



ASSOCIATE ANALYTICS FACILITATORS GUIDE MODULE 1



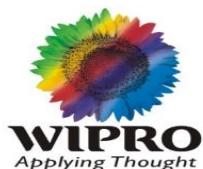
This Facilitators Guidebook for the Associate Analytics program contains detailed facilitation guidelines as well as the exhaustive course material for the Associate Analytics program.

Facilitator's Guide



Associate - Analytics

Powered by:



Copyright © 2014

NASSCOM

4E-Vandana Building (4th Floor)
11, Tolstoy Marg, Connaught Place
New Delhi 110 001, India
T 91 11 4151 9230; F 91 11 4151 9240
E ssc@nasscom.in
W www.nasscom.in

Published by



Building Domain | Enhancing Careers

T: 91 70365 88888
E info@mindmapconsulting.com
W www.mindmapconsulting.com

Disclaimer

The information contained herein has been obtained from sources reliable to NASSCOM. NASSCOM disclaims all warranties as to the accuracy, completeness or adequacy of such information. NASSCOM shall have no liability for errors, omissions, or inadequacies, in the information contained herein, or for interpretations thereof. Every effort has been made to trace the owners of the copyright material included in the book. The publishers would be grateful for any omissions brought to their notice for acknowledgements in future editions of the book.

No entity in NASSCOM shall be responsible for any loss whatsoever, sustained by any person who relies on this material. The material in this publication is copyrighted. No parts of this report can be reproduced either on paper or electronic media, unless authorized by NASSCOM.

Foreword

The Indian IT-ITeS industry has built its reputation in the global arena on several differentiators, chief among them being the availability of manpower. Organizations across the world recognize the value India brings to every engagement with its vast and readily available pool of IT professionals. Global entities have found it extremely effective to leverage this significant resource in order to enjoy a competitive edge and innovation benefits.

In the coming years, the landscape is expected to shift in ways that reveal more exciting opportunities. The world will require people with advanced technology skills and domain knowledge, set against a backdrop of heightened labour mobility across occupations and markets. India is largely acknowledged to be heir apparent to the benefits of a demographic dividend over the coming decades, which has the potential to see the nation emerge as one of the world's largest population base of employable youth. With many other countries set to face the effects of an aging and retirement-ready workforce, India is poised to become a sought after destination for those seeking higher value add and specialized services.

Global markets are on their way towards revival and recovery, and this is well reflected in the proactive recruitment measures taken by IT-ITeS organizations in India in recent times. India's IT-BPM industry is on track to achieve its target of USD 225 billion by 2020. From a base on about 3.1 million employees in FY2014, the industry is expected to add another 2 million additional employees by 2020. Indirect employment generated by 2020 is expected to be 3X the total direct employment number is between 13-16 million by 2020.

To realize India's potential of emerging as a skills hub of the world, a significant amount of foresight and work is requisite. It is imperative that stakeholders engage in a concerted effort to undertake the transformation of the labour pool estimated to enter the market into skilled and employable talent. Enabling the creation of a future industry-ready cohort will give the IT-ITeS industry an edge in leadership and sustainability.

One of the burgeoning areas of governance and strategy relates to leveraging big data and analytics. This led to the identification of the "hot skills" du jour, resulting in the formal creation of a qualification pack (QP) or job role framework for the role of Associate Analytics. The QP is designed to capture the skills demanded by the IT-BPM Industry for an entry level position in this field.

To ensure the creation of an academic course that is both relevant and viable, NASSCOM partnered with key industry stakeholders, including Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. In addition, the program addresses the need for faculty support, and achieves this by acquainting trainers with the latest advancements in pedagogy.

We wish the universities and colleges all the very best in their endeavor.

R Chandrashekhar
President
NASSCOM

Acknowledgements

NASSCOM would like to thank its member company representatives within the Analytics Special Interest Group (SIG) Council for believing in our vision to enhance the employability of the available engineering student pool. SSC NASSCOM facilitates this by developing and enabling the implementation of courses relevant to projected industry needs. The aim is to address two key requirements, of closing the industry-academia skill gap, and of creating a talent pool that can reasonably weather future externalities in the IT-BPM industry.

NASSCOM believes that this is an initiative of great importance for all stakeholders concerned – the industry, academia, and the students. The tremendous amount of work and ceaseless support offered by the members of this SIG in developing a meaningful strategy for the content and design of program training materials has been truly commendable.

We would like to particularly thank Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, Fractal Analytics, First America, Genpact, Infosys BPO, Insights of Data, Karvy Analytics, Wipro, WNS and Wells Fargo for bringing much needed focus to this effort.

NASSCOM recognizes the fantastic contributions of Mr. Ashok Polapragada, Mr. Ranjit Kumar and Mr. Prakash Devarakonda at Karvy Analytics; Mr. Dwaraka Ramana K at First American; Mr. Amit Agarwal, Mr. Sidhartha Shishoo and team at Genpact; Ms. Snigdha Ray and Mr. Amit Sharma at ADP; Mr. Manoj Koundinya at Capgemini, and Mr. Ashish Mediratta at Wipro.

We acknowledge with sincere gratitude the immense contribution of the SIG member companies, Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. For their part in the creation of this course and its accompanying training materials.

We extend our thanks to Mindmap Consulting Pvt. Ltd. for producing this course publication.

Dr Sandhya Chintala
Executive Director – Sector Skill Council
Vice President - NASSCOM

Table of Contents – Module 1

Introduction to QP Associate Analytics

Introduction to Associate Analytics	8
Career growth in Analytics	11
Qualification pack - Q/2101 Associate Analytics	12
Overall Associate Analytics Content Structure	20
Glossary of terms	22

CORE CONTENT

UNIT 1.1 Introduction to R and R Programming	27
UNIT 1.2 Manage your work to meet requirements	40
UNIT 2.1 Summarizing Data and Revisiting Probability	60
UNIT 2.2 Work effectively with Colleagues	73
UNIT 3.0 SQL using R	99
UNIT 4.0 Correlation and Regression Analysis	105
UNIT 5.0 Understanding the Verticals and Requirements Gathering	119

ADDITIONAL READING MATERIAL FROM KARVY ANALYTICS

Handbook on Big Data Overview	127
Handbook on Big Data Tools Alternatives	176

Introduction

Qualifications Pack-Associate – Associate Analytics SSC/Q2101

SECTOR: IT-ITeS

SUB-SECTOR: Business Process Management

OCCUPATION: Analytics

REFERENCE ID: SSC/Q2101

ALIGNED TO NCO CODE: TBD

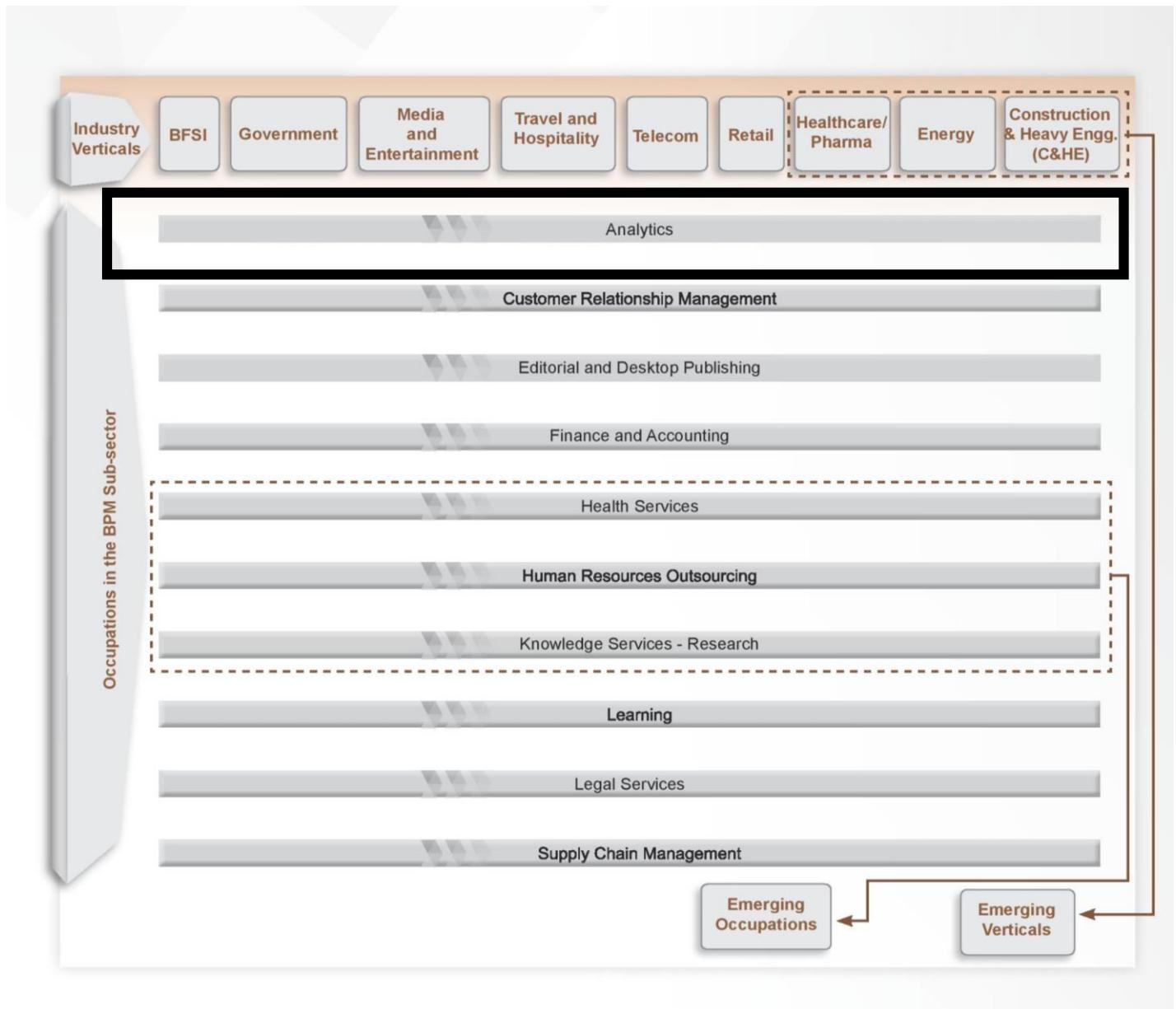
Brief Job Description: Individuals at this job are responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools

Personal Attributes: This job requires the individual to follow detailed instructions and procedures with an eye for detail. The individual should be analytical and result oriented and should demonstrate logical thinking.

Eligibility: Bachelor's Degree in Statistics/ Science/Technology, Master's Degree in Science/Technology/Statistics

Work Experience: 0-1 years of work experience/internship in analytics roles

Analytics is a key occupation in the structure of the ITS Sub-Sector



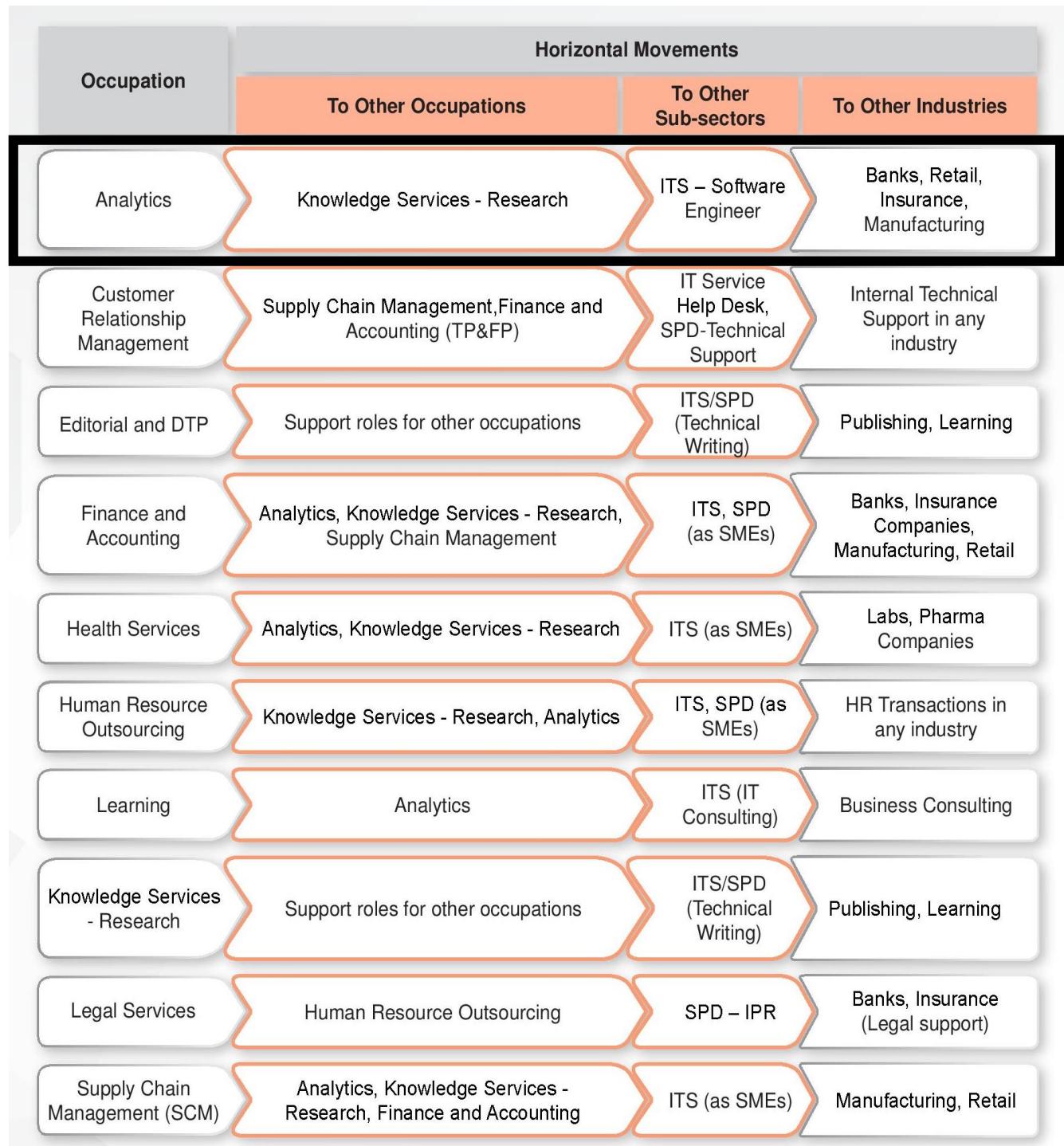
Analytics excellent Vertical and Horizontal movements in their

tracks

Occupation	Tracks	Entry-level Job Roles
Analytics	MIS - Reporting Analytics - Modelling and Analysis	Associate - Analytics
Customer Relationship Management (CRM)	Customer Care (Non-voice)	Associate - Customer Care (Non-voice) Associate - CRM
	Customer Care (Voice)	
	Sales/Telesales	
	Technical Support/IT Help Desk	
	Collections (Business to Customer)	
Editorial and Desktop Publishing (DTP)	Editorial	Associate - Editorial
	DTP and Design	Associate - DTP
Finance and Accounting (F&A)	Transaction Processing (Includes B2B Collections)	Associate - Transactional F&A
	Credit Analysis	Associate - F&A Complex
	Audit and Accounting	
	Financial Reporting	
	Financial Planning and Analysis (Includes Budgeting and Forecasting)	
Health Services	Clinical Data Management	Associate - Clinical Data Management
	Medical Transcription	Associate - Medical Transcription
Human Resource Outsourcing (HRO)	Recruitment	Associate - Recruitment
	Learning and Development	Associate - HRO
	Compensation and Benefits Management	
	Employee Relations	
Knowledge Services - Research	Secondary Research and Market Research	Analyst - Research
	Investment Banking Research	
Learning	Content Managemnet	Associate - Learning
	Instructional Design	
Legal Services	Legal Services	Document Coder/Processor Legal Associate
Supply Chain Management	Procurement Operations (Including Strategic Sourcing)	Associate - SCM
	Sales and Fulfilment (Including Inventory Mangement)	

Movement to Other Occupations, Sub-sectors and Industries:

Given the dynamic range of services that the BPM sub-sector is increasingly offering to its clients in the industry, there are a variety of roles that employees are performing across the entire spectrum of offerings. As such they become a valuable asset not only to the BPM sub-sector, but also to all the client industries they are associated with.



OVERALL QUALIFICATION PACK DETAILS

Job Details	Qualifications Pack Code	SSC/Q2101		
	Job Role	Associate - Analytics This job role is applicable in both national and international scenarios		
	Credits(NVEQF/NVQF/NSQF)		Version number	0.1
	Sector	IT-ITeS	Drafted on	30/04/13
	Sub-sector	Business Process Management	Last reviewed on	30/04/13
	Occupation	Analytics	Next review date	30/06/14

KA1. Job Role	KA2. Associate - Analytics (Business Analytics Associate/ Analyst)
KA3. Role Description	KA4. Responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools
KA5. NVEQF/NVQF level KA6.	KA10. 7
KA7. Minimum Educational Qualifications	KA11. Bachelor's Degree in Statistics/ Science/Technology or any other course
KA8.	KA12. Master's Degree in Science/Technology/Statistics or any other course
KA9. Maximum Educational Qualifications	
KA13. Training KA14. (Suggested but not mandatory)	KA15. Courses in SPSS, SAS, STATA and/or Spreadsheets KA16. RDBMS concepts, PL\SQL, OCA certification KA17. Financial and accounting terminologies in respective language & various accounting standards and GAAPs
KA18. Experience KA19.	KA20. 0-1 years of work experience/internship in analytics roles
KA21. Applicable National Occupational Standards (NOS)	KA22. Compulsory: 1. SSC/ N 0703 (Create documents for knowledge sharing) 2. SSC/ N 2101 (Carry out rule-based statistical analysis) 3. SSC/ N 9001 (Manage your work to meet requirements) 4. SSC/ N 9002 (Work effectively with colleagues) 5. SSC/ N 9003 (Maintain a healthy, safe and secure working environment) 6. SSC/ N 9004 (Provide data/information in standard formats) 7. SSC/ N 9005 (Develop your knowledge, skills and competence) KA23. KA24. Optional: KA25. Not Applicable
KA26. Performance Criteria	KA27. As described in the relevant OS units

SSC/ N0703 - Create Documents for knowledge sharing

Session Overview

In the Associate Analytics “Working with Documents”, the participant will learn about the most prominently used documentation techniques in corporate organizations. The Documentation types covered would include case studies, best practices, project artifacts, reports, minutes, policies, procedures, work instructions etc.

This session is NOT intended to cover technical documents or documents to support the deployment and use of products/applications, which are dealt with in different standards.

Session Goal

Participants should be able to have a good hands on understanding of MS Word and MS Visio, where there will be required to draft various documents/reports. The goal of the session is for the participant to be aware of the various documentation techniques which are used prominently in organizations.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish with **appropriate people** the purpose, scope, formats and target audience for the documents
- PC2. access existing documents, language standards, templates and documentation tools from your organization’s knowledge base
- PC3. liaise with **appropriate people** to obtain and verify the information required for the documents
- PC4. confirm the content and structure of the documents with **appropriate people**
- PC5. create documents using standard templates and agreed language standards
- PC6. review documents with **appropriate people** and incorporate their inputs
- PC7. submit documents for approval by **appropriate people**
- PC8. publish documents in agreed formats
- PC9. update your organization’s knowledge base with the documents
- PC10. comply with your organization’s policies, procedures and guidelines when creating documents for knowledge sharing

Note: The material for this NOS has been covered in the Associate Analytics Module 3 Book (book 3) in Unit 5

SSC/ N 2101 – Carry out rule-based statistical analysis

Session Overview

In the Associate Analytics *Carry out rule based statistical analysis*, the participants will go through Business Analytics using R tool. The participants will also learn Applied Statistical concepts like Descriptive Statistics and find their usage along with R. Furthermore, they will also have an overview of Big Data tools and their basic functioning.

Then they will learn about Machine Learning algorithm and their use in Data Mining and Predictive Analytics. Finally the participants will learn about Data Visualization and gather knowledge on Graphical representation of Data as well as results and reports.

Session Goal

The primary goal of the session is for the participants to learn the R tools and its various functions and features. Then also learn about Big Data tools and Big Data Analytics. Students will also learn about basic applied statistical concepts.

Session Objectives

To be competent, participants must be able to:

- PC1. establish clearly the objectives and scope of the **analysis**
- PC2. obtain guidance from **appropriate people** to identify suitable **data sources** to agree the methodological approach
- PC3. obtain and structure data using standard templates and tools
- PC4. validate data accurately and identify **anomalies**
- PC5. obtain guidance from **appropriate people** on how to handle **anomalies** in data
- PC6. carry out rule-based **analysis** of the data in line with the analysis plan
- PC7. validate the results of your **analysis** according to statistical guidelines
- PC8. review the results of your **analysis** with **appropriate people**
- PC9. undertake modifications to your **analysis** based on inputs from **appropriate people**
- PC10. draw justifiable inferences from your **analysis**
- PC11. present the results and inferences from your analysis using standard templates and tools
- PC12. comply with your organization's policies, procedures and guidelines when carrying out rule-based quantitative **analysis**

Note: The material for this NOS has been covered in all the three Modules of Associate Analytics

SSC/ N 9001: Manage Your Work to Meet Requirement

Session Overview

The Associate Analytics *Manage your work to meet requirement* module is designed to help participants understand the importance of time in a professional environment and how to manage multiple time bound requirements. It emphasizes on how time management is critical to work management and completing requirements/deliverables.

Participants learn how to manage work and how to ensure deliverables are completed in stipulated time in an organization by following tested principles to prevent/handle slippages on timelines. The module also emphasizes the need to respect time for self as well as colleagues.

Time management cannot override the qualitative aspect of the deliverable.

Session Goal

The primary goal of the session is for the participants to learn and manage time to be able to complete their work as required. The requirements of a work unit may be further classified into; activities, deliverable, quantity, standards and timelines. The session makes participants to be aware of defining requirements of every work unit and then ensuring delivery.

Additionally, this session discusses practical application of planning and execution of work plans to enable the participants to effectively deal with the failure points, minimize the impact, if any. Equally critical is the escalation plan and root cause analysis of exceptions.

Successful candidates will be able to understand the inter-relationship of time, effort, impact and cost.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Establish and agree your work requirements with appropriate people

PC2. Keep your immediate work area clean and tidy

PC3. Utilize your time effectively

PC4. Use resources correctly and efficiently

PC5. Treat confidential information correctly

PC6. Work in line with your organization's policies and procedures

PC7. Work within the limits of your job role

PC8. Obtain guidance from appropriate people, where necessary

PC9. Ensure your work meets the agreed requirements

Note: The material for this NOS has been covered in Unit 1 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9002: Work Effectively With Colleagues

Session Overview

The Associate Analytics *Work Effectively with Colleagues* module is designed to help participants understand the importance of teamwork in a professional environment. It emphasizes on how relationship management is critical to work management. It also focuses on the importance of personal grooming.

Participants learn how to manage cross functional relationships and how to nurture a good working environment. The module also stresses on the need to respect colleagues.

Session Goal

The primary goal of the session is for the participants to understand the importance of professional relationships with colleagues. Additionally, this session discusses importance of personal grooming.

Successful candidates will be able to understand the inter-relationship of professionalism and team-work.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Communicate with colleagues clearly, concisely and accurately.

PC2. Work with colleagues to integrate your work effectively with theirs.

PC3. Pass on essential information to colleagues in line with organizational requirements.

PC4. Work in ways that show respect for colleagues.

PC5. Carry out commitments you have made to colleagues.

PC6. Let colleagues know in good time if you cannot carry out your commitments, explaining the reasons.

PC7. Identify any problems you have working with colleagues and take the initiative to solve these problems.

PC8. Follow the organization's policies and procedures for working with colleagues.

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9003: Maintain a Healthy, Safe and Secure working Environment

Session Overview

The Associate Analytics *Health, Safety and Security* module is designed to help participants understand the importance of following safety rules and regulations at workplace.

Participants learn how to work safely in an organization by following guidelines to prevent/handle any accidents or emergencies. The module also emphasizes the need of security and the entities that can pose a threat to it.

Session Goal

The primary goal of the session is for the participants to be aware about the various hazards that they may come across at workplace and what are the defined health, safety and security measures that should be followed at the time of occurrence of such unpredictable events. Additionally, this session discusses practical application of the health and safety procedures to enable the participants to effectively deal with the hazardous events to minimize the impact, if any.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. Comply with your organization's current health, safety and security policies and procedures
- PC2. Report any identified breaches in health, safety, and security policies and procedures to the designated person
- PC3. Identify and correct any hazards that you can deal with safely, competently and within the limits of your authority
- PC4. Report any hazards that you are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected
- PC5. Follow your organization's emergency procedures promptly, calmly, and efficiently
- PC6. Identify and recommend opportunities for improving health, safety, and security to the designated person
- PC7. Complete any health and safety records legibly and accurately

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9004: Provide data/information in standard formats

Session Overview

The Associate Analytics *Provide data/information in standard formats* module is designed to help participants understand the standard operating procedures in organizations pertaining to reporting data in a logical sequence and arriving at conclusive decisions models after analysis of data. This module is aimed at developing the sense of understanding in an individual when the individual works with data, of how to take the data and present it as relevant information in standardized formats.

Participants learn how to share information with other people inside or outside a specified work group and also how to arrive at decisions regarding certain problem types.

Session Goal

The primary goal of the session is for the participants to analyze data and present it in a suitable format, as is suitable for the given process or organization.

Successful candidates will be able to understand the process of standardized reporting and the nuances of a publishing a report with a specified end objective in mind.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish and agree with appropriate people the data/information you need to provide, the formats in which you need to provide it, and when you need to provide it
- PC2. obtain the data/information from reliable sources
- PC3. check that the data/information is accurate, complete and up-to-date
- PC4. obtain advice or guidance from appropriate people where there are problems with the data/information
- PC5. carry out rule-based analysis of the data/information, if required
- PC6. insert the data/information into the agreed formats
- PC7. check the accuracy of your work, involving colleagues where required
- PC8. report any unresolved anomalies in the data/information to appropriate people
- PC9. provide complete, accurate and up-to-date data/information to the appropriate people in the required formats on time

SSC/ N 9005: Develop your knowledge, skills and competence

Session Overview

The Associate Analytics *develop your knowledge, skills and competence* module is designed to help participants understand the importance of skill development in a professional environment and how to enhance skills in order to excel. It emphasizes on how enhance skills and knowledge in a diversified professional environment.

Session Goal

The primary goal of the session is to give a overview on how skills and competency can be enhanced in a professional environment. It gives knowledge on organizational context, technical knowledge, core skills/geneic skills, professional skills and technical skills. The session makes participants to understand the need of skills improvement for personal and organizational growth.

Successful candidates will be able ro understand the relationship between skill enhancement and growth.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. obtain advice and guidance from appropriate people to develop their knowledge, skills and competence
- PC2. identify accurately the knowledge and skills you need for their job role
- PC3. identify accurately their current level of knowledge, skills and competence and any learning and development needs
- PC4. agree with appropriate people a plan of learning and development activities to address their learning needs
- PC5. undertake learning and development activities in line with their plan
- PC6. apply their new knowledge and skills in the workplace, under supervision
- PC7. obtain feedback from appropriate people on their knowledge and skills and how effectively they apply them
- PC8. review their knowledge, skills and competence regularly and take appropriate action

Overall Associate Analytics Content Structure

Module 1 – Book 1

Subject I / SSC NASSCOM - NOS- 2101, 9001, 9002	NOS	Hours	Minutes
Unit - 1	NOS 2101/9001		
Introduction to Analytics & R programing		6	360
Manage your work to meet requirements		4	240
Unit - 2	NOS 2101/9002		
Summarizing Data & Revisiting Probability		6	360
Work effectively with Colleagues		4	240
Unit - 3	NOS 2101		
SQL using R		9	510
Unit - 4	NOS 2101		
Correlation and Regression Analysis		9	510
Unit - 5	NOS 2101		
Understanding Verticals - Engg, Financial, others		6	390
Requirements Gathering		6	390
Total Hrs/Minutes		50	3000

Module 2 – Book 2

Subject II / SSC NASSCOM - NOS- 2010, 9003, 9004	NOS	Hours	Minutes
Unit - 1	NOS 2101/9003		
Data Management		7	420
Maintain Healthy, Safe & Secure Working environment		4	240
Unit - 2	NOS 2101/9004		
Big Data Tools		7	420
Provide Data/Information in Standard formats		4	240
Unit - 3	NOS 2101		
Big Data Analytics		8	480
Unit - 4	NOS 2101		
Machine Learning Algorithms		8	480
Unit - 5	NOS 2101		
Data Visualization		6	360
Product Implementation		6	360
Total Hrs/Minutes		50	3000

Module 3 – Book 3

Subject III / SSC NASSCOM - NOS - 0703, 2101, 9005	NOS	Hours	Minutes
Unit - 1	NOS 2101		
Introduction to Predictive Analytics		6	360
Linear Regression		6	360
Unit - 2	NOS 2101		
Logistics Regression		9	540
Unit - 3	NOS 2101/9005		
Objective Segmentation		6	360
Develop Knowledge Skill and competences		3	180
Unit - 4	NOS 2101		
Time Series Methods/Forecasting, Feature Extraction		5	300
Project		5	300
Unit - 5	NOS 0703		
Working with documents		10	600
Total Hrs/Minutes		50	3000

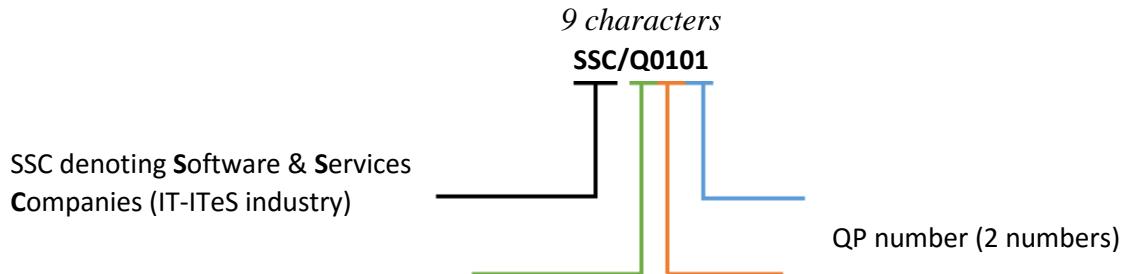
Glossary of Terms

Definitions	Keywords /Terms	Description
	Sector	Sector is a conglomeration of different business operations having similar businesses and interests. It may also be defined as a distinct subset of the economy whose components share similar characteristics and interests.
	Sub-sector	Sub-sector is derived from a further breakdown based on the characteristics and interests of its components.
	Vertical	Vertical may exist within a sub-sector representing different domain areas or the client industries served by the industry.
	Occupation	Occupation is a set of job roles, which perform similar/related set of functions in an industry.
	Function	Function is an activity necessary for achieving the key purpose of the sector, occupation, or area of work, which can be carried out by a person or a group of persons. Functions are identified through functional analysis and form the basis of OS.
	Sub-functions	Sub-functions are sub-activities essential to fulfill the achieving the objectives of the function.
	Job role	Job role defines a unique set of functions that together form a unique employment opportunity in an organisation.
	Occupational Standards (OS)	OS specify the standards of performance an individual must achieve when carrying out a function in the workplace, together with the knowledge and understanding they need to meet that standard consistently. Occupational Standards are applicable both in the Indian and global contexts.
	Performance Criteria	Performance Criteria are statements that together specify the standard of performance required when carrying out a task.
	National Occupational Standards (NOS)	NOS are Occupational Standards which apply uniquely in the Indian context.
	Qualifications Pack Code	Qualifications Pack Code is a unique reference code that identifies a qualifications pack.
	Qualifications Pack(QP)	Qualifications Pack comprises the set of OS, together with the educational, training and other criteria required to perform a job role. A Qualifications Pack is assigned a unique qualification pack code.
	Unit Code	Unit Code is a unique identifier for an OS unit, which can be denoted with either an 'O' or an 'N'.
	Unit Title	Unit Title gives a clear overall statement about what the incumbent should be able to do.

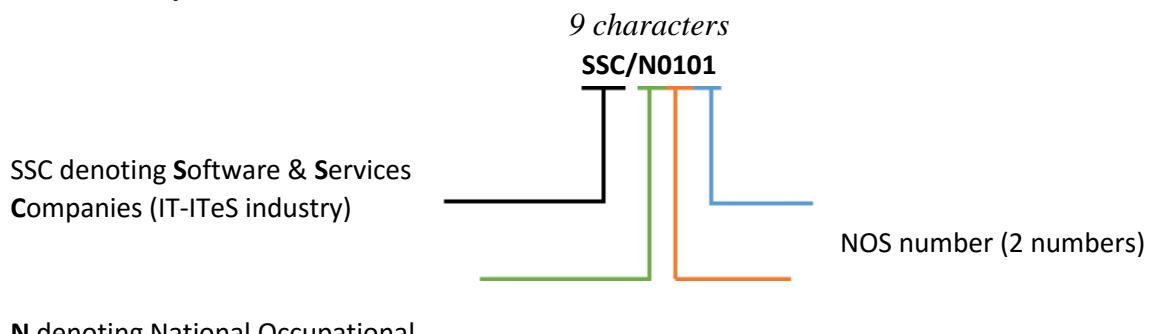
Description	Description gives a short summary of the unit content. This would be helpful to anyone searching on a database to verify that this is the appropriate OS they are looking for.
Scope	Scope is the set of statements specifying the range of variables that an individual may have to deal with in carrying out the function which have a critical impact on the quality of performance required.
Knowledge and Understanding	Knowledge and Understanding are statements which together specify the technical, generic, professional and organisational specific knowledge that an individual needs in order to perform to the required standard.
Organisational Context	Organisational Context includes the way the organisation is structured and how it operates, including the extent of operative knowledge managers have of their relevant areas of responsibility.
Technical Knowledge	Technical Knowledge is the specific knowledge needed to accomplish specific designated responsibilities.
Core Skills/Generic Skills	Core Skills or Generic Skills are a group of skills that are key to learning and working in today's world. These skills are typically needed in any work environment. In the context of the OS, these include communication related skills that are applicable to most job roles.
Helpdesk	Helpdesk is an entity to which the customers will report their IT problems. IT Service Helpdesk Attendant is responsible for managing the helpdesk.
Keywords /Terms	Description
IT-ITeS	Information Technology - Information Technology enabled Services
BPM	Business Process Management
BPO	Business Process Outsourcing
KPO	Knowledge Process Outsourcing
LPO	Legal Process Outsourcing
IPO	Information Process Outsourcing
BCA	Bachelor of Computer Applications
B.Sc.	Bachelor of Science
OS	Occupational Standard(s)
NOS	National Occupational Standard(s)
QP	Qualifications Pack
UGC	University Grants Commission
MHRD	Ministry of Human Resource Development
MoLE	Ministry of Labour and Employment
NVEQF	National Vocational Education Qualifications Framework
NVQF	National Vocational Qualifications Framework
NSQF	National Skill Qualification Framework

Nomenclature for QP & NOS UNITS

Qualifications Pack

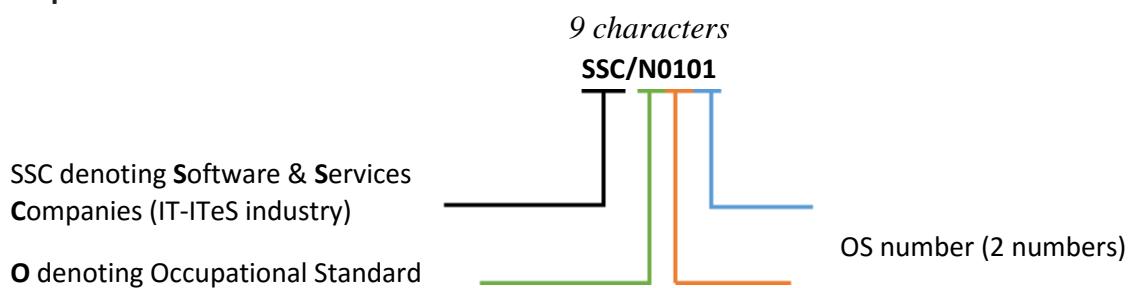


National Occupational Standard



N denoting National Occupational

Occupational Standard



It is important to note that an OS unit can be denoted with either an '**O**' or an '**N**'.

- If an OS unit denotes '**O**', it is an OS unit that is an international standard. An example of OS unit denoting '**O**' is **SSC/O0101**.
- If an OS unit denotes '**N**', it is an OS unit that is a national standard and is applicable only for the Indian IT-ITeS industry. An example of OS unit denoting '**N**' is **SSC/N0101**

The following acronyms/codes have been used in the nomenclature above:

Sub-Sector	Range of Occupation numbers
IT Service (ITS)	01-20
Business Process Management (BPM)	21-40
Engg. and R&D (ERD)	41-60
Software Products (SPD)	61-80

Sequence	Description	Example
Three letters	Industry name (Software & Service Companies)	SSC
Slash	/	/
Next letter	Whether QP or NOS	N
Next two numbers	Occupation Code	01
Next two numbers	OS number	01

Module 1: Unit – 1.1

Introduction to Analytics or R programming

Topic	Session Goals
Introduction to Analytics or R programming	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Understand R 2. Use functions of R

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Knowing language R	Classroom
Using R as calculator	Classroom
Understanding components of R	Classroom
Reading database using R	Classroom
Importing & Exporting CSV	Classroom
Working on Variables	Classroom
Outliers and Missing Data treatment	Classroom
Combining Data sets in R	Classroom
Discuss Function and Loops	Classroom
Check your understanding	Classroom
Summary	Classroom

Unit 1.1 : Step-by-Step

Key Points

Knowing the language “R”

R can be defined as below:-

- Programming language for graphics and statistical computations
- Available freely under the GNU public license
- Used in data mining and statistical analysis
- Included time series analysis, linear and nonlinear modeling among others
- Very active community and package contributions
- Very little programming language knowledge necessary
- Can be downloaded from <http://www.r-project.org/>

Did you know?



- R Studio is an IDE for R with advanced and more user-friendly GUI.
- R is the substrate on which we can mount various features using PACKAGES like RCMDR- R Commander or R-Studio.
- R was started by Bell Laboratories as “S” for Fortran Library.

Look at R!

R Commander

File Data Statistics Graphs Models Distributions Help

Data : Survey attach(S)

- Summaries
- Contingency tables
- Means
- Proportions
- Variances
- Nonparametric tests
- Dimensional analysis
- Fit models

Linear regression...
Linear model...
Generalized linear model...

Model: <No active model>

R-Commander Interface

RStudio

Source on US.states.R*

```

1 library(maps)
2 library(ggplot2)
3 data(us.cities)
4
5 aplot(long, lat, data = choro, group = group,
6 fill = assault, geom = "polygon")
7 aplot(long, lat, data = choro, group = group,
+ fill = assault / murder, geom = "polygon")
8 aplot(long, lat, data = choro, group = group,
fill = assault, geom = "polygon")

```

Console ~ /

```

> chorono <- chorono[order(chorono$order), ]
> aplot(long, lat, data = chorono, group = group,
+ fill = assault, geom = "polygon")
> aplot(long, lat, data = chorono, group = group,
+ fill = assault / murder, geom = "polygon")
> aplot(long, lat, data = chorono, group = group,
fill = assault, geom = "polygon")

```

Workspace History

Data

Capitals	2x47 double matrix
Distance	47x47 double matrix
Host	47x47 double matrix
LanguageDistance	47x47 double matrix
Multilingual	12x47 double matrix
Neighbours	47x47 double matrix

Files Plots Packages Help

US map showing assault rates across US states.

R-Studio Interface

Using R as calculator

R can be used as a calculator. For example, if we have to know what is 2+2 then-

[CODE]

> 2+2
Press enter and we get the answer as
[1] 4

Similarly we can calculate anything as if done on calculator.



Calculate the following using R:

1. Log of 2
2. $2^3 \times 3^2$
3. e^3

Understanding components of R

1. Data Type:

There are two types of data classified on very broad level. They are Numeric and Character data.

- Numeric Data: - It includes 0~9, “.” and “- ve” sign.
- Character Data: - Everything except Numeric data type is Character. For Example, Names, Gender etc.

Data is also classified as Quantitative and Qualitative.

For Example, “1,2,3...” are Quantitative Data while “Good”, “Bad” etc. are Qualitative Data.

Although we can convert Qualitative Data into Quantitative Data using Ordinal Values.

For Example, “Good” can be rated as 9 while “Average” can be rated as 5 and “Bad” can be rated as 0.

2. Data Frame:

A data frame is used for storing data tables. It is a list of vectors of equal length.

For example, here is a built-in data frame in R, called **mtcars**.

```
> mtcars
      mpg cyl disp  hp drat    wt
Mazda RX4     21.0   6 160 110 3.90 2.62
Mazda RX4 Wag 21.0   6 160 110 3.90 2.88
Datsun 710    22.8   4 108  93 3.85 2.32
```

The top line of the table, called the header, contains the column names. Each horizontal line afterward denotes a data row, which begins with the name of the row, and then followed by the actual data. Each data member of a row is called a cell. To retrieve data in a cell, we would enter its row and column coordinates in the single square bracket "[]" operator. The two coordinates are separated by a comma. In other words, the coordinates begins with row position, then followed by a comma, and ends with the column position. The order is important.

For Example,

Here is the cell value from the first row, second column of **mtcars**.

```
> mtcars[1, 2]
[1] 6
```

3. Array and Matrices:

We have two different options for constructing matrices or arrays. Either we use the creator functions **matrix ()** and **Array ()**, or you simply change the dimensions using the **dim ()** function.

For example, you make an array with four columns, three rows, and two “tables” like this:

[CODE]

```
> my.array <- array(1:24, dim=c(3,4,2))
```

In the above example, “my.array” is the name of the array we have given. And “ \leftarrow ” is the assignment operator.

There are 24 units in this array mentioned as “1:24” and are divided in three dimensions “(3, 4, 2)”.

Note: - Although the rows are given as the first dimension, the tables are filled column-wise. So, for arrays, R fills the columns, then the rows, and then the rest.

Alternatively, you could just add the dimensions using the **dim ()** function. This is a little hack that goes a bit faster than using the **array ()** function; it's especially useful if you have your data already in a vector. (This little trick also works for creating matrices, by the way, because a matrix is nothing more than an array with only two dimensions.)

Say you already have a vector with the numbers 1 through 24, like this:

[CODE]

```
> my.vector <- 1:24
```

You can easily convert that vector to an array exactly like my.array simply by assigning the dimensions, like this:

[CODE]

```
> dim(my.vector) <- c(3,4,2)
```



Create an Array with name “MySales” with 30 observations using following methods:

1. Defining the dimensions of the array as 3, 5 and 2.
2. By using Vector method.

You can check whether two objects are identical by using the **identical ()** function.

Reading Database using R

We can import Datasets from various sources having various files types for example,

- .csv format
- Big data tool – Impala
- **CSV File**

The sample data can also be in comma separated values (CSV) format. Each cell inside such data file is separated by a special character, which usually is a comma, although other characters can be used as well. The first row of the data file should contain the column names instead of the actual data. Here is a sample of the expected format.

Col1,Col2,Col3
100,a1,b1
200,a2,b2
300,a3,b3

After we copy and paste the data above in a file named "mydata.csv" with a text editor, we can read the data with the function `read.csv`.

[CODE]

```
> mydata = read.csv("mydata.csv") # read csv file
> mydata
   Col1 Col2 Col3
1 100  a1  b1
2 200  a2  b2
3 300  a3  b3
```

In various European locales, as the comma character serves as the decimal point, the function `read.csv2` should be used instead. For further detail of the `read.csv` and `read.csv2` functions, please consult the R documentation.

```
> help(read.csv)
  ➤ Big data tool – Impala
```

Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine runs natively in Apache Hadoop.

R package, **RImpala**, connects Impala to R.

RImpala enables querying the data residing in HDFS and Apache HBase from R, which can be further processed as an R object using R functions. RImpala is now available for download from the Comprehensive R Archive Network (CRAN) under GNU General Public License (GPL3).

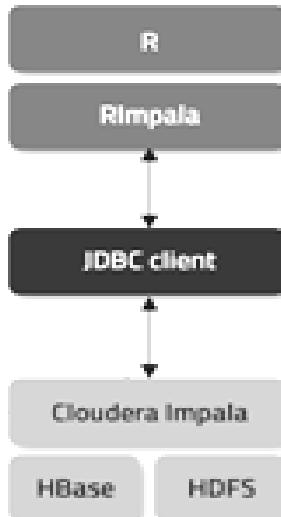
This package is developed and maintained by MuSigma.

To install RImpala :

We use following code to install RImpala package.

```
>install.packages("RImpala")
```

How RImpala works



Importing and Exporting CSV

- Loading data – `data(dataset_name)`
- read and write functions
- `getwd()` and `setwd(dir)`
- read and write functions use full path name

Example:

`read.csv ("C:/Rtutorials/Sampleddata.csv")`. Similarly for writing dataset we use the `write ()` function.

“Getwd” means get the working directory (wd) and “setwd” is used to set the working directory.

- **Help - ?function_name -> It is used to get help on any function in R.**

Activity



Import a CSV file in R and check the output.

Name, Age, Sex

Shaan, 21, M

Ritu, 24, F

Raj, 31, M

Working on Variables

Before learning about creating and modifying variables in R we will know the various operators in R.

There are 2 types of operators –Arithmetic and Logical.

Operator	Description
<code>+</code>	Addition
<code>-</code>	Subtraction
<code>*</code>	Multiplication
<code>/</code>	Division
<code>^ or **</code>	Exponentiation
<code>x %% y</code>	modulus (x mod y) 5%%2 is 1
<code>x %/% y</code>	integer division 5%/%2 is 2

Arithmetic Operators

Operator	Description
<code><</code>	less than
<code><=</code>	less than or equal to
<code>></code>	greater than
<code>>=</code>	greater than or equal to
<code>==</code>	exactly equal to
<code>!=</code>	not equal to
<code>!x</code>	Not x
<code>x y</code>	x OR y
<code>x & y</code>	x AND y
<code>isTRUE(x)</code>	test if X is TRUE

Logical Operators

➤ **Creating New variables:-**

Use the assignment operator “`<-`” to create new variables.

For example,

```
mydata$sum <- mydata$x1 + mydata$x2
```

New variable is created using two already available variables.

➤ **Modifying existing variable:-**

We can rename the existing variable by `rename()` function.

For examples,

```
mydata<- rename(mydata, c(oldname="newname"))
```

We can also recode variables in R.

For example,

If we want to rename variable based on some criteria like below

```
mydata$agecat<- ifelse(mydata$age> 70, c("older"), c("younger"))
```

Activity



- Create a new variable Total_Sales using variables Sales_1 and Sales_2.
- Then, modify the above new variable name from Total_Sales to Sales_Total.

Outliers and Missing Data treatment

Inputting missing data using standard methods and algorithmic approaches (mice package):

- In R, missing values are represented by the symbol **NA** (not available).
- Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number).
- Unlike SAS, R uses the same symbol for character and numeric data.

To test if there is any missing in the dataset we use `is.na ()` function.

For Example,

We have defined “y” and then checked if there is any missing value. T or True means that there is a missing value.

```
y <- c(1,2,3,NA)
is.na(y)
# returns a vector (F FF T)
```

Arithmetic functions on missing values yield missing values.**For Example,**

```
x <- c(1,2,NA,3)
mean(x)
# returns NA
```

To remove missing values from our dataset we use *na.omit()* function.

For Example,

We can create new dataset without missing data as below: -

```
newdata<- na.omit(mydata)
```

Or, we can also use “na.rm=TRUE” in argument of the operator. From above example we use na.rm and get desired result.

```
x <- c(1,2,NA,3)
mean(x, na.rm=TRUE)

# returns 2
```

MICE Package -> Multiple Imputation by Chained Equations

MICE uses PMM to impute missing values in a dataset.

PMM-> Predictive Mean Matching (PMM) is a semi-parametric imputation approach. It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model.

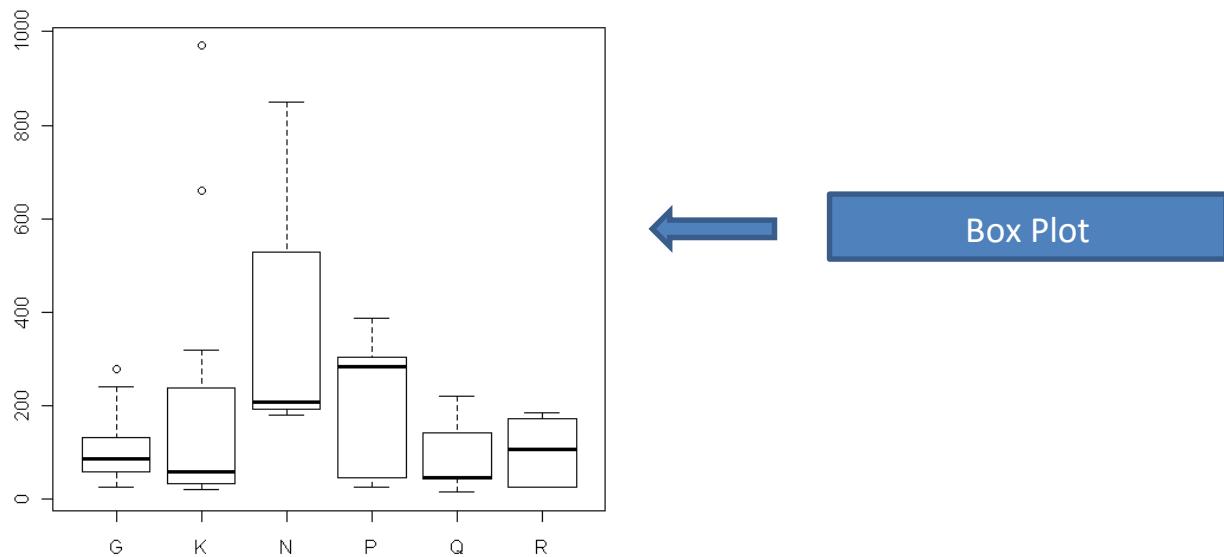
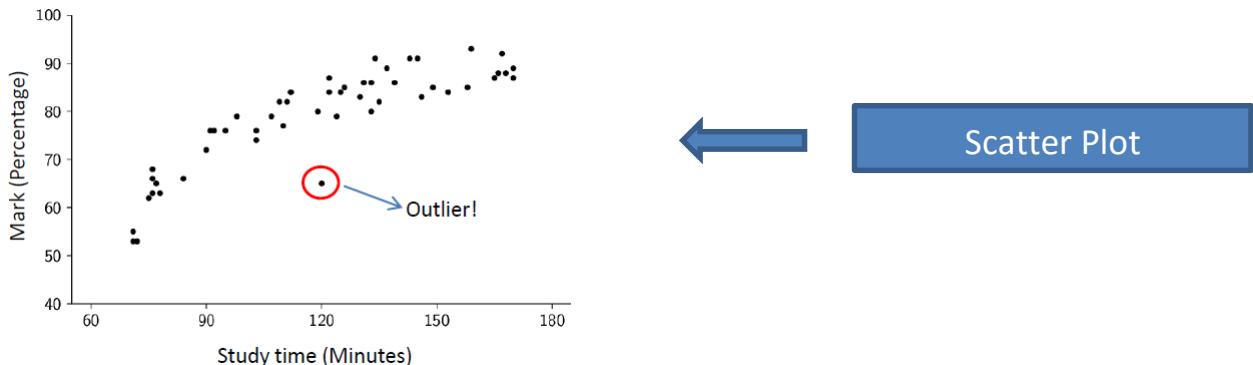
Outliers:

Outlier is a point or an observation that deviates significantly from the other observations.

- Due to experimental errors or “special circumstances”
- Outlier detection tests to check for outliers
- Outlier treatment –
 - Retention
 - Exclusion
 - Other treatment methods
 -

We have “OUTLIER” package in R to detect and treat outliers in Data.

Normally we use BOX Plot and Scatter plot to find outliers from graphical representation.



Combining Data sets in R

- To merge two data frames (datasets) horizontally, use the **merge** function. In most cases, you join two data frames by one or more common key variables (i.e., an inner join).

For example,

To merge two data frames by ID:

```
total <- merge(data frameA,data frameB,by="ID")
```
- To merge on more than one criteria we pass the argument as follows

To merge two data frames by ID and Country:

```
total <- merge(data frameA,data frameB,by=c("ID","Country"))
```

- To join two data frames (datasets) vertically, use the **rbind** function. The two data frames **must** have the same variables, but they do not have to be in the same order.

For example,

```
total <- rbind(data frameA, data frameB)
```

Note:-

If data frameA has variables that data frameB does not, then either:

- Delete the extra variables in data frameA or
- Create the additional variables in data frameB and set them to NA (missing)

before joining them with rbind().

We use **cbind()** function to combine data by column the syntax is same as **rbind()**.

Plyr package: Tools for Splitting, Applying and Combining Data.

We use rbind.fill() in plyr package in R. It binds or combines a list of data frames filling missing columns with NA.

For example,

```
rbind.fill(mtcars[c("mpg", "wt")], mtcars[c("wt", "cyl")])
```

In this all the missing value will be filled with NA.

Discuss Function and Loops

Using for and ifelse in R :

“FOR” Loop:-

To repeat an action for every value in a vector by using a “for” loop.

We construct a “for” loop in R as follows:

```
for(i in values){  
... do something ...  
}
```

This for loop consists of the following parts:

- The keyword for, followed by parentheses.
- An identifier between the parentheses. In this example, we use i, but that can be any object name you like.
- The keyword in, which follows the identifier.
- A vector with values to loop over. In this example code, we use the object values, but that again can be any vector you have available.
- A code block between braces that has to be carried out for every value in the object values.

In the code block, you can use the identifier. Each time R loops through the code, R assigns the next value in the vector with values to the identifier.

“IFELSE” Function:-

When using R, sometimes we need our function to do something if a condition is true and something else if it is not.

You could do this with two if statements, but there's an easier way in R: an if...else statement.

An if...else statement contains the same elements as an if statement), and then some extra:

- The keyword else, placed after the first code block
- A second block of code, contained within braces, that has to be carried out if and only if the result of the condition in the if() statement is FALSE

For example,

```
if(hours > 100) net.price<- net.price * 0.9  
if(public) {tot.price<- net.price * 1.06 } else  
{tot.price<- net.price * 1.12} round(tot.price)}
```

Or it can be written as also,

```
if(public) tot.price<- net.price * 1.06 else  
tot.price<- net.price * 1.12
```

Check Your Understanding

1. Which of the following is not a functionality of R?
 - a. Linear modeling
 - b. Nonlinear modeling
 - c. Developing webapplications
2. Which of the following is the function to display the first 5 rows of data?
3. What is the difference between rbind and cbind?
4. What are 2 ways of looping in R?

Case Study:

Segregate Sepal length based on the “Sepal Length” and “Sepal Width”. Use the dataset named “IRIS”.

Data Set

Sepal length	Sepal width	Petal length	Petal width	Species
7.9	3.8	6.4	2	I. virginica
7.7	3	6.1	2.3	I. virginica
5.6	2.5	3.9	1.1	I. versicolor
5.6	2.8	4.9	2	I. virginica
5.5	4.2	1.4	0.2	I. setosa
5.5	3.5	1.3	0.2	I. setosa
7.1	3	5.9	2.1	I. virginica
7	3.2	4.7	1.4	I. versicolor

Summary

- R is managed and maintained by CRAN.
- It is a freeware as well as open source software.
- We use “>” to start ant new code in R.
- We use “<-” as an assignment operator.
- A data frame is used for storing data tables.
- Rbind and Cbind are used to join 2 or more datasets.
- Outliers are extraordinary situations.
- IFELSE is used to 2 conditions simultaneously.

1. Divide the class into groups of 4-5 participants
2. Give the Dataset to the participants.
3. Give 10 minutes to the class for each group to discuss the steps that they would follow for Missing data treatment and Outlier Treatment process along with a discussion on the methods type which they would like to use
4. Each group presents their steps (5 min each)

Activity



Module 1: Unit– 1.2

Manage your work to meet requirements

Topic	Session Goals
Manage your work to meet requirements	<p>By the end of this session, you will be able learn about:</p> <ol style="list-style-type: none"> 1. Time Management 2. Planning and organizing your work 3. Expectation setting and measuring output

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Activity	Location
<ul style="list-style-type: none"> ✓ Discuss the significance of time management ✓ Create awareness on basic time management techniques ✓ Summarize the appropriate discussion points from the breakout sessions 	✓ Classroom
<ul style="list-style-type: none"> ✓ Discuss importance prioritization and planning. ✓ How to operationalize the plan. ✓ Create awareness on how to monitor performance 	✓ Classroom
<ul style="list-style-type: none"> ✓ Discuss importance of expectation setting ✓ Develop understanding on defining activities to be performed, deliverables and yardsticks of measuring output. ✓ Create awareness on the common Service Level Agreements 	✓ Classroom

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided: reflect on your own experiences and determine when to share them.**
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.**
- ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.**
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they

have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussion questions independently, rather than providing answer. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Topic: Welcome and Introduction

Time Management

Welcome the participants to the course and move to the introductions.

Introductions



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

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

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]



“Why are you here today?”

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

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

Suggested Responses/Benefits to Debrief:

The benefits of this course include:

- Efficient and Effective time management
- Efficient – Meeting timelines
- Effective – Meeting requirement for desired output
- Awareness of the SSC environment and time zone understanding
- Awareness of the SSC environment and importance of meeting timelines to handoffs

Review the course objectives listed above.

“To fulfill these objectives today, we’ll be conducting a number of hands-on activities. Hopefully we can open up some good conversations and some of you can share your experiences so that we can make this session as interactive as possible. Your participation will be crucial to your learning experience and that of your peers here in the session today.”

Ice Breaker – Open Discussion of 2 points

“Please share your thoughts on following.”



After participants give their views, debrief and bring to consensus

Question: Please share your thoughts on following?

- A. Time is perishable – Cannot be created or recovered
- B. Managing is only option – Prioritize



“Yes, I have room in my schedule to attend a Time Management Seminar...the day after I retire!”

Importance of Time Management

Provide a brief overview of the session. Discuss the importance of better utilization of time as the only tool to prevention of slippages on timelines.



Open up the discussion for the session and ask participants to share their thoughts on “time management”?

The first part of this session discusses the following:

- “Plan better avoid wastage”
- Understanding the timelines of the deliverables. Receiving the hand off from upstream teams at right time is critical to start self contribution and ensure passing the deliverables to downstream team.
- It is important to value others’ time as well to ensure overall organizational timelines are met
- Share the perspective of how important is time specifically in a global time zone mapping scenario



Why Time Management?



Ask the question to the participants and gather responses.

Discuss the responses with the group to understand the significance of time management.

Share the SSC model and how working along several time zones is important for the Shared Services Center.

Activity Description:

1. Refer to the **Aspects of Time Management** table in the Student Workbook and identify the rules that employees/workers must follow.
2. Refer to the **Vocabulary Words** table if you do not understand the meaning of a word/term.

Suggested Responses:

- Time management has to be looked at an organizational level and not just individual level
- These Aspects teach us how to build the blocks of time management.

Time Management Aspects

Prompt participants to come up with some aspects and relate them back to here.

- Planning and goal setting
- Managing yourself
- Dealing with other people
- Your time
- Getting results

The first 4 Interconnect and Interact to give the 5th one – Results

Differentiate between Urgent and Important task

Urgent task

- Assume importance as they demand immediate attention

Important Task

- May become urgent if left undone
- Usually have a long term effect

To judge importance vs. urgency, gauge tasks in terms of

- Impact of doing them
- Effect of not doing them

Main aim of prioritization is to avoid a crisis

We must

**Schedule our Priorities
as opposed to
Prioritizing our Schedule**

Time Management quadrants

1. Urgent and Important – Do Now
2. Not Urgent and Important – Schedule on your calendar
3. Urgent and Not Important – Delegate, Automate or Decline
4. Not Urgent Not Important – Delegate, Automate or Decline

Check Your Understanding



1. True or False? Time can be stored.
 - a. True
 - b. False

Suggested Responses:

False – Time once lost cannot be gotten back – hence important to plan time utilization properly



2. True or False? Time is perishable
 - a. True
 - b. False

Suggested Responses:

True – Time lost is lost for every – lost moments cannot be gotten back



3. True or False? Time management is required both at individual level and organizational level.
 - a. True
 - b. False

Suggested Responses:

True – plan for activities organizational level and also at individual level



4. True or False? Activities should be judged basis Urgency and Importance
 - c. True
 - d. False

Suggested Responses:

True – prioritization should be based on 2x2 matrix of urgency and importance

Team Exercise



List the items and ask participants to classify them as per the quadrant.

Ask the participants to pick up the items listed below and place them in the Urgent/Important quadrant. Discuss the rationale of their thoughts and categorization.

Activity Description:

1. Refer to the Time Management Quadrant on the display / in the Student Workbook and categorize the below items.
2. Refer to the Vocabulary Words table if you do not understand the meaning of a word/term.

Create teams of 2 participants and share the below list

Categorize the below items in the Time Management Quadrant

1. Wildly important goal
2. Last minute assignments from boss
3. Busy work
4. Personal health
5. Pressing problems
6. Crises
7. Planning
8. Time wasters
9. Professional development
10. Win-win performance agreement
11. Too many objectives
12. Vital customer call
13. Major Deadlines
14. Unimportant pre scheduled meetings
15. Meaningless management reports

16. Coaching and mentoring team
17. Low priority email
18. Other people's minor issues
19. Workplace gossip
20. Exercise
21. Needless interruptions
22. Defining contribution
23. Aimless Internet surfing
24. Irrelevant phone calls

Suggested Answers:

Depends on rationale shared

1. Wildly important goal – Q1
2. Last minute assignments from boss – Q1
3. Busy work – Q4 – Consumes time however not pressing
4. Personal health – Q4 – requires planning and care not pressing
5. Pressing problems – Q1 – has to be solved immediately
6. Crises – Q1 – have to tended to immediately
7. Planning – Q2 – Important but not urgent; should be done before crisis
8. Time wasters – Q4
9. Professional development – Q2
10. Win-win performance agreement – Q2 – Expectation setting part of planning
11. Too many objectives – Q3 – Prioritize further to establish which are important and pressing
12. Vital customer call – Q1 – Customer centricity
13. Major Deadlines – Q1
14. Unimportant pre scheduled meetings – Q3
15. Meaningless management reports – Q3 – Prioritize further to establish which are important and pressing
16. Coaching and mentoring team – Q2
17. Low priority email – Q3 – Prioritize further to establish which are important and pressing
18. Other people's minor issues – Q3 – May not be urgent but important for team building
19. Workplace gossip – Q4 – Non value add; occasionally creates negativity
20. Exercise – Q4 – Important for health and personal well being. To be done in spare and leisure time. Cannot be ignored.
21. Needless interruptions – Q3
22. Defining contribution – Q2
23. Aimless Internet surfing – Q4

24. Irrelevant phone calls – Q4 – Reserve and avoid



Summary

- It is important to manage time.
- To manage time one must:
 - Prioritize
 - Define Urgency
 - Define Importance

Work Management and Prioritization



Ask participants to define a job and split it into activities.

Preparing morning tea is a good example. Define time, no of family members, preparation required at night and then in the morning. Perfect execution to ensure good morning tea !!! with family.

Gather responses.

Start the session by connecting the course content to the candidate responses.

Work Management

Six steps for expectation setting with the stakeholders

1. Describe the jobs in terms of major outcomes and link to the organization's need

The first step in expectation setting is to describe the job to the employees. Employees need to feel there is a greater value to what they do. We need to feel out individual performance has an impact on the organization's mission.

Answer this question: My work is the key to ensuring the organization's success because...

While completing the answer link it to

- Job Description
- Team and Organization's need
- Performance Criteria

2. Share expectations in terms of work style

While setting expectation, it's not only important to talk about the "what we do" but also on "how we expect to do it". What are the ground rules for communication at the organization?

Sample ground rules

- Always let your team know where are the problems. Even if you have a solution, no one likes surprises.
- Share concerns openly and look for solutions
- If you see your colleagues doing something well, tell them. If you see them doing something poorly, tell them.

Sample work style questions

- Do you like to think about issues by discussing them in a meeting or having quite time alone?
- How do you prefer to plan your day?

3. Maximize Performance - Identify what is required to complete the work: Supervisor needs / Employee needs. Set input as well as output expectations

In order to ensure employees are performing at their best, the supervisor needs to provide not only the resource (time, infrastructure, desk, recognition etc.) but also the right levels of direction (telling how to do the task) and support (engaging with employees about the task).

4. Establish priorities. Establish thresholds and crisis plan

Use the time quadrant to establish priorities. Refer to earlier session.

5. Revalidate understanding. Create documentation and communication plan to establish all discussion

When you are having a conversation about expectations with stakeholders, you're covering lot of details so you'll need to review to make sure you both have a common understanding of the commitments you have made.

6. Establish progress check

No matter how careful you have been in setting expectations, you'll want to follow up since there will be questions as work progresses.

Schedule an early progress check to get things started the right way, and agreed on scheduled/unscheduled further checks. Acknowledge good performance and point your ways to improve



Check Your Understanding



1. True or False? Setting expectations is best done after the employee has worked for 6 months.
- a. True
 - b. False

Suggested Responses:

False, work expectations have to be set from day 1 – so that roles & responsibilities are clear



2. True or False? Do not provide too many details when setting expectations.
- a. True
 - b. False

Suggested Responses:

False, as much details with examples can help clarify all the expectations



3. True or False? Always check to make sure there is a common understanding of expectations.
- a. True
 - b. False

Suggested Responses:

True, asking the person to rearticulate the understanding of expectations is best way to ensure there is clear understanding on both sides



4. True or False? Try not to ask too many questions while setting expectations.
- a. True
 - b. False

Suggested Responses:

False, questions always to be encouraged to ensure any clarifications are responded



5. True or False? Employees need to know what tasks to do and how to communicate, appreciating work styles.
- a. True
 - b. False

Suggested Responses:

True, provides clarity and enables response based on work styles



6. True or False? Employees do not need to know how their work contributes to organizational results.
- a. True
 - b. False

Suggested Responses:

False, linking efforts with common goals is very motivating and develops team effort



7. True or False? Employees need to know what their team members' performance problems are.
- a. True
 - b. False

Suggested Responses:

True, knowing common problems bring teams focused towards solutions.



8. True or False? Employees how have work style different from the Boss/Peers need to change.
- a. True
 - b. False

Suggested Responses:

False, they need to adapt and respond based on the partners work style – understanding the work styles is very critical to enhance team operating performance.

Summary

- Define work and activities:
 - What
 - How
- Define Stakeholders and participants:
 - Whom to serve
 - Who all are serving
- Plan, Execute and Monitor

Quality Standards Adherence

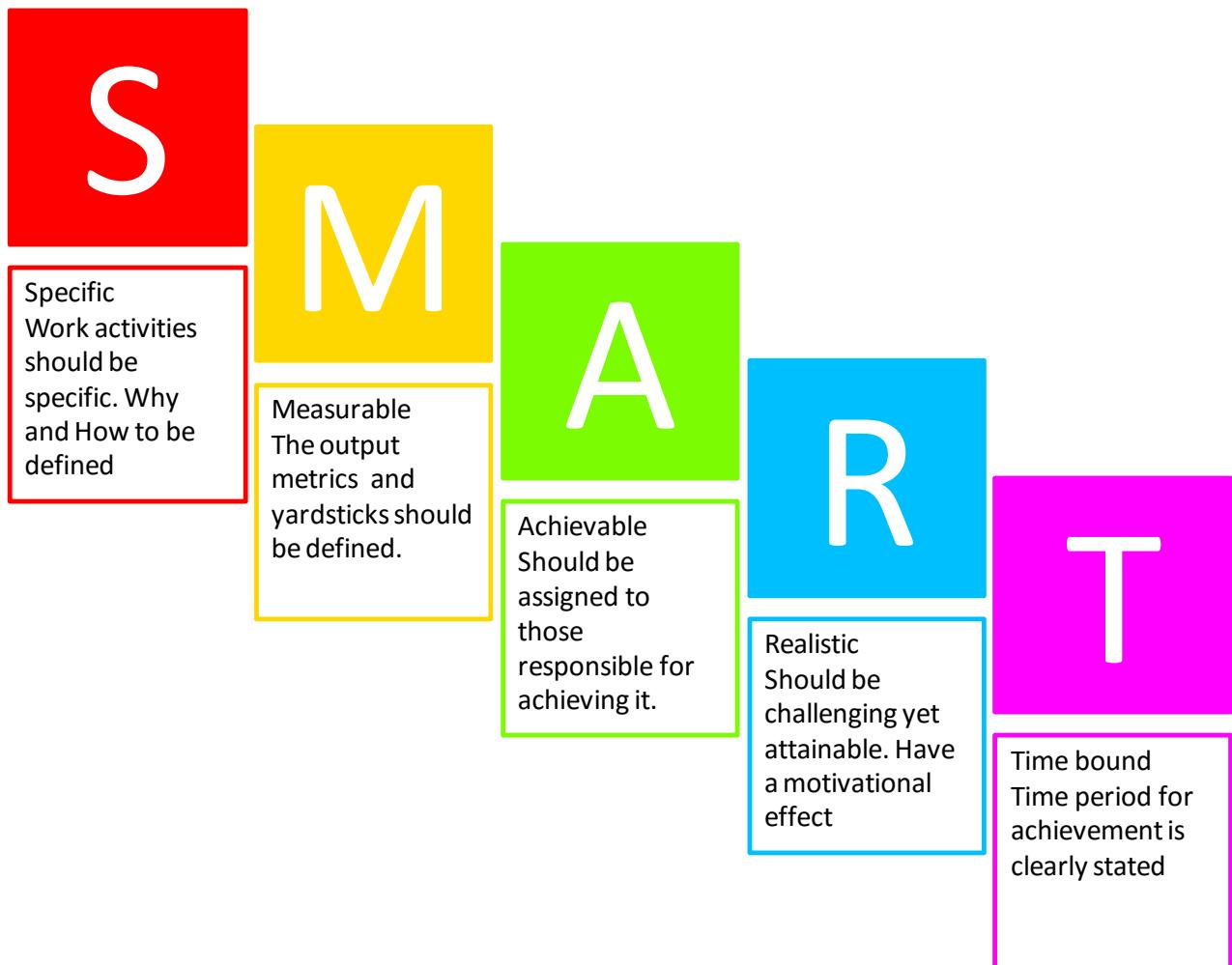
Let's Get Started

Provide a brief overview of the session. Discuss key points on importance of defining quality output. Re



- Iterate importance of Efficiency and Effectiveness
 - Efficiency – Performing activities well
 - Effectiveness – Performing activities right

Goals and Objectives compliant to SMART



Efficiency vs Effectiveness

		In efficient	Efficient
		Pursuing right goals but in efficient	Pursuing right goals and efficient
Pursuit of Appropriate Goals / Doing the Right Thing	Effective	Pursuing wrong goals and inefficient.	Pursuing Wrong goals but is Efficient
	In Effective	In efficient	Efficient

Use of Resources / Doing Things Right

Service Level Agreements

Service Level Agreement (SLA) is a contract between a service provider and its internal or external customers that documents what services the provider will furnish

SLA measures the service provider's performance and quality in a number of ways.

Some sample metrics SLAs may specify or include

- Availability and uptime – the percentage of the time services will be available
- The no of users being served, the bandwidth or volume being addressed or the quantum of work being performed in work units
- Specify performance benchmarks to which actual performance will be periodically compared
- Turnaround time

In addition to establishing performance metrics, an SLA may include a plan for addressing downtime and documentation for how the service provider will compensate customers in the event of a contract breach. SLAs, once established, should be periodically reviewed and updated to reflect changes in technology and the impact of any new regulatory directives



Summary

- Every activity must have defined goals and objectives. These goals and objectives should be SMART complaint (Specific, Measurable, Achievable, Realistic and Time-bound).
- One must balance the efficiency and effectiveness while performing the tasks to achieve the desired objectives.
- The Service Level Agreements should be clearly laid out to measure the quality and performance.

Course Conclusion

Course Conclusion



“We’ve almost reached the end of the course! Before we wrap up, let’s review what we’ve learned today”

Ask the participants to recall key learning points from the session and map these learning points to the course objectives.

Thank You Note

Module 1 - Unit – 2.1

Summarizing Data and Revisiting Probability

Topic	Session Goals
Summarizing Data, and Revisiting Probability	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Summarize Data 2. Work on Probability.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Summary statistics- summarizing data with R	Classroom
Probability	Classroom
Expected value	Classroom
Random & Bivariate Random Variables	Classroom
Probability Distribution	Classroom
Normal Distribution	Classroom
Central Limit Theorem	Classroom
Random walk	Classroom
Check your understanding	Classroom
Summary	Classroom

Summary statistics- summarizing data with R

- summary(data_frame)
- summary(iris)
- Output : Mean, Median, Minimum, Maximum, 1st and 3rd quartile

[CODE]

>summary(dataset)

For example ,

```
> summary(iris)
   Sepal.Length    Sepal.Width     Petal.Length     Petal.Width
Min.   :4.300    Min.   :2.000    Min.   :1.000    Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
Species
setosa      :50
versicolor  :50
virginica   :50
```

Probability

Probability is the chance of occurrence anything.

$$P(A) = S/P$$

Where S is sample size or no of positive outcomes and P is the population size or total no of outcomes.

A **probability distribution** describes how the values of a random variable are distributed.

For example, the collection of all possible outcomes of a sequence of coin tossing is known to follow the binomial distribution. Whereas the means of sufficiently large samples of a data population are known to resemble the normal distribution. Since the characteristics of these theoretical distributions are well understood, they can be used to make

Statistical inferences on the entire data population as a whole.

For example,

Probability of ace of Diamond in a pack of 52 cards when 1 card is pulled out at random.

Now, “At Random” means that there is no biased treatment with any card and the result will be totally at random.

So, No. of Ace of Diamond in a pack = S = 1

Total no of possible outcomes = Total no. of cards in pack = 52

Probability of positive outcome = S/P = 1/52

That is we have 1.92% chance that we will get positive outcome.

Expected value

The expected value of a random variable is intuitively the long-run average value of repetitions of the experiment it represents.

For example, the expected value of a dice roll is 3.5 because, roughly speaking, the average of an extremely large number of dice rolls is practically always nearly equal to 3.5.

Less roughly, the law of large numbers guarantees that the arithmetic mean of the values almost surely converges to the expected value as the number of repetitions goes to infinity.

The expected value is also known as the expectation, mathematical expectation, EV, mean, or first moment.

More practically, the expected value of a discrete random variable is the probability-weighted average of all possible values. In other words, each possible value the random variable can assume is multiplied by its probability of occurring, and the resulting products are summed to produce the expected value. The same works for continuous random variables, except the sum is replaced by an integral and the probabilities by probability densities. The formal definition subsumes both of these and also works for distributions which are neither discrete nor continuous: the expected value of a random variable is the integral of the random variable with respect to its probability measure. The expected value is a key aspect of how one characterizes a probability distribution; it is a location parameter.

Random & Bivariate Random Variables

➤ **Random Variable:**

- A random variable, aleatory variable or stochastic variable is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense). A random variable can take on a set of possible different values (similarly to other mathematical variables), each with an associated probability, in contrast to other mathematical variables.
- A *random variable* is a real-valued function defined on the points of a sample space.

Random variables can be discrete, that is, taking any of a specified finite or countable list of values, endowed with a probability mass function, characteristic of a probability distribution; or continuous, taking any numerical value in an interval or collection of intervals, via a probability density function that is characteristic of a probability distribution; or a mixture of both types. The realizations of a random variable, that is, the results of randomly choosing values according to the variable's probability distribution function, are called random variates.

For Example,

If we toss a coin for 10 times and we get heads 8 times then we cannot say that the 11th time if coin is tossed then we get a head or a tail. But we are sure that we will either get a head or a tail.

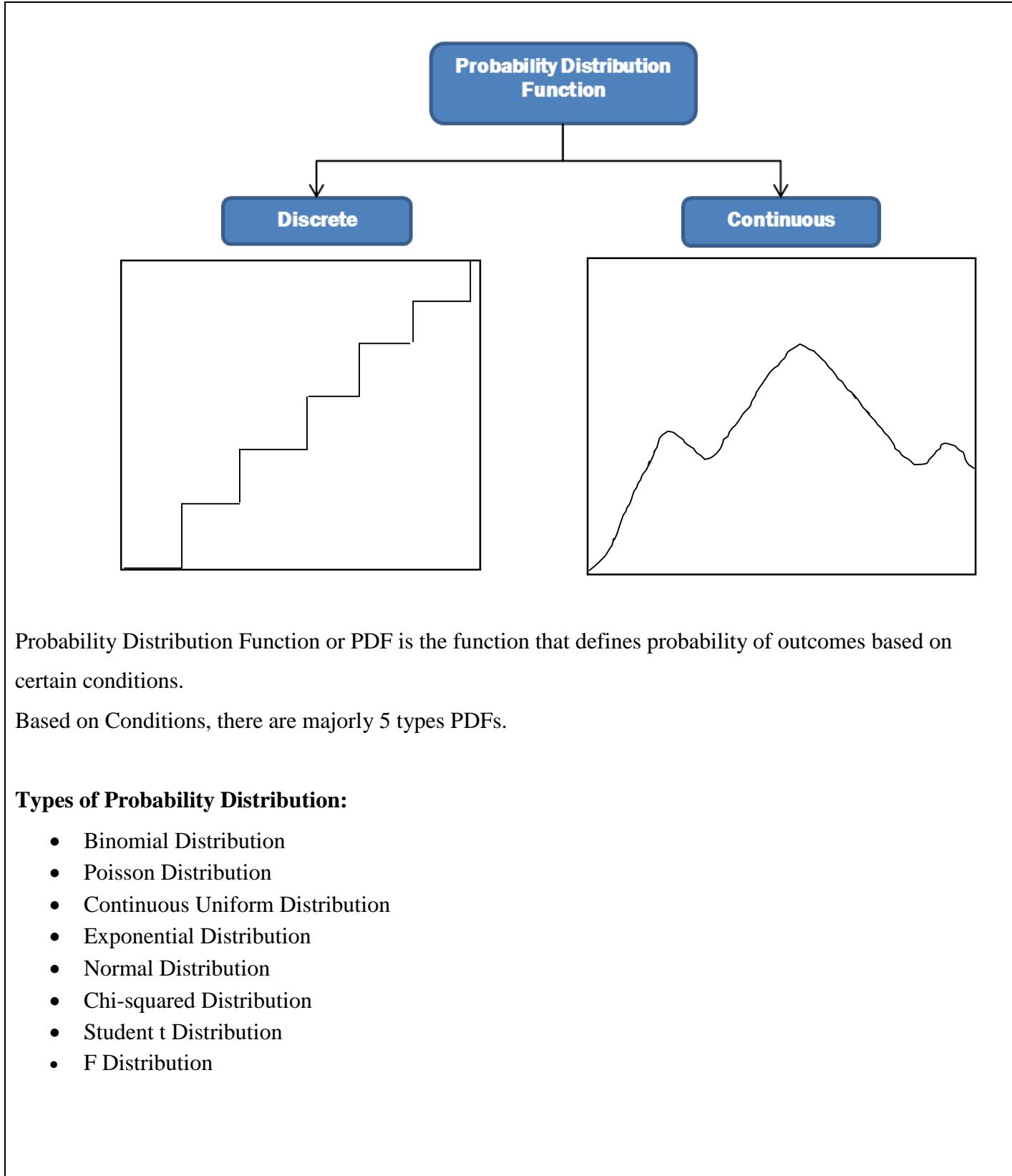
➤ **Bivariate Random Variable:**

Bivariate Random Variables are those variables having only 2 possible outcomes. For example flip of coin.

Probability Distribution

There are 2 types of Distribution Functions:-

1. Discrete
2. Continuous



Normal Distribution

We come now to the most important continuous probability density function and perhaps the most important probability distribution of any sort, the normal distribution.

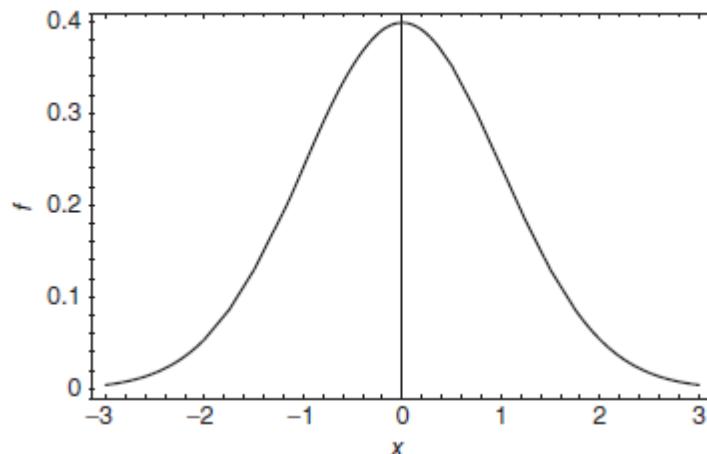
On several occasions, we have observed its occurrence in graphs from, apparently, widely differing sources: the sums when three or more dice are thrown; the binomial distribution for large values of n ; and in the hyper geometric distribution.

There are many other examples as well and several reasons, which will appear here, to call this distribution “normal.”

If

$$f(x) = \frac{1}{b \cdot \sqrt{2 \cdot \pi}} e^{-\frac{1}{2b^2}(x-a)^2}, \quad -\infty < x < \infty, \quad -\infty < a < \infty, \quad b > 0,$$

We say that X has a *normal* probability distribution. A graph of a normal distribution, where we have chosen $a = 0$ and $b = 1$, appears in figure below:

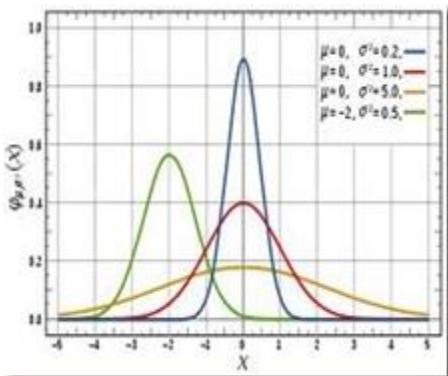


Normal Curve or Bell Curve

The shape of a normal curve is highly dependent on the standard deviation.

Importance of Normal Distribution:

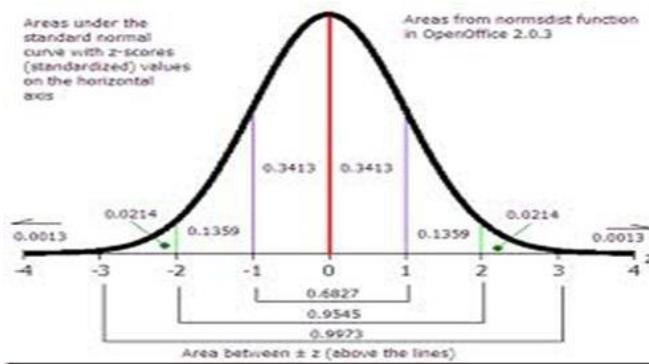
Normal distribution is a continuous distribution that is “bell-shaped”. Data are often assumed to be normal. Normal distributions can estimate probabilities over a *continuous interval* of data values.



Normal Distribution curves with different conditions

Also we can say A standard normal distribution is a normal distribution with :

- ✓ Mean=0 and
- ✓ Standard deviation = 1.



Standard Normal Distribution

The normal distribution $f(x)$, with any mean μ and any positive deviation σ , has the following properties:

- ❖ It is symmetric around the point $x = \mu$, which is at the same time the mode, the median and the mean of the distribution.
- ❖ It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$.
- ❖ Its density has two inflection points (where the second derivative of is zero and changes sign), located one standard deviation away from the mean, namely at $x = \mu - \sigma$ and $x = \mu + \sigma$.
- ❖ Its density is log-concave.
- ❖ Its density is infinitely differentiable, indeed super smooth of order 2.
- ❖ Its second derivative $f''(x)$ is equal to its derivative with respect to its variance σ^2 .

Test of Normal Distribution:

Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

- More precisely, the tests are a form of model selection, and can be interpreted several ways, depending on one's interpretations of probability:
- In descriptive statistics terms, one measures a goodness of fit of a normal model to the data – if the fit is poor then the data are not well modeled in that respect by a normal distribution, without making a judgment on any underlying variable.

- In frequentist statistics statistical hypothesis testing, data are tested against the null hypothesis that it is normally distributed.
- In Bayesian statistics, one does not "test normality" per se, but rather computes the likelihood that the data come from a normal distribution with given parameters μ, σ (for all μ, σ), and compares that with the likelihood that the data come from other distributions under consideration, most simply using a Bayes factor (giving the relative likelihood of seeing the data given different models), or more finely taking a prior distribution on possible models and parameters and computing a posterior distribution given the computed likelihoods.

1. Graphical methods:

An informal approach to testing normality is to compare a histogram of the sample data to a normal probability curve. The empirical distribution of the data (the histogram) should be bell-shaped and resemble the normal distribution. This might be difficult to see if the sample is small. In this case one might proceed by regressing the data against the quartiles of a normal distribution with the same mean and variance as the sample. Lack of fit to the regression line suggests a departure from normality.(see Anderson Darling coefficient and Minitab)

A graphical tool for assessing normality is the normal probability plot, a quantile-quantile plot (QQ plot) of the standardized data against the standard normal distribution. Here the correlation between the sample data and normal quartiles (a measure of the goodness of fit) measures how well the data are modeled by a normal distribution. For normal data the points plotted in the QQ plot should fall approximately on a straight line, indicating high positive correlation. These plots are easy to interpret and also have the benefit that outliers are easily identified.

Back-of-the-envelope test

Simple back-of-the-envelope test takes the sample maximum and minimum and computes their z-score, or more properly t-statistic (number of sample standard deviations that a sample is above or below the sample mean), and compares it to the 68–95–99.7 rule: if one has a 3σ event (properly, a $3s$ event) and substantially fewer than 300 samples, or a $4s$ event and substantially fewer than 15,000 samples, then a normal distribution will underestimate the maximum magnitude of deviations in the sample data.

This test is useful in cases where one faces kurtosis risk – where large deviations matter – and has the benefits that it is very easy to compute and to communicate: non-statisticians can easily grasp that " 6σ events are very rare in normal distributions".

2. Frequentist tests:

Tests of univariate normality include D'Agostino's K-squared test, the Jarque–Bera test, the Anderson–Darling test, the Cramér–von Mises criterion, the Lilliefors test for normality (itself an adaptation of the Kolmogorov–Smirnov test), the Shapiro–Wilk test, the Pearson's chi-squared test, and the Shapiro–Francia test. A 2011 paper from The Journal of Statistical Modeling and Analytics concludes that Shapiro-Wilk has the best power for a given significance, followed closely by Anderson-Darling when comparing the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests.

Some published works recommend the Jarque–Bera test. But it is not without weakness. It has low power for distributions with short tails, especially for bimodal distributions. Other authors have declined to include its data in their studies because of its poor overall performance.

Historically, the third and fourth standardized moments (skewness and kurtosis) were some of the earliest tests for normality. The Jarque–Bera test is itself derived from skewness and kurtosis estimates. Mardia's multivariate skewness and kurtosis tests generalize the moment tests to the multivariate case. Other early test statistics include the ratio of the mean absolute deviation to the standard deviation and of the range to the standard deviation.

More recent tests of normality include the energy test (Székely and Rizzo) and the tests based on the empirical characteristic function (ecf) (e.g. Epps and Pulley, Henze–Zirkler, BHEP test). The energy and the ecf tests are powerful tests that apply for testing univariate or multivariate normality and are statistically consistent against general alternatives.

The normal distribution has the highest entropy of any distribution for a given standard deviation. There are a number of normality tests based on this property, the first attributable to Vasicek.

3. Bayesian tests:

Kullback–Leibler divergences between the whole posterior distributions of the slope and variance do not indicate non-normality. However, the ratio of expectations of these posteriors and the expectation of the ratios give similar results to the Shapiro–Wilk statistic except for very small samples, when non-informative priors are used.

Spiegelhalter suggests using a Bayes factor to compare normality with a different class of distributional alternatives. This approach has been extended by Farrell and Rogers-Stewart.

Central Limit Theorem

The central limit theorem states that under certain (fairly common) conditions, the sum of many random variables will have an approximately normal distribution.

More specifically, where X_1, \dots, X_n are independent and identically distributed random variables with the same arbitrary distribution, zero mean, and variance σ^2 ; and Z is their mean scaled by

$$Z = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)$$

Then, as n increases, the probability distribution of Z will tend to the normal distribution with zero mean and variance (σ^2).

The central limit theorem also implies that certain distributions can be approximated by the normal distribution, for example:

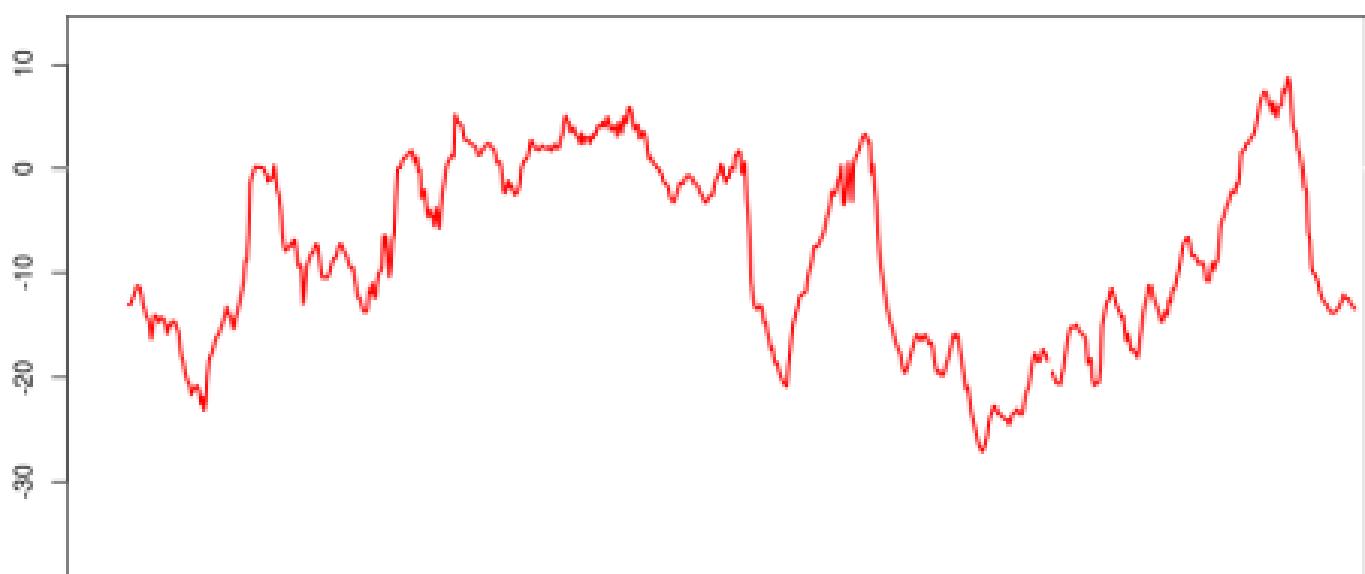
- The binomial distribution $B(n, p)$ is approximately normal with mean np and variance $np(1-p)$ for large n and for p not too close to zero or one.
- The Poisson distribution with parameter λ is approximately normal with mean λ and variance λ , for large values of λ .
- The chi-squared distribution $\chi^2(k)$ is approximately normal with mean k and variance $2k$, for large k .
- The Student's t-distribution $t(v)$ is approximately normal with mean 0 and variance 1 when v is large.

