



Chapter 1 : Introduction : Data Science and Big Data	1-1 to 1-44
1.1 Introduction to Data Science and Big Data.....	1-1
1.1.1 The Five Vs (Characteristics) of Big Data (Data Explosion)	1-2
1.1.2 Major Applications of Big Data	1-4
1.1.3 Data Formats	1-5
1.1.4 Comparison between Data Formats.....	1-6
1.1.5 DIKW Pyramid	1-6
1.1.6 Big Data Examples (Sources of Big Data)	1-8
1.1.7 Categories of Data Analytics	1-11
1.1.8 Comparison between Categories of Data Analytics.....	1-13
1.1.9 Drivers (Motivation) of Big Data Analytics.....	1-13
1.1.10 Emerging Big Data Ecosystem and New Approach	1-15
1.1.11 Key Roles in the New Big Data Ecosystem.....	1-16
1.1.12 Key Roles for a Successful Analytics Project.....	1-17
1.2 Big Data Infrastructure Challenges	1-18
1.3 Big Data Processing Architectures	1-20
1.3.1 Data Warehouse.....	1-20
1.3.2 Data Warehouse Architecture	1-21
1.3.3 Benefits of a Data Warehouse	1-22
1.3.4 Steps in Data Warehousing.....	1-22
1.3.5 Characteristics of a Data Warehouse.....	1-22
1.3.6 Schemas in Data Warehouses.....	1-22
1.3.6(A) Star Schema.....	1-22
1.3.6(B) Snowflake Schema	1-23
1.3.7 Data Warehouse vs Data Lake.....	1-24
1.3.8 Data Warehouse vs Database	1-24
1.3.9 Data Warehouse vs Data Mart.....	1-25
1.3.10 Re-Engineering the Data Warehouse	1-25
1.4 Shared-Everything and Shared-Nothing Architecture.....	1-27
1.4.1 Shared-Nothing Architecture	1-27
1.4.2 Shared-Everything Architecture.....	1-28
1.4.3 Comparison between Shared-Nothing and Shared-Everything Architecture	1-28



1.5	Data Analytics Life Cycle (Big Data Analytics).....	1-29
1.5.1	Phase 1 : Discovery.....	1-29
1.5.2	Phase 2 : Data Preparation	1-30
1.5.3	Phase 3 : Model Planning.....	1-30
1.5.4	Phase 4 : Model Building.....	1-31
1.5.5	Phase 5 : Communicate Results.....	1-31
1.5.6	Phase 6 : Operationalise.....	1-32
1.6	Data Science – The Big Picture	1-32
1.6.1	Business Intelligence (BI) vs Data Science.....	1-33
1.6.2	Relationship between Data Science and Information Science	1-33
1.7	Introduction to Machine Learning.....	1-34
1.7.1	How does Machine Learning Work ?	1-36
1.7.2	Key Terms Associated with Machine Learning.....	1-38
1.8	Types of Machine Learning (Big Data Learning Approaches)	1-38
1.8.1	Supervised Learning.....	1-38
1.8.2	Unsupervised Learning	1-39
1.8.3	Reinforcement Learning.....	1-40
1.8.4	How to Choose the Right Machine Learning Algorithm?	1-41
1.9	Statistical Learning.....	1-42

Chapter 2 : Mathematical Foundation of Big Data
2-1 to 2-36

2.1	Need of Statistics in Data Science and Big Data Analytics	2-1
2.1.1	Sampling Distributions	2-2
2.1.2	General Statistics	2-3
2.1.2(A)	Mean	2-3
2.1.2(B)	Median	2-3
2.1.2(C)	Mode.....	2-4
2.1.2(D)	Mid-range.....	2-4
2.1.2(E)	Range.....	2-4
2.1.3	Standard Deviation.....	2-5
2.1.4	Variance	2-6
2.1.5	Covariance	2-6
2.1.6	Mean Absolute Deviation.....	2-8
2.1.7	ANOVA (Analysis of Variance).....	2-9



2.2	Concepts of Probability.....	2-9
2.2.1	Fundamental Rules of Probability.....	2-10
2.3	Tail Bounds.....	2-11
2.4	Pair-Wise Independence and Universal Hashing.....	2-12
2.5	Approximate Counting.....	2-13
2.6	Approximate Median.....	2-14
2.7	Random Variables.....	2-15
2.7.1	Discrete Random Variables.....	2-15
2.7.2	Continuous Random Variables.....	2-16
2.7.3	Multiple Random Variables.....	2-16
2.7.4	Markov Models.....	2-16
2.7.5	Random Walk.....	2-18
2.7.5(A)	Steady State.....	2-20
2.7.5(B)	Hidden Markov Model.....	2-23
2.8	Flajolet-Martin (FM) Algorithm (LogLog Counting).....	2-29
2.9	Bloom Filters.....	2-29
2.10	Distance Sampling and Random Projections.....	2-32

Chapter 3 : Big Data Processing**3-1 to 3-36**

3.1	Big Data Analytics-Ecosystem and Technologies.....	3-1
3.2	Introduction to Google File System (GFS).....	3-1
3.2.1	Characteristics and Features of GFS.....	3-2
3.2.2	Architecture of GFS.....	3-2
3.2.3	Apache Hadoop.....	3-3
3.2.3(A)	MapReduce.....	3-3
3.2.3(B)	Hadoop Distributed File System (HDFS).....	3-7
3.2.3(C)	YARN (Yet Another Resource Negotiator).....	3-9
3.3	Common Hadoop Shell commands.....	3-10
3.3.1	Cluster Setup - SSH and Hadoop Configuration.....	3-13
3.3.2	Operating the Hadoop Cluster.....	3-14
3.3.3	Anatomy of File Read.....	3-15
3.3.4	Anatomy of File Write.....	3-15
3.4	Introduction to NoSQL.....	3-16
3.4.1	Reasons for Choosing NoSQL Databases.....	3-16
3.4.2	Types of NoSQL Databases.....	3-19



