B.TECH. CSE with specialization in Big Data Analytics

Departmental Elective-I	Introduction to Big data(BCS 049)
Departmental Elective-II	Cloud computing and virtualization(BCS 059)
Departmental Elective-III	Analytics and statistical modeling for big data(BCS 069)
Departmental Elective-IV	Application Development in Cloud(BCS 078)
Open Elective	Machine Learning (BOE 078)
Departmental Elective-V	Data Visualization(BCS 088)
Departmental Elective-VI	Hadoop and MapReduce(BCS 091)

Department Elective-I

BCS-049: Introduction to Big data

L T P **Credit-4**

3 1 2

UNIT I **08 hours**

Introduction to Big Data Analytics: Big Data overview, State of the practice in analytics role of data scientists, Big Data Analytics in industry verticals.

UNIT II **08 hours**

End-to-end Data Analytics Life Cycle: key roles for successful analytic project, main phases of life cycle, developing core deliverables for stakeholders.

UNIT III **08 hours**

Basic Analytic Methods: introduction to "R", analyzing and exploring data with "R", statistics for model building and evaluation

UNIT IV **08 hours**

Advanced Analytics and Statistical Modeling for Big Data: Naïve Bayseian Classifier, K-means Clustering, Association Rules, Decision Trees, Linear and Logistic Regression, Time Series Analysis, Text Analytics;

UNIT V **08 hours**

Technology and Tools – MapReduce/Hadoop , In- database Analytics, MADlib and advanced SQL Tools

References:

1. Noreen Burlingame ,The little book on Big Data, New Street publisher(eBook)
2. http://www.prlog.org/11800911-just-published-the-little-book-of-big-data-2012-edition.html
3. Norman Matloff ,The Art of R Programming: A Tour of Statistical Software Design , ISBN 13: 978-1-59327-384-2; ISBN-10: 1-59327-384-3
4. http://www.johndcook.com/R_language_for_programmers.html
5. http://bigdatauniversity.com/
6. http://home.ubalt.edu/ntsbarsh/stat-data/topics.htm#rintroduction

Department Elective-II

BCS-059: Cloud Computing and Virtualization

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Cloud Computing Fundamental: Cloud computing definition, private, public and hybrid cloud. Cloud types; IaaS, PaaS, SaaS. Benefits and challenges of cloud computing, public vs private clouds, role of virtualization in enabling the cloud; Business Agility: Benefits and challenges to Cloud architecture. Application availability, performance, security and disaster recovery; next generation Cloud Applications.

UNIT-II **08 hours**

Cloud Applications: Technologies and the processes required when deploying web services; Deploying a web service from inside and outside a cloud architecture, advantages and disadvantages.

UNIT-III **08 hours**

Cloud Services Management: Reliability, availability and security of services deployed from the cloud. Performance and scalability of services, tools and technologies used to manage cloud services deployment; Cloud Economics: Cloud Computing infrastructures available for implementing cloud based services. Economics of choosing a Cloud platform for an organization, based on application requirements, economic Constraints and business needs (e.g Amazon, Microsoft and Google, Salesforce.com, Ubuntu and Redhat

UNIT-IV **08 hours**

Application Development: Service creation environments to develop cloud based applications. Development environments for service development; Amazon, Azure, Google App.

UNIT-V **08 hours**

Best Practice Cloud IT Model: Analysis of Case Studies when deciding to adopt cloud computing architecture. How to decide if the cloud is right for your requirements. Cloud based service, applications and development platform deployment so as to improve the total cost of ownership (TCO).

References

1. Gautam Shroff, *Enterprise Cloud Computing Technology Architecture Applications* [ISBN: 978-0521137355]
2. Toby Velte, Anthony Velte, Robert Elsenpeter, *Cloud Computing, A Practical Approach* [ISBN: 0071626948]
3. Dimitris N. Chorafas, *Cloud Computing Strategies* [ISBN: 1439834539]

Department Elective-III

BCS-069: Analytics and statistical modeling for big data

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Basics: Quality assurance and management. Quality costs. Aims and objectives of statistical process control. Chance and assignable causes of variation. Statistical quality control. Process control, Rational subgroups. product control. Importance of statistical quality control in Industry.

UNIT-II **08 hours**

Charts for variables: Theoretical basis and practical background of control charts for variables. 3 sigma limits, warning limits and probability limits. Criteria for detecting lack of control. Derivation of limits and construction X , R and s charts and interpretation. Group control charts and sloping control charts. Natural tolerance limits and specification limits. Process capability studies. O.C and ARL curve for variable charts.

UNIT-III **08 hours**

Control charts for attributes: np chart, p chart, c chart and u chart. Basis, construction and interpretation. OC and ARL curve for attribute charts.

UNIT-IV **08 hours**

Product Control: Sampling inspection and 100 percent inspection. AQL, LTPD, Producer's risk and consumer's risk. Acceptance sampling. Sampling plans-single and double sampling plans by attributes.

