असतो मा सद्गमय ।
तमसो मा ज्योतिर्गमय ।।
SANDIP FOUNDATION

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

Industry-Academia Innovative Practice

Proceeding of

10 Days Training Programme

on

"Big Data Analytics"

2nd May to 11th May 2016

Sponsored by

Department of Science and Technology Government of India

under

"Big Data Initiative"

Organized by

Sandip Foundation's

Sandip Institute of Technology & Research Centre, NashiK

Department of Computer Engineering

Objectives:

- To promote and foster Big Data Science, Technology and Applications in the country and to develop core generic technologies, tools and algorithms for wider applications in Govt.
- To understand the present status of the industry in terms of market size, different players providing services across sectors/ functions, opportunities, SWOT of industry, policy framework (if any), present skill levels available etc.
- To carryout market landscape survey to assess the future opportunities and demand for skill levels in next 10 years
- To carryout gap analysis in terms of skills levels and policy framework.
- To evolve a strategic Road Map and micro level action plan clearly defining of roles of various stakeholders - Government, Industry, Academia, Industry Associations and others with clear timelines and outcome for the next 10 years.

http://www.sitrc.sandipfoundation.org/ug-courses-computer-engineering/

BIG DATA ANALYTICS TRAINING PROGRAMME

2nd May to 11th May 2016

Activity Report

Sponsored By

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

Big Data Initiative Division
Department of Science and Technology, Government of India

Organized By

असतो मा सद्गमय ।
तमसो मा ज्योतिर्गमय ।।
SANDIP FOUNDATION

Department of Computer Engineering
Sandip Institute of Technology & Research Centre, NashiK

“Torture the Data Long enough and they will confess to anything”

-Statisticians Quip; Computer Scientist Whip

Ref: “Data data everywhere” Report by The economist and EMC Corporation
http://www.economist.com/node/15557443
https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

AT A GLANCE

11 days programme, featuring
33 hours of lectures (excluding breaks),
44 hours of lab assignments and
1 special workshop on proposal writing
8 Lab Assignments completed
4 case study demonstrated
10 software packages presented

1 panel discussion,
1 event banquet
9 academic speakers

100 applications received
97 applications shortlisted,
87 participants registered and
75 participants attended from
36 institutions
1 ITeS industrie
15 cities and
2 states

About DST's Big Data Analytics Training Programme

BDI: An R&D Perspective

By definition, Big Data, is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. In other words, big data is characterised by volume, variety (structured and unstructured data) velocity (high rate of changing) and veracity (uncertainty and incompleteness).

In the Big Data research context, so called analytics over Big Data is playing a leading role. Analytics cover a wide family of problems mainly arising in the context of Database, Data Warehousing and Data Mining research. Analytics research is intended to develop complex procedures running over large-scale, enormous in-size data repositories with the objective of extracting useful knowledge hidden in such repositories. One of the most significant application scenarios where Big Data arise is, without doubt, scientific computing. Here, scientists and researchers produce huge amounts of data per-day via experiments (e.g., disciplines like high-energy physics, astronomy, biology, bio-medicine, and so forth). But extracting useful knowledge for decision making purposes from these massive, large-scale data repositories is almost impossible for actual DBMS-inspired analysis tools. From a methodological point of view, there are also research challenges. A new methodology is required for transforming Big Data stored in heterogeneous and different-in-nature data sources (e.g., legacy systems, Web, scientific data repositories, sensor and stream databases, social networks) into a structured, hence well-interpretable format for target data analytics. As a consequence, data-driven approaches, in biology, medicine, public policy, social sciences, and humanities, can replace the traditional hypothesis-driven research in science.

Big Data: Science & Technology - Challenges

Some of the S&T challenges that researchers across the globe and as well as in India facing are related to data deluge pertaining to Astrophysics, Materials Science, Earth & atmospheric observations, Energy, Fundamental Science, Computational Biology, Bioinformatics & Medicine, Engineering & Technology, GIS and Remote Sensing, Cognitive science and Statistical data. These challenges requires development of advanced algorithms, visualization techniques, data streaming methodologies and analytics. The overall constraints that community facing are:

1. **The IT Challenge:** Storage and computational power
2. **The computer science:** Algorithm design, visualization, scalability (Machine Learning, network & Graph analysis, streaming of data and text mining), distributed data, architectures, data dimension reduction and implementation
3. **The mathematical science:** Statistics, Optimisation, uncertainty quantification, model development (statistical, Ab Initio, simulation) analysis and systems theory
4. **The multi-disciplinary approach:** Contextual problem solving

Big data analytics and the India equation

To tap the analytics momentum, India now needs to build a sustainable analytics eco-system that brings in a strong partnership across the industry players, government, and academia. Some of the key actions for analytics eco-system in India would be around.

1. **Talent Pool -** Create industry academia partnership to groom the talent pool in universities as well as develop strong internal training curriculum to advance analytical depth.
2. **Collaborate -** Form analytics forum across organization boundaries to discuss the pain-points of the practitioner community and share best practices to scale analytics organizations.
3. **Capability Development -** Invest in long term skills and capabilities that forms the basis for differentiation and value creation. There needs to be an innovation culture that will facilitate IP creation and asset development.
4. **Value Creation -** Building rigor to measure the impact of analytics deployment is very critical to earn legitimacy within the organization.

Big Data and analytics offers tremendous untapped potential to drive big business outcomes. For organizations to leverage India as a global analytics hub can be one of the key levers to move up their analytics maturity curve.

Broad contours of DST initiated BDI programme

- To promote and foster Big Data Science, Technology and Applications in the country and to develop core generic technologies, tools and algorithms for wider applications in Govt.
- To understand the present status of the industry in terms of market size, different players providing services across sectors/ functions, opportunities, SWOT of industry, policy framework (if any), present skill levels available etc.
- To carryout market landscape survey to assess the future opportunities and demand for skill levels in next 10 years
- To carryout gap analysis in terms of skills levels and policy framework.
- To evolve a strategic Road Map and micro level action plan clearly defining of roles of various stakeholders – Govt., Industry, Academia, Industry Associations and others with clear timelines and outcome for the next 10 years.

Course Objectives

- To promote and foster Big Data Science, Technology and Applications in the country and to develop core generic technologies, tools and algorithms for wider applications in Govt.
- To understand the present status of the industry in terms of market size, different players providing services across sectors/ functions, opportunities, SWOT of industry, policy framework (if any), present skill levels available etc.
- To carryout market landscape survey to assess the future opportunities and demand for skill levels in next 10 years
- To carryout gap analysis in terms of skills levels and policy framework.
- To evolve a strategic Road Map and micro level action plan clearly defining of roles of various stakeholders - Government, Industry, Academia, Industry Associations and others with clear timelines and outcome for the next 10 years.

Sandip Foundation's
Sandip Institute of Technology & Research Center, NashiK
DEPARTMENT OF COMPUTER ENGINEERING

Technical Advisory Committee and Resource Persons

Dr K. R. Murli Mohan
Head (Big Data Initiative), Department of Science & Technology, Ministry of Science & Technology New Delhi

Dr. S. G. Bhirud
Professor, Computer Engineering Department, VJTI, Mumbai
Former Adviser-I, e-GOVERNANCE CELL and LEGAL CELL, AICTE, New Delhi
Former Advisor-I and Chief Vigilance Officer VIGILANCE CELL, AICTE, New Delhi

Dr. V. M. Thakare
Professor and Head in Computer Science, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, SGB Amravati University, Amravati

Dr.K.Thirupathi Rao
Principal Investigator and Professor, Department of Computer Science and Engineering, Associate Dean Academics, K L University, Guntur District, Aandhra Pradesh

Dr. V. M. Wadhai
Principal, Sinhgad Academy of Engineering, Pune

Dr S T Gandhe
Principal, Sandip Institute of Technology and Research Centre, Nashik

Mr. Piyush Somani
Chief Executive Officer,
ESDS Software Solutions Pvt. Ltd, Nashik

Dr. S. S. Sane
Professor and Head, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik

Dr. P. N. Mahalle
Professor & Head (Computer Engg.)
Smt. Kashibai Navale College of Engineering, Pune

Dr. S. R. Sakhare
Professor & Head, Department of Information Technology, Vishwakarma Institute of Information Technology, Pune

Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

DEPARTMENT OF COMPUTER ENGINEERING

Dr. D.V.Patil
Professor, GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik

Dr. R.S. Tiwari
Ex-Director, Head Computer Department,YCMOU Nashik

Mr. Shrikant Pande,
Co-founder
Ctronics Solutions Pvt. Ltd, Amravati

Dr. M.U. Kharat
Professor & Head
Department of Computer Engineering
MET~BKC, Institute of Engineering, Nashik

Prof. N.M Shahane
Associate Professor, Department of Computer Engineering,
K. K. Wagh Institute of Engineering Education & Research, Nashik

Prof. S. R. Dhore
Professor and Head, Department of Computer Engineering,
Army Institute of Technology, Pune

Prof. T. B. Kute
Assistant Professor, Sandip Institute of Technology and Research Centre, Nashik

Mr. Prasad Chandane
ERP, Database, Big Data Evangelist, IBM, Pune

Mr Amol Kute
Torna Informatics,Pune

Dr Rajeev Papneja
Chief Technical Officer, ESDS Software Solutions Pvct. Ltd., Nashik

Mr. Pavan Tiwari
Co-founder
Deepdive Infotech, Badlapur, mumbai

Dr. P.M. Jawandhiya
Principal, Pankaj Laddhad Institute of Technology and Management Studies,Buldana

Dr. M. R. Sanghavi
Associate Professor and Head, Department of Computer Engineering,
SNJB's KBJ College of Engineering, Chandwad, Nashik

Prof. A. R. Kalugade
Assistant Professor, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune

Prof G. K. Bhamare
Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies and Research, Nashik

Nirmit Kale and Rushikesh Jadhav
Technical Architect, eNight Cloud R & D Division,
ESDS Software Solutions Pvct. Ltd., Nashik

Organizing Committee

Chief Patron

Dr. SandipKumar N Jha
Hon. Chairman, Sandip Foundation

Prof. Mohini P. Patil
Hon. General Manager, Sandip Foundation

Prof. Prakash I. Patil
Hon. Mentor,Sandip Foundation

Dr. Sanjay T. Gandhe
Principal
Sandip Institute of Technology and Research Centre

Coordinator

Prof. Amol D. Potgantwar
Head, Department of Computer Engineering and Information Technology

Convener
Prof. (Mrs.) Namrata D. Ghuse

Co-Convener
Prof. (Mrs.) Shweta N. Patil

Assistant Professor, Department of Computer Engineering

Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

DEPARTMENT OF COMPUTER ENGINEERING

Department of Computer Enginering	**Assistant Professors** • Prof. Santosh D Kumar • Prof. Naresh C. Thoutam • Prof. Sandip M. Walunj • Prof. Rakesh S. Shirsath • Prof. Amit R. Gadekar • Prof. Sandeep S. Jore • Prof Amit H. Palve • Prof. Samadhan A. Sonawane • Prof.(Mrs.) Bharati A. Patil • Prof. Mahesh V. Korade • Prof.(Mrs.) Anjali P. Deore • Prof. Kanchan B. Mahajan • Er. Nikhil L. Kulkarni • Prof. Harish P. Patil • Prof. Pradynesh J. Bhisikar • Prof. (Ms.)Bhagyashree F. More • Prof. (Mrs.) Ankita V. Karale • Prof. Jayvant N Rajole	**Department of Information Technology**	**Assistant Professors** • Prof. Vivek M. Waghamare • Prof. Bhushan S Chaudhari • Prof. (Mrs.) Swati Rahul Khokale • Prof. Pravin R. Pachorkar • Prof. Vivek D.Patil • Prof. Tushar J Surwade • Prof. Mohan C Nikam • Prof. (Ms.) Sneha Khaire • Prof. (Ms.) Priyanka Salunke • Prof. (Ms.) Neha Kale • Prof. (Mrs.) Saba Shaikh • Prof. Pavan S Ahire • Prof. Tushar P Jagtap • Prof. Deepak Landge • Prof. (Mrs.) Prajkta Shirke • Prof. Yogesh Kolhe
	Technical Assistant • Mr. Sachin Lokhande • Mrs. Swati Thakare		**Technical Assistant** • Mr. Salim Khan • Mr. Amol Nalge

List of Talks

Academic

Sr. No.	Name and Affiliation	Talk Title
1	Dr. S.G.Bhirud	• Data formats,Data write, cleaning transformation • R basics, lists, arrays, matrices, tables, function • R scatterplotm, biplot, correlation, histogram
2	Dr. V. M Thakare	• Handling large matrices • Sampling from given distributors
3	Dr. S.S.Sane	• Linear models and generalized linear models • Modeling and prediction with R package lm • Plotting models fit to data
4	Prof. N.M Shahane	• Diagnostics plots Finding significant variables • Classification and Regression Tree(CART) • Very Fast Decision Trees(VFDT)
5	Dr. P. N.Mahalle	• Model performance via sensitivity • Specificity, precision and recall • ROC curves
6	Dr. S.R.Sakhare	• Ensemble methods • Bagging, Boosting and Random Forests KNN • SVM and Neural networks
7	Dr. D.V.Patil	• High Dimensional data matrix maipulation • Multi-Dimesinal Scaling(MDS) • Singular value decomposition(SVD)
8	Dr. M.U. Kharat	• Complex Event Processing • Basics of One-pass computing, Online algorithms
9	Dr. P.M.Jawandhiya	• Typical case studies may include Text Analysis • Sentiments analysis, Social media mining
10	Prof. (Dr.) S. S. Prabhune	• Boosting and Ransom Forest KNN • Neural Network
11	Prof. (Dr.) M. R. Sanghavi	• Understanding MapReduce architechture • Hadoop Distributed File System(HDFS)
12	Prof. S.R. Dhore	• Principal Components Analysis (PCA) • Singular Value Decomposition (SVD)
13	Prof. Amol Kalugade	• Features of various databases • R packages RmySQL,Rexcel, RmongoDB, Rhive
14	Prof. T. B. Kute	• Web mining • Internet of things; Data fusion • Geo-informatics and spatial statistical analysis

Industrial & Hands On Sessions

Sr. No.	Name and Affiliation
1	**Dr Rajeev Papneja** ESDS Software Solutions Pvt Ltd, Nashik
2	**Mr Shrikant Pande** Ctronics Solutions Pvt Ltd, Amravati
3	**Mr Pavan Tiwari** Deepdive Infotech, Badlapur
4	**Mr Mukund Mishra** Technical Architect Big Data Hadoop, L&T Infotech
5	**Mr Prasad Chandane** ERP, Database, Big Data Evangelist, IBM, Pune
6	**Mr Amol Kute** Torna Informatics,Pune
7	**Mr Nirmit Kale** Technical Architect, eNight Cloud R & D Division, ESDS Software Solutions Pvct. Ltd., Nashik
8	**Mr Rushikesh Jadhav** Technical Architect, eNight Cloud R & D Division, ESDS Software Solutions Pvct. Ltd., Nashik

Laboratory based Practices Demonstrated

Day	Name of Expert	Topics
2th May	Shrikant Pandey	R Programming Basics
3th May	Shrikant Pandey	Graphics & Statistics
5th May	Prof S S Dhore	Theory
6th May	Pavan Tiwari	Installation of Ambari & Hadoop
7th May	Tushar Kute	Installation of Hadoop 2.6.3,Hive,Pig,
8th May	Rushikesh Jadhav(ESDS)	Introduction to Cloud Computing & Registration
9th May	Prof. Amol Kalgoude	Basics of RDBMS & Basics MongoDB Commands
10th May	Mr Amol Kute & Prof Tushar Kute	File Upload & Download on Hadoop & Assignments
11th May	Mr Amol Kute & Prof Tushar Kute	File Upload & Download on Hadoop & Assignments

List of Participants

Sr.No.	Name	Affiliation
1	Prof Abhay Gaidhani	Sandip Institute of Engineering and Management, Nashik
2	Prof Amit G Patil	
3	Prof (Ms) Nikhita Amit Patil	R.M.D Sinhgad School of Engineering ,Pune
4	Prof (Ms) Abolee Hemant Patil	Matoshri College of Engineering & Research Centre, Nashik
5	Prof Bhushan Pawar	MES College of Engineering, Pune
6	Prof Rutul D. Dhomse	Imperial College of Engineering And Research, Nashik
7	Prof Vishal B shinde	Sandip Institute of Engineering and Management, Nashik
8	Prof Mubin Tamboli	Amrutvahini College of Engineering, Sangamner
9	Mr. Ganesh D Puri	
	Prof Sachin A Thanekar	
10	Mr. Sandip Ramkrishna Pandit	
11	Prof (Mrs) Sonal Gore	Pimpri Chinchwad College of Engineering,Pune
13	Prof (Mrs) Smita Patil	Pravara Rural Engineering College, Loni
14	Prof. Dnyaneshwar Wavhal	JCEI's Jaihind College of Engineering, Kuran
15	Prof. Amrut V Kanade	
16	Prof Chandrakant R Barde	GES'S R. H. Sapat College of Engineeriing Management Studies &Research, Nashik
17	Prof Bhavesh B. Shah	Suman Ramesh Tulsiani Technical Campus - Faculty of Engineering, Pune
18	Prof Amit Arun Gawade	Saraswati College of Engineering, Kharghar
19	Prof (Mrs) Abha P Nayak	Dr. D. Y. Patil Institute Of Engineering & Technology, Pimpri, Pune
20	Prof (Ms) Sujata S Wakchaure	MIT's COE, Pune
21	Prof (Ms) Neha Bhikchand Birla	
22	Prof B S Sonawane	
23	Prof Anil V Turukmane	P E S College of Engineering Auranagbad
24	Ms.Prajakta Suhasrao Kale	Government College of Engineering, Auranagabad
25	Prof (Mrs) Vaishali Wangikar	MIT Academy of Engineering, Pune
26	Prof N R Wankhade	Late G N Sapkal College of Engineering, Nashik
27	Prof. Mohan Nikam	
28	Prof (Mrs) Rupali Tajanpure	NDMVPS's KBT COE, Nashik

29	Prof P D Jadhav	MET's Institute of Engineering, Nashik
30	Prof (Mrs) Apeksha R Gawande	
31	Prof Gurunath G Machhale	TKIET, Warananagar, Kolhapur
32	Prof Satish S Banait	KKWIEER, Nashik
33	Prof Swati A Joshi	Sinhgad Academy of Engineering
34	Prof (Mrs) Rupali R Shewale	NDMVP'S KBTCOE, Nashik
35	Prof Sharad M Rokade	SVIT, Chincholi, Nashik
36	Prof Uttam R Patole	
37	Prof Swapnil N Dixit	Ashoka Center For business and Computer studies, Nashik
38	Prof (Mrs) Yogita K Desai	SNJB's LS KBJ COE, Chandwad
39	Prof Ranjit M Gawande	Research Schloar, Matoshri COE&R, Nashik
40	Prof. Bhushan Chaudhari	
41	Prof Madhuri D Kawade	SNJB's LS KBJ COE,Chandwad.
42	Mr Shakil Shaikh	JMN Infotech Pvt Ltd, Nashik
43	Miss Priya Lunkad	
44	Mr Mahesh V. Korde	PG Student, MIT Aurangabad
45	Mr Harish Patil	PG Student, COE, Jalgoan
46	Er Nikhil L Kulkarni	PG Student, Matoshri COE&R, Nashik
47	Mr Vivek D Patil	
48	Mr Pavan Ahire	PG Student, MET's IOE, Nashik
49	Miss Kaveri Sonawane	
50	Mr Vaibhav Bhor	PG Student, Late G N Sapkal College of Engineering, Nashik
51	Miss Sneha A Khaire	
52	Mr Jayvant N Rajole	PG Student, Everest COE, Aurangabad
53	Prof Naresh C Thoutam	PG Student, Anna University
54	Mrs Swati M Thakre	PG Student, BVCOE&R, Nashik
55	Mr Pravin R Pachorkar	PG Student, Sanjivani College of Engineering, Kopergaon
56	Mr Amol Nalge	PG Student, SNDCOE, Yeola

57	Mrs Swati Gawand	PG Student, SITRC, Nashik
59	Miss Bhagyashree F More	
60	Miss Neha A Kale	
61	Miss Saba A Shaikh	
62	Miss Priyanka Salunkhe	
63	Miss Gayatri M Tambe	
64	Mr Ajit Patil	
65	Miss Shivani Gaikar	
66	Mr Vaibhav Desale	
67	Mr Tushar P Jagtap	
68	Mrs. Ankita V. Karale	
69	Prof Amol D Potgantwar	Research Scholar, Sant Gadge Baba Amravati University, Amravati
70	Prof Sandeep S Jore	Assistant Professor, Sandip Institute of Technology and Research Centre, Nashik
71	Prof Samadhan A Sonavane	
72	Prof Rakesh S Shirsath	
73	Prof Amit H Palve	
74	Prof Pradynesh J Bhisikar	
75	Dr. Sachin Babar	Professor, Sinhgad Institute of Technology, Lonavala

Media Coverage

इंजिनिअरिंगच्या विद्यार्थ्यांना 'बिग डाटा' तंत्रज्ञानाचे धडे

केंद्र सरकारच्या 'विज्ञान-तंत्रज्ञान'तर्फे २ मेपासून विशेष कार्यशाळा

प्रतिनिधी | नाशिक

केंद्र सरकारच्या विज्ञान व तंत्रज्ञान विभागाच्या वतीने संदीप फाउंडेशन संचलित संदीप इन्स्टिट्यूट ऑफ टेक्नॉलॉजी ॲण्ड रिसर्च सेंटर महाविद्यालयात इंजिनिअरिंगच्या विद्यार्थ्यांसाठी 'बिग डाटा ॲनालिटिक्स' या विषयावरील राष्ट्रीय कार्यशाळेचे आयोजन करण्यात आले आहे. दहा दिवसीय ही कार्यशाळा २ ते ११ मे या कालावधीत होणार असून, त्यासाठी भारत सरकारतर्फे सहा लाखांचे अनुदान मिळाले आहे.

