2nd EDITION

BIG DATA AND ANALYTICS

AIML
005.7 SEE
10134325

Seema Acharya
Subhashini Chellappan

WILEY

10134325

2nd EDITION

BIG DATA AND ANALYTICS

NEW HORIZON COLLEGE OF ENGINEERING
NHCE
IN PURSUIT OF EXCELLENCE
LIBRARY
Bengaluru - 560 103

Seema Acharya
Infosys Limited

Subhashini Chellappan

WILEY

NEW HORIZON COLLEGE OF ENGINEERING
LIBRARY
Acc. No. 101343
Call No.
Bengaluru - 560 103

Big Data and Analytics

SECOND EDITION

Copyright © 2019 by Wiley India Pvt. Ltd., 4436/7, Ansari Road, Daryaganj, New Delhi-110002.

Cover Image: © aleksandarvelasevic/Getty Images

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or scanning without the written permission of the publisher.

Limits of Liability: While the publisher and the author have used their best efforts in preparing this book, Wiley and the author make no representation or warranties with respect to the accuracy or completeness of the contents of this book, and specifically disclaim any implied warranties of merchantability or fitness for any particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by sales representatives or written sales materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every individual. Neither Wiley India nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Disclaimer: The contents of this book have been checked for accuracy. Since deviations cannot be precluded entirely, Wiley or its author cannot guarantee full agreement. As the book is intended for educational purpose, Wiley or its author shall not be responsible for any errors, omissions or damages arising out of the use of the information contained in the book. This publication is designed to provide accurate and authoritative information with regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services.

Trademarks: All brand names and product names used in this book are trademarks, registered trademarks, or trade names of their respective holders. Wiley is not associated with any product or vendor mentioned in this book.

Other Wiley Editorial Offices:

John Wiley & Sons, Inc. 111 River Street, Hoboken, NJ 07030, USA

Wiley-VCH Verlag GmbH, Pappellaee 3, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 1 Fusionopolis Walk #07-01 Solaris, South Tower, Singapore 138628

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada, M9W 1L1

First Edition: 2015

Second Edition : 2019

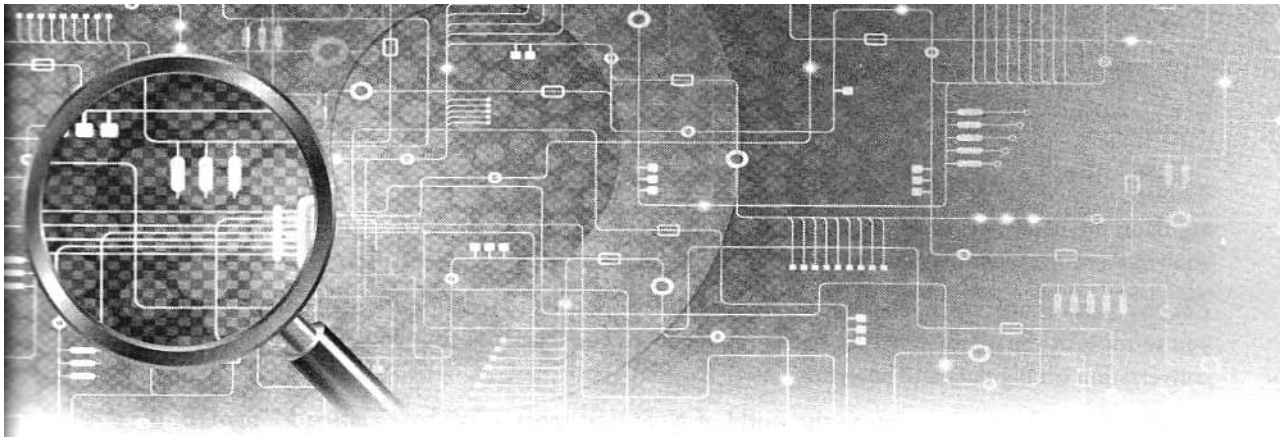
Reprint: 2022

ISBN: 978-81-265-7951-8

ISBN: 978-81-265-8836-7 (ebk)

www.wileyindia.com

Printed at: Printways



Preface

The last few years have been witness to a burgeoning growth of data. We have heard it being called Big Data! So what really is this big data? Big data is an evolving term used to describe any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. There is data everywhere from the sensors that gather weather information, to the likes, posts and comments on social media sites, to digital pictures, audios and videos that get circulated, to the conversations in a chat room, etc. All this and more is big data.

Need for this Book

We felt the need to compose a book, egged on by the enthusiasm and inquisitiveness of the students and instructors fraternity alike. A book which can take the readers through an easy comprehension of the big data technology landscape. Ours is an attempt to cover a plethora of technologies from NoSQL databases such as MongoDB and Cassandra to components of the Hadoop Ecosystem such as MapReduce, Pig and Hive to delving into analytics with association rule mining on one hand and decision trees on the other hand.

The Audience

This book is for all interested in learning about Big Data, Hadoop and Analytics. The only criteria is the willingness to learn and the ability to stretch yourself in learning to limits that you have not done before. The book is for all those who are new to big data irrespective of the field/background that you come from.

The book will be equally useful to an engineering graduate as it would be to a management graduate. The book has been designed and crafted such that it caters to the knowledge requirements of an IT person as well as a business user with ease.

Organization of the Book

This book has a total of 14 chapters. Here is a sneak peek into the chapters of our book…

Chapters 1–4 of the book provide a basic understanding of the types of digital data, the characteristics of big data, the challenges confronting the enterprises embracing big data, the sudden hype around big data analytics and the technologies that make up the big data landscape.

Chapter 5 introduces the open source software framework called Hadoop. We have attempted to introduce you to most of the major concepts and components to empower you to hold your own in any meaningful conversation on big data and analytics.

Chapters 6 and 7 introduce you to the world of NoSQL databases. We have chosen MongoDB, the document-oriented database, and Cassandra, the wide column store, to get you a feel of NoSQL databases. In explaining the NoSQL databases, we have built on the familiarity that the readers will have with RDBMS (Relational Database Management System).

Chapter 8 introduces you to the nitty-gritties of MapReduce Programming. The merits and challenges have been dealt with for a clearer appreciation.

Chapters 9 and 10 cover two major components of the Hadoop Ecosystem, namely "Pig" and "Hive".

Chapter 11 introduces to you an open source tool to draw out reports by pulling data from NoSQL databases.

Chapter 12 is focused on introducing you to the world of machine learning and analysis with algorithms under both supervised and unsupervised learning categories.

Chapter 13 is focused on bringing out the differences between various Hadoop ecosystem components for an easy lookup and remembrance. It will be good to read this chapter sequentially for better absorption. Starting with data warehouse versus data lakes, HDFS versus the first non-batch component – Hbase, it builds further to explain the differences between HDFS and RDBMS, then goes onto to highlight differences of MapReduce with Pig, Spark and finally delves into the differences between Pig and Hive.

Chapter 14 discusses the big data trends in 2019 and beyond. The years ahead will see an increase in the adoption of open-source technologies. Hadoop is and will remain fundamental, although there will be increased usage of the in-memory Spark. The years ahead will also awake to the container(ed) revolution. The last half a decade has been a witness to the commoditization of visualization. The rising wave of IoT (Internet of Things) will lead to processing being done on the edge of the network before moving it to the central data center in the cloud. The world will witness the power of empowered computing – edge and quantum. It is time to utilize and draw value/insight from the abundant dark data. Also bots will mature and get smarter in the coming years.

Glossary: A glossary of terms frequently used in the big data and analytics parlance is given at the end of the book. Although we strive to define terms as we introduce them in this book, we think you'll find the glossary a useful resource.

To get most out of this Book

We have included sections such as "POINT ME", "CONNECT ME", "TEST ME" to enable you to further your learning and comprehension.

The section "*POINT ME*" provides a list of books that you as a reader should check out to further your learning.

The section "*CONNECT ME*" provides a list of reference links which will feed you with good content on topics covered in the chapter.

The section "*TEST ME*" has a gamut of self-assessments such as "Crossword" puzzles, "Fill in the blanks", "Match the columns", etc. We have provided solved and unsolved exercises to better your learning.

There are *HANDS-ON ASSIGNMENTS* provided with MongoDB, Cassandra, MapReduce, Pig, Hive and JasperReports. We sincerely urge you to attempt these to gain good hands-on practice on these major technologies.

Next Steps...

We have endeavored to create an overview of big data and introduced you to all its significant components. We recommend you to read the book from cover to cover, but if you are not that kind of person, we have made an attempt to keep the chapters self-contained so that you can go straight to the topics that interest you most.

Whichever approach you may choose, we wish you well!

Available with the Book (www.wileyindia.com)

We have put together an installation guide to help our learners with easy steps to install and configure a Hadoop cluster. The steps to setting up the components of the Hadoop ecosystem such as MapReduce, Pig and Hive have also been explained in easy, DIY (Do It Yourself) steps.

We have provided a Microsoft Access Database (.accdb) and a text file on which we have based an assignment that when attempted and solved will surely challenge and satiate you.

A Quick Word for the Instructors' Fraternity

Attention has been paid in arriving at the sequence of chapters and also to the flow of topics within each chapter. This is done particularly with an objective to assist our fellow instructors and academicians in carving out a syllabi from the Table of Contents (TOC) of the book. The complete TOC can qualify as the syllabi for a semester or if the college has an existing syllabus on big data and analytics, a few chapters can be added to the syllabi to make it more robust. We leave it to your discretion on how you wish to use the same for your students.

We have ensured that each tool/component discussed in the book is with adequate hands-on content to enable you to teach better and provide ample hands-on practice to your students.

The easy-to-follow installation guide provided on the website should help you set up the lab environment for practice.

We have also provided Instructor Resources (IR) that can be procured directly from our publisher, Wiley India by visiting their website or writing to acadmktg@wiley.com. These Instructor Resources are presentation decks (one for each chapter) which can be taken to the class directly or can be customized as per your requirements.

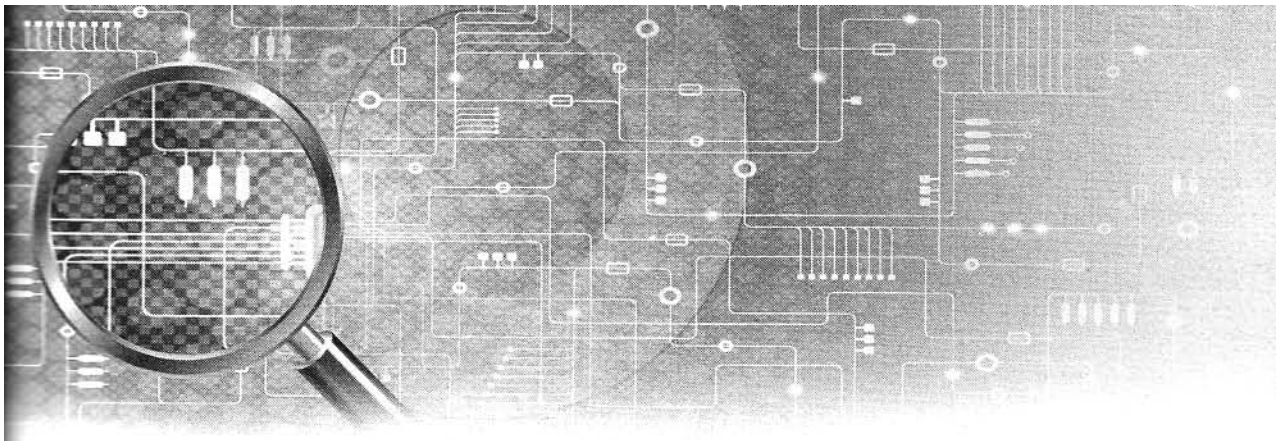
Connect with Authors

To stay connected with the students and instructors fraternity, we run a group on LinkedIn titled, "*Exploring big data and analytics*". Join us to discuss, share and learn!!!

Happy Learning!!!

Seema Acharya

Subhashini Chellappan



Acknowledgements

The making of the book was like a journey that we had undertaken for several months. We had our families, friends, colleagues, and well-wishers onboard this journey and we wish to express our heartfelt gratitude to each one of them. Without their unflinching support and affection, we could not have pulled it off.

We are grateful to the student and teacher community who kept us on our toes with their constant bombardment of queries which prompted us to learn more, simplify our learnings and findings, and place them neatly in the book. This book is for them.

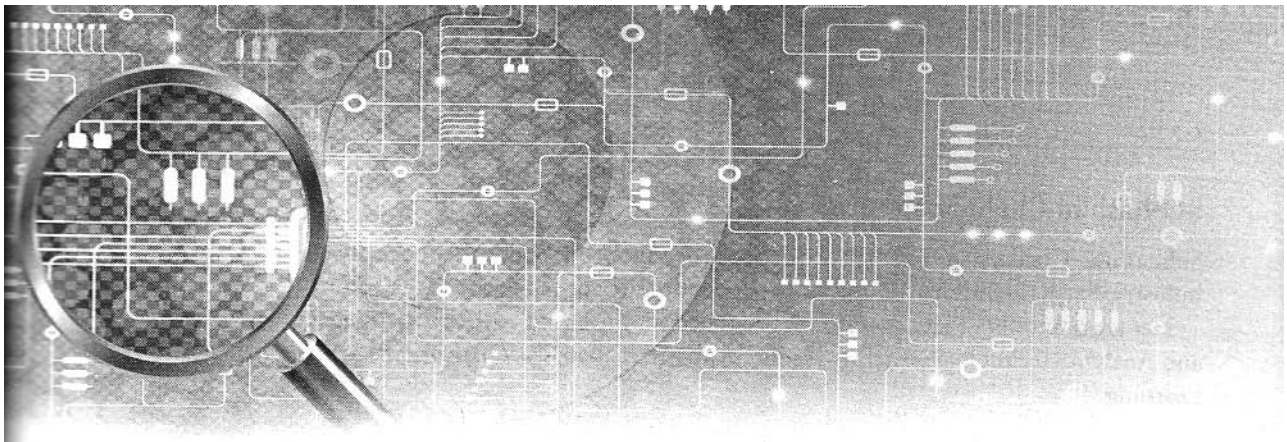
We wish to thank our friends – the practitioners from the field for filling us in on the latest in the big data field and sharing with us valuable insights on the best practices and methodologies followed therein.

A special thanks to R N Prasad for his encouragement and vigilant review.

We have been fortunate to have the support of our teams who sometimes knowingly and at other times unknowingly contributed to the making of the book by lending us their unwavering support.

We consider ourselves very fortunate for the editorial assistance provided by Wiley India. We wish to acknowledge and appreciate Meenakshi Sehrawat, Associate Publisher and her team of associates who adeptly guided us through the entire process of preparation and publication. Appreciation is also due to Rakesh Poddar and his team for working with us through the entire production process.

And finally we can never sufficiently thank our families and friends who have been our pillars of strength, our stimulus, and our soundboards all through the process, and endured patiently our crazy schedules as we assembled the book.



Author Profile

Seema Acharya

Seema Acharya is a Senior Lead Principal with the Education, Training and Assessment department of Infosys Limited. She is a technology evangelist, a learning strategist, and an author with over 15+ years of IT experience in learning/education services. She has designed and delivered several large-scale competency development programs across the globe involving organizational competency need analysis, conceptualization, design, development and deployment of competency development programs. She is an educator by choice and vocation, and has rich experience in both academia and the software industry.

She is also the author of the following books:

1. "Fundamentals of Business Analytics", ISBN: 978-81-265-3203-2, publisher – Wiley India.
2. "Pro Tableau – A Step by Step Guide", ISBN; 978-1484223512, publisher – Apress.
3. "Data Analytics using R", ISBN: 9789352605248, publisher – McGraw Hill Higher Education Society (2018).

She has co-authored a paper on "Collaborative Engineering Competency Development" for ASEE (American Society for Engineering Education). She holds the patent on "Method and System for Automatically Generating Questions for a Programming Language".

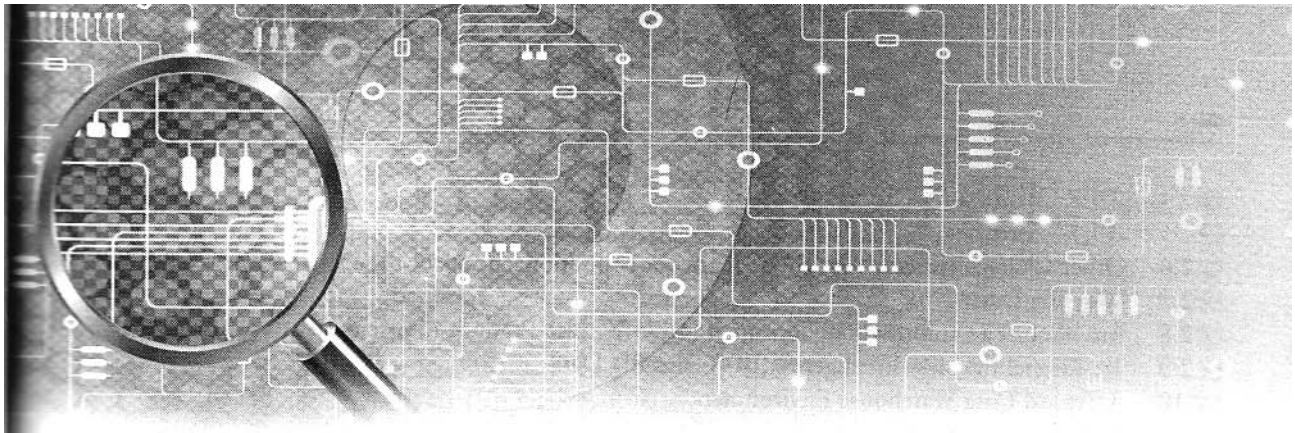
Her areas of interest and expertise are centered on Business Intelligence, Big Data and Analytics, technologies such as Data Warehousing, Data Mining, Data Analytics, Text Mining and Data Visualization.

She is passionate about exploring new paradigms of learning and also dabbles into creating e-learning content to facilitate learning anytime and anywhere.

Subhashini Chellappan

Subhashini Chellappan has rich experience in both academia and the software industry. She has published couple of papers in various Journals and Conferences.

Her areas of interest and expertise are centered on Business Intelligence, Big Data and Analytics technologies such as Hadoop, NoSQL Databases, Spark and Machine Learning.



Contents

Preface iii

Acknowledgements vii

Author Profile ix

Chapter 1 Types of Digital Data 1

What's in Store? 1
1.1 Classification of Digital Data 2
1.1.1 Structured Data 2
1.1.2 Semi-Structured Data 5
1.1.3 Unstructured Data 7
Remind Me 10
Point Me (Book) 11
Connect Me (Internet Resources) 11
Test Me 11
Scenario-Based Question 15

Chapter 2 Introduction to Big Data 17

What's in Store? 17
2.1 Characteristics of Data 18
2.2 Evolution of Big Data 19
2.3 Definition of Big Data 19
2.4 Challenges with Big Data 21
2.5 What is Big Data? 22
2.5.1 Volume 22

	2.5.2 Velocity	24
	2.5.3 Variety	25
2.6	Other Characteristics of Data Which are not Definitional Traits of Big Data	25
2.7	Why Big Data?	25
2.8	Are We Just an Information Consumer or Do We also Produce Information?	26
2.9	Traditional Business Intelligence (BI) versus Big Data	26
2.10	A Typical Data Warehouse Environment	27
2.11	A Typical Hadoop Environment	27
2.12	What is New Today?	28
	2.12.1 Coexistence of Big Data and Data Warehouse	28
2.13	What is Changing in the Realms of Big Data?	29
	Remind Me	30
	Point Me (Book)	30
	Connect Me (Internet Resources)	30
	Test Me	31
	Challenge Me	32

Chapter 3 Big Data Analytics **35**

What's in Store?		35
3.1	Where do we Begin?	36
3.2	What is Big Data Analytics?	37
3.3	What Big Data Analytics Isn't?	37
3.4	Why this Sudden Hype Around Big Data Analytics?	39
3.5	Classification of Analytics	39
	3.5.1 First School of Thought	40
	3.5.2 Second School of Thought	40
3.6	Greatest Challenges that Prevent Businesses from Capitalizing on Big Data	41
3.7	Top Challenges Facing Big Data	41
3.8	Why is Big Data Analytics Important?	42
3.9	What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?	42
3.10	Data Science	43
	3.10.1 Business Acumen Skills	43
	3.10.2 Technology Expertise	43
	3.10.3 Mathematics Expertise	44
3.11	Data Scientist…Your New Best Friend!!!	44
	3.11.1 Responsibilities of a Data Scientist	44
3.12	Terminologies Used in Big Data Environments	45
	3.12.1 In-Memory Analytics	45
	3.12.2 In-Database Processing	45
	3.12.3 Symmetric Multiprocessor System (SMP)	46
	3.12.4 Massively Parallel Processing	46
	3.12.5 Difference Between Parallel and Distributed Systems	46
	3.12.6 Shared Nothing Architecture	47

3.12.7 CAP Theorem Explained 49
3.13 Basically Available Soft State Eventual Consistency (BASE) 52
3.14 Few Top Analytics Tools 52
3.14.1 Open Source Analytics Tools 53
Remind Me 53
Connect Me (Internet Resources) 53
Test Me 54

Chapter 4 The Big Data Technology Landscape 57

What's in Store? 57
4.1 NoSQL (Not Only SQL) 58
4.1.1 Where is it Used? 58
4.1.2 What is it? 58
4.1.3 Types of NoSQL Databases 59
4.1.4 Why NoSQL? 60
4.1.5 Advantages of NoSQL 60
4.1.6 What We Miss With NoSQL? 61
4.1.7 Use of NoSQL in Industry 62
4.1.8 NoSQL Vendors 63
4.1.9 SQL versus NoSQL 63
4.1.10 NewSQL 64
4.1.11 Comparison of SQL, NoSQL, and NewSQL 64
4.2 Hadoop 65
4.2.1 Features of Hadoop 65
4.2.2 Key Advantages of Hadoop 65
4.2.3 Versions of Hadoop 66
4.2.4 Overview of Hadoop Ecosystems 68
4.2.5 Hadoop Distributions 74
4.2.6 Hadoop versus SQL 74
4.2.7 Integrated Hadoop Systems Offered by Leading Market Vendors 75
4.2.8 Cloud-Based Hadoop Solutions 75
Remind Me 75
Point Me (Books) 76
Connect Me (Internet Resources) 76
Test Me 76

Chapter 5 Introduction to Hadoop 79

What's in Store? 80
5.1 Introducing Hadoop 80
5.1.1 Data: The Treasure Trove 80
5.2 Why Hadoop? 81
5.3 Why not RDBMS? 82
5.4 RDBMS versus Hadoop 83

5.5 Distributed Computing Challenges 83
5.5.1 Hardware Failure 83
5.5.2 How to Process This Gigantic Store of Data? 84
5.6 History of Hadoop 84
5.6.1 The Name "Hadoop" 84
5.7 Hadoop Overview 85
5.7.1 Key Aspects of Hadoop 85
5.7.2 Hadoop Components 86
5.7.3 Hadoop Conceptual Layer 86
5.7.4 High-Level Architecture of Hadoop 86
5.8 Use Case of Hadoop 87
5.8.1 ClickStream Data 87
5.9 Hadoop Distributors 88
5.10 HDFS (Hadoop Distributed File System) 88
5.10.1 HDFS Daemons 89
5.10.2 Anatomy of File Read 91
5.10.3 Anatomy of File Write 92
5.10.4 Replica Placement Strategy 93
5.10.5 Working with HDFS Commands 93
5.10.6 Special Features of HDFS 95
5.11 Processing Data with Hadoop 95
5.11.1 MapReduce Daemons 96
5.11.2 How Does MapReduce Work? 96
5.11.3 MapReduce Example 98
5.12 Managing Resources and Applications with Hadoop YARN (Yet Another Resource Negotiator) 101
5.12.1 Limitations of Hadoop 1.0 Architecture 101
5.12.2 HDFS Limitation 101
5.12.3 Hadoop 2: HDFS 101
5.12.4 Hadoop 2 YARN: Taking Hadoop beyond Batch 102
5.13 Interacting with Hadoop Ecosystem 104
5.13.1 Pig 104
5.13.2 Hive 105
5.13.3 Sqoop 105
5.13.4 HBase 105
Remind Me 106
Point Me (Books) 106
Connect Me (Internet Resources) 106
Test Me 107
Challenge Me 113

Chapter 6 Introduction to MongoDB 115

What's in Store? 115
6.1 What is MongoDB? 116
6.2 Why MongoDB? 116

6.2.1 *Using Java Script Object Notation (JSON)* 116
6.2.2 *Creating or Generating a Unique Key* 118
6.2.3 *Support for Dynamic Queries* 118
6.2.4 *Storing Binary Data* 119
6.2.5 *Replication* 119
6.2.6 *Sharding* 120
6.2.7 *Updating Information In-Place* 120
6.3 Terms Used in RDBMS and MongoDB 121
6.3.1 *Create Database* 122
6.3.2 *Drop Database* 122
6.4 Data Types in MongoDB 122
6.5 MongoDB Query Language 126
6.5.1 *Insert Method* 127
6.5.2 *Save() Method* 131
6.5.3 *Adding a New Field to an Existing Document – Update Method* 132
6.5.4 *Removing an Existing Field from an Existing Document – Remove Method* 133
6.5.5 *Finding Documents based on Search Criteria – Find Method* 133
6.5.6 *Dealing with NULL Values* 142
6.5.7 *Count, Limit, Sort, and Skip* 144
6.5.8 *Arrays* 150
6.5.9 *Aggregate Function* 158
6.5.10 *MapReduce Function* 160
6.5.11 *Java Script Programming* 161
6.5.12 *Cursors in MongoDB* 162
6.5.13 *Indexes* 166
6.5.14 *MongoImport* 168
6.5.15 *MongoExport* 169
6.5.16 *Automatic Generation of Unique Numbers for the "_id" Field* 170
Remind Me 171
Point Me (Book) 171
Connect Me (Internet Resources) 171
Test Me 172
Assignments for Hands-On Practice 175

Chapter 7 Introduction to Cassandra 177

What's in Store? 178
7.1 Apache Cassandra – An Introduction 178
7.2 Features of Cassandra 179
7.2.1 *Peer-to-Peer Network* 179
7.2.2 *Gossip and Failure Detection* 180
7.2.3 *Partitioner* 180
7.2.4 *Replication Factor* 180
7.2.5 *Anti-Entropy and Read Repair* 180
7.2.6 *Writes in Cassandra* 181

	7.2.7 Hinted Handoffs	181
	7.2.8 Tunable Consistency	182
7.3	CQL Data Types	183
7.4	CQLSH	184
	7.4.1 Logging into cqlsh	184
7.5	Keyspaces	184
7.6	CRUD (Create, Read, Update, and Delete) Operations	188
7.7	Collections	195
	7.7.1 Set Collection	195
	7.7.2 List Collection	196
	7.7.3 Map Collection	196
	7.7.4 More Practice on Collections (SET and LIST)	198
	7.7.5 Using Map: Key, Value Pair	204
7.8	Using a Counter	205
7.9	Time to Live (TTL)	206
7.10	Alter Commands	207
	7.10.1 Alter Table to Change the Data Type of a Column	208
	7.10.2 Alter Table to Delete a Column	208
	7.10.3 Drop a Table	209
	7.10.4 Drop a Database	209
7.11	Import and Export	209
	7.11.1 Export to CSV	209
	7.11.2 Import from CSV	210
	7.11.3 Import from STDIN	211
	7.11.4 Export to STDOUT	212
7.12	Querying System Tables	213
7.13	Practice Examples	216
	Remind Me	218
	Point Me (Book)	219
	Connect Me (Internet Resources)	219
	Test Me	219
	Assignments for Hands-On Practice	219

Chapter 8 Introduction to MAPREDUCE Programming 221

What's in Store?		221
8.1	Introduction	222
8.2	Mapper	222
8.3	Reducer	223
8.4	Combiner	224
8.5	Partitioner	225
8.6	Searching	228
8.7	Sorting	230
8.8	Compression	232

Remind Me 232
Point Me (Book) 232
Connect Me (Internet Resources) 233
Test Me 233
Assignment for Hands-On practice 233

Chapter 9 Introduction to Hive 235

What's in Store? 235
9.1 What is Hive? 236
9.1.1 History of Hive and Recent Releases of Hive 237
9.1.2 Hive Features 237
9.1.3 Hive Integration and Work Flow 238
9.1.4 Hive Data Units 238
9.2 Hive Architecture 239
9.3 Hive Data Types 241
9.3.1 Primitive Data Types 241
9.3.2 Collection Data Types 242
9.4 Hive File Format 242
9.4.1 Text File 242
9.4.2 Sequential File 242
9.4.3 RCFile (Record Columnar File) 242
9.5 Hive Query Language (HQL) 243
9.5.1 DDL (Data Definition Language) Statements 243
9.5.2 DML (Data Manipulation Language) Statements 243
9.5.3 Starting Hive Shell 244
9.5.4 Database 244
9.5.5 Tables 247
9.5.6 Partitions 252
9.5.7 Bucketing 255
9.5.8 Views 257
9.5.9 Sub-Query 258
9.5.10 Joins 259
9.5.11 Aggregation 260
9.5.12 Group By and Having 260
9.6 RCFile Implementation 260
9.7 SerDe 261
9.8 User-Defined Function (UDF) 262
Remind Me 263
Point Me (Books) 263
Connect Me (Internet Resources) 264
Test Me 264
Assignments for Hands-On Practice 265

Chapter 10 Introduction to Pig 269

What's in Store? 270
10.1 What is Pig? 270
10.1.1 Key Features of Pig 270
10.2 The Anatomy of Pig 270
10.3 Pig on Hadoop 271
10.4 Pig Philosophy 271
10.5 Use Case for Pig: ETL Processing 271
10.6 Pig Latin Overview 272
10.6.1 Pig Latin Statements 272
10.6.2 Pig Latin: Keywords 272
10.6.3 Pig Latin: Identifiers 272
10.6.4 Pig Latin: Comments 273
10.6.5 Pig Latin: Case Sensitivity 273
10.6.6 Operators in Pig Latin 273
10.7 Data Types in Pig 273
10.7.1 Simple Data Types 273
10.7.2 Complex Data Types 273
10.8 Running Pig 274
10.8.1 Interactive Mode 274
10.8.2 Batch Mode 275
10.9 Execution Modes of Pig 275
10.9.1 Local Mode 275
10.9.2 MapReduce Mode 275
10.10 HDFS Commands 275
10.11 Relational Operators 276
10.11.1 FILTER 276
10.11.2 FOREACH 276
10.11.3 GROUP 277
10.11.4 DISTINCT 277
10.11.5 LIMIT 278
10.11.6 ORDER BY 278
10.11.7 JOIN 279
10.11.8 UNION 279
10.11.9 SPLIT 280
10.11.10 SAMPLE 281
10.12 Eval Function 281
10.12.1 AVG 281
10.12.2 MAX 282
10.12.3 COUNT 282
10.13 Complex Data Types 283
10.13.1 TUPLE 283
10.13.2 MAP 284

10.14 Piggy Bank 284
10.15 User-Defined Functions (UDF) 285
10.16 Parameter Substitution 286
10.17 Diagnostic Operator 286
10.18 Word Count Example using Pig 287
10.19 When to use Pig? 288
10.20 When not to use Pig? 288
10.21 Pig at Yahoo! 288
10.22 Pig versus Hive 288
Remind Me 289
Point Me (Book) 289
Connect Me (Internet Resources) 289
Test Me 289
Assignments for Hands-On Practice 290

Chapter 11 JasperReport using Jaspersoft 293

What's in Store? 293
11.1 Introduction to JasperReports 293
11.1.1 JasperReports 293
11.1.2 Jaspersoft Studio 294
11.2 Connecting to MongoDB NoSQL Database 294
11.2.1 Syntax of Few MongoDB Query Language 301
11.2.2 Elements and Attributes 302
11.2.3 Creating Variables 302
11.2.4 Creating Report Parameters 304
11.3 Connecting to Cassandra NoSQL Database 305
Remind Me 308
Point Me (Book) 309
Connect Me (Internet Resources) 309
Assignment for Hands-On Practice 309

Chapter 12 Introduction to Machine Learning 311

What's in Store? 311
12.1 Introduction to Machine Learning 312
12.1.1 Machine Learning Definition 312
12.2 Machine Learning Algorithms 312
12.2.1 Regression Model – Linear Regression 313
12.2.2 Clustering 315
12.2.3 Collaborative Filtering 317
12.2.4 Association Rule Mining 322
12.2.5 Decision Tree 325
Remind Me 329
Point Me (Book) 329

Connect Me (Internet Resources) 329
Test Me 329
Assignment for Hands-On Practice 330

Chapter 13 Few Interesting Differences 331

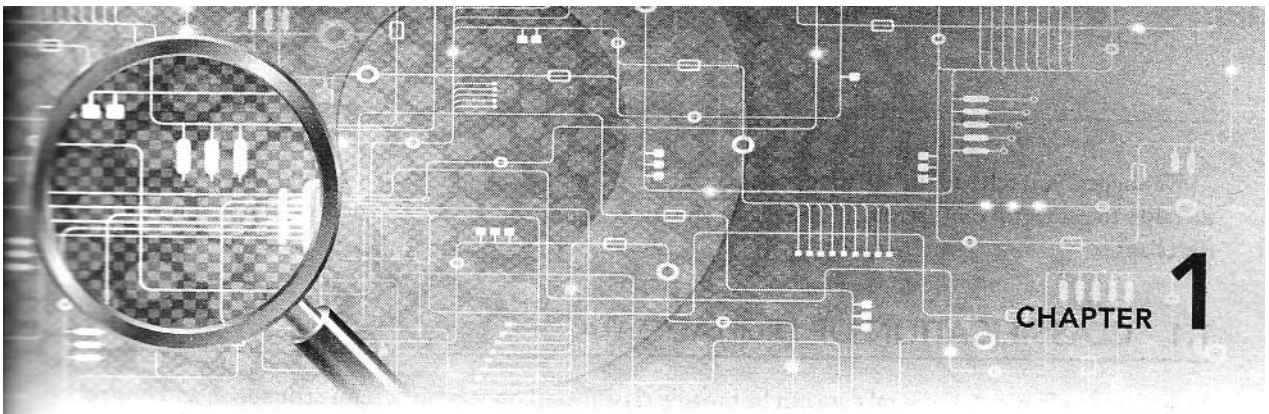
What's in Store? 331
13.1 Difference between Data Warehouse and Data Lake 331
13.2 Difference between RDBMS and HDFS 333
13.3 Difference between HDFS and HBase 334
13.4 Hadoop MapReduce versus Pig 335
13.5 Difference between Hadoop MapReduce and Spark 335
13.6 Difference between Pig and Hive 337

Chapter 14 Big Data Trends in 2019 and Beyond 339

What's in Store? 339
14.1 Rise of the New Age "Data Curators" 340
14.2 CDOs are Stepping Up 340
14.3 Dark Data in the Cloud 341
14.4 Streaming the IoT for Machine Learning 342
14.5 Edge Computing 343
14.6 Open Source 344
14.7 Hadoop is Fundamental and will Remain So! 344
14.8 Chatbots will Get Smarter 344
14.9 Container(ed) Revolution 344
14.10 Commoditization of Visualization 345

Glossary 347

Index 359



CHAPTER 1

Types of Digital Data

BRIEF CONTENTS

- What's in Store?
- Classification of Digital Data
- Structured Data
  - Sources of Structured Data
  - Ease of Working with Structured Data
- Semi-Structured Data
  - Sources of Semi-Structured Data
- Unstructured Data
  - Issues with "Unstructured" Data
  - How to Deal with Unstructured Data

"In God we trust, all others must bring data."

– W. Edwards Deming

WHAT'S IN STORE?

Irrespective of the size of the enterprise (big or small), data continues to be a precious and irreplaceable asset. Data is present internal to the enterprise and also exists outside the four walls and firewalls of the enterprise. Data is present in homogeneous sources as well as in heterogeneous sources. The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights.

Data → Information

Information → Insights

This chapter is a "must read" for first-time learners interested in understanding the role of data in business intelligence and business analysis and businesses at large. This chapter will introduce you to the various formats of digital data (structured, semi-structured, and unstructured data), the sources of each format of data, the issues with the terminology of unstructured data, etc.

We suggest you refer to the learning resources suggested at the end of this chapter and also attempt all the exercises to get a grip on this topic. We suggest you make your own notes/bookmarks while reading through the chapter.

1.1 CLASSIFICATION OF DIGITAL DATA

As depicted in Figure 1.1, digital data can be broadly classified into structured, semi-structured, and unstructured data.

1. **Unstructured data:** This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80–90% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.
2. **Semi-structured data:** This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program; for example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
3. **Structured data:** This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

Ever since the 1980s most of the enterprise data has been stored in relational databases complete with rows/records/tuples, columns/attributes/fields, primary keys, foreign keys, etc. Over a period of time Relational Database Management System (RDBMS) matured and the RDBMS, as they are available today, have become more robust, cost-effective, and efficient. We have grown comfortable working with RDBMS – the storage, retrieval, and management of data has been immensely simplified. The data held in RDBMS is typically structured data. However, with the Internet connecting the world, data that existed beyond one's enterprise started to become an integral part of daily transactions. This data grew by leaps and bounds so much so that it became difficult for the enterprises to ignore it. All of this data was not structured. A lot of it was unstructured. In fact, Gartner estimates that almost 80% of data generated in any enterprise today is unstructured data. Roughly around 10% of data is in the structured and semi-structured category. Refer Figure 1.2.

1.1.1 Structured Data

Let us begin with a very basic question – When do we say that the data is structured? The simple answer is when data conforms to a pre-defined schema/structure we say it is structured data.

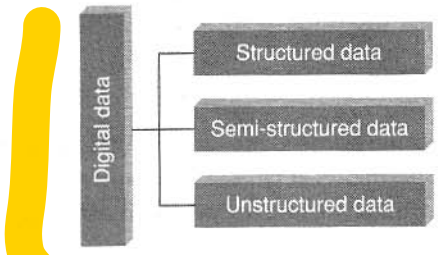


Figure 1.1 Classification of digital data.

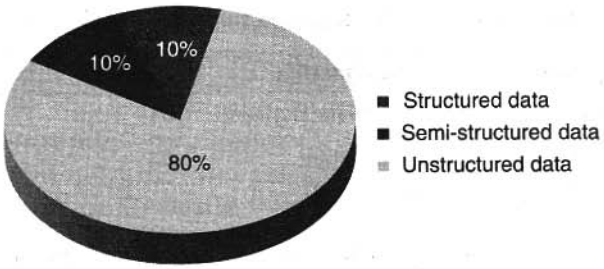


Figure 1.2 Approximate percentage distribution of digital data.

Think structured data, and think data model – a model of the types of business data that we intend to store, process, and access. Let us discuss this in the context of an RDBMS. Most of the structured data is held in RDBMS. An RDBMS conforms to the relational data model wherein the data is stored in rows/columns. Refer Table 1.1.

The number of rows/records/tuples in a relation is called the *cardinality of a relation* and the number of columns is referred to as the *degree of a relation.*

The first step is the design of a relation/table, the fields/columns to store the data, the type of data that will be stored [number (integer or real), alphabets, date, Boolean, etc.]. Next we think of the constraints that we would like our data to conform to (constraints such as UNIQUE values in the column, NOT NULL values in the column, a business constraint such as the value held in the column should not drop below 50, the set of permissible values in the column such as the column should accept only "CS", "IS", "MS", etc., as input).

To explain further, let us design a table/relation structure to store the details of the employees of an enterprise. Table 1.2 shows the structure/schema of an "Employee" table in a RDBMS such as Oracle.

Table 1.2 is an example of a good structured table (complete with table name, meaningful column names with data types, data length, and the relevant constraints) with absolute adherence to relational data model.

Table 1.1 A relation/table with rows and columns

	Column 1	Column 2	Column 3	Column 4
Row 1				

Table 1.2 Schema of an "Employee" table in a RDBMS such as Oracle

Column Name	Data Type	Constraints
EmpNo	Varchar(10)	PRIMARY KEY
EmpName	Varchar(50)	
Designation	Varchar(25)	NOT NULL
DeptNo	Varchar(5)	
ContactNo	Varchar(10)	NOT NULL

Table 1.3 Sample records in the "Employee" table

EmpNo	EmpName	Designation	DeptNo	ContactNo
E101	Allen	Software Engineer	D1	0999999999
E102	Simon	Consultant	D1	0777777777

It goes without saying that each record in the table will have exactly the same structure. Let us take a look at a few records in Table 1.3.

The tables in an RDBMS can also be related. For example, the above "Employee" table is related to the "Department" table on the basis of the common column, "DeptNo". It is not mandatory for the two tables that are related to have exactly the same name for the common column. On the contrary, the two tables are related on the basis of values held within the column, "DeptNo". Given in Figure 1.3 is a depiction of referential integrity constraint (primary – foreign key) with the "Department" table being the referenced table and "Employee" table being the referencing table.

1.1.1.1 Sources of Structured Data

If your data is highly structured, one can look at leveraging any of the available RDBMS [Oracle Corp. – Oracle, IBM – DB2, Microsoft – Microsoft SQL Server, EMC – Greenplum, Teradata – Teradata, MySQL (open source), PostgreSQL (advanced open source), etc.] to house it. Refer Figure 1.4. These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities. In other words, the data of the On-Line Transaction Processing (OLTP) systems are generally quite structured.

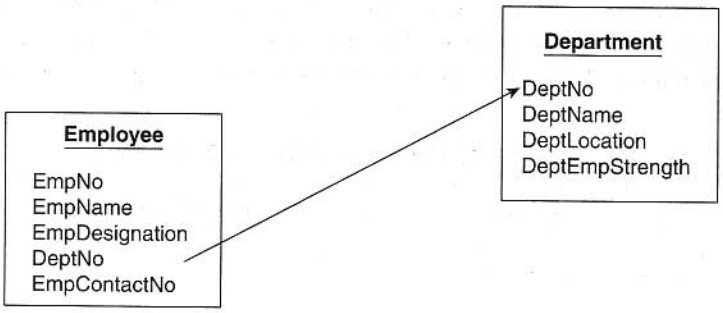


Figure 1.3 Relationship between "Employee" and "Department" tables.

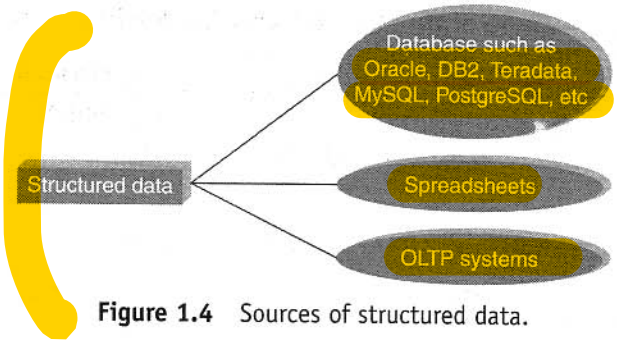


Figure 1.4 Sources of structured data.

1.1.1.2 Ease of Working with Structured Data

Structured data provides the ease of working with it. Refer Figure 1.5. The ease is with respect to the following:

1. **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
2. **Security:** How does one ensure the security of information? There are available staunch encryption and tokenization solutions to warrant the security of information throughout its lifecycle. Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.
3. **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.
4. **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server (increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.).
5. **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction. Given next is a quick explanation of the ACID properties:
   - ***Atomicity:*** A transaction is atomic, means that either it happens in its entirety or none of it at all.
   - ***Consistency:*** The database moves from one consistent state to another consistent state. In other words, if the same piece of information is stored at two or more places, they are in complete agreement.
   - ***Isolation:*** The resource allocation to the transaction happens such that the transaction gets the impression that it is the only transaction happening in isolation.
   - ***Durability:*** All changes made to the database during a transaction are permanent and that accounts for the durability of the transaction.

1.1.2 Semi-Structured Data

Semi-structured data is also referred to as self-describing structure. Refer Figure 1.6. It has the following features:

1. It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
2. It uses tags to segregate semantic elements.

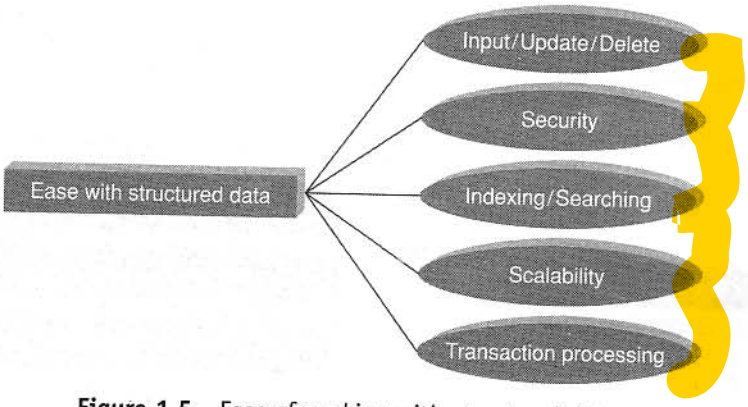


Figure 1.5 Ease of working with structured data.

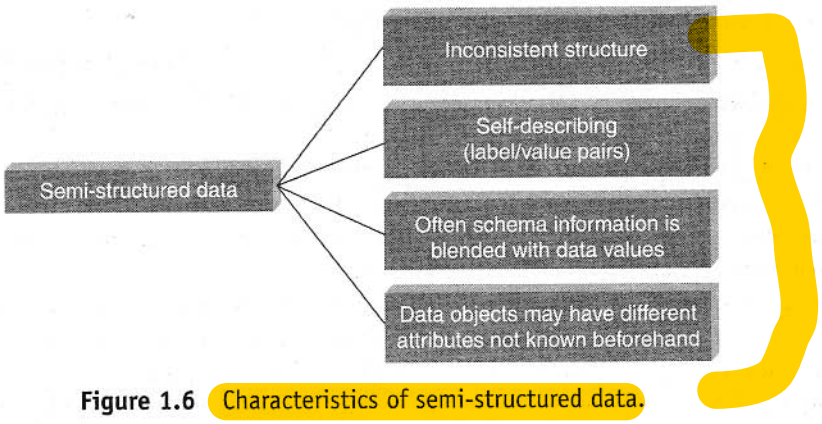


Figure 1.6 Characteristics of semi-structured data.

3. Tags are also used to enforce hierarchies of records and fields within data.
4. There is no separation between the data and the schema. The amount of structure used is dictated by the purpose at hand.
5. In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes. And if at all, they have the same set of attributes, the order of attributes may not be similar and for all practical purposes it is not important as well.

1.1.2.1 Sources of Semi-Structured Data

Amongst the sources for semi-structured data, the front runners are "XML" and "JSON" as depicted in Figure 1.7.

1. **XML:** eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.
2. **JSON:** Java Script Object Notation (JSON) is used to transmit data between a server and a web application. JSON is popularized by web services developed utilizing the Representational State Transfer (REST) – an architecture style for creating scalable web services. MongoDB (open-source, distributed, NoSQL, documented-oriented database) and Couchbase (originally known as Membase, open-source, distributed, NoSQL, document-oriented database) store data natively in JSON format.

An example of HTML is as follows:

```
<HTML>
<HEAD>
<TITLE>Place your title here</TITLE>
</HEAD>
<BODY BGCOLOR="FFFFFF">
```

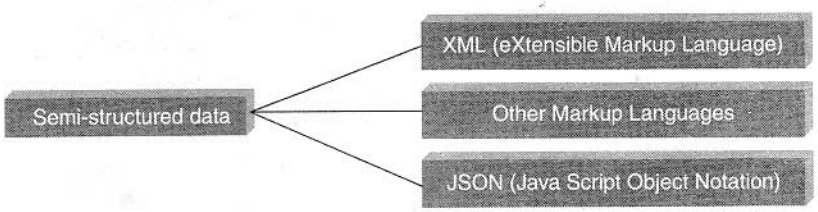


Figure 1.7 Sources of semi-structured data.

```
<CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"></CENTER>
<HR>
<a href="http://bigdatauniversity.com">Link Name</a>
<H1>this is a Header</H1>
<H2>this is a sub Header</H2>
Send me mail at <a href="mailto:support@yourcompany.com">
support@yourcompany.com</a>.
<P>a new paragraph!
<P><B>a new paragraph!</B>
<BR><B><I>this is a new sentence without a paragraph break, in bold italics.</I></B>
<HR>
</BODY>
</HTML>
```

Sample JSON document

```
{
_id:9,
BookTitle: "Fundamentals of Business Analytics",
AuthorName: "Seema Acharya",
Publisher: "Wiley India",
YearofPublication: "2011"
}
```

1.1.3 Unstructured Data

Unstructured data does not conform to any pre-defined data model. In fact, to explain things a little more, let us take a closer look at the various kinds of text available and the possible structure associated with it. As can be seen from the examples quoted in Table 1.4, the structure is quite unpredictable. In Figure 1.8 we look at the other sources of unstructured data.

1.1.3.1 Issues with "Unstructured" Data

Although unstructured data is known NOT to conform to a pre-defined data model or be organized in a pre-defined manner, there are incidents wherein the structure of the data (placed in the unstructured category) can still be implied. As mentioned in Figure 1.9, there could be few other reasons behind placing data in the unstructured category despite it having some structure or being highly structured.

There are situations where people argue that a text file should be in the category of semi-structured data and not unstructured data. Let us look at where they are coming from. Well, the text file does have a name,

Table 1.4 Few examples of disparate unstructured data

Twitter message	Feeling miffed ☹. Victim of twishing.
Facebook post	LOL. C ya. BFN
Log files	127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I; Nav)"
Email	Hey Joan, possible to send across the first cut on the Hadoop chapter by Friday EOD or maybe we can meet up over a cup of coffee. Best regards, Tom

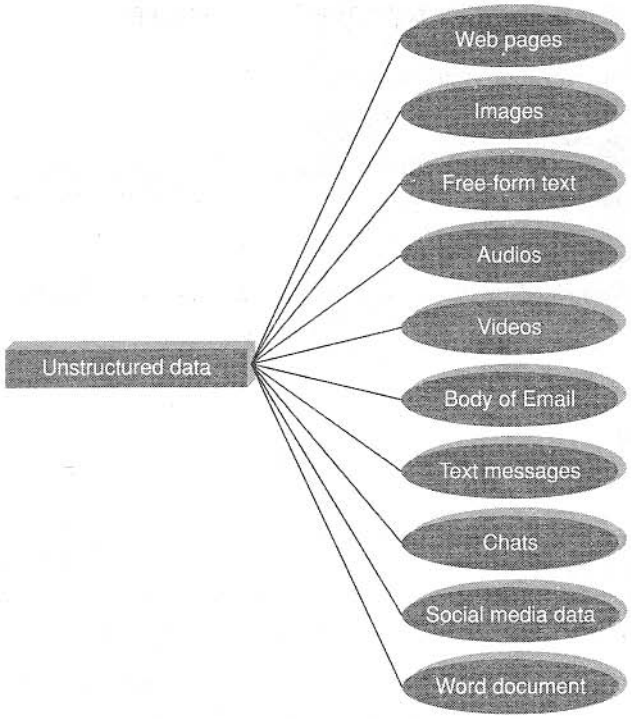


Figure 1.8 Sources of unstructured data.

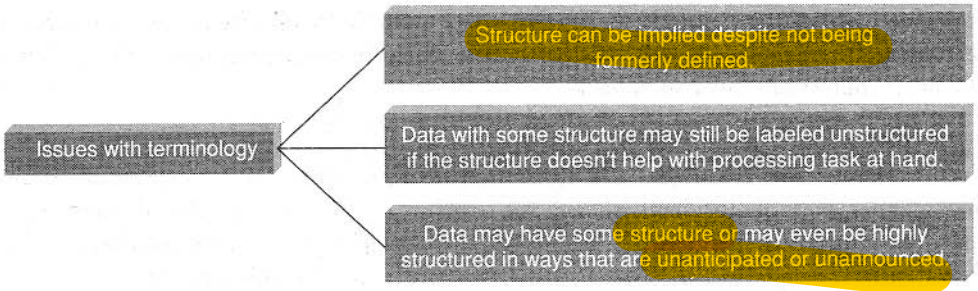


Figure 1.9 Issues with terminology of unstructured data.

one can easily look at the properties to get information such as the owner of the file, the date on which the file was created, the size of the file, etc. Okay, we do have little metadata. But when it comes to analysis, we are more concerned with the content of the text file rather than the name or any of the other properties. In fact, the other properties may not in any way contribute to the processing/analysis task at hand. Therefore, it is fair to place it in the unstructured data category.

1.1.3.2 How to Deal with Unstructured Data?

Today, unstructured data constitutes approximately 80% of the data that is being generated in any enterprise. The balance is clearly shifting in favor of unstructured data as shown in Figure 1.10. It is such a big percentage that it cannot be ignored. Figure 1.11 states a few ways of dealing with unstructured data.

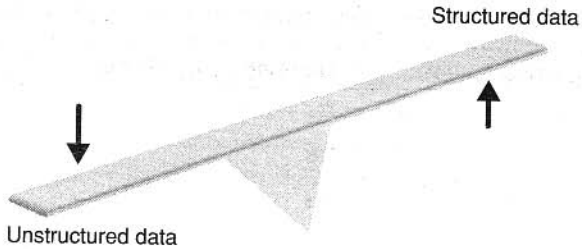


Figure 1.10 Unstructured data clearly constitutes a major percentage of enterprise data.

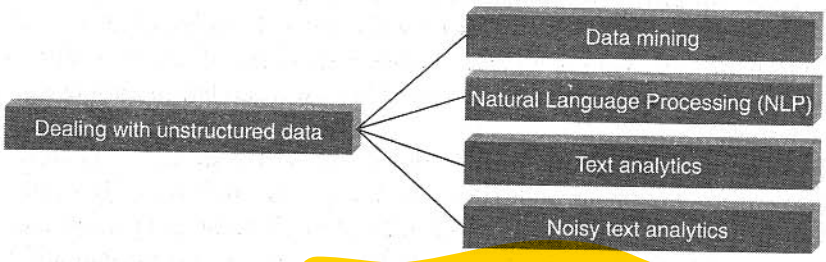


Figure 1.11 Dealing with unstructured data.

The following techniques are used to find patterns in or interpret unstructured data:

1. **Data mining:** First, we deal with large data sets. Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables. It is the analysis step of the "knowledge discovery in databases" process.

   Few popular data mining algorithms are as follows:

   - ***Association rule mining:*** It is also called "market basket analysis" or "affinity analysis". It is used to determine "What goes with what?" It is about when you buy a product, what is the other product that you are likely to purchase with it. For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.
   - ***Regression analysis:*** It helps to predict the relationship between two variables. The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.

PICTURE THIS...

You are interested in purchasing real estate. You have been looking at a few good sites. You have come to the conclusion that cost of the real estate depends on the location (outskirts or prime locale), the amenities provided by the builder (joggers track, senior citizen zone, gymnasium, swimming pools, etc.), the built up area, etc. The cost of the real estate is the dependent variable and the location, amenities, built-up area are called the independent variables.

Table 1.5 Sample records depicting learners' preferences for modes of learning

	Learning using Audios	Learning using Videos	Textual Learners
User 1	Yes	Yes	No
User 2	Yes	Yes	Yes
User 3	Yes	Yes	No
User 4	Yes	?	?

- ***Collaborative filtering:*** It is about predicting a user's preference or preferences based on the preferences of a group of users. For example, take a look at Table 1.5.
  We are looking at predicting whether User 4 will prefer to learn using videos or is a textual learner depending on one or a couple of his or her known preferences. We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

2. **Text analytics or text mining:** Compared to the structured data stored in relational databases, text is largely unstructured, amorphous, and difficult to deal with algorithmically. Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text. It includes tasks such as text categorization, text clustering, sentiment analysis, concept/entity extraction, etc.
3. **Natural language processing (NLP):** It is related to the area of human computer interaction. It is about enabling computers to understand human or natural language input.
4. **Noisy text analytics:** It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc. The noisy unstructured data usually comprises one or more of the following: Spelling mistakes, abbreviations, acronyms, non-standard words, missing punctuation, missing letter case, filler words such as "uh", "um", etc.
5. **Manual tagging with metadata:** This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
6. **Part-of-speech tagging:** It is also called POS or POST or grammatical tagging. It is the process of reading text and tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective", etc.
7. **Unstructured Information Management Architecture (UIMA):** It is an open source platform from IBM. It is used for real-time content analytics. It is about processing text and other unstructured data to find latent meaning and relevant relationship buried therein. Read up more on UIMA at the link: http://www.ibm.com/developerworks/data/downloads/uima/

REMIND ME

- ***Structured data:*** It conforms to a data model. For example, RDBMS conforms to relational data model. It has a pre-defined schema.
- ***Semi-structured data:*** For this format of data, little metadata is available, but is insufficient. Semi-structured data have a self-describing structure. There is little or no separation between data and schema.

- ***Unstructured data:*** This data is growing by the day and growing by leaps and bounds. It has innumerable sources such as human generated (social media data, emails, word documents, presentations, audio and video files that we create and share every day, etc.) and machine generated data (sensors, web server logs, call data records, etc.).

POINT ME (BOOK)

- Chapter 2: Types of Digital Data, "Fundamentals of Business Analytics", Wiley India; Authors – RN Prasad and Seema Acharya, 2011.

CONNECT ME (INTERNET RESOURCES)

- http://data-magnum.com/the-big-deal-about-big-data-whats-inside-structured-unstructured-and-semi-structured-data/
- http://www.webopedia.com/TERM/S/structured_data.html
- http://en.wikipedia.org/wiki/UIMA
- Matching unstructured data and structured data by Bill Inmon: http://www.tdan.com/view-articles/5009
- Semi-structured data analytics: Relational or Hadoop platform? – IBM: http://www.ibmbigdatahub.com/blog/semi-structured-data-analytics-relational-or-hadoop-platform-part-1

TEST ME

A. Place Me in the Basket

Structured	Unstructured	Semi-Structured

Following words are to be placed in the relevant basket:

Email
MS Access
Images
Database
Chat conversations
Relations/Tables
Facebook
Videos
MS Excel
XML

Answer:

Structured	Unstructured	Semi-Structured
MS Access	Email	XML
Database	Images	
Relations/Tables	Chat conversations	
MS Excel	Facebook	
	Videos	

B. Match the Following

1.

Column A	Column B
NLP	Content analytics
Text analytics	Text messages
UIMA	Chats
Noisy unstructured data	Text mining
Data mining	Comprehend human or natural language input
Noisy unstructured data	Uses methods at the intersection of statistics, AI, machine learning & DBs
IBM	UIMA

Answer:

Column A	Column B
NLP	Comprehend human or natural language input
Text analytics	Text mining
UIMA	Content analytics
Noisy unstructured data	Text messages
Data mining	Uses methods at the intersection of statistics, AI, machine learning & DBs
Noisy unstructured data	Chats
IBM	UIMA

2.

Column A	Column B
JSON	SOAP
MongoDB	REST
XML	JSON
Flexible structure	CouchDB
JSON	XML

Answer:

Column A	Column B
JSON	REST
MongoDB	JSON
XML	SOAP
Flexible structure	XML
JSON	CouchDB

C. Solve Me

You are a senior faculty at a premier engineering institute of the city. The Head of the Department has asked you to take a look at the institute's learning website and make a list of the unstructured data that gets generated on the website that can then be stored and analyzed to improve the website to facilitate and enhance the student's learning. You log into the institute's learning website and observe the following features on it:

- Presentation decks (.pdf files)
- Laboratory Manual (.doc files)
- Discussion forum
- Student's blog
- Link to Wikipedia
- A survey questionnaire for the students
- Student's performance sheet downloadable into an .xls sheet
- Student's performance sheet downloadable into a .txt file
- Audio/Video learning files (.wav files)
- .xls sheet having a compiled list of FAQs

From this list, you select the following as sources of unstructured data:

1.
2.
3.
4.
5.
6.

You have just finished making your list when your colleague comes in looking for you. Both of you decide to go away to the cafeteria in the vicinity of the institute's campus. You have forever liked this cafeteria. And you have reasons for the same. There are a couple of machines in the cafeteria's reception area that the customers can use to feed in their orders from a selection of menu items. Once the order is done, you are given a token number. Once your order is ready for serving, the display flashes your token number. It goes without saying

that the billing is also automated. You being in the IT department cannot refrain from thinking about the data that gets collected by these automatic applications. Here's your list:

1.
2.
3.
4.
5.
6.

You are thinking of the analysis that you can perform on this data. Here's your list:

1.
2.
3.
4.
5.
6.

D. Solved Exercises

1. Why is an email placed in the "unstructured category"?

   Answer: Let us take a look at what we can place in the body of the email. We can have any or more of the following:

   - Hyperlink
   - PDFs/DOCs/XLS/etc. attachments
   - Emoticons
   - Images
   - Audio/video attachments
   - Free flowing text, etc.

   The above are reasons behind placing the email in the "unstructured category".

2. What category will you place a CCTV footage into?

   Answer: Unstructured

3. You have just got a book issued from the library. What are the details about the book that can be placed in an RDBMS table?

   Answer:

   - Title of the book
   - Author of the book
   - Publisher of the book
   - Year of Publication
   - No. of pages in the book
   - Type of book such as whether hardbound or paperback
   - Price of the book
   - ISBN No. of the book
   - Attachments such as With CD or Without CD, etc.

4. Which category would you place the consumer complaints and feedback?
   Answer: Unstructured data

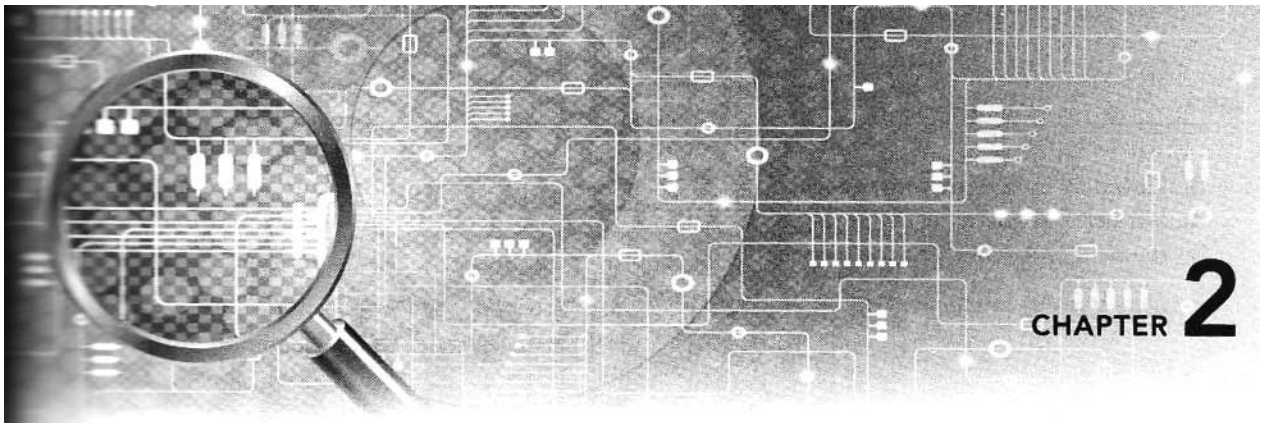
E. Unsolved Exercises

1. Which category (structured, semi-structured, or unstructured) will you place a web page in?
2. What according to you are the challenges with unstructured data?
3. Which category (structured, semi-structured, or unstructured) will you place a PowerPoint presentation in?
4. Which category (structured, semi-structured, or unstructured) will you place a Word Document in?
5. State a few examples of human generated and machine-generated data.

SCENARIO-BASED QUESTION

You are at the university library. You see a few students browsing through the library catalog on a kiosk. You observe the librarians busy at work issuing and returning books. You see a few students fill up the feedback form on the services offered by the library. Quite a few students are learning using the e-learning content.

Think for a while on the different types of data that are being generated in this scenario. Support your answer with logic.



CHAPTER 2

Introduction to Big Data

BRIEF CONTENTS

- What's in Store?
- Characteristics of Data
- Evolution of Big Data
- Definition of Big Data
- Challenges with Big Data
- What is Big Data?
  - Volume
  - Velocity
  - Variety
- Other Characteristics of Data Which are Not Definitional Traits of Big Data
- Why Big Data?
- Are We Just an Information Consumer or Do We Also Produce Information?
- Traditional Business Intelligence (BI) versus Big Data
- A Typical Data Warehouse Environment
- A Typical Hadoop Environment
- What is New Today?
  - Coexistence of Big Data and Data Warehouse
- What is Changing in the Realms of Big Data?

"Data is the new science. Big Data holds the answers."

– Pat Gelsinger, the Chief Executive Officer of VMware, Inc. and former Chief Operating Officer of EMC Corporation

WHAT'S IN STORE?

This chapter focuses on defining and explaining big data. The "Internet of Things" and its widely ultra-connected nature are leading to a burgeoning rise in big data. There is no dearth of data for today's enterprise. On the contrary, they are mired in data and quite deep at that. That brings us to the following questions:

1. Why is it that we cannot forego big data?
2. How has it come to assume such magnanimous importance in running business?

3. How does it compare with the traditional Business Intelligence (BI) environment?
4. Is it here to replace the traditional, relational database management system and data warehouse environment or is it likely to complement their existence?"

Data is widely available. What is scarce is the ability to extract wisdom from it.

Hal Varian, Google's Chief Economist, 2010

PICTURE THIS...

You recently availed the opportunity to attend a virtual classroom session from a leading training institute. You are reflecting back on the experience. Since the session was on big data, it gets you thinking on the types and volume of data that was created before, during, and after the session. It all began with you registering online a week ago for the "Big Data" course. You remember having received an acknowledgment confirming your registration. They had also stated that they will send across some reading contents two days prior to the session. And true to their word, they did. When you logged into the session, you saw that there were 493 other participants. The presenter was introducing the process on smooth learning through the session. During the session, the participants could converse with the presenter as well as with other participants using the chat facility. They had also activated a discussion forum for participants to share their learnings/views/opinions/experiences, etc. There were assignments, which would have to be attempted and submitted on their site. There was an assessment towards the end of the session that was graded. There was a feedback form that was made available at the end of the session to hear back from the participants. They also provided additional reading contents in the form of references to white papers/research papers. The lecture was recorded and made available for better learning and comprehension of the participants.

It was a good experience and you are already thinking of being part of another such experience very soon.

There is no dearth of such virtual classroom sessions being conducted today. There is a huge learning community out there eager to learn. Just think on the volume of data that gets generated, and the variety (the list of attendees, their scores and grades, their chat conversations, their assignments, the polling questions put forth by the instructor to gauge the level of understanding and participation from the learners, etc.) of data that we produce as well consume as we become part of these virtual training sessions.

2.1 CHARACTERISTICS OF DATA

Let us start with the characteristics of data. As depicted in Figure 2.1, data has three key characteristics:

1. **Composition:** The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.
2. **Condition:** The condition of data deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"
3. **Context:** The context of data deals with "Where has this data been generated?" "Why was this data generated?" "How sensitive is this data?" "What are the events associated with this data?" and so on.

Small data (data as it existed prior to the big data revolution) is about certainty. It is about fairly known data sources; it is about no major changes to the composition or context of data.

Most often we have answers to queries like why this data was generated, where and when it was generated, exactly how we would like to use it, what questions will this data be able to answer, and so on. Big data is

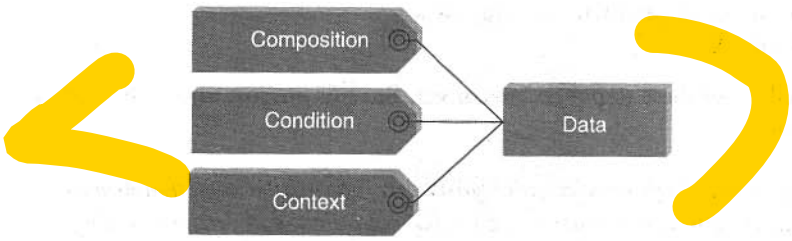


Figure 2.1 Characteristics of data.

about complexity... complexity in terms of multiple and unknown datasets, in terms of exploding volume, in terms of the speed at which the data is being generated and the speed at which it needs to be processed, and in terms of the variety of data (internal or external, behavioral or social) that is being generated.

2.2 EVOLUTION OF BIG DATA

1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia data. Refer Table 2.1.

Table 2.1 The evolution of big data

	Data Generation and Storage	Data Utilization	Data Driven
Complex and Unstructured			Structured data, unstructured data, multimedia data
Complex and Relational		Relational databases: Data-intensive applications	
Primitive and Structured	Mainframes: Basic data storage		
	1970s and before	**Relational (1980s and 1990s)**	**2000s and beyond**

2.3 DEFINITION OF BIG DATA

If we were to ask you the simple question: "Define Big Data", what would your answer be? Well, we will give you a few responses that we have heard over time:

1. Anything beyond the human and technical infrastructure needed to support storage, processing, and analysis.
2. Today's BIG may be tomorrow's NORMAL.

3. Terabytes or petabytes or zettabytes of data.
4. I think it is about 3 Vs.

Refer Figure 2.2. Well, all of these responses are correct. But it is not just one of these; in fact, big data is all of the above and more.

> *Big data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.*
>
> *Source:* Gartner IT Glossary

The 3Vs concept was proposed by the Gartner analyst Doug Laney in a 2001 MetaGroup research publication, titled, *3D Data Management: Controlling Data Volume, Variety and Velocity.*

Source: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

For the sake of easy comprehension, we will look at the definition in three parts. Refer Figure 2.3.

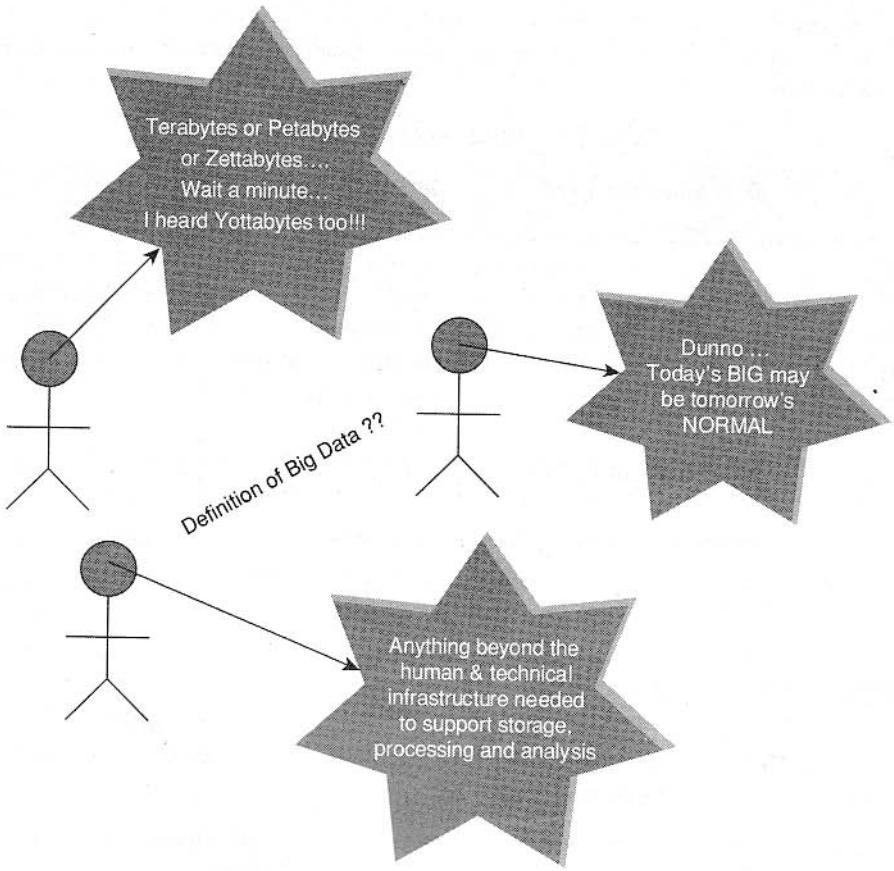


Figure 2.2 Definition of big data.