UNIT-V **08 hours**

Reliability: Reliability concepts. Reliability of components and systems. Life distributions, reliability functions, hazard rate, common life distributions-Exponential, Gamma and Weibull. System reliability, Series, parallel, standby systems, r/n systems. Complex systems

Text Books

1. Montgomery D. C., Introduction to Statistical Quality Control. Wiley International edition, (1985)
2. K. S. Krishnamurthy, Reliability Methods for Engineers, ASQ Press, (1992)

Reference Books

1. Grant E. L. and Leavenworth R. S., Statistical Quality control, , McGrawHill, 6th edition (1988)
2. Gupta R. C., Statistical Quality Control, Khanna Pub. Co.

Department Elective-IV

BCS-077: Application Development in Cloud

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Cloud Based Applications: Introduction, Contrast traditional software development and development for the cloud. Public v private cloud apps. Understanding Cloud ecosystems – what is SaaS/PaaS, popular APIs, mobile;

UNIT-II **08 hours**

Designing code for the Cloud: Class and Method design to make best use of the Cloud infrastructure; Web Browsers and the Presentation Layer: Understanding Web browsers attributes and differences. Building blocks of the presentation layer: HTML, HTML5, CSS, Silverlight, and Flash.

UNIT-III **08 hours**

Web Development Techniques and Frameworks : Building Ajax controls, introduction to Javascript using JQuery, working with JSON, XML, REST. Application developement Frameworks e.g. Ruby on Rails , .Net, Java API's or JSF; Deployment Environments – Platform As A Service (PAAS) ,Amazon, vmForce, Google App Engine, Azure, Heroku, AppForce

UNIT-IV **08 hours**

Use Case 1: Building an Application using the LAMP stack: Setting up a LAMP development environment. Building a simple Web app demonstrating an understanding of the presentation layer and connectivity with persistence.

UNIT-V **08 hours**

Use Case 2: Developing and Deploying an Application in the Cloud: Building on the experience of the first project students will study the design, development, testing and deployment of an application in the cloud using a development framework and deployment platform.

References:

1. Chris Hay, Brian Prince, *Azure in Action* [ISBN: 978-1935182481]
2. Henry Li, *Introducing Windows Azure* [ISBN: 978-1-4302-2469-3]
3. Eugenio Pace, Dominic Betts, Scott Densmore, Ryan Dunn, Masashi Narumoto, MatiasWoloski, *Developing Applications for the Cloud on the Microsoft Windows Azure Platform* [ISBN: 9780735656062]
4. Eugene Ciurana, *Developing with Google App Engine* [ISBN: 978-1430218319]
5. Charles Severance, *Using Google App Engine* [ISBN: 978-0596800697]
6. George Reese 2009, *Cloud application architectures*, O'Reilly Sebastopol, CA [ISBN: 978-0596156367]
7. Dan Sanderson, *Programming Google App Engine* [ISBN: 978-0596522728]
8. Paul J. Deitel, Harvey M. Deitel 2008, *Ajax, rich Internet applications, and web development for programmers*, Prentice Hall Upper Saddle River, NJ [ISBN: 978-0-13-158738-0]

Open Elective

Machine Learning

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Basics: Introduction to machine learning - different forms of learning; Basics of probability theory, linear algebra and optimization.

UNIT-II **08 hours**

Regression Analysis: Linear regression, ridge regression, Lasso, Bayesian regression, regression with basic functions.

UNIT-III **08 hours**

Classification Methods: Linear Discriminant Analysis, Logistic regression,Perceptrons, Large margin classification,Kernel methods, Support Vector Machines.Classification and Regression Trees, Multi-layer Perceptrons and Back propagation

Graphical Models: Bayesian Belief Networks, Markov Random Fields, Exact inference methods, approximate inference methods.

UNIT-IV **08 hours**

Ensemble Methods: Boosting - Adaboost, Gradient Boosting; Bagging - Simple methods, Random Forest.

Computational Learning Theory: PAC Learning, VC Dimension, Bias/Variance Tradeoff.

UNIT-V **08 hours**

Clustering: Partitional Clustering - k-means, k-medoids; Hierarchical Clustering - Agglomerative, Divisive, Distance measures; Density based clustering - DBScan; Spectral clustering.

Frequent Pattern Mining: Apriori Algorithm; FP-Growth

References:

1. Elements of Statistical Learning. Hastie, Tibshirani, and Friedman. Springer
2. Pattern Recognition and Machine Learning. Christopher Bishop.
3. Data Mining: Tools and Techniques, 3rd Edition. Jiawei Han and Michelline Kamber.

Department Elective-V

BCS-098: Data Visualization

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Data and Image Models, Value of Visualization, Graphical Excellence, Graphical Integrity, Sources of Graphical Integrity, Visual Display of Quantitative Information, Visualization Design, The Power of Representation, Data-Ink and Graphical Redesign, Data-Ink Maximization and Graphical Design.

