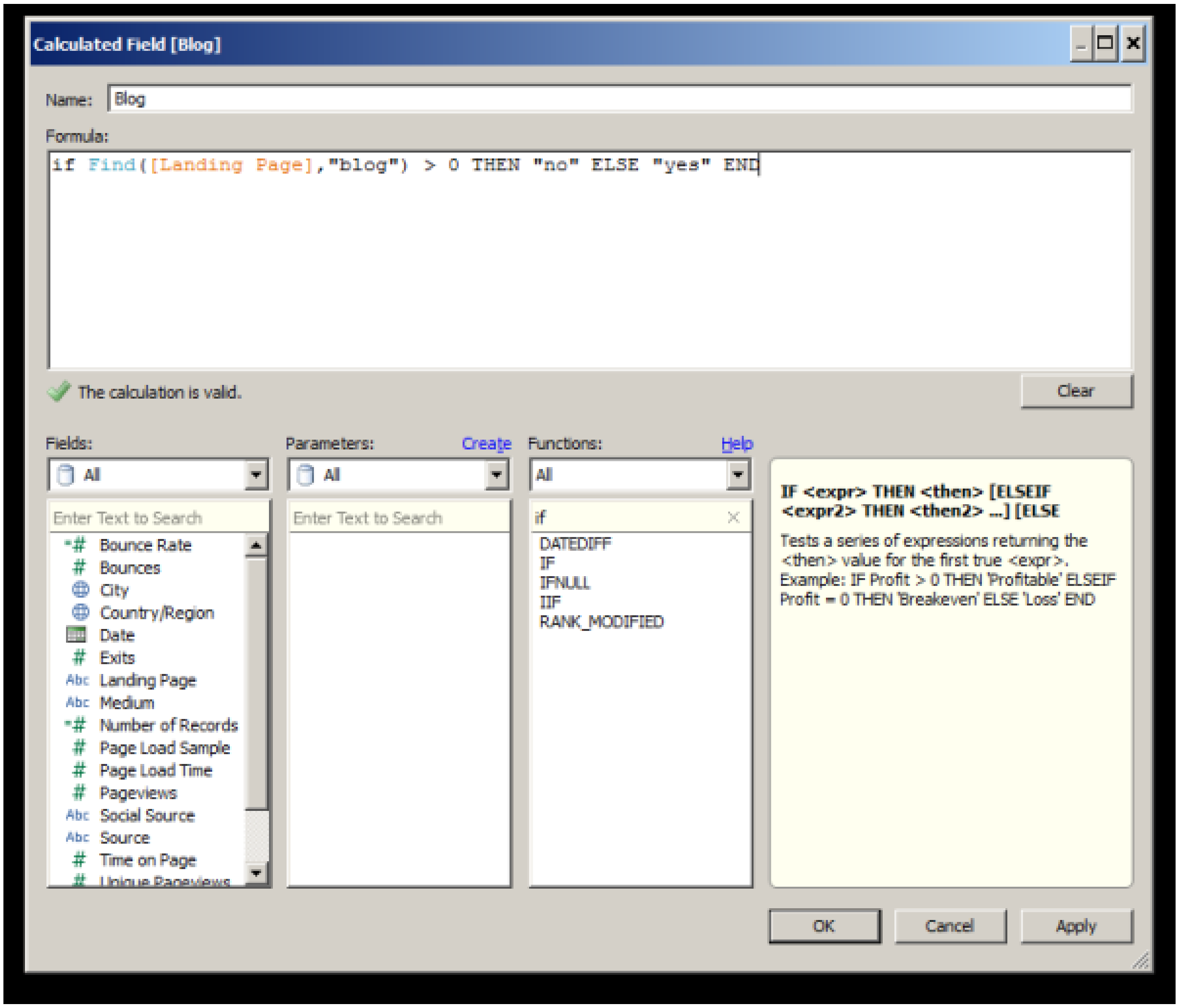
Another great way to look at the data is to filter down to only social source traffic, and then look at site engagement (PPV or time on-site). This will show you when your social traffic is performing best and when you can get the most ROI out of your social activities.