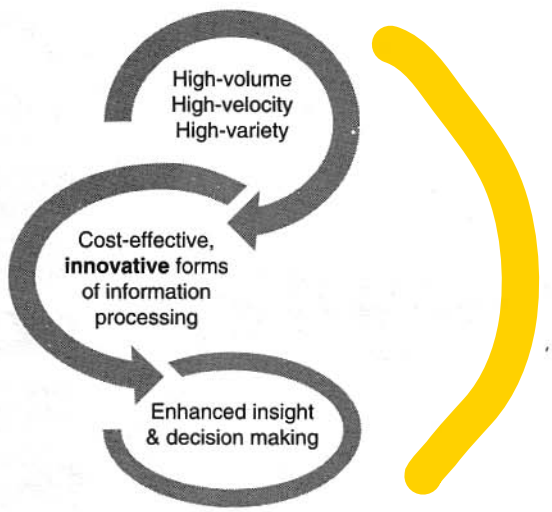


Figure 2.3 Definition of big data – Gartner.

Part I of the definition "big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of structured, semi-structured, and unstructured data) and will require a good speed/pace for storage, preparation, processing, and analysis.

Part II of the definition "cost effective, innovative forms of information processing" talks about embracing new techniques and technologies to capture (ingest), store, process, persist, integrate, and visualize the high-volume, high-velocity, and high-variety data.

Part III of the definition "enhanced insight and decision making" talks about deriving deeper, richer, and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge.

Data → Information → Actionable intelligence → Better decisions → Enhanced business value

2.4 CHALLENGES WITH BIG DATA

Refer Figure 2.4. Following are a few challenges with big data:

1. Data today is growing at an exponential rate. Most of the data that we have today has been generated in the last 2–3 years. This high tide of data will continue to rise incessantly. The key questions here are: "Will all this data be useful for analysis?", "Do we work with all this data or a subset of it?", "How will we separate the knowledge from the noise?", etc.
2. Cloud computing and virtualization are here to stay. Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity, and easy upgrading/downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.
3. The other challenge is to decide on the period of retention of big data. Just how long should one retain this data? A tricky question indeed as some data is useful for making long-term decisions, whereas in few cases, the data may quickly become irrelevant and obsolete just a few hours after having being generated.

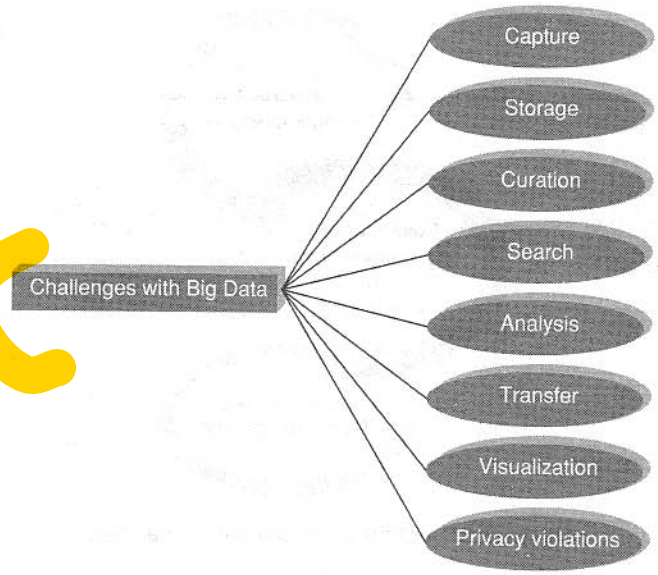


Figure 2.4 Challenges with big data.

4. There is a dearth of skilled professionals who possess a high level of proficiency in data sciences that is vital in implementing big data solutions.
5. Then, of course, there are other challenges with respect to capture, storage, preparation, search, analysis, transfer, security, and visualization of big data. Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how big the dataset should be for it to be considered "big data." Here we are to deal with data that is just too big, moves way to fast, and does not fit the structures of typical database systems. The data changes are highly dynamic and therefore there is a need to ingest this as quickly as possible.
6. Data visualization is becoming popular as a separate discipline. We are short by quite a number, as far as business visualization experts are concerned.

2.5 WHAT IS BIG DATA?

Big data is data that is big in volume, velocity, and variety. Refer Figure 2.5.

2.5.1 Volume

We have seen it grow from bits to bytes to petabytes and exabytes. Refer Table 2.2 and Figure 2.6.

Bits → Bytes → Kilobytes → Megabytes → Gigabytes → Terabytes → Petabytes → Exabytes → Zettabytes → Yottabytes

2.5.1.1 Where Does This Data get Generated?

There are a multitude of sources for big data. An XLS, a DOC, a PDF, etc. is unstructured data; a video on YouTube, a chat conversation on Internet Messenger, a customer feedback form on an online retail website

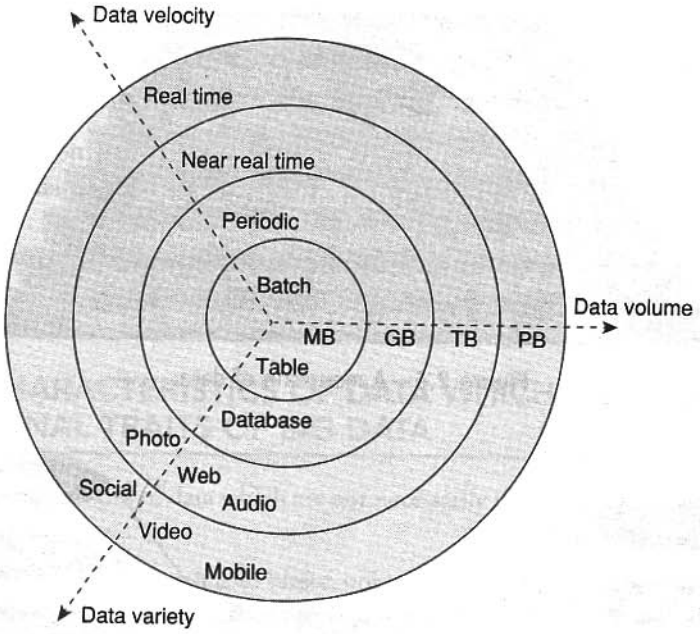


Figure 2.5 Data: Big in volume, variety, and velocity.

Table 2.2 Growth of data

Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

unstructured data; a CCTV coverage, a weather forecast report is unstructured data too. Refer Figure 2.7 for the sources of big data.

1. **Typical internal data sources:** Data present within an organization's firewall. It is as follows:
   - ***Data storage:*** File systems, SQL (RDBMSs – Oracle, MS SQL Server, DB2, MySQL, PostgreSQL, etc.), NoSQL (MongoDB, Cassandra, etc.), and so on.
   - ***Archives:*** Archives of scanned documents, paper archives, customer correspondence records, patients' health records, students' admission records, students' assessment records, and so on.

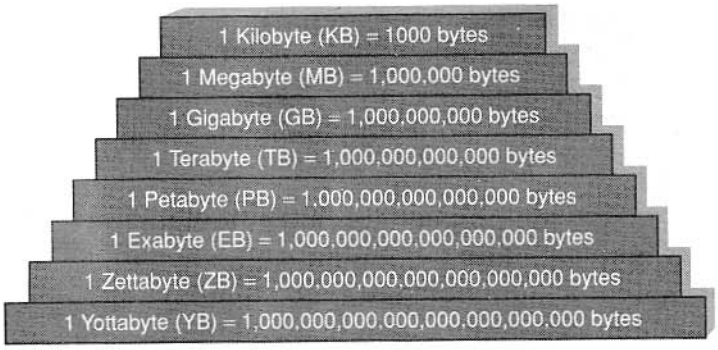


Figure 2.6 A mountain of data.

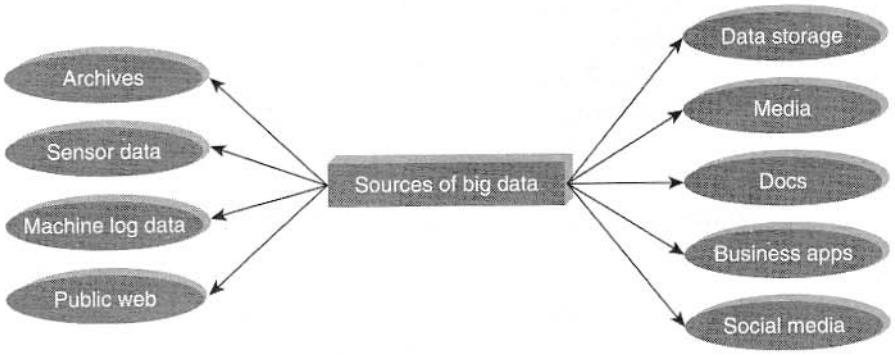


Figure 2.7 Sources of big data.

2. **External data sources:** Data residing outside an organization's firewall. It is as follows:
   - ***Public Web:*** Wikipedia, weather, regulatory, compliance, census, etc.
3. **Both (internal + external data sources)**
   - ***Sensor data:*** Car sensors, smart electric meters, office buildings, air conditioning units, refrigerators, and so on.
   - ***Machine log data:*** Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
   - ***Social media:*** Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
   - ***Business apps:*** ERP, CRM, HR, Google Docs, and so on.
   - ***Media:*** Audio, Video, Image, Podcast, etc.
   - ***Docs:*** Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

2.5.2 Velocity

We have moved from the days of batch processing (remember our payroll applications) to real-time processing.

Batch → Periodic → Near real time → Real-time processing

2.5.3 Variety

Variety deals with a wide range of data types and sources of data. We will study this under three categories: Structured data, semi-structured data and unstructured data.

1. **Structured data:** From traditional transaction processing systems and RDBMS, etc.
2. **Semi-structured data:** For example Hyper Text Markup Language (HTML), eXtensible Markup Language (XML).
3. **Unstructured data:** For example unstructured text documents, audios, videos, emails, photos, PDFs, social media, etc.

2.6 OTHER CHARACTERISTICS OF DATA WHICH ARE NOT DEFINITIONAL TRAITS OF BIG DATA

There are yet other characteristics of data which are not necessarily the definitional traits of big data. Few of these are listed as follows:

1. **Veracity and validity:** *Veracity* refers to biases, noise, and abnormality in data. The key question here is: "Is all the data that is being stored, mined, and analyzed meaningful and pertinent to the problem under consideration?" *Validity* refers to the accuracy and correctness of the data. Any data that is picked up for analysis needs to be accurate. It is not just true about big data alone.
2. **Volatility:** Volatility of data deals with, how long is the data valid? And how long should it be stored? There is some data that is required for long-term decisions and remains valid for longer periods of time. However, there are also pieces of data that quickly become obsolete minutes after their generation.
3. **Variability:** Data flows can be highly inconsistent with periodic peaks.

PICTURE THIS...

An online retailer announces the "big sale day" for a particular week. The retailer is likely to experience an upsurge in customer traffic to the website during this week. In the same way, he/she might experience a slump in his/her business immediately after the festival season. This reemphasizes the point that one might witness spikes in data at some point in time and at other times, the data flow can go flat.

2.7 WHY BIG DATA?

The more data we have for analysis, the greater will be the analytical accuracy and also the greater would be the confidence in our decisions based on these analytical findings. This will entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services, and optimizing existing services. Refer Figure 2.8.

More data → More accurate analysis → Greater confidence in decision making → Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.

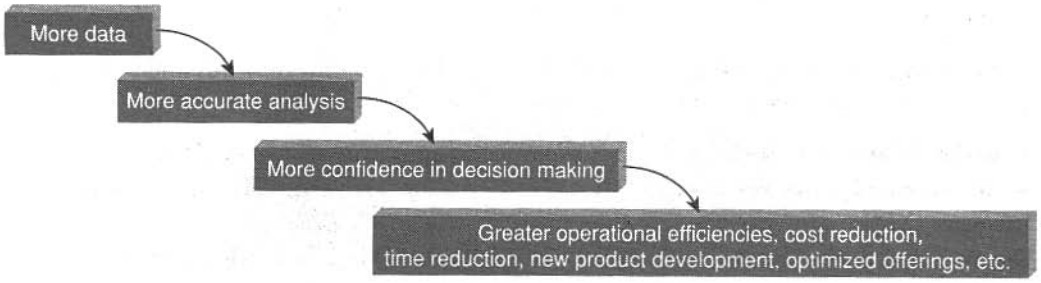


Figure 2.8 Why big data?

2.8 ARE WE JUST AN INFORMATION CONSUMER OR DO WE ALSO PRODUCE INFORMATION?

PICTURE THIS...

You have been invited to your friend's promotion party. You are happy and excited to join your friend at this important milestone in her career. You send in your confirmation through a text message. You get ready and leave for your friend's residence. On the way, you stop at a gas station to refuel. You pay using your credit card. You stop at an upmarket Archie's store to pick a good greeting card and a gift. You get the items billed at the Point of Sale system and pay cash at the counter. While at the party, you click photographs and post it on Facebook, Flickr, and the likes. Within minutes, you start to get likes and comments on your posts.

Mention the places in this scenario where data was generated:

1. Text message to send in the confirmation to attend the promotion bash.
2. Use of credit card to pay for gas/fuel at the gas station.
3. Point of Sale system at Archie's where your transaction gets recorded.
4. Photographs and posts on social networking sites.
5. Likes and comments to your post.

Likewise, there are several instances everyday where you generate data. Think about cases where you are a consumer of information.

2.9 TRADITIONAL BUSINESS INTELLIGENCE (BI) VERSUS BIG DATA

Let us take a sneak peek into some of the differences that one encounters dealing with traditional BI and big data.

1. In traditional BI environment, all the enterprise's data is housed in a central server whereas in a big data environment data resides in a distributed file system. The distributed file system scales by scaling in or out horizontally as compared to typical database server that scales vertically.
2. In traditional BI, data is generally analyzed in an offline mode whereas in big data, it is analyzed in both real time as well as in offline mode.

3. Traditional BI is about structured data and it is here that data is taken to processing functions (move data to code) whereas big data is about variety: Structured, semi-structured, and unstructured data and here the processing functions are taken to the data (move code to data).

2.10 A TYPICAL DATA WAREHOUSE ENVIRONMENT

Let us look at a typical Data Warehouse (DW) environment. Operational or transactional or day-to-day business data is gathered from Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM), legacy systems, and several third party applications. The data from these sources may differ in format [data could have been housed in any RDBMS such as Oracle, MS SQL Server, DB2, MySQL, and Teradata, and so on or in spreadsheet (.xls, .xlsx, etc.) or .csv or txt]. Data may come from data sources located in the same geography or different geographies. This data is then integrated, cleaned up, transformed, and standardized through the process of Extraction, Transformation, and Loading (ETL). The transformed data is then loaded into the enterprise data warehouse (available at the enterprise level) or data marts (available at the business unit/ functional unit or business process level). A host of market leading business intelligence and analytics tools are then used to enable decision making from the use of ad-hoc queries, SQL, enterprise dashboards, data mining, etc. Refer Figure 2.9.

2.11 A TYPICAL HADOOP ENVIRONMENT

Let us now study the Hadoop environment. Is it very different from the data warehouse environment and where exactly is this difference?

As is fairly obvious from Figure 2.10, the data sources are quite disparate from web logs to images, audios, and videos to social media data to the various docs, pdfs, etc. Here the data in focus is not just the data within the company's firewall but also data residing outside the company's firewall. This data is placed in Hadoop Distributed File System (HDFS). If need be, this can be repopulated back to operational systems or fed to the enterprise data warehouse or data marts or Operational Data Store (ODS) to be picked for further processing and analysis.

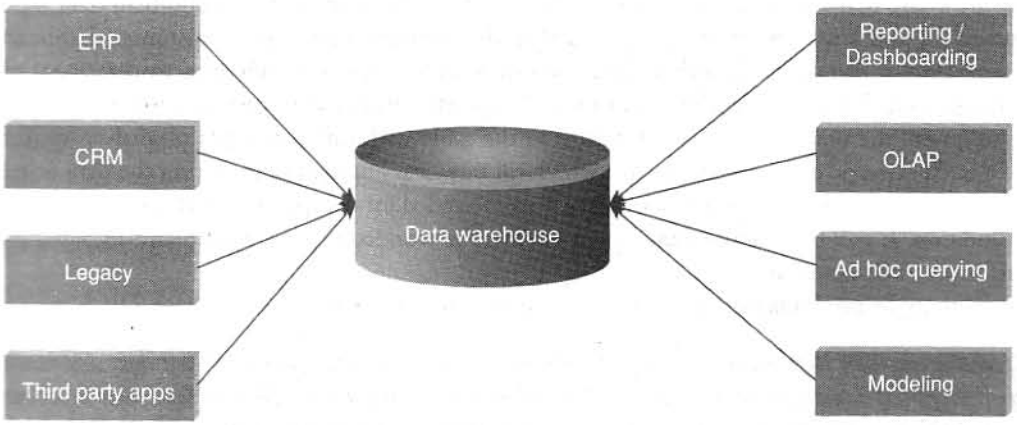


Figure 2.9 A typical data warehouse environment.

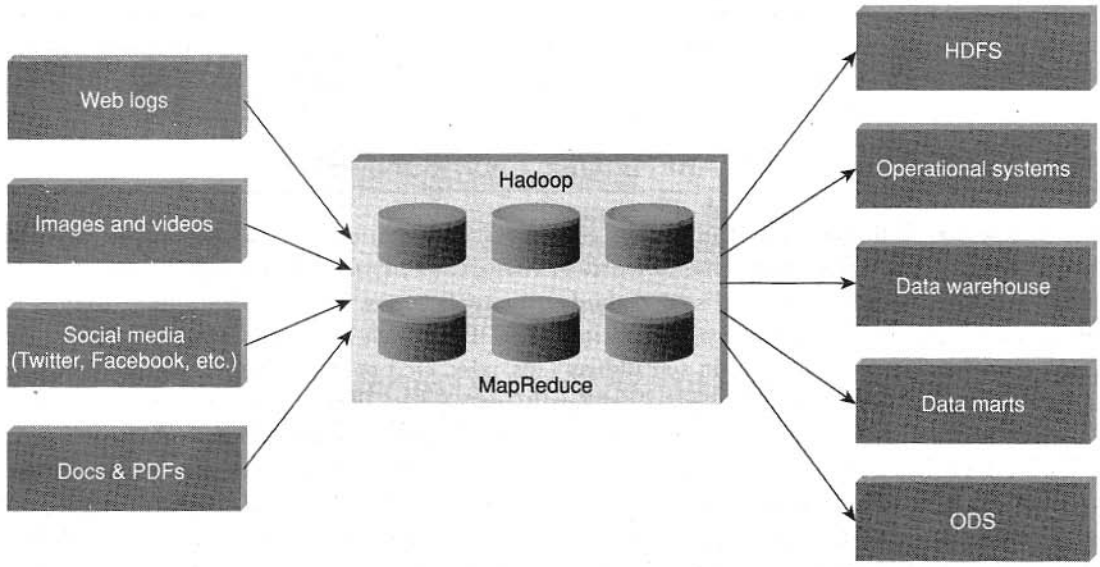


Figure 2.10 A typical Hadoop environment.

2.12 WHAT IS NEW TODAY?

A coexistence strategy that combines the best of legacy data warehouse and analytics environment with the new power of big data solutions is the best of both the worlds. Refer Figure 2.11.

2.12.1 Coexistence of Big Data and Data Warehouse

It is NOT about rip and replace. It will not be possible to get rid of RDBMS or massively parallel processing (MPP), but instead use the right tool for the right job.

As we are aware that few companies are a wee bit comfortable working with incumbent data warehouse for standard BI and analytics reporting, for example the quarterly sales report, customer dashboard, etc. The data warehouse can continue with its standard workload drawing data from legacy operational systems, storing the historical data to provision traditional BI reporting and analytics needs. However, one will not be able to ignore the power that Hadoop brings to the table with different types of analysis on different types of data. The same operational systems, which till now was engaged in powering the data warehouse, can also populate the big data environment when they're needed for computation-rich processing or for raw data exploration. It will be a tight balancing act to steer the workload to the right platform based on what that platform was designed to do.

Here is a thought-provoking piece from Ralph Kimball at a cloudera webinar:

> *"Here's a question that made me laugh a little bit, but it's a serious question: 'Well does this mean that relational databases are going to die?'. I think that there was a sense, three or four years ago, that maybe this was all a giant zero sum game between Hadoop and relational databases, and that has*

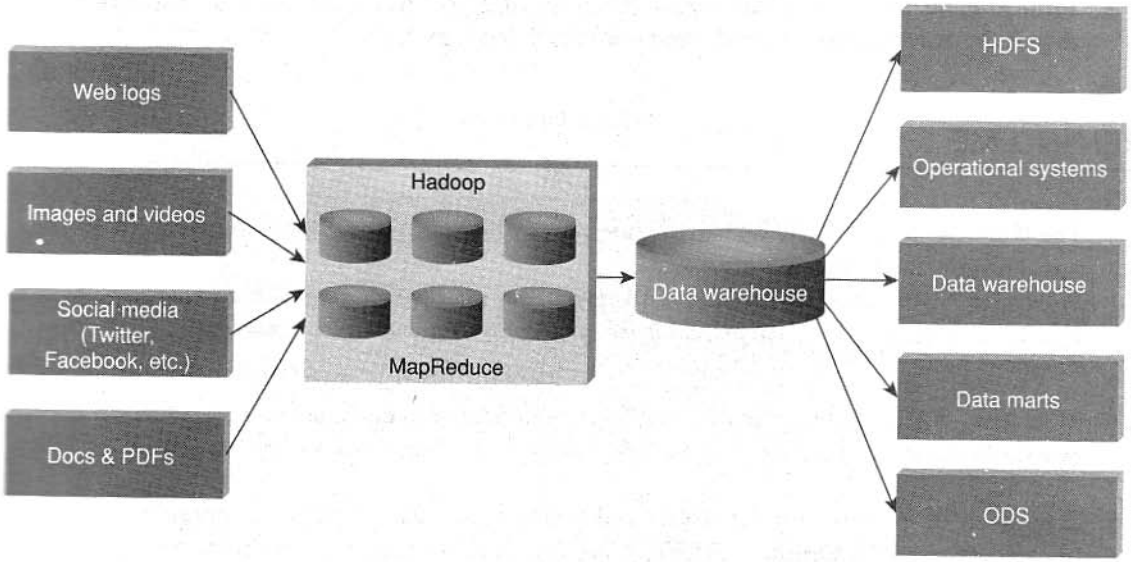


Figure 2.11 Big data and data warehouse coexistence.

simply gone away. Everyone has now realized that there's a huge legacy value in relational databases for the purposes they are used for. Not only transaction processing, but for all the much focused, index-oriented queries on that kind of data, and that will continue in a very robust way forever. Hadoop, therefore, will present this alternative kind of environment for different types of analysis for different kinds of data, and the two of them will coexist. And they will call each other. There may be points at which the business user isn't actually quite sure which one of them they are touching at any point of time."

Just as one cannot ignore the powerful analytics capability of Hadoop, one will not be able to ignore the revolutionary developments in RDBMS such as in-memory processing, etc. The need of the hour is to have both data warehouse and Hadoop co-exist in today's environment.

2.13 WHAT IS CHANGING IN THE REALMS OF BIG DATA?

Gone are the days when IT and business could work in silos and still see the business through. Today, it is an era of a tight handshake between business, IT, and yet another class called *Data Scientists* (more on it in Chapter 3 on "*Big Data Analytics*"). We are citing three very important reasons why companies should compulsorily consider leveraging big data:

1. **Competitive advantage:** The most important resource with any organization today is their data. What they do with it will determine their fate in the market.
2. **Decision making:** Decision making has shifted from the hands of the elite few to the empowered many. Good decisions play a significant role in furthering customer engagement, reducing operating margins in retail, cutting cost and other expenditures in the health sector.

3. **Value of data:** The value of data continues to see a steep rise. As the all-important resource, it is time to look at newer architecture, tools, and practices to leverage this.

REMIND ME

- The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia data.
- *Big data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.*

  Source: Gartner IT Glossary
- More data → More accurate analysis → Greater confidence in decision making → Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.
- Traditional BI is about structured data and it is here that data is taken to processing functions (move data to code). On the other hand, big data is about variety: Structured, semi-structured, and unstructured data and here the processing functions are taken to the data (move code to data).

POINT ME (BOOK)

- Big Data for Dummies – Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, Wiley India Pvt. Ltd.

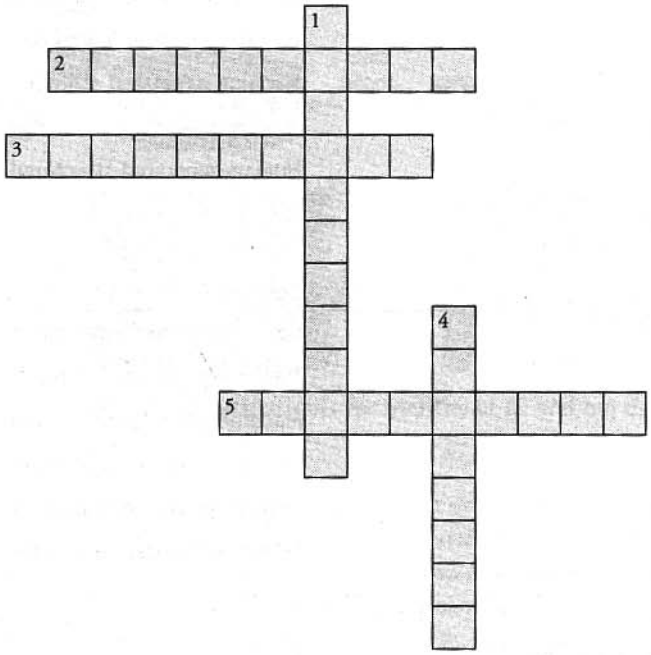
CONNECT ME (INTERNET RESOURCES)

- http://en.wikipedia.org/wiki/Big_data
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- https://www.oracle.com/bigdata/
- http://bigdatauniversity.com/
- http://www.sap.com/solution/big-data/software/overview.html
- http://www.ibm.com/software/data/bigdata/
- http://www.ibm.com/big-data/us/en/
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- http://timoelliott.com/blog/2014/04/no-hadoop-isnt-going-to-replace-your-data-warehouse.html

TEST ME

A. Crossword

Puzzle on Big Data



Across

2. _____, a Gartner analyst coined the term, 'Big Data'
3. _____, is the characteristic of data dealing with its retention.
5. _____, is a large data repository that stores data in its native format until it is needed.

Down

1. _____ characteristic of data explains the spikes in data.
4. Near real time processing or real time processing deals with _____ characteristic of data.

Answer:

Across

2. Doug Laney
3. Volatility
5. Data Lakes

Down

1. Variability
4. Velocity

B. Fill Me

1. Big data is high-volume, high-velocity, and high-variety information assets that demand _____, _____ forms of information processing for enhanced _____ and _____.

Answer: Cost-effective, Innovative, Insight, Decision making

C. Match the Following

Column A	Column B
PostgreSql	Machine generated unstructured data
Scientific data	Open source relational database
Point-of-sale	Human-generated unstructured data
Social Media data	Machine-generated structured data
Gaming-related data	Human-generated unstructured data
Mobile data	Human-generated structured data

Answer:

Column A	Column B
PostgreSql	Open source relational database
Scientific data	Machine generated unstructured data
Point-of-sale	Machine-generated structured data
Social Media data	Human-generated unstructured data
Gaming-related data	Human-generated structured data
Mobile data	Human-generated unstructured data

D. Unsolved Exercises

1. Share your understanding of big data.
2. How is traditional BI environment different from the big data environment?
3. *Big data (Hadoop) will replace the traditional RDBMS and data warehouse.* Comment.
4. Share your experience as a customer on an e-commerce site. Comment on the big data that gets created on a typical e-commerce site.
5. What is your understanding of "Big Data Analytics"?

CHALLENGE ME

1. What is Internet of Things and why does it matter?

   Answer: See http://www.sas.com/en_us/insights/big-data/internet-of-things.html

2. Can the same visualization tool that we run over conventional data warehouse, be used in big data environment?

   Answer: Let us look at Figure 2.12 to understand the solution:

   As per Figure 2.12, structured data is stored in Relational Database Management System (RDBMS) whereas big data (largely unstructured data) is stored in NoSQL databases. Structured data after cleansing, transforming, and converting to a uniform standard format are placed in the enterprise data warehouse

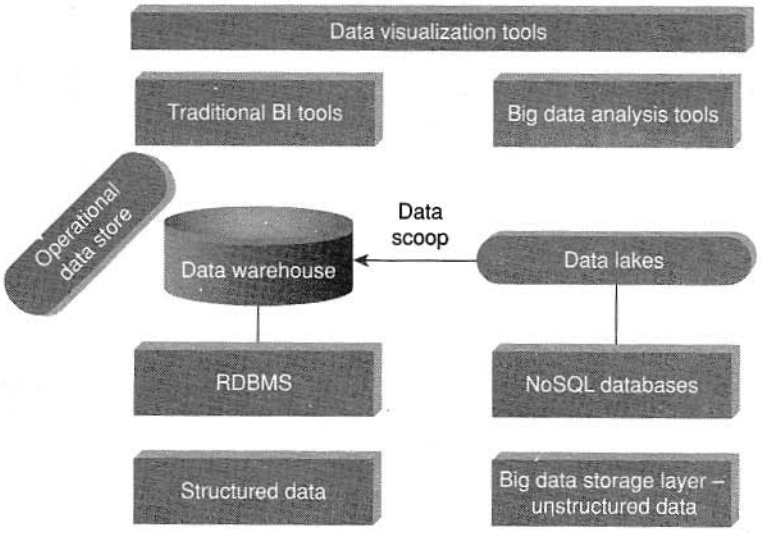
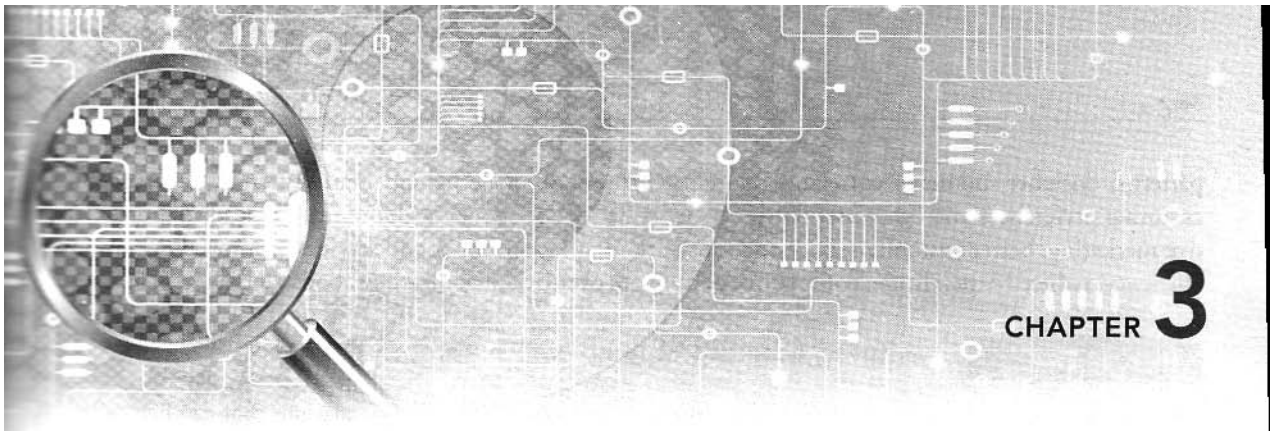


Figure 2.12 Visualization tools for traditional BI and big data.

(at the enterprise level) or the data marts (at the business unit or function level) or operational data stores (almost the complete operational data of an enterprise is housed here) whereas the good variety of data (structured, semi-structured, and unstructured data) is placed in data lakes (a large data repository that stores raw data in its native format until it is needed). Data can then be scooped from data lakes to data warehouses and traditional BI tools can then be run over them. A common set of data visualization tools can then be used to present results after analysis. This goes to emphasize the point, that it makes sense to use the tool that is a specialist for a particular function for example RDBMS for structured data and NoSQL for voluminous data that may be schema less.



CHAPTER 3

Big Data Analytics

BRIEF CONTENTS

- What's in Store?
- Where do we Begin?
- What is Big Data Analytics?
- What Big Data Analytics isn't?
- Why this Sudden Hype Around Big Data Analytics?
- Classification of Analytics
- Greatest Challenges that Prevent Businesses from Capitalizing on Big Data
- Top Challenges Facing Big Data
- Why is Big Data Analytics Important?
- What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?
- Data Science
- Data Scientist … Your New Best Friend!!!
- Terminologies Used in Big Data Environment
  - In-Memory Analytics
  - In-Database Processing
  - Symmetric Multiprocessor System
  - Massively Parallel Processing
  - Difference between Parallel and Distributed Systems
  - Shared Nothing Architecture
  - Consistency, Availability, Partition Tolerance (CAP) Theorem Explained
  - Basically Available Soft State Eventual Consistency (BASE)
- Few Top Analytics Tools

"If you do not know how to ask the right question, you discover nothing."

– W. Edwards Deming

WHAT'S IN STORE?

This chapter is about understanding "Big Data Analytics." We have taken you through the comprehension of the term Big Data – datasets which are voluminous, rich in variety, and calls for processing at a great speed. Big data analytics is the process of examining these large datasets of big data – to unearth hidden

patterns, decipher unknown correlations, understand the rationale behind market trends, and recognize customer preferences and other useful business information. The analytical findings can lead to more effective marketing, better customer service and satisfaction, newer products and services, improved operational efficiency, reduced expenditure, competitive advantages over rival organizations, boosted business gains, etc.

PICTURE THIS...

Scenario 1: You have heard a lot from your friends about the deals on offer on the Amazon site. You decide to register on www.amazon.co.in to avail their discount offers and bumper sales.

A couple of days later, you make a purchase on their site. You landed yourself a good deal by going for books by your favorite author. There is something that does not escape your attention. Amazon has made a few suggestions (of books on similar topics or books by the same author) to you to help with your next or future purchases. You wonder how Amazon's recommendation engine was able to do this for you. Is it something that they do for all their customers?

Well, Amazon's recommendation engine churns out these sort of good suggestions for customers like you, day in and day out. The company gathers all information about your past purchases together with what it knows about you, studies your buying patterns, and the buying patterns of customers like you to arrive at the recommendations that can help with your future purchase. At the core they have big data analytics working for them.

Scenario 2: You are the owner of a trucks transport company. Your company has 500 trucks plying several routes and carrying cargo from one place to another. It is one of those busy days where almost all the trucks are engaged in carrying cargo. You get a call to help with a cargo delivery. They are ready to pay double the charge. You do not want to miss this opportunity. But which truck should you engage. The one that is the nearest but is facing the heaviest traffic or the second nearest one but that is occupied to 75% and will not be able to take more load. There is a need to analyze the truck load, the fuel consumption, the traffic on various routes, etc. before deciding on which truck to select to pick up the new delivery.

3.1 WHERE DO WE BEGIN?

Raw data is collected, classified, and organized. Associating it with adequate metadata and laying bare the context converts this data into meaningful information. It is then aggregated and summarized so that it becomes easy to consume it for analysis. Gradual accumulation of such meaningful information builds a knowledge repository. This, in turn, helps with actionable insights which prove useful for decision making. Refer Figure 3.1.

Organizations have realized that they will not be able to ignore big data if they want to be competitive enough and make those timely decisions to make well of the fleeting opportunities. They will have to analyze

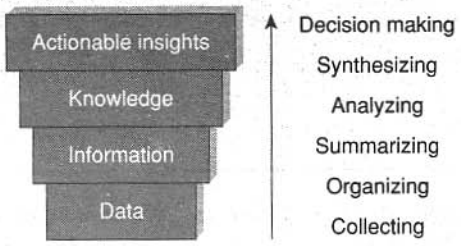


Figure 3.1 Transformation of data to yield actionable insights.

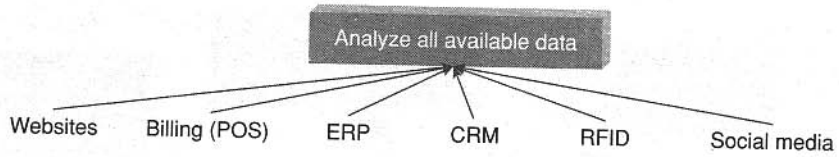


Figure 3.2 Types of unstructured data available for analysis.

big time and also take into consideration big data that makes it to the organization at unprecedented level in terms of volume, velocity, and variety.

Big data analytics is the process of examining big data to uncover patterns, unearth trends, and find unknown correlations and other useful information to make faster and better decisions. Analytics begin with analyzing all available data. Refer Figure 3.2.

3.2 WHAT IS BIG DATA ANALYTICS?

Big Data Analytics is…

1. **Technology-enabled analytics:** Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.
2. About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction, understanding the customer's demographics to cross-sell and up-sell to them, better leveraging the services of your vendors and suppliers, etc.

   Author's experience: The other day I was pleasantly surprised to get a few recommendations via email from one of my frequently visited online retailers. They had recommended clothing line from my favorite brand and also the color suggested was one to my liking. How did they arrive at this? In the recent past, I had been buying clothing line of a particular brand and the color preference was pastel shades. They had it stored in their database and pulled it out while making recommendations to me.
3. About a competitive edge over your competitors by enabling you with findings that allow quicker and better decision-making.
4. A tight handshake between three communities: IT, business users, and data scientists. Refer Figure 3.3.
5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.
6. About moving code to data. This makes perfect sense as the program for distributed processing is tiny (just a few KBs) compared to the data (Terabytes or Petabytes today and likely to be Exabytes or Zettabytes in the near future).

3.3 WHAT BIG DATA ANALYTICS ISN'T?

We have often asked participants of our learning programs as what comes to mind when you hear the term "Big Data." And we are not surprised by the answer… it is "Volume." But now that we have a clear understanding of big data, we know it isn't only about volume but the variety and velocity too are very important factors.

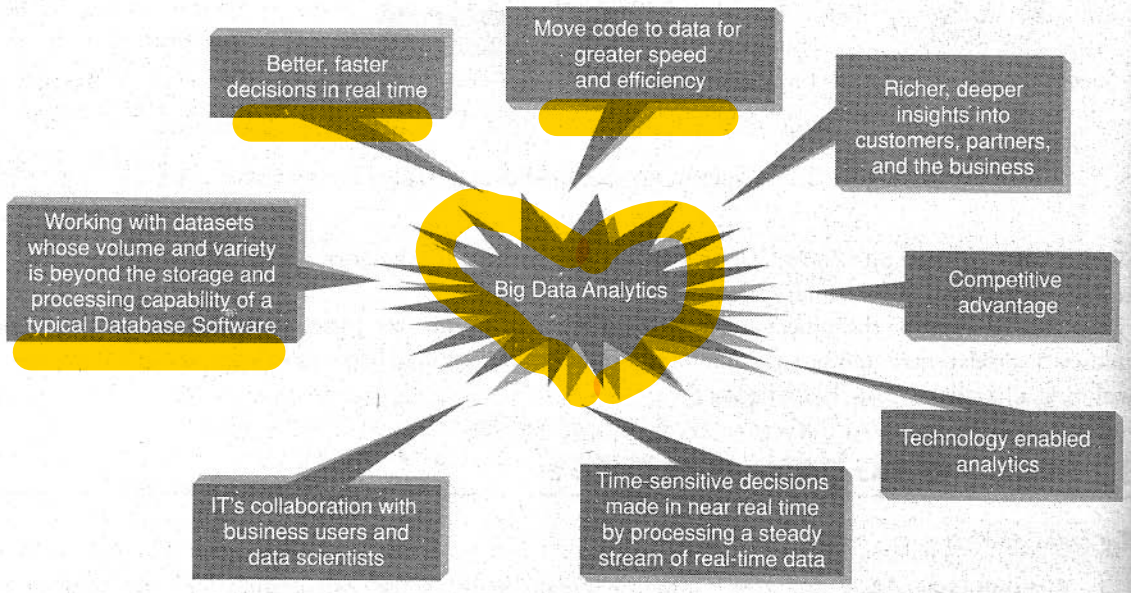


Figure 3.3 What is big data analytics?

Refer Figure 3.4. Big data isn't just about technology. It is about understanding what the data is saying to us. It is about understanding relationships that we thought never existed between datasets. It is about patterns and trends waiting to be unveiled.

And of course, big data analytics is not here to replace our now very robust and powerful Relational Database Management System (RDBMS) or our traditional Data Warehouse. It is here to coexist with both RDBMS and Data Warehouse, leveraging the power of each to yield business value. Big data analytics is not "One-size fits all" traditional RDBMS built on shared disk and memory.

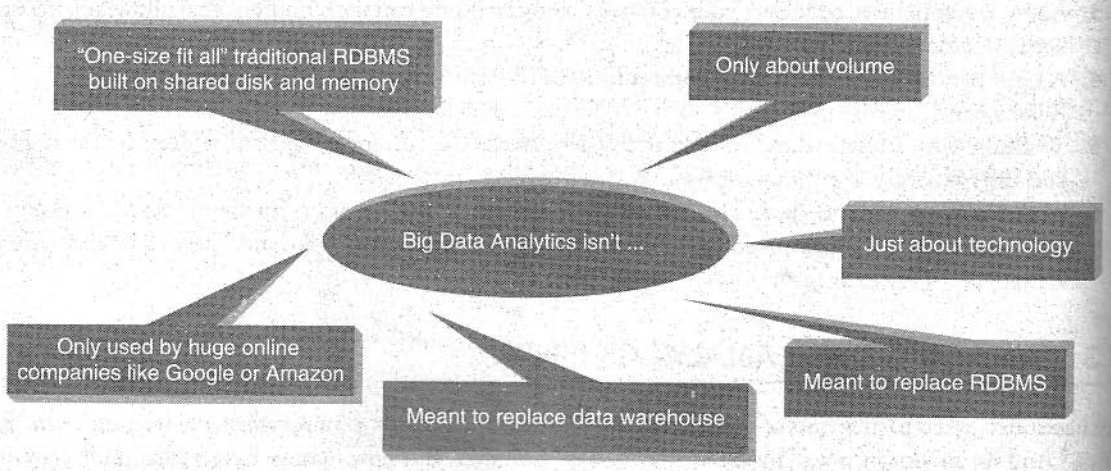


Figure 3.4 What big data analytics isn't?

And before we think it is only used by huge online companies like a Google or Amazon, let us clear the myth. It is for any business and any industry that needs actionable insights out of their data (both internal and external).

3.4 WHY THIS SUDDEN HYPE AROUND BIG DATA ANALYTICS?

If we go by the industry buzz, every place there seems to be talk about big data and big data analytics. Why this sudden hype? Refer Figure 3.5.

Let us put it down to three foremost reasons:

1. Data is growing at a 40% compound annual rate, reaching nearly 45 ZB by 2020. In 2010, almost about 1.2 trillion Gigabyte of data was generated. This amount doubled to 2.4 trillion Gigabyte in 2012 and to about 5 trillion Gigabytes in the year 2014. The volume of business data worldwide is expected to double every 1.2 years. Wal-Mart, the world retailer, processes one million customer transactions per hour. 500 million "tweets" are posted by Twitter users every day. 2.7 billion "Likes" and comments are posted by Facebook users in a day. Every day 2.5 quintillion bytes of data is created, with 90% of the world's data created in the past 2 years alone.

   Source:

   (a) http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

   (b) http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
2. Cost per gigabyte of storage has hugely dropped.
3. There are an overwhelming number of user-friendly analytics tools available in the market today.

3.5 CLASSIFICATION OF ANALYTICS

There are basically two schools of thought:

1. Those that classify analytics into basic, operationalized, advanced, and monetized.
2. Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0.

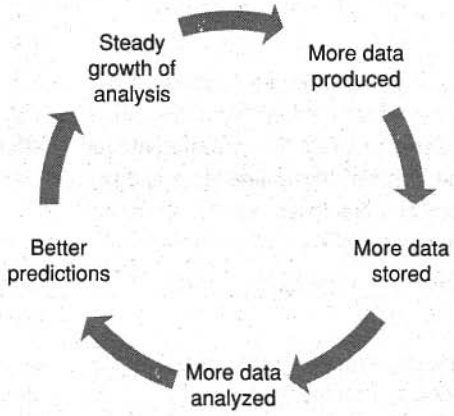


Figure 3.5 What big data entails?

3.5.1 First School of Thought

1. **Basic analytics:** This primarily is slicing and dicing of data to help with basic business insights. This is about reporting on historical data, basic visualization, etc.
2. **Operationalized analytics:** It is operationalized analytics if it gets woven into the enterprise's business processes.
3. **Advanced analytics:** This largely is about forecasting for the future by way of predictive and prescriptive modeling.
4. **Monetized analytics:** This is analytics in use to derive direct business revenue.

3.5.2 Second School of Thought

Let us take a closer look at analytics 1.0, analytics 2.0, and analytics 3.0. Refer Table 3.1.

Table 3.1 Analytics 1.0, 2.0, and 3.0

Analytics 1.0	Analytics 2.0	Analytics 3.0
Era: mid 1950s to 2009	**2005 to 2012**	**2012 to present**
Descriptive statistics (report on events, occurrences, etc. of the past)	Descriptive statistics + predictive statistics (use data from the past to make predictions for the future)	Descriptive + predictive + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situation to one's advantage)
Key questions asked: What happened? Why did it happen?	Key questions asked: What will happen? Why will it happen?	Key questions asked: What will happen? When will it happen? Why will it happen? What should be the action taken to take advantage of what will happen?
Data from legacy systems, ERP, CRM, and 3rd party applications.	Big data	A blend of big data and data from legacy systems, ERP, CRM, and 3rd party applications.
Small and structured data sources. Data stored in enterprise data warehouses or data marts.	Big data is being taken up seriously. Data is mainly unstructured, arriving at a much higher pace. This fast flow of data entailed that the influx of big volume data had to be stored and processed rapidly, often on massive parallel servers running Hadoop.	A blend of big data and traditional analytics to yield insights and offerings with speed and impact.
Data was internally sourced.	Data was often externally sourced.	Data is both being internally and externally sourced.
Relational databases	Database appliances, Hadoop clusters, SQL to Hadoop environments, etc.	In memory analytics, in database processing, agile analytical methods, machine learning techniques, etc.

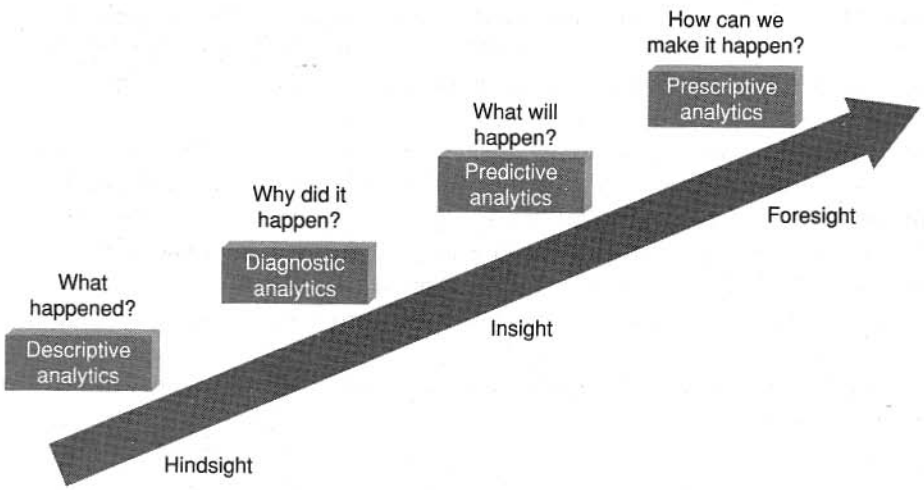


Figure 3.6 Analytics 1.0, 2.0, and 3.0.

Figure 3.6 shows the subtle growth of analytics from Descriptive → Diagnostic → Predictive → Prescriptive analytics.

3.6 GREATEST CHALLENGES THAT PREVENT BUSINESSES FROM CAPITALIZING ON BIG DATA

1. Obtaining executive sponsorships for investments in big data and its related activities (such as training, etc.).
2. Getting the business units to share information across organizational silos.
3. Finding the right skills (business analysts and data scientists) that can manage large amounts of structured, semi-structured, and unstructured data and create insights from it.
4. Determining the approach to scale rapidly and elastically. In other words, the need to address the storage and processing of large volume, velocity, and variety of big data.
5. Deciding whether to use structured or unstructured, internal or external data to make business decisions.
6. Choosing the optimal way to report findings and analysis of big data (visual presentation and analytics) for the presentations to make the most sense.
7. Determining what to do with the insights created from big data.

3.7 TOP CHALLENGES FACING BIG DATA

1. **Scale:** Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should you scale vertically or should you scale horizontally?

2. **Security:** Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information, and other sensitive data.
3. **Schema:** Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are passé.
4. **Continuous availability:** The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.
5. **Consistency:** Should one opt for consistency or eventual consistency?
6. **Partition tolerant:** How to build partition tolerant systems that can take care of both hardware and software failures?
7. **Data quality:** How to maintain data quality – data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

3.8 WHY IS BIG DATA ANALYTICS IMPORTANT?

Let us study the various approaches to analysis of data and what it leads to.

1. **Reactive – Business Intelligence:** What does Business Intelligence (BI) help us with? It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in the right format. It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc. It has support for both pre-specified reports as well as ad hoc querying.
2. **Reactive – Big Data Analytics:** Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.
3. **Proactive – Analytics:** This is to support futuristic decision making by the use of data mining, predictive modeling, text mining, and statistical analysis. This analysis is not on big data as it still uses the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.
4. **Proactive – Big Data Analytics:** This is sieving through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

3.9 WHAT KIND OF TECHNOLOGIES ARE WE LOOKING TOWARD TO HELP MEET THE CHALLENGES POSED BY BIG DATA?

1. The first requirement is of cheap and abundant storage.
2. We need faster processors to help with quicker processing of big data.
3. Affordable open-source, distributed big data platforms, such as Hadoop.
4. Parallel processing, clustering, virtualization, large grid environments (to distribute processing to a number of machines), high connectivity, and high throughputs rather than low latency.
5. Cloud computing and other flexible resource allocation arrangements.

3.10 DATA SCIENCE

Data science is the science of extracting knowledge from data. In other words, it is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques. It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability models, statistical learning, pattern recognition and learning, etc.

Today we have a plethora of use-cases for "Data Science" that are already exploring massive datasets (Peta to Zetta bytes of Information) for weather predictions, oil drillings, seismic activities, financial frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, and so many others beyond standard retail, manufacturing use-cases such as customer churn, market basket analytics (associative mining), collaborative filtering, regression analysis, etc. Data science is multi-disciplinary. Refer to Figure 3.7.

3.10.1 Business Acumen Skills

A data scientist should have the prowess to counter the pressures of business. A firm understanding of business domain further helps. The following is a list of traits that needs to be honed to play the role of data scientist.

1. Understanding of domain.
2. Business strategy.
3. Problem solving.
4. Communication.
5. Presentation.
6. Inquisitiveness.

3.10.2 Technology Expertise

It goes without saying that technology expertise will come in handy if one is to play the role of a data scientist. Cited below are few skills required as far as technical expertise is concerned.

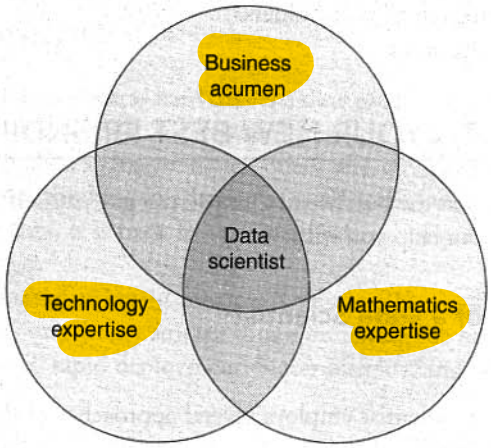


Figure 3.7 Data scientist.

1. Good database knowledge such as RDBMS.
2. Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
3. Programming languages such as Java, Python, C++, etc.
4. Open-source tools such as Hadoop.
5. Data warehousing.
6. Data mining.
7. Visualization such as Tableau, Flare, Google visualization APIs, etc.

3.10.3 Mathematics Expertise

Since the core job of the data scientist will require him to comprehend data, interpret it, make sense of it, and analyze it, he/she will have to dabble in learning algorithms. The following are the key skills that a data scientist will have to have in his arsenal.

1. Mathematics.
2. Statistics.
3. Artificial Intelligence (AI).
4. Algorithms.
5. Machine learning.
6. Pattern recognition.
7. Natural Language Processing.

To sum it up, the data science process is

1. Collecting raw data from multiple disparate data sources.
2. Processing the data.
3. Integrating the data and preparing clean datasets.
4. Engaging in explorative data analysis using model and algorithms.
5. Preparing presentations using data visualizations (commonly called Infographics, or BizAnalytics, or VizAnalytics, etc.)
6. Communicating the findings to all stakeholders.
7. Making faster and better decisions.

3.11 DATA SCIENTIST...YOUR NEW BEST FRIEND!!!

In today's data age, a data scientist is the best friend that you can gift yourself. Refer Figure 3.8 to learn about the tasks that the data scientist can help you with.

3.11.1 Responsibilities of a Data Scientist

Refer Figure 3.8.

1. **Data Management:** A data scientist employs several approaches to develop the relevant datasets for analysis. Raw data is just "RAW," unsuitable for analysis. The data scientist works on it to prepare it to reflect the relationships and contexts. This data then becomes useful for processing and further analysis.

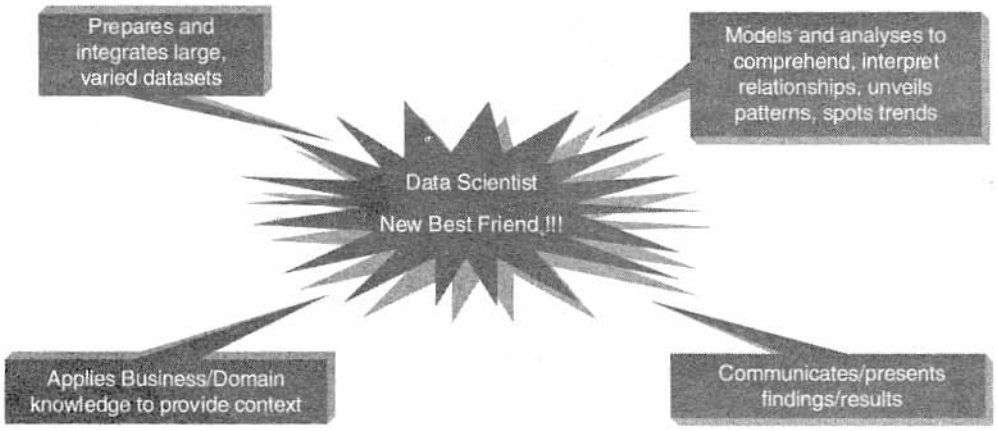


Figure 3.8 Data scientist: your new best friend!!!

2. **Analytical Techniques:** Depending on the business questions which we are trying to find answers to and the type of data available at hand, the data scientist employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and unveil patterns.
3. **Business Analysts:** A data scientist is a business analyst who distinguishes cool facts from insights and is able to apply his business acumen and domain knowledge to see the results in the business context. He is a good presenter and communicator who is able to communicate the results of his findings in a language that is understood by the different business stakeholders.

3.12 TERMINOLOGIES USED IN BIG DATA ENVIRONMENTS

In order to get a good handle on the big data environment, let us get familiar with a few key terminologies in this arena.

3.12.1 In-Memory Analytics

Data access from non-volatile storage such as hard disk is a slow process. The more the data is required to be fetched from hard disk or secondary storage, the slower the process gets. One way to combat this challenge is to pre-process and store data (cubes, aggregate tables, query sets, etc.) so that the CPU has to fetch a small subset of records. But this requires thinking in advance as to what data will be required for analysis. If there is a need for different or more data, it is back to the initial process of pre-computing and storing data or fetching it from secondary storage.

This problem has been addressed using in-memory analytics. Here all the relevant data is stored in Random Access Memory (RAM) or primary storage thus eliminating the need to access the data from hard disk. The advantage is faster access, rapid deployment, better insights, and minimal IT involvement.

3.12.2 In-Database Processing

In-database processing is also called as *in-database analytics*. It works by fusing data warehouses with analytical systems. Typically the data from various enterprise On Line Transaction Processing (OLTP) systems after

cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Data Warehouse (EDW) or data marts. The huge datasets are then exported to analytical programs for complex and extensive computations. With in-database processing, the database program itself can run the computations eliminating the need for export and thereby saving on time. Leading database vendors are offering this feature to large businesses.

3.12.3 Symmetric Multiprocessor System (SMP)

In SMP, there is a single common main memory that is shared by two or more identical processors. The processors have full access to all I/O devices and are controlled by a single operating system instance.

SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus. Refer Figure 3.9.

3.12.4 Massively Parallel Processing

Massive Parallel Processing (MPP) refers to the coordinated processing of programs by a number of processors working parallel. The processors, each have their own operating systems and dedicated memory. They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface. The MPP systems are more difficult to program as the application must be divided in such a way that all the executing segments can communicate with each other. MPP is different from Symmetrically Multiprocessing (SMP) in that SMP works with the processors sharing the same operating system and same memory. SMP is also referred to as *tightly-coupled multiprocessing*.

3.12.5 Difference Between Parallel and Distributed Systems

The next two terms that we discuss are parallel and distributed systems.

As is evident from Figure 3.10, a parallel database system is a tightly coupled system. The processors co-operate for query processing. The user is unaware of the parallelism since he/she has no access to a specific

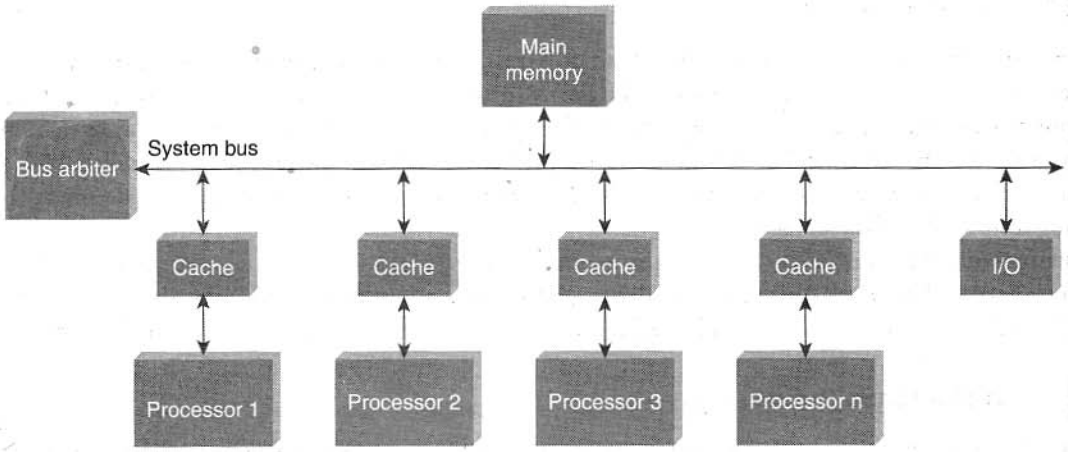


Figure 3.9 Symmetric Multiprocessor System.

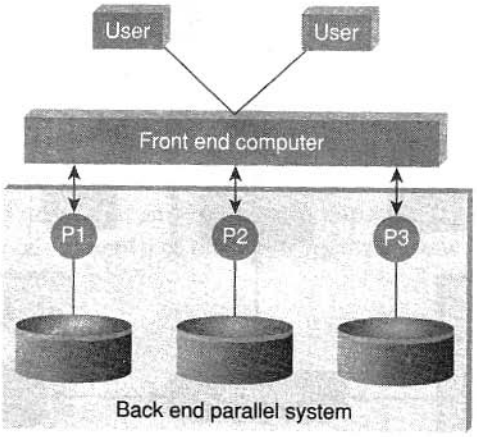


Figure 3.10 Parallel system.

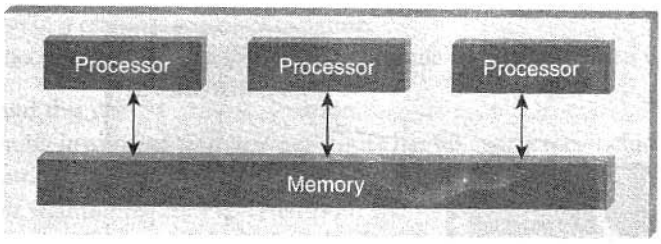


Figure 3.11 Parallel system.

processor of the system. Either the processors have access to a common memory (Refer Fig 3.11) or make use of message passing for communication.

Distributed database systems are known to be loosely coupled and are composed by individual machines. Refer Figure 3.12. Each of the machines can run their individual application and serve their own respective user. The data is usually distributed across several machines, thereby necessitating quite a number of machines to be accessed to answer a user query. Refer Figure 3.13.

3.12.6 Shared Nothing Architecture

Let us look at the three most common types of architecture for multiprocessor high transaction rate systems. They are:

1. Shared Memory (SM).
2. Shared Disk (SD).
3. Shared Nothing (SN).

In shared memory architecture, a common central memory is shared by multiple processors. In shared disk architecture, multiple processors share a common collection of disks while having their own private memory. In shared nothing architecture, neither memory nor disk is shared among multiple processors.

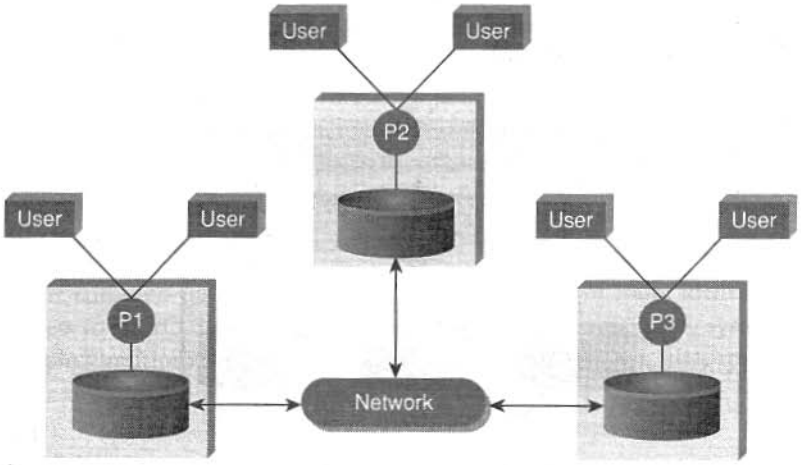


Figure 3.12 Distributed system.

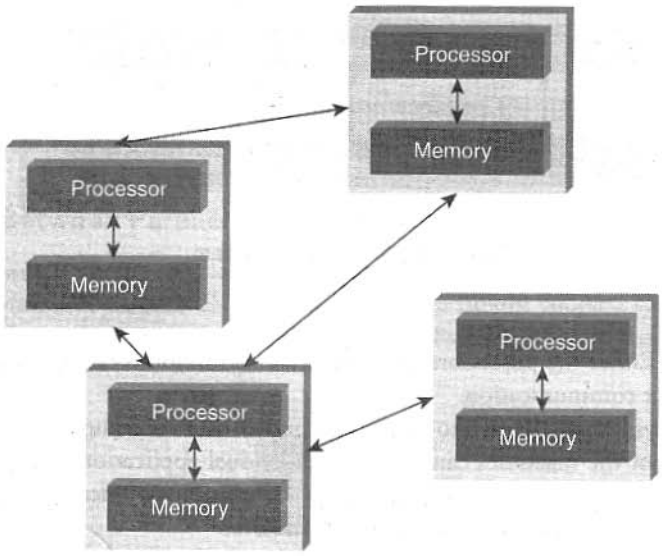


Figure 3.13 Distributed system.

3.12.6.1 Advantages of a "Shared Nothing Architecture"

1. **Fault Isolation:** A "Shared Nothing Architecture" provides the benefit of isolating fault. A fault in a single node is contained and confined to that node exclusively and exposed only through messages (or lack of it).
2. **Scalability:** Assume that the disk is a shared resource. It implies that the controller and the disk bandwidth are also shared. Synchronization will have to be implemented to maintain a consistent shared state. This would mean that different nodes will have to take turns to access the critical data. This

imposes a limit on how many nodes can be added to the distributed shared disk system, thus compromising on scalability.

3.12.7 CAP Theorem Explained

The CAP theorem is also called the *Brewer's Theorem*. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. Refer Figure 3.14. At best you can have two of the following three – one must be sacrificed.

1. Consistency
2. Availability
3. Partition tolerance

3.12.7.1 CAP Theorem

Let us spend some time understanding the earlier mentioned terms.

1. Consistency implies that every read fetches the last write.
2. Availability implies that reads and writes always succeed. In other words, each non-failing node will return a response in a reasonable amount of time.
3. Partition tolerance implies that the system will continue to function when network partition occurs.

Let us try to understand this using a real-life situation.

You work for a training institute, "XYZ." The institute has 50 instructors including you. All of you report to a training coordinator. At the end of the month, all the instructors together with the training coordinator peruse through the training requests received from the various corporate houses and prepare a training schedule for each instructor. These training schedules (one for each instructor) are shared with "Amey," the office administrator. Each morning, you either call the office helpdesk (essentially Amey's desk) or check in-person with Amey for your schedule for the day. In case a training request has been cancelled or updated (updates can be in the form of change in course, change in duration, change of the training timings, etc.), Amey is informed of the updates and the schedules are subsequently updated by him.

Things were good until now. Few corporate houses were your clients and the schedules of each instructor could be smoothly managed without any major hiccups. But your training institute has been implementing promotion campaigns to expand the business. As a result of advertising in the media and word of mouth publicity by your existing clients, you suddenly see an upsurge in training requests from existing and new clients. In consequence of that, more instructors have been recruited. Few trainers/consultants have also been roped in from other training institutes to help tackle the load.

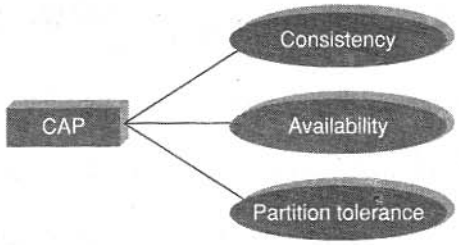


Figure 3.14 Brewer's CAP.

Now when you go to Amey to check your schedule or call in at the helpdesk, you are prepared for a wait in the queue. Looking at the current state of affairs, the training coordinator decides to recruit an additional office administrator "Joey." The helpdesk number will remain the same and will be shared by both the office administrators.

This arrangement works well for a couple of days. Then one day…

You: Hey Amey!

Amey: Hi! How can I help?

You: I think I am scheduled to anchor a training at 3:00 pm today. Can I please have the details?

Amey: Sure! Just a minute.

Amey browses through the file where he maintains the schedules. He does not see a training scheduled against your name at 3:00 pm today and responds back, "You do not have any training to conduct at 3:00 pm."

You: How is that possible? The training coordinator called up yesterday evening to inform of the same and said he has updated the office administrators of the same.

Amey: Oh! Did he say which office administrator? It could have been Joey. Please check with Joey.

Amey: Hey Joey! Please check the schedule for Paul here… Do you see something scheduled at 3:00 pm today?

Joey: Sure enough! He is anchoring the training for client "Z" today at 3:00 pm.

A clear case of inconsistent system!!! The updates in the schedule were shared by the training coordinator with Joey and you were checking for your schedule with Amey.

You share this incident with the training coordinator and that gets him thinking. The issue has to be addressed immediately otherwise it will be difficult to avoid a chaotic situation. He comes up with a plan and shares it with both the office administrators the following day.

Training Coordinator: Folks, each time that either an instructor or me calls any one of you to update a schedule, make sure that both of you update it in your respective files. This way the instructor will always get the most recent and consistent information irrespective of whom amongst the two of you he/she speaks to.

Joey: But that could mean a delay in answering either a phone call or sharing the schedule with the instructor waiting in queue.

Training Coordinator: Yes, I understand. But there is no way that we can give incorrect information.

Amey: There is this other problem as well. Suppose one of us is on leave on a particular day. That would mean that we cannot take any update related calls as we will not be able to simultaneously update both the files (my file and Joey's).

Training Coordinator: Well, good point! *That's the availability problem!!!* But I have thought about that as well. Here is the plan:

1. If one of you receives the update call (any updates to any schedule), ensure that you inform the other person if he is available.
2. In case the other person is not available, ensure that you inform him of all the updates to all schedules via email. It is a must!!!
3. When the other person resumes duty, the first thing he will do is update his file with all the updates to all schedules that he has received via email.

Wow!!! That is sure a Consistent and Available system!!!

Looks like everything is in control. Wait a minute! There is a tiff that has taken place between the office administrators. The two are pretty much available but are not talking to each other which, in other words, means that the updates are not flowing from one to the other. *We have to be partition tolerant!!!* As a training coordinator, you instruct them saying that none of you are taking any calls requesting for schedules or updates to schedules till you patch up. This implies that the system is partition tolerant but not available at that time.

In summary, one can at most decide to go with two of the three.

1. **Consistent:** The instructors or the training coordinator, once they have updated information with you, will always get the most updated information when they call subsequently.
2. **Availability:** The instructors or the training coordinators will always get the schedule if any or both of the office administrators have reported to work.
3. **Partition Tolerance:** Work will go on as usual even if there is communication loss between the office administrators owing to a spat or a tiff!

When to choose consistency over availability and vice-versa…

1. Choose availability over consistency when your business requirements allow some flexibility around when the data in the system synchronizes.
2. Choose consistency over availability when your business requirements demand atomic reads and writes.

Examples of databases that follow one of the possible three combinations

1. Availability and Partition Tolerance (AP)
2. Consistency and Partition Tolerance (CP)
3. Consistency and Availability (CA)

NEW HORIZON COLLEGE OF ENGINEERING ★ Bengaluru-560 103 ★ LIBRARY Acc. No. 0134325 Call No.:

Refer Figure 3.15 to get a glimpse of databases that adhere to two of the three characteristics of CAP theorem.

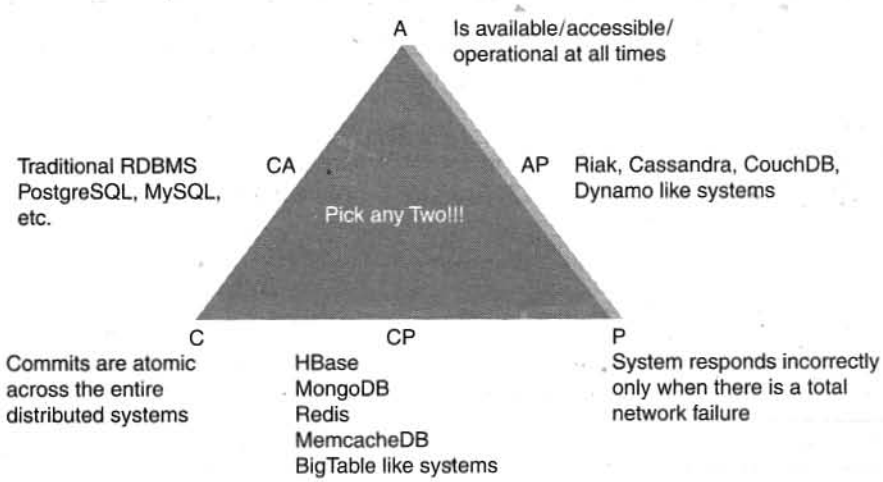


Figure 3.15 Databases and CAP.