बिग डाटा तंत्रज्ञानाचा वापर व प्रसार व्हावा, या उद्देशाने देशभर कार्यशाळांचे आयोजन करण्यात येत आहे. डाटा सायन्स, डाटा मॉडेल्स, मॅप रिड्यूस, बिग डाटा स्टडीज, आर प्रोग्रॅमिंग अशा विविध विषयांवर या कार्यशाळेत प्रात्यक्षिके सादर

देशभरातील विद्यार्थी होणार कार्यशाळेत सहभागी...

अभियांत्रिकीतील संगणक शाखेचे शिक्षण घेणारे देशभरातील विद्यार्थी व प्राध्यापक या कार्यशाळेत सहभागी होणार आहेत. या विद्यार्थ्यांसाठी संदीप फाउंडेशनतर्फे सर्व सुविधा उपलब्ध करून दिली असल्याची माहिती संस्थेतर्फे देण्यात आली आहे.

केली जाणार आहेत. बिट डाटा तंत्रज्ञान क्षेत्रात संशोधन करणारे अनेक तज्ज्ञ या कार्यशाळेत सहभागी होणार असून, विद्यार्थ्यांना मार्गदर्शन करतील. फाउंडेशनचे अध्यक्ष डॉ. संदीपकुमार झा, मेंटॉर पी. आय. पाटील, सरव्यवस्थापिका मोहिनी पाटील, प्राचार्य डॉ. संजय गंधे, डॉ. राकेश पाटील, डॉ. प्रसाद बाविस्कर, विभागप्रमुख प्रा. अमोल पोटगंटवार, प्रा. नम्रता घुसे यांच्या मार्गदर्शनाखाली ही कार्यशाळा होणार आहे.

22.04.2016 Divya Marathi pg 2

22.04.2016 Divya Marathi (City)

मंथन | केंद्र सरकारच्या विज्ञान व तंत्रज्ञानतर्फे राष्ट्रीय कार्यशाळा

बिग डेटा तंत्रज्ञानाच्या प्रसाराची गरज : भिरुड

इंजिनिअरिंगच्या विद्यार्थ्यांना बिग डेटा तंत्रज्ञानाचे धडे

प्रतिनिधी । नाशिक

सर्वच क्षेत्रात बिग डेटा तंत्रज्ञान फायदेशीर असून, त्याचा प्रचार व प्रसार करण्याची गरज आहे. डेटा सायन्स, डेटा मॉडेल्स, मॅप रिड्यूस, बिग डेटा स्टडीज, आर प्रोग्रॅमिंग अशा विविध विषयांच्या माध्यमातून इंजिनिअरिंगच्या विद्यार्थ्यांनी बिग डेटा तंत्रज्ञान समजून घेऊन त्याचा वापर करण्याचे आवाहन एआयसीटीइचे माजी सल्लागार प्रा. डॉ. एस. जी. भिरुड यांनी येथे केले.

केंद्र सरकारच्या विज्ञान व तंत्रज्ञान विभागाच्या वतीने संदीप फाउंडेशन संचलित संदीप इन्स्टिट्यूट ऑफ टेक्नॉलॉजी ॲण्ड रिसर्च सेंटर महाविद्यालयात इंजिनिअरिंगच्या विद्यार्थ्यांसाठी 'बिग डेटा ॲनालिटिक्स' या विषयावरील राष्ट्रीय कार्यशाळेचे आयोजन करण्यात आले आहे. दहा दिवसीय ही कार्यशाळा २ ते ११ मे या कालावधीत होत असून, त्यासाठी भारत सरकारतर्फे सहा लाखांचे अनुदान मिळाले आहे. संत गाडगेबाबा अमरावती विद्यापीठातील विभागप्रमुख प्रा. डॉ. एम. व्ही. ठाकरे यांच्या उपस्थितीत उद्घाटन झाले. या वेळी भिरुड बोलत होते. या प्रसंगी फाउंडेशनचे अध्यक्ष डॉ. संदीपकुमार झा, मेंटॉर पी. आय. पाटील, सरव्यवस्थापिका मोहिनी पाटील, प्राचार्य डॉ. संजय गंधे, डॉ. राकेश पाटील, डॉ. प्रसाद बाविस्कर, विभागप्रमुख प्रा. अमोल पोटगंटवार, प्रा. नम्रता घुसे आदी उपस्थित होते.

देशभरातील विद्यार्थी

अभियांत्रिकीतील संगणक शाखेचे शिक्षण घेणारे देशभरातील १०० हून अधिक विद्यार्थी व प्राध्यापक या कार्यशाळेत सहभागी झाले आहे. या विद्यार्थ्यांसाठी सर्व सुविधा उपलब्ध करून देण्यात आल्या आहेत. तसेच बिग डेटा तंत्रज्ञान क्षेत्रातील तज्ज्ञांच्या व्याख्यानांचे आयोजन करण्यात आले आहे.

04.05.2016 Divya Marathi

Sandip Foundation's
Sandip Institute of Technology & Research Center, Nashik
DEPARTMENT OF COMPUTER ENGINEERING
Big Data Analytics Training Programme
Minutes of Day – I

Department of Sciences & Technology
Government of India

Date: 02/05/2016

Day 1: Big Data Analytics session inaugurated with a great zeal along with sessions on "Data Science Basics"

- Inauguration Event for 10 days Department of Science and Technology, Govt. of India sponsored training programme started with lightening the Lamp in front of Saraswati Devi (Godess of Knowledge).
- Dignitories present on the Dias included:
  - Prof. (Dr.) S. G. Bhirud, Professor, Computer Engineering Department, VJTI, Mumbai. (*Chief Guest*)
  - Prof. (Dr.) V. M. Thakare, Professor and Head in Computer Science, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, SGB Amravati University, Amravati(*Guest of Honour*)
  - Prof. (Dr.) S. T. Gandhe, Principal, SITRC, Nashik
  - Prof. (Dr.) A. G. Jadhav, Principal, SIPS, Nashik
  - Prof. S. D. Pawar, Principal, SIP, Nashik
  - Prof. A. D. Potgantwar, Head, Department of Computer Engineering and Information Technology, SITRC, Nashik
  - Prof. (Mrs.) N.D. Ghuse, Convener, Big Data Analytics Training Programme-2016, SITRC, Nashik
- Felicitation of Chief Guest and Guest of Honour done by Prof. (Dr.) S. T. Gandhe and Prof. (Dr.) A. G. Jadhav respectively along with intorduction to their portfolio by Prof. (Mrs.) S. N. Patil as follows:
  - *Dr. S. G. Bhirud*, Professor, Computer Engineering Department, VJTI, Mumbai.
    Pursued Ph.D. in EC and Computer Science from Swami Ramanad Teerth Marathwada University, Nanded in 2001 and continuted to serve the education sector with all of possbile skills. Sir has numerous publications in Electronics and Computer, Image processing, Neural network and Digital Signal processing.
    AICTE's Adviser-I for e-GOVERNANCE CELL and LEGAL CELL. Also AICTE's Advisor-I and Chief Vigilance Officer for VIGILANCE CELL.
    His continous contribution in diversified research fields motivates all of us for thinking about the current buzzwords SMAC(*Social, Mobility, Analytics and Cloud*).

  - *Dr. V. M. Thakare*, Professor and Head in Computer Science, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, SGB Amravati University, Amravati
    Awarded with 2 UGC fellowships in 10th Plan of Govt of India.
    Statistical Views on his Career Excellence:
    21+ years of Experience
    Member of Expert Committee, Advisory Board, Board of Studies and Selection Committees in eminent universities across the nation.
    10+ Research Scholars awarded Ph.D.
    Guided 300+ projects at M.E./M.S./M.Phil./M.C.A. level
    85 + Research Paper published in Internatinal Journals and International/National Conferences
    55+ Keynote addresses and invited talks delivered in India and overseas.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – I

Department of Sciences & Technology
Government of India

Date: 02/05/2016

45+ subjects as a Proactive Academician

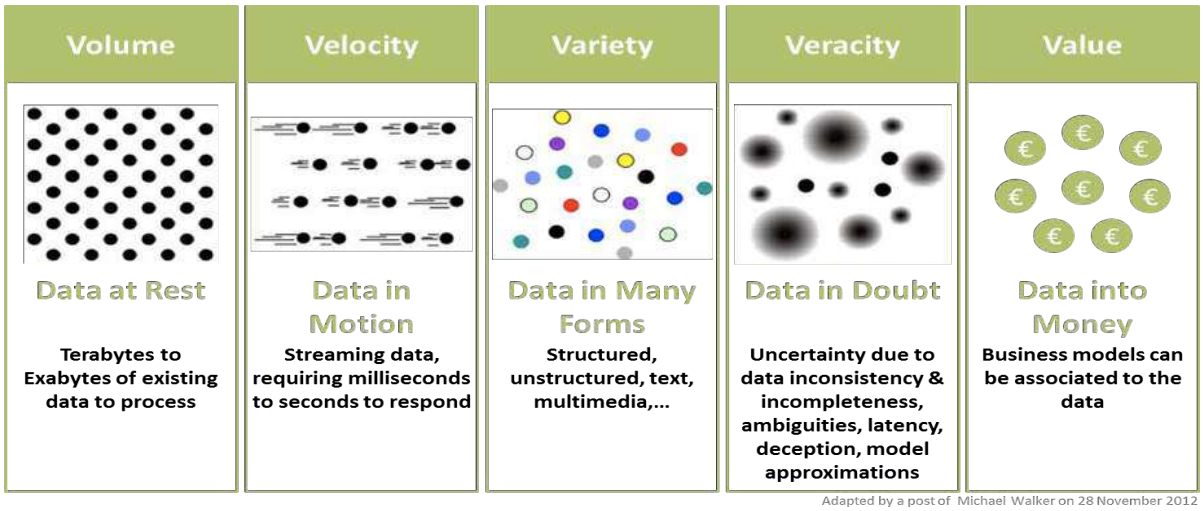
- His Thesis entitled "Study of Computer Architectures & Related AI Techniques for trajectories of Robot" keeps us directing towards the generation of Data at huge amount and need of analyzing it.

Key points from Prof. (Dr.) S T Gandhe's Welcome Speech:

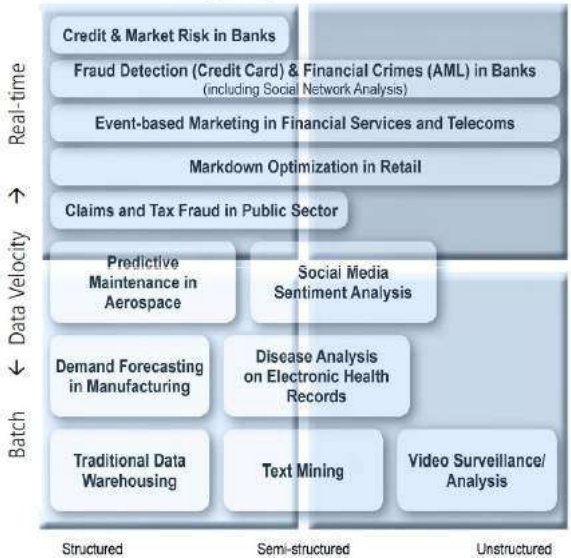
- Upcoming 4th Industrial revolution will affect 5 million jobs
- Currently, as an developing nation, India's 65% revenue generation is form service sector
- Big Data Analytics has not been limited to Computer/IT; it has widespread culture where data is in motion and generated streamline manner.

Key points from Dr. S. G. Bhirud's Keynote Speech:

- According to Gartner's definition, there are top 10 V's related to Big Data:
  Volume, Velocity, Variety, Veracity, Value, Validity, Variability, Venue, Vocabulary, Vagueness.
- Focus on first 5 V's:



- By end of 2020, there will be huge paradigm shift to automobile sector; it will become data driven
- Structure of Big data can be divided into
  - STRUCTURED: Most traditional data sources
  - SEMI-STRUCTURED: Many sources of big data
  - UNSTRUCTURED: Video data, audio data



Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – I

Department of Sciences & Technology
Government of India

Date: 02/05/2016

- How Big data impacts on IT
  - Big data is a troublesome force presenting opportunities with challenges to IT organizations.
  - Presently 4.4 million IT jobs in Big Data ; 1.9 million is in US itself
  - India will require a minimum of 1 lakh data scientists in the next couple of years in addition to data analysts and data managers to support the Big Data space.
- India - Big Data
  - Gaining attraction
  - Huge market opportunities for IT services (82.9% of revenues) and analytics firms (17.1 %)
  - Current market size is $200 million. By 2015 $1 billion
  - The opportunity for Indian service providers lies in offering services around Big Data implementation and analytics for global multinationals
- Broader view of Big Data:

- The roots of Big Data Analytics lies in Data Mining and Knowledge Extraction:

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – I

Department of Sciences & Technology
Government of India

Date: 02/05/2016

Handouts from Mr Shrikant Pande's Hands on Session:

- R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.
- The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

Features of R

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

With above features, R is world's most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications.

Date: 02/05/2016

Example:

You can perform R programming in two ways as follows:

1. With R Command Prompt

Type R on console

```
$ R
```

This will launch R interpreter and you will get a prompt > where you can start typing your program as follows:

```
> myString <- "Hello, World!"
> print ( myString)

Output:
[1] "Hello, World!"
```

For more details, please refer:

`Material/Day I/Mr Shrikant Pande/`

Date: 03/05/2016

Day 2: The Training Intensifies with Heavy Expert Lectures

Handouts from Prof N M Shahane's Expert Talk:

Contents to Cover:

a) Linear regression modeling and diagnostics
b) Multiple linear regression modeling
c) Logistic regression and binary classification

The use of probability to measure uncertainty and variability dates back hundreds of years. Probability has found application in areas as diverse as medicine, gambling, weather forecasting, and the law.

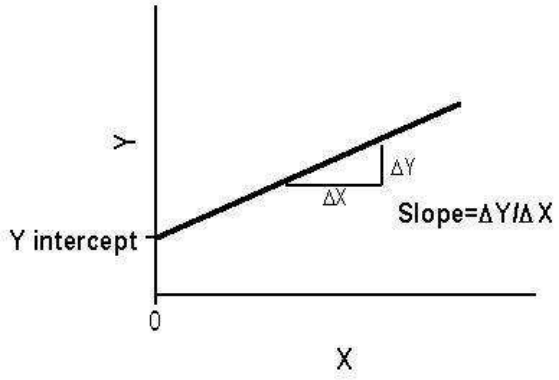
- The concepts of chance and uncertainty are as old as civilization itself. People have always had to cope with uncertainty about the weather, their food supply, and other aspects of their environment, and have striven to reduce this uncertainty and its effects. Even the idea of gambling has a long history. By about the year 3500 b.c., games of chance played with bone objects that could be considered precursors of dice were apparently highly developed in Egypt and elsewhere. Cubical dice with markings virtually identical to those on modern dice have been found in Egyptian tombs dating from 2000 b.c. We know that gambling with dice has been popular ever since that time and played an important part in the early development of probability theory.
- It is generally believed that the mathematical theory of probability was started by the French mathematicians Blaise Pascal (1623-1662) and Pierre Fermat (1601-1665) when they succeeded in deriving exact probabilities for certain gambling problems involving dice. Some of the problems that they solved had been outstanding for about 300 years. However, numerical probabilities of various dice combinations had been calculated previously by Girolamo Cardano (1501-1576) and Galileo Galilei (1564-1642).
- The theory of probability has been developed steadily since the seventeenth century and has been widely applied in diverse fields of study. Today, probability theory is an important tool in most areas of engineering, science, and management. Many research workers are actively engaged in the discovery and establishment of new applications of probability in fields such as medicine, meteorology, photography from satellites, marketing, earthquake prediction, human behavior, the design of computer systems, finance, genetics, and law. In many legal proceedings involving antitrust violations or employment discrimination, both sides will present probability and statistical calculations to help support their cases.

Date: 03/05/2016

Regression:

- Regression is an incredibly powerful statistical tool that, when used correctly, has the ability to help you predict certain values. When used with a controlled experiment, regression can actually help you predict the future. Businesses use it like crazy to help them build models to explain customer behavior. You're about to see that the judicious use of regression can be very profitable indeed.
- The word "regression" is more a historical artifact than something analytically illuminating. The guy who discovered the method, Sir Francis Galton (1822-1911), was studying how the height of fathers predicted the height of their sons. His data showed that, on average, short fathers had taller sons, and tall fathers had shorter sons. He called this phenomenon "regression to mediocrity."

- Linear regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a linear model.
- A linear regression equation can be written as $Y_p = mX + b$, where Yp is the predicted value of the dependent variable, m is the slope of the regression line, and b is the Y-intercept of the regression line.



- In statistics, linear regression is a method of estimating the conditional expected value of one variable y given the values of some other variable or variables x. The variable of interest, y, is conventionally called the "dependent variable". The terms "endogenous variable" and "output variable" are also used. The other variables x are called the "independent variables". The terms "exogenous variables" and "input variables" are also used. The dependent and independent variables may be scalars or vectors. If the independent variable is a vector, one speaks of multiple linear regression.

Statement of the linear regression model

- A linear regression model is typically stated in the form $y = ? + \beta x + ?$
- The right hand side may take other forms, but generally comprises a linear combination of the parameters, here denoted - and β. The term ε represents the unpredicted or unexplained variation in the dependent variable; it is conventionally

called the "error" whether it is really a measurement error or not. The error term is conventionally assumed to have expected value equal to zero, as a nonzero expected value could be absorbed into - . See also errors and residuals in statistics; the difference between an error and a residual is also dealt with below. It is also assumed that is ε independent of x.

Robust regression

- A useful alternative to linear regression is robust regression in which mean absolute error is minimized instead of mean squared error as in linear regression. Robust regression is computationally much more intensive than linear regression and is somewhat more difficult to implement as well.
- Robust regression usually means linear regression with robust (Huber-White) standard errors (e.g. relaxing the assumption of homoskedasticity).

An equivalent formulation which explicitly shows the linear regression as a model of conditional expectation is with the conditional distribution of y given x essentially the same as the distribution of the error term. A linear regression model need not be affine, let alone linear, in the independent variables x.

Advantages / Limitations of Linear Regression Model :

1. Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.
2. Linear regression is often inappropriately used to model non-linear relationships.
3. Linear regression is limited to predicting numeric output.
4. A lack of explanation about what has been learned can be a problem.

Multiple linear regression modeling

Refer Material/Day II/Prof N M Shahane/ Day II NMS Multiple Linear Regression Model.pdf

Date: 03/05/2016

Handouts from Mr Shrikant Pande's Hands on session:

Introduction to R

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

Features of R

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

With above features, R is world's most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications.