Random walk

A random walk is a mathematical formalization of a path that consists of a succession of random steps.

For example, the path traced by a molecule as it travels in a liquid or a gas, the search path of a foraging animal, the price of a fluctuating stock and the financial status of a gambler can all be modeled as random walks, although they may not be truly random in reality. The term random walk was first introduced by Karl Pearson in 1905.

Random walks have been used in many fields: ecology, economics, psychology, computer science, physics, chemistry, and biology.



Random Walk Graphical Representation

Application of Random Walk:

Applying the random walk theory to finance and stocks suggests that stock prices change randomly, making it impossible to predict stock prices. The random walk theory corresponds to the belief that markets are efficient, and that it is not possible to beat or predict the market because stock prices reflect all available information and the occurrence of new information is seemingly random as well

Check your understanding



1. What is central Limit Theorem?
2. What is Normal Distribution Curve and Why it is called as Bell Curve?
3. How to find Summary statistics in R?
4. What are the various types of Probability Distribution curves?
5. Why is Normal curve so widely used?

Summary

- •summary(data_frame)
- summary (iris)
- Output: Mean, Median, Minimum, Maximum, 1st and 3rd
- Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.
- A random walk is a mathematical formalization of a path that consists of a succession of random steps.
- Probability Distribution Function or PDF is the function that defines probability of outcomes based on certain conditions.
- A random variable, aleatory variable or stochastic variable is a variable whose value is subject to variations due to chance

Activity



1. Divide the class into groups of 4-5 participants
2. Give the Dataset to the participants.
3. Give 10 minutes to the class for each group to discuss the various random variable examples process along with a discussion on the methods type which they would like to use
4. Each group presents their examples with justification. (5 min)



CASE STUDY – Binomial Dist.

Sachin buys a chocolate bar every day during a promotion that says one out of six chocolate bars has a gift coupon within. Answer the following questions:

- What is the distribution of the number of chocolates with gift coupons in seven days?
- What is the probability that Amir gets no chocolates with gift coupons in seven days?
- Amir gets no gift coupons for the first six days of the week. What is the chance that he will get a one on the seventh day?

- Amir buys a bar every day for six weeks. What is the probability that he gets at least three gift coupons?
- How many days of purchase are required so that Amir's chance of getting at least one gift coupon is 0.95 or greater?

Solution:

Steps:

$$\text{Formula} = {}^nC_r p^r q^{n-r}$$

Where

- n is the no. of trials
- r is the number of successful outcomes
- p is the probability of success, and
- q is the probability of failure.

Other important formulae include

$$p + q = 1$$

$$\text{Hence, } q = 1 - p$$

$$\text{Thus, } p = 1/6$$

$$q = 5/6$$

1. Distribution of number of chocolates with gift coupons in 7 days: ${}^7C_r (1/6)^r (5/6)^{7-r}$
2. Probability of failing 7 days: $P(X=0) = (5/6)^7$
3. Probability of winning a coupon on the 7th day: $1/6$
4. The number of winning at least 3 wrappers in six weeks:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - (P(X=0) + P(X=1) + P(X=2)) \\ &= 1 - (0.0005 + 0.0040 + 0.0163) \\ &= 0.979 \end{aligned}$$

5. Number of purchase days required so that probability of success is greater than 0.95:

$$\begin{aligned} P(X \geq 1) \geq 0.95 &= 1 - P(X \leq 0) \geq 0.95 \\ &= 1 - P(X=0) \leq 0.05 \\ &= n \geq 16.43 \text{ (applying log function)} \\ &= 17 \text{ days.} \end{aligned}$$

Module 1 - Unit – 2.2

Work Effectively with Colleagues

Topic	Session Goals
Manage your work to meet requirements	<p>By the end of this session, you will be able learn about:</p> <ol style="list-style-type: none"> 1. Professionalism 2. Team Work 3. Effective Communication

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Topic description	Location
✓ Welcome participants to the course ✓ Recap of core skills through questions and Polling Questions ✓ Review learning objectives	✓ Classroom
✓ Importance of team above individual ✓ How to unlock group potential ✓ Compete as a team not within the team	✓ Classroom
✓ Being a Professional ✓ Importance of Grooming	✓ Classroom
✓ What is effective communication ✓ Verbal and Written Communication ✓ Common Communication Barriers	✓ Classroom
✓ Validate learning objectives have been met ✓ Make final summary remarks	✓ Classroom

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided:** reflect on your own experiences and determine when to share them.
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content.** Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.
- ✓ **Note that all examples are in italics to emphasize key learning points;** however, you may use your own professional experience to enhance the learning.
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussion questions independently, rather than providing answer. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Topic: Team Work

Working Effectively

Welcome the participants to the course and move to the introductions.

Introductions



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

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

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]



“Why are you here today?”

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

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

Debrief the following:

Why are teams more popular??

- Teams outperform individuals
- Teams use employee talent better
- Teams are more flexible and responsive to environmental changes in the organization.
- Teams facilitate employee involvement
- Teams are an excellent way to democratize an organization and increase motivation.

Topic: Team Work

Team Work



Ask participants to share their thoughts on:

- What is team work?
- How is it more advantageous?

What is a Team?

A team comprises a group of people linked in a common purpose.

Teams are especially appropriate for conducting tasks that are high in complexity and have many interdependent subtasks.



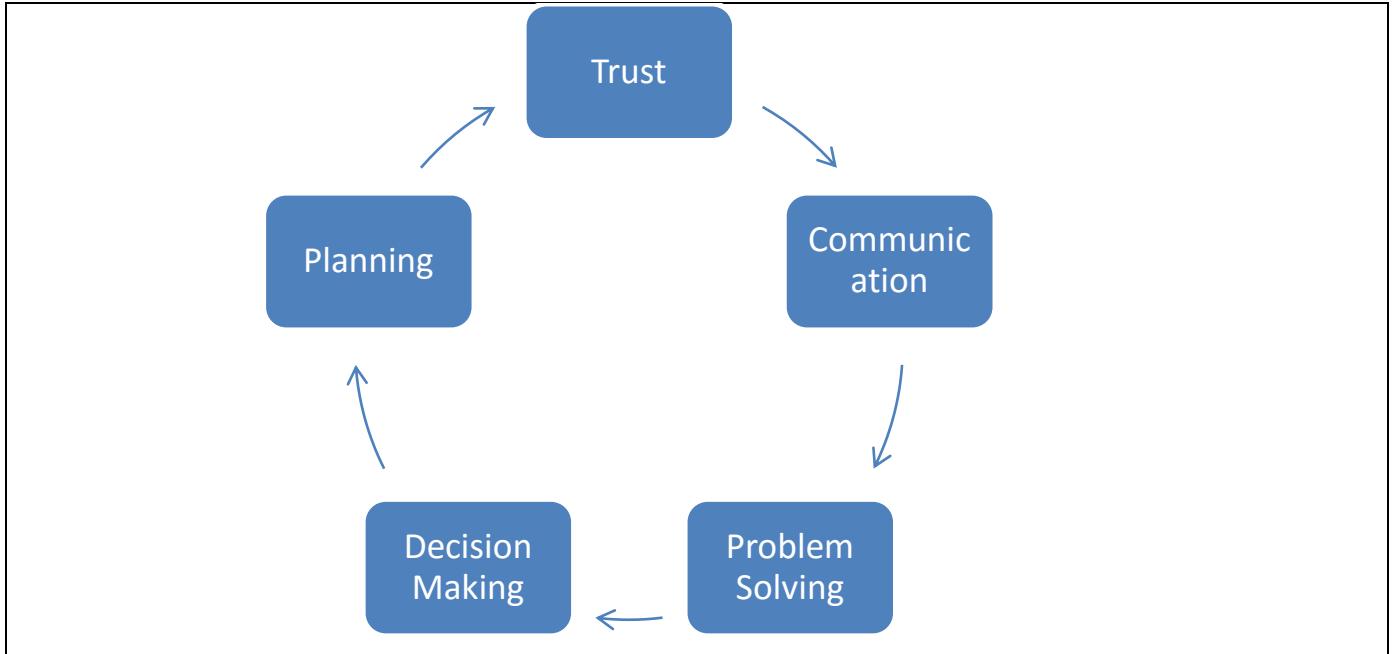
Team

Work

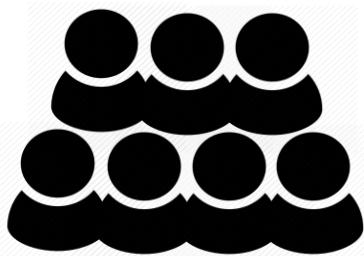
Coming together is a beginning, keeping together is progress and working together is success. A team is a number of people associated together in work or activity. In a good team members create an environment that allows everyone to go beyond their limitation.

Why do we need teamwork – The overriding need of all people working together for the same organization is to make the organization profitable.

Team Building



Team work vs. Individual work



- Team Work  • Individual work
- Team Work
 - Agree on goals/ milestones
 - Establish tasks to be completed
 - Communicate / monitor progress
 - Solve Problem
 - Interpret Results
 - Agree completion of projects
 - Individual work
 - Work on tasks
 - Work on new / revised tasks

Team Development

Team building is any activity that builds and strengthens the team as a team. The teams that are integrated in Spirit, Enthusiasm, Cohesiveness and Camaraderie are vitally important.

Team building fundamentals

- Clear Expectations – Vision/Mission
- Context – Background – Why participation in Teams?
- Commitment – dedication – Service as valuable to Organization & Own
- Competence – Capability – Knowledge
- Charter – agreement – Assigned area of responsibility
- Control – Freedom & Limitations
- Collaboration – Team work
- Communication
- Creative Innovation
- Consequences – Accountable for rewards
- Coordination
- Cultural Change

Roles of team member

- Communicate
- Don't Blame Others
- Support Group Member's Ideas
- No Bragging – No Full of yourself
- Listen Actively
- Get Involved
- Coach, Don't Demonstrate
- Provide Constructive Criticism
- Try To Be Positive
- Value Your Group's Ideas



Ice Breaker – Open Discussion of below points



“Please share your thoughts on following.”

After participants give their views, debrief and bring to consensus

Team Work: Pros and Cons

Check Your Understanding



1. True or False. The organizations that display a higher level of team work are generally more successful.

Suggested Answer:

True



2. Which one of the following is NOT a key attribute of Team Work?
 - a. Commitment
 - b. Communication
 - c. Vague expectations
 - d. Transparency

Suggested Answer:

c. Vague expectations

Summary

- A team comprises a group of people linked in a common purpose.
- Team work is essential to the success of every organization. In a good team, members create an environment that allows everyone to go beyond their limitation.
- Some of the fundamentals on which a team is built are: Collaboration, Clear Expectations and Commitment.

Key Points

Importance of Professionalism

Provide a brief overview of the session. Discuss the importance of professional behavior in the organization.

Activity

Activity Description:

Ask the candidates “What does professionalism mean to you?”

Take a few minutes and write down your thoughts... as a definition or description.

Summary

Who is professional?

A person who has achieved an acclaimed level of proficiency in any trade and whose competencies can be measured against fixed set of standards or guidelines. The following are the key characteristics in a person that make him stand out as a professional.

- **Positively proactive.** Professionals demonstrate behaviours that are positive, proactive instead of negative, and reactive.
- **Respect.** Through this ethic and value of respect, professionals are known and trusted within and without their respective organizations.
- **Opportunities to help others.** Those who avow before understand they have a responsibility to help others whether it is to grow self-leadership skills or provide some expert advice.
- **Follow-up.** No one likes to wait for un-returned phone calls or emails. Professionals make it a habit to follow-up on everything and accept responsibility when they fail to engage in that behavior.
- **Empathy.** Professionals know how to be empathetic. This characteristic is a one of the signs of high emotional intelligence and a predictor for leadership success.
- **Self-confident.** When individuals are self confident, they do not have to put others down at their own expense. These individuals have a high sense of balanced self-esteem and role awareness.
- **Sustainable.** Professionals are truly sustainable in that they can continue forward when times become difficult. Their ethics and beliefs keep them focused.
- **Integrity.** Integrity is putting your values into action; doing the right thing when no one else is looking without personal gain or benefit; and accepting a potential personal cost.
- **Optimize all interactions.** This is critical because professionals do not negate the value of people. They look to see how one interaction can benefit someone else even before himself or herself.

- **Nimble.** Being flexible and open to change allows these individuals to be quick on their feet and nimble to the opportunities that they encounter on a daily basis.
- **Awareness.** Having a high level of awareness of themselves, the marketplace, the community and even the world helps these individuals continually stay on top of things.
- **Leadership.** Last, but not least, professionals demonstrate exceptional leadership skills and even more importantly self-leadership skill. For if you cannot lead yourself, you cannot lead others.

What is professionalism?

- Professionalism is the competence or set of skills that are expected from a professional.
- Professionalism determines how a person is perceived by his employer, co-workers, and casual contacts.

How long does it take for someone to form an opinion about you?

- Studies have proved that it just takes six seconds for a person to form an opinion about another person.

How does someone form an opinion about you?

Eye Contact – Maintaining eye contact with a person or the audience says that you are confident. It says that you are someone who can be trusted and hence can maintain contact with you.

Handshake – Grasp the other person's hand firmly and shake it a few times. This shows that you are enthusiastic.

Posture – Stand straight but not rigid, this will showcase that you are receptive and not very rigid in your thoughts.

Clothing – Appropriate clothing says that you are a leader with a winning potential.

How to exhibit professionalism?

- Empathy
- Positive Attitude
- Teamwork
- Professional Language
- Knowledge
- Punctual
- Confident
- Emotionally stable

Grooming

What are the colours that one can opt for work wear?

A good rule of thumb is to have your pants, skirts and blazers in neutral colours. Neutrals are not only restricted to grey brown and off white - you can also take advantage of the beautiful navies, forest greens, burgundies, tans and caramel

tones around. Pair these neutrals with blouses, scarves or other accessories in accent colours - ruby red, purple, teal blue, soft metallic and pinks are some examples.

Things to remember

- Wear neat clothes at work which are well ironed and do not stink.
- Ensure that the shoes are polished and the socks are clean
- Cut your nails on a regular basis and ensure that your hair is in place.
- Women should avoid wearing revealing clothes at work.
- Remember that the way one presents oneself plays a major role in the professional world.

Check Your Understanding



1. True or False? Polo T-Shirt is professional dress.
 - a. True
 - b. False

Suggested Responses:

False



2. True or False? I can wear sandals to office
 - a. True
 - b. False

Suggested Responses:

False



3. True or False? Well tailored Salwar Suit is not professional.
 - a. True
 - b. False

Suggested Responses:

False

Activity Description:

Ask the participants to pick up the items listed below and place them in the Acceptable / Unacceptable category. Discuss the rationale of their thoughts and categorization.

Categorize the below items in the Acceptable / Unacceptable

1. Polo T Shirt –
2. Golf Shoes –
3. Collared Shirt -
4. Suede Shoes –
5. Leather laced Shoes –
6. Matching Socks –
7. Backpacks –
8. Lynards in Pockets –
9. Jeans on weekdays –
10. Rolled Up Sleeves –
11. Matching Belt and Shoes –
12. Pressed Suit –
13. Knee Length Skirt –
14. Short Skirts –
15. Obvious Tatoos –

Suggested Answers:

1. Polo T Shirt – Unacceptable
2. Golf Shoes – Unacceptable
3. Collared Shirt - Acceptable
4. Suede Shoes – Unacceptable
5. Leather laced Shoes – Acceptable
6. Matching Socks – Acceptable
7. Backpacks – Unacceptable
8. Lynards in Pockets – Unacceptable
9. Jeans on weekdays – Unacceptable
10. Rolled Up Sleeves – Unacceptable
11. Matching Belt and Shoes – Acceptable

12. Pressed Suit – Acceptable
13. Knee Length Skirt – Acceptable
14. Short Skirts – Unacceptable
15. Obvious Tatoos – Unacceptable

Summary

- Professionalism determines how a person is perceived by his employer and co-workers.
- Empathy, Positive Attitude, Teamwork, Professional Language, Knowledge, Punctuality, Confidence are some of the key characteristics that determine the professionalism of a person.
- The type of clothes you wear and grooming also plays an important role in forming an impression of the person in his/her work environment.

Key Points

Effective Communication



Provide a brief overview of the session.

Prompt candidates to discuss the consequences of ineffective or unclear communication.



We would probably all agree that effective communication is essential to workplace effectiveness. And yet, we probably don't spend much time thinking about how we communicate, and how we might improve our communication skills. The purpose of building communication skills is to achieve greater understanding and meaning between people and to build a climate of trust, openness, and support. To a large degree, getting our work done involves working with other people. And a big part of working well with other people is communicating effectively. Sometimes we just don't realize how critical effective communication is to getting the job done. So, let's have an experience that reminds us of the importance of effective communication. Actually, this experience is a challenge to achieve a group result without any communication at all! Let's give it a shot.

Activity Description:

Ask the participants to share an experience that reminds them of the significance of effective communication OR consequences of ineffective communication.

What is Effective Communication?

We cannot not communicate.

The question is: Are we communicating what we intend to communicate?

Does the message we send match the message the other person receives?

Impression = Expression

Real communication or understanding happens only when the receiver's impression matches what the sender intended through his or her expression. So the goal of effective communication is a mutual understanding of the message.

In simple terms, effective communication means this . . .

You say it.

I get it.

But how do we know if the other person “gets” our message? We don’t know until we complete the communication. Until a message is complete, the best we can say about its meaning is this:



The meaning of a message is not what is intended by the sender, but what is understood by the receiver.

So what does it take to complete communication? It takes completing the loop. It takes adding one more step. It takes feedback.

In simple terms, complete or effective communication means . . .

You say it.

I get it.

You get that I got it.

So far, then, we’ve defined effective communication and what makes it complete. Let’s now explore the process, or circle, of communication to see how, where, and why it breaks down.

Forms of Communication:

The most common way in which we communicate is by talking to the other person. What are the other possible ways in which you communicate?



There are three main forms of Communication:

1. Verbal communication
2. Non verbal communication
3. Written communication



Verbal Communication

Verbal communication refers to the use of sounds and language to relay a message. It serves as a vehicle for expressing desires, ideas and concepts and is vital to the processes of learning and teaching. In combination with nonverbal forms of communication, verbal communication acts as the primary tool for expression between two or more people.

Types of verbal communication:

Interpersonal communication and public speaking are the two basic types of verbal communication. Whereas public speaking involves one or more people delivering a message to a group, interpersonal communication generally refers to a two-way exchange that involves both talking and listening.

Verbal communication has many purposes, but its main function is relaying a message to one or more recipients. It encompasses everything from simple one-syllable sounds to complex discussions and relies on both language and emotion to produce the desired effect. Verbal communication can be used to inform, inquire, argue and discuss topics of all kinds.

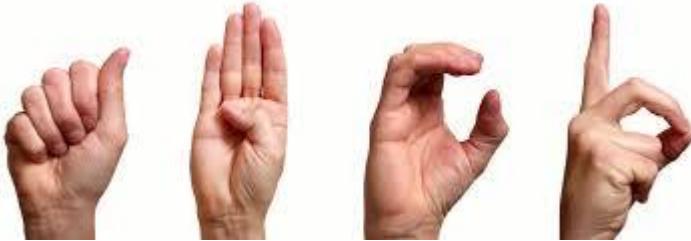


Non Verbal Communication

How do we communicate without words???

- We communicate a lot to each other outside what we say.
- We create confusion when our verbal and nonverbal messages don't match & When verbal and nonverbal messages don't match, we tend to “listen” to the nonverbal one.
(Intuitively, we generally view others’ “body language” as a more reliable indicator of their attitudes and feelings than their words.)
- We can learn to read the meanings of nonverbal behaviors.
 - The key is discovering an individual’s behavior patterns—there is predictability to their meaning.
 - However, be careful—people can mask their feelings.

- Also, trying to read something into every movement others make can get in the way of effective interactions.



Forms of non verbal communication

1. **Ambulation** is the way one walks. Whether the person switches, stomps, or swaggers can indicate how that person experiences the environment.
2. **Touching** is possibly the most powerful nonverbal communication form. People communicate trust, compassion, tenderness, warmth, and other feelings through touch. Also, people differ in their willingness to touch and to be touched. Some people are “touchers” and others emit signals not to touch them.
3. **Eye contact** is used to size up the trustworthiness of another. Counselors use this communication method as a very powerful way to gain understanding and acceptance. Speakers use eye contact to keep the audience interested.
4. **Posturing** can constitute a set of potential signals that communicate how a person is experiencing the environment. It is often said that a person who sits with his/her arms folded and legs crossed is defensive or resistant. On the other hand, the person may just be cold.
5. **Tics** are involuntary nervous spasms that can be a key to indicate one is being threatened. For example, some people stammer or jerk when they are threatened. But these mannerisms can easily be misinterpreted.
6. **Sub-vocals** are the non-words one says, such as “ugh” or “um.” They are used when one is trying to find the right word. People use a lot of non-words trying to convey a message to another person. Another example is the use of "you know." It is used in place of the "ugh" and other grunts and groans commonly used.
7. **Distancing** is a person’s psychological space. If this space is invaded, one can become somewhat tense, alert, or “jammed up.” People may try to move back to reestablish their personal space. The kind of relationship and the motives toward one another determines this personal space.

8. **Gesturing** carries a great deal of meaning between people, but different gestures can mean different things to the sender and the receiver. This is especially true between cultures. Still, gestures are used to emphasize our words and to attempt to clarify our meaning.
9. **Vocalism** is the way a message is packaged and determines the signal that is given to another person. For example, the message, “I trust you,” can have many meanings. “I trust you” could imply that someone else does not. “I trust you” could imply strong sincerity. “I trust you” could imply that the sender does not trust others.

Written Communication

Written communication involves any type of message that makes use of the written word. Written communication is the most important and the most effective of any mode of business communication.

Examples of written communications generally used with clients or other businesses include email, Internet websites, letters, proposals, telegrams, faxes, postcards, contracts, advertisements, brochures, and news releases.



Advantages and disadvantages of written communication:

Advantages

- Creates permanent record
- Allows to store information for future reference
- Easily distributed
- All recipients receive the same information
- Written communication helps in laying down apparent principles, policies and rules for running on an organization.
- It is a permanent means of communication. Thus, it is useful where record maintenance is required.
- Written communication is more precise and explicit
- Effective written communication develops and enhances organization's image
- It provides ready records and references
- Written communication is more precise and explicit.
- Effective written communication develops and enhances an organization's image
- Necessary for legal and binding documents

Disadvantages of Written Communication

- Written communication does not save upon the costs. It costs huge in terms of stationery and the manpower employed in writing/typing and delivering letters.
- Also, if the receivers of the written message are separated by distance and if they need to clear their doubts, the response is not spontaneous.
- Written communication is time-consuming as the feedback is not immediate. The encoding and sending of message takes time.
- Effective written communication requires great skills and competencies in language and vocabulary use. Poor writing skills and quality have a negative impact on organization's reputation.
- Too much paper work and e-mails burden is involved

Common Etiquettes In Written Communication

Continuing with the series of etiquettes in communication, language expert Preeti Shirodkar tells us about what we need to keep in mind while communicating in writing.

While written communication affords greater flexibility, since it can be edited and both composed and read at leisure or at one's pace, a great deal of care needs to be taken, in order to ensure its effectiveness; as it can serve as a point of reference, which one can turn to time and again, thus creating a more lasting impact.

1 – Structuring of the Content

Introduction, Body and Conclusion: While writing one should ensure that the content is well organized, with the overview/basic details comprising the introduction; all major points with their explanation and exemplification constituting the body (preferably divided into a separate paragraph each for every new point, with titles and subtitles, if necessary).

2 – Ensuring Connectivity

The content that comprises a piece of writing should reflect fluency and should be connected through a logical flow of thought, in order to prevent misinterpretation and catch the attention of the reader.

Moreover, care should be taken to ensure that the flow is not brought about through a forced/deliberate use of connectives, as this make the piece extremely uninteresting and artificial.

3 – Steering Clear of Short Form

People may not be aware of the meaning of various short forms and may thus find it difficult to interpret them. Moreover, short forms can at time be culture specific or even organization specific and may thus unnecessarily complicate the communication.

4 – Importance of Grammar, Spelling and Punctuation

Improper grammar can at worst cause miscommunication and at least results in unwanted humour and should be thus avoided. So too, spellings can create the same effect or can even reflect a careless attitude on part of the sender.

Finally, effective use of punctuations facilitates reading and interpretation and can in rare cases even prevent a completely different meaning, which can result in miscommunication.

5 – Sensitivity to the Audience

One needs to be aware of and sensitive to the emotions, need and nature of the audience in choosing the vocabulary, content, illustrations, formats and medium of communication, as a discomfort in the audience would hamper rather than facilitate communication.

6 – Importance of Creativity

In order to hold the readers' attention one needs to be creative to break the tedium of writing and prevent monotony from creeping in.

This is especially true in the case of all detailed writing that seeks to hold the readers' attention.

7 – Avoidance Excessive use of Jargons

Excessive use of jargon can put off a reader, who may not read further, as, unlike a captive audience, the choice of whether to participate in the communication rests considerably with the reader.

Some Do's and Don'ts of Writing

- Be Specific: Just like a reporter, communicate the “who, what, where, why, when and how” of what needs to done. Stay objective and specific.
- Avoid the Passive Voice: Instead of writing “The program was planned by Dane,” write, “Dane planned the program.”
- Be Concise: There’s no need to be long-winded. Get to the point. You’ll lose readers if you spout off too long!
- Get Things Right: Take great care when spelling people’s names,, and other specifics. And also make sure that you do a careful proof of your work.
- Know When Formal Language is Required: If you’re writing an informal note to group members, it’s fine to use contractions (“don’t” instead of “do not”).However, if you’re writing for a formal audience, like a proposal to the board of directors, be more formal with your language.
- Read It Out Loud: One very effective way to self-proof your work is to read it out loud. This will help you determine if you’ve used incorrect words, if your sentences run on too long, if your tenses don’t match, and more.

Communication Barriers

Ask to the candidates “What Communication Barrier means to you?”



Take a few minutes and share your thoughts/ examples.

Common barriers to effective Communication:

1. **The use of jargons.** Over Over-complicated, unfamiliar and/or technical terms.
2. **Emotional barriers and taboos.** Some people may find it difficult to express their emotions and some topics may be completely 'off-limits' or taboo.
3. **Lack of attention, interest, distractions, or irrelevance to the receiver.**
4. **Differences in perception and viewpoint.**
5. **Physical disabilities** such as hearing problems or speech difficulties.
6. **Physical barriers to non verbal communication.**
Not being able to see the non-verbal cues, gestures, posture and general body language can make communication less effective. Accents.
7. **Language differences and the difficulty in understanding unfamiliar accents.**
8. **Expectations and prejudices which may lead to false assumptions or stereotyping.** People often hear what they expect to hear rather than what is actually said and jump to incorrect conclusions.
9. **Cultural differences.** The norms of social interaction vary greatly in different cultures, as do the way in which emotions are expressed. For example, the concept of personal space varies between cultures and between different social settings.



Check Your Understanding



1. True or False? A good definition of communication is the sending of information from one person to another.
 - a. True
 - b. False

Suggested Responses:

False



2. True or False? Good working relationships between people form an important foundation for effective communication
- a. True
 - b. False

Suggested Responses:

True



3. True or False? Empathy is one of the most important concepts in communication.
- a. True
 - b. False

Suggested Responses:

True



4. True or False? The best way to get feedback is to ask, “Do you have any questions?”
- a. True
 - b. False

Suggested Responses:

False



5. True or False? A person’s attitude toward the value of communication is more important than the skills or methods used to communicate.
- a. True
 - b. False

Suggested Responses:

True



6. True or False? Everyone should be responsible for effective upward, downward, and horizontal communication.
- a. True
 - b. False

Suggested Responses:

True



7. True or False? A sender has failed to communicate unless the receiver understands the message the way the sender intended it.
- a. True
 - b. False

Suggested Responses:

True



8. True or False? The grapevine is usually an accurate source of information, and should be used intentionally to communicate.
- a. True
 - b. False

Suggested Responses:

False

9. True or False? If people don't understand, they will usually indicate so by asking questions or by saying they don't understand.
- a. True
 - b. False

Suggested Responses:

False



10. True or False? In order to have an effective communication program, top management must take an active part

a. True

b. False

Suggested Responses:

True



11. True or False? Where persuasion is needed, oral communication is better than written communication.

a. True

b. False

Suggested Responses:

True



12. True or False? In keeping others informed, it is better to under-communicate than over communicate

a. True

b. False

Suggested Responses:

False

13. True or False? The best way to be sure we understand a communication is to repeat it back to the communicator.

a. True

b. False



Suggested Responses:

True



14. True or False? The use of effective visual aids by a speaker usually provides a significant increase in the audience's understanding of the message
- a. True
 - b. False

Suggested Responses:

True



15. True or False? The use of a large vocabulary helps greatly in a person's communication effectiveness.
- a. True
 - b. False

Suggested Responses:

True



16. True or False? Most people can listen approximately four times faster than they speak.
- a. True
 - b. False

Suggested Responses:

True

17. True or False? Information is usually distorted when it is orally communicated through more than two people.

- a. True



b. False

Suggested Responses:

True



18. True or False? In getting people to listen, subject content is more important than the manner in which the subject is communicated.

a. True

b. False

Suggested Responses:

True



19. True or False? People will accept a logical explanation even if it ignores their personal feelings.

a. True

b. False

Suggested Responses:

True

Summary

- The purpose of effective communication skills is to achieve greater understanding between people that builds a climate of trust, openness, and support.
- There are three commonly used forms of communication: Verbal, Non verbal and Written.
- Lack of attention and interest, use of jargons and language differences are some of the common communication barriers. One must watch out for these in order to effectively communicate.



Module 1 : Unit – 3

SQL using R

Topic	Activities
SQL using R	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Understand NOSQL 2. Work on Excel and R integration.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
NO SQL	Classroom
Excel and R integration with R connector	Classroom
Check your understanding	Classroom
Summary	Classroom

Step-by-Step

NO SQL

Before we understand about NO SQL we will see how SQL is used in R.

SQL using R:

It is sqldf, an R package for running SQL statements on data frames.

To load the “SQLDF” package we use below step

```
Library (sqldf)
```

```
# Use the titanic data set
```

```
data(titanic3, package="PASWR")
colnames(titanic3)
head(titanic3)
```

NO SQL:

A NoSQL (originally referring to "non SQL" or "non-relational") database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

There have been various approaches to classify NoSQL databases, each with different categories and subcategories, some of which overlap.

The Benefits of NoSQL

When compared to relational databases, NoSQL databases are more scalable and provide superior performance, and their data model addresses several issues that the relational model is not designed to address:

- Large volumes of structured, semi-structured, and unstructured data
- Agile sprints, quick iteration, and frequent code pushes
- Object-oriented programming that is easy to use and flexible
- Efficient, scale-out architecture instead of expensive, monolithic architecture

A basic classification based on data model, with examples:

- **Column:** Accumulo, Cassandra, Druid, HBase, Vertica
- **Document:** Clusterpoint, Apache CouchDB, Couchbase, DocumentDB, HyperDex, Lotus Notes, MarkLogic, MongoDB, OrientDB, Qizx
- **Key-value:** CouchDB, Oracle NoSQL Database, Dynamo, FoundationDB, HyperDex, MemcacheDB, Redis, Riak, FairCom c-treeACE, Aerospike, OrientDB, MUMPS
- **Graph:** Allegro, Neo4J, InfiniteGraph, OrientDB, Virtuoso, Stardog
- **Multi-model:** OrientDB, FoundationDB, ArangoDB, Alchemy Database, CortexDB

NoSQL vs. SQL Summary

	SQL Databases	NOSQL Databases
Types	One type (SQL database) with minor variations	Many different types including key-value stores, <u>document databases</u> , wide-column stores, and graph databases
Development History	Developed in 1970s to deal with first wave of data storage applications	Developed in 2000s to deal with limitations of SQL databases, particularly concerning scale, replication and unstructured data storage
Examples	MySQL, Postgres, Oracle Database	MongoDB, Cassandra, HBase, Neo4j
Data Storage Model	Individual records (e.g., "employees") are stored as rows in tables, with each column storing a specific piece of data about that record (e.g., "manager," "date hired," etc.), much like a spreadsheet. Separate data types are stored in separate tables, and then joined together when more complex queries are executed. For example, "offices" might be stored in one table, and "employees" in another. When a user wants to find the work address of an employee, the database engine joins the "employee" and	Varies based on database type. For example, key-value stores function similarly to SQL databases, but have only two columns ("key" and "value"), with more complex information sometimes stored within the "value" columns. Document databases do away with the table-and-row model altogether, storing all relevant data together in single "document" in JSON, XML, or another format, which can nest values hierarchically.

	"office" tables together to get all the information necessary.	
Schemas	Structure and data types are fixed in advance. To store information about a new data item, the entire database must be altered, during which time the database must be taken offline.	Typically dynamic. Records can add new information on the fly, and unlike SQL table rows, dissimilar data can be stored together as necessary. For some databases (e.g., wide-column stores), it is somewhat more challenging to add new fields dynamically.
Scaling	Vertically, meaning a single server must be made increasingly powerful in order to deal with increased demand. It is possible to spread SQL databases over many servers, but significant additional engineering is generally required.	Horizontally, meaning that to add capacity, a database administrator can simply add more commodity servers or cloud instances. The database automatically spreads data across servers as necessary.
Development Model	Mix of open-source (e.g., Postgres, MySQL) and closed source (e.g., Oracle Database)	Open-source
Supports Transactions	Yes, updates can be configured to complete entirely or not at all	In certain circumstances and at certain levels (e.g., document level vs. database level)
Data Manipulation	Specific language using Select, Insert, and Update statements, e.g. SELECT fields FROM table WHERE...	Through object-oriented APIs
Consistency	Can be configured for strong consistency	Depends on product. Some provide strong consistency (e.g., MongoDB) whereas others offer eventual consistency (e.g., Cassandra)

Excel and R integration with R connector

1 – Read Excel spreadsheet in R

gdata: it requires you to install additional Perl libraries on Windows platforms but it's very powerful.

```
require(gdata)
```

```
myDf<- read.xls ("myfile.xlsx"), sheet = 1, header = TRUE)
```

RODBC: This is reported for completeness only. It's rather dated; there are better ways to interact with Excel nowadays.

XLConnect: It might be slow for large dataset but very powerful otherwise.

```
require (XLConnect)
```

```
wb<- loadWorkbook("myfile.xlsx")
```

```
myDf<- readWorksheet(wb, sheet = "Sheet1", header = TRUE)
```

xlsx: Prefer the read.xlsx2() over read.xlsx(), it's significantly faster for large dataset.

```
require(xlsx)
```

```
read.xlsx2("myfile.xlsx", sheetName = "Sheet1")
```

xlsReadWrite: Available for Windows only. It's rather fast but doesn't support .xlsx files which is a serious drawback. It has been removed from CRAN lately.

read.table("clipboard"): It allows to copy data from Excel and read it directly in R. This is the quick and dirty R/Excel interaction but it's very useful in some cases.

```
myDf<- read.table("clipboard")
```

2 – Read R output in Excel

First create a csv output from an R data.frame then read this file in Excel. There is one function that you need to know it's write.table. You might also want to consider: write.csv which uses “.” for the decimal point and a comma for the separator and write.csv2 which uses a comma for the decimal point and a semicolon for the separator.

```
x <- cbind(rnorm(20),runif(20))
colnames(x) <- c("A ","B ")
write.table(x,"your_path",sep=",",row.names=FALSE)
```

3 – Execute R code in VBA

RExcel is from my perspective the best suited tool but there is at least one alternative. You can run a batch file within the VBA code. If R.exe is in your PATH, the general syntax for the batch file (.bat) is:

R CMD BATCH [options] myRScript.R

Here's an example of how to integrate the batch file above within your VBA code.

4 – Execute R code from an Excel spreadsheet

RExcel is the only tool I know for the task. Generally speaking once you installed RExcel you insert the excel code within a cell and execute from RExcel spreadsheet menu. See the RExcel references below for an example.

5 – Execute VBA code in R

This is something I came across but I never tested it myself. This is a two steps process. First write a VBscript wrapper that calls the VBA code. Second run the VBscript in R with the system or shell functions. The method is described in full details here.

6 – Fully integrate R and Excel

RExcel is a project developed by Thomas Baier and Erich Neuwirth, “making R accessible from Excel and allowing to use Excel as a frontend to R”. It allows communication in both directions: Excel to R and R to Excel and covers most of what is described above and more. I’m not going to put any example of RExcel use here as the topic is largely covered elsewhere but I will show you where to find the relevant information. There is a wiki for installing RExcel and an excellent tutorial available here. I also recommend the following two documents: RExcel – Using R from within Excel and High-Level Interface Between R and Excel. They both give an in-depth view of RExcel capabilities.

Check your understanding



1. How do you read excel dataset in R?
2. What are the types of No SQL tools based on Data Models?
3. Why do we use No SQL?
4. Is No SQL a query language like SQL?

Summary

- For integration of SQL and R we use SQLDF package.
- A NoSQL (originally referring to "non SQL" or "non-relational") database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.
- NoSQL support SQL like query language.
- NoSQL is used primarily in compliment to Big Data tools.
- Excel can be integrated with R using R-Connector.
- How to execute VBA code in R tool?



- Activity**
1. Divide the class into groups of 4-5 participants
 2. Give the Dataset to the participants.
 3. Give 10 minutes to the class for each group to discuss the various change points between SQL and NO SQL along with a discussion on the methods type which they would like to use
 4. Each group presents their examples with justification. (5 min each)

Module 1 UNIT – 4

Correlation and Regression

Topic	Session Goals
Correlation and Regression	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Make Regression models 2. Find Correlation 3. Understand Multi Collinearity 4. Work on Multiple Regression 5. Work with Dummy variables
Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Basic Regression Analysis	Classroom
OLS Regression	Classroom
Regression Modeling	Classroom
Regression residuals	Classroom
Correlation	Classroom
Heteroscedasticity	Classroom
Autocorrelation& Multicollinearity	Classroom
Introduction to Multiple Regression	Classroom
Dummy Variables	Classroom
Check your understanding	Classroom

Basic Regression Analysis

Regression analysis is the statistical method you use when both the response variable and the explanatory variable are continuous variables (i.e. real numbers with decimal places – things like heights, weights, volumes, or temperatures).

In simple regression, we try to determine whether there is a relationship between two variables. It is assumed that there is a high degree of correlation between the two variables chosen for use in regression.

In R we use lm () function to do simple regression modeling.

For example,

```
> fit <- lm(data$petal_length ~ data$petal_width)
```

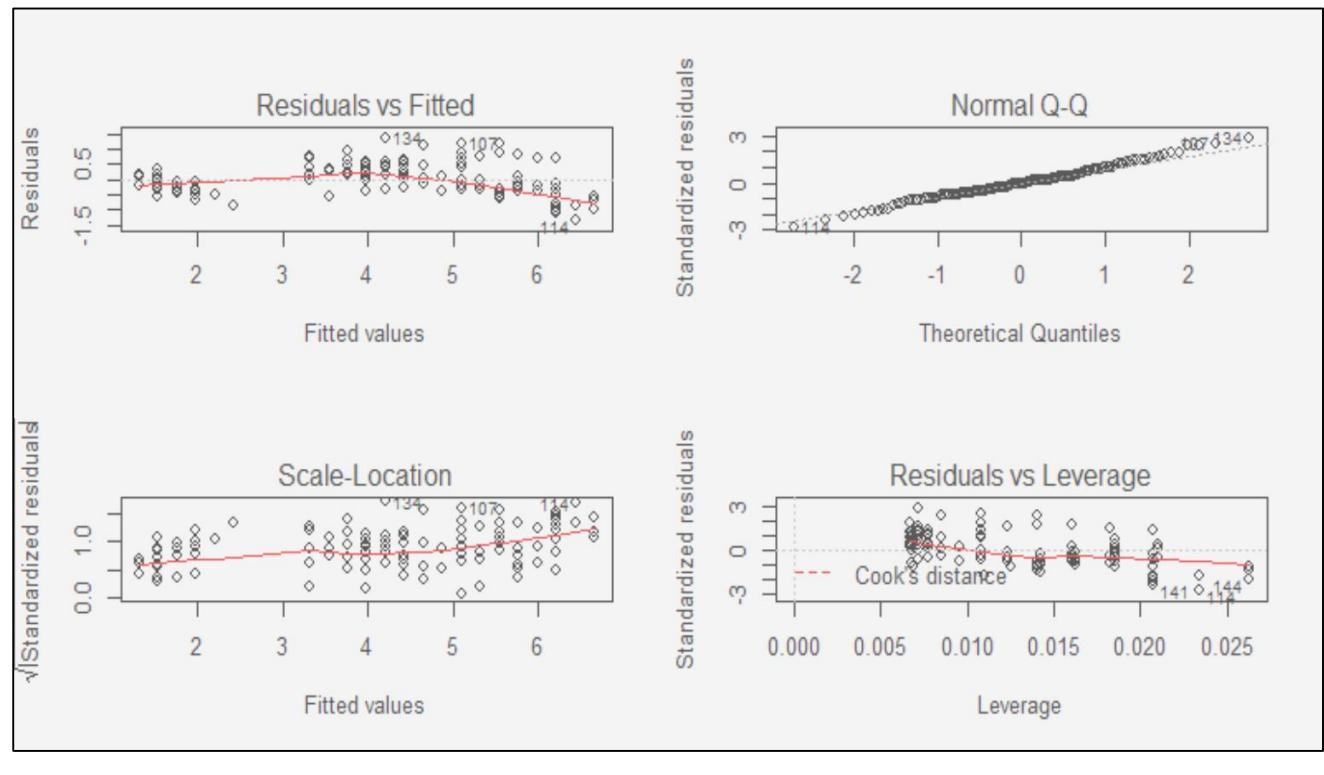
When we call “fit” as below

```
>fit
```

We get the intercept “C” and the slope “m” of the equation – $Y = mX + C$

The fit information displays four charts: Residuals vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs. Leverage.

Below are the various graphs representing values of regression



OLS Regression

OLS:- Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data.

This is applied in both simple linear and multiple regression where the common assumptions are

- (1) The model is linear in the coefficients of the predictor with an additive random error term
- (2) The random error terms are

- normally distributed with 0 mean and
- a variance that doesn't change as the values of the predictor covariates (i.e. IVs) change.

Regression Modeling

- Regression modeling or analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.
- Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.
- Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the

regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

- The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.
- In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification. The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.

Regression residuals

The residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest.

Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.

➤ **Residuals**

The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the **residual** (e). Each data point has one residual.

$$\begin{aligned} \text{Residual} &= \text{Observed value} - \text{Predicted value} \\ e &= y - \hat{y} \end{aligned}$$

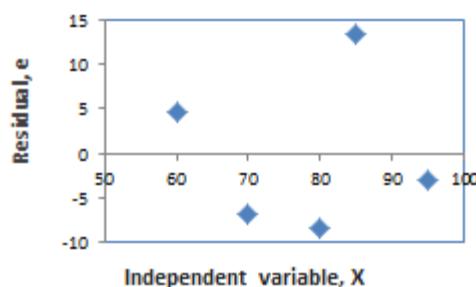
Both the sum and the mean of the residuals are equal to zero. That is, $\sum e = 0$ and $e = 0$.

➤ **Residual Plots**

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

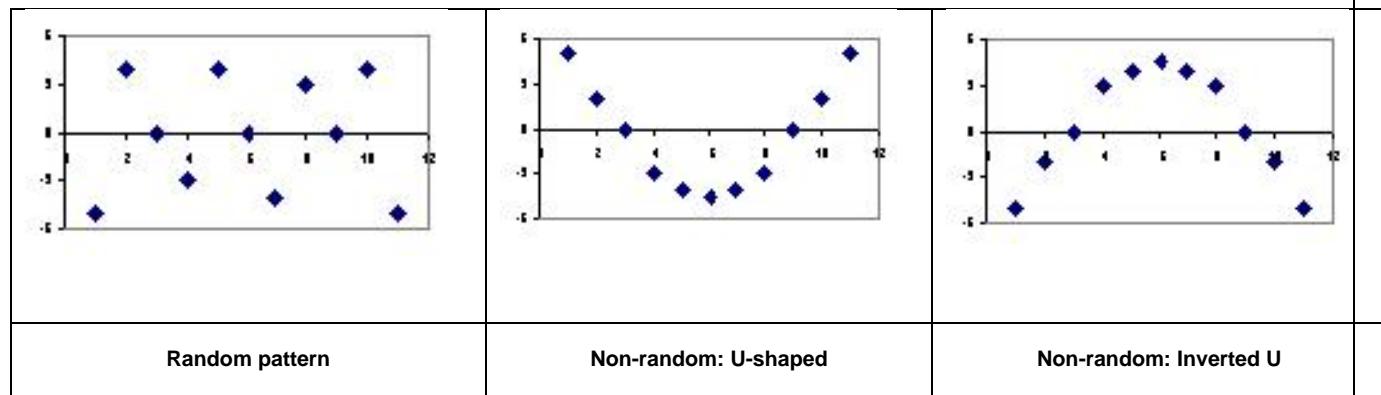
Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

x	60	70	80	85	95
y	70	65	70	95	85
\hat{y}	65.411	71.849	78.288	81.507	87.945
e	4.589	-6.849	-8.288	13.493	-2.945



The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.



Correlation

- Measure of association between variables
- Positive and negative correlation, ranging between +1 and -1
- Positive correlation example:
 - Earning and expenditure
 - Negative correlation example
 - Speed and time
 - Parametric – normal distribution and homogenous variance
 - Pearson correlation
 - Non parametric – no assumptions, nominal variables
 - Spearman correlation

Correlation Coefficients:-

- r : correlation coefficient
- +1 : Perfectly positive
- -1 : Perfectly negative
- 0 – 0.2 : No or very weak association
- 0.2 – 0.4 : Weak association
- 0.4 – 0.6 : Moderate association
- 0.6 – 0.8 : Strong association
- 0.8 – 1 : Very strong to perfect association

Correlation and Covariance:

With two continuous variables, x and y, the question naturally arises as to whether their values are correlated with each other (remembering, of course, that correlation does not imply causation). Correlation is defined in terms of the variance of x, the variance of y, and the covariance of x and y (the way the two vary together; the way they co-vary) on the assumption that both variables are normally distributed. We have symbols already for the two variances, s^2_x and s^2_y .

We denote the covariance of x and y by $\text{cov}(x, y)$, after which the correlation coefficient r is defined as

$$r = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

Heteroscedasticity:

A collection of random variables is heteroscedastic (or 'heteroskedastic' from Ancient Greek hetero "different" and skedasis "dispersion") if there are sub-populations that have different variabilities from others. Here "variability" could be quantified by the variance or any other measure of statistical dispersion. Thus heteroscedasticity is the absence of homoscedasticity.

The existence of heteroscedasticity is a major concern in the application of regression analysis, including the analysis of variance, as it can invalidate statistical tests of significance that assume that the modeling errors are uncorrelated and uniform—hence that their variances do not vary with the effects being modeled. For instance, while the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient because the true variance and covariance are underestimated. Similarly, in testing for differences between sub-populations using a location test, some standard tests assume that variances within groups are equal.

Test of Heteroscedasticity:-

Tests in regression

- Levene's test
- Goldfeld–Quandt test
- Park test
- Glejser test
- Brown–Forsythe test
- Harrison–McCabe test
- Breusch–Pagan test
- White test
- Cook–Weisberg test

Tests for grouped data

- F-test of equality of variances
- Cochran's C test
- Hartley's test

These tests consist of a test statistic (a mathematical expression yielding a numerical value as a function of the data), a hypothesis that is going to be tested (the null hypothesis), an alternative hypothesis, and a statement about the distribution of statistic under the null hypothesis.

Fixes:-

There are four common corrections for heteroscedasticity. They are:

- View logarithmized data. Non-logarithmized series that are growing exponentially often appear to have increasing variability as the series rises over time. The variability in percentage terms may, however, be rather stable.
- Use a different specification for the model (different X variables, or perhaps non-linear transformations of the X variables).
- Apply a weighted least squares estimation method, in which OLS is applied to transformed or weighted values of X and Y. The weights vary over observations, usually depending on the changing error variances. In one variation the weights are directly related to the magnitude of the dependent variable, and this corresponds to least squares percentage regression.
- Heteroscedasticity-consistent standard errors (HCSE), while still biased, improve upon OLS estimates. HCSE is a consistent estimator of standard errors in regression models with heteroscedasticity. This method corrects for heteroscedasticity without altering the values of the coefficients. This method may be superior to regular OLS because if heteroscedasticity is present it corrects for it, however, if the data is homoscedastic, the standard errors are equivalent to conventional standard errors estimated by OLS. Several modifications of the White method of computing heteroscedasticity-consistent standard errors have been proposed as corrections with superior finite sample properties.

Autocorrelation

Autocorrelation, also known as serial correlation or cross-autocorrelation, is the cross-correlation of a signal with itself at different points in time (that is what the cross stands for). Informally, it is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

In statistics, the autocorrelation of a random process describes the correlation between values of the process at different times, as a function of the two times or of the time lag. Let X be some repeatable process, and i be some point in time after the start of that process. (i may be an integer for a discrete-time process or a real number for a continuous-time process.) Then X_i is the value (or realization) produced by a given run of the process at time i . Suppose that the process is further known to have defined values for mean μ_i and variance σ_i^2 for all times i . Then the definition of the autocorrelation between times s and t is

$$R(s, t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s},$$

where "E" is the expected value operator.

Test: -

- The traditional test for the presence of first-order autocorrelation is the Durbin–Watson statistic or, if the explanatory variables include a lagged dependent variable, Durbin's h statistic. The Durbin-Watson can be linearly mapped however to the Pearson correlation between values and their lags.
- A more flexible test, covering autocorrelation of higher orders and applicable whether or not the regressors include lags of the dependent variable, is the Breusch–Godfrey test. This involves an auxiliary regression, wherein the residuals obtained from estimating the model of interest are regressed on (a) the original regressors and (b) k lags of the residuals, where k is the order of the test. The simplest version of the test statistic from this auxiliary regression is TR^2 , where T is the sample size and R^2 is the coefficient of determination. Under the null hypothesis of no autocorrelation, this statistic is asymptotically distributed as χ^2 with k degrees of freedom.

Multicollinearity

In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regressions may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

In case of perfect multicollinearity the predictor matrix is singular and therefore cannot be inverted. Under these circumstances, for a general linear model $y=X\beta+\epsilon$, the ordinary estimator

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

does not exist.

Test:-

Indicators that multicollinearity may be present in a model:

- 1) Large changes in the estimated regression coefficients when a predictor variable is added or deleted
- 2) Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an F-test)
- 3) If a multivariable regression finds an insignificant coefficient of a particular explanator, yet a simple linear regression of the explained variable on this explanatory variable shows its coefficient to be significantly different from zero, this situation indicates multicollinearity in the multivariable regression.
- 4) Some authors have suggested a formal detection-tolerance or the variance inflation factor (VIF) for multicollinearity:

$$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{\text{tolerance}},$$

Where R_j^2 is the coefficient of determination of a regression of explanator j on all the other explanators. A tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicates a multicollinearity problem.

- 5) Condition number test: The standard measure of ill-conditioning in a matrix is the condition index. It will indicate that the inversion of the matrix is numerically unstable with finite-precision numbers (standard computer floats and doubles). This indicates the potential sensitivity of the computed inverse to small changes in the original matrix. The Condition Number is computed by finding the square root of (the maximum eigenvalue divided by the minimum eigenvalue). If the Condition Number is above 30, the regression may have significant multicollinearity; multicollinearity exists if, in addition, two or more of the variables related to the high condition number have high proportions of variance explained. One advantage of this method is that it also shows which variables are causing the problem.
- 6) Farrar–Glauber test: If the variables are found to be orthogonal, there is no multicollinearity; if the variables are not orthogonal, then multicollinearity is present. C. Robert Wichers has argued that Farrar–Glauber partial correlation test is ineffective in that a given partial correlation may be compatible with different multicollinearity patterns. The Farrar–Glauber test has also been criticized by other researchers.
- 7) Perturbing the data. Multicollinearity can be detected by adding random noise to the data and re-running the regression many times and seeing how much the coefficients change.
- 8) Construction of a correlation matrix among the explanatory variables will yield indications as to the likelihood that any given couplet of right-hand-side variables is creating multicollinearity problems. Correlation values (off-diagonal elements) of at least .4 are sometimes interpreted as indicating a multicollinearity problem. This procedure is, however, highly problematic and cannot be recommended. Intuitively, correlation describes a bivariate relationship, whereas collinearity is a multivariate phenomenon.

Introduction to Multiple Regression

The general purpose of multiple regressions (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

For example,

A real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating).

Dummy Variables

In regression analysis, a dummy variable (also known as an indicator variable, design variable, Boolean indicator, categorical variable, binary variable, or qualitative variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories (such as smoker/non-smoker, etc.).

In other words, Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

For example,

Suppose Gender is one of the qualitative variables relevant to a regression. Then, female and male would be the categories included under the Gender variable. If female is arbitrarily assigned the value of 1, then male would get the value 0. Then the intercept (the value of the dependent variable if all other explanatory variables hypothetically took on the value zero) would be the constant term for males but would be the constant term plus the coefficient of the gender dummy in the case of females.

Check your understanding



In the context of regression analysis, which of the following statements are true?

- I. When the sum of the residuals is greater than zero, the data set is nonlinear.
 - II. A random pattern of residuals supports a linear model.
 - III. A random pattern of residuals supports a non-linear model.
- (A) I only
(B) II only
(C) III only
(D) I and II
(E) I and III

Summary

- Regression is method of establishing relation between two or more variables.
- Correlation shows the extent of relation
- Correlation coefficient lies between -1 and 1.
- Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model.
- Multiple regressions find relationship between several independent or predictor variables and a dependent or criterion variable.
- Multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.
- A dummy variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect.
- Autocorrelation, also known as serial correlation or cross-autocorrelation

Module 1 UNIT – 5

Understand the verticals and requirements gathering

Topic	Activities
Understand the Verticals and Requirement Gathering	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Solve Engg. & Manufac. Issues 2. Create Business Models

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Understand systems viz. Engineering Design, Manufacturing, smart utilities, production lines, Automotive industries, Tech system	Classroom
Understand the business problem Related to engineering, Identify the critical issues. Set business objectives.	Classroom
Requirement gathering	Classroom
Summary	Classroom

Step-by-Step

Understand systems viz. Engineering Design, Manufacturing, smart utilities, production lines, Automotive industries, Tech system

Engineering Design:

The **engineering design** process is a methodical series of steps that engineers use in creating functional products and processes. The process is highly iterative - parts of the process often need to be repeated many times before production phase can be entered - though the part(s) that get iterated and the number of such cycles in any given project can be highly variable.

One framing of the engineering design process delineates the following stages: research, conceptualization, feasibility assessment, establishing design requirements, preliminary design, detailed design, production planning and tool design, and production.

Manufacturing:

Manufacturing is the production of merchandise for use or sale using labour and machines, tools, chemical and biological processing, or formulation. The term may refer to a range of human activity, from handicraft to high tech, but is most commonly applied to industrial production, in which raw materials are transformed into finished goods on a large scale. Such finished goods may be used for manufacturing other, more complex products, such as aircraft, household appliances or automobiles, or sold to wholesalers, who in turn sell them to retailers, who then sell them to end users – the "consumers".

Manufacturing takes turns under all types of economic systems. In a free market economy, manufacturing is usually directed toward the mass production of products for sale to consumers at a profit. In a collectivist economy, manufacturing is more frequently directed by the state to supply a centrally planned economy. In mixed market economies, manufacturing occurs under some degree of government regulation.

Modern manufacturing includes all intermediate processes required for the production and integration of a product's components. Some industries, such as semiconductor and steel manufacturers use the term fabrication instead.

The manufacturing sector is closely connected with engineering and industrial design. Examples of major manufacturers in North America include General Motors Corporation, General Electric, Procter & Gamble, General Dynamics, Boeing, Pfizer, and Precision Cast parts. Examples in Europe include Volkswagen Group, Siemens, and Michelin. Examples in Asia include Sony, Huawei, Lenovo, Toyota, Samsung, and Bridgestone.

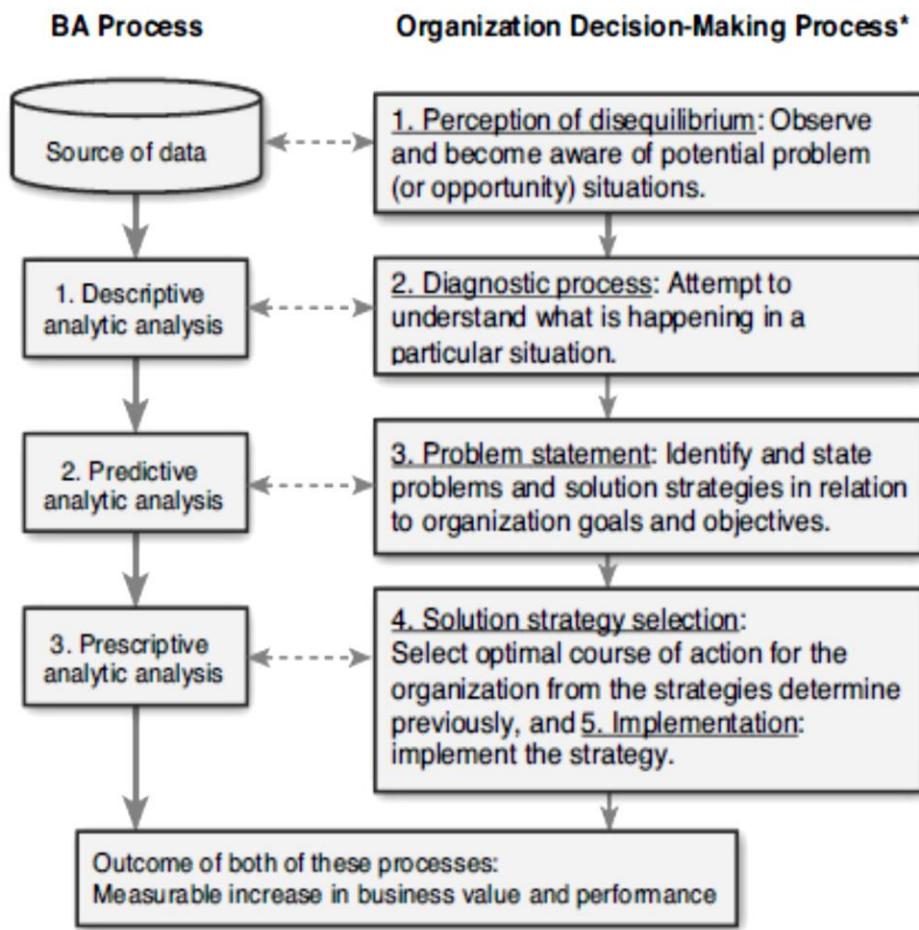
SMART Utilities:

S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology; often written as SMART) is a monitoring system included in computer hard disk drives (HDDs) and solid-state drives (SSDs) that detects and reports on various indicators of drive reliability, with the intent of enabling the anticipation of hardware failures.

When S.M.A.R.T. data indicates a possible imminent drive failure, software running on the host system may notify the user so stored data can be copied to another storage device, preventing data loss, and the failing drive can be replaced.

Understand the business problem related to engineering, Identify the critical issues. Set business objectives.

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities can be viewed as a problem of strategy choice requiring a solution.



Comparison of business analytics and organization decision-making processes

Requirements gathering : Gather all the Data related to Business objective

There are many different approaches that can be used to gather information about a business. They include the following:

- ❖ Review business plans, existing models and other documentation
- ❖ Interview subject area experts
- ❖ Conduct fact-finding meetings
- ❖ Analyze application systems, forms, artifacts, reports, etc.

The business analyst should use one-on-one interviews early in the business analysis project to gage the strengths and weaknesses of potential project participants and to obtain basic information about the business. Large meetings are not a good use of time for data gathering.

Facilitated work sessions are a good mechanism for validating and refining “draft” requirements. They are also useful to prioritize final business requirements. Group dynamics can often generate even better ideas.

Primary or local data is collected by the business owner and can be collected by survey, focus group or observation. Third party static data is purchased in bulk without a specific intent in mind. While easy to get (if you have the cash) this data is not specific to your business and can be tough to sort through as you often get quite a bit more data than you need to meet your objective. Dynamic data is collected through a third party process in near real-time from an event for a specific purpose (read into that VERY expensive).

Three key questions you need to ask before making a decision about the best method for your firm.

- What is the timeline required to accomplish your business objective?
- What is your required return on investment?
- Is the data collection for a stand-alone event or for part of a broader data collection effort?

How to interpret Data to make it useful for Business:-

Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.



Business intelligence is made up of an increasing number of components including:

- Multidimensional aggregation and allocation
- Denormalization, tagging and standardization
- Realtime reporting with analytical alert
- A method of interfacing with unstructured data sources
- Group consolidation, budgeting and rolling forecasts
- Statistical inference and probabilistic simulation
- Key performance indicators optimization
- Version control and process management
- Open item management

Business intelligence can be applied to the following business purposes, in order to drive business value.

- Measurement – program that creates a hierarchy of performance metrics (see also Metrics Reference Model) and benchmarking that informs business leaders about progress towards business goals (business process management).
- Analytics – program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, complex event processing and prescriptive analytics.
- Reporting/enterprise reporting – program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
- Collaboration/collaboration platform – program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
- Knowledge management – program to make the company data-driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge. Knowledge management leads to learning management and regulatory compliance.

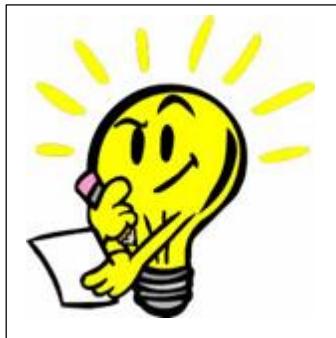
In addition to the above, business intelligence can provide a pro-active approach, such as alert functionality that immediately notifies the end-user if certain conditions are met. For example, if some business metric exceeds a pre-defined threshold, the metric will be highlighted in standard reports, and the business analyst may be alerted via e-mail or another monitoring service. This end-to-end process requires data governance, which should be handled by the expert.

Data can be always gathered using surveys.

Your surveys should follow a few basic but important rules:

1. **Keep it VERY simple.** I recommend one page with 3-4 questions maximum. Customers are visiting to purchase or to have an experience, not to fill out surveys.
2. **Choose only one objective for the survey.** Don't try to answer too many questions, ultimately you won't get much useful data that way because your customer will get confused and frustrated.
3. **Don't give the respondent any wiggle room.** Open ended questions are tough to manage. Specific choices that are broad enough to capture real responses gives you data that is much easier to use.
4. **Always gather demographics.** Why not? But rather than name and e-mail (leading to concerns with confidentiality and often less than truthful answers) gather gender, age and income; you might be surprised at who is actually buying what.

Check your understanding



1. What are various steps involved Organization Decision making?
2. What are examples of SMART utilities?
3. What do you understand by Production Lines?
4. What are various components of Descriptive Statistics?

Summary

- Engineering design process is a methodical series of steps that engineers use in creating functional products and processes.
- Manufacturing is the production of merchandise for use or sale using labor and machines, tools, chemical and biological processing, or formulation.
- Assembly line or Production line concept was first used by Henry Ford in automobile industry. It reduces production time drastically.
- Most of the critical Business problems are solved with help of Data Analytics.

Activity



1. Divide the class into groups of 4-5 participants
2. Give the Dataset to the participants.
3. Give 10 minutes to the class for each group to discuss on an issue of an automobile giant X where the quality of Bikes is not as per standard. So solve this issue by taking some Business Decisions. Along with a discussion on the methods type which they would like to use
4. Each group presents their examples with justification. (5 min each)



Hand book on Big Data Analytics Basics

By

Prakash D Devarakonda

Member of the board and ceo

karvy analytics

WWW.KARVY-ANALYTICS.COM

contactus@karvy-analytics.com

Linkedin: <http://in.linkedin.com/in/prakashdurgad>

About Author



Durga Prakash Devarakonda
CEO & Member of the Board
Karvy Analytics Limited

Durga Prakash (Prakash) heads big data analytics business and drive verticalization of Analytics solutions in partnership with Vertical competencies. Prakash is based out of Hyderabad.

Prakash has been contributing to global IT services business for the last 20 years in Data Analytics space, and has played key roles with companies like Accenture, Oracle, Capgemini etc., He was instrumental in setting up the Data Analytics team at Accenture and later on at Tech Mahindra.

Prakash carries a wealth of experience in Data Analytics across sectors like Defence, Aerospace etc., and had worked in different roles - client facing, Pre-sales, delivery etc., He has worked across geographies viz., US, Canada, UK and the Middle East and has built connects there over the years.

Prakash was Instrumental in setting up of Incubation team with Accenture Analytics in India. He contributed to Analytics practice strategy, service offerings, Industry and academia partnerships apart from nurturing several key North American client relationships. He is an active member of International Institute for Analytics (IIA).

Prakash won accolades from Bill Gates for achieving a status of “youngest Indian who completed all possible Microsoft certifications in the year 1998. He contributed to several leadership development programs internally in the past. He attended Harvard Business

School's executive education program on "Managing and transforming Professional services firms in India".

He is an MIT and has also attended an executive education program from Harvard Business School. Prakash is an avid reader of books and is big movie buff. He is an active member of various CSR initiatives.

NoSQL and big data: State of the art

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 have 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 insufficient capacity to distribute treatments over a large quantity of machines, to face picks of charge: this is **the scaling of treatment**,
- The insufficient capacity to store data over a large quantity of machines, to address their **volume**: this is **the scaling of data**,
- The insufficient capacity to distribute the database over several data-centers to insure the continuity of service: This is **the redundancy**,
- the insufficient capacity to adjust its data model to the nature of the data, due to the **velocity** of web data,
- the bad performance to access data as soon as it requires the combination several tables due to the **variety** of web data and the **complexity** of unstructured data.

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

As a relational database, it provides a set of functionalities to access data across several entities (tables) by complex queries. It provides also an 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.

In one word, both the **3rd normal form** and the **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:

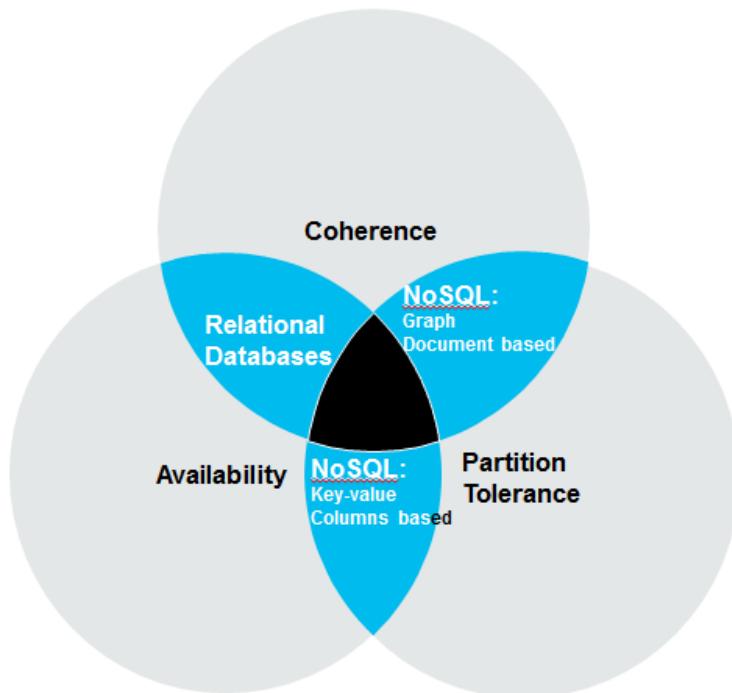
- The **Coherence**: All the nodes of the system have to see exactly the same data at the same time
- The **Availability**: The system must stay up and running even if one of its node is failing down
- The **Partition Tolerance**: each subnet-works must be autonomous

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, relational databases 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.



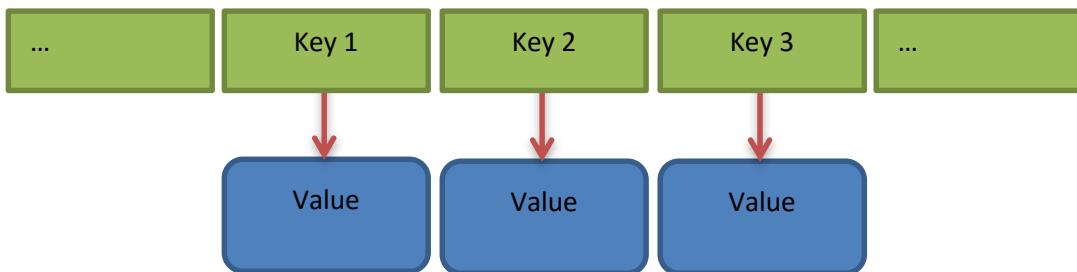
The four categories of NOSQL databases

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.



A few providers

Solution	Distribution	Comment
Redis http://redis.io	BSD licence	Certainly the more mature, Providing some functionalities of to manipulate and store strings and collections No real mechanism of partitioning but some functionalities of master/slave replication Sponsored by VMware
Riak http://wiki.basho.com	Apache licence 2.0	Open source implementation of Amazon Dynamo Completely distributed Map/reduce enabled
Voldemort http://project-voldemort.com	Apache licence 2.0	Initially developed by LinkedIn Optimizing the communication between nodes of the network

Columns based databases

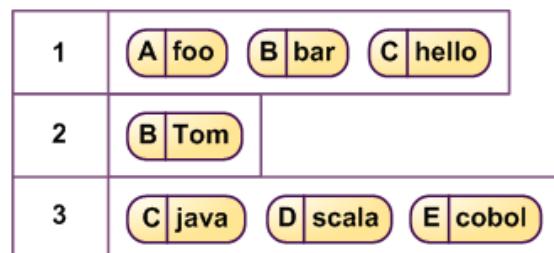
Concept

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 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.

	A	B	C	D	E
1	foo	bar	hello		
2		Tom			
3			java	scala	cobol

Structure of a table in a relational database



Structure of a table in a columns oriented database

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 were 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 is also possible to query the database from the content of the columns.

A few providers

Solution	Distribution	Comment
Cassandra http://cassandra.apache.org	Apache licence 2.0	Very popular open source database Most online services are using it such as Facebook
Goole Big table	SaaS	Database of the Google App Engine, the Google web development environment Only accessible from the API of google App engine
HBase http://hbase.apache.org	Apache licence 2.0	Open source implementation of Google Big Table based on Hadoop

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 an 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.

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).

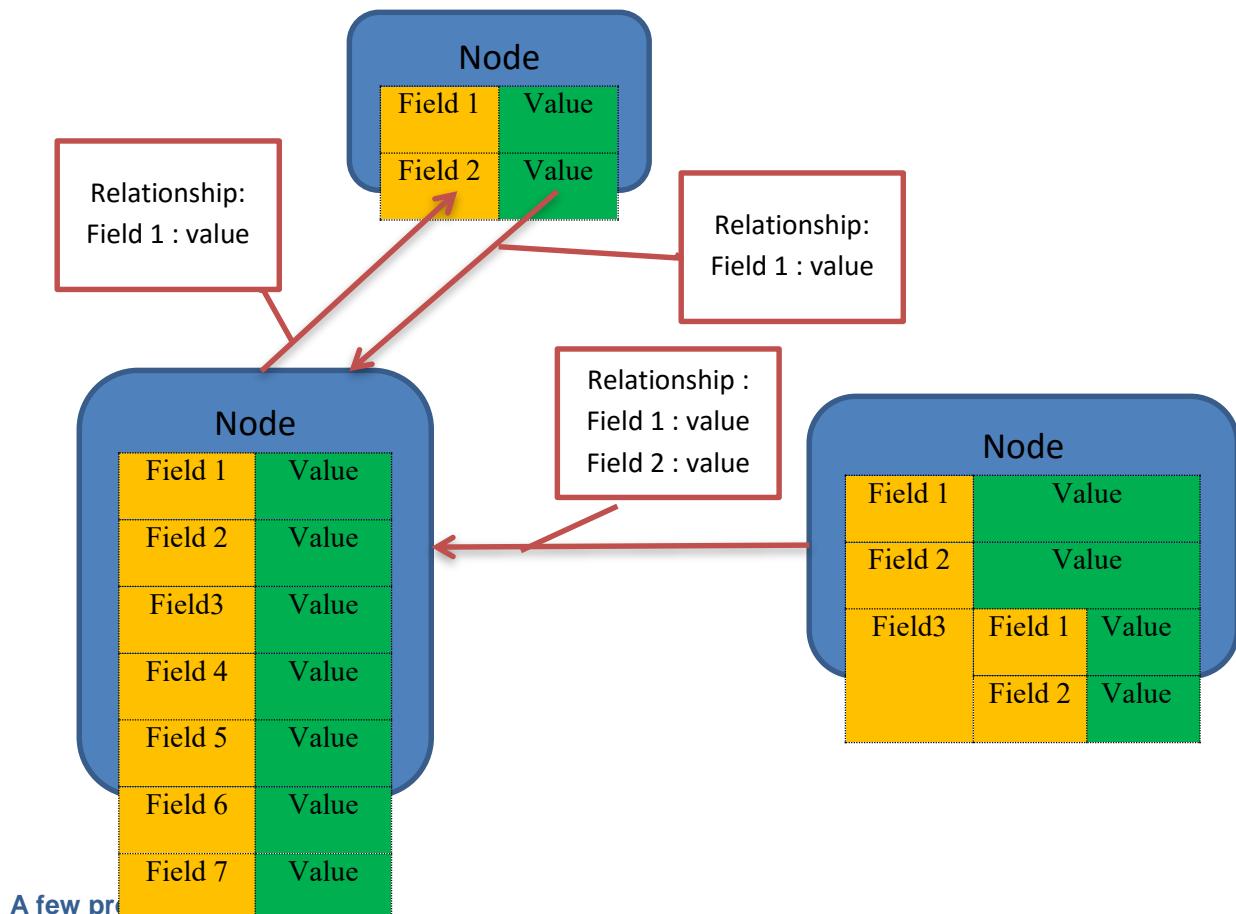
A few providers

Solution	Description	Comment
MongoDB http://www.mongodb.org	Licence GNU AGPL v3.0	Very popular open source database Mature and stable solution Good functional coverage and professional support
CouchDB http://couchdb.apache.org	Apache licence 2.0	Metadata and versioning oriented
Terrastore code.google.com/p/terrastore	Apache licence 2.0	This solution and its querying language can be easily extended

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.



Solution	Distribution	Comment
Neo4j http://neo4j.org	Licence GNU AGPL v3.0/ Commercial version	Stable since January 2010 professional support
OrientDB http://orienttechnologies.com	Apache licence 2.0	Possible to query the database using SQL language

New practices for new paradigms

The implementation of such large and distributed databases implies new methods in the design of the data model.

The de-normalization

At the top of these practices, the de-normalization of the data model is concretizing the rupture with the world of relational databases. The designer/developer has the responsibility to decide if he will materialize the association between two entities by:

- Embedding the data of the child entity in the record of the parent entity: this approach is optimal in term of response time, but makes complex any “update” query due to the multiple references,

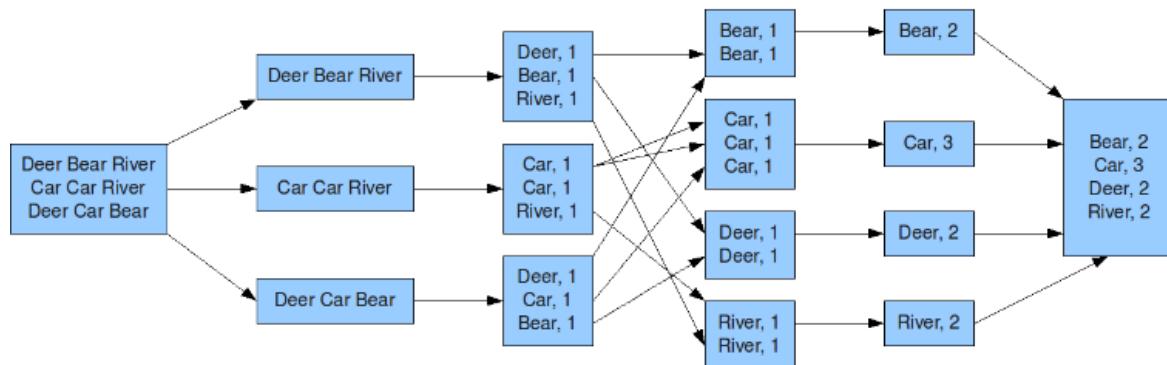
Or

- Keeping two distinct entities linked together in a one-to-one, one-to-many or many-to-many relationship: This approach increases the consistency of the database but downgrades the response time.

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.



Hadoop Installation

Hadoop is a framework written in Java for running applications on large clusters of commodity hardware and incorporates features similar to those of the Google File System and of MapReduce. HDFS is a highly fault-tolerant distributed file system and like Hadoop designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets

- 1) Download and install VMWare Player if you are trying to configure on windows system.
If you already using linux or ubuntu no need to download VMWare.
- 2) Download ubuntu for VMWare from
<http://www.trendsigma.net/vmware/ubuntu1104.html>
- 3) Open ubuntu from VMWare player .
- 4) Check weather jdk installed in ubuntu if not please install it (Jdk 1.5 and above.)

```
user@ubuntu:~$ su - hduser
hduser@ubuntu:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
9b:82:ea:58:b4:e0:35:d7:ff:19:66:a6:ef:ae:0e:d2 hduser@ubuntu
The key's randomart image is:
[...snipp...]
hduser@ubuntu:~$
```

- 5) Configure SSH :
- 6) Create RSA key pair and check SSH configuration

```
hduser@ubuntu:~$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

The final step is to test the SSH setup by connecting to your local machine with the `hduser` user. The step is also needed to save your local machine's host key fingerprint to the `hduser` user's `known_hosts` file. If you have any special SSH configuration for your local machine like a non-standard SSH port, you can define host-specific SSH options in `$HOME/.ssh/config` (see `man ssh_config` for more information).

```
hduser@ubuntu:~$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
RSA key fingerprint is d7:87:25:47:ae:02:00:eb:1d:75:4f:bb:44:f9:36:26.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (RSA) to the list of known hosts.
Linux ubuntu 2.6.32-22-generic #33-Ubuntu SMP Wed Apr 28 13:27:30 UTC 2010 i686 GNU/Linux
Ubuntu 10.04 LTS
[...snipp...]
hduser@ubuntu:~$
```

7) Disabling IPv6 :

To disable IPv6 on Ubuntu 10.04 LTS, open `/etc/sysctl.conf` in the editor of your choice and add the following lines to the end of the file:

```
#disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

You can check whether IPv6 is enabled on your machine with the following command

```
"$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

- 8) Download Hadoop from <http://www.apache.org/dyn/closer.cgi/hadoop/core> and extract at following location.

```
$ cd /usr/local
$ sudo tar xzf hadoop-1.0.3.tar.gz
$ sudo mv hadoop-1.0.3 hadoop
$ sudo chown -R hduser:hadoop hadoop
```

Update \$HOME/.bashrc

Add the following lines to the end of the `$HOME/.bashrc` file of user `hduser`. If you use a shell other than bash, you should of course update its appropriate configuration files instead of `.bashrc`.

```
# Set Hadoop-related environment variables
export HADOOP_HOME=/usr/local/hadoop

# Set JAVA_HOME (we will also configure JAVA_HOME directly for Hadoop later on)
export JAVA_HOME=/usr/lib/jvm/java-6-sun

# Some convenient aliases and functions for running Hadoop-related commands
unalias fs &> /dev/null
alias fs="hadoop fs"
unalias hls &> /dev/null
alias hls="fs -ls"

# If you have LZO compression enabled in your Hadoop cluster and
# compress job outputs with LZOP (not covered in this tutorial):
# Conveniently inspect an LZOP compressed file from the command
# line; run via:
#
# $ lzohead /hdfs/path/to/lzop/compressed/file.lzo
#
# Requires installed 'lzop' command.
#
lzohead () {
    hadoop fs -cat $1 | lzop -dc | head -1000 | less
}

# Add Hadoop bin/ directory to PATH
export PATH=$PATH:$HADOOP_HOME/bin
```

You can repeat this exercise also for other users who want to use Hadoop.

Configuration Hadoop .sh files:

Hadoop-env.sh:

Change

```
# The java implementation to use. Required.

# export JAVA_HOME=/usr/lib/j2sdk1.5-sun
```

To

```
# The java implementation to use. Required.
```

```
export JAVA_HOME=/usr/lib/jvm/java-6-sun
```

conf/*-site.xml

```
$ sudo mkdir -p /app/hadoop/tmp  
$ sudo chown hduser:hadoop /app/hadoop/tmp  
# ...and if you want to tighten up security, chmod from 755 to 750...  
$ sudo chmod 750 /app/hadoop/tmp
```

conf/core-site.xml:

```
<!-- In: conf/core-site.xml -->  
  
<property>  
  
  <name>hadoop.tmp.dir</name>  
  
  <value>/app/hadoop/tmp</value>  
  
  <description>A base for other temporary directories.</description>  
  
</property>  
  
  
<property>  
  
  <name>fs.default.name</name>  
  
  <value>hdfs://localhost:54310</value>  
  
  <description>The name of the default file system. A URI whose  
  scheme and authority determine the FileSystem implementation. The  
  Uri's scheme determines the config property (fs.SCHEME.impl) naming
```

the FileSystem implementation class. The uri's authority is used to determine the host, port, etc. for a filesystem.</description>

</property>

Conf/mapred-site.xml:

```
<!-- In: conf/mapred-site.xml -->

<property>

    <name>mapred.job.tracker</name>
    <value>localhost:54311</value>
    <Description>The host and port that the MapReduce job tracker runs at. If "local", then jobs are run in-process as a single map and reduce task.
    </description>
</property>
```

Conf/hdfs-site.xml:

```
<!-- In: conf/hdfs-site.xml -->

<property>

    <name>dfs.replication</name>
    <value>1</value>
    <description>Default block replication.
    The actual number of replications can be specified when the file is created.
    The default is used if replication is not specified in create time.
    </description>
</property>
```

```
</description>  
  
</property>
```

Formatting the HDFS filesystem via the NameNode :

```
hduser@ubuntu:~$ /usr/local/hadoop/bin/hadoop namenode -format
```

The output will look like this:

```
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop namenode -format  
10/05/08 16:59:56 INFO namenode.NameNode: STARTUP_MSG:  
/*****  
STARTUP_MSG: Starting NameNode  
STARTUP_MSG: host = ubuntu/127.0.1.1  
STARTUP_MSG: args = [-format]  
STARTUP_MSG: version = 0.20.2  
STARTUP_MSG: build =  
https://svn.apache.org/repos/asf/hadoop/common/branches/branch-0.20 -r  
911707; compiled by 'chrисdo' on Fri Feb 19 08:07:34 UTC 2010  
*****/  
10/05/08 16:59:56 INFO namenode.FSNamesystem: fsOwner=hduser,hadoop  
10/05/08 16:59:56 INFO namenode.FSNamesystem: supergroup=supergroup  
10/05/08 16:59:56 INFO namenode.FSNamesystem: isPermissionEnabled=true  
10/05/08 16:59:56 INFO common.Storage: Image file of size 96 saved in  
0 seconds.  
10/05/08 16:59:57 INFO common.Storage: Storage directory .../hadoop-  
hduser/dfs/name has been successfully formatted.
```

```
10/05/08 16:59:57 INFO namenode.NameNode: SHUTDOWN_MSG:  
*****  
SHUTDOWN_MSG: Shutting down NameNode at ubuntu/127.0.1.1  
*****  
hduser@ubuntu:/usr/local/hadoop$
```

Starting single-node cluster:

```
hduser@ubuntu:~$ /usr/local/hadoop/bin/start-all.sh
```

This will start up a Namenode, Datanode, Jobtracker and a Tasktracker on your machine.

The output will look like this:

```
hduser@ubuntu:/usr/local/hadoop$ bin/start-all.sh  
  
starting namenode, logging to /usr/local/hadoop/bin/..../logs/hadoop-  
hduser-namenode-ubuntu.out  
  
localhost: starting datanode, logging to  
/usr/local/hadoop/bin/..../logs/hadoop-hduser-datanode-ubuntu.out  
  
localhost: starting secondarynamenode, logging to  
/usr/local/hadoop/bin/..../logs/hadoop-hduser-secondarynamenode-  
ubuntu.out  
  
starting jobtracker, logging to /usr/local/hadoop/bin/..../logs/hadoop-  
hduser-jobtracker-ubuntu.out  
  
localhost: starting tasktracker, logging to  
/usr/local/hadoop/bin/..../logs/hadoop-hduser-tasktracker-ubuntu.out  
  
hduser@ubuntu:/usr/local/hadoop$
```

Stopping your single-node cluster:

```
hduser@ubuntu:~$ /usr/local/hadoop/bin/stop-all.sh
```

Hbase

Configuration Document:

This Document describes setup of a standalone HBase instance that uses the local file system.
Steps to configure Hbase

- 1) Download Hbase stable released from the site <http://www.apache.org/dyn/closer.cgi/hbase/>.
- 2) Unpack .tar.gz file preferred location is to put at hadoop folder location
- 3) Before starting Hbase we have to update few configuration files

`conf/hbase-site.xml :`

<pre><property> </property></pre>	<pre><name>hbase.rootdir</name> <value>/app/hadoop/tmp</value></pre>
---	--

Replace DIRECTORY in the above with a path to a directory where you want HBase to store its data. By default, hbase.rootdir is set to /tmp/hbase-\${user.name} which means you'll lose all your data whenever your server reboots (Most operating systems clear /tmp on restart).

`conf/hbase-env.sh :`

<pre>export export HBASE_MANAGES_ZK=true</pre>	<pre>JAVA_HOME=/usr/lib/jvm/jre-1.6.0</pre>
--	---

- 4) **Start Hbase** : Start Hadoop first before starting Hbase once Hadoop clusters starts
Start hbase with command

```
start-hbase.sh
```

5) Hbase Shell Exercises :

```
Hbase shell
hbase(main):003:0> create 'test', 'cf'
0 row(s) in 1.2200 seconds
hbase(main):003:0> list 'test'
..
1 row(s) in 0.0550 seconds
hbase(main):004:0> put 'test', 'row1', 'cf:a', 'value1'
0 row(s) in 0.0560 seconds
hbase(main):005:0> put 'test', 'row2', 'cf:b', 'value2'
0 row(s) in 0.0370 seconds
hbase(main):006:0> put 'test', 'row3', 'cf:c', 'value3'
    0  row(s) in 0.0450 seconds
```

6) Retrieve data from Hbase :

```
hbase(main):007:0> scan 'test'
ROW      COLUMN+CELL
row1    column=cf:a, timestamp=1288380727188, value=value1
row2    column=cf:b, timestamp=1288380738440, value=value2
row3    column=cf:c, timestamp=1288380747365, value=value3
3 row(s) in 0.0590 seconds
hbase(main):007:0> scan 'test'
ROW      COLUMN+CELL
row1    column=cf:a, timestamp=1288380727188, value=value1
row2    column=cf:b, timestamp=1288380738440, value=value2
row3    column=cf:c, timestamp=1288380747365, value=value3
```

Get a single row as follows

```
hbase(main):008:0> get 'test', 'row1'  
COLUMN      CELL  
cf:a        timestamp=1288380727188, value=value1  
1 row(s) in 0.0400 seconds
```

Now, disable and drop your table. This will clean up all done above.

```
hbase(main):012:0> disable 'test'  
0 row(s) in 1.0930 seconds  
hbase (main):013:0> drop 'test'  
0 row(s) in 0.0770 seconds
```

Exit the shell by typing exit.