3.13 BASICALLY AVAILABLE SOFT STATE EVENTUAL CONSISTENCY (BASE)

A few basic questions to start with:

1. ***Where is it used?***
   In distributed computing.
2. ***Why is it used?***
   To achieve high availability.
3. ***How is it achieved?***
   Assume a given data item. If no new updates are made to this given data item for a stipulated period of time, eventually all accesses to this data item will return the updated value. In other words, if no new updates are made to a given data item for a stipulated period of time, all updates that were made in the past and not applied to this given data item and the several replicas of it will percolate to this data item so that it stays as current/recent as is possible.
4. ***What is replica convergence?***
   A system that has achieved eventual consistency is said to have converged or achieved *replica convergence*.
5. ***Conflict resolution*: How is the conflict resolved?**
   (a) **Read repair:** If the read leads to discrepancy or inconsistency, a correction is initiated. It slows down the read operation.
   (b) **Write repair:** If the write leads to discrepancy or inconsistency, a correction is initiated. This will cause the write operation to slow down.
   (c) **Asynchronous repair:** Here, the correction is not part of a read or write operation.

3.14 FEW TOP ANALYTICS TOOLS

There is no dearth of analytical tools in the market. Please find below our list of few top analytics tools. We have also provided the links after each tool for you to explore more…

1. MS Excel
   https://support.office.microsoft.com/en-in/article/Whats-new-in-Excel-2013-1cbc42cd-bfaf-43d7-9031-5688ef1392fd?CorrelationId=1a2171cc-191f-47de-8a55-08a5f2e9c739&ui=en-US&rs=en-IN&ad=IN
2. SAS
   http://www.sas.com/en_us/home.html
3. IBM SPSS Modeler
   http://www-01.ibm.com/software/analytics/spss/products/modeler/
4. Statistica
   http://www.statsoft.com/

5. Salford systems (World Programming Systems)
   http://www.salford-systems.com/
6. WPS
   http://www.teamwpc.co.uk/products/wps

3.14.1 Open Source Analytics Tools

Let us look at a couple of open source analytics tools. We have also provided the links after each tool for you to explore more…

1. R analytics
   http://www.revolutionanalytics.com/
2. Weka
   http://www.cs.waikato.ac.nz/ml/weka/

REMIND ME

- Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.
- Big data analytics is a about a tight handshake between three communities: IT, business users, and data scientists.
- *Data science* is the science of extracting knowledge from data.
- The CAP theorem is also called the Brewer's Theorem. It states that in a distributed computing environment (a collection of interconnected nodes that share data), it is impossible to provide the following guarantees. At best you can have two of the following three – one must be sacrificed.
  - Consistency
  - Availability
  - Partition tolerance

CONNECT ME (INTERNET RESOURCES)

- http://en.wikipedia.org/wiki/Data_science
- http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/
- http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html
- http://spotfire.tibco.com/blog/?p=4240
- http://reports.informationweek.com/abstract/106/1255/Financial/tech-center-taking-advantage-of-in-memory-analytics.html
- http://www.informationweek.com/software/information-management/oracle-analytics-package-expands-in-database-processing-options/d/d-id/1102712?

TEST ME

A. Fill Me

1. The _____ technology helps query data that resides in a computer's random access memory (RAM) rather than data stored on physical disks.
2. Eventual consistency is a consistency model used in distributed computing to achieve high _____.
3. A coordinated processing of a program by multiple processors, each working on different parts of the program and using its own operating system and memory is called _____.
4. A collection of independent computers that appear to its users as a single coherent system is _____.

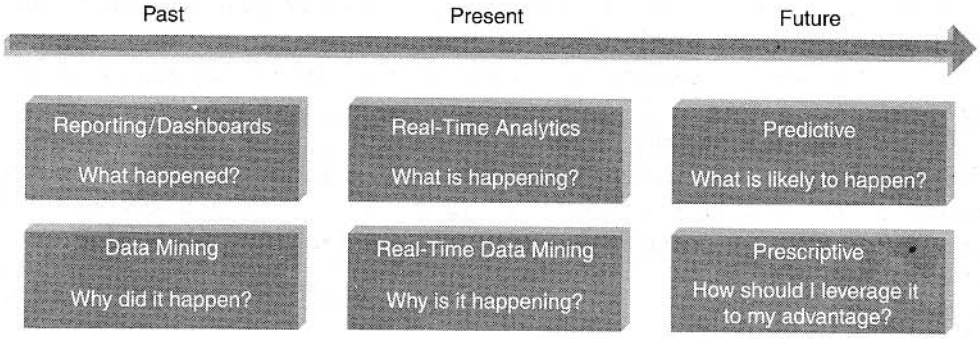
Answers:

1. In-memory analytics
2. Availability
3. Massively parallel processing
4. Distributed systems

B. Answer Me

1. What are the various types of analytics?

Answer:



2. What are the key questions to be answered by all organizations stepping into analytics?

Answer: The key questions for any organization stepping into analytics are:

- Should you be storing all of your big data? If "Yes", where are you going to store it? If "No", how will you know what to store and what to discard?
- How will you sieve through your massive data to filter out the relevant from the irrelevant?
- How long will you store this data?
- How will you accommodate the peaks (variability in terms of data influx) in your data?
- How will you analyze? Will you analyze all the data that is stored or analyze a sample?
- What will you do with the insights generated from this analysis?

3. What can one expect from analytics 3.0?

Answer:

- In-memory analytics.
- In-database processing.
- Leveraging analytics to improve operational, tactical, and strategic decision making.

- Coupling the in-memory analytics and in-database processing with agile analytical methods and machine learning techniques.
- Appropriate tools to effectively support decision-making at the front lines, such as mobile and self-serve analytical applications.

4. Which industries will be affected most by analytics 3.0? Who will benefit the most?

Answer: Almost all the firms in all the industries and not just online firms will be affected by analytics 3.0. A lot of analytics have already been done in the Transport, Retail, and Banking sector. Telecom, entertainment, and health sectors have a bit of catching up to do.

5. What is predictive and prescriptive analytics?

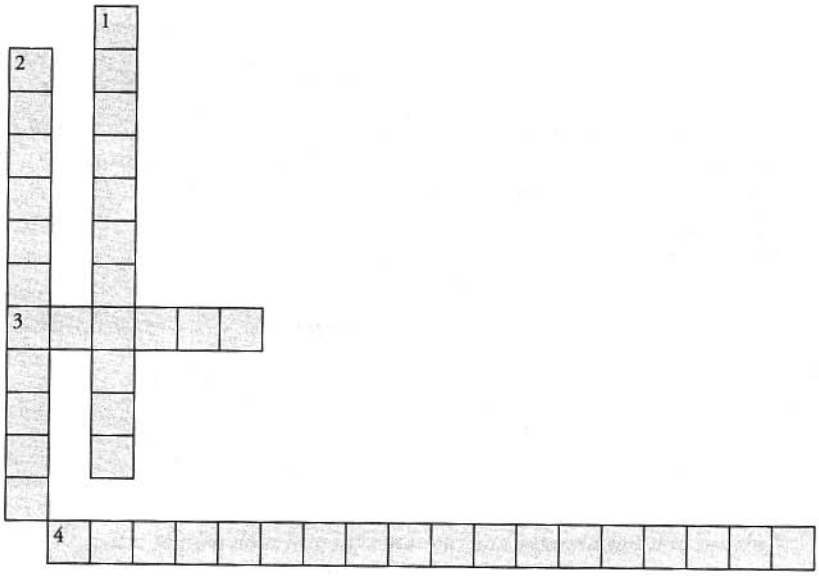
Answer:

Predictive analytics helps you answer the questions: "What will happen?" and "Why will it happen"?

Prescriptive analytics goes beyond "What will happen?" "Why will it happen?" and "When will it happen?" to answer "What should be the action taken to take advantage of what will happen"?

C. Crossword

1. Puzzle on CAP Theorem



Across

3. CAP theorem is also called as _____ theorem.
4. System will continue to function even when network partition occurs.

Down

1. Every read fetches the most recent write.
2. A non-failing node will return a reasonable response within a reasonable amount of time.

Solution:

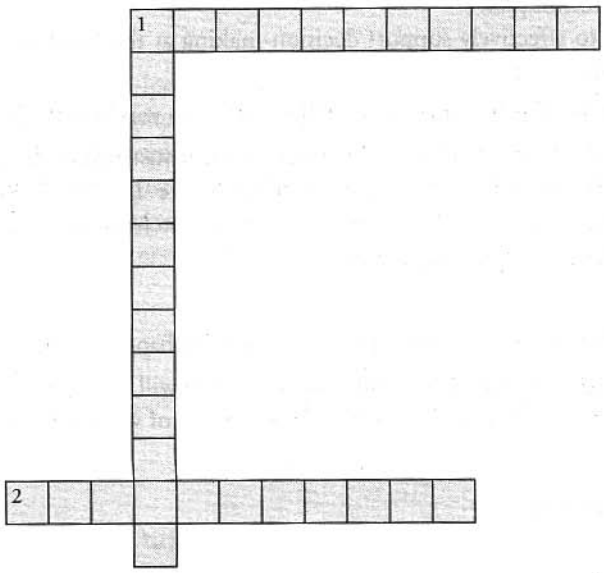
Across

3. Brewer
4. Partition Tolerant

Down

1. Consistency
2. Availability

2. Puzzle on Architecture



Across

1. _____ is an important advantage of shared nothing architecture.
2. In this architecture, multiple processors have their own private memory.

Down

1. In this architecture, central memory is shared by multiple processors.

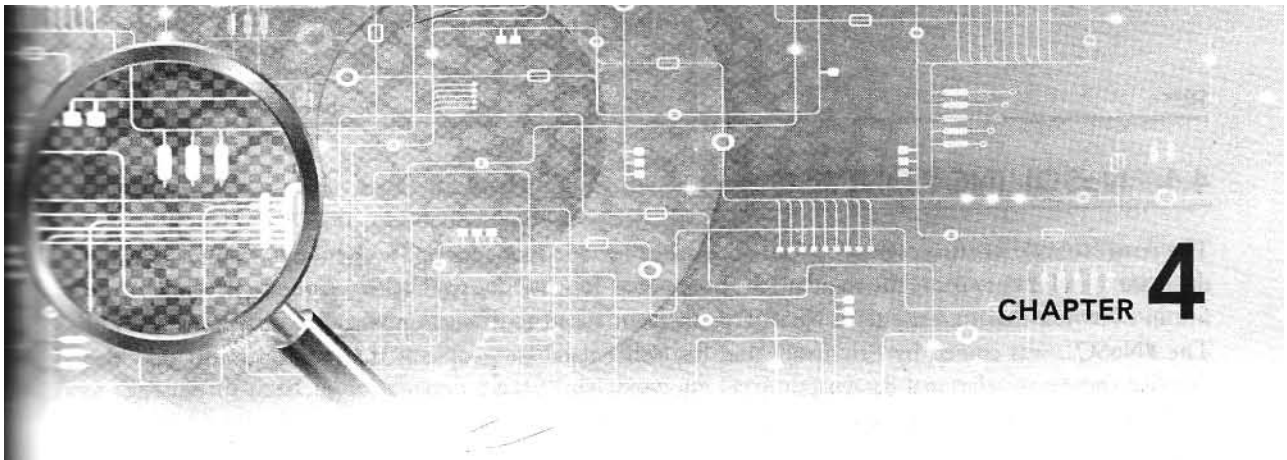
Answer:

Across

1. Scalability
2. Shared Disk

Down

1. Shared Memory



CHAPTER 4

The Big Data Technology Landscape

BRIEF CONTENTS

- What's in Store?
- NoSQL (Not Only SQL)
  - Where is it used?
  - What is it?
  - Types of NoSQL Databases
  - Why NoSQL?
  - Advantages of NoSQL
  - What we miss with NoSQL?
  - Use of NoSQL in Industry
  - NoSQL Vendors
  - SQL versus NoSQL
  - NewSQL
  - Comparison of SQL, NoSQL, and NewSQL
- Hadoop
  - Features of Hadoop
  - Key Advantages of Hadoop
  - Versions of Hadoop
    - Hadoop 1.0
    - Hadoop 2.0
- Overview of Hadoop Ecosystems
- Hadoop Distributions
- Hadoop versus SQL
- Integrated Hadoop Systems Offered by Leading Market Vendors
- Cloud-Based Hadoop Solutions

"The goal is to turn data into information, and information into insight."

– Carly Fiorina, former CEO, Hewlett-Packard Co

WHAT'S IN STORE?

The focus of this chapter is on understanding "big data technology landscape". This chapter is an overview on NoSQL and Hadoop. There are separate chapters on NoSQL (MongoDB and Cassandra) as well as Hadoop in the book.

The big data technology landscape can be majorly studied under two important technologies:

1. NoSQL
2. Hadoop

4.1 NoSQL (NOT ONLY SQL)

The term NoSQL was first coined by Carlo Strozzi in 1998 to name his lightweight, open-source, relational database that did not expose the standard SQL interface. Johan Oskarsson, who was then a developer at last.fm, in 2009 reintroduced the term NoSQL at an event called to discuss open-source distributed network. The #NoSQL was coined by Eric Evans and few other database people at the event found it suitable to describe these non-relational databases.

Few features of NoSQL databases are as follows:

1. They are open source.
2. They are non-relational.
3. They are distributed.
4. They are schema-less.
5. They are cluster friendly.
6. They are born out of 21st century web applications.

4.1.1 Where is it Used?

NoSQL databases are widely used in big data and other real-time web applications. Refer Figure 4.1. NoSQL databases is used to stock log data which can then be pulled for analysis. Likewise it is used to store social media data and all such data which cannot be stored and analyzed comfortably in RDBMS.

4.1.2 What is it?

NoSQL stands for Not Only SQL. These are non-relational, open source, distributed databases. They are hugely popular today owing to their ability to scale out or scale horizontally and the adeptness at dealing with a rich variety of data: structured, semi-structured and unstructured data. Refer Figure 4.2 for additional features of NoSQL. NoSQL databases.

1. **Are non-relational:** They do not adhere to relational data model. In fact, they are either key–value pairs or document-oriented or column-oriented or graph-based databases.

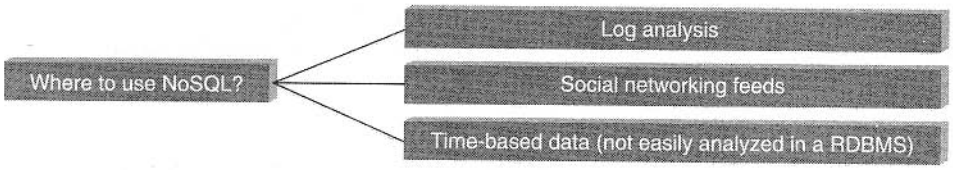


Figure 4.1 Where to use NoSQL?

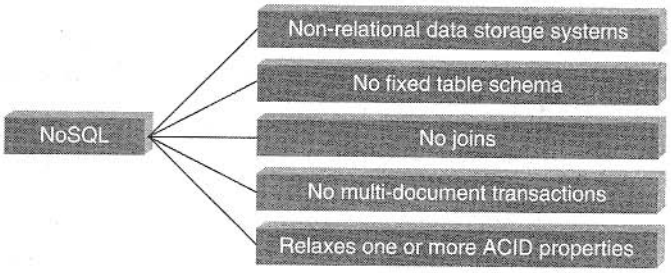


Figure 4.2 What is NoSQL?

2. **Are distributed:** They are distributed meaning the data is distributed across several nodes in a cluster constituted of low-cost commodity hardware.
3. **Offer no support for ACID properties (Atomicity, Consistency, Isolation, and Durability):** They do not offer support for ACID properties of transactions. On the contrary, they have adherence to Brewer's CAP (Consistency, Availability, and Partition tolerance) theorem and are often seen compromising on consistency in favor of availability and partition tolerance.
4. **Provide no fixed table schema:** NoSQL databases are becoming increasing popular owing to their support for flexibility to the schema. They do not mandate for the data to strictly adhere to any schema structure at the time of storage.

4.1.3 Types of NoSQL Databases

We have already stated that NoSQL databases are non-relational. They can be broadly classified into the following:

1. Key–value or the big hash table.
2. Schema-less.

Refer Figure 4.3. Let us take a closer look at key–value and few other types of schema-less databases:

1. **Key–value:** It maintains a big hash table of keys and values. For example, Dynamo, Redis, Riak, etc.
   Sample Key-Value Pair in Key-Value Database

Key	Value
First Name	Simmonds
Last Name	David

2. **Document:** It maintains data in collections constituted of documents. For example, MongoDB, Apache CouchDB, Couchbase, MarkLogic, etc.
   Sample Document in Document Database

```
{
"Book Name":    "Fundamentals of Business Analytics",
"Publisher":    "Wiley India",
"Year of Publication":    "2011"
}
```

3. **Column:** Each storage block has data from only one column. For example: Cassandra, HBase, etc.

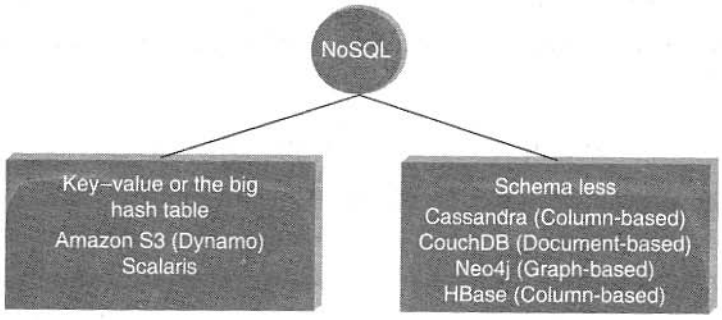
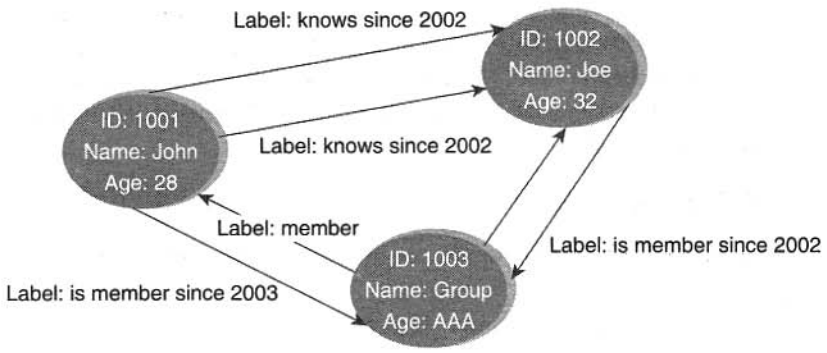


Figure 4.3 Types of NoSQL databases.

4. **Graph:** They are also called network database. A graph stores data in nodes. For example, Neo4j, HyperGraphDB, etc.
   Sample Graph in Graph Database



Refer Table 4.1 for popular schema-less databases.

4.1.4 Why NoSQL?

1. It has scale out architecture instead of the monolithic architecture of relational databases.
2. It can house large volumes of structured, semi-structured, and unstructured data.
3. **Dynamic schema:** NoSQL database allows insertion of data without a pre-defined schema. In other words, it facilitates application changes in real time, which thus supports faster development, easy code integration, and requires less database administration.
4. **Auto-sharding:** It automatically spreads data across an arbitrary number of servers. The application in question is more often not even aware of the composition of the server pool. It balances the load of data and query on the available servers; and if and when a server goes down, it is quickly replaced without any major activity disruptions.
5. **Replication:** It offers good support for replication which in turn guarantees high availability, fault tolerance, and disaster recovery.

4.1.5 Advantages of NoSQL

Let us enumerate the advantages of NoSQL. Refer Figure 4.4.

1. **Can easily scale up and down:** NoSQL database supports scaling rapidly and elastically and even allows to scale to the cloud.

Table 4.1 Popular schema-less databases

Key–Value Data Store	Column-Oriented Data Store	Document Data Store	Graph Data Store
• Riak	• Cassandra	• MongoDB	• InfiniteGraph
• Redis	• HBase	• CouchDB	• Neo4j
• Membase	• HyperTable	• RavenDB	• AllegroGraph

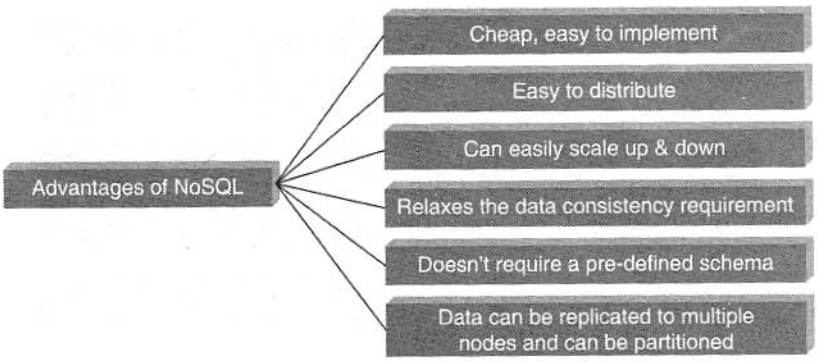


Figure 4.4 Advantages of NoSQL.

(a) ***Cluster scale:*** It allows distribution of database across 100+ nodes often in multiple data centers.
(b) ***Performance scale:*** It sustains over 100,000+ database reads and writes per second.
(c) ***Data scale:*** It supports housing of 1 billion+ documents in the database.

2. **Doesn't require a pre-defined schema:** NoSQL does not require any adherence to pre-defined schema. It is pretty flexible. For example, if we look at MongoDB, the documents (equivalent of records in RDBMS) in a collection (equivalent of table in RDBMS) can have different sets of key–value pairs.

   {_id: 101,"BookName": "Fundamentals of Business Analytics", "AuthorName": "Seema Acharya", "Publisher": "Wiley India"}
   {_id:102, "BookName":"Big Data and Analytics"}

3. **Cheap, easy to implement:** Deploying NoSQL properly allows for all of the benefits of scale, high availability, fault tolerance, etc. while also lowering operational costs.
4. **Relaxes the data consistency requirement:** NoSQL databases have adherence to CAP theorem (Consistency, Availability, and Partition tolerance). Most of the NoSQL databases compromise on consistency in favor of availability and partition tolerance. However, they do go for eventual consistency.
5. **Data can be replicated to multiple nodes and can be partitioned:** There are two terms that we will discuss here:

   (a) ***Sharding:*** Sharding is when different pieces of data are distributed across multiple servers. NoSQL databases support auto-sharding; this means that they can natively and automatically spread data across an arbitrary number of servers, without requiring the application to even be aware of the composition of the server pool. Servers can be added or removed from the data layer without application downtime. This would mean that data and query load are automatically balanced across servers, and when a server goes down, it can be quickly and transparently replaced with no application disruption.
   (b) ***Replication:*** Replication is when multiple copies of data are stored across the cluster and even across data centers. This promises high availability and fault tolerance.

4.1.6 What We Miss With NoSQL?

With NoSQL around, we have been able to counter the problem of scale (NoSQL scales out). There is also the flexibility with respect to schema design. However there are few features of conventional RDBMS that are greatly missed. Refer Figure 4.5.

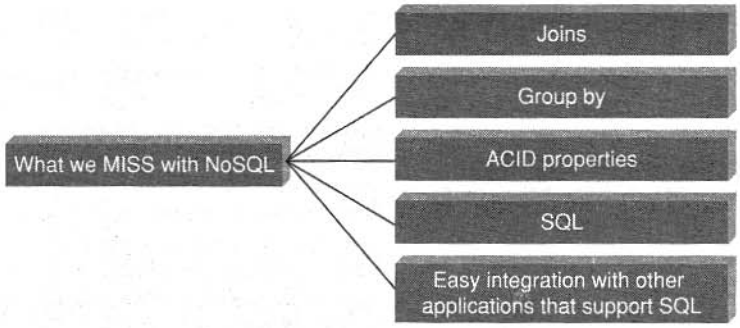


Figure 4.5 What we miss with NoSQL?

NoSQL does not support joins. However, it compensates for it by allowing embedded documents as in MongoDB. It does not have provision for ACID properties of transactions. However, it obeys the Eric Brewer's CAP theorem. NoSQL does not have a standard SQL interface but NoSQL databases such as MongoDB and Cassandra have their own rich query language [MongoDB query language and Cassandra query language (CQL)] to compensate for the lack of it. One thing which is dearly missed is the easy integration with other applications that support SQL.

4.1.7 Use of NoSQL in Industry

NoSQL is being put to use in varied industries. They are used to support analysis for applications such as web user data analysis, log analysis, sensor feed analysis, making recommendations for upsell and cross-sell, etc. Refer Figure 4.6.

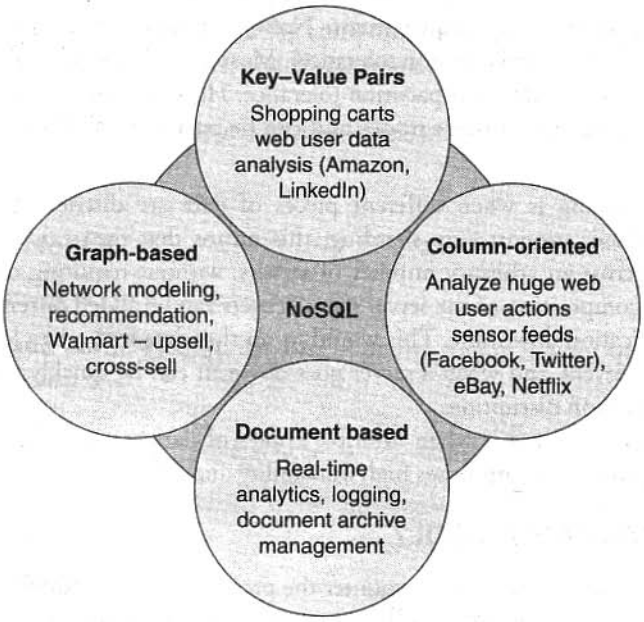


Figure 4.6 Use of NoSQL in industry.

4.1.8 NoSQL Vendors

Refer Table 4.2 for few popular NoSQL vendors.

Table 4.2 Few popular NoSQL vendors

Company	Product	Most Widely Used by
Amazon	DynamoDB	LinkedIn, Mozilla
Facebook	Cassandra	Netflix, Twitter, eBay
Google	BigTable	Adobe Photoshop

4.1.9 SQL versus NoSQL

Refer Table 4.3 for few salient differences between SQL and NoSQL.

Table 4.3 SQL versus NoSQL

SQL	NoSQL
Relational database	Non-relational, distributed database
Relational model	Model-less approach
Pre-defined schema	Dynamic schema for unstructured data
Table based databases	Document-based or graph-based or wide column store or key–value pairs databases
Vertically scalable (by increasing system resources)	Horizontally scalable (by creating a cluster of commodity machines)
Uses SQL	Uses UnQL (Unstructured Query Language)
Not preferred for large datasets	Largely preferred for large datasets
Not a best fit for hierarchical data	Best fit for hierarchical storage as it follows the key–value pair of storing data similar to JSON (Java Script Object Notation)
Emphasis on ACID properties	Follows Brewer's CAP theorem
Excellent support from vendors	Relies heavily on community support
Supports complex querying and data keeping needs	Does not have good support for complex querying
Can be configured for strong consistency	Few support strong consistency (e.g., MongoDB), some others can be configured for eventual consistency (e.g., Cassandra)
Examples: Oracle, DB2, MySQL, MS SQL, PostgreSQL, etc.	Examples: MongoDB, HBase, Cassandra, Redis, Neo4j, CouchDB, Couchbase, Riak, etc.

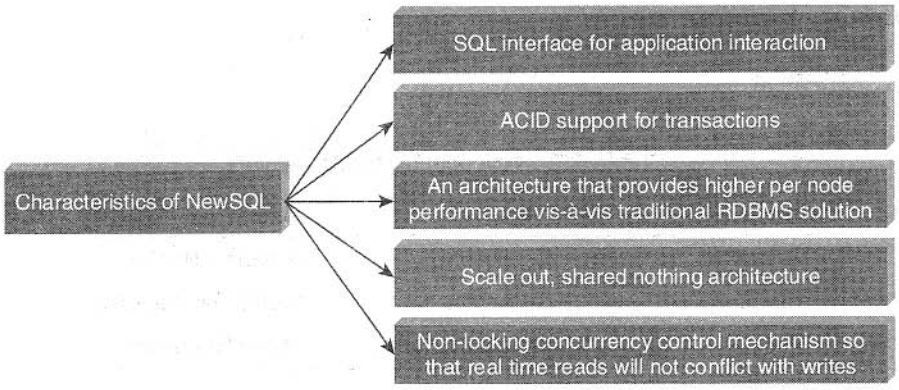


Figure 4.7 Characteristics of NewSQL.

4.1.10 NewSQL

There is yet another new term doing the rounds – "NewSQL". So, what is NewSQL and how is it different from SQL and NoSQL?

What is that we love about NoSQL and is not there with our traditional RDBMS and what is that we love about SQL that NoSQL does not have support for? You guessed it right!!! We need a database that has the same scalable performance of NoSQL systems for On Line Transaction Processing (OLTP) while still maintaining the ACID guarantees of a traditional database. This new modern RDBMS is called NewSQL. It supports relational data model and uses SQL as their primary interface.

4.1.10.1 Characteristics of NewSQL

Refer Figure 4.7 to learn about the characteristics of NewSQL. NewSQL is based on the shared nothing architecture with a SQL interface for application interaction.

4.1.11 Comparison of SQL, NoSQL, and NewSQL

Refer Table 4.4 for a comparative study of SQL, NoSQL and NewSQL.

Table 4.4 Comparative study of SQL, NoSQL and NewSQL

	SQL	NoSQL	NewSQL
Adherence to ACID properties	Yes	No	Yes
OLTP/OLAP	Yes	No	Yes
Schema rigidity Adherence to data model	Yes Adherence to relational model	No	Maybe
Data Format Flexibility	No	Yes	Maybe
Scalability	Scale up Vertical Scaling	Scale out Horizontal Scaling	Scale out
Distributed Computing	Yes	Yes	Yes
Community Support	Huge	Growing	Slowly growing

4.2 HADOOP

Hadoop is an open-source project of the Apache foundation. It is a framework written in Java, originally developed by Doug Cutting in 2005 who named it after his son's toy elephant. He was working with Yahoo then. It was created to support distribution for "Nutch", the text search engine. Hadoop uses Google's MapReduce and Google File System technologies as its foundation. Hadoop is now a core part of the computing infrastructure for companies such as Yahoo, Facebook, LinkedIn, Twitter, etc. Refer Figure 4.8.

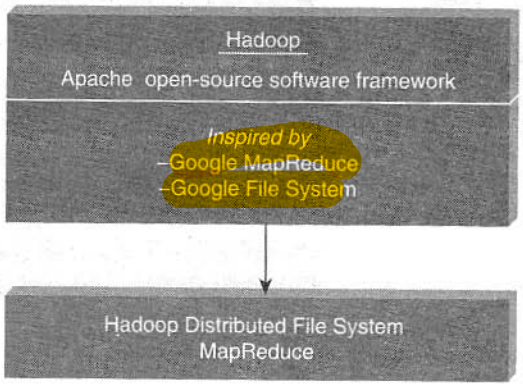


Figure 4.8 Hadoop.

4.2.1 Features of Hadoop

Let us cite a few features of Hadoop:

1. It is optimized to handle massive quantities of structured, semi-structured, and unstructured data, using commodity hardware, that is, relatively inexpensive computers.
2. Hadoop has a shared nothing architecture.
3. It replicates its data across multiple computers so that if one goes down, the data can still be processed from another machine that stores its replica.
4. Hadoop is for high throughput rather than low latency. It is a batch operation handling massive quantities of data; therefore the response time is not immediate.
5. It complements On-Line Transaction Processing (OLTP) and On-Line Analytical Processing (OLAP). However, it is not a replacement for a relational database management system.
6. It is NOT good when work cannot be parallelized or when there are dependencies within the data.
7. It is NOT good for processing small files. It works best with huge data files and datasets.

4.2.2 Key Advantages of Hadoop

Refer Figure 4.9 for a quick look at the key advantages of Hadoop. Some of them are as follows:

1. **Stores data in its native format:** Hadoop's data storage framework (HDFS – Hadoop Distributed File System) can store data in its native format. There is no structure that is imposed while keying in data or storing data. HDFS is pretty much schema-less. It is only later when the data needs to be processed that structure is imposed on the raw data.
2. **Scalable:** Hadoop can store and distribute very large datasets (involving thousands of terabytes of data) across hundreds of inexpensive servers that operate in parallel.

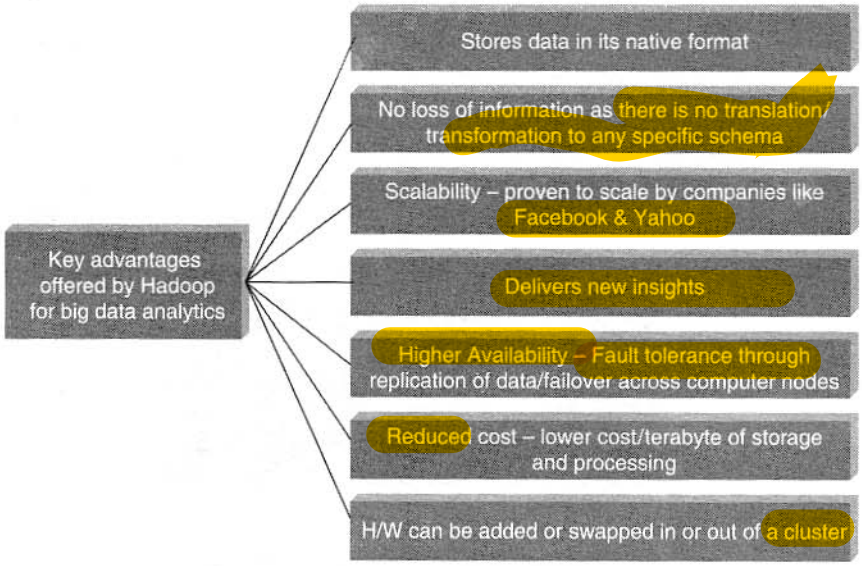


Figure 4.9 Key advantages of Hadoop.

3. **Cost-effective:** Owing to its scale-out architecture, Hadoop has a much reduced cost/terabyte of storage and processing.
4. **Resilient to failure:** Hadoop is fault-tolerant. It practices replication of data diligently which means whenever data is sent to any node, the same data also gets replicated to other nodes in the cluster, thereby ensuring that in the event of a node failure, there will always be another copy of data available for use.
5. **Flexibility:** One of the key advantages of Hadoop is its ability to work with all kinds of data: structured, semi-structured, and unstructured data. It can help derive meaningful business insights from email conversations, social media data, click-stream data, etc. It can be put to several purposes such as log analysis, data mining, recommendation systems, market campaign analysis, etc.
6. **Fast:** Processing is extremely fast in Hadoop as compared to other conventional systems owing to the "move code to data" paradigm.

   Hadoop has a shared-nothing architecture.

4.2.3 Versions of Hadoop

There are two versions of Hadoop available:

1. Hadoop 1.0
2. Hadoop 2.0

Let us take a look at the features of both. Refer Figure 4.10.

4.2.3.1 Hadoop 1.0

It has two main parts:

1. **Data storage framework:** It is a general-purpose file system called Hadoop Distributed File System (HDFS). HDFS is schema-less. It simply stores data files. These data files can be in just about any

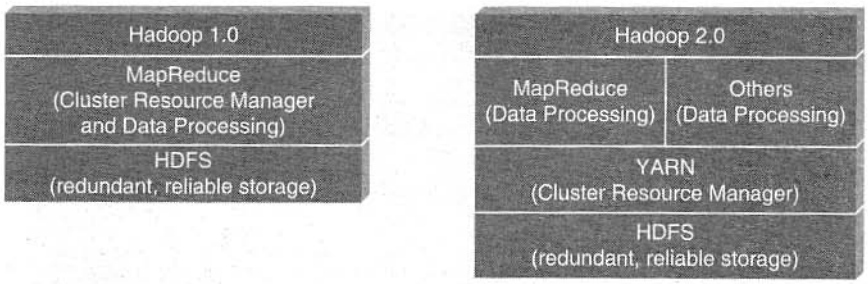


Figure 4.10 Versions of Hadoop.

format. The idea is to store files as close to their original form as possible. This is turn provides the business units and the organization the much needed flexibility and agility without being overly worried by what it can implement.

2. **Data processing framework:** This is a simple functional programming model initially popularized by Google as MapReduce. It essentially uses two functions: the MAP and the REDUCE functions to process data. The "Mappers" take in a set of key–value pairs and generate intermediate data (which is another list of key–value pairs). The "Reducers" then act on this input to produce the output data. The two functions seemingly work in isolation from one another, thus enabling the processing to be highly distributed in a highly-parallel, fault-tolerant, and scalable way.

There were, however, a few limitations of Hadoop 1.0. They are as follows:

1. The first limitation was the requirement for MapReduce programming expertise along with proficiency required in other programming languages, notably Java.
2. It supported only batch processing which although is suitable for tasks such as log analysis, large-scale data mining projects but pretty much unsuitable for other kinds of projects.
3. One major limitation was that Hadoop 1.0 was tightly computationally coupled with MapReduce, which meant that the established data management vendors were left with two options: Either rewrite their functionality in MapReduce so that it could be executed in Hadoop or extract the data from HDFS and process it outside of Hadoop. None of the options were viable as it led to process inefficiencies caused by the data being moved in and out of the Hadoop cluster.

Let us look at whether these limitations have been wholly or in parts resolved by Hadoop 2.0.

4.2.3.2 *Hadoop 2.0*

In Hadoop 2.0, HDFS continues to be the data storage framework. However, a new and separate resource management framework called **Y**et **A**nother **R**esource **N**egotiator (YARN) has been added. Any application capable of dividing itself into parallel tasks is supported by YARN. YARN coordinates the allocation of subtasks of the submitted application, thereby further enhancing the flexibility, scalability, and efficiency of the applications. It works by having an ApplicationMaster in place of the erstwhile JobTracker, running applications on resources governed by a new NodeManager (in place of the erstwhile TaskTracker). ApplicationMaster is able to run any application and not just MapReduce.

This, in other words, means that the MapReduce Programming expertise is no longer required. Furthermore, it not only supports batch processing but also real-time processing. MapReduce is no longer the only data processing option; other alternative data processing functions such as data standardization, master data management can now be performed natively in HDFS.

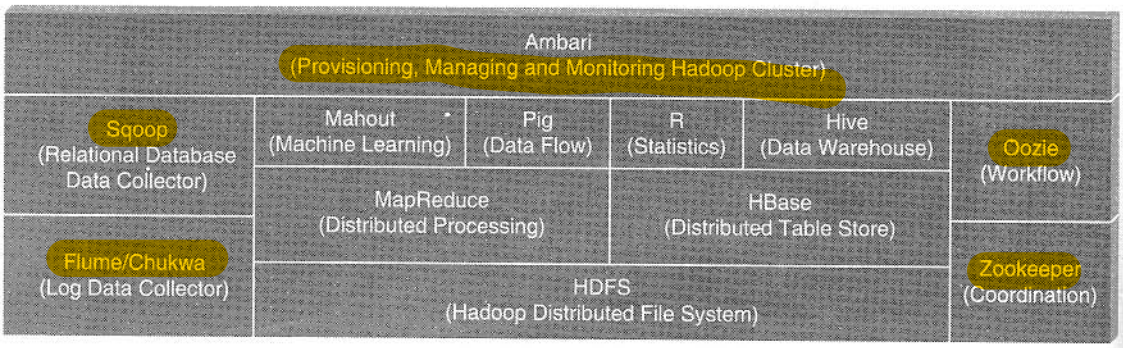


Figure 4.11 Hadoop ecosystem.

4.2.4 Overview of Hadoop Ecosystems

The components of the Hadoop ecosystem are shown in Figure 4.11.

There are components available in the Hadoop ecosystem for data ingestion, processing, and analysis.

Data Ingestion → Data Processing → Data Analysis

Components that help with Data Ingestion are:

1. Sqoop
2. Flume

Components that help with Data Processing are:

1. MapReduce
2. Spark

Components that help with Data Analysis are:

1. Pig
2. Hive
3. Impala

HDFS

It is the distributed storage unit of Hadoop. It provides streaming access to file system data as well as file permissions and authentication. It is based on GFS (Google File System). It is used to scale a single cluster node to hundreds and thousands of nodes. It handles large datasets running on commodity hardware. HDFS is highly fault-tolerant. It stores files across multiple machines. These files are stored in redundant fashion to allow for data recovery in case of failure.

PICTURE THIS...

An e-commerce website stores millions of customers' data in a distributed manner. Data has been collected over 4–5 years. It then runs batch analytics on the archived data to analyze customer's behavior, buying patterns, their preferences, their requirements, etc. This helps to understand which products are purchased by customers in which months, etc.

HBase

It stores data in HDFS. It is the first non-batch component of the Hadoop Ecosystem. It is a database on top of HDFS. It provides a quick random access to the stored data. It has very low latency compared to HDFS. It is a NoSQL database, is non-relational and is a column-oriented database. A table can have thousands of columns. A table can have multiple rows. Each row can have several column families. Each column family can have several columns. Each column can have several key values. It is based on Google BigTable. This is widely used by Facebook, Twitter, Yahoo, etc.

PICTURE THIS...

The same e-commerce website as in the HDFS case above also stores millions of product data. To search for a product among millions of products and to produce the result immediately (or you can say in real time), it needs to optimize the request and search process. HBase supports real-time analytics. Given the huge velocity of data, they opted for HBase over HDFS, as HDFS does not support real-time writes. The results were overwhelming; it reduced the query time from 3 days to 3 minutes.

Difference between HBase and Hadoop/HDFS

1. HDFS is the file system whereas HBase is a Hadoop database. It is like NTFS and MySQL.
2. HDFS is WORM (Write once and read multiple times or many times). Latest versions support appending of data but this feature is rarely used. However, HBase supports real-time random read and write.
3. HDFS is based on Google File System (GFS) whereas HBase is based on Google Big Table.
4. HDFS supports only full table scan or partition table scan. Hbase supports random small range scan or table scan.
5. Performance of Hive on HDFS is relatively very good but for HBase it becomes 4–5 times slower.
6. The access to data is via MapReduce job only in HDFS whereas in HBase the access is via Java APIs, Rest, Avro, Thrift APIs.
7. HDFS does not support dynamic storage owing to its rigid structure whereas HBase supports dynamic storage.
8. HDFS has high latency operations whereas HBase has low latency operations.
9. HDFS is most suitable for batch analytics whereas HBase is for real-time analytics.

Hadoop Ecosystem Components for Data Ingestion

1. **Sqoop:** Sqoop stands for SQL to Hadoop. Its main functions are
   a) Importing data from RDBMS such as MySQL, Oracle, DB2, etc. to Hadoop file system (HDFS, HBase, Hive).
   b) Exporting data from Hadoop File system (HDFS, HBase, Hive) to RDBMS (MySQL, Oracle, DB2).

   Uses of Sqoop

   a) It has a connector-based architecture to allow plug-ins to connect to external systems such as MySQL, Oracle, DB2, etc.

b) It can provision the data from external system on to HDFS and populate tables in Hive and HBase.
c) It integrates with Oozie allowing you to schedule and automate import and export tasks.

2. **Flume:** Flume is an important log aggregator (aggregates logs from different machines and places them in HDFS) component in the Hadoop ecosystem. Flume has been developed by Cloudera. It is designed for high volume ingestion of event-based data into Hadoop. The default destination in Flume (called as sink in flume parlance) is HDFS. However it can also write to HBase or Solr.

PICTURE THIS...

There is a bank of web servers. Flume moves log events from those files into new aggregated files in HDFS for processing.

Hadoop Ecosystem Components for Data Processing

1. **MapReduce:** It is a programing paradigm that allows distributed and parallel processing of huge datasets. It is based on Google MapReduce. Google released a paper on MapReduce programming paradigm in 2004 and that became the genesis of Hadoop processing model. The MapReduce framework gets the input data from HDFS. There are two main phases: Map phase and the Reduce phase. The map phase converts the input data into another set of data (key–value pairs). This new intermediate dataset then serves as the input to the reduce phase. The reduce phase acts on the datasets to combine (aggregate and consolidate) and reduce them to a smaller set of tuples. The result is then stored back in HDFS.
2. **Spark:** It is both a programming model as well as a computing model. It is an open-source big data processing framework. It was originally developed in 2009 at UC Berkeley's AmpLab and became an open-source project in 2010. It is written in Scala. It provides in-memory computing for Hadoop. In Spark, workloads execute in memory rather than on disk owing to which it is much faster (10 to 100 times) than when the workload is executed on disk. However, if the datasets are too large to fit into the available system memory, it can perform conventional disk-based processing. It serves as a potentially faster and more flexible alternative to MapReduce. It accesses data from HDFS (Spark does not have its own distributed file system) but bypasses the MapReduce processing.

   Spark can be used with Hadoop coexisting smoothly with MapReduce (sitting on top of Hadoop YARN) or used independently of Hadoop (standalone). As a programming model, it works well with Scala, Python (it has API connectors for using it with Java or Python) or R programming language.

   The following are the Spark libraries:

   a) ***Spark SQL:*** Spark also has support for SQL. Spark SQL uses SQL to help query data stored in disparate applications.
   b) ***Spark streaming:*** It helps to analyze and present data in real time.
   c) ***MLib:*** It supports machine learning such as applying advanced statistical operations on data in Spark Cluster.
   d) ***GraphX:*** It helps in graph parallel computation.

   Spark and Hadoop are usually used together by several companies. Hadoop was primarily designed to house unstructured data and run batch processing operations on it. Spark is used extensively for its

high speed in memory computing and ability to run advanced real-time analytics. The two together have been giving very good results.

Hadoop Ecosystem Components for Data Analysis

1. **Pig:** It is a high-level scripting language used with Hadoop. It serves as an alternative to MapReduce. It has two parts:
   (a) ***Pig Latin:*** It is SQL-like scripting language. Pig Latin scripts are translated into MapReduce jobs which can then run on YARN and process data in the HDFS cluster. It was initially developed by Yahoo. It is immensely popular with developers who are not comfortable with MapReduce. However, SQL developers may have a preference for Hive.

   How it works? There is a "Load" command available to load the data from "HDFS" into Pig. Then one can perform functions such as grouping, filtering, sorting, joining etc. The processed or computed data can then be either displayed on screen or placed back into HDFS.

   It gives you a platform for building data flow for ETL (Extract, Transform and Load), processing and analyzing huge data sets.
   (b) ***Pig runtime:*** It is the runtime environment.
2. **Hive:** Hive is a data warehouse software project built on top of Hadoop. Three main tasks performed by Hive are summarization, querying and analysis. It supports queries written in a language called HQL or HiveQL which is a declarative SQL-like language. It converts the SQL-style queries into MapReduce jobs which are then executed on the Hadoop platform.

Difference between Hive and RDBMS

Both Hive and traditional databases such as MySQL, MS SQL Server, PostgreSQL support SQL interface. However, Hive is better known as a datawarehouse (D/W) rather than a database.

Let us look at the difference between Hive and traditional databases as regards the schema.

1. Hive enforces schema on Read Time whereas RDBMS enforces schema on Write Time. In RDBMS, at the time of loading/inserting data, the table's schema is enforced. If the data being loaded does not conform to the schema then it is rejected. Thus, the schema is enforced on write (loading the data into the database). Schema on write takes longer to load the data into the database; however it makes up for it during data retrieval with a good query time performance. However, Hive does not enforce the schema when the data is being loaded into the D/W. It is enforced only when the data is being read/retrieved. This is called schema on read. It definitely makes for fast initial load as the data load or insertion operation is just a file copy or move.
2. Hive is based on the notion of write once and read many times whereas the RDBMS is designed for read and write many times.
3. Hadoop is a batch-oriented system. Hive, therefore, is not suitable for OLTP (Online Transaction Processing) but, although not ideal, seems closer to OLAP (Online Analytical Processing). The reason being that there is quite a latency between issuing a query and receiving a reply as the query written in HiveQL will be converted to MapReduce jobs which are then executed on the Hadoop cluster. RDBMS is suitable for housing day-to-day transaction data and supports all OLTP operations with frequent insertions, modifications (updates), deletions of the data.

4. Hive handles static data analysis which is non-real-time data. Hive is the data warehouse of Hadoop. There are no frequent updates to the data and the query response time is not fast. RDBMS is suited for handling dynamic data which is real time.
5. Hive can be easily scaled at a very low cost when compared to RDMS. Hive uses HDFS to store data, thus it cannot be considered as the owner of the data, while on the other hand RDBMS is the owner of the data responsible for storing, managing and manipulating it in the database.
6. Hive uses the concept of parallel computing, whereas RDBMS uses serial computing.

We summarize the difference in Table 4.5.

Table 4.5 Hive versus RDBMS

	Hadoop	RDBMS
Data Variety	Used for structured, semi-structured and unstructured data. Hadoop supports a variety of data formats in real time such as XML, JSON, and text-based flat file formats.	Used for structured data
Data Storage	Usually datasets of size terabytes, petabytes	Usually datasets of size gigabytes
Querying	HiveQL	SQL
Query Response	In Hadoop, there is latency due to batch processing.	In RDBMS, query response time is immediate.
Schema	Schema required on read	Schema required on write
Speed	Writes are faster compared to reads as there is no adherence to schema required at the time of inserting or writing data. Schema is enforced at read time	Reads are very fast (supported by building indexes on required columns).
	Hadoop is designed for write once read many times. It does not work for random reading and writing of a few records like RDBMS.	RDBMS is designed for read and write many times.
Cost	Apache Hadoop is open-source, large-scale, distributed, scalable, data intensive computing.	Available as proprietary RDBMS such as Oracle, MS SQL Server, IBM DB2, etc. Also open-source RDBMS are available such as MySQL, PostgreSQL, etc.
Use Cases	Analytics, data discovery	OLTP (Online Transaction Processing). Mainly used to store and process day-to-day business data.

(Continued)

Table 4.5 (Continued)

	Hadoop	RDBMS
Throughput	High	Low
Scalability	Horizontal (Hadoop scales by adding nodes to a Hadoop cluster of easily available commodity machines).	Vertical: RDBMS scales vertically by increasing the horsepower (CPU, Hard Disk Capacity, RAM, etc.) of the machine.
Hardware	Commodity/Utility Hardware	High End Servers
Integrity	Low	High. Obeys ACID properties A – Atomicity C – Consistency I – Integrity D – Durability

Difference between Hive and HBase

1. Hive is a MapReduce-based SQL engine that runs on top of Hadoop. HBase is a key–value NoSQL database that runs on top of HDFS.
2. Hive is for batch processing of big data. HBase is for real-time data streaming.

Impala

It is a high performance SQL engine that runs on Hadoop cluster. It is ideal for interactive analysis. It has very low latency measured in milliseconds. It supports a dialect of SQL called Impala SQL.

ZooKeeper

It is a coordination service for distributed applications.

Oozie

It is a workflow scheduler system to manage Apache Hadoop jobs.

Mahout

It is a scalable machine learning and data mining library.

Chukwa

It is a data collection system for managing large distributed systems.

Ambari

It is a web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.

4.2.5 Hadoop Distributions

Hadoop is an open-source Apache project. Anyone can freely download the core aspects of Hadoop. The core aspects of Hadoop include the following:

1. Hadoop Common
2. **H**adoop **D**istributed **F**ile **S**ystem (HDFS)
3. Hadoop YARN (**Y**et **A**nother **R**esource **N**egotiator)
4. Hadoop MapReduce

There are few companies such as IBM, Amazon Web Services, Microsoft, Teradata, Hortonworks, Cloudera, etc. that have packaged Hadoop into a more easily consumable distributions or services. Although each of these companies have a slightly different strategy, the key essence remains its ability to distribute data and workloads across potentially thousands of servers thus making big data manageable data. A few Hadoop distributions are given in Figure 4.12.

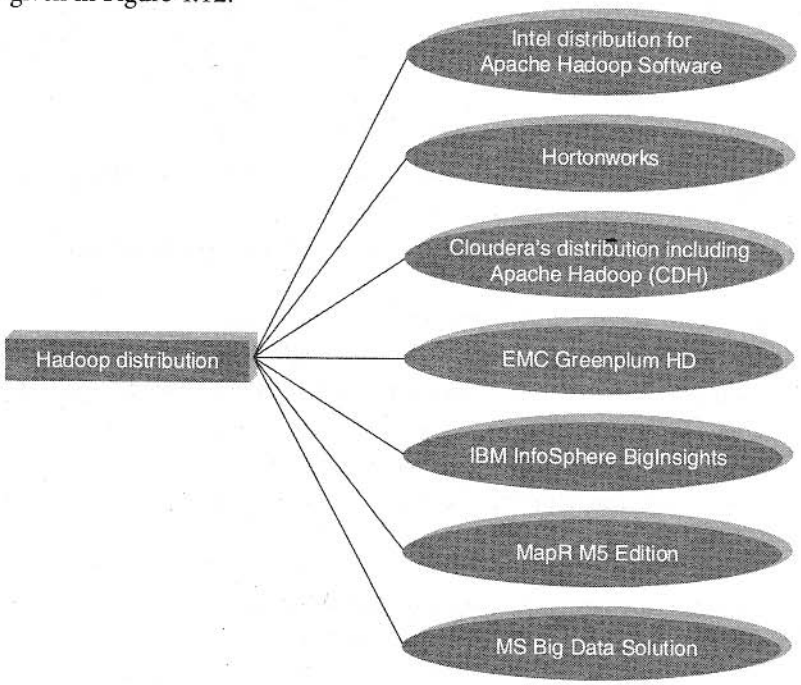


Figure 4.12 Hadoop distributions.

4.2.6 Hadoop versus SQL

Table 4.6 lists the differences between Hadoop and SQL.

Table 4.6 Hadoop versus SQL

Hadoop	SQL
Scale out	Scale up
Key–Value pairs	Relational table
Functional Programming	Declarative Queries
Offline batch processing	Online transaction processing

4.2.7 Integrated Hadoop Systems Offered by Leading Market Vendors

Refer Figure 4.13 to get a glimpse of the leading market vendors offering integrated Hadoop systems.

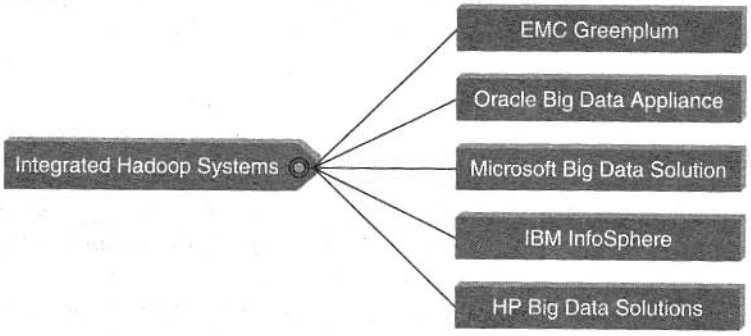


Figure 4.13 Integrated Hadoop systems.

4.2.8 Cloud-Based Hadoop Solutions

Amazon Web Services holds out a comprehensive, end-to-end portfolio of cloud computing services to help manage big data. The aim is to achieve this and more along with retaining the emphasis on reducing costs, scaling to meet demand, and accelerating the speed of innovation.

The Google Cloud Storage connector for Hadoop empowers one to perform MapReduce jobs directly on data in Google Cloud Storage, without the need to copy it to local disk and running it in the Hadoop Distributed File System (HDFS). The connector simplifies Hadoop deployment, and at the same time reduces cost and provides performance comparable to HDFS, all this while increasing reliability by eliminating the single point of failure of the name node. Refer Figure 4.14.

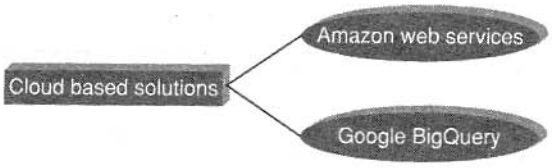


Figure 4.14 Cloud-based solutions.

REMIND ME

- NoSQL databases are non-relational, open source, distributed databases.
- NoSQL database allows insertion of data without a pre-defined schema.
- Hadoop has a shared nothing architecture.

- Hadoop 1.0 has two main parts:
  - Data storage framework
  - Data processing framework
- In Hadoop 2.0, a new and separate resource management framework called **Y**et **A**nother **R**esource **N**egotiator (YARN) has been added.

POINT ME (BOOKS)

- Hadoop for Dummies, Dirk Deroos, Paul C. Zikopoulos, Roman B. Melnyk, Bruce Brown, Wiley India Pvt. Ltd.
- NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Pramod J. Sadalage and Martin Fowler.

CONNECT ME (INTERNET RESOURCES)

- http://www.mongodb.com/nosql-explained
- http://nosql-database.org/
- http://www.techrepublic.com/blog/10-things/10-things-you-should-know-about-nosql-databases/
- http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduce_Compatibility_Hadoop1_Hadoop2.html
- http://hadoop.apache.org/

TEST ME

A. Fill Me

1. The expansion for CAP is _____, _____ and _____.
2. The expansion of BASE is _____.
3. MongoDB is _____ and _____.
4. Cassandra is _____ and _____.
5. _____ has no support for ACID properties of transactions.
6. _____ is a robust database that supports ACID properties of transactions and has the scalability of NoSQL.

Answers:

1. Consistency, Availability and Partition Tolerance
2. Basically Available Soft State Eventual Consistency
3. Consistent and Partition Tolerant
4. Available and Partition Tolerant
5. NoSQL
6. NewSQL

1. Place it in the Basket

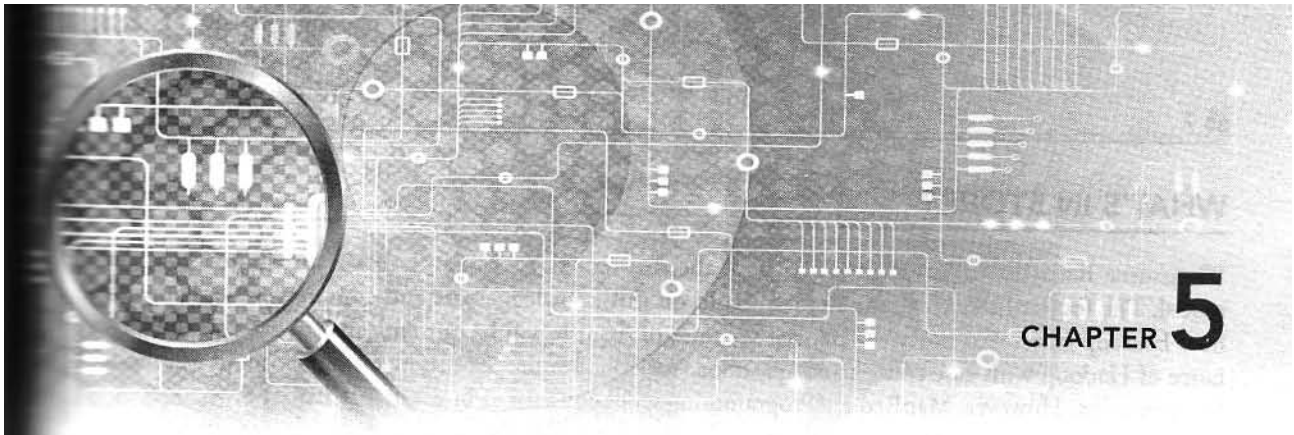
SQL	NoSQL

Following words are to be placed in the relevant basket:

(a) Relational
(b) Distributed
(c) Predefined schema
(d) Wide-column stores
(e) Vertically scalable
(f) Key–value pairs
(g) MySQL
(h) CouchDB
(i) Neo4j
(j) Cassandra
(k) Large dataset
(l) ACID properties
(m) Brewers CAP theorem
(n) Document based database
(o) Scales horizontally
(p) Avoids join operations
(q) JSON data
(r) Table or relations

Answers:

SQL	NoSQL
Relational	Distributed
Predefined schema	Wide-column stores
Vertically scalable	Key–value pairs
MySQL	CouchDB
ACID properties	Neo4j
Table or relations	Cassandra
	Large dataset
	Brewers CAP theorem
	Document based database
	Scales horizontally
	Avoids join operations
	JSON data



CHAPTER 5

Introduction to Hadoop

BRIEF CONTENTS

- What's in Store?
- Introducing Hadoop
  - Data: The Treasure Trove
- Why Hadoop?
- Why not RDBMS?
- RDBMS versus Hadoop
- Distributed Computing Challenges
  - Hardware Failure
  - How to Process this Gigantic Store of Data?
- History of Hadoop
  - The Name "Hadoop"
- Hadoop Overview
  - Key Aspects of Hadoop
  - Hadoop Components
  - Hadoop Conceptual Layer
  - High-Level Architecture of Hadoop
- Use Case for Hadoop
  - ClickStream Data
- Hadoop Distributors
- HDFS
  - HDFS Daemons
  - Anatomy of File Read
  - Anatomy of File Write
  - Replica Placement Strategy
  - Working with HDFS Commands
  - Special Features of HDFS
- Processing Data with Hadoop
  - MapReduce Daemons
  - How does MapReduce Work?
  - MapReduce Example
- Managing Resources and Applications with Hadoop YARN
  - Limitations of Hadoop 1.0 Architecture
  - HDFS Limitation
  - Hadoop 2: HDFS
  - Hadoop 2 YARN: Taking Hadoop Beyond Batch
- Interacting with Hadoop Ecosystem
  - Pig
  - Hive
  - Sqoop
  - HBase

"There were 5 exabytes of information created between the dawns of civilization through 2003, but that much information is now created every 2 days."

– Eric Schmidt, of Google, said in 2010

WHAT'S IN STORE?

We assume that you are already familiar with the distributed file system and the distributed computing model. The focus of this chapter will be to build on this knowledge base and comprehend and appreciate how Hadoop stores and processes colossal volumes of data. It will be our endeavor to get you the importance of Hadoop with case studies and scenarios. We will also discuss HDFS commands and MapReduce Programming. However, MapReduce Programming will be discussed in detail in Chapter 8.

We suggest you refer to some of the learning resources provided at the end of this chapter and also complete the "Test Me" exercises.

5.1 INTRODUCING HADOOP

Today, Big Data seems to be the buzz word! Enterprises, the world over, are beginning to realize that there is a huge volume of untapped information before them in the form of structured, semi-structured, and unstructured data. This varied variety of data is spread across the networks.

Let us look at few statistics to get an idea of the amount of data which gets generated every day, every minute, and every second.

1. **Every day:**
   (a) NYSE (New York Stock Exchange) generates 1.5 billion shares and trade data.
   (b) Facebook stores 2.7 billion comments and Likes.
   (c) Google processes about 24 petabytes of data.
2. **Every minute:**
   (a) Facebook users share nearly 2.5 million pieces of content.
   (b) Twitter users tweet nearly 300,000 times.
   (c) Instagram users post nearly 220,000 new photos.
   (d) YouTube users upload 72 hours of new video content.
   (e) Apple users download nearly 50,000 apps.
   (f) Email users send over 200 million messages.
   (g) Amazon generates over $80,000 in online sales.
   (h) Google receives over 4 million search queries.
3. **Every second:**
   (a) Banking applications process more than 10,000 credit card transactions.

5.1.1 Data: The Treasure Trove

1. Provides business advantages such as generating product recommendations, inventing new products, analyzing the market, and many, many more, ...
2. Provides few early key indicators that can turn the fortune of business.
3. Provides room for precise analysis. If we have more data for analysis, then we have greater precision of analysis.

To process, analyze, and make sense of these different kinds of data, we need a system that scales and addresses the challenges shown in Figure 5.1.

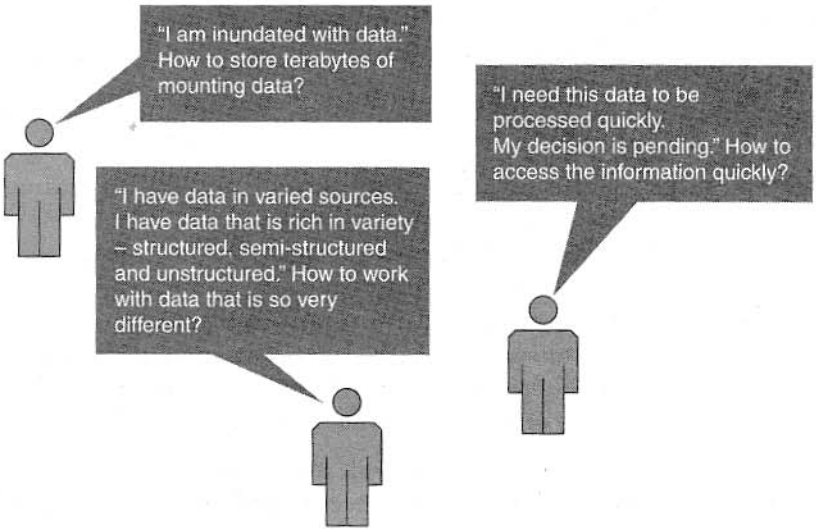


Figure 5.1 Challenges with big volume, variety, and velocity of data.

5.2 WHY HADOOP?

Ever wondered why Hadoop has been and is one of the most wanted technologies!!

The key consideration (the rationale behind its huge popularity) is:

Its capability to handle massive amounts of data, different categories of data – fairly quickly.

The other considerations are (Figure 5.2):

1. **Low cost:** Hadoop is an open-source framework and uses commodity hardware (commodity hardware is relatively inexpensive and easy to obtain hardware) to store enormous quantities of data.

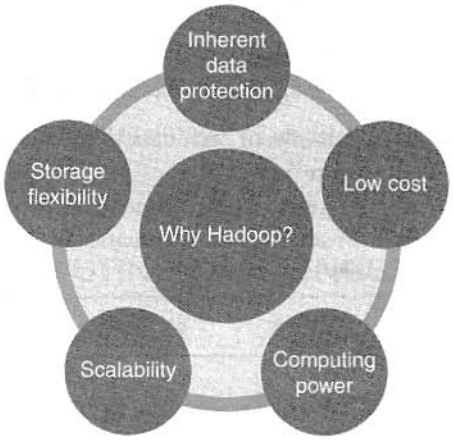


Figure 5.2 Key considerations of Hadoop.

2. **Computing power:** Hadoop is based on distributed computing model which processes very large volumes of data fairly quickly. The more the number of computing nodes, the more the processing power at hand.
3. **Scalability:** This boils down to simply adding nodes as the system grows and requires much less administration.
4. **Storage flexibility:** Unlike the traditional relational databases, in Hadoop data need not be pre-processed before storing it. Hadoop provides the convenience of storing as much data as one needs and also the added flexibility of deciding later as to how to use the stored data. In Hadoop, one can store unstructured data like images, videos, and free-form text.
5. **Inherent data protection:** Hadoop protects data and executing applications against hardware failure. If a node fails, it automatically redirects the jobs that had been assigned to this node to the other functional and available nodes and ensures that distributed computing does not fail. It goes a step further to store multiple copies (replicas) of the data on various nodes across the cluster.

Hadoop makes use of commodity hardware, distributed file system, and distributed computing as shown in Figure 5.3. In this new design, groups of machine are gathered together; it is known as a **Cluster**.

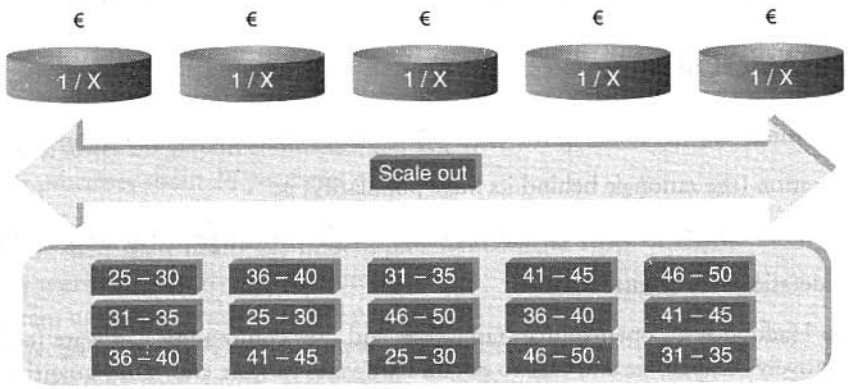


Figure 5.3 Hadoop framework (distributed file system, commodity hardware).

With this new paradigm, the data can be managed with **Hadoop** as follows:

1. Distributes the data and duplicates chunks of each data file across several nodes, for example, 25–30 is one chunk of data as shown in Figure 5.3.
2. Locally available compute resource is used to process each chunk of data in parallel.
3. Hadoop Framework handles failover smartly and automatically.

5.3 WHY NOT RDBMS?

RDBMS is not suitable for storing and processing large files, images, and videos. RDBMS is not a good choice when it comes to advanced analytics involving machine learning. Figure 5.4 describes the RDBMS system with respect to cost and storage. It calls for huge investment as the volume of data shows an upward trend.

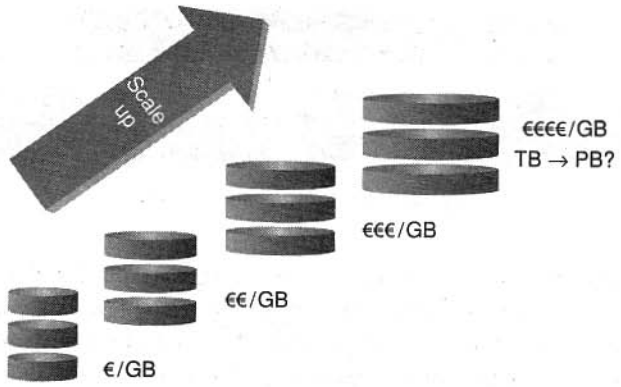


Figure 5.4 RDBMS with respect to cost/GB of storage.

5.4 RDBMS versus HADOOP

Table 5.1 describes the difference between RDBMS and Hadoop.

Table 5.1 RDBMS versus Hadoop

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System.	Node Based Flat Structure.
Data	Suitable for structured data.	Suitable for structured, unstructured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats, etc.
Processing	OLTP	Analytical, Big Data Processing
Choice	When the data needs consistent relationship.	Big Data processing, which does not require any consistent relationships between data.
Processor	Needs expensive hardware or high-end processors to store huge volumes of data.	In a Hadoop Cluster, a node requires only a processor, a network card, and few hard drives.
Cost	Cost around $10,000 to $14,000 per terabytes of storage.	Cost around $4,000 per terabytes of storage.

5.5 DISTRIBUTED COMPUTING CHALLENGES

Although there are several challenges with distributed computing, we will focus on two major challenges.

5.5.1 Hardware Failure

In a distributed system, several servers are networked together. This implies that more often than not, there may be a possibility of hardware failure. And when such a failure does happen, how does one retrieve the

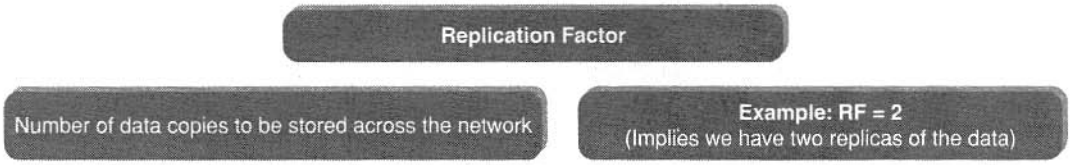


Figure 5.5 Replication factor.

data that was stored in the system? Just to explain further – a regular hard disk may fail once in 3 years. And when you have 1000 such hard disks, there is a possibility of at least a few being down every day.

Hadoop has an answer to this problem in **Replication Factor (RF)**. **Replication Factor** connotes the number of data copies of a given data item/data block stored across the network. Refer Figure 5.5.

JUST TO UNDERSTAND REPLICATION FURTHER, PICTURE THIS...