Sandip Foundation's
Sandip Institute of Technology & Research Center, Nashik
DEPARTMENT OF COMPUTER ENGINEERING
Big Data Analytics Training Programme
Minutes of Day – III

Department of Sciences & Technology
Government of India

Date: 04/05/2016

DAY 3 - A Comprehensive demonstration of SPSS predictive analytics software for Receiver Operating Characteristics (ROC) Curves

Handouts From Dr. Parikshit N. Mahalle's Expert Talk:

DATA MODEL PERFORMANCE-"ROC PERSPECTIVE"

- IN 2015 - TODAY, A PETABYTE OF DATA --1,024 TERABYTES, TO BE EXACT --PROBABLY MEETS MANY PEOPLE'S DEFINITION OF "BIG DATA."
- IN 2025 - FAST-FORWARD 10 YEARS, HOWEVER, AND A PETABYTE NO LONGER WILL QUALIFY AS BIG --AT LEAST NOT IN THE ENTERPRISE.

DATA EXPLOSION

Each day the following happens:
There are 1440 minutes per day...that means there are approximately

- 294 BILLION emails sent every day!
- 6 BILLION Google Searches each day!
- 3.5 BILLION Facebook messages posted daily!
- 40 Million Tweets shared each day!
- Average Retailer Generates 2.5 Petabytes Per Day

Source: https://www.gwava.com/blog/internet-data-created-daily-2014

When the size of the data, itself, becomes a problem
When the "old way" of processing data just doesn't work effectively
It's "big" when we have to rethink:

- How we store that much data?
- How we move that much data?
- How we extract, load & transform that much data?
- How we explore and analyze that much data?
- How we process and get meaningful insights from that much data?
- Data Ingestion, Storage, Processing, Reporting and Action

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

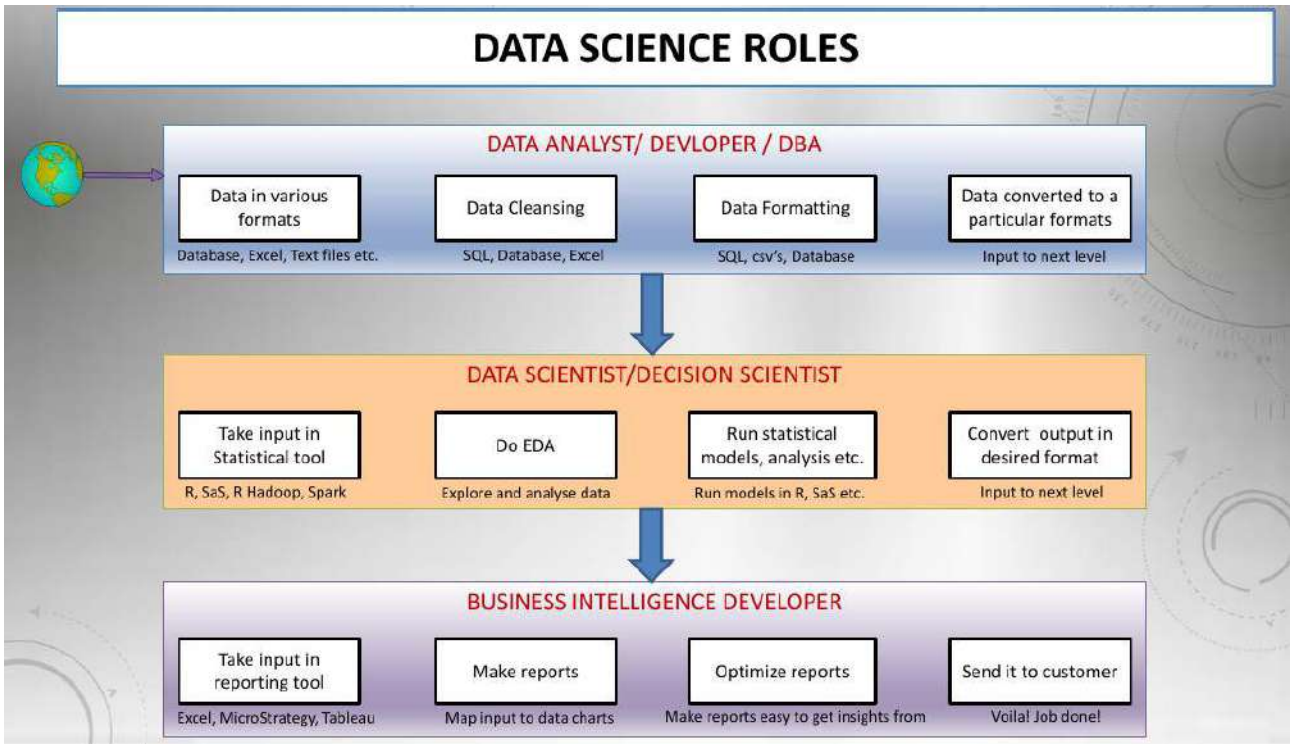
DEPARTMENT OF COMPUTER ENGINEERING

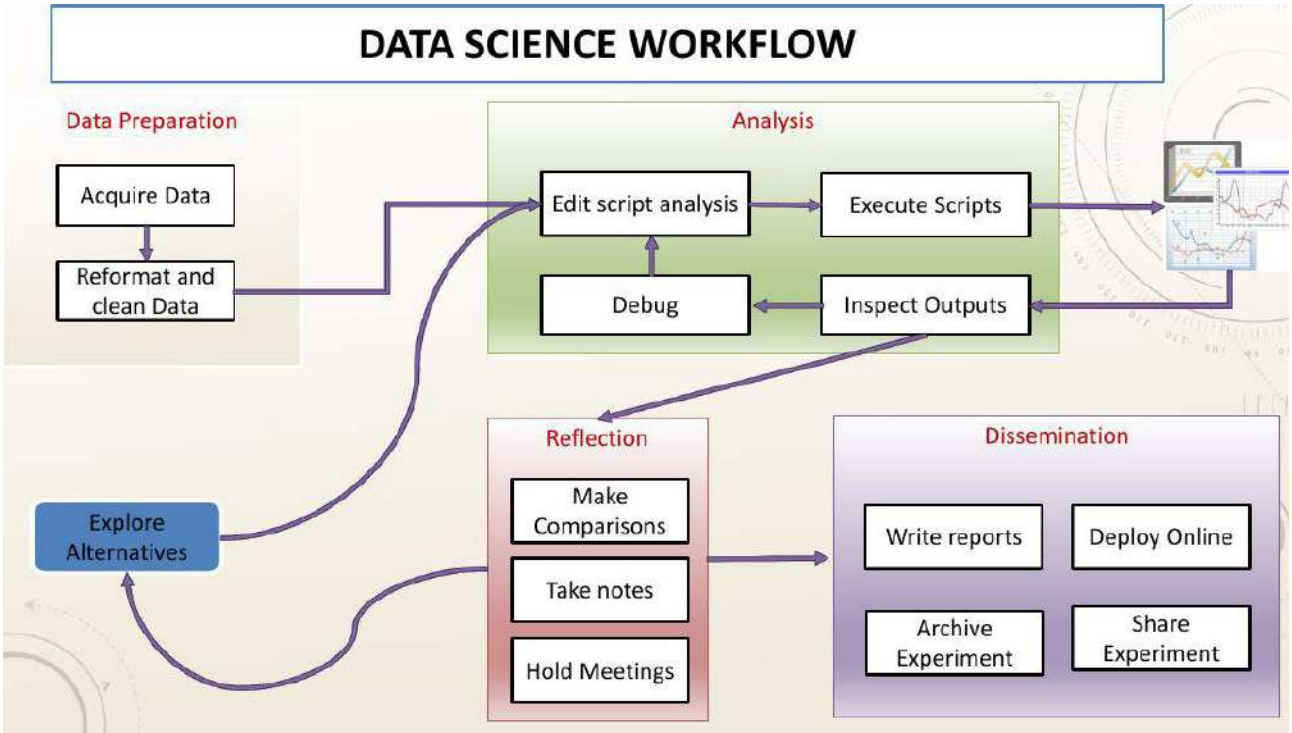
Big Data Analytics Training Programme

Minutes of Day – III

Department of Sciences & Technology Government of India

Date: 04/05/2016





Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – III

Department of Sciences & Technology
Government of India

Date: 04/05/2016

WHY DATA MODEL

- Receiver Operating Characteristics curves are used in machine learning applications to assess classifiers.
- Assessing classifiers means deciding a cutoff value for given data set.
- Consider the medical example, where 200 is considered as cutoff value for B12. All patients below 200 B12 value are considered as deficient and above 200 are considered as normal. However, it is also possible that the patients with B12 value more than 200 are deficient (False Positive) and below 200 are normal (False Negative).
- Receiver operating characteristic (ROC) curves are used in medicine to determine a cutoff value for a clinical test. For example, the cutoff value of 4.0 ng/ml was determined for the prostate specific antigen (PSA) test for prostate cancer. A test value below 4.0 is considered to be normal and above 4.0 to be abnormal. Clearly there will be patients with PSA values below 4.0 that are abnormal (false negative) and those above 4.0 that are normal (false positive). The goal of an ROC curve analysis is to determine the cutoff value.

Date: 04/05/2016

Handouts From Dr. Dipak V Patil's Expert Talk:

Data Analytics is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

- **Valid:** The patterns hold in general.
- **Novel:** We did not know the pattern before.
- **Useful:** We can devise actions from the patterns.
- **Understandable:** We can interpret and understand the patterns.

Supervised learning

Classification and regression

Unsupervised learning

Clustering
Dependency modeling
Associations, summarization, causality
Outlier and deviation detection
Trend analysis and change detection

Ensemble Methods: Increasing the Accuracy

Ensemble methods

Use a combination of models to increase accuracy
Combine a series of k learned models, M_1, M_2, …, M_k, with the aim of creating an improved model M*
Popular ensemble methods

Bagging: averaging the prediction over a collection of classifiers
Boosting: weighted vote with a collection of classifiers
Ensemble: combining a set of heterogeneous classifiers

Bagging: Boostrap Aggregation
Analogy: Diagnosis based on multiple doctors' majority vote
Training

Given a set D of d tuples, at each iteration i, a training set Di of d tuples is sampled with replacement from D (i.e., bootstrap)
A classifier model Mi is learned for each training set Di

Classification: classify an unknown sample X
Each classifier Mi returns its class prediction
The bagged classifier M* counts the votes and assigns the class with the most votes to X
Prediction: can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple

Accuracy
Often significantly better than a single classifier derived from D
For noise data: not considerably worse, more robust
Proved improved accuracy in prediction

B)KNN

The KNN has been reported as one of the widely used text classification approaches due to its simplicity and efficiency in handling various types of text classification tasks. However, there exists a major problem of the KNN in determining the appropriate value for parameter K in order to guarantee high classification effectiveness. This is due to the fact that the selection of the value of parameter K has high impact on the accuracy of the KNN classifier. Other than determining the optimal value of parameter K, the KNN is also a lazy learning method which keeps the entire training samples until classification time. Hence, the computational process of the KNN has become intensive when the value of parameter K increases.

- **Properties**

k-NN is a special case of a variable-bandwidth, kernel density "balloon" estimator with a uniform kernel.

The naïve version of the algorithm is easy to implement by computing the distances from the test example to all stored examples, but it is computationally intensive for large training sets. Using an appropriate nearest neighbor search algorithm makes *k*-NN computationally tractable even for large data sets. Many nearest neighbor search algorithms have been proposed over the years; these generally seek to reduce the number of distance evaluations actually performed.

C)SVM and Neural networks

The SVM is utilized to reduce the training samples for each of the available categories to their support vectors (SVs). The SVs from different categories are used as the training data of nearest neighbor classification algorithm in which the Euclidean distance function is used to calculate the average distance between the testing data point to each set of SVs of different categories. The classification decision is made based on the category which has the shortest average distance between its SVs and the testing data point. The experiments on several benchmark text datasets show that the classification accuracy of the SVM-NN approach has low impact on the value of parameter, as compared to the conventional KNN classification model.

An artificial neural network (or neural network for short) is a predictive model moti- vated by the way the brain operates. Think of the brain as a collection of neurons wired together. Each neuron looks at the outputs of the other neurons that feed into it, does a calculation, and then either fires (if the calculation exceeds some thresh- hold) or doesn't (if it doesn't).
Accordingly, artificial neural networks consist of artificial neurons, which perform similar calculations over their inputs. Neural networks can solve a wide variety of problems like handwriting recognition and face detection, and they are used heavily in deep learning, one of the trendiest subfields of data

science. However, most neural networks are "black boxes"—inspecting their details doesn't give you much under- standing of how they're solving a problem. And large neural networks can be difficult to train. For most problems you'll encounter as a budding data scientist, they're probably not the right choice. Someday, when you're trying to build an artificial intelli- gence to bring about the Singularity, they very well might be. Perceptrons Pretty much the simplest neural network is the perceptron, which approximates a single neuron with n binary inputs. It computes a weighted sum of its inputs and "fires" if that weighted sum is zero or greater:

```
def step_function(x):
return 1 if x >= 0 else 0
def perceptron_output(weights, bias, x):
"""returns 1 if the perceptron 'fires', 0 if not"""
213calculation = dot(weights, x) + bias
return step_function(calculation)
```

The perceptron is simply distinguishing between the half spaces separated by the hyperplane of points x for which:

```
dot(weights,x) + bias == 0
```

With properly chosen weights, perceptrons can solve a number of simple problems (Figure 18-1). For example, we can create an AND gate (which returns 1 if both its inputs are 1 but returns 0 if one of its inputs is 0) with:

```
weights = [2, 2]
bias = -3
```

If both inputs are 1, the calculation equals 2 + 2 - 3 = 1, and the output is 1. If only one of the inputs is 1, the calculation equals 2 + 0 - 3 = -1, and the output is 0. And if both of the inputs are 0, the calculation equals -3, and the output is 0.
Similarly, we could build an OR gate with:

```
weights = [2, 2]
bias = -1
```

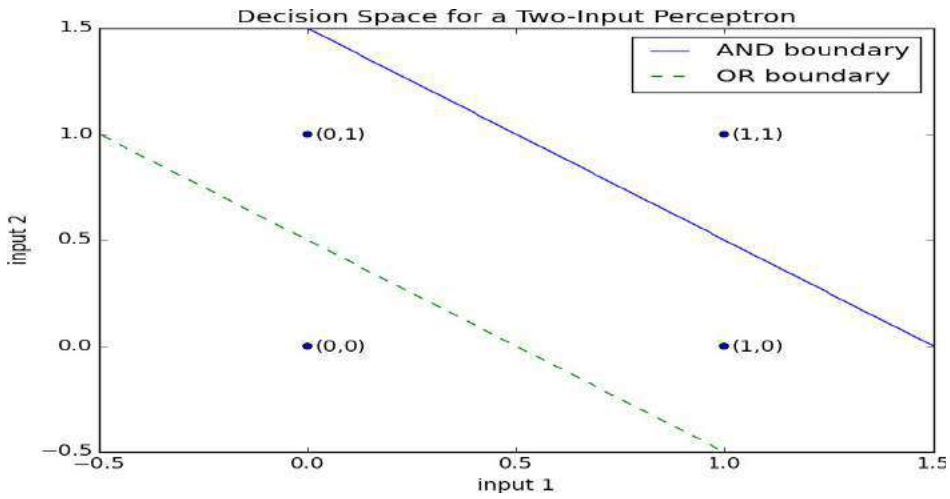


Figure 18-1. Decision space for a two-input perceptr

And we could build a NOT gate (which has one input and converts 1 to 0 and 0 to 1) with:

weights = [-2]
bias = 1
However, there are some problems that simply can't be solved by a single perceptron. For example, no matter how hard you try, you cannot use a perceptron to build an XOR gate that outputs 1 if exactly one of its inputs is 1 and 0 otherwise. This is where we start needing more-complicated neural networks.
Of course, you don't need to approximate a neuron in order to build a logic gate:
and_gate = min
or_gate = max
xor_gate = lambda x, y: 0 if x == y else 1
Like real neurons, artificial neurons start getting more interesting when you start connecting them together.

D)High dimensional data matrix manipulation:

A Data Matrix code is a two-dimensional matrix barcode consisting of black and white "cells" or modules arranged in either a square or rectangular pattern.

A)Variable selection by penalized regression such as LASSO and lars

- **LASSO**

In statistics and machine learning, lasso (least absolute shrinkage and selection operator) (also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's Nonnegative Garrote.Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates need not be unique if covariates are collinear.

- **LARS**

In statistics, least-angle regression (LARS) is an algorithm for fitting linear regression models to high-dimensional data, developed by Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani.
Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients.
Instead of giving a vector result, the LARS solution consists of a curve denoting the solution for each value of the L1 norm of the parameter vector. The algorithm is similar to forward stepwise regression, but instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual.

B)MDS

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. An MDS algorithm aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the N dimensions. The number of dimensions of an MDS plot N can exceed 2 and is specified a priori. Choosing N=2 optimizes the object locations for a two-dimensional scatterplot.

C)Principal components analysis(PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of

linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables.

D)Singular value decomposition(SVD)

In linear algebra, the singular value decomposition (SVD) is a factorization of a real or complex matrix. It is the generalization of the eigendecomposition of a positive semidefinite normal matrix (for example, a symmetric matrix with positive eigenvalues) to any $m \times n$matrix via an extension of polar decomposition. It has many useful applications in signal processing and statistics.

Date: 05/05/2016

DAY 4 - Live Demonstration of Finding structures in data through Clustering

Handouts from Prof Dr S.R. Dhore Expert Talk:

Contents to Cover:

a) Clustering methods

- K-means, Partitioninng around medoids(PAM),
- Visualization of clustering results,
- clustering evaluation using Silhouette coeficients and other indices,
- BIRCH clustering for large datasets, Anomoly detection, Finding frequent Itemsets using A-Priori Algorithm and variants, e.g.SON; Basics of networks centrality measures, Network link analysis, PageRank algorithm.

b) Outlier analysis

- Anomoly detection,
- Finding frequent Itemsets using A-Priori Algorithm and variants,

c) Association analysis

- e.g.SON

d) Network analysis and optimization

- Basics of networks centrality measures,
- Network link analysis
- PageRank algorithm.

What is Clustering?

Clustering, in the context of databases, refers to the ability of several servers or instances to connect to a single database. An instance is the collection of memory and processes that interacts with a database, which is the set of physical files that actually store data. Clustering offers two major advantages, especially in high-volume database environments:

Fault tolerance: Because there is more than one server or instance for users to connect to, clustering offers an alternative, in the event of individual server failure.

Load balancing: The clustering feature is usually set up to allow users to be automatically allocated to the server with the least load.

Date: 05/05/2016

- **Simulated Annealing:**

An additional empirical study has compared the performance of the following clustering algorithms: SA, GA, TS, randomized branch-and-bound (RBA), and hybrid search(HS).The convergence pace of SA is too slow; RBA and TS performed best; and HS is good for high dimensional data.for example, ANNs work better in classifying images represented using extracted features rather than with raw images, and hybrid classifiers work better than ANNs.

Clustering methods

The main reason for having many clustering methods is the fact that the notion of "cluster" is not precisely defined Consequently many clustering methods have been developed, each of which uses a different induction principle. Farley and Raftery(1998) suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods.

1.Hierarchical Methods

These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be subdivided as following:

1.1 Agglomerative hierarchical clustering — Each object initially represents a cluster of its own.

1.2 Divisive hierarchical clustering — All objects initially belong to one cluster.

1.3 Single-link clustering (also called the connectedness, the minimum method or the nearest neighbor method) — methods that consider the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

1.4 Complete-link clustering (also called the diameter, the maximum method or the furthest neighbor method) - methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster

1.5 Average-link clustering (also called minimum variance method) - methods that consider the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Such clustering algorithms may be found in (Ward, 1963) and (Murtagh, 1984).

1.6 The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single link methods are more versatile.

1. Versatility - The single-link methods, for example, maintain good performance on data sets containing non-isotropic clusters, including well-separated, chain-like and concentric clusters.

2. Multiple partitions - Hierarchical methods produce not one partition,but multiple nested partitions, which allow different users to choose different partitions, according to the desired similarity level.

2. Partitioning Methods

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning.