3.4.2(A)	Key-Value	3-19
3.4.2(B)	Document.....	3-20
3.4.2(C)	Column.....	3-21
3.4.2(D)	Graph.....	3-22
3.4.3	Comparison between Relational Database and NoSQL Database	3-23
3.4.4	CAP Theorem (Brewer’s Theorem).....	3-24
3.5	Textual ETL Processing (Introduction to Text Analysis).....	3-25
3.5.1	Challenges in Text Analysis	3-25
3.5.2	Steps in Text Analysis.....	3-26
3.5.3	Text Pre-Processing Techniques.....	3-27
3.5.4	Bag-of-Words	3-30
3.5.5	Bag-of-n-Grams	3-31
3.6	Term Frequency - Inverse Document Frequency (TFIDF)	3-31
3.6.1	Term Frequency (TF)	3-32
3.6.2	Inverse Document Frequency (IDF).....	3-32

Chapter 4 : Big Data Analytics**4-1 to 4-40**

4.1	Big Data Analytics – Architecture and Life Cycle.....	4-1
4.2	Types of Analysis – Analytical Approaches.....	4-1
4.3	Data Ingestion from Different Sources	4-1
4.3.1	Pandas.....	4-1
4.4	Data Quality and Remediation (Data Pre-Processing)	4-13
4.4.1	Common Data Quality Issues	4-13
4.4.2	Remediating (Fixing) Data Quality Issues.....	4-14
4.5	Data Wrangling	4-16
4.5.1	Need for Data Wrangling.....	4-17
4.5.1(A)	Data	4-17
4.5.1(B)	Tasks.....	4-17
4.5.1(C)	Models.....	4-18
4.5.1(D)	Features.....	4-18
4.5.1(E)	Feature Engineering.....	4-19
4.5.1(F)	Data Engineering -vs- Feature Engineering.....	4-21
4.6	Data Wrangling Methods (Data Analytics with Mathematical Manipulations)	4-22
4.6.1	Feature Scaling or Normalisation.....	4-26
4.6.2	Min-Max Scaling.....	4-27



4.6.3	Standardisation (Variance Scaling)	4-27
4.6.4	Encoding Categorical Variables.....	4-29
4.6.5	One-Hot Encoding.....	4-30
4.6.6	Dummy Coding.....	4-31
4.6.7	Feature Hashing.....	4-31
4.7	Multi-class Classification Techniques (Handling Categorical Data with 2 and More Categories).....	4-33
4.7.1	One vs One (OvO).....	4-33
4.7.2	One vs Rest (OvR) (Ove vs All).....	4-34
4.7.3	Comparison between OvO and OvR.....	4-34
4.8	Hive Data Analytics.....	4-35
4.8.1	Characteristics and Features of Hive.....	4-35
4.8.2	Architecture of Hive.....	4-36
4.9	Statistical and Graphical Analysis Methods.....	4-36

Chapter 5 : Big Data Visualization**5-1 to 5-41**

5.1	Introduction to Data Visualisation.....	5-1
5.1.1	Goals (Objectives) of Data Visualisation.....	5-2
5.2	Challenges (Difficulties) with Big Data Visualisation.....	5-4
5.3	Techniques for Visual Data Representations.....	5-5
5.3.1	Conventional Data Visualisation Tools.....	5-5
5.3.2	Types of Data Visualisation.....	5-6
5.3.2(A)	Comparative Plots.....	5-7
5.3.2(B)	Statistical Plots.....	5-12
5.3.2(C)	Topology Plots.....	5-19
5.3.2(D)	Spatial Plots.....	5-21
5.4	Data Visualisation Taxonomy.....	5-24
5.5	Visualizing Big Data.....	5-26
5.6	General Workflow of Analytics and Visualisation.....	5-28
5.7	Tools Used in Data Visualisation.....	5-29
5.7.1	Tableau.....	5-29
5.7.2	Microsoft Power BI.....	5-30
5.7.3	Qlik.....	5-31
5.7.4	ThoughtSpot.....	5-32
5.7.5	Candela.....	5-33



5.7.6	D3.js.....	5-33
5.7.7	Google Charts.....	5-34
5.8	Analytical Techniques Used in Big Data Visualisation.....	5-35
5.9	Case Study - Analysis of a Business Problem of Zomato Using Visualisation.....	5-36

Chapter 6 : Big Data Technologies Application and Impact
6-1 to 6-14

6.1	Social Network Analysis (SNA) (Social Media Analytics).....	6-1
6.1.1	Need (Applications) of Social Network Analysis	6-2
6.2	Text Mining	6-4
6.3	Mobile Analytics	6-4
6.4	Data Analytics Life Cycle of Case Studies	6-5
6.5	Big Data Value Creation Drivers	6-6
6.6	Big Data Analytics Challenges and Research directions	6-6
6.7	Michael Porter’s Valuation Creation Models	6-6
6.8	Organisational Impact of Big Data.....	6-7
6.9	Understanding Decision Theory.....	6-8
6.10	Creating Big Data Strategy.....	6-9
6.11	Big Data User Experience Ramifications	6-10
6.12	Identifying Big Data Use Cases	6-11