You work in a project team. There are six other members in the team. Each time there is an update related to the project work or an input received from the client, the project manager, Alex, ensures that he keeps at least three team members aware of the developments. You have been wondering at this style of working of your project manager. One day during the coffee break, when the project manager joins for coffee, you hesitantly ask him the question. Alex, "I had this question for you. Why is that each time we have an input from the client or any important piece of information, you leave it with at least three of our team members?" Alex smiled as he answered, "The reason is very simple. Assume that the client called and suggested some modification to the project. I shared it with just one person, let us say, person X. Tomorrow, when the suggested changes have to be incorporated, person X calls in sick. He is indisposed and not in office. Will that lead to our project coming to a standstill? Yes, isn't it? Therefore I share it with at least three team members, so that even if one is on leave or out of office for some reason, our work will not be stalled."

5.5.2 How to Process This Gigantic Store of Data?

In a distributed system, the data is spread across the network on several machines. A key challenge here is to integrate the data available on several machines prior to processing it.

Hadoop solves this problem by using **MapReduce** Programming. It is a programming model to process the data (MapReduce programming will be discussed a little later).

5.6 HISTORY OF HADOOP

Hadoop was created by Doug Cutting, the creator of Apache Lucene (a commonly used text search library). Hadoop is a part of the Apache Nutch (Yahoo) project (an open-source web search engine) and also a part of the Lucene project. Refer Figure 5.6 for more details.

5.6.1 The Name "Hadoop"

The name Hadoop is not an acronym; it's a made-up name. The project creator, Doug Cutting, explains how the name came about: "*The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid's term*".

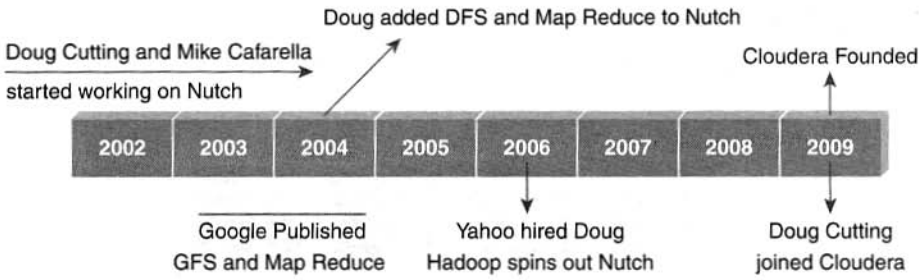


Figure 5.6 Hadoop history.

Subprojects and "contrib" modules in Hadoop also tend to have names that are unrelated to their function, often with an elephant or other animal theme ("Pig", for example).

Reference: Hadoop, The Definitive Guide, 3rd Edition, O'Reilly Publication Page. No. 9.

5.7 HADOOP OVERVIEW

Open-source software framework to store and process massive amounts of data in a distributed fashion on large clusters of commodity hardware. Basically, Hadoop accomplishes two tasks:

1. Massive data storage.
2. Faster data processing.

5.7.1 Key Aspects of Hadoop

Figure 5.7 describes the key aspects of Hadoop.

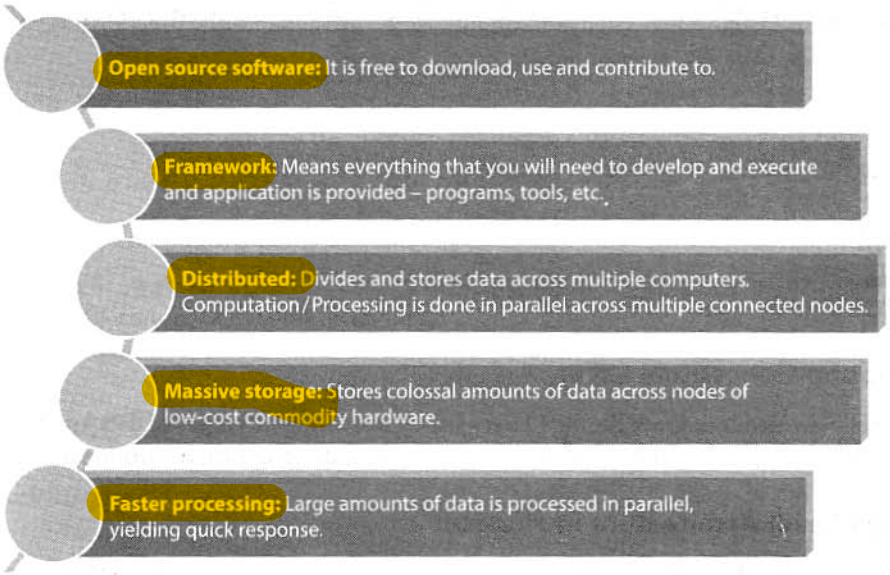


Figure 5.7 Key aspects of Hadoop.

5.7.2 Hadoop Components

Figure 5.8 depicts the Hadoop components.

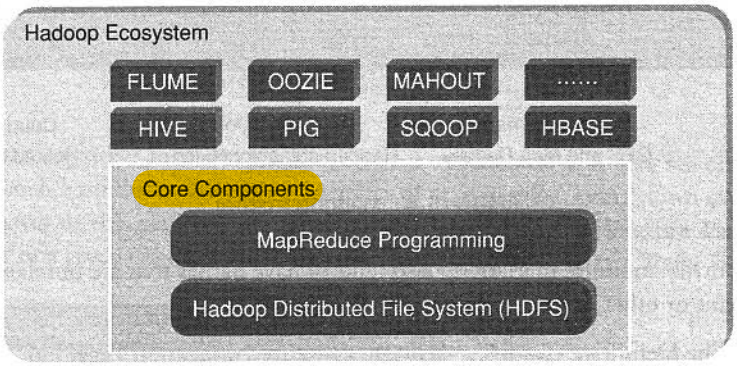


Figure 5.8 Hadoop components.

Hadoop Core Components

1. **HDFS:**
   (a) Storage component.
   (b) Distributes data across several nodes.
   (c) Natively redundant.
2. **MapReduce:**
   (a) Computational framework.
   (b) Splits a task across multiple nodes.
   (c) Processes data in parallel.

Hadoop Ecosystem: Hadoop Ecosystem are support projects to enhance the functionality of Hadoop Core Components. The Eco Projects are as follows:

1. HIVE
2. PIG
3. SQOOP
4. HBASE
5. FLUME
6. OOZIE
7. MAHOUT

5.7.3 Hadoop Conceptual Layer

It is conceptually divided into **Data Storage Layer** which stores huge volumes of data and **Data Processing Layer** which processes data in parallel to extract richer and meaningful insights from data (Figure 5.9).

5.7.4 High-Level Architecture of Hadoop

Hadoop is a distributed **Master-Slave** Architecture. Master node is known as **NameNode** and slave nodes are known as **DataNodes.** Figure 5.10 depicts the Master–Slave Architecture of Hadoop Framework.

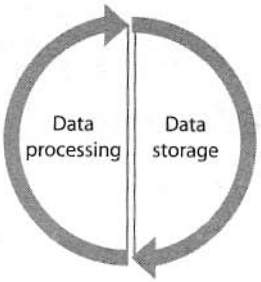


Figure 5.9 Hadoop conceptual layer.

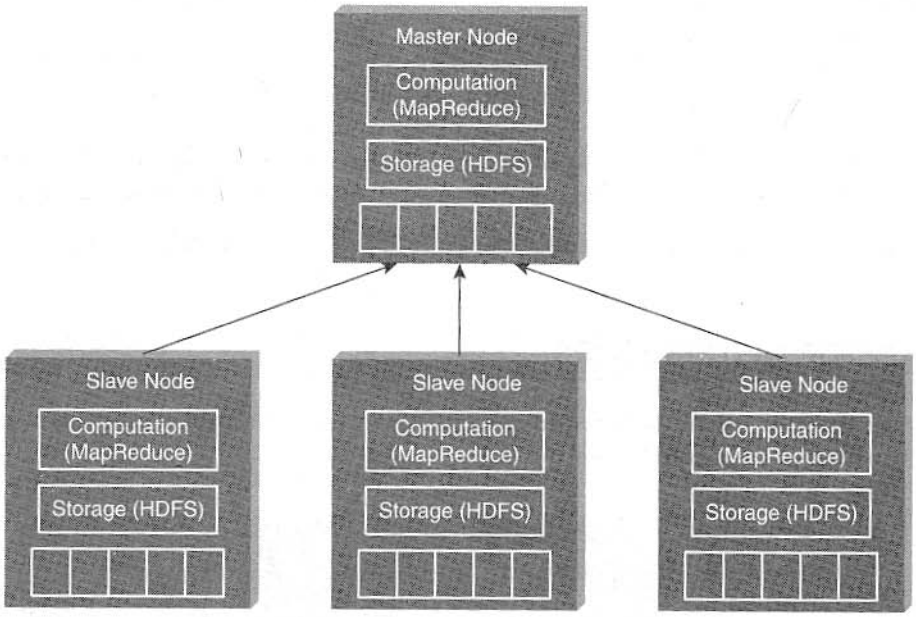


Figure 5.10 Hadoop high-level architecture.
Reference: Hadoop in Practice, Alex Holmes.

Let us look at the key components of the Master Node.

1. **Master HDFS:** Its main responsibility is partitioning the data storage across the slave nodes. It also keeps track of locations of data on DataNodes.
2. **Master MapReduce:** It decides and schedules computation task on slave nodes.

5.8 USE CASE OF HADOOP

5.8.1 ClickStream Data

ClickStream data (mouse clicks) helps you to understand the purchasing behavior of customers. ClickStream analysis helps online marketers to optimize their product web pages, promotional content, etc. to improve their business.

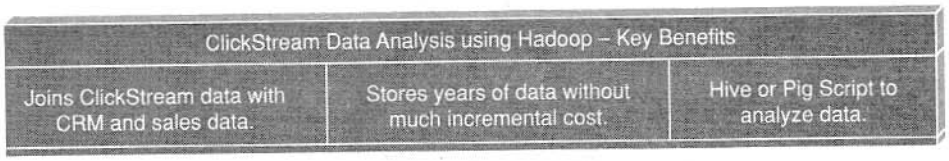


Figure 5.11 ClickStream data analysis.

The ClickStream analysis (Figure 5.11) using Hadoop provides **three key benefits**:

1. Hadoop helps to join ClickStream data with other data sources such as Customer Relationship Management Data (Customer Demographics Data, Sales Data, and Information on Advertising Campaigns). This additional data often provides the much needed information to understand customer behavior.
2. Hadoop's scalability property helps you to store years of data without ample incremental cost. This helps you to perform temporal or year over year analysis on ClickStream data which your competitors may miss.
3. Business analysts can use **Apache Pig** or **Apache Hive** for website analysis. With these tools, you can organize ClickStream data by user session, refine it, and feed it to visualization or analytics tools.

Reference: http://hortonworks.com/wp-content/uploads/2014/05/Hortonworks.BusinessValueofHadoop.v1.0.pdf

5.9 HADOOP DISTRIBUTORS

The companies shown in Figure 5.12 provide products that include Apache Hadoop, commercial support, and/or tools and utilities related to Hadoop.

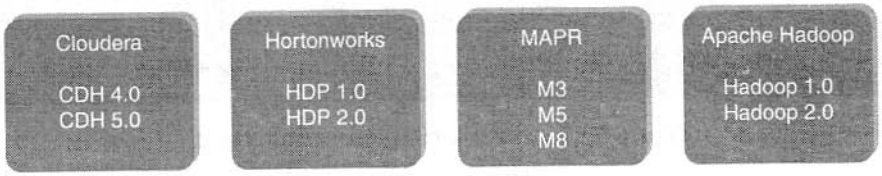


Figure 5.12 Common Hadoop distributors.

5.10 HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

Some key Points of Hadoop Distributed File System are as follows:

1. Storage component of Hadoop.
2. Distributed File System.
3. Modeled after Google File System.
4. Optimized for high throughput (HDFS leverages large block size and moves computation where data is stored).
5. You can replicate a file for a configured number of times, which is tolerant in terms of both software and hardware.

6. Re-replicates data blocks automatically on nodes that have failed.
7. You can realize the power of HDFS when you perform read or write on large files (gigabytes and larger).
8. Sits on top of native file system such as ext3 and ext4, which is described in Figure 5.13.

Figure 5.14 describes important key points of HDFS. Figure 5.15 describes Hadoop Distributed File System Architecture. Client Application interacts with NameNode for metadata related activities and communicates with DataNodes to read and write files. DataNodes converse with each other for pipeline reads and writes.

Let us assume that the file "Sample.txt" is of size **192 MB**. As per the default data block size (64 MB), it will be split into three blocks and replicated across the nodes on the cluster based on the default replication factor.

5.10.1 HDFS Daemons

5.10.1.1 NameNode

HDFS breaks a large file into smaller pieces called **blocks**. NameNode uses a **rack ID** to identify DataNodes in the rack. A rack is a collection of DataNodes within the cluster. NameNode keeps tracks of blocks of a file as it is placed on various DataNodes. NameNode manages file-related operations such as read, write, create, and delete. Its main job is managing the **File System Namespace**. A file system namespace is collection of files in the cluster. NameNode stores HDFS namespace. File system namespace includes mapping of blocks to file, file properties and is stored in a file called **FsImage**. NameNode uses an EditLog (transaction log) to record every transaction that happens to the file system metadata. Refer Figure 5.16. When NameNode starts up, it reads FsImage and EditLog from disk and applies all transactions from the EditLog to in-memory representation of the FsImage. Then it flushes out new version of FsImage on disk and truncates the old EditLog because the changes are updated in the FsImage. There is a single NameNode per cluster.

Reference: http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html

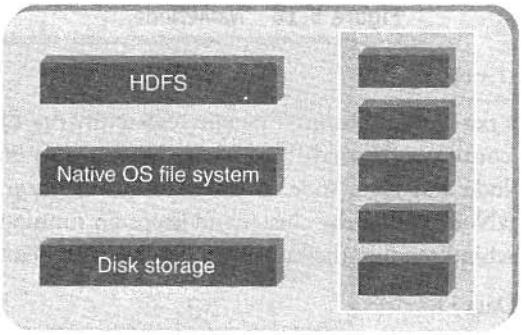


Figure 5.13 Hadoop Distributed File System.

Figure 5.14 Hadoop Distributed File System – key points.

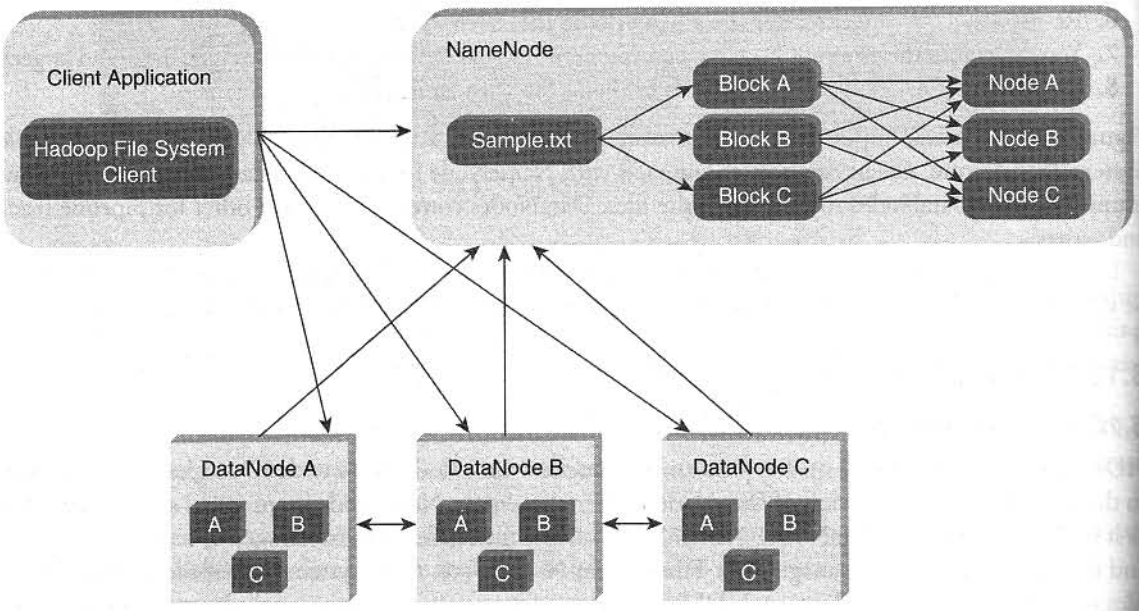


Figure 5.15 Hadoop Distributed File System Architecture.
Reference: Hadoop in Practice, Alex Holmes.

NameNode – Manages File Related Operations	
FsImage – File, in which entire file system is stored.	EditLog – Records every transaction that occurs to file system metadata.

Figure 5.16 NameNode.

5.10.1.2 DataNode

There are multiple DataNodes per cluster. During Pipeline read and write DataNodes communicate with each other. A DataNode also continuously sends **"heartbeat"** message to NameNode to ensure the connectivity between the NameNode and DataNode. In case there is no heartbeat from a DataNode, the NameNode replicates that DataNode within the cluster and keeps on running as if nothing had happened.

Let us explain the concept behind sending the heartbeat report by the DataNodes to the NameNode.

Reference: Wrox Certified Big Data Developer.

PICTURE THIS...

You work for a renowned IT organization. Every day when you come to office, you are required to swipe in to record your attendance. This record of attendance is then shared with your manager to keep him posted on who all from his team have reported for work. Your manager is able to allocate tasks to the team members who are present in office. The tasks for the day cannot be allocated to team members who have not turned in. Likewise heartbeat report is a way by which DataNodes inform the NameNode that they are up and functional and can be assigned tasks. Figure 5.17 depicts the above scenario.

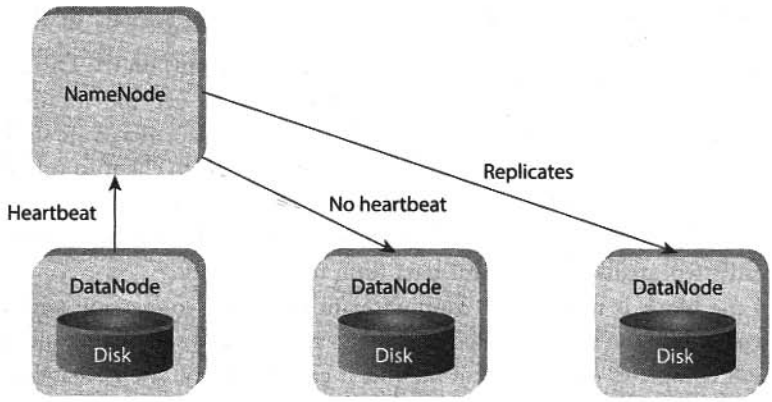


Figure 5.17 NameNode and DataNode Communication.

5.10.1.3 *Secondary NameNode*

The Secondary NameNode takes a snapshot of HDFS metadata at intervals specified in the Hadoop configuration. Since the memory requirements of Secondary NameNode are the same as NameNode, it is better to run NameNode and Secondary NameNode on different machines. In case of failure of the NameNode, the Secondary NameNode can be configured manually to bring up the cluster. However, the Secondary NameNode does not record any real-time changes that happen to the HDFS metadata.

5.10.2 Anatomy of File Read

Figure 5.18 describes the anatomy of File Read.

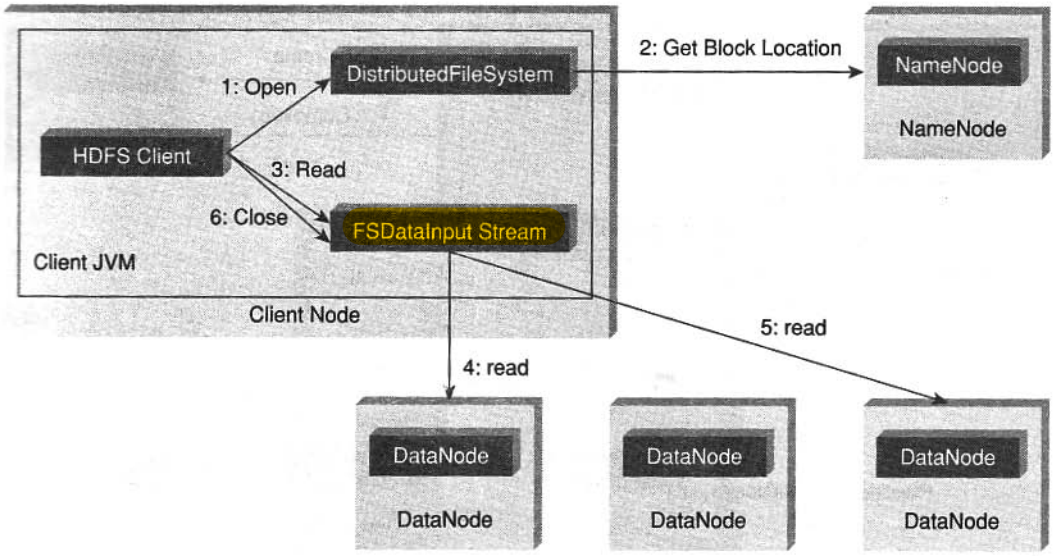


Figure 5.18 File Read.

The steps involved in the File Read are as follows:

1. The client opens the file that it wishes to read from by calling open() on the DistributedFileSystem.
2. DistributedFileSystem communicates with the NameNode to get the location of data blocks. NameNode returns with the addresses of the DataNodes that the data blocks are stored on. Subsequent to this, the DistributedFileSystem returns an FSDataInputStream to client to read from the file.
3. Client then calls read() on the stream DFSInputStream, which has addresses of the DataNodes for the first few blocks of the file, connects to the closest DataNode for the first block in the file.
4. Client calls read() repeatedly to stream the data from the DataNode.
5. When end of the block is reached, DFSInputStream closes the connection with the DataNode. It repeats the steps to find the best DataNode for the next block and subsequent blocks.
6. When the client completes the reading of the file, it calls close() on the FSDataInputStream to close the connection.

Reference: Hadoop, The Definitive Guide, 3rd Edition, O'Reilly Publication.

5.10.3 Anatomy of File Write

Figure 5.19 describes the anatomy of File Write. The steps involved in anatomy of File Write are as follows:

1. The client calls create() on DistributedFileSystem to create a file.
2. An RPC call to the NameNode happens through the DistributedFileSystem to create a new file. The NameNode performs various checks to create a new file (checks whether such a file exists or not). Initially, the NameNode creates a file without associating any data blocks to the file. The DistributedFileSystem returns an FSDataOutputStream to the client to perform write.
3. As the client writes data, data is split into packets by DFSOutputStream, which is then written to an internal queue, called *data queue*. DataStreamer consumes the data queue. The DataStreamer requests

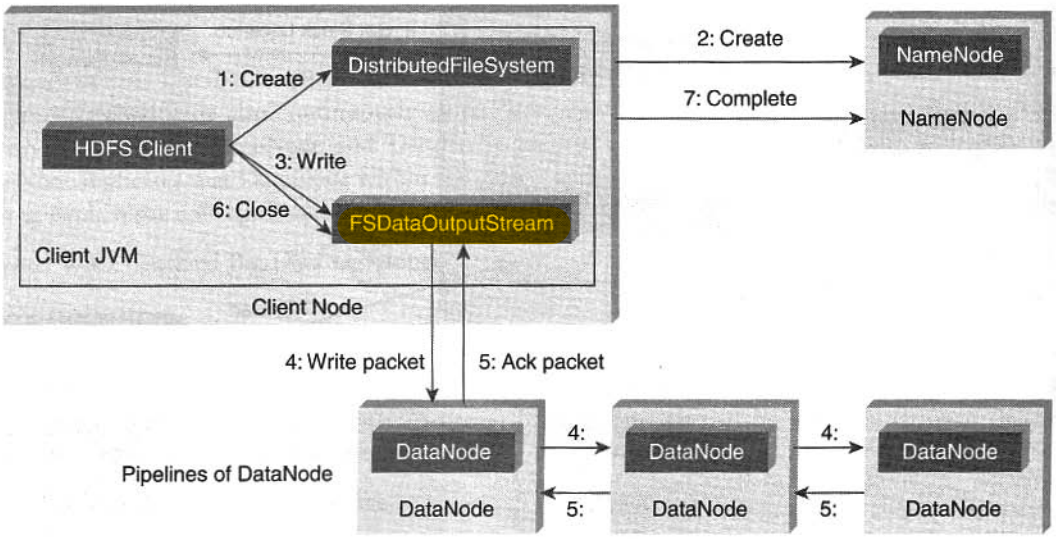


Figure 5.19 File Write.

the NameNode to allocate new blocks by selecting a list of suitable DataNodes to store replicas. This list of DataNodes makes a pipeline. Here, we will go with the default replication factor of three, so there will be three nodes in the pipeline for the first block.

4. DataStreamer streams the packets to the first DataNode in the pipeline. It stores packet and forwards it to the second DataNode in the pipeline. In the same way, the second DataNode stores the packet and forwards it to the third DataNode in the pipeline.
5. In addition to the internal queue, DFSOutputStream also manages an "Ack queue" of packets that are waiting for the acknowledgement by DataNodes. A packet is removed from the "Ack queue" only if it is acknowledged by all the DataNodes in the pipeline.
6. When the client finishes writing the file, it calls close() on the stream.
7. This flushes all the remaining packets to the DataNode pipeline and waits for relevant acknowledgments before communicating with the NameNode to inform the client that the creation of the file is complete.

Reference: Hadoop, The Definitive Guide, 3rd Edition, O'Reilly Publication.

5.10.4 Replica Placement Strategy

5.10.4.1 *Hadoop Default Replica Placement Strategy*

As per the Hadoop Replica Placement Strategy, first replica is placed on the same node as the client. Then it places second replica on a node that is present on different rack. It places the third replica on the same rack as second, but on a different node in the rack. Once replica locations have been set, a pipeline is built. This strategy provides good reliability. Figure 5.20 describes the typical replica pipeline.

Reference: Hadoop, the Definite Guide, 3rd Edition, O'Reilly Publication.

5.10.5 Working with HDFS Commands

Objective: To get the list of directories and files at the root of HDFS.

Act:

hadoop fs -ls /

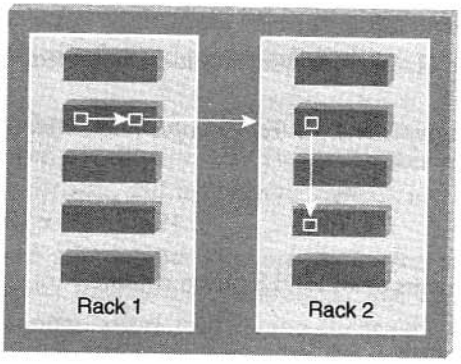


Figure 5.20 Replica Placement Strategy.

Objective: To get the list of complete directories and files of HDFS.
Act:
hadoop fs -ls -R /

Objective: To create a directory (say, sample) in HDFS.
Act:
hadoop fs -mkdir /sample

Objective: To copy a file from local file system to HDFS.
Act:
hadoop fs -put /root/sample/test.txt /sample/test.txt

Objective: To copy a file from HDFS to local file system.
Act:
hadoop fs -get /sample/test.txt /root/sample/testsample.txt

Objective: To copy a file from local file system to HDFS via copyFromLocal command.
Act:
hadoop fs -copyFromLocal /root/sample/test.txt /sample/testsample.txt

Objective: To copy a file from Hadoop file system to local file system via copyToLocal command.
Act:
hadoop fs -copyToLocal /sample/test.txt /root/sample/testsample1.txt

Objective: To display the contents of an HDFS file on console.
Act:
hadoop fs -cat /sample/test.txt

Objective: To copy a file from one directory to another on HDFS.
Act:
hadoop fs -cp /sample/test.txt /sample1

Objective: To remove a directory from HDFS.
Act:
hadoop fs-rm-r /sample1

5.10.6 Special Features of HDFS

1. **Data Replication:** There is absolutely no need for a client application to track all blocks. It directs the client to the nearest replica to ensure high performance.
2. **Data Pipeline:** A client application writes a block to the first DataNode in the pipeline. Then this DataNode takes over and forwards the data to the next node in the pipeline. This process continues for all the data blocks, and subsequently all the replicas are written to the disk.

Reference: Wrox Certified Big Data Developer.

5.11 PROCESSING DATA WITH HADOOP

MapReduce Programming is a software framework. MapReduce Programming helps you to process massive amounts of data in parallel.

In MapReduce Programming, the input dataset is split into independent chunks. **Map tasks** process these independent chunks completely in a parallel manner. The output produced by the map tasks serves as intermediate data and is stored on the local disk of that server. The output of the mappers are automatically shuffled and sorted by the framework. MapReduce Framework sorts the output based on **keys**. This sorted output becomes the input to the **reduce tasks**. Reduce task provides reduced output by combining the output of the various mappers. Job inputs and outputs are stored in a file system. MapReduce framework also takes care of the other tasks such as scheduling, monitoring, re-executing failed tasks, etc.

Hadoop Distributed File System and MapReduce Framework run on the same set of nodes. This configuration allows effective scheduling of tasks on the nodes where data is present (**Data Locality**). This in turn results in very high throughput.

There are two daemons associated with MapReduce Programming. A single master **JobTracker** per cluster and one slave **TaskTracker** per cluster-node. The JobTracker is responsible for scheduling tasks to the TaskTrackers, monitoring the task, and re-executing the task just in case the TaskTracker fails. The TaskTracker executes the task. Refer Figure 5.21.

The MapReduce functions and input/output locations are implemented via the MapReduce applications. These applications use suitable interfaces to construct the job. The application and the job parameters together are known as **job configuration**. Hadoop **job client** submits job (jar/executable, etc.) to the JobTracker. Then it is the responsibility of JobTracker to schedule tasks to the slaves. In addition to scheduling, it also monitors the task and provides status information to the job-client.

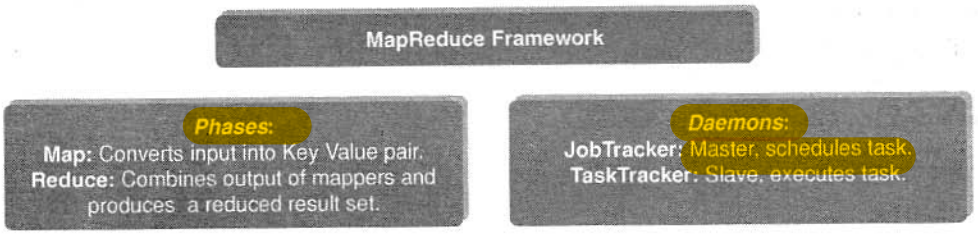


Figure 5.21 MapReduce Programming phases and daemons.

Reference: http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html

5.11.1 MapReduce Daemons

1. **JobTracker:** It provides connectivity between Hadoop and your application. When you submit code to cluster, JobTracker creates the execution plan by deciding which task to assign to which node. It also monitors all the running tasks. When a task fails, it automatically re-schedules the task to a different node after a predefined number of retries. JobTracker is a master daemon responsible for executing overall MapReduce job. There is a single JobTracker per Hadoop cluster.
2. **TaskTracker:** This daemon is responsible for executing individual tasks that is assigned by the JobTracker. There is a single TaskTracker per slave and spawns multiple Java Virtual Machines (JVMs) to handle multiple map or reduce tasks in parallel. TaskTracker continuously sends heartbeat message to JobTracker. When the JobTracker fails to receive a heartbeat from a TaskTracker, the JobTracker assumes that the TaskTracker has failed and resubmits the task to another available node in the cluster. Once the client submits a job to the JobTracker, it partitions and assigns diverse MapReduce tasks for each TaskTracker in the cluster. Figure 5.22 depicts JobTracker and TaskTracker interaction.

Reference: Hadoop in Action, Chuck Lam.

5.11.2 How Does MapReduce Work?

MapReduce divides a data analysis task into two parts – **map** and **reduce**. Figure 5.23 depicts how the MapReduce Programming works. In this example, there are two mappers and one reducer. Each mapper

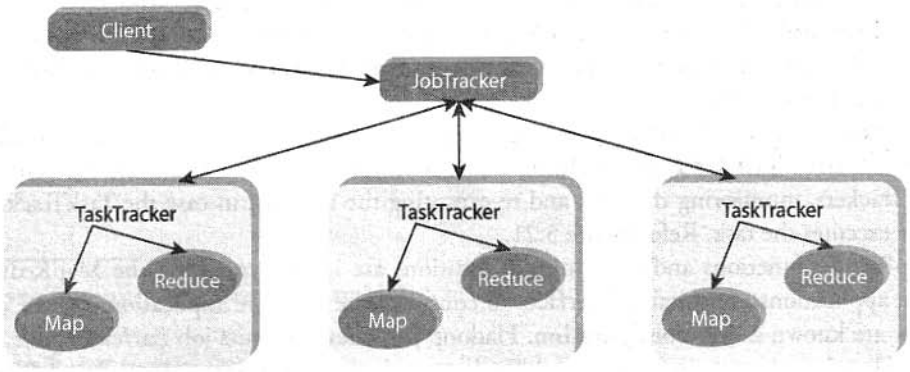


Figure 5.22 JobTracker and TaskTracker interaction.

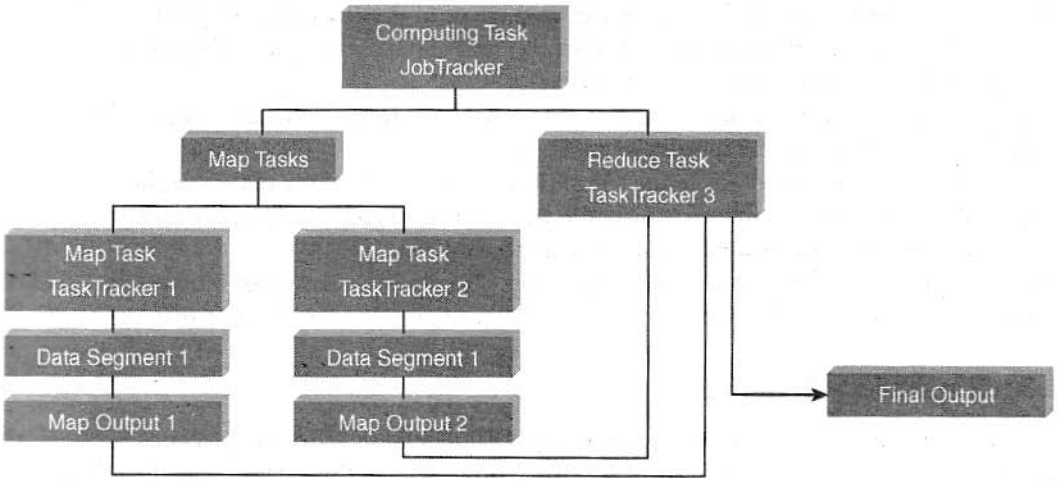


Figure 5.23 MapReduce programming workflow.

works on the partial dataset that is stored on that node and the reducer combines the output from the mappers to produce the reduced result set.

Reference: Wrox Big Data Certification Materials.

Figure 5.24 describes the working model of MapReduce Programming. The following steps describe how MapReduce performs its task.

1. First, the input dataset is split into multiple pieces of data (several small subsets).
2. Next, the framework creates a master and several workers processes and executes the worker processes remotely.

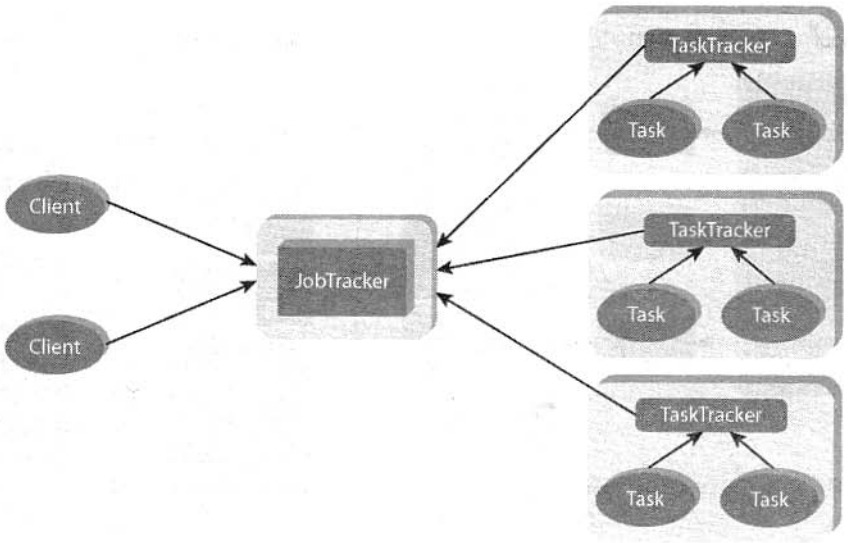


Figure 5.24 MapReduce programming architecture.

3. Several map tasks work simultaneously and read pieces of data that were assigned to each map task. The map worker uses the map function to extract only those data that are present on their server and generates key/value pair for the extracted data.
4. Map worker uses partitioner function to divide the data into regions. Partitioner decides which reducer should get the output of the specified mapper.
5. When the map workers complete their work, the master instructs the reduce workers to begin their work. The reduce workers in turn contact the map workers to get the key/value data for their partition. The data thus received is shuffled and sorted as per keys.
6. Then it calls reduce function for every unique key. This function writes the output to the file.
7. When all the reduce workers complete their work, the master transfers the control to the user program.

5.11.3 MapReduce Example

The famous example for MapReduce Programming is **Word Count.** For example, consider you need to count the occurrences of similar words across 50 files. You can achieve this using MapReduce Programming. Refer Figure 5.25.

Word Count MapReduce Programming using Java

The MapReduce Programming requires three things.

1. **Driver Class:** This class specifies **Job Configuration** details.
2. **Mapper Class:** This class overrides the **Map Function** based on the problem statement.
3. **Reducer Class:** This class overrides the **Reduce Function** based on the problem statement.

Wordcounter.java: Driver Program

```
package com.app;
import java.io.IOException;
```

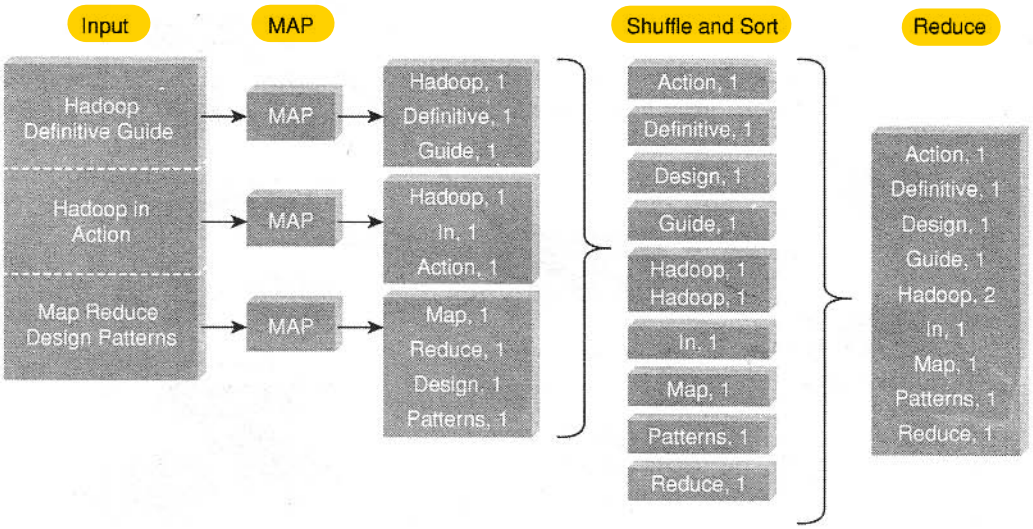


Figure 5.25 Wordcount example.

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordCounter {

public static void main (String [] args) throws IOException,
InterruptedException, ClassNotFoundException {
          Job job = new Job ();
          job.setJobName ("wordcounter");
          job.setJarByClass (WordCounter.class);
          job.setMapperClass (WordCounterMap.class);
          job.setReducerClass (WordCounterRed.class);
          job.setOutputKeyClass (Text.class);
          job.setOutputValueClass (IntWritable.class);

          FileInputFormat.addInputPath (job, new Path ("/sample/word.
txt"));
          FileOutputFormat.setOutputPath (job, new Path ("/sample/
wordcount"));
          System.exit (job.waitForCompletion (true)? 0: 1);

     }
}
```

WordCounterMap.java: Map Class

```
package com.app;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class WordCounterMap extends Mapper<LongWritable, Text, Text,
IntWritable> {
     @Override
```

```
    protected void map (LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        String [] words=value.toString ().split (",");
        for (String word: words) {
            context.write (new Text (word), new IntWritable (1));
        }
    }
}
```

WordCountReduce.java: Reduce Class

```
package com.infosys;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class WordCounterRed extends Reducer<Text, IntWritable, Text,
IntWritable>{

    @Override
    protected void reduce(Text word, Iterable<IntWritable> values,
Context context)
            throws IOException, InterruptedException {
        Integer count = 0;
        for(IntWritable val: values){
            count += val.get();
        }
        context.write(word, new IntWritable(count));
    }
}
```

Table 5.2 describes differences between SQL and MapReduce.

Table 5.2 SQL versus MapReduce

	SQL	MapReduce
Access	Interactive and Batch	Batch
Structure	Static	Dynamic
Updates	Read and write many times	Write once, read many times
Integrity	High	Low
Scalability	Nonlinear	Linear

5.12 MANAGING RESOURCES AND APPLICATIONS WITH HADOOP YARN (YET ANOTHER RESOURCE NEGOTIATOR)

Apache Hadoop YARN is a sub-project of Hadoop 2.x. Hadoop 2.x is YARN-based architecture. It is a general processing platform. YARN is not constrained to MapReduce only. You can run multiple applications in Hadoop 2.x in which all applications share a common resource management. Now Hadoop can be used for various types of processing such as Batch, Interactive, Online, Streaming, Graph, and others.

5.12.1 Limitations of Hadoop 1.0 Architecture

In Hadoop 1.0, HDFS and MapReduce are Core Components, while other components are built around the core.

1. Single NameNode is responsible for managing entire namespace for Hadoop Cluster.
2. It has a restricted processing model which is suitable for batch-oriented MapReduce jobs.
3. Hadoop MapReduce is not suitable for interactive analysis.
4. Hadoop 1.0 is not suitable for machine learning algorithms, graphs, and other memory intensive algorithms.
5. **MapReduce** is responsible for **cluster resource management and data processing**.

In this Architecture, **map slots might be "full", while the reduce slots are empty and vice versa**. This causes **resource utilization issues**. This needs to be improved for proper resource utilization.

5.12.2 HDFS Limitation

NameNode saves all its file metadata in main memory. Although the main memory today is not as small and as expensive as it used to be two decades ago, still there is a limit on the number of objects that one can have in the memory on a single NameNode. The NameNode can quickly become overwhelmed with load on the system increasing.

In Hadoop 2.x, this is resolved with the help of **HDFS Federation**.

5.12.3 Hadoop 2: HDFS

HDFS 2 consists of two major components: **(a) namespace, (b) blocks storage service**. Namespace service takes care of file-related operations, such as creating files, modifying files, and directories. The block storage service handles data node cluster management, replication.

HDFS 2 Features

1. Horizontal scalability.
2. High availability.

HDFS Federation uses multiple independent NameNodes for horizontal scalability. NameNodes are independent of each other. It means, NameNodes does not need any coordination with each other. The DataNodes are common storage for blocks and shared by all NameNodes. All DataNodes in the cluster registers with each NameNode in the cluster.

High availability of NameNode is obtained with the help of **Passive Standby NameNode**. In Hadoop 2.x, Active–Passive NameNode handles failover automatically. All namespace edits are recorded to a shared NFS storage and there is a single writer at any point of time. Passive NameNode reads edits from shared storage

and keeps updated metadata information. In case of Active NameNode failure, Passive NameNode becomes an Active NameNode automatically. Then it starts writing to the shared storage. Figure 5.26 describes the Active–Passive NameNode interaction.

Reference: http://www.edureka.co/blog/introduction-to-hadoop-2-0-and-advantages-of-hadoop-2-0/

Figure 5.27 depicts Hadoop 1.0 and Hadoop 2.0 architecture.

5.12.4 Hadoop 2 YARN: Taking Hadoop beyond Batch

YARN helps us to store all data in one place. We can interact in multiple ways to get predictable performance and quality of services. This was originally architected by **Yahoo**. Refer Figure 5.28.

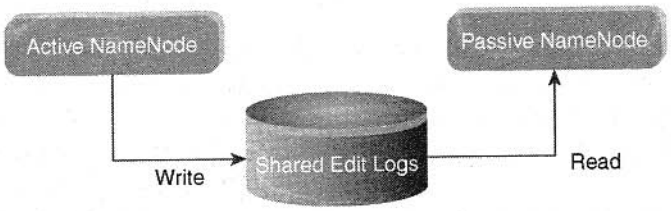


Figure 5.26 Active and Passive NameNode interaction.

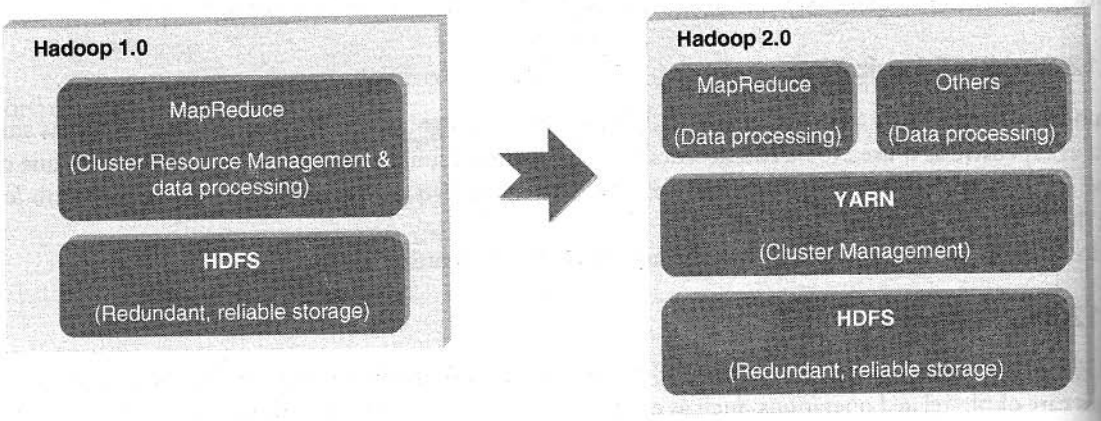


Figure 5.27 Hadoop 1.x versus Hadoop 2.x.

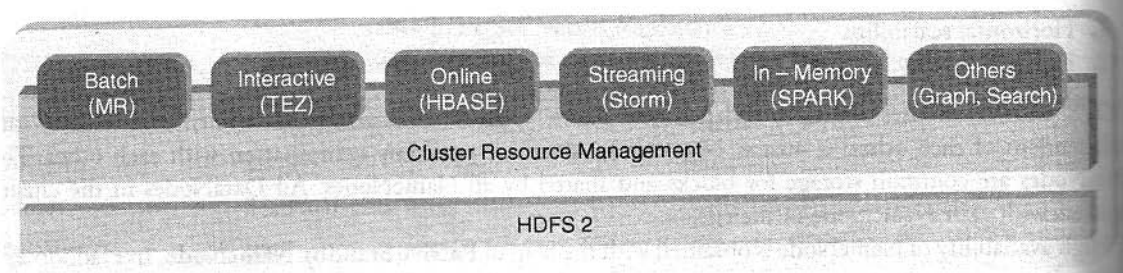


Figure 5.28 Hadoop YARN.

5.12.4.1 Fundamental Idea

The fundamental idea behind this architecture is splitting the JobTracker responsibility of resource management and Job Scheduling/Monitoring into separate daemons. Daemons that are part of YARN Architecture are described below.

1. **A Global ResourceManager:** Its main responsibility is to distribute resources among various applications in the system. It has two main components:
   (a) ***Scheduler:*** The pluggable scheduler of ResourceManager decides allocation of resources to various running applications. The scheduler is just that, a pure scheduler, meaning it does NOT monitor or track the status of the application.
   (b) ***ApplicationManager:*** ApplicationManager does the following:
   - Accepting job submissions.
   - Negotiating resources (container) for executing the application specific ApplicationMaster.
   - Restarting the ApplicationMaster in case of failure.
2. **NodeManager:** This is a per-machine slave daemon. NodeManager responsibility is launching the application containers for application execution. NodeManager monitors the resource usage such as memory, CPU, disk, network, etc. It then reports the usage of resources to the global ResourceManager.
3. **Per-application ApplicationMaster:** This is an application-specific entity. Its responsibility is to negotiate required resources for execution from the ResourceManager. It works along with the NodeManager for executing and monitoring component tasks.

5.12.4.2 Basic Concepts

Application:

1. Application is a job submitted to the framework.
2. Example – **MapReduce Job**.

Container:

1. Basic unit of allocation.
2. Fine-grained resource allocation across multiple resource types (Memory, CPU, disk, network, etc.)
   (a) container_0 = 2GB, 1CPU
   (b) container_1 = 1GB, 6 CPU
3. Replaces the fixed map/reduce slots.

YARN Architecture:

Figure 5.29 depicts YARN architecture. The steps involved in YARN architecture are as follows:

1. A client program submits the application which includes the necessary specifications to launch the application-specific **ApplicationMaster** itself.
2. The ResourceManager launches the ApplicationMaster by assigning some container.
3. The ApplicationMaster, on boot-up, registers with the ResourceManager. This helps the client program to query the ResourceManager directly for the details.
4. During the normal course, ApplicationMaster negotiates appropriate resource containers via the resource-request protocol.

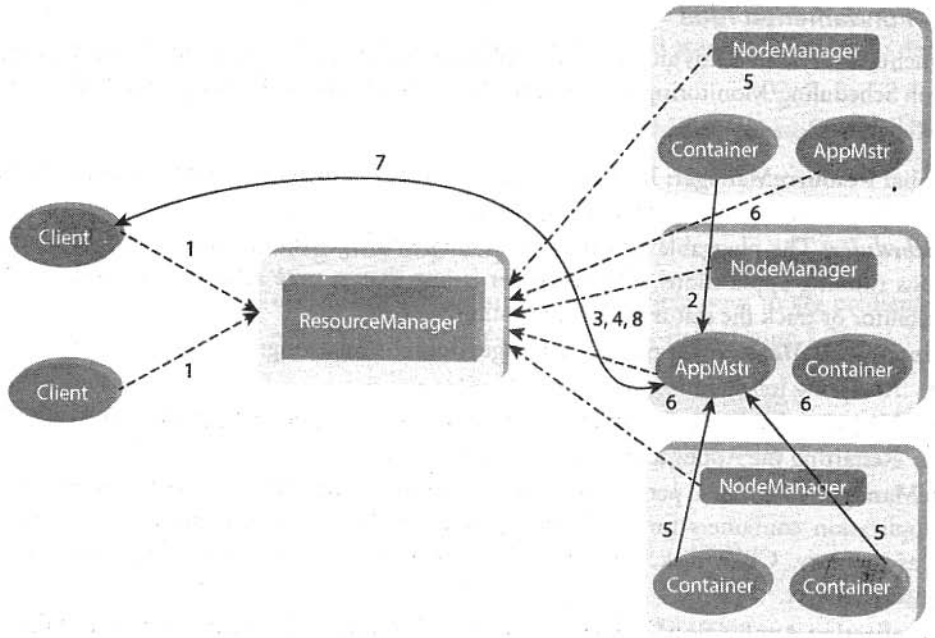


Figure 5.29 YARN architecture.

5. On successful container allocations, the ApplicationMaster launches the container by providing the container launch specification to the NodeManager.
6. The NodeManager executes the application code and provides necessary information such as progress, status, etc. to it's ApplicationMaster via an application-specific protocol.
7. During the application execution, the client that submitted the job directly communicates with the ApplicationMaster to get status, progress updates, etc. via an application-specific protocol.
8. Once the application has been processed completely, ApplicationMaster deregisters with the ResourceManager and shuts down, allowing its own container to be repurposed.

Reference: http://hortonworks.com/blog/apache-hadoop-yarn-background-and-an-overview/

5.13 INTERACTING WITH HADOOP ECOSYSTEM

Hadoop ecosystem was introduced in Chapter 4. Here we will look at it in more detail.

5.13.1 Pig

Pig is a data flow system for Hadoop. It uses Pig Latin to specify data flow. Pig is an alternative to MapReduce Programming. It abstracts some details and allows you to focus on data processing. It consists of two components.

1. **Pig Latin:** The data processing language.
2. **Compiler:** To translate Pig Latin to MapReduce Programming.

Figure 5.30 depicts the Pig in the Hadoop ecosystem.

5.13.2 Hive

Hive is a Data Warehousing Layer on top of Hadoop. Analysis and queries can be done using an SQL-like language. Hive can be used to do ad-hoc queries, summarization, and data analysis. Figure 5.31 depicts Hive in the Hadoop ecosystem.

5.13.3 Sqoop

Sqoop is a tool which helps to transfer data between Hadoop and Relational Databases. With the help of Sqoop, you can import data from RDBMS to HDFS and vice-versa. Figure 5.32 depicts the Sqoop in Hadoop ecosystem.

5.13.4 HBase

HBase is a NoSQL database for Hadoop. HBase is column-oriented NoSQL database. HBase is used to store **billions of rows and millions of columns.** HBase provides random read/write operation. It also supports record level updates which is not possible using HDFS. HBase sits on top of HDFS. Figure 5.33 depicts the HBase in Hadoop ecosystem.

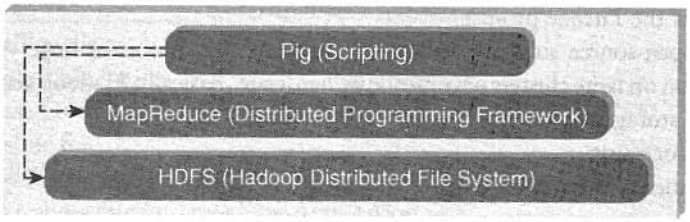


Figure 5.30 Pig in the Hadoop ecosystem.

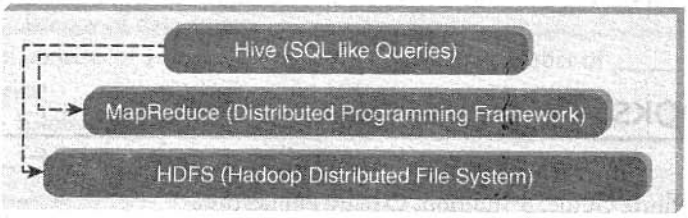


Figure 5.31 Hive in the Hadoop ecosystem.

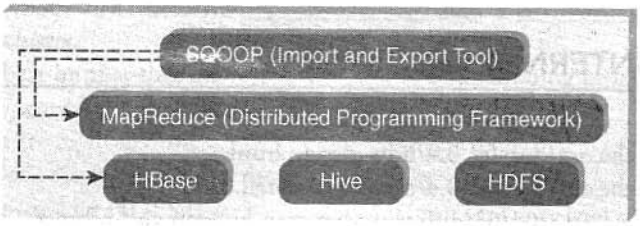


Figure 5.32 Sqoop in the Hadoop ecosystem.

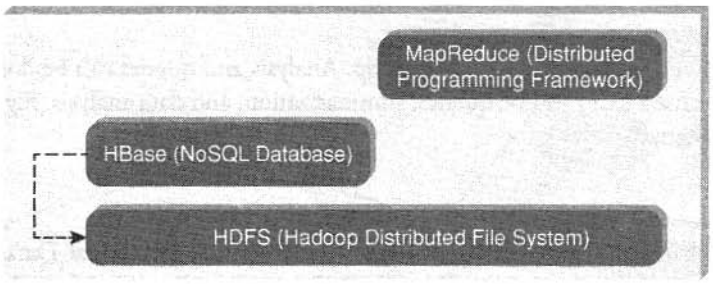


Figure 5.33 HBase in the Hadoop Ecosystem.

REMIND ME

- The key consideration (the rationale behind the huge popularity of Hadoop) is: ***Its capability to handle massive amounts of data, different categories of data – fairly quickly.***
- Hadoop was created by Doug Cutting, the creator of Apache Lucene (a commonly used text search library). Hadoop is a part of the Apache Nutch (Yahoo) project (an open-source web search engine) and also a part of the Lucene project.
- Hadoop is an open-source software framework. It stores and processes huge volumes of data in a distributed fashion on large clusters of commodity hardware. Basically, Hadoop accomplishes two tasks:
  - Massive data storage.
  - Faster data processing.
- The core components of Hadoop are:
  - HDFS
  - MapReduce
- Apache Hadoop YARN is a sub-project of Hadoop 2.x. Hadoop 2.x is YARN-based architecture. It provides general processing platform which is not constrained to MapReduce only.

POINT ME (BOOKS)

- Hadoop, the Definite Guide, 3rd Edition, O'reilly Publication.
- Hadoop in Practice, Alex Holmes.
- Hadoop in Action, Chuck Lam.

CONNECT ME (INTERNET RESOURCES)

- http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
- http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html
- http://oraclesys.com/2013/04/03/difference-between-hadoop-and-rdbms/

- http://hortonworks.com/blog/apache-hadoop-yarn-background-and-an-overview/
- http://www.tomsitpro.com/articles/hadoop-2-vs-1,2-718.html
- http://www.wikidifference.com/difference-between-hadoop-and-rdbms/
- http://www.edureka.co/blog/introduction-to-hadoop-2-0-and-advantages-of-hadoop-2-0/

TEST ME

A. Fill Me

1. Hadoop is _____ based flat structure.
2. RDBMS is best choice when _____ is the main concern.
3. Hadoop supports _____, _____ and _____ data formats.
4. RDBMS supports _____ data formats.
5. In Hadoop, data is processed in _____.
6. HDFS can be deployed on _____.
7. NameNode uses _____ to store file system namespace.
8. NameNode uses _____ to record every transaction.
9. Secondary NameNode is a _____ daemon.
10. DataNode is responsible for _____ file operation.
11. Hadoop 2.x is based on _____ architecture.
12. YARN is responsible for _____.
13. Global ResourceManager distributes _____ among applications.
14. NodeManager is responsible for launching Application _____.
15. Application is a _____ submitted to framework.
16. _____ is an open-source framework managed by Apache Software Foundations.
17. The emphasis of HDFS is on _____ throughput of data access rather than _____ latency of data access.
18. An HDFS cluster consists of a single _____ and a number of _____.
19. Complete the series:
    Bits → Bytes → Kilobytes → Megabytes → Gigabytes → _____ → _____ → _____ → _____ → Yottabytes
20. HDFS has a _____/_____ architecture.
21. HDFS is built using the _____ language.
22. The _____ maintains the file system Namespace.
23. The number of copies of a file is called the _____ of that file.
24. The NameNode periodically receives a _____ and a _____ from each of the DataNodes in the cluster.
25. Receipt of a Heartbeat implies that the _____ is functioning properly.
26. A _____ contains a list of all blocks on a DataNode.
27. The blocks of a file are replicated for _____ tolerance.
28. When the NameNode starts up, it reads the _____ and _____ from disk.
29. A typical block size used by HDFS is _____.
30. _____ are responsible for serving read and write requests from the file system's clients.

31. _____ perform block creation, deletion and replication upon instruction from the _____.
32. _____ was the first to publicize MapReduce – a system they had used to scale their data processing needs.
33. _____ developed an open-source version of MapReduce system called _____.
34. Hadoop is an open-source framework for writing and running _____ applications that process large amounts of data.
35. The key distinctions of Hadoop are _____, _____ and _____.
36. Hadoop runs on large clusters of _____.
37. Hadoop scales _____ to handle larger data by adding more _____ to the cluster.
38. Hadoop focusses on moving _____ to _____.
39. The move-code-to-data philosophy makes sense for _____ intensive processing.
40. Hadoop is designed to be a scale _____ architecture operating on cluster of commodity PC machines.
41. Hadoop uses _____ as its basic data unit, which is flexible enough to work with less-structured data types.
42. Hadoop is best used as a _____ once and _____ many times type of data store.
43. Under SQL we have _____ statements; under MapReduce we have _____ and _____.
44. Under the MapReduce Model, data processing primitives are called _____ and _____.
45. The Mapper is meant to _____ and _____ the input into something that the reducer can _____ over.
46. _____ and _____ are common design patterns that go along with mapping and reducing.
47. _____ is the official development and production platform for Hadoop.
48. _____ started out as a sub-project of _____, which in turn was a sub-project of _____.
49. _____ is a single point of failure of Hadoop cluster.
50. _____ is the bookkeeper of HDFS.
51. _____ keeps track of how your files are broken down into file blocks, which nodes store those blocks, and the overall health of the distributed file system.
52. _____ communicates with the NameNode to take snapshots of the HDFS metadata at intervals defined by the cluster configuration.
53. There is only one _____ daemon per Hadoop cluster.
54. There is a single _____ per slave node.

Answers:

1. Node
2. Consistency
3. Structured, semi-structured and unstructured
4. Structured
5. Parallel
6. Low cost hardware
7. FsImage
8. EditLog

9. Helper or House Keeping
10. Read/Write
11. YARN
12. Cluster Management
13. Resources
14. Containers
15. Job
16. Hadoop
17. High, Low
18. NameNode, DataNodes
19. Terabytes, Petabytes, Exabytes, Zettabytes
20. Master/slave
21. Java
22. NameNode
23. Replication factor
24. Heartbeat, Blockreport
25. DataNode
26. Blockreport
27. Fault
28. FsImage, EditLog
29. 64MB
30. DataNodes
31. DataNodes, NameNode
32. Google
33. Doug Cutting, Hadoop
34. Distributed
35. Accessible, Robust, and Scalable
36. Commodity machines
37. Linearly, nodes
38. Code, Data
39. Data
40. Out
41. Key/value
42. Write, read
43. Query, Scripts, and Code
44. Mappers, Reducers
45. Filter and transform, aggregate
46. Partitioning and Shuffling
47. Linux
48. Hadoop, Nutch, Apache Lucene
49. NameNode
50. NameNode
51. NameNode
52. Secondary NameNode
53. JobTracker
54. TaskTracker

Match Me

Column A	Column B
HDFS	DataNode
MapReduce Programming	NameNode
Master node	Processing Data
Slave node	Google File System and MapReduce
Hadoop Implementation	Storage

Answer:

Column A	Column B
HDFS	Storage
MapReduce Programming	Processing Data
Master node	NameNode
Slave node	DataNode
Hadoop Implementation	Google File System and MapReduce

2.

Column A	Column B
JobTracker	Executes Task
MapReduce	Schedules Task
TaskTracker	Programming Model
Job Configuration	Converts input into Key Value pair
Map	Job Parameters

Answer:

Column A	Column B
JobTracker	Schedules Task
MapReduce	Programming Model
TaskTracker	Executes Task
Job Configuration	Job Parameters
Map	Converts input into Key Value pair

3.

Column A	ColumnB
NameNode	Handles processing on master
JobTracker	Handles storage on slave
DataNode	Handles storage on master
TaskTracker	Handles processing on slave

Answer:

Column A	Column B
NameNode	Handles storage on master
JobTracker	Handles processing on master
DataNode	Handles storage on slave
TaskTracker	Handles processing on slave

C. True or False

1. For using Hadoop to process your data, the data has to be moved/ingested into HDFS.
2. Sqoop is used to query HDFS data.

3. Oozie is to import/export data from RDBMS.
4. "hadoop fs -ls /" will show the contents for the HDFS root directory.
5. Master node in Hadoop can be low on disk space but needs to have good amount of RAM.
6. In Production, NameNode preferably runs on Red Hat OS.
7. Hadoop configurations are stored in CSV format.

Answers:

1. True
2. False
3. False
4. True
5. True
6. True
7. False

D. Pick the Right Choice

1. Which of the two are components of Hadoop?
   (a) HDFS
   (b) MapReduce
   (c) Secondary NameNode
   (d) Shuffler
   (e) Sqoop
2. How many blocks will be created for a file that is 300 MB? The default block size is 64 MB and the replication factor is 3.
   (a) 30
   (b) 5
   (c) 15
   (d) 100
3. Pig is a
   (a) Data flow language
   (b) Scheduling engine
   (c) Import export tool
   (d) Shuffler
4. What does JobTracker do?
   (a) Stores blocks of data
   (b) Coordinates and schedules the job
   (c) Stores metadata
   (d) Acts as a mini reducer
5. Which ecosystem project is ideal for use when we have multiple MapReduce and Pig programs to run in a sequence?
   (a) Oozie
   (b) Pig
   (c) Hive
   (d) Sqoop
6. Which file is used for updating MapReduce settings?
   (a) core-site
   (b) hdfs-site
   (c) mapred-site
   (d) hadoop-env.sh

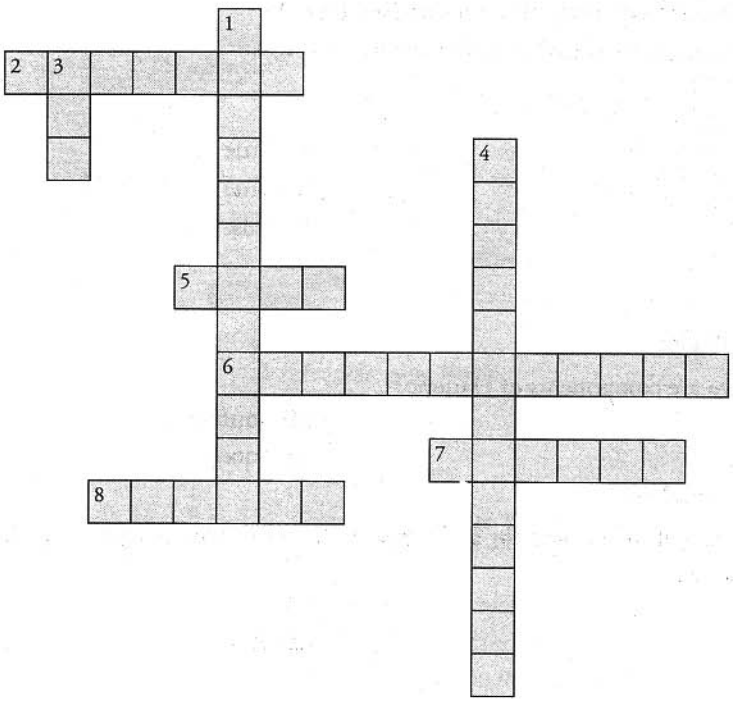
Answers:

1. (a) and (b)
2. (c)
3. (a)
4. (b)
5. (a)
6. (c)

E. Crossword

Puzzle on Big Data and Hadoop

Complete the crossword below



Across

2. One _____ Gigabytes are there in one Exabyte.
5. _____ is Splunk's new product to search, access and report on Hadoop data sets.
6. _____ gave Hadoop its name
7. _____ open-source software was developed from Google's MapReduce concept.
8. The MapReduce programming model widely used in analytics was developed at _____.

Down

1. _____ created the popular Hadoop software framework for storage and processing of large datasets.
3. _____ traditional IT Company is the largest Big Data vendor in the world.
4. According to a study by IBM, approximately _____ amount of data existed in the digital universe in 2012.

Answer:

Across

2. Billion
5. Hunk
6. Toy Elephant
7. Hadoop
8. Google

Down

1. Doug cutting
3. IBM
4. 2.7 Zettabytes

CHALLENGE ME

There are questions on topics that are not covered in the chapter. We will need you to read up on your own.

1. **What are the four modules that make up the Apache Hadoop framework?**

   Answer:

   - Hadoop Common, which contains the common utilities and libraries necessary for Hadoop's other modules.
   - Hadoop YARN, the framework's platform for resource management.
   - Hadoop Distributed File System, or HDFS, which stores information on commodity machines.
   - Hadoop MapReduce, a programming model used to process large-scale sets of data.

2. **Which modes can Hadoop be run in? List a few features for each mode.**

   Answer:

   - Standalone, or local mode, which is one of the least commonly used environments. When it is used, it's usually only for running MapReduce programs. Standalone mode lacks a distributed file system and uses a local file system instead.
   - Pseudo-distributed mode, which runs all daemons on a single machine. It is most commonly used in QA and development environments.
   - Fully distributed mode, which is most commonly used in production environments. Unlike pseudo-distributed mode, fully distributed mode runs all daemons on a cluster of machines rather than a single one.

3. **Where are Hadoop's configuration files located?**

   Answer: Hadoop's configuration files can be found inside the conf sub-directory.

4. **List Hadoop's three configuration files.**

   Answer:

   - hdfs-site.xml
   - core-site.xml
   - mapred-site.xml

5. **How many NameNodes can run on a single Hadoop cluster?**

   Answer: Only one NameNode process can run on a single Hadoop cluster. The file system will go offline if this NameNode goes down.

6. **What is a DataNode?**

   Answer: Unlike NameNode, a DataNode actually stores data within the Hadoop distributed file system. DataNodes run on their own Java virtual machine process.

7. **How many data nodes can run on a single Hadoop cluster?**

   Answer: Hadoop slave nodes contain only one data node process each.

8. **What is JobTracker in Hadoop?**

   Answer: JobTracker is used to submit and track jobs in MapReduce.

9. **How many JobTracker processes can run on a single Hadoop cluster?**

   Answer: There can only be one JobTracker process running on a single Hadoop cluster. JobTracker processes run on their own Java virtual machine process. If JobTracker goes down, all currently active jobs stop.

10. What is the difference between replication and sharding?

Answer: Replication essentially takes the same data and copies it over several machines/nodes (the number of copies it makes, depends on the defined replication factor).

Sharding takes different data and places it on different machines. It is particularly valuable for performance as it can help with read and write operations. Replication is for fault tolerance.

11. What is polyglot persistence?

Answer: The official definition of polyglot is "a person who has the ability to speak, read, and write several languages". Now consider an organization that has grown over 35 years. It has a lot of applications which write to a number of data sources (RDBMSs, Flat files, .xls, csv files, etc.). The organization also has several data marts, content management server, etc. This is a typical polyglot situation as an analytics application may require the data to be read from all of these different types of data sources.

Consider a scenario: You are one of the sponsors of an online retail firm.

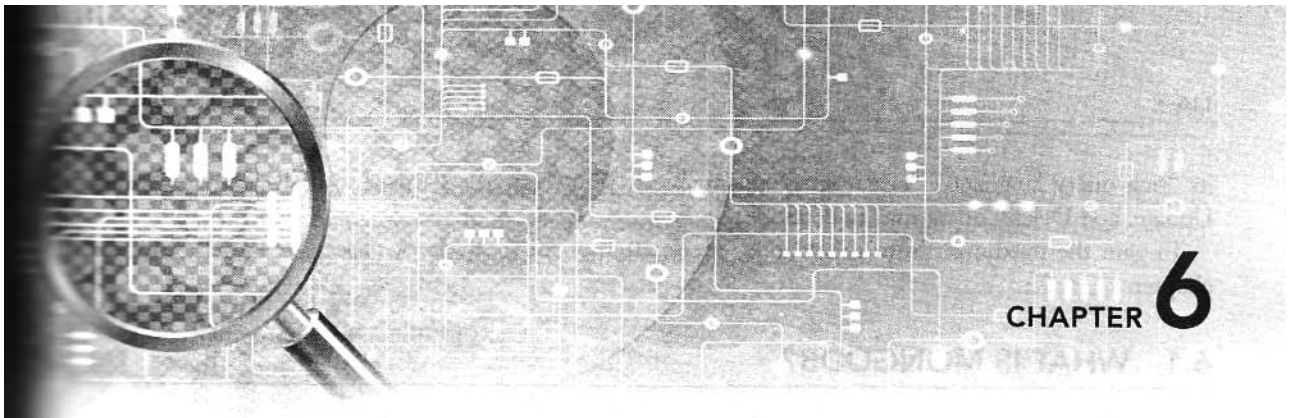
You have a few questions which you need answered:

- Who are the customers who have purchased a product X in the last 12 months?
- Do you have comments left by these customers on social network site?
- Are there repeat customers on the company's website?
- Have they recommended your product to their friends, colleagues, and relatives?
- Did they go to check the product elsewhere?

This calls for data to be collected from varied disparate data sources (relational and non-relational) and analyzed. The above is a typical case of polyglot persistence.

12. What is BigTable?

Answer: It is a compressed, proprietary data storage system built on Google File System. It is not distributed outside of Google, although it underlies the Google Datastore.



CHAPTER 6

Introduction to MongoDB

BRIEF CONTENTS

- What's in Store?
- What is MongoDB?
- Why MongoDB?
  - Using JSON
  - Creating or Generating a Unique Key
  - Support for Dynamic Queries
  - Storing Binary Data
  - Replication
  - Sharding
  - Updating Information In-Place
- Terms used in RDBMS and MongoDB
- Data Types in MongoDB
- MongoDB Query Language: CRUD (Create, Read, Update, and Delete)
  - Insert(), Update(), Save(), Remove(), find()
  - Null Values
  - Count, Limit, Sort, and Skip
  - Arrays
  - Aggregate Function
  - MapReduce Function
  - Java Script Programming
  - Cursors in MongoDB
  - Indexes
  - MongoImport
  - MongoExport
  - Automatic Generation of Unique Numbers for the "_id" Field

"*You can have data without information, but you cannot have information without data.*"

Daniel Keys Moran, computer programmer and science fiction author

WHAT'S IN STORE?

The relational database model has prevailed for decades. Of late a new kind of database is gaining ground in the enterprise called NoSQL (Not only SQL). The focus of this chapter will be on exploring a NoSQL database called "MongoDB". We bring to you the features of MongoDB such as "Auto Sharding", "Replication",

its "rich query language", "fast in-place update", etc. The chapter will cover the CRUD (Create, Read, Update, and Delete) operations in detail.

To gain the maximum from the chapter, please attempt the Test Me exercises given at the end of the chapter.

6.1 WHAT IS MONGODB?

MongoDB is

1. Cross-platform.
2. Open source.
3. Non-relational.
4. Distributed.
5. NoSQL.
6. Document-oriented data store.

6.2 WHY MONGODB?

Few of the major challenges with traditional RDBMS are dealing with large volumes of data, rich variety of data – particularly unstructured data, and meeting up to the scale needs of enterprise data. The need is for a database that can scale out or scale horizontally to meet the scale requirements, has flexibility with respect to schema, is fault tolerant, is consistent and partition tolerant, and can be easily distributed over a multitude of nodes in a cluster. Refer Figure 6.1.

6.2.1 Using Java Script Object Notation (JSON)

JSON is extremely expressive. MongoDB actually does not use JSON but BSON (pronounced Bee Son) – it is Binary JSON. It is an open standard. It is used to store complex data structures.

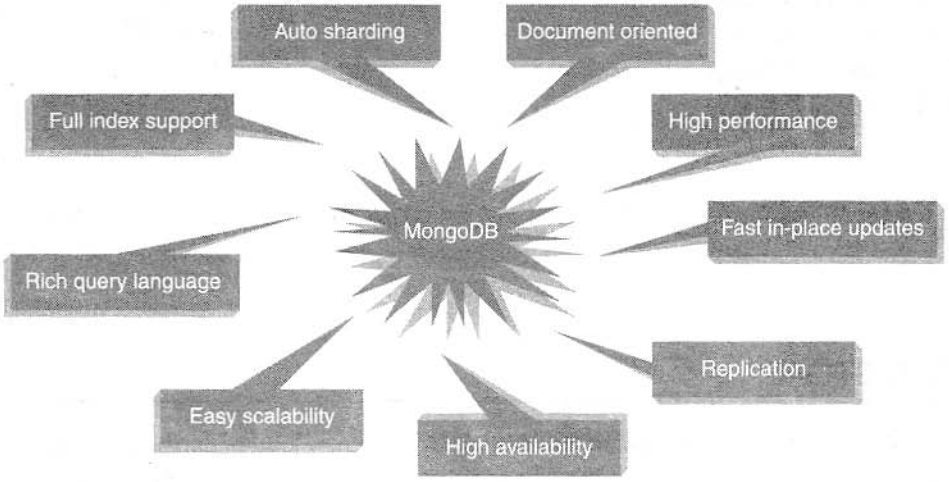


Figure 6.1 Why MongoDB?

Let us trace the journey from .csv to XML to JSON: Let us look at how data is stored in .csv file. Assume that this data is about the employees of an organization named "XYZ". As can be seen below, the column values are separated using commas and the rows are separated by a carriage return.

John, Mathews, +123 4567 8900
Andrews, Symmonds, +456 7890 1234
Mable, Mathews, +789 1234 5678

This looks good! However let us make it slightly more legible by adding column heading.

FirstName, LastName, ContactNo
John, Mathews, +123 4567 8900
Andrews, Symmonds, +456 7890 1234
Mable, Mathews, +789 1234 5678

Now assume that few employees have more than one ContactNo. It can be neatly classified as OfficeContactNo and HomeContactNo. But what if few employees have more than one OfficeContactNo and more than one HomeContactNo? Ok, so this is the first issue we need to address.

Let us look at just another piece of data that you wish to store about the employees. You need to store their email addresses as well. Here again we have the same issues, few employees have two email addresses, few have three and there are a few employees with more than three email addresses as well.

As we come across these fields or columns, we realize that it gets messy with .csv. CSV are known to store data well if it is flat and does not have repeating values.

The problem becomes even more complex when different departments maintain the details of their employees. The formats of .csv (columns, etc.) could vastly differ and it will call for some efforts before we can merge the files from the various departments to make a single file.

This problem can be solved by XML. But as the name suggests XML is highly extensible. It does not just call for defining a data format, rather it defines how you define a data format. You may be prepared to undertake this cumbersome task for highly complex and structured data; however, for simple data exchange it might just be too much work.

Enter JSON! Let us look at how it reacts to the problem at hand.

{
FirstName: John,
LastName: Mathews,
ContactNo: [+123 4567 8900, +123 4444 5555]
}
{
FirstName: Andrews,
LastName: Symmonds,
ContactNo: [+456 7890 1234, +456 6666 7777]
}
{
FirstName Mable,
LastName: Mathews,
ContactNo: +789 1234 5678
}

As you can see it is quite easy to read a JSON. There is absolutely no confusion now. One can have a list of *n* contact numbers, and they can be stored with ease.

JSON is very expressive. It provides the much needed ease to store and retrieve documents in their real form. The binary form of JSON is BSON. BSON is an open standard. In most cases it consumes less space as compared to the text-based JSON. There is yet another advantage with BSON. It is much easier and quicker to convert BSON to a programming language's native data format. There are MongoDB drivers available for a number of programming languages such as C, C++, Ruby, PHP, Python, C#, etc., and each works slightly differently. Using the basic binary format enables the native data structures to be built quickly for each language without going through the hassle of first processing JSON.

6.2.2 Creating or Generating a Unique Key

Each JSON document should have a unique identifier. It is the _id key. It is similar to the primary key in relational databases. This facilitates search for documents based on the unique identifier. An index is automatically built on the unique identifier. It is your choice to either provide unique values yourself or have the mongo shell generate the same.

0	1	2	3	4	5	6	7	8	9	10	11
Timestamp				Machine ID			Process ID		Counter		

6.2.2.1 Database

It is a collection of collections. In other words, it is like a container for collections. It gets created the first time that your collection makes a reference to it. This can also be created on demand. Each database gets its own set of files on the file system. A single MongoDB server can house several databases.

6.2.2.2 Collection

A collection is analogous to a table of RDBMS. A collection is created on demand. It gets created the first time that you attempt to save a document that references it. A collection exists within a single database. A collection holds several MongoDB documents. A collection does not enforce a schema. This implies that documents within a collection can have different fields. Even if the documents within a collection have same fields, the order of the fields can be different.

6.2.2.3 Document

A document is analogous to a row/record/tuple in an RDBMS table. A document has a dynamic schema. This implies that a document in a collection need not necessarily have the same set of fields/key–value pairs. Shown in Figure 6.2 is a collection by the name "students" containing three documents.

6.2.3 Support for Dynamic Queries

MongoDB has extensive support for dynamic queries. This is in keeping with traditional RDBMS wherein we have static data and dynamic queries. CouchDB, another document-oriented, schema-less NoSQL database and MongoDB's biggest competitor, works on quite the reverse philosophy. It has support for dynamic data and static queries.

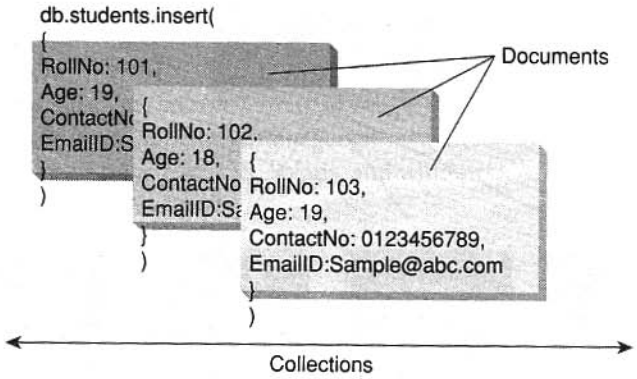


Figure 6.2 A collection "students" containing 3 documents.

6.2.4 Storing Binary Data

MongoDB provides GridFS to support the storage of binary data. It can store up to 4 MB of data. This usually suffices for photographs (such as a profile picture) or small audio clips. However, if one wishes to store movie clips, MongoDB has another solution.

It stores the metadata (data about data along with the context information) in a collection called "file". It then breaks the data into small pieces called chunks and stores it in the "chunks" collection. This process takes care about the need for easy scalability.

6.2.5 Replication

Why replication? It provides data redundancy and high availability. It helps to recover from hardware failure and service interruptions. In MongoDB, the replica set has a single primary and several secondaries. Each write request from the client is directed to the primary. The primary logs all write requests into its Oplog (operations log). The Oplog is then used by the secondary replica members to synchronize their data. This way there is strict adherence to consistency. Refer Figure 6.3. The clients usually read from the primary. However, the client can also specify a read preference that will then direct the read operations to the secondary.

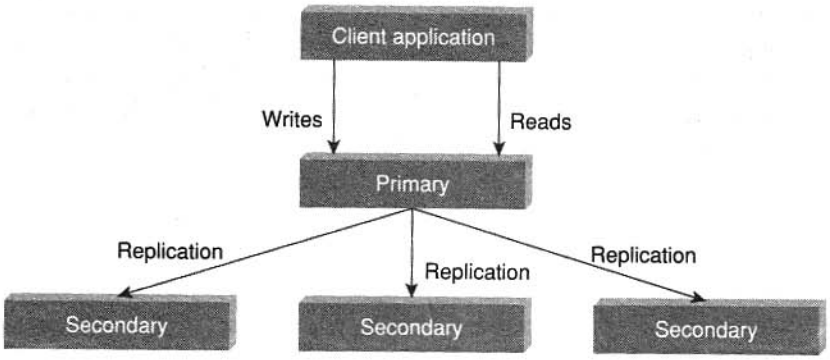


Figure 6.3 The process of **REPLICATION** in MongoDB.

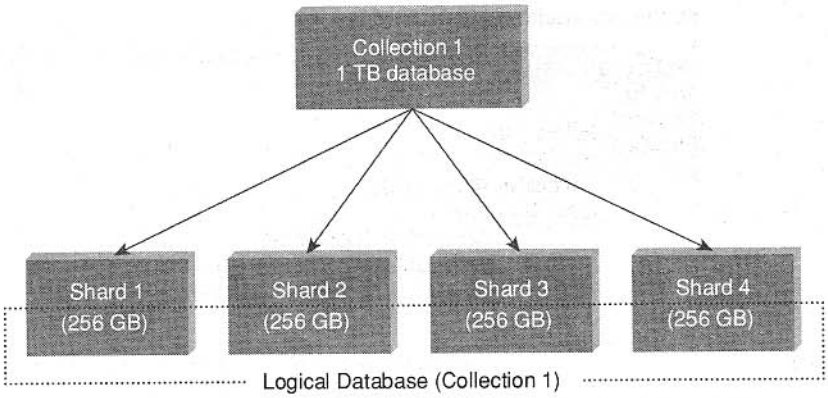


Figure 6.4 The process of **SHARDING** in MongoDB.

6.2.6 Sharding

Sharding is akin to horizontal scaling. It means that the large dataset is divided and distributed over multiple servers or shards. Each shard is an independent database and collectively they would constitute a logical database.

The prime advantages of sharding are as follows:

1. Sharding reduces the amount of data that each shard needs to store and manage. For example, if the dataset was 1 TB in size and we were to distribute this over four shards, each shard would house just 256 GB data. Refer Figure 6.4. As the cluster grows, the amount of data that each shard will store and manage will decrease.
2. Sharding reduces the number of operations that each shard handles. For example, if we were to insert data, the application needs to access only that shard which houses that data.

6.2.7 Updating Information In-Place

MongoDB updates the information in-place. This implies that it updates the data wherever it is available. It does not allocate separate space and the indexes remain unaltered.

MongoDB is all for lazy-writes. It writes to the disk once every second. Reading and writing to disk is a slow operation as compared to reading and writing from memory. The fewer the reads and writes that we perform to the disk, the better is the performance. This makes MongoDB faster than its other competitors who write almost immediately to the disk. However, there is a tradeoff. MongoDB makes no guarantee that data will be stored safely on the disk.

Guess Me

A. Who am I?

- I am blindingly fast
- I am massively scalable
- I am easy to use
- I work with documents rather than rows

B. Who am I?

- I am not for everyone
- I am good with complex data structures such as blog posts and comments
- I am good with analytics such as a real time google analytics
- I am comfortable with Linux, Mac OS, Solaris, and windows

C. Who am I?

- I have support for transactions
- I have static data
- I allow dynamic queries to be run on me

D. Who am I?

- I am one of the biggest competitor for MongoDB
- I have dynamic data
- Only static queries can be run on me
- I am document-oriented too

Answers:

A. MongoDB
B. MongoDB
C. Traditional RDBMS
D. CouchDB

6.3 TERMS USED IN RDBMS AND MONGODB

RDBMS	MongoDB
Database	Database
Table	Collection
Record	Document
Columns	Fields / Key Value pairs
Index	Index
Joins	Embedded documents
Primary Key	Primary key (_id is a identifier)

	MySQL	Oracle	MongoDB
Database Server	MySqld	Oracle	Mongod
Database Client	MySql	SQL Plus	mongo

6.3.1 Create Database

The syntax for creating database is as follows:

use DATABASE_Name

To create a database by the name "myDB" the syntax is

use myDB