2.1 Error Minimization Algorithms.

These algorithms, which tend to work well with isolated and compact clusters, are the most intuitive and frequently used methods.

K-means Algorithm

The simplest and most commonly used algorithm, employing a squared error criterion is the **K-means algorithm**. This algorithm partitions the data into K clusters (C 1 , C 2 , . . . , C K), represented by their centers or means. The center of each cluster is calculated as the mean of all the instances belonging to that cluster. The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two.Then the cluster centers are re-calculated.The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

where N k is the number of instances belonging to cluster k and μ k is the mean of the cluster k.

A number of convergence conditions are possible. For example, the search may stop when the partitioning error is not reduced by the relocation of the centers.This indicates that the present partition is locally optimal. Other stopping criteria can be used also such as exceeding a pre-defined number of iterations.

Input: S (instance set),

- **K (number of cluster)**

Output: clusters

1: Initialize K cluster centers.
2: while termination condition is not satisfied do
3: Assign instances to the closest cluster center.
4: Update cluster centers based on the assignment.
5: end while

The K-means algorithm may be viewed as a gradient-decent procedure, which begins with an initial set of K cluster-centers and iteratively updates it so as to decrease the error function.The complexity of T iterations of the K-means algorithm performed on a sample size of m instances, each characterized by N attributes, is: $\mathbf{O(T * K * m * N)}$.

This linear complexity is one of the reasons for the popularity of the K-means algorithms. Even if the number of instances is substantially large (which often is the case nowadays), this algorithm is com**1.6 The complete-link** clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single link methods are more versatile. Generally, hierarchical methods are characterized with the following strengths:

Versatility - The single-link methods, for example, maintain good performance on data sets containing non-isotropic clusters, including well-separated, chain-like and concentric clusters.

Multiple partitions - hierarchical methods produce not one partition,but multiple nested partitions, which allow different users to choose different partitions, according to the desired similarity level. The hierarchical partition is presented using the dendrogram.

2. Partitioning Methods

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning.

2.1 Error Minimization Algorithms.

These algorithms, which tend to work well with isolated and compact clusters, are the most intuitive and frequently used methods. The basic idea is to find a clustering structure that minimizes a certain error criterion which measures the "distance" of each instance to its representative value.

The Achilles heel of the K-means algorithm involves the selection of the initial partition.

PAM (partition around medoids — (Kaufmann and Rousseeuw,

Date: 05/05/2016

1987)). This algorithm is very similar to the K-means algorithm. It differs from the latter mainly in its representation of the different clusters. Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster.The K-medoids method is more robust than the K-means algorithm in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the K-means method. Both methods require the user to specify K, the number of clusters.

PAM (partition around medoids) Algorithm

The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into aset S of selected objects. If O is the set of objects that the set U = O – S is the set of unselected objects.

The algorithm has two phases

(i) In the first phase, BUILD, a collection of k objects are selected for an initial set S.

(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

2.2 Graph-Theoretic Clustering

Graph theoretic methods are methods that produce clusters via graphs. The edges of the graph connect the instances represented as nodes.

Single-link clusters are subgraphs of the MST of the data instances.

3 Density-based Methods

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution. The overall distribution of the data is assumed to be a mixture of several distributions.

parameters. These methods are designed for discovering clusters of arbitrary shape which are not necessarily convex, namely: $x_i, x_j \in C_k$. This does notnecessarily imply that: $\alpha \cdot x_i + (1 - \alpha) \cdot x_j \in C_k$.

The idea is to continue growing the given cluster as long as the density (number of objects or data points) in the neighborhood exceeds some threshold.Namely, the neighborhood of a given radius has to contain at least a minimum number of objects.

4 . Model-based Clustering Methods

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class. The most frequently used induction methods are decision trees and neural

networks.

Anomaly Detection

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains.

What are anomalies?

Anomalies are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 illustrates anomalies in a simple 2-dimensional data set. The data has two normal regions, N 1 and N 2 , since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., points o 1 and o 2 , and points in region O 3 , are anomalies.Anomalies might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but all of the reasons have a common characteristic that they are interesting to the analyst. The "interestingness" or real life relevance of anomalies is a key
feature of anomaly detection.

SON:-

A **Self-Organizing Network** (**SON**) is an automation technology designed to make the planning, configuration, management, optimization and healing of mobile radio access networks simpler and faster. SON functionality and behavior has been defined and specified in generally accepted mobile industry recommendations produced by organizations such as 3GPP (3rd Generation Partnership Project) and the NGMN (Next Generation Mobile Networks).

SON architectural types

Self-organizing networks are commonly divided into three major architectural types.

1. Distributed SON:-

In this type of SON (D-SON), functions are distributed among the network elements at the edge of the network, typically the ENodeB elements. This implies a certain degree of localization of functionality and is normally supplied by the network equipment vendor manufacturing the radio cell.

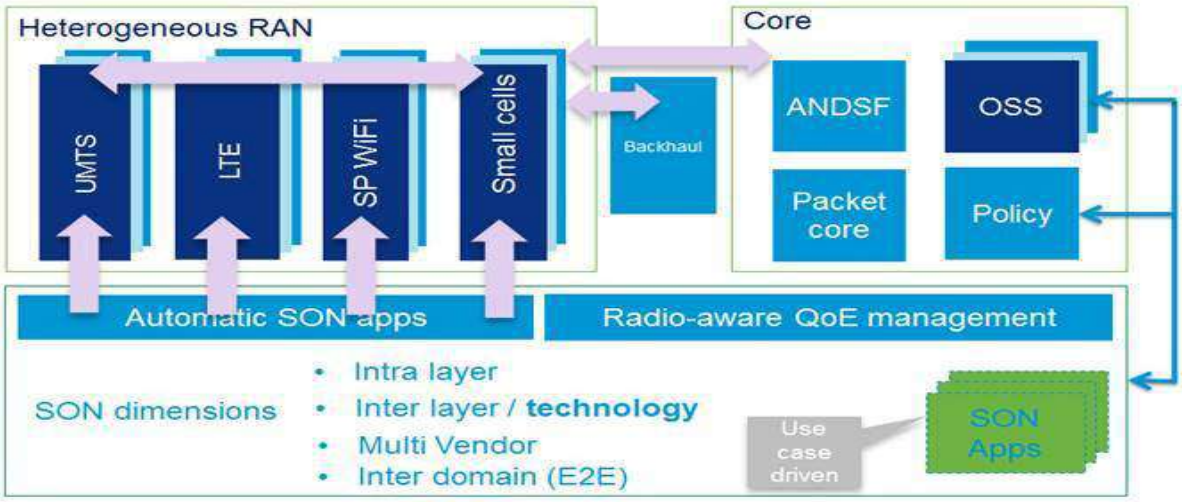


Figure. Architecture of SON

2. Centralized SON:-

In centralized SON (C-SON), function is more typically concentrated closer to higher-order network nodes or the network OSS, to allow a broader overview of more edge elements and coordination of e.g. load across a wide geographic area.

3. Hybrid SON:-

Hybrid SON is a mix of centralized and distributed SON, combining elements of each in a hybrid solutions.

SON sub-functions

Self-organizing network functionalities are commonly divided into three major sub-functional groups, each containing a wide range of decomposed use cases.

1. Self-configuration functions:-

Self-configuration strives towards the "plug-and-play" paradigm in the way that new base stations shall automatically be configured and integrated into the network.

2. Self-optimization functions

Every base station contains hundreds of configuration parameters that control various aspects of the cell site. Each of these can be altered to change network behavior, based on observations of both the base station itself and measurements at the mobile station or handset.

3. Self-healing functions

When some nodes in the network become inoperative, self-healing mechanisms

aim at reducing the impacts from the failure, for example by adjusting parameters and algorithms in adjacent cells so that other nodes can support the users that were supported by the failing node. In legacy networks, the failing base stations are at times hard to identify and a significant amount of time and resources is required to fix it.

Network Link Analysis Concepts

Link analysis is a data-analysis technique used to evaluate relationships (connections) between nodes. Relationships may be identified among various types of nodes (objects), including organizations, people and transactions. Link analysis has been used for investigation of criminal activity (fraud detection, counterterrorism, and intelligence), computer security analysis, search engine optimization, market research, medical research, and art.

Knowledge discovery is an iterative and interactive process used to identify, analyze and visualize patterns in data.[Network analysis, link analysis and social network analysis are all methods of knowledge discovery, each a corresponding subset of the prior method. Most knowledge discovery methods follow these steps (at the highest level):

1. Data processing
2. Transformation
3. Analysis
4. Visualization

Applications

1. FBI Violent Criminal Apprehension Program (ViCAP)
2. Iowa State Sex Crimes Analysis System
3. Minnesota State Sex Crimes Analysis System (MIN/SCAP)

Page Rank Algoithm

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

Description:-

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element

A PageRank results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page.

Date: 06/05/2016

DAY 5 - Industry Showers its Expertise in the Hadoop Ecosystem

Handouts from Dr. M. R. Sanghavi Expert Talk:

Contents to Cover:

a) MapReduce basics
b) Hadoop MapReduce
c) HDFS basics
d) Haddop Ecosystem

MapReduce :-

The MapReduce algorithm contains two important tasks, namely Map and Reduce.The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. This simple scalability is what has attracted many programmers to use the MapReduce model.The Algorithm MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage. Algorithm MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – V

Department of Sciences & Technology
Government of India

Date: 06/05/2016

Handouts from Mr. Pawan Tiwari Expert Talk:

Contents to Cover:

a) MapReduce basics
b) Hadoop MapReduce
c) HDFS basics
d) Haddop Ecosystem

Hadoop Distributed File System(HDFS):-

The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. An important characteristic of Hadoop is the partitioning of data and computation across many (thousands) of hosts, and the execution of application computations in parallel close to their data. An important characteristic of Hadoop is the partitioning of data and computation across many (thousands) of hosts, and the execution of application computations in parallel close to their data.

NameNode: The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes. Inodes record attributes like permissions, modification and access times, namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file), and each block of the file is independently replicated at multiple DataNodes (typically three, but user selectable file-by-file). The NameNode maintains the namespace tree and the mapping of blocks to DataNodes. The current design has a single NameNode for each cluster. The cluster can have thousands of DataNodes and tens of thousands of HDFS clients per cluster, as each DataNode may execute multiple application tasks concurrently. The NameNode is a multithreaded system and processes requests simultaneously from multiple clients.

DataNodes: Each block replica on a DataNode is represented by two files in the local native filesystem. The first file contains the data itself and the second file records the block's metadata including checksums for the data and the generation stamp. The size of the data file equals the actual length of the block and does not require extra space to round it up to the nominal block size as in traditional filesystems. Thus, if a block is half full it needs only half of the space of the full block on the local drive.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – V

Department of Sciences & Technology
Government of India

Date: 06/05/2016

Basics of Sqoop,Hbase,Hive,Pig etc:-

Sqoop: Here, Sqoop occupies a place in the Hadoop ecosystem to provide feasible interaction between relational database server and Hadoop's HDFS. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.

Hbase: HBase is an open-source, distributed, column-oriented database built on top of HDFS based on BigTable. Distributed data store that can scale horizontally to 1,000s of commodity servers and petabytes of indexed storage.Designed to operate on top of the Hadoop distributed file system (HDFS) or Kosmos File System (KFS, aka ClouAdstore) for scalability, fault tolerance, and high availability.

Hive: Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

Pig: Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Apache Pig. To write data analysis programs, Pig provides a high- level language known as Pig Latin. To analyze data using Apache Pig, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

Date: 07/05/2016

DAY 6 - Showcasing of Applications of large data in High-performance

Handouts from Prof T. B. Kute Expert Talk:

Contents to Cover:

a) Integrating R and Hadoop
b) RHIPE and RHadoop
c) Applications on large data
d) High-performance and parallel R

Distributed R :-

Distributed R is an open source, high-performance platform for the R language. It splits tasks between multiple processing nodes to reduce execution time and analyze large data sets. Distributed R enhances R by adding distributed data structures, parallelism primitives to run functions on distributed data, a task scheduler, and multiple data loaders. It is mostly used to implement distributed versions of machine learning tasks. Distributed R is written in C++ and R, and retains the familiar look and feel of R. As of February 2015, Hewlett Packard (HP) provides enterprise support for Distributed R with proprietary additions such as a fast data loader from the Vertica database.Distributed R was begun in 2011 by Indrajit Roy, Shivaram Venkataraman, Alvin AuYoung, and Robert S. Schreiber as a research project at HP Labs. It was open sourced in 2014 under the GPLv2 license and is available at Github.In February 2015, Distributed R reached its first stable version 1.0, along with enterprise support from HP.Distributed R is a platform to implement and execute distributed applications in R. The goal is to extend R for distributed computing, while retaining the simplicity and look-and-feel of R.

Parallel Computing in R:-

When working with R, you will often encounter situations in which you need to repeat a computation, or a series of computations, many times. This can be accomplished through the use of a for loop. However, if there are a large number of computations that need to be carried out (i.e. many thousands), or if those individual computations are time-consuming (i.e. many minutes), a for loop can be very slow That said, almost all computers now have multicore processors, and as long as these computations do not need to communicate they can be spread across multiple cores and executed in parallel, reducing computation time.

A parallel computer is a "Collection of processing elements that communicate and co-operate to solve large problems fast".

Hadoop :-

HDFS has a master/slave architecture. An HDFS cluster consists of a single Name Node, a master server that manages the file system name space and regulates access to files by clients. In addition, there are a number of Data Nodes, usually one per node in the cluster, which manage

storage attached to the nodes that they run on. HDFS exposes a file system name space and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of Data Nodes. The Name Node executes file system name space operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data nodes. The Data Nodes are responsible for serving read and write requests from the file system's clients. The Data Nodes also perform block creation, deletion, and replication upon instruction from the Name Node.

The Name Node and Data Node are pieces of software designed to run on commodity machines. These machines typically run a GNU/Linux operating system (OS). HDFS is built using the Java language; any machine that supports Java can run the Name Node or the Data Node software. Usage of the highly portable Java language means that HDFS can be deployed on a wide range of machines. A typical deployment has a dedicated machine that runs only the Name Node software. Each of the other machines in the cluster runs one instance of the Data Node software. The architecture does not preclude running multiple Data Nodes on the same machine but in a real deployment that is rarely the case. The existence of a single Name Node in a cluster greatly simplifies the architecture of the system. The Name Node is the arbitrator and repository for all HDFS meta data. The system is designed in such a way that user data never flows through the Name Node.

Integrating R and Hadoop:-

Many modern enterprises are collecting data at the most detailed level possible, creating data repositories ranging from terabytes to petabytes in size. The ability to apply sophisticated statistical analysis methods to this data is becoming essential for marketplace competitiveness. This need to perform deep analysis over huge data repositories poses a significant challenge to existing statistical software and data management systems. On the one hand, statistical software provides rich functionality for data analysis and modeling,but can handle only limited amounts of data; e.g., popular packages like R and SPSS operate entirely in main memory. On the other hand, data management systems such as MapReduce-based systems can scale to petabytes of data, but provide insufficient analytical functionality. We report our experiences in building Ricardo, a scalable platform for deep analytics. Ricardo is part of the eXtreme Analytics Platform (XAP) project at the IBM Almaden Research Center, and rests on a decomposition of data-analysis algorithms into parts executed by the R statistical analysis system and parts handled by the Hadoop data management system. This decomposition attempts to minimize the transfer of data across system boundaries.

Date: 07/05/2016

RHIPE architecture:-

RHIPE is a Java package that integrates the R environment with Hadoop, the open source implementation of Google's Map Reduce. Using Rhipe, it is possible to write Map Reduce algorithms in R. To run the Map Reduce jobs on Hadoop, two functions namely Map and Reduce functions are used.In "R" the developer writes R expressions to achieve map and reduce functionality. Rhipe encapsulates this map and reduce expressions and other Hadoop related parameters (like combiner, num of reduce and map tasks etc) provided by the user and submits the job to Hadoop. The encapsulation is done by Rhipe function "rhmr", and the job is submitted by "rhex" function. For more information on these functions refer the supporting document for R.RHIPE stands for R and Hadoop Integrated Programming Environment. It means "in a moment" in Greek and is a merger of R and Hadoop.The RHIPE package uses the Divide and Recombine technique to perform data analytics over Big Data.RHIPE has mainly been designed to accomplish two goals.Allowing you to perform in-depth analysis of large as well as small data.Allowing users to perform the analytics operations within R using a lower-level language. RHIPE is a lower-level interface as compared to HDFS and MapReduce operation.

Hadoop streaming R package :-

hsCmdLineArgs

Offers several command line arguments useful for Hadoop streaming. Allows specifying input and output files, column separators, and much more. Optionally opens the I/O connections

hsCmdLineArgs(spec=c(),openConnections=TRUE,args=commandArgs(TRUE))

The spec vector has length 6*n, where n is the number of command line arguments specified. The spec has the same format as the spec parameter in the getopt function of the getopt package, though we have one additional entry specifying a defaut value. The six entries per argument are the following:

long flag name (a multi-character string)

short flag name (a single character)

Argument specification: 0=no arg, 1=required arg, 2=optional arg

Data type ('logical', 'integer', 'double', 'complex', or 'character')

A string describing the option

The default value to be assigned to this parameter

hsKeyValReader

Uses scan to read in chunkSize lines at a time, where each line consists of a key string and a value string. The first skip lines of input are skipped. Each group of key/value pairs are passed to FUN as a character vector of keys and character vector of values

hsKeyValReader(file = "", chunkSize = -1, skip = 0, sep = "\t",FUN = function(k, v) cat(paste(k, v, sep = ": "), se

file
A connection object or a character string, as in scan.

chunkSize
The (maximal) number of lines to read at a time. The default is -1, which speci-
fies that the whole file should be read at once.

skip
Number of lines to ignore at the beginning of the file

FUN
A function that takes a character vector as input

sep
The character separating the key and the value strings

hsLineReader

This function repeatedly reads chunkSize lines of data from file and passes a character vector of these strings to FUN. The first skip lines of input are ignored.

hsLineReader(file = "", chunkSize = -1, skip = 0, FUN = function(x) cat(x, sep = "\n")

file
A connection object or a character string, as in readLines.

chunkSize
The (maximal) number of lines to read at a time. The default is -1, which speci-
fies that the whole file should be read at once.

skip
Number of lines to ignore at the beginning of the file

Date: 07/05/2016

FUN
A function that takes a character vector as input

hsTableReader

This function repeatedly reads chunks of data from an input connection, packages the data as a data frame, optionally ensures that all the rows for certain keys are contained in the data frame, and passes the data frame to a handler for processing. This continues until the end of file.

file
Any file specification accepted by scan

cols
A list of column names, as accepted by the 'what' arg to scan

chunkSize
Number of lines to read at a time

FUN
A function accepting a data frame with columns given by cols

ignoreKey
If TRUE, always passes chunkSize rows to FUN, regardless of whether the chunk has only some of the rows for a given key. If TRUE, the singleKey arg is ignored.

SingleKey

If TRUE, then each data frame passed to FUN will contain all rows corresponding to a single key. If FALSE, then will contain several complete keys.

skip
Number of lines to skip at the beginning of the file.

sep
Any separator character accepted by scan

keyCol
The column name of the column with the keys.