UNIT-II **08 hours**

Exploratory Data Analysis, Data Density and Small Multiples, Macro/Micro Readings, In Envisioning Information, Low-Level Components of Analytic Activity in Information Visualization, HTML/CSS, JavaScript, SVG, Technology Fundamentals, In Interactive Data Visualization for the Web, Multidimensional Data, A System for Query, Analysis and Visualization of Multi-dimensional Relational Databases, Graphical Perception, Perception in visualization, Layering and Separation, Interactive Data Visualization for the Web,

UNIT-III **08 hours**

Interaction, Interactive Dynamics for Visual Analysis, Animation, Animated Transitions in Statistical Data Graphics, Effectiveness of Animation in Trend Visualization, Animated Exploration of Graphs with Radial Layout, Color, Color and Information, Color Use Guidelines for Data Representation, Color Naming Models for Color Selection, Image Editing and Palette Design,

UNIT-IV **08 hours**

Design Critiques, the Cartogram: Value-by-Area Mapping. In Cartography: Thematic Map Design, Adaptive Composite Map Projections, Narrative Visualization, A Deeper Understanding of Sequence in Narrative Visualization.

UNIT-V **08 hours**

Text Visualization, Information Visualization for Search Interfaces, Information Visualization for Text Analysis, Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis.

References

1. Visualizing Data: Exploring and Explaining Data with the Processing Environment by Ben Fry, O'Reilly Media, 2007.
2. Interactive Data Visualization for the Web by Scott Murray, O'Reilly Media, 2012.
3. The Visual Display of Quantitative Information by Edward Tufte, 2001.
4. Network Science Book Project By Albert-László Barabási (Visualizations by Mauro Martino, Analysis by Márton Pósfai), E-Book, 2012.
5. The Nature of Code (Uses Processing) By Daniel Shiffman, Interactive HTML Book, 2012.

Department Elective-VI

BCS-092: Hadoop and MapReduce

L T P **Credit-4**

3 1 2

UNIT-I **08 hours**

Introduction to big data and Hadoop, Hadoop Architecture, Installing Ubuntu with Java 1.8 on VM Workstation 11, Hadoop Versioning and Configuration Single Node Hadoop, installation on Ubuntu, Multi Node Hadoop, Linux commands and Hadoop commands, Cluster architecture and block placement, Modes in Hadoop, Local Mode,Pseudo Distributed Mode, Fully Distributed Mode, Hadoop Daemon, Master Daemons(Name Node, Secondary Name Node, Job Tracker), Slave Daemons(Job tracker, Task tracker) Task Instance, Hadoop HDFS Commands, Accessing HDFS, CLI Approach, Java Approach.

UNIT-II **08 hours**

Map-Reduce, Understanding Map Reduce Framework, Inspiration to Word-Count Example, Developing Map-Reduce Program using Eclipse Luna, HDFS Read-Write Process, Map-Reduce Life Cycle Method, Serialization(Java), Datatypes, Comparator and Comparable(Java), Custom Output File, Analysing Temperature dataset using Map-Reduce, Custom Partitioner & Combiner, Running Map-Reduce in Local and Pseudo Distributed Mode.

UNIT-III **08 hours**

Advanced Map-Reduce, Enum(Java), Custom and Dynamic Counters, Running Map-Reduce in Multi-node Hadoop Cluster, Custom Writable, Site Data Distribution, Using Configuration, Using DistributedCache, Using stringifie, Input Formatters, NLine Input Formatter, XML Input Formatter,

UNIT-IV **08 hours**

Sorting, Primary Reverse Sorting,Secondary Sorting, Compression Technique, Working with Sequence File Format, Working with AVRO File Format, Testing MapReduce with MR Unit, Working with NYSE DataSets, Working with Million Song DataSets, Running Map-Reduce in Cloudera Box

UNIT-V **08 hours**

Hive Introduction & Installation, Data Types in Hive, Commands in Hive, Exploring Internal and External Table, Partitions, Complex data types, UDF in Hive, Built-in UDF, Custom UDF, Thrift Server, Java to Hive Connection, Joins in Hive, Working with HWI, Bucket Map-side Join, More commands, View, SortBy, Distribute By, Lateral View, Running Hive in Cloudera.

References

1. Berlińska, Joanna; Drozdowski, Maciej "Scheduling divisible MapReduce computations". Journal of Parallel and Distributed Computing. doi:10.1016/j.jpdc.2010.12.004. Retrieved 2016-01-14.
2. MapReduce: Simplified Data Processing on Large Clusters http://www.mcs.anl.gov/research/projects/mpi/mpi-standard/mpi-report-2.0/mpi2-report.htm
3. Ullman, J. D. (2012). "Designing good MapReduce algorithms". XRDS: Crossroads, The ACM Magazine for Students (Association for Computing Machinery) doi:10.1145/2331042.2331053