```
> use myDB;
switched to db myDB
>
```

To confirm the existence of your database, type the command at the MongoDB shell:

db

```
> db;
myDB
>
```

To get a list of all databases, type the below command:

show dbs

```
> show dbs;
admin  (empty)
local  0.078GB
test   0.078GB
>
```

Notice that the newly created database, "myDB" does not show up in the list above. The reason is that the database needs to have at least one document to show up in the list.

The default database in MongoDB is test. If one does not create any database, all collections are by default stored in the test database.

6.3.2 Drop Database

The syntax to drop database is as follows:

db.dropDatabase();

To drop the database, "myDB", first ensure that you are currently placed in "myDB" database and then use the db.dropDatabase() command to drop the database.

use myDB;
db.dropDatabase();

Confirm if the database "myDB" has been dropped.

```
> db.dropDatabase();
{ "dropped" : "myDB", "ok" : 1 }
>
```

If no database is selected, the default database "test" is dropped.

6.4 DATA TYPES IN MONGODB

The following are various data types in MongoDB.

String	Must be UTF-8 valid. Most commonly used data type.
Integer	Can be 32-bit or 64-bit (depends on the server).
Boolean	To store a true/false value.
Double	To store floating point (real values).
Min/Max keys	To compare a value against the lowest or highest BSON elements.
Arrays	To store arrays or list or multiple values into one key.
Timestamp	To record when a document has been modified or added.
Null	To store a NULL value. A NULL is a missing or unknown value.
Date	To store the current date or time in Unix time format. One can create object of date and pass day, month and year to it.
Object ID	To store the document's id.
Binary data	To store binary data (images, binaries, etc.).
Code	To store javascript code into the document.
Regular expression	To store regular expression.

A few commands worth looking at are as follows (try them!!!).

To report the name of the current database:

```
C:\windows\system32\cmd.exe - mongo
> db
test
>
```

To display the list of databases:

```
C:\windows\system32\cmd.exe - mongo
> show dbs
admin  (empty)
local  0.078GB
myDB1  0.078GB
>
```

To switch to a new database, for example, myDB1:

```
C:\windows\system32\cmd.exe - mongo
> use myDB1
switched to db myDB1
>
```

To display the list of collections (tables) in the current database:

```
C:\windows\system32\cmd.exe - mongo
> show collections
system.indexes
system.js
>
```

To display the current version of the MongoDB server:

```
C:\windows\system32\cmd.exe - mongo
> db.version()
2.6.1
>
```

To display the statistics that reflect the use state of a database:

```
C:\windows\system32\cmd.exe - mongo
> db.stats()
{
        "db" : "myDB1",
        "collections" : 3,
        "objects" : 6,
        "avgObjSize" : 122.66666666666667,
        "dataSize" : 736,
        "storageSize" : 24576,
        "numExtents" : 3,
        "indexes" : 1,
        "indexSize" : 8176,
        "fileSize" : 67108864,
        "nsSizeMB" : 16,
        "dataFileVersion" : {
                "major" : 4,
                "minor" : 5
        },
        "extentFreeList" : {
                "num" : 14,
                "totalSize" : 974848
        },
        "ok" : 1
}
>
```

Type in db.help() in the MongoDB client to get a list of commands:

```
C:\windows\system32\cmd.exe - mongo
> db.help();
DB methods:
        db.adminCommand(nameOrDocument) - switches to 'admin' db, and runs comma
nd [ just calls db.runCommand(...) ]
        db.auth(username, password)
        db.cloneDatabase(fromhost)
        db.commandHelp(name) returns the help for the command
        db.copyDatabase(fromdb, todb, fromhost)
        db.createCollection(name, { size : ..., capped : ..., max : ... } )
        db.createUser(userDocument)
        db.currentOp() displays currently executing operations in the db
        db.dropDatabase()
        db.eval(func, args) run code server-side
        db.fsyncLock() flush data to disk and lock server for backups
        db.fsyncUnlock() unlocks server following a db.fsyncLock()
        db.getCollection(cname) same as db['cname'] or db.cname
        db.getCollectionNames()
        db.getLastError() - just returns the err msg string
        db.getLastErrorObj() - return full status object
        db.getMongo() get the server connection object
        db.getMongo().setSlaveOk() allow queries on a replication slave server
        db.getName()
        db.getPrevError()
        db.getProfilingLevel() - deprecated
        db.getProfilingStatus() - returns if profiling is on and slow threshold
        db.getReplicationInfo()
        db.getSiblingDB(name) get the db at the same server as this one
        db.getWriteConcern() - returns the write concern used for any operations
 on this db, inherited from server object if set
        db.hostInfo() get details about the server's host
        db.isMaster() check replica primary status
        db.killOp(opid) kills the current operation in the db
        db.listCommands() lists all the db commands
        db.loadServerScripts() loads all the scripts in db.system.js
        db.logout()
        db.printCollectionStats()
        db.printReplicationInfo()
        db.printShardingStatus()
        db.printSlaveReplicationInfo()
        db.dropUser(username)
        db.repairDatabase()
        db.resetError()
        db.runCommand(cmdObj) run a database command.  if cmdObj is a string, tu
rns it into { cmdObj : 1 }
        db.serverStatus()
        db.setProfilingLevel(level,<slowms>) 0=off 1=slow 2=all
        db.setWriteConcern( <write concern doc> ) - sets the write concern for w
rites to the db
        db.unsetWriteConcern( <write concern doc> ) - unsets the write concern f
or writes to the db
        db.setVerboseShell(flag) display extra information in shell output
        db.shutdownServer()
        db.stats()
        db.version() current version of the server
>
```

Consider a table "Students" with the following columns:

1. StudRollNo
2. StudName
3. Grade
4. Hobbies
5. DOJ

Before we get into the details of CRUD operations in MongoDB, let us look at how the statements are written in RDBMS and MongoDB.

	RDBMS	MongoDB
Insert	Insert into Students (StudRollNo, StudName, Grade, Hobbies, DOJ) Values ('S101', 'Simon David', 'VII', 'Net Surfing', '10-Oct-2012')	db.Students.insert({_id:1, StudRollNo: 'S101', StudName: 'Simon David', Grade: 'VII', Hobbies: 'Net Surfing', DOJ: '10-Oct-2012'});
Update	Update Students set Hobbies ='Ice Hockey' where StudRollNo ='S101'	db.Students.update({StudRollNo: 'S101'}, {$set: {Hobbies : 'Ice Hockey'}})
	Update Students Set Hobbies ='Ice Hockey'	db.Students.update({},{$set: {Hobbies: 'Ice Hockey' }}, {multi:true})
Delete	Delete from Students where StudRollNo = 'S101'	db.Students.remove ({StudRollNo : 'S101'})
	Delete From Students	db.Students.remove({})
Select	Select * from Students	db.Students.find() db.Students.find().pretty()
	Select * from students where StudRollNo = 'S101'	db.Students.find({StudRollNo: 'S101'})
	Select StudRollNo, StudName, Hobbies from Students	db.Students.find({},{StudRollNo:1, StudName:1, Hobbies:1, _id:0})
	Select StudRollNo, StudName, Hobbies from Students where StudRollNo = 'S101'	db.Students.find({StudRollNo: 'S101'}, {StudRollNo : 1, StudName: 1, Hobbies : 1, _id:0})

RDBMS	MongoDB
Select StudRollNo, StudName, Hobbies From Students Where Grade ='VII' and Hobbies ='Ice Hockey'	db.Students.find({Grade: 'VII' , Hobbies: 'Ice Hockey'}, {StudRollNo : 1, StudName: 1, Hobbies : 1, _id:0})
Select StudRollNo, StudName, Hobbies From Students Where Grade ='VII' or Hobbies = 'Ice Hockey'	db.Students.find({ $or: [{Grade: 'VII' , Hobbies: 'Ice Hockey'}], StudRollNo : 1, StudName: 1, Hobbies : 1, _id:0})
Select * From Students Where StudName like 'S%'	db.Students.find({ StudName: /^S/}).pretty()

6.5 MONGODB QUERY LANGUAGE

CRUD (Create, Read Update, and Delete) operations in MongoDB

Create → Creation of data is done using insert() or update() or save() method.
Read → Reading the data is performed using the find() method.
Update → Update to data is accomplished using the update() method with UPSERT set to false.
Delete → a document is Deleted using the remove() method.

We will present the various methods available in MongoDB shell to deal with data in the next few sections. The sections have been designed as follows:

Objective: What is it that we are trying to achieve here?
Input: What is the input that has been given to us to act upon?
Act: The actual statement/command to accomplish the task at hand.
Outcome: The result/output as a consequence of executing the statement.

At few places we have also provided the equivalent in RDBMS such as Oracle.

Objective: To create a collection by the name "Person". Let us take a look at the collection list prior to the creation of the new collection "Person":

```
C:\windows\system32\cmd.exe - mongo
> show collections
Students
food
system.indexes
system.js
>
```

Act: The statement to create the collection is
db.createCollection("Person")

```
C:\windows\system32\cmd.exe - mongo
> db.createCollection("Person");
{ "ok" : 1 }
>
```

Outcome: Below is the collection list after the creation of the new collection "Person":

```
C:\windows\system32\cmd.exe - mongo
> show collections;
Person
Students
food
system.indexes
system.js
>
```

Objective: To drop a collection by the name "food".
Take a look at the current collection list:

```
C:\windows\system32\cmd.exe - mongo
> show collections;
Person
Students
food
system.indexes
system.js
>
```

Act: The statement to drop the collection is

db.food.drop();

```
C:\windows\system32\cmd.exe - mongo
> db.food.drop();
true
>
```

Outcome: The collection list after the execution of the statement is as follows:

```
C:\windows\system32\cmd.exe - mongo
> show collections
Person
Students
system.indexes
system.js
>
```

6.5.1 Insert Method

We now explain the syntax of insert method.

```
db.students.insert(          <------ Collection
{
RollNo: 101,                 <------ Field: value
Age: 19,                     <------ Field: value
ContactNo:0123456789,        <------ Field: value
EmailID:Sample@abc.com       <------ Field: value
}
)
```

Objective: To create a collection by the name "Students" and insert documents.

Input: Check the list of existing collections.

```
C:\windows\system32\cmd.exe - mongo
> show collections
system.indexes
system.js
>
```

Act: Create a collection by the name "Students" and store the following data in it.

db.Students.insert({_id:1, StudName:"Michelle Jacintha", Grade: "VII", Hobbies: "Internet Surfing"});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.insert({_id:1, StudName:"Michelle Jacintha", Grade: "VII", Hobbies: "Internet Surfing"});
WriteResult({ "nInserted" : 1 })
>
```

Outcome: Check if the collection has been successfully created.

```
C:\windows\system32\cmd.exe - mongo
> show collections
Students
system.indexes
system.js
>
```

Check if the document for Student "Michelle Jacintha" has been successfully inserted into the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find();
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : "Internet Surfing" }
>
```

To format the result, one can add the pretty() method to the operation.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
>
```

Objective: Insert another document into the collection.

Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
>
```

Act:

db.Students.insert({_id:2, StudName:"Mabel Mathews", Grade: "VII", Hobbies: "Baseball"});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.insert({_id:2, StudName:"Mabel Mathews", Grade: "VII", Hobbies: "Baseball"});
WriteResult({ "nInserted" : 1 })
>
```

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

Objective: Insert the document for "Aryan David" into the Students collection only if it does not already exist in the collection. However, if it is already present in the collection, then update the document with new values. (Update his Hobbies from "Skating" to "Chess".) Use "**Update else insert**" (if there is an existing document, it will attempt to update it, if there is no existing document then it will insert it).

Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

Act:

db.Students.update({_id:3, StudName:"Aryan David", Grade: "VII"},{$set:{Hobbies: "Skating"}},{upsert:true});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:3, StudName:"Aryan David", Grade: "VII"},{$set:{Hobbies: "Skating"}},{upsert:true});
WriteResult({ "nMatched" : 0, "nUpserted" : 1, "nModified" : 0, "_id" : 3 })
>
```

Outcome: Confirm the presence of the document of "Aryan David" in the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:3});
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Skating" }
>
```

Objective: Insert the document for "Aryan David" into the Students collection only if it does not already exist in the collection. However, if it is already present in the collection, then update the document with new values. (Update his Hobbies from "Skating" to "Chess"). Use "**Update else insert**" (if there is an existing document, it will attempt to update it, if there is no existing document then it will insert it). Try the UPSERT operator by setting it first to "true" and then to "false" and observe the output.

Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Skating"
}
>
```

Act:

db.Students.update({_id:3, StudName:"Aryan David", Grade: "VII"},{$set:{Hobbies: "Chess"}},{upsert:true});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:3, StudName:"Aryan David", Grade: "VII"},{$set:{Hobbies: "Chess"}},{upsert:true});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: Confirm that the required changes have been made to the document of "Aryan David" in the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:3});
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess" }
>
```

Objective: To demonstrate Save method to insert a document for student "Vamsi Bapat" in the "Students" collection. Omit providing value for the _id key.

Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
>
```

Act:

db.Students.save({StudName:"Vamsi Bapat",Grade:"VI"})

```
C:\windows\system32\cmd.exe - mongo
> db.Students.save({StudName:"Vamsi Bapat",Grade:"VI"})
WriteResult({ "nInserted" : 1 })
>
```

Outcome: Confirm the presence of the document of "Vamsi Bapat" in the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : ObjectId("546dd0e0a7fba710799bb94d"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
>
```

6.5.2 Save() Method

We now explain the save() method. The save() method will insert a new document if the document with the specified _id does not exist. However, if a document with the specified id exists, it replaces the existing document with the new one.

Objective: Insert the document of "Hersch Gibbs" into the Students collection using the Update method. First try with upsert set to false and then with upsert set to true.

Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find();
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : "Internet Surfing" }
{ "_id" : 2, "StudName" : "Mabel Mathews", "Grade" : "VII", "Hobbies" : "Baseball" }
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess" }
{ "_id" : ObjectId("546dd0e0a7fba710799bb94d"), "StudName" : "Vamsi Bapat", "Grade" : "VI" }
>
```

Act: *Update method with upsert set to false.*

db.Students.update({_id:4, StudName:"Hersch Gibbs", Grade: "VII"},{$set:{Hobbies: "Graffiti"}}, {upsert:false});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:4, StudName:"Hersch Gibbs", Grade: "VII"},{$set:{Hobbies: "Graffiti"}}, {upsert:false});
WriteResult({ "nMatched" : 0, "nUpserted" : 0, "nModified" : 0 })
>
```

As evident from the above display (nUpserted : 0), no document has been inserted because upsert is set to false.

Update method with upsert set to true.

db.Students.update({_id:4, StudName:"Hersch Gibbs", Grade: "VII"},{$set:{Hobbies: "Graffiti"}}, {upsert:true});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:4, StudName:"Hersch Gibbs", Grade: "VII"},{$set:{Hobbies: "Graffiti"}}, {upsert:true});
WriteResult({ "nMatched" : 0, "nUpserted" : 1, "nModified" : 0, "_id" : 4 })
>
```

nUpserted: 1 implies that a document with _id:4 has been inserted.

Outcome: Confirm the presence of the document of "Hersch Gibbs" in the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find();
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : "Internet Surfing" }
{ "_id" : 2, "StudName" : "Mabel Mathews", "Grade" : "VII", "Hobbies" : "Baseball" }
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess" }
{ "_id" : ObjectId("546dd0e0a7fba710799bb94d"), "StudName" : "Vamsi Bapat", "Grade" : "VI" }
{ "_id" : 4, "Grade" : "VII", "StudName" : "Hersch Gibbs", "Hobbies" : "Graffiti" }
>
```

6.5.3 Adding a New Field to an Existing Document – Update Method

We now discuss the syntax of update method.

```
db.students.update(      ←———— Collection
{Age: {$gt 18}},         ←———— Update Criteria
{$set: {Status: "A" }}   ←———— Update Action
{multi:true}             ←———— Update Option
)
```

Objective: To add a new field "Location" with value "Newark" to the document (_id:4) of "Students" collection.

Input: Check the document (_id:4) in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:4}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
>
```

Act:

db.Students.update({_id:4},{$set:{Location:"Newark"}});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:4},{$set:{Location:"Newark"}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: Confirm that the new field "Location" with value "Newark" has been added to document (_id:4) in the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:4}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti",
        "Location" : "Newark"
}
>
```

6.5.4 Removing an Existing Field from an Existing Document – Remove Method

In this section we will explain the syntax of remove method.

```
db.students.remove(        ←———— Collection
{Age: {$gt 18}},           ←———— Remove Criteria
)
```

Objective: To remove the field "Location" with value "Newark" in the document (_id:4) of "Students" collection.

Input: Check the document (_id:4) in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:4}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti",
        "Location" : "Newark"
}
>
```

Act:

db.Students.update({_id:4},{$unset:{Location:"Newark"}});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:4},{$unset:{Location:"Newark"}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: Confirm if the stated field ("Location") has been dropped from the document (_id:4) of the "Students" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:4}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
>
```

6.5.5 Finding Documents based on Search Criteria – Find Method

The syntax of find method is as follows:

```
db.students.find(          ←———— Collection
{Age: {$gt 18}},           ←———— Selection Criteria
{RollNo:1,Age:1,_id:1}     ←———— Projection
).limit(10)                ←———— Cursor Modifier
```

Objective: To search for documents from the "Students" collection based on certain search criteria.
Input: Check the documents in the "Students" collection before proceeding.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

Act: Find the document wherein the "StudName" has value "Aryan David".

db.Students.find({StudName:"Aryan David"});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({StudName:"Aryan David"});
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess" }
>
```

To format the above output, use the pretty() method:

db.Students.find({StudName:"Aryan David"}).pretty();

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({StudName:"Aryan David"}).pretty();
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
>
```

RDBMS equivalent:

Select *
    From Students
        Where StudName like 'Aryan David';

```
SQL> select * from Students where StudName like 'Aryan David';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
3     Aryan David               VII   Chess
SQL>
```

Objective: To display only the StudName from all the documents of the Student's collection. The identifier "__id" should be suppressed and NOT displayed.

Act:

db.Students.find({},{StudName:1,_id:0});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({},{StudName:1,_id:0});
{ "StudName" : "Michelle Jacintha" }
{ "StudName" : "Aryan David" }
{ "StudName" : "Hersch Gibbs" }
{ "StudName" : "Vamsi Bapat" }
{ "StudName" : "Mabel Mathews" }
>
```

RDBMS equivalent:

Select StudName
From Students;

```
SQL Plus
SQL> select StudName from Students;

STUDNAME
------------------------
Michelle Jacintha
Aryan David
Mabel Mathews
Hersch Gibbs
Vamsi Bapat

SQL>
```

Objective: To display only the StudName and Grade from all the documents of the Students collection. The identifier _id should be suppressed and NOT displayed.

Act:

db.Students.find({},{StudName:1,Grade:1,_id:0});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({},{StudName:1,Grade:1,_id:0});
{ "StudName" : "Michelle Jacintha", "Grade" : "VII" }
{ "Grade" : "VII", "StudName" : "Aryan David" }
{ "Grade" : "VII", "StudName" : "Hersch Gibbs" }
{ "StudName" : "Vamsi Bapat", "Grade" : "VI" }
{ "StudName" : "Mabel Mathews", "Grade" : "VII" }
>
```

RDBMS equivalent:

Select StudName, Grade
From Students;

```
SQL Plus
SQL> select StudName, Grade from Students;

STUDNAME                 GRADE
------------------------ ------
Michelle Jacintha        VII
Aryan David              VII
Mabel Mathews            VII
Hersch Gibbs             VII
Vamsi Bapat              VI

SQL>
```

Objective: To display the StudName, Grade as well the identifier, _id from the document of the Students collection where the _id column is 1.

Act:

db.Students.find({_id:1},{StudName:1,Grade:1});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:1},{StudName:1,Grade:1});
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII" }
>
```

RDBMS equivalent:

Select StudRollNo, StudName, Grade
From Students
Where StudRollNo = '1';

```
SQL> select StudRollNo, StudName, Grade from Students where StudRollNo = '1';
STUDR STUDNAME                    GRADE
----- ------------------------    -----
1     Michelle Jacintha           VII
SQL>
```

Objective: To display the StudName and Grade from the document of the Students collection where the _id column is 1. The _id field should NOT be displayed.

Act:

db.Students.find({_id:1},{StudName:1,Grade:1,_id:0});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({_id:1},{StudName:1,Grade:1,_id:0});
{ "StudName" : "Michelle Jacintha", "Grade" : "VII" }
>
```

RDBMS equivalent:

Select StudName, Grade
From Students
Where StudRollNo like '1';

```
SQL> select StudName, Grade from Students where StudRollNo like '1';
STUDNAME                    GRADE
------------------------    -----
Michelle Jacintha           VII
SQL>
```

Relational operators available to use in the search criteria:

$eq	→ equal to
$ne	→ not equal to
$gte	→ greater than or equal to
$lte	→ less than or equal to
$gt	→ greater than
$lt	→ less than

Objective: To find those documents where the Grade is set to 'VII'.

Act:

db.Students.find({Grade:{$eq:'VII'}}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({Grade:{$eq:'VII'}}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where Grade like 'VII';

```
SQL Plus
SQL> select * from Students where Grade like 'VII';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ------ ------------------------------
1     Michelle Jacintha           VII   Internet surfing
3     Aryan David                 VII   Chess
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
SQL>
```

Objective: To find those documents where the Grade is NOT set to 'VII'.

Act:

db.Students.find({Grade:{$ne:'VII'}}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({Grade:{$ne:'VII'}}).pretty();
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
>
```

There is just one document that meets the above criteria of Grade NOT EQUAL to 'VII'.

RDBMS Equivalent:

Select *

From Students

Where Grade <> 'VII';

```
SQL Plus
SQL> select * from Students where Grade <> 'VII';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ------ ------------------------------
      Vamsi Bapat                 VI
SQL>
```

OR

```
SQL Plus
SQL> select * from Students where Grade != 'VII';
STUDR STUDNAME                    GRADE HOBBIES
----- -------------------------- ----- ------------------------------
      Vamsi Bapat                 VI
SQL>
```

Objective: To find those documents from the Students collection where the Hobbies is set to either 'Chess' or is set to 'Skating'.

Act:

db.Students.find ({Hobbies :{ $in: ['Chess','Skating']}}).pretty ();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find ({Hobbies :{ $in: ['Chess','Skating']}}).pretty ();
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where Hobbies in ('Chess', 'Skating');

```
SQL Plus
SQL> select * from Students where Hobbies in ('Chess', 'Skating');
STUDR STUDNAME                    GRADE HOBBIES
----- -------------------------- ----- ------------------------------
3     Aryan David                 VII   Chess
SQL>
```

Objective: To find those documents from the Students collection where the Hobbies is set neither to 'Chess' nor is set to 'Skating'.

Act:

db.Students.find ({Hobbies :{ $nin: ['Chess','Skating']}}).pretty ();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find ({Hobbies :{ $nin: ['Chess','Skating']}}).pretty ();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where Hobbies not in ('Chess', 'Skating');

```
SQL> select * from Students where Hobbies not in ('Chess','Skating');
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
1     Michelle Jacintha           VII   Internet surfing
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
SQL>
```

Objective: To find those documents from the Students collection where the Hobbies is set to 'Graffiti' and the StudName is set to 'Hersch Gibbs' (AND condition).

Act:

db.Students.find({Hobbies:'Graffiti', StudName: 'Hersch Gibbs'}).pretty();

Outcome:

```
> db.Students.find({Hobbies:'Graffiti', StudName: 'Hersch Gibbs'}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where Hobbies like 'Graffiti' and StudName like 'Hersch Gibbs';

```
SQL> select * from Students where Hobbies like 'Graffiti' and StudName like 'Her
sch Gibbs';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
4     Hersch Gibbs                VII   Graffiti
SQL>
```

Objective: To find documents from the Students collection where the StudName begins with "M".

Act:

db.Students.find({StudName:/^M/}).pretty();

Outcome:

```
> db.Students.find({StudName:/^M/}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where StudName like 'M%';

```
SQL> select * from Students where StudName like 'M%';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
1     Michelle Jacintha           VII   Internet surfing
2     Mabel Mathews               VII   Baseball
SQL>
```

Objective: To find documents from the Students collection where the StudName ends in "s".

Act:

db.Students.find({StudName:/s$/}).pretty();

Outcome:

```
> db.Students.find({StudName:/s$/}).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where StudName like '%s';

```
SQL> select * from Students where StudName like '%s';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
SQL>
```

Objective: To find documents from the Students collection where the StudName has an "e" in any position.

Act:

db.Students.find({StudName:/e/}).pretty();

OR

db.Students.find({StudName:/.*e.*/}).pretty();

OR

db.Students.find({StudName:{$regex:"e"}}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({StudName:/e/}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where StudName like '%e%';

```
SQL> select * from Students where StudName like '%e%';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
1     Michelle Jacintha         VII   Internet surfing
2     Mabel Mathews             VII   Baseball
4     Hersch Gibbs              VII   Graffiti
SQL>
```

Objective: To find documents from the Students collection where the StudName ends in "a".

Act:

db.Students.find({StudName:{$regex:"a$"}}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({StudName:{$regex:"a$"}}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where StudName like '%a';

```
SQL> select * from Students where StudName like '%a';
STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
1     Michelle Jacintha         VII   Internet surfing
SQL>
```

Objective: To find documents from the Students collection where the StudName begins with "M".

Act:

db.Students.find({StudName:{$regex:"^M"}}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({StudName:{$regex:"^M"}}).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *
From Students
Where StudName like 'M%';

```
SQL Plus
SQL> select * from Students where StudName like 'M%';

STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------- ----- ------------------------------
1     Michelle Jacintha         VII   Internet surfing
2     Mabel Mathews             VII   Baseball

SQL>
```

6.5.6 Dealing with NULL Values

Objective: To add a new field with null value in existing documents (_id:3 and _id:4) of Students collection. A NULL is a missing or unknown value. When we place NULL as a value for a field, it implies that currently we do not know the value or the value is missing. We can always update the value of the field once we know it.

Input: Before we execute the commands to update documents with a null value in a column, let us first view the two documents.

db.Students.find({$or:[{_id:3},{_id:4}]})

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({$or:[{_id:3},{_id:4}]});
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess" }
{ "_id" : 4, "Grade" : "VII", "StudName" : "Hersch Gibbs", "Hobbies" : "Graffiti" }
>
```

Act: Update the documents with NULL values in the "Location" column.

db.Students.update({_id:3},{$set:{Location:null}});

db.Students.update({_id:4},{$set:{Location:null}});

```
C:\windows\system32\cmd.exe - mongo
> db.Students.update({_id:3},{$set:{Location:null}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Students.update({_id:4},{$set:{Location:null}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

RDBMS Equivalent:

Update Students
    Set Location = null
        Where StudRollNo in ('3','4');

```
SQL> update Students set Location = null where StudRollNo in ('3','4');
2 rows updated.
SQL>
```

Outcome: To search for NULL values in Location column.

db.Students.find({Location:{$eq:null}});

```
> db.Students.find({Location:{$eq:null}});
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : "Internet Surfing" }
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess", "Location" : null }
{ "_id" : 4, "Grade" : "VII", "StudName" : "Hersch Gibbs", "Hobbies" : "Graffiti", "Location" : null }
{ "_id" : ObjectId("5464849889ad1ab07d489b7f"), "StudName" : "Vamsi Bapat", "Grade" : "VI" }
{ "_id" : 2, "StudName" : "Mabel Mathews", "Grade" : "VII", "Hobbies" : "Baseball" }
>
```

The above statement displays documents which either have NULL values in the location column or do not have the location column at all.

RDBMS Equivalent:

Select *
    From Students
        Where Location is Null;

```
SQL> select * from Students where Location is NULL;
STUDR STUDNAME                    GRADE HOBBIES                          LOCATION
----- ------------------------- ----- ------------------------------ -----------
---------
1     Michelle Jacintha           VII   Internet surfing
3     Aryan David                 VII   Chess
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
      Vamsi Bapat                 VI
SQL>
```

Objective: To remove "Location" field having "NULL" values from the documents (_id:3 and _id:4) from the Students collection.

Input: Document from the "Students" collection having "NULL" values in the "Location" column.

```
> db.Students.find({Location:{$eq:null}});
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : "Internet Surfing" }
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess", "Location" : null }
{ "_id" : 4, "Grade" : "VII", "StudName" : "Hersch Gibbs", "Hobbies" : "Graffiti", "Location" : null }
{ "_id" : ObjectId("5464849889ad1ab07d489b7f"), "StudName" : "Vamsi Bapat", "Grade" : "VI" }
{ "_id" : 2, "StudName" : "Mabel Mathews", "Grade" : "VII", "Hobbies" : "Baseball" }
>
```

Act:

db.Students.update({_id:3},{$unset:{Location:null}});
db.Students.update({_id:4},{$unset:{Location:null}});

```
> db.Students.update({_id:3},{$unset:{Location:null}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 0 })
> db.Students.update({_id:4},{$unset:{Location:null}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 0 })
>
```

Outcome: Let us confirm if the changes have been made by running find method on the Students collection.

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({})
{ "_id" : 1, "StudName" : "Michelle Jacintha", "Grade" : "VII", "Hobbies" : '
{ "_id" : 3, "Grade" : "VII", "StudName" : "Aryan David", "Hobbies" : "Chess'
{ "_id" : 4, "Grade" : "VII", "StudName" : "Hersch Gibbs", "Hobbies" : "Graf'
{ "_id" : ObjectId("5464849889ad1ab07d489b7f"), "StudName" : "Vamsi Bapat",
```

6.5.7 Count, Limit, Sort, and Skip

Objective: To find the number of documents in the Students collection.

Act:

db.Students.count()

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.count()
5
>
```

Objective: To find the number of documents in the Students collection wherein the Grade is VII.

Act:

db.Students.count({Grade:"VII"});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.count({Grade:"VII"});
4
>
```

Objective: To retrieve the first 3 documents from the Students collection wherein the Grade is VII.

Act:

db.Students.find({Grade:"VII"}).limit(3).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find({Grade:"VII"}).limit(3).pretty();
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
>
```

RDBMS Equivalent:

Select *

From Students

Where Grade like 'VII' and rownum < 4;

```
SQL> select * from Students where Grade like 'VII' and rownum < 4;
STUDR STUDNAME                    GRADE HOBBIES                          LOCATION
----- ------------------------- ----- ------------------------------ ----------
1     Michelle Jacintha         VII   Internet surfing
3     Aryan David               VII   Chess
2     Mabel Mathews             VII   Baseball
SQL>
```

Objective: To sort the documents from the Students collection in the ascending order of StudName.

Act:

db.Students.find().sort({StudName:1}).pretty();

Outcome:

```
> db.Students.find().sort({StudName:1}).pretty();
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
>
```

RDBMS Equivalent:

Select *

From Students

Order by StudName asc;

```
SQL> select * from Students order by StudName asc;
STUDR STUDNAME                    GRADE HOBBIES                          LOCATION
----- ------------------------- ----- ------------------------------ ----------
3     Aryan David               VII   Chess
4     Hersch Gibbs              VII   Graffiti
2     Mabel Mathews             VII   Baseball
1     Michelle Jacintha         VII   Internet surfing
      Vamsi Bapat               VI
SQL>
```

Objective: To sort the documents from the Students collection in the descending order of StudName.

Act:

db.Students.find().sort({StudName:-1}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().sort({StudName:-1}).pretty();
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
>
```

RDBMS Equivalent:

Select *

From Students

Order by StudName desc;

```
SQL Plus
SQL> select * from Students order by StudName desc;
STUDR STUDNAME                    GRADE HOBBIES                         LOCATION
----- ------------------------- ----- ----------------------------- -----------
---------
      Vamsi Bapat                 VI
1     Michelle Jacintha           VII   Internet surfing
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
3     Aryan David                 VII   Chess
SQL>
```

Objective: To sort the documents from the Students collection first on Grade in ascending order and then on Hobbies in descending order.

Act:

db.Students.find().sort({Grade:1, Hobbies:-1}).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().sort({Grade:1, Hobbies:-1}).pretty();
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select *

From Students

Order by Grade asc, hobbies desc;

```
SQL> select * from Students order by Grade asc, Hobbies desc;
STUDR STUDNAME                    GRADE HOBBIES                          LOCATION
----- ------------------------- ----- ------------------------------ ------------
---------
      Vamsi Bapat                 VI
1     Michelle Jacintha           VII   Internet surfing
4     Hersch Gibbs                VII   Graffiti
3     Aryan David                 VII   Chess
2     Mabel Mathews               VII   Baseball

SQL>
```

Objective: To sort the documents from the Students collection first on Grade in ascending order and then on Hobbies in ascending order.

Act:

db.Students.find().sort({Grade:1, Hobbies:1}).pretty();

Outcome:

```
> db.Students.find().sort({Grade:1, Hobbies:1}).pretty();
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "Aryan David",
        "Hobbies" : "Chess"
}
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
>
```

RDBMS Equivalent:

Select *

From Students

Order by Grade asc, Hobbies asc;

```
SQL> select * from Students order by Grade asc, Hobbies asc;
STUDR STUDNAME                    GRADE HOBBIES                          LOCATION
----- ------------------------- ----- ------------------------------ ------------
---------
      Vamsi Bapat                 VI
2     Mabel Mathews               VII   Baseball
3     Aryan David                 VII   Chess
4     Hersch Gibbs                VII   Graffiti
1     Michelle Jacintha           VII   Internet surfing

SQL>
```

Objective: To skip the first 2 documents from the Students collection.

Act:

db.Students.find().skip(2).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().skip(2).pretty();
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

RDBMS Equivalent:

Select StudRollNo, StudName, Grade, Hobbies
From (Select StudRollNo, StudName, Grade, Hobbies, RowNum as TheRowNum
From Students)
Where TheRowNum > 2;

```
SQL Plus
SQL> Select StudRollNo, StudName, Grade, Hobbies from (Select StudRollno, StudNa
me, Grade, Hobbies, Rownum as therownum from Students) where theRownum > 2;

STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------    ----- ------------------------------
2     Mabel Mathews               VII   Baseball
4     Hersch Gibbs                VII   Graffiti
      Vamsi Bapat                 VI

SQL>
```

Objective: To sort the documents from the Students collection and skip the first document from the output.

Act:

db.Students.find().skip(1).pretty().sort({StudName:1});

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().skip(1).pretty().sort({StudName:1});
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
{
        "_id" : 1,
        "StudName" : "Michelle Jacintha",
        "Grade" : "VII",
        "Hobbies" : "Internet Surfing"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
>
```

RDBMS Equivalent:

Select StudRollNo, StudName, Grade, Hobbies
From (Select StudRollNo, StudName, Grade, Hobbies, RowNum as TheRowNum
From Students)
Where TheRowNum > 1
Order by StudName;

```
SQL> Select StudRollNo, StudName, Grade, Hobbies from (Select StudRollno, StudNa
me, Grade, Hobbies, Rownum as therownum from Students) where theRownum >1 order
by StudName;

STUDR STUDNAME                    GRADE HOBBIES
----- ------------------------    ----- ------------------------------
3     Aryan David                 VII   Chess
4     Hersch Gibbs                VII   Graffiti
2     Mabel Mathews               VII   Baseball
      Vamsi Bapat                 VI

SQL>
```

Objective: To display the last 2 records from the Students collection.

Act:

db.Students.find().pretty().skip(db.Students.count()-2);

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.Students.find().pretty().skip(db.Students.count()-2);
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

Objective: To retrieve the third, fourth, and fifth document from the Students collection.

Act:

db.Students.find().pretty().skip(2).limit(3);

Outcome:

```
C:\windows\system32\cmd.exe - mongo
  db.Students.find().pretty().skip(2).limit(3);
{
        "_id" : 4,
        "Grade" : "VII",
        "StudName" : "Hersch Gibbs",
        "Hobbies" : "Graffiti"
}
{
        "_id" : ObjectId("5464849889ad1ab07d489b7f"),
        "StudName" : "Vamsi Bapat",
        "Grade" : "VI"
}
{
        "_id" : 2,
        "StudName" : "Mabel Mathews",
        "Grade" : "VII",
        "Hobbies" : "Baseball"
}
>
```

6.5.8 Arrays

Objective: To create a collection by the name "food" and then insert documents into the "food" collection. Each document should have a "fruits" array.

Act:

db.food.insert({_id:1,fruits:['banana','apple','cherry'] })
db.food.insert({_id:2,fruits:['orange','butterfruit','mango']})
db.food.insert({_id:3,fruits:['pineapple','strawberry','grapes']});
db.food.insert({_id:4,fruits:['banana','strawberry','grapes']});
db.food.insert({_id:5,fruits:['orange','grapes']});

```
C:\windows\system32\cmd.exe - mongo
> db.food.insert({_id:1,fruits:[ 'banana','apple','cherry' ] })
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:2,fruits:[ 'orange','butterfruit','mango' ]})
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:3,fruits:[ 'pineapple','strawberry','grapes']});
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:4,fruits:[ 'banana','strawberry','grapes']});
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:5,fruits:[ 'orange','grapes']});
WriteResult({ "nInserted" : 1 })
>
```

Outcome: Let us check if these documents are now in the "food" collection.

db.food.find({})

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({})
{ "_id" : 1, "fruits" : [ "banana", "apple", "cherry" ] }
{ "_id" : 2, "fruits" : [ "orange", "butterfruit", "mango" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "strawberry", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection which has the "fruits array" constituted of "banana", "apple" and "cherry".

Act:

db.food.find({fruits:['banana','apple','cherry']}).pretty()

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({fruits:['banana','apple','cherry']}).pretty()
{ "_id" : 1, "fruits" : [ "banana", "apple", "cherry" ] }
>
```

Objective: To find those documents from the "food" collection which has the "fruits" array having "banana", as an element.

Act:

db.food.find({fruits:'banana'})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({fruits:'banana'})
{ "_id" : 1, "fruits" : [ "banana", "apple", "cherry" ] }
{ "_id" : 4, "fruits" : [ "banana", "strawberry", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection which have the "fruits" array having "grapes" in the first index position. The index position begins at 0.

Act:

db.food.find({'fruits.1':'grapes'})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({'fruits.1':'grapes'})
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection where "grapes" is present in the 2nd index position of the "fruits" array.

Act:

db.food.find({'fruits.2':'grapes'})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({'fruits.2':'grapes'})
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "strawberry", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection where the size of the array is two. The size implies that the array holds only 2 values.

Act:

db.food.find({"fruits":{$size:2}})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({"fruits":{$size:2}})
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection where the size of the array is three. The size implies that the array holds only 3 values.

Act:

db.food.find({"fruits":{$size:3}})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({"fruits":{$size:3}});
{ "_id" : 1, "fruits" : [ "banana", "apple", "cherry" ] }
{ "_id" : 2, "fruits" : [ "orange", "butterfruit", "mango" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "strawberry", "grapes" ] }
>
```

Objective: To find the document with (_id: 1) from the "food" collection and display the first two elements from the array "fruits".

Act:

db.food.find({_id:1},{"fruits":{$slice:2}})

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:1},{"fruits":{$slice:2}})
{ "_id" : 1, "fruits" : [ "banana", "apple" ] }
>
```

Objective: To find all documents from the "food" collection which have elements "orange" and "grapes" in the array "fruits".

Act:

db.food.find ({fruits: {$all: ["orange", "grapes"]}}).pretty ();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find ({fruits: {$all: ["orange", "grapes"]}}).pretty ();
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
>
```

Objective: To find those documents from the "food" collection which have the element "orange" in the 0th index position in the array "fruits".

Act:

db.food.find({ "fruits.0" : "orange" }).pretty();

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({ "fruits.0" :  "orange" }).pretty();
{ "_id" : 2, "fruits" : [ "orange", "butterfruit", "mango" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
>
```

Objective: To find the document with (_id: 1) from the "food" collection and display two elements from the array "fruits", starting with the element at 0th index position.

Act:

db.food.find({_id:1},{"fruits":{$slice:[0,2]}})

Outcome:

```
> db.food.find({_id:1},{"fruits":{$slice:[0,2]}})
{ "_id" : 1, "fruits" : [ "banana", "apple" ] }
>
```

Objective: To find the document with (_id: 1) from the "food" collection and display two elements from the array "fruits", starting with the element at 1st index position.

Act:

db.food.find({_id:1},{"fruits":{$slice:[1,2]}})

Outcome:

```
> db.food.find({_id:1},{"fruits":{$slice:[1,2]}});
{ "_id" : 1, "fruits" : [ "apple", "cherry" ] }
>
```

Objective: To find the document with (_id: 1) from the "food" collection and display three elements from the array "fruits", starting with the element at 2nd index position. Since we have only 3 elements in the array "fruits" for the document with _id:1, it displays only one element, the element at 2nd index position, that is, "cherry".

Act:

db.food.find({_id:1},{"fruits":{$slice:[2,3]}})

Outcome:

```
> db.food.find({_id:1},{"fruits":{$slice:[2,3]}});
{ "_id" : 1, "fruits" : [ "cherry" ] }
>
```

6.5.8.1 *Update on the Array*

Before we begin the update operations on the "fruits" array of the documents of "food" collection, let us take a look at the documents that we have in the "food" collection:

```
> db.food.find({});
{ "_id" : 1, "fruits" : [ "banana", "apple", "cherry" ] }
{ "_id" : 2, "fruits" : [ "orange", "butterfruit", "mango" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "strawberry", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
```

Objective: To update the document with "_id:4" and replace the element present in the 1st index position of the "fruits" array with "apple".

Act:

db.food.update({_id:4},{$set:{'fruits.1': 'apple'}})

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:4},{$set:{'fruits.1': 'apple'}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: Let us take a look at how this update has changed our document.

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:4});
{ "_id" : 4, "fruits" : [ "banana", "apple", "grapes" ] }
>
```

Objective: To update the document with "_id:1" and replace the element "apple" of the "fruits" array with "An apple".

Act:

db.food.update({_id:1, 'fruits':'apple'},{$set:{'fruits.$': 'An apple' }})

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:1, 'fruits':'apple'},{$set:{'fruits.$': 'An apple' }})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The document after update is as follows.

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:1});
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
>
```

Objective: To update the document with "_id:2" and push new key value pairs in the "fruits" array.

Act:

db.food.update({_id:2},{$push:{price:{orange:60,butterfruit:200,mango:120}}})

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:2},{$push:{price:{orange:60,butterfruit:200,mango:120}}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

6.5.8.2 *Further Updates to the Array "fruits" ...*

Before we do the updates to the documents in the food collection, let us look at the current state:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Objective: To update the document with "_id:4" by adding an element "orange" to the list of elements in the array "fruits".

Act:

db.food.update({_id:4},{$addToSet:{fruits:"orange"}});

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:4},{$addToSet:{fruits:"orange"}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The result after the execution of the statement is as follows.

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "apple", "grapes", "orange" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Objective: To update the document with "_id:4" by popping an element from the list of elements present in the array "fruits". The element popped is the one from the end of the array.

Act:

db.food.update({_id:4},{$pop:{fruits:1}});

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:4},{$pop:{fruits:1}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The "food" collection after the execution of the statement is as follows.

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "banana", "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Objective: To update the document with "_id:4" by popping an element from the list of elements present in the array "fruits". The element popped is the one from the beginning of the array.

Act:

db.food.update({_id:4},{$pop:{fruits:-1}});

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:4},{$pop:{fruits:-1}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The "food" collection after the execution of the above update statement is as follows.

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
{ "_id" : 4, "fruits" : [ "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Objective: To update the document with "_id:3" by popping two elements from the list of elements present in the array "fruits". The elements popped are "pineapple" and "grapes".

The document with "_id:3" before the update is

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:3});
{ "_id" : 3, "fruits" : [ "pineapple", "strawberry", "grapes" ] }
>
```

Act:

db.food.update({_id:3},{$pullAll:{fruits: ['pineapple','grapes']}});

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:3},{$pullAll:{fruits: [ 'pineapple','grapes' ]}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The document with "_id:3" after the update is as follows:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:4});
{ "_id" : 4, "fruits" : [ "apple", "grapes" ] }
>
```

Objective: To update the documents having "banana" as an element in the array "fruits" and pop out the element "banana" from those documents.

The "food" collection before the update is as follows:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "banana", "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "strawberry" ] }
{ "_id" : 4, "fruits" : [ "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Act:

db.food.update({fruits:'banana'}, {$pull:{fruits:'banana'}})

```
C:\windows\system32\cmd.exe - mongo
> db.food.update({fruits:'banana'}, {$pull:{fruits:'banana'}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: The "food" collection after the update is as follows:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find().pretty();
{ "_id" : 1, "fruits" : [ "An apple", "cherry" ] }
{ "_id" : 3, "fruits" : [ "strawberry" ] }
{ "_id" : 4, "fruits" : [ "apple", "grapes" ] }
{ "_id" : 5, "fruits" : [ "orange", "grapes" ] }
{
        "_id" : 2,
        "fruits" : [
                "orange",
                "butterfruit",
                "mango"
        ],
        "price" : [
                {
                        "orange" : 60,
                        "butterfruit" : 200,
                        "mango" : 120
                }
        ]
}
>
```

Objective: To pull out an array element based on index position.

There is no direct way of pulling the array elements by looking up their index numbers. However a workaround is available. The document with "_id:4" in the food collection prior to the update is as follows:

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:4}).pretty();
{ "_id" : 4, "fruits" : [ "apple", "grapes" ] }
>
```

Act: The update statement is

db.food.update({_id:4}, {$unset : {"fruits.1" : null }}) ;

db.food.update({_id:4}, {$pull : {"fruits" : null}});

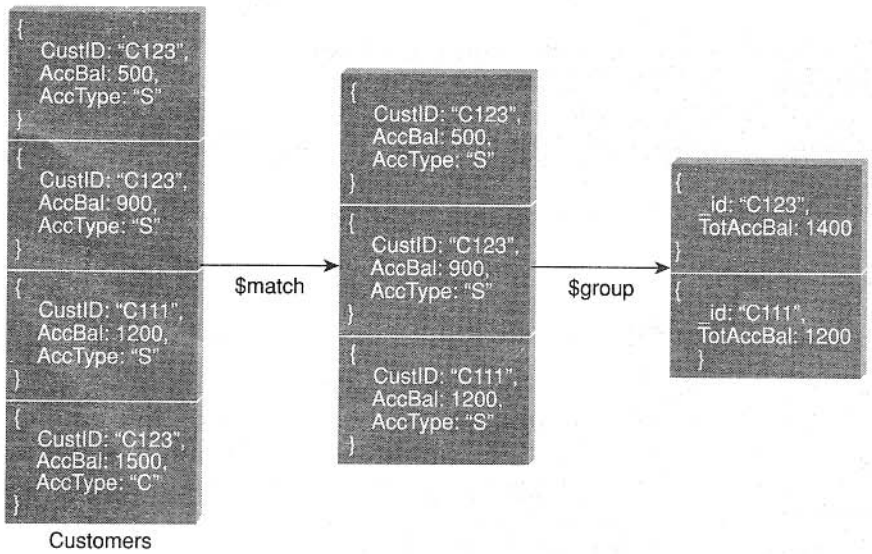
```
C:\windows\system32\cmd.exe - mongo
> db.food.update({_id:4}, {$unset : {"fruits.1" : null }}) ;
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.food.update({_id:4}, {$pull : {"fruits" : null}});
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
>
```

Outcome: After update, the document with _id:4 in the food collection is

```
C:\windows\system32\cmd.exe - mongo
> db.food.find({_id:4}).pretty();
{ "_id" : 4, "fruits" : [ "apple" ] }
>
```

6.5.9 Aggregate Function

Objective: Consider the collection "Customers" as given below. It has four documents. We would like to filter out those documents where the "AccType" has a value other than "S". After the filter, we should be left with three documents where the "Acctype": "S". It is then required to group the documents on the basis of CustID and sum up the "AccBal" for each unique "CustID". This is similar to the output received with group by clause in RDBMS. Once the groups have been formed [as per the example below, there will be only two groups: (a) "CustID" : "C123" and (b) "CustID" : "C111"], filter and display that group where the "TotAccBal" column has a value greater than 1200.



Let us start off by creating the collection "Customers" with the above displayed four documents:

```
db.Customers.insert([{CustID:"C123",AccBal:500,AccType:"S"},
{CustID:"C123", AccBal: 900, AccType:"S"},
{CustID:"C111", AccBal: 1200, AccType:"S"},
{CustID:"C123", AccBal: 1500, AccType:"C"}]);
```

```
> db.Customers.insert([{CustID:"C123",AccBal:500,AccType:"S"},{CustID:"C123",AccBal:900,AccType:"S"},
... {CustID:"C111",AccBal:1200,AccType:"S"}, {CustID:"C123",AccBal:1500,AccType:"C"}]);
BulkWriteResult({
        "writeErrors" : [ ],
        "writeConcernErrors" : [ ],
        "nInserted" : 4,
        "nUpserted" : 0,
        "nMatched" : 0,
        "nModified" : 0,
        "nRemoved" : 0,
        "upserted" : [ ]
})
>
```

To confirm the presence of four documents in the "Customers" collection, use the below syntax:

```
db.Customers.find().pretty();
```

```
> db.Customers.find().pretty();
{
        "_id" : ObjectId("54993269f4263d0150bfa72c"),
        "CustID" : "C123",
        "AccBal" : 500,
        "AccType" : "S"
}
{
        "_id" : ObjectId("54993269f4263d0150bfa72d"),
        "CustID" : "C123",
        "AccBal" : 900,
        "AccType" : "S"
}
{
        "_id" : ObjectId("54993269f4263d0150bfa72e"),
        "CustID" : "C111",
        "AccBal" : 1200,
        "AccType" : "S"
}
{
        "_id" : ObjectId("54993269f4263d0150bfa72f"),
        "CustID" : "C123",
        "AccBal" : 1500,
        "AccType" : "C"
}
>
```

To group on "CustID" and compute the sum of "AccBal", use the below syntax:

```
db.Customers.aggregate( { $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } });
```

```
> db.Customers.aggregate( { $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } });
{ "_id" : "C111", "TotAccBal" : 1200 }
{ "_id" : "C123", "TotAccBal" : 2900 }
>
```

In order to first filter on "AccType:S" and then group it on "CustID" and then compute the sum of "AccBal", use the below syntax:

```
db.Customers.aggregate( { $match : {AccType : "S" } },
{ $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } });
```

```
> db.Customers.aggregate( { $match : {AccType : "S" } },
... { $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } });
{ "_id" : "C111", "TotAccBal" : 1200 }
{ "_id" : "C123", "TotAccBal" : 1400 }
>
```

In order to first filter on "AccType:S" and then group it on "CustID" and then to compute the sum of "AccBal" and then filter those documents wherein the "TotAccBal" is greater than 1200, use the below syntax:

```
db.Customers.aggregate( { $match : {AccType : "S" } },
{ $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } },
{ $match : {TotAccBal : { $gt : 1200 } }});
```

```
> db.Customers.aggregate( { $match : {AccType : "S" } },
... { $group : { _id : "$CustID",TotAccBal : { $sum : "$AccBal" } } },
... { $match : {TotAccBal : { $gt : 1200 } }});
{ "_id" : "C123", "TotAccBal" : 1400 }
>
```

To group on "CustID" and compute the average of the "AccBal" for each group:

db.Customers.aggregate({ $group : { _id : "$CustID",TotAccBal : { $avg : "$AccBal" } } });

```
> db.Customers.aggregate( { $group : { _id : "$CustID",TotAccBal : { $avg : "$AccBal" } } });
{ "_id" : "C111", "TotAccBal" : 1200 }
{ "_id" : "C123", "TotAccBal" : 966.6666666666666 }
>
```

To group on "CustID" and determine the maximum "AccBal" for each group:

db.Customers.aggregate({ $group : { _id : "$CustID",TotAccBal : { $max : "$AccBal" } } });

```
>
> db.Customers.aggregate( { $group : { _id : "$CustID",TotAccBal : { $max : "$AccBal" } } });
{ "_id" : "C111", "TotAccBal" : 1200 }
{ "_id" : "C123", "TotAccBal" : 1500 }
>
>
```

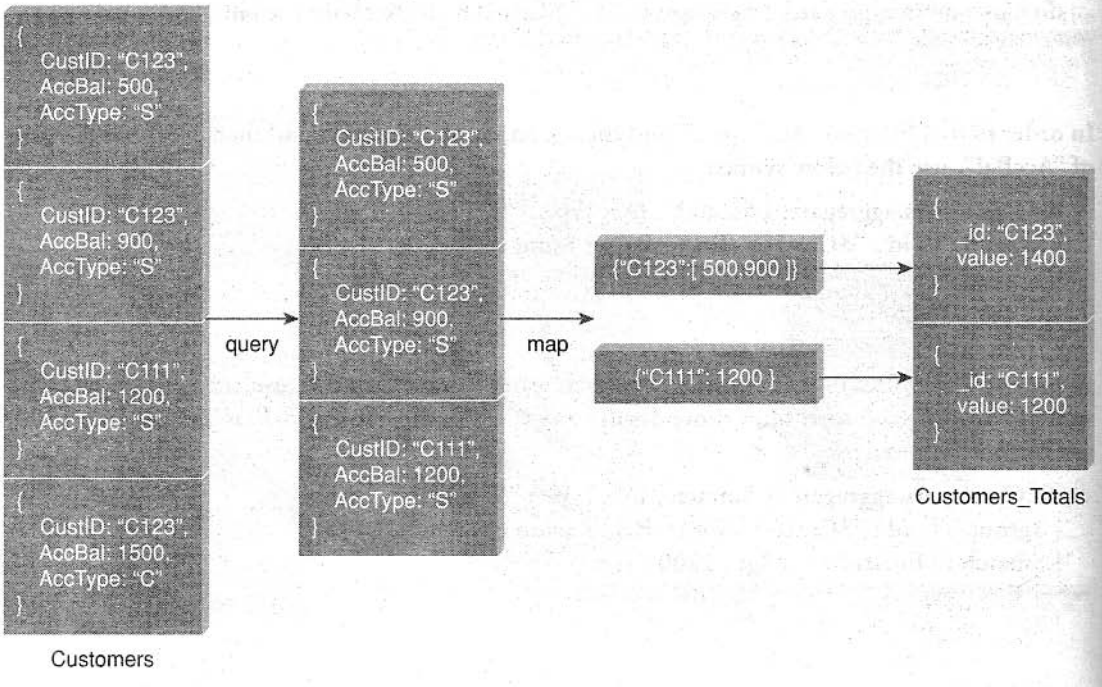
To group on "CustID" and determine the minimum "AccBal" for each group:

db.Customers.aggregate({ $group : { _id : "$CustID",TotAccBal : { $min : "$AccBal" } } });

```
> db.Customers.aggregate( { $group : { _id : "$CustID",TotAccBal : { $min : "$AccBal" } } });
{ "_id" : "C111", "TotAccBal" : 1200 }
{ "_id" : "C123", "TotAccBal" : 500 }
>
```

6.5.10 MapReduce Function

Objective: Consider the collection "Customers" below. There are four documents. Run a query to filter out those documents where the key "AccType" has a value other than "S". Then for each unique CustID, prepare a list of AccBal values. For example, for CustID: "C123", the AccBals are 500,900. This task will be assigned to the mapper function. The output from the mapper function serves as the input to the reducer function. The reducer function then aggregates the AccBal for each CustID. For example, for CustID: "C123", the value is 1400, etc.



Given below is the syntax that we will use to accomplish the objective.

```
db.Customers.mapReduce (
map        →    function() { emit ( this.CustID, this.AccBal ); },
reduce     →    function(key, values) { return Array.sum (values ) },
                {
query      →        query: { AccType: "S"},
output     →        out: "Customer_Totals"
                }
                )
```

Map Function

var map = function(){
emit (this.CustID, this.AccBal);}

```
> var map = function(){
... emit ( this.CustID, this.AccBal );}
>
```

Reduce Function

var reduce = function(key, values){ return Array.sum(values) ; }

```
> var reduce = function(key, values){ return Array.sum(values) ; }
>
```

To execute the query

db.Customers.mapReduce(map, reduce,{out: "Customer_Totals", query:{AccType:"S"}});

```
> db.Customers.mapReduce(map, reduce,{out: "Customer_Totals", query:{AccType:"S"}});
{
        "result" : "Customer_Totals",
        "timeMillis" : 7,
        "counts" : {
                "input" : 3,
                "emit" : 3,
                "reduce" : 1,
                "output" : 2
        },
        "ok" : 1,
}
>
```

The output as archived in "Customer_Totals" collection:

```
> db.Customer_Totals.find().pretty();
{ "_id" : "C111", "value" : 1200 }
{ "_id" : "C123", "value" : 1400 }
>
```

6.5.11 Java Script Programming

Objective: To compute the factorial of a given positive number. The user is required to create a function by the name "factorial" and insert it into the "system.js" collection.

Before we proceed, a quick check on what is contained in the "system.js" collection:

```
> db.system.js.find();
>
```

As per the screenshot above, currently there are no functions in the system.js collection.