PFUN
Same as FUN, except handles incomplete keys. See below.

carryMemLimit
Max memory used for values of a single key

carryMaxRows
Max number of values allowed for a key.

stringsAsFactors
Whether strings converted to factors.

debug
Whether to print debug messages.

hsWriteTable

Calls write.table without row names or column names, without string quotes, and with tab as the default separator.

hsWriteTable(d, file = '"', sep = "\t")

d
A data frame

file
A connection, as taken by write.table()

sep
The column separator, defaults to a tab character

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – VII

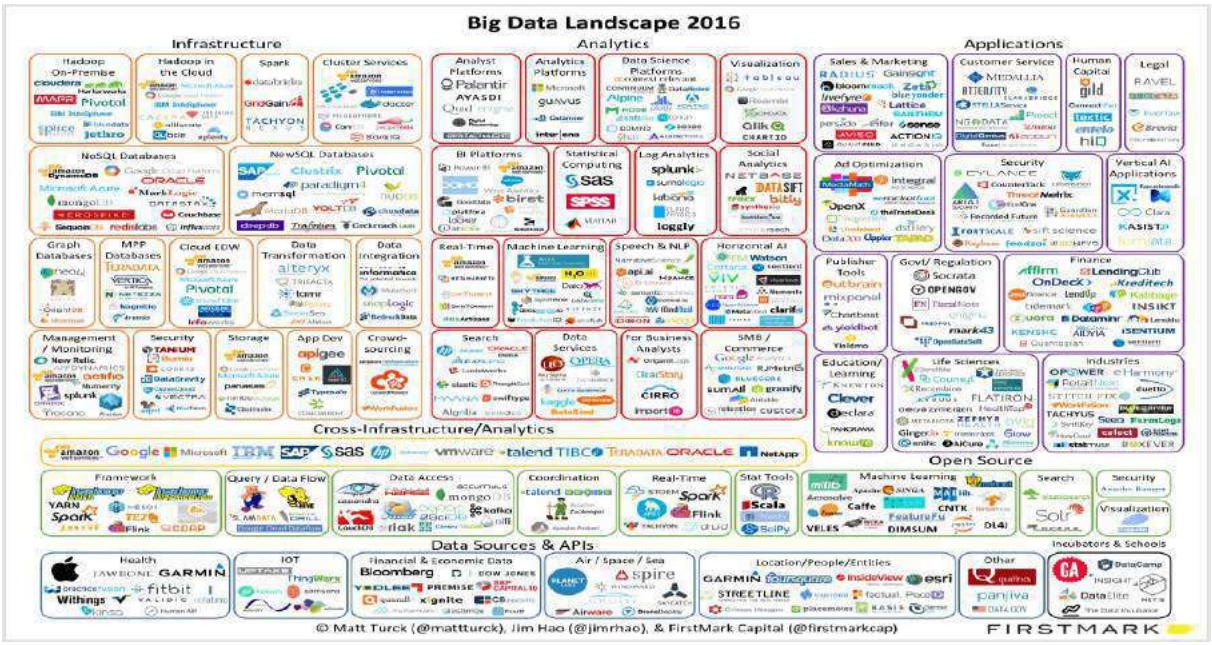
Department of Sciences & Technology
Government of India

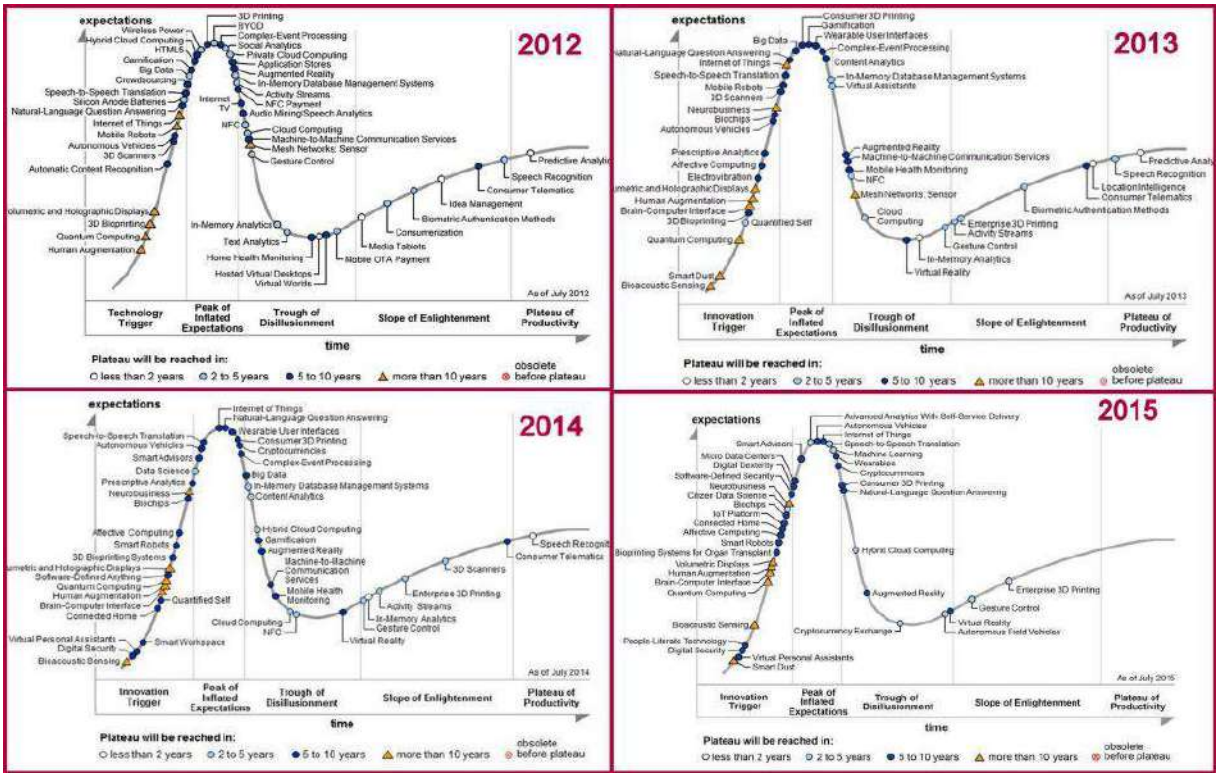
Date: 08/05/2016

DAY 7 - Trainees experienced the Analysis of data in motion

Handouts from Mr Prasad Chandane Expert Lecture:

(*Session Made through Pecha-Kucha Presentation Style*)





Date: 08/05/2016

Handouts from Prof G. K. Bhamare Hands-on Session:

Contents to Cover:

a) Real time data streams
b) Stream data basics

Real Time Data Streams :-

Complex Event Processing: Complex event processing, also known as event, stream or event stream processing is a technique used for querying data prior to its being stored within a database or, in some cases, without it ever being so stored. If event processing involves tracking and analysing streams of data about things that happen (events) and deriving appropriate conclusions, complex event processing, or CEP, processes 'complex events', derived from multiple source events (which could be from the same source or not), combined to generate further downstream events or patterns. Complex event processing, or CEP, is event processing that combines data from multiple sources to infer events or patterns that suggest more complicated circumstances.

Stram Data Basics:-

Online Algorithm: In Computer Science, an online algorithm is one that can process its input piece-by-piece in a serial fashion, i.e, in the order that the input is fed to the algorithm, without having the entire input available from the start. n contrast, an offline algorithm is given the whole problem data from the beginning and is required to output an answer which solves the problem at hand. As an example, consider the sorting algorithm selection and insertion sort. Selection sort repeatedly selects the minimum element from the unsorted remainder and places it at the front, which requires access to the entire input; it is thus an offline algorithm. On the other hand, insertion sort considers one input element per iteration and produces a partial solution without considering future elements. Thus insertion sort is an online algorithm. The competitive ratio of an online problem is the best competitive ratio achieved by an online algorithm. Intuitively, the competitive ratio of an algorithm gives a measure on the quality of solutions produced by this algorithm, while the competitive ratio of a problem shows the importance of knowing the future for this problem.

Some online Algorithms:

Insertion Sort: Insertion sort is a simple sorting algorithm that builds the final sorted array (or list) one item at a time. It is much less efficient on large lists than more advanced algorithms such as quicksort, heapsort or merge sort.

Greedy algorithm: A greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum. In many problems, a greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time.

Metrical task systems: ask systems are mathematical objects used to model the set of possible configuration of online algorithm. A task system determines a set of states and costs to change states. Task systems obtain as input a sequence of

requests such that each request assigns processing times to the states. The objective of an online algorithm for task systems is to create a schedule that minimizes the overall cost incurred due to processing the tasks with respect to the states and due to the cost to change states.

Odds algorithm: The odds-algorithm is a mathematical method for computing optimal strategies for a class of problems that belong to the domain of optimal stopping problems. Their solution follows from the odds-strategy, and the importance of the odds-strategy lies in its optimality, as explained below. This was used to devise betting strategies called martingales.

Page replacement algorithm: A computer operating system that uses paging for virtual memory management, page replacement algorithms decide which memory pages to page out (swap out, write to disk) when a page of memory needs to be allocated. Paging happens when a page fault occurs and a free page cannot be used to satisfy the allocation, either because there are none, or because the number of free pages is lower than some threshold.

Algorithm for calculating variance: Algorithms for calculating variance play a major role in computational statistics. A key difficulty in the design of good algorithm for this problem is that formulas for the variance may involve sums of squares, which can lead to numerical instability as well as to arithmetic oveflow when dealing with large values.

Ukkonen's algorithm: Ukkonen's algorithm is a linear-time, online algorithm for constructing suffix trees. The algorithm begins with an implicit suffix tree containing the first character of the string. Then it steps through the string adding successive characters until the tree is complete. This order addition of characters gives Ukkonen's algorithm its "on-line" property.

•

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – VII

Department of Sciences & Technology Government of India

Date: 08/05/2016

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – VII

Department of Sciences & Technology Government of India

Date: 08/05/2016

Handouts from Mr.Prasad Chandane Expert Talk:

Contents to Cover:

c) Data stream analysis platforms

d) R based stream analysis

Data stream analysis platforms :-

SAMPLING STREAM: We obtain a smaller data set with the same structure. Estimating on a sample is often straightforward. Run the analysis on the sample that you would on the full data. Some rescaling/reweighting may be necessary. Sampling is general and agnostic to the analysis to be done Other summary methods only work for certain computations. Though sampling can be tuned to optimize some criteria Sampling is (usually) easy to understand. So prevalent that we have an intuition about sampling.

CONCEPT DRIFTS: In predictive analytics and the **concept drift** means that the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes problems because the predictions become less accurate as time passes. The term *concept* refers to the quantity to be predicted. More generally, it can also refer to other phenomena of interest besides the target concept, such as an input, but, in the context of concept drift, the term commonly refers to the target variable.

MASSIVE ONLINE ANALYSIS: MOA is an open-source framework software that allows to build and run experiments of machine learning or data mining on evolving data streams. It includes a set of learners and stream generators that can be used from the Graphical User Interface (GUI), the command-line, and the Java API. MOA contains several collections of machine learning algorithms. MOA performs BIG DATA stream mining in real time, and large scale machine learning. MOA can be extended with new mining algorithms, and new stream generators or evaluation measures. The goal is to provide a benchmark suite for the stream mining community. Massive Online Analysis (MOA) is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. MOA is designed to deal with the challenging problems of scaling up the implementation of state of the art algorithms to real world dataset sizes and of making algorithms comparable in benchmark streaming settings.

STORM: Storm is a distributed real-time computation system for processing large volumes of high-velocity data. Storm is extremely fast, with the ability to process over a million records per second per node on a cluster of modest size. Enterprises harness this speed and combine it with other data access applications in Hadoop to prevent undesirable events or to optimize positive outcomes.

SPARK: Industries are using Hadoop extensively to analyze their data sets. The reason is that Hadoop framework is based on a simple programming model (MapReduce) and it enables a computing solution that is scalable, flexible, fault-tolerant and cost effective. Here, the main concern is to maintain speed in

processing large datasets in terms of waiting time between queries and waiting time to run the program. Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.As against a common belief, Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management. Hadoop is just one of the ways to implement Spark.

R Based Stram Analysis:-

STREAM: Typical statistical and data mining methods (e.g., clustering, regression, classification and frequent pattern mining) work with "static" data sets, meaning that the complete data set is available as a whole to perform all necessary computations. Well known methods like k- means clustering, linear regression, decision tree induction and the APRIORI algorithm to find frequent itemsets scan the complete data set repeatedly to produce their results However, in recent years more and more applications need to work with data which are not static, but are the result of a continuous data generating process which is likely to evolve over time. Some examples are web click-stream data, computer network monitoring data, telecommunication connection data, readings from sensor nets and stock quotes.

- **Stream Framework:** The stream framework provides an R-based alternative to MOA which seamlessly integrates with the extensive existing R infrastructure. Since R can interface code written in many different programming languages (e.g., C/C++, Java, Python), data stream mining algorithms in any of these languages can be easily integrated into stream.

RSTORM: While streaming processing provides opportunities to deal with extremely large and ever growing data sets in (near) real time, the development of streaming algorithms for complex models is often cumbersome: the software packages that facilitate streaming processing in production environments do not provide statisticians with the simulation, estimation, and plotting tools they are used to. Developers of streaming algorithms would thus benefit from the flexibility of [R] to create, plot and compute data while developing streaming algorithms. Package RStorm implements a streaming architecture modeled on Storm for easy development and testing of streaming algorithms in [R]. RStorm is not intended as a production package, but rather a development tool for streaming algorithms. RStorm is introduced using the canonical streaming example used often for the introduction of Storm: a streaming word count. For RStorm the basic terminology and concepts from Storm 3 are adapted, which are briefly explained before discussing the implementation of a streaming word count in RStorm. The aim of the streaming word count algorithm is to, given a stream of sentences such as posts to a web service like Twitter – count the frequency of occurrence of each word.

Date: 09/05/2016

DAY 8 - Cloud in the context of Big Data - eNightCloud

Handouts from Prof.Amol Kalugade Expert Talk:

Contents to Cover:

a) Relational and Non-relational databases
b) R interface to databases
c) Managing data security and variety
d) Cloud in the context of Big Data

Relational databases usually work with structured data, while non relational databases usually work with semistructured data (i.e. However, relational databases haven't necessarily adapted well to changes in the types and quantities of data now being generated, such as the unstructured data that is prevalent in big data applications.Most NoSQL databases support data replication, storing multiple copies of data across the cluster or even across data centers, to ensure high availability and disaster recovery Relational model databases can be tweaked and set up to run large-scale read-only operations through data warehousing, and thus potentially serve a large amount of users who are querying a large amount of data,especially when using relational MPP architectures like Analytics Platform System, Teradata, Oracle Exadata, or IBM Netezza, which all support scaling.If you're working with data in DB2, you can use the IBM Data Studio tool or the web console within dashDB for Cloud to examine the database schema or define new views to simplify data access from your R scripts.

This package enables the R user to perform common data manipulation operations, as found in popular packages such as plyr and reshape2, on very large data sets stored on Hadoop.Where to find these data are out of the scope of this tutorial, so for now it's enough to mention this blog post, which explains well how to find data on the internet, and Data Camp's interactive tutorial, which deals with how to import and manipulate Quandl data sets.Most organizations with traditional data platforms typically relational database management systems (RDBMS) coupled to enterprise data warehouses (EDW) using ETL tools find that their legacy infrastructure is either technically incapable or financially impractical for storing and analyzing big data.A traditional ETL process extracts data from multiple sources, then cleanses, formats, and loads it into a data warehouse for analysis.When the source data sets are large, fast, and unstructured, traditional ETL can become the bottleneck, because it is too complex to develop, too expensive to operate, and takes too long to execute.By most accounts, 80 percent of the development effort in a big data project goes into data integration and only 20 percent goes toward data analysis.First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computingplatform, cloud architecture, cloud database and data storage scheme.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day –VIII

Department of Sciences & Technology Government of India

Date: 09/05/2016

Sandip Foundation's
Sandip Institute of Technology & Research Center, Nashik
DEPARTMENT OF COMPUTER ENGINEERING
Big Data Analytics Training Programme
Minutes of Day – IX

Department of Sciences & Technology
Government of India

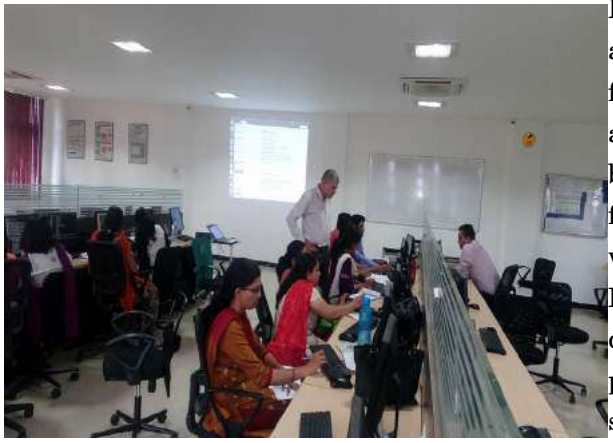
Date: 10/05/2016

DAY 9 - The Training Programme comes to an end with heavy expectations

Handouts from Prof T.B.Kute Expert Talk:

Contents to Cover:

a) Integrating R and Hadoop
b) RHIPE and RHadoop
c) Applications on large data
d) High-performance and parallel R



Data transfer (matrices and data frames) between R and Excel in both directions RHadoop is a collection of five R packages that allow users to manage and analyze data with Hadoop. R programmers can browse, read, write, and modify files stored in HDFS from within R. R programmers can browse, read, write, and modify tables stored in HBASE from within R.This package enables the R user to perform common data manipulation operations, as found in popular packages such as plyr and reshape2, on very large data sets stored on Hadoop. A package that allows R developer to perform statistical analysis in R via Hadoop Map Reduce functionality on a Hadoop cluster. By default the rhbase uses "native" R serialization (serialize/unserialize) to read and write data from hbase Importing data into R and writing back results To import data into R, you first need to have data.This data can be saved in a file onto your computer Where to find these data are out of the scope of this tutorial, so for now it's enough to mention this blog post, which explains well how to find data on the internet, and Data Camp's interactive tutorial, which deals with how to import and manipulate Quandl datasets.

Tip: before you move on and discover how to load your data into R, it might be useful to go over the following checklist that will make it easier to import the data correctly into R: Since this function has no argument, it is assumed that you mean the data sets and functions that you as a user have defined.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – IX

Department of Sciences & Technology Government of India

Date: 10/05/2016

And you might consider changing the path that you get as a result of this function, maybe to the folder in which you have stored your data set: Getting Data from Common Sources into R You will see that the

following basic R functions focus on getting spreadsheets into R, rather than Excel or other type of files. Function cat underlies the functions for exporting data. The most common task is to write a matrix or data frame to file as a rectangular grid of numbers, possibly with row and column labels. Table is more convenient, and writes out a data frame (or an object that can be coerced to a data frame) with row and column labels. There are a number of issues that need to be considered in writing out a data frame to a text file. Text files do not contain metadata on their encodings, so for non ASCII data the file needs to be targetted to the application intended to read it. Function write.foreign in package foreign uses write.table to produce a text file and also writes a code file that will read this text file into another statistical package.Managing data security and variety Managing big data and navigating today's threat environment is challenging.This evolving threat landscape, the number of sophisticated tools and computing power that cybercriminals now have at their disposal, and the proliferation of big data mean software security companies are wrestling with challenges on an unprecedented scale.Successful protection relies on the right combination of methodologies, human insight, an expert understanding of the threat landscape, and the efficient processing of big data to create actionable intelligence. components are not thoroughly examined here, this white paper summarizes how big data is analyzed In January 2008, the industry saw more malware in one month than had been seen in Big Data Integration the previous 15 years combined.The need to manage, maintain and process this huge volume and variety of data on a regular basis presents security vendors with an unprecedented velocity challenge.

Date: 11/05/2016

DAY 10 – Heavy Discussions and Interactions at the last days of the Training Programme

Handouts from Prof T.B.Kute Expert Talk:

Contents to Cover:

a) Integrating R and Hadoop
b) RHIPE and RHadoop
c) Applications on large data
d) High-performance and parallel R

Big data integration Fig: Big data integration ITgroups may find their skill sets, workload, and budgets over stretched by the need to manage terabytes or petabytes of data in a way that delivers genuine value to business users. Big Data projects have fascinated business executives with the promise of higher business returns and greater customer understanding. However, far less talked about is the large number of big data projects that have hit a plateau within businesses that have not been able to deliver the promised pot of gold. The success of any big data project fundamentally depends on an enterprise's ability to capture, store and govern its data. The better an enterprise can provide fast, trustworthy and secure data to business decision maker's the higher the chances of success in exploiting big data, obtaining planned return on investments and justifying further investments. In this paper, we focus on big data integration and take a look at the top five most common mistakes enterprises make when approaching big data integration initiatives and how to avoid them.