7. Segmenting Your Data. DVTs are also very effective in data segmentation (which is a big passion of mine). So let's look at our sample data. The website I am using in this sample has a blog section. Generally, user behavior on blog pages differs dramatically from that on general brand/product pages. (Blog visits via search are generally one page and extended time on-page.) Therefore, we really do not want to judge engagement as an average across all pages.

One way around this is to segment your data by landing page. In our sample site, my URLs are: http://www.brand.com/blog/topicXYZ. In order to separate the blog pages from the rest, I would insert another calculated field and add the following expression:

```
if Find([Landing Page],"/blog/") > 0 THEN "no" ELSE "yes" END
```

This expression would check if the landing page contains the string /blog/ and if it does, it adds the word "yes" into ournewly calculated field (column). This will give me another dimension to segment my data against.



Calculated Field [Blog]

Name: Blog

Formula:

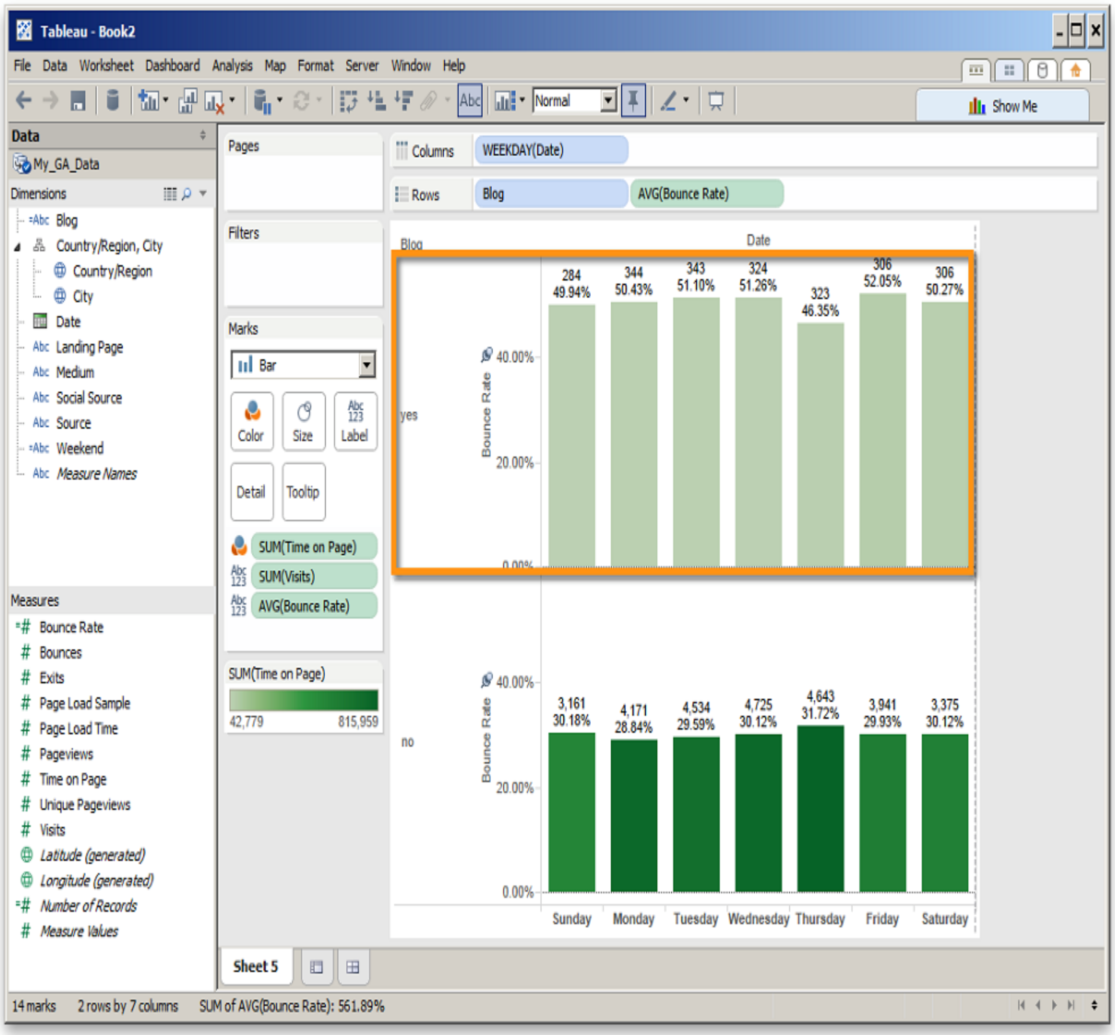
```
if Find([Landing Page],"blog") > 0 THEN "no" ELSE "yes" END
```