Act:

db.system.js.insert({_id:"factorial",

```
value:function(n)
{
        if (n==1)
        return 1;
        else
        return n * factorial(n-1);
}
        }
);
```

```
C:\windows\system32\cmd.exe - mongo
> db.system.js.insert({_id:"factorial",
... value:function(n)
... {
...  if (n==1)
... return 1;
...  else
...  return n * factorial(n-1);
...  }
... }
...  );
WriteResult({ "nInserted" : 1 })
>
```

Confirm the presence of the "factorial" function in the system.js collection.

```
C:\windows\system32\cmd.exe - mongo
> db.system.js.find();
{ "_id" : "factorial", "value" : function (n)
{
 if (n==1)
return 1;
 else
 return n * factorial(n-1);
 } }
>
```

To execute the function "factorial", use the eval() method.

db.eval("factorial(3)");

```
C:\windows\system32\cmd.exe - mongo
> db.eval("factorial(3)");
6
>
```

db.eval("factorial(5)");

```
C:\windows\system32\cmd.exe - mongo
> db.eval("factorial(5)");
120
>
```

db.eval("factorial(1)");

```
C:\windows\system32\cmd.exe - mongo
> db.eval("factorial(1)");
1
>
```

6.5.12 Cursors in MongoDB

Objective: To create a collection by the name "alphabets" and insert documents in it containing two fields, "_id" and "alphabet". The values stored in the "alphabet" field should be "a", "b", "c", "d", etc. with one value stored per document. There should be 26 documents in all. We need to use cursor to iterate through the "alphabets" collection.

Note: "Alphabets" is the name of the collection and "alphabet" is the name of the field.

Act: To create the collection "alphabets" with its 26 documents.

db.alphabets.insert({_id:1,alphabet:"a"});
db.alphabets.insert({_id:2,alphabet:"b"});

```
db.alphabets.insert({_id:3,alphabet:"c"});
db.alphabets.insert({_id:4,alphabet:"d"});
db.alphabets.insert({_id:5,alphabet:"e"});
db.alphabets.insert({_id:6,alphabet:"f"});
db.alphabets.insert({_id:7,alphabet:"g"});
db.alphabets.insert({_id:8,alphabet:"h"});
db.alphabets.insert({_id:9,alphabet:"i"});
db.alphabets.insert({_id:10,alphabet:"j"});
db.alphabets.insert({_id:11,alphabet:"k"});
db.alphabets.insert({_id:12,alphabet:"l"});
db.alphabets.insert({_id:13,alphabet:"m"});
db.alphabets.insert({_id:14,alphabet:"n"});
db.alphabets.insert({_id:15,alphabet:"o"});
db.alphabets.insert({_id:16,alphabet:"p"});
db.alphabets.insert({_id:17,alphabet:"q"});
db.alphabets.insert({_id:18,alphabet:"r"});
db.alphabets.insert({_id:19,alphabet:"s"});
db.alphabets.insert({_id:20,alphabet:"t"});
db.alphabets.insert({_id:21,alphabet:"u"});
db.alphabets.insert({_id:22,alphabet:"v"});
db.alphabets.insert({_id:23,alphabet:"w"});
db.alphabets.insert({_id:24,alphabet:"x"});
db.alphabets.insert({_id:25,alphabet:"y"});
db.alphabets.insert({_id:26,alphabet:"z"});
```

```
C:\windows\system32\cmd.exe - mongo
> db.alphabets.insert({_id:1,alphabet:"a"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:2,alphabet:"b"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:3,alphabet:"c"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:4,alphabet:"d"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:5,alphabet:"e"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:6,alphabet:"f"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:7,alphabet:"g"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:8,alphabet:"h"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:9,alphabet:"i"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:10,alphabet:"j"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:11,alphabet:"k"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:12,alphabet:"l"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:13,alphabet:"m"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:14,alphabet:"n"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:15,alphabet:"o"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:16,alphabet:"p"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:17,alphabet:"q"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:18,alphabet:"r"});
WriteResult({ "nInserted" : 1 })
```

```
> db.alphabets.insert({_id:19,alphabet:"s"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:21,alphabet:"u"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:22,alphabet:"v"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:23,alphabet:"w"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:24,alphabet:"x"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:25,alphabet:"y"});
WriteResult({ "nInserted" : 1 })
> db.alphabets.insert({_id:26,alphabet:"z"});
WriteResult({ "nInserted" : 1 })
>
```

Confirm the presence of 26 documents in the "alphabets" collection.

```
C:\windows\system32\cmd.exe - mongo
> db.alphabets.find();
{ "_id" : 1, "alphabet" : "a" }
{ "_id" : 2, "alphabet" : "b" }
{ "_id" : 3, "alphabet" : "c" }
{ "_id" : 4, "alphabet" : "d" }
{ "_id" : 5, "alphabet" : "e" }
{ "_id" : 6, "alphabet" : "f" }
{ "_id" : 7, "alphabet" : "g" }
{ "_id" : 8, "alphabet" : "h" }
{ "_id" : 9, "alphabet" : "i" }
{ "_id" : 10, "alphabet" : "j" }
{ "_id" : 11, "alphabet" : "k" }
{ "_id" : 12, "alphabet" : "l" }
{ "_id" : 13, "alphabet" : "m" }
{ "_id" : 14, "alphabet" : "n" }
{ "_id" : 15, "alphabet" : "o" }
{ "_id" : 16, "alphabet" : "p" }
{ "_id" : 17, "alphabet" : "q" }
{ "_id" : 18, "alphabet" : "r" }
{ "_id" : 19, "alphabet" : "s" }
{ "_id" : 20, "alphabet" : "t" }
Type "it" for more
>
```

A quick word on how the db.collection.find() method works. This is the primary method for read operation. In other words, it allows one to fetch the documents from the collection. To be able to access the documents, one needs to iterate the cursor.

However, in the mongo shell, if the returned cursor is not assigned to a variable using the var keyword, then cursor is automatically iterated up to 20 times to print the first 20 documents in the result.

```
C:\windows\system32\cmd.exe - mongo
> var myCursor = db.alphabets.find();
> myCursor;
{ "_id" : 1, "alphabet" : "a" }
{ "_id" : 2, "alphabet" : "b" }
{ "_id" : 3, "alphabet" : "c" }
{ "_id" : 4, "alphabet" : "d" }
{ "_id" : 5, "alphabet" : "e" }
{ "_id" : 6, "alphabet" : "f" }
{ "_id" : 7, "alphabet" : "g" }
{ "_id" : 8, "alphabet" : "h" }
{ "_id" : 9, "alphabet" : "i" }
{ "_id" : 10, "alphabet" : "j" }
{ "_id" : 11, "alphabet" : "k" }
{ "_id" : 12, "alphabet" : "l" }
{ "_id" : 13, "alphabet" : "m" }
{ "_id" : 14, "alphabet" : "n" }
{ "_id" : 15, "alphabet" : "o" }
{ "_id" : 16, "alphabet" : "p" }
{ "_id" : 17, "alphabet" : "q" }
{ "_id" : 18, "alphabet" : "r" }
{ "_id" : 19, "alphabet" : "s" }
{ "_id" : 20, "alphabet" : "t" }
Type "it" for more
>
```

Let us now look at designing manual cursors to iterate through the documents in the "alphabets" collection. We will use two methods with manual cursors: hasNext() and next(). We now quickly explain the two methods.

Method 1: hasNext() method. Return value: Boolean. The hasNext() method returns true if the cursor returned by the db.Collection.find() query can iterate further to return more documents.

Method 2: next() method. The next() method returns the next document in the cursor as returned by the db.collection.find() method.

```
C:\windows\system32\cmd.exe - mongo
> var myCur=db.alphabets.find({});
> while(myCur.hasNext()){
... var myRec=myCur.next();
... print("The alphabet is : " + myRec.alphabet);
... }
The alphabet is : a
The alphabet is : b
The alphabet is : c
The alphabet is : d
The alphabet is : e
The alphabet is : f
The alphabet is : g
The alphabet is : h
The alphabet is : i
The alphabet is : j
The alphabet is : k
The alphabet is : l
The alphabet is : m
The alphabet is : n
The alphabet is : o
The alphabet is : p
The alphabet is : q
The alphabet is : r
The alphabet is : s
The alphabet is : t
The alphabet is : u
The alphabet is : v
The alphabet is : w
The alphabet is : x
The alphabet is : y
The alphabet is : z
>
```

The same result can be obtained by iterating through the cursor using a forEach loop.

```
C:\windows\system32\cmd.exe - mongo
> var cur=db.alphabets.find({});
> var myRec;
> cur.forEach( function(myRec) {
... print("The alphabet is : " + myRec.alphabet);
...
... });
The alphabet is : a
The alphabet is : b
The alphabet is : c
The alphabet is : d
The alphabet is : e
The alphabet is : f
The alphabet is : g
The alphabet is : h
The alphabet is : i
The alphabet is : j
The alphabet is : k
The alphabet is : l
The alphabet is : m
The alphabet is : n
The alphabet is : o
The alphabet is : p
The alphabet is : q
The alphabet is : r
The alphabet is : s
The alphabet is : t
The alphabet is : u
The alphabet is : v
The alphabet is : w
The alphabet is : x
The alphabet is : y
The alphabet is : z
>
```

6.5.13 Indexes

Assume the collection with the following documents:

```
> db.books.find().pretty();
{
        "_id" : 6,
        "Category" : "Machine Learning",
        "Bookname" : "Machine Learning for Hackers",
        "Author" : "Drew Conway",
        "qty" : 25,
        "price" : 400,
        "rol" : 30,
        "pages" : 350
}
{
        "_id" : 7,
        "Category" : "Web Mining",
        "Bookname" : "Mining the Social Web",
        "Author" : "Matthew A.Russell",
        "qty" : 55,
        "price" : 500,
        "rol" : 30,
        "pages" : 250
}
{
        "_id" : 8,
        "Category" : "Python",
        "Bookname" : "Python for Data Analysis",
        "Author" : "Wes McKinney",
        "qty" : 8,
        "price" : 150,
        "rol" : 20,
        "pages" : 150
}
{
        "_id" : 9,
        "Category" : "Visualization",
        "Bookname" : "Visualizing Data",
        "Author" : "Ben Fry",
        "qty" : 12,
        "price" : 325,
        "rol" : 6,
        "pages" : 450
}
{
        "_id" : 10,
        "Category" : "Web Mining",
        "Bookname" : "Algorithms for the intelligent web",
        "Author" : "Haralambos Marmanis",
        "qty" : 5,
        "price" : 850,
        "rol" : 10,
        "pages" : 120
}
>
```

Create an index on the key "Category" in the "books" collection.

```
> db.books.ensureIndex({"Category":1});
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
>
```

Check on the status, that is, number and name of the indexes:

```
> db.books.stats();
{
        "ns" : "test.books",
        "count" : 5,
        "size" : 1200,
        "avgObjSize" : 240,
        "storageSize" : 8192,
        "numExtents" : 1,
        "nindexes" : 2,
        "lastExtentSize" : 8192,
        "paddingFactor" : 1,
        "systemFlags" : 1,
```

```
        "userFlags" : 1,
        "totalIndexSize" : 16352,
        "indexSizes" : {
                "_id_" : 8176,
                "Category_1" : 8176
        },
        "ok" : 1
>
```

Get the list of all indexes on the "books" collection:

```
> db.books.getIndexes();
        {
                "v" : 1,
                "key" : {
                        "_id" : 1
                },
                "name" : "_id_",
                "ns" : "test.books"
        },
        {
                "v" : 1,
                "key" : {
                        "Category" : 1
                },
                "name" : "Category_1",
                "ns" : "test.books"
        }
>
```

To use the index on "Category" in the "books" collection, use the hint method:

```
> db.books.find({"Category":"Web Mining"}).pretty().hint({"Category":1});
        "_id" : 7,
        "Category" : "Web Mining",
        "Bookname" : "Mining the Social Web",
        "Author" : "Matthew A.Russell",
        "qty" : 55,
        "price" : 500,
        "rol" : 30,
        "pages" : 250

        "_id" : 10,
        "Category" : "Web Mining",
        "Bookname" : "Algorithms for the intelligent web",
        "Author" : "Haralambos Marmanis",
        "qty" : 5,
        "price" : 850,
        "rol" : 10,
        "pages" : 120
>
```

Check the explain plan to get a deeper understanding on the use of index.

```
> db.books.find({"Category":"Web Mining"}).pretty().hint({"Category":1}).explain();
        "cursor" : "BtreeCursor Category_1",
        "isMultiKey" : false,
        "n" : 2,
        "nscannedObjects" : 2,
        "nscanned" : 2,
        "nscannedObjectsAllPlans" : 2,
        "nscannedAllPlans" : 2,
        "scanAndOrder" : false,
        "indexOnly" : false,
        "nYields" : 0,
        "nChunkSkips" : 0,
```

```
        "millis" : 0,
        "indexBounds" : {
                "Category" : [
                                [
                                        "Web Mining",
                                        "Web Mining"
                                ]
                        ]
        },
        "server" : "PUNITP123103L:27017",
        "filterSet" : false
}
>
```

Let us look at the case of covered index. Observe that the "indexOnly" property will be set to true for covered index.

```
> db.books.find({"Category":"Web Mining"},{"Category":1,_id:0}).pretty().hint({"Category":1}).explain();
{
        "cursor" : "BtreeCursor Category_1",
        "isMultiKey" : false,
        "n" : 2,
        "nscannedObjects" : 0,
        "nscanned" : 2,
        "nscannedObjectsAllPlans" : 0,
        "nscannedAllPlans" : 2,
        "scanAndOrder" : false,
        "indexOnly" : true,
        "nYields" : 0,
        "nChunkSkips" : 0,
        "millis" : 0,
        "indexBounds" : {
                "Category" : [
                                [
                                        "Web Mining",
                                        "Web Mining"
                                ]
                        ]
        },
        "server" : "PUNITP123103L:27017",
        "filterSet" : false
}
>
```

In order to have the index cover the query, ensure that only those columns are projected on which the index is built. In the above example, the index is built on the "Category" column, and "Category" is the only column that is projected. Even the identifier (_id) is suppressed.

6.5.14 MongoImport

This command used at the command prompt imports CSV (Comma Separated Values) or TSV (Tab Separated Values) files or JSON (Java Script Object Notation) documents into MongoDB.

Objective: Given a CSV file "sample.txt" in the D: drive, import the file into the MongoDB collection, "SampleJSON". The collection is in the database "test".

The "sample.txt" file is as follows:

```
_id,FName,LName
1,Samuel,Jones
2,Virat,Kumar
3,Raul,"A Simpson"
4,,"Andrew Simon"
```

Act:

At the command prompt, execute the following command:

Mongoimport --db test --collection SampleJSON --type csv --headerline --file d:\sample.txt

On successful execution of the command, the message at the prompt will be as follows:

```
connected to: 127.0.0.1
2015-02-20T21:09:27.301+0530 imported 4 objects
```

Output: To confirm the output, log into MongoDB shell and navigate to the "SampleJSON" collection in the "test" database.

The following are the JSON documents in the collection:

```
> db
test
> show collections
Customers
SampleJSON
books
fs.chunks
fs.files
persons
system.indexes
usercounters
users
> db.SampleJSON.find().pretty();
{ "_id" : 1, "FName" : "Samuel", "LName" : "Jones" }
{ "_id" : 2, "FName" : "Virat", "LName" : "Kumar" }
{ "_id" : 3, "FName" : "Raul", "LName" : "A Simpson" }
{ "_id" : 4, "FName" : "", "LName" : "Andrew Simon" }
>
```

6.5.15 MongoExport

This command used at the command prompt exports MongoDB JSON documents into CSV (Comma Separated Values) or TSV (Tab Separated Values) files or JSON (Java Script Object Notation) documents.

Objective: This command used at the command prompt exports MongoDB JSON documents from "Customers" collection in the "test" database into a CSV file "Output.txt" in the D: drive.

Given below is a snapshot of the JSON documents in the "Customers" collection of the "test" database.

```
> db
test
> show collections
Customers
SampleJSON
books
fs.chunks
fs.files
persons
system.indexes
usercounters
users
> db.Customers.find().pretty();
{
        "_id" : ObjectId("54df6d4f46a31d28183b9a5b"),
        "CustID" : "C123",
        "AccBal" : 500,
        "AccType" : "S"
}
{
        "_id" : ObjectId("54df6d4f46a31d28183b9a5c"),
        "CustID" : "C123",
        "AccBal" : 900,
        "AccType" : "S"
}
```

```
{
        "_id" : ObjectId("54df6d4f46a31d28183b9a5d"),
        "CustID" : "C111",
        "AccBal" : 1200,
        "AccType" : "S"
}
{
        "_id" : ObjectId("54df6d4f46a31d28183b9a5e"),
        "CustID" : "C123",
        "AccBal" : 1500,
        "AccType" : "C"
}
>
```

Act: At the command prompt, execute the following command:

Mongoexport --db test --collection Customers --csv --fieldFile d:\fields.txt --out d:\output.txt

Before executing this command, ensure that you have created a "fields.txt" with a format defined as follows. The "fields.txt" file:

CustID
AccBal
AccType

For the MongoExport command to execute successfully, ensure that the fields are spelt as is in the MongoDB collection. The case also has to be maintained. It is mandatory to ensure that only one field name is placed per line.

On successful execution of the command, the message at the prompt will be as follows:

```
connected to: 127.0.0.1
exported 4 records
```

Output: To confirm the output, navigate to the D: drive and check the file "Output.txt".

"Output.txt"
CustID,AccBal,AccType
"C123",500.0,"S"
"C123",900.0,"S"
"C111",1200.0,"S"
"C123",1500.0,"C"

6.5.16 Automatic Generation of Unique Numbers for the "_id" Field

Step 1: Run the insert() method on a new collection "usercounters". This is to start off with an initial value of 0 for the "seq" field.

```
db.usercounters.insert(
{
_id: "empid",
seq:0
}
)
```

Step 2: Create a user-defined function "getnextseq". This method will invoke "findAndModify()" method on the "usercounters" collection. This is to increment the value of seq field by 1 and update the same in "usercounters" collection.

```
function getnextseq(name) {
var ret=db.usercounters.findAndModify(
{
query: {_id:name},
update: {$inc:{seq:1}},
new:true
}
);
return ret.seq;
}
```

Step 3: Run the insert() method on the collection where you need to have the "_id" field and get the uniquely generated number. Notice the call to getnextseq() method as value to _id. The return value from the getnextseq() method becomes the value of _id.

```
db.users.insert(
{
_id:getnextseq("empid"),
Name: "sarah jane"
}
)
```

REMIND ME

- MongoDB is a non-relational, open source, distributed database.
- It stores data into JSON (Java Script Object Notation) documents.
- It adheres to CP (Consistency and Partition Tolerant) traits of Brewer's CAP theorem.
- It has NO support for multi-statement transactions.
- It supports embedded documents.
- It practices automatic sharding.

POINT ME (BOOK)

- MongoDB: The definitive guide by Kristina Chodorow, Michael Dirolf, O'Reilly Media.

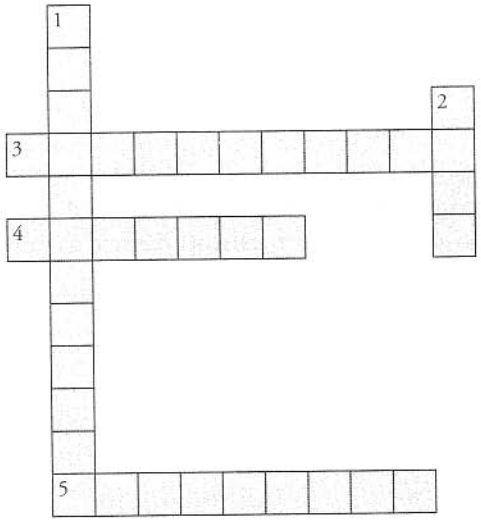
CONNECT ME (INTERNET RESOURCES)

- http://www.mongodb.org/
- https://university.mongodb.com/
- http://www.tutorialspoint.com/mongodb/

TEST ME

A. Crossword

Puzzle on MongoDB



Across

3. MongoDB database stores its data is _____.
4. MongoDB uses _____ schemes.
5. A collection holds one or more _____.

Down

1. MongoDB uses _____ files.
2. MongoDB uses _____, a binary object format similar to, but more expensive than JSON.

Answer:

Across

3. Collections
4. Dynamic
5. Documents

Down

1. Memorymapped
2. BSON

B. Pick the Right Choice

1. MongoDB supports dynamic schema design.
   (a) False (b) True
2. MongoDB supports query joins between collections.
   (a) False (b) True
3. Which of the following MongoDB conditional operator is not a valid operator?
   (a) $lt (c) $gt
   (b) $ltu (d) $lte
4. '$unset' is used with
   (a) Insert (c) Delete
   (b) Update

5. MongoDB is supported by
   (a) Perl
   (b) Python
   (c) PHP
   (d) All the above
6. MongoDB is _____.
   (a) RDBMS
   (b) Object-oriented DBMS
   (c) Document-oriented DBMS
   (d) Key–value store
7. 'MongoImport' command is used _____.
   (a) For multiple command insertion.
   (b) To import content from a JSON, CSV, or TSV export created by MongoExport.
   (c) For multiple command import.
8. Which of the following is the correct command to insert data into MongoDB? Assume that document is a valid JSON document.
   (a) db.Students.insert(document)
   (b) db.Students.insert().(document)
   (c) Students.insert(document)
9. MongoDB documents are represented as _____.
   (a) XML
   (b) JSON
   (c) DOCUMENT
10. MongoDB supports unique indexes just like most other relational databases.
   (a) True
   (b) False
11. Which of the following command creates an index, where mobile_no is a field in the collection, employees?
   (a) db.employees.SetIndex({ "mobile_no": 1 })
   (b) db.employees.ensureIndex({ "mobile_no": 1 })
   (c) employees.SetIndex({ "mobile_no": 1 })
12. Which of the following command is correct when you want to fetch documents from a collection for "only those employees whose salary is either 8500 or 10,000"?
   (a) db.employees.find.sort({"salary" :{$in :[8500,10000]}})
   (b) db.employees.find({"salary" :{"$in :[8500,10000]"}})
   (c) db.employees.find({"salary" :{$in :[8500,10000]}})
13. Which of the following is the command equivalent to
   Select first_name,salary,date_of_join from employees where designation="Manager";
   (a) db.employees.find({"designation":"Manager"},{"first_name" : 1;"salary":1;"date_of_join":1})
   (b) db.employees.find({"designation:Manager"},{"first_name" : 1,"salary":1,"date_of_join":1})
   (c) db.employees.find({"designation":"Manager"},{"first_name" : 1,"salary":1,"date_of_join":1})
14. MongoDB enforces attribute similarity across documents in a collection.
   (a) True
   (b) False
15. The maximum BSON document size is
   (a) 8 megabytes
   (b) 4 megabytes
   (c) 32 megabytes
   (d) 16 megabytes
16. Core MongoDB operations are
   (a) Create, Select, Update, Delete
   (b) Create, Read, Update, Delete
   (c) Create, Read, Update, Drop

17. Which of the following command provides you with a list of all the databases in MongoDB?
    (a) show databases
    (b) show dbs
    (c) show all dbs
    (d) None of the above
18. If we want to remove the document from the collection 'employees' which contains the 'first_name' as "John" then which of the following MongoDB command can be used:
    (a) db.userdetails.remove({})
    (b) db.employees.remove({ "first_name : John" })
    (c) db.employees.remove({ "first_name" : "John" })
19. What does the following command do?
    db.sample.find().limit(10)
    (a) Show 10 documents randomly from the collection sample
    (b) Show only first 10 documents from the collection sample
    (c) Repeats the first document 10 times
20. Which one of the following is equivalent to
    Select * from employees order by salary
    (a) db.employees.find().sort({"salary:1"})
    (b) db.employees.sort({"salary":1})
    (c) db.employees.find().sort({"salary":1})
21. Which command in MongoDB is equivalent to SQL select?
    (a) search()
    (b) find()
    (c) document()
22. Which of the following is equivalent to
    select first_name,salary from employees where designation="Manager";
    Assume that there are three columns first_name,salary,date_of_join.
    (a) db.employees.find({"designation:Manager"},{"date_of_join" : 0})
    (b) db.employees.find({"designation:Manager"},{"date_of_join" : 1})
    (c) db.employees.find({"designation":"Manager"},{"date_of_join" : 0})
23. Which of the following statement is equivalent to the SQL command – Select * from employees where designation = "Manager"?
    (a) employees.find({"designation":"manager"}
    (b) db.employees.find({"designation:manager"}
    (c) db.employees.find({"designation": "manager"}
24. Which of the following is the correct command to update a document?
    (a) db.books.update({ item: "book", qty: { $gt: 7 } }, { $set: { x: 5 }, $inc: { y: 8} })
    (b) db.books.find().update({ item: "book", qty: { $gt: 7 } }, { $set: { x: 5 }, $inc: { y: 8} })
    (c) db.books.update({ item: "book", qty: { $gt: 7 } }, { $set: { x: 5 }, $inc: { y: 8} } , { multi: true })
25. Which one of the following is equivalent to
    Select * from employees order by salary desc;
    (a) db.employees.find().sort({"salary":1})
    (b) db.employees.find().sort({"salary":-1})
    (c) db.employees.sort({"salary":-1})
26. Which of the following command is correct when you want to fetch documents from a collection for only those employees whose salary is either 7500 or date of joining is 17/10/2009?
    (a) db.employees.find({ "salary" : "7500" , ($or : [{ "date_of_join" : "17/10/2009" }] }))
    (b) db.employees.find({ "salary" : "7500" , $or : [{ "date_of_join" : "17/10/2009" }] })
    (c) db.employees.find().sort({ "salary" : "7500" , $or : [{ "date_of_join" : "17/10/2009" }] })

27. What does the following command do?
    db.employees.find().skip(5).limit(5)
    (a) Skips first five documents and then shows just next five documents
    (b) Shows just next five documents
    (c) Skips first five documents and then shows the sixth one five times
28. Which of the following command is correct when you want to fetch documents from collection demo, where value of a field 'interest' is null?
    (a) db.demo.find({ "interest" : null })
    (b) db.demo.find().sort({ "interest" : null })
    (c) db.demo.find({ "interest : null" })

Answers:

1. True
2. False
3. (b)
4. (b)
5. (d)
6. (c)
7. (b)
8. (a)
9. (b)
10. (a)
11. (b)
12. (c)
13. (c)
14. (b)
15. (d)
16. (b)
17. (b)
18. (c)
19. (b)
20. (c)
21. (b)
22. (c)
23. (c)
24. (a)
25. (b)
26. (c)
27. (a)
28. (a)

C. Unsolved Exercises

1. Enumerate few features of MongoDB.
2. List the difference between SQL and MongoDB.
3. Explain Map Reduce programming in MongoDB with a suitable example.
4. What is a cursor? How is a cursor implemented in MongoDB. Explain with a suitable example.
5. What is the significance of _id key in MongoDB?

ASSIGNMENTS FOR HANDS-ON PRACTICE

ASSIGNMENT 1

Objective: **To practice MapReduce programming in MongoDB.**

Step 1: Execute the below statements at the MongoDB shell prompt.

db.books.save({ _id:1,Category:"Machine Learning", Bookname:"Machine Learning for Hackers", Author:"Drew Conway",qty:25,price:400,rol:30,pages:350});

db.books.save({ _id:2,Category:"Business Intelligence", Bookname:"Fundamentals of Business Analytics", Author:"Seema Acharya",qty:55,price:500,rol:30,pages:250});

db.books.save({ _id:3,Category:"Analytics", Bookname:" Competing on Analytics", Author:"Thomas Davenport",qty:8,price:150,rol:20,pages:150});

db.books.save({ _id:4,Category:"Visualization", Bookname:"Visualizing Data", Author:"Ben Fry",qty:12, price:325,rol:6,pages:450});

db.books.save({ _id:5,Category:"Web Mining", Bookname:" Learning R ", Author:" Richard Cotton",qty:5, price:850,rol:10,pages:120});

Step 2: Confirm the presence of the above documents in the "books" collection.

Step 3: Write map and reduce functions to split the books into the following two categories:

(a) Big books

(b) Small books

Books which have more than 300 pages should be in the big book category. Books which have less than 300 pages should be in the small book category.

Step 4: Count the number of books in each category.

Step 5: Store the output as follows as documents in a new collection, called, "Book_Result".

Book Category	Count of the Books
(a) Big books	2
(b) Small books	3

ASSIGNMENT 2

Objective: To practice import, export, and aggregation in MongoDB.

Step 1: Pick any public dataset from the site www.kdnuggets.com. Convert it into CSV format. Make sure that you have at least two numeric columns.

Step 2: Use MongoImport to import data from the CSV format file into MongoDB collection. "MongoDBHandsOn" in test database.

Step 3: Identify a grouping column.

Step 4: Compute the sum of the values in the first numeric column.

Step 5: Compute the average of the values in the second numeric column.

ASSIGNMENT 3

Objective: To copy the JSON documents from one MongoDB collection to another MongoDB collection.

ASSIGNMENT 4

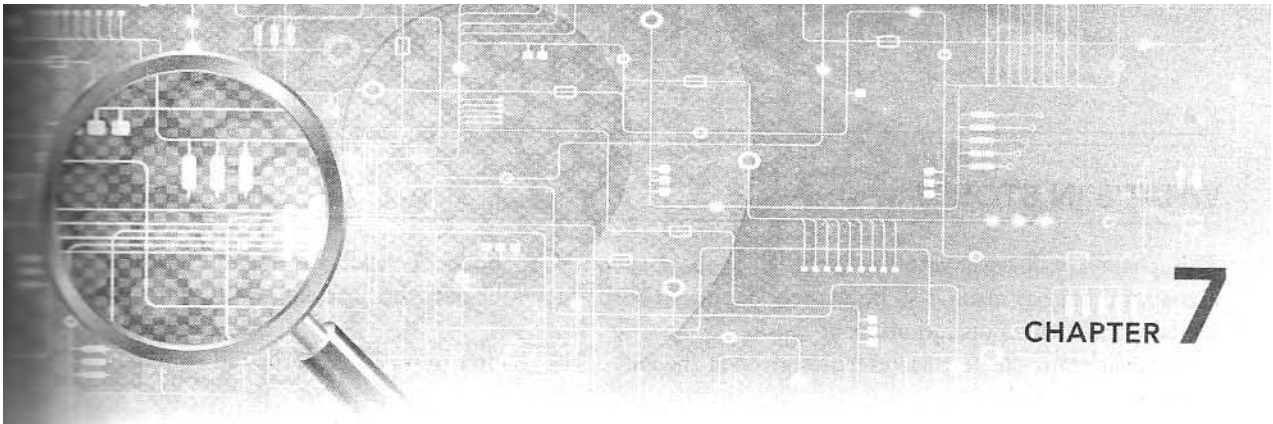
Objective: Write the insert method to store the following document in MongoDB.

Name: "Stephen More"

Address:

```
{ "City"        : "Bangalore",
 "Street"       : "Electronics City",
 "Affiliation"  : "XYZ Ltd"
}
```

Hobbies: Chess, Lawn Tennis, Base ball



CHAPTER 7

Introduction to Cassandra

BRIEF CONTENTS

- What's in Store?
- Apache Cassandra – An Introduction
- Features of Cassandra
  - Peer-to-Peer Network
  - Gossip and Failure Detection
  - Partitioner
  - Replication Factor
  - Anti-Entropy and Read Repair
  - Writes in Cassandra
  - Hinted Handoffs
  - Tunable Consistency: Read Consistency and Write Consistency
- CQL Data Types
- CQLSH
- Keyspaces
- CRUD Operations
  - Insert
  - Update
  - Delete
  - Select
- Collections
  - Set Collection
  - List Collection
  - Map Collection
- Using a Counter
- Time To Live (TTL)
- Alter Commands
  - Alter Table to Change the Data Type of a Column
  - Alter Table to Delete a Column
  - Drop a Table
  - Drop a Database
- Import and Export
  - Export to CSV
  - Import from CSV
  - Import from STDIN
  - Export to STDOUT
- Querying System Tables
- Practice Examples

"Data is a precious thing and will last longer than the systems themselves."

– Tim Berners-Lee, inventor of the World Wide Web.

WHAT'S IN STORE?

This chapter will cover another NoSQL database called "Cassandra". We will explore the features of Cassandra that has made it so immensely popular. The chapter will cover the basic CRUD (Create, Read, Update, and Delete) operations using cqlsh.

Please attempt the Test Me exercises given at the end of the chapter to practice, learn, and comprehend Cassandra effectively.

7.1 APACHE CASSANDRA – AN INTRODUCTION

We shall start this chapter with few points that a reader should know about Cassandra.

1. Apache Cassandra was born at Facebook. After Facebook open sourced the code in 2008, Cassandra became an Apache Incubator project in 2009 and subsequently became a top-level Apache project in 2010.
2. It is built on Amazon's dynamo and Google's BigTable.
3. Cassandra does NOT compromise on availability. Since it does not have a master-slave architecture, there is no question of single point of failure. This proves beneficial for business critical applications that need to be up and running always and cannot afford to go down ever.
4. It is highly scalable (it scales out), high performance distributed database. It distributes and manages gigantic amount of data across commodity servers.

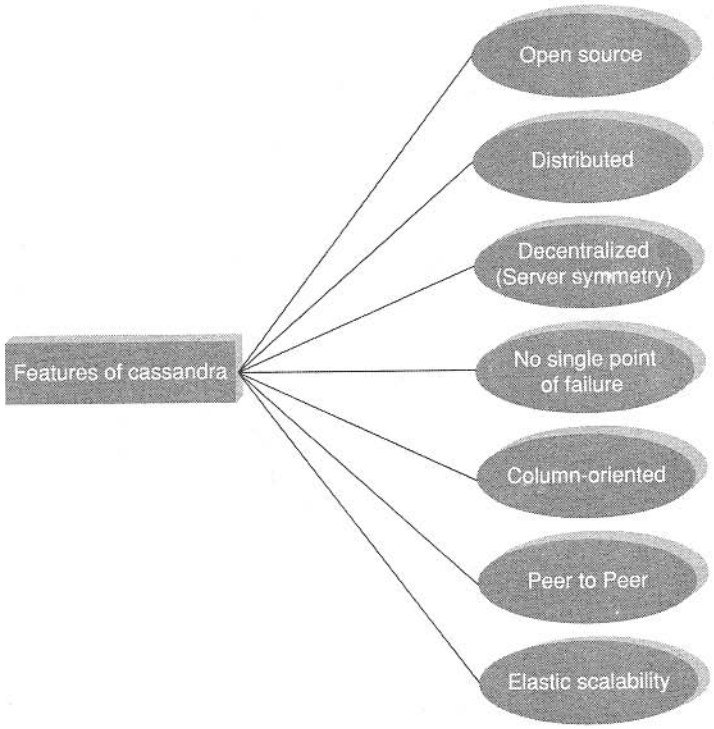


Figure 7.1 Features of Cassandra.

5. It is a column-oriented database designed to support peer-to-peer symmetric nodes instead of the master-slave architecture.
6. It has adherence to the Availability and Partition Tolerance properties of CAP theorem. It takes care of consistency using BASE (Basically Available Soft State Eventual Consistency) approach.

Refer Figure 7.1. Few companies that have successfully deployed Cassandra and have benefitted immensely from it are as follows:

1. Twitter
2. Netflix
3. Cisco
4. Adobe
5. eBay
6. Rackspace

7.2 FEATURES OF CASSANDRA

7.2.1 Peer-to-Peer Network

As with any other NoSQL database, Cassandra is designed to distribute and manage large data loads across multiple nodes in a cluster constituted of commodity hardware. Cassandra does NOT have a master-slave architecture which means that it does NOT have single point of failure. A node in Cassandra is structurally identical to any other node. Refer Figure 7.2. In case a node fails or is taken offline, it definitely impacts the throughput. However, it is a case of graceful degradation where everything does not come crashing at any given instant owing to a node failure. One can still go about business as usual. It tides over the problem of failure by employing a peer-to-peer distributed system across homogeneous nodes. It ensures that data is distributed across all nodes in the cluster. Each node exchanges information across the cluster every second.

Let us look at how a Cassandra node writes. Each write is written to the commit log sequentially. A write is taken to be successful only if it is written to the commit log. Data is then indexed and pushed to an in-memory structure called "Memtable". When the in-memory data structure, "the Memtable", is full, the contents are flushed to "SSTable" (Sorted String) data file on the disk. The SSTable is immutable and is

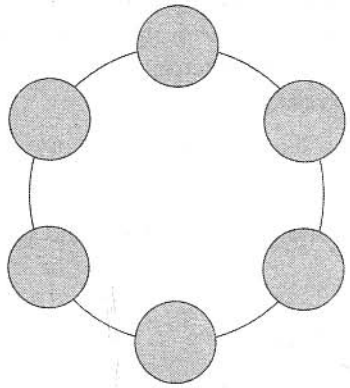


Figure 7.2 Sample Cassandra cluster.

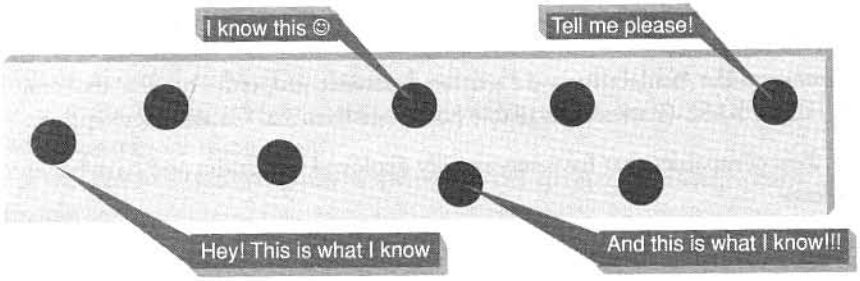


Figure 7.3 Gossip protocol.

append-only. It is stored on disk sequentially and is maintained for each Cassandra table. The partitioning and replication of all writes are performed automatically across the cluster.

7.2.2 Gossip and Failure Detection

Gossip protocol is used for intra-ring communication. It is a peer-to-peer communication protocol which eases the discovery and sharing of location and state information with other nodes in the cluster. Refer Figure 7.3. Although there are quite a few subtleties involved, but at its core it's a simple and robust system. A node only has to send out the communication to a subset of other nodes. For repairing unread data, Cassandra uses what's called an anti-entropy version of the gossip protocol.

7.2.3 Partitioner

A partitioner takes a call on how to distribute data on the various nodes in a cluster. It also determines the node on which to place the very first copy of the data. Basically a partitioner is a hash function to compute the token of the partition key. The partition key helps to identify a row uniquely.

7.2.4 Replication Factor

The replication factor determines the number of copies of data (replicas) that will be stored across nodes in a cluster. If one wishes to store only one copy of each row on one node, they should set the replication factor to one. However, if the need is for two copies of each row of data on two different nodes, one should go with a replication factor of two. The replication factor should ideally be more than one and not more than the number of nodes in the cluster. A replication strategy is employed to determine which nodes to place the data on. Two replication strategies are available:

1. SimpleStrategy.
2. NetworkToplogyStrategy.

The preferred one is NetworkTopologyStrategy as it is simple and supports easy expansion to multiple data centers, should there be a need.

7.2.5 Anti-Entropy and Read Repair

A cluster is made up of several nodes. Since the cluster is constituted of commodity hardware, it is prone to failure. In order to achieve fault tolerance, a given piece of data is replicated on one or more nodes. A client

can connect to any node in the cluster to read data. How many nodes will be read before responding to the client is based on the consistency level specified by the client. If the client-specified consistency is not met, the read operation blocks. There is a possibility that few of the nodes may respond with an out-of-date value. In such a case, Cassandra will initiate a read repair operation to bring the replicas with stale values up to date.

For repairing unread data, Cassandra uses an anti-entropy version of the gossip protocol. Anti-entropy implies comparing all the replicas of each piece of data and updating each replica to the newest version. The read repair operation is performed either before or after returning the value to the client as per the specified consistency level.

7.2.6 Writes in Cassandra

Let us look at behind the scene activities. Here is a client that initiates a write request. Where does his write get written to? It is first written to the commit log. A write is taken as successful only if it is written to the commit log. The next step is to push the write to a memory resident data structure called Memtable. A threshold value is defined in the Memtable. When the number of objects stored in the Memtable reaches a threshold, the contents of Memtable are flushed to the disk in a file called SSTable (Sorted String Table). Flushing is a non-blocking operation. It is possible to have multiple Memtables for a single column family. One out of them is current and the rest are waiting to be flushed.

7.2.7 Hinted Handoffs

The first question that arises is: Why Cassandra is all for availability? It works on the philosophy that it will always be available for writes.

Assume that we have a cluster of three nodes – Node A, Node B, and Node C. Node C is down for some reason. Refer Figure 7.4. We are maintaining a replication factor of 2 which implies that two copies of each row will be stored on two different nodes. The client makes a write request to Node A. Node A is the coordinator and serves as a proxy between the client and the nodes on which the replica is to be placed. The client

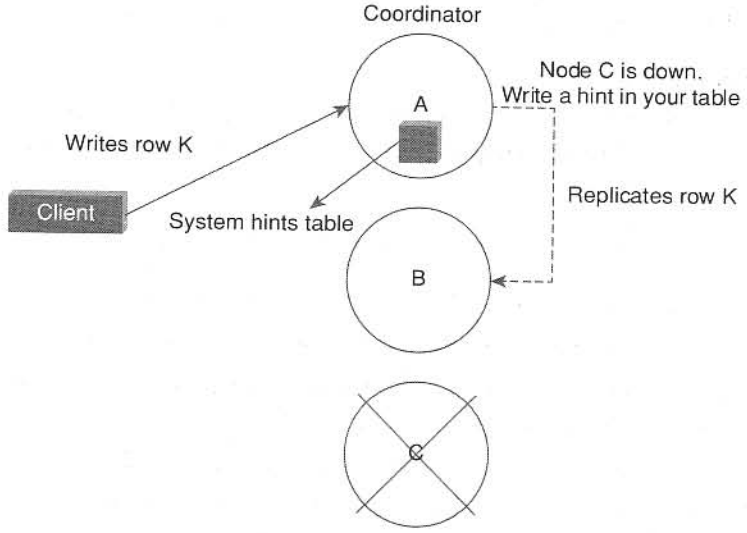


Figure 7.4 Depiction of hinted handoffs.

writes Row K to Node A. Node A then writes Row K to Node B and stores a hint for Node C. The hint will have the following information:

1. Location of the node on which the replica is to be placed.
2. Version metadata.
3. The actual data.

When Node C recovers and is back to the functional self, Node A reacts to the hint by forwarding the data to Node C.

7.2.8 Tunable Consistency

One of the features of Cassandra that has made it immensely popular is its ability to utilize tunable consistency. The database systems can go for either strong consistency or eventual consistency. Cassandra can cash in on either flavor of consistency depending on the requirements. In a distributed system, we work with several servers in the system. Few of these servers are in one data center and others in other data centers. Let us take a look at what it means by strong consistency and eventual consistency.

1. **Strong consistency:** If we work with strong consistency, it implies that each update propagates to all locations where that piece of data resides. Let us assume a single data center setup. Strong consistency will ensure that all of the servers that should have a copy of the data, will have it, before the client is acknowledged with a success. If we are wondering whether it will impact performance, yes it will. It will cost a few extra milliseconds to write to all servers.
2. **Eventual consistency:** If we work with eventual consistency, it implies that the client is acknowledged with a success as soon as a part of the cluster acknowledges the write. When should one go for eventual consistency? The choice is fairly obvious... when application performance matters the most. Example: A single server acknowledges the write and then begins propagating the data to other servers.

7.2.8.1 Read Consistency

Let us understand what the read consistency level means. It means how many replicas must respond before sending out the result to the client application. There are several read consistency levels as mentioned in Table 7.1.

7.2.8.2 Write Consistency

Let us understand what the write consistency level means. It means on how many replicas write must succeed before sending out an acknowledgement to the client application. There are several write consistency levels as mentioned in Table 7.2.

Table 7.1 Read consistency levels in Cassandra

ONE	Returns a response from the closest node (replica) holding the data.
QUORUM	Returns a result from a quorum of servers with the most recent timestamp for the data.
LOCAL_ QUORUM	Returns a result from a quorum of servers with the most recent timestamp for the data in the same data center as the coordinator node.
EACH_QUORUM	Returns a result from a quorum of servers with the most recent timestamp in all data centers.
ALL	This provides the highest level of consistency of all levels and the lowest level of availability of all levels. It responds to a read request from a client after all the replica nodes have responded.

Table 7.2 Write consistency levels in Cassandra

ALL	This is the highest level of consistency of all levels as it necessitates that a write must be written to the commit log and Memtable on all replica nodes in the cluster.
EACH_QUORUM	A write must be written to the commit log and Memtable on a quorum of replica nodes in *all* data centers.
QUORUM	A write must be written to the commit log and Memtable on a quorum of replica nodes.
LOCAL_QUORUM	A write must be written to the commit log and Memtable on a quorum of replica nodes in the same data center as the coordinator node. This is to avoid latency of inter-data center communication.
ONE	A write must be written to the commit log and Memtable of at least one replica node.
TWO	A write must be written to the commit log and Memtable of at least two replica nodes.
THREE	A write must be written to the commit log and Memtable of at least three replica nodes.
LOCAL_ONE	A write must be sent to, and successfully acknowledged by, at least one replica node in the local data center.

7.3 CQL DATA TYPES

Refer Table 7.3 for built-in data types for columns in CQL.

Table 7.3 Built-in data types in Cassandra

Int	32 bit signed integer
Bigint	64 bit signed long
Double	64-bit IEEE-754 floating point
Float	32-bit IEEE-754 floating point
Boolean	True or false
Blob	Arbitrary bytes, expressed in hexadecimal
Counter	Distributed counter value
Decimal	Variable – precision integer
List	A collection of one or more ordered elements
Map	A JSON style array of elements
Set	A collection of one or more elements
Timestamp	Date plus time
Varchar	UTF 8 encoded string
Varint	Arbitrary-precision integers
Text	UTF 8 encoded string

7.4 CQLSH

7.4.1 Logging into cqlsh

The below screenshot depicts the cqlsh command prompt after logging in, using cqlsh succeeds.

```
C:\windows\system32\cmd.exe - Python cqlsh

d:\apache-cassandra-2.0.0\apache-cassandra-2.0.0\apache-cassandra-2.0.0\bin>Python cqlsh
Connected to Test Cluster at localhost:9160.
[cqlsh 4.0.0 | Cassandra 2.0.0 | CQL spec 3.1.0 | Thrift protocol 19.37.0]
Use HELP for help.
cqlsh>
```

The upcoming sections have been designed as follows:

Objective: What is it that we are trying to achieve here?
Input (optional): What is the input that has been given to us to act upon?
Act: The actual statement /command to accomplish the task at hand.
Outcome: The result/output as a consequence of executing the statement.

Objective: To get help with CQL.
Act:

Help

Outcome:

```
C:\windows\system32\cmd.exe - Python cqlsh

d:\apache-cassandra-2.0.0\apache-cassandra-2.0.0\apache-cassandra-2.0.0\bin>Python cqlsh
Connected to Test Cluster at localhost:9160.
[cqlsh 4.0.0 | Cassandra 2.0.0 | CQL spec 3.1.0 | Thrift protocol 19.37.0]
Use HELP for help.
cqlsh> help

Documented shell commands:
==========================
CAPTURE      COPY  DESCRIBE  EXPAND  SHOW    TRACING
CONSISTENCY  DESC  EXIT      HELP    SOURCE

CQL help topics:
================
ALTER                        CREATE_TABLE_OPTIONS  REVOKE
ALTER_ADD                    CREATE_TABLE_TYPES    SELECT
ALTER_ALTER                  CREATE_USER           SELECT_COLUMNFAMILY
ALTER_DROP                   DELETE                SELECT_EXPR
ALTER_RENAME                 DELETE_COLUMNS        SELECT_LIMIT
ALTER_USER                   DELETE_USING          SELECT_TABLE
ALTER_WITH                   DELETE_WHERE          SELECT_WHERE
APPLY                        DROP                  TEXT_OUTPUT
ASCII_OUTPUT                 DROP_COLUMNFAMILY     TIMESTAMP_INPUT
BEGIN                        DROP_INDEX            TIMESTAMP_OUTPUT
BLOB_INPUT                   DROP_KEYSPACE         TRUNCATE
BOOLEAN_INPUT                DROP_TABLE            TYPES
CREATE                       DROP_USER             UPDATE
CREATE_COLUMNFAMILY          GRANT                 UPDATE_COUNTERS
CREATE_COLUMNFAMILY_OPTIONS  INSERT                UPDATE_SET
CREATE_COLUMNFAMILY_TYPES    LIST                  UPDATE_USING
CREATE_INDEX                 LIST_PERMISSIONS      UPDATE_WHERE
CREATE_KEYSPACE              LIST_USERS            USE
CREATE_TABLE                 PERMISSIONS           UUID_INPUT

cqlsh>
```

7.5 KEYSPACES

What is a keyspace? A keyspace is a container to hold application data. It is comparable to a relational database. It is used to group column families together. Typically, a cluster has one keyspace per application. Replication is controlled on a per keyspace basis. Therefore, data that has different replication requirements should reside on different keyspaces.

When one creates a keyspace, it is required to specify a strategy class. There are two choices available with us. Either we can specify a "SimpleStrategy" or a "NetworkTopologyStrategy" class. While using Cassandra for evaluation purpose, go with "SimpleStrategy" class and for production usage, work with the "NetworkTopologyStrategy" class.

Objective: To create a keyspace by the name "Students".

Act:

```
CREATE KEYSPACE Students WITH REPLICATION = {
        'class':'SimpleStrategy',
        'replication_factor':1
};
```

Outcome:

```
C:\windows\system32\cmd.exe - Python cqlsh
cqlsh> CREATE KEYSPACE Students WITH REPLICATION = {
   ...             'class':'SimpleStrategy',
   ...             'replication_factor':1
   ...  };
cqlsh>
```

The replication factor stated above in the syntax for creating keyspace is related to the number of copies of keyspace data that is housed in a cluster.

Objective: To describe all the existing keyspaces.

Act:

```
DESCRIBE KEYSPACES;
```

Outcome:

```
C:\windows\system32\cmd.exe - python cqlsh

d:\apache-cassandra-2.0.0\apache-cassandra-2.0.0\apache-cassandra-2.0.0\bin>python cq
Connected to Test Cluster at localhost:9160.
[cqlsh 4.0.0 | Cassandra 2.0.0 | CQL spec 3.1.0 | Thrift protocol 19.37.0]
Use HELP for help.
cqlsh> describe keyspaces;

system  students  system_traces

cqlsh>
```

Objective: To get more details on the existing keyspaces such as keyspace name, durable writes, strategy class, strategy options, etc.

Act:

```
SELECT *
        FROM system.schema_keyspaces;
```

Outcome:

```
cqlsh> SELECT * FROM system.schema_keyspaces;

 keyspace_name | durable_writes | strategy_class                                    | strategy_options
---------------+----------------+---------------------------------------------------+-----------------------------
      demo_con |           True | org.apache.cassandra.locator.SimpleStrategy       | {"replication_factor":"3"}
        system |           True |  org.apache.cassandra.locator.LocalStrategy       | {}
 system_traces |           True | org.apache.cassandra.locator.SimpleStrategy       | {"replication_factor":"1"}
      students |           True | org.apache.cassandra.locator.SimpleStrategy       | {"replication_factor":"1"}

(4 rows)
```

Note: Cassandra converted the Students keyspace to lowercase as quotation marks were not used.

Objective: To use the keyspace "Students", use the following command:

Use keyspace_name

Use connects the client session to the specified keyspace.

Act:

USE Students;

Outcome:

```
C:\windows\system32\cmd.exe - python cqlsh
cqlsh> use Students;
cqlsh:students>
```

Objective: To create a column family or table by the name "student_info".

Act:

CREATE TABLE Student_Info (
RollNo int PRIMARY KEY,
StudName text,
DateofJoining timestamp,
LastExamPercent double
);

Outcome:

```
cqlsh> use Students;
cqlsh:students> CREATE TABLE Student_Info (
            ...   RollNo int PRIMARY Key,
            ...   StudName text,
            ...   DateofJoining timestamp,
            ...   LastExamPercent double
            ...   );
```

The table "student_info" gets created in the keyspace "students".

Note: Tables can have either a single or compound primary key. Always ensure that there is exactly one primary key definition. The primary key, however, can be simple (consisting of a single attribute) or composite (comprising two or more attributes).

Explanation about the composite PRIMARY KEY:

Primary key (column_name1, *column_name2, column_name3* ...)
Primary key ((column_name4, *column_name5*), column_name6, *column_name7* ...)

In the above syntax,
column_name1 is the partition key
column_name2 and column_name3 are the clustering columns.
column_name4 and column_name5 are the partitioning keys
column_name6 and column_name7 are the clustering columns.

The partition key is used to distribute the data in the table across various nodes that constitute the cluster. The clustering columns are used to store data in ascending order on the disk.

Objective: To lookup the names of all tables in the current keyspace, or in all the keyspaces if there is no current keyspace.

Act:

DESCRIBE TABLES;

Outcome:

```
cqlsh:students> describe tables;

student_info

cqlsh:students>
```

Objective: To describe the table "student_info" use the below command.

Act:

DESCRIBE TABLE student_info;

Note: The output is a list of CQL commands with the help of which the table "student_info" can be recreated.

Outcome:

```
C:\windows\system32\cmd.exe - python cqlsh
cqlsh:students> describe table student_info;

CREATE TABLE student_info (
  rollno int,
  dateofjoining timestamp,
  lastexampercent double,
  studname text,
  PRIMARY KEY (rollno)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

7.6 CRUD (CREATE, READ, UPDATE, AND DELETE) OPERATIONS

Objective: To insert data into the column family "student_info".

An insert writes one or more columns to a record in Cassandra table atomically. An insert statement does not return an output. One is not required to place values in all the columns; however, it is mandatory to specify all the columns that make up the primary key. The columns that are missing do not occupy any space on disk.

Internally insert and update operations are equal. However, insert does not support counters but update does. Counters will be discussed later in the chapter.

Act:

```
BEGIN BATCH
INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
VALUES (1,'Michael Storm','2012-03-29', 69.6)
INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
VALUES (2,'Stephen Fox','2013-02-27', 72.5)
INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
VALUES (3,'David Flemming','2014-04-12', 81.7)
INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
VALUES (4,'Ian String','2012-05-11', 73.4)
APPLY BATCH;
```

Outcome:

```
cqlsh:students> BEGIN BATCH
           ...   INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
           ... VALUES (1,'Michael Storm','2012-03-29',69.6)
           ...   INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
           ... VALUES (2,'Stephen Fox','2013-02-27',72.5)
           ...   INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
           ...  VALUES (3,'David Flemming ','2014-04-12',81.7)
           ...   INSERT INTO student_info (RollNo,StudName,DateofJoining,LastExamPercent)
           ...  VALUES (4,'Ian String','2012-05-11',73.4)
           ... APPLY BATCH;
cqlsh:students>
```

Objective: To view the data from the table "student_info".

Act:

```
SELECT *
      FROM student_info;
```

The above select statement retrieves data from the "student_info" table.

Outcome:

```
cqlsh:students> select * from student_info;

 rollno | dateofjoining                          | lastexampercent | studname
--------+----------------------------------------+-----------------+----------------
      1 | 2012-03-29 00:00:00India Standard Time |            69.6 |   Michael Storm
      2 | 2013-02-27 00:00:00India Standard Time |            72.5 |     Stephen Fox
      4 | 2012-05-11 00:00:00India Standard Time |            73.4 |      Ian String
      3 | 2014-04-12 00:00:00India Standard Time |            81.7 | David Flemming

(4 rows)

cqlsh:students>
```

Objective: To view only those records where the RollNo column either has a value 1 or 2 or 3.

Act:

```
SELECT *
    FROM student_info
        WHERE RollNo IN(1,2,3);
```

Note: For the above statement to execute successfully, ensure that the following criteria are satisfied:

1. Either the partition key definition includes the column that is used in the where clause i.e. search criteria.
2. OR the column being used in the where clause, that is, search criteria, has an index defined on it using the CREATE INDEX statement.

Outcome:

```
cqlsh:students> Select * from student_info where RollNo IN(1,2,3);

 rollno | dateofjoining                          | lastexampercent | studname
--------+----------------------------------------+-----------------+-----------------
      1 | 2012-03-29 00:00:00India Standard Time |            69.6 |   Michael Storm
      2 | 2013-02-27 00:00:00India Standard Time |            72.5 |     Stephen Fox
      3 | 2014-04-12 00:00:00India Standard Time |            81.7 | David Flemming

(3 rows)

cqlsh:students>
```

Let us try running a query with "studname" in the where clause. Since "studname" is neither the primary key column nor a column in the primary key definition and also does not have an index defined on it, such a query will lead to error.

We set the stage to resolve the error by creating an index on the "studname" column of the "student_info" table and then subsequently executing the query.

To create an index on the "studname" column of the "student_info" column family use

```
CREATE INDEX ON student_info(studname)
```

To execute the query using the index defined on "studname" column use

```
SELECT *
    FROM student_info
        WHERE studname='Stephen Fox' ;
```

Outcome:

```
cqlsh:students> create index on student_info(studname);
cqlsh:students> select * from student_info where studname='Stephen Fox' ;

 rollno | dateofjoining                          | lastexampercent | studname
--------+----------------------------------------+-----------------+-------------
      2 | 2013-02-27 00:00:00India Standard Time |            72.5 | Stephen Fox

(1 rows)

cqlsh:students>
```

Objective: Let us create another index on the "LastExamPercent" column of the "student_info" column family.

Act:

CREATE INDEX ON student_info(LastExamPercent);

Outcome:

```
cqlsh:students> create index on student_info(Last
cqlsh:students> select * from student_info where LastExamPercent = 81.7;
 rollno | dateofjoining                          | lastexampercent | studname
--------+----------------------------------------+-----------------+-----------------
      3 | 2014-04-12 00:00:00India Standard Time |            81.7 | David Flemming

(1 rows)
```

Objective: To specify the number of rows returned in the output using limit.

Act:

SELECT rollno, hobbies, language, lastexampercent
FROM student_info LIMIT 2;

Outcome:

```
cqlsh:students> select rollno, hobbies, language, lastexampercent from student_info limit 2;
 rollno | hobbies                              | language          | lastexampercent
--------+--------------------------------------+-------------------+-----------------
      1 |                  {'Chess, Table Tennis'} | ['Hindi, English'] |            69.6
      4 | {'Lawn Tennis, Table Tennis, Golf'} | ['Hindi, English'] |            73.4

(2 rows)
```

Objective: To use column alias for the column "language" in the "student_info" table. We would like the column heading to be "knows language".

Act:

SELECT rollno, language AS "knows language"
FROM student_info;

Outcome:

```
cqlsh:students> select rollno, language as "knows language" from student_info;
 rollno | knows language
--------+---------------------------
      1 |          ['Hindi, English']
      4 |          ['Hindi, English']
      3 | ['Hindi,English,French']

(3 rows)
```

Objective: To update the value held in the "StudName" column of the "student_info" column family to "David Sheen" for the record where the RollNo column has value = 2.

Note: An update updates one or more column values for a given row to the Cassandra table. It does not return anything.

Act:

UPDATE student_info SET StudName = 'David Sheen' WHERE RollNo = 2;

Outcome:

```
cqlsh:students> UPDATE student_info  SET StudName = 'David Sheen'  WHERE RollNo = 2;
cqlsh:students> select * from student_info where rollno = 2;

 rollno | dateofjoining                                | lastexampercent | studname
--------+----------------------------------------------+-----------------+-------------
      2 | 2013-02-27 00:00:00India Standard Time |            72.5 | David Sheen

(1 rows)

cqlsh:students>
```

Objective: Let us try updating the value of a primary key column.

Act:

UPDATE student_info SET rollno=6 WHERE rollno=3;

Outcome:

```
cqlsh:students> update student_info set rollno=6 where rollno=3;
Bad Request: PRIMARY KEY part rollno found in SET part
cqlsh:students>
```

Note: It does not allow update to a primary key column.

Objective: Updating more than one column of a row of Cassandra table.

Act:

Step 1: Before the update

```
cqlsh:students> select rollno, studname, lastexampercent from student_info where rollno=3;

 rollno | studname        | lastexampercent
--------+-----------------+-----------------
      3 | David Flemming  |            81.7

(1 rows)
```

Step 2: Applying the update

```
cqlsh:students> select rollno, studname, lastexampercent from student_info where rollno=3;

 rollno | studname | lastexampercent
--------+----------+-----------------
      3 |  Samaira |              85

(1 rows)
```

Step 3: After the update

```
cqlsh:students> DELETE LastExamPercent FROM student_info where RollNo=2;
cqlsh:students> select * from student_info where rollno = 2;

 rollno | dateofjoining                                | lastexampercent | studname
--------+----------------------------------------------+-----------------+-------------
      2 | 2013-02-27 00:00:00India Standard Time |            null | David Sheen

(1 rows)

cqlsh:students>
```

Objective: To delete the column "LastExamPercent" from the "student_info" table for the record where the RollNo = 2.

Note: Delete statement removes one or more columns from one or more rows of a Cassandra table or removes entire rows if no columns are specified.

Act:

DELETE LastExamPercent FROM student_info WHERE RollNo=2;

Outcome:

```
cqlsh:students> DELETE LastExamPercent FROM student_info where RollNo=2;
cqlsh:students> select * from student_info where rollno = 2;

 rollno | dateofjoining                          | lastexampercent | studname
--------+----------------------------------------+-----------------+-------------
      2 | 2013-02-27 00:00:00India Standard Time |            null | David Sheen

(1 rows)

cqlsh:students>
```

Objective: To delete a row (where RollNo = 2) from the table "student_info".

Act:

DELETE FROM student_info WHERE RollNo=2;

Outcome:

```
cqlsh:students> DELETE  FROM student_info where RollNo=2;
cqlsh:students> select * from student_info where rollno=2;

(0 rows)

cqlsh:students>
```

Objective: To create a table "project _details" with primary key as (project_id, project_name).

Act:

CREATE TABLE project_details (
project_id int,
project_name text,
stud_name text,
rating double,
duration int,
PRIMARY KEY (project_id, project_name));

Outcome:

```
cqlsh:students> CREATE TABLE project_details (
            ...   project_id int,
            ...   project_name text,
            ...   stud_name text,
            ...   rating double,
            ...   duration int,
            ...   PRIMARY KEY (project_id, project_name));
cqlsh:students>
```

Objective: To insert data into the column family "project_details".

Act:

```
BEGIN BATCH
INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
VALUES (1,'MS data migration','David Sheen',3.5,720)
INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
VALUES (1,'MS Data Warehouse','David Sheen',3.9,1440)
INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
VALUES (2,'SAP Reporting','Stephen Fox',4.2,3000)
INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
VALUES (2,'SAP BI DW','Stephen Fox',4,9000)
APPLY BATCH;
```

Outcome:

```
cqlsh:students> BEGIN BATCH
            ...   INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
            ... VALUES (1,'MS data Migration','David Sheen',3.5,720)
            ...   INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
            ... VALUES (1,'MS Data Warehouse','David Sheen',3.9,1440)
            ...   INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
            ... VALUES (2,'SAP Reporting','Stephen Fox',4.2,3000)
            ...   INSERT INTO project_details (project_id,project_name,stud_name,rating,duration)
            ... VALUES (2,'SAP BI DW','Stephen Fox',4,9000)
            ... APPLY BATCH;
```

Objective: To view all the rows of the "project_details" table.

Act:

```
SELECT *
      FROM project_details;
```

Outcome:

```
cqlsh:students> select * from project_details;

 project_id | project_name       | duration | rating | stud_name
------------+--------------------+----------+--------+-------------
          1 | MS Data Warehouse  |     1440 |    3.9 | David Sheen
          1 | MS data Migration  |      720 |    3.5 | David Sheen
          2 |          SAP BI DW |     9000 |      4 | Stephen Fox
          2 |      SAP Reporting |     3000 |    4.2 | Stephen Fox

(4 rows)
```

Objective: To view row/record from the "project_details" table wherein the project_id=1.

Act:

```
SELECT *
    FROM project_details
        WHERE project_id=1;
```

Outcome:

```
cqlsh:students> Select * from project_details where project_id=1;

 project_id | project_name       | duration | rating | stud_name
------------+--------------------+----------+--------+-------------
          1 | MS Data Warehouse  |     1440 |    3.9 | David Sheen
          1 | MS data Migration  |      720 |    3.5 | David Sheen

(2 rows)
```

Objective: To use "allow filtering" with the Select statement.

Note: When one attempts a potentially expensive query that might involve searching a range of rows, a prompt such as the one shown below appears:

> Bad Request: Cannot execute this query as it might involve data filtering and thus may have unpredictable performance. If you want to execute this query despite the performance unpredictability, use ALLOW FILTERING.

Act:

```
SELECT *
    FROM project_details
        WHERE project_name='MS Data Warehouse' ALLOW FILTERING;
```

Outcome:

```
cqlsh:students> Select * from project_details where project_name='MS Data Warehouse' allow filtering;

 project_id | project_name       | duration | rating | stud_name
------------+--------------------+----------+--------+-------------
          1 | MS Data Warehouse  |     1440 |    3.9 | David Sheen

(1 rows)
```

Objective: To sort or order the rows/records of the "project_details" column in ascending order of project_name.

Act:

```
SELECT *
    FROM project_details
        WHERE project_id IN (1,2)
            ORDER BY project_name DESC;
```

Outcome:

```
cqlsh:students> SELECT * FROM project_details WHERE project_id IN (1,2) ORDER BY project_name DESC;

 project_id | project_name       | duration | rating | stud_name
------------+--------------------+----------+--------+-------------
          2 |        SAP Reporting |     3000 |    4.2 | Stephen Fox
          2 |             SAP BI DW |     9000 |      4 | Stephen Fox
          1 | MS data Migration |      720 |    3.5 | David Sheen
          1 | MS Data Warehouse |     1440 |    3.9 | David Sheen

(4 rows)
```

Note: By default sorting or ordering is done in ascending order. The user can specify the order by using the keyword "ASC" for ascending or "DESC" for descending.

ORDER BY clause can select a single column only. This column is the second column of the compound primary key. This applies even when the compound primary key has more than two columns.

When specifying the ORDER BY clause, use only the column name and not the column alias.

7.7 COLLECTIONS

7.7.1 Set Collection

A column of type set consists of unordered unique values. However, when the column is queried, it returns the values in sorted order. For example, for text values, it sorts in alphabetical order.

PICTURE THIS...

You are required to store details about users of service "xyz". The details of the user include: User_ID, User_Name, User_Contact_Nos, User_Email_Ids. A user may have n number of Contact Nos and also may have n number of Email IDs. How do we accomplish this task in RDBMS?

We would create a table, let us say "Users", to store details such as "User_ID", "User_Name" and another table "UsersContactDetails". The relationship between "UsersContactDetails" and "Users" is many-to-one. Likewise, we would create a table "UsersEmailIDs" and establish a many-to-one relationship between "UserEmailIDs" and the "Users" table.

However, the multiple Contact Nos and multiple Email IDs problem can be solved by defining a column as a collection. The usage of collection types for columns is not only convenient but intuitive as well.

CQL makes use of the following collection types:

- Set
- List
- Map

When to use collection?
Use collection when it is required to store or denormalize a small amount of data.

What is the limit on the values of items in a collection?
The values of items in a collection are limited to 64K.

Where to use collections?
Collections can be used when you need to store the following:

1. Phone numbers of users.
2. Email ids of users.

When should one refrain from using a collection?
One should refrain from using a collection when the data has unbound growth potential such as all the messages posted by a user or all the event data as captured by a sensor. When faced with such a situation, use a table with compound primary key with data being held in clustering columns.

7.7.2 List Collection

When the order of elements matter, one should go for a list collection. For example, when you store the preferences of places to visit by a user, you would like to respect his preferences and retrieve the values in the order in which he has entered rather than in sorted order. A list also allows one to store the same value multiple times.

7.7.3 Map Collection

As the name implies, a map is used to map one thing to another. A map is a pair of typed values. It is used to store timestamp related information. Each element of the map is stored as a Cassandra column. Each element can be individually queried, modified, and deleted.

Objective: To alter the schema for the table "student_info" to add a column "hobbies".

Act:

ALTER TABLE student_info ADD hobbies set<text>;

Outcome:

```
cqlsh:students> ALTER TABLE student_info ADD hobbies set<text>;
```

Confirm the structure of the table after the change has been made:

```
cqlsh:students> describe table student_info;

CREATE TABLE student_info (
  rollno int,
  dateofjoining timestamp,
  hobbies set<text>,
  lastexampercent double,
  studname text,
  PRIMARY KEY (rollno)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};

CREATE INDEX student_info_lastexampercent_idx ON student_info (lastexampercent);

CREATE INDEX student_info_studname_idx ON student_info (studname);
```

Objective: To alter the schema of the table "student_info" to add a list column "language".

Act:

ALTER TABLE student_info ADD language list<text>;

Outcome:

```
cqlsh:students> ALTER TABLE student_info ADD language list<text>;
cqlsh:students>
```

Confirm the structure of the table after the change has been made:

```
cqlsh:students> describe table student_info;

CREATE TABLE student_info (
  rollno int,
  dateofjoining timestamp,
  hobbies set<text>,
  language list<text>,
  lastexampercent double,
  studname text,
  PRIMARY KEY (rollno)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};

CREATE INDEX student_info_lastexampercent_idx ON student_info (lastexampercent);

CREATE INDEX student_info_studname_idx ON student_info (studname);

cqlsh:students>
```

Objective: To update the table "student_info" to provide the values for "hobbies" for the student with Rollno =1.

Act:

```
UPDATE student_info
    SET hobbies = hobbies + {'Chess, Table Tennis'}
        WHERE RollNo=1;
```

Outcome:

```
cqlsh:students> UPDATE student_info
           ...   SET hobbies = hobbies + {'Chess, Table Tennis'} WHERE RollNo=1;
```

To confirm the values in the hobbies column, use the below command:

```
SELECT *
    FROM student_info
        WHERE RollNo=1;
```

```
cqlsh:students> select * from student_info where RollNo=1;

 rollno | dateofjoining                        | hobbies                 | language | lastexampercent | studname
--------+--------------------------------------+-------------------------+----------+-----------------+----------------
      1 | 2012-03-29 00:00:00India Standard Time | {'Chess, Table Tennis'} |     null |            69.6 | Michael Storm

(1 rows)
```

Likewise update a few more records:

```
cqlsh:students> UPDATE student_info
           ...   SET hobbies = hobbies + {'Chess, Badminton'} WHERE RollNo=3;
cqlsh:students>
cqlsh:students> UPDATE student_info
           ...   SET hobbies = hobbies + {'Lawn Tennis, Table Tennis, Golf'} WHERE RollNo=4;
cqlsh:students>
```

Records after the updation:

```
cqlsh:students> select * from student_info;

 rollno | dateofjoining                          | hobbies                                | language | lastexampercent | studna
--------+----------------------------------------+----------------------------------------+----------+-----------------+-------
      1 | 2012-03-29 00:00:00India Standard Time |                {'Chess, Table Tennis'} |     null |            69.6 |   Mich
      4 | 2012-05-11 00:00:00India Standard Time | {'Lawn Tennis, Table Tennis, Golf'} |     null |            73.4 |      I
      3 | 2014-04-12 00:00:00India Standard Time |                 {'Chess, Badminton'} |     null |            81.7 | David

(3 rows)
```

Objective: To update values in the list column, "language" of the table "student_info".

Act:

```
UPDATE student_info
    SET language = language + ['Hindi, English']
        WHERE RollNo=1;
```

Outcome:

```
cqlsh:students> UPDATE student_info
           ...   SET language = language + ['Hindi, English'] WHERE RollNo=1;
cqlsh:students>
```

Likewise update the remaining records.