The challenge of extracting value from big data is similar in many ways to the age old problem of distilling business intelligence from transactional data. At the heart of this challenge is the process used to extract data from multiple sources, transform it to fit your analytical needs, and load it into a data warehouse for subsequent analysis, a process known as "Extract, Transform & Load" (ETL). The nature of big data requires that the infrastructure for this process can scale cost

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – X

Department of Sciences & Technology Government of India

Date: 11/05/2016

effectively. Apache Hadoop* has emerged as the de facto standard for managing big data. The ETL Bottleneck in Big Data Analytics Big Data refers to the large amounts, at least terabytes, of poly structured data that flows continuously through and around organizations, including video, text, sensor logs, and trans actional records. By analyzing all the data available, decision makers can better assess competitive threats, anticipate changes in customer behavior, strengthen supply chains, improve the effectiveness of marketing campaigns, and enhance business continuity However, most organizations have yet to take full advantage of new technologies for handling big data. Put simply, the cost of the technologies needed to store and analyze large volumes of diverse data has dropped, thanks to open source software running on industry standard hardware.The cost has dropped so much, in fact, that the key strategic question is no longer what data is relevant, but rather how to extract the most value from all the available data. Rapidly ingesting, storing, and processing big data requires a cost effective infrastructure that can scale with the amount of data and the scope of analysis. Most organizations with traditional data platforms typically relational database management systems (RDBMS) coupled to enterprise data warehouses (EDW) using ETL tools find that their legacy infrastructure is either technically incapable or financially impractical for storing and analyzing big data.

Sandip Foundation's

Sandip Institute of Technology & Research Center, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

Big Data Analytics Training Programme

Minutes of Day – X

Department of Sciences & Technology Government of India

Date: 11/05/2016

A traditional ETL process extracts data from multiple sources, then cleanses, formats, and loads it into a data warehouse for analysis. When the source data sets are large, fast, and unstructured, traditional ETL can become the bottleneck, because it is too complex to develop, too expensive to operate, and takes too long to execute.By most accounts, 80 percent of the development effort in a big data project goes into data integration and only 20 percent goes toward data analysis. ETL emerged as an alternative approach in which data is extracted from the sources, loaded into the target database,

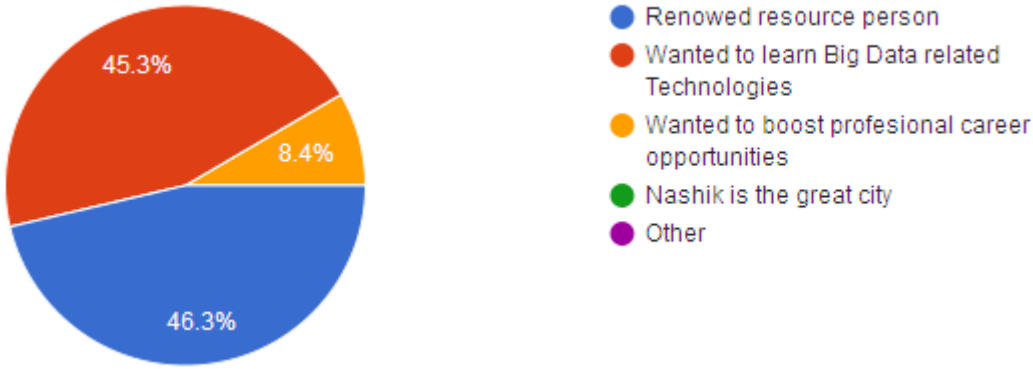
and then transformed and integrated into the desired format. C) Cloud in the context of Big Data With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Effective management and analysis of large scale data poses an interesting but critical challenge. This paper introduces several big data processing technics from system and applicationaspects.First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computing platform, cloud architecture, cloud database

and data storage scheme. Finally, we discuss the open issues and challenges, and deeply explore the research directions in the future on big data processing in cloud computing environments. Security is one of the major issues which reduces the growth of cloud computing and complications with data privacy and data protection continue to plague the market. Cloud service users need to be vigilant in understanding the risks of data breaches in this new environment.

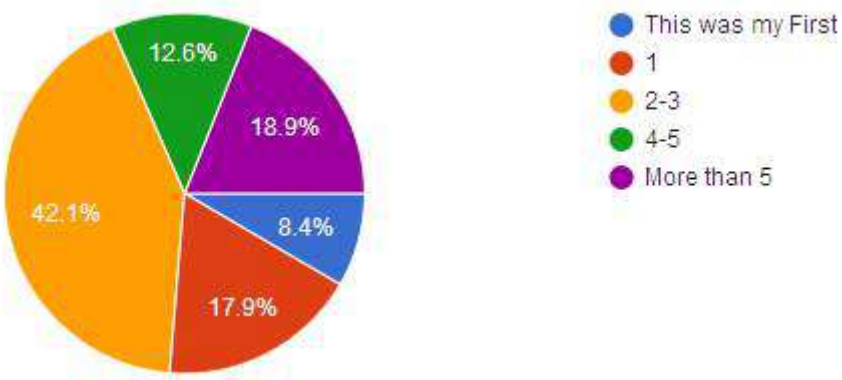
Feedback form Analysis

Note:
Questions where the participants were unanimously in agreement of a good organizational aspect are suppressed.
Sex Ratio: 50:25 (from 75)
Age Average: 30, **Median:** 34 (from 75)

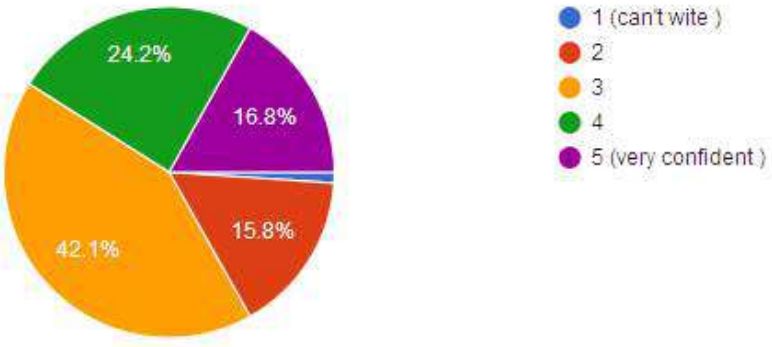
1. What was the prime reason for you attend this workshop?



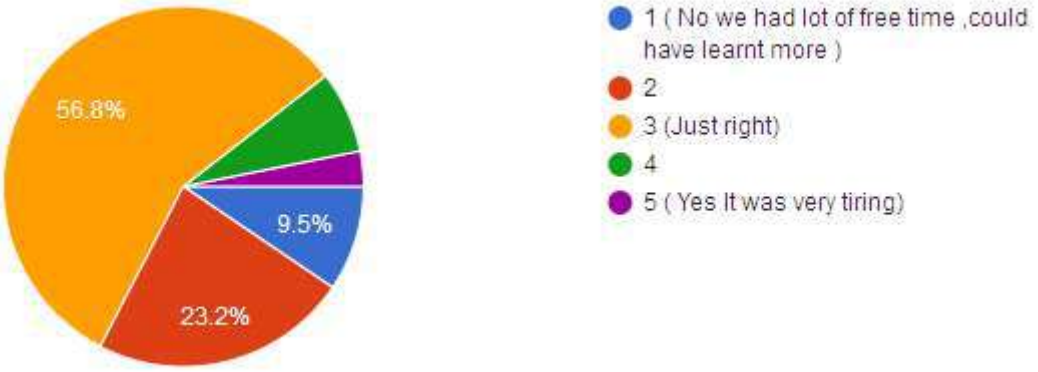
2. How many other Workshops have you attended in the last 5 year?



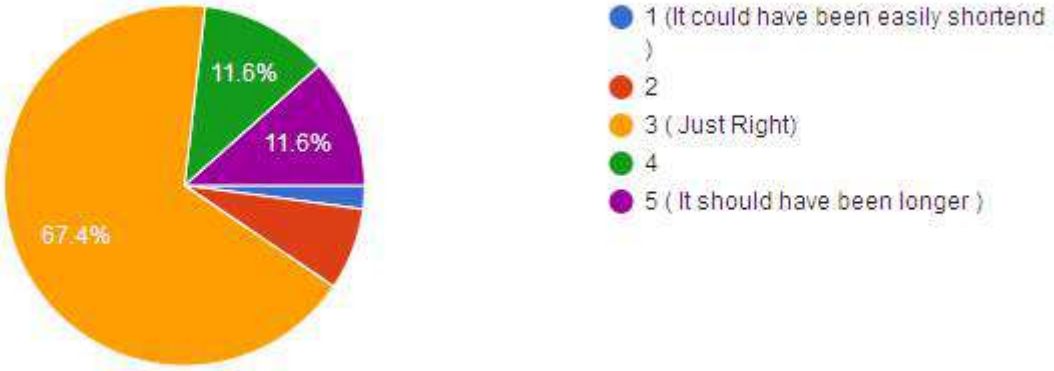
3. To what extent do you feel you have this workshop?



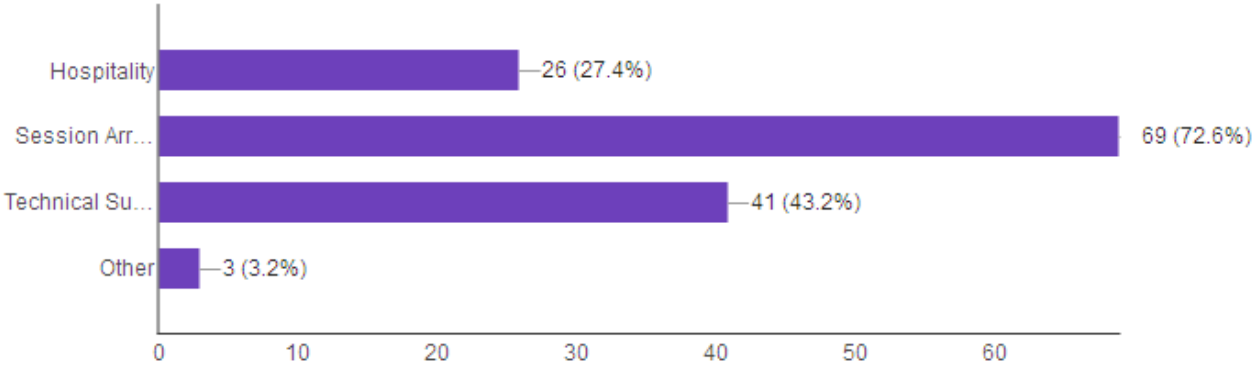
4. Do you think the schedule was too heavy for the specified contents?



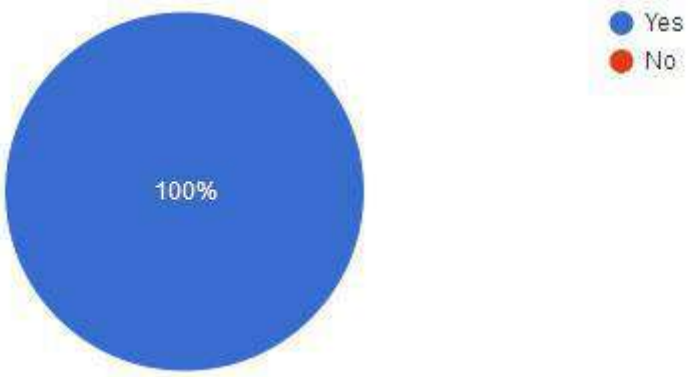
5. Do you think 10 days duration was enough for workshop?



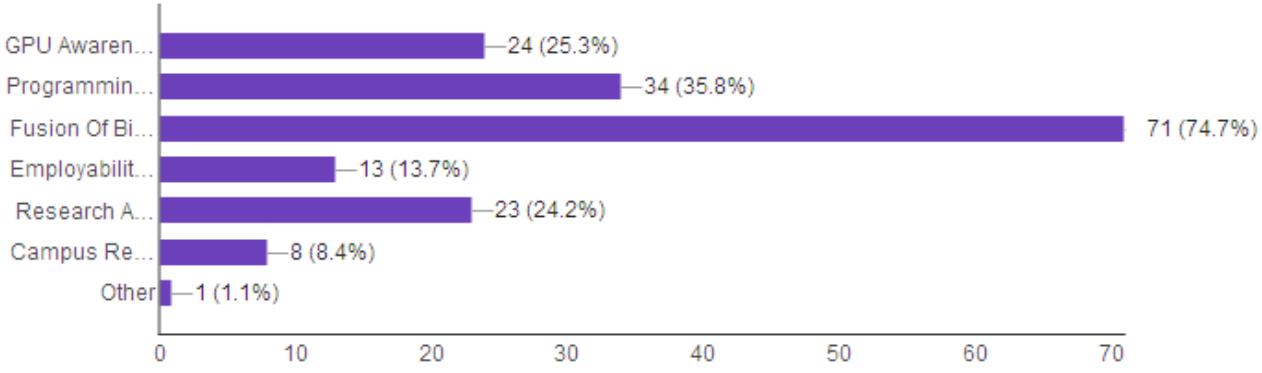
6. Would you have loved more workshop ?
   1. *Hospitality*
   2. *Session Arrangment*
   3. *Technical Support during hands on sesion*
   4. *Others*



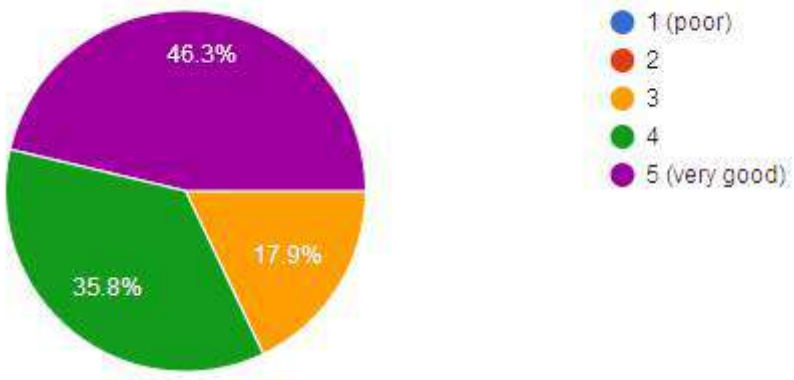
7. Did you attend the more workshop in near future?



8. If Offered ,Which of the following workshop would like to attend?
   1. *GPU Awareness in High Performance Computing*
   2. *Programming and Computing in the current Scenario*
   3. *Fusion of Big Data Analytics and Cloud Computing*
   4. *Employability Skill Development*
   5. *Research Awareness in wireless technologies- RFID,NFC etc.*
   6. *Campus Recruitment Training Programme*



9. How do you rate the hospitality during the event ?



Sandip Foundation's

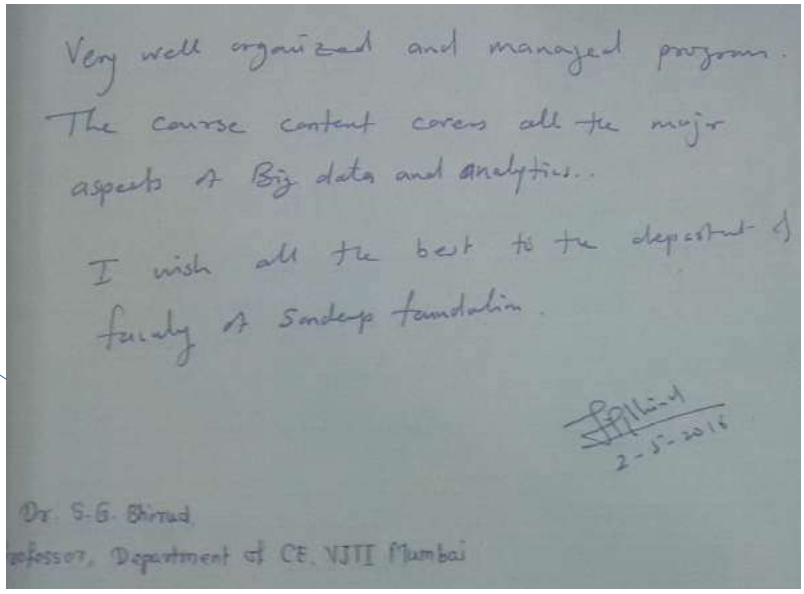
Sandip Institute of Technology & Research Center, NashiK

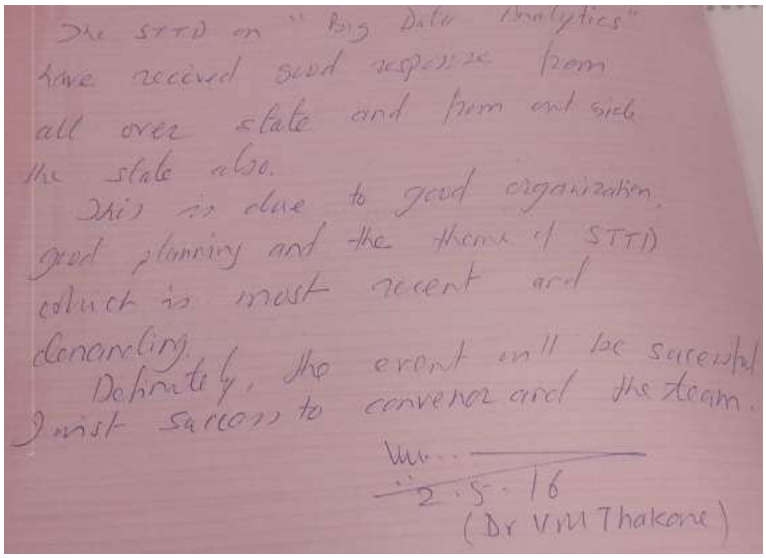
DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

FeedBack from Experts

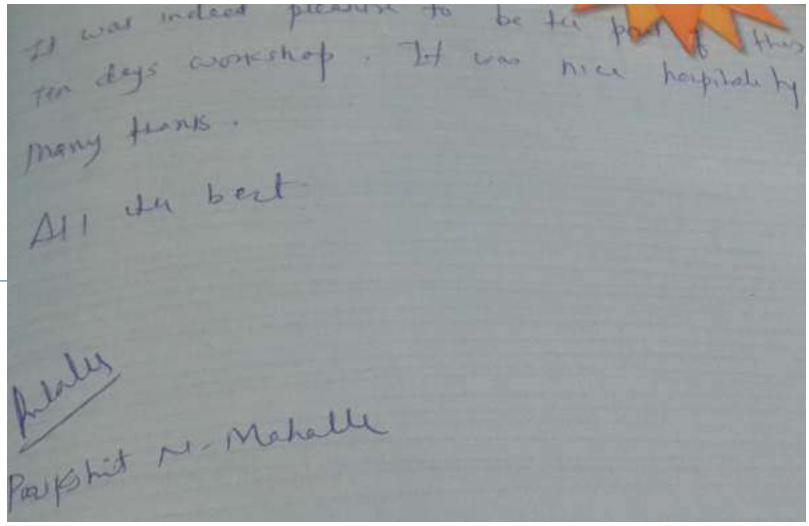
Dr. S. G. Bhirud
Professor, Computer Engineering Department, VJTI, Mumbai
Former Adviser-I, e-GOVERNANCE CELL and LEGAL CELL, AICTE, New Delhi
Former Advisor-I and Chief Vigilance Officer VIGILANCE CELL, AICTE, New Delhi

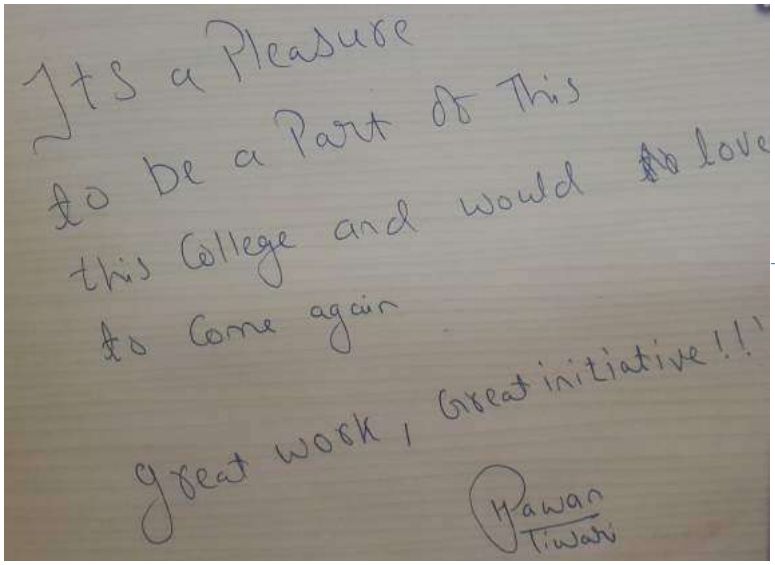




Dr. V. M. Thakare
Professor and Head in Computer Science, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, SGB Amravati University, Amravati