The calculation is valid.

IF <expr> THEN <then> [ELSEIF <expr2> THEN <then2> ...] [ELSE

Tests a series of expressions returning the <then> value for the first true <expr>. Example: IF Profit > 0 THEN 'Profitable' ELSEIF Profit = 0 THEN 'Breakeven' ELSE 'Loss' END

Now, we can look at the engagement by blog pages and non-blog pages, and even divide it by day of week. As you can see, the upper section (blog pages) has a much higher bounce rate than the non-blog pages. (It also seems that there was some special activity on Thursdays that affected the bounce rate).



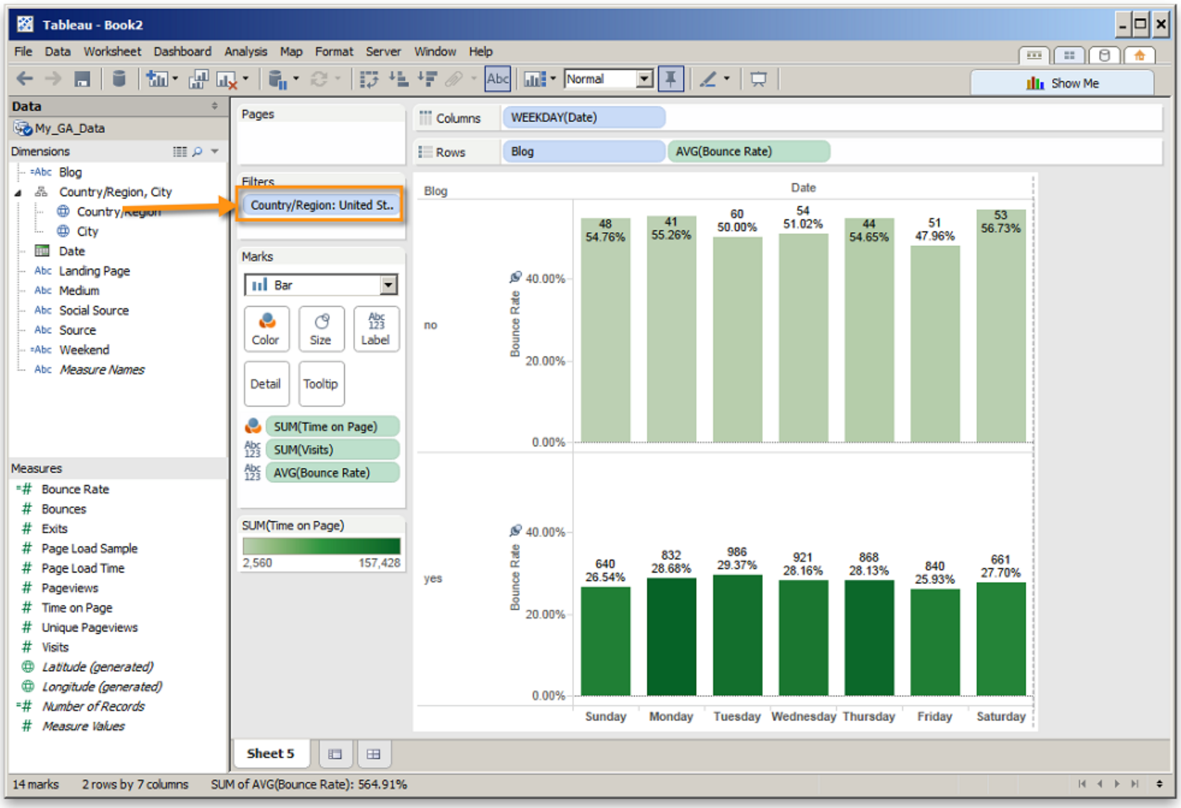
Columns: WEEKDAY(Date)