```
cqlsh:students> UPDATE student_info
           ...   SET language = language + ['Hindi,English,French'] WHERE RollNo=3;
cqlsh:students>
cqlsh:students> UPDATE student_info
           ...   SET language = language + ['Hindi, English'] WHERE RollNo=4;
cqlsh:students>
```

To view the updates to the records, use the below statement:

```
cqlsh:students> select rollNo, studname, hobbies, language from student_info;

 rollno | studname        | hobbies                                | language
--------+-----------------+----------------------------------------+-------------------------
      1 |   Michael Storm |                {'Chess, Table Tennis'} |       ['Hindi, English']
      4 |      Ian String | {'Lawn Tennis, Table Tennis, Golf'} |       ['Hindi, English']
      3 | David Flemming  |                 {'Chess, Badminton'} | ['Hindi,English,French']

(3 rows)
```

7.7.4 More Practice on Collections (SET and LIST)

Objective: To create a table "users" with an "emails" column. The type of this column "emails" is "set".

Act:

```
CREATE TABLE users (
        user_id text PRIMARY KEY,
        first_name text,
        last_name text,
        emails set<text>
);
```

Outcome:

```
cqlsh:students> CREATE TABLE users (
            ...   user_id text PRIMARY KEY,
            ...   first_name text,
            ...   last_name text,
            ...   emails set<text>
            ... );
cqlsh:students>
```

Objective: To insert values into the "emails" column of the "users" table.

Note: Set values must be unique.

Act:

```
INSERT INTO users
        (user_id, first_name, last_name, emails)
                VALUES('AB', 'Albert', 'Baggins', {'a@baggins.com', 'baggins@gmail.com'});
```

Outcome:

```
cqlsh:students> INSERT INTO users (user_id, first_name, last_name, emails)
            ...   VALUES('AB', 'Albert', 'Baggins', {'a@baggins.com', 'baggins@gmail.com'});
cqlsh:students>
```

Objective: Add an element to a set using the UPDATE command and the addition (+) operator.

Act:

```
UPDATE users
        SET emails = emails + {'ab@friendsofmordor.org'}
                WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> UPDATE users
            ...   SET emails = emails + {'ab@friendsofmordor.org'} WHERE user_id = 'AB';
cqlsh:students>
```

Objective: To retrieve email addresses for Albert from the set.

Act:

```
SELECT user_id, emails
    FROM users
        WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> SELECT user_id, emails FROM users WHERE user_id = 'AB';

 user_id | emails
---------+-----------------------------------------------------------------
      AB | {'a@baggins.com', 'ab@friendsofmordor.org', 'baggins@gmail.com'}

(1 rows)
```

Objective: To remove an element from a set using the subtraction (–) operator.

Act:

```
UPDATE users
    SET emails = emails - {'ab@friendsofmordor.org'}
        WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> UPDATE users
           ...   SET emails = emails - {'ab@friendsofmordor.org'} WHERE user_id = 'AB';
```

To view the records from the "users" table:

```
cqlsh:students> select * from users;

 user_id | emails                                   | first_name | last_name
---------+------------------------------------------+------------+-----------
      AB | {'a@baggins.com', 'baggins@gmail.com'} |     Albert |   Baggins

(1 rows)
```

Objective: To remove all elements from a set by using the UPDATE or DELETE statement.

Act:

```
UPDATE users
    SET emails = {}
        WHERE user_id = 'AB';
cqlsh:students> UPDATE users SET emails = {} WHERE user_id = 'AB';
```

Outcome: The above command removes the emails column. Here is the confirmation:

```
cqlsh:students> select * from users;

 user_id | emails | first_name | last_name
---------+--------+------------+-----------
      AB |   null |     Albert |   Baggins

(1 rows)
```

OR

DELETE emails
FROM users
WHERE user_id = 'AB';

```
cqlsh:students> DELETE emails FROM users WHERE user_id = 'AB';
```

The above command removes the emails column. Here is the confirmation:

```
cqlsh:students> select * from users;

 user_id | emails | first_name | last_name
---------+--------+------------+-----------
      AB |   null |     Albert |   Baggins

(1 rows)
```

Objective: To alter the "users" table to add a column, "top_places" of type list.

Act:

ALTER TABLE users ADD top_places list<text>;

```
cqlsh:students> ALTER TABLE users ADD top_places list<text>;
```

Outcome:

The above command alters the structure of the table, "users". Here is the confirmation.

```
cqlsh:students> describe table users;

CREATE TABLE users (
  user_id text,
  emails set<text>,
  first_name text,
  last_name text,
  top_places list<text>,
  PRIMARY KEY (user_id)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

Objective: To update the list column "top_places" in the "users" table for user_id = 'AB'.

Act:

```
UPDATE users
        SET top_places = [ 'Lonavla', 'Khandala' ]
                WHERE user_id = 'AB';
```

```
cqlsh:students> UPDATE users
           ...   SET top_places = [ 'Lonavla', 'Khandala' ] WHERE user_id = 'AB';
cqlsh:students>
```

Outcome:

```
cqlsh:students> select * from users where user_id = 'AB';

 user_id | emails | first_name | last_name | top_places
---------+--------+------------+-----------+--------------------------
      AB |   null |     Albert |   Baggins | ['Lonavla', 'Khandala']

(1 rows)
```

Objective: Prepend an element to the list by enclosing it in square brackets and using the addition (+) operator.

Act:

```
UPDATE users
        SET top_places = [ 'Mahabaleshwar' ] + top_places
                WHERE user_id = 'AB';
```

```
cqlsh:students> UPDATE users
           ...   SET top_places = [ 'Mahabaleshwar' ] + top_places WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> select * from users;

 user_id | emails | first_name | last_name | top_places
---------+--------+------------+-----------+-----------------------------------------
      AB |   null |     Albert |   Baggins | ['Mahabaleshwar', 'Lonavla', 'Khandala']

(1 rows)
```

Objective: To append an element to the list by switching the order of the new element data and the list name in the update command.

Act:

```
UPDATE users
        SET top_places = top_places + [ 'Tapola' ]
                WHERE user_id = 'AB';
```

```
cqlsh:students> UPDATE users
           ...   SET top_places = top_places + [ 'Tapola' ] WHERE user_id = 'AB';
cqlsh:students>
```

Outcome:

```
cqlsh:students> select * from users;

 user_id | emails | first_name | last_name | top_places
---------+--------+------------+-----------+-----------------------------------------------------
      AB |   null |     Albert |   Baggins | ['Mahabaleshwar', 'Lonavla', 'Khandala', 'Tapola']

(1 rows)
```

Objective: To query the database for a list of top places.

Act:

```
SELECT user_id, top_places
        FROM users
                WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> SELECT user_id, top_places FROM users WHERE user_id = 'AB';

 user_id | top_places
---------+-----------------------------------------------------
      AB | ['Mahabaleshwar', 'Lonavla', 'Khandala', 'Tapola']

(1 rows)
```

Objective: To remove an element from a list using the DELETE command and the list index position in square brackets.

The record as it exists prior to deletion is

```
cqlsh:students> SELECT user_id, top_places FROM users WHERE user_id = 'AB';

 user_id | top_places
---------+-----------------------------------------------------
      AB | ['Mahabaleshwar', 'Lonavla', 'Khandala', 'Tapola']

(1 rows)
```

Act:

```
DELETE top_places[3]
        FROM users
                WHERE user_id = 'AB';
```

```
cqlsh:students> DELETE top_places[3] FROM users WHERE user_id = 'AB';
```

Outcome: The status after deletion is

```
cqlsh:students> select * from users;

 user_id | emails | first_name | last_name | top_places
---------+--------+------------+-----------+-------------------------------------------
      AB |   null |     Albert |   Baggins | ['Mahabaleshwar', 'Lonavla', 'Khandala']

(1 rows)
```

7.7.5 Using Map: Key, Value Pair

Objective: To alter the "users" table to add a map column "todo".

Act:

```
ALTER TABLE users
        ADD todo map<timestamp, text>;
```

```
cqlsh:students> ALTER TABLE users ADD todo map<timestamp, text>;
```

Outcome:

```
cqlsh:students> describe table users;

CREATE TABLE users (
  user_id text,
  emails set<text>,
  first_name text,
  last_name text,
  todo map<timestamp, text>,
  top_places list<text>,
  PRIMARY KEY (user_id)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};

cqlsh:students>
```

Objective: To update the record for user (user_id = 'AB') in the "users" table.

Act: The record from user_id = 'AB' as it exists in the "users" table is

```
cqlsh:students> select * from users where user_id='AB';

 user_id | emails | first_name | last_name | todo | top_places
---------+--------+------------+-----------+------+-------------------------------------------
      AB |   null |     Albert |   Baggins | null | ['Mahabaleshwar', 'Lonavla', 'Khandala']

(1 rows)

cqlsh:students> UPDATE users
            ...   SET todo =
            ...   {'2014-9-24':'Cassandra Session',
            ...    '2014-10-2 12:00' :'MongoDB Session'}
            ...   WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> select user_id, todo from users where user_id='AB';

r_id | todo
-----+-----------------------------------------------------------------------------------------------------------------------
  AB | {'2014-09-24 00:00:00India Standard Time': 'Cassandra Session', '2014-10-02 12:00:00India Standard Time': 'MongoDB Se
```

Objective: To delete an element from the map using the DELETE command and enclosing the timestamp of the element in square brackets.

Act:

DELETE todo['2014-9-24']
FROM users
WHERE user_id = 'AB';

```
cqlsh:students> DELETE todo['2014-9-24'] FROM users WHERE user_id = 'AB';
```

Outcome:

```
cqlsh:students> select user_id, todo from users where user_id='AB';
 user_id | todo
---------+-----------------------------------------------------------------
      AB | {'2014-10-02 12:00:00India Standard Time': 'MongoDB Session'}
(1 rows)
```

7.8 USING A COUNTER

A counter is a special column that is changed in increments. For example, we may need a counter column to count the number of times a particular book is issued from the library by the student.

Step 1:

CREATE TABLE library_book (
counter_value counter,
book_name varchar,
stud_name varchar,
PRIMARY KEY (book_name, stud_name)
);

```
cqlsh:students> CREATE TABLE library_book
           ...   (     counter_value counter,
           ...         book_name varchar,
           ...         stud_name varchar,
           ...   PRIMARY KEY (book_name, stud_name)
           ... );
```

Step 2: Load data into the counter column.

UPDATE library_book
SET counter_value = counter_value + 1
WHERE book_name='Fundamentals of Business Analytics' AND stud_name='jeet';

```
cqlsh:students> UPDATE library_book
           ...  SET counter_value = counter_value + 1
           ... WHERE book_name='Fundamentals of Business Analytics' AND stud_name='jeet';
```

Step 3: Take a look at the counter value.

SELECT *

FROM library_book;

Output is:

```
cqlsh:students> select * from library_book;

 book_name                          | stud_name | counter_value
------------------------------------+-----------+---------------
 Fundamentals of Business Analytics |      jeet |             1

(1 rows)
```

Step 4: Let us increase the value of the counter.

UPDATE library_book

SET counter_value = counter_value + 1

WHERE book_name='Fundamentals of Business Analytics' AND stud_name='shaan';

```
cqlsh:students> UPDATE library_book
           ...  SET counter_value = counter_value + 1
           ... WHERE book_name='Fundamentals of Business Analytics' AND stud_name='shaan';
```

Step 5: Again, take a look at the counter value.

```
cqlsh:students> select * from library_book;

 book_name                          | stud_name | counter_value
------------------------------------+-----------+---------------
 Fundamentals of Business Analytics |      jeet |             1
 Fundamentals of Business Analytics |     shaan |             1

(2 rows)
```

Step 6: Update another record for Stud_name "Jeet".

UPDATE library_book

SET counter_value = counter_value + 1

WHERE book_name='Fundamentals of Business Analytics' AND stud_name='jeet';

```
cqlsh:students> UPDATE library_book
           ...  SET counter_value = counter_value + 1
           ... WHERE book_name='Fundamentals of Business Analytics' AND stud_name='jeet';
```

Step 7: Let us take a look at the counter value, one last time.

```
cqlsh:students> select * from library_book;

 book_name                          | stud_name | counter_value
------------------------------------+-----------+---------------
 Fundamentals of Business Analytics |      jeet |             2
 Fundamentals of Business Analytics |     shaan |             1

(2 rows)
```

7.9 TIME TO LIVE (TTL)

Data in a column, other than a counter column, can have an optional expiration period called TTL (time to live). The client request may specify a TTL value for the data. The TTL is specified in seconds.

CREATE TABLE userlogin(

userid int primary key, password text

);

```
cqlsh:students> CREATE TABLE  userlogin(
            ... userid int primary key, password text
            ... );
```

INSERT INTO userlogin (userid, password)

VALUES (1,'infy') USING TTL 30;

```
cqlsh:students> INSERT INTO userlogin (userid, password)
            ... VALUES (1,'infy') USING TTL 30;
```

SELECT TTL (password)

FROM userlogin

WHERE userid=1;

```
cqlsh:students> SELECT TTL (password)
            ... FROM userlogin
            ... WHERE userid=1;

 ttl(password)
---------------
             18

(1 rows)
```

7.10 ALTER COMMANDS

Let us look at a few Alter commands to bring about changes to the structure of the table/column family.

1. Create a table "sample" with columns "sample_id" and "sample_name".

   CREATE TABLE sample(
   sample_id text,
   sample_name text,
   PRIMARY KEY (sample_id)
   );

```
cqlsh:students> Create table sample (sample_id text, sample_name text, primary key (sample_id));
cqlsh:students>
```

2. Insert a record into the table "sample".

   INSERT INTO sample(
   sample_id, sample_name)
   VALUES ('S101', 'Big Data');

```
cqlsh:students> Insert into sample(sample_id, sample_name) values( 'S101', 'Big Data' );
```

3. View the records of the table "sample".

   SELECT *
   FROM sample;

```
cqlsh:students> select * from sample;

 sample_id | sample_name
-----------+-------------
      S101 |    Big Data

(1 rows)
```

7.10.1 Alter Table to Change the Data Type of a Column

1. Alter the schema of the table "sample". Change the data type of the column "sample_id" to integer from text.

ALTER TABLE sample
ALTER sample_id TYPE int;

```
cqlsh:students> alter table sample alter sample_id type int;
```

2. After the data type of the column "sample_id" is changed from text to integer, try inserting a record as follows and observe the error message:

INSERT INTO sample(sample_id, sample_name)
VALUES('S102', 'Big Data');

```
cqlsh:students> Insert into sample(sample_id, sample_name) values( 'S102', 'Big Data' );
Bad Request: Invalid STRING constant (S102) for sample_id of type int
cqlsh:students>
```

3. Try inserting a record as given below into the table "sample".

INSERT INTO sample(sample_id, sample_name)
VALUES(102, 'Big Data');

```
cqlsh:students> Insert into sample(sample_id, sample_name) values( 102, 'Big Data' );
cqlsh:students> select * from sample;

 sample_id  | sample_name
------------+-------------
 1395732529 |    Big Data
        102 |    Big Data

(2 rows)
```

4. Alter the data type of the "sample_id" column to varchar from integer.

ALTER TABLE sample
ALTER sample_id TYPE varchar;

```
cqlsh:students> alter table sample alter sample_id type varchar;
```

5. Check the records after the data type of "sample_id" has been changed to varchar from integer.

```
cqlsh:students> select * from sample;

 sample_id      | sample_name
----------------+-------------
           S101 |    Big Data
 \x00\x00\x00f  |    Big Data

(2 rows)
```

7.10.2 Alter Table to Delete a Column

1. Drop the column "sample_id" from the table "sample".

ALTER TABLE sample
DROP sample_id;

```
cqlsh:students> alter table sample drop sample_id;
Bad Request: Cannot drop PRIMARY KEY part sample_id
```

Note: The request to drop the "sample_id" column from the table "sample" does not succeed as it is the primary key column.

2. Drop the column "sample_name" from the table "sample".

   ALTER TABLE sample
   DROP sample_name;

```
cqlsh:students> alter table sample drop sample_name;
```

   Note: the above request to drop the column "sample_name" from table "sample" succeeds.

7.10.3 Drop a Table

1. Drop the column family/table "sample".

   DROP columnfamily sample;

```
cqlsh:students> drop columnfamily sample;
```

   The above request succeeds. The table/column family no longer exists in the keyspace.

2. Confirm the non-existence of the table "sample" in the keyspace by giving the following command:

```
cqlsh:students> describe table sample;

Column family 'sample' not found
```

7.10.4 Drop a Database

1. Drop the keyspace "students".

   DROP keyspace students;

```
cqlsh:students> drop keyspace students;
```

2. Confirm the non-existence of the keyspace "students" by issuing the following command:

```
cqlsh:students> describe keyspace students;

Keyspace 'students' not found.
```

7.11 IMPORT AND EXPORT

7.11.1 Export to CSV

Objective: Export the contents of the table/column family "elearninglists" present in the "students" database to a CSV file (d:\elearninglists.csv).

Act:

Step 1: Check the records of the table "elearninglists" present in the "students" database.

SELECT *
FROM elearninglists;

```
cqlsh:students> select * from elearninglists;

 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+----------------
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra
 101 |            2 |      1002 |       Seema |   NoSQL MongoDB
 101 |            3 |      1003 |       Seema |    Hadoop Sqoop
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume

(4 rows)
```

Step 2: Execute the below command at the cqlsh prompt:

COPY elearninglists (id, course_order, course_id, courseowner, title) TO 'd:\elearninglists.csv';

```
cqlsh:students> copy elearninglists (id, course_order, course_id, courseowner, title) to 'd:\elearninglists.csv';
4 rows exported in 0.000 seconds.
cqlsh:students>
```

Step 3: Check the existence of the "elearninglists.csv" file in "D:\". Given below is the content of the "d:\elearninglists.csv" file.

	A	B	C	D	E
1	101	1	1001	Subhashini	NoSQL Cassandra
2	101	2	1002	Seema	NoSQL MongoDB
3	101	3	1003	Seema	Hadoop Sqoop
4	101	4	1004	Subhashini	Hadoop Flume

7.11.2 Import from CSV

Objective: To import data from "D:\elearninglists.csv" into the table "elearninglists" present in the "students" database.

Step 1: Check for the table "elearninglists" in the "students" database. If the table is already present, truncate the table. This will remove all records from the table but retain the structure of the table. In our case, the table "elearninglists" is already present in the "students" database. Let us take a look at the records of the "elearninglists" before we run the truncate command on it.

```
cqlsh:students> select * from elearninglists;

 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+-----------------
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra
 101 |            2 |      1002 |       Seema |   NoSQL MongoDB
 101 |            3 |      1003 |       Seema |    Hadoop Sqoop
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume

(4 rows)
```

Truncate the table using the below command:

TRUNCATE elearninglists;

```
cqlsh:students> Truncate  elearninglists;
cqlsh:students>
```

Note: No record is present in the table "elearninglists". The structure/schema is however preserved. We confirm it by executing the below command at the cqlsh prompt.

```
cqlsh:students> select * from elearninglists;

(0 rows)

cqlsh:students>
```

Step 2: Check for the content of the "D:\elearninglists.csv" file.

	A	B	C	D	E
1	101	1	1001	Subhashini	NoSQL Cassandra
2	101	2	1002	Seema	NoSQL MongoDB
3	101	3	1003	Seema	Hadoop Sqoop
4	101	4	1004	Subhashini	Hadoop Flume

Note: The content in the CSV agrees with the structure of the table "elearninglists" in the "students" database. The structure should be such that the content from the CSV can be housed within it without any issues.

Step 3: Execute the below command to import data from "d:\elearninglists.csv" into the table "elearninglists" in the database "students".

COPY elearninglists (id, course_order, course_id, courseowner, title) FROM 'd:\elearninglists.csv';

```
cqlsh:students> copy elearninglists (id, course_order, course_id, courseowner, title) from 'd:\elearninglists.csv';
4 rows imported in 0.031 seconds.
cqlsh:students>
```

Step 4: Confirm that records have been imported into the table.

SELECT *
FROM elearninglists;

```
cqlsh:students> select * from elearninglists;

 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+-----------------
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra
 101 |            2 |      1002 |       Seema |   NoSQL MongoDB
 101 |            3 |      1003 |       Seema |    Hadoop Sqoop
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume

(4 rows)

cqlsh:students>
```

7.11.3 Import from STDIN

Objective: To import data into an existing table "persons" present in the "students" database. The data is to be provided by the user using the standard input device.

Step 1: Ensure that the table "persons" exists in the database "students".

DESCRIBE TABLE persons;

```
cqlsh:students> describe table persons;

CREATE TABLE persons (
  id int,
  fname text,
  lname text,
  PRIMARY KEY (id)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  index_interval=128 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};

cqlsh:students>
```

Step 2:

COPY persons (id, fname, lname) FROM STDIN;

```
cqlsh:students> COPY persons (id, fname, lname) FROM STDIN;
[Use \. on a line by itself to end input]
[copy] 1,"Samuel","Jones"
[copy] 2,"Virat","Kumar"
[copy] 3,"Andrew","Simon"
[copy] 4,"Raul","A Simpson"
[copy] \.

4 rows imported in 1 minute and 24.336 seconds.
cqlsh:students>
```

Step 3: Confirm that the records from the standard input device are loaded into the "persons" table existing in the "students" database.

SELECT *
FROM persons;

```
cqlsh:students> select * from persons;

 id | fname  | lname
----+--------+-----------
  1 | Samuel |     Jones
  2 |  Virat |     Kumar
  4 |   Raul | A Simpson
  3 | Andrew |     Simon

(4 rows)

cqlsh:students>
```

7.11.4 Export to STDOUT

Objective: Export the contents of the table/column family "elearninglists" present in the "students" database to the standard output device (STDOUT).

Act:

Step 1: Check the records of the table "elearninglists" present in the "students" database.

SELECT *
FROM elearninglists;

```
cqlsh:students> select * from elearninglists;

 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+----------------
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra
 101 |            2 |      1002 |       Seema |   NoSQL MongoDB
 101 |            3 |      1003 |       Seema |    Hadoop Sqoop
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume

(4 rows)
```

Step 2: Execute the below command at the cqlsh prompt.

COPY elearninglists (id, course_order, course_id, courseowner, title) TO STDOUT;

```
cqlsh:students> copy elearninglists (id, course_order, course_id, courseowner, title) to STDOUT;
101,1,1001,Subhashini,NoSQL Cassandra
101,2,1002,Seema,NoSQL MongoDB
101,3,1003,Seema,Hadoop Sqoop
101,4,1004,Subhashini,Hadoop Flume
4 rows exported in 0.031 seconds.
cqlsh:students>
```

7.12 QUERYING SYSTEM TABLES

There are quite a few system tables such as schema_keyspaces, schema_columnfamilies, schema_columns, local, peers, etc. Let us look at what each of these system tables store in them.

SELECT *

FROM system.schema_keyspaces;

```
cqlsh:system> describe table system.schema_keyspaces;

CREATE TABLE schema_keyspaces (
  keyspace_name text,
  durable_writes boolean,
  strategy_class text,
  strategy_options text,
  PRIMARY KEY (keyspace_name)
) WITH COMPACT STORAGE AND
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='keyspace definitions' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=8640 AND
  index_interval=128 AND
  read_repair_chance=0.000000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

SELECT *

FROM system.schema_columnfamilies;

```
CREATE TABLE schema_columnfamilies (
  keyspace_name text,
  columnfamily_name text,
  bloom_filter_fp_chance double,
  caching text,
  column_aliases text,
  comment text,
  compaction_strategy_class text,
  compaction_strategy_options text,
  comparator text,
  compression_parameters text,
  default_time_to_live int,
  default_validator text,
```

```
  dropped_columns map<text, bigint>,
  gc_grace_seconds int,
  index_interval int,
  key_aliases text,
  key_validator text,
  local_read_repair_chance double,
  max_compaction_threshold int,
  memtable_flush_period_in_ms int,
  min_compaction_threshold int,
  populate_io_cache_on_flush boolean,
  read_repair_chance double,
  replicate_on_write boolean,
  speculative_retry text,
  subcomparator text,
  type text,
  value_alias text,
  PRIMARY KEY (keyspace_name, columnfamily_name)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='ColumnFamily definitions' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=8640 AND
  index_interval=128 AND
  read_repair_chance=0.000000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
```

```
compaction={'class': 'SizeTieredCompactionStrategy'} AND
compression={'sstable_compression': 'LZ4Compressor'};
```

SELECT *

FROM system.schema_columns;

```
cqlsh:system> describe table system.schema_columns;

CREATE TABLE schema_columns (
  keyspace_name text,
  columnfamily_name text,
  column_name text,
  component_index int,
  index_name text,
  index_options text,
  index_type text,
  type text,
  validator text,
  PRIMARY KEY (keyspace_name, columnfamily_name, column_name)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='ColumnFamily column attributes' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=8640 AND
  index_interval=128 AND
  read_repair_chance=0.000000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

SELECT *

FROM system.local;

```
CREATE TABLE local (
  key text,
  bootstrapped text,
  cluster_name text,
  cql_version text,
  data_center text,
  gossip_generation int,
  host_id uuid,
  native_protocol_version text,
  partitioner text,
  rack text,
  release_version text,
  schema_version uuid,
  thrift_version text,
  tokens set<text>,
  truncated_at map<uuid, blob>,
  PRIMARY KEY (key)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='information about the local node' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=0 AND
  index_interval=128 AND
  read_repair_chance=0.000000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

SELECT *

FROM system.peers;

```
CREATE TABLE peers (
  peer inet,
  data_center text,
  host_id uuid,
  preferred_ip inet,
  rack text,
  release_version text,
  rpc_address inet,
  schema_version uuid,
  tokens set<text>,
  PRIMARY KEY (peer)
) WITH
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='known peers in the cluster' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=0 AND
  index_interval=128 AND
  read_repair_chance=0.000000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  default_time_to_live=0 AND
  speculative_retry='NONE' AND
  memtable_flush_period_in_ms=0 AND
  compaction={'class': 'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression': 'LZ4Compressor'};
```

7.13 PRACTICE EXAMPLES

Objective: To create table "elearninglist" with columns: id, course_order, course_id, title, courseowner.

Act:

```
CREATE TABLE elearninglists (
    id int,
    course_order int,
    course_id int,
    title text,
    courseowner text,
    PRIMARY KEY (id, course_order )
    );
```

Here, id ==> Partition Key, course_order ==> Clustering Column. The combination of the id and course_order in the elearninglists table uniquely identifies a row in the elearninglists table. You can have more than one row with the same id as long as the rows contain different course_order values.

Outcome:

```
cqlsh:students> CREATE TABLE elearninglists (
            ...   id int,
            ...   course_order int,
            ...   course_id int,
            ...   title text,
            ...   courseowner text,
            ...   PRIMARY KEY  (id, course_order ) );
cqlsh:students>
```

Objective: To insert rows into the table "elearninglists".

Act:

```
INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
VALUES (101, 1, 1001,'NoSQL Cassandra','Subhashini');

INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
VALUES (101, 2, 1002,'NoSQL MongoDB','Seema');

INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
VALUES (101, 3, 1003,'Hadoop Sqoop','Seema');

INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
VALUES (101, 4, 1004,'Hadoop Flume', 'Subhashini');
```

Outcome:

```
cqlsh:students> INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
            ...   VALUES (101,1,1001,'NoSQL Cassandra','Subhashini');
cqlsh:students>
cqlsh:students> INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
            ...   VALUES (101,2,1002,'NoSQL MongoDB','Seema');
cqlsh:students>
cqlsh:students> INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
            ...   VALUES (101,3,1003,'Hadoop Sqoop','Seema',);
Bad Request: line 2:44 no viable alternative at input ')'
cqlsh:students>
cqlsh:students> INSERT INTO elearninglists (id, course_order, course_id, title, courseowner)
            ...   VALUES (101, 4,1004,'Hadoop Flume', 'Subhashini');
cqlsh:students>
```

Objective: To query the table "elearninglists".

Act:

SELECT *

FROM elearninglists;

Outcome:

```
cqlsh:students> SELECT * FROM elearninglists;

 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+-----------------
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra
 101 |            2 |      1002 |       Seema |   NoSQL MongoDB
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume

(3 rows)
```

Objective: To query the "elearninglist" table on "courseowner" as a filter.

Act:

SELECT *

FROM elearninglists

WHERE courseowner = 'Seema';

Outcome: The query returns an error stating "No indexed columns present in by-columns clause with Equal operator".

```
cqlsh:students> SELECT * FROM elearninglists WHERE courseowner = 'Seema';
Bad Request: No indexed columns present in by-columns clause with Equal operator
cqlsh:students>
```

Solution: Create an index on courseowner column.

CREATE INDEX ON

Elearninglists(courseowner);

```
cqlsh:students> CREATE INDEX ON elearninglists(courseowner);
cqlsh:students>
```

Executing the same query now shows the record:

```
cqlsh:students> SELECT * FROM elearninglists WHERE courseowner = 'Seema';
 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+---------------
 101 |            2 |      1002 |       Seema | NoSQL MongoDB

(1 rows)

cqlsh:students>
```

Objective: To order all the rows of elearninglists with id = 101 in descending order of "course_order". The maximum number of records to retrieve is 50.

Act:

```
SELECT *
        FROM elearninglists
                WHERE id = 101
                        ORDER BY course_order DESC LIMIT 50;
```

Outcome:

```
cqlsh:students> SELECT * FROM elearninglists WHERE id = 101 ORDER BY course_order DESC LIMIT 50;
 id  | course_order | course_id | courseowner | title
-----+--------------+-----------+-------------+-----------------
 101 |            4 |      1004 |  Subhashini |    Hadoop Flume
 101 |            2 |      1002 |       Seema |  NoSQL MongoDB
 101 |            1 |      1001 |  Subhashini | NoSQL Cassandra

(3 rows)
```

REMIND ME

- Apache Cassandra was born at Facebook. After Facebook open sourced the code in 2008, Cassandra became an Apache Incubator project in 2009 and subsequently became a top-level Apache project in 2010.
- Cassandra does NOT compromise on availability. Since it does not have a master-slave architecture, there is no question of single point of failure. This proves beneficial for business critical applications that need to be up and running always and cannot afford to go down ever.
- A replication strategy is employed to determine which nodes to place the data on. Two replication strategies are available:
  - SimpleStrategy.
  - NetworkToplogyStrategy.
- One of the features of Cassandra that has made it immensely popular is its ability to utilize tunable consistency. The database systems can go for either strong consistency or eventual consistency.
- Read consistency means how many replicas must respond before sending out the result to the client application.
- Write consistency means on how many replicas write must succeed before sending out an acknowledgement to the client application.

POINT ME (BOOK)

- ***Cassandra:*** The Definitive Guide By Eben Hewitt, Publisher: O'Reilly Media.

CONNECT ME (INTERNET RESOURCES)

- http://www.datastax.com/documentation/cassandra/2.0/cassandra/gettingStartedCassandraIntro.html
- http://www.datastax.com/documentation/cql/3.1/pdf/cql31.pdf
- http://www.datastax.com/documentation/cassandra/2.0/cassandra/dml/dml_config_consistency_c.html
- http://www.datastax.com/docs/1.0/cluster_architecture/about_client_requests
- http://www.datastax.com/docs/datastax_enterprise3.1/solutions/about_pig

TEST ME

A. Unsolved Exercises

1. What is Cassandra?
2. Comment on Cassandra writes.
3. What is your understanding of tunable consistency?
4. What are collections in CQLSH? Where are they used?
5. Explain hinted handoffs.
6. What is Cassandra – cli?
7. Explain Cassandra's data model.
8. Explain the replication strategy in Cassandra.
9. Cassandra adheres to the Availability and Partition tolerant traits as stated by the CAP theorem. Explain.

ASSIGNMENTS FOR HANDS-ON PRACTICE

ASSIGNMENT 1: COLLECTIONS

Objective: To learn about the various collection types: Set, List and Map.
Problem Description: Design a table/column family to support the following requirements.

- Store the basic information about students such as Student Roll No, Student Name, Student Date of Birth, and Student Address.
- Store the subject preferences of each student. There should be a minimum of two subject preferences and a maximum of four. The order of preferences as given by the student should be preserved.
- Store the hobbies of each student. There should be a minimum of two hobbies and a maximum of four. The hobbies as given by the student should be arranged in alphabetical order.

ASSIGNMENT 2: TIME TO LIVE

Objective: To learn about the TTL type (Time To Live).

Problem Description: Design a table/column family to support the following requirements.

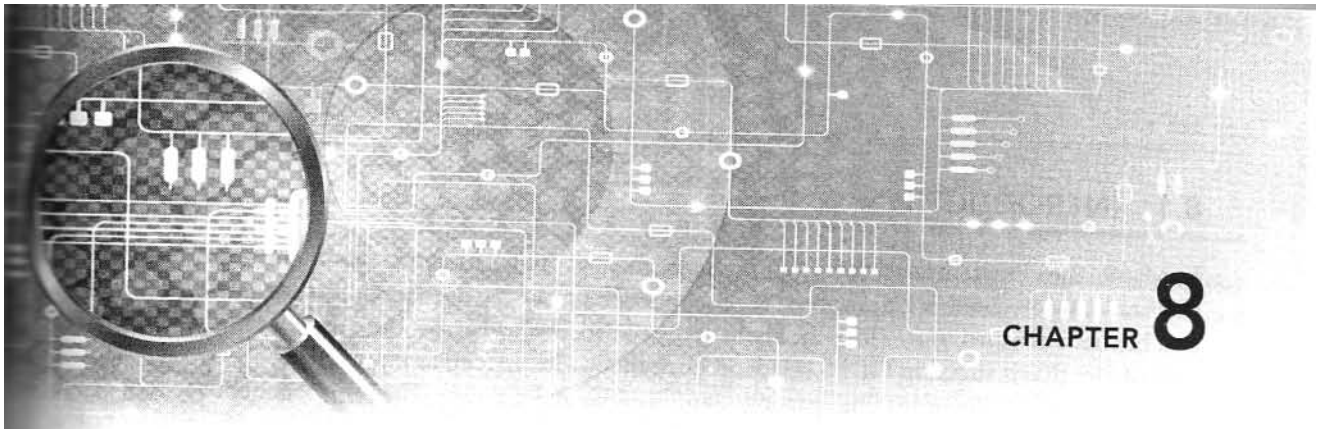
Store the login details of the user such as UserID and Password. The information stored should expire in a day's time.

ASSIGNMENT 3: IMPORT FROM CSV

Objective: To learn about the import from CSV to Cassandra table/column family.

Problem Description: Read a public dataset from the site www.kdnuggets.com. If not already in CSV format, first convert to CSV format and then import into a Cassandra table/column family by the name "PublicDataSet" in the "Sample" database.

Confirm the presence of data in the table "PublicDataSet" in the "Sample" database.



CHAPTER 8

Introduction to MAPREDUCE Programming

BRIEF CONTENTS

- What's in Store?
- Introduction
- Mapper
  - RecordReader
  - Map
  - Combiner
  - Partitioner
- Reducer
  - Shuffle
  - Sort
  - Reduce
  - Output Format
- Combiner
- Partitioner
- Searching
- Sorting
- Compression

"The alchemists in their search for gold discovered many other things of greater value."

– Arthur Schopenhauer, German Philosopher

WHAT'S IN STORE?

We assume that you are familiar with the basic concepts of HDFS and MapReduce Programming discussed in Chapters 4 and 5. The focus of this chapter will be to build on this knowledge to understand optimization techniques of MapReduce Programming such as combiner, partitioner, and compression. We will also discuss how to write MapReduce Programming for sorting and searching.

We suggest you refer to some of the learning resources provided at the end of this chapter for better learning and comprehension.

8.1 INTRODUCTION

In MapReduce Programming, Jobs (Applications) are split into a set of map tasks and reduce tasks. Then these tasks are executed in a distributed fashion on Hadoop cluster. Each task processes small subset of data that has been assigned to it. This way, Hadoop distributes the load across the cluster. MapReduce job takes a set of files that is stored in HDFS (Hadoop Distributed File System) as input.

Map task takes care of loading, parsing, transforming, and filtering. The responsibility of reduce task is grouping and aggregating data that is produced by map tasks to generate final output. Each map task is broken into the following phases:

1. RecordReader.
2. Mapper.
3. Combiner.
4. Partitioner.

The output produced by map task is known as intermediate keys and values. These intermediate keys and values are sent to reducer. The reduce tasks are broken into the following phases:

1. Shuffle.
2. Sort.
3. Reducer.
4. Output Format.

Hadoop assigns map tasks to the DataNode where the actual data to be processed resides. This way, Hadoop ensures data locality. Data locality means that data is not moved over network; only computational code is moved to process data which saves network bandwidth.

8.2 MAPPER

A mapper maps the input key–value pairs into a set of intermediate key–value pairs. Maps are individual tasks that have the responsibility of transforming input records into intermediate key–value pairs.

1. **RecordReader:** RecordReader converts a byte-oriented view of the input (as generated by the InputSplit) into a record-oriented view and presents it to the Mapper tasks. It presents the tasks with keys and values. Generally the key is the positional information and value is a chunk of data that constitutes the record.
2. **Map:** Map function works on the key–value pair produced by RecordReader and generates zero or more intermediate key–value pairs. The MapReduce decides the key–value pair based on the context.
3. **Combiner:** It is an optional function but provides high performance in terms of network bandwidth and disk space. It takes intermediate key–value pair provided by mapper and applies user-specific aggregate function to only that mapper. It is also known as local reducer.
4. **Partitioner:** The partitioner takes the intermediate key–value pairs produced by the mapper, splits them into shard, and sends the shard to the particular reducer as per the user-specific code. Usually, the key with same values goes to the same reducer. The partitioned data of each map task is written to the local disk of that machine and pulled by the respective reducer.

8.3 REDUCER

The primary chore of the Reducer is to reduce a set of intermediate values (the ones that share a common key) to a smaller set of values. The Reducer has three primary phases: Shuffle and Sort, Reduce, and Output Format.

1. **Shuffle and Sort:** This phase takes the output of all the partitioners and downloads them into the local machine where the reducer is running. Then these individual data pipes are sorted by keys which produce larger data list. The main purpose of this sort is grouping similar words so that their values can be easily iterated over by the reduce task.
2. **Reduce:** The reducer takes the grouped data produced by the shuffle and sort phase, applies reduce function, and processes one group at a time. The reduce function iterates all the values associated with that key. Reducer function provides various operations such as aggregation, filtering, and combining data. Once it is done, the output (zero or more key–value pairs) of reducer is sent to the output format.
3. **Output Format:** The output format separates key–value pair with tab (default) and writes it out to a file using record writer.

Figure 8.1 describes the chores of Mapper, Combiner, Partitioner, and Reducer for the word count problem. The Word Count problem has been discussed under "Combiner" and "Partitioner".

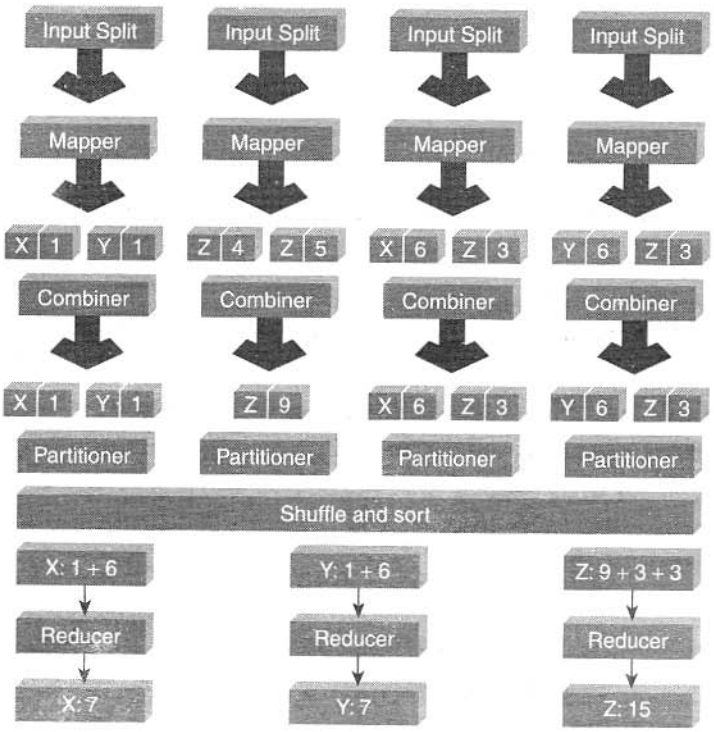


Figure 8.1 The chores of Mapper, Combiner, Partitioner, and Reducer.

8.4 COMBINER

It is an optimization technique for MapReduce Job. Generally, the reducer class is set to be the combiner class. The difference between combiner class and reducer class is as follows:

1. Output generated by combiner is intermediate data and it is passed to the reducer.
2. Output of the reducer is passed to the output file on disk.

The sections have been designed as follows:

Objective: What is it that we are trying to achieve here?
Input Data: What is the input that has been given to us to act upon?
Act: The actual statement/command to accomplish the task at hand.
Output: The result/output as a consequence of executing the statement.

Objective: Write a MapReduce program to count the occurrence of similar words in a file. Use combiner for optimization.

Note: Refer Chapter 5 – Hadoop for Mapper Class and Reduce Class and Driver Program.

Input Data:

```
Welcome to Hadoop Session
Introduction to Hadoop
Introducing Hive
Hive Session
Pig Session
```

Act: In the driver program, set the combiner class as shown below.

```
job.setCombinerClass(WordCounterRed.class);

// Input and Output Path
FileInputFormat.addInputPath(job, new Path("/mapreducedemos/lines.txt"));
FileOutputFormat.setOutputPath(job, new Path("/mapreducedemos/output/wordcount/"));
```

hadoop jar <<jar name>> <<driver class>> <<input path>> <<output path>>

Here driver class name, input path, and output path are optional arguments.

Output:

```
[root@volgalnx010 mapreducedemos]# hadoop jar wordcount.jar
```

Contents of directory /mapreducedemos

Goto : /mapreducedemos go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
lines.txt	file	91 B	3	128 MB	2015-03-01 21:05	rw-r--r--	root	supergroup
output	dir				2015-03-01 23:21	rwxr-xr-x	root	supergroup

Go back to DFS home

Local logs

Contents of directory /mapreducedemos/output

Goto : /mapreducedemos/output go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
wordcount	dir				2015-03-01 23:21	rwxr-xr-x	root	supergroup

Go back to DFS home

Local logs

The reducer output will be stored in part-r-00000 file by default.

Contents of directory /mapreducedemos/output/wordcount

Goto : /mapreducedemos/output/wo go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 B	3	128 MB	2015-03-01 23:21	rw-r--r--	root	supergroup
part-r-00000	file	76 B	3	128 MB	2015-03-01 23:21	rw-r--r--	root	supergroup

Go back to DFS home

Local logs

File: /mapreducedemos/output/wordcount/part-r-00000

Goto : /mapreducedemos/output/wo go

Go back to dir listing
Advanced view/download options

```
Hadoop  2
Hive    2
Introducing     1
Introduction    1
Pig     1
Session 3
Welcome 1
to      2
```

8.5 PARTITIONER

The partitioning phase happens after map phase and before reduce phase. Usually the number of partitions are equal to the number of reducers. The default partitioner is hash partitioner.

Objective: Write a MapReduce program to count the occurrence of similar words in a file. Use partitioner to partition key based on alphabets.
Note: Refer Chapter 5 – Hadoop for Mapper Class and Reduce Class and Driver Program.

Input Data:

Welcome to Hadoop Session
Introduction to Hadoop
Introducing Hive
Hive Session
Pig Session

Act:

WordCountPartitioner.java

```
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Partitioner;

public class WordCountPartitioner extends Partitioner<Text, IntWritable> {
        @Override
        public int getPartition(Text key, IntWritable value, int numPartitions) {
                String word = key.toString();
                char alphabet = word.toUpperCase().charAt(0);
                int partitionNumber = 0;
                switch(alphabet) {
                case 'A': partitionNumber = 1; break;
                case 'B': partitionNumber = 2; break;
                case 'C': partitionNumber = 3; break;
                case 'D': partitionNumber = 4; break;
                case 'E': partitionNumber = 5; break;
                case 'F': partitionNumber = 6; break;
                case 'G': partitionNumber = 7; break;
                case 'H': partitionNumber = 8; break;
                case 'I': partitionNumber = 9; break;
                case 'J': partitionNumber = 10; break;
                case 'K': partitionNumber = 11; break;
                case 'L': partitionNumber = 12; break;
                case 'M': partitionNumber = 13; break;
                case 'N': partitionNumber = 14; break;
                case 'O': partitionNumber = 15; break;
                case 'P': partitionNumber = 16; break;
                case 'Q': partitionNumber = 17; break;
                case 'R': partitionNumber = 18; break;
                case 'S': partitionNumber = 19; break;
                case 'T': partitionNumber = 20; break;
                case 'U': partitionNumber = 21; break;
                case 'V': partitionNumber = 22; break;
                case 'W': partitionNumber = 23; break;
                case 'X': partitionNumber = 24; break;
                case 'Y': partitionNumber = 25; break;
                case 'Z': partitionNumber = 26; break;
                default: partitionNumber = 0; break;
                }
                return partitionNumber;
        }
}
```

In the driver program, set the partitioner class as shown below:

```
job.setNumReduceTasks(27);
job.setPartitionerClass(WordCountPartitioner.class);

// Input and Output Path
FileInputFormat.addInputPath(job, new Path("/mapreducedemos/lines.txt"));
FileOutputFormat.setOutputPath(job, new Path("/mapreducedemos/output/
wordcountpartitioner/"));
```

Output:

You can see 27 partitions in the below output.

Contents of directory /mapreducedemos/output/wordcountpartitioner

Goto : /mapreducedemos/output/wo go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00000	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00001	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00002	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00003	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00004	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00005	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00006	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00007	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00008	file	16 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00009	file	29 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00010	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00011	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00012	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00013	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00014	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00015	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00016	file	6 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00017	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00018	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00019	file	10 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00020	file	5 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00021	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00022	file	0 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00023	file	10 B	3	128 MB	2015-03-01 23:40	rw-r--r--	root	supergroup
part-r-00024	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00025	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup
part-r-00026	file	0 B	3	128 MB	2015-03-01 23:39	rw-r--r--	root	supergroup

The output file part-r-00008 is associated with alphabet **'H'**.

File: /mapreducedemos/output/wordcountpartitioner/part-r-00008

Goto : /mapreducedemos/output/wo go

Go back to dir listing

Advanced view/download options

Hadoop 2

Hive 2

8.6 SEARCHING

Objective: To write a MapReduce program to search for a specific keyword in a file.

Input Data:

```
1001,John,45
1002,Jack,39
1003,Alex,44
1004,Smith,38
1005,Bob,33
```

Act:

WordSearcher.java

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordSearcher {

        public static void main(String[] args) throws IOException,
                        InterruptedException, ClassNotFoundException {
                Configuration conf = new Configuration();
                Job job = new Job(conf);
                job.setJarByClass(WordSearcher.class);
                job.setOutputKeyClass(Text.class);
                job.setOutputValueClass(Text.class);
                job.setMapperClass(WordSearchMapper.class);
                job.setReducerClass(WordSearchReducer.class);
                job.setInputFormatClass(TextInputFormat.class);
                job.setOutputFormatClass(TextOutputFormat.class);
                job.setNumReduceTasks(1);
                job.getConfiguration().set("keyword", "Jack");
                FileInputFormat.setInputPaths(job, new Path("/mapreduce/student.csv"));
```

```
                FileOutputFormat.setOutputPath(job, new Path("/mapreduce/output/search"));
                System.exit(job.waitForCompletion(true) ? 0 : 1);
        }
}
```

WordSearchMapper.java

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.InputSplit;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.lib.input.FileSplit;

public class WordSearchMapper extends Mapper<LongWritable, Text, Text, Text> {

        static String keyword;
        static int pos = 0;

        protected void setup(Context context) throws IOException,
                        InterruptedException {

                Configuration configuration = context.getConfiguration();
                keyword = configuration.get("keyword");
        }

        protected void map(LongWritable key, Text value, Context context)
                        throws IOException, InterruptedException {

                InputSplit i = context.getInputSplit(); // Get the input split for this map.
                FileSplit f = (FileSplit) i;
                String fileName = f.getPath().getName();
                Integer wordPos;
                pos++;
                if (value.toString().contains(keyword)) {
                        wordPos = value.find(keyword);
                        context.write(value, new Text(fileName + ","+ new IntWritable(pos).
toString() + ", "+ wordPos.toString()));
                }
        }
}
```

WordSearchReducer.java

```
import java.io.IOException;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class WordSearchReducer extends Reducer<Text, Text, Text, Text> {

        protected void reduce(Text key, Text value, Context context)
                        throws IOException, InterruptedException {

                context.write(key, value);
        }
}
```

Output:

File: /mapreduce/output/search/part-r-00000

Goto : /mapreduce/output/search go

Go back to dir listing
Advanced view/download options

```
1002,Jack,39    student.csv,2, 5
```

8.7 SORTING

Objective: To write a MapReduce program to sort data by student name (value).

Input Data:

```
1001,John,45
1002,Jack,39
1003,Alex,44
1004,Smith,38
1005,Bob,33
```

Act:

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
```

```
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class SortStudNames {

        public static class SortMapper extends
                        Mapper<LongWritable, Text, Text, Text> {

                protected void map(LongWritable key, Text value, Context context)
                                throws IOException, InterruptedException {

                        String[ ] token = value.toString().split(",");
                        context.write(new Text(token[1]), new Text(token[0]+" - "+token[1]));
                }
        }
    // Here, value is sorted...
        public static class SortReducer extends
                        Reducer<Text, Text, NullWritable, Text> {

                public void reduce(Text key, Iterable<Text> values, Context context)
                                throws IOException, InterruptedException {

                        for (Text details : values) {
                                context.write(NullWritable.get(), details);
                        }
                }
        }

        public static void main(String[ ] args) throws IOException,
                        InterruptedException, ClassNotFoundException {

                Configuration conf = new Configuration();
                Job job = new Job(conf);
                job.setJarByClass(SortEmpNames.class);
                job.setMapperClass(SortMapper.class);
                job.setReducerClass(SortReducer.class);
                job.setOutputKeyClass(Text.class);
                job.setOutputValueClass(Text.class);
                FileInputFormat.setInputPaths(job, new Path("/mapreduce/student.csv"));
                FileOutputFormat.setOutputPath(job, new
Path("/mapreduce/output/sorted/"));
                System.exit(job.waitForCompletion(true) ? 0 : 1);
        }

}
```

Output:

File: /mapreduce/output/search/part-r-00000

Goto : /mapreduce/output/search go

Go back to dir listing
Advanced view/download options

```
1002,Jack,39    student.csv,2, 5
```

8.8 COMPRESSION

In MapReduce programming, you can compress the MapReduce output file. Compression provides two benefits as follows:

1. Reduces the space to store files.
2. Speeds up data transfer across the network.

You can specify compression format in the Driver Program as shown below:

```
conf.setBoolean("mapred.output.compress", true);
conf.setClass("mapred.output.compression.codec", GzipCodec.class,CompressionCodec.class);
```

Here, codec is the implementation of a compression and decompression algorithm. GzipCodec is the compression algorithm for gzip. This compresses the output file.

REMIND ME

- Mapper maps the input key–value pairs to intermediate key–value pairs.
- Reducer then reduces the set of key–value pairs that share a common key to a smaller set of values.
- The Reducer has three primary phases:
  - Shuffle and Sort
  - Reduce
  - Output Format
- Combiner and Partitioner are optimization techniques.

POINT ME (BOOK)

- MapReduce Design Patterns, O'REILLY, Donald Miner and Adam Shook.

CONNECT ME (INTERNET RESOURCES)

- http://hadooptutorial.wikispaces.com/MapReduce
- http://bigdataanalyticsnews.com/anatomy-mapreduce-job/
- http://bigdataconsultants.blogspot.in/2013/11/secondary-sort-in-hadoop-actor.html

TEST ME

A. Fill Me

1. Partitioner phase belongs to _____ task.
2. Combiner is also known as _____.
3. RecordReader converts byte-oriented view into _____ view.
4. MapReduce sorts the intermediate value based on _____.
5. In MapReduce Programming, reduce function is applied _____ group at a time.

Answers:

1. map
2. local reducer
3. record-oriented
4. keys
5. one

ASSIGNMENT FOR HANDS-ON PRACTICE

ASSIGNMENT 1

Objective: To learn about MapReduce Programming using Java.

Problem Description: Write a MapReduce Program to arrange the data on user id, then within the user id sort them in increasing order of the page count.

Input:

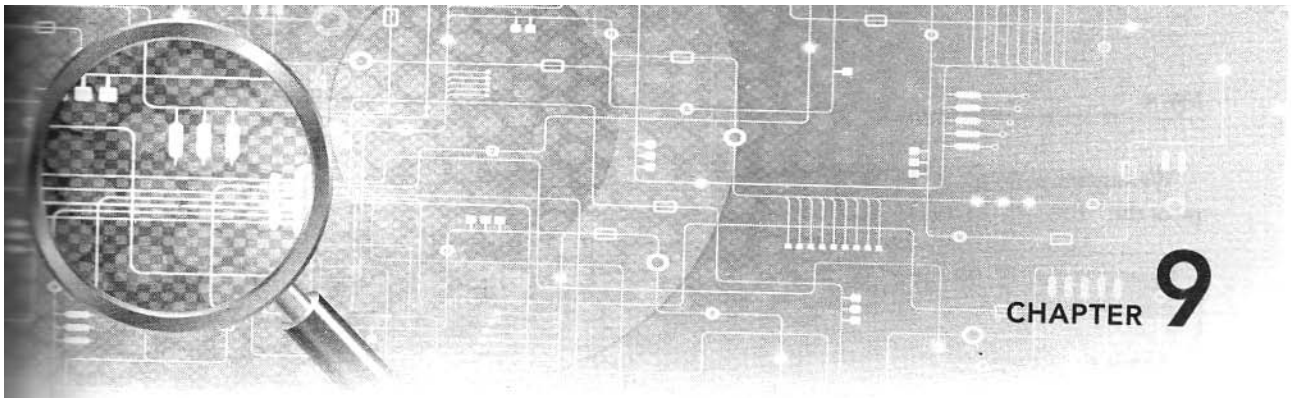
User_id	count	URL
12398	5	http://www.cbtnuggets.com/
23487	9	http://www.xda-developers.com/
34576	3	http://www.w3schools.com/
45665	6	https://www.google.co.in/
56754	4	http://www.encyclopedia.com/
67843	6	http://tutorialspoint.com/
78932	7	http://stackoverflow.com/
89021	3	http://www.wikipedia.org/
91210	2	http://www.cisce.org/results
82391	4	http://www.slideshare.net/

ASSIGNMENT 2

Objective: To learn about MapReduce Programming using Java.
Problem Description: Write a MapReduce Program to find unitwise salary.

Input:

Empno	Empname	Unit	Designation	Salary	Location
1001	John	IMST	TA	30000	Trivandrum
1002	Jack	CLOUD	PM	80000	Bangalore
1003	Joshi	FNPR	TA	35000	Trivandrum
1004	Josh	ECSSAP	PM	75000	Bangalore
1005	Jim	FSADM	SPM	60000	Bangalore
1006	Smith	ICS	TA	24000	Chandigarh
1007	Tiger	IMST	SPM	56000	Trivandrum
1008	Kate	FNPR	PM	76000	Chennai
1009	Cassy	MFGADM	TA	40000	Bangalore
1010	Ronald	ECSSAP	SPM	65000	Chennai



CHAPTER 9

Introduction to Hive

BRIEF CONTENTS

- What's in Store?
- What is Hive?
  - History of Hive and Recent Releases of Hive
  - Hive Features
  - Hive Integration and Work Flow
  - Hive Data Units
- Hive Architecture
- Hive Data Types
  - Primitive Data Types
  - Collection Data Types
- Hive File Format
  - Text File
  - Sequential File
  - RCFile (Record Columnar File)
- Hive Query Language (HQL)
  - DDL (Data Definition Language) Statements
  - DML (Data Manipulation Language) Statements
  - Starting Hive Shell
  - Database
  - Tables
  - Partitions
  - Bucketing
  - Views
  - Sub-Query
  - Joins
  - Aggregation
  - Group By and Having
- RCFILE Implementation
- SERDE
- User-Defined Function (UDF)

"Information is the oil of the 21st century, and analytics is the combustion engine."

– Peter Sondergaard, Gartner Research

WHAT'S IN STORE?

We assume that you are already familiar with commercial database systems. In this chapter, we will try to use that knowledge as our base to build a structure on Hadoop for effective analysis. We will discuss the importance of Hive with the help of use cases. We will also enrich your knowledge by working with Hive Query Language.

We suggest you refer to some of the learning resources suggested at the end of this chapter and also complete the "Test Me" exercises.

CASE STUDY: RETAIL LOG PROCESSING

About the Company

TENTOTEN is a Retail Store which has a chain of hypermarkets in India. They have 250+ stores across 95 cities and towns. About 45,000+ people are working in **TENTOTEN**. **TENTOTEN** deals in a wide range of products including fashion apparels, food products, books, furniture, etc. Around 1500+ customers visit and/or purchase products every day from each of these stores.

Problem Scenario

The approximate size of **TENTOTEN** log datasets is 12 TB. Information about the various stores is stored in the form of semi-structured data. Traditional Business Intelligence (BI) tools are good when data is present in pre-defined schema and datasets are just several hundreds of gigabytes. But the **TENTOTEN** dataset is mostly log dataset, which does not conform to any particular schema. Querying such large dataset is difficult and immensely time consuming.

The challenges are:

1. Moving the log dataset to HDFS (Hadoop Distributed File System).
2. Performing analysis on HDFS data.

Hadoop MapReduce can be used to resolve these issues. However we will still have to deal with the below constraints:

1. Writing complex MapReduce jobs in Java can be tedious and error prone.
2. Joining across large datasets is quite tricky.

Enter Hive to counter the above challenges.

9.1 WHAT IS HIVE?

Hive is a Data Warehousing tool that sits on top of Hadoop. Refer Figure 9.1. Hive is used to process structured data in Hadoop. The three main tasks performed by Apache Hive are:

1. Summarization
2. Querying
3. Analysis

Facebook initially created Hive component to manage their ever-growing volumes of log data. Later Apache software foundation developed it as open-source and it came to be known as Apache Hive.

Hive makes use of the following:

1. HDFS for Storage.
2. MapReduce for execution.
3. Stores metadata/schemas in an RDBMS.

Hive provides HQL (Hive Query Language) or HiveQL which is similar to SQL. Hive compiles SQL queries into MapReduce jobs and then runs the job in the Hadoop Cluster. It is designed to support

Hive – Suitable For		
Data warehousing applications	Processes batch jobs on huge data that is immutable (data whose structure cannot be changed after it is created is called immutable data).	Examples: Web Logs, Application Logs

Figure 9.1 Hive – a data warehousing tool.

OLAP (Online Analytical Processing). Hive provides extensive data type functions and formats for data summarization and analysis.

Note:

1. Hive is not RDBMS.
2. It is not designed to support OLTP (Online Transaction Processing).
3. It is not designed for real-time queries.
4. It is not designed to support row-level updates.

9.1.1 History of Hive and Recent Releases of Hive

The history of Hive and recent releases of Hive are illustrated pictorially in Figures 9.2 and 9.3, respectively.

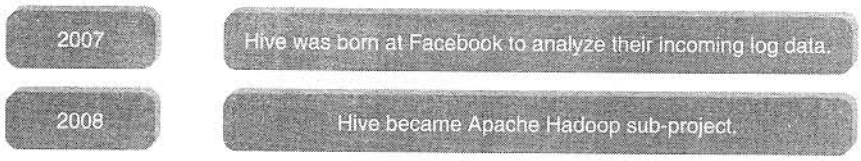


Figure 9.2 History of Hive.

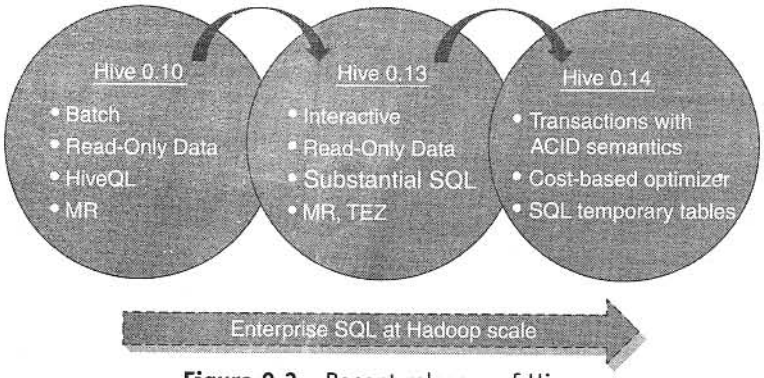


Figure 9.3 Recent releases of Hive.

9.1.2 Hive Features

1. It is similar to SQL.
2. HQL is easy to code.
3. Hive supports rich data types such as structs, lists and maps.
4. Hive supports SQL filters, group-by and order-by clauses.
5. Custom Types, Custom Functions can be defined.

9.1.3 Hive Integration and Work Flow

Figure 9.4 depicts the flow of log file analysis.

Explanation of the workflow. Hourly Log Data can be stored directly into HDFS and then data cleansing is performed on the log file. Finally, Hive table(s) can be created to query the log file.

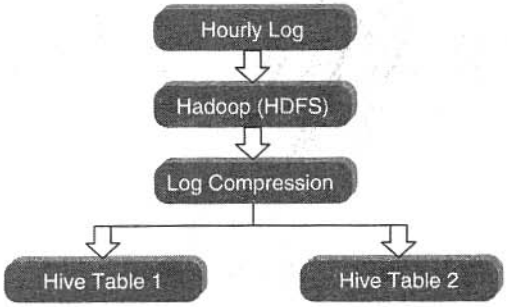


Figure 9.4 Flow of log analysis file.

9.1.4 Hive Data Units

1. **Databases:** The namespace for tables.
2. **Tables:** Set of records that have similar schema.
3. **Partitions:** Logical separations of data based on classification of given information as per specific attributes. Once hive has partitioned the data based on a specified key, it starts to assemble the records into specific folders as and when the records are inserted.
4. **Buckets (or Clusters):** Similar to partitions but uses hash function to segregate data and determines the cluster or bucket into which the record should be placed.

Let us take an example to understand partitioning and bucketing.

PICTURE THIS...

"XYZ Corp" has their customer base spread across 190+ countries. There are 5 million records/entities available. If it is required to fetch the entities pertaining to a particular country, in the absence of partitioning, there is no choice but to go through all of the 5 million entities. This despite the fact our query will eventually result in few thousand entities of the particular country. However, creating partitions based on country will greatly help to alleviate the performance issue by checking the data belonging to the partition for the country in question.

Partitioning tables changes how Hive structures the data storage. Hive will create subdirectories reflecting the partitioning structure like

.../customers/*country=ABC*

Although partitioning helps in enhancing performance and is recommended, having too many partitions may prove detrimental for few queries.

Bucketing is another technique of managing large datasets. If we partition the dataset based on customer_ID, we would end up with far too many partitions. Instead, if we bucket the customer table and use customer_id as the bucketing column, the value of this column will be hashed by a user-defined number

into buckets. Records with the same customer_id will always be placed in the same bucket. Assuming we have far more customer_ids than the number of buckets, each bucket will house many customer_ids. While creating the table you can specify the number of buckets that you would like your data to be distributed in using the syntax "CLUSTERED BY (customer_id) INTO XX BUCKETS"; here XX is the number of buckets.

When to Use Partitioning/Bucketing?

Bucketing works well when the field has high cardinality (cardinality is the number of values a column or field can have) and data is evenly distributed among buckets. Partitioning works best when the cardinality of the partitioning field is not too high. Partitioning can be done on multiple fields with an order (Year/Month/Day) whereas bucketing can be done on only one field.

Figure 9.5 shows how these data units are arranged in a Hive Cluster. Figure 9.6 describes the semblance of Hive structure with database.

A database contains several tables. Each table is constituted of rows and columns. In Hive, tables are stored as a folder and partition tables are stored as a sub-directory. Bucketed tables are stored as a file.

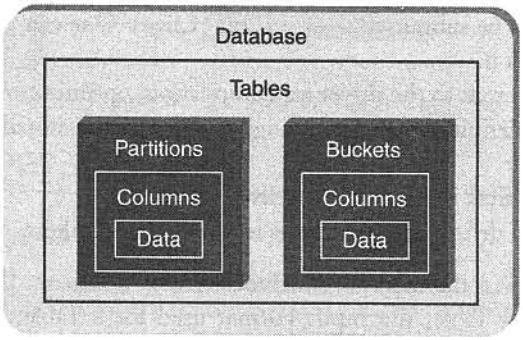


Figure 9.5 Data units as arranged in a Hive.

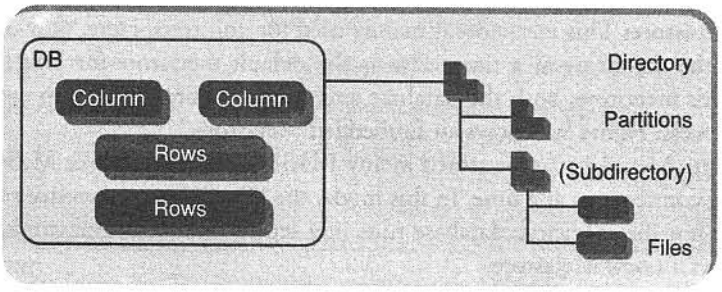


Figure 9.6 Semblance of Hive structure with database.

9.2 HIVE ARCHITECTURE

Hive Architecture is depicted in Figure 9.7. The various parts are as follows:

1. **Hive Command-Line Interface (Hive CLI):** The most commonly used interface to interact with Hive.
2. **Hive Web Interface:** It is a simple Graphic User Interface to interact with Hive and to execute query.
3. **Hive Server:** This is an optional server. This can be used to submit Hive Jobs from a remote client.

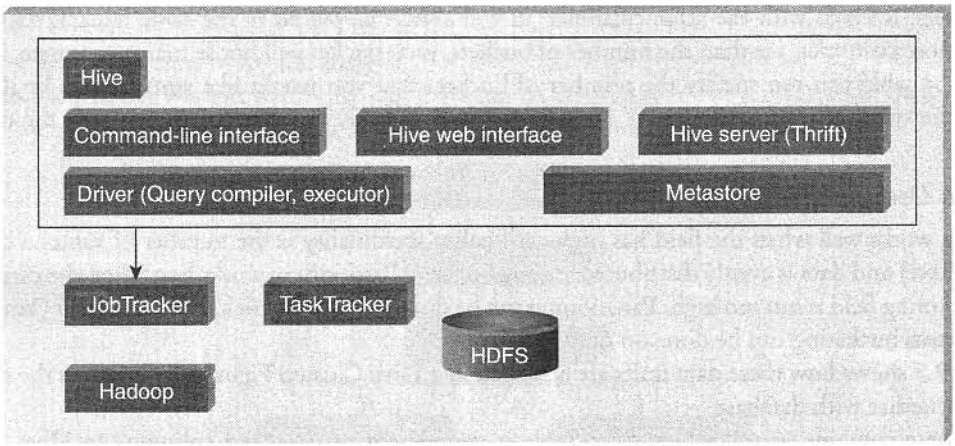


Figure 9.7 Hive architecture.

4. **JDBC/ODBC:** Jobs can be submitted from a JDBC Client. One can write a Java code to connect to Hive and submit jobs on it.
5. **Driver:** Hive queries are sent to the driver for compilation, optimization and execution.
6. **Metastore:** Hive table definitions and mappings to the data are stored in a Metastore. A Metastore consists of the following:
   - **Metastore service:** Offers interface to the Hive.
   - **Database:** Stores data definitions, mappings to the data and others.

The metadata which is stored in the metastore includes IDs of Database, IDs of Tables, IDs of Indexes, etc., the time of creation of a Table, the Input Format used for a Table, the Output Format used for a Table, etc. The metastore is updated whenever a table is created or deleted from Hive. There are three kinds of metastore.

1. **Embedded Metastore:** This metastore is mainly used for unit tests. Here, only one process is allowed to connect to the metastore at a time. This is the default metastore for Hive. It is Apache Derby Database. In this metastore, both the database and the metastore service run embedded in the main Hive Server process. Figure 9.8 shows an Embedded Metastore.
2. **Local Metastore:** Metadata can be stored in any RDBMS component like MySQL. Local metastore allows multiple connections at a time. In this mode, the Hive metastore service runs in the main Hive Server process, but the metastore database runs in a separate process, and can be on a separate host. Figure 9.9 shows a Local Metastore.
3. **Remote Metastore:** In this, the Hive driver and the metastore interface run on different JVMs (which can run on different machines as well) as in Figure 9.10. This way the database can be fire-walled from the Hive user and also database credentials are completely isolated from the users of Hive.

Figure 9.8 Embedded Metastore.

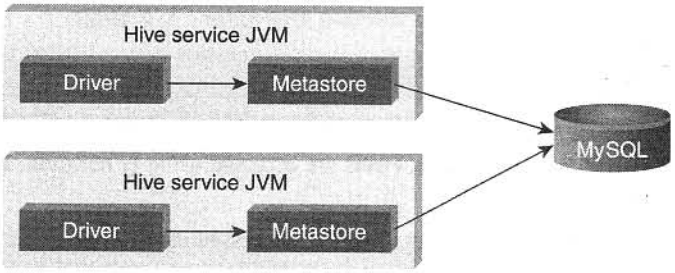


Figure 9.9 Local Metastore.

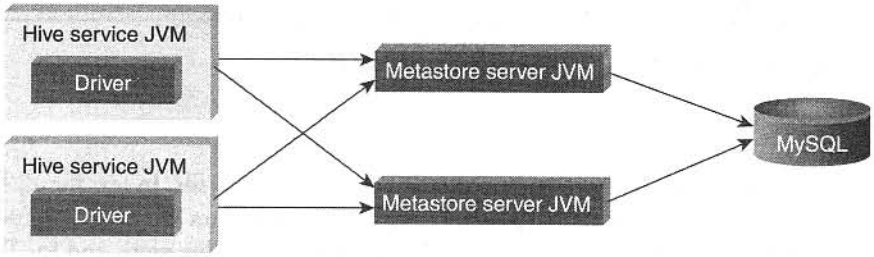


Figure 9.10 Remote Metastore.

9.3 HIVE DATA TYPES

9.3.1 Primitive Data Types

Numeric Data Type	
TINYINT	1-byte signed integer
SMALLINT	2-byte signed integer
INT	4-byte signed integer
BIGINT	8-byte signed integer
FLOAT	4-byte single-precision floating-point
DOUBLE	8-byte double-precision floating-point number

String Types	
STRING	
VARCHAR	Only available starting with Hive 0.12.0
CHAR	Only available starting with Hive 0.13.0

Strings can be expressed in either single quotes (') or double quotes (")

Miscellaneous Types	
BOOLEAN	
BINARY	Only available starting with Hive

9.3.2 Collection Data Types

Collection Data Types	
STRUCT	Similar to 'C' struct. Fields are accessed using dot notation. E.g.: struct('John', 'Doe')
MAP	A collection of key–value pairs. Fields are accessed using [] notation. E.g.: map('first', 'John', 'last', 'Doe')
ARRAY	Ordered sequence of same types. Fields are accessed using array index. E.g.: array('John', 'Doe')

9.4 HIVE FILE FORMAT

The file formats in Hive specify how records are encoded in a file.