Dr. P. N. Mahalle
Professor & Head (Computer Engg.)
Smt. Kashibai Navale College of Engineering, Pune





Mr. Pavan Tiwari
Co-founder
Deepdive Infotech, Badlapur, mumbai

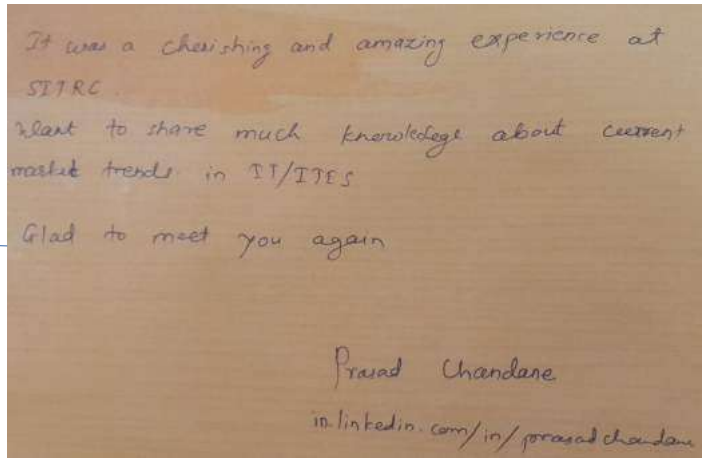
Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

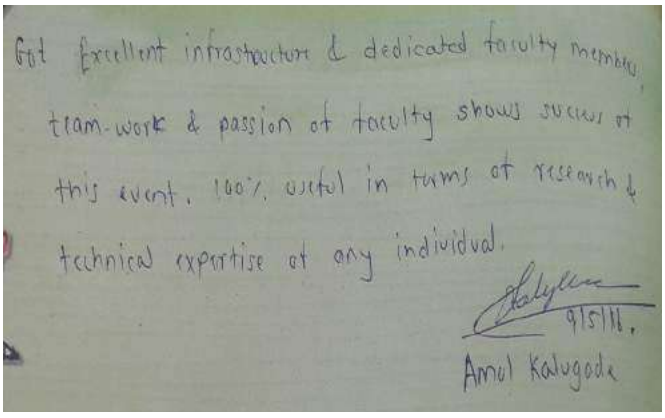
DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

Mr. Prasad Chandane
ERP, Database, Big Data Evangelist, IBM, Pune



Prof. A. R. Kalugade
Assistant Professor, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune

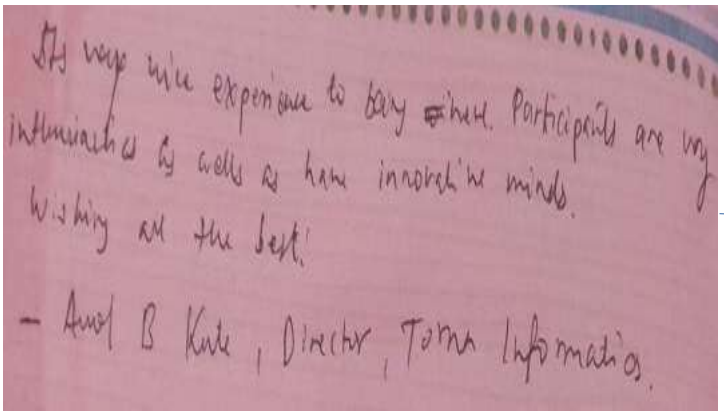


Sandip Foundation's

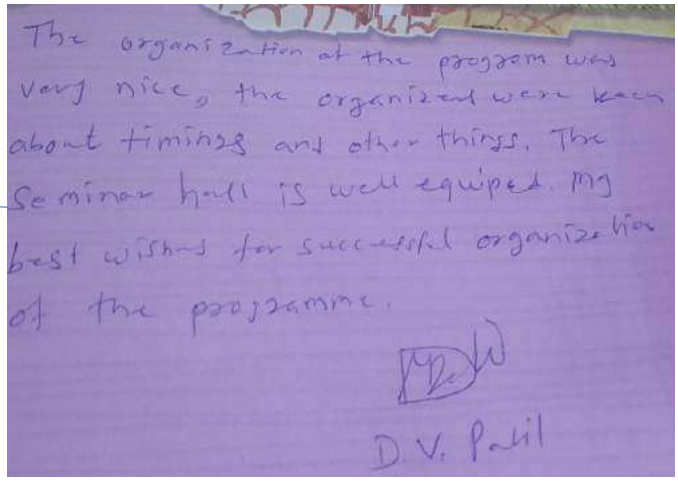
Sandip Institute of Technology & Research Center, NashiK

DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India



Prof. T. B. Kute
Assistant Professor, Sandip Institute of Technology and Research Centre, Nashik



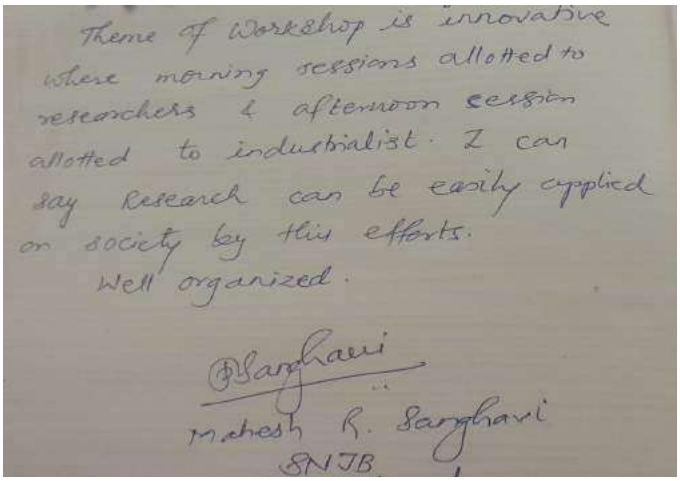
Dr. D.V.Patil
Professor, GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik

Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

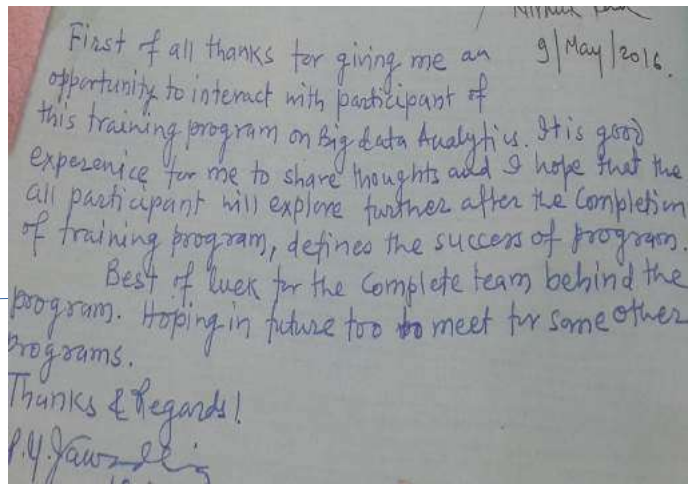
DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

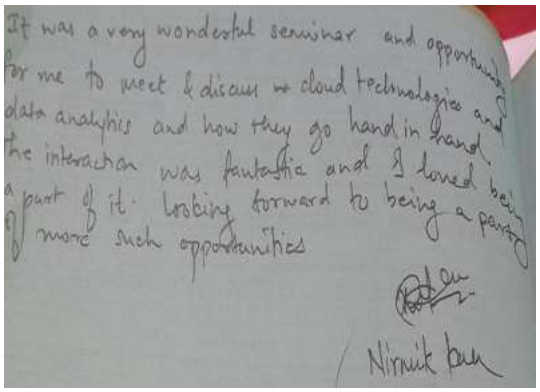

Theme of Workshop is innovative where morning sessions allotted to researchers & afternoon session allotted to industrialist. I can say Research can be easily applied on society by this efforts.
Well organized.

@Sanghavi
mahesh R. Sanghavi
SNJB

Dr. M. R. Sanghavi
Associate Professor and Head, Department of Computer Engineering,
SNJB's KBJ College of Engineering, Chandwad, Nashik

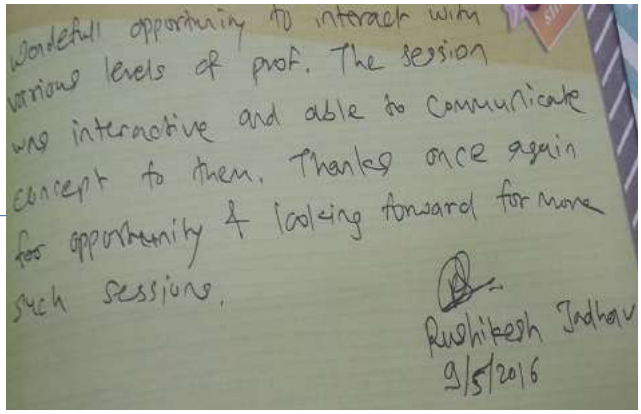

9/May/2016.
First of all thanks for giving me an opportunity to interact with participant of this training program on Big data Analytics. It is good experience for me to share thoughts and I hope that the all participant will explore further after the Completion of training program, defines the success of program.
Best of luck for the Complete team behind the program. Hoping in future too to meet for some other programs.
Thanks & Regards!
P.M. Jawandhiya

Dr. P.M.Jawandhiya
Principal, Pankaj Laddhad Institute of Technology and Management Studies,Buldana



Nirmit Kale
Technical Architect, eNight Cloud R & D Division,
ESDS Software Solutions Pvct. Ltd., Nashik

Rushikesh Jadhav
Technical Architect, eNight Cloud R & D Division,
ESDS Software Solutions Pvct. Ltd., Nashik

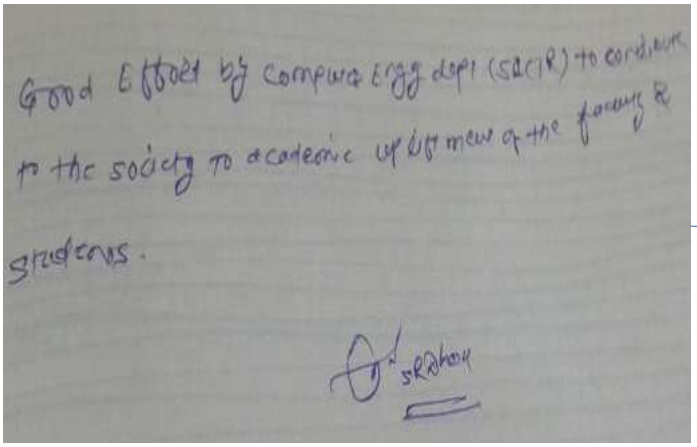


Sandip Foundation's

Sandip Institute of Technology & Research Center, NashiK

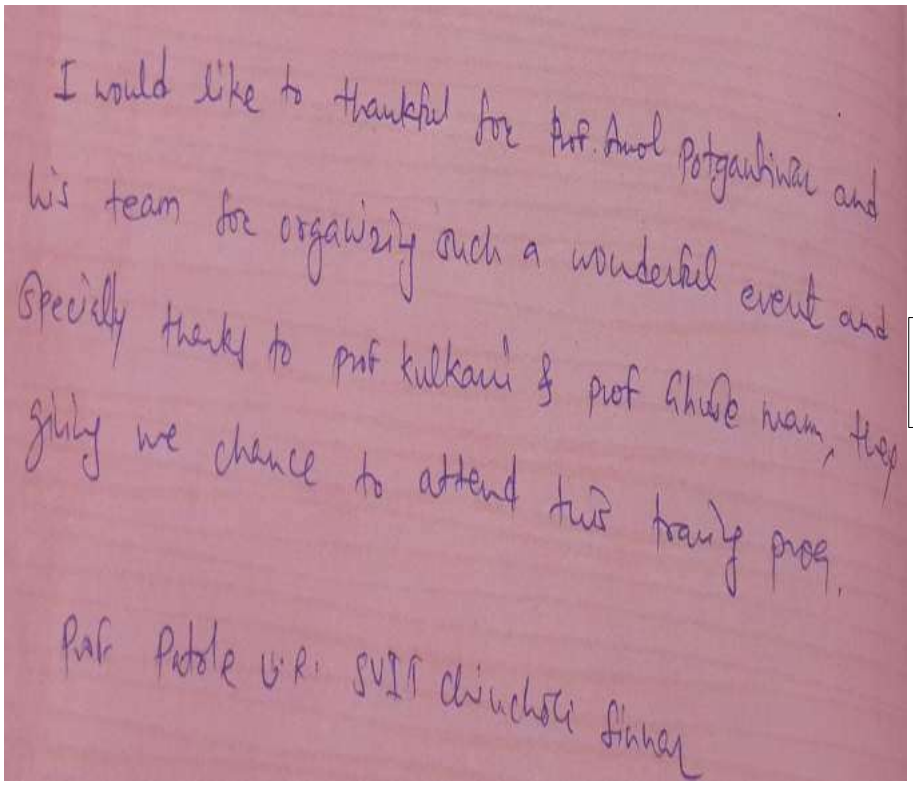
DEPARTMENT OF COMPUTER ENGINEERING

सत्यमेव जयते
Department of Science and Technology
Ministry of Science and Technology
Government of India

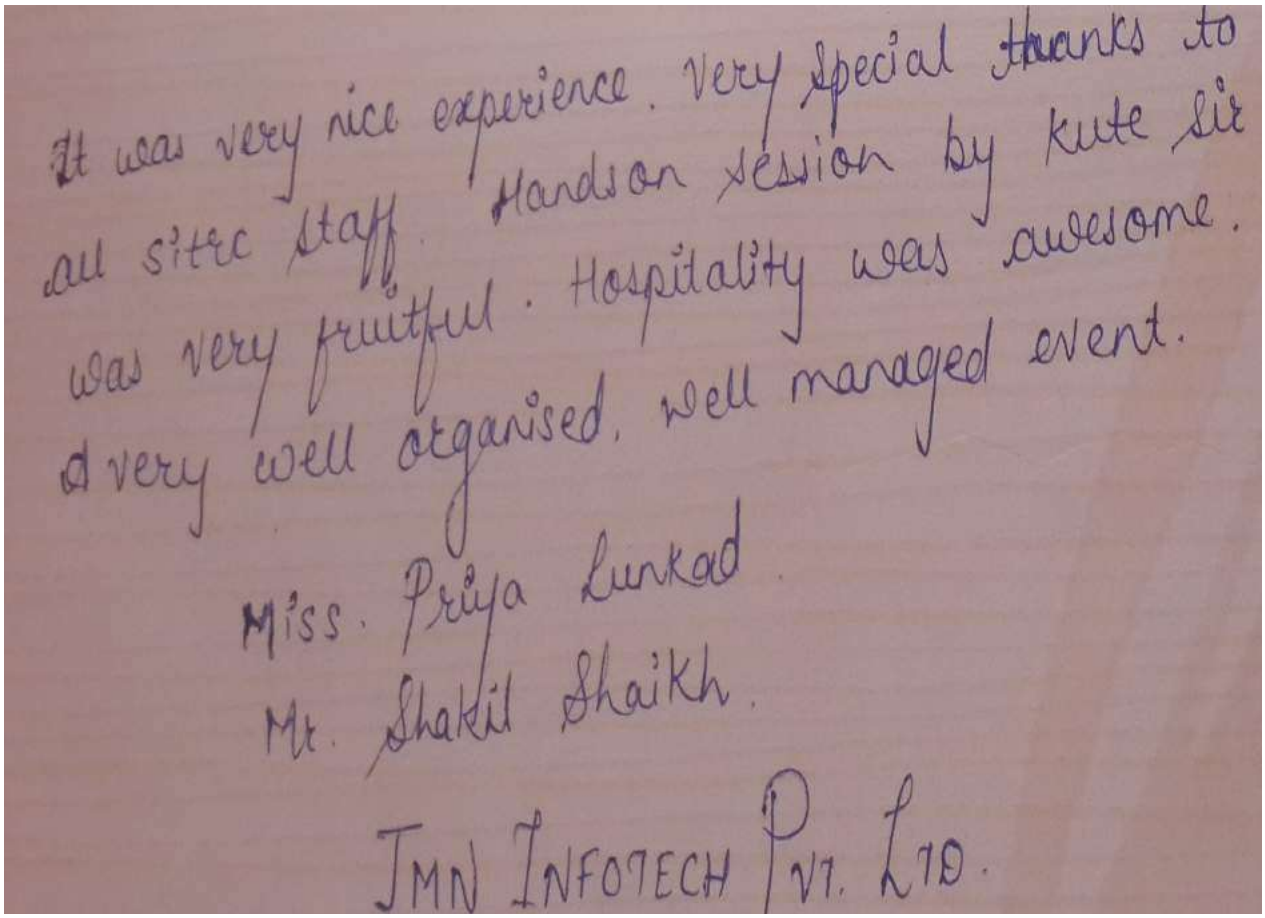


Prof. S. R. Dhore
Professor and Head, Department of Computer Engineering,
Army Institute of Technology, Pune