```
hbase (main):014:0> exit
```

7) Stopping Hbase :

```
$ ./bin/stop-hbase.sh  
stopping hbase.....
```

HDFS components properties and description

tHDFSConnection properties

Function	tHDFSConnection provides connection to the Hadoop distributed file system (HDFS) .
Basic Settings	1)Hadoop Version : Apache 1.0.0 2)NameNode URI : "http://localhost:50075/" 3)User Name : "hduser" 4)Hadoop Property : To use custom configuration for the Hadoop of interest, complete this table with the property or properties to be customized.

Scenario: Connect to HDFS

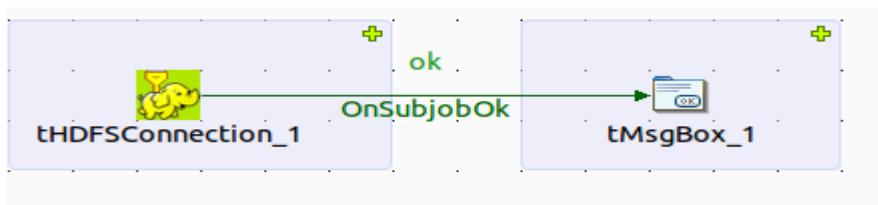
The following scenario describes a simple Job that connect to HDFS and on successful connection display a message.

Setting up the Job

1.Drop the following components from the **Palette** onto the design workspace:

tHDFSConnection, tMsgBox.

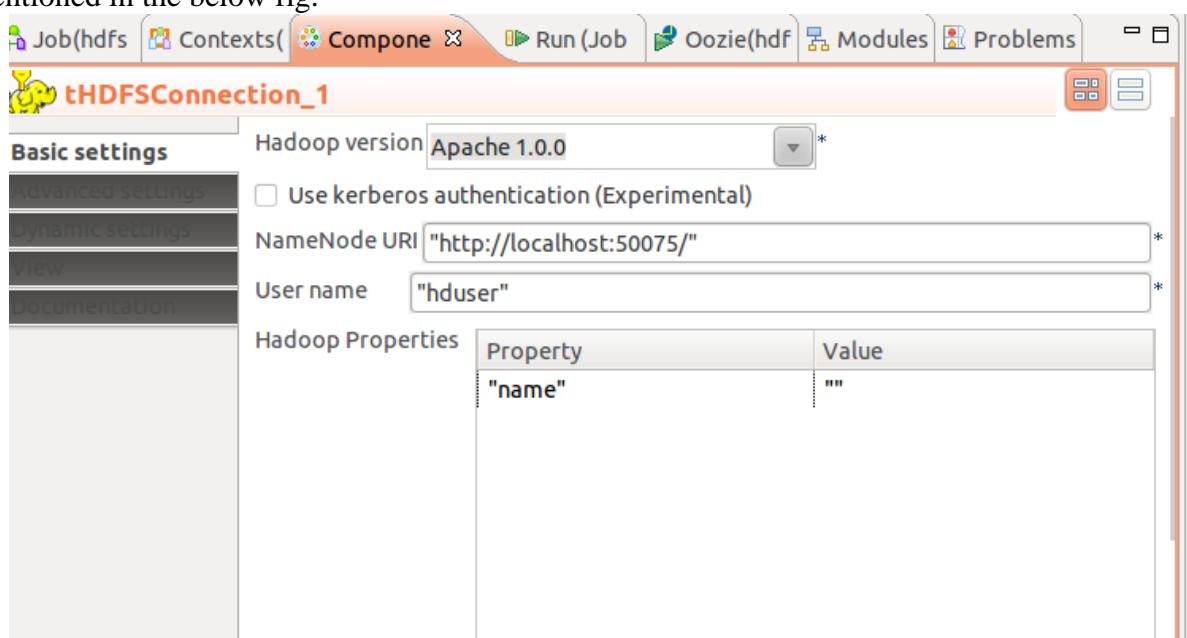
2)Connect **tHDFSConnection** to **tMsgBox** using an **OnSubjobOk** connection.



Configuring

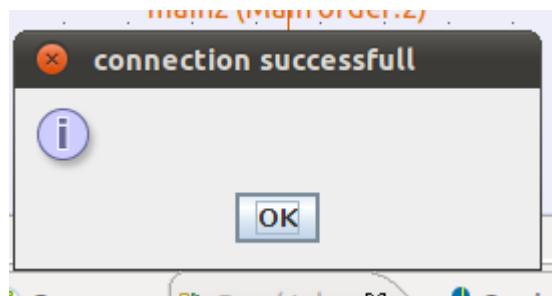
the components

1)Double-click **tHDFSConnection** to define the component in its **Basic settings** view as mentioned in the below fig.



Executing the Job

Save the Job and press **F6** to execute it.



tHDFSInput properties

Function	tHDFSInput reads a file located on a given Hadoop distributed file system (HDFS) and puts the data of interest from this file into a Talend schema. Then it passes the data to the component that follows.
Basic Settings	1)Schema : Define the number of fields that will be processed and passed on to the next component. 2)Hadoop Version : Apache 1.0.0 3)NameNode URI : "hdfs://localhost:54310/" 4)User Name : "hduser" 5)File Name : Browse to, or enter the path to the source files in HDFS.

tHDFSOutput properties

Function	tHDFSOutput writes data flows it receives onto a given Hadoop distributed file system (HDFS).
Basic Settings	1)Schema : Define the number of fields that will be processed and passed on to the next component. 2)Hadoop Version : Apache 1.0.0 3)NameNode URI : "hdfs://localhost:54310/" 4)User Name : "hduser" 5)File Name : Browse to, or enter the path to the source files in HDFS. 6>Action : Select an operation in HDFS: Create: Creates a file with data using the file name defined in the File Name field. Overwrite: Overwrites the data in the file specified in the File Name field. Append: Inserts the data into the file specified in the File Name field. The specified file is created automatically if it does not exist.

Scenario: Reading and writing files from HDFS

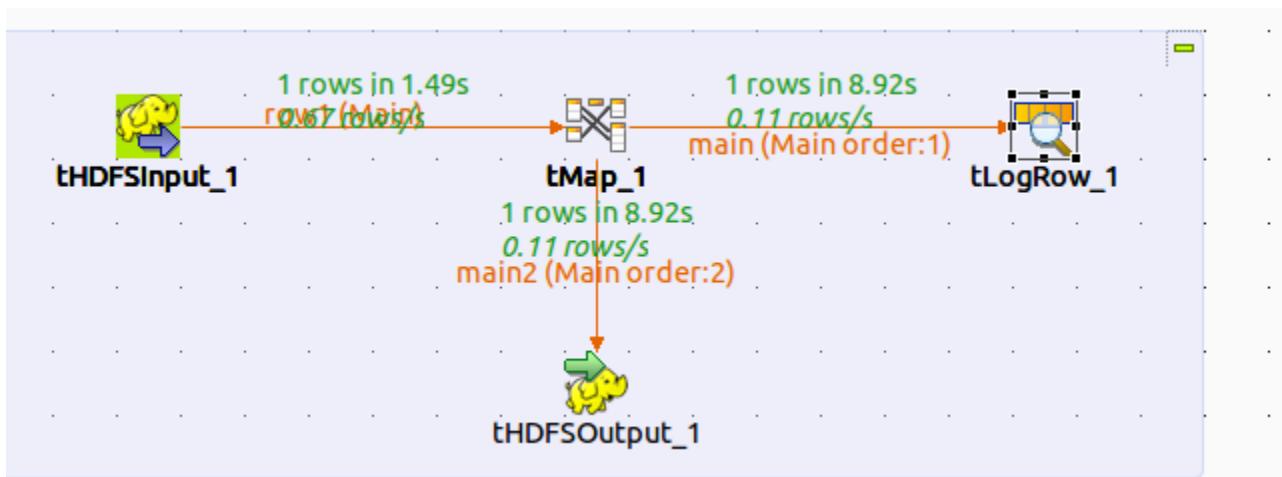
The following scenario describes a simple Job that will read and write the file into the hdfs.

Setting up the Job

1. Drop the following components from the **Palette** onto the design workspace:

tHDFSInput, tHDFSOutput,tMap,tLogRow.

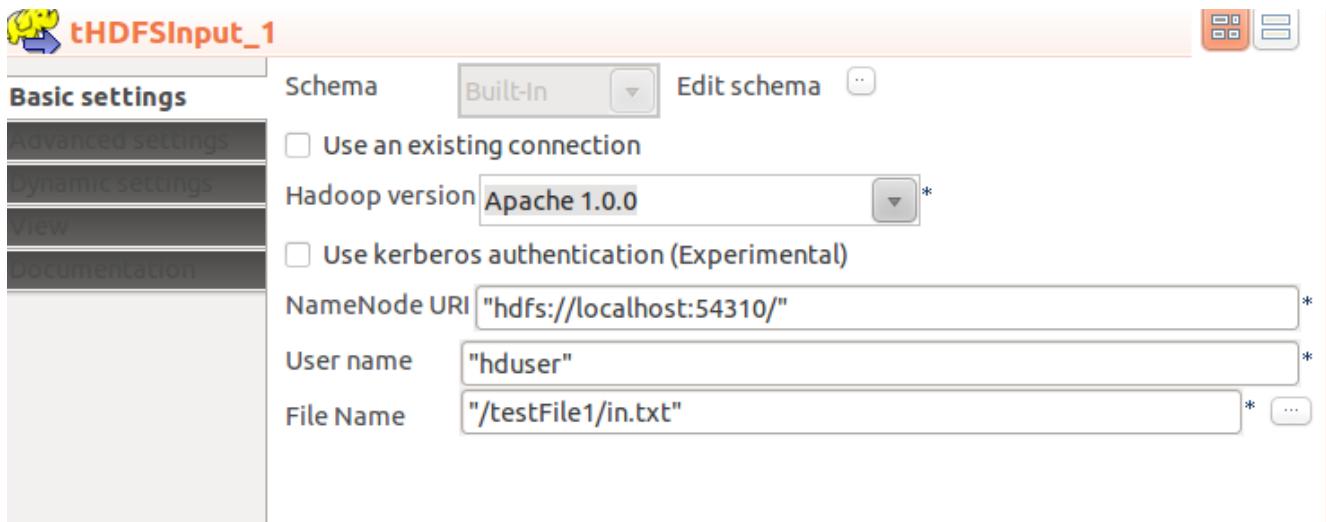
2. Connect **tHDFSInput** to **tMap** using a **Row > Main** connection.
3. Connect **tMap** to **tLogRow** using a **Row > Main** connection
4. Connect **tMap** to **tHDFSOutput** using a **Row > Main** connection



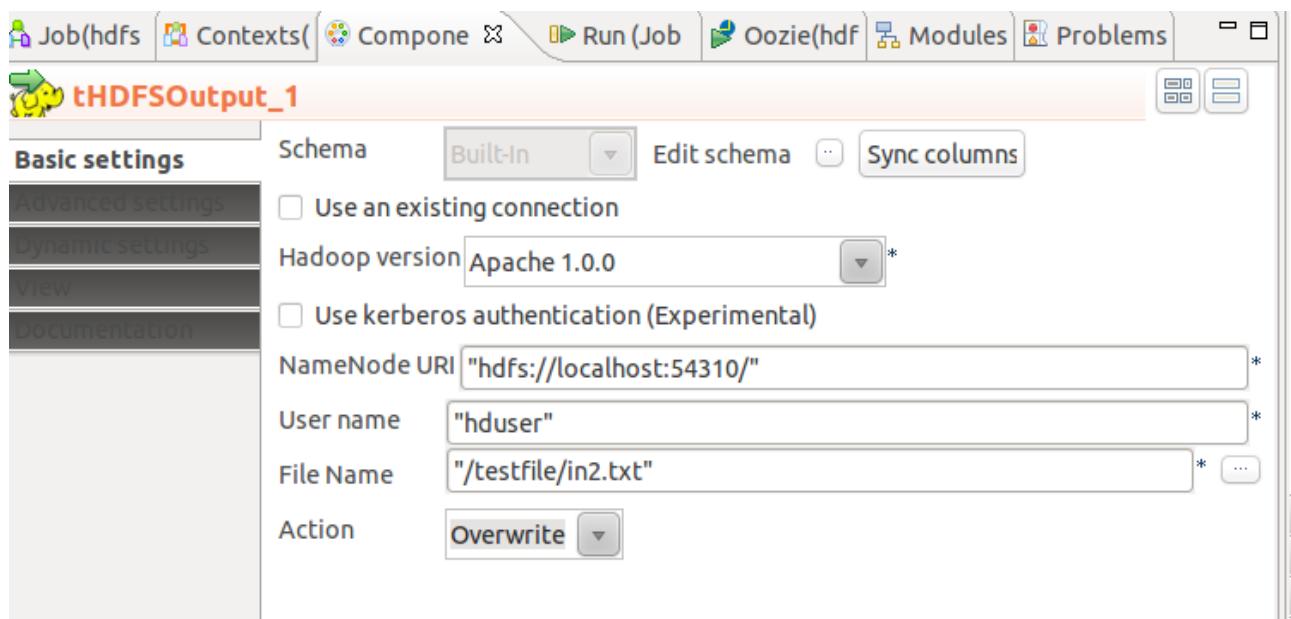
Configuring the component

Configuring the components

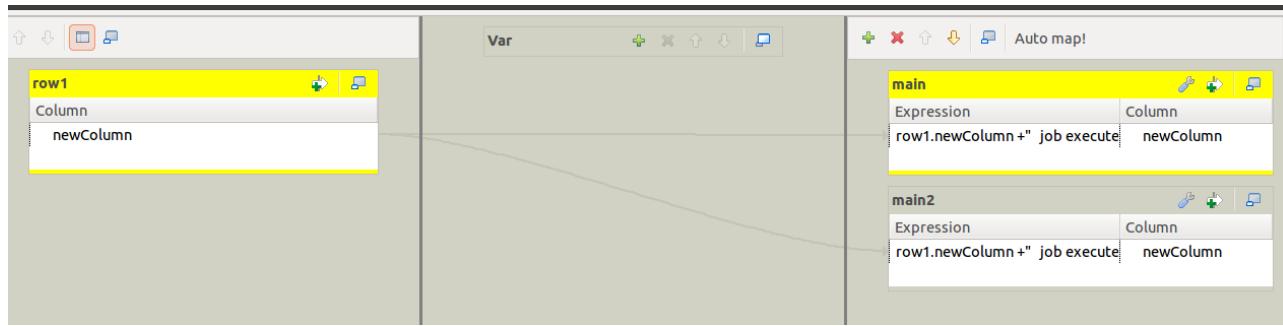
- 1) Double-click **tHDFSInput** to define the component in its **Basic settings** view as mentioned in the below fig.



2) Double-click **tHDFSOutput** to define the component in its **Basic settings** view as mentioned in the below fig.



3) Double-click **tMap** and Set the **Schema** to **Built-In** and click the three-dot [...] button next to **Map Editor** to describe the data structure you want to create In this scenario, the schema contains one column. Change the Expression of output column by clicking on three-dot [...] button next to output column.



Save the Job and press **F6** to execute it.

The *in2.txt* file is created and loaded into the HDFS.

File: [/testfile/in2.txt](#)

Goto :

[Go back to dir listing](#)

[Advanced view/download options](#)

Hello world!, ia am in pune job executed successfully

tHDFSPut properties

Function	tHDFSPut loads data into Hadoop distributed file system(HDFS).
Basic Settings	1)Hadoop Version : Apache 1.0.0 2)Host : “localhost” 3)port : 54310 4)NameNode URI : "hdfs://localhost:54310/" 5)User Name : "hduser" 6)Local Dir : Local directory where are stored the files to be loaded into HDFS. 7)HDFS Dir : Location to store the files in HDFS. 8)Overwrite File : Options to overwrite or not the existing file with the new one. 9)Files : - File mask: the file name to be selected from the local directory. - New name: a new name to the loaded file.

tHDFSGet properties

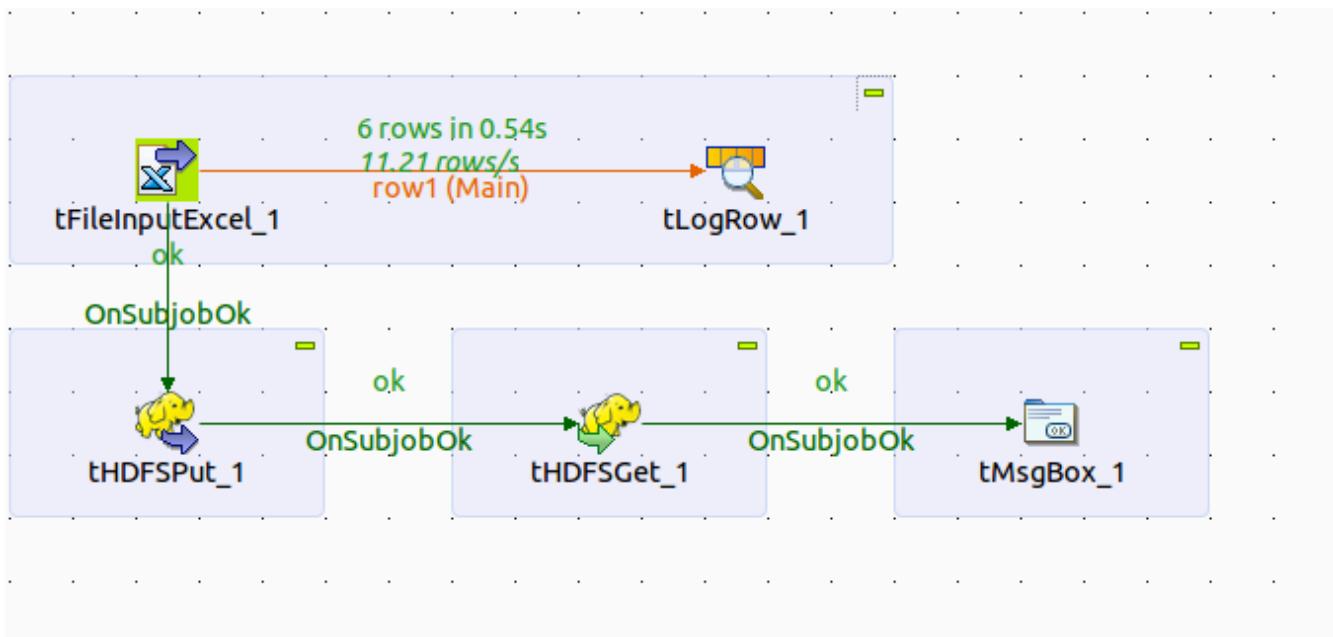
Function	tHDFSGet gets data from Hadoop distributed file system(HDFS).
Basic Settings	1)Hadoop Version : Apache 1.0.0 2)Host : “localhost” 3)port : 54310 4)NameNode URI : "hdfs://localhost:54310/" 5)User Name : "hduser" 6)Local Dir : Local directory where are stored the files to be loaded into HDFS. 7)HDFS Dir : Location to store the files in HDFS. 8)Overwrite File : Options to overwrite or not the existing file with the new one. 9)Append : Select this check box to add the new rows at the end of the records 10)Files : - File mask: the file name to be selected from the local directory. - New name: a new name to the loaded file.

Scenario: To Load and get the data from HDFS

The following scenario describes a simple Job that will load and get the data from the hdfs.

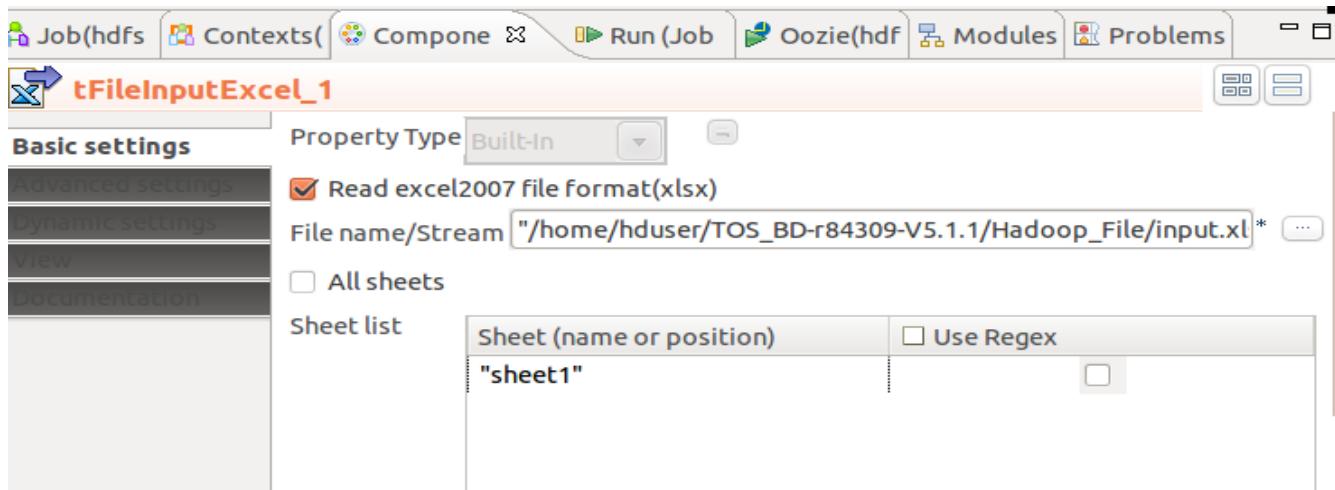
Setting up the Job

1. Drop the following components from the **Palette** onto the design workspace:
tFileInputExcel, tHDFSPut,tHDFSGet,tLogRow,tMsgBox.
5. Connect **tFileInputExcel** to **tLogRow** using a Row > Main connection.
6. Connect **tFileInputExcel** to **tHDFSPut** using an **OnSubjobOk** connection.
7. Connect **tHDFSPut** to **tHDFSGet** using an **OnSubjobOk** connection.
8. Connect **tHDFSGet** to **tMsgBox** using an **OnSubjobOk** connection.



Configuring the component

- 1) Double-click **tFileInputExcel** to define the component in its **Basic settings** view as mentioned in the below fig.



- 2) Set the Schema to Built-In and click the three-dot [...] button next to Schema Editor to describe the data structure you want to create In this scenario,

Schema of tFileInputExcel_1

FileInputExcel_1

Column	Key	Type	NotNull	Date Pattern	Length	Precision	Default	Comments
name	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
level	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
tech	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					

Buttons:

OK **Cancel**

3) Double-click **tHDFSPut** to define the component in its Basic settings view as mentioned in the below fig.

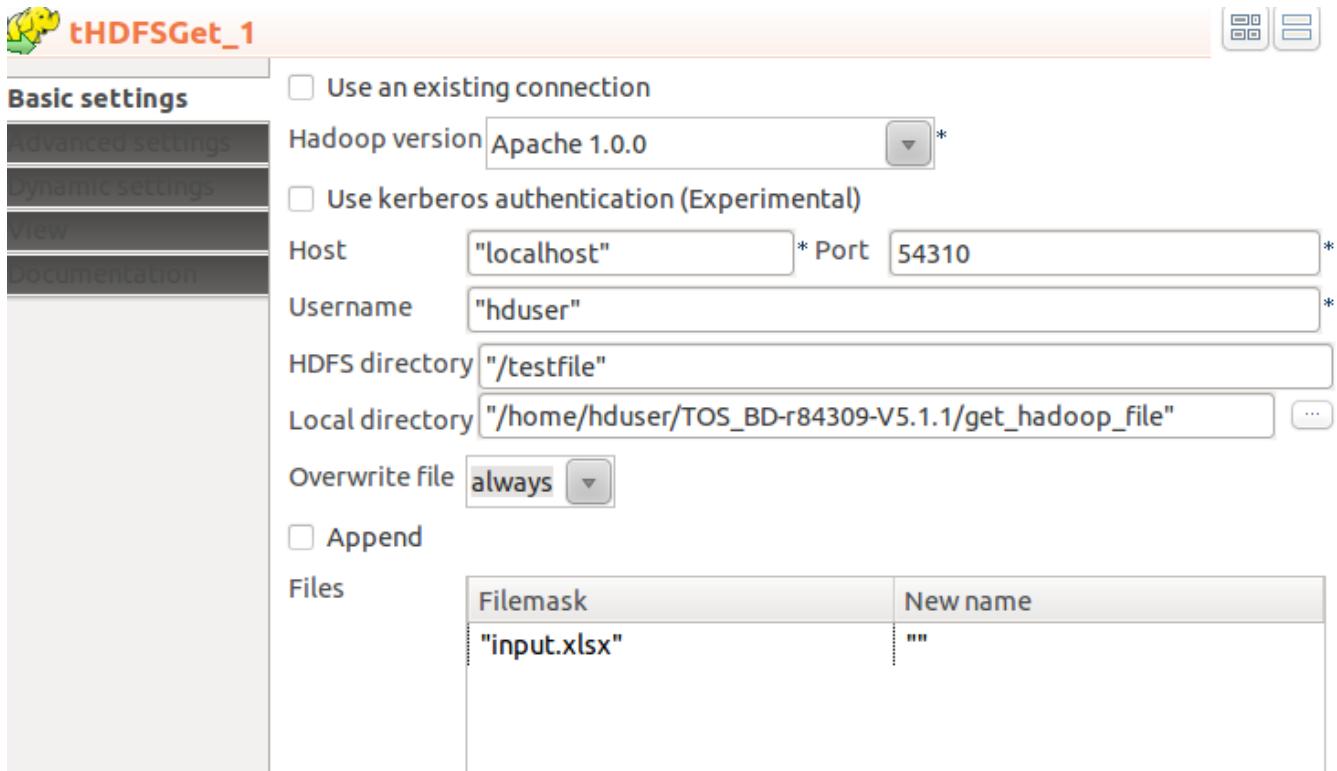
tHDFSPut_1

Basic settings

- Use an existing connection
- Hadoop version **Apache 1.0.0**
- Use kerberos authentication (Experimental)
- Host **"localhost"** * Port **54310**
- Username **"hduser"**
- Local directory **"/home/hduser/TOS_BD-r84309-V5.1.1/Hadoop_File"**
- HDFS directory **"/testfile"**
- Overwrite file **always**
- Files

Filemask	New name
"input.xlsx"	""

4) Double-click **tHDFSGet** to define the component in its Basic settings view as mentioned in the below fig.



5)Double-click tMsgBox to define the component in its Basic settings view.

Executing the Job

Save the Job and press F6 to execute it.

The *input.xlsx* file is created and loaded into the HDFS.

Contents of directory /testfile

Goto : go

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
in2.txt	file	0.05 KB	3	64 MB	2012-06-11 02:55	rw-r--r--	hduser	supergroup
input.xlsx	file	8.22 KB	3	64 MB	2012-06-11 00:41	rw-r--r--	hduser	supergroup

[Go back to DFS home](#)



Apache Hive

Configuration Document:

This Document describes setup of a standalone Hive instance that uses the local file system. Steps to configure Hive with Apache Hadoop.

- 8) Download Pig stable released from the site <http://www.apache.org/dyn/closer.cgi/hive/>.
- 9) Unpack tar.gz file. preferred location to put at hadoop folder location
- 10) Set the environment variable HIVE_HOME to point to the installation directory
 - a) \$ cd hive-x.y.z
 - b) \$ export HIVE_HOME={{pwd}}**
 - c) add \$HIVE_HOME/bin to your PATH
\$ export PATH=\$HIVE_HOME/bin:\$PATH
- 11) /tmp and /user/hive/warehouse (aka hive.metastore.warehouse.dir) and set them chmod g+w in HDFS before a table can be created in Hive
- 12) Commands to perform this setup :


```
$ $HADOOP_HOME/bin/hadoop fs -mkdir          /tmp
$ $HADOOP_HOME/bin/hadoop fs -mkdir          /user/hive/warehouse
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w      /tmp
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w      /user/hive/warehouse
```

 - a) Hive default configuration is stored in <install-dir>/conf/hive-default.xml Configuration variables can be changed by (re-)defining them in <install-dir>/conf/hive-site.xml
 - b) The location of the Hive configuration directory can be changed by setting the HIVE_CONF_DIR environment variable.
 - c) Log4j configuration is stored in <install-dir>/conf/hive-log4j.properties
 - d) Hive configuration is an overlay on top of hadoop - meaning the hadoop configuration variables are inherited by default.
 - e) Hive configuration can be manipulated by:
 - f) Editing hive-site.xml and defining any desired variables (including hadoop variables) in it
 - g) From the cli using the set command (see below)
 - h) By invoking hive using the syntax:

- \$ bin/hive -hiveconf x1=y1 -hiveconf x2=y2
this sets the variables x1 and x2 to y1 and y2 respectively
- i) By setting the HIVE_OPTS environment variable to "-hiveconf x1=y1 -hiveconf x2=y2" which does the same as above

DDL Operations :

Creating Hive tables and browsing through them

```
hive> CREATE TABLE pokes (foo INT, bar STRING);
```

Creates a table called pokes with two columns, the first being an integer and the other a string

```
hive> CREATE TABLE invites (foo INT, bar STRING) PARTITIONED BY (ds STRING);
```

Creates a table called invites with two columns and a partition column called ds. The partition column is a virtual column. It is not part of the data itself but is derived from the partition that a particular dataset is loaded into.

By default, tables are assumed to be of text input format and the delimiters are assumed to be ^A(ctrl-a).

```
hive> SHOW TABLES;
```

lists all the tables

```
hive> SHOW TABLES '.*s';
```

lists all the table that end with 's'. The pattern matching follows Java regular expressions. Check out this link for documentation <http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html>

```
hive> DESCRIBE invites;
```

shows the list of columns

As for altering tables, table names can be changed and additional columns can be dropped:

```
hive> ALTER TABLE pokes ADD COLUMNS (new_col INT);
hive> ALTER TABLE invites ADD COLUMNS (new_col2 INT COMMENT 'a comment');
hive> ALTER TABLE events RENAME TO 3koobecaf;
```

Dropping tables:

```
hive> DROP TABLE pokes;
```

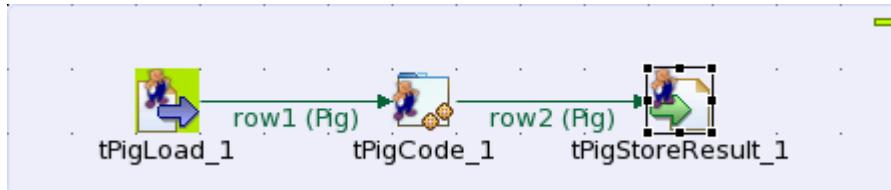
Pig Components

1) tPigCode

Function This component allows you to enter personalized Pig code to integrate it in **Talend** program. You can execute this code only once.

Scenario: Selecting a column of data from an input file and store it into a local file

This scenario describes a three-component Job that selects a column of data that matches filter condition defined in **tPigCode** and stores the result into a local file.

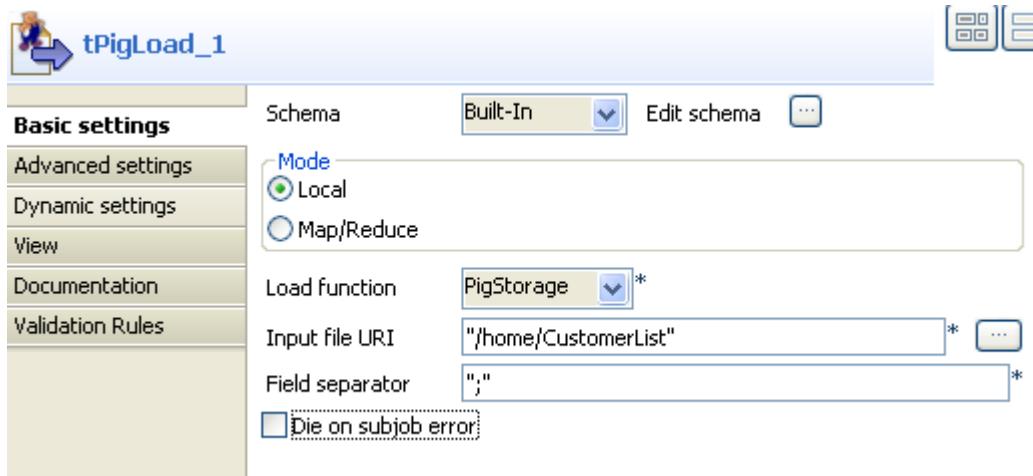


Setting up the Job

1. Drop the following components from the **Palette** to the design workspace: **tPigCode**, **tPigLoad**, **tPigStoreResult**.
2. Right-click **tPigLoad** to connect it to **tPigCode** using a **Row > Pig Combine** connection.
3. Right-click **tPigCode** to connect it to **tPigStoreResult** using a **Row > Pig Combine** connection.

Loading the data

1. Double-click **tPigLoad** to open its **Basic settings** view.



2. Click the three-dot button next to **Edit schema** to add columns for **tPigLoad**.

Column	Key	Type	N..	Dat...	L..	O...	P...	I...
Name	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
Country	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
Age	<input type="checkbox"/>	Character	<input checked="" type="checkbox"/>					

Buttons at the bottom: +, -, Up, Down, Save, Load, Delete, Help.

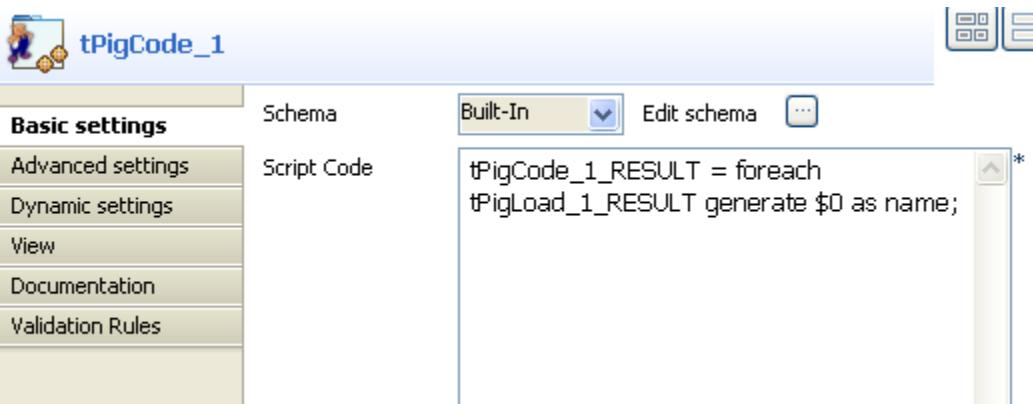
3. Click the plus button to add *Name*, *Country* and *Age* and click **OK** to save the setting.
4. Select **Local** from the **Mode** area.
5. Fill in the **Input filename** field with the full path to the input file.

In this scenario, the input file is *CustomerList* which contains rows of names, country names and age.

6. Select **PigStorage** from the **Load function** list.
7. Leave rest of the settings as they are.

Configuring the tPigCode component

1. Double-click **tPigCode** component to open its **Basic settings** view.



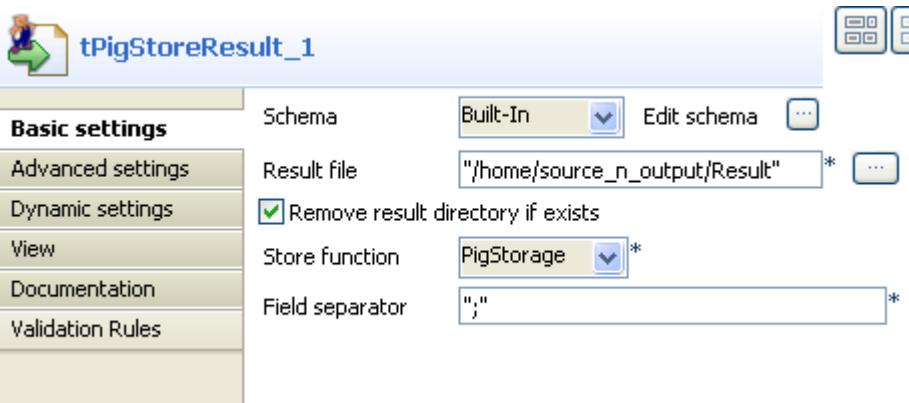
2. Click **Sync columns** to retrieve the schema structure from the preceding component.
3. Fill in the **Script Code** field with following expression:

```
tPigCode_1_RESULT = foreach tPigLoad_1_RESULT generate $0 as name;
```

This filter expression selects column *Name* from *CustomerList*.

Saving the result data to a local file

1. Double-click **tPigStoreResult** to open its **Basic settings** view.



2. Click **Sync columns** to retrieve the schema structure from the preceding component.
3. Fill in the **Result file** field with the full path to the result file.

In this scenario, the result is saved in *Result* file.

4. Select **Remove result directory if exists**.
5. Select **PigStorage** from the **Store function** list.
6. Leave rest of the settings as they are.

Executing the Job

Save your Job and press **F6** to run it.

Result	
Name	
Mike	
Silvia	
Romeo	
Ahmad	
Toyota	
Manik	
Natasha	
Billy	
Eminem	
Bill	
Fenricka	
Huamei	
Selena	
Julio	
Pantalion	
Simao	
Nancy	
Gaddafi	
Zidane	
Didi	
Juan	
Bob	
Mario	
Ricky	

The *Result* file is generated containing the selected column of data.

2) tPigFilterRow

Function	The tPigFilterRow component filters the input flow in a Pig chain based on conditions set on given column(s).
-----------------	--

Scenario: Filtering rows of data based on a condition and saving the result to a local file

This scenario describes a four-component Job that filters a list of customers to find out customers from a particular country, and saves the result list to a local file. Before the input data is filtered, duplicate entries are first removed from the list.

The input file contains three columns: Name, Country, and Age, and it has some duplicate entries, as shown below:

Mario;PuertoRico;49

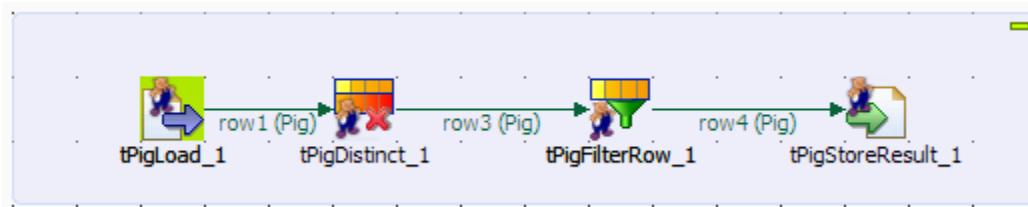
```

Mike;USA;22
Ricky;PuertoRico;37
Silvia;Spain;20
Billy;Canada;21
Ricky;PuertoRico;37
Romeo;UK;19
Natasha;Russia;25
Juan;Cuba;23
Bob;Jamaica;55
Mario;PuertoRico;49

```

Dropping and linking components

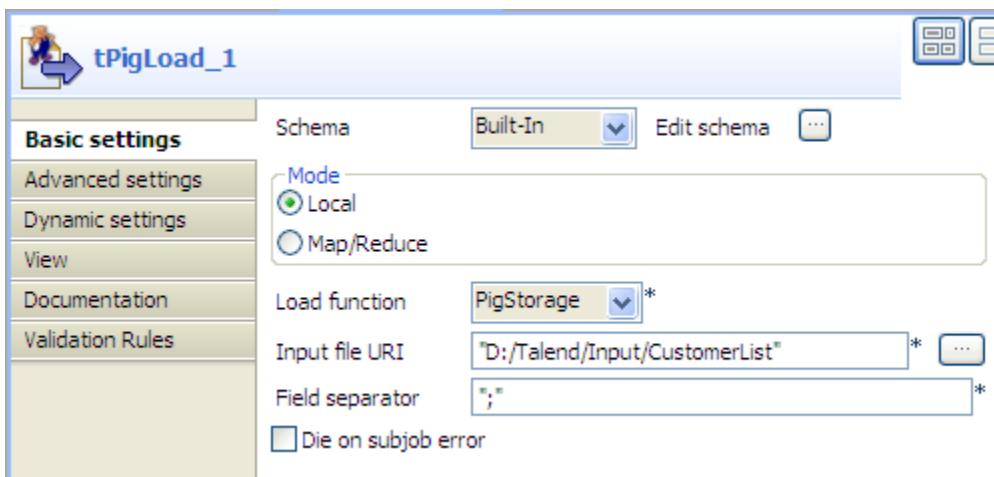
1. Drop the following components from the **Palette** to the design workspace: **tPigLoad**, **tPigDistinct**, **tPigFilterRow**, and **tPigStoreResult**.
2. Right-click **tPigLoad**, select **Row > Pig Combine** from the contextual menu, and click **tPigDistinct** to link these two components.
3. Repeat this operation to link **tPigDistinct** to **tPigFilterRow**, and **tPigFilterRow** to **tPigStoreResult** using **Row > Pig Combine** connections to form a Pig chain.



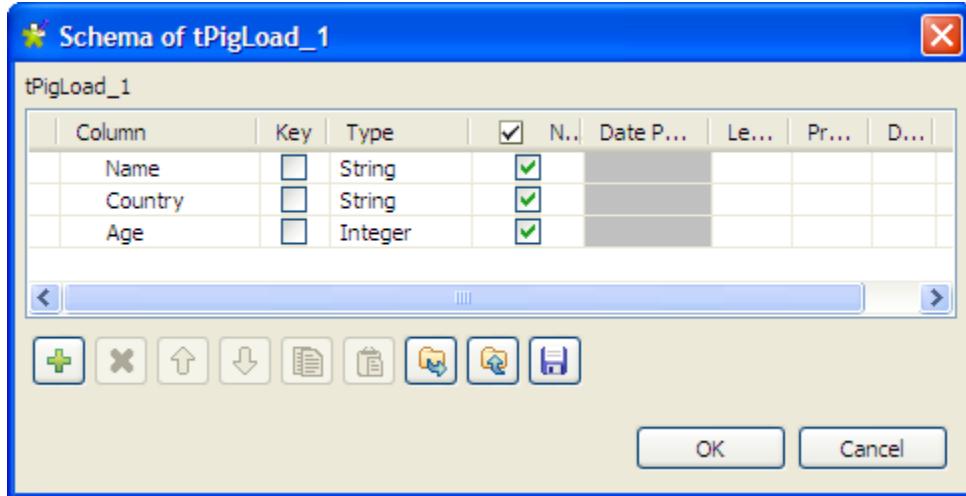
Configuring the components

Procedure 1.1. Loading the input data and removing duplicates

1. Double-click **tPigLoad** to open its **Basic settings** view.



2. Click the [...] button next to **Edit schema** to open the **[Schema]** dialog box.

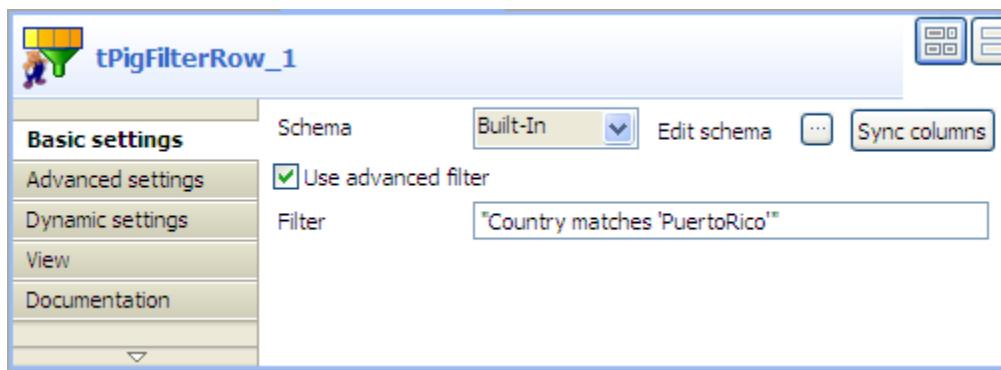


3. Click the [+] button to add three columns according to the data structure of the input file: *Name* (string), *Country* (string) and *Age* (integer), and then click **OK** to save the setting and close the dialog box.
4. Click **Local** in the **Mode** area.
5. Fill in the **Input file URI** field with the full path to the input file.
6. Select **PigStorage** from the **Load function** list, and leave rest of the settings as they are.
7. Double-click **tPigDistinct** to open its **Basic settings** view, and click **Sync columns** to make sure that the input schema structure is correctly propagated from the preceding component.

This component will remove any duplicates from the data flow.

Procedure 1.2. Configuring the filter

1. Double-click **tPigFilterRow** to open its **Basic settings** view.



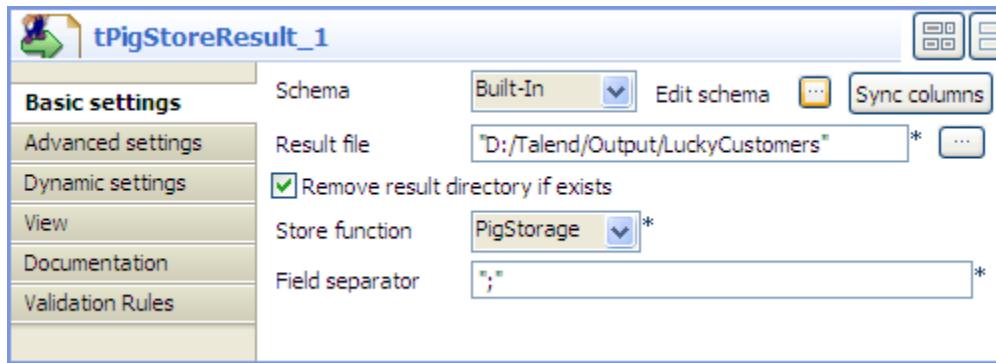
2. Click **Sync columns** to make sure that the input schema structure is correctly propagated from the preceding component.
3. Select **Use advanced filter** and fill in the **Filter** field with filter expression:

"Country matches 'PuertoRico'"

This filter expression selects rows of data that contains "PuertoRico" in the Country column.

Procedure 1.3. Configuring the file output

1. Double-click **tPigStoreResult** to open its **Basic settings** view.



2. Click **Sync columns** to make sure that the input schema structure is correctly propagated from the preceding component.
3. Fill in the **Result file** field with the full path to the result file.
4. If the target file already exists, select the **Remove result directory if exists** check box.
5. Select **PigStorage** from the **Store function** list, and leave rest of the settings as they are.

Saving and executing the Job

1. Press **Ctrl+S** to save your Job.
2. Press **F6** or click the **Run** button on the **Run** tab to run the Job.

The result file contains the information of customers from the specified country.

LuckyCustomers	
1	Mario;PuertoRico;49
2	Ricky;PuertoRico;37
3	

3) tPigJoin

Function	This component allows you to perform join of two files based on join keys.
-----------------	--

Scenario: Joining two files based on an exact match and saving the result to a local file

This scenario describes a four-component Job that combines data of an input file and a reference file that matches a given join key, removes unwanted columns, and then saves the final result to a local file.

The main input file contains the information about people's IDs, first names, last names, group IDs, and salaries, as shown below:

```
1;Woodrow;Johnson;3;1013.39
2;Millard;Monroe;2;8077.59
3;Calvin;Eisenhower;3;6866.88
4;Lyndon;Wilson;3;5726.28
5;Ronald;Garfield;2;4158.58
6;Rutherford;Buchanan;3;2897.00
7;Calvin;Coolidge;1;6650.66
8;Ulysses;Roosevelt;2;7854.78
9;Grover;Tyler;1;5226.88
10;Bill;Tyler;2;8964.66
```

The reference file contains only the information of group IDs and group names:

```
1;group_A
2;group_B
```

Dropping and linking the components

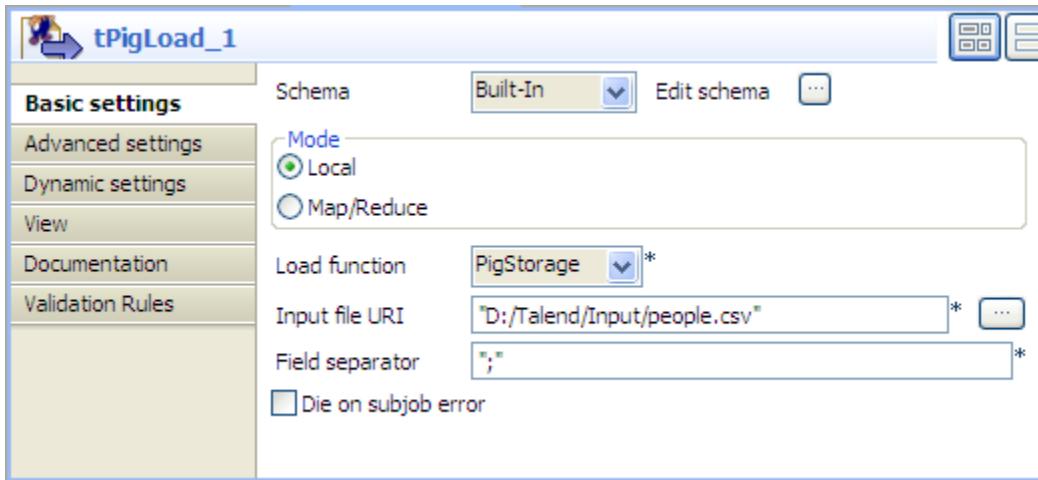
1. Drop the following components from the **Palette** to the design workspace: **tPigLoad**, **tPigJoin**, **tPigFilterColumns**, and **tPigStoreResult**.
2. Connect these components in a series using **Row > Pig Combine** connections.



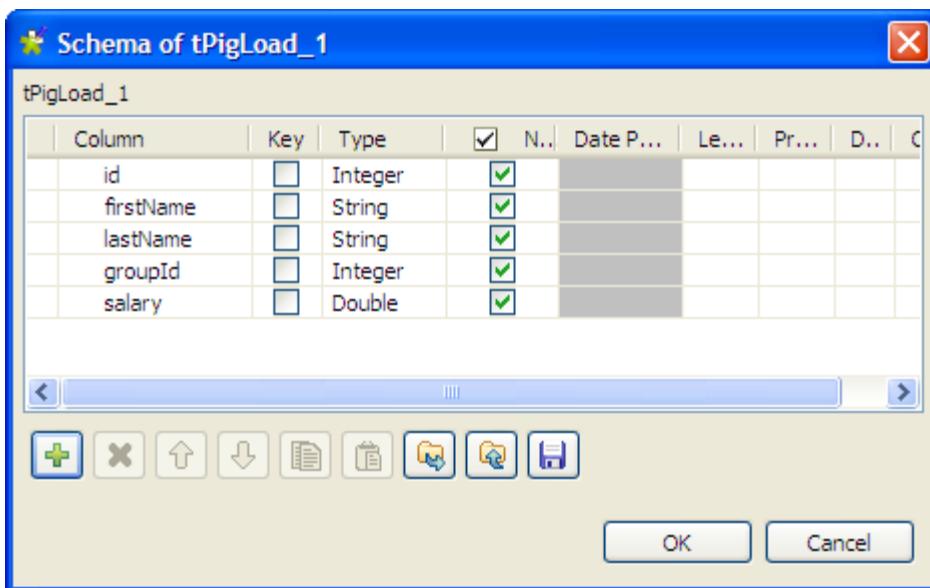
Configuring the components

Procedure 1.4. Loading the main input file

1. Double-click **tPigLoad** to open its **Basic settings** view.



2. Click the [...] button next to **Edit schema** to open the **[Schema]** dialog box.



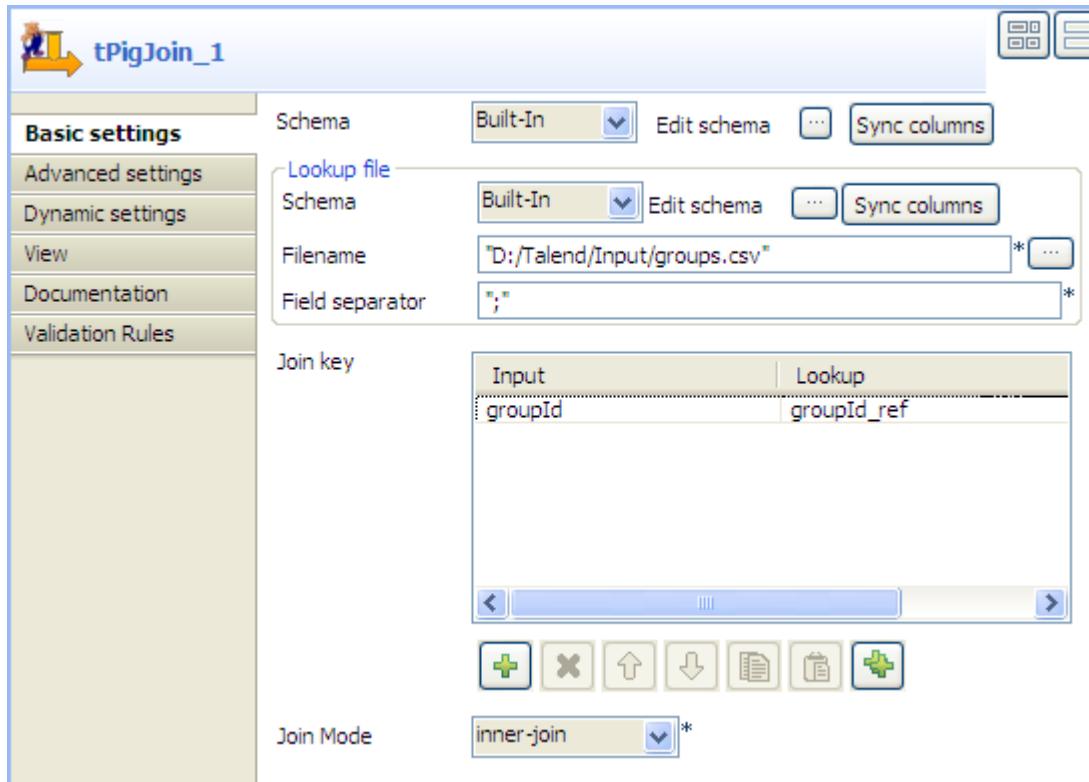
3. Click the [+] button to add columns, name them and define the column types according to the structure of the input file. In this example, the input schema has five columns: id (integer), firstName (string), lastName (string), groupId (integer), and salary (double).

Then click **OK** to validate the setting and close the dialog box.

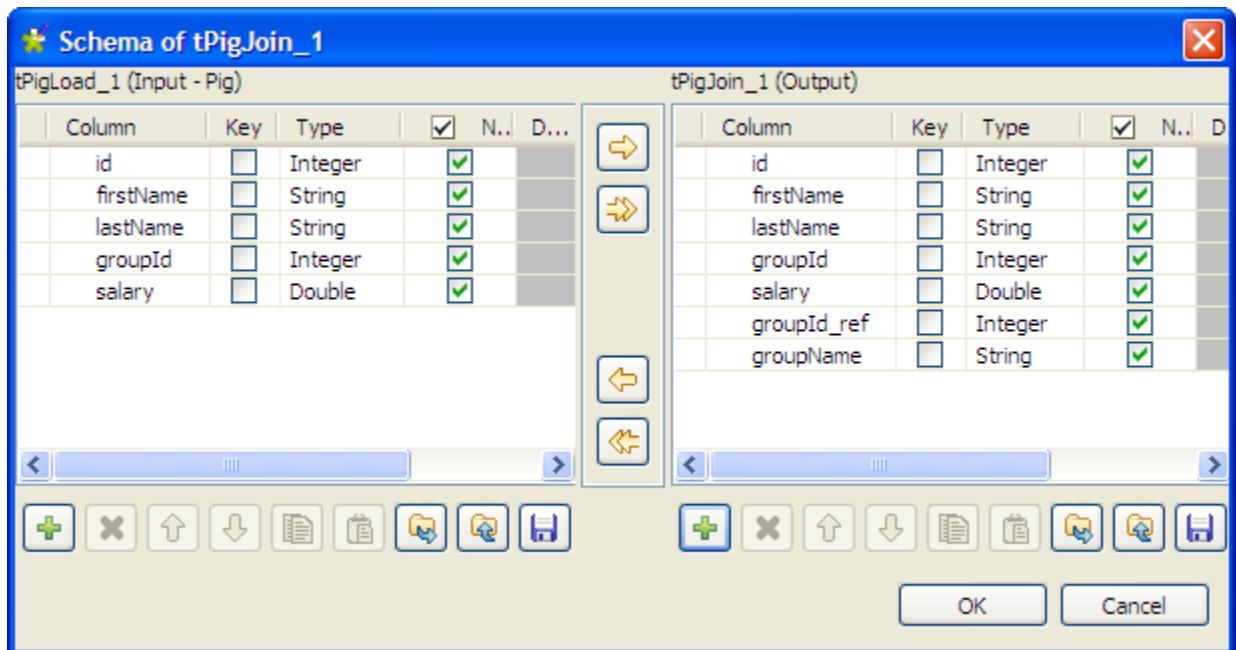
4. Click **Local** in the **Mode** area.
5. Select **PigStorage** from the **Load function** list.
6. Fill in the **Input file URI** field with the full path to the input file, and leave the rest of the setting as they are.

Procedure 1.5. Loading the reference file and setting up an inner join

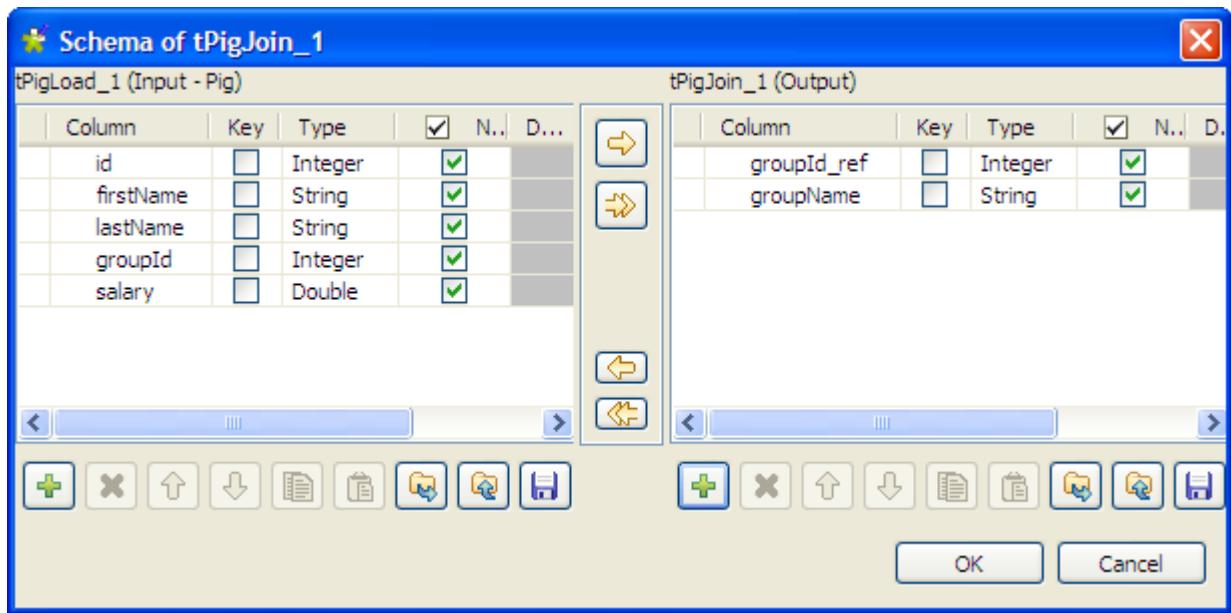
1. Double-click **tPigJoin** to open its **Basic settings** view.



2. Click the [...] for the main schema to open the [Schema] dialog box.



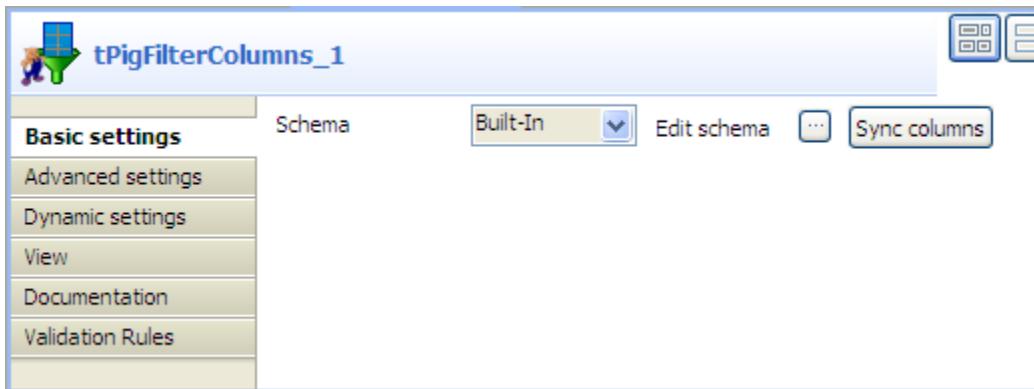
3. Check that input schema is correctly retrieved from the preceding component. If needed, click the [->] button to copy all the columns of the input schema to the output schema.
4. Click the [+] button under the output panel to add new columns according to the data structure of the reference file, groupId_ref (integer) and groupName (string) in this example. Then click **OK** to close the dialog box.
5. Click the [...] for the schema lookup flow to open the **[Schema]** dialog box.



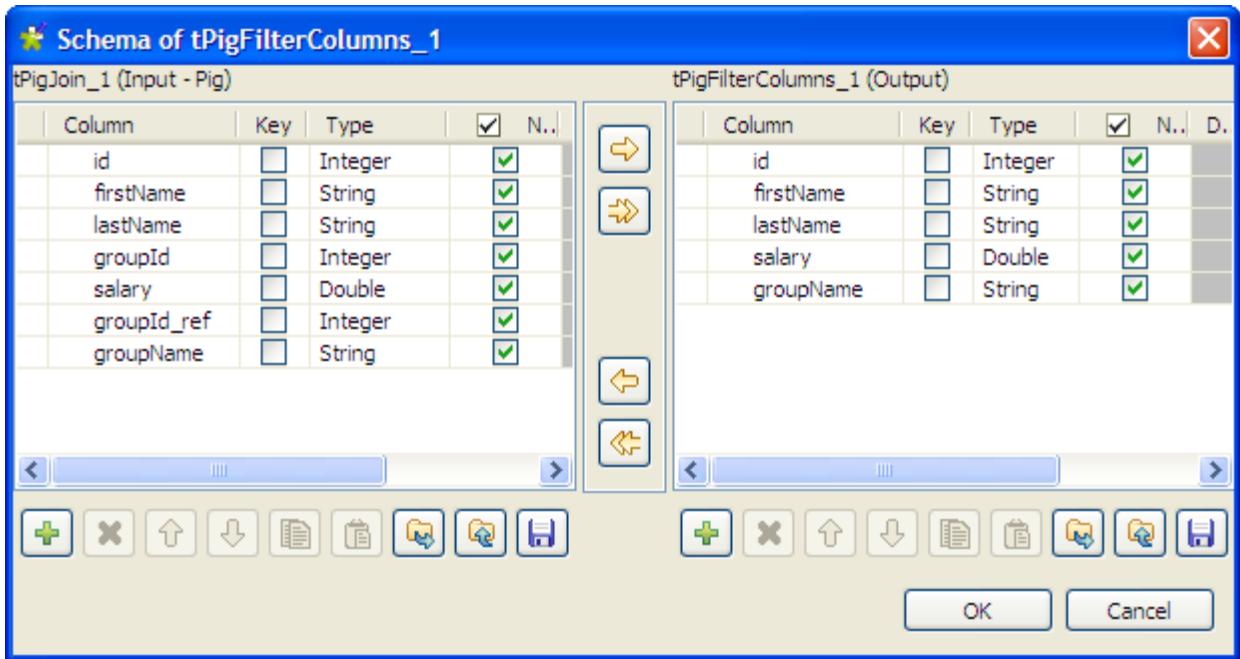
6. Click the [+] button under the output panel to add two columns: groupId_ref (integer) and groupName (string), and then click **OK** to close the dialog box.
7. In the **Filename** field, specify the full path to the reference file.
8. Click the [+] button under the **Join key** table to add a new line, and select *groupId* and *groupId_ref* respectively from the **Input** and **Lookup** lists to match data from the main input flow with data from the lookup flow based on the group ID.
9. From the **Join Mode** list, select **inner-join**.

Procedure 1.6. Defining the final output schema and the output file

1. Double-click **tPigFilterColumns** to open its **Basic settings** view.



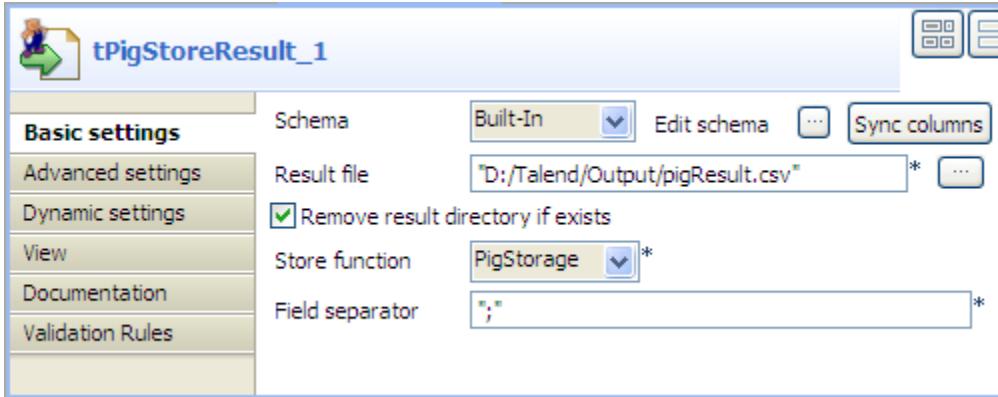
- Click the [...] button next to **Edit schema** to open the [Schema] dialog box.



- From the input schema, select the columns you want to include in your result file by clicking them one after another while pressing the **Shift** key, and click the [->] button to copy them to the output schema. Then, click **OK** to validate the schema setting and close the dialog box.

In this example, we want the result file to include all the information except the group IDs.

- Double-click **tPigStoreResult** to open its **Basic settings** view.



5. Click **Sync columns** to retrieve the schema structure from the preceding component.
6. Fill in the **Result file** field with the full path to the result file, and select the **Remove result file directory if exists** check box.
7. Select **PigStorage** from the **Store function** list, and leave rest of the settings as they are.

Saving and executing the Job

1. Press **Ctrl+S** to save your Job.
2. Press **F6** or click **Run** on the **Run** tab to run the Job.

The result file includes all the information related to people of group A and group B, except their group IDs.

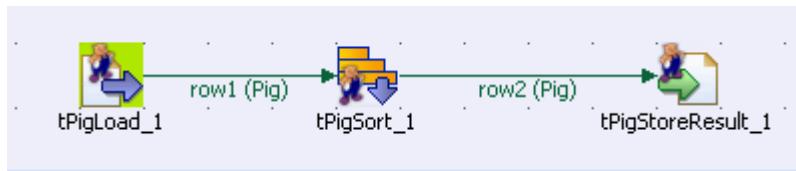
	pigResult.csv
1	7;Calvin;Coolidge;6650.66;group_A
2	9;Grover;Tyler;5226.88;group_A
3	2;Millard;Monroe;8077.59;group_B
4	5;Ronald;Garfield;4158.58;group_B
5	8;Ulysses;Roosevelt;7854.78;group_B
6	10;Bill;Tyler;8964.66;group_B
7	

4) tPigSort

Function	This component allows you to sort a relation based on one or more defined sort keys.
-----------------	--

Scenario: Sorting data in ascending order

This scenario describes a three-component Job that sorts rows of data based on one or more sorting conditions and stores the result into a local file.

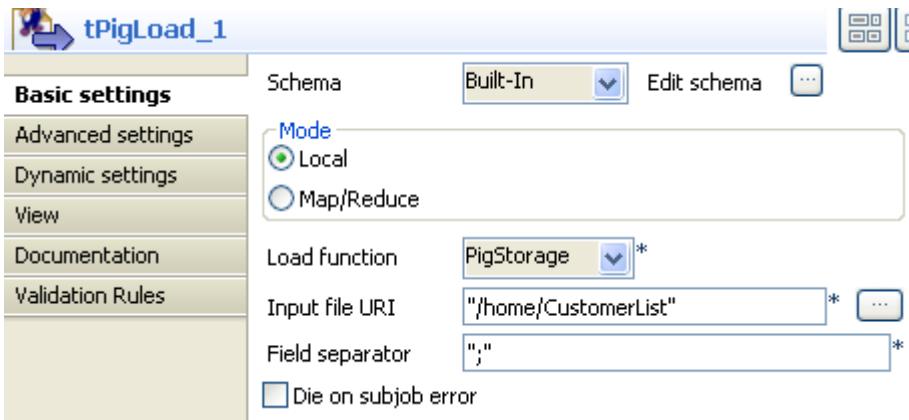


Setting up the Job

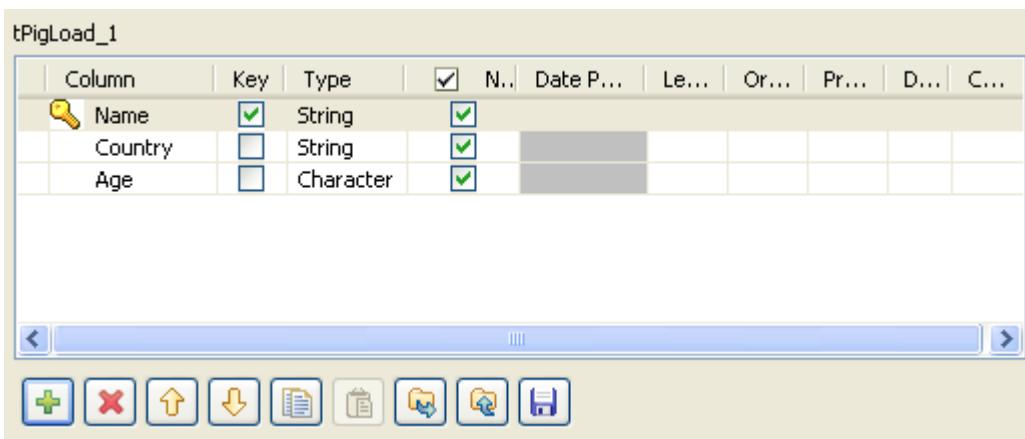
1. Drop the following components from the **Palette** to the design workspace: **tPigSort**, **tPigLoad**, **tPigStoreResult**.
2. Connect **tPigLoad** to **tPigFilterRow** using a **Row > Pig Combine** connection.
3. Connect **tPigFilterRow** to **tPigStoreResult** using a **Row > Pig Combine** connection.

Loading the data

1. Double-click **tPigLoad** to open its **Basic settings** view.



2. Click the [...] button next to **Edit schema** to add columns for **tPigLoad**.



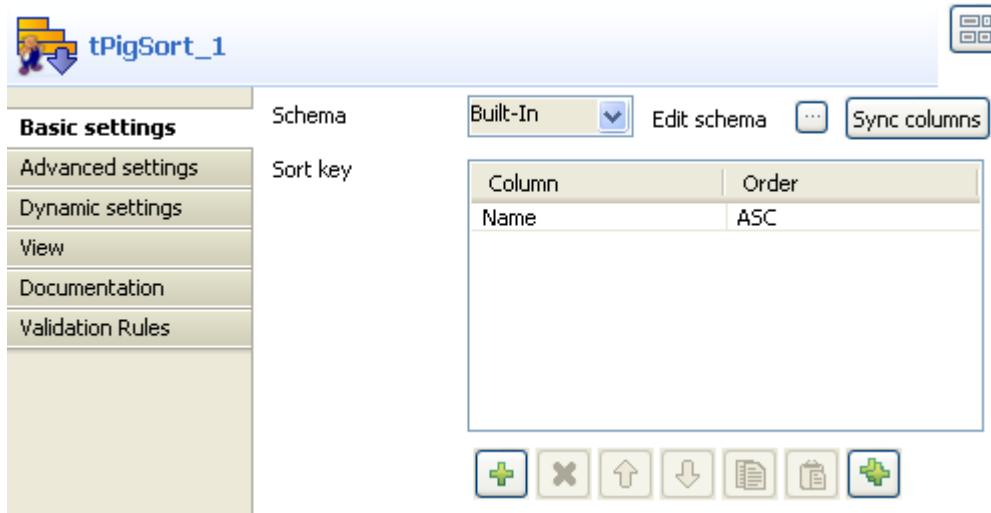
3. Click the [+] button to add *Name*, *Country* and *Age* and click **OK** to save the setting.
4. Select **Local** from the **Mode** area.
5. Fill in the **Input filename** field with the full path to the input file.

In this scenario, the input file is *CustomerList* that contains rows of names, country names and age.

6. Select **PigStorage** from the **Load function** list.
7. Leave rest of the settings as they are.

Setting the sorting condition

1. Double-click **tPigSort** to open its **Basic settings** view.

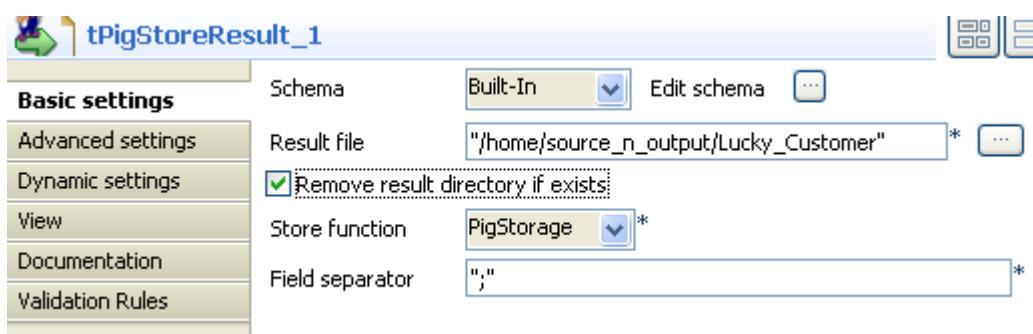


2. Click **Sync columns** to retrieve the schema structure from the preceding component.
3. Click the [+] button beneath the **Sort key** table to add a new sort key. Select **Age** from the **Column** list and select **ASC** from the **Order** list.

This sort key will sort the data in *CustomerList* in ascending order based on *Age*.

Saving the data to a local file

1. Double-click **tPigStoreResult** to open its **Basic settings** view.



2. Click **Sync columns** to retrieve the schema structure from the preceding component.

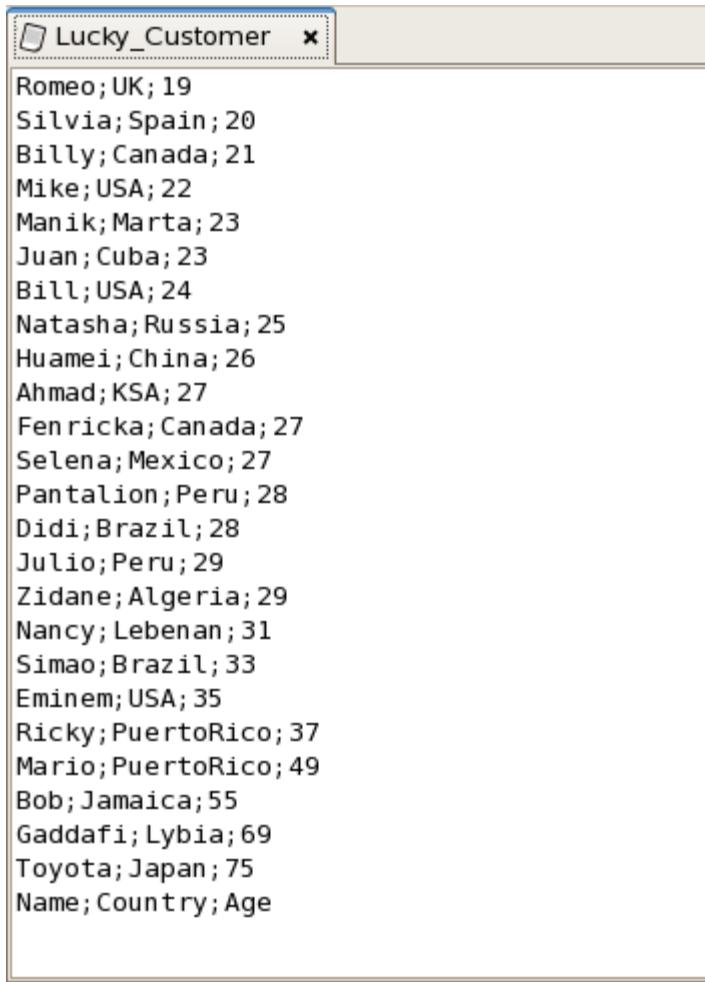
3. Select **Remove result directory if exists**.
4. Fill in the **Result file** field with the full path to the result file.

In this scenario, the result of filter is saved in *Lucky_Customer* file.

5. Select **PigStorage** from the **Store function** list.
6. Leave rest of the settings as they are.

Executing the Job

Save your Job and press **F6** to run it.



The screenshot shows a terminal window with the title bar 'Lucky_Customer'. The window contains a list of 30 entries, each consisting of a name, a country, and an age, separated by semicolons. The entries are sorted in ascending order based on age. The last three entries are headers: 'Name;Country;Age'.

Name	Country	Age
Romeo	UK	19
Silvia	Spain	20
Billy	Canada	21
Mike	USA	22
Manik	Marta	23
Juan	Cuba	23
Bill	USA	24
Natasha	Russia	25
Huamei	China	26
Ahmad	KSA	27
Fenricka	Canada	27
Selena	Mexico	27
Pantalion	Peru	28
Didi	Brazil	28
Julio	Peru	29
Zidane	Algeria	29
Nancy	Lebenan	31
Simao	Brazil	33
Eminem	USA	35
Ricky	PuertoRico	37
Mario	PuertoRico	49
Bob	Jamaica	55
Gaddafi	Lybia	69
Toyota	Japan	75
Name	Country	Age

The *Lucky_Customer* file is generated containing the data in ascending order based on *Age*.

Apache Pig

Configuration Document:

This Document describes setup of a standalone Pig instance that uses the local file system.

Steps to configure Pig

- 1) Download Pig stable released from the site <http://apache.techartifact.com/mirror/pig/>.
- 2) Unpack tar.gz file. preferred location to put at hadoop folder location
- 3) Update pig.properties from pig/conf folder and update PIG_CONF_DIR environment variable (`export PATH=/<my-path-to-pig>/pig-n.n.n/bin:$PATH`)
- 4) Start Hadoop clusters and now check pig installation by following command :
`Pig/bin/ pig -help`

Sqoop Installation and Setup

Configuration Document:

This Document describes setup of a standalone sqoop instance that uses the local file system.
Steps to configure sqoop.

- 1) Download sqoop stable released from the site <http://sqoop.apache.org/>
- 2) Unpack .tar.gz file preferred location is to put at hadoop folder location
- 3) Set the environment variables as :
Set environment variable to **sqoop-env-template.sh**
 - ❖ HADOOP_HOME : Hadoop Home directory
 - ❖ HBASE_HOME : HBase Home directory
 - ❖ HIVE_HOME : Hive Home directory
- 4) Sqoop can be connected to various types of databases For example it can talk to mysql, Oracle, Postgress databases. It uses JDBC to connect. JDBC driver for each of databases is needed by sqoop to connect to them.
- 5) Download jars for JDBC drivers selected database and paste it to sqoop/lib folder.
- 6) Need to check sqoop configured properly check with following command.
 - **\$ sqoop help**



HAND BOOK ON
BIG DATA ANALYSIS TOOLS ALTERNATIVES
By
PRAKASH D DEVARAKONDA

MEMBER OF THE BOARD AND CEO

KARVY ANALYTICS

WWW.KARVY-ANALYTICS.COM

CONTCTUS@KARVY-ANALYTICS.COM

Linkedin: <http://in.linkedin.com/in/prakashdurgad>

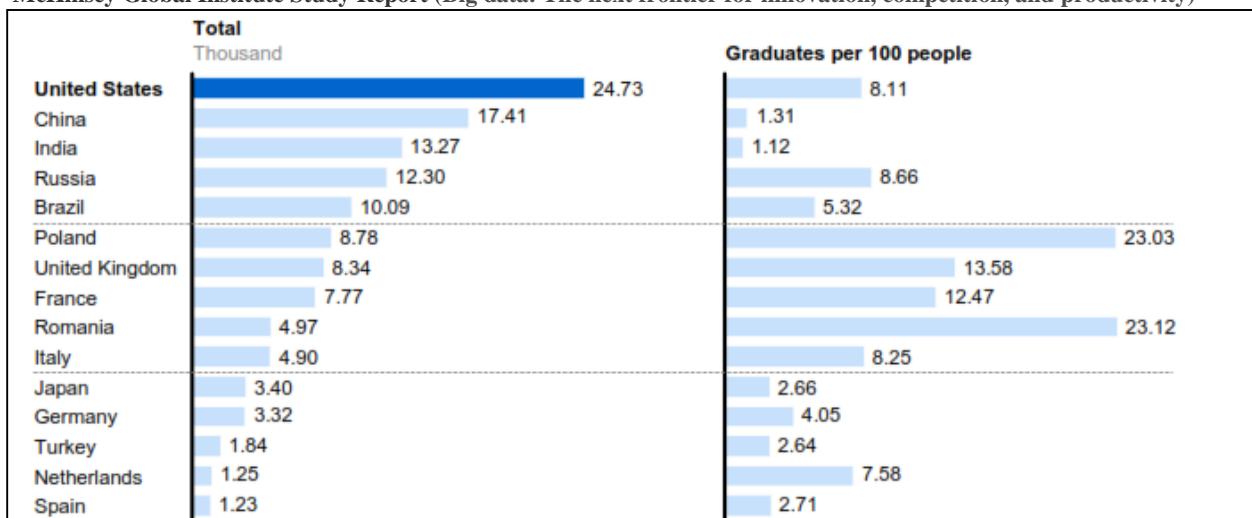
1.0 Executive Summary

INDIA'S HIGH DEMAND FOR BIG DATA WORKERS

The biggest fallout of the big data revolution -- where every type of business gathers and analyzes data -- is a massive human resources shortage. Across the globe, thousands of data analytics jobs are going begging because of a shortage of qualified manpower. A McKinsey Global Institute Study Report (Big data: The next frontier for innovation, competition, and productivity) projects that the US alone will face a shortage of about 190,000 data scientists by 2018 and, further, a shortfall of 1.5 million managers and analysts who can understand and make decisions using big data. As per this report India is producing third largest absolute numbers of BA Professionals after USA and China; however India is producing only 1.12 BA professionals per 100 as compared to USA's 8.11 and that of China's 1.31. The worry is that India's figure of 1.12 is smaller than most of the countries. See Figure 1.6.

FIGURE 1.6: NUMBER OF GRADUATES WITH DEEP ANALYTICAL TRAINING

McKinsey Global Institute Study Report (Big data: The next frontier for innovation, competition, and productivity)



Three key types of talent are required to capture value from big data:

- Deep Analytical Talent -people with technical skills in statistics and machine learning, for example, who are capable of analyzing large volumes of data to derive business insights;
- Data-Savvy Managers and Analysts - who have the skills to be effective consumers of big data insights—i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions; and
- Supporting Technology Personnel - who develop, implement, and maintain the hardware and software tools such as databases and analytic programs needed to make use of big data.

Data analytics as a job discipline became main stream almost a decade ago, and the demand for trained professionals has been growing steadily since. Given India's reputation for the availability of professionals in varied disciplines at reasonable costs, global banks and financial

services firms were the first to migrate their analytics work to India, followed by pharmaceutical and life sciences companies. Global retailers, consumer firms, logistics firms, consultancies, and engineering firms have all begun routing their data analytics work to IT services providers and specialized analytics service providers in India.

The talent deficit is on two fronts, data scientists who can perform analytics and analytics consultants who can understand and use the data. The first, big data engineers and scientists are extremely scarce and in the second category, better quality is needed, and India is going to be short of a million data consultants soon.

Global analytics market

As firms gain access to greater volumes and newer varieties of data, and as they unearth more innovative ways of generating insights for improved customer engagement, implementing analytics is gaining in importance. The global analytics market (software products and outsourced services) is growing at over 12 per cent since 2012. The 2014 market size is estimated at USD 96 billion and is projected to reach USD 121 billion by 2016. Outsourced services around analytics is growing at a faster CAGR of over 14 per cent vis-à-vis analytics software (CAGR ~10 per cent). This growth is being driven by a host of factors – cloud, in-memory computing; mobile devices, social media; emergence of different business units across an organization as consumers of analytics, etc. With analytics being consistently recognized as the top priority for CXOs, firms are also industrializing analytics within the organizational culture and this in turn, is seeing the emergence of the Chief Data Officers' role.

India analytics market

Compared to the global market, the overall India analytics market size is minuscule and currently accounts for only 1 per cent share. The India market (exports and domestic) is growing at double the rate of global market at 24 per cent CAGR. In FY2014, the total market was USD 954 million and is expected to reach nearly USD 2.3 billion by FY2016. The ratio of exports-to-domestic is likely to remain steady at 85:15 during this period. Currently, this segment has over 600 firms offering analytics-related products and services and it employs about 29,000 people. Of this, India is the primary target market for ~50 per cent of these firms. The fact that India's Top 100 IT-BPM (integrated) firms and about 500+ start-up firms are focused on analytics is statement of proof of this technology's increasing relevance.

India is rapidly emerging as the analytics hub for the world. It has the complete range of ecosystem players from GICs, integrated IT-BPM firms, pure-play analytics firms to BPM-KPOs and a vibrant analytics product firms. In terms of geographic density, Bengaluru has the highest number of analytics firms – 29 per cent, followed by Mumbai and Pune – 24 per cent. Apart from this, many Tier II/III cities are also emerging hubs - Trivandrum, Kochi, Mysore, Indore, etc.

Analytics in the India domestic market

There is also a pull factor from the user side – firms in India are beginning to realize the value of implementing analytics. Potential impact can be operational (cost control, process efficiencies), end customers (user insights, targeted marketing) and strategic (driving sales, improved decision making).

Firms in the BFSI, telecom and ecommerce verticals have so far been taking the lead in adopting and applying analytics to a wide range of business areas – portfolio analytics, risk & compliance analytics, customer loyalty, subscriber profiling, churn management, etc. Emerging verticals that are still in the pilot phase of adoption include retail, manufacturing and media & entertainment. One of the key verticals that is showing great promise is the Government – SEBI (fraud detection), NATGRID (anti-terrorism) – and state level initiatives - Maharashtra Sales Tax Department and Hyderabad's intelligent transport system.

SWOT ANALYSIS OF THE BIG DATA ANALYTICS

THE NEED

A SWOT analysis helps in understanding the strengths and weaknesses and helps in identification of open opportunities and the threat that can come along. It provides with a vision to differentiate between marginal and valuable opportunities. It also helps in deciding what to exploit and what to ignore. SWOT analysis gives a taste of what are the threats and their intensity. It facilitates with options to keep an eye on the unlikely to cause damage and beware of increasingly dangerous threats. Finally provides it an opportunity to indentify the **GAPs** that will lead to preparation of a strong and structured Strategic Roadmap for Big Data Analytics. Below is the SWOT analysis of big data analytics in India.

SWOT ANALYSIS – BIG DATA ANALYTICS, INDIA

Strengths

- There is a growing interest in archiving, sensing, behavioral data, and personal data.
- There is a large amount of content and data available – the issue is accessing and making use of it.
- There is a broad and detailed domain know-how as well as process know-how available.
- Many domains have innovative technology and skilled people.
- There are many universities/institutions with high capacity where skills can be developed.
- Avenues where good science/engineering /domain specific education can be obtained.
- Immense growth opportunity in the analytics market: Indian product firms have shown a growth rate of 20-40 per cent in the last few years; several emerging players have witnessed over 100 per cent growth within the first year of launch. (NASSCOM)
- Analytics – a definite market for India: Over 100 Indian analytics focused software product firms have successfully developed and launched products catering to niche business needs, cut across vertical-specific, horizontal process-centric and niche applications and platforms. (NASSCOM)
- Growing start-up base accelerating the growth: Four-fold increase in analytics start-ups in the last four years. (NASSCOM)
- Innovative offerings focusing on end-to-end customer business needs. (NASSCOM)

Weaknesses

- There are no established cooperation networks between content providers in several domains.

- Computer clusters and cloud resources are readily available and accessible to the users/stakeholders such as Researchers in the Institutes and Research Labs.
- There are not many SMEs that are dynamic and flexible and can react quickly to market changes.
- Geospatial and environmental data sets and supporting infrastructure data sets are not readily available.
- There is no existing and strong content/data market in India.
- There is a lack of a solid start-up culture because of risk aversion and intolerance of failure.
- There are few large companies to lead the market, and many small sized companies that need nurturing.
- There is a lack of access to Big Data facilities that make data more easily accessible.
- There is no visibility of ecosystem service offerings.
- It is unclear what data should be preserved, and for how long, in all the different sectors and markets.
- Lack of processable linked data, and of aggregated/combined data.
- Lack of seamless data access and inter-connectivity, and low levels of interoperability: data is often in silos and data sharing is difficult due to an ineffective Data Sharing Policy as well as standards e.g. formats and semantics.
- Migration of data between systems, versions or partners is challenging.
- Access and processing of data sets those are too big to be given to the end user.
- Public data in the country is not available to the extent it should be.
- The quality of data in even in open data portals is often very low.
- The different languages within the country create a barrier (multilingualism) during data processing. Structural data sources often lack precise semantics.
- Poor and inconsistent use or management of metadata.
- There is a lack of specialized education programs for data analysts.
- There are not enough skilled people to participate in capacity building training programs.
- Legislative restrictions on data sharing decrease availability across the country and makes nationally/industry/domain focused initiatives that address these issues more difficult.
- Rules and regulations are fragmented across the country/industry/domain.
- There are high security/sensitivity/confidentiality demands that can be difficult to address.
- There is no well-designed data governance: Data governance is a must-have, and no longer merely a good-to-have. In today's extremely hyper-competitive markets, insightful knowledge means the difference between success and being overwhelmed. But it has to be based on the right data, based on business requirements.
- Data protection Policy: "Ignoring data security, data quality and data access can cost organizations millions of dollars, hurting enterprise agility, efficiency and reputation."

Opportunities

- Being a multi-cultural society, various cultures/practices/strengths/approaches can result in creative thinking if they are mixed.
- The proposed topics by the DST/BDI and best practice examples in other initiatives can lead to synergies.
- Strengthening the Indian market, e.g. by fusing the emerging start-up nucleus.
- Create lots of SMEs for the low hanging fruits of Big Data for which agility is required.
- Investment in the entire innovation chain, beyond basic research.

- Investment support mechanisms for SMEs/Research/ Institutions/Students/Scholars/Entrepreneurs.
- Collaboration within Industry/Academia/DST/Service Providers/Data Generators.
- Improve and encourage innovation & creativity to create cost-effective solutions.
- There is the opportunity to open up completely new and different business areas and services.
- New applications can be created throughout the Big Data ecosystem, ranging over acquisition, data extraction, analysis, visualization and utilization.
- Easier syndication of data and content across industry/domains
- Micropayments for processed data or the results from analytics.
- Wearable sensors and sensor technologies become mainstream generating more data.
- The explosion of device types opens up access to any data from any device for greater and more varied usage.
- Development of APIs for access becoming standardized and available.
- Interoperability tools and standardized APIs to facilitate data exchange.
- Greater visibility and increased use of directory services for data sources.
- Use semantics to align content from various data sources.
- Providing facilities to better navigate and curate data.
- Contextualization and personalization of data.
- The evolution of different sectors and the increased volume of data enable innovative applications to be developed.
- Exploring new research areas.
- Training focused on innovation in DST/BDI.
- Use and exploration of Big Data to be ubiquitous in education and training.
- Address the safe and secure storage of data on the national basis.
- User generated and crowd-sourced content increasingly available that will help variety of recurring problems solved once for all.
- Data-as-a-service can significantly lower the market entry barriers (in particular to new markets).
- Shift from technology push to end-user engagement.
- Create rich and complex data value chains.
- Develop strong and workable policies for data access in the country across private and public data to help build comprehensive capabilities.
- By 2020, information will be used to reinvent, digitalize or eliminate 80% of business processes and products from a decade earlier: As the presence of the Internet of Things (IoT) — such as connected devices, sensors and smart machines — grows, the ability of things to generate new types of real-time information and to actively participate in an industry's value stream will also grow. (GARTNER)
- By 2017, more than 30% of enterprise access to broadly based big data will be via intermediary data broker services, serving context to business decisions: Digital business demands real-time situation-awareness. This includes insights into what goes on both inside and outside the organization. How do weather patterns impact inventory? More so, how do this season's customer preferences as expressed in social media suggest greater or lesser inventory? (GARTNER)
- By 2017, more than 20% of customer-facing analytic deployments will provide product tracking information leveraging the IoT: Fueled by the Nexus of Forces (mobile, social, cloud and information), customers now demand a lot more information from their vendors. The rapid dissemination of the IoT will create a new style of customer-facing analytics — product tracking

- where increasingly less expensive sensors will be embedded into all types of products. (GARTNER)
- Analytics – Opening up a gamut of opportunities for Indian software product firms (NASSCOM)
- Big Data as a service (BDaaS): That is the delivery of Statistical Analysis tools or information by an outside provider that helps organizations understand and use insights gained from large information sets in order to gain a competitive advantage.

Threats

- Many skilled professionals leave the country to work in other regions; adding to the risk of a “Brain Drain”.
- Acute lack of skilled professionals and graduates.
- Non standardization of the ‘contents’, ‘duration’, ‘mode of delivery’ and ‘certification’ of the skilling and or up skilling efforts made by the education/training ecosystem of the society.
- There are no existing ecosystems and portals where reliable data sets are available, however, there is a need to create them.
- Policies are often too connected to the ‘old data’ world.
- Complete analysis of ethical and privacy issues is needed.
- Risk of over-regulation and protectionism in the country as compared to elsewhere in the developed world.
- Policies of data availability; for example companies are not willing to make data available ‘just-in-case’ it may cause a legal action or result in competition.
- Technology & Techniques: To capture value from big data the organizations will have to deploy new technologies e.g. storage, computing and analytical software. The range of technology and technique challenges and priorities set for tackling them will differ depending on the data maturity of the institution.
- Organizational Change and Talent: Organizational leaders may not fully understand and appreciate the value in big data as well as how to unlock this value.
- Shortage of Skills: There are a wide range of skills relevant for businesses wanting to use data analytics, including knowledge of statistical techniques, the ability to program and use software, market-specific knowledge and communication. These skills may not be available in required quantity and quality.
- Business-Education Collaboration: One way to provide the multi-disciplinary skills required for big data analysis is for students to work closely with a company during their studies. Collaboration between a university/institution with analysis expertise and a business with real world data can be beneficial for both parties.
- Trying to rush all data out to everyone all at once: Consider the whole cycle from the acquisition of data to the extraction of information, and consider the hygiene factors along this path. There is a time in which data should be immediately available to decision makers, and there is a time when it can be retired.
- BDaaS requires a coordinated effort: Successful Big Data-as-a-Service implementation would require close collaboration between Enterprise Architects, Data Architects, Database admin, BI and DW SMEs, SOA experts, InfoSec representatives and business strategists.
- Data Sharing Policy: The recommendations made by CODATA on Capacity Building and the Data Sharing Principles in Developing Countries are as given below. Unless these are not implemented the use of Big Data Analytics may not takeoff as desired.

- Data should be open and unrestricted.
- Data should be free to the user.
- Data should be informative and assessed for quality.
- Data sharing should be timely.
- Data should be easy to find and access.
- Data should be interoperable
- Data should be sustainable.
- Data contributors should be given credit.
- Data access should be equitable.
- Data may be restricted, in exceptional cases, if adequately justified.

A large government agency (*<THE AGENCY>*) tasked Karvy Analytics with assessing the tools available to establish and support a platform for performing “Big Data” predictive analytics.

In conducting this assessment, Karvy Analytics employed a three-tier approach.

1. Reviews of literature from industry-leading information technology (IT) research and advisory companies such as Gartner, Inc. and the Forrester Group.
2. Research into open-source software (OSS) extensively used to support analytics. Many OSS products were initially developed by such companies as Google, Amazon, Facebook, and Twitter to support the daily analysis of terabytes of collected data. The software was then made freely available to the IT community for use and further development.
3. Evaluations of product offerings from market leaders in analytics, with a particular emphasis on vendor offerings that run on top of OSS. The use of such products reduces the potential for vendor lock-in, since the source code of the underlying non-proprietary technology is freely accessible to *<THE AGENCY>*.

1.1 Summary of findings and Recommendations

Karvy Analytics recommends that *<THE AGENCY>* build its Big Data environment on a Hadoop Core foundation. Of the several platforms that *<THE AGENCY>* could use for its predictive analytics infrastructure, Hadoop receives the greatest support both from the open-source community and from analytics product and service vendors. This translates into greater flexibility for *<THE AGENCY>* in choosing products and services for its analytics initiatives, and indicates that Hadoop support and improvements will continue to grow.

Karvy Analytics further recommends that *<THE AGENCY>* give primary consideration to the following OSS tools when evaluating candidates for its predictive analytics infrastructure. These tools are among the best available, even when compared to proprietary products, and are supported by several vendors in the open-source industry.

- **CentOS Linux distribution** (see section [4.1.1](#)). CentOS is a community-supported clone of RedHat Linux and is the distribution used by Cloudera for the virtual machine (VM) versions of its CDH (Cloudera's Distribution Including Apache Hadoop).
- **Cloudera CDH Hadoop distribution** (see section [5.1.1](#)). Hadoop is by far the most widely accepted Big Data platform today, and Cloudera is a recognized leader in Hadoop support

and training. The CDH distribution is 100% open source and is the most complete distribution in integrating Hadoop support products. Cloudera also provides major support to many open-source projects, helping to ensure that a healthy infrastructure continues to develop and improve.

- **R with RHadoop plugins analytics package** (see section [4.6.1](#)). R is an extremely powerful open-source analytics package, on a par with commercial products such as SAS and IBM SPSS. It is widely used in colleges and universities, and many recently graduated analysts will be comfortable with its use.
- **Mahout and Python with Hadoop plugins for machine learning** (see section [4.7.1](#) for a discussion of Mahout and section [4.5.2](#) for a discussion of Python). Mahout is the largest machine learning (ML) library for Hadoop, and controls the mindshare in ML research. Python is a very powerful programming language that is less cumbersome than Java for developing complex analysis programs.
- **Pentaho BI Community Edition for reporting, dashboards, and statistical and predictive analytics** (see section [5.2.1](#)). Pentaho BI is competitive with many commercial BI (business intelligence) products (although not in the same league as Microstrategy or IBM). It is, however, 100% open source and should serve *<THE AGENCY>*'s needs for the foreseeable future.

Once the Big Data infrastructure has been established and *<THE AGENCY>* has gained some familiarity with these basic products and their use, the agency will be in a better position to evaluate any additional products that might be needed for its Big Data initiatives.

Karvy Analytics suggests that *<THE AGENCY>* select one of the “pure” OSS vendors (Cloudera, Hortonworks) to aid in the construction and maintenance of the agency’s Big Data environment and analytics foundation. All of the products recommended above are open-source community projects, and there is a great deal of freely available documentation and help on the web. However, contracting with an open-source service provider with expertise in these projects will reduce the amount of time needed by *<THE AGENCY>* to keep track of package updates and compatibility issues.

Finally, because this is a new and rapidly changing field, it is in *<THE AGENCY>*’s best interests to conduct regular surveys of Big Data products and vendors to determine which new products to incorporate into the agency’s Big Data environment.

2.0 Overview

2.1 Assessment Strategy

Karvy Analytics conducted extensive research of the Big Data field in developing this paper. This research included:

- Reviewing Gartner and Forrester documents on Big Data and business analytics;
- Reading published information from major corporations that use Big Data such as Yahoo, Google, and Facebook, as well as documentation on Big Data products; and
- Attending webinars and seminars related to Big Data, and in some cases downloading products for evaluation.

Despite the extensive effort, however, this research and accompanying analysis cannot be characterized as either exhaustive or final. Big Data analytics is a relatively new field, having only come into its own within the last few years. The products reviewed for this effort are all undergoing extensive development, and many new Big Data products appear every month. Moreover, many smaller vendors have been acquired by big vendors within the last two years as the bigger players buy a stake in the Big Data market.

2.2 Big Data Concepts

The term “Big Data” originally referred to large streaming datasets, such as audio and video files, that could not be stored in a database management system (DBMS). The term has since expanded to include files of large numbers of unstructured and semi-structured records that could not be processed by DBMSs, and currently seems to refer to “whatever your shop will be processing three years from now.” Gartner’s definition of Big Data is data exhibiting the following attributes.

- **Volume** – This encompasses large streaming datasets, such as audio and video files, and files of large numbers of unstructured and semi-structured records. “Large” is a relative term, and in the last decade it has gone from meaning gigabytes (10^9) to terabytes (10^{12}), and may very well refer to petabytes (10^{15}) within the next year or so.
- **Velocity** – This includes both the rate at which data is generated and the need to process the data in a timely fashion. Data may be created at a rate of terabytes per second or more, either by huge numbers of users via the Internet or by large numbers of sensors that automatically collect and forward data at subsecond rates. Moreover, because of its ephemeral nature, some data must be analyzed within days, hours, or minutes of its creation to be useful.
- **Variety** – Data collected from multiple sources may be in incompatible formats, use inconsistent semantics, or be collected at different reference points within a dataspace.
- **Complexity** – This attribute covers individual data types, such as free-form text strings, audio and video files, and digital instrument readings.

Big Data has existed in IT for some time, although the term may be relatively new. There has always been a class of data that was either too voluminous to store affordably or too complex for affordable analysis relative to the value of the knowledge gained. However, the cost of both

storage and processing power has dropped dramatically over the last decade, to a point where the processing capability of a multimillion-dollar 90s era datacenter can now be had for a few hundred dollars and comfortably fit in one's pocket.

2.3 Why Open Source Software?

The relational database management systems (RDBMS) available from vendors today are built using a paradigm based in 1970s technology: expensive (and centralized) computers and disk drives; small (24 X 80 character) display systems; slow (<10 pages per minute) printers; and long lead times crafting custom applications. In such an environment, disk usage is reduced by sharing the data between several applications. At first this was accomplished by single-threading batch jobs to manipulate the data and produce reports overnight, but as users grew more sophisticated in their data requirements, the ability to access sets of data online and relatively quickly became paramount.

RDBMS technology evolved to perform several primary functions:

- Ensure that partial transactions do not corrupt the data,
- Provide a consistent view of the data to existing applications while allowing new applications to incorporate new data items,
- Ensure that transactions do not interfere with one another, and
- Ensure that the integrity of the database is maintained in the case of software or hardware failures.

These attributes are collectively referred to as ACID (atomicity, consistency, isolation, and durability). They are implemented in RDBMSs by enforcing a rigid definition of the data structure, controlled by a central “traffic cop” that regulates what transactions can or cannot do at any one time and capturing *before* and *after* images of the changes made to the database.

Not long after the turn of the millennium, several Internet giants began to find it increasingly difficult to process in a timely manner the huge amounts of data they were collecting. These corporations began to develop in-house alternatives to RDBMS databases, collectively referred to as “NoSQL,” which may be taken to stand for “No SQL needed for access” or “Not Only SQL.” Several of these products have found their way into the OSS community, either as donations by the companies that created them or via open-source projects based upon the originals. Examples of these are provided in [Table 1](#).

- *Table 1: NoSQL Products/Solutions*

Company	Product	OSS Equivalent
Amazon	DynamoDB	M/DB
Facebook	Cassandra	Apache Cassandra
Google	BigTable	Apache HBase
Yahoo	Hadoop	Apache Hadoop

This is by no means an exhaustive list of OSS NoSQL products. Other notables include CouchDB, Membase, and MongoDB.

While the companies that built the Big Data applications still actively contribute to their enhancement, there is now a cottage industry of others who freely provide additional enhancements or sell value-added products that work with Big Data OSS.

In addition, some proprietary DBMSs have recently been offered in the Big Data arena, such as Greenplum (recently acquired by EMC), Infobright, and Aster data.

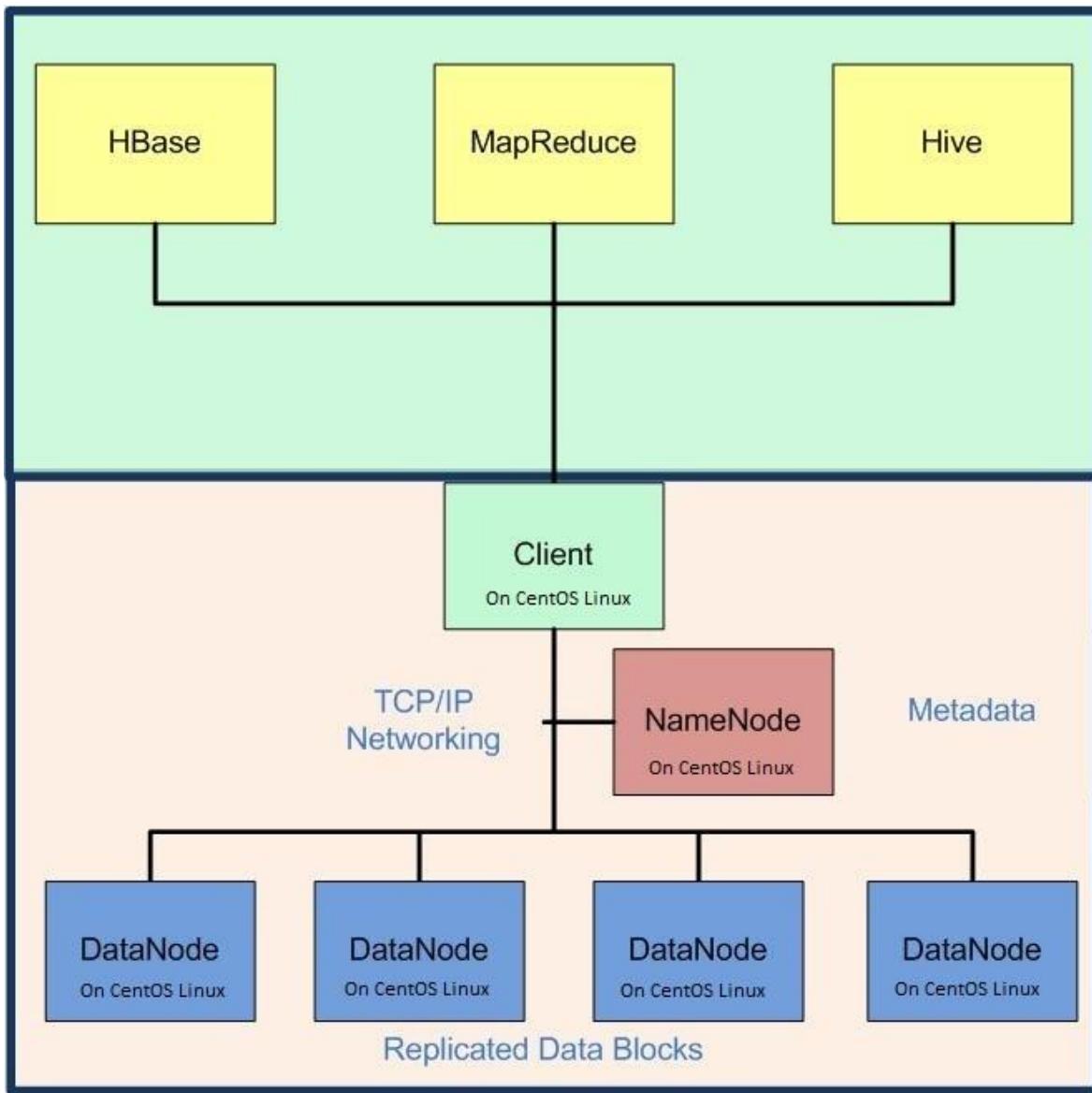
Of all of the NoSQL products available today, the one with the greatest mindshare is Apache Hadoop. It is the most mature of the NoSQL products and is supported by the OSS community, by major vendors such as IBM and EMC, and by smaller vendors such as Cloudera, MapR, and Pentaho. By building its Big Data environment on Apache Hadoop, <THE AGENCY> will gain access to the wealth of tools that support Hadoop. Furthermore, the potential for vendor lock-in will be significantly reduced, since all of these tools rest on a common, freely available foundation.

3.0 General Assessment Findings

Based on its research for this paper, Karvy Analytics recommends that *<THE AGENCY>* build an OSS Hadoop-based architecture to support its Big Data and predictive analytics initiatives. Karvy Analytics feels that an architecture built upon OSS will provide *<THE AGENCY>* with the greatest flexibility for a current implementation while preventing the agency from being locked out of future developments by dependence on proprietary commercial products. The use of OSS will provide *<THE AGENCY>* with some of the best software available for Big Data processing as well as access to myriad applications that have been developed by the open-source community. Since the source for these applications is freely available for download and modification, *<THE AGENCY>* can build upon work done by others to develop its own custom applications much more rapidly and with greater sophistication than if it had to build these applications from scratch with a limited pool of programming experience. Finally, the open-source products suggested in this paper currently hold the greatest mindshare of the Big Data market, ensuring both a ready supply of third-party vendor products and support, and a larger stable of young programming talent with which to fill its ranks.

The Big Data environment recommended by Karvy Analytics is shown in [Figure 1](#):

- *Figure 1: Hadoop Core*



The lower portion of this figure represents the Hadoop Data File System (HDFS). Several proprietary commercial packages are available to replace or enhance HDFS in order to provide high availability, higher performance, and automatic recovery, and to eliminate the NameNode as a single point of failure. However, even though they are proprietary, these products all replicate HDFS APIs and so are “plug compatible” with the rest of the Hadoop universe.

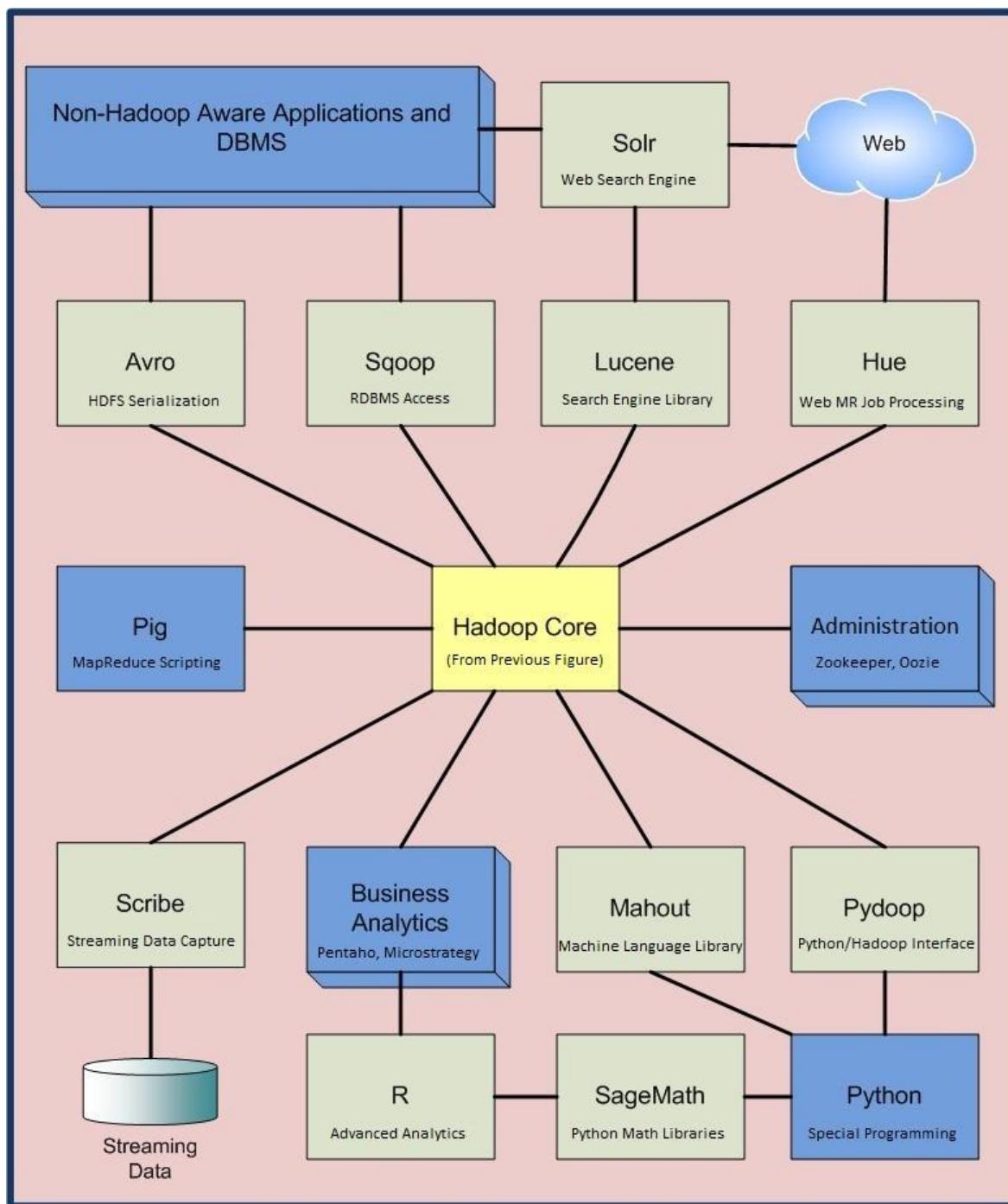
On top of HDFS sit the primary Hadoop data components. Since most applications do not understand HDFS processing, they will make use of these components for data storage, processing, and retrieval.

- **MapReduce**, the component most associated with the word “Hadoop,” provides parallel processing features.

- **HBase** is a column-oriented NoSQL DBMS that takes advantage of HDFS for storing columns.
- **Hive** is a data warehouse application that can access HDFS files directly, as well as HBase tables.

The core components form the foundation of a Hadoop “sandbox,” an integrated platform on which to perform Big Data processing. On top of the core are applications for performing advanced analytics, and APIs to connect Hadoop to other applications, as shown in [Figure 2](#):

- *Figure 2: Hadoop Support Environment*



The products that form the Hadoop environment shown in Figure 2 are described below.

- **Avro** provides APIs to support programming languages such as C, C++, C#, Java, PHP, Python, and so on (see section [4.5.3](#) for further discussion).

- **Lucene** is a powerful search library that can access Hadoop. **Solr** is a search server built on Lucene that can be called from web applications or programs written in non-Hadoop-aware languages (see sections [4.8.4](#) and [4.8.5](#) for further discussion).
- **Hue** is a web server portal that can directly access Hadoop. It is used to create and submit Hadoop jobs, monitor their execution, and review the results (see section [4.8.3](#) for further discussion).
- **Sqoop** provides interoperability with RDBMS and NoSQL databases outside of Hadoop (see section [4.4.1](#) for further discussion).
- **Scribe** provides access to streaming data from Hadoop to support complex event processing (CEP) and predictive analytics processing (see section [4.4.2](#) for further discussion).
- **Mahout** is an ML library that can take advantage of Hadoop. It too supports CEP and predictive analytics processing (see section [4.7.1](#) for further discussion).

The following applications can access Hadoop directly and are included in the Hadoop sandbox shown in Figure 2.

- **Pig** is a scripting platform for writing and executing MapReduce jobs (see section [4.5.1](#) for further discussion).
- **Administration applications**, such as ZooKeeper, Oozie, and Starfish, provide job scheduling, process control, and cluster configuration (see sections [Error! Reference source not found.](#), [4.8.2](#), and [4.8.6](#) for further discussion of these products).
- **R, with Hadoop extensions**, provides complex analytics tools for knowledgeable data analysts (see section [4.6.1](#) for further discussion).
- **Business analytics applications** available from several open-source and commercial vendors provide tools for performing analysis, visualization, and reporting of data processed in Hadoop (see section [5.2](#) for further discussion of open-source products and section [5.3](#) for proprietary products).

Karvy Analytics recommends that *<THE AGENCY>* use **Python**, a powerful and flexible programming language, to create complex or custom applications. Python is especially well-suited for creating analytics programs because of its mathematics support. It is also ideal for *<THE AGENCY>*'s Big Data analysis projects because programs can be written much faster in Python than in Java or C-based languages.

Karvy Analytics recommends that *<THE AGENCY>* use Python release 2.7 for the agency's Big Data test environment. While a newer version of Python is available (release 3.x), the newer release is not backward compatible, and many of the support packages needed by Python are not yet available for the newer release. This situation will need to be reassessed when *<THE AGENCY>* is ready to establish its production environment. Many Python support applications

will be available for the newer release sometime in 2012 or 2013. If the support applications needed by <*THE AGENCY*> have all been ported to Python 3.x by the time the production Big Data environment is to be built, then <*THE AGENCY*> should install the latest release of Python at that time.

Karvy Analytics recommends that <*THE AGENCY*> utilize the following Python support packages in its Big Data environment.

- **Pydoop**, an API designed to provide Python-like access to Hadoop from Python programs.
- **RPy2**, to support calling R functions from Python programs.
- **JPipe**, to support calling Java from Python.
- **Sage Math**, an all-encompassing mathematics and statistics library to support complex analysis in Python programs.

See section [4.5.2](#) for more details on Python and its support packages.

4.0 Open-Source Software

This section provides more detailed information on the OSS available for Big Data and predictive analytics. It focuses primarily on software that runs on top of Hadoop, since it is the way that the market is moving.

4.1 Operating System Platforms

All of the Big Data products available today, as well as much of the infrastructure to support Big Data, are designed to run on the Linux operating system. Karvy Analytics reviewed several GNU General Public License Linux distributions to determine their suitability as a platform on which to base the Big Data test environment at <THE AGENCY>. This section presents some of the highlights of these distributions.

4.1.1 CentOS

CentOS is a Linux community-supported distribution that is maintained to be an exact clone of Red Hat Enterprise Linux (RHEL). RHEL is the foremost Linux distribution in production use today, and there is an extremely large amount of support for RHEL. For example, nearly every application available for Linux can be found in an RPM (Red Hat Package Manager) format for installation using RPM or Yum, the RHEL package administration applications. RHEL is certified to run on IBM System z hardware.

Since CentOS is a clone of RHEL, it can make use of RPM packages built for RHEL as well, and will have the same dependencies and installation procedures. In addition, a large pool of individual and corporate consultancies provides support for RHEL – and by way of extension, CentOS. A final benefit of using the CentOS distribution is that Cloudera releases a CDH VM image that is built on CentOS. This guarantees that the CDH package will install cleanly on CentOS, since Cloudera has already created a version of its Hadoop environment on this distribution.

Karvy Analytics recommends that <THE AGENCY> use the CentOS distribution to build its Big Data test environment and either CentOS or RHEL to build its production Big Data environment.

4.1.2 Arch Linux

The Arch Linux distribution is designed to be a lightweight Linux environment. The official package set contains the bare minimum number of applications to get Linux up and running. Anything beyond the basics must be downloaded separately and installed.

Arch Linux is an ideal distribution to use when the exact needs for the server are known and the administrator wishes to restrict the products installed to only those needed. An example of this might be a production data server that will only be accessed remotely. In this scenario there is no need to install developer tools or productivity tools, since no one will be performing the type of work that needs these tools on the server.

The disadvantage of Arch Linux is that the administrator must have a clear understanding of exactly what applications will be needed, including the network of packages needed to support these packages (that is, the dependencies). This can lead to lost time in the test environment while these needs are determined. It therefore would be more expedient to select a Linux distribution that provides a fuller set of applications “out of the box.”

Karvy Analytics recommends that *<THE AGENCY>* not use Arch Linux for its Big Data environment at this time. The product may, however, be of interest to *<THE AGENCY>* sometime in the future when the exact application needs of the production Hadoop cluster are known.

4.1.3Fedora

The Fedora distribution of Linux is a community-supported distribution that is based on Red Hat Linux. It is designed for use as a desktop alternative to Microsoft Windows or Apple OS X. As such, it includes a wealth of applications that are of interest to home and small-to-medium business uses as well as to developers. Fedora is supported by a large community of developers, including many who also work for Red Hat on RHEL.

While Fedora provides many of the benefits of CentOS because of its common heritage with Red Hat, it is less desirable for use in a server environment because it contains many applications that are not of interest on a server. It also does not attempt to maintain compatibility with RHEL, so there is not a one-to-one correspondence to the work needed to support the Fedora environment compared to the RHEL environment. This makes the lessons learned on the Fedora platform less applicable to production, given that the production environment will most likely be RHEL or similar to RHEL.

Karvy Analytics recommends that *<THE AGENCY>* not use Fedora for its Big Data test environment. However, *<THE AGENCY>* may be interested in installing Fedora on its developers' workstations. Fedora provides most of the features found on a Microsoft Windows desktop while providing easier integration with the Linux Big Data environment.

4.1.4OpenSUSE

OpenSUSE is a community-supported Linux distribution that is based on Novell's SUSE Linux Enterprise distribution. Like Fedora, the OpenSUSE distribution is designed to be a desktop replacement for Microsoft Windows or Apple OS X. It is quite possibly the most complete assemblage of Linux applications available on the market, and is a particularly good distribution for use on development workstations. OpenSUSE also incorporates many performance enhancements; in a recent performance comparison by Phoronix, OpenSUSE performed as well as or better than its competitors, including RHEL 6.0.¹

The disadvantage of OpenSUSE is that it is not based on RHEL but rather on Novell SLES. Novell has fallen into disfavor with the open-source community in recent years, and consequently OpenSUSE does not enjoy the level of support provided for distributions based on Red Hat.

Karvy Analytics recommends that *<THE AGENCY>* not use OpenSUSE for its Big Data platform.

4.1.5Ubuntu

Ubuntu is a community-supported Linux distribution that is based on Debian. It is specifically designed to be a desktop replacement for Microsoft Windows or Apple OS X, and is often chosen for its ease of use by users familiar with these other desktop environments. While a server version is available, it is best suited for home and office use. It lacks correspondence to any production-quality server distribution.

¹ [Red Hat Enterprise Linux 6.0 Benchmarks](#), Michael Larabel, November 29, 2010

Karvy Analytics recommends that *<THE AGENCY>* not use Ubuntu for its Big Data platform.

4.2 Big Data File Systems

4.2.1 Hadoop

Hadoop is a framework that allows for the distributed processing of large datasets across clusters of computers using a simple programming model. Hadoop Core comprises HDFS (Hadoop Distributed File System), a distributed, scalable, and portable file system written in [Java](#) for the Hadoop framework; **MapReduce**, a framework for performing analysis of HDFS files in parallel; and utilities to support other Hadoop subprojects, such as **Hive** and **HBase**.

Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Data segments are propagated to each available machine, and work for each segment is processed on the local computer (the “map” phase). Results are then consolidated from each of the participants’ mapped files (the “reduce” phase) and returned to the caller.

Rather than rely on hardware to deliver high availability, the library itself is designed to detect and handle failures at the application layer. This delivers a highly available service on top of a cluster of computers, each of which may be prone to failure. The HDFS file system creates several copies of each segment and places them on different servers. (It is “rack aware” and will attempt to distribute copies on different racks if possible.) If any map task does not return within an expected period of time, a new map task is initiated against one of the available copies.

Of all of the Big Data products available on the market today, Hadoop has by far the largest support base, both from the OSS community and from commercial vendors.

Karvy Analytics recommends that *<THE AGENCY>* build its Big Data platform on Hadoop.

4.2.2 Cassandra

Cassandra is a Big Data system originally built by Facebook, which open-sourced it in 2008. Cassandra is designed to be distributed, highly scalable, and “eventually consistent” (i.e., changes are made immediately to a local copy and are then propagated to remote copies, so that eventually all changes will appear in all copies of the data). It merges design ideas from Amazon’s Dynamo system for cloud computing and Google’s BigTable NoSQL DBMS, which is a column-oriented datastore.

The Dynamo features built into Cassandra support distribution of the data and computing (similar to Hadoop’s HDFS). However, unlike Hadoop, Cassandra does not implement a special “name server” and does not have any component that creates a single point of failure.

The column-oriented datastore implemented in Cassandra can be thought of as a relational table turned “sideways”: instead of storing all of the columns that make up a row, together with indexes to provide access to the rows by column values, each column is stored in its own table with vectors that are used to reassemble a row from its constituent column values. This structure is better suited to data analysis, since all of the values for a given attribute (column) can be assembled quickly.

Cassandra is used most notably by Netflix and Twitter. It does not have the volume of third-party support that Hadoop has. This may be due to the relative newness of its availability (July 2009 versus September 2007 for Hadoop).

4.2.3 CouchDB

Apache CouchDB is a document-oriented database; that is, it is designed to store entire documents as single entities, rather than decompose them into some sort of structure as is common in object-relational databases. It supports distribution of storage and work across a farm of commodity servers and like Cassandra is designed to be “eventually consistent.” It supports MapReduce processing and is designed for easy integration with the web. It currently has a relatively small user list and support community. CouchBase, a major supporter of CouchDB, has recently begun work on creating a hybrid incorporating CouchDB with Membase, an in-memory key/value database system. The new hybrid, Couchbase, will be designed to support both key/value and document-oriented database designs and make extensive use of in-memory processing to provide faster performance.

4.2.4 MongoDB

MongoDB is a document-based database management system created and maintained by 10gen and released as OSS under a creative commons license. It supports distribution and replication across servers using a process called autosharding, which is similar in nature to HDFS. MongoDB supports MapReduce processing. It also provides support for a wide array of programming languages including C, C++, Java, JavaScript, PHP, and Python, among others.

MongoDB has a small support community. Notable users include Sourceforge, Foursquare, and Craigslist.

4.2.5 Disco

Disco is an open-source, large-scale data analysis platform originally created by Nokia for use in processing Big Data. It includes MapReduce processing and provides extensive support for Python programming. Its architecture is similar to Hadoop’s: a master server coordinates the distribution of data and processing over one or more servants. The data is processed locally, and data can be transferred between servants over HTTP. Disco’s main claim to fame is that it is written in Erlang, a functional language created by Ericsson for programming fault-tolerant networks. Disco’s only user of note is Nokia, and no third-party vendor support appears to be available.

4.2.6 Spark

Spark is an open-source cluster computing system developed at UC Berkeley. It is designed to load as much data as possible into memory to reduce disk access, running much faster than Hadoop MapReduce. Spark is written in the functional object-oriented language Scala, which was developed at École Polytechnique Fédérale de Lausanne (EPFL). It does not appear to have any large users or any third-party support.

4.3 Hadoop-based Databases

HDFS is designed primarily to support large MapReduce jobs. This design works very well for very large datasets that will be read sequentially. On the other hand, basic HDFS is not well-suited for small datasets, data that requires random access, or data that needs to support update-in-place. For these types of data, a DBMS that exploits HDFS capabilities is desired.

In the traditional database environment, DBMS offerings are dominated by row-oriented RDBMSs. These are designed to provide support for structured data and are well-suited to providing high availability, fault tolerance, and supporting data used by high-volume/low-content

online transaction processing (OLTP). RDBMSs are designed to handle fairly large amounts of data (tables in the gigabyte range), but scale in a non-linear fashion once Big Data volumes are introduced.

In the Big Data area, DBMSs have evolved that are better designed to support the processing of extremely large amounts of data in reasonable amounts of time. These DBMSs provide both flexibility in data structure definition to reduce the amount of time needed to design and maintain the databases, and structures that support high data throughput. Their attributes are well-designed to support ad hoc analysis of extremely large datasets, but not OLTP applications.

The reliability of data stored in a database depends on the DBMS's ACID (atomicity, consistency, isolation, durability) attributes.

- **Atomicity:** For a given unit of work, either all of the changes made to the database are made permanent or none of the changes are made permanent.
- **Consistency:** Any update made to the data in the database must comply with any and all rules defined for the data.
- **Isolation:** One unit of work will not interfere with another unit of work (this usually means that only one transaction can update a unit of data at any one time).
- **Durability:** Once a unit of work's changes have been committed to the database, those changes will be maintained regardless of later hardware or software failure.

While ACID attributes are strongly implemented in RDBMSs, they may be weakly implemented or not implemented at all in NoSQL databases. These are deliberate trade-offs made to ensure that the large amounts of data housed in the database can be processed in a reasonable amount of time without regard to scaling. Rather than adhering to ACID, NoSQL DBMSs (NDBMSs) tend to exhibit BASE (basic, adaptable, speedy, extendible) attributes.²

- **Basic:** A minimum of “bells and whistles” make maintenance and administration simpler. Unlike an RDBMS, an NDBMS requires very few DBAs or other support staff;
- **Adaptable:** No rigid, predefined structure is imposed on the data. In many cases, all data is treated by the NDBMS as strings. NDBMSs that implement row and column architectures allow the addition of new columns as well as new rows programmatically without the need to restructure existing data.
- **Speedy:** Whenever trade-offs are needed between data integrity and performance, performance is chosen.
- **Extendible:** NDBMSs are architected to support horizontal extension across commodity hardware without degradation in performance.

² In much of the literature on NoSQL databases, the acronym BASE is defined as “**B**asically **A**vailable, **S**oft state, **E**ventual consistency.” However, Karvy Analytics feels that the definition used in this paper is a more accurate description of NoSQL database attributes.

Several NoSQL databases are available on the market, both open source and proprietary. Listed below are the most popular in use with Hadoop as of the writing of this document.

4.3.1 HBase

HBase is an open-source, distributed, versioned, column-oriented DBMS modeled after Google's BigTable database. (BigTable is used by Google to support its large (multi-petabyte) data projects such as Google Analytics and Google Earth.) HBase provides BigTable-like capabilities on top of Hadoop and HDFS.

Conceptually, an HBase table can be pictured as an infinite Excel spreadsheet, where the row and column labels are arbitrary strings, and the cells contain one or more versions of the data (labeled by default by timestamp). A cell is uniquely addressed by these three dimensions (row name, column name, version), and the contents of the cell are treated by HBase as though they were binary strings. If a version is not explicitly specified, the most recent version in the cell (row, column) is returned. Column values are classified as belonging to column "families," and the name of the column family is the only structure that must be predefined in the table. These column families are important to HBase security and administration: access control, disk, and memory accounting are performed at the column-family level. They also come into play in the physical implementation of the HBase table, since storage and tuning are performed based on column family.

HBase supports both random and sequential reads and random writes and updates. HBase supports the ACID attributes as well as the BASE attributes listed above.

Data residing in HBase tables can be both read from and written to directly from MapReduce jobs. HBase tables can also be accessed from non-Hadoop-aware programs via APIs in Java, a REST-full web service, or a Thrift gateway.

Karvy Analytics recommends that <THE AGENCY> include HBase in its Big Data platform.

4.3.2 Hive

Apache Hive is a data warehouse platform that is built on top of Hadoop. Data definitions for structures to be processed in Hive are defined using SQL-like syntax and stored in a metastore (a catalog database, normally implemented in MySQL). Hive can process Hadoop or HBase files and can process its own data directly using HDFS. Hive can also access any data that is available via ODBC, JDBC, or Thrift.

Hive does not enforce any particular data structure on files it creates: the user defines the data format using SERDE (serialization/deserialization) format descriptions. Hive uses these data format descriptions to store and retrieve the data. SERDE provides easy-to-use import and export functions for Hive. Several SERDES are defined for common data types, and additional ones can be created (as Java plugins) for unique situations. Once a SERDE has been assigned to a table in the Hive metastore, it is invoked automatically whenever accessing the data structure.

In addition to SERDE, customized user-defined functions (UDFs) can be created. These are written as Java plugins, and are called like any other function defined in Hive.

Hive uses a modified version of SQL, called HiveQL, for data access and manipulation. It is interesting to note that Microstrategy uses HiveQL to interact with Hadoop data and MapReduce. Additional APIs are provided for programmatic access to Hive via JDBC, ODBC, Python, PHP, and Java and C++ via Thrift.

Karvy Analytics recommends that *<THE AGENCY>* include Hive in its Big Data platform.

4.3.3Giraph

Apache Giraph is a graph (network) database framework that runs on Hadoop. This project is still in the incubator stage, but is worth keeping an eye on for future reference. The project announced its very first incubator release (0.1) on February 6, 2012.

4.4 Data Capture/Input Tools

Hadoop structures do not exist in a data vacuum. Rather, large numbers of operational datastores, historical datastores, and possibly electronic data warehouses will have been in use for some time. In the case of *<THE AGENCY>*, data is stored in MADAM, IDMS, DB2 and Oracle RDBMSs, plus QSAM and VSAM files on IBM mainframes. Data will often need to be exported from these data sources into Hadoop for analysis. Conversely, the results from the Hadoop analysis may need to be populated back into structures in these DBMSs for use in generating management reports or additional analysis by tools that do not interact with Hadoop.

In addition to the data already collected in legacy datastores, vast amounts of data are generated by both hardware and software instrumentation. This data, generically referred to as “log data,” is characterized by large numbers of small data records from a multitude of sources that are generated in response to triggering events. Hadoop is not well-designed for capturing this data directly; it was designed with the batch processing of large datasets in mind. To be able to use log data, the log records need to be consolidated into larger “chunks” for Hadoop processing.

The tools described in this section are designed to import data from these other sources into the Big Data environment. In many cases, little or no customized coding is required, especially when dealing with RDBMS tables. In those cases where customization is required, it is generally in the form of a Java plugin that describes the structure of the data and the methods for extracting and updating the data.

4.4.1Sqoop

The Sqoop application is designed to import and export data from structured datastores such as relational databases, enterprise data warehouses, and NoSQL systems. Data from the source system is extracted and populated to HDFS files or to Hive or HBase tables. The syntax is easy to use; Sqoop reads the database metadata to extract the source data and creates a map-only MapReduce job to load the data into Hadoop.

Similarly, Sqoop can be used to extract data from Hadoop and populate it into other DBMSs. Sqoop comes prepackaged with special connectors for MySQL, DB2, Oracle, and MS SQL Server, and a generic connector that can work with any JDBC-accessible DBMS. Customized connectors can be created to support NoSQL DBMSs as well as specialized data sources.

Sqoop was created by Cloudera and donated to Apache. Cloudera includes Sqoop in its Hadoop distributions. The name is a mash-up of “SQL” and “scoop.”

Karvy Analytics recommends that *<THE AGENCY>* include Sqoop in its Big Data platform.

4.4.2Scribe

Scribe is a server for aggregating streaming log data. It is designed to scale to a very large number of nodes and to be tolerant of network and node failures. A Scribe server instance runs on each system node where messages are to be collected. These collection nodes aggregate the locally

received messages and send the aggregated messages to a central Scribe server (or servers). If the central Scribe server isn't available, the local Scribe server writes the messages to a file on local disk and sends them when the central server recovers. The central Scribe server(s) can write the messages to the files that are their final destination, such as an HBase database or directly to HDFS files.

Messages processed by Scribe are prefixed with a high-level description of the intended destination server, which is provided from a configuration file on the source server. Datastores can be relocated by changing the local Scribe configuration, without the need to modify client code. Configuration entries can be changed at runtime without stopping the server. This permits dynamic configuration of the Scribe network.

Scribe was developed and is maintained by Facebook. The software was made available to the general public under the Apache License, version 2.0. It is a mature product and is used by Facebook to process billions of messages daily in near real time.

Karvy Analytics recommends that *<THE AGENCY>* use Scribe to capture messages for analysis by the agency's Big Data framework.

4.4.3Chukwa

Chukwa is a Hadoop-based platform for capturing and analyzing log data. While data is collected in real time, Chukwa is designed to process large amounts of data in batches, not real time. Chukwa comprises the following components.

- **Agents:** These are “wrapper” programs that capture log data on local machines and pass the data to collectors in the expected format.
- **Collectors:** Collectors receive data from the agents and write the data to what is called a sink file. Sink files are periodically made available to MapReduce jobs for archiving.
- **MapReduce Jobs:** These jobs parse (“demux”) and archive the data in the sink files. The processed data is then loaded into a MySQL database.
- **HICC (Hadoop Infrastructure Care Center):** This is a web-portal-style interface for displaying data from the MySQL database.

As this is a new application (it is still in Apache incubator status), Karvy Analytics recommends that *<THE AGENCY>* not include Chukwa in its Big Data platform at this time.

4.4.4Flume

Flume is a Hadoop-based platform for the capture, movement, and aggregation of event and log-structure data. Flume's architecture comprises the following components.

- **Agent Nodes:** These are applications that retrieve log data on a local machine and pass it on to collector nodes.
- **Collector Nodes:** These applications receive data from agents, aggregate the data, and then pass the aggregated data to Hadoop.

- **Master Nodes:** These applications monitor the health of the Flume architecture, dynamically reconfiguring to support additional nodes or reroute data through the system in the event of a node loss.

Flume is similar to Chukwa in its structure and capabilities, but unlike Chukwa, Flume's architecture supports real-time streaming analysis. Flume data can be written directly into a DBMS (for example, HBase) and does not require MapReduce preprocessing like Chukwa. The data streams in Flume do not need to be "serialized" (that is, written to a disk file), but can be passed from node-to-node via queues or pipes. (The product is aptly named, as a flume is an artificial channel or diverted stream that carries logs down a mountainside to saw mills for processing.)

It is designed to be fault tolerant, and it supports dynamic reconfiguration to provide ongoing processing in the event of the loss of one or more nodes. If there is a slowdown or loss of an upstream node, the local node will store the streaming data to a file until either the upstream node is available or the local node is notified by the master node to reroute it.

Flume was created by Cloudera and donated to Apache. Cloudera includes Flume in its Hadoop distributions.

As this is a new application (it is still in Apache incubator status), Karvy Analytics recommends that <THE AGENCY> not include Flume in its Big Data platform at this time. However, Karvy Analytics recommends that <THE AGENCY> consider the use of Flume once it has matured.

4.5 Language and APIs

Hadoop MapReduce is written in Java and provides Java APIs. This section presents scripting language platforms and APIs that can be used to provide seamless access to Hadoop from other languages that are often used for machine learning and analytics, such as Python, Perl, Ruby, and Scala (among others).

4.5.1 Pig

Apache Pig is a compiler that translates scripts written in "Pig Latin" into Hadoop MapReduce programs. The language is designed to be easy to understand and can be extended by the creation of UDFs. These scripts are then run through a compiler that translates the script into an optimized MapReduce job. A single Pig script may generate several MapReduce tasks that all run in parallel, but this is of no concern to the script writer.

Pig can be executed from Grunt (a command-line shell for Pig), from Eclipse using the Piggen plugin, and from Java programs via APIs. In addition, plugins are available for Emacs, Vim, and Mac TextMate, and wrappers are available to embed Pig in Python (PigPy) and Ruby (Piglet).

Pig has an active user community, and user-contributed UDFs are available from Piggybank.

4.5.2 Python Language and Support

Python is a flexible and powerful general-purpose programming language. It is designed to be easy to understand and it supports many different programming paradigms, such as object-oriented, functional, and aspect-oriented programming. Python libraries provide extensive mathematics and statistics capabilities. It is also well-suited for use in artificial intelligence applications as well as machine learning (for example, Python was used to create the Natural

Language Toolkit). Its flexibility and power make it an ideal language for developing specialized programs needed to support Big Data and predictive analytics.

Karvy Analytics recommends that *<THE AGENCY>* use Python for developing specialized programs in the agency's Big Data environment.

In addition to the standard Python distribution, Karvy Analytics also recommends that *<THE AGENCY>* use the following Python support packages. These packages extend Python's utility in the proposed Big Data environment.

- **Pydoop** is a Python-based abstraction layer designed to simplify the writing and running of Hadoop programs and to create Python code for the map and reduce phases. Embedded within the Python program is a call to execute the MapReduce methods that indicates when within the program the MapReduce job should execute. The process is similar to the COBOL SORT function, defining input and output processing to wrap around an external utility. This permits the programmer to execute code in Hadoop as well as on the client so that (for example) the Python program can invoke MapReduce to do the “heavy lifting” data processing and then process the result set locally, such as storing the data into an RDBMS table. When the program is run, Pydoop ships the necessary code and support libraries to nodes in the Hadoop cluster for use by MapReduce.

Pydoop makes use of the Hadoop pipes implementation. This makes Pydoop much faster than many of its competitors by reducing the amount of data that needs to be written to disk. Pydoop is written in C++ and compares favorably in processing time to similar products written in Java. It was created and is maintained by Simone Leo and Gianluigi Zanetti and is available to the general public under Apache License version 2.0. Similar products include **Dumbo** and **Happy**. However, Pydoop runs much faster than Dumbo and provides greater flexibility than Happy. Karvy Analytics therefore recommends that *<THE AGENCY>* use Pydoop for its Python/Hadoop interface.

- **RPy2** is a Python interface to R. It allows Python programs to call R functions to perform complex computations. It is available to the general public under GNU General Public License (GPL) and Mozilla Public License (MPL).
- **Sage Math** composes a multitude of open-source statistics and mathematics packages, including R, under a single Python wrapper. It includes modules to support algebra, geometry, calculus, number theory, cryptography, and numerical computation. It incorporates support to print complex mathematical symbols in LaTeX, and supports both two- and three-dimensional plotting.
- **JPipe** is a Python-to-Java interface. It is rather unique in that rather than reinventing Python in Java (as Jython does) JPipe supports interaction between Python and Java at the JVM level using Java Native Interface (JNI). It is a very convenient extension for calling Java applications from Python, such as **Mahout** and **Weka**. JPipe is written and maintained by Steve Menard, who has made it available to the general public under Apache License version 2.0.

4.5.3 Avro

Apache Avro is a remote procedure call and serialization framework. While originally developed to support access to Hadoop data, Avro has grown to become a “universal” framework for all serialized data. It uses JSON for defining data types and protocols, and serializes data in a compact binary format. JSON embeds metadata in the file describing the data, so Avro files can be read by any program that understands the JSON format. It provides a serialization format for persistent data and an API to present Hadoop services to client programs. APIs are currently defined for C, C++, C#, Java, PHP, Python, and Ruby. Support for Avro’s file structure has been integrated into Flume, Hive, Pig, and Sqoop.

4.5.4 Thrift

Apache Thrift is an interface library and code definition language that is used to define and create communications links between programs written in different languages including C#, C++, Erlang, OCaml, Perl, PHP, Python, Ruby, and Smalltalk. A script file is created that defines the data and services to be created. The server and client components are all generated from the same file, with the data types and protocols unique for each language type. Thrift does not actually provide a server itself; rather, it provides the following:

- Structure definitions for the data to be passed back and forth.
- Subroutines to prepare the data for shipping across the assigned transport type.
- Subroutines for sending and receiving the data.
- Subroutines to implement synchronization.

These features are all wrapped in language-dependent interfaces that are used by the programs themselves to implement the client/server conversation. It is not an ESB (Enterprise Service Bus) nor does it directly support Hadoop. However, Thrift-based servers that support HBase are available.

Thrift was originally developed by Facebook, which donated the code to Apache. Facebook now uses the Apache version of Thrift for its internal use.

4.6 Analytics Tools

Hadoop’s MapReduce is designed to process large volumes of data in a reasonably short amount of time. To meet these requirements, MapReduce provides little in the way of complex analysis capability. The tools defined in this section can be used for the analysis of data residing in HDFS or returned from MapReduce. Some of the products will work equally well on non-Hadoop data platforms.

4.6.1 R

R is an open-source analytics package that is a big player in academia. It provides many if not all of the capabilities found in SAS and SPSS, as well as features not found in either. A large number of enhancements and scripts are available from the R open-source community. The most notable of these is **RHadoop**, a set of packages that support access to HDFS, Hive, and HBase natively from R. Revolution Analytics actively supports R development (for example, RHadoop was developed by Revolution Analytics) and offers product support subscriptions as well as value-added editions of R. While R itself is distributed with only a command-line-based shell, several

GUIs are available for less technical analysts or those analysts who prefer a graphical interface. These GUIs include the open-source products JGR, R Commander, RStudio, RKward, and SciViews-R, and an Eclipse plugin, StaTE for R. Revolution Analytics markets a GUI with its Enterprise Edition that currently supports Windows only; a Linux version is expected to be available with the next release.

4.6.2Pentaho

Pentaho Business Analytics is a complete end-to-end solution that includes authentication, logging, auditing, web services, engines, analysis, OLAP, dashboards, reporting, data integration, and data mining capabilities. Pentaho Business Analytics enables business users to access, discover, and analyze their data without needing to understand programming. Pentaho offers open-source software (Pentaho BI Community Edition, Pentaho Reporting, Kettle, Mondrian, and WEKA) as well as value-added products (Pentaho Enterprise). Pentaho's products run on Hadoop, HDFS, and almost all RDBMSs on the market.

4.6.3KNIME

KNIME (Konstanz Information Miner) is an open-source data integration, processing, analysis, and exploration platform developed at the University of Konstanz (Germany). KNIME Desktop is available as open-source software (under GPLv3), while more advanced versions are available as commercial offerings. A spinoff company, KNIME AG, offers the commercially licensed versions with performance enhancements (running on servers or server clusters), collaboration, and technical support. KNIME AG was designated a “Cool Vendor” by Gartner. KNIME appears to be used primarily in chemistry, pharmaceuticals, and bioinformatics fields, either directly or imbedded within other products.

4.7 Machine Learning Tools

These tools are libraries that are used to build programs that “learn” about data relations, as opposed to programs that have all processing logic predefined. Machine learning is important in Big Data environments because often the data is very dynamic and the volumes are too large to be processed in a timely fashion by human analysts.

Machine learning comes in two classes: “supervised” and “unsupervised.”

- **Supervised learning** is performed by providing the ML program with a subset of data deliberately chosen to illustrate relationships that are to be found in the full dataset. It is useful for testing theories and program behavior before unleashing the program on extremely large datasets, and for data where the relationships are well-defined and stable.
- **Unsupervised learning** is performed by defining a few basic rules and running a program against the full dataset so the program can discover the data relationships on its own. This type of learning is suitable for data where the data relationships are not known beforehand or are in a constant state of flux.

4.7.1Mahout

Apache Mahout is a distributed or otherwise scalable machine-learning library for the Hadoop platform. It supports both supervised and unsupervised learning. Mahout is a work in progress;

the number of implemented algorithms is large, but there are still various multivariate analytics algorithms missing. Mahout currently supports the following use cases.

- **Recommendation Mining:** analyzing a user's behavior to predict other items of interest.
- **Clustering:** grouping documents by topic.
- **Classification:** learning attributes of previously categorized documents and applying this learning to unlabeled documents.
- **Frequent Item Set Mining:** determining items that are generally found together and defining them into groups.

4.7.2 Weka

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine-learning schemes, and an R interface is available (RWekaJars). Weka was developed by the University of Waikato (New Zealand) under the GPL. The Weka project is extensively supported by Pentaho and is incorporated into the Pentaho Business Analytics suite.

4.7.3 Natural Language Toolkit

NLTK is a set of Python-written tools that support natural language processing, also known as computational linguistics. It consists of open-source Python modules, linguistic data, and documentation for research and development in natural language processing and text analytics. It was created and is maintained by Edward Loper and Steven Bird and is available to the general public under Apache License version 2.0.

4.8 Other Tools of Interest

4.8.1 ZooKeeper

ZooKeeper is a centralized service written in Java for providing or maintaining configuration information, naming, distributed synchronization, and group services. It does not perform these services itself, but instead provides a reliable platform on which these more complex services can be built. Originally a subproject of Hadoop, ZooKeeper is used by Hadoop MapReduce and many Hadoop-based applications. It provides both high performance and high availability by implementing its registry in memory, cloned across several servers. The ZooKeeper service guarantees the following.

- **Sequential Consistency:** Updates from a client are applied in the order in which they were sent.
- **Atomicity:** Updates either succeed or fail (no partial results).
- **Single System Image:** A client will see the same view of the service regardless of the server that it connects to.

- **Reliability:** Once an update has been applied, it will persist from that time forward until a client overwrites the update.
- **Timeliness:** The client's view of the system is guaranteed to be up-to-date within a certain time bound.

4.8.2 Oozie

Oozie is a workflow/coordination system for managing Apache Hadoop jobs. Oozie Workflow jobs are directed acyclic graphs (or in simple terms, non-looping flowcharts) of actions. Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability. Oozie is integrated with the rest of the Hadoop stack, supporting several types of Hadoop jobs out-of-the-box (Java MapReduce, Streaming MapReduce, Pig, etc.).

Oozie is largely text oriented; for example, workflows are created in XML. While it includes a web-based console, the console only supports the monitoring of already-submitted jobs. However, Oozie is still in the Apache incubator stage, and may include a GUI front-end by the time it graduates to a full-fledged Apache project.

4.8.3 Hue

Hue is a web interface and application framework for Hadoop. It is an open-source project supported by Cloudera. It includes a file browser for HDFS files, an application for designing and submitting Hadoop jobs, the Beeswax GUI for Hive, and a job monitor. It supports Python, Django (a high-level Python web framework), Mako (a template language), and MooTools (JavaScript framework).

4.8.4 Lucene

Apache Lucene is a high-performance text search library, written in Java. It provides APIs to incorporate its search engine into other projects but is not an out-of-the-box search tool in and of itself. It is incorporated into a huge number of applications (such as Solr) and web sites.

4.8.5 Solr

Apache Solr is a search server that is built on top of Lucene. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, rich document (e.g., Word, PDF) handling, and geospatial search. Solr is highly scalable, providing distributed search and index replication. It provides APIs for XML/HTTP and for JSON, Ruby, and Python.

4.8.6 Starfish

Starfish is a cost-based query optimizer for Hadoop under development at Duke University. It comprises the following components.

- **Profiler:** The Profiler uses dynamic instrumentation to learn performance models, called job profiles, for unmodified MapReduce programs. The Profiler also exposes an interface for analyzing past MapReduce executions.

- **What-if Engine:** The What-if Engine uses a mix of simulation and model-based estimation at the task-phase level of MapReduce job execution, in order to predict the performance of a MapReduce job before the job's execution on a Hadoop cluster.
- **Cost-based Optimizer:** The Cost-based Optimizer enumerates and searches through the high-dimensional space of job configuration settings, making appropriate calls to the What-if Engine, in order to find the optimal configuration settings to use for executing a MapReduce job.
- **Visualizer:** The Visualizer provides a GUI that allows the user to (a) analyze past MapReduce job executions, (b) ask hypothetical (what-if) questions on how the job behavior will change when parameter settings, cluster resources, or data properties change, and (c) ultimately optimize the job.

Starfish is available from Duke University for research purposes under a limited license. A commercial license is also available for full, non-exclusive commercial use. Starfish should not be incorporated into an *<THE AGENCY>* Hadoop structure at this time. It is documented here for future reference, should the need for a cost-based optimizer arise.

4.8.7 Talend Open Studio for Big Data

Talend Open Studio for Big Data is an integrated development environment designed for use with Hadoop. It provides a layer of abstraction over Hadoop and related applications that contains a GUI. The Open Studio GUI allows users without expertise in Pig, HBase, Sqoop, or Hive to create Big Data jobs and MapReduce tasks. It is available without cost from Talend under the Apache License version 2.0.

5.0 Evaluation of Vendor Technologies

This section provides an overview of several vendors that provide and support Hadoop distributions, or that provide useful applications that run within the Hadoop infrastructure. It by no means represents a complete list of vendors in the Hadoop arena, but rather presents an evaluation of top-tier vendors whose products/capabilities will be needed by <THE AGENCY> to perform Big Data and predictive analysis.

- The vendors selected as candidates for further evaluation are presented below in Table 2: Candidates for Vendor Evaluation*

- Table 2: Candidates for Vendor Evaluation*

Candidates for Vendor Evaluation	
Vendor	Product/Service
Cloudera	OSS Hadoop platform and support, Cloudera Manager for monitoring and administration.
EMC Greenplum	OSS Hadoop platform and support, Greenplum data warehouse platform.
IBM	OSS Hadoop platform and support; InfoSphere products for data capture, analysis, monitoring and administration; Cognos BI for analytics.
MapR	OSS Hadoop platform and support, MapR Control System for job monitoring and management.
Hortonworks	OSS Hadoop platform, training, and support. Includes Ambari for cluster management and HCatalog for metadata.
Pentaho	Business analytics and data integration products built on Hadoop.
Pervasive Software	High-performance data analytics development platform built on Hadoop.
Jaspersoft	Business analytics and data integration products that run against Hadoop and most RDBMS and NoSQL products currently on the market.
Revolution Analytics	R statistics package with Hadoop performance optimization and support.

5.1 Hadoop Distributions

Karvy Analytics has selected five Hadoop distributors to be evaluated for use at <THE AGENCY>: Cloudera, EMC Greenplum, Hortonworks, IBM, and MapR. These distros were chosen for the following reasons.

- **Maturity:** Stable versions of the distros have been available for a year or more. One distro that was still in beta testing at the time of this paper, from Hortonworks, will be of interest to <THE AGENCY>. While Hortonworks is a “new” distributor, it is a spin-off from Yahoo, the creator of Hadoop. It has been included in this review because of that pedigree.
- **Open-Source Edition:** The distros chosen are all available solely as open-source software. The selected vendors all freely offer an OSS edition as a “basic” version consisting of software with little or no support. A second-tier offering featuring enhancements and support is available, but in each case builds on the OSS foundation of the basic distro. By selecting only OSS-based software, <THE AGENCY>’s chance of vendor lock-in is reduced.
- **Production Support:** Each vendor selected for evaluation offers a production version of its distribution that includes product support as well as additional applications needed to support a production environment.
- **Recognized Industry Leader:** Each selected vendor has been recognized by industry analysts such as Gartner or Forrester as a leader in the Big Data market.

It should be repeated that the above list is not exhaustive, nor is there a requirement to use any vendor’s distribution. <THE AGENCY> can create its own distro by downloading all of the required software from OSS repositories, including both source code and binaries. Creating a Hadoop environment in this manner will provide the greatest flexibility while preventing <THE AGENCY> from being locked into a single software vendor. However, this approach will require <THE AGENCY> to conduct more research prior to applying maintenance to ensure that updates to any component of the environment are compatible with any other updates. If the agency were to rely on a vendor-provided distribution, then that vendor would be responsible for maintaining compatibility, reducing the time and resources <THE AGENCY> would need to maintain its Big Data environment. Lock-in to a particular vendor’s distro can still be minimized by mainly using OSS components of the selected distribution, avoiding features of a proprietary nature.

Table 3 below presents a comparison of the selected vendors and their products.

- *Table 3: Comparison of Hadoop Distribution Vendors*

Hadoop Distribution Vendor Comparison					
Application \ Vendor	Cloudera	EMC Greenplum	Hortonworks	IBM	MapR
Free Edition Name	CDH	Community Edition	HDP	Basic Edition	M3 Edition
Enterprise Edition Name	Cloudera Enterprise	Enterprise Edition		Enterprise Edition	M5 Edition
HDFS	•	•	•	•	•
Hadoop MapReduce	•	•	•	•	•
HBase	•	•	•	•	•
Hive	•	•	•	•	•
Pig	•	•	•	•	•
ZooKeeper	•	•	•	•	•
Oozie	•			•	•
Avro	•			•	•
Lucene				•	
Mahout	•	•			•
Administration Tools	Cloudera Manager	MapR Control System	Ambari	BigInsights Console	MapR Control System
Security	Kerberos			LDAP	
Other OSS	Flume, Hue, Sqoop, Whirr		HCatalog	Flume	
Other Proprietary					Direct Access MFS

5.1 Cloudera

Cloudera is found in the Forrester Wave Leaders Circle for Hadoop distributions, in Gartner's "How to Choose the Right Apache Hadoop Distribution," and in O'Reilly Radar's "[Big data market survey: Hadoop Solutions](#)." Cloudera is a major contributor to the Hadoop open-source community as well as the most responsible for the acceptance of Hadoop in the marketplace today. As a result, Cloudera is the most established of the Hadoop distro vendors, both in market share and longevity.

The Cloudera Distribution for Hadoop (CDH) is entirely composed of open-source components, including HDFS, Hadoop, HBase, Hive, Pig, and ZooKeeper for core Hadoop processing; Flume for log data input; Sqoop for RDBMS interconnection; Avro for application connectivity; Fuse-DFS to support access to HDFS files from non-HDFS-aware applications; Mahout for machine learning; and Hue and Oozie for administration. It is so complete that the Enterprise Edition adds

only an unrestricted version of Cloudera Manager for improved administration, and full technical support.

Karvy Analytics recommends that <THE AGENCY> use the Cloudera CDH distribution to establish its Hadoop environment.

5.1.2EMC (Greenplum)

EMC acquired Greenplum in July 2010. EMC Greenplum is found in the Forrester Wave Leaders Circle for Hadoop distributions, in Gartner’s “How to Choose the Right Apache Hadoop Distribution,” and in O’Reilly Radar’s “[Big data market survey: Hadoop Solutions](#).” The Greenplum Community and Greenplum HD editions are both based on the corresponding MapR distributions. EMC adds value to these base distributions by providing performance improvements and integration with the Greenplum NoSQL DBMS. The Greenplum distros include HDFS, Hadoop, HBase, Hive, Pig, and ZooKeeper for core Hadoop processing, and Mahout for machine learning. Greenplum HD includes additional fault-tolerance and automatic recovery features as well as product support.

5.1.3Hortonworks

Hortonworks was created in June 2010 as a joint venture between Yahoo! and Benchmark Capital, a venture capital firm. Hortonworks states on its website that HDP (Hortonworks Data Platform) will contain 100% open-source code and that any extensions created by Hortonworks will be donated to Apache Hadoop.

Release 1.0 of HDP is due out in May 2012 and thus was not available at the time of this writing. The following information is taken directly from the Hortonworks website.

Hortonworks Data Platform includes the most popular and essential Apache Hadoop projects including the Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase and Zookeeper. In addition to these components, Hortonworks Data Platform includes open source technologies that make the Hadoop platform more ***manageable, open, and extensible***.

Unlike other Hadoop solutions that lock away management features within proprietary extensions, Hortonworks Data Platform includes ***Ambari***, an open source installation and management system out of the box. Hortonworks Data Platform also includes ***HCatalog***, a metadata management service for simplifying data sharing between Hadoop and other enterprise information systems, along with a complete set of ***open APIs***, including WebHDFS and those for Ambari and HCatalog, to make it easier for ISVs to integrate and extend Apache Hadoop.

All of these components have been integrated and tested as part of the Hortonworks Data Platform release process. Installation and configuration tools have also been included to make it easier to install, deploy and use the Hortonworks Data Platform.

The initial release of Hortonworks Data Platform (version 1) is based on Apache Hadoop 0.20.205, a stable release and the first Apache Hadoop release to support security and HBase. There has also been significant progress on Next Generation MapReduce in Apache Hadoop 0.23, and within the coming weeks we expect to release an early

technology preview of Hortonworks Data Platform version 2 which includes this important and emerging technology.

5.1.4IBM

IBM BigInsights is found in Gartner’s Magic Quadrant for Business Intelligence Platforms, Forrester’s Wave Leaders Circle for Hadoop distributions, Gartner’s “How to Choose the Right Apache Hadoop Distribution,” and O’Reilly Radar’s “[Big data market survey: Hadoop Solutions](#).” The IBM BigInsights Basic Edition is a free Hadoop distribution, while the IBM BigInsights Enterprise Edition includes extensive analytics capabilities and so is discussed in [section 5.3, Proprietary Analytics Products](#). The BigInsights Basic Edition comprises HDFS, Hadoop, HBase, Hive, Pig, and ZooKeeper for core Hadoop processing; Flume for log data input; Avro for application connectivity; Jaql for RDBMS interconnectivity; Lucene for text searching; and Oozie for administration. In addition to providing analytics software and product support, the Enterprise Edition replaces HDFS with GPFS, a POSIX-compliant file system that provides greater performance and recoverability than HDFS while still appearing like HDFS to other Hadoop components.

5.1.5MapR