Rows: Blog, AVG(Bounce Rate)

14 marks 2 rows by 7 columns SUM of AVG(Bounce Rate): 561.89%

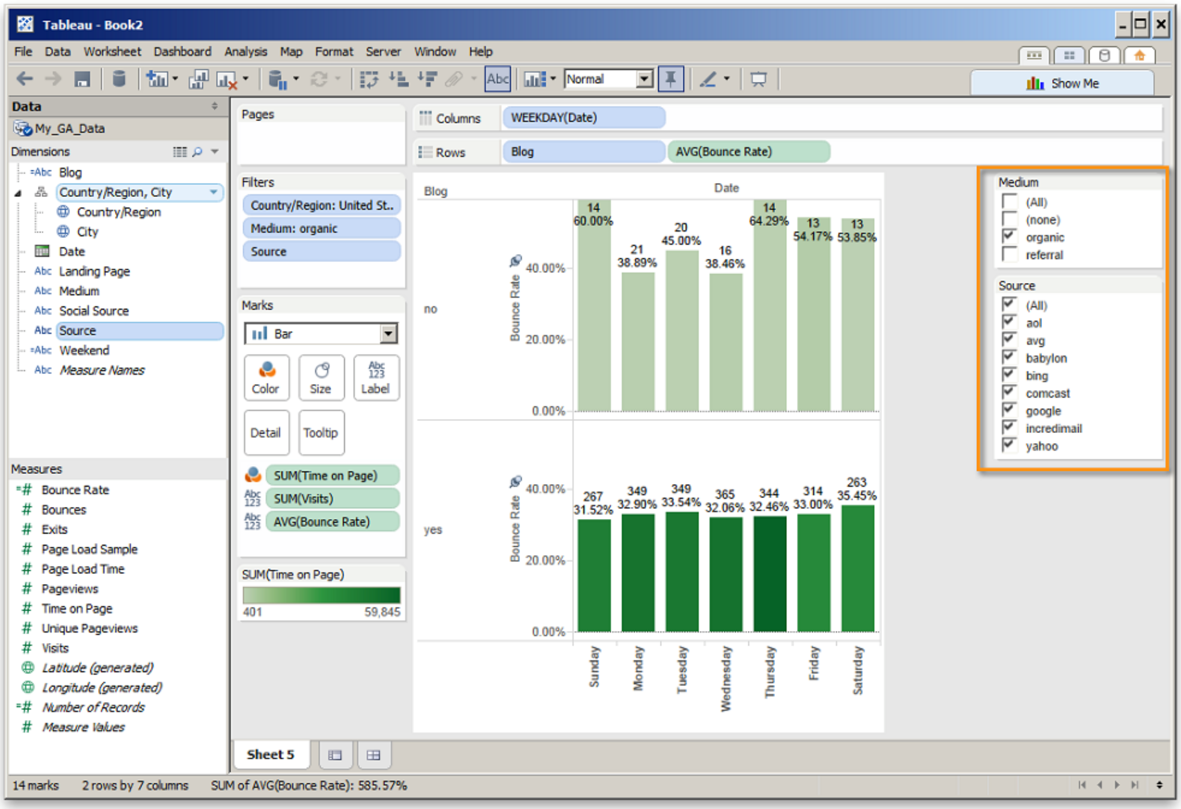
The calculated field option is one of my favorite features in Tableau, as it allows you to dynamically extend and segment your data. We have done some amazing calculations with this functionality, and once you start playing around with the calculated field dialog, you will see the large variety of powerful functions available to you. We are using it from score calculations all the way to a form of tagging. The beauty here is that if you would refresh your data or even swap your data sources, all these fields will be recalculated.