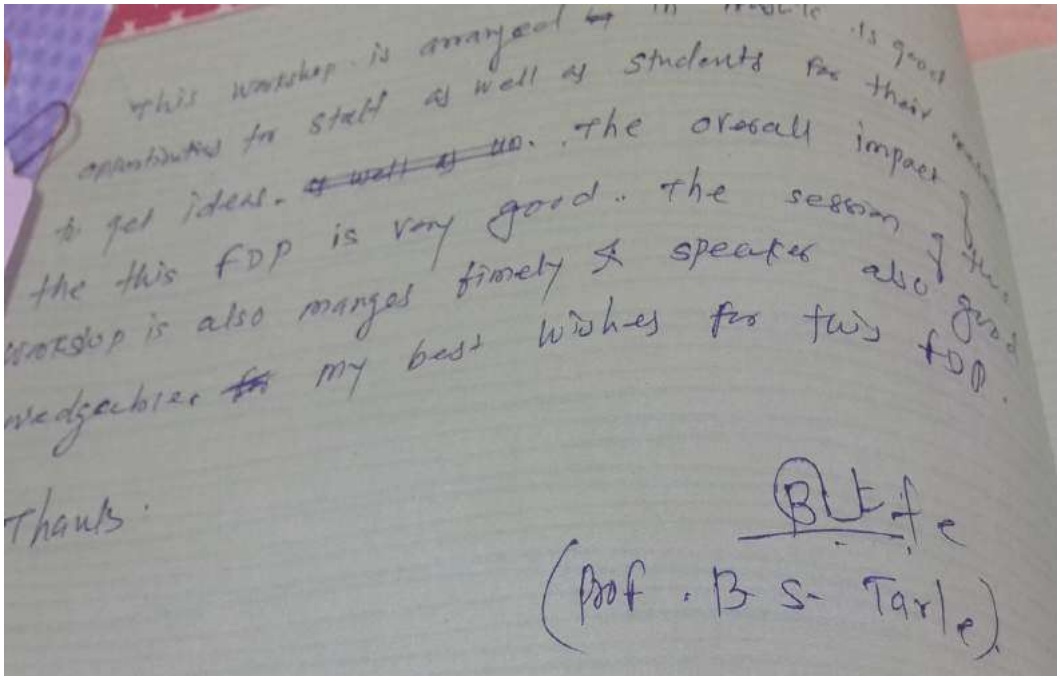
Quotes from Participants



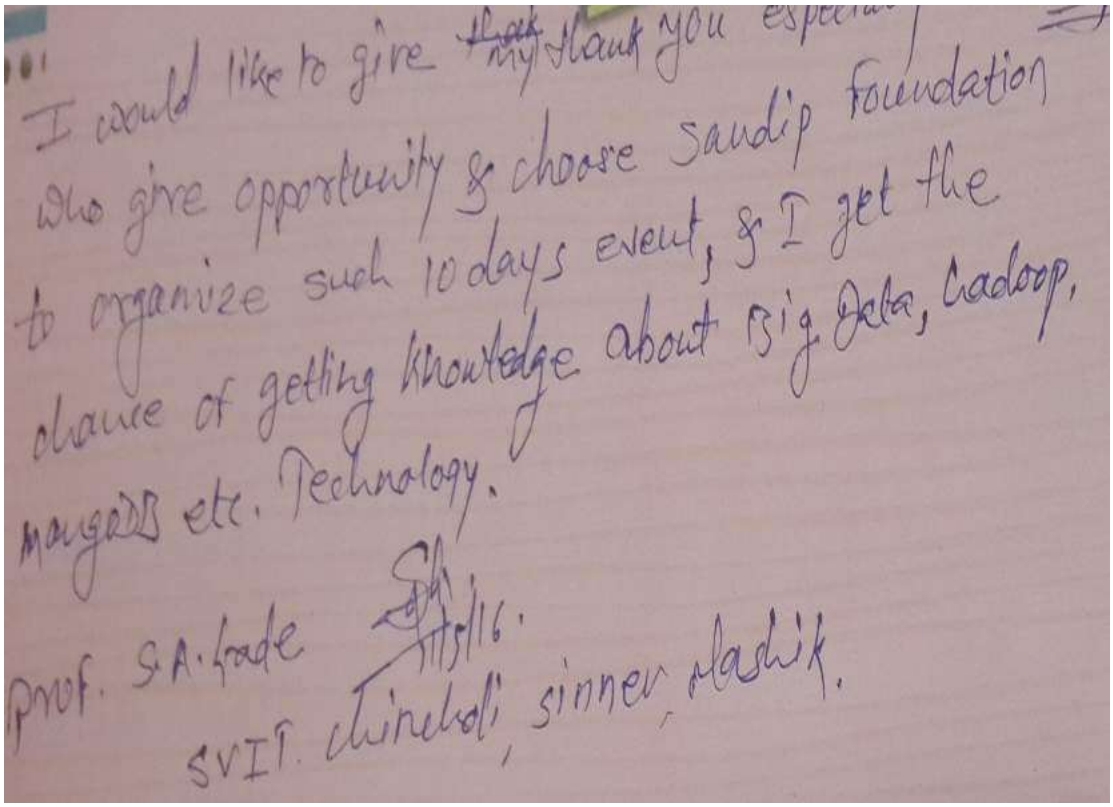
Prof. Uttam Patole
SVIT Chincholi



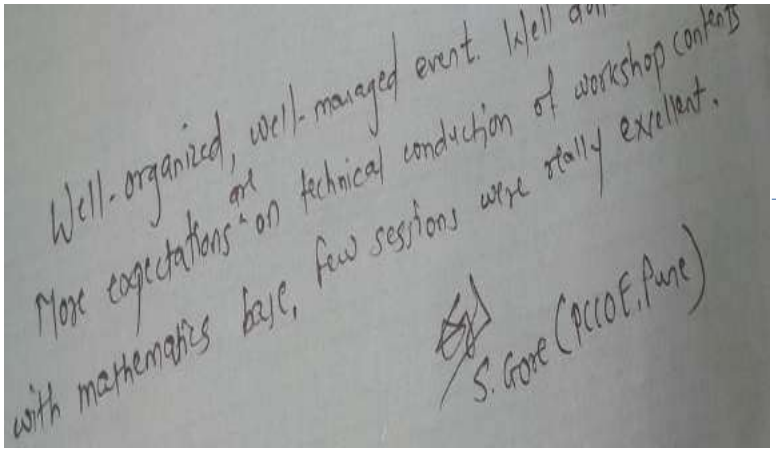
Miss.Priya Lunkad
Mr.Shakil Shaikh
JMN Infotech Pvt.ltd



Prof.B.S.Tarle
HOD Computer Engg.
KBTCOE,Nashik.

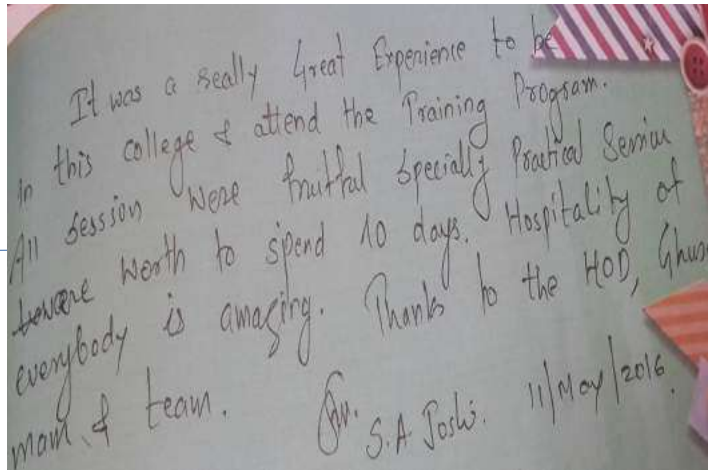


Prof.S.A.Gade
SVIT Chincholi.



Prof.S.G.Gore
PCCOE,Pune..

Prof.S.A.Joshi



Outcome

- Able to Harness data mining methods to answer crucial business questions from internal and external data sources
- Able to Create competitive advantage from both structured and unstructured data
- Able to Predict outcomes with supervised machine learning techniques
- Able to Unearth patterns in customer behavior with unsupervised techniques
- Able to Work with R and RHadoop to analyze structured, unstructured and Big Data
- Able to Model and implement efficient big data solutions for various application areas using appropriately selected algorithms and data structures.
- Able to analyse methods and algorithms, to compare and evaluate them with respect to time and space requirements, and make appropriate design choices when solving real-world problems.
- Able to motivate and explain trade-offs in big data processing technique design and analysis in written and oral form.
- Able to Explain the Big Data Fundamentals, including the evolution of Big Data, the characteristics of Big Data and the challenges introduced.
- Able to Apply non-relational databases, the techniques for storing and processing large volumes of structured and unstructured data, as well as streaming data.
- Able to apply the novel architectures and platforms introduced for Big data, in particular Hadoop and MapReduce.
- Able to apply the basic principles behind Data Warehouses and their use in data analytics, with emphasis on the design of Data Warehousing applications, systems used and algorithms for processing vast amounts of data.

Industry Contribution

Allied industry, "JMN Infotech Pvt. Ltd, Nashik" sooner opening up division of IoT based product manufacturing with Keeping focus on customers from Healthcare, Automobile and Education sector. 2 participants were attended the training programme in the view of inculcating Big Data Methodologies for better customer satisfaction and service.

Big data is about data, plain and simple. Yes, you can add all sorts of adjectives when talking about "big" data, but at the end of the day, it's all data.

IoT is about data, devices, and connectivity. Data – big and small – is front and center in the IoT world of connected devices.Productive Research Contribution has been decided from mutual ends as follows:

- Publishing Research Articles in renowed journals and International Conferences
- Filing patents based prototypes of both technologies
- Producing Customer demand-driven and Research based products and services

Conclusions and Recommendations

The training programme has largely benefitted the participants, who to re-stress, came from institutes of national importance and wide geographical locations. The objectives, clearly laid down before the start of the programme, can be seen to be clearly achieved with great success. The programme has not only given an insight into the various research challenges and areas of work in big data, it has also enabled them to understand the basic technology from a greater perspective and get a practical insight into the problems. Further, the stress on hands on workshop, as well as through repeated instructions at the workshop has played a major role in orienting the participants and encouraging them for proposal writing.

Based on the experience of the summer programme, it is recommended that such programmes be organized in large numbers, thereby benefitting more participants. The research scholars particularly showed active interest for the programme, despite being non-funded for the programme. Considering the excitement, some of them had to be accepted for participation. A larger number of funded seats for research scholars from leading Government institutes is recommendable.

A problem faced at the summer programme was the interdisciplinary nature of big data, thereby incorporating participants from multiple domains and with multiple expectations. The future versions of the training programme may well focus upon a particular specialized segment of applications of Big Data,rather than attempting a programme on the entire domain. This would also help in improving the duration of the programme, which some participants have stated.

While this training programme showed both the therotical and application based aspects, and the techniques to work over with the demonstrated softwares, thus aiming that the participants get a good insight into the whole concept.

Feedbacks indicate that they instead want to play and program the applications themselves, and setups are not very acceptable. The future versions of the programme need to aim at such workshops, including having updated software technologies in sufficient numbers for the participants to play around.

Schedule for Big Data Analytics training program

Date	Day_Wise Curriculum	Name of Speaker	Contents
2nd May 16	**Day 1:Data Science basics** a) Data Preparation and processing	**Dr.S.G.Bhirud**	Keynote Address
	b) Introduction to R and visualization	**Dr.V.M Thakare**	Data formats,Data read and write,Data cleaning and transformation;R basics, lists, arrays, matrices, tables, function,objects;
	c)Basics of Matrices and Linear Algebra d) Basics of Probability and Statistics	**Mr.Shrikant Pande**	R scatterplotm, biplot, correlation, histogram, boxplot, barplot, pair-plot, etc;Handling large matrices; Basic results in probability; Basic defintion in statistics; sampling from given distributors, random sampling from data, summarizing random samples using statistics
3Rd May 16	**Day 2:Basic data models** a) Linear regression modeling and diagnostics	**Dr.S.S.Sane**	Linear models and generalized linear models;
	b) Multiple linear regression modeling c) Logistic regression and binary	**Prof. N.M Shahane**	Modeling and prediction with R package lm, glm and biglm; Plotting models fit to data; diagnostics plots; Finding significant variables inmultiple regression;
	classification d) Decision tree modelling	**Mr.Shrikant Pande**	Classification and Regression Tree(CART); Plotting decision trees; Very Fast Decision Trees(VFDT)
4th May 16	**Day 3:Advanced data models** a) Model evaluation and improvement	**Dr.P.N.Mahalle**	Model performance via sensitivity, specificity, precision and recall, ROC curves, cross validation
	b) Nonlinear classification methods	**Dr.S.R.Sakhare**	Ensemble methods, Bagging, Boosting and Random Forests KNN, SVM and Neural networks;High Dimensional data matrix maipulation;
	c) Dimensionality reduction methods d)Matrix decomposition methods	**Dr.D.V.Patil**	Variable selection by penalized regression such as LASSO and lars ; Multi-Dimesinal Scaling(MDS); Principal components analysis(PCA); Singular value decomposition(SVD)
5th May 16	**Day 4:Finding structures in data** a) Clustering methods b) Outlier analysis c) Association analysis d) Network analysis and optimization	**Mr.Ganesh Bhosale**	K-means, Partitioninng around medoids(PAM), Visualization of clustering results, clustering evaluation using Silhouette coeficients and other indices, BIRCH clustering for large datasets, Anomoly detection, Finding frequent Itemsets using A-Priori Algorithm and variants, e.g.SON; Basics of networks centrality measures, Network link analysis, PageRank algorithm
6th May 16	**Day 5:MapReduce and Hadoop** a) MapReduce basics b) Hadoop MapReduce	**Dr.M.R.Sanghavi**	Understanding MapReduce architechture and dataflow; Hadoop MapReduce entities; Hadoop Distributed File System(HDFS);
	c) HDFS basics d) Haddop Ecosystem	**Mr.Pawan Tiwari**	NameNodes and DataNodes, Writing Hadoop MapReduce programs;Basics of Hbase, Hive, Pig,Sqoop,etc

7th May 16	**Day 6:Distributed and parallel computing using R** a) Integrating R and Hadoop b) RHIPE and RHadoop	**Dr.S.S.Prabhune**	RHIPE architecture; R hadoop architecture;R hadoop examples; HadoopStreaming R package;
	c) Applications on large data d)High-performance and parallel R	**Prof.T.B.Kute**	Writing programs for linear and logistics regression, and clustering and classification, using R and Hadoop; bigmemory , biglars, speedflm, bigrf R packages; Introducing the various parallel computing packages in R(e.g Rmpi, parallel, snow, snowfall)
8th May 16	**Day 7:Analyzing data in motion** a) Real time data streams b) Stream data basics	**Dr.M.U. Kharat**	Complex Event Processing; Basics of One-pass computing, Online algorithms,
	c) Data stream analysis platforms d) R based stream analysis	**Mr.Prasad Chandane**	Sampling streams, Concept drift; Massive online Analysis(MOA); Storm ; Spark; R package like stream, Rstorm; Graphics for real-time analytics
9th May 16	**Day 8:Databases and operationalizing Big Data** a)Relational and Non-relational databases	**Mr.Amol Kalugade**	Features of various databases including relational, document, column-based, graph, spatial, etc;
	b) R interface to databases	**Dr. Sandeep Chaware**	RpackagesRmySQL,Rexcel,RmongoDB,Rhive,RHBase,etc
	c) Managing data security and variety d) Cloud in the context of Big Data	**Mr.Rajeev Papneja**	Importing data into R and writing back results;Data security and privacy issues;Big Data Integration; Hadop as ETL; Cloud deployment and delivery models for Big Data
10 & 11th May 16	**Day 9 & Day 10:Big Data case studies** (practical lab sessions)	**Dr.P.M.Jawandhiya**	Must involve 2 or more significantly large datasets for reading,analytics and writing;Typical case studies may include Text Analysis;Sentiments analysis; Social media mining;
		Prof.Amol Kute	Web mining; Unstructured and image data analysis; Recommnedation systems; Fraud detection; Financial data analysis; Predictive analysis in time series data,Health and environmental analytics ;Genomics data analysis;Agriculture; Sensor and Uncertain data analysis; Internet of things; Data fusion; Geo-informatics and spatial statistical analysis; E-governance applications;etc

About Foundation

Sandip foundation, formed on 07.05.2005 in Mumbai as charitable trust under the chairmanship of honorable Dr. Sandip Jha with a vision to render selfless, dedicated and yeomen services to higher education in the fields of engineering, management and paramedical sciences. The trust has proved its proficiency in education for last fifteen years. At sandip foundation, the focus is on interactive teaching learning, industrial projects, on job training, faculty empowerment, industrial visits so that learning is applicable oriented and students develop the necessary skills.

About Institute

Sandip foundation was established with the primary objective of rendering selfless, dedicated and yeomen service to higher education in engineering branches, management, arts, science and paramedical sciences. The trustees are already involved in the field of higher education for last fifteen years. During the period the trustee have rendered valuable services to thousands of students. Sandip institute of technology and research (SITRC) is offering from the sandip foundation to the people of Nasik. SITRC is located in the scenic, eco-friendly and conductive to study campus at an elevation off the trimbak road mahiravani, nasik. SITRC is committed to imparting quality education in an atmosphere that will ensure that its students are confident, self motivated and industry ready. Towards this goal, we are giving importance to qualified and experienced faculty for effective teaching-learning process, equipped our laboratories with best in -class machines and instruments and developing overall personality of our students.

Scope of Training Programme

- To promote new areas of Science & Technology and to play the role of a nodal department for organizing, coordinating and promoting Science and technology activities in the country.
- To create employment opportunities in the knowledge economy sectors including promotion of semi conductor /micro/ nano/ Bio -technology based manufacturing units in the State.
- Providing an opportunity for teachers/ scholars to familiarize themselves with modern engineering practices including the latest technological advances adopted by industry.

Advisory Committee

Dr. S.T.Gandhe	Principal,SITRC
Dr. P.R.Baviskar	Dean Academic, Mech. Department
Dr. R.S. Patil	Dean Admin & Head MBA Department
Dr. S. N. Patil	Head Applied Science Department
Dr. P. G. Burade	Head Electrical Engg. Department
Dr. G.M.Phade	Head E & TC Engg.Department
Prof. A.S.Maheshwari	Head Mech Engg. Department
Prof. J.G.Nayak	Head Civil Engg. Department

Organizing Committee

Prof. Santosh Kumar	Asst. Prof. Computer Engg.
Prof. N. C. Thoutam	Asst. Prof. Computer Engg.
Prof. S. M. Walunj	Asst. Prof. Computer Engg.
Prof. R. S. Shirsath	Asst. Prof. Computer Engg.
Prof. A. R. Gadekar	Asst. Prof. Computer Engg.
Prof. S. S. Jore	Asst. Prof. Computer Engg.
Prof. A. H. Palve	Asst. Prof. Computer Engg.
Prof. S. A. Sonawane	Asst. Prof. Computer Engg.
Prof. B. A. Patil	Asst. Prof. Computer Engg.
Prof. K. B. Mahajan	Asst. Prof. Computer Engg.
Prof. N. L. Kulkarni	Asst. Prof. Computer Engg.
Prof. H. P. Patil	Asst. Prof. Computer Engg.
Prof. A. P. Deore	Asst. Prof. Computer Engg.
Prof. B. F. More	Asst. Prof. Computer Engg.
Prof. P. J. Bhisikar	Asst. Prof. Computer Engg.
Prof. A. V. Karale	Asst. Prof. Computer Engg.

Convener

Prof. N. D. Ghuse (Computer Engg Dept.)

Email: namrata.ghuse@sitrc.org

Co-convener

Prof. S. N. Patil (Computer Engg Dept.)

Email: shweta.patil@sitrc.org

असतो मा सद्गमय ।
तमसो मा ज्योतिर्गमय ।।
SANDIP FOUNDATION

सत्यमेव जयते
Department of Science & Technology
Govt. of India

Sponsored By

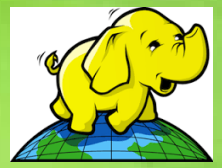
DST ,Government of India, under Big Data Initiative

10 Days Training Program

On

Big Data Analytics

(2nd May To 11th May 2016)



Chief Patron

Dr. Sandip Kumar Jha

(Chairman, Sandip Foundation)

Patrons

Prof. Mohini Patil

(G.M. Sandip Foundation)

Prof. P. I. Patil

(Mentor,Sandip Foundation)

Dr. S.T. Gandhe

(Principal)

Organized By

Prof. A. D. Potgantwar

Head

DEPARTMENT OF COMPUTER ENGINEERING

SANDIP FOUNDATION'S

SANDIP INSTITUTE OF TECHNOLOGY AND RESEARCH CENTER

MAHIRAVANI, NASHIK – 422 213

Phone: (02594) 222 554,222 5555

Who Can Apply?

Ph.D/PG research scholars, scientist/researcher in relevant area, teachers from educational institutes, industry sectors functioning in relevant area , entrepreneur willing like to work on this area.

Course Contents

- **Day 1**: Data Science Basics
- **Day 2**: Basic Data Models
- **Day 3**: Advanced Data Models
- **Day4**: Finding Structures in Data
- **Day 5**: Map Reduce & Hadoop
- **Day 6**: Distributed and parallel computing using R
- **Day 7**: Analyzing data in motion
- **Day 8**: Databases and operationalizing Big Data
- **Day 9**: Big Data case studies
- **Day 10**: Big Data case studies(cont'd)

Venue

Sandip Foundation's SITRC
Computer Engineering Dept
Buildng "C"

Registration Details

- **NO REGISTRATION FEES FOR PARTICIPANTS**
- **NO TA/DA FOR PARTICIPANTS**
- **MAX 30 REGISTRATIONS ON FIRST COME FIRST SERVE BASIS**

Participants must visit: "https://docs.google.com/forms/d/1FrV296q & fill online "Registration Form".

Registration Committee

Prof. M. V. Korade
Email:mahesh.korade@sitrc.org
Mob.No:9545554474

Prof. B. A. Patil
Email:bharati.patil@sitrc.org
Mob.No:9960430010

Accommodation

Accommodation will be provided on request and first come first serve basis in college hostel,at cost.

Contact(Accomodation)

Prof. P. J. Bhisikar
Email:pradnyesh.bhisikar@sitrc.org
Mob.No: 8550902002

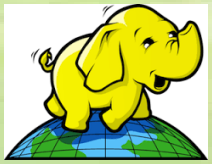
Mr. T. P. Jagtap
Email:tushar.jagtap@sitrc.org
Mob. No 9021110053

Important Dates

Last date of receiving Registration form
29th April 2016

Speakers

Eminent resourse persons from research institutes & IT industries like IBM, L&T, Prygma etc .



Sandip Foundation's

Sandip Institute of Technology & Research Centre , Nashik

Sponsored By
DST ,Government of India, under Big Data Initiative
10 Days Training Program
On

Big Data Analytics

From 2nd To 11th May 2016

Registration Form

FullNameMr/Mrs/Dr.----------
---------- ----------
---------- ----------
Designation:----------
Department:----------
Organization:----------
HighestQualification:----------
Accommodation required: Yes/No
Address for correspondance:----------

Pin----------
Email:----------
Date:----------
Signature:----------

Recommendation of Sponsoring Authority.

The Applicant is hereby recommended and will be permitted to attend the above Ten Days Training Programme & notified duration, if selected.

Date: Signature & Seal of Sponsoring Authority