MapR is found in the Forrester Wave Leaders Circle for Hadoop distributions, in Gartner’s “How to Choose the Right Apache Hadoop Distribution,” and in O’Reilly Radar’s “[Big data market survey: Hadoop Solutions](#).” The M3 (free) edition contains HDFS, Hadoop, HBase, Hive, Pig, and ZooKeeper for core Hadoop processing; Flume for log data input; Sqoop for RDBMS interconnection; Avro for application connectivity; Mahout for machine learning; and Oozie for administration. Both editions include the MapR Control System for administration and Direct Access NFS, which allows the Hadoop file system to be mounted like a network file system for access from non-HDFS-aware software. The M5 edition includes full product support and high availability, performance, and recovery features not found in the M3 version. MapR products are often embedded in OEM offerings, such as EMC Greenplum.

5.2 Open-Source Analytics Products

Karvy Analytics has selected four open-source analytics distributions for this study. As with the selection of Hadoop distributions in the previous section, this is neither an exhaustive list nor a recommendation of a specific vendor. Rather, these vendors’ products have been included to provide an example of what is available from the open-source community. The vendors were selected based on the following criteria.

- **Maturity:** Stable versions of the distros have been available for a year or more.
- **Open-Source Edition:** The distros chosen are all available solely as open-source software. The selected vendors all freely offer their OSS edition as a “basic” version consisting of software with little or no support. A second-tier offering featuring enhancements and support is available, but in each case builds on the OSS foundation of the basic distro. By selecting only OSS-based software, <THE AGENCY> can reduce the chance of vendor lock-in.

- **Production Support:** Each vendor discussed in this section offers a production version of its distribution that includes product support as well as additional applications needed to support a production environment.
- **Recognized Industry Leader:** Each vendor has been recognized by industry analysts such as Gartner or Forrester as a leader in the Big Data market.

Table 4 below presents a comparison of OSS analytics product vendors.

- **Table 4: Comparison of OSS Analytics Vendors**

Open Source Analytics Vendors Comparison				
Vendor Feature	Jaspersoft	Pentaho	Pervasive Software	Revolution Analytics
Product Name	Jaspersoft BI Suite	Pentaho Analytics	DataRush for KNIME	Revolution R
Free Edition	jasperforge.org	community.pentaho.com	KNIME.org	r-project.org
Reporting	●	●	●	●
Dashboards	●	●	●	○
Ad Hoc Query	●	●	●	●
Interactive Visualization		●	●	●
OLAP	●	●	—	○
ETL	●	●	—	○
Data Mining	●	●	●	○
Statistical Analytics	—	●	●	●
Predictive Analytics	—	●	●	●

Legend:

- Supported by product
- Supported by CRAN package
- Not supported

5.2.1 Pentaho Business Analytics

Pentaho Business Analytics is a complete end-to-end solution that includes business intelligence, data integration, and data mining capabilities. According to Gartner in *Who's Who in Open Source*, "Pentaho provides a breadth of functionality that can be considered the closest match to commercial offerings such as Business Objects, Cognos or Oracle BI." Pentaho Business Analytics supports reporting, analysis, dashboarding, data mining, OLAP and ETL. Advanced statistics and predictive analytics are supported by the incorporation of OSS from the Weka project, which Pentaho actively supports. Pentaho runs on Hadoop, HDFS, and almost all of the RDBMS products on the market. The OSS edition is available for free download from community.pentaho.com

Karvy Analytics recommends that <THE AGENCY> use Pentaho's OSS edition for use in its Big Data test environment. This will provide the business analytics capabilities needed for the initial testing of the <THE AGENCY> Big Data environment without requiring the purchase of a business analytics tool that may not be amenable to <THE AGENCY>'s analyst's requirements.

5.2.2 Jaspersoft

Jaspersoft provides a complete BI platform for reporting and analysis, including a metadata and content repository, report management, security integration, and scheduling. The BI platform has been designed for stand-alone deployments and integration with other applications through web services such as Java, HTTP, and REST-based APIs. Jaspersoft provides native connectivity to NoSQL engines such as Hadoop and MongoDB. The Community Edition is OSS and available from jasperforge.org. It contains a substantial subset of the Enterprise Edition.

The Jaspersoft BI suite includes JasperReports Library, JasperReports Server, Jaspersoft iReports Designer, Jaspersoft ETL, and Jaspersoft OLAP.

- **JasperReports Server:** Supports data retrieval from JDBC, HBase, HIVE, EJB, POJO, Hibernate, MDX, XML, CSV, and custom data sources. It provides ad hoc query, in-memory and Big Data analysis, dashboarding, and mash-up capability.
- **iReports Designer:** Provides a visual tool for building reports.
- **JasperReports Library:** Provides a repository for storing the reports built with **iReports Designer**.
- **Jaspersoft OLAP:** Provides OLAP features via a web browser or Excel interface.
- **Jaspersoft ETL:** Provides data integration, cleaning, and enrichment.

Jaspersoft does not provide advanced statistics or predictive analytics natively, but it can be connected to R via **RevoConnectR for JasperReports Server**, an OSS project jointly maintained with Revolution Analytics. It is available at jasperforge.org

5.2.3 Pervasive DataRush for KNIME

Pervasive's products are designed primarily to "turbo charge" processing on Hadoop and or SMP clusters. The company's analytics offering is based on KNIME, incorporated with its performance products. The free KNIME package is available directly from knime.org. Pervasive claims that the combination of its DataRush technology with KNIME reduces the time needed to perform analysis from hours to seconds by use of parallel processing on 100 nodes or more. The features listed are actually for the KNIME software. KNIME itself is a GPL OSS application that supports advanced analytics, predictive analysis, visualization, and reporting. It has interfaces to both Weka and R.

5.2.4 Revolution Analytics

Revolution Analytics software is built on R, an open-source-software statistics package that compares favorably with SAS and SPSS, offering as many or more capabilities as these commercial packages. The free R software package is available directly from the company's website. Also available is RHadoop, an open-source package of plugins for R that support direct access from R to HDFS, Hive, and HBase. RevolutionR, the company's premium offering, integrates seamlessly with Hadoop and can take advantage of parallel processing using MapReduce. RevolutionR also comes with product support. In addition, RevolutionR includes a GUI for Windows users. A Linux version of the GUI is expected to be available sometime in 2012.

5.3 Proprietary Analytics Products

Karvy Analytics has selected proprietary analytics products from four vendors as candidates for use at <THE AGENCY>. These are presented below in **Table 5**.

- **Table 5: Comparison of Proprietary Software Analytics Vendors**

Proprietary Analytics Vendor Comparison				
Vendor Feature	IBM	Microstrategy	SAS	TIBCO
Product Name	InfoSphere BigInsights	Microstrategy BI	SAS Analytics	Spotfire
Reporting	●	●	●	●
Dashboards	●	●	●	●
Ad Hoc Query	●	●	●	●
Interactive Visualization	●	●	●	●
OLAP	●	●	●	●
ETL	●	●	●	●
Data Mining	●	●	●	●
Statistical Analytics	●	●	●	●
Predictive Analytics	●	●	●	●

Legend:

- Supported by product
- Supported by 3rd Party Vendor
- Not supported

5.3.1 IBM

IBM BigInsights is in Gartner's Magic Quadrant and Forrester's Wave Leaders Circle. The IBM BigInsights Enterprise Edition includes extensive analytics capabilities.

- **BigSheets:** This is a browser-based spreadsheet application that can pull data from diverse locations such as local files or the web and supply them to Hadoop MapReduce for processing, while hiding the mechanics (from the user) behind the spreadsheet interface. It includes several options for visualizing the results, such as pie charts, bar charts, heat maps, geographic maps, and tag clouds, and supports macro creation and execution.
- **Advanced Analytics Toolkit:** This toolkit integrates with MapReduce to perform text analytics. Scripts (referred to as “extractors” by IBM) are written in SQL-like language called AQL (Annotator Query Language). The extractors are compiled and distributed across the Hadoop cluster for processing. Prebuilt extractors are provided for common text types, such as name, address, phone number, etc.

- **BigIndex:** This application is built on top of Lucene and is used to build indexes for text search capability. It supports partitioned and distributed indexes, as well as index building/modification in near real time.
- **Jaql:** This is an IBM-developed OSS query language that can process both structured and unstructured data. It is an SQL-like declarative language. Jaql is extensible via modules, and IBM is leveraging this capability to include useful prebuilt Jaql modules in BigInsights to enable integration with (for example) text analytics and Netezza warehouses.

InfoSphere BigInsights supports open development standards and works with existing systems, including IBM Netezza, InfoSphere Warehouse, IBM Smart Analytics System, and IBM DB2 for Linux, UNIX, and Windows.

5.3.2 Microstrategy BI

Microstrategy BI is in Gartner's Leaders Magic Quadrant for BI Analytics. It works on a large number of data platforms and is cited by its customer base for providing high-performance processing.

Microstrategy BI provides a full complement of analysis tools, including reporting, visualization, dashboarding, OLAP and ROLAP, integration between data sources to support drilldown from the data warehouse into Hadoop, and support for mobile computing platforms. It also provides an extensive library of statistical functions, metrics, and filter objects to support the creation of complex analytical applications and reports.

It interacts with Hadoop via Hive, providing both point-and-click access to Hive-defined data as well as direct submission of HiveQL scripts. It does not currently integrate with MapReduce, nor can it use native HDFS files.

5.3.3 SAS

SAS is the 500-pound gorilla of the analytics world. Gartner perennially places SAS in the Leaders Magic Quadrant for business intelligence platforms, although both IBM and Microstrategy score higher overall. SAS is rated as both expensive and difficult to implement by its customer base. However, *<THE AGENCY>* already owns and makes use of SAS, so these drawbacks may not be of concern.

SAS provides advanced analytical techniques for data mining and predictive analytics, and is rated by its customers as well above the offerings of other vendors in its advanced analytics capability. Despite this mindshare, products available from IBM, TIBCO, and Microstrategy all provide similar capabilities, as does the R open-source statistics package.

SAS added native support for Hadoop databases in March 2012. It supports MapReduce, direct HDFS access, and Hive data structures. Karvy Analytics recommends that *<THE AGENCY>* consider testing this new version of SAS in its Big Data sandbox to determine its suitability for production analytics use.

5.3.4 TIBCO Spotfire

The Spotfire analytics platform supports ad hoc analysis, interactive reporting and dashboards, domain-specific applications, event-driven real-time analysis, and statistical analysis. Gartner placed TIBCO Spotfire in its challenger Magic Quadrant for business intelligence platforms in

both 2011 and 2012. In Gartner's survey, Spotfire was rated as having some of the most complex analysis capabilities and yet one of the easiest platforms to use. Spotfire also provides an option to integrate R and S with Spotfire's visualization tools, using scripts written to utilize these powerful analytics packages. The R and S scripts can be written and verified by data scientists and made available for use to others who do not write in these languages.

As of the time of this writing, TIBCO did not offer native integration with Hadoop.

6.0 Summary of Recommendations

The following table lists the major components that Karvy Analytics recommends <THE AGENCY> use to establish their Big Data environment.

- *Table 6: Recommended Software*

<THE AGENCY>'s Big Data Test Environment	
Component	Product
Operating System	CentOS Linux 6.2
Hadoop Distribution	Cloudera CDH 3
Analytics Platform	R; Pentaho BI Community Edition
Programming Language	Python 2.7
Machine Learning Library	Mahout

7.0 References

7.1 Websites for Products Mentioned in this Document

Vendor/Product	URL
Apache Ambari	http://incubator.apache.org/ambari/
Apache Avro	http://avro.apache.org/
Apache Cassandra	http://cassandra.apache.org/
Apache Chukwa	http://incubator.apache.org/chukwa/
Apache CouchDB	http://couchdb.apache.org/
Apache Giraph	http://incubator.apache.org/giraph/
Apache Hadoop	http://hadoop.apache.org/
Apache HBase	http://hbase.apache.org/
Apache HCatalog	http://incubator.apache.org/hcatalog/
Apache Hive	http://hive.apache.org/
Apache Lucene	http://lucene.apache.org/core/
Apache Mahout	http://mahout.apache.org/
Apache Oozie	http://incubator.apache.org/oozie/
Apache Pig	http://pig.apache.org/
Apache Piggybank	https://cwiki.apache.org/confluence/display/PIG/PiggyBank
Apache Software Foundation	http://apache.org/
Apache Solr	http://lucene.apache.org/solr/
Apache Sqoop	http://incubator.apache.org/Sqoop/
Apache Thrift	http://thrift.apache.org/
Apache Whirr	http://wiki.apache.org/incubator/WhirrProposal
Apache ZooKeeper	http://zookeeper.apache.org/
Arch Linux	http://www.archlinux.org/
CentOS Linux	http://www.centos.org/
Cloudera Community Edition	http://www.cloudera.com/community/
Cloudera Enterprise	http://www.cloudera.com/products-services/enterprise/
Cloudera Manager	http://www.cloudera.com/products-services/tools/
Couchbase	http://www.couchbase.com/
CRAN	http://cran.r-project.org/
Disco	http://discoproject.org/
Dumbo	https://github.com/klbostee/dumbo/wiki/
EMC Greenplum	http://www.greenplum.com/
Fedora Linux	http://fedoraproject.org/
Flume	https://github.com/cloudera/flume
Happy	http://code.google.com/p/happy/

Hortonworks	http://hortonworks.com/
Hue	https://github.com/cloudera/hue
IBM BigInsights	http://www-01.ibm.com/software/data/infosphere/biginsights/
Jaspersoft BI Suite	http://www.jaspersoft.com/bi-platform
Jaspersoft	
Community	
Edition	http://jasperforge.org/
JGR	http://rforge.net/JGR/index.html
JPipe	http://jpipe.sourceforge.net/
Kerberos	http://web.mit.edu/kerberos/
KNIME	http://www.knime.org/
MapR	http://www.mapr.com/
Microstrategy BI	http://microstrategy.com/software/businessintelligence/
MongoDB	http://www.mongodb.org/
Natural Language Toolkit	http://www.nltk.org/
OpenSUSE Linux	http://www.opensuse.org/en/
Pentaho Analytics	http://www.pentaho.com/
Pentaho BI	
Community	
Edition	http://community.pentaho.com/
Pervasive	
DataRush for KNIME	http://www.pervasivebigdata.com/Products/PervasiveDataRushforKNIME.aspx http://sourceforge.net/apps/mediawiki/pydoop/index.php?title=Main_Page
Pydoop	http://Python.org/
Python	
R	http://www.r-project.org/
R Commander	
Revolution	http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/
Analytics	
RHadoop	
RKward	http://www.revolutionanalytics.com/
RPy2	https://github.com/RevolutionAnalytics/Rhadoop
RStudio	Main_Page">http://sourceforge.net/apps/mediawiki/rkward/index.php?title>Main_Page
Sage Math	http://sourceforge.net/projects/rpy/
SAS	http://www.rstudio.org/
Scala	http://www.sagemath.org/
SciViews-R	http://www.sas.com/
Scribe	http://www.scala-lang.org/
Spark	http://www.sciviews.org/SciViews-R/index.html
Starfish	https://github.com/facebook/scribe
StaTE for R	http://www.spark-project.org/
	http://www.cs.duke.edu/starfish/index.html
	http://www.walware.de/goto/statet

Talend Open Studio for	http://www.talend.com
Tibco Spifire	http://www.tibco.com/products/business-optimization/analytics-visualization/default.jsp
Ubuntu Linux	http://www.ubuntu.com/ubuntu
Weka	http://www.cs.waikato.ac.nz/ml/weka/

7.2 Bibliography

Market Survey: Hadoop Solutions

Dumbill, Edd

O'Reilly Radar, January 19, 2012

<http://radar.oreilly.com/2012/01/big-data-ecosystem.html>

The Forrester Wave: Enterprise Hadoop Solutions, Q1 2012

Kobiels, James G.

Forrester Research, Inc., February 2, 2012

How to Choose the Right Apache Hadoop Distribution

Adrian, Merv

Gartner, Inc. February 9, 2012

Who's Who in Open-Source Business Intelligence

Bitterer, Andreas

Gartner, Inc., April 16, 2008

Magic Quadrant for Business Intelligence Platforms

Hagerty, John; Sallam, Rita L.; Richardson, James

Gartner, Inc. February 10, 2012

Predicts 2012: Information Infrastructure and

Gartner, Inc., November 29, 2011

Blechar, Mike; Adrian, Merv; Friedman, Ted; Schulte, W. Roy; Laney, Douglas

7.3 Recommended reading

Maverick Research: Judgment Day, or Why We Should Let Machines Automate Decision Making

Rayner, Nigel

Gartner, Inc., October 7, 2011

The next frontier for innovation, competition, and productivity

Manyika, James; Chui, Michael; Brown , Brad; Bughin, Jacques; Dobbs, Richard;

Roxburgh, Charles; Byers, Angela Hung

McKinsey Global Institute, June, 2011

The Age of Big Data

Lohr, Steve

New York Times, February 11, 2012

Hadoop Spurs Big Data Revolution

Henschen, Doug

InformationWeek, November 09, 2011

<http://www.informationweek.com/news/development/database/231902466>

Does the 21st-Century "" Warehouse Mean the End of the Enterprise Data Warehouse?

Beyer, Mark A; Feinberg, Donald

Gartner, Inc., August 25, 2011

Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data

Eaton, Chris; Deroos, Dirk; Deutsch, Tom; Lapis, George; Zikopoulos, Paul

McGraw-Hill, 2012

From Data to Decisions: The Power of Analytics

Partnership for Public Service

IBM Center for the Business of Government, November, 2011

Should I use Python 2 or Python 3 for my development activity?

Pitrou, Antoine

December 9, 2011

<http://wiki.Python.org/moin/Python2orPython3>



ASSOCIATE ANALYTICS FACILITATORS GUIDE MODULE 2



This Facilitators Guidebook for the Associate Analytics program contains detailed facilitation guidelines as well as the exhaustive course material for the Associate Analytics program.

Facilitator's Guide



Associate - Analytics

Powered by:



Copyright © 2014

NASSCOM

4E-Vandana Building (4th Floor)
11, Tolstoy Marg, Connaught Place
New Delhi 110 001, India
T 91 11 4151 9230; F 91 11 4151 9240
E ssc@nasscom.in
W www.nasscom.in

Published by



Building Domain | Enhancing Careers

T: 91 70365 88888
E info@mindmapconsulting.com
W www.mindmapconsulting.com

Disclaimer

The information contained herein has been obtained from sources reliable to NASSCOM. NASSCOM disclaims all warranties as to the accuracy, completeness or adequacy of such information. NASSCOM shall have no liability for errors, omissions, or inadequacies, in the information contained herein, or for interpretations thereof. Every effort has been made to trace the owners of the copyright material included in the book. The publishers would be grateful for any omissions brought to their notice for acknowledgements in future editions of the book.

No entity in NASSCOM shall be responsible for any loss whatsoever, sustained by any person who relies on this material. The material in this publication is copyrighted. No parts of this report can be reproduced either on paper or electronic media, unless authorized by NASSCOM.

Foreword

The Indian IT-ITeS industry has built its reputation in the global arena on several differentiators, chief among them being the availability of manpower. Organizations across the world recognize the value India brings to every engagement with its vast and readily available pool of IT professionals. Global entities have found it extremely effective to leverage this significant resource in order to enjoy a competitive edge and innovation benefits.

In the coming years, the landscape is expected to shift in ways that reveal more exciting opportunities. The world will require people with advanced technology skills and domain knowledge, set against a backdrop of heightened labour mobility across occupations and markets. India is largely acknowledged to be heir apparent to the benefits of a demographic dividend over the coming decades, which has the potential to see the nation emerge as one of the world's largest population base of employable youth. With many other countries set to face the effects of an aging and retirement-ready workforce, India is poised to become a sought after destination for those seeking higher value add and specialized services.

Global markets are on their way towards revival and recovery, and this is well reflected in the proactive recruitment measures taken by IT-ITeS organizations in India in recent times. India's IT-BPM industry is on track to achieve its target of USD 225 billion by 2020. From a base on about 3.1 million employees in FY2014, the industry is expected to add another 2 million additional employees by 2020. Indirect employment generated by 2020 is expected to be 3X the total direct employment number is between 13-16 million by 2020.

To realize India's potential of emerging as a skills hub of the world, a significant amount of foresight and work is requisite. It is imperative that stakeholders engage in a concerted effort to undertake the transformation of the labour pool estimated to enter the market into skilled and employable talent. Enabling the creation of a future industry-ready cohort will give the IT-ITeS industry an edge in leadership and sustainability.

One of the burgeoning areas of governance and strategy relates to leveraging big data and analytics. This led to the identification of the "hot skills" du jour, resulting in the formal creation of a qualification pack (QP) or job role framework for the role of Associate Analytics. The QP is designed to capture the skills demanded by the IT-BPM Industry for an entry level position in this field.

To ensure the creation of an academic course that is both relevant and viable, NASSCOM partnered with key industry stakeholders, including Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. In addition, the program addresses the need for faculty support, and achieves this by acquainting trainers with the latest advancements in pedagogy.

We wish the universities and colleges all the very best in their endeavor.

R Chandrashekhar
President
NASSCOM

Acknowledgements

NASSCOM would like to thank its member company representatives within the Analytics Special Interest Group (SIG) Council for believing in our vision to enhance the employability of the available engineering student pool. SSC NASSCOM facilitates this by developing and enabling the implementation of courses relevant to projected industry needs. The aim is to address two key requirements, of closing the industry-academia skill gap, and of creating a talent pool that can reasonably weather future externalities in the IT-BPM industry.

NASSCOM believes that this is an initiative of great importance for all stakeholders concerned – the industry, academia, and the students. The tremendous amount of work and ceaseless support offered by the members of this SIG in developing a meaningful strategy for the content and design of program training materials has been truly commendable.

We would like to particularly thank Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, Fractal Analytics, First America, Genpact, Infosys BPO, Insights of Data, Karvy Analytics, Wipro, WNS and Wells Fargo for bringing much needed focus to this effort.

NASSCOM recognizes the fantastic contributions of Mr. Ashok Polapragada, Mr. Ranjit Kumar and Mr. Prakash Devarakonda at Karvy Analytics; Mr. Dwaraka Ramana K at First American; Mr. Amit Agarwal, Mr. Sidhartha Shishoo and team at Genpact; Ms. Snigdha Ray and Mr. Amit Sharma at ADP; Mr. Manoj Koundinya at Capgemini, and Mr. Ashish Mediratta at Wipro.

We acknowledge with sincere gratitude the immense contribution of the SIG member companies, Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. For their part in the creation of this course and its accompanying training materials.

We extend our thanks to Mindmap Consulting Pvt. Ltd. for producing this course publication.

Dr Sandhya Chintala
Executive Director – Sector Skill Council
Vice President - NASSCOM

Table of Contents – Module 2

Introduction to QP Associate Analytics

Introduction to Associate Analytics	8
Career growth in Analytics	11
Qualification pack - Q/2101 Associate Analytics	12
Overall Associate Analytics Content Structure	20
Glossary of terms	22

CORE CONTENT

UNIT 1.1 Data Management	27
UNIT 1.2 Maintain Healthy, Safe & Secure working Environment	36
UNIT 2.1 Big Data Tools	69
UNIT 2.2 Provide Data/Information in Standard Formats	76
UNIT 3.0 Big Data Analytics	104
UNIT 4.0 Machine Learning Algorithms	115
UNIT 5.1 Data Visualization	121
UNIT 5.2 Analytics Application to various domains	134
UNIT 6.0 Case Study in R (Recap from Module 1)	139

Introduction

Qualifications Pack-Associate –Associate Analytics SSC/Q2101

SECTOR: IT-ITeS

SUB-SECTOR: Business Process Management

OCCUPATION: Analytics

REFERENCE ID: SSC/Q2101

ALIGNED TO NCO CODE: TBD

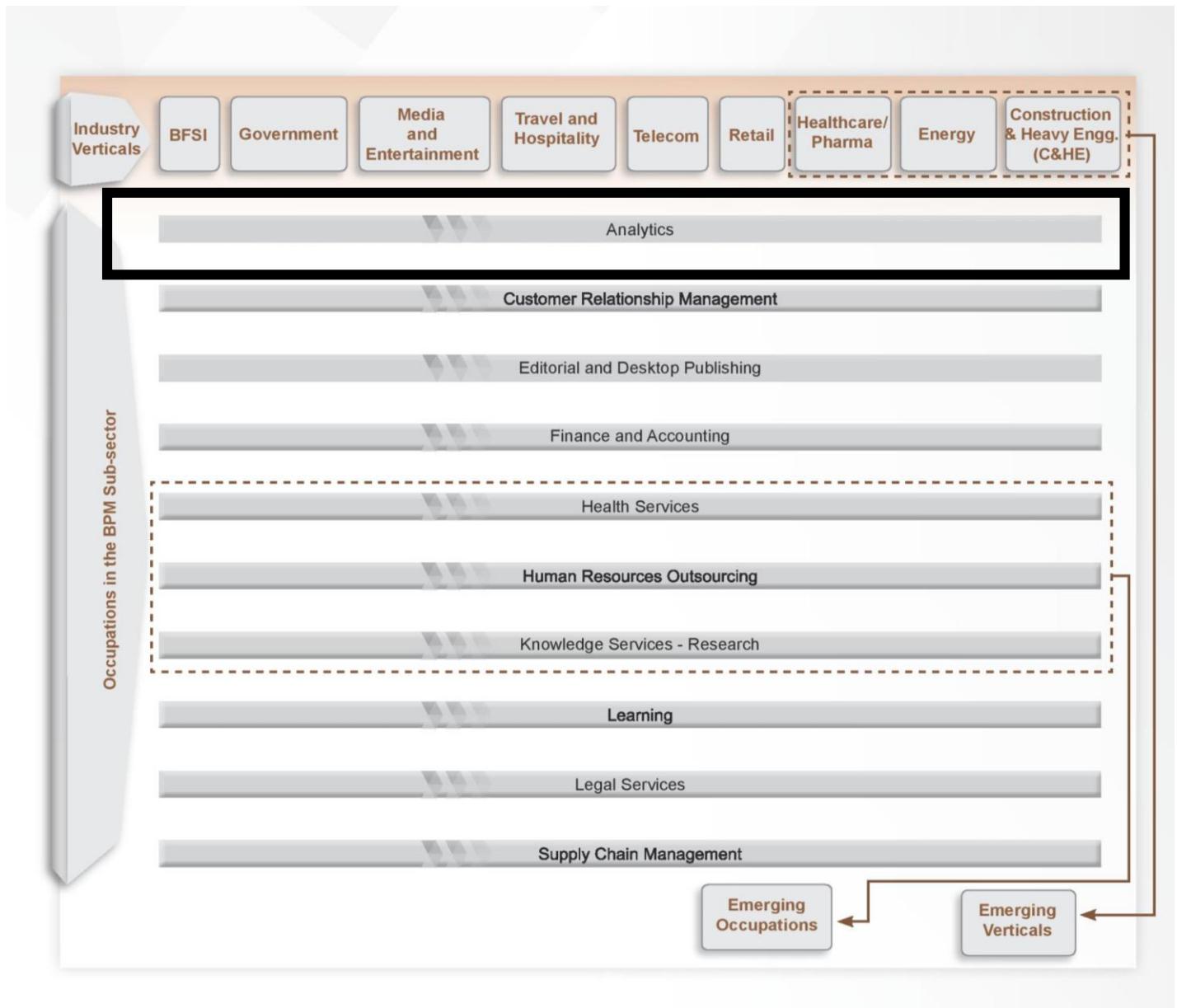
Brief Job Description: Individuals at this job are responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools

Personal Attributes: This job requires the individual to follow detailed instructions and procedures with an eye for detail. The individual should be analytical and result oriented and should demonstrate logical thinking.

Eligibility: Bachelor's Degree in Statistics/ Science/Technology, Master's Degree in Science/Technology/Statistics

Work Experience: 0-1 years of work experience/internship in analytics roles

Analytics is a key occupation in the structure of the ITS Sub-Sector



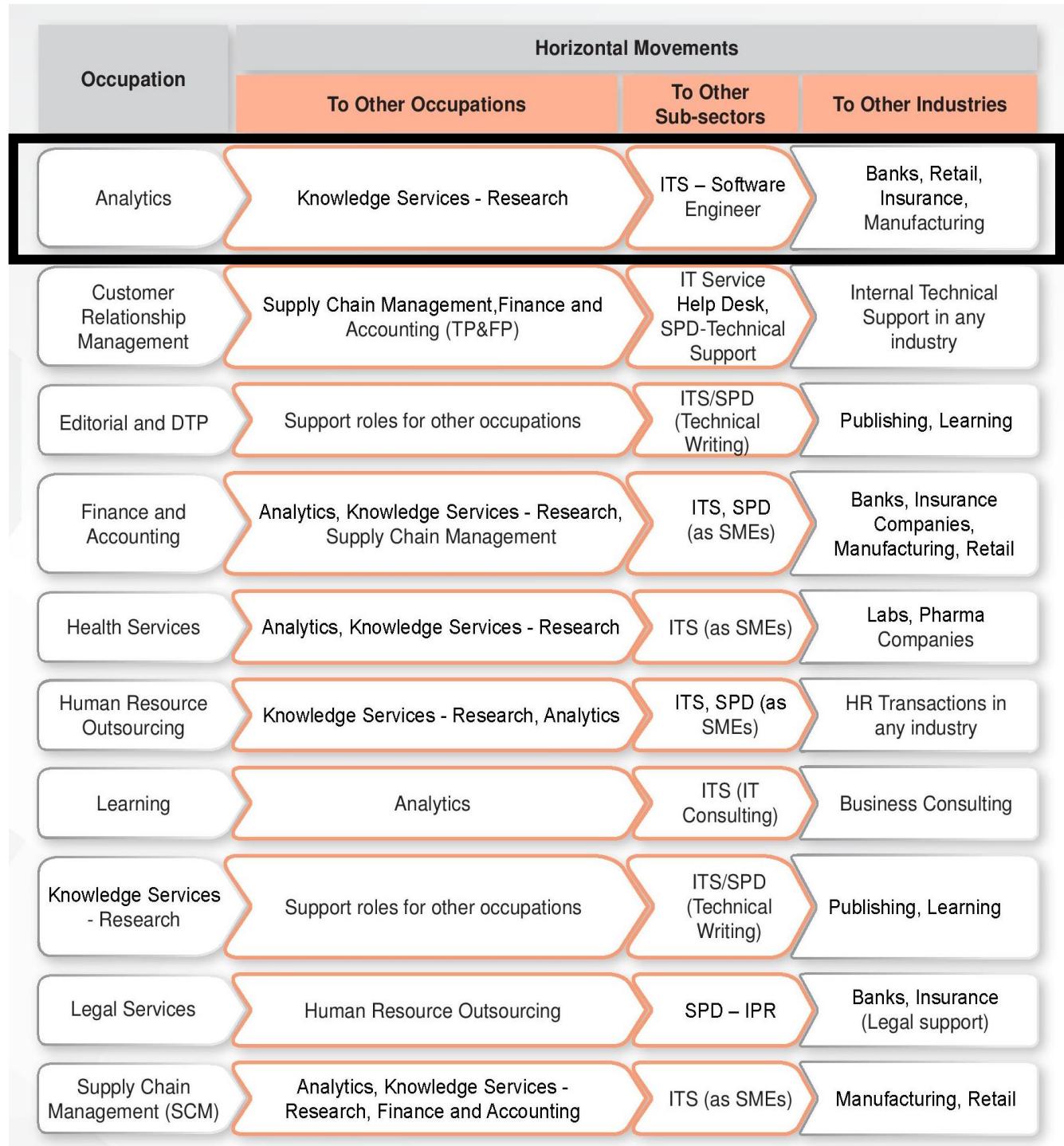
Analytics excellent Vertical and Horizontal movements in their

tracks

Occupation	Tracks	Entry-level Job Roles
Analytics	MIS - Reporting Analytics - Modelling and Analysis	Associate - Analytics
	Customer Care (Non-voice)	Associate - Customer Care (Non-voice)
Customer Relationship Management (CRM)	Customer Care (Voice)	Associate - CRM
	Sales/Telesales	
	Technical Support/IT Help Desk	
	Collections (Business to Customer)	
Editorial and Desktop Publishing (DTP)	Editorial	Associate - Editorial
	DTP and Design	Associate - DTP
Finance and Accounting (F&A)	Transaction Processing (Includes B2B Collections)	Associate - Transactional F&A
	Credit Analysis	Associate - F&A Complex
	Audit and Accounting	
	Financial Reporting	
	Financial Planning and Analysis (Includes Budgeting and Forecasting)	
Health Services	Clinical Data Management	Associate - Clinical Data Management
	Medical Transcription	Associate - Medical Transcription
Human Resource Outsourcing (HRO)	Recruitment	Associate - Recruitment
	Learning and Development	Associate - HRO
	Compensation and Benefits Management	
	Employee Relations	
Knowledge Services - Research	Secondary Research and Market Research	Analyst - Research
	Investment Banking Research	
Learning	Content Managemnet	Associate - Learning
	Instructional Design	
Legal Services	Legal Services	Document Coder/Processor Legal Associate
Supply Chain Management	Procurement Operations (Including Strategic Sourcing)	Associate - SCM
	Sales and Fulfilment (Including Inventory Mangement)	

Movement to Other Occupations, Sub-sectors and Industries:

Given the dynamic range of services that the BPM sub-sector is increasingly offering to its clients in the industry, there are a variety of roles that employees are performing across the entire spectrum of offerings. As such they become a valuable asset not only to the BPM sub-sector, but also to all the client industries they are associated with.



OVERALL QUALIFICATION PACK DETAILS

Job Details	Qualifications Pack Code	SSC/Q2101					
	Job Role	Associate - Analytics This job role is applicable in both national and international scenarios					
	Credits(NVEQF/NVQF/NSQF)		Version number	0.1			
	Sector	IT-ITeS	Drafted on	30/04/13			
	Sub-sector	Business Process Management	Last reviewed on	30/04/13			
	Occupation	Analytics	Next review date	30/06/14			
	KA1. Job Role	KA2. Associate - Analytics (Business Analytics Associate/ Analyst)					
KA3. Role Description	KA4. Responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools						
KA5. NVEQF/NVQF level KA6.	KA10. 7						
KA7. Minimum Educational Qualifications	KA11. Bachelor's Degree in Statistics/ Science/Technology or any other course						
KA8.	KA12. Master's Degree in Science/Technology/Statistics or any other course						
KA9. Maximum Educational Qualifications							
KA13. Training KA14. (Suggested but not mandatory)	KA15. Courses in SPSS, SAS, STATA and/or Spreadsheets KA16. RDBMS concepts, PL\SQL, OCA certification KA17. Financial and accounting terminologies in respective language & various accounting standards and GAAPs						
KA18. Experience KA19.	KA20. 0-1 years of work experience/internship in analytics roles						
KA21. Applicable National Occupational Standards (NOS)	KA22. Compulsory: 1. SSC/ N 0703 (Create documents for knowledge sharing) 2. SSC/ N 2101 (Carry out rule-based statistical analysis) 3. SSC/ N 9001 (Manage your work to meet requirements) 4. SSC/ N 9002 (Work effectively with colleagues) 5. SSC/ N 9003 (Maintain a healthy, safe and secure working environment) 6. SSC/ N 9004 (Provide data/information in standard formats) 7. SSC/ N 9005 (Develop your knowledge, skills and competence) KA23. KA24. Optional: KA25. Not Applicable						
KA26. Performance Criteria	KA27. As described in the relevant OS units						

SSC/ N0703 - Create Documents for knowledge sharing

Session Overview

In the Associate Analytics “Working with Documents”, the participant will learn about the most prominently used documentation techniques in corporate organizations. The Documentation types covered would include case studies, best practices, project artifacts, reports, minutes, policies, procedures, work instructions etc.

This session is NOT intended to cover technical documents or documents to support the deployment and use of products/applications, which are dealt with in different standards.

Session Goal

Participants should be able to have a good hands on understanding of MS Word and MS Visio, where there will be required to draft various documents/reports. The goal of the session is for the participant to be aware of the various documentation techniques which are used prominently in organizations.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish with **appropriate people** the purpose, scope, formats and target audience for the documents
- PC2. access existing documents, language standards, templates and documentation tools from your organization’s knowledge base
- PC3. liaise with **appropriate people** to obtain and verify the information required for the documents
- PC4. confirm the content and structure of the documents with **appropriate people**
- PC5. create documents using standard templates and agreed language standards
- PC6. review documents with **appropriate people** and incorporate their inputs
- PC7. submit documents for approval by **appropriate people**
- PC8. publish documents in agreed formats
- PC9. update your organization’s knowledge base with the documents
- PC10. comply with your organization’s policies, procedures and guidelines when creating documents for knowledge sharing

Note: The material for this NOS has been covered in the Associate Analytics Module 3 Book (book 3) in Unit 5

SSC/ N 2101 – Carry out rule-based statistical analysis

Session Overview

In the Associate Analytics *Carry out rule based statistical analysis*, the participants will go through Business Analytics using R tool. The participants will also learn Applied Statistical concepts like Descriptive Statistics and find their usage along with R. Furthermore, they will also have an overview of Big Data tools and their basic functioning.

Then they will learn about Machine Learning algorithm and their use in Data Mining and Predictive Analytics. Finally the participants will learn about Data Visualization and gather knowledge on Graphical representation of Data as well as results and reports.

Session Goal

The primary goal of the session is for the participants to learn the R tools and its various functions and features. Then also learn about Big Data tools and Big Data Analytics. Students will also learn about basic applied statistical concepts.

Session Objectives

To be competent, participants must be able to:

- PC1. establish clearly the objectives and scope of the **analysis**
- PC2. obtain guidance from **appropriate people** to identify suitable **data sources** to agree the methodological approach
- PC3. obtain and structure data using standard templates and tools
- PC4. validate data accurately and identify **anomalies**
- PC5. obtain guidance from **appropriate people** on how to handle **anomalies** in data
- PC6. carry out rule-based **analysis** of the data in line with the analysis plan
- PC7. validate the results of your **analysis** according to statistical guidelines
- PC8. review the results of your **analysis** with **appropriate people**
- PC9. undertake modifications to your **analysis** based on inputs from **appropriate people**
- PC10. draw justifiable inferences from your **analysis**
- PC11. present the results and inferences from your analysis using standard templates and tools
- PC12. comply with your organization's policies, procedures and guidelines when carrying out rule-based quantitative **analysis**

Note: The material for this NOS has been covered in all the three Modules of Associate Analytics

SSC/ N 9001: Manage Your Work to Meet Requirement

Session Overview

The Associate Analytics *Manage your work to meet requirement* module is designed to help participants understand the importance of time in a professional environment and how to manage multiple time bound requirements. It emphasizes on how time management is critical to work management and completing requirements/deliverables.

Participants learn how to manage work and how to ensure deliverables are completed in stipulated time in an organization by following tested principles to prevent/handle slippages on timelines. The module also emphasizes the need to respect time for self as well as colleagues.

Time management cannot override the qualitative aspect of the deliverable.

Session Goal

The primary goal of the session is for the participants to learn and manage time to be able to complete their work as required. The requirements of a work unit may be further classified into; activities, deliverable, quantity, standards and timelines. The session makes participants to be aware of defining requirements of every work unit and then ensuring delivery.

Additionally, this session discusses practical application of planning and execution of work plans to enable the participants to effectively deal with the failure points, minimize the impact, if any. Equally critical is the escalation plan and root cause analysis of exceptions.

Successful candidates will be able to understand the inter-relationship of time, effort, impact and cost.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Establish and agree your work requirements with appropriate people

PC2. Keep your immediate work area clean and tidy

PC3. Utilize your time effectively

PC4. Use resources correctly and efficiently

PC5. Treat confidential information correctly

PC6. Work in line with your organization's policies and procedures

PC7. Work within the limits of your job role

PC8. Obtain guidance from appropriate people, where necessary

PC9. Ensure your work meets the agreed requirements

Note: The material for this NOS has been covered in Unit 1 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9002: Work Effectively With Colleagues

Session Overview

The Associate Analytics *Work Effectively with Colleagues* module is designed to help participants understand the importance of teamwork in a professional environment. It emphasizes on how relationship management is critical to work management. It also focuses on the importance of personal grooming.

Participants learn how to manage cross functional relationships and how to nurture a good working environment. The module also stresses on the need to respect colleagues.

Session Goal

The primary goal of the session is for the participants to understand the importance of professional relationships with colleagues. Additionally, this session discusses importance of personal grooming.

Successful candidates will be able to understand the inter-relationship of professionalism and team-work.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Communicate with colleagues clearly, concisely and accurately.

PC2. Work with colleagues to integrate your work effectively with theirs.

PC3. Pass on essential information to colleagues in line with organizational requirements.

PC4. Work in ways that show respect for colleagues.

PC5. Carry out commitments you have made to colleagues.

PC6. Let colleagues know in good time if you cannot carry out your commitments, explaining the reasons.

PC7. Identify any problems you have working with colleagues and take the initiative to solve these problems.

PC8. Follow the organization's policies and procedures for working with colleagues.

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9003: Maintain a Healthy, Safe and Secure working Environment

Session Overview

The Associate Analytics *Health, Safety and Security* module is designed to help participants understand the importance of following safety rules and regulations at workplace.

Participants learn how to work safely in an organization by following guidelines to prevent/handle any accidents or emergencies. The module also emphasizes the need of security and the entities that can pose a threat to it.

Session Goal

The primary goal of the session is for the participants to be aware about the various hazards that they may come across at workplace and what are the defined health, safety and security measures that should be followed at the time of occurrence of such unpredictable events. Additionally, this session discusses practical application of the health and safety procedures to enable the participants to effectively deal with the hazardous events to minimize the impact, if any.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. Comply with your organization's current health, safety and security policies and procedures
- PC2. Report any identified breaches in health, safety, and security policies and procedures to the designated person
- PC3. Identify and correct any hazards that you can deal with safely, competently and within the limits of your authority
- PC4. Report any hazards that you are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected
- PC5. Follow your organization's emergency procedures promptly, calmly, and efficiently
- PC6. Identify and recommend opportunities for improving health, safety, and security to the designated person
- PC7. Complete any health and safety records legibly and accurately

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9004: Provide data/information in standard formats

Session Overview

The Associate Analytics *Provide data/information in standard formats* module is designed to help participants understand the standard operating procedures in organizations pertaining to reporting data in a logical sequence and arriving at conclusive decisions models after analysis of data. This module is aimed at developing the sense of understanding in an individual when the individual works with data, of how to take the data and present it as relevant information in standardized formats.

Participants learn how to share information with other people inside or outside a specified work group and also how to arrive at decisions regarding certain problem types.

Session Goal

The primary goal of the session is for the participants to analyze data and present it in a suitable format, as is suitable for the given process or organization.

Successful candidates will be able to understand the process of standardized reporting and the nuances of a publishing a report with a specified end objective in mind.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish and agree with appropriate people the data/information you need to provide, the formats in which you need to provide it, and when you need to provide it
- PC2. obtain the data/information from reliable sources
- PC3. check that the data/information is accurate, complete and up-to-date
- PC4. obtain advice or guidance from appropriate people where there are problems with the data/information
- PC5. carry out rule-based analysis of the data/information, if required
- PC6. insert the data/information into the agreed formats
- PC7. check the accuracy of your work, involving colleagues where required
- PC8. report any unresolved anomalies in the data/information to appropriate people
- PC9. provide complete, accurate and up-to-date data/information to the appropriate people in the required formats on time

SSC/ N 9005: Develop your knowledge, skills and competence

Session Overview

The Associate Analytics *develop your knowledge, skills and competence* module is designed to help participants understand the importance of skill development in a professional environment and how to enhance skills in order to excel. It emphasizes on how enhance skills and knowledge in a diversified professional environment.

Session Goal

The primary goal of the session is to give a overview on how skills and competency can be enhanced in a professional environment. It gives knowledge on organizational context, technical knowledge, core skills/geneic skills, professional skills and technical skills. The session makes participants to understand the need of skills improvement for personal and organizational growth.

Successful candidates will be able ro understand the relationship between skill enhancement and growth.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. obtain advice and guidance from appropriate people to develop their knowledge, skills and competence
- PC2. identify accurately the knowledge and skills you need for their job role
- PC3. identify accurately their current level of knowledge, skills and competence and any learning and development needs
- PC4. agree with appropriate people a plan of learning and development activities to address their learning needs
- PC5. undertake learning and development activities in line with their plan
- PC6. apply their new knowledge and skills in the workplace, under supervision
- PC7. obtain feedback from appropriate people on their knowledge and skills and how effectively they apply them
- PC8. review their knowledge, skills and competence regularly and take appropriate action

Overall Associate Analytics Content Structure

Module 1 – Book 1

Subject I / SSC NASSCOM - NOS- 2101, 9001, 9002	NOS	Hours	Minutes
Unit - 1	NOS 2101/9001		
Introduction to Analytics & R programing		6	360
Manage your work to meet requirements		4	240
Unit - 2	NOS 2101/9002		
Summarizing Data & Revisiting Probability		6	360
Work effectively with Colleagues		4	240
Unit - 3	NOS 2101		
SQL using R		9	510
Unit - 4	NOS 2101		
Correlation and Regression Analysis		9	510
Unit - 5	NOS 2101		
Understanding Verticals - Engg, Financial, others		6	390
Requirements Gathering		6	390
Total Hrs/Minutes		50	3000

Module 2 – Book 2

Subject II / SSC NASSCOM - NOS- 2010, 9003, 9004	NOS	Hours	Minutes
Unit - 1	NOS 2101/9003		
Data Management		7	420
Maintain Healthy, Safe & Secure Working environment		4	240
Unit - 2	NOS 2101/9004		
Big Data Tools		7	420
Provide Data/Information in Standard formats		4	240
Unit - 3	NOS 2101		
Big Data Analytics		8	480
Unit - 4	NOS 2101		
Machine Learning Algorithms		8	480
Unit - 5	NOS 2101		
Data Visualization		6	360
Analytics Application to various domains		6	360
Case Study			
Total Hrs/Minutes		50	3000

Module 3 – Book 3

Subject III / SSC NASSCOM - NOS - 0703, 2101, 9005	NOS	Hours	Minutes
Unit - 1	NOS 2101		
Introduction to Predictive Analytics		6	360
Linear Regression		6	360
Unit - 2	NOS 2101		
Logistics Regression		9	540
Unit - 3	NOS 2101/9005		
Objective Segmentation		6	360
Develop Knowledge Skill and competences		3	180
Unit - 4	NOS 2101		
Time Series Methods/Forecasting, Feature Extraction		5	300
Project		5	300
Unit - 5	NOS 0703		
Working with documents		10	600
Total Hrs/Minutes		50	3000

Glossary of Terms

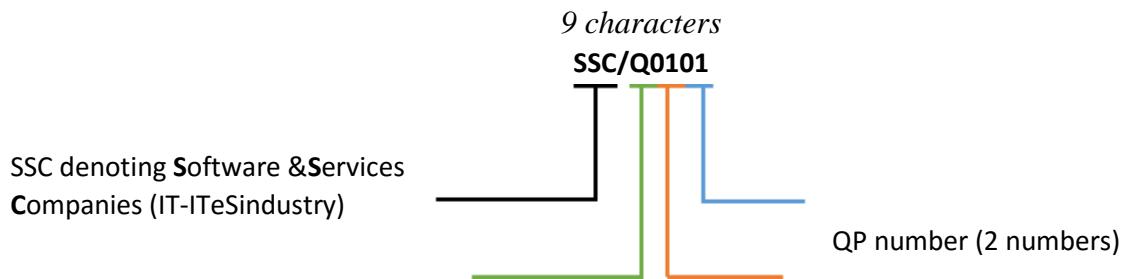
Definitions	Keywords /Terms	Description
	Sector	Sector is a conglomeration of different business operations having similar businesses and interests. It may also be defined as a distinct subset of the economy whose components share similar characteristics and interests.
	Sub-sector	Sub-sector is derived from a further breakdown based on the characteristics and interests of its components.
	Vertical	Vertical may exist within a sub-sector representing different domain areas or the client industries served by the industry.
	Occupation	Occupation is a set of job roles, which perform similar/related set of functions in an industry.
	Function	Function is an activity necessary for achieving the key purpose of the sector, occupation, or area of work, which can be carried out by a person or a group of persons. Functions are identified through functional analysis and form the basis of OS.
	Sub-functions	Sub-functions are sub-activities essential to fulfill the achieving the objectives of the function.
	Job role	Job role defines a unique set of functions that together form a unique employment opportunity in an organisation.
	Occupational Standards (OS)	OS specify the standards of performance an individual must achieve when carrying out a function in the workplace, together with the knowledge and understanding they need to meet that standard consistently. Occupational Standards are applicable both in the Indian and global contexts.
	Performance Criteria	Performance Criteria are statements that together specify the standard of performance required when carrying out a task.
	National Occupational Standards (NOS)	NOS are Occupational Standards which apply uniquely in the Indian context.
	Qualifications Pack Code	Qualifications Pack Code is a unique reference code that identifies a qualifications pack.
	Qualifications Pack(QP)	Qualifications Pack comprises the set of OS, together with the educational, training and other criteria required to perform a job role. A Qualifications Pack is assigned a unique qualification pack code.
	Unit Code	Unit Code is a unique identifier for an OS unit, which can be denoted with either an 'O' or an 'N'.
	Unit Title	Unit Title gives a clear overall statement about what the incumbent should be able to do.

Description	Description gives a short summary of the unit content. This would be helpful to anyone searching on a database to verify that this is the appropriate OS they are looking for.
Scope	Scope is the set of statements specifying the range of variables that an individual may have to deal with in carrying out the function which have a critical impact on the quality of performance required.
Knowledge and Understanding	Knowledge and Understanding are statements which together specify the technical, generic, professional and organisational specific knowledge that an individual needs in order to perform to the required standard.
Organisational Context	Organisational Context includes the way the organisation is structured and how it operates, including the extent of operative knowledge managers have of their relevant areas of responsibility.
Technical Knowledge	Technical Knowledge is the specific knowledge needed to accomplish specific designated responsibilities.
Core Skills/Generic Skills	Core Skills or Generic Skills are a group of skills that are key to learning and working in today's world. These skills are typically needed in any work environment. In the context of the OS, these include communication related skills that are applicable to most job roles.
Helpdesk	Helpdesk is an entity to which the customers will report their IT problems. IT Service Helpdesk Attendant is responsible for managing the helpdesk.
Keywords /Terms	Description
IT-ITeS	Information Technology - Information Technology enabled Services
BPM	Business Process Management
BPO	Business Process Outsourcing
KPO	Knowledge Process Outsourcing
LPO	Legal Process Outsourcing
IPO	Information Process Outsourcing
BCA	Bachelor of Computer Applications
B.Sc.	Bachelor of Science
OS	Occupational Standard(s)
NOS	National Occupational Standard(s)
QP	Qualifications Pack
UGC	University Grants Commission
MHRD	Ministry of Human Resource Development
MoLE	Ministry of Labour and Employment
NVEQF	National Vocational Education Qualifications Framework
NVQF	National Vocational Qualifications Framework
NSQF	National Skill Qualification Framework

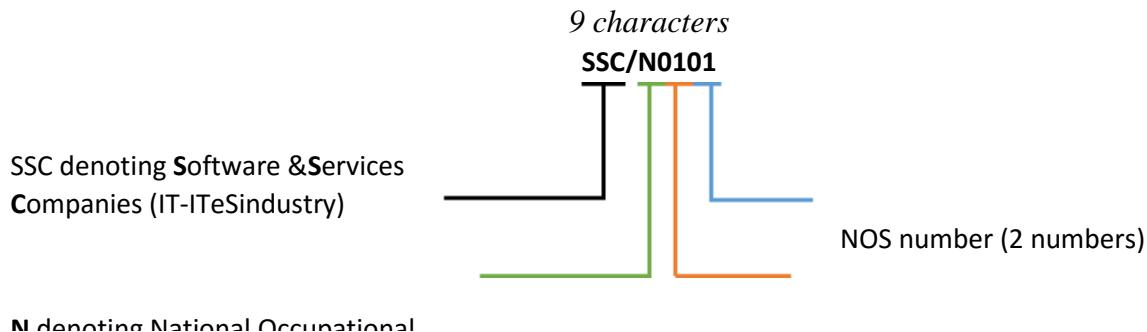
Nomenclature for QP & NOS

UNITS

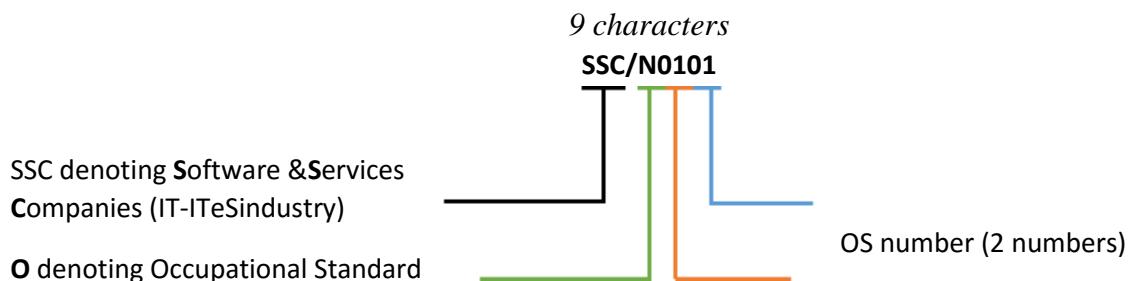
Qualifications Pack



National Occupational Standard



Occupational Standard



It is important to note that an OS unit can be denoted with either an 'O' or an 'N'.

- If an OS unit denotes 'O', it is an OS unit that is an international standard. An example of OS unit denoting 'O' is **SSC/O0101**.
- If an OS unit denotes 'N', it is an OS unit that is a national standard and is applicable only for the Indian IT-ITeS industry. An example of OS unit denoting 'N' is **SSC/N0101**

The following acronyms/codes have been used in the nomenclature above:

Sub-Sector	Range of Occupation numbers
IT Service (ITS)	01-20
Business Process Management (BPM)	21-40
Engg. and R&D (ERD)	41-60
Software Products (SPD)	61-80

Sequence	Description	Example
Three letters	Industry name (Software & Service Companies)	SSC
Slash	/	/
Next letter	Whether QP or NOS	N
Next two numbers	Occupation Code	01
Next two numbers	OS number	01

Module 2: Unit 1.1

Data Management

Topic	Activities
Data Management	<p>By the end of this session, you will be able to:</p> <ul style="list-style-type: none"> 1. Design Data Architecture 2. Understand various Data Sources 3. Export Data to Amazon S3

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Design Data Architecture and manage the Data for analysis	Classroom
Understand various sources of the Data.	Classroom
Export all the Data onto the cloud	Classroom
Check your understanding	Classroom
Summary	Classroom

Step-by-Step

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 an 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 doses 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.

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

Check your understanding



1. Quote 5 practical examples of various types of Data Sources?
2. What are the steps involved in Export job process In Amazon S3?
3. What are other cloud storage vendors?
4. Which source of Data is more reliable and Why?

Summary

- 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.
- The sources of generating primary data are -
 1. Observation Method
 2. Survey Method
 3. Experimental Method
- We usually export our data to cloud for purposes like safety, multiple access and real time simultaneous analysis.

Module 2: Unit- 1.2

Maintain a Healthy, Safe and Secure Working Environment

Topic	Activities
Maintain a Healthy, Safe and Secure Working Environment	<p>By the end of this session, you will be able to learn about:</p> <ol style="list-style-type: none"> 1. Workplace safety 2. Reporting accidents and emergencies 3. Protecting health and safety as you work
Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Classroom Session Map

Topic description	Location
<ul style="list-style-type: none"> ✓ Welcome participants to the course ✓ Introduce facilitators ✓ Recap of core skills through questions and Polling Questions ✓ Review learning objectives 	✓ Classroom
<ul style="list-style-type: none"> ✓ Discuss the significance of work place safety ✓ Create awareness on basic safety guidelines ✓ Summarize the appropriate discussion points from the breakout sessions 	✓ Classroom
<ul style="list-style-type: none"> ✓ Discuss accidents and emergencies and how to identify one. ✓ How to address risks and threats and handle accidents ✓ Create awareness around how to handle general emergencies 	✓ Classroom

- | | |
|--|-------------|
| <ul style="list-style-type: none">✓ Develop understanding on the potential health and safety hazards found at work place✓ Create awareness on the common safety signs used at workplace | ✓ Classroom |
|--|-------------|

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided: reflect on your own experiences and determine when to share them.**
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.**
- ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.**
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussion questions independently, rather than providing answers. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Session : 1 - Welcome and Introduction

Topic:Welcome and Introduction

Health, Safety and Security

Welcome the participants to the course and move to the introductions.

Introductions



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

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

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]



“Why are you here today?”

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

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

Suggested Responses/Benefits to Debrief:

The benefits of this course include:

- Impact of workplace disasters and need for workplace safety
- Clear understanding of the basic guidelines to be followed in the event of a risk or a hazardous event
- Awareness of the common security threats and risks and actions to be taken to address them.

Review the course objectives listed above.

“To fulfill these objectives today, we’ll be conducting a number of hands-on activities. Hopefully we can open up some good conversations and some of you can share your

Topic:Welcome and Introduction

experiences so that we can make this session as interactive as possible. Your participation will be crucial to your learning experience and that of your peers here in the session today.”

Knowledge Check Question 1

“Please answer the following question.” Discuss and debrief the correct answer

Question: What are some of the hazardous events that may happen at your workplace?



- A. Fire break-out (fire accident)
- B. Terrorist/ Bomb threat
- C. Tripping accidents
- D. All of the above

Answer – All of the above – each of the situation A,B,C can lead to an hazardous event and needs to have a safely plan in place to mitigate and manage the risk if it occurs.

Session: 2– Workplace Safety

Key Points

Let's Get Started

Importance of prevention of disasters/ risk events



Provide a brief overview of the session. Discuss the importance of prevention of disasters than be sorry after the event.

Open up the discussion for the session and ask participants to share their thoughts on “workplace safety”?

The first part of this session discusses the following:

- “Prepare and prevent, don't repair and repent.”
- It's better to prevent disasters from happening than be sorry and suffer after the accident.
- It is important to follow safety rules in any office and as future employees, you should know about these safety rules.

Why Workplace Safety



Ask the question to the participants and gather responses.

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

1. Refer to the **Workplace Safety Rules table** in the material later and identify the rules that employees/workers must follow.
2. Refer to the **Vocabulary Words** table in the material later if you do not understand the meaning of a word/term.

Key Points

Suggested Responses:

- Safety rules in the workplace protect workers from injury or death.
- These rules teach workers how to work safely – use rules in tables as outlined later to discuss .

Basic Workplace Safety Guidelines



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

➤ **Fire Safety**

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

➤ **Falls and Slips**

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

➤ **First Aid**

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

➤ **Security**

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

➤ **Electrical Safety**

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

Key Points

Check Your Understanding



1. True or False? The employer and employees are responsible for workplace safety.
 - a. True
 - b. False

Suggested Responses:

Yes – It is the joint responsibility of both employer and employees to ensure that the workplace is safe and secure.



2. True or False? Any injury at work should be reported to the supervisor immediately.
 - a. True
 - b. False

Suggested Responses:

True, always keep the management informed on any potential injury or health, Safety and Security events or risks noticed in an organization.



3. True or False? No matter how big or small the injury; the injured person should receive medical attention.
 - a. True
 - b. False

Suggested Responses:

True, No matter what the size of the injury – it is critical that medical help is sought. Sometimes physical injury may be minimal but internal injury cannot be assessed, which could be critical.

Key Points



4. True or False? While working with machines and equipment, employees must follow the safety guidelines set by the company.
- c. True
 - d. False

Suggested Responses:

True, all guidelines set by the company takes into account the potentials risks to the employees and measures to encounter those risks. While there is temptations to indentify shortcuts or alternative ways of working with machines and equipment, the pre-defined protocol should not be changed without proper authorization by requisite experts.



5. True or False? At any office, the first-aid kit should always be available for use in an emergency.
- a. True
 - b. False

Suggested Responses:

True, at times when a medical emergency hits – the medical aid could take time to reach. First aid kits can provide relief in the interim and prevent increased risks.



6. True or False? It is optional to participate in the random fire drills conducted by the Offices from time-to-time.
- a. True
 - b. False

Suggested Responses:

False, fire drills are critical activities that everyone should participate unless permissions have been taken prior. When emergency hits knowing the process to follow is very critical to provide safety support to the employees. It is always better to be prepared for such situations.

7. True or False? The "Wet Floor" sign is not needed and causes problems for people. Wet floor can be identified easily, without the signs.
- a. True
 - b. False

Key Points**Suggested Responses:**

False, Wet floors are hazards waiting to happen unless clearly marked. It is not very easy from distance to know if a surface is wet. Slips can be fatal and should be avoided with proper signage on the floor when they are wet



8. True or False? It is okay to place heavy and light items on the same shelf.
- True
 - False

Suggested Responses:

False, Heavy and light items should be clearly demarcated. If heavy items are placed by error in light items they can lead to breakage and other related accidents. Further limits of weight for each shelf should be defined so that it does not exceed acceptable limits.



9. True or False? There is no need to train employees on how to use the fire extinguisher. They can operate extinguishers following the instruction written on the extinguisher case, when needed.
- True
 - False

Suggested Responses:

False, In case of situations like fire panic is created and there will not be enough time to react. Reading instructions will be a challenge leading to more disaster. Being prepared knowing how to operate will enable employees to react fast and prevent further damage that could be caused due to fire

10. True or False? The cleaning supplies, especially chemical products, can be left in the bathrooms or in any of the cupboards in the office.

- True
- False

Key Points



Suggested Responses:

False, Cleaning supplies and other chemical products should be kept safe and secure with only authorized staff. If consumed by error can cause harmful impact.

Create Your Own Checklist

Activity Description:

Based on what you have learnt, create safety checklists for yourself.

These checklists will be discussed in the next session.



Summary

- It is important to follow safety rules to prevent accidents and protect workers.
- Employees must follow safety guidelines for the following:
 - Fire safety
 - Falls and slips
 - Electrical safety
 - Use of first aid

Key Points

Case studies of hazardous events

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

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

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

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

Module 3 - Unit: 1.2

Session: 3- Report Accidents and Emergencies

Key Points

Accidents and Emergencies



Ask participants to define accidents and emergencies.

Gather responses.

Start the session by connecting the course content to the candidate responses.

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.

Key Points

Get help promptly and in the most suitable way: Follow the procedure for handling a particular type of accident and emergency. Promptly act as per the guidelines. Ensure that you provide the required help and support as laid down in the policies. Do not act outside the guidelines and policies laid down for your role even if your actions are motivated by the best intention. Remember that only properly trained and certified professionals may be authorized to take decisions beyond the organization's policies and guidelines, if the situation requires.



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

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

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

Types of Accidents

The following are some of commonly occurring accidents in organizations:

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

Key Points

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.

Key Points



Activity Description:

1. Refer to the **Workplace Safety Rules table** in the Student Workbook and identify the rules that employees/workers must follow.
2. Refer to the **Vocabulary Words** table if you do not understand the meaning of a word/term.

Workplace Safety Rules

#	Workplace Safety Rules	Followed by workers	Followed by employers
1	Keep the floor dry all the time.		
2	Regularly check safety equipment such as fire extinguishers to make sure they are in working condition.		
3	Mark fire exit doors clearly.		
4	Know where fire extinguishers and fire alarms are kept.		
5	Conduct mock drills regularly.		
6	Find out the fire escape routes in a building.		
7	Keep first-aid kits where they can be easily found.		
8	Make sure that first-aid kits are stocked with all necessary things.		
9	Check and service all electrical equipment regularly.		
10	Repair faulty machinery immediately.		

Key Points

11	Make sure there is proper lighting in all areas.		
12	Make sure that the office layout and furniture are designed and arranged so that they do not cause injury to workers.		

Vocabulary Words

Mock Drill/Fire Drill

Practice how to respond/react in case of an emergency, such as a fire

Fire Extinguisher

A small container usually filled with special chemicals for putting out a fire.

Exit

The way to go out of a building or room

First Aid Kit

A container, which has medicines and ointments

Fire Escape Route

The way out in case of a fire

Emergency

A sudden, urgent and unexpected event

Spilt Liquid

Soft drink/water/coffee/tea etc. that has fallen on the floor

Routine inspections –

Regular checking

Damaged equipment

Torn wires or broken plugs

Stairways

Staircase/ stairs to go to the next floor

Key Points**Light fixtures**

Bulbs, tube lights etc.

Injury

Getting hurt/bleeding

Kitchen equipment

Vessels used in the kitchen, such as wok, knives, cutting board etc.

Cleaning Supplies

Liquid soap, dish washing liquid etc.

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.

Key Points



Activity Description:

1. Present a scenario where there is a physical injury to a colleague at workplace.
2. Ask participants on how they would react/attend to the emergency.

Types of Emergencies



- Discuss the various types of emergencies that one may come across at the workplace.
- Share some examples.

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

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

Substance emergencies, such as fire, chemical spills, and explosions:

Substance emergency is an unfavourable situation caused by a toxic, hazardous, or inflammable substance that has the capability of doing mass scale damage to properties and people.



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

Key Points

Security emergencies, such as armed robberies, intruders, and mob attacks or civil disorder:

Security emergency is an unfavourable situation caused by a breach in security posing a significant danger to life and property.

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

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.

Key Points

- Regularly check that all emergency handling equipments 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.

Check Your Understanding

1. True or False? An accident is a serious or crisis situation that needs immediate attention and action.



- a. True
- b. False

Suggested Responses:

True, you need to attend to the person immediately, inform to the supervisor and assist the supervisor

2. Which of the following are appropriate actions for handling accidents and emergencies?
Select the two correct actions.



a)

You should give medical treatment or first aid to the injured even if you are not properly trained in such procedures because such treatments should be given promptly.

b)

Take decisions beyond the organization's policies and guidelines, if the situation requires.

c)

Get help promptly and in the most suitable way.

d)

Follow instructions given by senior staff and the emergency services.

Suggested Responses:

Key Points

c and d – You should provide first aid only if you are qualified to do such treatments. While attending the accident or emergencies it is critical that all policy and guidelines needs to be adhered to.

3. Match each type of emergency with its corresponding example.



Type of Emergency	Example
A. Medical	iv. An expectant mother in labor
B. Substance	v. Chemical spills
C. Structural	ii. Power failure
D. Security	iii. Armed robbery
E. Natural Disaster	i. Earthquake

Type of Emergency	Example
A. Medical	i. Earthquake
B. Substance	ii. Power failure
C. Structural	iii. Armed robbery
D. Security	iv. An expectant mother in labor
E. Natural Disaster	v. Chemical spills

**Suggested
Responses:**

Key Points

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.

Session: 4 – Protect Health & Safety as You Work

Key Points

Let's Get Started



Provide a brief overview of the session. Discuss the key points/guidelines to protect health and safety as you work.

- Each year, an estimated 2 million people die because of occupational accidents and work-related diseases.
- Across the globe, there are almost 270 million occupational accidents and 160 million work-related diseases each year.

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.



Key Points

- 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 rotating objects can cause serious harm.

Potential Sources of Hazards in an Organization



Ask participants to come up with examples of sources/items/places that can possibly be the root cause of the hazard.

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

Key Points

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

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



After discussing the course content, ask candidates to prompt the key points on their understanding of the evacuation procedures at their current organization.

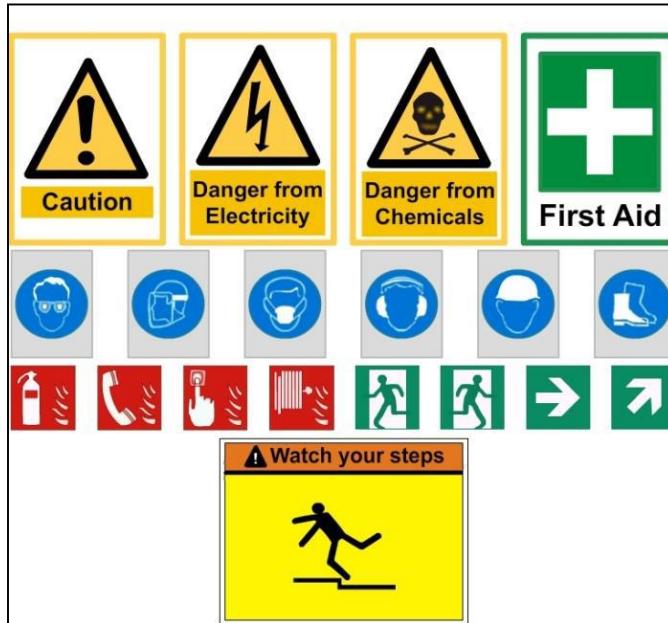
Safety Signs



Some of the common safety signs are given below. Note down the labels for each sign.

Key Points

Discuss and check the participants understanding of the various safety signs given in the picture above.



Review: Safety Guidelines Checklist

Key Points

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

Find the Problem

Key Points



In this activity, you will be shown some pictures. Observe the displayed pictures carefully and identify the problems in each of the pictures that could cause accidents.

Situation

Picture 1



Possible Answers:

There are many sources of hazard in the picture among others discuss about – painter on ladder without any support being provided, people walking on the stairs with wet paint on the floor, person picking a heavy box without any support, ladder too close and in the way of the stairs.

Key Points

Situation

Picture 2



Possible Answers:

There are many sources of hazard in the picture among others discuss about – Women climbing on the chair and handling electric equipment. Heater on the floor could be a cause of fire, Computer wires hanging out could easily get entangled in legs and make people fall, there is material on the floor on the mat where a person could trip, coffee kettle on a tall drawer chest could

Key Points

Situation

Picture 3



Possible Answers:

There are many sources of hazard in the picture among others discuss about:

- Hard hat area but staff not wearing hat
- Water spilled on the floor
- Smoking in No smoking zone
- Picture frame on Health & Safety instructions

Key Points

Healthy Living

What constitutes healthy living?



Eating a balanced diet: A balanced diet is a meal that provides you the right amount of carbohydrate, fat, protein, vitamins, and minerals. A balanced diet helps to keep you physically fit and provides stamina to work.

Having proper sleep: Good sleep reduces stress, reduces risk for developing diseases, and keeps you alert. You need to get 6 or 7 hours of sleep each night. Lack of sleep increases the chances of high blood pressure and cholesterol, and stroke.

Exercising regularly: Exercise is a physical activity that keeps your body fit. Exercising helps prevent development of disease conditions and makes you energetic.

Avoiding bad habits, such as smoking and drinking: It's not too late to identify and change bad habits such as smoking, drinking, over-eating, and more. Understanding the harmful routines is the first step to reversing these. The next step is realizing ways to correct them and embracing new ones, which help adopt healthier behaviours and start living a happier, healthier life.

Key Points

Ergonomics: Ergonomics is the science concerned with designing and arranging things so that people can use them easily and safely. Applying ergonomics can reduce the potential for accidents, potential for injury and ill health, and improve performance and productivity.

Activity Description:

1. Make groups of 4-5.
2. Ask participants to discuss within group – and present their thoughts on “healthy living”.

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.

Key Points

Module 2 - Unit: 2.1

Big Data tools

Topic	Activities
Big Data Tools	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Know the basics of Big Data Tools. 2. Understand gaps in data.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Introduction to the Big Data tools like spark, Scala, Impala	Classroom
Identify gaps in the data and follow-up for decision making	Classroom
Check your understanding	Classroom
Summary	Classroom

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

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

1.Apache Spark :-

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

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



2.Scala:-

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

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

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, RCFFile, 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: $X \perp\!\!\!\perp R_y | (R_x, Z)$.

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 $R_x = 0$ and $R_y = 0$ 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. R_y), conditional on R_x can be submitted to the following refutation 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.

Check your understanding



1. What is the difference between Spark, Scala and Impala?
2. What is the platform of Scala?
3. What are the benefits of using Spark, Scala and Impala?

Summary

- 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.
- Cloudera Impala is Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop.
- Scala is a programming language for general software applications. Scala has full support for functional programming and a very strong static type system.

Module 2 - Unit: 2.2

Provide data/information in standard formats

Topic	Activities
Provide data/information in standard formats	<p>By the end of this session, you will be able to learn about</p> <ol style="list-style-type: none"> 1. Knowledge Management 2. Standardized reporting and compliance 3. Decision Models

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Topic description	Location
✓ Welcome participants to the course ✓ Recap of core skills through questions and Polling Questions	✓ Classroom
✓ What is knowledge management ✓ Emerging trends in Product Development and Knowledge management ✓ KM approach of a few organizations	✓ Classroom
✓ Standard Reporting templates and whitepapers ✓ Organizing data/information ✓ Policies and procedures for recording and sharing information ✓ Importance of Compliance	✓ Classroom
✓ Deciding how to decide ✓ The Vroom-Yetton-Jago Decision Model	✓ Classroom

- | | |
|---|--|
| <ul style="list-style-type: none">✓ The Kepner-Tregoe Matrix✓ OODA Loops | |
|---|--|

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided: reflect on your own experiences and determine when to share them.**
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.**
- ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.**
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussion questions independently, rather than providing answer. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Step by Step

Topic: Knowledge Management

Knowledge Management

Welcome the participants to the course and move to the introductions.

Introductions



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

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

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]



“Why are you here today?”

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

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

Debrief the following:

Why is knowledge management so important?

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

Key Points

Let's Get Started

Importance of knowledge management



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

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

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

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

Why Knowledge Management



Ask the question to the participants and gather responses.

Discuss the responses with the group to understand the significance of knowledge management

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



Key Points



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

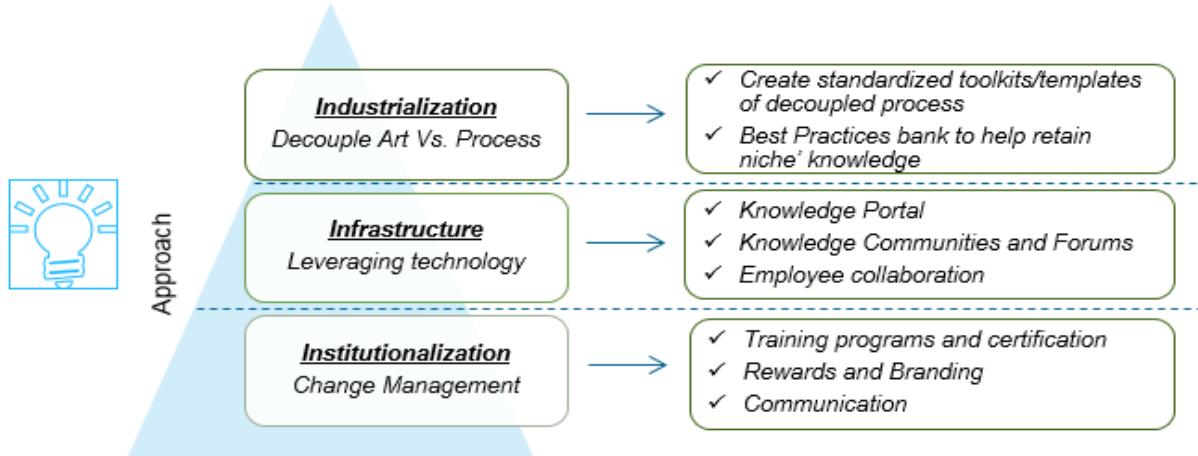


Fig : Knowledge Management Approach

Emerging Trends in Product Development & Knowledge Management Process

Key Points	
Trends	Examples/Implications
Crowdsourcing and social media playing a big role in collaborative product development	<ul style="list-style-type: none"> Fiat Mio is the world's first crowd-sourced car, developed based on new ideas sourced from users via a social media platform. While this trend will change the way products will be designed in the future, organizations will also have to collate, classify, and assess large amount of data, and feed it back into the product development process.
Big data analytics close looping with knowledge base	<ul style="list-style-type: none"> Many product development organizations, including auto OEMs, are critically analyzing Big Data (product failure data, service data, warranty data, historical design data, materials data etc.) to extract information patterns CAD/CAE/PLM technologies now have capabilities to integrate various extracted knowledge elements into the product design process.
Internal and external collaboration for knowledge sharing	<ul style="list-style-type: none"> Many leading corporates in North America and Europe have developed collaboration platforms with features including blogs, wikis, chats, communities of practice, expert corners and alerts among others to allow employees to share knowledge through discussion threads, idea voting etc. While collaboration platforms are becoming extremely popular within organizations, in some cases they are also being extended to external users with appropriate security features. Organizations will need to have business processes and supporting infrastructure in place to enable such collaboration platforms and process the information shared on these platforms.
Use of telematics in assimilating knowledge for product development	<ul style="list-style-type: none"> Sensing devices integrated with remote monitoring devices are being used to gather intelligence on product performance, health diagnostics, usage patterns etc. Manufacturing organizations must work on integrating the knowledge and data generated in this way into the core product development process to enable faster and right decision making.

Key Points

Baking knowledge into the product development process

- Manufacturing organizations the world over are realizing that tacit knowledge available among current employees should be formalized and captured for posterity.
- Global auto OEMs are gearing up to put business processes and the required infrastructure in place to capture this tacit knowledge in terms of best practices/lessons learned etc., and close-loop the knowledge management process by integrating these best practices with CAD/PLM tools.

Activity

Activity Description:

Make groups of 3-5 people and ask them to discuss and come up with ideas on how the Indian Census board can device good Knowledge Management practices

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.

Key Points

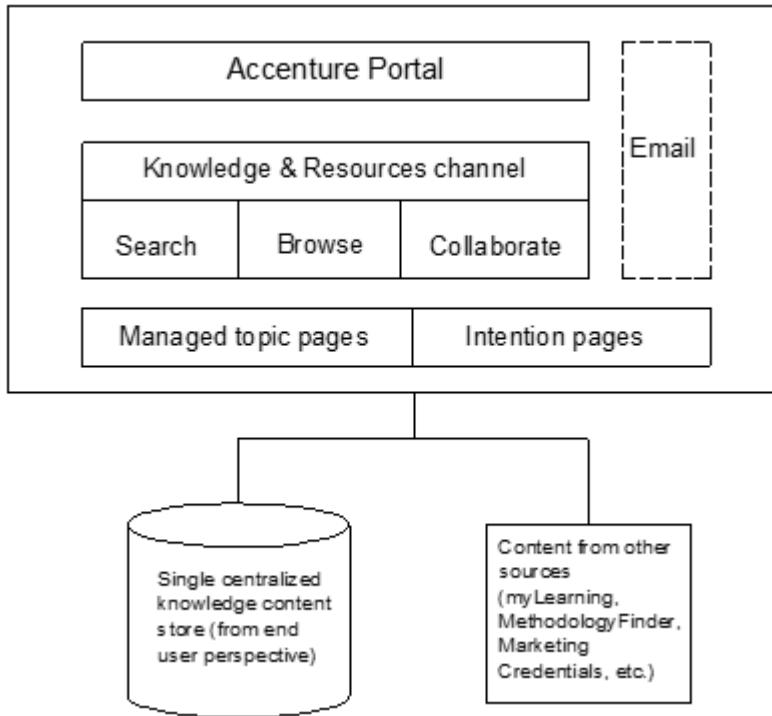


Fig : Accenture Knowledge Management solution structure

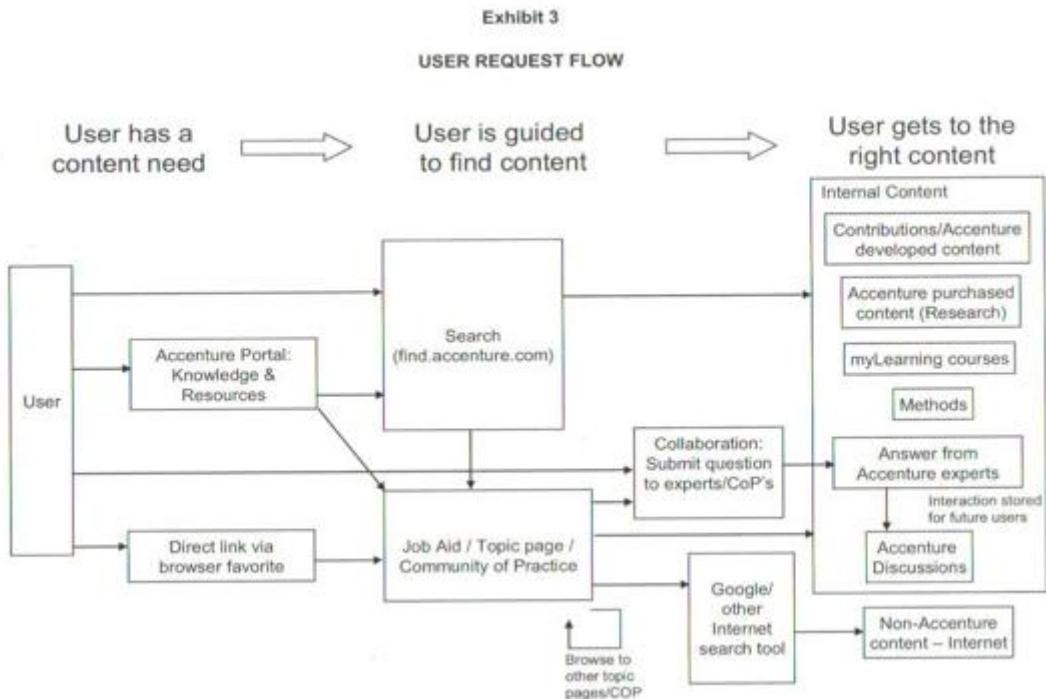


Fig : Accenture Knowledge Management report extraction

Key Points

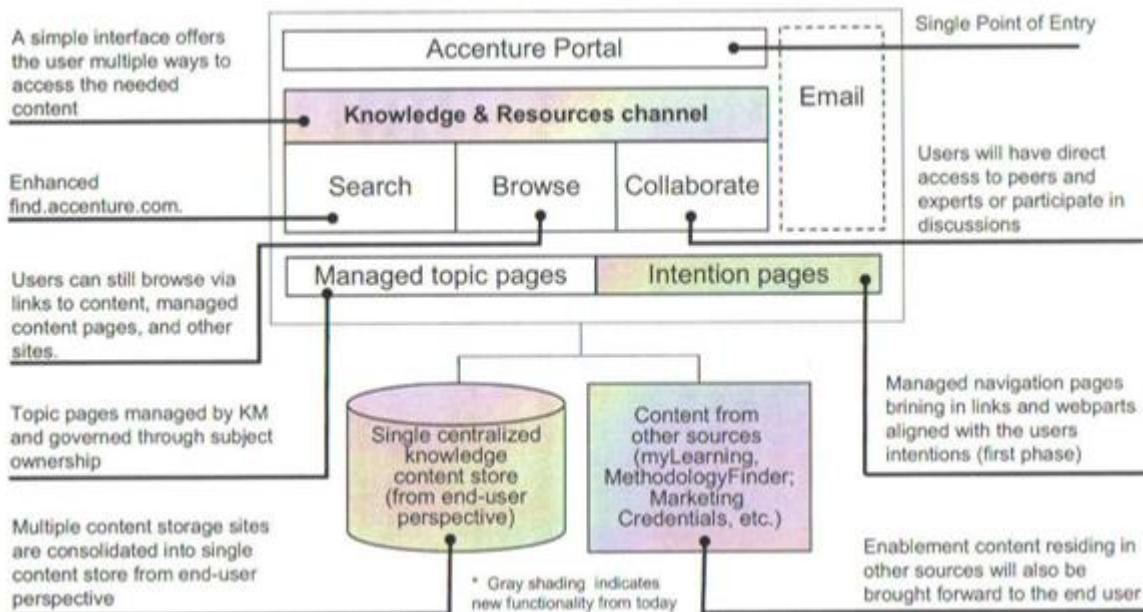


Fig : Accenture Knowledge Management Portal

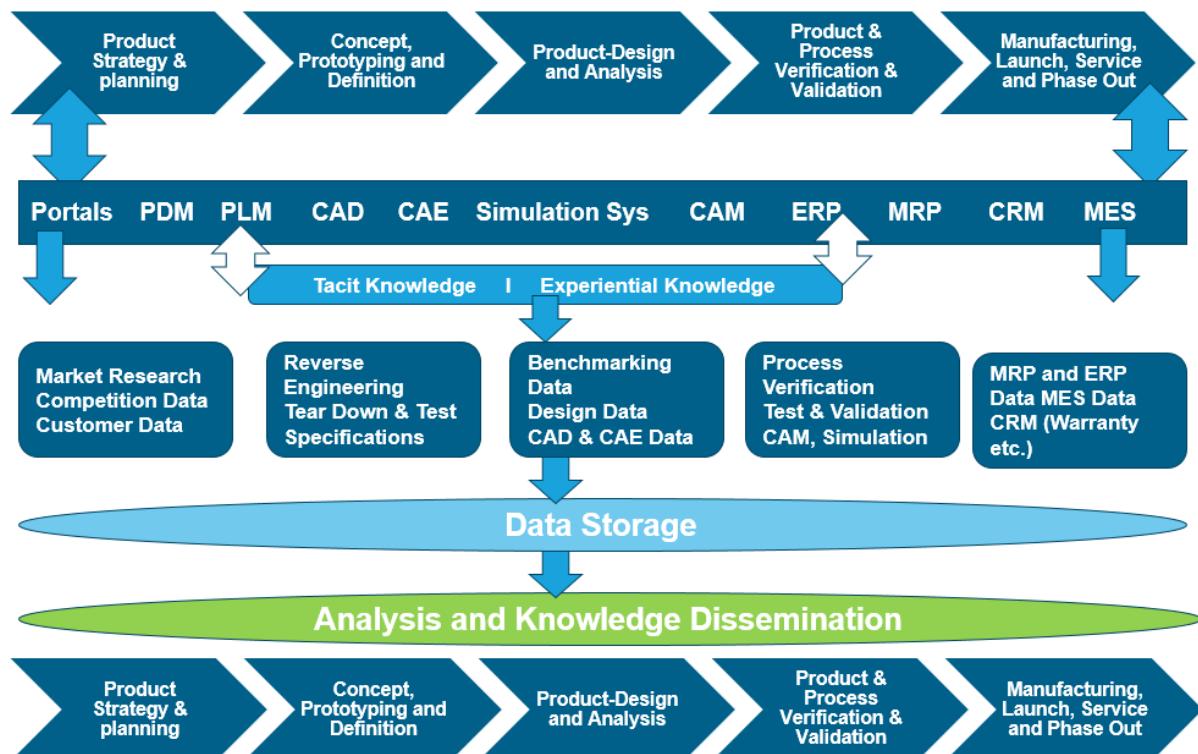


Fig : TCS Enterprise View of Big Data Sources & Knowledge Management

Key Points

Annexure of terms used:

PDM- Product development management

PLM- Product life management

CAD –Computer aided design

CAE- Computer aided engineering

CAM- Computer aided manufacturing

ERP – enterprise resource planning

MRP – Material requirements planning

CRM- Customer relationship management

MES- Manufacturing execution systems



Take the participants through the various knowledge management solution platforms for major organizations and facilitate an interactive discussion with them on the same

Check your Understanding



1. True or False? Knowledge Management is an evolving process and does not need to adhere to stringent rules.
c. True
d. False

Suggested Responses:

True, but this is a trick question. The knowledge management needs to be done within a specified framework for an organization, but is still an evolving process.



2. True or False? It is not important to maintain adequate documentation for a new organization process since documentation is a very time consuming process.
a. True
b. False

Suggested Responses:

Key Points

False, without the right documentation, nobody can ever have end to end visibility of the entire process. This is critical because often employees move in and out of a particular role and adequate documentation help in the right transitioning.

Create Your Own Knowledge Management Framework

Activity Description:

Based on what you have learnt, create a Knowledge Management framework for a particular kind of analytical report to ensure that a smooth transition takes place to the new joinee when you move out of the reporting role.

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

Session:3– Standardized Reporting and compliance

Key Points



Let's Get Started

Provide a brief overview of the session. Discuss the key points of a creating template based reports.

- One of the biggest mistakes people do when they join a corporate organization, is that they try to reinvent the wheel. This is however, quite needless since most organizations use standardized templates for reporting.
- These standard templates help in saving a lot of time and efforts in getting standardized reports out.

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

Key Points

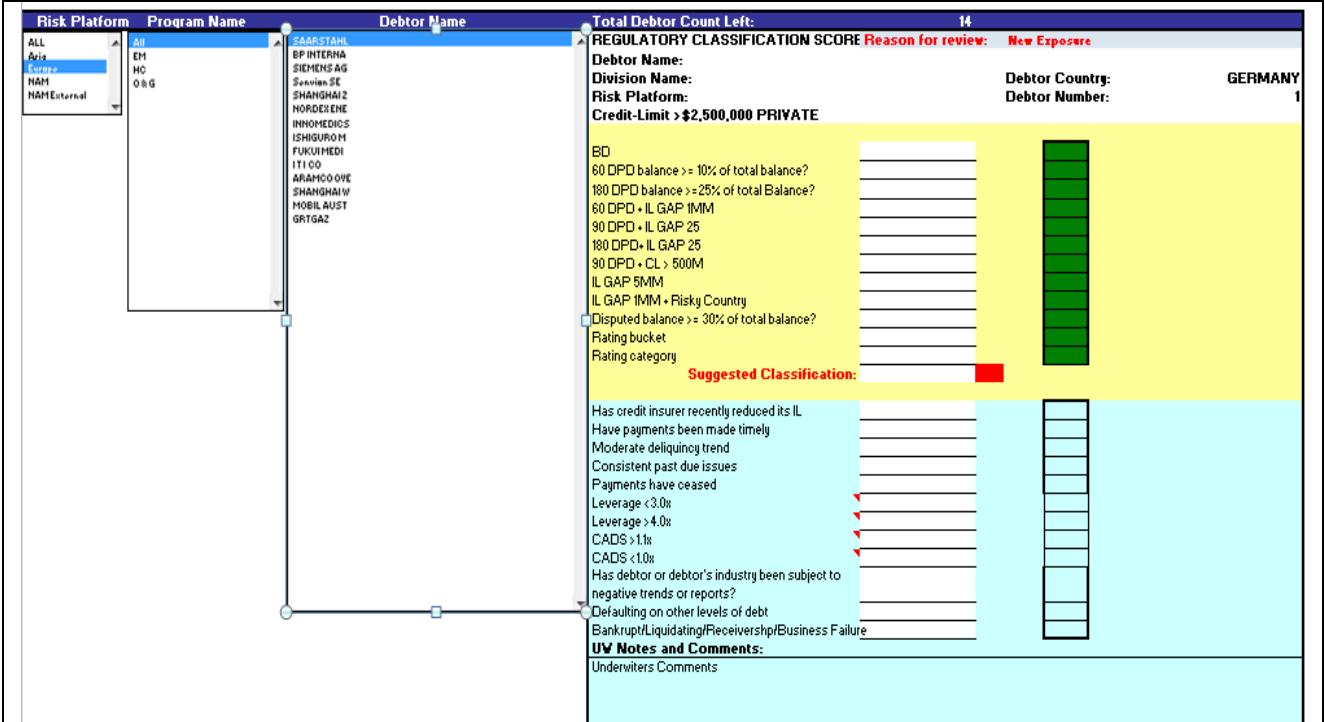


Fig : Reporting Templates Sample– Debtor Information & Score Card

GSO Collections Metrics Overview																																
		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	Mar-13	Apr-13	May-13	Jun-13	Jul-13	Aug-13	Sep-13	Oct-13	Nov-13	Dec-13	Jan-14	Feb-14	Mar-14	Apr-14	May-14	Jun-14	Jul-14	Aug-14	Sep-14	Oct-14	Nov-14	Dec-14	Jan-15	Feb-15	Mar-15	Apr-15	May-15	Jun-15	Jul-15
Performance	AR (#)																															
	PD (#)																															
	PD%																															
	Disputed PD (#)																															
	DPD of AR (#)																															
	DPD to PD (%)																															
	Customer #																															
	Invoices #																															
	PD Invoices #																															
	Disputed #																															
CIM/CA	Completed actions #																															
	Completed actions %																															
	Strategic adherence %																															
	DTL Incoming Volume																															
	Payment promises #																															
Resourcing	Klpt promises # (Conversion)																															
	Quality Audit Score																															
	Collectors																															
	SMEs																															
Fig : Reporting Templates Sample– Collection Metrics Template																																

Key Points

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

The image shows a Microsoft Word document with a 'GENPACT' watermark. On the left, there is a large black rectangular redaction box covering most of the page content. Above this box, the word 'GENPACT' is visible. Below the redaction box, the title 'A mathematical manpower planning model for after-sales field services support' is centered, followed by the date 'Date: 27th June 2013'. At the top of the page, there is a toolbar with various icons and the text 'Genpact Analytics & Research'. On the right side, the 'GENPACT' watermark is repeated. Below the watermark, a 'Table of Contents' is listed:

Table of Contents	
Abstract.....	1
Introduction.....	1
Problem setup.....	3
Problem Definition.....	3
Problem Statement.....	3
Task Prioritization and Resource Allocation.....	3
Sub Function Analysis.....	4
Preference Scores.....	5
Top-down decision tree.....	7
Algorithm.....	7
Numerical Example.....	7
Conclusion and Future Scope.....	8
References.....	10
Annexure.....	11

Key Points

GENPACT

Let's say δ satisfies the above equation. Then if distance to travel is more than δ_0 , 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 condition of after-sales service are written in the Benefit Level Agreement (BLA) 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 welder 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. Tasks and Technicians 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:



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 i^{th} waiting tasks corresponding to j^{th} SubFunction
B _(j,k)	Attribute value of j^{th} Technician corresponding to k^{th} Sub-Function
W _(j)	Weight of j^{th} sub-function
D _(i,j)	Attribute Value of the i^{th} waiting tasks corresponding to j^{th} Sub-Function
D _(j,k)	Function value of j^{th} Technician corresponding to k^{th} Sub-Function
PS _(i)	preference score of i^{th} waiting task
PSP _(i)	preference score of i^{th} Technician

$$PS(i) = \sum_j^M D_{(i,j)} W(j) A_{(i,j)}$$

$$PSP(i) = \sum_j^M D_{(i,j)} W(j) B_{(L)} \sum_k^M W(k)$$

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 Ω^1 , Ω^1 , and Ω^4 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:

$$\begin{pmatrix} \Omega^1 - d \\ \Omega^2 - d \\ \Omega^3 - d \end{pmatrix} = \begin{pmatrix} \delta_1^3 & \delta_1^2 & \delta_1 \\ \delta_2^3 & \delta_2^2 & \delta_2 \\ \delta_3^3 & \delta_3^2 & \delta_3 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

6

Activity

Activity Description:

Make groups of 4-5 and discuss the following case study:

You are required to create a whitepaper for a technical process of your choice. What are the required steps to create and publish the whitepaper.

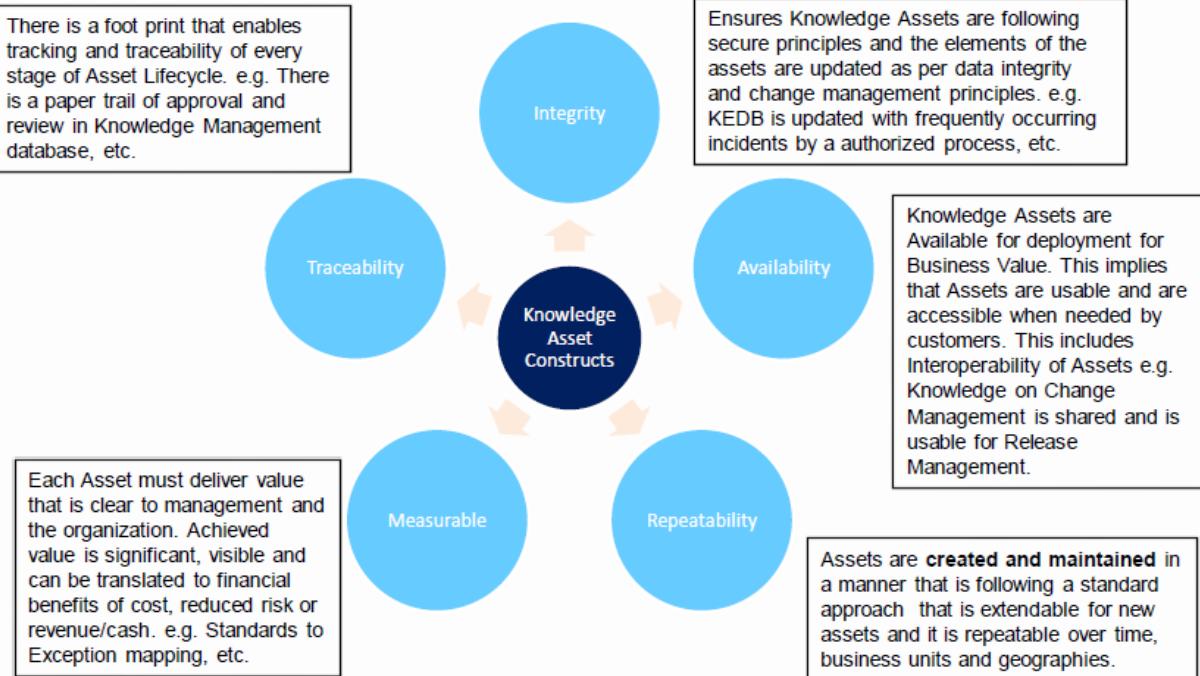
Organizing data/information

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 formal 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

Key Points

Policies and procedures for recording and sharing information



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.

Key Points

- 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 the 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.



Activity

Activity Description:

Make groups of 4-5 and discuss the following case study:

You are launching a new food product in the market for a processed food manufacturer. What are the external compliance norms that you need to keep in mind regarding the quality of food product, the launch of the product and internal compliance norms to ensure good quality delivery.

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.

Session: 4 – Decision Models

Key Points

Deciding how to decide!



Deciding which decision-making method to use can seem like a puzzle in itself.

How you go about making a decision can involve as many choices as the decision itself. Sometimes you have to take charge and decide what to do on your own. Other times it's better to make a decision using group consensus. How do you decide which approach to use?

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.

You don't want to make autocratic decisions when team acceptance is crucial for a successful outcome. Nor do you want be involving your team in every decision you make, because that is an ineffective use of time and resources. What this means is you have to adapt your leadership style to the situation and decision you are facing. Autocratic styles work some of the time, highly participative styles work at other times, and various combinations of the two work best in the times in between.

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

Key Points

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:

Key Points

	<p>Style: Autocratic – you make the decision and inform others of it.</p> <p>There are two separate processes for decision making in an autocratic style:</p> <p><i>Autocratic 1 (A1)</i> – you use the information you already have and make the decision</p> <p><i>Autocratic 2 (A2)</i> – 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.</p> <p>Processes:</p> <p>Style: Consultative – you gather information from the team and other and then make the decision.</p> <p><i>Consultative 1 (C1)</i> – 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.</p> <p><i>Consultative 2 (C2)</i> – you are responsible for making the decision, however, you get together as a group to discuss the situation, hear other perspectives, and solicit suggestions.</p> <p>Processes:</p> <p>Style: Collaborative – you and your team work together to reach a consensus.</p> <p>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.</p>
--	--

Activity Description:

The class is going to be divided 4 groups (each group representing one of the styles of the decision making model). There has been a major security lapse in your organization which deals in trading in securities. It was discovered that there were unauthorized trades to the value of Rs 7.5 Crores which took place in the last month. How would you resolve this issue? Please make your decision in

Key Points



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.

Key Points

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.

Fig : Accenture Knowledge Management solution structure

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

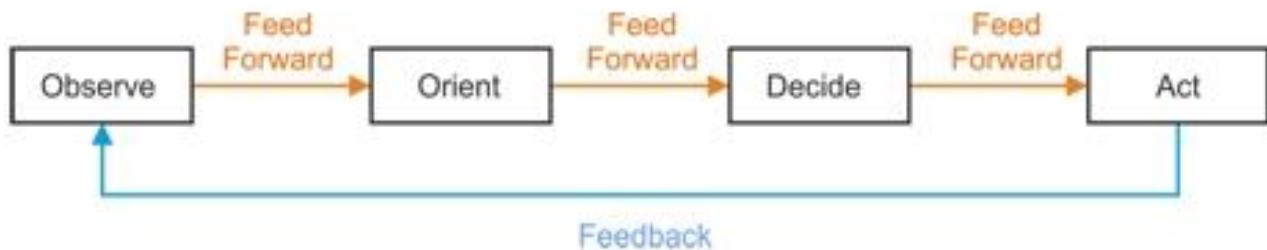
Key Points

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 Act**. Both highlight the importance of analyzing a situation accurately, checking that your actions are having the intended results, and making changes as needed.

Stage 1. Observe

At this initial point in the loop, you should be on the look-out for new information, and need to be aware of unfolding circumstances. The more information you can take in here, the more accurate

Key Points

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

Key Points

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.

Session: 5 - Course Conclusion

Key Points

Course Conclusion



“We’ve almost reached the end of the course! Before we wrap up, let’s review what we’ve learned today”

Ask the participants to recall key learning points from the session and map these learning points to the course objectives.

Thank You Note

Module 2 - Unit: 3

Big Data Analytics

Topic	Activities
Big Data Analytics	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Execute Descriptive analytics on Big Data tools. 2. Detect outlier and eliminate them. 3. Prepare data for analysis.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Introduction to Big Data Analytics	Classroom
Run the descriptive statistics for all the variables and observe the data ranges	Classroom
Outlier detection and elimination	Classroom
Data preprocessing for the analysis	Classroom
Check your understanding	Classroom
Summary	Classroom

Step-by-Step

Key Points

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 insufficient capacity to distribute treatments over a large quantity of machines, to face peaks of charge: this is **the scaling of treatment**,
- The insufficient capacity to store data over a large quantity of machines, to address their **volume**: this is **the scaling of data**,
- The insufficient capacity to distribute the database over several data-centres to insure the continuity of service: This is **the redundancy**,
- the insufficient capacity to adjust its data model to the nature of the data, due to the **velocity** of web data,
- The bad performance to access data as soon as it requires the combination of several tables due to the **variety** of web data and the **complexity** of unstructured data.

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

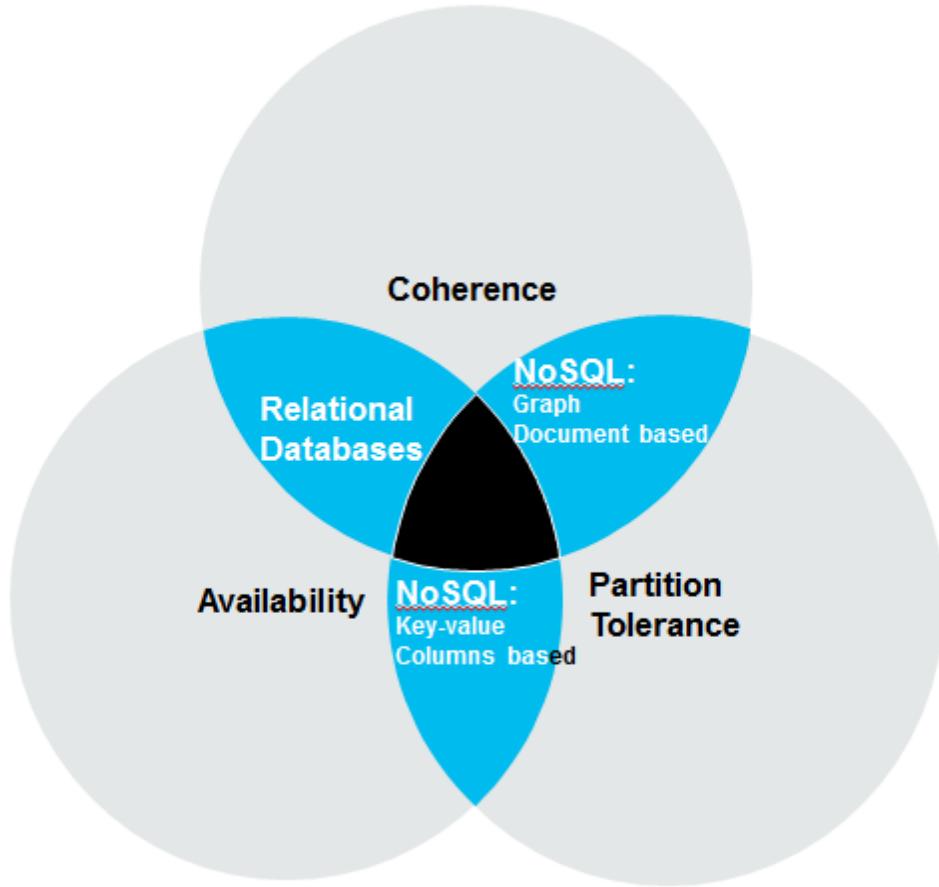
As a relational database, it provides a set of functionalities to access data across several entities (tables) by complex queries. It provides also integrity referential to insure the constant validity of the links between entities. Such mechanisms are extremely costly and complex to implement in distributed 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 of 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:

- The **Coherence**: All the nodes of the system have to see exactly the same data at the same time
- The **Availability**: The system must stay up and running even if one of its nodes is failing down

- The **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*

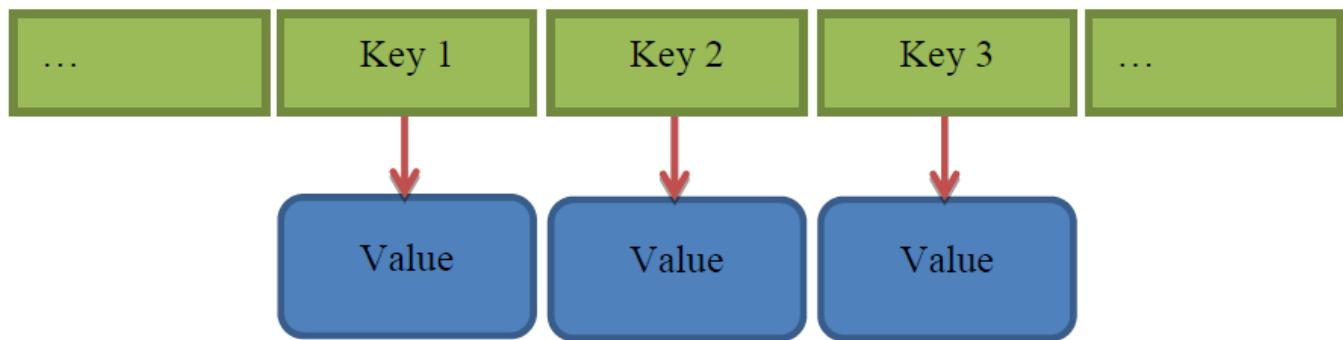
databases are more focused on the complexity of data, and thus on the *coherence* of the database. NoSQL and Big Data

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.

A few providers Solution	Distribution	Comment
Redis http://redis.io	BSD licence	Certainly the more mature, Providing some functionalities of to manipulate and store strings and collections No real mechanism of partitioning but some functionalities of master/slave replication Sponsored by VMware
Riak http://basho.com	Apache licence 2.0	Open source implementation of Amazon Dynamo Completely distributed Map/reduce enabled
Voldemort http://www.project-voldemort.com/voldemort/	Apache licence 2.0	Initially developed by LinkedIn Optimizing the communication between nodes of the network



Columns based databases

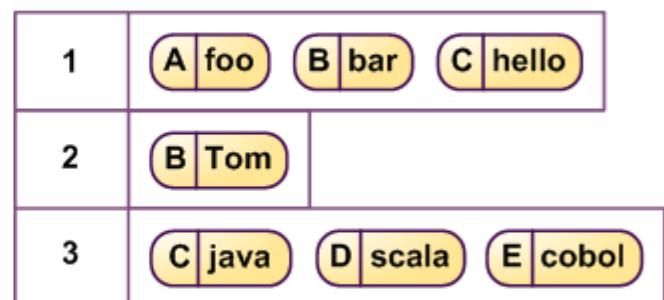
Concept

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.

	A	B	C	D	E
1	foo	bar	hello		
2			Tom		
3			java	scala	cobol

Structure of a table in a relational database



Structure of a table in a columns oriented database

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.

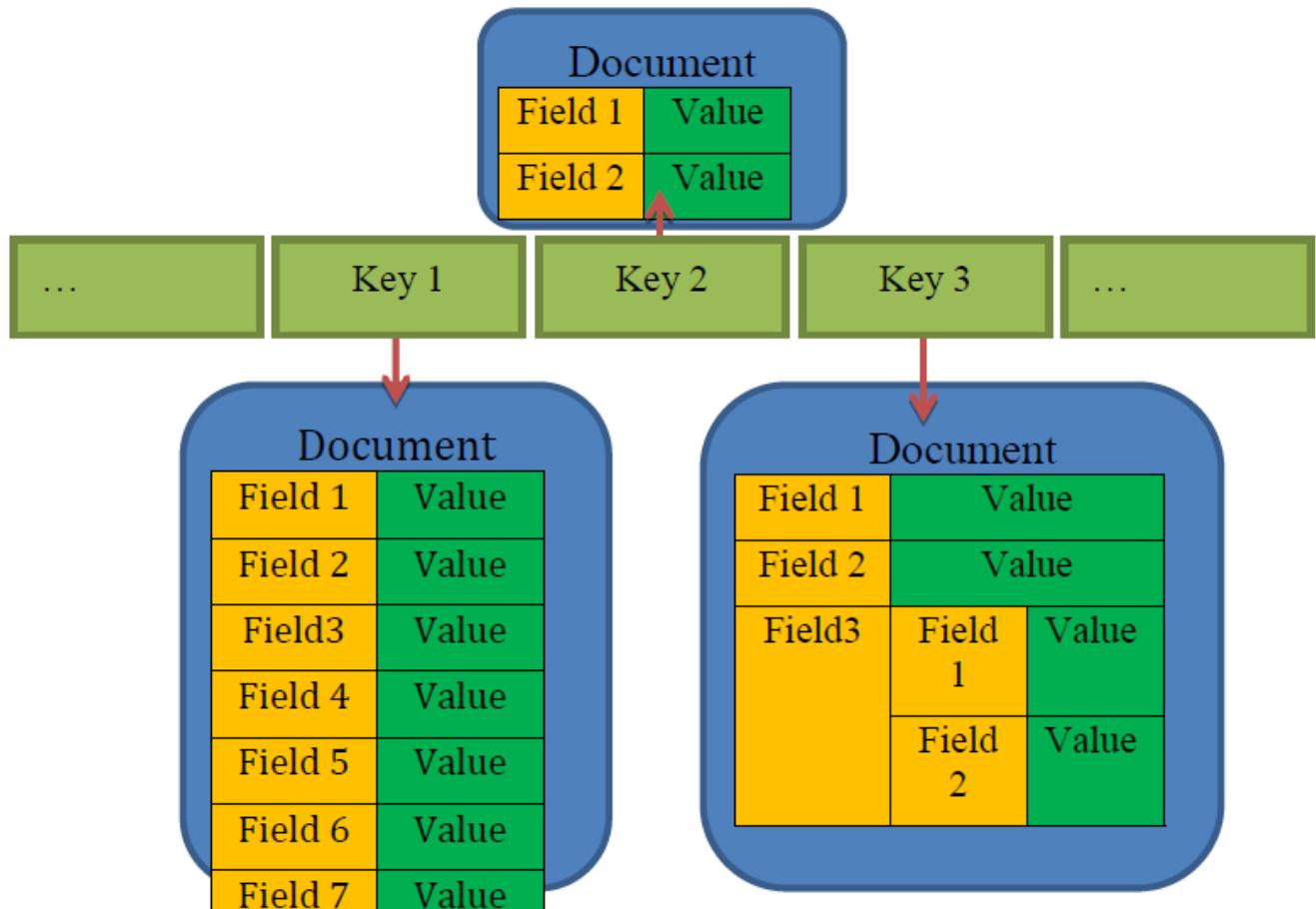
3.2.2 A few providers Solution	Distribution	Comment
Cassandra http://cassandra.apache.org	Apache licence 2.0	Very popular open source database Most online services are using it such as Facebook
Goole Big table	SaaS	Database of the Google App Engine, the Google web development environment Only accessible from the API of Google App engine
HBase http://hbase.apache.org	Apache licence 2.0	Open source implementation of Google Big Table based on Hadoop

Document based databases

Concept

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

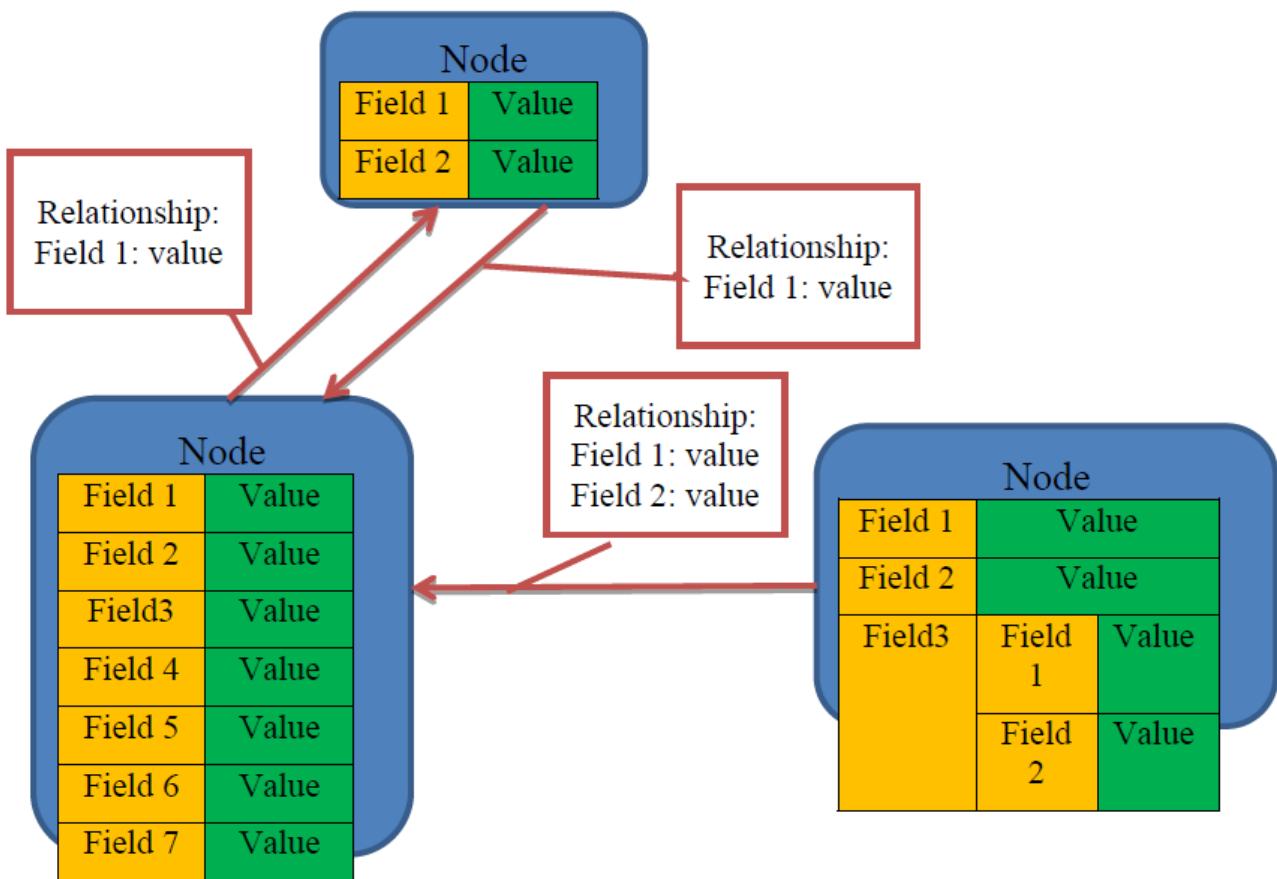
Even if the documents are structured, these databases are *schemaless*, meaning that it is not necessary to previously determine the structure of the document. The simplicity and flexibility of this data model makes it particularly applicable to *Content Management Systems* (CMS).



Graph databases

Concept

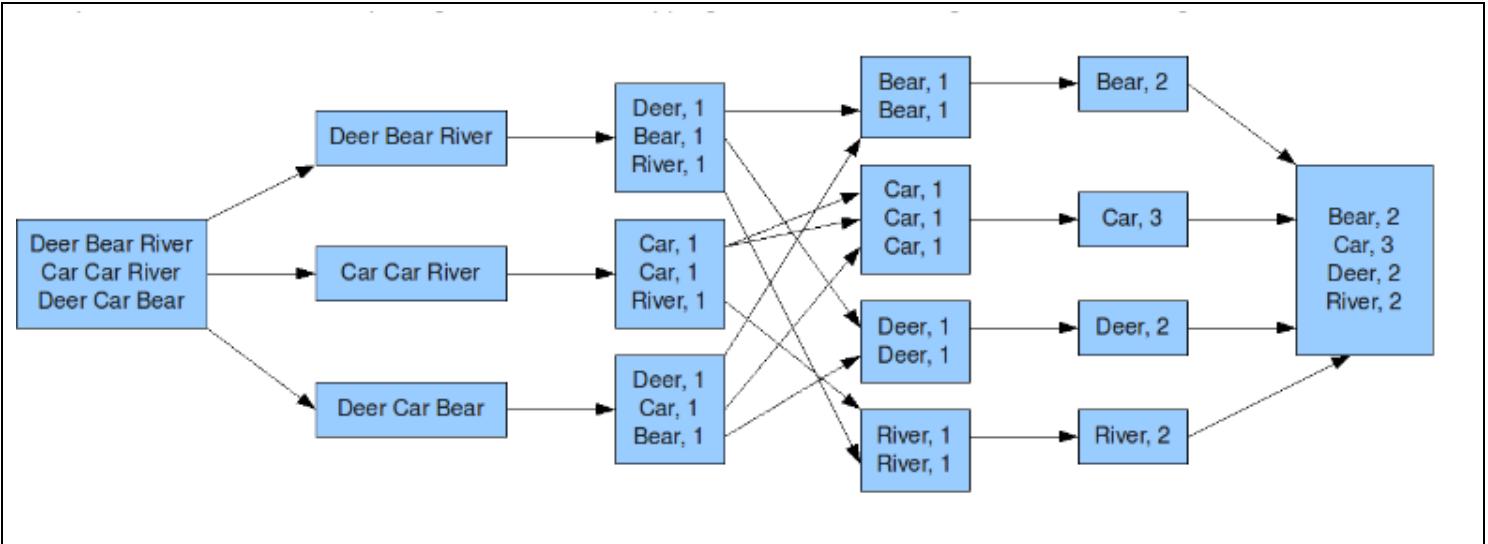
The graph paradigm is a data model in which entities are *nodes* and associations between entities are *arcs* or *relationships*. Both nodes and relationships are characterized by a set of properties. This category of databases is typically designed to address the complexity of databases more than their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



MapReduce

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

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



Descriptive Statistics

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

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

Outlier detection and elimination

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

Data 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.

Check your understanding



1. What are examples of places where Big Data analytics is used?
2. What is the need of Data Preprocessing?
3. What are Outliers and how do we detect them?
4. What is Descriptive Analytics in Big Data Analytics?

Summary

- Big data analytics is the process of examining large data sets containing a variety of data types.
- The primary goal of big data analytics is to help companies make more informed business decisions.
- 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.
- Data pre-processing is an important step in the data mining process

Module 2: Unit – 4

Machine learning Algorithms

Topic	Activities
Machine Learning Algorithms	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Do Hypothesis Testing 2. Determine multiple analytical methodologies. 3. Train model no 2/3 sample data. 4. Predict Sample. 5. Explore chosen algorithms for accuracy.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Hypothesis testing and Determining the multiple analytical methodologies	Classroom
Train model using statistical/machine learning algorithms, Test model	Classroom
Sample for prediction	Classroom
Explore the chosen algorithms for more accuracy	Classroom
Check your understanding	Classroom
Summary	Classroom

Step-by-Step

Key Points

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.

Check your understanding



1. What is Machine Learning?
2. What are the types of Machine Learning?
3. What are practical uses of Machine Learning?
4. Why is Machine Learning joined with Data Analytics?
5. What is Train Model and Test Model?
6. What are steps followed in Machine Learning Algorithm?

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.

Module 2: Unit – 5.1

Data visualization

Topic	Activities
Data Visualization	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Prepare Data for visualization. 2. Draw insights out of visualization tools.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Prepare the data for visualization	Classroom
Draw insights out of the visualization tool	Classroom
Check your understanding	Classroom
Summary	Classroom

Step-by-Step

Key Points

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 statisticalcomputations.

Data Visualization in Tablue

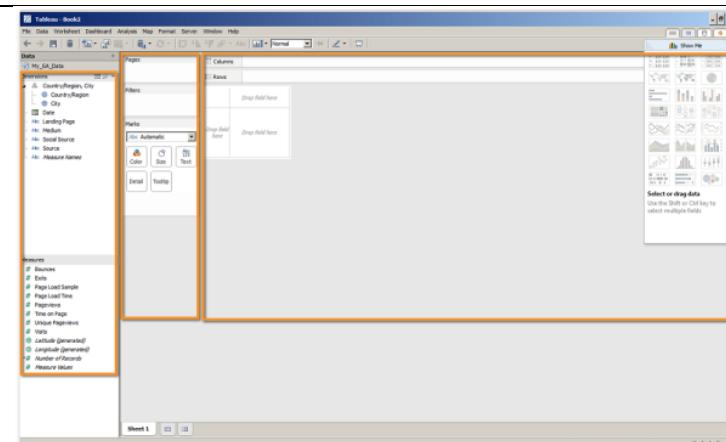
Extract The Data.

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.

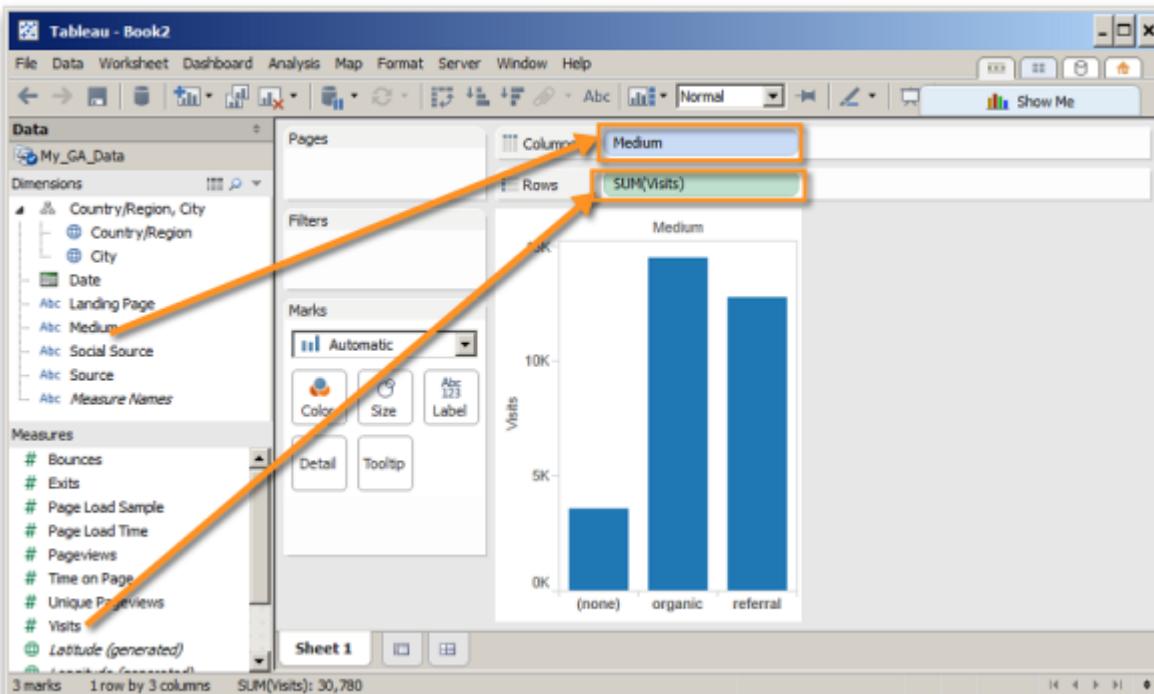
Keep in mind that the more dimensions you add, the larger the data set will get. For example, adding a device type will give you one row of measures for each device. You can think of it this way: if your default data has 10,000 rows, and you add the hour dimension, you would have $10,000 \times 24$ (hours).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

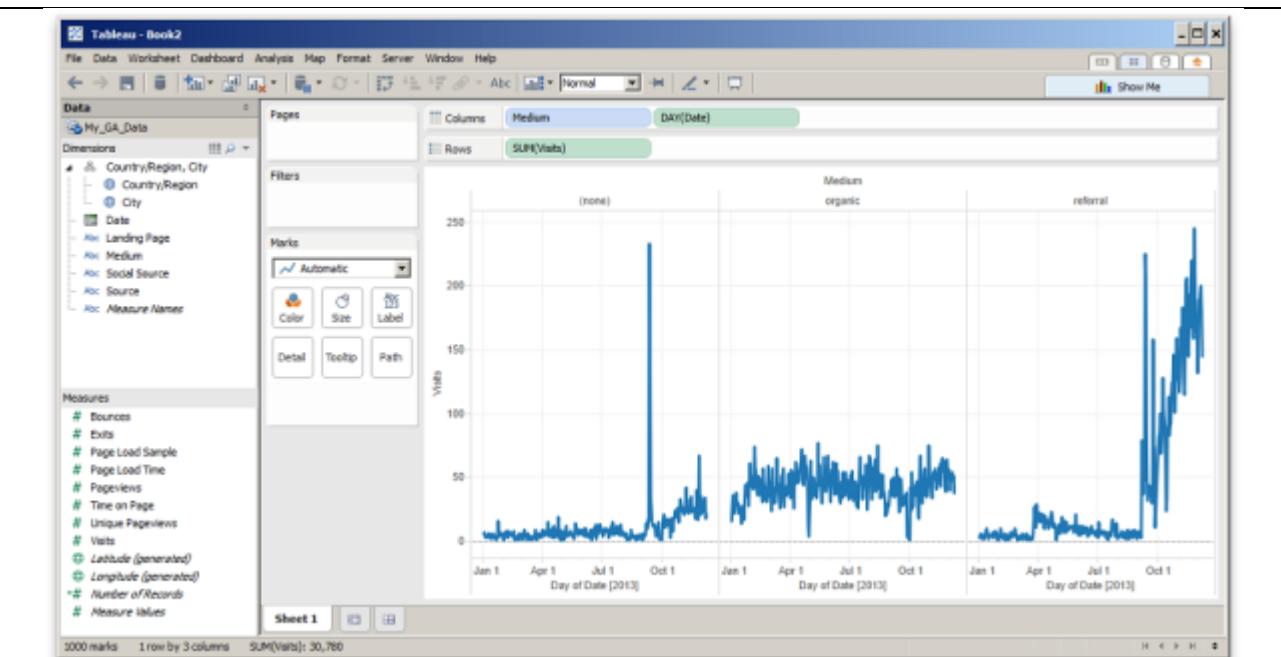


Your First Data Visualization. For our first effort, let's say we want to see what the traffic by medium looks like. To accomplish this, all you need to do is drag and drop icons from the dimensions and measures sections over to the columns and rows spots at the top. Specifically, drag the mediums icon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

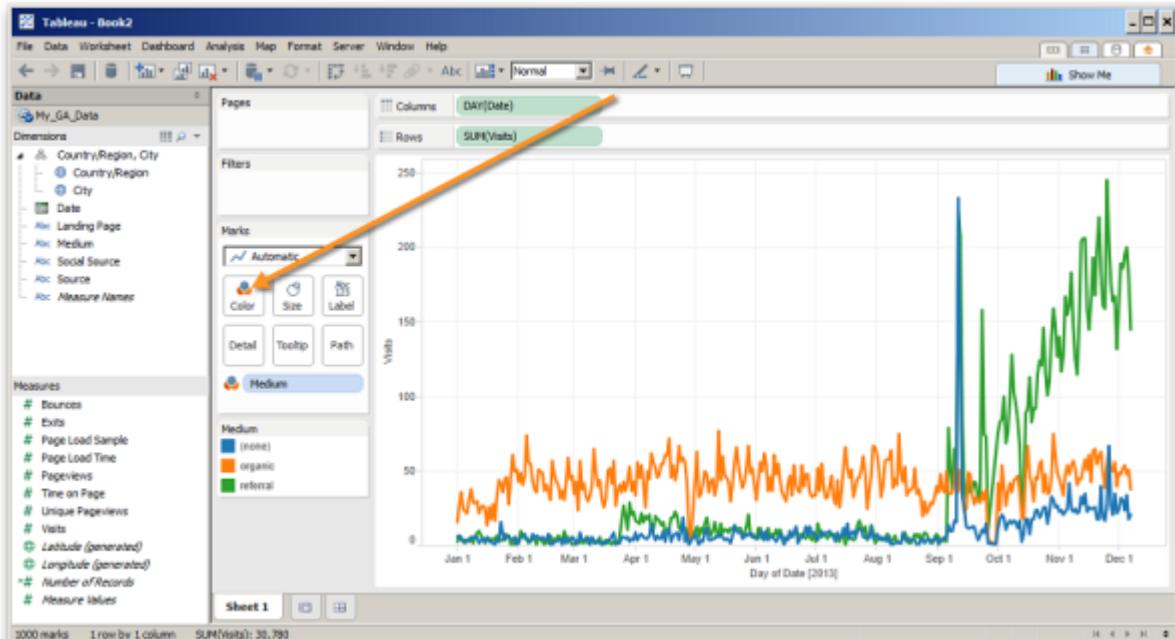


Now, how can we utilize this to take things to the next level? Let's look at historic performance. To do so, drag the date icon into the end of the column line. This will show us the performance by medium by year.

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.



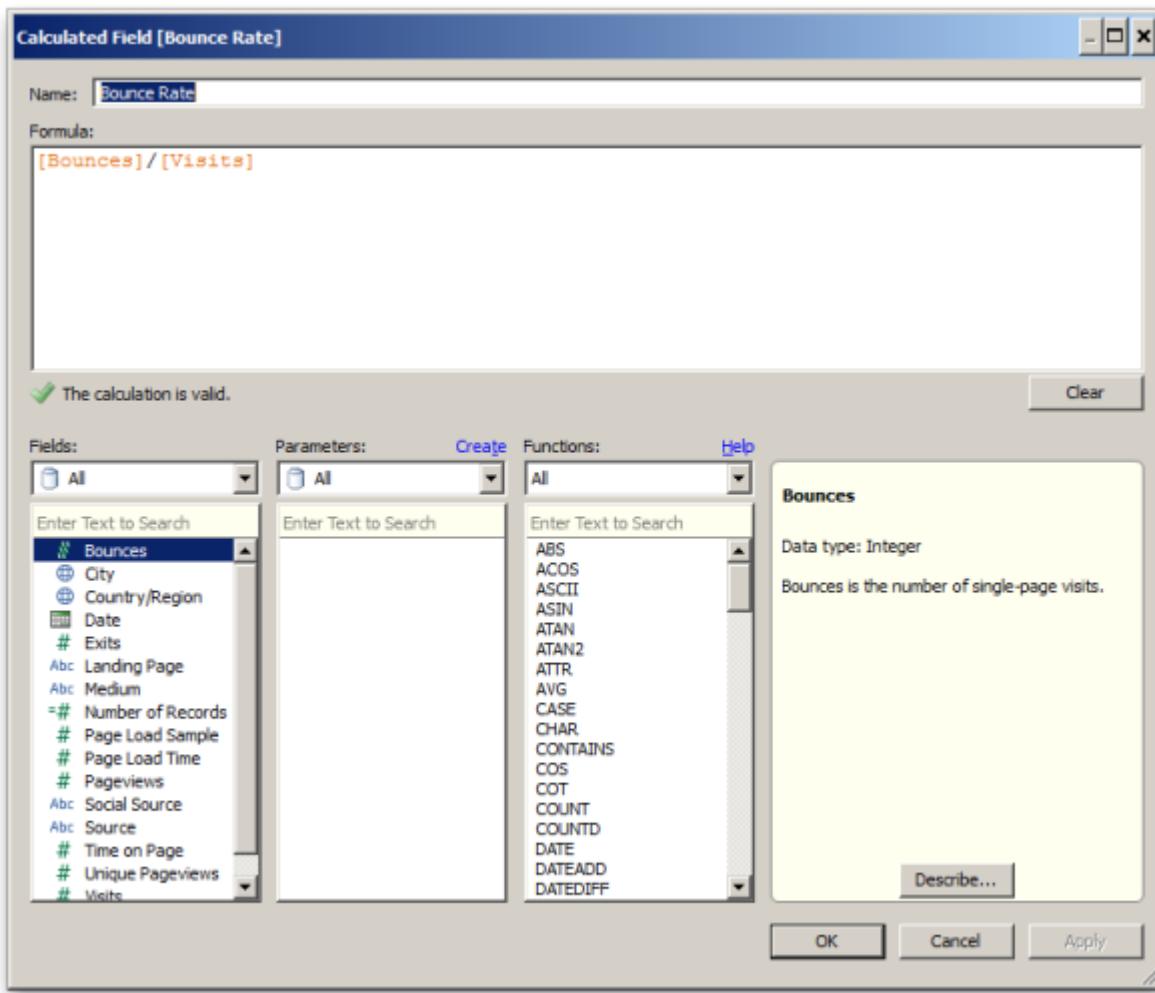
But here is where the real power of DVTs comes into play. By simply dragging the medium from the columns area to the color area, you are instantly removing the columns for medium, combining it into one chart area, and then coloring by medium. This allows you to easily compare the data in a more visually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



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].



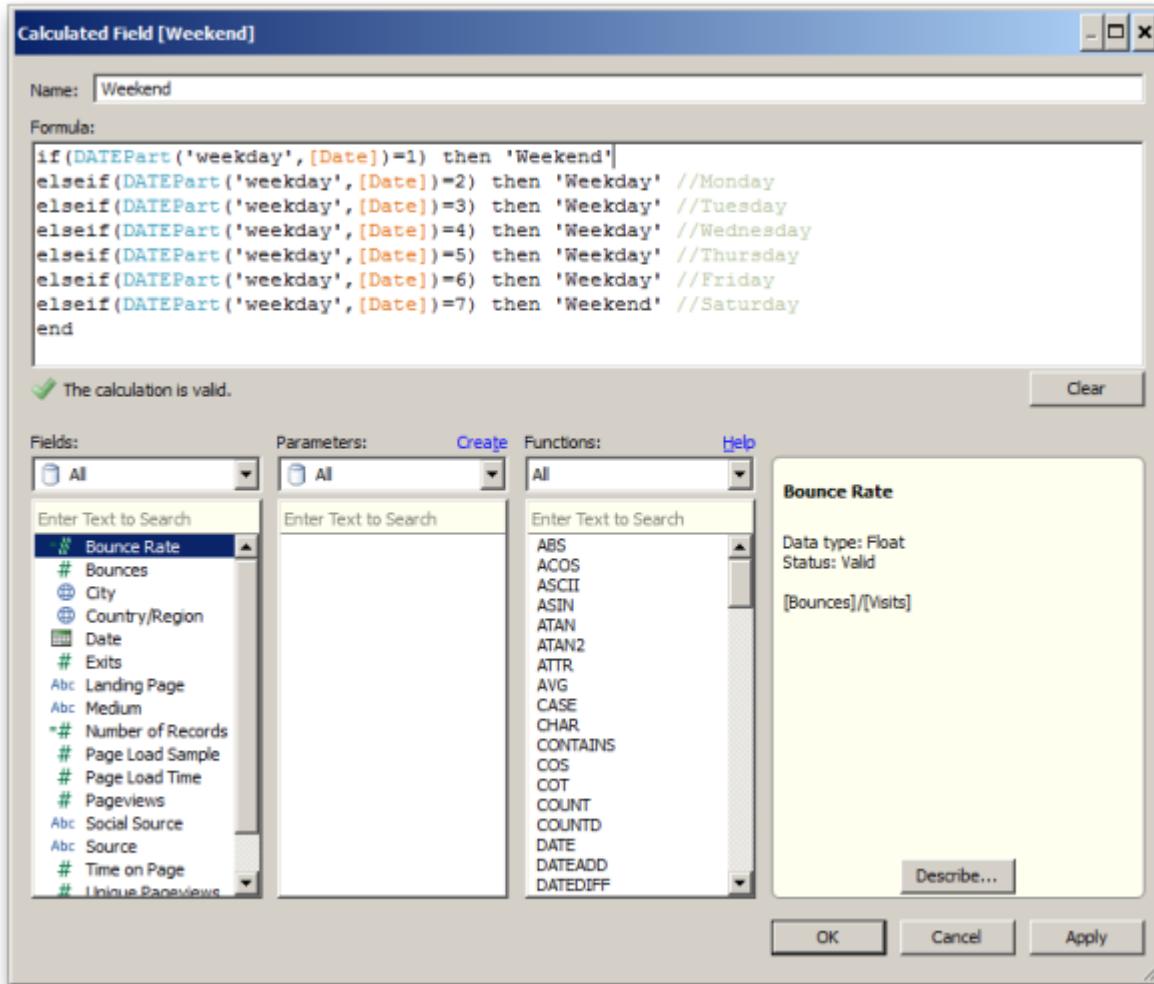
The same approach can be used to do a variety of calculations. For instance, the code below could be used to distinguish weekend vs. weekday traffic:

```
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

```



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.

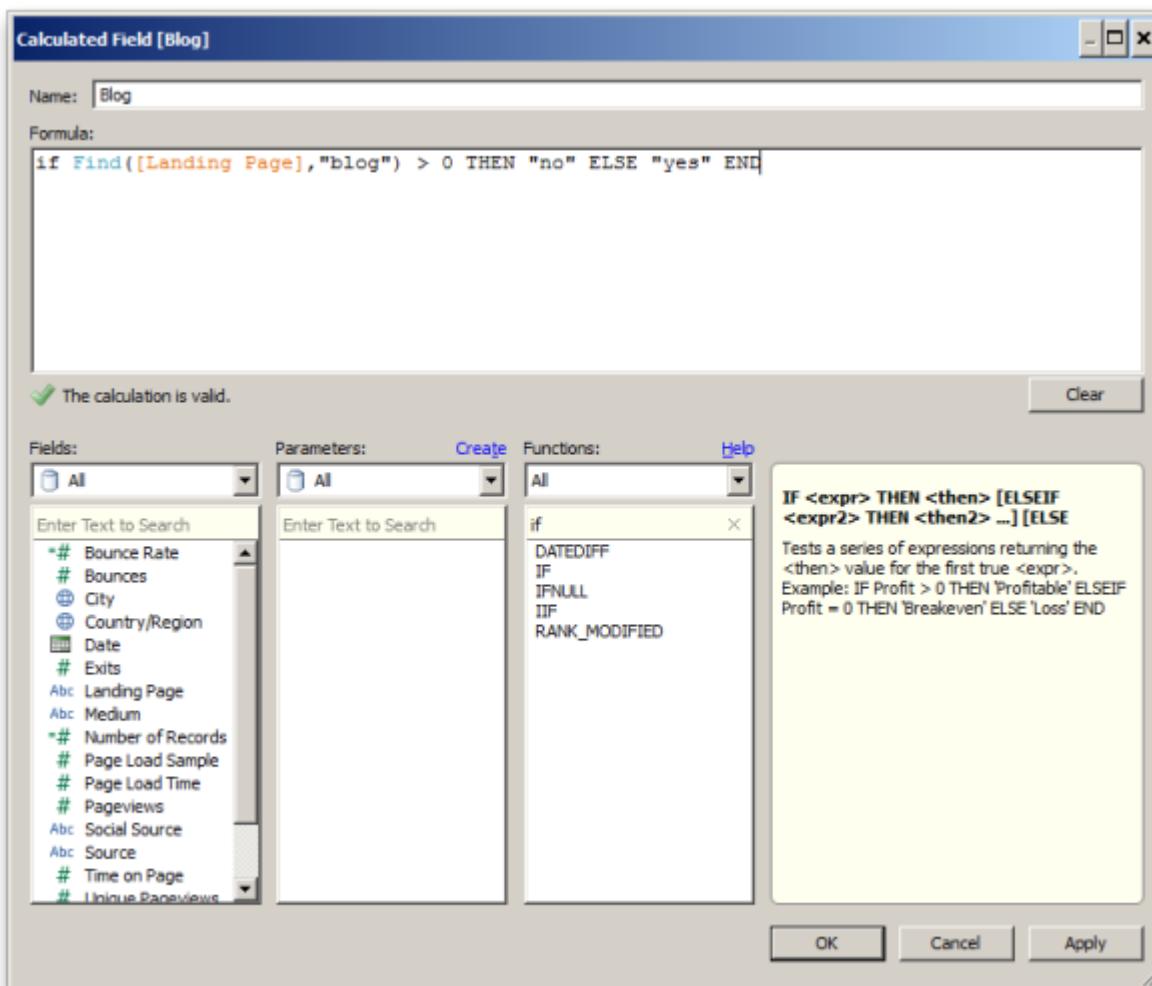
Another great way to look at the data is to filter down to only social source traffic, and then look at site engagement (PPV or time on-site). This will show you when your social traffic is performing best and when you can get the most ROI out of your social activities.

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:

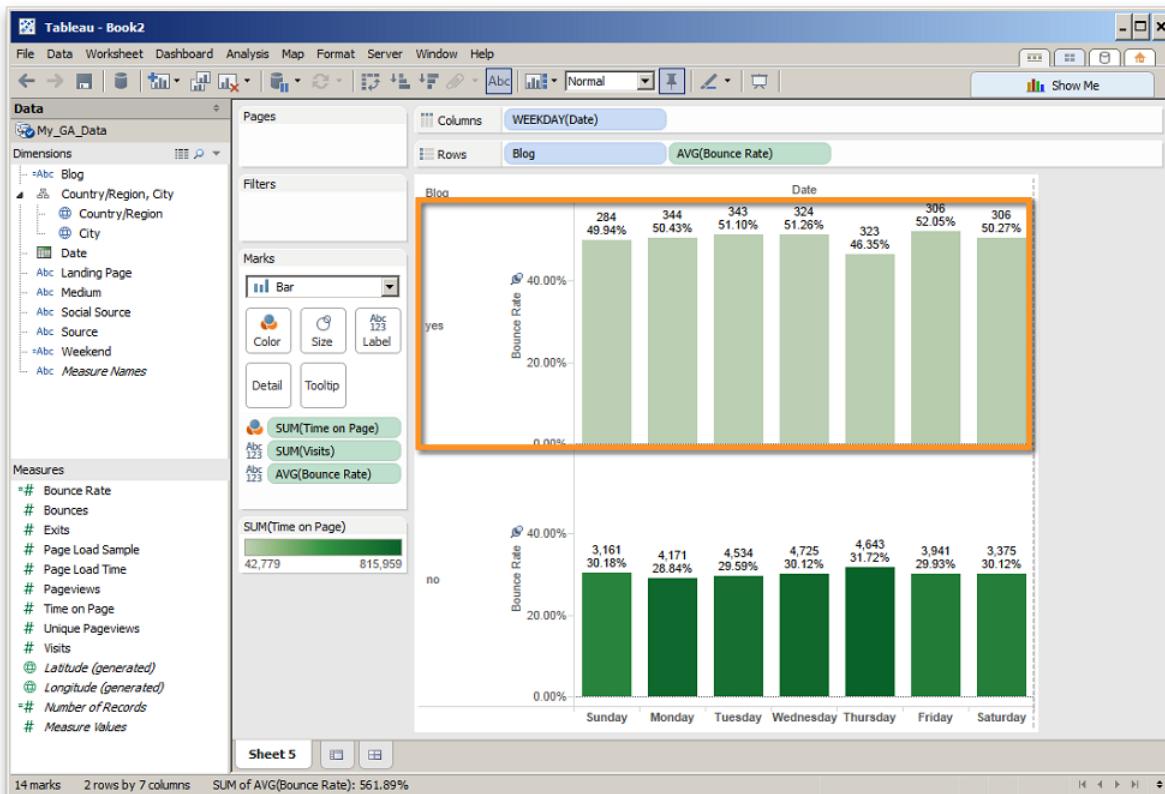
if Find([Landing Page],"/blog/") > 0 THEN "no" ELSE "yes" END

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



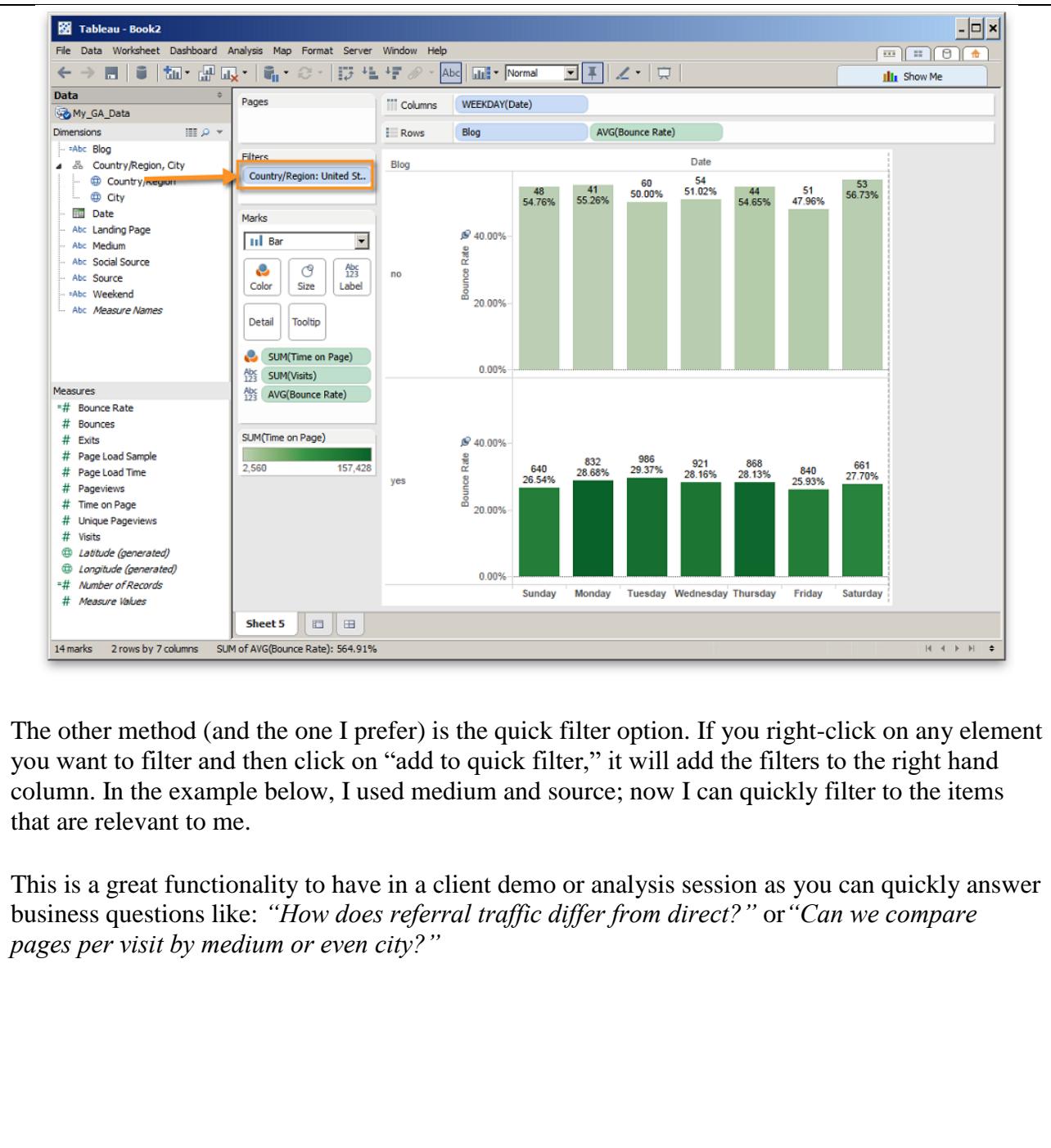
Now, we can look at the engagement by blog pages and non-blog pages, and even divide it by day of week. As you can see, the upper section (blog pages) has a much higher bounce rate than the

non-blog pages. (It also seems that there was some special activity on Thursdays that affected the bounce rate).



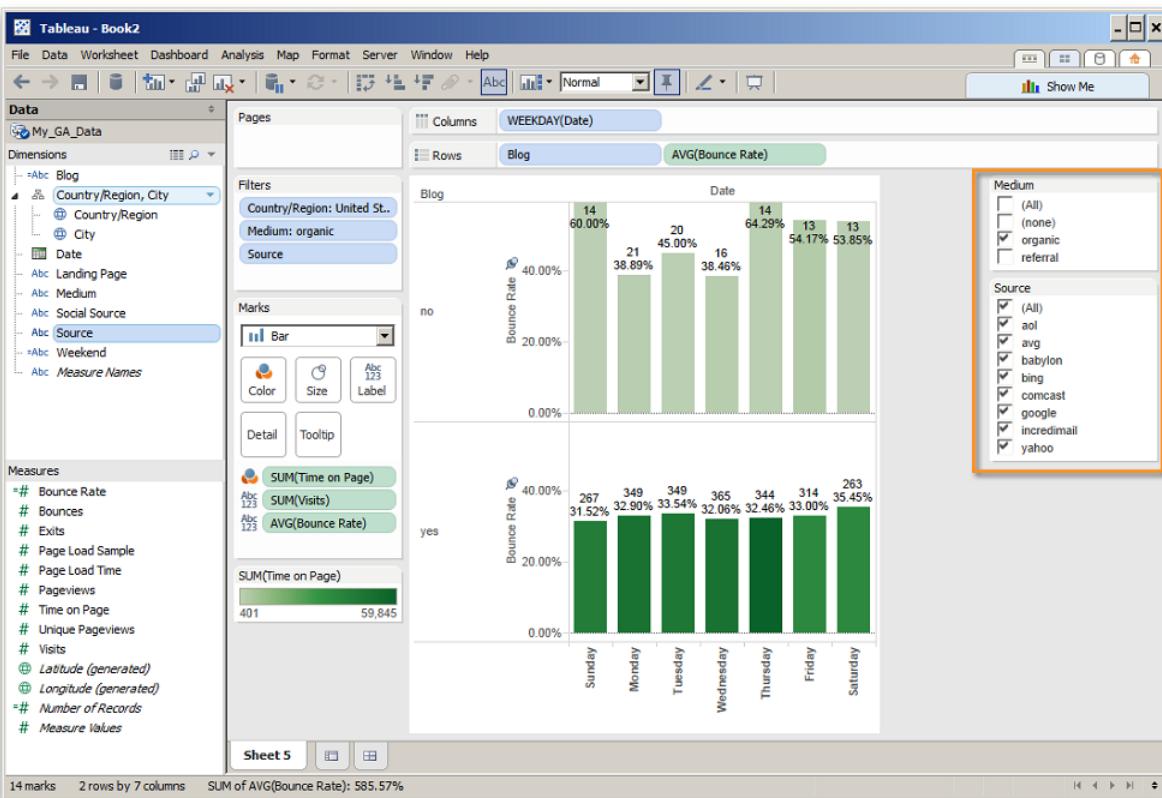
The calculated field option is one of my favorite features in Tableau, as it allows you to dynamically extend and segment your data. We have done some amazing calculations with this functionality, and once you start playing around with the calculated field dialog, you will see the large variety of powerful 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.

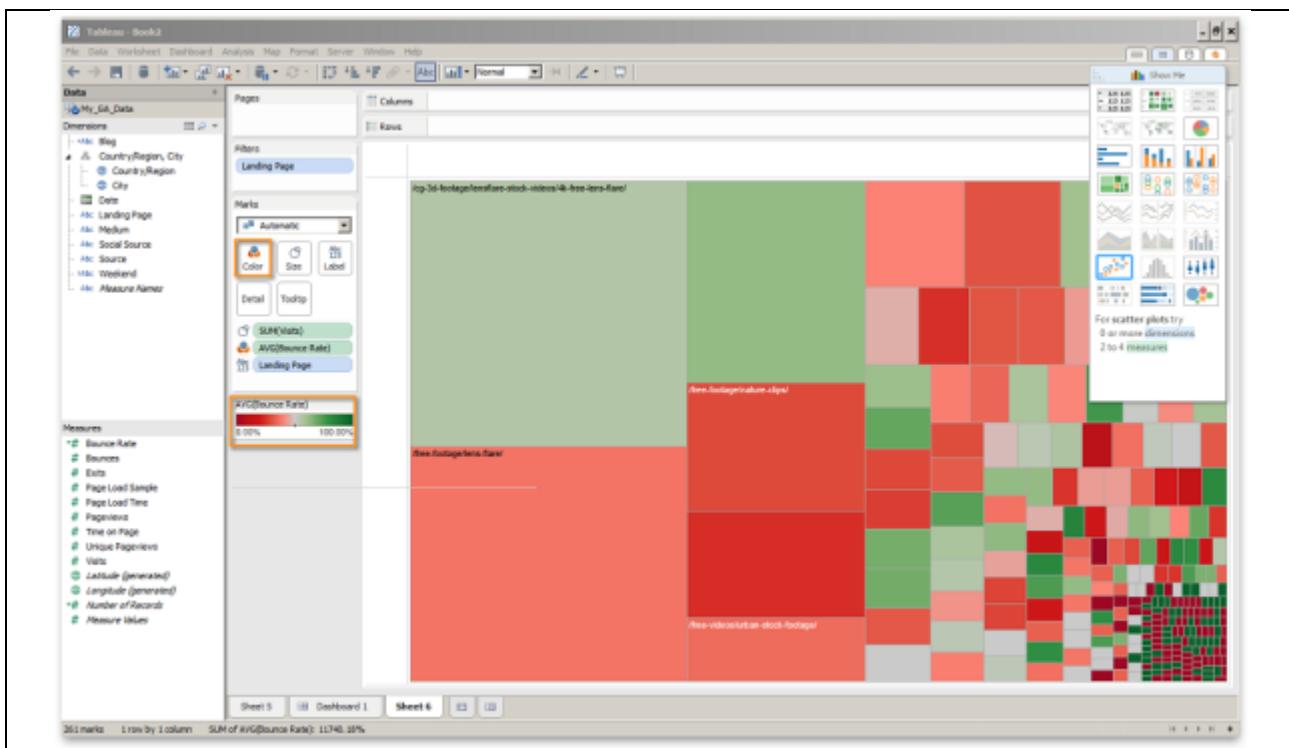
For instance, let's select landing page and visits from the measure and dimensions area, and then select treemap for the visual type. Immediately, it shows you squares that represent individual landing pages, each one sized by the number of visits it has received.

The screenshot shows the Tableau interface with a treemap visualization. The left pane displays the data source 'My_Edt_Data' with dimensions like 'Flag', 'Country/Region, City', 'Social Source', 'Medium', 'Source', 'Weekend', and 'Landing Page'. The measures pane includes 'Bounce Rate', 'Bounces', 'Exits', 'Page Load Sample', 'Page Load Time', 'Pageviews', 'Time on Page', 'Unique Pageviews', and 'Visits'. The 'Visits' item is highlighted with an orange box. The treemap itself has a large green area representing the homepage ('/'). A tooltip window titled 'Show Me' is open over the green area, containing icons for various Tableau functions and the text 'For treemaps try 1 or more dimensions 1 or 2 measures'.

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.



Check your understanding



1. What is DPA?
2. Why do we use Data Visualization?
3. What is the role of Tableau in Data Visualization?
4. Write down steps involved in Data Visualization InTableau.

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.

Unit – 5.2

Analytics Application to various domains

Topic	Activities
Analytics application to various Domains	By the end of this session, you will be able to:

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Planning and estimation	Classroom
Drivers of Asset/ engine risk categories analytics	Classroom
Different approaches to asset scoring models	Classroom
Validation and maintenance of asset scoring models	Classroom
Understanding the current Engineering/Manufacturing/Asset System	Classroom
Creating the Business Understanding document	Classroom
Understanding the data and creation of Data Dictionary	Classroom
Preparing the data, analysis and modeling	Classroom

Step-by-Step

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.

Unit – 6

CASE STUDY

Exploring Employee Attrition data in R

Problem Statement

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

In this case study, we explore the impact of various employee/employment aspects which can possibly have an impact on attrition. Using the sample dataset available from <https://community.watsonanalytics.com/hr-employee-attrition/>, explore the data, understand each attribute and its contribution to the attrition using visualizations and eventually come up with a fit using various models which summarizes the significant attributes which most influence attrition.

Loading and understanding data

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

Data exploration and visualization

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

- Explore the correlation between numeric variables
 - Identify the variables which seem most correlated to employee attrition
 - Plot the correlations
 - Do some independent variables appear more correlated to each other than others

Fitting a logistic regression model

- Specify the null hypothesis for this model
- Run a logistic regression fit for Attrition against remaining variables
- Explain the following from fit summary
 - AIC
 - Intercept and coefficients
 - Null and Residual deviance values
- Calculate the goodness of the model using log-likelihood function
- Try various combinations of encoding and scaling schemes to the data to improve the error score
- Finally, based on the results, should the null hypothesis be rejected

Applying different models

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

Present your findings and conclusion

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

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



ASSOCIATE ANALYTICS
FACILITATORS GUIDE
MODULE 3



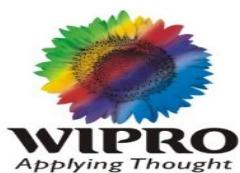
This Facilitators Guidebook for the Associate Analytics program contains detailed facilitation guidelines as well as the exhaustive course material for the Associate Analytics program.

Facilitator's Guide



Associate - Analytics

Powered by:



Copyright © 2014

NASSCOM

4E-Vandana Building (4th Floor)
11, Tolstoy Marg, Connaught Place
New Delhi 110 001, India
T 91 11 4151 9230; F 91 11 4151 9240
E ssc@nasscom.in
W www.nasscom.in

Published by



Building Domain | Enhancing Careers

T: 91 70365 88888
E info@mindmapconsulting.com
W www.mindmapconsulting.com

Disclaimer

The information contained herein has been obtained from sources reliable to NASSCOM. NASSCOM disclaims all warranties as to the accuracy, completeness or adequacy of such information. NASSCOM shall have no liability for errors, omissions, or inadequacies, in the information contained herein, or for interpretations thereof. Every effort has been made to trace the owners of the copyright material included in the book. The publishers would be grateful for any omissions brought to their notice for acknowledgements in future editions of the book.

No entity in NASSCOM shall be responsible for any loss whatsoever, sustained by any person who relies on this material. The material in this publication is copyrighted. No parts of this report can be reproduced either on paper or electronic media, unless authorized by NASSCOM.

Foreword

The Indian IT-ITeS industry has built its reputation in the global arena on several differentiators, chief among them being the availability of manpower. Organizations across the world recognize the value India brings to every engagement with its vast and readily available pool of IT professionals. Global entities have found it extremely effective to leverage this significant resource in order to enjoy a competitive edge and innovation benefits.

In the coming years, the landscape is expected to shift in ways that reveal more exciting opportunities. The world will require people with advanced technology skills and domain knowledge, set against a backdrop of heightened labour mobility across occupations and markets. India is largely acknowledged to be heir apparent to the benefits of a demographic dividend over the coming decades, which has the potential to see the nation emerge as one of the world's largest population base of employable youth. With many other countries set to face the effects of an aging and retirement-ready workforce, India is poised to become a sought after destination for those seeking higher value add and specialized services.

Global markets are on their way towards revival and recovery, and this is well reflected in the proactive recruitment measures taken by IT-ITeS organizations in India in recent times. India's IT-BPM industry is on track to achieve its target of USD 225 billion by 2020. From a base on about 3.1 million employees in FY2014, the industry is expected to add another 2 million additional employees by 2020. Indirect employment generated by 2020 is expected to be 3X the total direct employment number is between 13-16 million by 2020.

To realize India's potential of emerging as a skills hub of the world, a significant amount of foresight and work is requisite. It is imperative that stakeholders engage in a concerted effort to undertake the transformation of the labour pool estimated to enter the market into skilled and employable talent. Enabling the creation of a future industry-ready cohort will give the IT-ITeS industry an edge in leadership and sustainability.

One of the burgeoning areas of governance and strategy relates to leveraging big data and analytics. This led to the identification of the "hot skills" du jour, resulting in the formal creation of a qualification pack (QP) or job role framework for the role of Associate Analytics. The QP is designed to capture the skills demanded by the IT-BPM Industry for an entry level position in this field.

To ensure the creation of an academic course that is both relevant and viable, NASSCOM partnered with key industry stakeholders, including Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. In addition, the program addresses the need for faculty support, and achieves this by acquainting trainers with the latest advancements in pedagogy.

We wish the universities and colleges all the very best in their endeavor.

R Chandrashekhar
President
NASSCOM

Acknowledgements

NASSCOM would like to thank its member company representatives within the Analytics Special Interest Group (SIG) Council for believing in our vision to enhance the employability of the available engineering student pool. SSC NASSCOM facilitates this by developing and enabling the implementation of courses relevant to projected industry needs. The aim is to address two key requirements, of closing the industry-academia skill gap, and of creating a talent pool that can reasonably weather future externalities in the IT-BPM industry.

NASSCOM believes that this is an initiative of great importance for all stakeholders concerned – the industry, academia, and the students. The tremendous amount of work and ceaseless support offered by the members of this SIG in developing a meaningful strategy for the content and design of program training materials has been truly commendable.

We would like to particularly thank Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, Fractal Analytics, First America, Genpact, Infosys BPO, Insights of Data, Karvy Analytics, Wipro, WNS and Wells Fargo for bringing much needed focus to this effort.

NASSCOM recognizes the fantastic contributions of Mr. Ashok Polapragada, Mr. Ranjit Kumar and Mr. Prakash Devarakonda at Karvy Analytics; Mr. Dwaraka Ramana K at First American; Mr. Amit Agarwal, Mr. Sidhartha Shishoo and team at Genpact; Ms. Snigdha Ray and Mr. Amit Sharma at ADP; Mr. Manoj Koundinya at Capgemini, and Mr. Ashish Mediratta at Wipro.

We acknowledge with sincere gratitude the immense contribution of the SIG member companies, Accenture, ADP, Capgemini, Concentrix, Cyient Insights, EXL, First American, Fractal Analytics, GENPACT, Infosys BPO, Karvy Analytics, Wells Fargo, Wipro, and WNS. For their part in the creation of this course and its accompanying training materials.

We extend our thanks to Mindmap Consulting Pvt. Ltd. for producing this course publication.

Dr Sandhya Chintala
Executive Director – Sector Skill Council
Vice President - NASSCOM

Table of Contents – Module 3

Introduction to QP Associate Analytics

Introduction to Associate Analytics	8
Career growth in Analytics	11
Qualification pack - Q/2101 Associate Analytics	12
Overall Associate Analytics Content Structure	20
Glossary of terms	22

CORE CONTENT

UNIT 1.1 Introduction to Predictive Analytics	27
UNIT 1.2 Linear Regression	33
UNIT 2.0 Logistic Regression	37
UNIT 3.1 Objective Segmentation	51
UNIT 3.2 Develop knowledge, skills and competencies	57
UNIT 4.1 Time Series Methods/Forecasting, Feature extraction	80
UNIT 4.2 Project	88
UNIT 5.0 Working with documents	90

Introduction

Qualifications Pack-Associate –Associate Analytics SSC/Q2101

SECTOR: IT-ITeS

SUB-SECTOR: Business Process Management

OCCUPATION: Analytics

REFERENCE ID: SSC/Q2101

ALIGNED TO NCO CODE: TBD

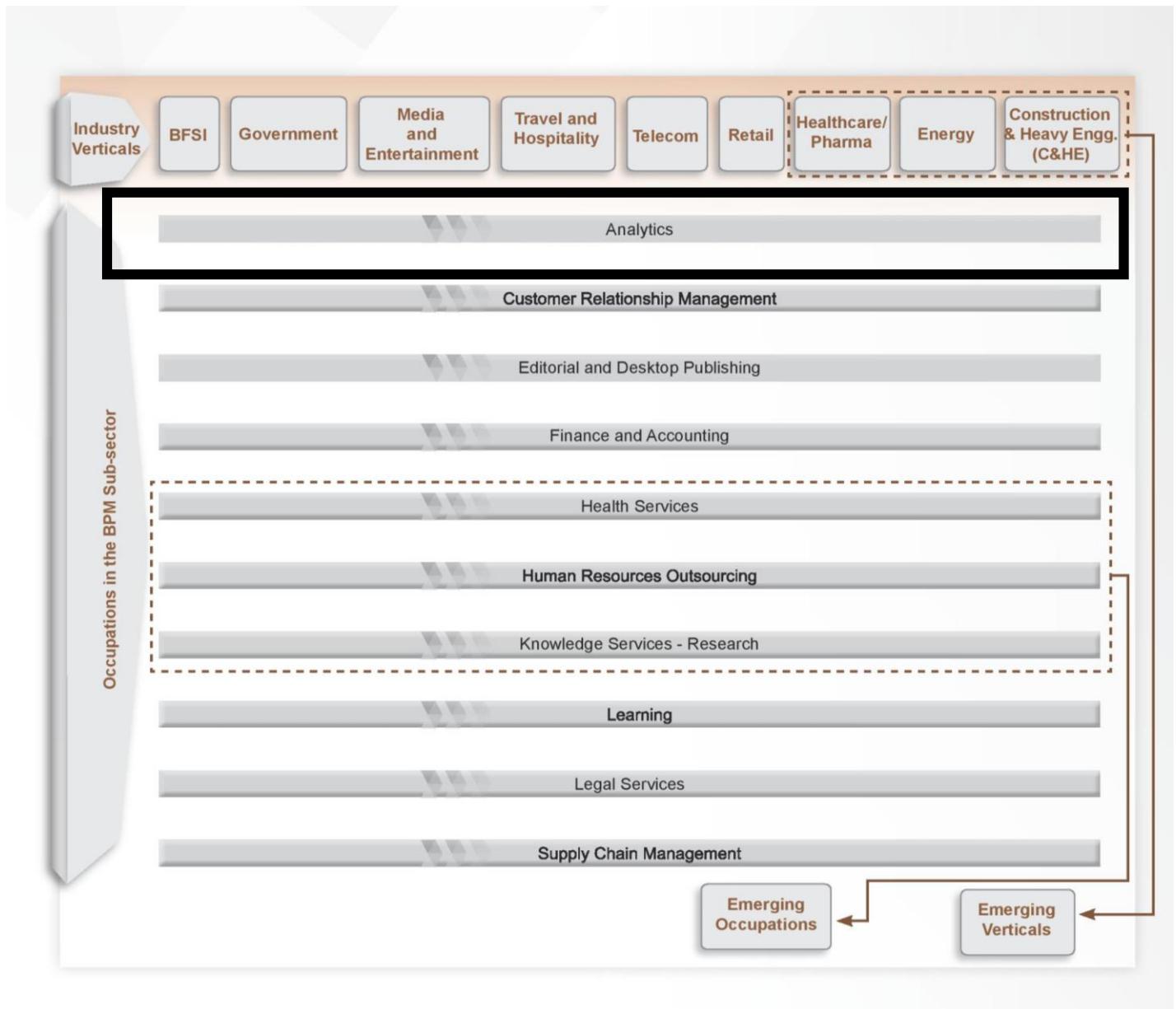
Brief Job Description: Individuals at this job are responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools

Personal Attributes: This job requires the individual to follow detailed instructions and procedures with an eye for detail. The individual should be analytical and result oriented and should demonstrate logical thinking.

Eligibility: Bachelor's Degree in Statistics/ Science/Technology, Master's Degree in Science/Technology/Statistics

Work Experience: 0-1 years of work experience/internship in analytics roles

Analytics is a key occupation in the structure of the ITS Sub-Sector



Analytics excellent Vertical and Horizontal movements in their

tracks

Occupation	Tracks	Entry-level Job Roles
Analytics	MIS - Reporting Analytics - Modelling and Analysis	Associate - Analytics
Customer Relationship Management (CRM)	Customer Care (Non-voice) Customer Care (Voice) Sales/Telesales Technical Support/IT Help Desk Collections (Business to Customer)	Associate - Customer Care (Non-voice) Associate - CRM
Editorial and Desktop Publishing (DTP)	Editorial DTP and Design	Associate - Editorial Associate - DTP
Finance and Accounting (F&A)	Transaction Processing (Includes B2B Collections) Credit Analysis Audit and Accounting Financial Reporting Financial Planning and Analysis (Includes Budgeting and Forecasting)	Associate - Transactional F&A Associate - F&A Complex
Health Services	Clinical Data Management Medical Transcription	Associate - Clinical Data Management Associate - Medical Transcription
Human Resource Outsourcing (HRO)	Recruitment Learning and Development Compensation and Benefits Management Employee Relations	Associate - Recruitment Associate - HRO
Knowledge Services - Research	Secondary Research and Market Research Investment Banking Research	Analyst - Research
Learning	Content Managemnet Instructional Design	Associate - Learning
Legal Services	Legal Services	Document Coder/Processor Legal Associate
Supply Chain Management	Procurement Operations (Including Strategic Sourcing) Sales and Fulfilment (Including Inventory Mangement)	Associate - SCM

Movement to Other Occupations, Sub-sectors and Industries:

Given the dynamic range of services that the BPM sub-sector is increasingly offering to its clients in the industry, there are a variety of roles that employees are performing across the entire spectrum of offerings. As such they become a valuable asset not only to the BPM sub-sector, but also to all the client industries they are associated with.

Occupation	Horizontal Movements		
	To Other Occupations	To Other Sub-sectors	To Other Industries
Analytics	Knowledge Services - Research	ITS – Software Engineer	Banks, Retail, Insurance, Manufacturing
Customer Relationship Management	Supply Chain Management, Finance and Accounting (TP&FP)	IT Service Help Desk, SPD-Technical Support	Internal Technical Support in any industry
Editorial and DTP	Support roles for other occupations	ITS/SPD (Technical Writing)	Publishing, Learning
Finance and Accounting	Analytics, Knowledge Services - Research, Supply Chain Management	ITS, SPD (as SMEs)	Banks, Insurance Companies, Manufacturing, Retail
Health Services	Analytics, Knowledge Services - Research	ITS (as SMEs)	Labs, Pharma Companies
Human Resource Outsourcing	Knowledge Services - Research, Analytics	ITS, SPD (as SMEs)	HR Transactions in any industry
Learning	Analytics	ITS (IT Consulting)	Business Consulting
Knowledge Services - Research	Support roles for other occupations	ITS/SPD (Technical Writing)	Publishing, Learning
Legal Services	Human Resource Outsourcing	SPD – IPR	Banks, Insurance (Legal support)
Supply Chain Management (SCM)	Analytics, Knowledge Services - Research, Finance and Accounting	ITS (as SMEs)	Manufacturing, Retail

KA1. Job Role	KA2. Associate - Analytics (Business Analytics Associate/ Analyst)
KA3. Role Description	KA4. Responsible for building analytical packages using Databases, Excel or other Business Intelligence (BI) tools
KA5. NVEQF/NVQF level <small>KA6.</small>	KA10. 7
KA7. Minimum Educational Qualifications	KA11. Bachelor's Degree in Statistics/ Science/Technology or any other course
KA8. KA9. Maximum Educational Qualifications	KA12. Master's Degree in Science/Technology/Statistics or any other course
KA13. Training KA14. (Suggested but not mandatory)	KA15. Courses in SPSS, SAS, STATA and/or Spreadsheets KA16. RDBMS concepts, PL\SQL, OCA certification KA17. Financial and accounting terminologies in respective language & various accounting standards and GAAPs
KA18. Experience KA19.	KA20. 0-1 years of work experience/internship in analytics roles
KA21. Applicable National Occupational Standards (NOS)	<p>KA22. Compulsory:</p> <ol style="list-style-type: none"> SSC/ N 0703 (Create documents for knowledge sharing) SSC/ N 2101 (Carry out rule-based statistical analysis) SSC/ N 9001 (Manage your work to meet requirements) SSC/ N 9002 (Work effectively with colleagues) SSC/ N 9003 (Maintain a healthy, safe and secure working environment) SSC/ N 9004 (Provide data/information in standard formats) SSC/ N 9005 (Develop your knowledge, skills and competence) <p>KA23. KA24. Optional: KA25. Not Applicable</p>

KA26. Performance Criteria		KA27. As described in the relevant OS units		
Qualifications Pack Code		SSC/Q2101		
Job Role		Associate - Analytics This job role is applicable in both national and international scenarios		
Credits(NVEQF/NVQF/NSQF)		Version number	0.1	
Sector	IT-ITeS	Drafted on	30/04/13	
Sub-sector	Business Process Management	Last reviewed on	30/04/13	
Occupation	Analytics	Next review date	30/06/14	

OVERALL QUALIFICATION PACK DETAILS

SSC/ N0703 - Create Documents for knowledge sharing

Session Overview

In the Associate Analytics “Working with Documents”, the participant will learn about the most prominently used documentation techniques in corporate organizations. The Documentation types covered would include case studies, best practices, project artifacts, reports, minutes, policies, procedures, work instructions etc.

This session is NOT intended to cover technical documents or documents to support the deployment and use of products/applications, which are dealt with in different standards.

Session Goal

Participants should be able to have a good hands on understanding of MS Word and MS Visio, where there will be required to draft various documents/reports. The goal of the session is for the participant to be aware of the various documentation techniques which are used prominently in organizations.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish with **appropriate people** the purpose, scope, formats and target audience for the documents
- PC2. access existing documents, language standards, templates and documentation tools from your organization’s knowledge base

- PC3. liaise with **appropriate people** to obtain and verify the information required for the documents
- PC4. confirm the content and structure of the documents with **appropriate people**
- PC5. create documents using standard templates and agreed language standards
- PC6. review documents with **appropriate people** and incorporate their inputs
- PC7. submit documents for approval by **appropriate people**
- PC8. publish documents in agreed formats
- PC9. update your organization's knowledge base with the documents
- PC10. comply with your organization's policies, procedures and guidelines when creating documents for knowledge sharing

Note: The material for this NOS has been covered in the Associate Analytics Module 3 Book (book 3) in Unit 5

SSC/ N 2101 – Carry out rule-based statistical analysis

Session Overview

In the Associate Analytics *Carry out rule based statistical analysis*, the participants will go through Business Analytics using R tool. The participants will also learn Applied Statistical concepts like Descriptive Statistics and find their usage along with R. Furthermore, they will also have an overview of Big Data tools and their basic functioning.

Then they will learn about Machine Learning algorithm and their use in Data Mining and Predictive Analytics. Finally the participants will learn about Data Visualization and gather knowledge on Graphical representation of Data as well as results and reports.

Session Goal

The primary goal of the session is for the participants to learn the R tools and its various functions and features. Then also learn about Big Data tools and Big Data Analytics. Students will also learn about basic applied statistical concepts.

Session Objectives

To be competent, participants must be able to:

- PC1. establish clearly the objectives and scope of the **analysis**
- PC2. obtain guidance from **appropriate people** to identify suitable **data sources** to agree the methodological approach
- PC3. obtain and structure data using standard templates and tools
- PC4. validate data accurately and identify **anomalies**
- PC5. obtain guidance from **appropriate people** on how to handle **anomalies** in data
- PC6. carry out rule-based **analysis** of the data in line with the analysis plan
- PC7. validate the results of your **analysis** according to statistical guidelines
- PC8. review the results of your **analysis** with **appropriate people**
- PC9. undertake modifications to your **analysis** based on inputs from **appropriate people**
- PC10. draw justifiable inferences from your **analysis**
- PC11. present the results and inferences from your analysis using standard templates and tools
- PC12. comply with your organization's policies, procedures and guidelines when carrying out rule-based quantitative **analysis**

Note: The material for this NOS has been covered in all the three Modules of Associate Analytics

SSC/ N 9001: Manage Your Work to Meet Requirement

Session Overview

The Associate Analytics *Manage your work to meet requirement* module is designed to help participants understand the importance of time in a professional environment and how to manage multiple time bound requirements. It emphasizes on how time management is critical to work management and completing requirements/deliverables.

Participants learn how to manage work and how to ensure deliverables are completed in stipulated time in an organization by following tested principles to prevent/handle slippages on timelines. The module also emphasizes the need to respect time for self as well as colleagues.

Time management cannot override the qualitative aspect of the deliverable.

Session Goal

The primary goal of the session is for the participants to learn and manage time to be able to complete their work as required. The requirements of a work unit may be further classified into; activities, deliverable, quantity, standards and timelines. The session makes participants to be aware of defining requirements of every work unit and then ensuring delivery.

Additionally, this session discusses practical application of planning and execution of work plans to enable the participants to effectively deal with the failure points, minimize the impact, if any. Equally critical is the escalation plan and root cause analysis of exceptions.

Successful candidates will be able to understand the inter-relationship of time, effort, impact and cost.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Establish and agree your work requirements with appropriate people

PC2. Keep your immediate work area clean and tidy

PC3. Utilize your time effectively

PC4. Use resources correctly and efficiently

PC5. Treat confidential information correctly

PC6. Work in line with your organization's policies and procedures

PC7. Work within the limits of your job role

PC8. Obtain guidance from appropriate people, where necessary

PC9. Ensure your work meets the agreed requirements

Note: The material for this NOS has been covered in Unit 1 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9002: Work Effectively With Colleagues

Session Overview

The Associate Analytics *Work Effectively with Colleagues* module is designed to help participants understand the importance of teamwork in a professional environment. It emphasizes on how relationship management is critical to work management. It also focuses on the importance of personal grooming.

Participants learn how to manage cross functional relationships and how to nurture a good working environment. The module also stresses on the need to respect colleagues.

Session Goal

The primary goal of the session is for the participants to understand the importance of professional relationships with colleagues. Additionally, this session discusses importance of personal grooming.

Successful candidates will be able to understand the inter-relationship of professionalism and team-work.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

PC1. Communicate with colleagues clearly, concisely and accurately.

PC2. Work with colleagues to integrate your work effectively with theirs.

PC3. Pass on essential information to colleagues in line with organizational requirements.

PC4. Work in ways that show respect for colleagues.

PC5. Carry out commitments you have made to colleagues.

PC6. Let colleagues know in good time if you cannot carry out your commitments, explaining the reasons.

PC7. Identify any problems you have working with colleagues and take the initiative to solve these problems.

PC8. Follow the organization's policies and procedures for working with colleagues.

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9003: Maintain a Healthy, Safe and Secure working Environment

Session Overview

The Associate Analytics *Health, Safety and Security* module is designed to help participants understand the importance of following safety rules and regulations at workplace.

Participants learn how to work safely in an organization by following guidelines to prevent/handle any accidents or emergencies. The module also emphasizes the need of security and the entities that can pose a threat to it.

Session Goal

The primary goal of the session is for the participants to be aware about the various hazards that they may come across at workplace and what are the defined health, safety and security measures that should be followed at the time of occurrence of such unpredictable events.

Additionally, this session discusses practical application of the health and safety procedures to enable the participants to effectively deal with the hazardous events to minimize the impact, if any.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. Comply with your organization's current health, safety and security policies and procedures
- PC2. Report any identified breaches in health, safety, and security policies and procedures to the designated person
- PC3. Identify and correct any hazards that you can deal with safely, competently and within the limits of your authority
- PC4. Report any hazards that you are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected
- PC5. Follow your organization's emergency procedures promptly, calmly, and efficiently
- PC6. Identify and recommend opportunities for improving health, safety, and security to the designated person
- PC7. Complete any health and safety records legibly and accurately

Note: The material for this NOS has been covered in Unit 2 of Module 1. Much of the material herein is going to be self-study for the participants

SSC/ N 9004: Provide data/information in standard formats

Session Overview

The Associate Analytics *Provide data/information in standard formats* module is designed to help participants understand the standard operating procedures in organizations pertaining to reporting data in a logical sequence and arriving at conclusive decisions models after analysis of data. This module is aimed at developing the sense of understanding in an individual when the individual works with data, of how to take the data and present it as relevant information in standardized formats.

Participants learn how to share information with other people inside or outside a specified work group and also how to arrive at decisions regarding certain problem types.

Session Goal

The primary goal of the session is for the participants to analyze data and present it in a suitable format, as is suitable for the given process or organization.

Successful candidates will be able to understand the process of standardized reporting and the nuances of publishing a report with a specified end objective in mind.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. establish and agree with appropriate people the data/information you need to provide, the formats in which you need to provide it, and when you need to provide it
- PC2. obtain the data/information from reliable sources
- PC3. check that the data/information is accurate, complete and up-to-date
- PC4. obtain advice or guidance from appropriate people where there are problems with the data/information
- PC5. carry out rule-based analysis of the data/information, if required
- PC6. insert the data/information into the agreed formats
- PC7. check the accuracy of your work, involving colleagues where required
- PC8. report any unresolved anomalies in the data/information to appropriate people
- PC9. provide complete, accurate and up-to-date data/information to the appropriate people in the required formats on time

SSC/ N 9005: Develop your knowledge, skills and competence

Session Overview

The Associate Analytics *develop your knowledge, skills and competence* module is designed to help participants understand the importance of skill development in a professional environment and how to enhance skills in order to excel. It emphasizes on how enhance skills and knowledge in a diversified professional environment.

Session Goal

The primary goal of the session is to give a overview on how skills and competency can be enhanced in a professional environment. It gives knowledge on organizational context, technical knowledge, core skills/geneic skills, professional skills and technical skills. The session makes participants to understand the need of skills improvement for personal and organizational growth.

Successful candidates will be able ro understand the relationship between skill enhancement and growth.

Session Objectives

Upon completion of both parts of this course, the participants will be able to:

- PC1. obtain advice and guidance from appropriate people to develop their knowledge, skills and competence
- PC2. identify accurately the knowledge and skills you need for their job role
- PC3. identify accurately their current level of knowledge, skills and competence and any learning and development needs
- PC4. agree with appropriate people a plan of learning and development activities to address their learning needs
- PC5. undertake learning and development activities in line with their plan
- PC6. apply their new knowledge and skills in the workplace, under supervision
- PC7. obtain feedback from appropriate people on their knowledge and skills and how effectively they apply them
- PC8. review their knowledge, skills and competence regularly and take appropriate action

Overall Associate Analytics Content Structure

Module 1 – Book 1

Subject I / SSC NASSCOM - NOS- 2101, 9001, 9002	NOS	Hours	Minutes
Unit - 1	NOS 2101/9001		
Introduction to Analytics & R programing		6	360
Manage your work to meet requirements		4	240
Unit - 2	NOS 2101/9002		
Summarizing Data & Revisiting Probability		6	360
Work effectively with Colleagues		4	240
Unit - 3	NOS 2101		
SQL using R		9	510
Unit - 4	NOS 2101		
Correlation and Regression Analysis		9	510
Unit - 5	NOS 2101		
Understanding Verticals - Engg, Financial, others		6	390
Requirements Gathering		6	390
Total Hrs/Minutes		50	3000

Module 2 – Book 2

Subject II / SSC NASSCOM - NOS- 2010, 9003, 9004	NOS	Hours	Minutes
Unit - 1	NOS 2101/9003		
Data Management		7	420
Maintain Healthy, Safe & Secure Working environment		4	240
Unit - 2	NOS 2101/9004		
Big Data Tools		7	420
Provide Data/Information in Standard formats		4	240
Unit - 3	NOS 2101		
Big Data Analytics		8	480
Unit - 4	NOS 2101		
Machine Learning Algorithms		8	480
Unit - 5	NOS 2101		
Data Visualization		6	360
Product Implementation		6	360
Total Hrs/Minutes		50	3000

Module 3 – Book 3

Subject III / SSC NASSCOM - NOS - 0703, 2101, 9005	NOS	Hours	Minutes
Unit - 1	NOS 2101		
Introduction to Predictive Analytics		6	360
Linear Regression		6	360
Unit - 2	NOS 2101		
Logistics Regression		9	540
Unit - 3	NOS 2101/9005		
Objective Segmentation		6	360
Develop Knowledge Skill and competences		3	180
Unit - 4	NOS 2101		
Time Series Methods/Forecasting, Feature Extraction		5	300
Project		5	300
Unit - 5	NOS 0703		
Working with documents		10	600
Total Hrs/Minutes		50	3000

Glossary of Terms

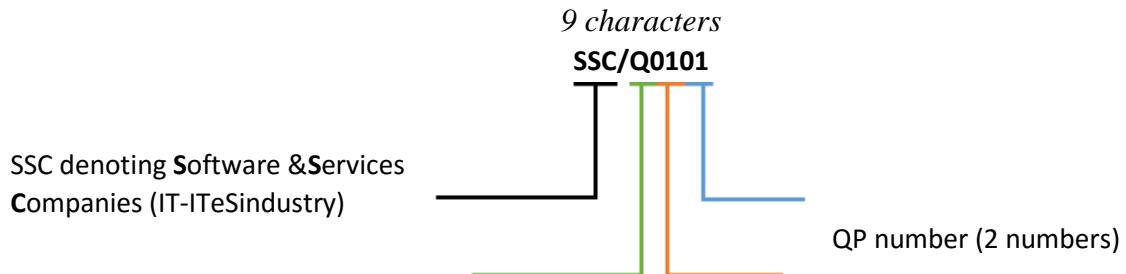
Definitions	Keywords /Terms	Description
	Sector	Sector is a conglomeration of different business operations having similar businesses and interests. It may also be defined as a distinct subset of the economy whose components share similar characteristics and interests.
	Sub-sector	Sub-sector is derived from a further breakdown based on the characteristics and interests of its components.
	Vertical	Vertical may exist within a sub-sector representing different domain areas or the client industries served by the industry.
	Occupation	Occupation is a set of job roles, which perform similar/related set of functions in an industry.
	Function	Function is an activity necessary for achieving the key purpose of the sector, occupation, or area of work, which can be carried out by a person or a group of persons. Functions are identified through functional analysis and form the basis of OS.
	Sub-functions	Sub-functions are sub-activities essential to fulfill the achieving the objectives of the function.
	Job role	Job role defines a unique set of functions that together form a unique employment opportunity in an organisation.
	Occupational Standards (OS)	OS specify the standards of performance an individual must achieve when carrying out a function in the workplace, together with the knowledge and understanding they need to meet that standard consistently. Occupational Standards are applicable both in the Indian and global contexts.
	Performance Criteria	Performance Criteria are statements that together specify the standard of performance required when carrying out a task.
	National Occupational Standards (NOS)	NOS are Occupational Standards which apply uniquely in the Indian context.
	Qualifications Pack Code	Qualifications Pack Code is a unique reference code that identifies a qualifications pack.
	Qualifications Pack(QP)	Qualifications Pack comprises the set of OS, together with the educational, training and other criteria required to perform a job role. A Qualifications Pack is assigned a unique qualification pack code.
	Unit Code	Unit Code is a unique identifier for an OS unit, which can be denoted with either an 'O' or an 'N'.
	Unit Title	Unit Title gives a clear overall statement about what the incumbent should be able to do.

Description	Description gives a short summary of the unit content. This would be helpful to anyone searching on a database to verify that this is the appropriate OS they are looking for.
Scope	Scope is the set of statements specifying the range of variables that an individual may have to deal with in carrying out the function which have a critical impact on the quality of performance required.
Knowledge and Understanding	Knowledge and Understanding are statements which together specify the technical, generic, professional and organisational specific knowledge that an individual needs in order to perform to the required standard.
Organisational Context	Organisational Context includes the way the organisation is structured and how it operates, including the extent of operative knowledge managers have of their relevant areas of responsibility.
Technical Knowledge	Technical Knowledge is the specific knowledge needed to accomplish specific designated responsibilities.
Core Skills/Generic Skills	Core Skills or Generic Skills are a group of skills that are key to learning and working in today's world. These skills are typically needed in any work environment. In the context of the OS, these include communication related skills that are applicable to most job roles.
Helpdesk	Helpdesk is an entity to which the customers will report their IT problems. IT Service Helpdesk Attendant is responsible for managing the helpdesk.
Keywords /Terms	Description
IT-ITeS	Information Technology - Information Technology enabled Services
BPM	Business Process Management
BPO	Business Process Outsourcing
KPO	Knowledge Process Outsourcing
LPO	Legal Process Outsourcing
IPO	Information Process Outsourcing
BCA	Bachelor of Computer Applications
B.Sc.	Bachelor of Science
OS	Occupational Standard(s)
NOS	National Occupational Standard(s)
QP	Qualifications Pack
UGC	University Grants Commission
MHRD	Ministry of Human Resource Development
MoLE	Ministry of Labour and Employment
NVEQF	National Vocational Education Qualifications Framework
NVQF	National Vocational Qualifications Framework
NSQF	National Skill Qualification Framework

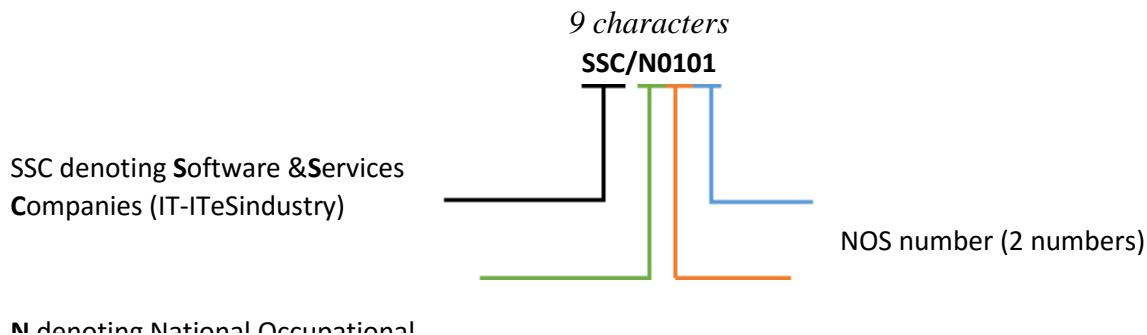
Nomenclature for QP & NOS

UNITS

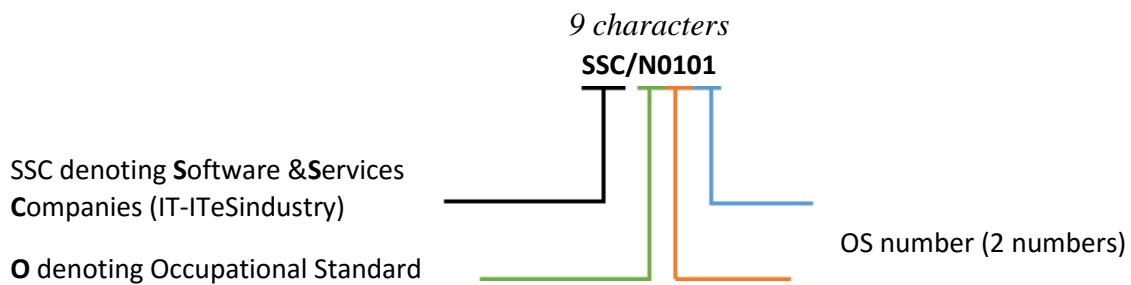
Qualifications Pack



National Occupational Standard



Occupational Standard



It is important to note that an OS unit can be denoted with either an '**O**' or an '**N**'.

- If an OS unit denotes '**O**', it is an OS unit that is an international standard. An example of OS unit denoting '**O**' is **SSC/O0101**.
- If an OS unit denotes '**N**', it is an OS unit that is a national standard and is applicable only for the Indian IT-ITeS industry. An example of OS unit denoting '**N**' is **SSC/N0101**

The following acronyms/codes have been used in the nomenclature above:

Sub-Sector	Range of Occupation numbers
IT Service (ITS)	01-20
Business Process Management (BPM)	21-40
Engg. and R&D (ERD)	41-60
Software Products (SPD)	61-80

Sequence	Description	Example
Three letters	Industry name (Software & Service Companies)	SSC
Slash	/	/
Next letter	Whether QP or NOS	N
Next two numbers	Occupation Code	01
Next two numbers	OS number	01

Module 3: Unit– 1.1

Introduction to Predictive Analytics and Linear Regression

Topic	Activities
<ol style="list-style-type: none"> 1. What and Why analytics 2. Introduction to tools and Environment 3. Application of Modeling in Business 4. Databases+Type of data and variables 5. Data Modeling Techniques Overview 6. Missing Imputations 	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Understand Basics of Predictive Analytics. 2. Understand Data types and Variable types. 3. Understand Basic Modeling. 4. And work on Missing Data.
Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
What and Why analytics	Classroom
Introduction to tools and Environment	Classroom
Application of Modeling in Business	Classroom
Databases+Type of data and variables	Classroom
Data Modeling Techniques Overview	Classroom
Missing Imputations	Classroom
Linear Regression	Classroom

Step-by-Step

Key Points

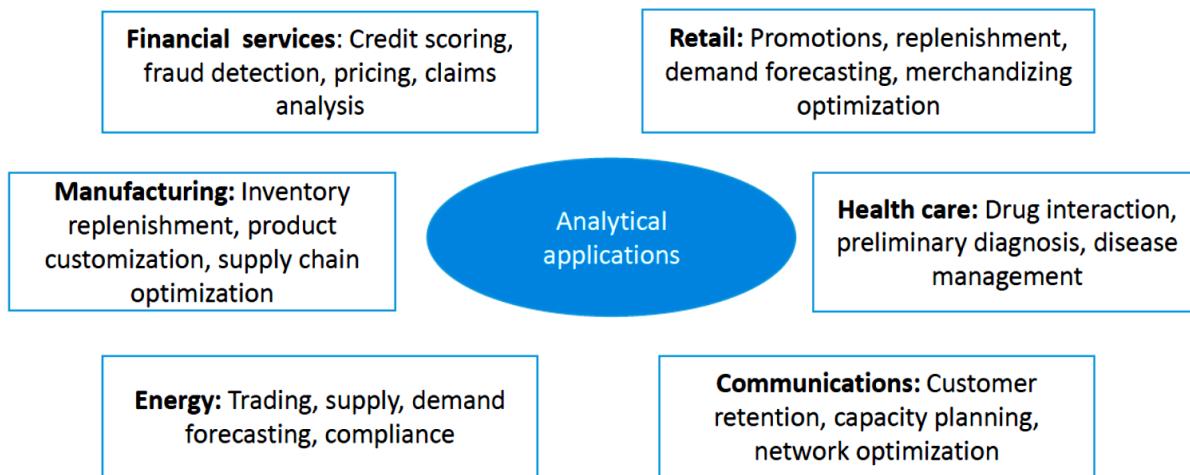
Introduction:

- ✓ Predictive Analytics is an art of predicting future on the basis of past trend.
- ✓ It is a branch of Statistics which comprises of Modeling Techniques, Machine Learning & Data Mining.
- ✓ Predictive Analytics is primarily used in Decision Making.

What and Why analytics:

Analytics is a journey that involves a combination of potential skills, advanced technologies, applications, and processes used by firm to gain business insights from data and statistics. This is done to perform business planning.

Places where Analytics is used:



Reporting Vs Analytics:

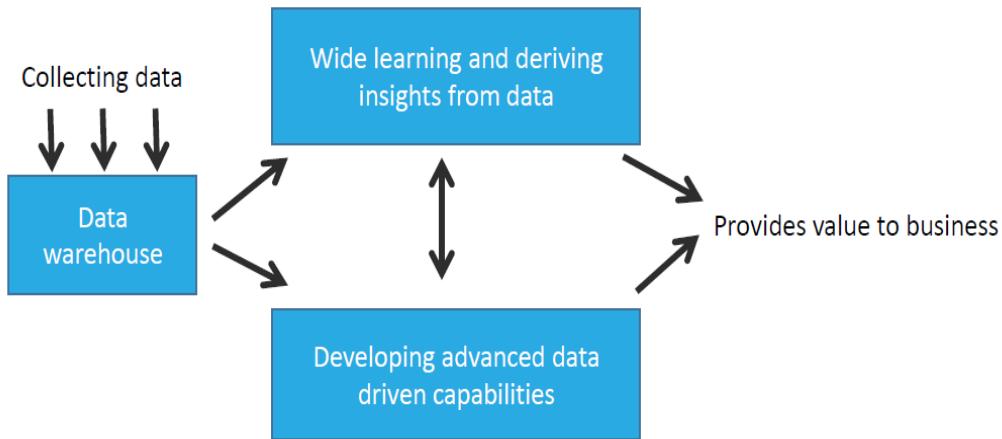
Reporting is presenting result of data analysis and Analytics is process or systems involved in analysis of data to obtain a desired output.

Introduction to tools and Environment:

- Analytics is now days used in all the fields ranging from Medical Science to Aero science to Government Activities.
- Data Science and Analytics are used by Manufacturing companies as well as Real Estate firms to develop their business and solve various issues by the help of historical data base.
- Tools are the softwares that can be used for Analytics like SAS or R. While techniques are the procedures to be followed to reach up to a solution.
- Various steps involved in Analytics:
 1. Access
 2. Manage
 3. Analyze
 4. Report



- Various Analytics techniques are:
 1. Data Preparation
 2. Reporting, Dashboards & Visualization
 3. Segmentation Icon
 4. Forecasting
 5. Descriptive Modeling
 6. Predictive Modeling
 7. Optimization



Application of Modeling in Business:

- A statistical model embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population.
- A model represents, often in considerably idealized form, the data-generating process.
- Signal processing is an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals.
- It uses mathematical, statistical, computational, heuristic, and linguistic representations, formalisms, and techniques for representation, modeling, analysis, synthesis, discovery, recovery, sensing, acquisition, extraction, learning, security, or forensics.
- In manufacturing statistical models are used to define Warranty policies, solving various conveyor related issues, Statistical Process Control etc.

Databases & Type of data and variables:

- A data dictionary, or metadata repository, as defined in the IBM Dictionary of Computing, is a "centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format".
- The term can have one of several closely related meanings pertaining to databases and database management systems (DBMS):
 - A document describing a database or collection of databases
 - An integral component of a DBMS that is required to determine its structure
 - A piece of middleware that extends or supplants the native data dictionary of a DBMS
- Data can be categorized on various parameters like Categorical, Type etc.
- Data is of 2 types – Numeric and Character. Again numeric data can be further divided into sub group of – Discrete and Continuous.
- Again, Data can be divided into 2 categories – Nominal and ordinal.
- Also based on usage data is divided into 2 categories – Quantitative and Qualitative
- Manufacturing industry also have their data divided in the groups discussed above. Like production quantity is a discrete quantity while production rate is a continuous data. Similarly quality parameter can be given ratings which ordinal data.

Data Modeling Techniques Overview:

- Regression analysis mainly focuses on finding a relationship between a dependent variable and one or more independent variables.
- Predict the value of a dependent variable based on the value of at least one independent variable.
- It explains the impact of changes in an independent variable on the dependent variable.

$$Y = f(X, \beta)$$

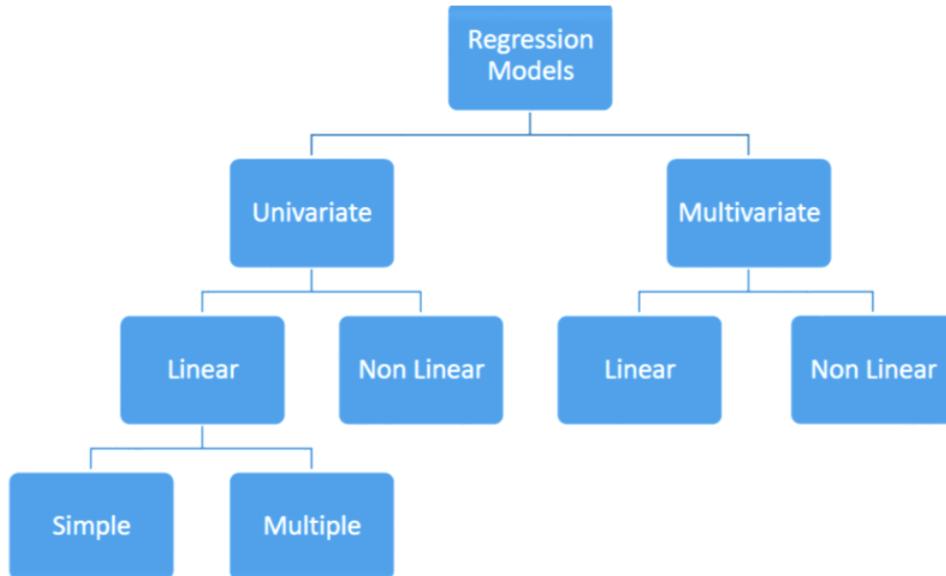
where Y is the dependent variable

X is the independent variable

β is the unknown coefficient

- Widely used in prediction and forecasting

Types of Regression model are as below:



Type of Regression	Conditions
Univariate	Only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one predictor variable
Multiple	Two or more predictor variables
Linear	All parameters enter the equation linearly, possibly after transformation of the data
Nonlinear	The relationship between the response and some of the predictors is nonlinear or some of the parameters appear nonlinearly, but no transformation is possible to make the parameters appear linearly
Analysis of variance	All predictors are qualitative variables
Analysis of covariance	Some predictors are quantitative variables and others are qualitative variables
Logistic	The response variable is qualitative

UNIT – 1.2

Session:2 Linear Regression

- It's a common technique to determine how one variable of interest is affected by another.
- Its used for three main purposes:
 - For describing the linear dependence of one variable on the other.
 - For prediction of values of other variable from the one which has more data.
 - Correction of linear dependence of one variable on the other.
- A line is fitted through the group of plotted data.

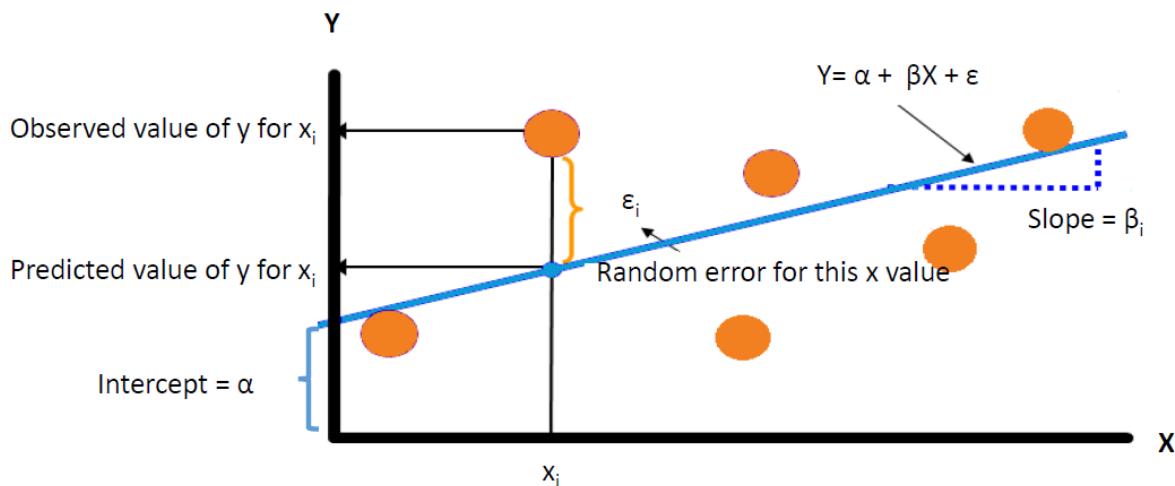
$$Y = \alpha + \beta X + \varepsilon$$

α = intercept coefficients

β = slope coefficients

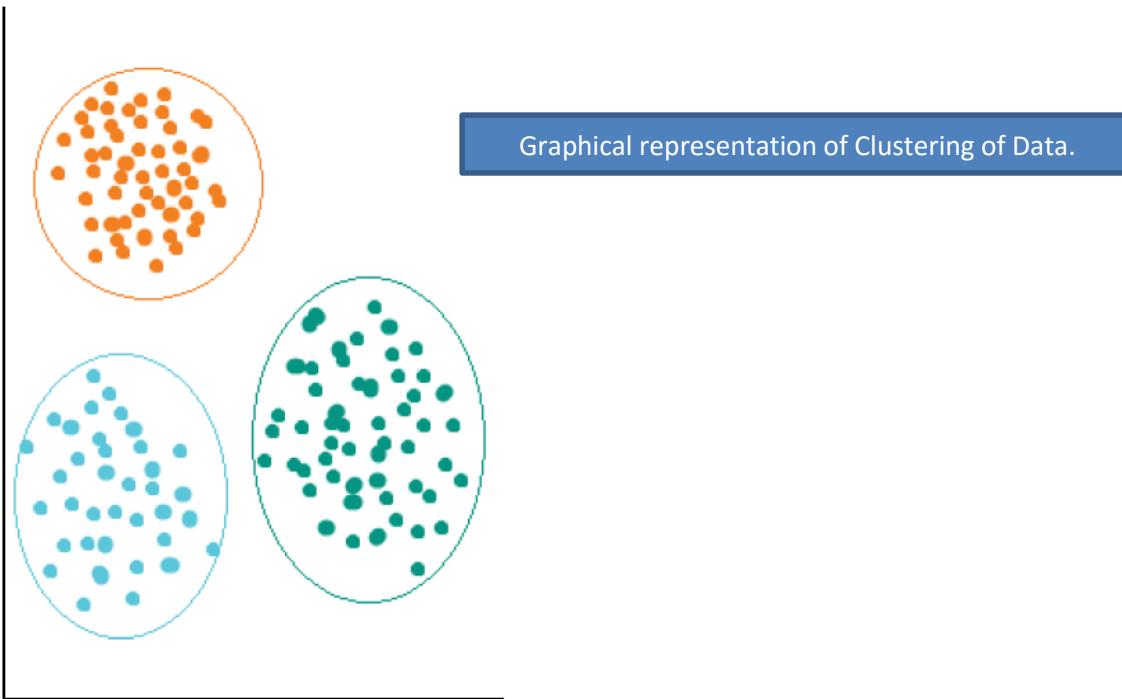
ε = residuals

- The residual value is a discrepancy between the actual and the predicted value.
- The distance of the plotted points from the line gives the residual value.
- The procedure to find the best fit is called the least-squares method.



➤ Cluster Analysis:

- Cluster Analysis is the process of forming groups of related variable for the purpose of drawing important conclusions based on the similarities within the group.
- The greater the similarity within a group and greater the difference between the groups, more distinct is the clustering.
- Often there are no assumptions about the underlying distribution of the data
- The reason for taking such an approach is that the objects in a group are similar to one another and are different from the objects in other groups. Therefore it is very easy to find pattern here.



➤ Time Series

- Time series data is an ordered sequence of observations on a quantitative variable measured over an equally spaced time interval.
- Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematic finance, weather forecasting, earthquake prediction electroencephalography, control engineering, astronomy, communications engineering and other places.

- Time series analysis is used in
 - Analyzing time series data
 - Forecasting the future value of the variable under consideration.
- In time series analysis it is assumed that the data consist of set of identifiable components and random errors which usually makes the pattern difficult to identify.

E.g. Sales of quilts and blankets in a store across a period of five years.

Missing Imputations:

- In **R**, missing values are represented by the symbol **NA** (not available). Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number). Unlike SAS, **R** uses the same symbol for character and numeric data.
- To test if there is any missing in the dataset we use *is.na()* function.

For Example,

We have defined “y” and then checked if there is any missing value. T or True means that there is a missing value.