8. Filtering Options. One of the great “show room” qualities Tableau has is its ability to filter data in real time, and it provides two primary ways to make it happen. The first method is to simply drag and drop the element you want to filter onto the filter region, and then pick your option.



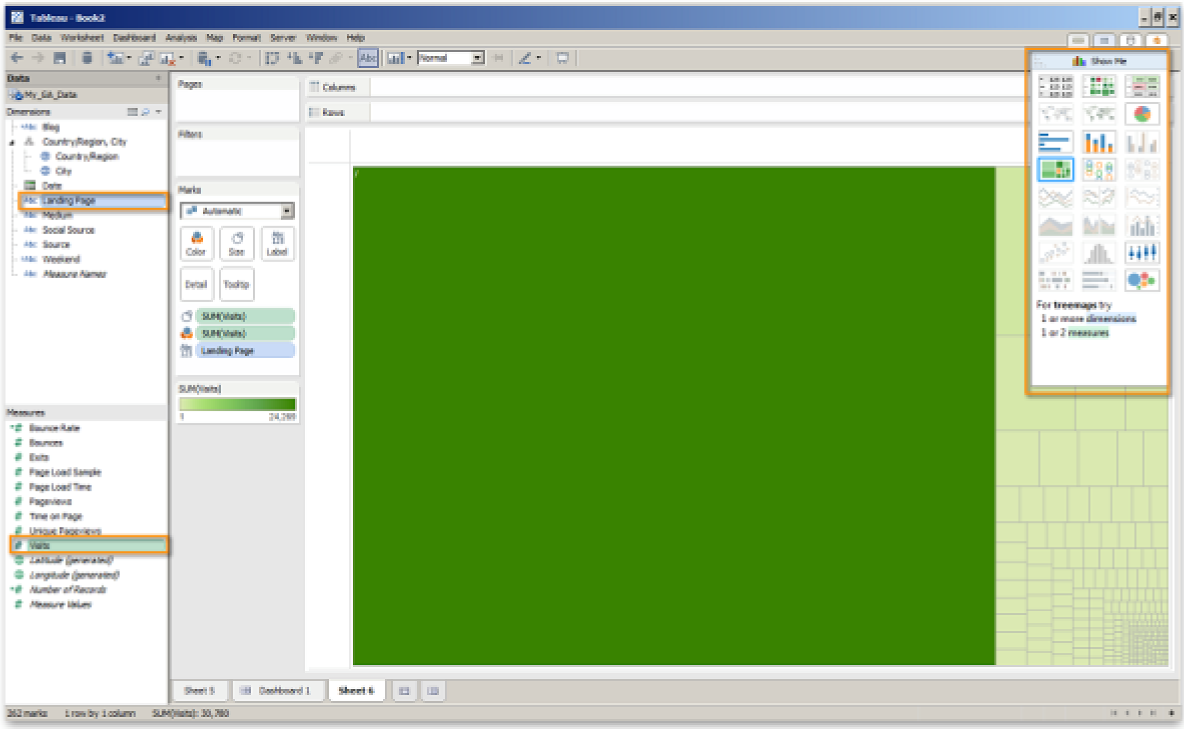
The other method (and the one I prefer) is the quick filter option. If you right-click on any element you want to filter and then click on "add to quick filter," it will add the filters to the right hand column. In the example below, I used medium and source; now I can quickly filter to the items that are relevant to me.

This is a great functionality to have in a client demo or analysis session as you can quickly answer business questions like: *"How does referral traffic differ from direct?"* or*"Can we compare pages per visit by medium or even city?"*



9. Quick Visual Options. In order to get started really quickly with your visualizations, Tableau has a feature called "Show me." It is located in the top right of the screen, and it shows the different types of visuals Tableau offers. When you hover over the visuals, it will tell you what is required for each.