```
y <- c(1,2,3,NA)
is.na(y)
# returns a vector (F FF T)
```

Arithmetic functions on missing values yield missing values.

For Example,

```
x <- c(1,2,NA,3)
mean(x)
# returns NA
```

To remove missing values from our dataset we use *na.omit()* function.

For Example,

We can create new dataset without missing data as below: -

```
newdata<- na.omit(mydata)
```

Or, we can also use “na.rm=TRUE” in argument of the operator. From above example we use *na.rm* and get desired result.

```
x <- c(1,2,NA,3)
```

```
mean(x, na.rm=TRUE)
```

returns 2

MICE Package -> Multiple Imputation by Chained Equations

MICE uses PMM to impute missing values in a dataset.

PMM-> Predictive Mean Matching (PMM) is a semi-parametric imputation approach. It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model.

Check Your Understanding



1. What is Linear Regression?
2. What are the steps involved in Analysis of Data?
3. What is the meaning of epsilon (ϵ) in regression equation?
4. What is cluster analysis of Data?
5. Is Clustering a Data Mining technique?
6. How to impute missing Data in R?

Summary

- Predictive modeling is a forecasting technique which uses software and applied statistics theories to predict future.
- Regression means looking into past and creating equation.
- $Y = \alpha + \beta X + \epsilon$ is the equation for Linear Regression Eqn.
- Cluster Analysis is grouping data on 1 common feature or criteria and establishing trend.

Select a data set of your own choice and you scatter plot to create cluster of data. Find the trend and present it in front of class.



Activity

Module 3: Unit– 2

Logistic Regression

Topic	Activities
1. Logistic Regression	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Understand Logistic Regression and its components 2. Understand Sigmoidal Function 3. Logistic Transformation and its components. 4. Execute variable transformation. 5. Understand Tableau visualization.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

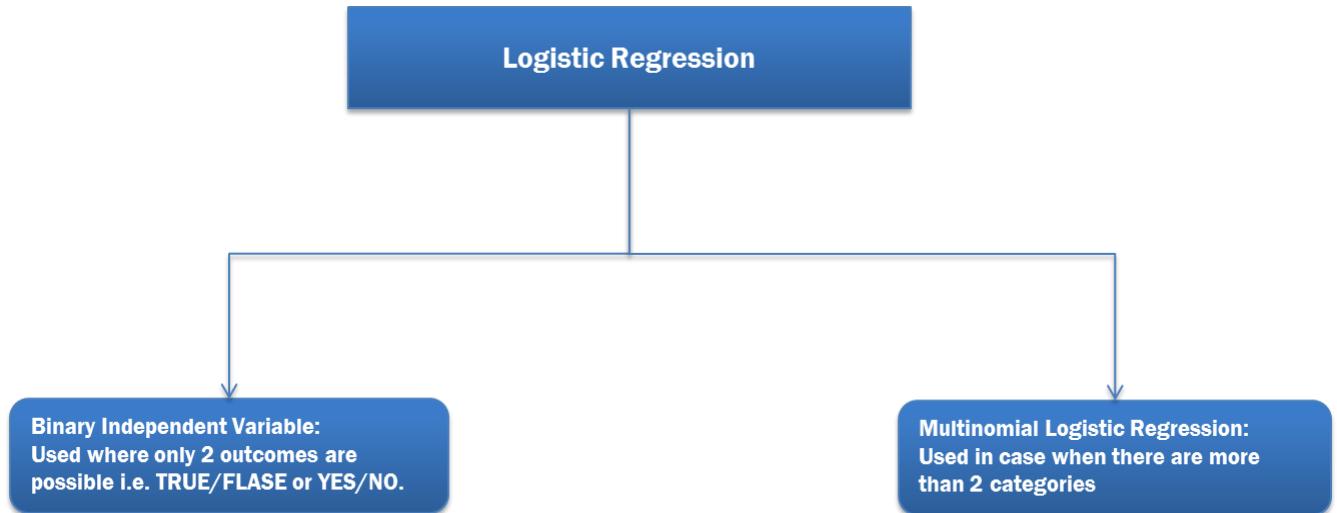
Session Plan:

Activity	Location
Model Theory	Classroom
The Classification problem	Classroom
The joint probability prediction	Classroom
Sigmoidal function	Classroom
The logistic transformation	Classroom
Likelihood function and Maximum likelihood estimation	Classroom
Difference between OLS & MLE	Classroom

Variable selection	Classroom
Interpretation of the model through maximum likelihood estimates	Classroom
Variable transformations/ Derived variables	Classroom
Introduction to data cleaning & Classing	Classroom
Deriving new variables from existing variables	Classroom
Model fit statistic	Classroom
Concordants and sommer's D	Classroom
Hosmer Lemeshow Test	Classroom
Error / confusion matrix	Classroom
Receiver operating characteristics	Classroom
Model conclusion- Tableau Visualization	Classroom
Check your understanding	Classroom
Summary	Classroom

Logistic Regression and its components:

- Logistic regression, or Logit regression, or Logit model is a regression model where the dependent variable (DV) is categorical.
- Logistic regression was developed by statistician David Cox in 1958.



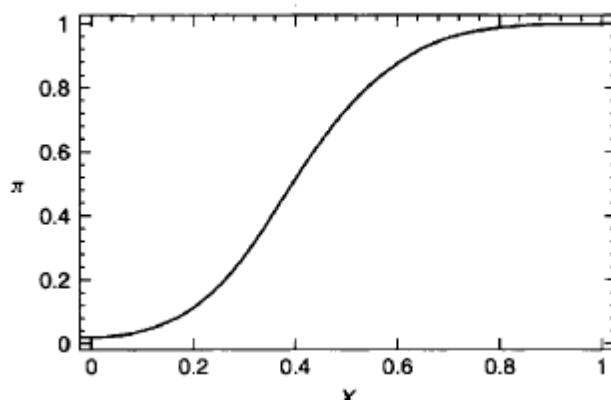
- In our discussion of regression analysis so far the response variable Y has been regarded as a continuous quantitative variable. The predictor variables, however, have been both quantitative, as well as qualitative. Indicator variables, which we have described earlier, fall into the second category.
- There are situations, however, where the response variable is qualitative. In this chapter we present methods for dealing with this situation. The methods presented in this chapter are very different from the method of least squares considered in earlier chapters. Consider a procedure in which individuals are selected on the basis of their scores in a battery of tests. After five years the candidates are classified as "good" or "poor." We are interested in examining the ability of the tests to predict the job performance of the candidates. Here the response variable, performance, is dichotomous. We can code "good" as 1 and "poor" as 0, for example. The predictor variables are the scores in the tests.
- In a study to determine the risk factors for cancer, health records of several people were studied. Data were collected on several variables, such as age, gender, smoking, diet, and the family's medical history. The response variable was the person had cancer ($Y = 1$) or did not have cancer ($Y = 0$).

$$\pi = \Pr(Y = 1|X = x) = \beta_0 + \beta_1 x.$$

The relationship between the probability π and X can often be represented by a logistic response function. It resembles an S-shaped curve. The probability π initially increases slowly with increase in X , and then the increase accelerates, finally stabilizes, but does not increase beyond 1. Intuitively this makes sense. Consider the probability of a questionnaire being returned as a function of cash reward, or the probability of passing a test as a function of the time put in studying for it.

The shape of the S-curve can be reproduced if we model the probabilities as follows:

$$\pi = \Pr(Y = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}},$$



- A sigmoid function is a bounded differentiable real function that is defined for all real input values and has a positive derivative at each point.
- It has an “S” shape. It is defined by below function:

$$S(t) = \frac{1}{1 + e^{-t}}.$$

- The process of linearization of logistic regression function is called Logit Transformation.

$$\ln\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

- Modeling the response probabilities by the logistic distribution and estimating the parameters of the model given below constitutes fitting a logistic regression. In logistic regression the fitting is carried out by working with the logits. The Logit transformation produces a model that is linear in the parameters. The method of estimation used is the maximum likelihood method. The maximum likelihood estimates are obtained numerically, using an iterative procedure.

$$\begin{aligned}\pi &= \Pr(Y = 1 | X_1 = x_1, \dots, X_p = x_p) \\ &= \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}.\end{aligned}$$

OLS and MLE:

OLS -> Ordinary Least Square

MLE -> Maximum Likelihood Estimation

The ordinary least squares, or OLS, can also be called the linear least squares. This is a method for approximately determining the unknown parameters located in a linear regression model. According to books of statistics and other online sources, the ordinary least squares is obtained by minimizing the total of squared vertical distances between the observed responses within the dataset and the responses predicted by the linear approximation. Through a simple formula, you can express the resulting estimator, especially the single regressor, located on the right-hand side of the linear regression model.

For example, you have a set of equations which consists of several equations that have unknown parameters. You may use the ordinary least squares method because this is the most standard approach in finding the approximate solution to your overly determined systems. In other words, it is your overall solution in minimizing the sum of the squares of errors in your equation. Data fitting can be your most suited application. Online sources have stated that the data that best fits the ordinary least squares minimizes the sum of squared residuals. “Residual” is “the difference between an observed value and the fitted value provided by a model.”

Maximum likelihood estimation, or MLE, is a method used in estimating the parameters of a statistical model, and for fitting a statistical model to data. If you want to find the height measurement of every basketball player in a specific location, you can use the maximum likelihood estimation. Normally, you would encounter problems such as cost and time constraints. If you could not afford to measure all of the basketball players’ heights, the maximum likelihood estimation would be very handy. Using the maximum likelihood estimation, you can estimate the mean and variance of the height of your subjects. The MLE would set the mean and variance as parameters in determining the specific parametric values in a given model.

Multinomial Logistic Regression

We have n independent observations with p explanatory variables. The qualitative response variable has k categories. To construct the logits in the multinomial case one of the categories is considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level. We will take category k as the base level in our description of the method. Since there is no ordering, it is apparent that any category may be labeled k. Let $7rj$ denote the multinomial probability of an observation falling in

the jth category. We want to find the relationship between this probability and the p explanatory variables, X₁, X₂, ..., X_p. The multiple logistic regression model then is

$$\ln \left(\frac{\pi_j(x_i)}{\pi_k(x_i)} \right) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}; \quad \begin{matrix} j = 1, 2, \dots, (k-1), \\ i = 1, 2, \dots, n. \end{matrix}$$

Since all the terms add to unity, this reduces to

$$\ln(\pi_j(x_i)) = \frac{\exp(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi})}{1 + \sum_{j=1}^{k-1} \exp(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi})},$$

For j = 1, 2, ..., (k - 1). The model parameters are estimated by the method of maximum likelihood. Statistical software is available to do this fitting.

Hosmer Lemeshow Test:

- The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models.
- It is used frequently in risk prediction models.
- The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population.
- The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values.
- Models for which expected and observed event rates in subgroups are similar are called well calibrated.
- The Hosmer–Lemeshow test statistic is given by:

$$H = \sum_{g=1}^G \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}.$$

Here O_g, E_g, N_g, and π_g denote the observed events, expected events, observations, predicted risk for the gth risk decile group, and G is the number of groups.

Error Matrix:

A confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

A **table of confusion** (sometimes also called a **confusion matrix**), is a table with two rows and two columns that reports the number of *false positives*, *false negatives*, *true positives*, and *true negatives*. This allows more detailed analysis than mere proportion of correct guesses (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the number of samples in different classes vary greatly). For example, if there were 95 cats and only 5 dogs in the data set, the classifier could easily be biased into classifying all the samples as cats. The overall accuracy would be 95%, but in practice the classifier would have a 100% recognition rate for the cat class but a 0% recognition rate for the dog class.

Assuming the confusion matrix above, its corresponding table of confusion, for the cat class, would be:

5 true positives (actual cats that were correctly classified as cats)	2 false positives (dogs that were incorrectly labeled as cats)
3 false negatives (cats that were incorrectly marked as dogs)	17 true negatives (all the remaining animals, correctly classified as non-cats)

Receiver Operating Characteristics:

A **receiver operating characteristic (ROC)**, or **ROC curve**, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or the sensitivity index d' , known as "d-prime" in signal detection and biomedical informatics, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as $(1 - \text{specificity})$. The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from $-\infty$ to $+\infty$) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability in x-axis.

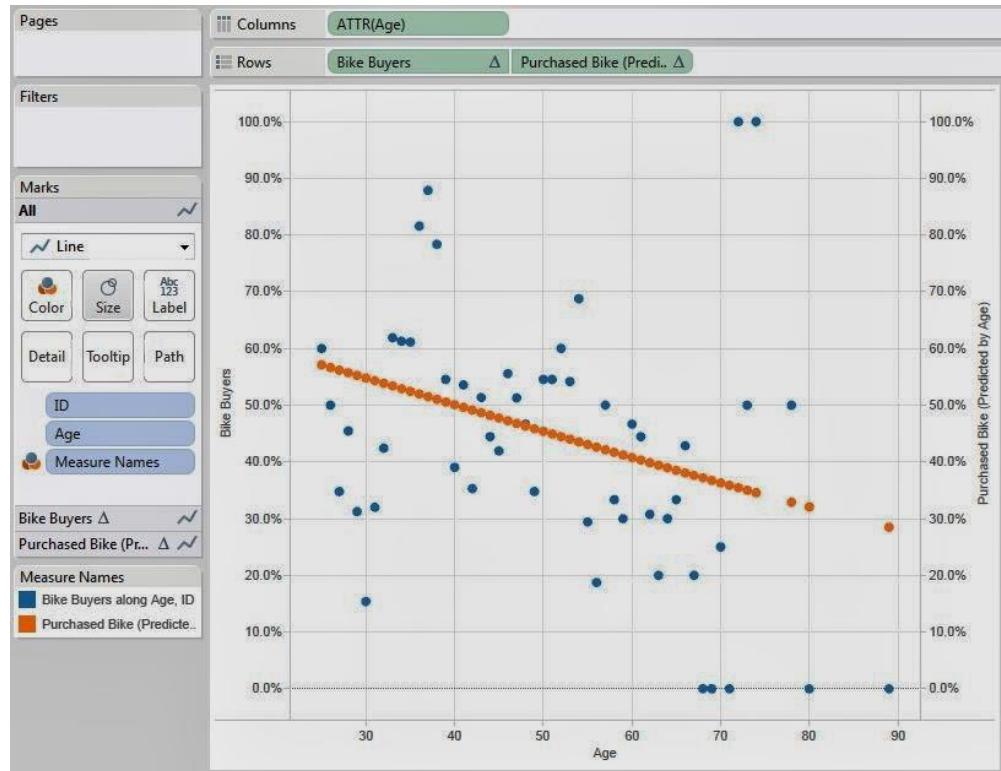
ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.

The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.

Data Visualization using Tableau:

- We use tools like Tableau to get graphical output from data predictive analytics results.
For example ,



Patient	RW	IR	SSPG	CC	Patient	RW	IR	SSPG	CC
1	0.81	124	55	3	46	0.91	106	56	3
2	0.95	117	76	3	47	0.95	118	122	3
3	0.94	143	105	3	48	0.95	112	73	3
4	1.04	199	108	3	49	1.03	157	122	3
5	1.00	240	143	3	50	0.87	292	128	3
6	0.76	157	165	3	51	0.87	200	233	3
7	0.91	221	119	3	52	1.17	220	132	3
8	1.10	186	105	3	53	0.83	144	138	3
9	0.99	142	98	3	54	0.82	109	83	3
10	0.78	131	94	3	55	0.86	151	109	3
11	0.90	221	53	3	56	1.01	158	96	3
12	0.73	178	66	3	57	0.88	73	52	3
13	0.96	136	142	3	58	0.75	81	42	3
14	0.84	200	93	3	59	0.99	151	122	2
15	0.74	208	68	3	60	1.12	122	176	3
16	0.98	202	102	3	61	1.09	117	118	3
17	1.10	152	76	3	62	1.02	208	244	2
18	0.85	185	37	3	63	1.19	201	194	2
19	0.83	116	60	3	64	1.06	131	136	3
20	0.93	123	50	3	65	1.20	162	257	2
21	0.95	136	47	3	66	1.05	148	167	2
22	0.74	134	50	3	67	1.18	130	153	3
23	0.95	184	91	3	68	1.01	137	248	3
24	0.97	192	124	3	69	0.91	375	273	3
25	0.72	279	74	3	70	0.81	146	80	3
26	1.11	228	235	3	71	1.10	344	270	2
27	1.20	145	158	3	72	1.03	192	180	3
28	1.13	172	140	3	73	0.97	115	85	3
29	1.00	179	145	3	74	0.96	195	106	3
30	0.78	222	99	3	75	1.10	267	254	3
31	1.00	134	90	3	76	1.07	281	119	3
32	1.00	143	105	3	77	1.08	213	177	2
33	0.71	169	32	3	78	0.95	156	159	3
34	0.76	263	165	3	79	0.74	221	103	3
35	0.89	174	78	3	80	0.84	199	59	3
36	0.88	134	80	3	81	0.89	76	108	3
37	1.17	182	54	3	82	1.11	490	259	3
38	0.85	241	175	3	83	1.19	143	204	2
39	0.97	128	80	3	84	1.18	73	220	3
40	1.00	222	186	3	85	1.06	237	111	2
41	1.00	165	117	3	86	0.95	748	122	2
42	0.89	282	160	3	87	1.06	320	253	2
43	0.98	94	71	3	88	0.98	188	211	2
44	0.78	121	29	3	89	1.16	607	271	2
45	0.74	73	42	3	90	1.18	297	220	2

Check your understanding



1. What is Logistic Regression?
2. Where is Logistic Regression concepts applied?
3. What is OLS?
4. What is MLE?
5. Why do we use data visualization?
6. Name some software for Data Visualization?

Summary

- Logistic Regression is an example of nonlinear regression equation.
- It can be bivariate as well as multivariate.
- The Logit transformation produces a model that is linear in the parameters.
- The method of estimation is called Maximum Likelihood Method.
- The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models.
- OLS -> Ordinary Least Square
- MLE -> Maximum Likelihood Estimation

Activity

Using the data on diabetes analyzed in Tables 1 :

- (a) Show that inclusion of the variable RW does not result in a substantial improvement in the classification rate from the multinomial logistic model using IR and SSPG.
- (b) Fit an ordinal logistic model using RW, IR, and SSPG to explain CC. Show that there is no substantial improvement in fit, and the correct



CASE STUDY

Detecting ailing financial and business establishments is an important function of audit and control. Systematic failure to do audit and control can lead to grave consequences, such as the savings-and-loan fiasco of the 1980s in the United States.

Table below gives some of the operating financial ratios of 33 firms that went bankrupt after 2 years and 33 that remained solvent during the same period.

A multiple logistic regression model is fitted using variables X 1, X 2, and X 3. Three financial ratios were

$$\begin{aligned}X_1 &= \frac{\text{Retained Earnings}}{\text{Total Assets}}, \\X_2 &= \frac{\text{Earnings Before Interest and Taxes}}{\text{Total Assets}}, \\X_3 &= \frac{\text{Sales}}{\text{Total Assets}}.\end{aligned}$$

Row	Y	X ₁	X ₂	X ₃	Row	Y	X ₁	X ₂	X ₃
1	0	-62.8	-89.5	1.7	34	1	43.0	16.4	1.3
2	0	3.3	-3.5	1.1	35	1	47.0	16.0	1.9
3	0	-120.8	-103.2	2.5	36	1	-3.3	4.0	2.7
4	0	-18.1	-28.8	1.1	37	1	35.0	20.8	1.9
5	0	-3.8	-50.6	0.9	38	1	46.7	12.6	0.9
6	0	-61.2	-56.2	1.7	39	1	20.8	12.5	2.4
7	0	-20.3	-17.4	1.0	40	1	33.0	23.6	1.5
8	0	-194.5	-25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	-4.3	1.0	42	1	68.6	13.8	1.6
10	0	-106.1	-22.9	1.5	43	1	37.3	33.4	3.5
11	0	-39.4	-35.7	1.2	44	1	59.0	23.1	5.5
12	0	-164.1	-17.7	1.3	45	1	49.6	23.8	1.9
13	0	-308.9	-65.8	0.8	46	1	12.5	7.0	1.8
14	0	7.2	-22.6	2.0	47	1	37.3	34.1	1.5
15	0	-118.3	-34.2	1.5	48	1	35.3	4.2	0.9
16	0	-185.9	-280.0	6.7	49	1	49.5	25.1	2.6
17	0	-34.6	-19.4	3.4	50	1	18.1	13.5	4.0
18	0	-27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	-48.2	6.8	1.6	52	1	21.5	-14.4	1.0
20	0	-49.2	-17.2	0.3	53	1	8.5	5.8	1.5
21	0	-19.2	-36.7	0.8	54	1	40.6	5.8	1.8
22	0	-18.1	-6.5	0.9	55	1	34.6	26.4	1.8
23	0	-98.0	-20.8	1.7	56	1	19.9	26.7	2.3
24	0	-129.0	-14.2	1.3	57	1	17.4	12.6	1.3
25	0	-4.0	-15.8	2.1	58	1	54.7	14.6	1.7
26	0	-8.7	-36.3	2.8	59	1	53.5	20.6	1.1
27	0	-59.2	-12.8	2.1	60	1	35.9	26.4	2.0
28	0	-13.1	-17.6	0.9	61	1	39.4	30.5	1.9
29	0	-38.0	1.6	1.2	62	1	53.1	7.1	1.9
30	0	-57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	-8.8	-9.1	0.9	64	1	59.5	7.0	2.0
32	0	-64.7	-4.0	0.1	65	1	16.3	20.4	1.0
33	0	-11.4	4.8	0.9	66	1	21.7	-7.8	1.6

Output of Logistic Regression using X₁, X₂ and X₃

Variable	Coeff.	s.e.	Z-Test	p-value	Odds	95% C.I.	
					Ratio	Lower	Upper
Constant	-10.15	10.84	-0.94	0.35			
X ₁	0.33	0.30	1.10	0.27	1.39	0.77	2.51
X ₂	0.18	0.11	1.69	0.09	1.20	0.97	1.48
X ₃	5.09	5.08	1.00	0.32	161.98	0.01	3.43 × 10 ⁶
Log-Likelihood = -2.906		G = 85.683		df = 3	p-value < 0.000		

The response variable is defined as

$$Y = \begin{cases} 0, & \text{if bankrupt after 2 years,} \\ 1, & \text{if solvent after 2 years.} \end{cases}$$

We now describe and interpret the output obtained from fitting a logistic regression. If π denotes the probability of a firm remaining solvent after 2 years, the fitted Logit is given by

$$\hat{g}(x_1, \dots, x_p) = -10.15 + 0.33 x_1 + 0.18 x_2 + 5.09 x_3.$$

This corresponds to the fitted regression equation in standard analysis. Here instead of predicting Y we obtain a model to predict the logits, $\log[\pi/(1 - \pi)]$. From the logits, after transformation, we can get the predicted probabilities. The constant and the coefficients are read directly from the second column in the table. The standard errors (s.e.) of the coefficients are given in the third column. The fourth column headed by Z is the ratio of the coefficient and the standard deviation. The Z is sometimes referred to as the Wald Statistic (Test). The Z corresponding to the coefficient of X2 is obtained from dividing 0.181 by 0.107. In the standard regression this would be the t-Test. This ratio for the logistic regression has a normal distribution as opposed to a t-distribution that we get in linear regression. The fifth column gives the p-value corresponding to the observed Z value, and should be interpreted like any p-value. These p-values are used to judge the significance of the coefficient. Values those are smaller than 0.05 would lead us to conclude that the coefficient is significantly different from 0 at the 5% significance level. From the p-values in Table 12.2, we see that none of the variables individually are significant for predicting the logits of the observations. In the standard regression output the regression coefficients have a simple interpretation. The regression coefficient of the jth predictor variable X_j is the expected change in Y for unit change in X_j when other variables are held fixed. The coefficient of X2 is the expected change in the logit for unit change in X2 when the other variables are held fixed. The coefficients of a logistic regression fit have another interpretation that is of major practical importance. Keeping X1 and X3 fixed, for unit increase in X2 the relative odds of

is multiplied by $e^{0.181} = 1.198$, that is, there is an increase of 20%. These values for each of the variables are given in the sixth column headed by Odds Ratio. They represent the change in odds ratio for unit change of a particular variable while the others are held constant. The change in odds ratio for unit change in variable X_j , while the other variables are held fixed, is e^{β_j} . If X_j was a binary variable, taking values 1 or 0, then e^{β_j} would be the actual value of the odds ratio rather than the change in the value of the odds ratio. The 95% confidence intervals of the

odds ratios are given in the last two columns of the table. If the confidence interval does not contain the value 1, the variable has a significant effect on the odds ratio. If the interval is below 1, the variable lowers significantly the relative odds. On the other hand, if the interval lies above 1, the relative odds is significantly increased by the variable. To see whether the variables collectively contribute in explaining the logits a test that examines whether the coefficients β_1, β_2, \dots etc. are all zero is performed. This corresponds to the case in multiple regression analysis where we test whether all the regression coefficients can be taken to be zero. The statistic G given at the bottom of the output table performs that task. The statistic G has a chi-square distribution. The p-value is considerably smaller than 0.05, and indicates that the variables collectively influence the logits.

Module 3: Unit– 3.1

Objective Segmentation

Topic	Activities
Objective segmentation	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Understand CHAID and CART 2. Understand how to build decision trees

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Session Plan:

Activity	Location
Regression vs. segmentation- supervised and unsupervised learning	Classroom
CHAID AND CART	Classroom
Impurity measures- Gini index and entropy, Information gain, decision tree algorithms	Classroom
Tree Building	Classroom
Multiple Decision Tree	Classroom
Check your understanding	Classroom
Summary	Classroom

What is Decision Tree:

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

CHAID:

CHAID stands for CHI-squared Automatic Interaction Detector.

Morgan and Sonquist (1963) proposed a simple method for fitting trees to predict a quantitative variable. They called the method AID, for Automatic Interaction Detection. The algorithm performs stepwise splitting. It begins with a single cluster of cases and searches a candidate set of predictor variables for a way to split this cluster into two clusters. Each predictor is tested for splitting as follows: sort all the n cases on the predictor and examine all $n-1$ ways to split the cluster in two. For each possible split, compute the within-cluster sum of squares about the mean of the cluster on the dependent variable. Choose the best of the $n-1$ splits to represent the predictor's contribution. Now do this for every other predictor. For the actual split, choose the predictor and its cut point which yields the smallest overall within-cluster sum of squares. Categorical predictors require a different approach. Since categories are unordered, all possible splits between categories must be considered. For deciding on one split of k categories into two groups, this means that $2k-1$ possible splits must be considered. Once a split is found, its suitability is measured on the same within-cluster sum of squares as for a quantitative predictor. Morgan and Sonquist called their algorithm AID because it naturally incorporates interaction among predictors. Interaction is not correlation. It has to do instead with conditional discrepancies. In the analysis of variance, interaction means that a trend within one level of a variable is not parallel to a trend within another level of the same variable. In the ANOVA model, interaction is represented by cross-products between predictors. In the tree model, it is represented by branches from the same nodes which have different splitting predictors further down the tree.



No interaction (left) and interaction (right) trees.

Regression trees parallel regression/ANOVA modeling, in which the dependent variable is quantitative. Classification trees parallel discriminant analysis and algebraic classification methods. Kass (1980) proposed a modification to AID called CHAID for categorized dependent and independent variables. His algorithm incorporated a sequential merge and split procedure based on a chi-square test statistic. Kass was concerned about computation time (although this has since proved an unnecessary worry), so he decided to settle for a sub-optimal split on each predictor instead of searching for all possible combinations of the categories. Kass's algorithm is like sequential cross-tabulation. For each predictor:

- 1) cross tabulate the m categories of the predictor with the k categories of the dependent variable,
- 2) find the pair of categories of the predictor whose $2 \times k$ sub-table is least significantly different on a chi-square test and merge these two categories;
- 3) if the chi-square test statistic is not “significant” according to a preset critical value, repeat this merging process for the selected predictor until no non-significant chi-square is found for a sub-table, and pick the predictor variable whose chi-square is largest and split the sample into subsets, where l is the number of categories resulting from the merging process on that predictor;
- 4) Continue splitting, as with AID, until no “significant” chi-squares result.

The CHAID algorithm saves some computer time, but it is not guaranteed to find the splits which predict best at a given step. Only by searching all possible category subsets can we do that. CHAID is also limited to categorical predictors, so it cannot be used for quantitative or mixed categorical-quantitative models.

CART:

CART stands for Classification And Regression Tree.

CART algorithm was introduced in Breiman et al. (1986). A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The CART growing method attempts to maximize within-node homogeneity. The extent to which a node does not represent a homogenous subset of cases is an indication of impurity. For example, a terminal node in which all cases have the same value for the dependent variable is a homogenous node that requires no further splitting because it is "pure." For

categorical (nominal, ordinal) dependent variables the common measure of impurity is Gini, which is based on squared probabilities of membership for each category. Splits are found that maximize the homogeneity of child nodes with respect to the value of the dependent variable.

Impurity Measure:

GINI Index

Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability f_i of each item being chosen times the probability $1-f_i$ of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

To compute Gini impurity for a set of items, suppose $i \in \{1, 2, \dots, m\}$, and let f_i be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i) = \sum_{i=1}^m (f_i - f_i^2) = \sum_{i=1}^m f_i - \sum_{i=1}^m f_i^2 = 1 - \sum_{i=1}^m f_i^2 = \sum_{i \neq k} f_i f_k$$

Tree Building:

Decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

There are many specific decision-tree algorithms. Notable ones include:

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHI-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
- MARS: extends decision trees to handle numerical data better.
- Conditional Inference Trees. Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid over fitting. This approach results in unbiased predictor selection and does not require pruning.

ID3 and CART were invented independently at around the same time (between 1970 and 1980), yet follow a similar approach for learning decision tree from training tuples.

Advantages of Decision Tree:

- Simple to understand and interpret. People are able to understand decision tree models after a brief explanation.

- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- Able to handle both numerical and categorical data. Other techniques are usually specialized in analysing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.)
- Uses a white box model. If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. (An example of a black box model is an artificial neural network since the explanation for the results is difficult to understand.)
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Robust. Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- Performs well with large datasets. Large amounts of data can be analyzed using standard computing resources in reasonable time.

Tools used to make Decision Tree:

Many data mining software packages provide implementations of one or more decision tree algorithms. Several examples include:

- Salford Systems CART (which licensed the proprietary code of the original CART authors)
- IBM SPSS Modeler
- Rapid Miner
- SAS Enterprise Miner
- Matlab
- R (an open source software environment for statistical computing which includes several CART implementations such as rpart, party and random Forest packages)
- Weka (a free and open-source data mining suite, contains many decision tree algorithms)
- Orange (a free data mining software suite, which includes the tree module orngTree)
- KNIME
- Microsoft SQL Server
- Scikit-learn (a free and open-source machine learning library for the Python programming language).

Real life example to understand Decision Tree:

Has it ever struck you just how many military terms have become everyday terms in business-speak? War and business are often compared and contrasted. As well as "fighting off threats" or "engaging in a price war", we talk about "gathering intelligence", "making a pre-emptive strike", and even trying to "out-maneuver" the competition.

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.

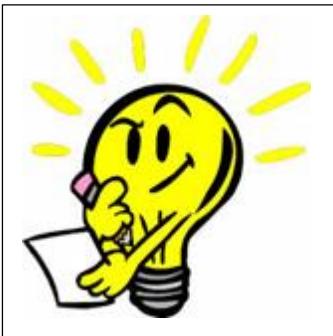
Boyd developed his model after analyzing the success of the American F-86 fighter plane compared with that of the Soviet MIG-15. Although the MIG was faster and could turn better, the American plane won more battles because, according to Boyd, the pilot's field of vision was far superior.

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.

Check your understanding



1. What is Decision Tree?
2. Why do we use Decision Tree?
3. What are various methods of creating a Decision Tree?
4. What is the difference between CHAID and CART?
5. Name some software to make Decision Tree.
6. What is the Measure of Impurity?

Summary

- Decision tree learning is a method commonly used in data mining.
- Decision trees used in data mining are of two types.
- CHAID stands for CHI-squared Automatic Interaction Detector.
- In the ANOVA model, interaction is represented by cross-products between predictors.
- CART stands for Classification and Regression Tree.
- Decision tree learning is the construction of a decision tree from class-labeled training tuples.
- Salford Systems CART is licensed the proprietary code of the original CART authors

Activity



Suppose you are running a coaching institute. Now you want to make steps for your employees to follow for better output and greater conversion of the enquiries.

Create a decision tree with assumptions and brainstorming.

Module 3: Unit– 3.2

Develop Knowledge, skills and competencies

Topic	Activities
Develop Knowledge, skills and competencies	<p>By the end of this session, you will be able to learn about:</p> <ol style="list-style-type: none"> 1. Knowledge, skills and competencies 2. Training and Development 3. Learning and Development policies and record keeping

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Classroom Session Map

Topic Description	Location
✓ Welcome and introduction session	✓ Classroom
✓ Let's get started	✓ Classroom
✓ Understanding knowledge, skills and competence	
✓ Importance of knowledge, skills and competence	
✓ Identifying training needs	✓ Classroom
✓ Work/task analysis and performance analysis	
✓ Feedback and Evaluation/Review of trainings	
✓ Sample L&D Policy - Genpact	✓ Classroom
✓ Individual record keeping	
✓ Continuous professional Development	

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided: reflect on your own experiences and determine when to share them.**
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.**
- ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.**
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussion questions independently, rather than providing answer. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Session1: Welcome and Introduction

Topic: Training, learning and development

Welcome the participants to the course and move to the introductions.

Introductions



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

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

Give a brief of your own experience and background.

Why are you here today? [Course Objectives]



“Why are you here today?”

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

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

Debrief the following:

The benefits of this course include:

- Needs of skills and knowledge enhancement
- Benefits of Training and development
- Training evaluation and its effectiveness

To fulfill these objectives today, we’ll be conducting a number of hands-on activities. Hopefully we can open up some good conversations and some of you can share your experiences so that we can make this session as interactive as possible. Your participation will be crucial to your learning experience and that of your peers here in the session today.

Knowledge, Skills & Competence

Key Points

Let's Get Started

Importance of knowledge management



Provide a brief overview of the session. Discuss the importance of knowledge, skills and development an individual's career standpoint.

Open up the discussion for the session and ask participants to share their thoughts on “why do you think skill development is important”?

Understanding knowledge, skills and competence

Knowledge – Mastery of facts, range of information in subject matter area.

Skills – Proficiency, expertise, or competence in given area; e.g., science, art, crafts

Competence– Demonstrated performance to use knowledge and skills when needed

➤ Some important definitions:



- **Interpersonal Skill:** Is aware of, responds to, and considers the needs, feelings, and capabilities of others. Deals with conflicts, confrontations, disagreements in a positive manner, which minimizes personal impact, to include controlling one's feelings and reactions. Deals effectively with others in both favorable and unfavorable situations regardless of status of position. Accepts interpersonal and cultural diversity.
- **Team Skill:** Establishes effective working relationships among team members. Participates in solving problems and making decisions.
- **Communications:** Presents and expresses ideas and information clearly and concisely in a manner appropriate to the audience, whether oral or written. Actively listens to what others are saying to achieve understanding. Shares information with others and facilitates the open exchange of ideas and information. Is open, honest, and straightforward with others.
- **Planning and Organizing:** Establishes courses of action for self to accomplish specific goals [e.g., establishes action plans]. Identifies need, arranges for, and obtains resources needed to accomplish own goals and objectives. Develops and uses tracking systems for monitoring own work progress. Effectively uses

Key Points

resources such as time and information.

- **Organizational Knowledge and Competence:** Acquires accurate information concerning the agency components, the mission[s] of each relevant organizational unit, and the principal programs in the agency. Interprets and utilizes information about the formal and informal organization, including the organizational structure, functioning, and relationships among units. Correctly identifies and draws upon source[s] of information for support.
- **Problem Solving and Analytical Ability:** Identifies existing and potential problems/issues. Obtains relevant information about the problem/issue, including recognizing whether or not more information is needed. Objectively evaluates relevant information about the problem/issue. Identifies the specific cause of the problem/issue. Develops recommendations, develops and evaluates alternative course of action, selects courses of action, and follows up.
- **Judgment:** Makes well-reasoned and timely decisions based on careful, objective review and informed analysis of available considerations and factors. Supports decisions or recommendations with accurate information or reasoning.
- **Direction and Motivation:** Sets a good example of how to do the job; demonstrates personal integrity, responsibility, and accountability. Provides advice and assistance to help others accomplish their work. Directs/motivates self.
- **Decisiveness:** Identifies when immediate action is needed, is willing to make decisions, render judgments, and take action. Accepts responsibility for the decision, including sustaining effort in spite of obstacles.
- **Self-Development:** Accurately evaluates own performance and identifies skills and abilities as targets of training and development activities related to current and future job requirements. Analyzes present career status. Sets goals [short and/or long term]. Identifies available resources and methods for self-improvement. Sets realistic time frames for goals and follows up.
- **Flexibility:** Modifies own behavior and work activities in response to new information, changing conditions, or unexpected obstacles. Views issues/problems from different perspectives. Considers a wide range of alternatives, including innovative or creative approaches. Strives to take actions that are acceptable to others having differing views.
- **Leadership:** Ability to make right decisions based on perceptive and analytical processes. Practices good judgment in gray areas. Acts decisively.

Key Points

Systematic Learning for the Individual

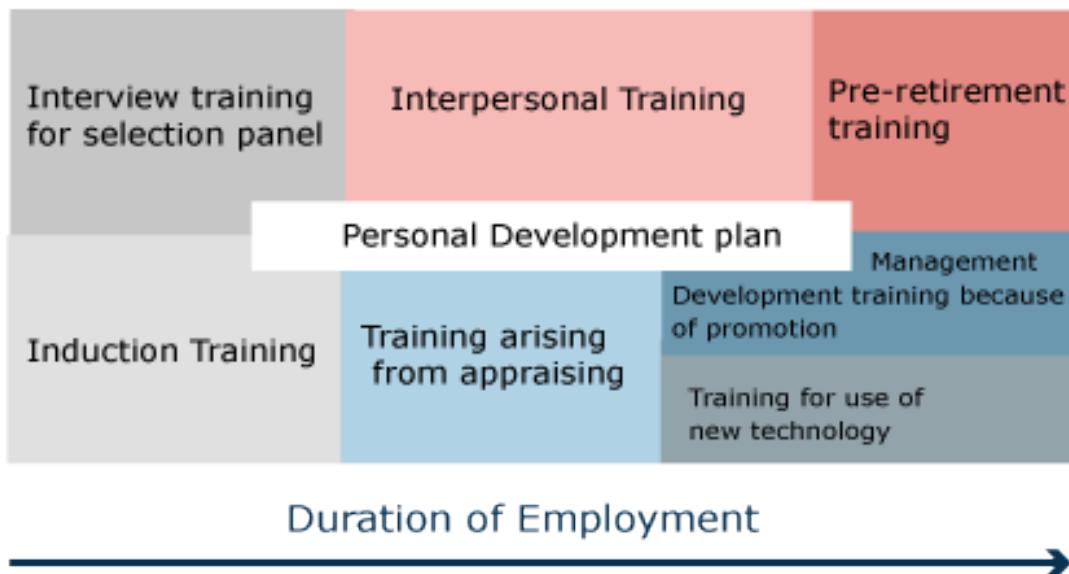


Fig : Systematic Learning for the Individual

Importance of Knowledge, Skills & Competence (KSC)

The primary purpose of KSC is to measure those qualities that will set one candidate apart from the others. KSC identify the better candidates from a group of persons basically qualified for a position. How well an applicant can show that he or she matches the position's defined KSAs determines whether that person will be seriously considered for the job.

Importance of developing skills:

More and more, job roles are requiring formal training qualifications either because of legislative requirements or to meet the requirements of specific employers.

Key Points

Developing your skills through further training provides significant benefits including:

Increased career development opportunities:

Developing a career in a chosen field is something many of us aspire to. Experience alone, in many cases does not suffice when employers are seeking to promote their staff. By undertaking further training, the opportunity to develop your career is enhanced.

Personal growth:

Training not only provides you with the skills in a particular area. By undertaking further training you build your networking, time management, communication and negotiation skills.



Increase your knowledge and understanding of the industry:

Trainings to know more about the industry & its development keeps the resource abreast with current industry trends & a better perspective to approach industry problems

Activity Description:

Make groups of 3-5 people and ask them to discuss and come up with ideas on how they would like to plan out their careers after they join an organization. The candidates will be required to create a career map showing where they stand in the organization and their individual career paths at 5 year intervals.

Check your Understanding



1. True or False? Personal skill development is equally important for an individual's career as is performing well in the organization
 - c. True
 - d. False

Key Points

Suggested Responses:

True, skills development is one of the most important things any fresh joinee in an organization needs to think about. Skills development helps out the individual in the long run.



2. True or False? After formal education is completed, one can lay free and doesn't need to engage in any additional self-training.
 - a. True
 - b. False

Suggested Responses:

False, one should never stop moving ahead in life; and one can move ahead in life only by continuous self improvement.

Summary



- **Knowledge** – Mastery of facts, range of information in subject matter area.
- **Skills** – Proficiency, expertise, or competence in given area; e.g., science, art, crafts
- **Competence** – Demonstrated performance to use knowledge and skills when needed
- More and more, job roles are requiring formal training qualifications either because of legislative requirements or to meet the requirements of specific employers.
- Training not only provides you with the skills in a particular area. By undertaking further training you build your networking, time management, and communication and negotiation skills.
- Trainings to know more about the industry & its development keeps the resource abreast with current industry trends & a better perspective to approach industry problems

Key Points

Training and Development

Key Points

Identifying Training Needs

Different methods are used by the organization to review skills and knowledge including:

- training need analysis
- skills need analysis
- performance appraisals

Training Need Analysis

- Training needs analysis is the first stage in training process and involves a procedure to determine whether training will indeed address the problem, which has been identified. Training can be described as “the acquisition of skills, concepts or attitudes that result in improved performance within the job environment”. Training analysis looks at each aspect of an operational domain so that the initial skills, concepts and attitudes of the human elements of a system can be effectively identified and appropriate training can be specified.
- Analysing what the training needs are is a vital prerequisite for an effective training programme or event. Simply throwing training at individuals may miss priority needs, or even cover areas that are not essential. TNA enables organisations to channel resources into the areas where they will contribute the most to employee development, enhancing morale and organizational performance. TNA is a natural function of appraisal systems and is key requirement for the award of Investors in People
- Training needs analysis involves:
 - monitoring current performance using techniques such as observation, interviews and questionnaires
 - anticipating future shortfalls or problems
 - identifying the type and level of training required and analysing how this can best be provided.

Key Points**Work / Task Analysis****Conducting a Work / Task Analysis**

- Interview subject matter experts (SME's) and high performing employees. Interview the supervisors and managers in charge. Review job descriptions and occupational information. Develop an understanding of what employees need to know in order to perform their jobs. Important questions to ask when conducting a Task Analysis:
 - What tasks are performed?
 - How frequently are they performed?
 - How important is each task?
 - What knowledge is needed to perform the task?
 - How difficult is each task?
 - What kinds of training are available?
- Observe the employee performing the job. Document the tasks being performed. When documenting the tasks, make sure each task starts with an action verb. How does this task analysis compare to existing job descriptions? Did the task analysis miss any important parts of the job description? Were there tasks performed that were omitted from the job description?
- Organize the identified tasks. Develop a sequence of tasks. Or list the tasks by importance.
- Are there differences between high and low performing employees on specific work tasks? Are there differences between Experts and Novices? Would providing training on those tasks improve employee job performance?
- Most employees are required to make decisions based on information. How is information gathered by the employee? What does the employee do with the information? Can this process be trained? Or, can training improve this process?

Key Points

Performance Analysis

- **Performance Analysis** is used to identify which employees need the training. Review performance appraisals. Interview managers and supervisors. Look for performance measures such as benchmarks and goals.
- Sources of performance data:
 - Performance Appraisals
 - Quotas met (un-met)
 - Performance Measures
 - Turnover
 - Shrinkage
 - Leakage
 - Spoilage
 - Losses
 - Accidents
 - Safety Incidents
 - Grievances
 - Absenteeism
 - Units per Day
 - Units per Week
 - Returns
 - Customer Complaints

Check your Understanding!



Are there differences between high and low performing employees on specific competencies? Would providing training on those competencies improve employee job performance?

Facilitator Notes: Yes, there can be significant differences between high and low performing employees on competencies, and that is why it is even more important to understand training needs more carefully!

Evaluation/Review of Trainings

Key Points

Evaluation of the impact of learning interventions may be carried out at a number of levels and involve a variety of factors:

Reaction: What did the participants think about the learning interventions? What did the providers think about the training interventions? What were their thoughts about the venue facilities?

Learning: What were the main areas which were remembered by the whole group of participants? What were the main areas which were forgotten by the whole group of participants?

Transfer: Which elements of the learning have been applied in the workplace? Which elements of the learning have not been applied in the workplace? Why do the participants apply some of the elements of the learning programme and not others?

Results: What were results of the changed work behavior? What effect did this have on productivity?

Return on Investment: What was the return on investment (ROI) of the training? How does the cost of training compare to the financial return on increased (decreased) productivity?

Key Points

Sl no	Questions	Agree	Neutral	Disagree
1	The objectives of the training were clearly defined			
2	Participation and interaction were encouraged			
3	The topics covered were relevant to me			
4	The content was organized and easy to follow			
5	The materials distributed were helpful			
6	This training experience will be useful in my work			
7	The trainer was knowledgeable about the training topics			
8	The trainer was well prepared			
9	The training objectives were met			
10	The time allotted for the training was sufficient.			
11	The meeting room and facilities were adequate and comfortable.			

Fig : Sample training feedback form

Donald L Kirkpatrick, Professor Emeritus, University Of Wisconsin (where he achieved his BBA, MBA and PhD), first published his ideas in 1959, in a series of articles in the Journal of American Society of Training Directors. The articles were subsequently included in Kirkpatrick's book Evaluating Training Programs (originally published in 1994; now in its 3rd edition - Berrett- Koehler Publishers)

Kirkpatrick's four levels of evaluation model

The four levels of Kirkpatrick's evaluation model essentially measure:

- reaction of student - what they thought and felt about the training
- learning - the resulting increase in knowledge or capability
- behaviour - extent of behaviour and capability improvement and implementation/application
- results - the effects on the business or environment resulting from the trainee's performance

Feedback

Feedback is an essential mean to understand and identify the right trainings & knowledge needed for the required job function.

What is a 360-degree feedback survey?

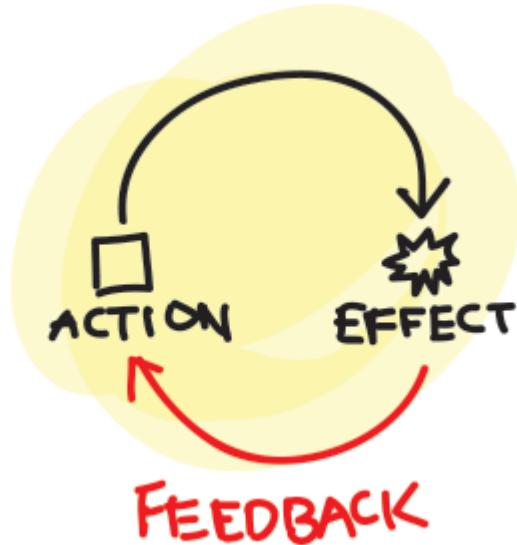
Key Points

One of the best feedback tools for professionals is 360-degree feedback it's also known as multi source feedback that comes from members of an employee's immediate work group. Most often, 360-degree feedback will include direct feedback from an employee's subordinates, supervisors and colleagues, as well as self-evaluation. In some cases it may also comprise feedback from external sources, such as customers, suppliers and other interested stakeholders which reveals how others perceive you.

It's used for planning and mapping specific paths in their development. In few organizations results are used in making administrative decisions related to pay and promotions. Usually it is best used for development "than evaluation".

How to go about 360-Degree Feedback survey?

To start the process 10-12 raters are to be pulled, out of which at least 6-8 (other than self) ratings must be obtained. Raters should offer confidential and anonymous feedback about the individuals. The accumulated report helps individuals to reflect and start working on their developmental aspects. Honest and realistic feedback will be much more valuable to the participants in their self-development. Each individual who receives feedback will then be encouraged to work on the development areas. It's advisable to re-run the survey after 9 months to know the progress and to know the extent of improvement.



The feedback includes parameters such as: job performance, behaviour at workplace, managerial effectiveness and skills like delegation, communication and team play. It also includes higher aspects like ethics, fairness, etiquette values, like professional courtesies. These are only an indicative list and we can customize the parameters for each organization.

A **Activity Description:**

Create a feedback form for a soft skills training. Identify what fields you will include and include a grading mechanism for the trainer on each parameter.

Key Points



Summary

- Different methods are used by the organizations to review skills and knowledge including:
 - training need analysis
 - skills need analysis
 - performance appraisals
- Training needs analysis is the first stage in training process and involves a procedure to determine whether training will indeed address the problem, which has been identified
- Performance Analysis is used to identify which employees need the training.
- Evaluation of the impact of learning interventions may be carried out at a number of levels and involve a variety of factors.

Learning and Development policies and record keeping

Key Points

Sample L&D Policy for Genpact



The Genpact logo features the word "GENPACT" in large, bold, blue capital letters.



The tagline "GENERATING IMPACT" is displayed in bold, dark gray capital letters, with a small "SM" superscript at the end of "IMPACT".

Genpact is completely committed towards continuous talent development, and our Learning and development framework is a key differentiator for us when it comes to employee retention. We have made significant investments in developing in-house capabilities in many training areas, both technical and non-technical, and have also partnered with several leading training providers, in order to ensure best-in-class training for our employees.

Our Learning & Development function delivers more than 6 million hours of training annually. The testimony to our commitment lies in a series of industry recognition that we have won over the years, such as recognition from American Society for Training & Development (ASTD) and multiple Brandon Hall Excellence in learning awards.

Training needs identification for each individual is done at the time of joining the organization / new process/ new role and during subsequent performance appraisals. Trainings provided cover all aspects of professional and personal development – business / process understanding, technical capabilities, domain knowledge, communication and interpersonal skills, and leadership potential development.

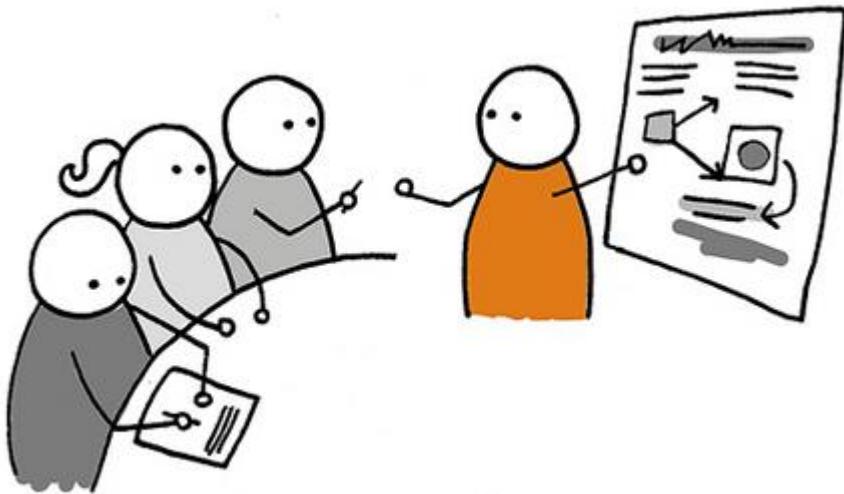
All new hires are required to attend a mandatory New Hire Orientation program, which familiarizes them with Genpact as an organization and its defining values, as well as various HR policies and processes.

This is followed by process and vertical specific new hire inductions to familiarize the new hire with the process and industry as well as provide overview of the work done in that space. In order to sensitize our employees with cultural nuances we also provide a certification program on cultural sensitivity, with modules for more than 130 countries in the world.

Key Points

For technical trainings, the focus is towards developing self-reliance and having internal experts to conduct trainings using case studies and practical examples as data points. This allows for imparting of technical trainings within our business context. In order to supplement the in-house trainer pool we also work extensively with reputed training providers to meet the training needs in a timely and effective manner.

Genpact also has a comprehensive leadership skills development curriculum, which focuses on each stage of an individual's professional growth, from the time the person starts leading a team for the first time, to gradually assuming greater responsibilities, both in terms of span as well as scope of work.



These programs are a mix of online modules, such as a suite of 42 elearnings from Harvard Business School (Harvard Manage Mentor®), and instructor-led classroom sessions.

We have a dedicated residential learning facility at Hyderabad where programs of longer duration are conducted centrally. For senior leadership programs, we have tied up with reputed providers of Executive Education such as Duke University, IIM Ahmedabad, Imparta and CapStone.



Check your Understanding!

What do you learn from the L&D policy for Genpact? Why do you think Genpact invests so heavily into resource upskilling and training

Facilitator Notes: This is because when a resource is hired, or is working in a specified process, there is always a continuous need to upskill the resource which would increase the productivity of the resource. Hence, organizations like to measure ROI on the resource trainings.

Key Points

Record Keeping

- As a professional, you have a responsibility to keep your skills and knowledge up to date. The learning & development helps you turn that accountability into a positive opportunity to identify and achieve your own career objectives.
 - At least once a year, we recommend you review your learning over the previous 12 months, and set your development objectives for the coming year. Reflecting on the past and planning for the future in this way makes your development more methodical and easier to measure. This is a particularly useful exercise prior to your annual appraisal!
 - Some people find it helpful to write things down in detail, while others record 'insights and learning points' in their diaries as they go along. This helps them to assess their learning continuously. These records and logs are useful tools for planning and reflection: it would be difficult to review your learning and learning needs yearly without regularly recording in some way your experiences.
- 
- Training is an investment that you make in yourself. It's a way of planning your development that links learning directly to practice. Trainings help you keep your skills up to date, and prepare you for greater responsibilities. It can boost your confidence, strengthen your professional credibility and help you become more creative in tackling new challenges. Trainings makes your working life more interesting and can significantly increase your job satisfaction. It can accelerate your career development and is an important part of upgrading to chartered membership.
 - It is strongly recommended that you maintain a personal portfolio. This will assist you in a number of key aspects related to your career:
 - ✓ You will be able to provide documented evidence of your commitment to your chosen profession; and of your continued competence

Key Points

- ✓ It will act as an excellent reference, both in the updating of your Curriculum Vitae and in recalling details of topics you have studied
- ✓ It will be a most useful aid in your career development, providing a means by which you can plan, record and review your relevant activities

Development record

NAME:		MEMBERSHIP NUMBER:	
COVERING THE PERIOD FROM:		TO:	

This record sheet is for your guidance only – you may present your development record in any other format.

Key dates	What did you do?	Why?	What did you learn from this?	How have/will you use this? Any further action?

Figures : A sample development record (top) and a sample development plan template (bottom)

Development plan

NAME:		MEMBERSHIP NUMBER:	
COVERING THE PERIOD FROM:		TO:	

This record sheet is for your guidance only – you may present your development plan in any other format.

Planned outcome

Where do I want to be by the end of this period? What do I want to be doing? (This may be evolutionary or "more of the same".)



What do I want/need to learn?	What will I do to achieve this?	What resources or support will I need?	What will my success criteria be?	Target dates for review and completion

Key Points

Continuous Professional Development (CPD) - Refers to the process of tracking and documenting the skills, knowledge and experience that you gain both formally and informally as you work, beyond any initial training. It's a record of what you experience, learn and then apply. The term is generally used to mean a physical folder or portfolio documenting your development as a professional.

CPD can help you to reflect, review and document your learning and to develop and update your professional knowledge and skills. It is also very useful to:

- ✓ provides an overview of your professional development to date
- ✓ reminds you of your achievements and how far you've progressed
- ✓ directs your career and helps you keep your eye on your goals
- ✓ uncovers gaps in your skills and capabilities
- ✓ Opens up further development needs
- ✓ provides examples and scenarios for a CV or interview
- ✓ demonstrates your professional standing to clients and employers
- ✓ helps you with your career development or a possible career change



How can you assess this? Answer these questions:

Where am I now? - Review and reflect on any learning experiences over the previous year or over the past three months. Write your thoughts down about what you learned, what insights it gave you and what you might have done differently. Include both formal training events and informal learning

Where do I want to be? - Write down your overall job skill requirements – immediate & in next 1-year.

What do I have to do to get there? - Make a note of what you need to do to achieve them. This could include further training, job or role progression or changes in direction.

For shorter term objectives, include the first step - what you can do today or tomorrow. For example, having a chat with your manager about a new responsibility or finding out about new technology from a colleague who has experience of it.

Key Points

Activity Description:

Create a sample development plan for your career after your hypothetical first year of employment in a large organization. Make sure the development plan is a result of development record which you've been maintaining throughout your tenure in the organization. Do create a template of development record as well.

Summary

- Every organization has detailed goals on learning and development needs, similar to the learning and development goals of Genpact, we saw in the sample policy.
- As a professional, you have a responsibility to keep your skills and knowledge up to date.
- At least once a year, we recommend you review your learning over the previous 12 months, and set your development objectives for the coming year.
- It is strongly recommended that you maintain a personal portfolio. This will assist you in a number of key aspects related to your career.
- **Continuous Professional Development** refers to the process of tracking and documenting the skills, knowledge and experience that you gain both formally and informally as you work, beyond any initial training.

Module 3: Unit– 4.1

Time series methods

Topic	Activities
Time series methods/ forecasting feature Extraction	<p>By the end of this session, you will be able to:</p> <ol style="list-style-type: none"> 1. Univariate stationary processes (ARMA) and forecasts 2. Univariate Non- stationary , integrated processes(ARIMA) and forecasts 3. Measures of forecast Accuracy 4. ETL Approach 5. Extract features from the generated model as height, Average, Energy etc. and analyze prediction

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

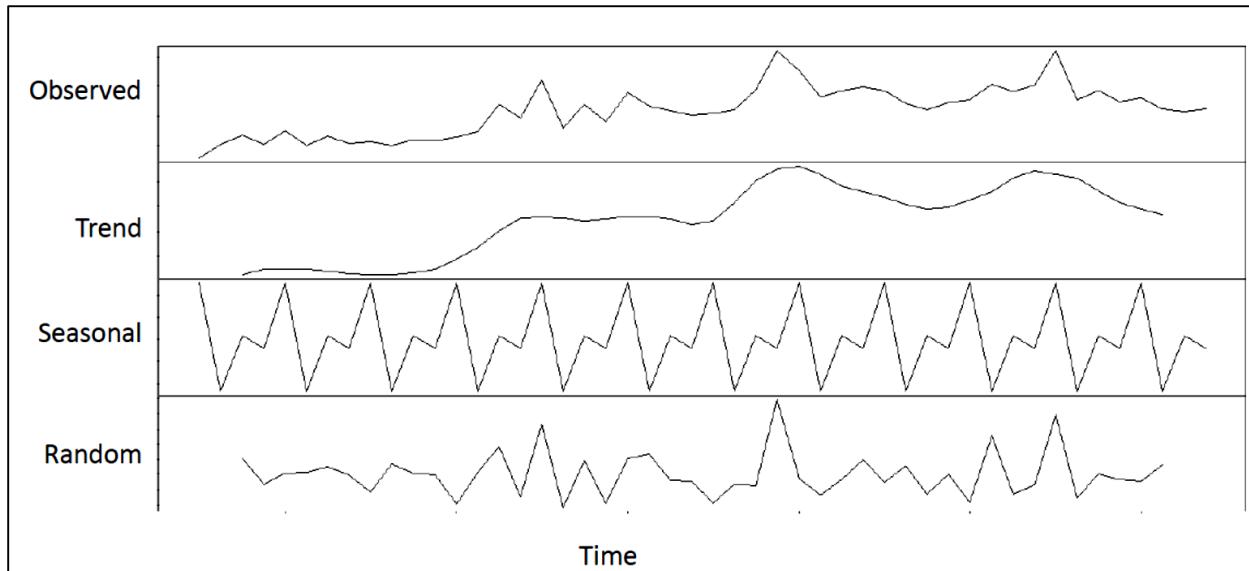
Session Plan:

Activity	Location
Univariate stationary processes (ARMA)	Classroom
Univariate Non- stationary processes(ARIMA)	Classroom
Measures of forecast Accuracy	Classroom
ETL Approach	Classroom
Check your understanding and Summary	Classroom

Step-by-Step

Components of Time Series

- Long term trend – The smooth long term direction of time series where the data can increase or decrease in some pattern.
- Seasonal variation – Patterns of change in a time series within a year which tends to repeat every year.
- Cyclical variation – It's much alike seasonal variation but the rise and fall of time series over periods are longer than one year.
- Irregular variation – Any variation that is not explainable by any of the three above mentioned components. They can be classified into – stationary and non – stationary variation.
- When the data neither increases nor decreases, i.e. it's completely random it's called stationary variation.
- When the data has some explainable portion remaining and can be analyzed further then such case is called non – stationary variation.



ARIMA & ARMA:

In time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series

(forecasting). They are applied in some cases where data show evidence of non-stationary, wherean initial differencing step (corresponding to the "integrated" part of the model) can be applied to reduce the non-stationary.

Non-seasonal ARIMA models are generally denoted ARIMA(p, d, q) where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is the order of the Moving-average model. Seasonal ARIMA models are usually denoted ARIMA(p, d, q)(P, D, Q)_m, where m refers to the number of periods in each season, and the uppercase P, D, Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model. ARIMA models form an important part of the Box-Jenkins approach to time-series modeling.

Univariate stationary processes (ARMA)

A covariance stationary process is an ARMA (p, q) process of autoregressive order p and moving average order q if it can be written as

$$\begin{aligned} y_t = & \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} \\ & + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \end{aligned}$$

For this process to be stationary the number of moving average coefficients q must be finite and the roots of the same characteristic equation as for the AR (p) process, must all lie inside the unit circle.

$$m^p - \phi_1 m^{p-1} - \cdots - \phi_p = 0$$

Measure of Forecast Accuracy:

Forecast Accuracy can be defined as the deviation of Forecast or Prediction from the actual results.

Error = Actual demand – Forecast

OR

$$e_i = A_t - F_t$$

We measure Forecast Accuracy by 2 methods :

1. Mean Forecast Error (MFE)

For n time periods where we have actual demand and forecast values:

$$MFE = \frac{\sum_{i=1}^n (e_i)}{n}$$

Ideal value = 0;

$MFE > 0$, model tends to under-forecast

$MFE < 0$, model tends to over-forecast

2. Mean Absolute Deviation (MAD)

For n time periods where we have actual demand and forecast values:

$$MAD = \frac{\sum_{i=1}^n |e_i|}{n}$$

While MFE is a measure of forecast model bias, MAD indicates the absolute size of the errors

Uses of Forecast error:

- Forecast model bias
- Absolute size of the forecast errors
- Compare alternative forecasting models
- Identify forecast models that need adjustment

ETL Approach:

Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that:

- Extracts data from homogeneous or heterogeneous data sources
- Transforms the data for storing it in proper format or structure for querying and analysis purpose
- Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)

Usually all the three phases execute in parallel since the data extraction takes time, so while the data is being pulled another transformation process executes, processing the already received data and prepares the data for loading and as soon as there is some data ready to be loaded into the target, the data loading kicks off without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The disparate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

Commercially available ETL tools include:

- Anatella
- Alteryx
- CampaignRunner
- ESF Database Migration Toolkit
- InformaticaPowerCenter
- Talend
- IBM InfoSphereDataStage
- Ab Initio
- Oracle Data Integrator (ODI)
- Oracle Warehouse Builder (OWB)
- Microsoft SQL Server Integration Services (SSIS)
- Tomahawk Business Integrator by Novasoft Technologies.
- Pentaho Data Integration (or Kettle) opensource data integration framework
- Stambia

- Diyotta DI-SUITE for Modern Data Integration
- FlyData
- Rhino ETL
- SAP Business Objects Data Services
- SAS Data Integration Studio
- SnapLogic
- Clover ETL opensource engine supporting only basic partial functionality and not server
- SQ-ALL - ETL with SQL queries from internet sources such as APIs
- North Concepts Data Pipeline

There are various steps involved in ETL. They are as below in detail:

Extract

The Extract step covers the data extraction from the source system and makes it accessible for further processing. The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible. The extract step should be designed in a way that it does not negatively affect the source system in terms of performance, response time or any kind of locking.

There are several ways to perform the extract:

- ❖ Update notification - if the source system is able to provide a notification that a record has been changed and describe the change, this is the easiest way to get the data.
- ❖ Incremental extract - some systems may not be able to provide notification that an update has occurred, but they are able to identify which records have been modified and provide an extract of such records. During further ETL steps, the system needs to identify changes and propagate it down. Note, that by using daily extract, we may not be able to handle deleted records properly.
- ❖ Full extract - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to be able to identify changes. Full extract handles deletions as well.
- ❖ When using Incremental or Full extracts, the extract frequency is extremely important. Particularly for full extracts; the data volumes can be in tens of gigabytes.

Clean

The cleaning step is one of the most important as it ensures the quality of the data in the data warehouse. Cleaning should perform basic data unification rules, such as:

- ❖ Making identifiers unique (sex categories Male/Female/Unknown, M/F/null, Man/Woman/Not Available are translated to standard Male/Female/Unknown)
- ❖ Convert null values into standardized Not Available/Not Provided value
- ❖ Convert phone numbers, ZIP codes to a standardized form
- ❖ Validate address fields, convert them into proper naming, e.g. Street/St/St./Str./Str
- ❖ Validate address fields against each other (State/Country, City/State, City/ZIP code, City/Street).

Transform

The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

Load

During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency.

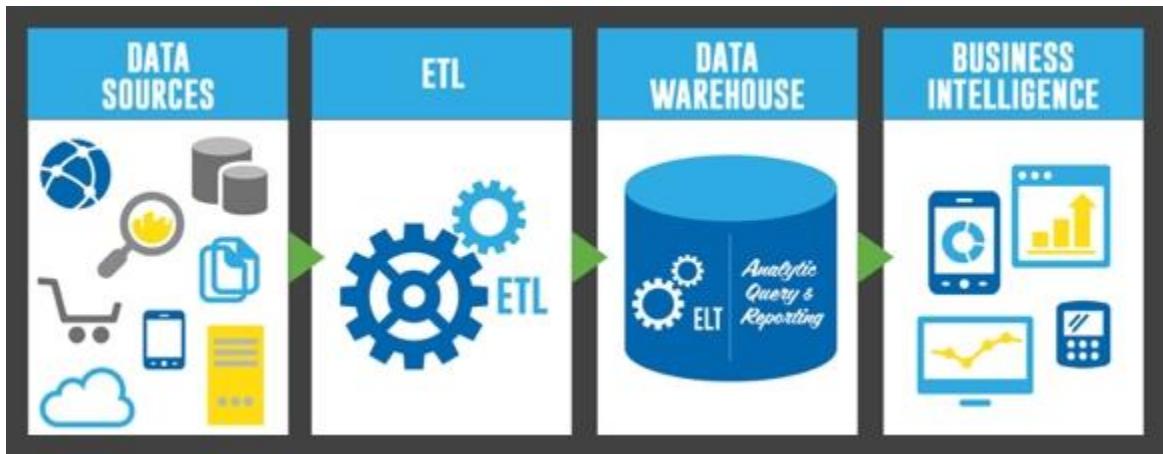
Managing ETL Process

The ETL process seems quite straight forward. As with every application, there is a possibility that the ETL process fails. This can be caused by missing extracts from one of the systems, missing values in one of the reference tables, or simply a connection or power outage. Therefore, it is necessary to design the ETL process keeping fail-recovery in mind.

Staging

It should be possible to restart, at least, some of the phases independently from the others. For example, if the transformation step fails, it should not be necessary to restart the Extract step. We can ensure this by implementing proper staging. Staging means that the data is simply dumped to the location (called the Staging Area) so that it can then be read by the next processing phase.

The staging area is also used during ETL process to store intermediate results of processing. This is ok for the ETL process which uses for this purpose. However, the staging area should be accessed by the load ETL process only. It should never be available to anyone else; particularly not to end users as it is not intended for data presentation to the end-user. May contain incomplete or in-the-middle-of-the-processing data.



Check your understanding



1. What is Time series?
2. What is ETL?
3. What is Measure of Forecast Accuracy?
4. What is the difference between ARMA and ARIMA?
5. What are the uses of Forecast error?

Summary

- There are 4 components of Time series.
- When the data neither increases nor decreases, i.e. it's completely random it's called stationary variation.
- When the data has some explainable portion remaining and can be analyzed further then such case is called non – stationary variation.
- ETL is Extract Transform and Load.
- Forecast Accuracy can be defined as the deviation of Forecast or Prediction from the actual results.
- In time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model.

Module 3: Unit– 4.2

Project

Case:- Analytics on Employee Health and Safety Data from USA Mining.

Source: USA Mining - Public Data (for 31 years) [1993 through 2013]

Introduction:

Data files on mining accidents, injuries, fatalities, employment, production, etc., are collected by the Mine Safety and Health Administration

(MSHA) under Part 50 (mandatory reporting of such incidents in mining) of the U.S. Code of Federal Regulations. Original raw data files are released periodically to the public on the MSHA web site. As a convenience, NIOSH has converted MSHA data to SPSS (includes labels and coding information) format.

We use this data to get two types of information:

1. Descriptive and Visualization
2. Prediction of injuries and fatalities

Tasks: Following are the tasks, learners have to complete

1. There are 30 files of yearly reported data in SPSS (.SAV) format, students have to combine them into one big file.
2. Learners have to clean and make this data file workable (no Missing Values, no NA's and no NaN's). Learners should apply "Imputation" method using appropriate statistical method (mean, median etc.)
3. Once the above tasks are completed, learners can check the various insights from the data such as the ones listed on MSHA website:

<http://www.cdc.gov/niosh/mining/statistics/allmining.html>

(please read carefully the adopted statistical analysis methodology
at <http://www.cdc.gov/niosh/mining/statistics/methodology.html>)

4. Summarizing the following:

- i. Number of mines with 5 or more death in a block of 5 years (e.g. 1983 - 88)
- ii. Year and mines with largest number of disaster (grouped as coal, iron etc.)
- iii. Year and mine with largest number of mine-disasters overall

(for more information and help

visit: <http://www.msha.gov/MSHAINFO/FactSheets/MSHAFCT8.HTM>)

5. Use historical death cases and predict the death for 5 years in future using ARIMA or any other suitable algorithm of your choice.

Course Conclusion



Ask the participants to recall key learning points from the session.

Map the Key Learning to the Course Objectives

Thank You Note

Module 3: Unit– 5

Working With Documents

Topic	Activities
Working with documents	<p>By the end of this session, you will be able to understand and learn:</p> <ol style="list-style-type: none"> 1. Procedures, Guidelines, Purpose & Scope of documentation. 2. Structure of documents 3. Tools for preparing Document.

Material and Handouts	
Facilitator Material	Participant Material and Handouts
Facilitator Guide, Handouts	✓ Participants' Guide

Course Goals	Location
<ul style="list-style-type: none"> ✓ Purpose for the documents ✓ Scope of documents ✓ Format of documents 	✓ Classroom
<ul style="list-style-type: none"> ✓ Case Study ✓ Technical reports. ✓ Client reports. ✓ Minutes of meeting. 	✓ Classroom
<ul style="list-style-type: none"> ✓ Visio ✓ PowerPoint ✓ MS-Word ✓ MS-Excel ✓ Case Study combining all tools 	<ul style="list-style-type: none"> ✓ Classroom ✓ Computer Laboratory

Facilitator Preparation

Responsibilities

- ✓ **Review examples provided: reflect on your own experiences and determine when to share them.**
- ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)**
- ✓ **Make sure you have copies of all the handouts.**
- ✓ **Make sure the learning resources are loaded on your computer.**
- ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.**
- ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.**
- ✓ **Make sure you create folders for all breakout activities.**

Principles of Facilitating

Personal Experiences

As a facilitator, you lead participants through prepared scenarios and discussions. During this process, relate your own professional experience to add realism. Often, personal experiences on how you helped a colleague through the career ownership process and guided them to achieving work satisfaction are more memorable than step-by-step instructions on following the career ownership process. Sharing experiences helps participants understand how professionals work and think, and gives them the opportunity to apply those lessons to their own work processes. Also, participants are more likely to remember answers if they have to think and explore on their own. Your goal is to foster independent thinking and action rather than having participants depend on your experience.

Experiential Learning

This workshop includes exercises designed to help participants discover the principles of guiding the participants through the career ownership process and career satisfaction. Encourage a free-wheeling discussion and call out important trends and insights. Make liberal use of the whiteboard to capture and display critical participant insights.

Socratic Questions

Your goal throughout the session is to guide participants towards thinking through the scenarios and discussing questions independently, rather than providing answer. For example:

Rather than saying...	Ask...
The Reality Check worksheet provides valuable information about how time is currently spent and what it would look like in the best case scenario.	What information can you gather from the Reality Check worksheet and how can the information be used to move towards career satisfaction?

Step-by-Step.

Key Points

Purpose of the documents

Provide a brief overview of the session.



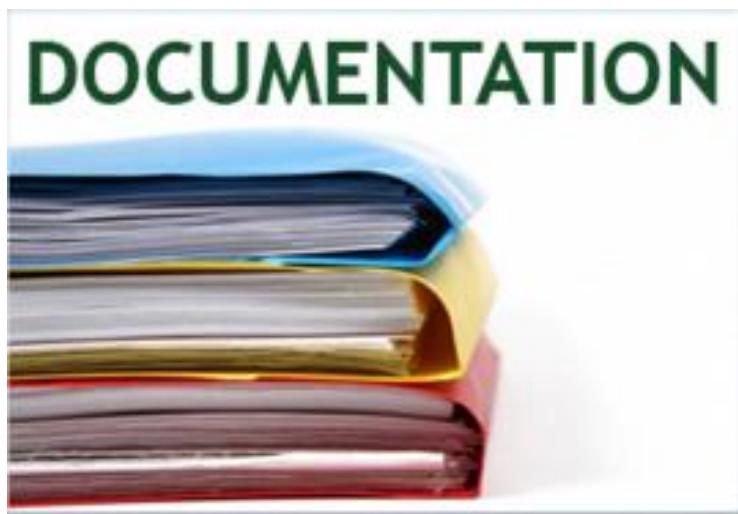
Open up the discussion for the session and ask participants to share their thoughts on:

- What is meant by documentation
- Why is documentation important

Suggested discussion:

While dealing with business processes, it is essential that one document the process as well as any improvements made to it. Most consultants will document both the “As-Is Process” as well as the “To-Be Process” (in case any changes are planned to the process). While many think about it as customary and do it for the same reasons, there are other important reasons to document the process. Documenting helps the organization gain long term primary and secondary benefits

Documentation is a set of documents provided on paper, or online, or on digital or analog media, such as audio tape or CDs. Examples are user guides, white papers, on-line help, quick-reference guides



Scope for the documents

A scope statement is one of the most critical pieces of a document, and writing one can be a difficult task for an individual – no matter what type of document management methodology is being used. But, an effectively written scope statement can help the rest of the document flow along with minimal problems. It is written after the document charter, and includes everything that the document is intended to produce.

A document's scope is usually used for three different reasons:

- Authorizing the document
- Providing a high level overview
- Identifying the main stakeholders

The scope often includes the name of the document owner as well as document sponsors. It also identifies objectives or goals, and constraints on resources or time. Finally, the scope is used as a focal point throughout the life of the document, which can be especially useful during change control meetings for minimizing scope creep. Here's a sample scope document

Format of Documents

Project Title: Information Technology (IT) Upgrade Project		
Project Start Date: March 4, 2007 Projected Finish Date: December 4, 2007		
Project Manager: Kim Nguyen, 691-2784, knguyen@course.com		
Project Objectives: Upgrade hardware and software for all employees (approximately 2,000) within nine months based on new corporate standards. See attached sheet describing the new standards. Upgrades may affect servers, as well as associated network hardware and software. Budgeted \$1,000,000 for hardware and software costs and \$500,000 for labor costs.		
Approach: <ul style="list-style-type: none"> ■ Update the information technology inventory database to determine upgrade needs ■ Develop detailed cost estimate for project and report to CIO ■ Issue a request for quote to obtain hardware and software ■ Use internal staff as much as possible for planning, analysis, and installation 		
ROLES AND RESPONSIBILITIES:		
NAME	ROLE	RESPONSIBILITY
Walter Schmidt	CEO	Project sponsor, monitor project
Mike Zwack	CIO	Monitor project, provide staff
Kim Nguyen	Project Manager	Plan and execute project
Jeff Johnson	Director of Information, Technology Operations	Mentor Kim
Nancy Reynolds	VP, Human Resources	Provide staff, issue memo to all employees about project
Steve McCann	Director of Purchasing	Assist in purchasing hardware and software

What are documentation formats?

A documentation format is a standard approach to the citation of sources that the author of a paper has consulted, abstracted, or quoted from. It prescribes methods for citing references within the text, providing a list of works cited at the end of the paper, and even formatting headings and margins.

Different organizations use different documentation formats; your instructor may require you to use a particular format, or may allow you use one of your choosing.

It is important to fully understand the documentation format to be used in your paper, and to apply it consistently. Furthermore, documentation formats allow you to give credit for secondary sources you have used in writing your paper.

Citing sources not only gives credit where it's due, but also allows your reader to locate the sources you have consulted. In short, the reader of your paper must be able to use the information you provide, both in the text and in appended list(s), to duplicate the research you have done.

What do I need to document?

In general, you must document information that originates in someone else's work. All of the following should be accompanied by a reference to the original:

- Direct quotations
- Paraphrases and summaries
- Information and ideas that are not common knowledge or are not available in a standard reference work
- Any borrowed material that might appear to be your own if there were no citation

Giving credit where it's due is a founding principle of academic inquiry, one that fosters the free exchange of ideas. Ultimately, you'll need to decide for yourself which ideas you can claim as your own and which should be attributed to others. Perhaps we should consider how we'd like our work to be credited, and use that as our guide.

How should I gather information for documenting sources?

You can make the process of applying any documentation format easier if you keep good notes while you perform research.

It's a good idea to put citations into your paper as you draft it. When you quote, put the source and page number directly after, perhaps marked with asterisks. When you refer, do the same. And when you place a citation in your text, add the source to your working bibliography.

When it comes time to put the finishing touches on your paper, the information you need will be available right in your text, and may be easily put into the proper format.

Which format should I use?

Choosing the appropriate documentation format for your paper may depend on three factors:

- The requirements of the particular course;
- The standard for the discipline in which you are studying; or
- Your individual preference.
- Documentation format required for a course

Your instructor may assign a documentation format for papers to be written for that course. This will often be indicated on the course syllabus or in the paper assignment, but may simply be mentioned during class. If no documentation format is prescribed, you should ask whether the instructor has a preference. If no preference is indicated, then you are free to choose a format.

What tools should I use for documentation?

- **Microsoft Word** : For all documents where a lot of textual documentation needs to be created
- **Microsoft Excel** : Where graphs/charts or spreadsheet based tables need to be created
- **Microsoft PowerPoint** : Where you need to create presentations with a combination of data types
- **Microsoft Visio** : Where flowcharts need to be created
- **Adobe PDF Format** : Any kind of document can be converted into un-editable PDF format by exporting a document type into PDF



QUIZ:

Which document formats are used for the specified applications as per the picture on your right?

Suggested Response: Word documents have the extension of .doc or .docx; Excel spreadsheets have the extension of .xls or .xlsx; PowerPoint presentations have the extension of .ppt or .pptx and Adobe PDF documents have the extension of .PDF

Summary

- .. Documenting helps the organization gain long term primary and secondary benefits
- Documentation is a set of documents provided on paper, or online, or on digital or analog media, such as audio tape or CDs. Examples are user guides, white papers, on-line help, quick-reference guides
- .. The scope is used as a focal point throughout the life of the document, which can be especially useful during change control meetings for minimizing scope creep
- . A documentation **format** is a standard approach to the citation of sources that the author of a paper has consulted, abstracted, or quoted from.

Step-by-Step.

What is a Case Study?

A **case study** is an account of an activity, event or problem that contains a real or hypothetical situation and includes the complexities you would encounter in the workplace

Here are the steps in writing your case study

1Determine which case study type, design or style is most suitable to your intended audience. Corporations may choose illustrative case study method to show what has been done for a client; schools, educators and students may select cumulative or critical case study method and legal teams may demonstrate exploratory (investigative) case study method as a way to provide factual evidence.

2Determine the topics of your case study. Once you've picked your angle, you need to determine what your research will be about and where it will take place (your case site). What have you talked about in class? Have you caught yourself coming up with questions during your reading?

3Set up interviews with subject matter experts (account managers in a corporation, clients and customers using applicable tools and services, etc.).

4Develop and write your case study using the data collected throughout the research, interviewing and analysis processes. Include at least four sections in your case study: an introduction, background information explaining why the case study was created, presentation of findings and a conclusion which clearly presents all of the data and references.

Format of a Sample Case Study:

Following is a suggested guideline for preparing your case study reports (remember to always use your specified corporate template for all formatting and referencing as per the course outline):

Cover Page

(Include employee names and employee Ids)

Executive Summary

(If appropriate – should be written last to focus on key points/findings)

Introduction

Current Situation Analysis and pertinent Background including a synopsis of the relevant information from the case analysis tool short form.

Body

May include:

- Target Market Identification
- Market Needs
- Analysis of Case
- Key Issues/Goals
- Recommendations

Should include:

- Decision Criteria
- Assumptions
- Data Analysis (analysis in appendix and summary info in body)
- Preferred Alternative with rationale.
- Justification/Predicted Outcome:

It is important that all guesstimates or creative ideas be founded upon some marketing rationale and a solid understanding of the metrics related to the target market and anticipated financial changes/impact. Using target market analysis and education estimation of population, \$, and units is appropriate.

Conclusion

References

Recommend that you source business journals, periodicals, and textual references as well as any online research. Make sure you support your ideas with facts and figures. Please try to use your own words and ideas based on research rather than copy and paste other's words from the internet. You should USE PROPER HARVARD style in-line citations, image source citations, and an alphabetical CITATION LIST in a references section.

Appendices

All charts, financials, visuals, and other related items can be placed here and referenced in the report.

Technical Reports and Client Reports

A technical report is a formal report designed to convey technical information in a clear and easily accessible format. It is divided into sections which allow different readers to access different levels of information. This guide explains the commonly accepted format for a technical report; explains the purposes of the individual sections; and gives hints on how to go about drafting and refining a report in order to produce an accurate, professional document.

Following is the structure of a technical report

Section	Details
Title page	Must include the title of the report. Reports for assessment, where the word length has been specified, will often also require the summary word count and the main text word count
Summary	A summary of the whole report including important features, results and conclusions
Contents	Numbers and lists all section and subsection headings with page numbers
Introduction	States the objectives of the report and comments on the way the topic of the report is to be treated. Leads straight into the report itself. Must not be a copy of the introduction in a lab handout.
The sections which make up the body of the report	Divided into numbered and headed sections. These sections separate the different main ideas in a logical order
Conclusions	A short, logical summing up of the theme(s) developed in the main text
References	Details of published sources of material referred to or quoted in the text (including any lecture notes and URL addresses of any websites used).
Bibliography	Other published sources of material, including websites, not referred to in the text but useful for background or further reading.
Acknowledgements	List of people who helped you research or prepare the report, including your proofreaders
Appendices (if appropriate)	Any further material which is essential for full understanding of your report (e.g. large scale diagrams, computer code, raw data, specifications) but not required by a casual reader

Candidates can see a sample technical report here:

<http://www3.ntu.edu.sg/home/cfcavallaro/Reports/Sample%20report.htm>

Client Reports

Client reports, simply put are reports that are designed for a client. These are always designed:

- As per the client template

- In line with periodicity defined by the client
- After double checking that all items to be included in the report have been included

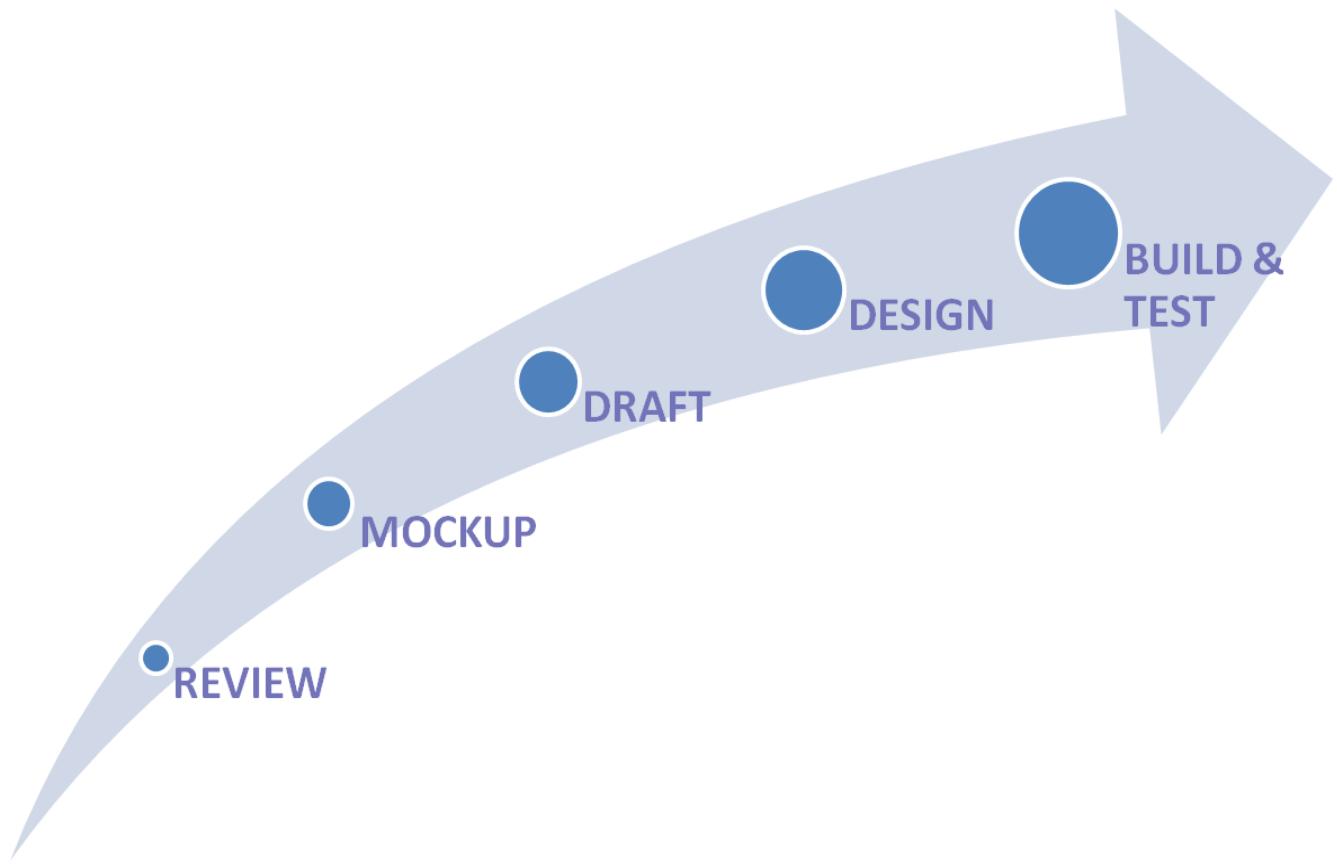
Client reports are one of the most important form of reports as they represent the services of the organization you work for, and utmost care needs to be taken to ensure that they come out correctly.

But these reports can be tedious to read, tedious to write, and can leave your client with a lot of unanswered questions, diminishing your hard-earned impact and undermining your recommendations. In this whitepaper, we present five ways to improve your client reporting strategy by using interactive data dashboards as a reporting tool:

- Author it once, use it every time.
- Ditch the long reports, and make your point clear with just a single screen.
- Empower clients to ask and answer their own questions.
- Give clients access to their data anywhere, anytime.
- Gain credibility by communicating your insights clearly.



Steps in creating a Client Report



Minutes of Meeting

Minutes are a tangible record of the meeting for its participants and a source of information for members who were unable to attend. In some cases, meeting minutes can act as a reference point, for example:

- when a meeting's outcomes impact other collaborative activities or projects within the organization
- minutes can serve to notify (or remind) individuals of tasks assigned to them and/or timelines

Using the MOM, it is important to capture the essence of the meeting, including details such as:

- Decisions made (motions made, votes, etc.)
- next steps planned
- identification and tracking of action items

Here are some important tips from International Association of Administrative Professionals (IAAP) on taking Minutes of Meeting

- *Be objective.*
- *Write in the same tense throughout*
- *Avoid using people's names except for motions or seconds. This is a business document, not about who said what.*
- *Avoid inflammatory or personal observations. The fewer adjectives or adverbs you use, the better.*
- *If you need to refer to other documents, attach them in an appendix or indicate where they may be found. Don't rewrite their intent or try to summarize them.*

Sample minutes of meeting document:

MINUTES OF MEETING

Meeting - 1.

Date :

Venue :

Attendees:

Minutes Taken By :

Issues	By	Discussion & Decision	Responsible	Deadline

Signature of Attendees: _____

Case Studies and Activities

Activity



Write a Case Study for the given project:

Organize the class into groups of 5 people and ask them to hold a meeting to collect data for a hypothetical study of its latest products in a few cities over a period of time. The group will be required to collect data in different cities in India. How need to collect data to measure success of the business model and draft a case study based on your findings.

Activity



Summary

is an account of an activity, event or problem that contains a real or hypothetical situation. It includes the complexities you would encounter in the workplace.

A report is a formal report designed to convey technical information in a clear and easily digestible format.

Reports are one of the most important forms of reports as they represent the services of the organization you work for, and utmost care needs to be taken to ensure that they come out well.

Meeting minutes are a tangible record of the meeting for its participants and a source of information for future reference.

Step-by-Step.

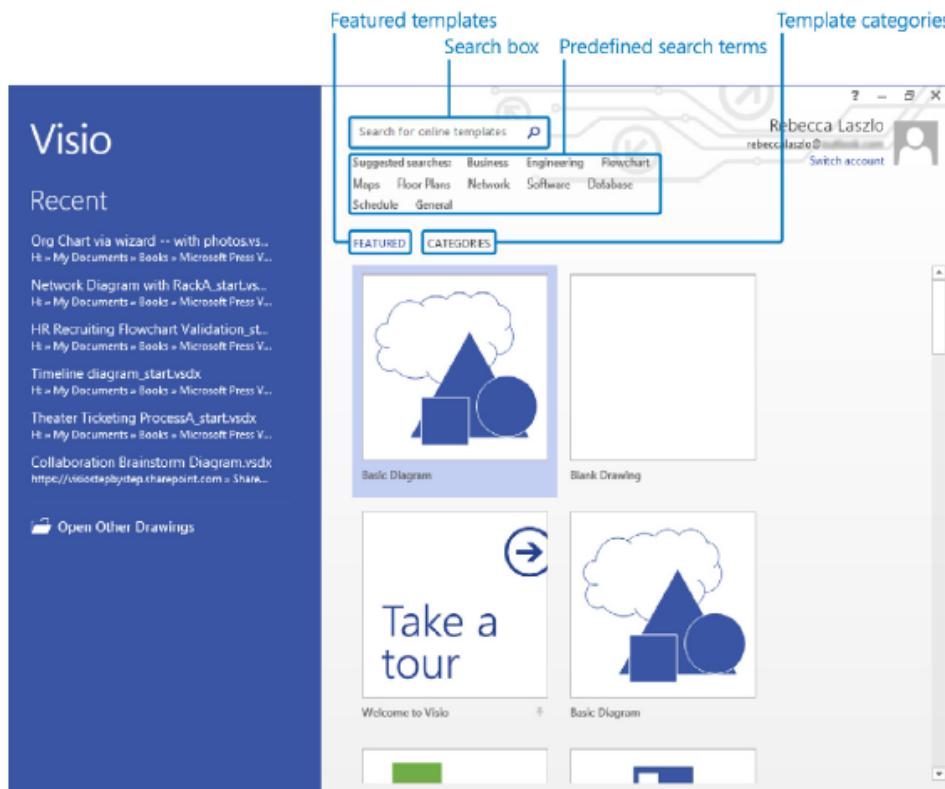


Facilitator Notes: For all the tools: Visio, MS Word, Excel and PowerPoint, please demonstrate the actions as mentioned in this book on the respective application, live in the class. The subject matter of this session is to be presented in a hands-on manner where the students will practice this alongside the facilitator

Microsoft Visio 2013

Microsoft Visio is the premier application for creating business diagrams of all types, ranging from flowcharts, network diagrams, and organization charts, to floor plans and brainstorming diagrams. You will often be required to create your flowcharts to document business processes. When this need arises, Microsoft VISIO is the best tool

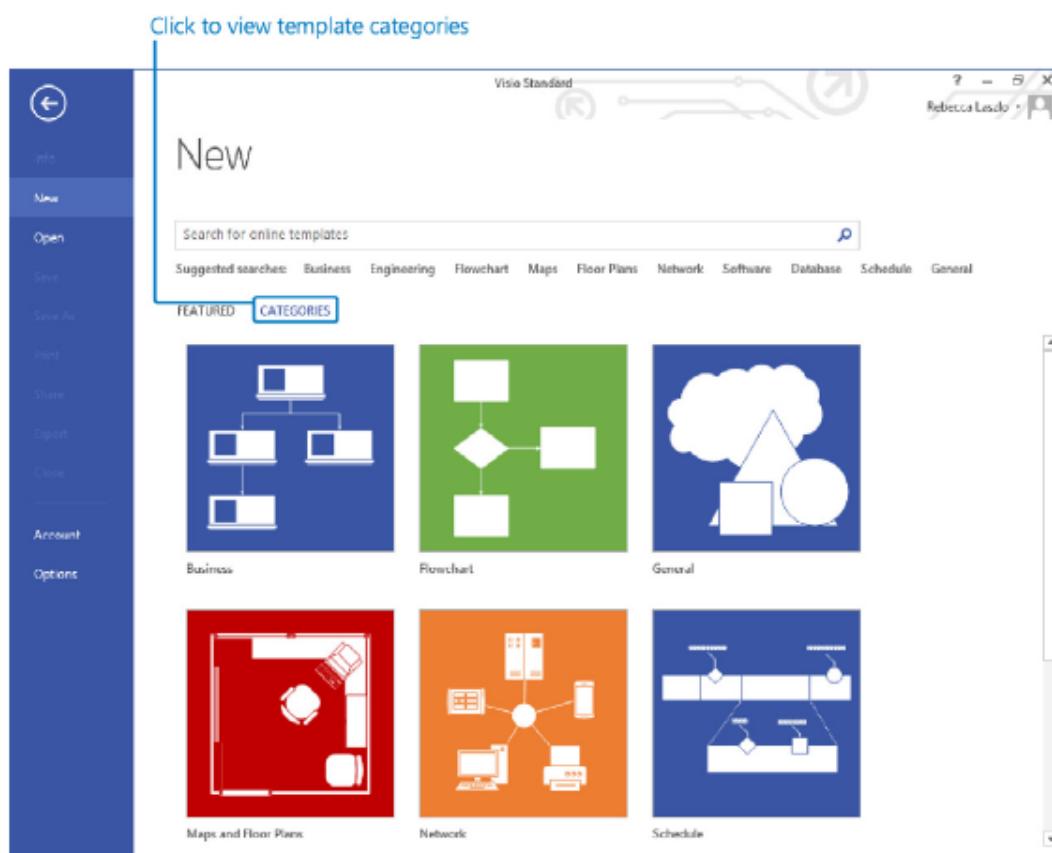
Getting Started with Visio 2013



Key sections of the start page are described in the following list:

- In the narrower left column is a list of recently opened diagrams. Clicking any diagram name opens it again.
If you want to open a diagram that is not on the recent list, click the Open Other Drawings button at the bottom of the list and Visio will take you to the Open page that is described in the next section.
- In the wider right column is a collection of thumbnails representing recently used or recommended templates.
- Above the template thumbnails are four important ways to find Visio templates.
 - You can type any words into the Search for online templates box and Visio will present templates that match your keywords.
 - You can click any word in the Suggested searches list to initiate an online search for matching templates.
- Featured** is the default selection for the template thumbnails that appear in the main part of the page (refer to the preceding graphic). The presentation of thumbnails is dynamic; the templates you use most frequently will rise to the top.
- Clicking **Categories** presents a set of template categories that are the same as the categories in previous versions of Visio: Business, Flowchart, General, Maps and Floor Plans, Network, and Schedule. The Professional edition also includes Engineering, and Software and Database categories.

Creating a New Document



Clicking any template category displays thumbnails for the diagrams in that category. If you click once on a diagram thumbnail, Visio displays information about that template. If you double-click a diagram thumbnail, Visio launches a new diagram.

Commonly used terminologies in Visio:

MasterAn object in a Visio stencil. The vast majority of people who create diagrams with Visio use the masters that ship with Visio or that they download from the Internet.

StencilA collection of masters.

ShapeAn object on a Visio drawing page. Often you create shapes by dragging a master from a stencil to the drawing page; however, you can also create shapes in other ways. A shape can be very simple: a line, a polygon, an image. A shape can also be a sophisticated object that changes appearance or behavior as data values change, as its position on the page changes, or as properties of another shape change—the possibilities are endless.



TemplateA Visio document that includes one or more drawing pages with preset dimensions and measurement units. A template may also include one or more stencils; it may include background pages and designs; its pages may contain shapes or text. A template may also include special software that only operates in that template.

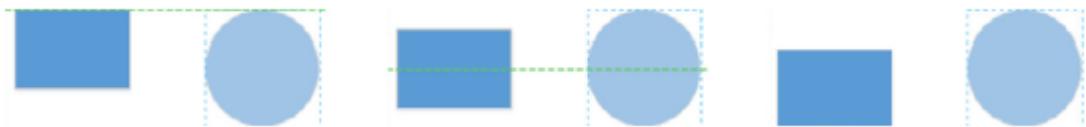
WorkspaceA collection of Visio windows and window settings. At minimum, the workspace consists of the drawing window and the zoom settings for the pages in the drawing; frequently, it also includes a Shapes window containing one or more stencils. The workspace can also include the Shape Data, Size & Position or Pan &Zoom windows. Unless you have changed the default action, Visio saves the on-screen workspace whenever you save the document. As a result, when you next open the same document, the same collection of windows is restored.

Activity: Creating Diagrams in Visio

Visio 2013 provides an enhanced *Dynamic Grid*. The purpose of the Dynamic Grid is to help you position a shape with greater accuracy as you drop it on the page or when you relocate it, thereby eliminating much of the need to drag and nudge the shape into alignment after you've placed it.

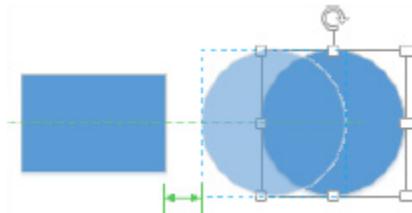
In this exercise, you will create a drawing from a stencil containing basic Visio shapes. In the process of doing so, you will use several types of Dynamic Grid feedback. You will also create several shapes using Visio's drawing tools.

1. Drag a Rectangle shape onto the drawing page and position it toward the upper-left corner of the page.
2. Drag a Circle shape onto the drawing page and position it to the right of the rectangle. Before you release the mouse button to drop the circle, move it up and down on the page. As you move the circle, a green, horizontal Dynamic Grid line

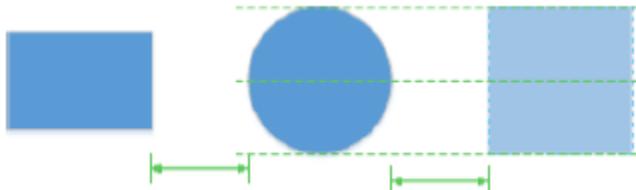


appears when the circle is in certain positions relative to the rectangle. From left to right in the following graphics, the Dynamic Grid line indicates when the circle is aligned with the top, center, and bottom of the existing rectangle.

3. Use the Dynamic Grid to align the circle with the middle of the rectangle and drop it so the space between the shapes is approximately 1 inch (2.5 cm).
4. Click on the circle and drag it closer to the rectangle. The Dynamic Grid centerline appears, and if you've located the circle at a certain distance from the rectangle, a second Dynamic Grid element appears. When the distance between the two shapes matches the default spacing interval for this page, a double-headed arrow appears.



5. Press Ctrl+Z to undo the shape movement and position the circle back where you originally dropped it.
6. Drag a Square shape onto the page and position it on the right side of the circle but don't release the mouse button yet.
7. Use the Dynamic Grid to align the square with the center of the circle and then move the square left and right until the green double-headed arrow appears. Notice that the double-headed arrow shown in the following graphic is longer than the double-headed arrow shown in the graphic after step 4 and that there are two of them, not one. In step 4, the double-headed arrow shows that the interval between your shapes matches the drawing's default spacing. In this example, the pair of double-headed arrows indicates that your new shape is the same distance from the circle that the circle is from the rectangle.

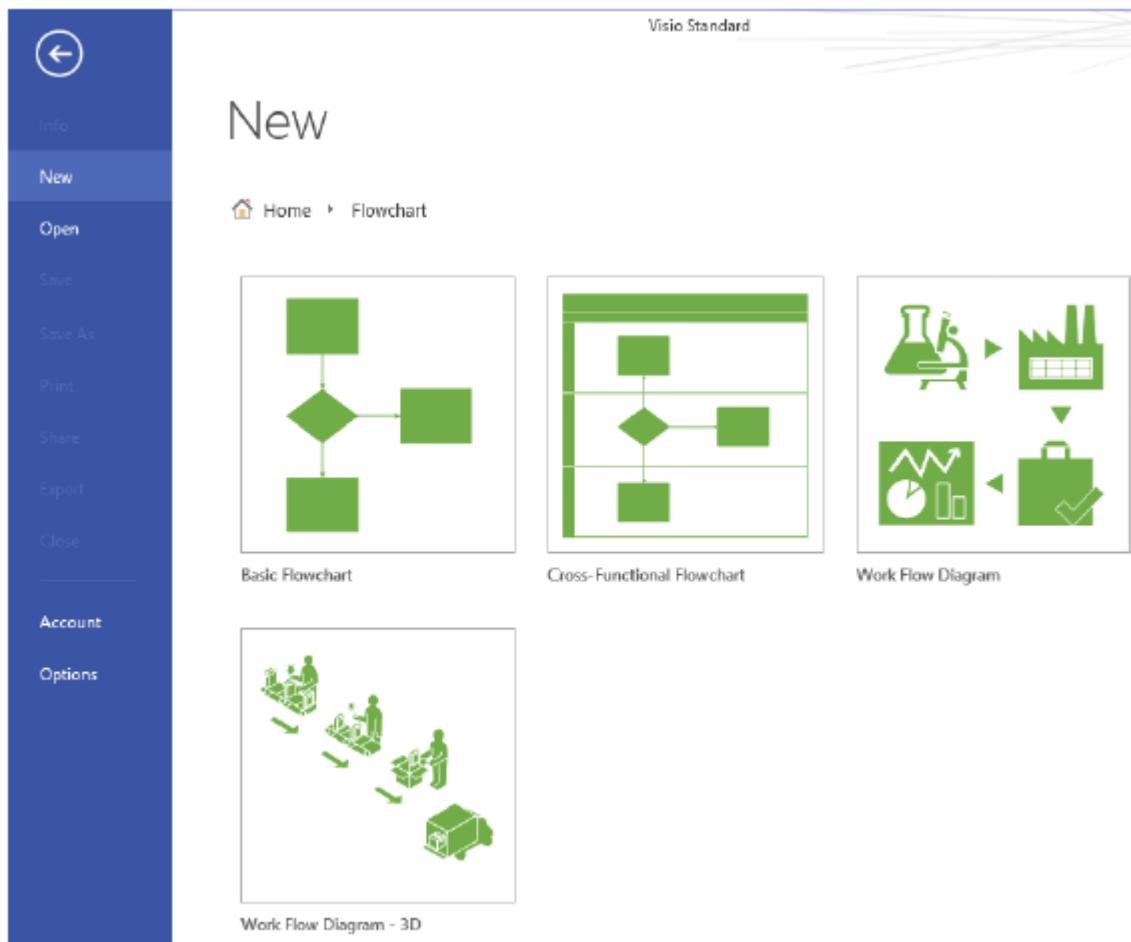


8. Release the left mouse button to drop the square.

Creating Flowcharts and Organization Charts in Visio

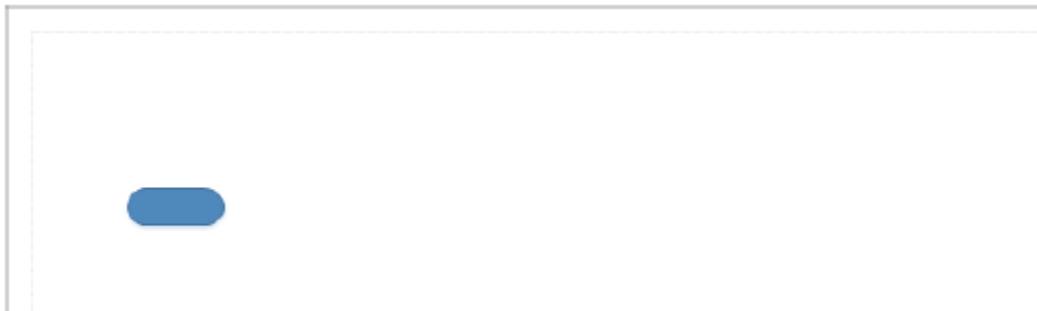
Microsoft Visio Standard 2013 includes four flowchart templates, as shown in the following graphic. You will work with the Basic Flowchart and Cross-Functional Flowchart templates in this chapter. The Work Flow Diagram template is a brand new, theme-capable template for

creating workflow diagrams. The corresponding template from previous versions of Visio was retained in Visio 2013 and is now called Work Flow Diagram - 3D.



In this exercise, you'll create a new flowchart for a simple human resources recruiting process. The flowchart will have seven process steps and one decision.

1. Drag a Start/End shape from the Basic Flowchart Shapes stencil onto the drawing page.

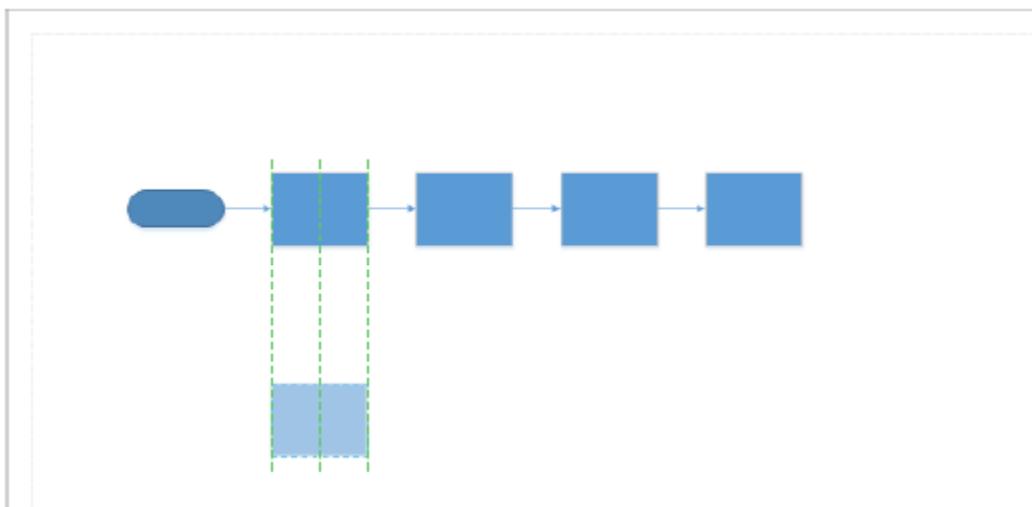


2. Point to the start shape you added to the drawing page, click the right-facing blue triangle that appears, and then click the Process shape from the Quick Shapes menu.

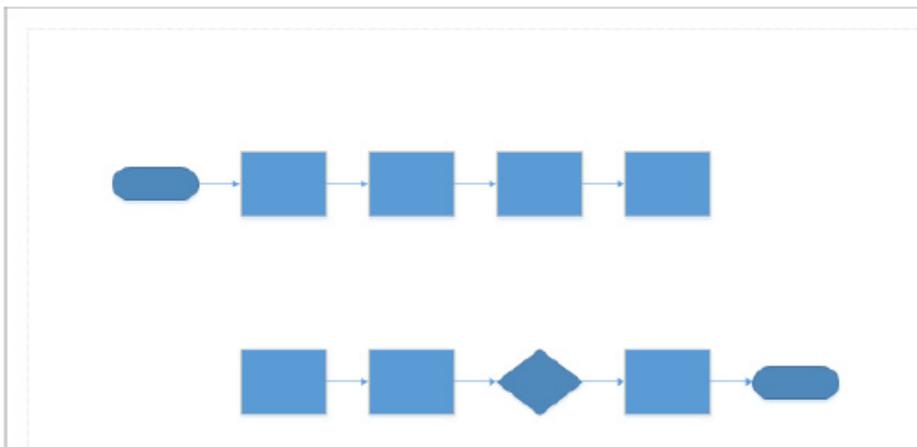
3. Use the same technique to add three more Process shapes to the page.



4. Drag a Process shape onto the drawing page. Then use the Dynamic Grid to position the new process shape below the leftmost process shape.



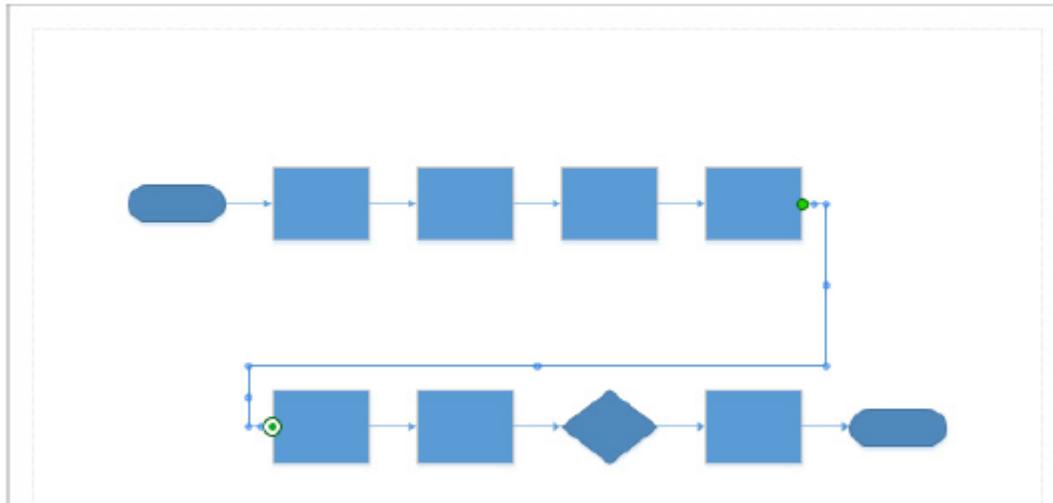
5. From the Quick Shapes menu, add the following four shapes:
Another Process shape to the right of the one from step 4
A Decision diamond to the right of the previous process shape
Another Process shape to the right of the decision diamond
A Start/End shape to the right of the final process shape



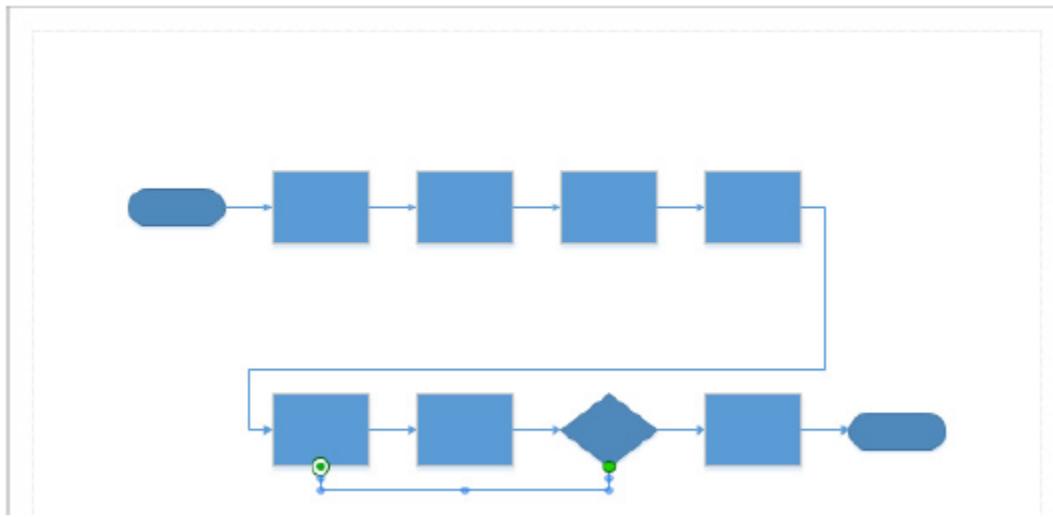
At this point, the flowchart is nearly complete with the exception of two connectors:

One that links the end of the first row to the beginning of the second row, and one that links the decision diamond back to a previous step in the flowchart.

6. Right-click anywhere on the drawing page, select the Connector Tool from the Mini Toolbar, and then move the cursor near the last shape in the first row.
7. Click the connection point on the right of the top-right process shape, and then drag to the leftmost connection point on the first process shape in the second row.

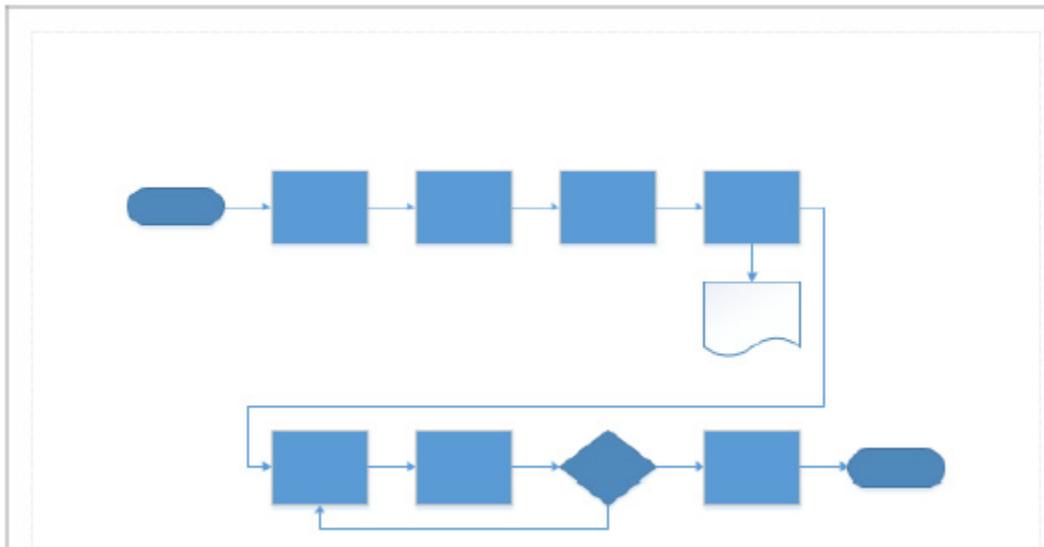


8. Click the blue Auto Connect arrow under the decision shape and drag it to the connection point on the bottom of the leftmost process shape in the same row.

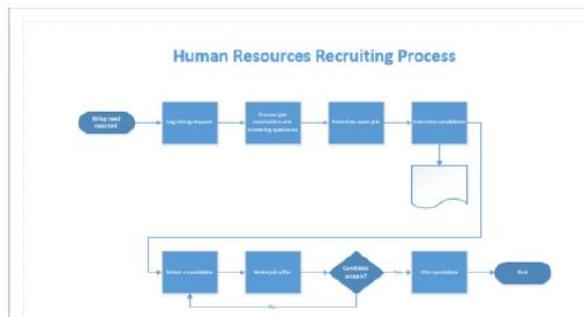


9. Drag a Document shape to just below the last process shape in the top row.

10. Drag a bounding box around all of the shapes in the bottom row. Then hold down the Shift key while you drag that row down to make more room. Once again, Visio will reposition the connector line to accommodate the new location of the bottom row.
11. Click the blue AutoConnect arrow under the upper-right process shape to connect it to the document shape. The layout of your flowchart is now complete.



12. Add the following labels to the flowchart:
 1. Start – Hiring need reported
 2. First 4 processes would be as follows:
 1. Prepare job description and screening questions
 2. Advertise open job
 3. Interview candidates
 3. The bottom row needs to have the following:
 1. Select a candidate
 2. Make job offer
 3. Candidate accepts?
 4. Hire candidate
 5. End
 4. Click the connector between the ‘Candidate accepts?’ shape and the Hire candidate shape and type Yes. Click the connector between the ‘Candidate accepts?’ shape and the Select a candidate shape and type ‘No’. Add a text box to the top of the page, type Human Resources Recruiting Process as a title for the flowchart, and then set the font to 24 pt. and bold. Your finished flowchart should look something like the following graphic.



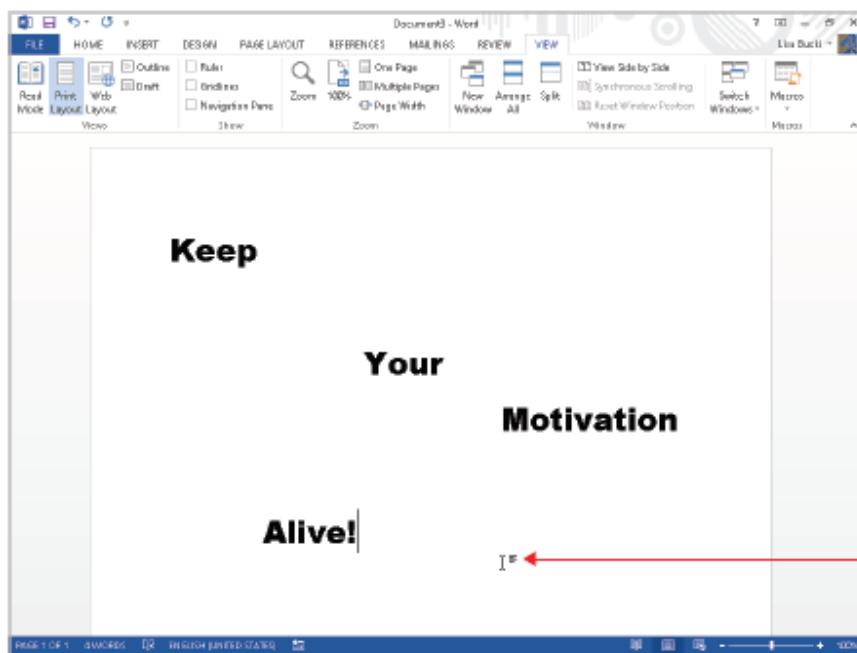
Microsoft Word 2013

Microsoft word is the most widely used word processor which is used to create formal as well as informal documents. Its advantages lie in the fact that it is very easy to use and at the same time offers a wide range of features

Basics of MS Word 2013

1. Click and Type:

Double-click and type anywhere on the page.



You can double click anywhere on the page to start writing. This is a very big advantage for MS Word.

2. Saving the Document

Saving initially: Before you begin to type, you should save your document. To do this, go to File Tab (Figure 2) > Save As. Microsoft Word will ask you to choose a location and then browse to a folder to save it in. After selecting these dialogue box (figure 3) will open and you can name the file. Once you have specified a name and a place for your new file, press the Save button.

Note: If you want to save your document on a Mac and then open it on a PC you must specify a file extension (i.e. .doc). Usually your computer will do this for you, but if it does not you must do this process while in Save As. Once you have titled your document, you can give it a file extension by clicking in the Format box. Click Microsoft Word Document for the correct file extension and make sure Append File Extension is checked.

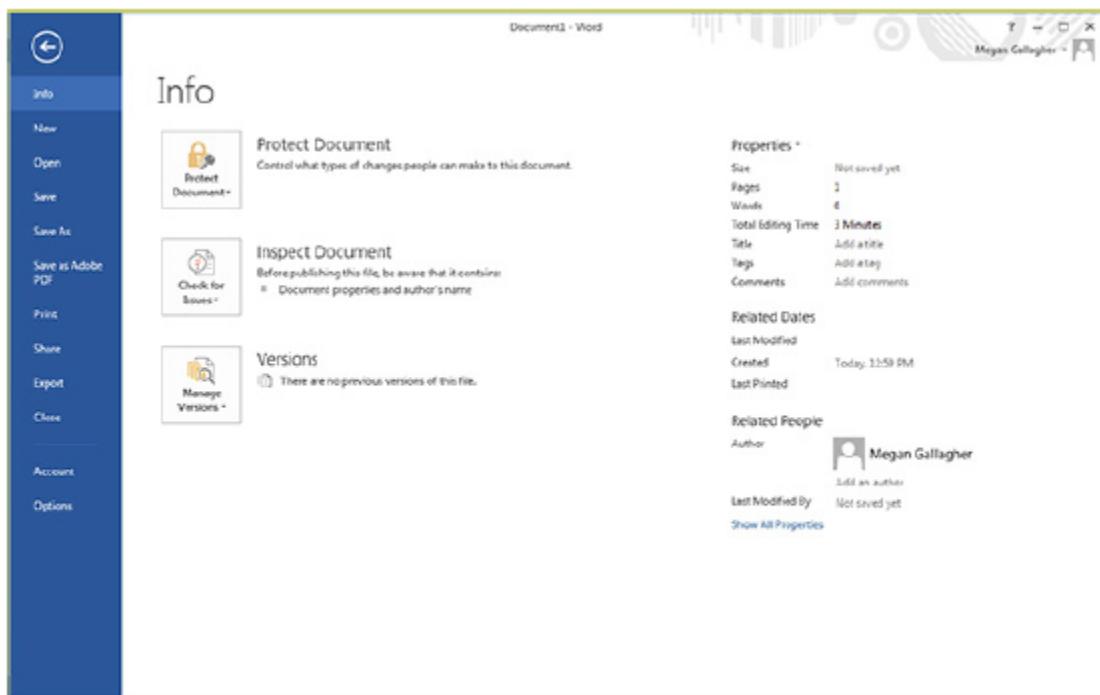


Figure 2. File Tab

Saving Later: After you have initially saved your blank document under a new name, you can begin writing on your paper. However, you will still want to periodically save your work as insurance against a computer freeze or a power outage. To save, click File tab > Save.

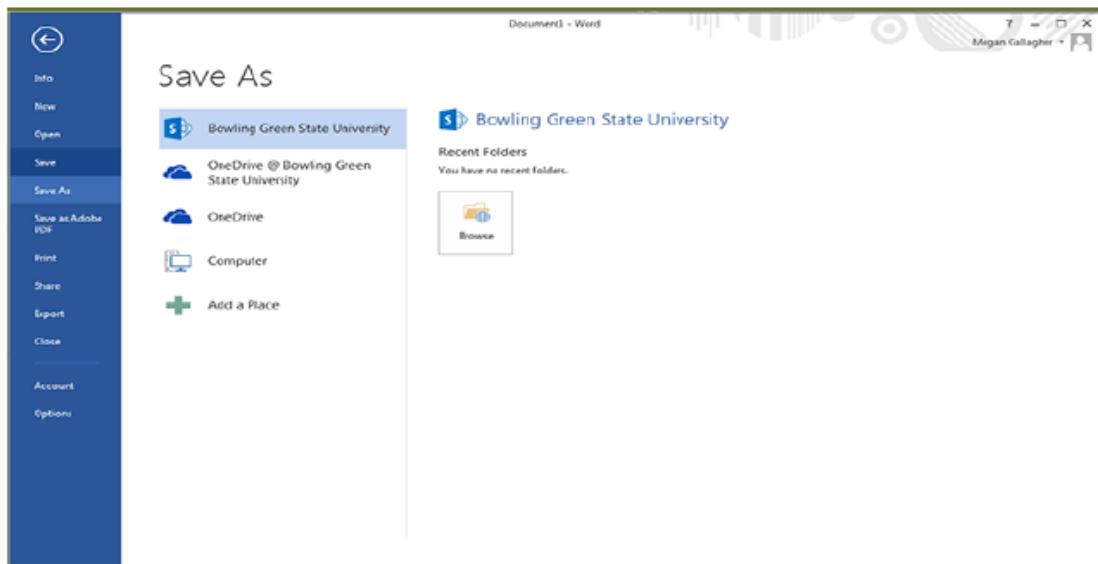


Figure 3. Saving dialog box.

3. Toolbars and Tabs

The new Microsoft Word uses one main toolbar to allow you to modify your document. Within this toolbar, you can switch between tabs to determine what you would like to do.

The Main Toolbar contains all the options available to you in Microsoft Word. The file tab (see figure 2) allows you to Save, Save As, Save as Adobe PDF, Open, Close, Print, Export, and Share.

The **Home Tab** (Figure 4) is Microsoft Words standard view. This is the view most widely used and allows you to format text by Font Style, Font Size, Bold, Italic, Underline, Alignment, Numbered List, Bulleted List, Indentation, Spacing, and Font Color.



Figure 4. Home Tab

The **Insert Tab** (Figure 5) contains any additives you want to place in your document, including but not limited to: Tables, Online Picture/Clip Art searches, Headers, and Footers. These icons are convenient and will bring up a dialogue box to give you further options when clicked.

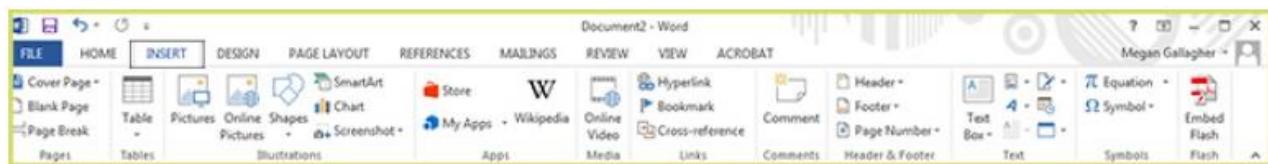


Figure 5. Insert Tab

The **Design** tab (Figure 6) contains different styles of page formatting. When you type, your layout will automatically match the format selected. It also allows you to change the color scheme, watermark, and paragraph spacing of the document.



Figure 6. Design Tab

The **Page Layout** Tab (Figure 7) contains icons for page setup and paragraph actions, such as Margin, Orientation, Size and Columns.

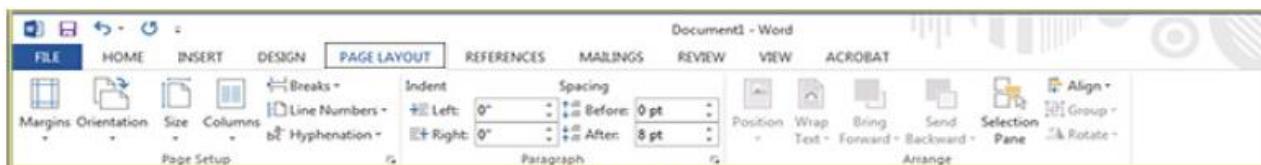


Figure 7. Page Layout Tab.

The **References** Tab (Figure 8) makes it especially simple to add Table of Contents, Footnotes, Bibliographic Information, Indexes and Citations.

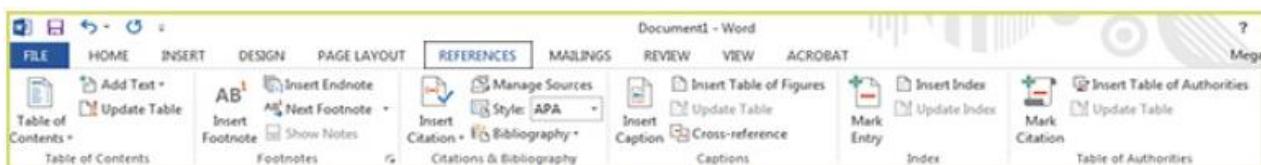


Figure 8. References Tab.

The **Mailings** Tab (Figure 9) is for post-office related uses. If you wanted to create custom Envelopes or Labels, this is where you would find such actions.



Figure 9. Mailings Tab.

The **Review** Tab (Figure 10) is where one can find Spelling & Grammar, the built in Thesaurus and Dictionary, you can Track Changes, Check Word Count, and Show/Add Comments.

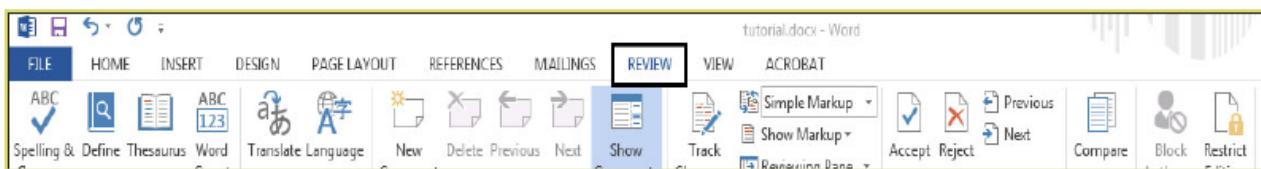


Figure 10. Review Tab.

The **View** tab (Figure 11) allows you to change the views of your document.

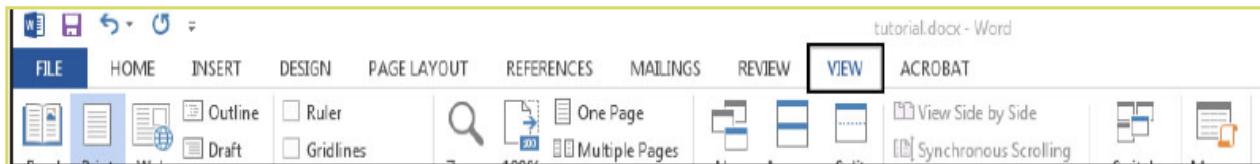
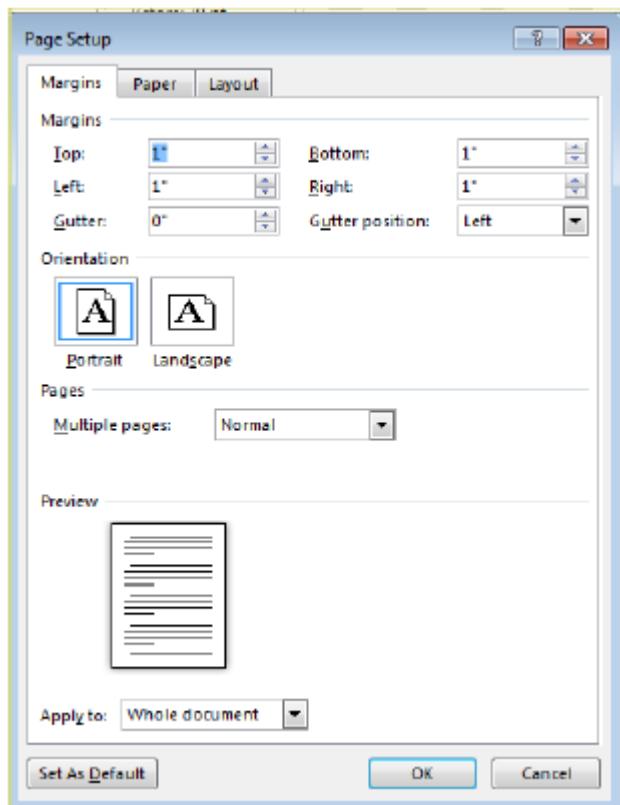


Figure 11. View Tab.

4. Formatting the Document

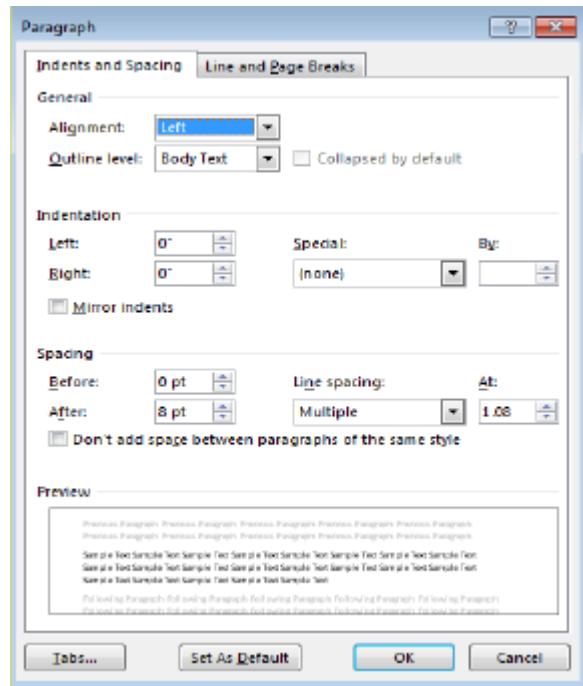
The default page margins for Microsoft Word documents are 1 inch, but you may want to change them for a project. To change the page margins on a PC, go to Page Layout Tab > Page Setup box > Margins button. On a PC, a dropdown will appear to give a set of standard options, but by clicking “Custom Margins,” a menu will appear where you can type irregular margins. From the same menu click Portrait if you want your document to be 8.5 x 11 inches (most common). Click Landscape if you want your document to be 11 x 8.5 inches. Landscape simply flips the page 90 degrees.



Formatting Paragraphs

To format your paragraph, first highlight the paragraph you wish to format. To highlight more than one paragraph, click at the beginning of the paragraph and drag the mouse over the text. To apply changes to the entire document, select all by hitting Ctrl + A. To specify Alignment, Line Spacing, Indentation, and Page Break expand the Paragraph section of the Home Tab. This will open up the Paragraph menu.

The Alignment option allows you to choose how you want your paragraph to look (i.e. justified, right, center, or left). The Line Spacing option allows you to set the desired spacing, such as single or double. The Indentation option allows you to tab/push the line(s) in your paragraph either left or right. The Page Break option is found in Paragraph menu, but you must first select the Line and Page Breaks tab. Page Break allows you to split a paragraph or a page up into sections. You can also bring up this menu by right clicking (or by hitting Ctrl + Click on a one button mouse) within the document and selecting Paragraph.



Cut, Copy, And Paste

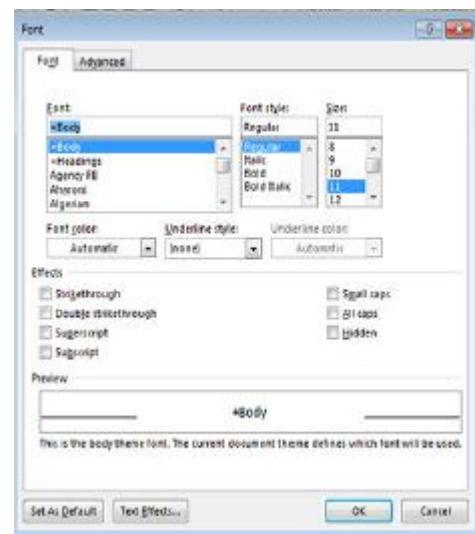
You can use the Cut, Copy and Paste features of Word to change the order of sections within your document, to move sections from other documents into new documents, and to save yourself the time of retyping repetitive sections in a document. Cut will actually remove the selection from the original location and allow it to be placed somewhere else. Copy allows you to leave the original selection where it is and insert a copy elsewhere. Paste is used to insert whatever has been cut or copied.

To Cut or Copy:

Highlight the text by clicking and dragging over the text to be cut or copied. Go to Home Tab > Clipboard box > Copy or Home Tab > Clipboard box > Cut. Click the location where the information should be placed. Go to Home Tab > Clipboard box > Paste.

Formatting Text

Before you type, you should select your font style, size, color and attributes (such as bold, italic and underline) in the Home Tab. You can expand the Font Menu box to get more options by clicking the down-arrow. However, if you wish to change text that has already been typed, click and drag over the text to be changed to highlight it (or go to Edit > Select All to select the entire document) and change it as before.



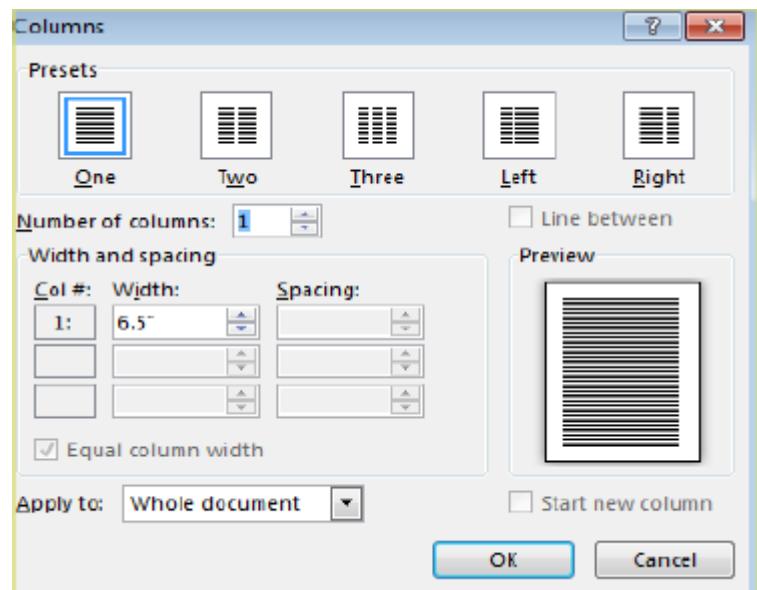
Numbered And Bulleted Lists

To create a simple numbered or bulleted list, click on the Numbering or Bullet button on the Paragraph toolbar in the Home Tab. To have more control over the format of your list, click the down arrows beside each style of list. Type the first item in the list and press Return to move to the next number or bullet. Press Return twice to exit the list.

Adding Columns

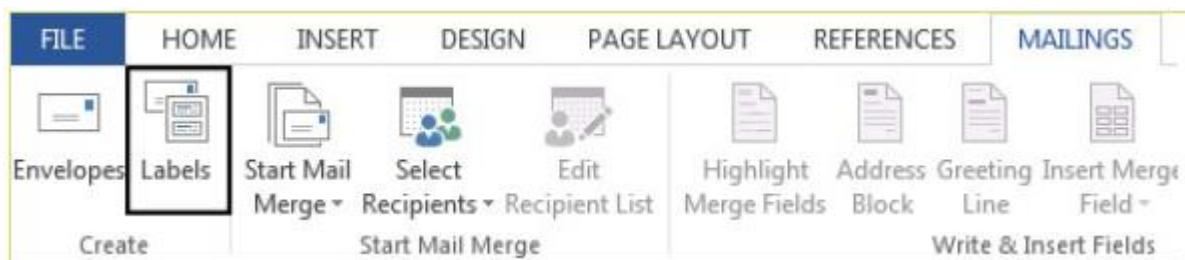
Columns can be used for a variety of document types, such as a tri-fold brochure. To do this, go to Page Layout Tab > Page Setup box > Columns.

From the Columns menu, you can choose the number of columns, or for more options, click More Columns where you can set column width and spacing. Once you select your preferred design, it will show up in the Preview box. This is a nice feature because it allows you to see what you are selecting before applying it to your word document. When you are happy with how your document looks, click Ok.



Headers And Footers

Headers and Footers can be used to give a uniform look to the pages of your document. To create one, go to Insert Tab > Header and Footer. Use this toolbar to insert and format words and objects in the header. When editing the header, a new Top View will appear that is specific to headers and footers



5. Inserting and adding objects

Clip Art

When trying to enhance your word document you may want to include Clip Art and/or Word Art. Microsoft Word comes with an Online Pictures button that contains a large variety of images including pictures, borders, and backgrounds. To find a desired image, you can either search through the Clip Art gallery or search the Internet with the Bing Image Search engine.

To insert Clip Art or Pictures: Go to Insert Tab > Illustrations box > Online Pictures. A dialogue box with the Clip Art gallery and the Bing Image Search engine will appear. If you wish to have a Clip Art photo, search the image's description in the Clip Art Gallery.

If you wish to have an image from the Internet, search the image's description in the Bing Image Search engine. The picture will be inserted at the location of your cursor within your document. If you need to modify your Clip Art, click on it once to select it, and small boxes will appear around the corners.

Once your Clip Art is selected, you can resize your picture by clicking and dragging on the boxes. Holding Shift while clicking and dragging will resize the Clip Art proportionately. You can use the Drawing toolbar to further modify your Clip Art.

To delete Clip Art, select it by clicking on it until the black boxes appear and then hit Delete.

WordArt

To insert WordArt, go to Insert Tab > Text box > WordArt button. Select the desired style and click OK. Type the desired text and click Ok. You can further modify your text by using the Drawing toolbar. To select your WordArt, click on it, and small boxes will appear in the corners. Moving the circle arrow anchor on the top of the WordArt allows the user to change the slant of the WordArt. WordArt can be resized and deleted similarly to ClipArt.

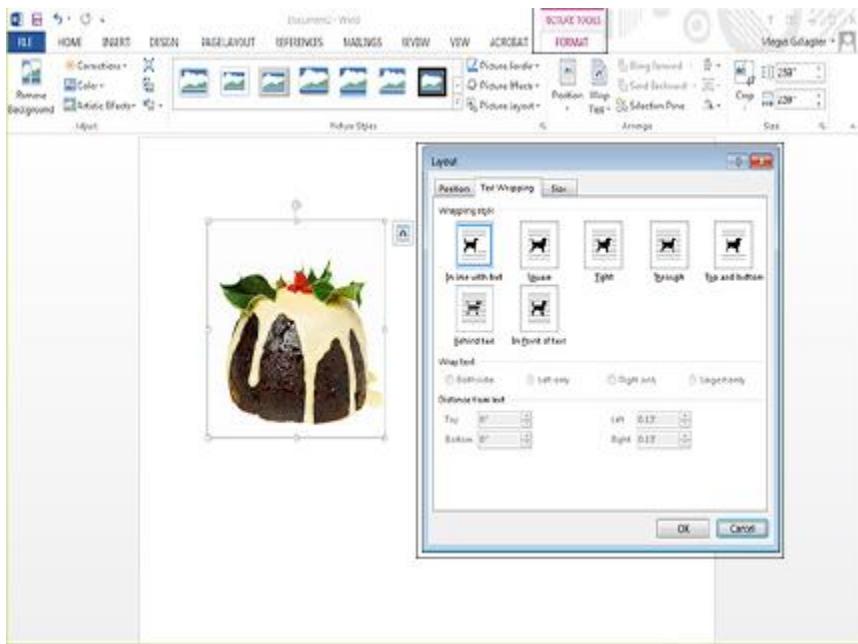
Word Wrap

Word Wrap is part of formatting pictures. To use Word Wrap, select your inserted image, and go to the Format Tab > Arrange box. Here users can choose the type of text wrapping style desired. For more options select More Layout Options

Insert Pictures

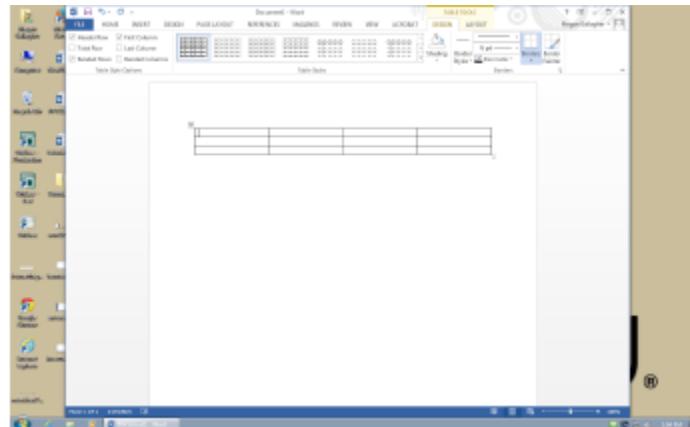


Sign in with your Microsoft account to insert photos and videos from Facebook, Flickr, and other sites.



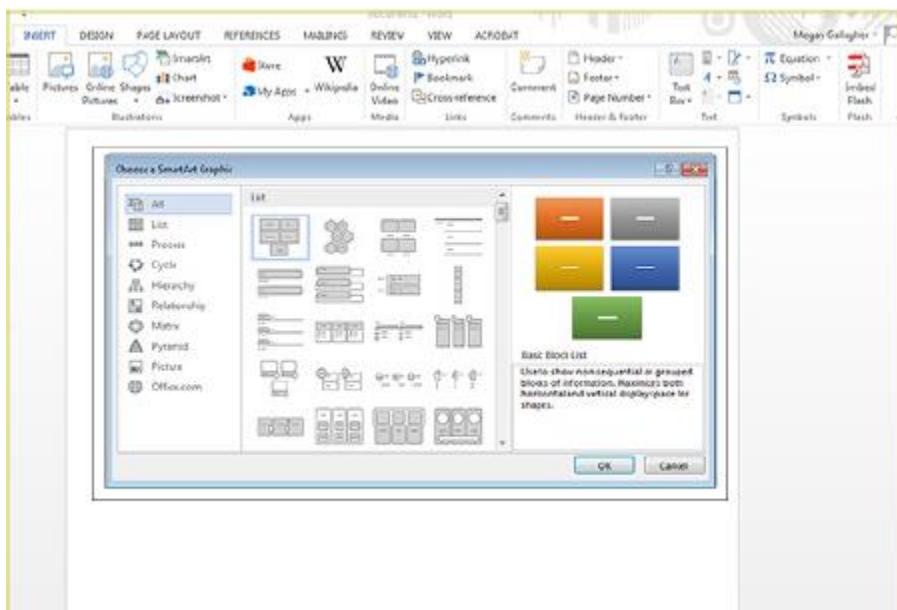
Creating a Table

To create a table within your document, go to Insert Tab > Tables box > Table button. Choose the desired table size and format by hovering over the boxes, and press Enter on the keyboard. The table will be inserted at the cursor's location within your document. To navigate within your table, use the arrow keys. To modify your table, when your table is selected, a Design Tab and Layout Tab will appear in the Toolbar. From here, you can add cells, columns or rows, merge or split cells, and further modify your table. To exit the table, click outside of it.



Smart Art

Smart Arts are used to create diagrams in Microsoft Word. If you want to create a custom flowchart, you can use the Drawing capabilities discussed earlier. To insert Smart Art, go to Insert Tab > Illustrations box > Smart Art. A dialogue box will open with basic choices.



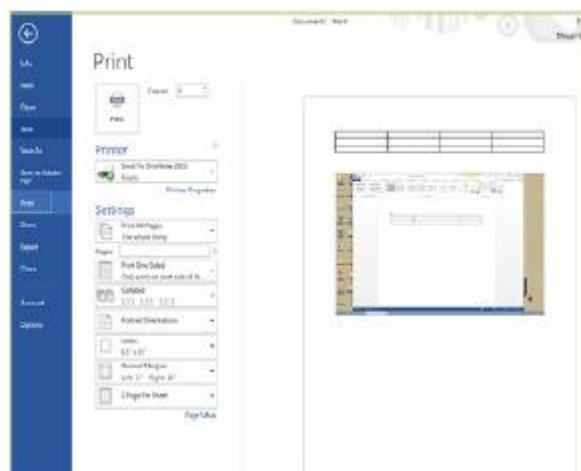
6. Printing

It is important to always save your document before you print!

Print Preview

Before you print your document, you may want to preview it to make sure you are happy with the page layout and appearance of your document. To do this, go to File Tab > Print.

This should open up a preview of your document. To zoom in on the page, find the scale in the bottom right corner of the screen. You can move it back and forth to adjust the magnification. If you are satisfied with the appearance of your document, you can click on the Print icon in the toolbar. If you need to make changes to the document or are not ready to print, select close on the toolbar.



Printing

To print your document, go to File Tab > Print > Print, select your desired settings, and then click Print again. It is also possible to print by clicking the drop down arrow in the top left corner. You can either quick print which will not preview or print preview and print, which will show your document. Below it is the option to print with the preview.

Saving As a PDF

Go to File Tab > Save As Adobe PDF. Or

Go to Mailings Tab > Acrobat box > Merge to Adobe PDF button.

7. Some other helpful functions

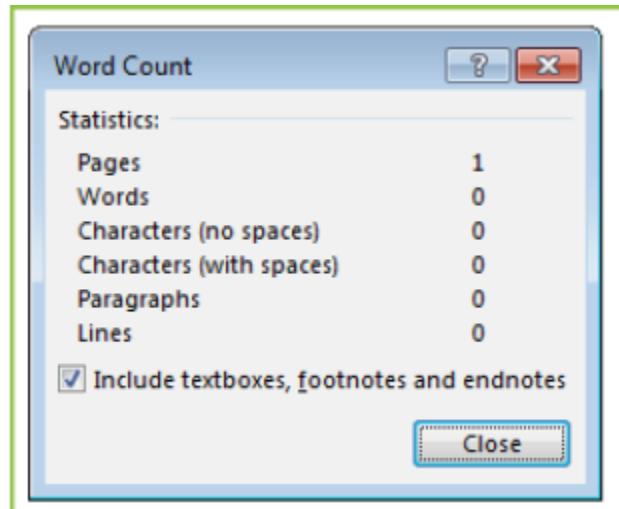
Undo and Redo

The easiest way to undo an action is with the key commands **Ctrl + Z** and to redo an action with **Ctrl + Y**. It is important to note that not all actions are undoable, thus it is important to save before you make any major changes in your document so you can revert back to your saved document. There are also two icons above the Main Toolbar near the Save Icon. The left icon is Undo and the right icon is Redo.



Word Count

To get an accurate word count of your document, go to Review Tab > Proofing Box > Word Count. This will give you the total number of words in your document. If you need to word-count a specific section, highlight that section first by clicking and dragging over it and then to go Review Tab > Proofing Box > Word Count as before.



Quitting

Before you quit, it's a good idea to save your document one final time. Go to File Tab > Exit Word. This is better than just closing the window, as it insures your document quits correctly.

8. Shortcut keys for Microsoft Office

The shortcut keys listed below can be a great help when using Microsoft Office products. Not only are they quick and easy, they are also amazing time savers.

CTRL+K create a hyperlink

CTRL+/ Display HTML tags

CTRL+T Create an Auto Thumbnail of the selected picture

CTRL+SHIFT+B Preview a page in a Web browser

SHIFT+ALT+F11 Display the Microsoft Script Editor

CTRL+N create a new page

CTRL+B Bold

CTRL+I Italic

CTRL+U Underline

CTRL+C Copy

CTRL+V Paste

CTRL+Z Undo

CTRL+S Save

CTRL+P Print

CTRL+O Open

Microsoft Excel 2013

Microsoft Excel is one of the most popular spreadsheet applications that helps you manage data, create visually persuasive charts, and thought-provoking graphs. Excel is supported by both Mac and PC platforms. Microsoft Excel can also be used to balance a checkbook, create an expense report, build formulas, and edit them.

Creating a new document

Begin by opening Microsoft Excel.

On a PC, click **Start >All Programs >Microsoft Office >Microsoft Excel 2013**.

When opened, a new spreadsheet will pop up on the screen. If this does not happen, click on the **File tab >New**. From here a dialog box with various different templates will appear on the screen that you can choose from. Once a template is chosen, click **Create**.

Three commonly used toolbars

The **Home Tab**: This is one of the most common tabs used in Excel. You are able to format the text in your document, cut, copy, and paste information. Change the alignment of your data, insert, delete, and format cells. The **Home Tab** also allows you to change the number of your data (i.e. currency, time, date).



The **Insert Tab**: This tab is mainly used for inserting visuals and graphics into your document. There are various different things that can be inserted from this tab such as pictures, clip art, charts, links, headers and footers, and word art.



The **Page Layout Tab**: Here you are able to add margins, themes to your document, change the orientation, page breaks, and titles. The scale fit of your document is also included as a feature within this tab, if needed.



Working with Cells

Cells are an important part of any project being used in Microsoft Excel. Cells hold all of the data that is being used to create the spreadsheet or workbook. To enter data into a cell you simply click once inside of the desired cell, a green border will appear around the cell. This border indicates that it is a selected cell. You may then begin typing in the data for that cell.

A	B	C	D
1			
2			
3			
4			
5			

Changing an Entry within a Cell

You may change an entry within a cell two different ways:

1. Click the cell one time and begin typing. The new information will replace any information that was previously entered.
2. Double click the cell and a cursor will appear inside. This allows you to edit certain pieces of information within the cells instead of replacing all of the data.

Cut, Copy, and Paste

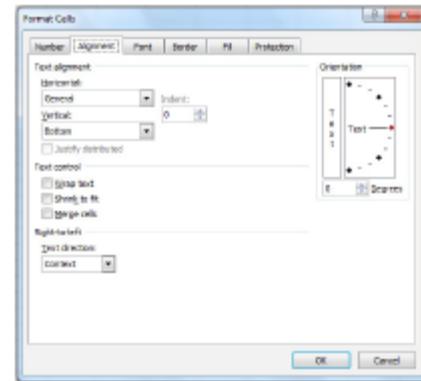
You can use the Cut, Copy and Paste features of Excel to change the data within your spreadsheet, to move data from other spreadsheets into new spreadsheets, and to save yourself the time of re-entering information in a spreadsheet. Cut will actually remove the selection from the original location and allow it to be placed somewhere else. Copy allows you to leave the original selection where it is and insert a copy elsewhere. Paste is used to insert data that has been cut or copied.

1. Highlight the data or text by selecting the cells that they are held within.
2. Go to the Home Tab >Copy (CTRL + C) or Home Tab >Cut (CTRL + X).
3. Click the location where the information should be placed.
4. Go to Home Tab >Paste (CTRL + V) to be able to paste your information.

Formatting Cells

There are various options that can be changed to format the spreadsheets cells. When changing the format within cells you must select the cells that you wish to format.

1. Drag and select the cells you wish to change.
2. Click Home Tab >Format >Format Cells. A box will appear on the screen with six different tab options.



The basic options in the format dialog box are as follows:

Number: Allows you to change the measurement in which your data is used. (If your data is concerned with money the number that you would use is currency)

Alignment: This allows you to change the horizontal and vertical alignment of your text within each cell. You can also change the orientation of the text within the cells and the control of the text within the cells as well.

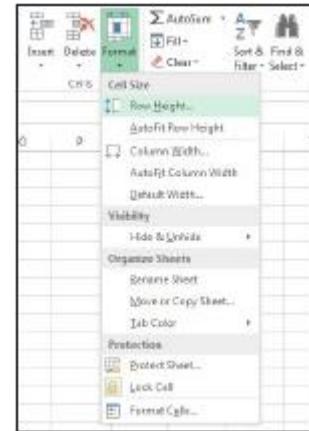
Font: Gives the option to change the size, style, color, and effects.

Border: Gives the option to change the design of the border around or through the cells.

Formatting Rows and Columns

When formatting rows and columns you can change the height, choose for your information to autofit to the cells, hide information within a row or column, un-hide the information. To format a row or column, proceed with the following steps:

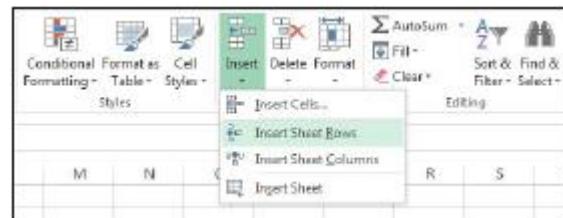
1. Select the cells which will be altered.
2. Go to **Home Tab >Row Height (or Column Height).**
3. Choose which height you are going to use.



Adding Rows and Columns

Rows are cells that run horizontally across the document. You can insert an extra row of cells like this:

1. Drag select along the row of cells where you want your new row to appear.
2. Click **Home Tab >Insert >Insert Sheet Rows.** The row will automatically be placed on the spreadsheet and any data that was selected in the original row will be moved down below the new row.



Columns are cells that run vertically down the document. You can insert an extra column of cells like this:

1. Drag select along the column of cells where you want your new column to appear.
2. Go to **Home Tab >Insert >Insert Sheet Column.** The column will automatically be placed on the spreadsheet and any data to the right of the new column will be moved more to the right.

Working With Charts

Charts are an important part to being able to create a visual for spreadsheet data.



1. In order to create a chart within Excel the data that is going to be used for it needs to be entered already into the spreadsheet document. Once the data is entered, the cells that are going to be used for the chart need to be highlighted so that the software knows what to include. Next, click on the **Insert Tab** that is located at the top of the screen.
2. You may choose the chart that is desired by clicking the category of the chart you will use. Once the category is chosen the charts will appear as small graphics within a drop down menu.

To choose a particular chart just click on its icon and it will be placed within the spreadsheet you are working on.

3. To move the chart to a page of its own, select the border of the chart and **Right Click**. This will bring up a drop down menu, navigate to the option that says **Move Chart**. This will bring up a dialog box that says **Chart Location**. From here you will need to select the circle next to **As a New Sheet** and name the sheet that will hold your chart. The chart will pop up larger in a separate sheet but in the same workbook as your entered data.

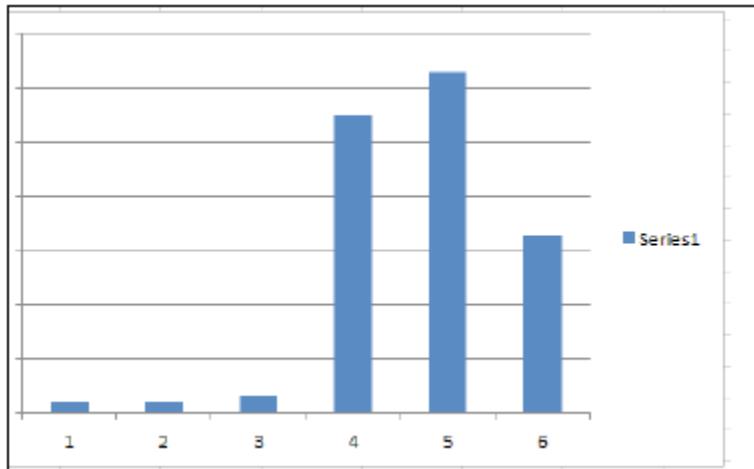


Chart Design

There are features that you can change to make your chart more appealing. To be able to make these changes you will need to have the chart selected or be viewing the chart page that is within your workbook. Once you have done that the **Design Tab** will appear highlighted with various different options to format your graphic.



Chart Options

Titles: Within the new chart Design tab, click the **Add Chart Element** icon. Here, you will see the option to title the chart as well as various components of the chart.

Change Chart Type: You can change your chart easily by selecting this icon and navigating to a more desirable chart. This feature is very convenient for someone who chose the wrong chart and doesn't wish to reselect all their data and go through the process a second time.

Format Chart Area: This allows for changes to be made to the chart's border, style, fill, shadows, and more. To get this option you will need to right click on the chart's border and navigate to the **Format Chart Area** option. Once this is clicked a dialog box will appear.

Creating Functions in Excel

Excel is a very powerful data analysis tool and has bunch of built in functions for statistical, financial or any other form of data analysis. In our course, we are concerned with excel only to the extent of using it as a documentation tool, and hence it is not in the scope to delve into the details of functions creation. We will nonetheless explore some fundamentals.



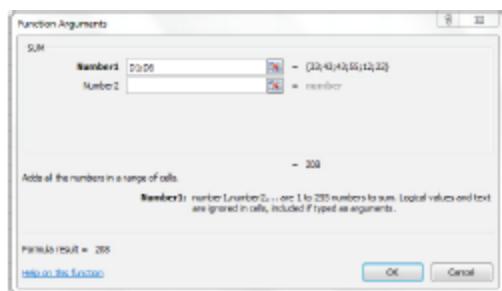
Creating Functions

When creating a function in Excel you must first have the data that you wish to perform the function with selected.

1. Select the cell that you wish for the calculation to be entered in (i.e.: if I want to know the sum of B1:B5 I will highlight cell B6 for my sum to be entered into).
2. Once you have done this you will need to select the **Formulas Tab** located at the top of the screen.
3. A list of Most Recently Used, Financial, Logical, Text, Date and Time, Math and Trig formulas will appear. To choose one of the formulas click the icon that holds the formula you are looking for.
4. Once you have clicked your formula this will display a dialog box on your screen.
5. To accept that calculation you can press **OK** and the result will show up in the selected cell.

A	B	C
	33	
	24	
	55	
	15	
	20	
	45	
	=SUM(B1:B6)	

Did you know? Many of the functions in Microsoft Office applications are common and hence other functions like printing, saving, opening, etc. will have the same flow as Microsoft Word.



Microsoft PowerPoint 2013

Microsoft PowerPoint is one of the most popular presentation programs supported by both Mac and PC platforms. Microsoft PowerPoint can be used to create interactive presentations for classroom, business, or personal use.

Most of the features of opening and saving would be the same as MS word and excel (which we've already covered). The additional features and learning areas will be covered now

The Design Ribbon toolbar:

The Design Ribbon toolbar contains several categories for formatting the design and elements of your presentation. These include: Home, Insert, Design, Transitions, Animations, Slide Show, Review and View. These icons are convenient but will not bring up dialogue boxes that allow you to change the settings of these actions. You should use whichever method works the best for you.



- **The Home bar:** This has features that allow you to edit Slides, Fonts, Paragraph, Drawing and Editing.
- **The Insert bar:** This has features that allow you to add Tables, Images, Illustrations, Links, WordArt, and Media Clips.
- **The Design bar:** This allows you to edit how your presentation will look. It includes features such as Page Setup, Themes, and Background Styles.
- **The Transitions bar:** This allows you to edit Slide Transitions, Effects, and Transition Timing.
- **The Animations bar:** This allows you to add custom animations to your presentation. You can select from Preview, Animations, Advance Animation, and Animation Timing.
- **The Slide Show bar:** This has features that allow you to select how your presentation will be displayed. From here, you can Start Slide Show, Set Up Slide Show, and Adjust Monitor Settings.
- **The Review bar:** This allows you to Proofread, Translate Languages, Comment Slides, and Compare your presentation.
- **The View bar:** This has features that allow you to set the View of your Presentation, Create Master Views, Show/Hide Features, Zoom, Switch from Color to Grayscale, Adjust Windows, and Add Macros.

Formatting your presentation

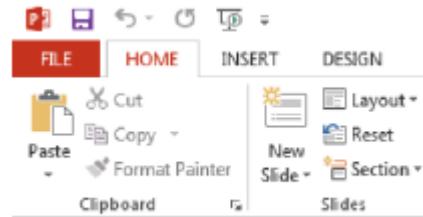
The default slide design for Microsoft PowerPoint documents is a blank slide.

Therefore, if you want your slides to have a specific design, you must add one. To do this on a PC, go to the Design Ribbon >Design. To change the theme of your presentation, click on the theme that you like and it will automatically apply to your slide.



To apply a Slide Layout, select the Design Ribbon > Home > Layout

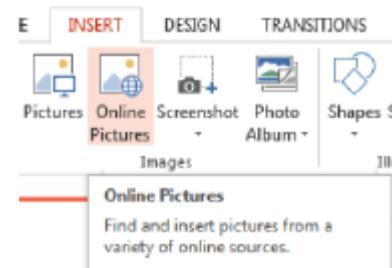
This helps to customize the layout of your document. From here, you can specify how the content on your slide is displayed. To apply a layout to your slide click the layout button, and choose your preferred style.



Inserting and adding objects

Online Pictures

When trying to enhance your word document you may want to include Online Pictures and/or Word Art. Microsoft PowerPoint comes with a Clip Gallery that contains a large variety of images including pictures, borders, and backgrounds. To find a desired image, you can either click on topics or type in the search box to find exactly what you are looking for.



Go to Insert > Online Pictures and then select the desired picture in the Picture Gallery.

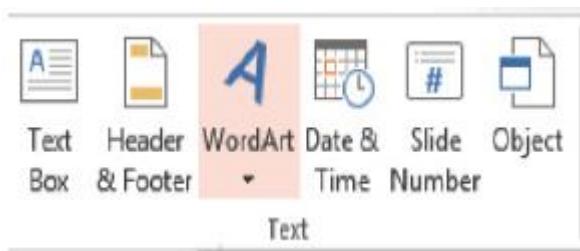
The picture will be inserted at the location of your cursor within your document. If you need to modify your Online Pictures click on it once to select it, and small boxes will appear around the corners. Once your Online Pictures is selected, you can resize your picture by clicking and dragging on the boxes. Holding shift while clicking and dragging will resize the Online Pictures proportionately.

Word Art

WordArt is inserted similarly to Clip Art.

To insert WordArt, go to Insert > WordArt.

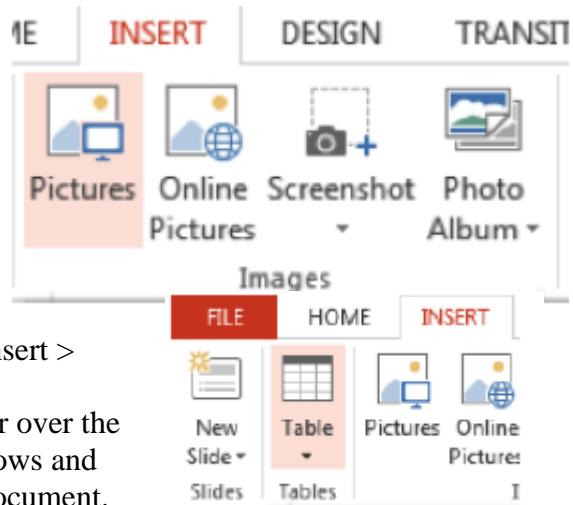
Afterwards, you will be prompted to insert your text. To select your WordArt, click on it, and small boxes will appear in the corners. Moving the small boxes will allow you to resize the WordArt. You can also change the effects of the WordArt by clicking the options in the design ribbon toolbar.



Pictures

To insert a picture into your presentation select Insert > Picture

You can choose either to get the photo from a browser or from a file or locate the image that you want to put into your document and click Insert.



Creating a table

To create a table within your document, go to the Insert > Table

Choose the desired table size by moving your cursor over the grid and until you have the appropriate amount of rows and columns and then click to insert the table into the document.

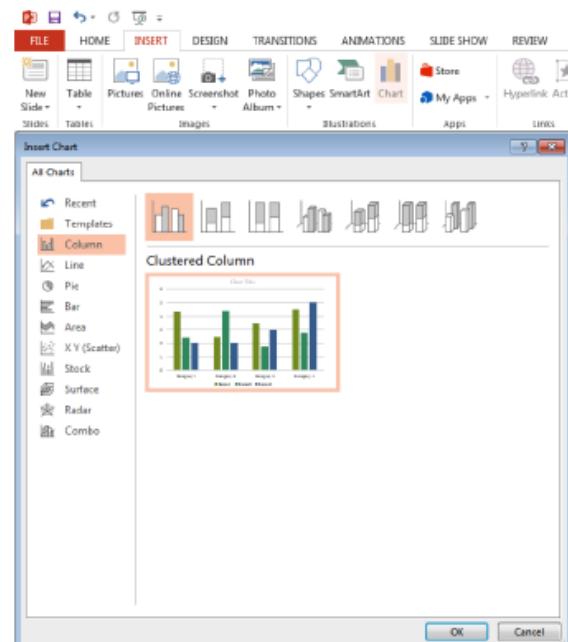
To navigate within your table, use the arrow keys.

Charts

Charts are used to create diagrams in Microsoft PowerPoint.

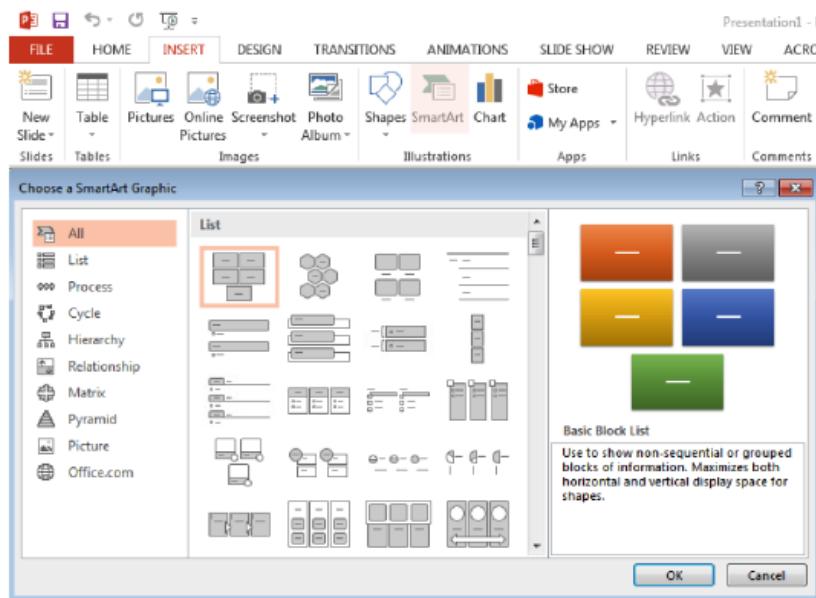
To insert a Chart in PowerPoint, click the Design Ribbon > Insert > Chart. Once the Chart button is expanded, you will have several options to choose from, such as Column, Line, Pie, etc.

To insert SmartArt, click the Design Ribbon > Insert > SmartArt. When the SmartArt button is expanded, you will have several options to choose from, such as List, Process, Cycle, Hierarchy, etc. To insert Shapes, click the Design Ribbon > Insert > Shapes. To resize the Shapes, simply click and drag any of the four corners on the object.



Smart Art

Steps to insert SmartArt are the same as the steps required to insert SmartArt in MS Word, which we have already gone through earlier.

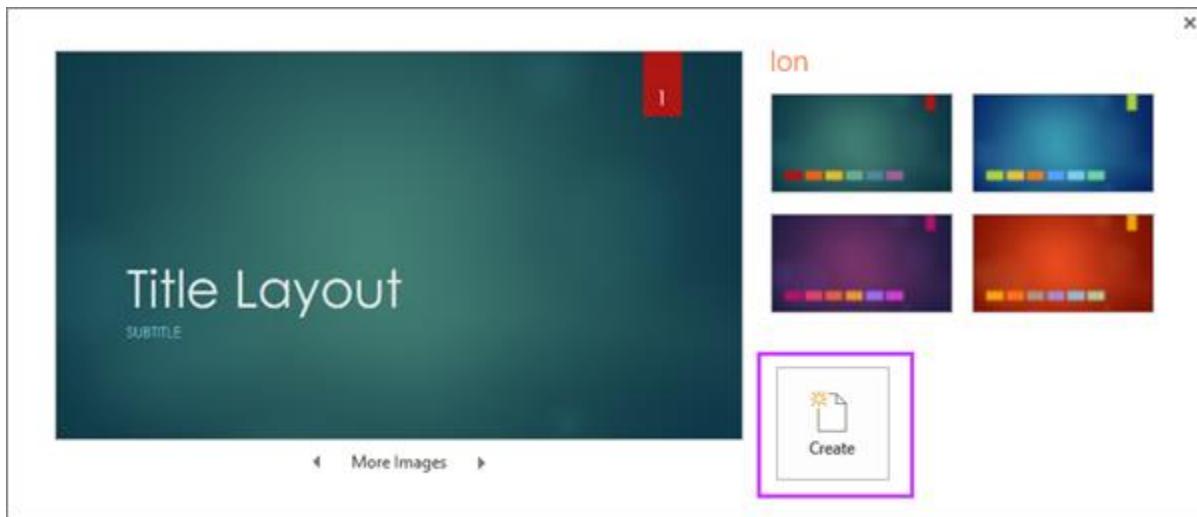


Giving your presentation (Steps)

Choose a theme

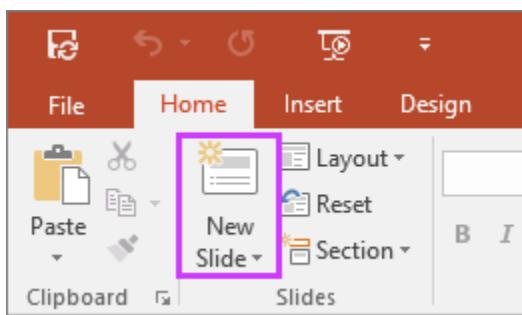
When you open PowerPoint, you'll see some built-in themes and templates. A theme is a slide design that contains matching colors, fonts, and special effects like shadows, reflections, and more.

1. Choose a theme.
2. Click **Create**, or pick a color variation and then click **Create**.



Insert a new slide

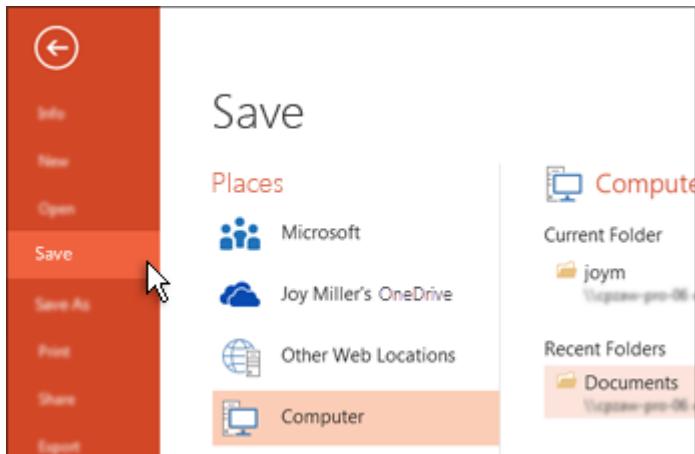
On the **Home** tab, click **New Slide**, and pick a slide layout.



Save your presentation

1. On the **File** tab, choose **Save**.

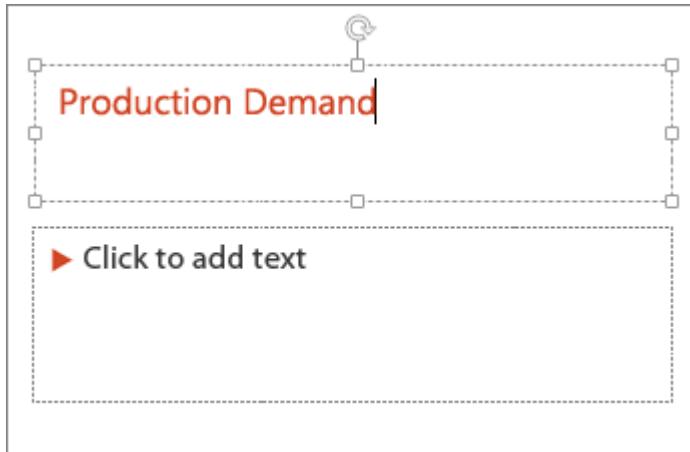
2. Pick or browse to a folder.
3. In the **File name** box, type a name for your presentation, and then choose **Save**.



Save your work as you go. Hit **Ctrl+S** often.

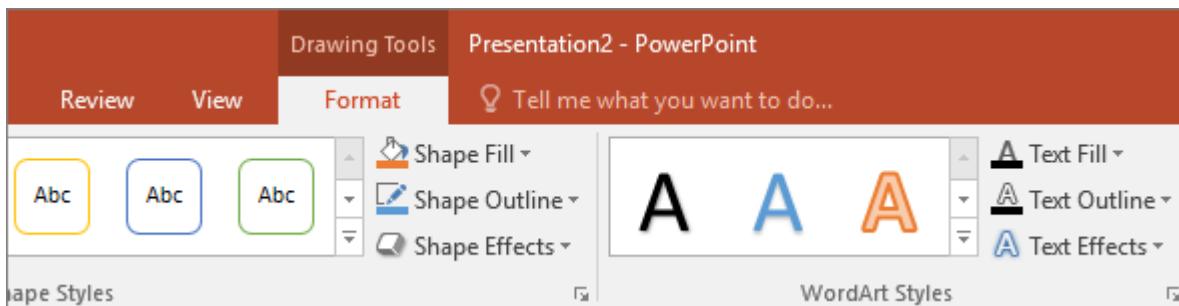
Add text

Select a text placeholder, and begin typing.



Format your text

1. Select the text.
2. Under **Drawing Tools**, choose **Format**.



3. Do one of the following:
4. To change the color of your text, choose **Text Fill**, and then choose a color.
5. To change the outline color of your text, choose **Text Outline**, and then choose a color.
6. To apply a shadow, reflection, glow, bevel, 3-D rotation, a transform, choose **Text Effects**, and then choose the effect you want.

Add speaker notes

Slides are best when you don't cram in too much information. You can put helpful facts and notes in the speaker notes, and refer to them as you present.

1. To open the notes pane, at the bottom of the window, click **Notes**
2. Click inside the **Notes** pane below the slide, and begin typing your notes.

WingTip Toys

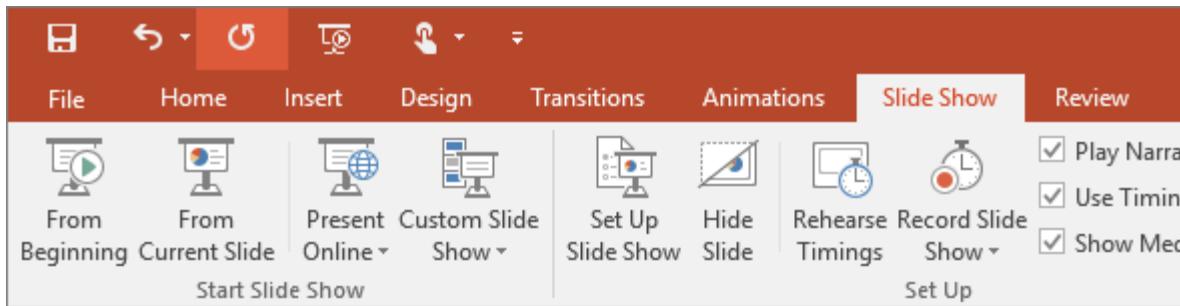
- ▶ Main sales page
- ▶ Support email address
- ▶ Summary

Introduce George
Q1 Sales results
New Product line

Give your presentation

On the **Slide Show** tab, do one of the following:

To start the presentation at the first slide, in the **Start Slide Show** group, click **From Beginning**.



If you're not at the first slide and want to start from where you are, click **From Current Slide**.

If you need to present to people who are not where you are, click **Present Online** to set up a presentation on the web, and then choose one of the following options:

To get out of Slide Show view at any time, on the keyboard, press **Esc**.

Activity



The class will be divided into groups of 5 people and then each group will work on the following case study presentation:

- You are all members of the marketing department of Coca Cola and are trying to investigate why the recent sales of Coca Cola in the state of Andhra Pradesh has dropped significantly.
- While carrying out your investigation, you find out that a new soft drink called “Fizz” has appeared in the state of Andhra Pradesh and is capturing the market by storm!
- It is your task to find out why is fizz performing well and what can be done to improve the market positioning of Coca Cola
- You are required:
 - To create a document your approach method for this activity in MS Visio using a flowchart
 - To create a sample database of 20 respondents interviewed by you in MS Excel
 - To present your findings in MS PowerPoint
 - To create a one page summary of your conclusion in MS Word
- You have 3 hours of time to do this activity after which, 2 members of the group will have to present their findings and the group will be judged on the same.



Check your Understanding

Please answer the following questions:

- 1.** Under which tab in Microsoft Word would you look under to change the page orientation?
A.Insert
B.Page Layout
C.File
D.View

- 2.** In which Microsoft Office product do you work with animations?
A.Excel
B.Word
C.PowerPoint
D.Publisher

- 3.** Explain in complete sentences how you would insert an image from the internet into a Word document or PowerPoint.

- 4.** Visio is used to create which of the following?
A.Organization Charts
B.Flowcharts
C.Floor plans
D.All of the above

- 5.** Which option must you select if you would like text to be typed around an image instead of just in front of or behind it?
A.Bring forward
B.Text Wrap
C.Insert Text Box
D.Crop

- 6.** In Excel, what are the boxes called?
A.Spreadsheet
B.Squares
C.Cells
D.Rectangles

- 7.** The "boxes" going from left to right are called...
A.Rows
B.Columns

C.Letters

D.Titles

8. _____ is the best Microsoft Office product to use to create a graph.

9. The best Microsoft Office product to use if you would like to create a calendar is...

A.Word

B.Excel

C.PowerPoint

D.Visio

Answer Key for facilitator:

1-B

2-C

4-D

5-B

6-C

7-A

8-Excel

9-B

Summary

- The Microsoft Office Package provides excellent options for documentation and has the right tools for creating all forms of documentation.
- MS Visio is mostly a tool used to create flowcharts, organization charts, brainstorm diagrams and floor plans.
- MS Word is the world's most widely used word processor. Most documents (including the book which you're reading currently) are developed on Microsoft Word
- MS Excel is a very powerful spreadsheet application and a great way to document all financial data, or large rows of data dumps. It also has a large number of inbuilt financial, mathematical and statistical computational functions
- MS PowerPoint is probably the best tool to present data of any form. Data from all other tools can be consolidated and exported into Microsoft PowerPoint in a very impressive format.