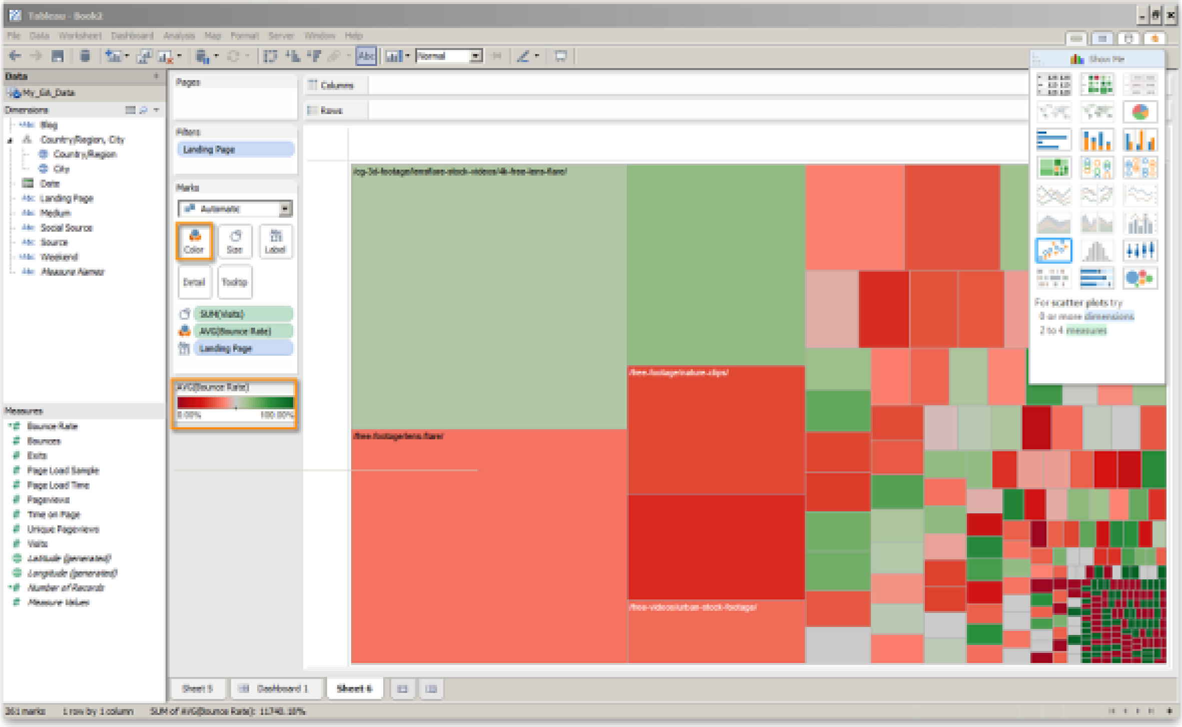
For instance, let's select landing page and visits from the measure and dimensions area, and then select treemap for the visual type. Immediately, it shows you squares that represent individual landing pages, each one sized by the number of visits it has received.



In my example, the homepage "/" is very dominant and prevents us from digging into the details. To make things easier, let's right-click on the homepage "/" and click exclude. By default, it is colored and sized by the amount of visits. This is great — but let's start evaluating our data on multiple dimensions. Drop bounce rate on the color icon. (***Note***: *I changed the color to a red/green gradient.*) Now it shows us the top performing pages, sized by the volume, and color by the bounce rate.

This allows us to look at what is driving volume and what is driving engagement. Now, we can actually prioritize which pages we want to work on first.

Of course, in order to really evaluate this, you want to make sure you are filtering to the correct country your content is targeting, as well as a specific medium you are interested in evaluating. Again, it will get interesting if you now add conversion rate or another value KPI.



Summary

- 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.
- Tableau automatically knows the settings for a Text File Connection.
- Data visualization is both an art and a science.

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 gage 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 gage 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.