9.4.1 Text File

The default file format is text file. In this format, each record is a line in the file. In text file, different control characters are used as delimiters. The delimiters are ^A (octal 001, separates all fields), ^B (octal 002, separates the elements in the array or struct), ^C (octal 003, separates key–value pair), and \n. The term **field** is used when overriding the default delimiter. The supported text files are CSV and TSV. JSON or XML documents too can be specified as text file.

9.4.2 Sequential File

Sequential files are flat files that store binary key–value pairs. It includes compression support which reduces the CPU, I/O requirement.

9.4.3 RCFile (Record Columnar File)

RCFile stores the data in **Column Oriented Manner** which ensures that **Aggregation** operation is not an expensive operation. For example, consider a table which contains four columns as shown in Table 9.1.

Instead of only partitioning the table horizontally like the row-oriented DBMS (row-store), RCFile partitions this table first horizontally and then vertically to serialize the data. Based on the user-specified value, first the table is partitioned into multiple row groups horizontally. Depicted in Table 9.2, Table 9.1 is partitioned into two row groups by considering three rows as the size of each row group.

Next, in every row group RCFile partitions the data vertically like column-store. So the table will be serialized as shown in Table 9.3.

Table 9.1 A table with four columns

C1	C2	C3	C4
11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44
51	52	53	54

Table 9.2 Table with two row groups

Row Group 1				Row Group 2			
C1	C2	C3	C4	C1	C2	C3	C4
11	12	13	14	41	42	43	44
21	22	23	24	51	52	53	54
31	32	33	34				

Table 9.3 Table in RCFile Format

Row Group 1	Row Group 2
11, 21, 31;	41, 51;
12, 22, 32;	42, 52;
13, 23, 33;	43, 53;
14, 24, 34;	44, 54;

9.5 HIVE QUERY LANGUAGE (HQL)

Hive query language provides basic SQL like operations. Here are few of the tasks which HQL can do easily.

1. Create and manage tables and partitions.
2. Support various Relational, Arithmetic, and Logical Operators.
3. Evaluate functions.
4. Download the contents of a table to a local directory or result of queries to HDFS directory.

9.5.1 DDL (Data Definition Language) Statements

These statements are used to build and modify the tables and other objects in the database. The DDL commands are as follows:

1. Create/Drop/Alter Database
2. Create/Drop/Truncate Table
3. Alter Table/Partition/Column
4. Create/Drop/Alter View
5. Create/Drop/Alter Index
6. Show
7. Describe

9.5.2 DML (Data Manipulation Language) Statements

These statements are used to retrieve, store, modify, delete, and update data in database. The DML commands are as follows:

1. Loading files into table.
2. Inserting data into Hive Tables from queries.

Note: Hive 0.14 supports update, delete, and transaction operations.

9.5.3 Starting Hive Shell

To start Hive, go to the installation path of Hive and type as below:

```
[root@volgalnx005 ~]# hive

Logging initialized using configuration in jar:file:/root/Desktop/VMDATA/Hive/hi
ve/lib/hive-common-0.14.0.jar!/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/root/Desktop/VMDATA/Hadoop/hadoop/share/hadoo
p/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/root/Desktop/VMDATA/Hive/hive/lib/hive-jdbc-0
.14.0-standalone.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive>
```

The sections have been designed as follows:

Objective:	What is it that we are trying to achieve here?
Input (optional):	What is the input that has been given to us to act upon?
Act:	The actual statement/command to accomplish the task at hand.
Outcome:	The result/output as a consequence of executing the statement.

9.5.4 Database

A database is like a container for data. It has a collection of tables which houses the data.

Objective: To create a database named "STUDENTS" with comments and database properties.

Act:

```
CREATE DATABASE IF NOT EXISTS STUDENTS COMMENT 'STUDENT Details'
WITH DBPROPERTIES ('creator' = 'JOHN');
```

Outcome:

```
hive> CREATE DATABASE IF NOT EXISTS STUDENTS COMMENT 'STUDENT Details' WITH DBPROPERTIES ('creato
r' = 'JOHN');
OK
Time taken: 0.536 seconds
hive>
```

Explanation of the syntax:

IF NOT EXIST: It is an optional clause. The create database statement with "IF Not EXISTS" clause creates a database if it does not exist. However, if the database already exists then it will notify the user that a database with the same name already exists and will not show any error message.

COMMENT: This is to provide short description about the database.

WITH DBPROPERTIES: It is an optional clause. It is used to specify any properties of database in the form of (key, value) separated pairs. In the above example, "Creator" is the "Key" and "JOHN" is the value.

We can use "**SCHEMA**" in place of "**DATABASE**" in this command.

Note: We have not specified the location where the Hive database will be created. By default all the Hive databases will be created under default warehouse directory (set by the property hive.metastore.warehouse.dir) as /user/hive/warehouse/database_name.db. But if we want to specify our own location, then the LOCATION clause can be specified. This clause is optional.

Objective: To display a list of all databases.

Act:

SHOW DATABASES;

Outcome:

```
hive> SHOW DATABASES;
OK
students
Time taken: 0.082 seconds, Fetched: 22 row(s)
hive>
```

By default, **SHOW DATABASES** lists all the databases available in the metastore. We can use "**SCHEMAS**" in place of "**DATABASES**" in this command. The command has an optional "Like" clause. It can be used to filter the database names using regular expressions such as "*" , "?", etc.

SHOW DATABASES LIKE "Stu*"

SHOW DATABASES like "Stud???s"

Objective: To describe a database.

Act:

DESCRIBE DATABASE STUDENTS;

Note: Shows only DB name, comment, and DB directory.

Outcome:

```
hive> DESCRIBE DATABASE STUDENTS;
OK
students        STUDENT Details hdfs://volgalnx010.ad.infosys.com:9000/user/hive/warehouse/studen
ts.db   root    USER
Time taken: 0.03 seconds, Fetched: 1 row(s)
hive>
```

Objective: To describe the extended database.

Act:

DESCRIBE DATABASE EXTENDED STUDENTS;

Note: Shows DB properties also.

Outcome:

```
hive> DESCRIBE DATABASE EXTENDED STUDENTS;
OK
students        STUDENT Details hdfs://volgalnx010.ad.infosys.com:9000/user/hive/warehouse/studen
ts.db   root    USER    {creator=JOHN}
Time taken: 0.027 seconds, Fetched: 1 row(s)
hive>
```

DESCRIBE DATABASE EXTENDED shows database's properties given under DBPROPERTIES argument at the time of creation.

We can use "**SCHEMA**" in place of "**DATABASE**", "**DESC**" in place of "**DESCRIBE**" in this command.

Objective: To alter the database properties.

Act:

```
ALTER DATABASE STUDENTS SET DBPROPERTIES ('edited-by' = 'JAMES');
```

Note: We can use the "ALTER DATABASE" command to

- Assign any new (key, value) pairs into DBPROPERTIES.
- Set owner user or role to the Database.

In Hive, it is not possible to unset the DB properties.

Outcome:

```
hive> ALTER DATABASE STUDENTS SET DBPROPERTIES ('edited-by' = 'JAMES');
OK
Time taken: 0.086 seconds
hive>
```

In the above example, the ALTER DATABASE command is used to assign new ('edited-by' = 'JAMES') pair into DBPROPERTIES. This can be verified by using the 'describe *extended*'.

Hive> DESCRIBE DATABASE Student EXTENDED

Objective: To make the database as current working database.

Act:

```
USE STUDENTS;
```

Outcome:

```
hive> USE STUDENTS;
OK
Time taken: 0.02 seconds
hive>
```

There is no command to show the current database, but use the below command statement to keep printing the current database name as suffix in the command line prompt.

set hive.cli.print.current.db=true;

Objective: To drop database.

Act:

```
DROP DATABASE STUDENTS;
```

Note: Hive creates database in the warehouse directory of Hive as shown below:

Contents of directory /user/hive/warehouse

Goto : /user/hive/warehouse go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
students.db	dir				2015-02-24 21:50	rwxr-xr-x	root	supergroup

Now assume that the database "STUDENTS" has 10 tables within it. How do we delete the complete database along with the tables contained therein?

Use the command:

```
DROP DATABASE STUDENTS CASCADE;
```

By default the mode is RESTRICT which implies that the database will NOT be dropped if it contains tables.

Note: The complete syntax is as follows:

DROP DATABASE [IF EXISTS] database_name [RESTRICT | CASCADE]

9.5.5 Tables

Hive provides two kinds of table:

1. Internal or Managed Table
2. External Table

9.5.5.1 *Managed Table*

1. Hive stores the Managed tables under the warehouse folder under Hive.
2. The complete life cycle of table and data is managed by Hive.
3. When the internal table is dropped, it drops the data as well as the metadata.

When you create a table in Hive, by default it is internal or managed table. If one needs to create an external table, one will have to use the keyword "EXTERNAL".

Objective: To create managed table named 'STUDENT'.

Act:

```
CREATE TABLE IF NOT EXISTS STUDENT(rollno INT,name STRING,gpa FLOAT) ROW
FORMAT DELIMITED FIELDS TERMINATED BY '\t';
```

Outcome:

```
hive> CREATE TABLE IF NOT EXISTS STUDENT(rollno INT,name STRING,gpa FLOAT) ROW FORMAT DELIMITED F
IELDS TERMINATED BY '\t';
OK
Time taken: 0.355 seconds
hive>
```

Objective: To describe the "STUDENT" table.

Act:

DESCRIBE STUDENT;

Outcome:

```
hive> DESCRIBE STUDENT;
OK
rollno                  int
name                    string
gpa                     float
Time taken: 0.163 seconds, Fetched: 3 row(s)
hive>
```

Note: Hive creates managed table in the warehouse directory of Hive as shown below:

Contents of directory /user/hive/warehouse/students.db

Goto : /user/hive/warehouse/student go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
student	dir				2015-02-24 22:03	rwxr-xr-x	root	supergroup

Go back to DFS home

Local logs

Log directory

Hadoop, 2015.

To check whether an existing table is managed or external, use the below syntax:

DESCRIBE FORMATTED tablename;

It displays complete metadata of a table. You will see one row called table type which will display either MANAGED_TABLE OR EXTERNAL_TABLE.

DESCRIBE FORMATTED STUDENT;

9.5.5.2 External or Self-Managed Table

1. When the table is dropped, it retains the data in the underlying location.
2. **External** keyword is used to create an external table.
3. **Location** needs to be specified to store the dataset in that particular location.

Objective: To create external table named 'EXT_STUDENT'.

Act:

```
CREATE EXTERNAL TABLE IF NOT EXISTS EXT_STUDENT(rollno INT,name STRING,gpa FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LOCATION '/STUDENT_INFO;
```

Outcome:

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS EXT_STUDENT(rollno INT,name STRING,gpa FLOAT) ROW FORMA
T DELIMITED FIELDS TERMINATED BY '\t'  LOCATION '/STUDENT_INFO';
OK
Time taken: 0.123 seconds
hive>
```

Note: Hive creates the external table in the specified location.

9.5.5.3 Loading Data into Table from File

Objective: To load data into the table from file named student.tsv.

Act:

```
LOAD DATA LOCAL INPATH '/root/hivedemos/student.tsv' OVERWRITE INTO TABLE EXT_STUDENT;
```

Note: Local keyword is used to load the data from the local file system. In this case, the file is copied. To load the data from HDFS, remove local key word from the statement. In this case, the file is moved from the original location.

Outcome:

```
hive> LOAD DATA LOCAL INPATH  '/root/hivedemos/student.tsv' OVERWRITE INTO TABLE EXT_STUDENT;
Loading data to table students.ext_student
Table students.ext_student stats: [numFiles=0, numRows=0, totalSize=0, rawDataSize=0]
OK
Time taken: 5.034 seconds
hive>
```

Hive loads the file in the specified location as shown below:

Contents of directory /STUDENT_INFO

Goto : STUDENT_INFO go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
student.tsv	file	121 B	3	128 MB	2015-02-24 22:19	rw-r--r--	root	supergroup

Go back to DFS home

Local logs

Log directory

Hadoop, 2015.

File: /STUDENT_INFO/student.tsv

Goto /STUDENT_INFO go

Go back to dir listing
Advanced view/download options

```
1001    John    3.0
1002    Jack    4.0
1003    Smith   4.5
1004    Jamesh  4.2
1005    Joshi   3.5
1006    Alex    4.0
1007    David   4.2
1008    Scott   3.9
```

Let us understand the difference between INTO TABLE and OVERWRITE TABLE with an example:

Assume the "EXT_STUDENT" table already had 100 records and the "student.tsv" file has 10 records. After issuing the LOAD DATA statement with the INTO TABLE clause, the table "EXT_STUDENT" will contain 110 records; however, the same LOAD DATA statement with the OVERWRITE clause will wipe out all the former content from the table and then load the 10 records from the data file.

9.5.5.4 Collection Data Types

Objective: To work with collection data types.

Input:
1001,John,Smith:Jones,Mark1!45:Mark2!46:Mark3!43
1002,Jack,Smith:Jones,Mark1!46:Mark2!47:Mark3!42

Act:

```
CREATE TABLE STUDENT_INFO (rollno INT,name String, sub ARRAY<STRING>,marks
MAP<STRING,INT>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
COLLECTION ITEMS TERMINATED BY ':'
MAP KEYS TERMINATED BY '!';
LOAD DATA LOCAL INPATH '/root/hivedemos/studentinfo.csv' INTO TABLE
STUDENT_INFO;
```

Outcome:

```
hive> CREATE TABLE STUDENT_INFO (rollno INT,name String, sub ARRAY<STRING>,marks MAP<STRING,FLOAT>)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > COLLECTION ITEMS TERMINATED BY ':'
    > MAP KEYS TERMINATED BY '|';
OK
Time taken: 0.112 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/studentinfo.csv' INTO TABLE STUDENT_INFO;
Loading data to table students.student_info
Table students.student_info stats: [numFiles=1, totalSize=109]
OK
Time taken: 0.397 seconds
hive>
```

9.5.5.5 Querying Table

Objective: To retrieve the student details from "EXT_STUDENT" table.

Act:

```
SELECT * from EXT_STUDENT;
```

Outcome:

```
hive> select * from EXT_STUDENT;
OK
1001    John    3.0
1002    Jack    4.0
1003    Smith   4.5
1004    Scott   4.2
1005    Joshi   3.5
1006    Alex    4.5
1007    David   4.2
1008    James   4.0
1009    John    3.0
1010    Joshi   3.5
Time taken: 0.054 seconds, Fetched: 10 row(s)
hive>
```

Objective: Querying Collection Data Types.

Act:

```
SELECT * from STUDENT_INFO;
SELECT NAME,SUB FROM STUDENT_INFO;
// To retrieve value of Mark1
SELECT NAME, MARKS['Mark1'] from STUDENT_INFO;
// To retrieve subordinate (array) value
SELECT NAME,SUB[0] FROM STUDENT_INFO;
```

Outcome:

```
hive> SELECT * from STUDENT_INFO;
OK
1001    John    ["Smith","Jones"]       {"Mark1":45,"Mark2":46,"Mark3":43}
1002    Jack    ["Smith","Jones"]       {"Mark1":46,"Mark2":47,"Mark3":42}
Time taken: 0.044 seconds, Fetched: 2 row(s)
```

```
hive> SELECT NAME,SUB FROM STUDENT_INFO;
OK
John    ["Smith","Jones"]
Jack    ["Smith","Jones"]
Time taken: 0.061 seconds, Fetched: 2 row(s)
hive>
```

```
hive> SELECT NAME, MARKS['Mark1']  from STUDENT_INFO;
OK
John    45
Jack    46
Time taken: 0.06 seconds, Fetched: 2 row(s)
hive>
```

```
hive> SELECT NAME,SUB[0] FROM STUDENT_INFO;
OK
John    Smith
Jack    Smith
Time taken: 0.071 seconds, Fetched: 2 row(s)
hive>
```

9.5.6 Partitions

In Hive, the query reads the entire dataset even though a where clause filter is specified on a particular column. This becomes a bottleneck in most of the MapReduce jobs as it involves huge degree of I/O. So it is necessary to reduce I/O required by the MapReduce job to improve the performance of the query. A very common method to reduce I/O is data partitioning.

Partitions split the larger dataset into more meaningful chunks.

We will try to understand partitioning with the help of a simple example.

PICTURE THIS...

"XYZ enterprise" has a wide customer base spread across several states in the US. The data has been fed to the central system. The senior leadership team would like to get a report providing the statewise Sales %.

The IT team has proposed two options to help service this request:

1. Run the query with the where clause of each state name (such as where StateName = "A"). This will mean that the entire dataset is scanned/read through for each and every state. If we have data for 25 states, the dataset is read 25 times (once for each state). Assuming we have 5 million records, it would mean that 5 million records are read 25 times. This can be a major performance bottle neck and will worsen as the dataset grows.
2. The second option stated here can help alleviate this problem. The IT team can start off with creating 25 folders (one for each state) and instruct to have the data pertaining to states place into the folder of the respective states. This arrangement will greatly help at the time of querying the data. To get the Sales % of a particular state, the folder belonging to that state only has to be scanned/read through.

This intelligent way of grouping data during data load is termed as PARTITIONING in Hive.

This brings us to the next question.

If you are looking through the folder for state = "A", is there any possibility of you coming across data for state = "B" or state = "C"? Think through! I am sure your answer will be No! And we know the reason.

A point to note here is that as we create the partitions, there is no need to include the partitioned column along with the other columns of the dataset as this is something that is automatically taken care of "BY PARTITIONED" clause.

In our example above, we will refrain from adding the partitioned column along with other columns of the dataset and trust Hive to automatically manage this.

Partition is of two types:

1. **STATIC PARTITION:** It is upon the user to mention the partition (the segregation unit) where the data from the file is to be loaded.
2. **DYNAMIC PARTITION:** The user is required to simply state the column, basis which the partitioning will take place. Hive will then create partitions basis the unique values in the column on which partition is to be carried out.

Points to consider as you create partitions:

1. STATIC PARTITIONING implies that the user controls everything from defining the PARTITION column to loading data into the various partitioned folders.
2. As in our example above, if STATIC partition is done over the STATE column and assume by mistake the data for state "B" is placed inside the partition for state "A", our query for data for state "B" is

bound to return zilch records. The reason is obvious. A Select fired on STATIC partition just takes into consideration the partition name, and does not consider the data held inside the partition.

3. DYNAMIC PARTITIONING means Hive will intelligently get the distinct values for partitioned column and segregate data into respective partitions. There is no manual intervention.

By default, dynamic partitioning is enabled in Hive. Also by default it is strictly implying that one is required to do one level of STATIC partitioning before Hive can perform DYNAMIC partitioning inside this STATIC segregation unit.

In order to go with full dynamic partitioning, we have to set below property to non-strict in Hive.

hive> set hive.exec.dynamic.partition.mode=nonstrict

9.5.6.1 *Static Partition*

Static partitions comprise columns whose values are known at compile time.

Objective: To create static partition based on "gpa" column.

Act:

CREATE TABLE IF NOT EXISTS STATIC_PART_STUDENT (rollno INT, name STRING) PARTITIONED BY (gpa FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

Outcome:

```
hive> CREATE TABLE IF NOT EXISTS STATIC_PART_STUDENT(rollno INT,name STRING) PARTITIONED BY (gpa FLOAT) RO
W FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.105 seconds
hive>
```

Objective: Load data into partition table from table.

Act:

INSERT OVERWRITE TABLE STATIC_PART_STUDENT PARTITION (gpa =4.0) SELECT rollno, name from EXT_STUDENT where gpa=4.0;

Outcome:

```
hive> INSERT OVERWRITE TABLE STATIC_PART_STUDENT PARTITION (gpa =4.0) SELECT rollno,name from EXT_STUDENT
where gpa=4.0;
Query ID = root_20150224230404_4500d58a-cb21-4912-ba40-788e5cf8f9da
Total jobs = 3
```

Hive creates the folder for the value specified in the partition.

Contents of directory /user/hive/warehouse/students.db

Goto /user/hive/warehouse/student go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
static_part_student	dir				2015-02-24 23:04	rwxr-xr-x	root	supergroup
student	dir				2015-02-24 22:03	rwxr-xr-x	root	supergroup
student_info	dir				2015-02-24 22:54	rwxr-xr-x	root	supergroup

Go back to DFS home

Local logs

Log directory

Contents of directory /user/hive/warehouse/students.db/static_part_student

Goto : /user/hive/warehouse/student go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
gpa=4.0	dir				2015-02-24 23:04	rwxr-xr-x	root	supergroup

Go back to DFS home

Local logs

Log directory

File: /user/hive/warehouse/students.db/static_part_student/gpa=4.0/000000_0

Goto : /user/hive/warehouse/student go

Go back to dir listing
Advanced view/download options

```
1002    Jack
1008    James
```

Objective: To add one more static partition based on "gpa" column using the "alter" statement.

Act:

ALTER TABLE STATIC_PART_STUDENT ADD PARTITION (gpa=3.5);
INSERT OVERWRITE TABLE STATIC_PART_STUDENT PARTITION (gpa =4.0) SELECT rollno,name from EXT_STUDENT where gpa=4.0;

Outcome:

```
hive> ALTER TABLE STATIC_PART_STUDENT ADD PARTITION (gpa=3.5);
OK
Time taken: 0.166 seconds
hive>
```

Contents of directory /user/hive/warehouse/students.db/static_part_student

Goto : /user/hive/warehouse/student go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
gpa=3.5	dir				2015-02-24 23:09	rwxr-xr-x	root	supergroup
gpa=4.0	dir				2015-02-24 23:11	rwxr-xr-x	root	supergroup

Go back to DFS home

9.5.6.2 Dynamic Partition

Dynamic partition have columns whose values are known only at Execution Time.

Objective: To create dynamic partition on column date.

Act:

CREATE TABLE IF NOT EXISTS DYNAMIC_PART_STUDENT(rollno INT,name STRING) PARTITIONED BY (gpa FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

Outcome:

```
hive> CREATE TABLE IF NOT EXISTS DYNAMIC_PART_STUDENT(rollno INT,name STRING) PARTITIONED BY (gpa FLOAT) R
OW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.166 seconds
hive>
```

Objective: To load data into a dynamic partition table from table.

Act:

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

Note: The dynamic partition strict mode requires at least one static partition column. To turn this off, set hive.exec.dynamic.partition.mode=nonstrict

INSERT OVERWRITE TABLE DYNAMIC_PART_STUDENT PARTITION (gpa) SELECT rollno,name,gpa from EXT_STUDENT;

Outcome:

Contents of directory /user/hive/warehouse/students.db/dynamic_part_student

Goto : /user/hive/warehouse/student go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
gpa=3.0	dir				2015-02-24 23:16	rwxr-xr-x	root	supergroup
gpa=3.5	dir				2015-02-24 23:16	rwxr-xr-x	root	supergroup
gpa=4.0	dir				2015-02-24 23:16	rwxr-xr-x	root	supergroup
gpa=4.2	dir				2015-02-24 23:16	rwxr-xr-x	root	supergroup
gpa=4.5	dir				2015-02-24 23:16	rwxr-xr-x	root	supergroup

Go back to DFS home

Note: Create partition for all values.

9.5.7 Bucketing

Bucketing is similar to partition. However, there is a subtle difference between partition and bucketing. In a partition, you need to create partition for each unique value of the column. This may lead to situations where you may end up with thousands of partitions. This can be avoided by using Bucketing in which you can limit the number of buckets that will be created. A bucket is a file whereas a partition is a directory.

Objective: To learn the concept of bucket in hive.

Act:

CREATE TABLE IF NOT EXISTS STUDENT (rollno INT,name STRING,grade FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

LOAD DATA LOCAL INPATH '/root/hivedemos/student.tsv' INTO TABLE STUDENT;

Set below property to enable bucketing.

set hive.enforce.bucketing=true;

```
// To create a bucketed table having 3 buckets
CREATE TABLE IF NOT EXISTS STUDENT_BUCKET (rollno INT,name STRING,grade
FLOAT)
CLUSTERED BY (grade) into 3 buckets;
// Load data to bucketed table
FROM STUDENT
INSERT OVERWRITE TABLE STUDENT_BUCKET
SELECT rollno,name,grade;
// To display content of first bucket
SELECT DISTINCT GRADE FROM STUDENT_BUCKET
TABLESAMPLE(BUCKET 1 OUT OF 3 ON GRADE);
```

Outcome:

```
hive> CREATE TABLE IF NOT EXISTS STUDENT (rollno INT,name STRING,grade FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.101 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/student.tsv' INTO TABLE STUDENT;
Loading data to table book.student
Table book.student stats: [numFiles=1, totalSize=145]
OK
Time taken: 0.536 seconds
hive>
```

```
hive> set hive.enforce.bucketing=true;
hive>
```

```
hive> CREATE TABLE IF NOT EXISTS STUDENT_BUCKET (rollno INT,name STRING,grade FLOAT)
    > CLUSTERED BY (grade) into 3 buckets;
OK
Time taken: 0.101 seconds
hive>
```

```
hive> FROM STUDENT
    > INSERT OVERWRITE TABLE STUDENT_BUCKET
    > SELECT rollno,name,grade;
```

3 buckets have been created as shown below:

Contents of directory /user/hive/warehouse/book.db/student_bucket

Goto : /user/hive/warehouse/book.db go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
000000_0	file	59 B	3	128 MB	2015-03-10 22:29	rw-r--r--	root	supergroup
000001_0	file	59 B	3	128 MB	2015-03-10 22:29	rw-r--r--	root	supergroup
000002_0	file	28 B	3	128 MB	2015-03-10 22:29	rw-r--r--	root	supergroup

Go back to DFS home

Local logs

Log directory

Hadoop, 2015.

```
hive>
    > SELECT DISTINCT GRADE FROM STUDENT_BUCKET
    > TABLESAMPLE(BUCKET 1 OUT OF 3 ON GRADE);
OK
4.0
4.2
Time taken: 21.117 seconds, Fetched: 2 row(s)
hive>
```

9.5.8 Views

In Hive, view support is available only in version starting from 0.6. Views are purely logical object.

Objective: To create a view table named "STUDENT_VIEW".

Act:

CREATE VIEW STUDENT_VIEW AS SELECT rollno, name FROM EXT_STUDENT;

Outcome:

```
hive> CREATE VIEW STUDENT_VIEW AS SELECT rollno,name FROM EXT_STUDENT;
OK
Time taken: 0.606 seconds
hive>
```

Objective: Querying the view "STUDENT_VIEW".

Act:

SELECT * FROM STUDENT_VIEW LIMIT 4;

Outcome:

```
hive> SELECT * FROM STUDENT_VIEW LIMIT 4;
OK
1001    John
1002    Jack
1003    Smith
1004    Scott
Time taken: 0.279 seconds, Fetched: 4 row(s)
hive>
```

Objective: To drop the view "STUDENT_VIEW".

Act:

DROP VIEW STUDENT_VIEW;

Outcome:

```
hive> DROP VIEW STUDENT_VIEW;
OK
Time taken: 0.452 seconds
hive>
```

9.5.9 Sub-Query

In Hive, sub-queries are supported only in the FROM clause (Hive 0.12). You need to specify name for sub-query because every table in a FROM clause has a name. The columns in the sub-query select list should have unique names. The columns in the subquery select list are available to the outer query just like columns of a table.

Objective: Write a sub-query to count occurrence of similar words in the file.

Act:

```
CREATE TABLE docs (line STRING);
LOAD DATA LOCAL INPATH '/root/hivedemos/lines.txt' OVERWRITE INTO TABLE docs;
CREATE TABLE word_count AS
SELECT word, count(1) AS count FROM
(SELECT explode (split (line, ' ')) AS word FROM docs) w
GROUP BY word
ORDER BY word;
SELECT * FROM word_count;
```

Outcome:

```
hive> CREATE TABLE docs (line STRING);
OK
Time taken: 0.118 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/lines.txt' OVERWRITE INTO TABLE docs;
Loading data to table students.docs
Table students.docs stats: [numFiles=1, numRows=0, totalSize=91, rawDataSize=0]
OK
Time taken: 2.697 seconds
hive>
```

```
hive> CREATE TABLE word_count AS
    > SELECT word, count(1) AS count FROM
    > (SELECT explode (split (line, ' ')) AS word FROM docs) w
    > GROUP BY word
    > ORDER BY word;
```

```
hive> SELECT * FROM word_count;
OK
Hadoop  2
Hive    2
Introducing     1
Introduction    1
Pig     1
Session 3
Welcome 1
to      2
Time taken: 0.062 seconds, Fetched: 8 row(s)
hive>
```

Note: The explode() function takes an array as input and outputs the elements of the array as separate rows.

In Hive 0.13, sub-queries are supported in the where clause as well.

9.5.10 Joins

Joins in Hive is similar to the SQL Join.

Objective: To create JOIN between Student and Department tables where we use RollNo from both the tables as the join key.

Act:

CREATE TABLE IF NOT EXISTS STUDENT(rollno INT,name STRING,gpa FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

LOAD DATA LOCAL INPATH '/root/hivedemos/student.tsv' OVERWRITE INTO TABLE STUDENT;

CREATE TABLE IF NOT EXISTS DEPARTMENT(rollno INT,deptno int,name STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

LOAD DATA LOCAL INPATH '/root/hivedemos/department.tsv' OVERWRITE INTO TABLE DEPARTMENT;

SELECT a.rollno, a.name, a.gpa, b.deptno FROM STUDENT a JOIN DEPARTMENT b ON a.rollno = b.rollno;

Outcome:

```
hive> CREATE TABLE IF NOT EXISTS STUDENT(rollno INT,name STRING,gpa FLOAT) ROW FORMAT D
ELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.115 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/student.tsv' OVERWRITE INTO TABLE STUDENT
;
Loading data to table students.student
Table students.student stats: [numFiles=1, numRows=0, totalSize=145, rawDataSize=0]
OK
Time taken: 0.723 seconds
hive>
```

```
hive> CREATE TABLE IF NOT EXISTS DEPARTMENT(rollno INT,deptno int,name STRING) ROW FORM
AT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.099 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/department.tsv' OVERWRITE INTO TABLE DEPA
RTMENT;
Loading data to table students.department
Table students.department stats: [numFiles=1, numRows=0, totalSize=120, rawDataSize=0]
OK
Time taken: 0.442 seconds
hive>
```

```
hive> SELECT a.rollno, a.name, a.gpa, b.deptno  FROM STUDENT  a JOIN DEPARTMENT b ON a.
rollno = b.rollno;
OK
1001    John    3.0     101
1002    Jack    4.0     102
1003    Smith   4.5     103
1004    Scott   4.2     104
1005    Joshi   3.5     105
1006    Alex    4.5     101
1007    David   4.2     104
1008    James   4.0     102
Time taken: 115.282 seconds, Fetched: 8 row(s)
hive>
```

9.5.11 Aggregation

Hive supports aggregation functions like avg, count, etc.

Objective: To write the average and count aggregation functions.

Act:

SELECT avg(gpa) FROM STUDENT;
SELECT count(*) FROM STUDENT;

Outcome:

```
hive> SELECT avg(gpa) FROM STUDENT;

OK
3.839999961853027
Time taken: 28.639 seconds, Fetched: 1 row(s)
hive>

hive> SELECT count(*) FROM STUDENT;

OK
10
Time taken: 26.218 seconds, Fetched: 1 row(s)
hive>
```

9.5.12 Group By and Having

Data in a column or columns can be grouped on the basis of values contained therein by using "Group By". "Having" clause is used to filter out groups NOT meeting the specified condition.

Objective: To write group by and having function.

Act:

SELECT rollno, name,gpa FROM STUDENT GROUP BY rollno,name,gpa HAVING gpa > 4.0;

Outcome:

```
1003    Smith   4.5
1004    Scott   4.2
1006    Alex    4.5
1007    David   4.2
Time taken: 78.972 seconds, Fetched: 4 row(s)
hive>
```

9.6 RCFILE IMPLEMENTATION

RCFile (Record Columnar File) is a data placement structure that determines how to store relational tables on computer clusters.

Objective: To work with RCFILE Format.

Act:

```
CREATE TABLE STUDENT_RC( rollno int, name string,gpa float ) STORED AS RCFILE;
INSERT OVERWRITE table STUDENT_RC SELECT * FROM STUDENT;
SELECT SUM(gpa) FROM STUDENT_RC;
```

Outcome:

```
hive> CREATE TABLE STUDENT_RC( rollno int, name string,gpa float ) STORED AS RCFILE;
OK
Time taken: 0.093 seconds
hive>
```

```
hive> INSERT OVERWRITE table STUDENT_RC SELECT * from STUDENT;

hive> SELECT SUM(gpa) from STUDENT_RC;

OK
38.39999961853027
Time taken: 25.41 seconds, Fetched: 1 row(s)
hive>
```

Note: Stores the data in column oriented manner.

File: /user/hive/warehouse/students.db/student_rc/000000_0

Goto : /user/hive/warehouse/student go

Go back to dir listing
Advanced view, download options

```
RCFile.io.rcfile.column.number3V
◆..    ◆◆◆((◆◆◆◆◆◆◆◆◆◆◆◆◆◆JohnJackSmithScottJoshiAlexDavidJamesJohnJoshi
```

9.7 SERDE

SerDe stands for Serializer/Deserializer.

1. Contains the logic to convert unstructured data into records.
2. Implemented using Java.
3. Serializers are used at the time of writing.
4. Deserializers are used at query time (SELECT Statement).

Deserializer interface takes a binary representation or string of a record, converts it into a java object that Hive can then manipulate. Serializer takes a java object that Hive has been working with and translates it into something that Hive can write to HDFS.

Objective: To manipulate the XML data.

Input:

<employee> <empid>1001</empid> <name>John</name> <designation>Team Lead</designation> </employee>

<employee> <empid>1002</empid> <name>Smith</name> <designation>Analyst</designation> </employee>

Act:

```
CREATE TABLE XMLSAMPLE(xmldata string);
LOAD DATA LOCAL INPATH '/root/hivedemos/input.xml' INTO TABLE XMLSAMPLE;

CREATE TABLE xpath_table AS
SELECT xpath_int(xmldata,'employee/empid'),
xpath_string(xmldata,'employee/name'),
xpath_string(xmldata,'employee/designation')
FROM xmlsample;

SELECT * FROM xpath_table;
```

Outcome:

```
hive> CREATE TABLE XMLSAMPLE(xmldata string);
OK
Time taken: 0.244 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/root/hivedemos/input.xml' INTO TABLE XMLSAMPLE;
Loading data to table students.xmlsample
Table students.xmlsample stats: [numFiles=1, totalSize=194]
OK
Time taken: 0.889 seconds
hive>
```

```
hive> CREATE TABLE xpath_table AS
    > SELECT xpath_int(xmldata,'employee/empid'),
    > xpath_string(xmldata,'employee/name'),
    > xpath_string(xmldata,'employee/designation')
    > FROM xmlsample;
```

```
hive> SELECT * FROM xpath_table;
OK
1001    John    Team Lead
1002    Smith   Analyst
Time taken: 0.064 seconds, Fetched: 2 row(s)
hive>
```

9.8 USER-DEFINED FUNCTION (UDF)

In Hive, you can use custom functions by defining the User-Defined Function (UDF).

Objective: Write a Hive function to convert the values of a field to uppercase.

Act:

```
package com.example.hive.udf;
import org.apache.hadoop.hive.ql.exec.Description;
import org.apache.hadoop.hive.ql.exec.UDF;
@Description(
 name="SimpleUDFExample")
```

```
public final class MyLowerCase extends UDF {
  public String evaluate(final String word) {
   return word.toLowerCase();
  }
}
```

Note: Convert this Java Program into Jar.

```
ADD JAR /root/hivedemos/UpperCase.jar;
CREATE TEMPORARY FUNCTION touppercase AS 'com.example.hive.udf.MyUpperCase';
SELECT TOUPPERCASE(name) FROM STUDENT;
```

Outcome:

```
hive> ADD JAR /root/hivedemos/UpperCase.jar;
Added [/root/hivedemos/UpperCase.jar] to class path
Added resources: [/root/hivedemos/UpperCase.jar]
hive> CREATE TEMPORARY FUNCTION touppercase AS 'com.example.hive.udf.MyUpperCase';
OK
Time taken: 0.014 seconds
hive>
```

```
hive> Select touppercase (name) from STUDENT;
OK
JOHN
JACK
SMITH
SCOTT
JOSHI
ALEX
DAVID
JAMES
JOHN
JOSHI
Time taken: 0.061 seconds, Fetched: 10 row(s)
hive>
```

REMIND ME

- Hive is a Data Warehousing tool.
- Hive is used to query structured data built on top of Hadoop.
- Hive provides HQL (Hive Query Language) which is similar to SQL.
- A Hive database contains several tables. Each table is constituted of rows and columns. In Hive, tables are stored as a folder and partition tables are stored as a sub-directory.
- Bucketed tables are stored as a file.

POINT ME (BOOKS)

- Programming Hive, Jason Rutherglen, O'Reilly Publication.

CONNECT ME (INTERNET RESOURCES)

- **http://en.wikipedia.org/wiki/RCFile**
- **https://cwiki.apache.org/confluence/display/Hive/DynamicPartitions**
- **https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL**
- **https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML**

TEST ME

A. Match Me

Column A	Column B
HQL	Web Logs
Database	struct, map
Complex Data Types	Set of records
Hive Application	Hive Query Language
Table	Namespace

Answers:

Column A	Column B
HQL	Hive Query Language
Database	Namespace
Complex Data Types	struct, map
Hive Application	Web Logs
Table	Set of records

B. Fill Me

1. The metastore consists of _____ and a _____.
2. The most commonly used interface to interact with Hive is _____.
3. The default metastore for Hive is _____.
4. Metastore contains _____ of Hive tables.
5. _____ is responsible for compilation, optimization, and execution of Hive queries.

Answers:

1. Metaservices, database
2. Command Line Interface
3. Derby
4. System Catalog
5. Driver

ASSIGNMENTS FOR HANDS-ON PRACTICE

ASSIGNMENT 1: PARTITION

Objective: To learn about partitions in hive.

Problem Description:

Create a partition table for customer schema to reward the customers based on their life time values.

Input:

Customer ID	Customers	Life Time Value
1001	Jack	25000
1002	Smith	8000
1003	David	12000
1004	John	15000
1005	Scott	12000
1006	Joshi	28000
1007	Ajay	12000
1008	Vinay	30000
1009	Joseph	21000

- Create a partition table if life time value is 12000.
- Create a partition table for all life time values.

ASSIGNMENT 2: PARTITION

1. Create Table:

```
CREATE EXTERNAL TABLE Emp_Proj (
   EmpID INT,
   TechnologyID INT,
   StartData date,
    EndDate date
) PARTITIONED BY (ProjectID INT)
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY '#'
  STORED AS TEXTFILE;
```

2. Load partitioned data:

```
LOAD DATA INPATH '/hive/data/Employee' INTO TABLE Emp_Proj PARTITION (ProjectID=1)
```

3. Verify data load:

```
SELECT * from Emp_Proj where ProjectID = 1;
```

Did you notice, it prints all data, that is, data with ProjectID = 2,3,4,5 as well.
Why did this happen? Was partitioning NOT performed?

4. Verify file location:

hadoop fs -ls /user/hive/warehouse/Employee.db/Emp_Proj/

It should have folder named ProjectID=1.

DYNAMIC PARTITIONING:

1. Create Table:

```
CREATE EXTERNAL TABLE Emp_Proj_DP (
    EmpID INT,
   TechnologyID INT,
   StartData date,
    EndDate date
) PARTITIONED BY (ProjectID INT)
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY '#'
  STORED AS TEXTFILE;
```

2. Create TEMP table:

```
CREATE EXTERNAL TABLE Temp_ProjectID (
    EmpID INT,
   TechnologyID INT,
   StartData date,
    EndDate date,
    ProjectID INT
)
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY '#'
  STORED AS TEXTFILE;
```

3. Load data in temp table from file:

LOAD DATA LOCAL INPATH '/home/Seema/bigdata/hivedata/Employee' INTO TABLE Temp_ProjectID;

4. Load data in dynamic partitioned table:

INSERT INTO TABLE Emp_Proj_DP PARTITION (ProjectID) SELECT EmpID, TechnologyID , StartDate , EndDate FROM Temp_ProjectID;

Note: The column order while select should be maintained except partitioned column, which should be selected last and if there are multiple partitioned columns then they should be cited in the order of creation.

5. Verify data load:

```
SELECT * from Emp_Proj_DP where ProjectID = 1;
SELECT * from Emp_Proj_DP where ProjectID = 2;
SELECT * from Emp_Proj_DP where ProjectID = 3;
SELECT * from Emp_Proj_DP where ProjectID = 4;
```

6. Verify file location:

```
hadoop fs -ls /user/hive/warehouse/Employee.db/Emp_Proj_DP/
```

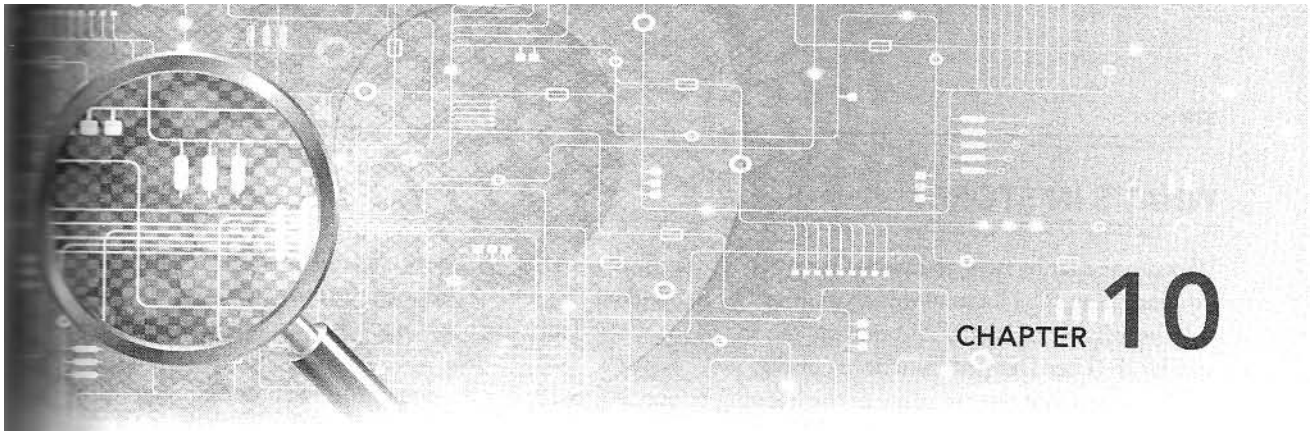
It should have folder named ProjectID=1, ProjectID=2, ProjectID=3, etc..

ASSIGNMENT 3: HIVEQL

Objective: To learn about HiveQL statement.
Problem Description:
Create a data file for below schemas:

- **Order:** CustomerId, ItemId, ItemName, OrderDate, DeliveryDate
- **Customer:** CustomerId, CustomerName, Address, City, State, Country

1. Create a table for Order and Customer Data.
2. Write a HiveQL to find number of items bought by each customer.



CHAPTER 10

Introduction to Pig

BRIEF CONTENTS

- What's in Store?
- What is Pig?
  - Key Features of Pig
- The Anatomy of Pig
- Pig on Hadoop
- Pig Philosophy
- Use Case for Pig: ETL Processing
- Pig Latin Overview
  - Pig Latin Statements
  - Pig Latin: Keywords
  - Pig Latin: Identifiers
  - Pig Latin: Comments
  - Pig Latin: Case Sensitivity
  - Operators in Pig Latin
- Data Types in Pig
  - Simple Data Types
  - Complex Data Types
- Running Pig
  - Interactive Mode
  - Batch Mode
- Execution Modes of Pig
  - Local Mode
  - MapReduce Mode
- HDFS Commands
- Relational Operators
- EVAL Function
- Complex Data Types
  - Tuple
  - Map
- Piggy Bank
- User-Defined Functions (UDF)
- Parameter Substitution
- Diagnostic Operator
- Word Count Example using Pig
- When to use Pig?
- When NOT to use Pig?
- Pig at Yahoo!
- Pig versus Hive

"If you can't explain it simply, you don't understand it well enough."

– Albert Einstein, Physicist

WHAT'S IN STORE?

We assume that by now you would have become familiar with the basic concepts of HDFS and MapReduce Programming. The focus of this chapter will be to build on this knowledge to perform analysis using Pig. We will discuss few relational and eval operators of Pig. We will also discuss Complex Data Types, Piggy Bank, and UDF (User Defined Functions) of Pig.

We suggest you refer to some of the learning resources provided at the end of this chapter for better learning. We also suggest you to practice "Test Me" exercises.

10.1 WHAT IS PIG?

Apache Pig is a platform for data analysis. It is an alternative to MapReduce Programming. Pig was developed as a research project at Yahoo.

10.1.1 Key Features of Pig

1. It provides an **engine** for executing **data flows** (how your data should flow). Pig processes data in parallel on the Hadoop cluster.
2. It provides a language called "**Pig Latin**" to express data flows.
3. Pig Latin contains operators for many of the traditional data operations such as join, filter, sort, etc.
4. It allows users to develop their own functions (User Defined Functions) for reading, processing, and writing data.

10.2 THE ANATOMY OF PIG

The main components of Pig are as follows:

1. Data flow language (**Pig Latin**).
2. Interactive shell where you can type Pig Latin statements (**Grunt**).
3. Pig interpreter and execution engine.

Refer Figure 10.1.

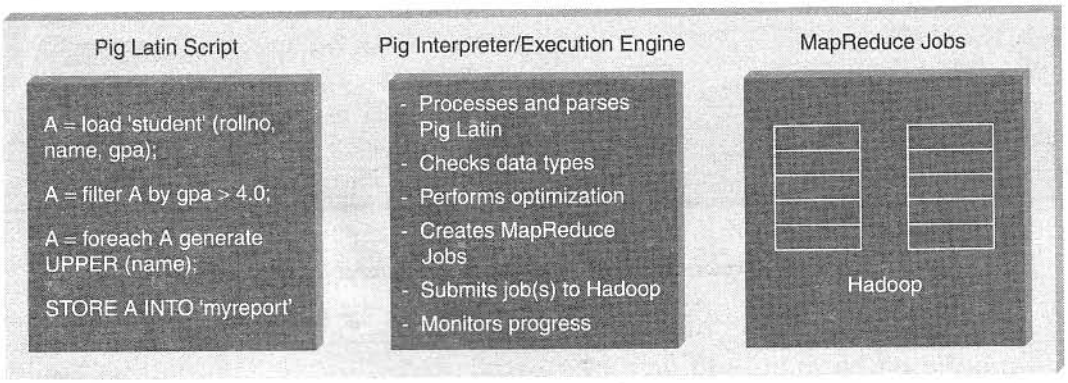


Figure 10.1 The anatomy of Pig.

10.3 PIG ON HADOOP

Pig runs on Hadoop. Pig uses both Hadoop Distributed File System and MapReduce Programming. By default, Pig reads input files from HDFS. Pig stores the intermediate data (data produced by MapReduce jobs) and the output in HDFS. However, Pig can also read input from and place output to other sources.

Pig supports the following:

1. HDFS commands.
2. UNIX shell commands.
3. Relational operators.
4. Positional parameters.
5. Common mathematical functions.
6. Custom functions.
7. Complex data structures.

10.4 PIG PHILOSOPHY

Figure 10.2 describes the Pig philosophy.

1. **Pigs Eat Anything:** Pig can process different kinds of data such as structured and unstructured data.
2. **Pigs Live Anywhere:** Pig not only processes files in HDFS, it also processes files in other sources such as files in the local file system.
3. **Pigs are Domestic Animals:** Pig allows you to develop user-defined functions and the same can be included in the script for complex operations.
4. **Pigs Fly:** Pig processes data quickly.

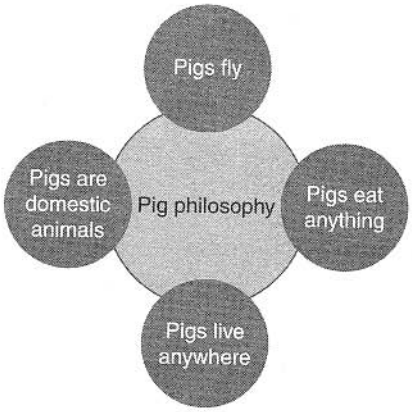


Figure 10.2 Pig philosophy.

10.5 USE CASE FOR PIG: ETL PROCESSING

Pig is widely used for "ETL" (Extract, Transform, and Load). Pig can extract data from different sources such as ERP, Accounting, Flat Files, etc. Pig then makes use of various operators to perform transformation on the data and subsequently loads it into the data warehouse. Refer Figure 10.3.

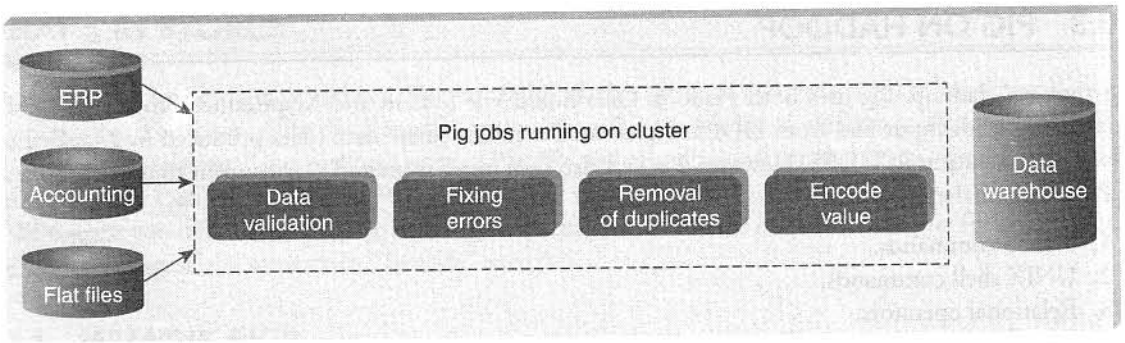


Figure 10.3 Pig: ETL Processing.

10.6 PIG LATIN OVERVIEW

10.6.1 Pig Latin Statements

1. Pig Latin statements are basic constructs to process data using Pig.
2. Pig Latin statement is an operator.
3. An operator in Pig Latin takes a relation as input and yields another relation as output.
4. Pig Latin statements include schemas and expressions to process data.
5. Pig Latin statements should end with a semi-colon.

Pig Latin Statements are generally ordered as follows:

1. **LOAD** statement that reads data from the file system.
2. Series of statements to perform transformations.
3. **DUMP** or **STORE** to display/store result.

The following is a simple Pig Latin script to load, filter, and store "student" data.

```
A = load 'student' (rollno, name, gpa);
A = filter A by gpa > 4.0;
A = foreach A generate UPPER (name);
STORE A INTO 'myreport'
```

Note: In the above example **A** is a relation and NOT a variable.

10.6.2 Pig Latin: Keywords

Keywords are reserved. It cannot be used to name things.

10.6.3 Pig Latin: Identifiers

1. Identifiers are names assigned to fields or other data structures.
2. It should begin with a letter and should be followed only by letters, numbers, and underscores.

Table 10.1 Valid and invalid identifiers

Valid Identifier	**Y**	**A1**	**A1_2014**	**Sample**
Invalid Identifier	**5**	**Sales$**	**Sales%**	**_Sales**

Table 10.1 describes valid and invalid identifiers.

10.6.4 Pig Latin: Comments

In Pig Latin two types of comments are supported:

1. Single line comments that begin with "--".
2. Multiline comments that begin with "/* and end with */".

10.6.5 Pig Latin: Case Sensitivity

1. Keywords are *not* case sensitive such as LOAD, STORE, GROUP, FOREACH, DUMP, etc.
2. Relations and paths are case-sensitive.
3. Function names are case sensitive such as PigStorage, COUNT.

10.6.6 Operators in Pig Latin

Table 10.2 describes operators in Pig Latin.

Table 10.2 Operators in Pig Latin

Arithmetic	Comparison	Null	Boolean
+	==	IS NULL	AND
−	! =	IS NOT NULL	OR
*	<		NOT
/	>		
%	<=		
	>=		

10.7 DATA TYPES IN PIG

10.7.1 Simple Data Types

Table 10.3 describes simple data types supported in Pig. In Pig, fields of unspecified types are considered as an array of bytes which is known as bytearray.

Null: In Pig Latin, **NULL** denotes a value that is unknown or is non-existent.

10.7.2 Complex Data Types

Table 10.4 describes complex data types in Pig.

Table 10.3 Simple data types supported in Pig

Name	Description
Int	Whole numbers
Long	Large whole numbers
Float	Decimals
Double	Very precise decimals
Chararray	Text strings
Bytearray	Raw bytes
Datetime	Datetime
Boolean	true or false

Table 10.4 Complex data types in Pig

Name	Description
Tuple	An ordered set of fields. **Example: (2,3)**
Bag	A collection of tuples. **Example: {(2,3),(7,5)}**
map	key, value pair **(open # Apache)**

10.8 RUNNING PIG

You can run Pig in two ways:

1. Interactive Mode.
2. Batch Mode.

10.8.1 Interactive Mode

You can run Pig in interactive mode by invoking **grunt** shell. Type pig to get grunt shell as shown below.

```
root@volgalnx010:~
[root@volgalnx010 ~]# pig
2015-02-23 21:07:38,916 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.
1.3 (rexported) compiled Sep 16 2014, 20:39:43
2015-02-23 21:07:38,917 [main] INFO  org.apache.pig.Main - Logging error messages to: /roo
t/pig_1424705858915.log
2015-02-23 21:07:38,934 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file
/root/.pigbootup not found
2015-02-23 21:07:39,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - ma
pred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2015-02-23 21:07:39,313 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs
.default.name is deprecated. Instead, use fs.defaultFS
2015-02-23 21:07:39,313 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecut
ionEngine - Connecting to hadoop file system at: hdfs://volgalnx010.ad.infosys.com:9000
2015-02-23 21:07:39,800 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to l
oad native-hadoop library for your platform... using builtin-java classes where applicable
2015-02-23 21:07:40,234 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs
.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Once you get the grunt prompt, you can type the Pig Latin statement as shown below.

```
grunt> A = load '/pigdemo/student.tsv' as (rollno, name, gpa);
grunt> DUMP A;
```

Here, the path refers to HDFS path and DUMP displays the result on the console as shown below.

```
(1001,John,3.0)
(1002,Jack,4.0)
(1003,Smith,4.5)
(1004,Scott,4.2)
(1005,Joshi,3.5)
grunt>
```

10.8.2 Batch Mode

You need to create **"Pig Script"** to run pig in batch mode. Write Pig Latin statements in a file and save it with **.pig** extension.

10.9 EXECUTION MODES OF PIG

You can execute pig in two modes:

1. Local Mode.
2. MapReduce Mode.

10.9.1 Local Mode

To run pig in local mode, you need to have your files in the local file system.

Syntax:

pig –x local filename

10.9.2 MapReduce Mode

To run pig in MapReduce mode, you need to have access to a Hadoop Cluster to read /write file. This is the default mode of Pig.

Syntax:

pig filename

10.10 HDFS COMMANDS

You can work with all HDFS commands in Grunt shell. For example, you can create a directory as shown below.

```
grunt> fs -mkdir /piglatindemos;
grunt>
```

The sections have been designed as follows:

Objective: What is it that we are trying to achieve here?
Input: What is the input that has been given to us to act upon?
Act: The actual statement/command to accomplish the task at hand.
Outcome: The result/output as a consequence of executing the statement.

10.11 RELATIONAL OPERATORS

10.11.1 FILTER

FILTER operator is used to select tuples from a relation based on specified conditions.

Objective: Find the tuples of those student where the GPA is greater than 4.0.
Input:
Student (rollno:int,name:chararray,gpa:float)
Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = filter A by gpa > 4.0;
DUMP B;
```

Output:

```
(1003,Smith,4.5)
(1004,Scott,4.2)
[root@volgalnx010 pigdemos]#
```

10.11.2 FOREACH

Use ***FOREACH*** when you want to do data transformation based on columns of data.

Objective: Display the name of all students in uppercase.
Input:
Student (rollno:int,name:chararray,gpa:float)
Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = foreach A generate UPPER (name);
DUMP B;
```

Output:

```
(JOHN)
(JACK)
(SMITH)
(SCOTT)
(JOSHI)
[root@volgalnx010 pigdemos]#
```

10.11.3 GROUP

GROUP operator is used to group data.

Objective: Group tuples of students based on their GPA.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = GROUP A BY gpa;
DUMP B;
```

Output:

```
(3.0,{(1001,John,3.0),(1001,John,3.0)})
(3.5,{(1005,Joshi,3.5),(1005,Joshi,3.5)})
(4.0,{(1008,James,4.0),(1002,Jack,4.0)})
(4.2,{(1007,David,4.2),(1004,Scott,4.2)})
(4.5,{(1006,Alex,4.5),(1003,Smith,4.5)})
[root@volgalnx010 pigdemos]#
```

10.11.4 DISTINCT

DISTINCT operator is used to remove duplicate tuples. In Pig, DISTINCT operator works on the entire tuple and NOT on individual fields.

Objective: To remove duplicate tuples of students.

Input:

Student (rollno:int,name:chararray,gpa:float)

Input:

1001	John	3.0
1002	Jack	4.0
1003	Smith	4.5
1004	Scott	4.2
1005	Joshi	3.5
1006	Alex	4.5
1007	David	4.2
1008	James	4.0
1001	John	3.0
1005	Joshi	3.5

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = DISTINCT A;
DUMP B;
```

Output:

```
(1001,John,3.0)
(1002,Jack,4.0)
(1003,Smith,4.5)
(1004,Scott,4.2)
(1005,Joshi,3.5)
(1006,Alex,4.5)
(1007,David,4.2)
(1008,James,4.0)
[root@volgalnx010 pigdemos]#
```

10.11.5 LIMIT

LIMIT operator is used to limit the number of output tuples.

Objective: Display the first 3 tuples from the "student" relation.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = LIMIT A 3;
DUMP B;
```

Output:

```
(1001,John,3.0)
(1002,Jack,4.0)
(1003,Smith,4.5)
[root@volgalnx010 pigdemos]#
```

10.11.6 ORDER BY

ORDER BY is used to sort a relation based on specific value.

Objective: Display the names of the students in Ascending Order.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = ORDER A BY name;
DUMP B;
```

Output:

```
(1006,Alex,4.5)
(1007,David,4.2)
(1002,Jack,4.0)
(1008,James,4.0)
(1001,John,3.0)
(1001,John,3.0)
(1005,Joshi,3.5)
(1005,Joshi,3.5)
(1004,Scott,4.2)
(1003,Smith,4.5)
[root@volgalnx010 pigdemos]#
```

10.11.7 JOIN

It is used to join two or more relations based on values in the common field. It always performs inner Join.

Objective: To join two relations namely, "student" and "department" based on the values contained in the "rollno" column.

Input:

Student (rollno:int,name:chararray,gpa:float)

Department(rollno:int,deptno:int,deptname:chararray)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = load '/pigdemo/department.tsv' as (rollno:int, deptno:int,deptname:chararray);
C = JOIN A BY rollno, B BY rollno;
DUMP C;
DUMP B;
```

Output:

```
(1001,John,3.0,1001,101,B.E.)
(1001,John,3.0,1001,101,B.E.)
(1002,Jack,4.0,1002,102,B.Tech)
(1003,Smith,4.5,1003,103,M.Tech)
(1004,Scott,4.2,1004,104,MCA)
(1005,Joshi,3.5,1005,105,MBA)
(1005,Joshi,3.5,1005,105,MBA)
(1006,Alex,4.5,1006,101,B.E)
(1007,David,4.2,1007,104,MCA)
(1008,James,4.0,1008,102,B.Tech)
[root@volgalnx010 pigdemos]#
```

10.11.8 UNION

It is used to merge the contents of two relations.

Objective: To merge the contents of two relations "student" and "department".

Input:

Student (rollno:int,name:chararray,gpa:float)

Department(rollno:int,deptno:int,deptname:chararray)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno, name, gp);
B = load '/pigdemo/department.tsv' as (rollno, deptno,deptname);
C = UNION A,B;
STORE C INTO '/pigdemo/uniondemo';
DUMP B;
```

Output:

"Store" is used to save the output to a specified path. The output is stored in two files: part-m-00000 contains "student" content and part-m-00001 contains "department" content.

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 B	3	128 MB	2015-02-24 17:23	rw-r--r--	root	supergroup
part-m-00000	file	146 B	3	128 MB	2015-02-24 17:23	rw-r--r--	root	supergroup
part-m-00001	file	114 B	3	128 MB	2015-02-24 17:23	rw-r--r--	root	supergroup

File: /pigdemo/uniondemo/part-m-00000

Goto : /pigdemo/uniondemo go

Go back to dir listing
Advanced view/download options

```
1001    John    3.0
1002    Jack    4.0
1003    Smith   4.5
1004    Scott   4.2
1005    Joshi   3.5
1006    Alex    4.5
1007    David   4.2
1008    James   4.0
1001    John    3.0
1005    Joshi   3.5
```

File: /pigdemo/uniondemo/part-m-00001

Goto : /pigdemo/uniondemo go

Go back to dir listing
Advanced view/download options

```
1001    101     B.E.
1002    102     B.Tech
1003    103     M.Tech
1004    104     MCA
1005    105     MBA
1006    101     B.E
1007    104     MCA
1008    102     B.Tech
```

10.11.9 SPLIT

It is used to partition a relation into two or more relations.

Objective: To partition a relation based on the GPAs acquired by the students.

- GPA = 4.0, place it into relation X.
- GPA is < 4.0, place it into relation Y.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
SPLIT A INTO X IF gpa==4.0, Y IF gpa<=4.0;
DUMP X;
```

Output: Relation X

```
(1002,Jack,4.0)
(1008,James,4.0)
[root@volgalnx010 pigdemos]#
```

Output: Relation Y

```
(1001,John,3.0)
(1002,Jack,4.0)
(1005,Joshi,3.5)
(1008,James,4.0)
(1001,John,3.0)
(1005,Joshi,3.5)
[root@volgalnx010 pigdemos]#
```

10.11.10 SAMPLE

It is used to select random sample of data based on the specified sample size.

Objective: To depict the use of ***SAMPLE***.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = SAMPLE A 0.01;
DUMP B;
```

10.12 EVAL FUNCTION

10.12.1 AVG

AVG is used to compute the average of numeric values in a single column bag.

Objective: To calculate the average marks for each student.

Input:

Student (studname:chararray,marks:int)

Act:

```
A = load '/pigdemo/student.csv' USING PigStorage (',') as (studname:chararray,marks:int);
B = GROUP A BY studname;
C = FOREACH B GENERATE A.studname, AVG(A.marks);
DUMP C;
```

Output:

```
({(Jack),(Jack),(Jack),(Jack)},39.75)
({(John),(John),(John),(John)},39.0)
[root@volgalnx010 pigdemos]#
```

Note: You need to use PigStorage function if you wish to manipulate files other than .tsv.

10.12.2 MAX

MAX is used to compute the maximum of numeric values in a single column bag.

Objective: To calculate the maximum marks for each student.

Input:

Student (studname:chararray,marks:int)

Act:

```
A = load '/pigdemo/student.csv' USING PigStorage (',') as (studname:chararray, marks:int);
B = GROUP A BY studname;
C = FOREACH B GENERATE A.studname, MAX(A.marks);
DUMP C;
```

Output:

```
({(Jack),(Jack),(Jack),(Jack)},46)
({(John),(John),(John),(John)},45)
[root@volgalnx010 pigdemos]#
```

Note: Similarly, you can try the MIN and the SUM functions as well.

10.12.3 COUNT

COUNT is used to count the number of elements in a bag.

Objective: To count the number of tuples in a bag.
Input:
Student (studname:chararray,marks:int)
Act:

```
A = load '/pigdemo/student.csv' USING PigStorage (',') as (studname:chararray, marks:int);
B = GROUP A BY studname;
C = FOREACH B GENERATE  A. studname,COUNT(A);
DUMP C;
```

Output:

```
({(Jack),(Jack),(Jack),(Jack)},4)
({(John),(John),(John),(John)},4)
[root@volgalnx010 pigdemos]#
```

Note: The default file format of Pig is .tsv file. Use PigStorage() to manipulate files other than .tsv file.

10.13 COMPLEX DATA TYPES

10.13.1 TUPLE

A *TUPLE* is an ordered collection of fields.

Objective: To use the complex data type "Tuple" to load data.
Input:

```
(John,12)      (Jack,13)
(James,7)      (Joseph,5)
(Smith,8)      (Scott,12)
```

Act:

```
A = LOAD '/root/pigdemos/studentdata.tsv' AS (t1:tuple(t1a:chararray,
t1b:int),t2:tuple(t2a:chararray,t2b:int));
B = FOREACH A GENERATE t1.t1a, t1.t1b,t2.$0,t2.$1;
DUMP B;
```

Output:

```
(John,12,Jack,13)
(James,7,Joseph,5)
(Smith,8,Scott,12)
[root@volgalnx010 pigdemos]#
```

Note: You can refer to the field using Positional Notation as shown above. The Positional Notation is denoted by $ sign and the position starts with 0 (e.g., $0).

10.13.2 MAP

MAP represents a key/value pair.

Objective: To depict the complex data type "map".

Input:

```
John    [city#Bangalore]
Jack    [city#Pune]
James   [city#Chennai]
```

Act:

```
A = load '/root/pigdemos/studentcity.tsv' Using PigStorage as
(studname:chararray,m:map[chararray]);
B = foreach A generate m#'city' as CityName:chararray;
DUMP B
```

Output:

```
(Bangalore)
(Pune)
(Chennai)
[root@volgalnx010 pigdemos]#
```

10.14 PIGGY BANK

Pig user can use Piggy Bank functions in Pig Latin script and they can also share their functions in Piggy Bank.

Objective: To use Piggy Bank string UPPER function.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
register '/root/pigdemos/piggybank-0.12.0.jar';
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
upper = foreach A generate
org.apache.pig.piggybank.evaluation.string.UPPER(name);
DUMP upper;
```

Output:

```
(JOHN)
(JACK)
(SMITH)
(SCOTT)
(JOSHI)
(ALEX)
(DAVID)
(JAMES)
(JOHN)
(JOSHI)
[root@volgalnx010 pigdemos]#
```

Note: You need to use the "**register**" keyword to use Piggy Bank jar function in your pig script.

10.15 USER-DEFINED FUNCTIONS (UDF)

Pig allows you to create your own function for complex analysis.

Objective: To depict user-defined function.

Java Code to convert name into uppercase:

```
package myudfs;
import java.io.IOException;
import org.apache.pig.EvalFunc;
import org.apache.pig.data.Tuple;
import org.apache.pig.impl.util.WrappedIOException;
public class UPPER extends EvalFunc<String>
  {
    public String exec(Tuple input) throws IOException {
      if (input == null || input.size() == 0)
        return null;
        try{
          String str = (String)input.get(0);
          return str.toUpperCase();
        }catch(Exception e){
          throw WrappedIOException.wrap ("Caught exception processing input row ", e);
      }
    }
  }
```

Note: Convert above java class into jar to include this function into your code.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
register /root/pigdemos/myudfs.jar;
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
B = FOREACH A GENERATE myudfs.UPPER(name);
DUMP B;
```

Output:

```
(JOHN)
(JACK)
(SMITH)
(SCOTT)
(JOSHI)
(ALEX)
(DAVID)
(JAMES)
(JOHN)
(JOSHI)
[root@volgalnx010 pigdemos]#
```

10.16 PARAMETER SUBSTITUTION

Pig allows you to pass parameters at runtime.

Objective: To depict parameter substitution.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '$student' as (rollno:int, name:chararray, gpa:float);
DUMP A;
```

Execute:

```
pig –param student=/pigdemo/student.tsv parameterdemo.pig
```

Output:

```
(1001,John,3.0)
(1002,Jack,4.0)
(1003,Smith,4.5)
(1004,Scott,4.2)
(1005,Joshi,3.5)
(1006,Alex,4.5)
(1007,David,4.2)
(1008,James,4.0)
(1001,John,3.0)
(1005,Joshi,3.5)
[root@volgalnx010 pigdemos]#
```

10.17 DIAGNOSTIC OPERATOR

It returns the schema of a relation.

Objective: To depict the use of ***DESCRIBE***.

Input:

Student (rollno:int,name:chararray,gpa:float)

Act:

```
A = load '/pigdemo/student.tsv' as (rollno:int, name:chararray, gpa:float);
DESCRIBE A;
```

Output:

```
A: {rollno: int,name: chararray,gpa: float}
[root@volgalnx010 pigdemos]#
```

10.18 WORD COUNT EXAMPLE USING PIG

Objective: To count the occurrence of similar words in a file.

Input:

```
Welcome to Hadoop Session
Introduction to Hadoop
Introducing Hive
Hive Session
Pig Session
```

Act:

```
lines = LOAD '/root/pigdemos/lines.txt' AS (line:chararray);
words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
grouped = GROUP words BY word;
wordcount = FOREACH grouped GENERATE group, COUNT(words);
DUMP wordcount;
```

Output:

```
(to,2)
(Pig,1)
(Hive,2)
(Hadoop,2)
(Session,3)
(Welcome,1)
(Introducing,1)
(Introduction,1)
[root@volgalnx010 pigdemos]#
```

Note:

TOKENIZE splits the line into a field for each word.

FLATTEN will take the collection of records returned by TOKENIZE and produce a separate record for each one, calling the single field in the record word.

10.19 WHEN TO USE PIG?

Pig can be used in the following situations:

1. When your data loads are time sensitive.
2. When you want to process various data sources.
3. When you want to get analytical insights through sampling.

10.20 WHEN NOT TO USE PIG?

Pig should not be used in the following situations:

1. When your data is completely in the unstructured form such as video, text, and audio.
2. When there is a time constraint because Pig is slower than MapReduce jobs.

10.21 PIG AT YAHOO!

Yahoo uses Pig for two things:

1. **In Pipelines**, to fetch log data from its web servers and to perform cleansing to remove companies interval views and clicks.
2. **In Research**, script is used to test a theory. Pig provides facility to integrate Perl or Python script which can be executed on a huge dataset.

10.22 PIG versus HIVE

Features	Pig	Hive
Used By	Programmers and Researchers	Analyst
Used For	Programming	Reporting
Language	Procedural data flow language	SQL Like
Suitable For	Semi - Structured	Structured
Schema/Types	Explicit	Implicit
UDF Support	YES	YES
Join/Order/Sort	YES	YES
DFS Direct Access	YES (Implicit)	YES (Explicit)
Web Interface	YES	NO
Partitions	YES	NO
Shell	YES	YES

REMIND ME

- Apache Pig is a platform for data analysis. It is an alternative to MapReduce Programming.
- It provides an **engine** for executing **data flows** (how your data should flow). Pig processes data in parallel on the Hadoop cluster.
- It provides a language called "**Pig Latin**" to express data flows.
- The main components of Pig are as follows:
  - Data flow language (**Pig Latin**).
  - Interactive shell where you can type Pig Latin statements (**Grunt**).
  - Pig interpreter and execution engine.
- You can run Pig in two ways:
  - Interactive Mode.
  - Batch Mode.

POINT ME (BOOK)

- Programming Pig, Alan Gates, O'REILY.

CONNECT ME (INTERNET RESOURCES)

- http://pig.apache.org/docs/r0.12.0/index.html
- http://www.edureka.co/blog/introduction-to-pig/
- http://www.edureka.co/blog/pig-vs-hive/

TEST ME

A. Fill Me

1. Pig is a _____ language.
2. In Pig, _____ is used to specify data flow.
3. Pig provides an _____ to execute data flow.
4. _____, _____ are execution modes of Pig.
5. The interactive mode of Pig is _____.
6. _____ and _____ are case sensitive in Pig.
7. _____, _____, _____ are Complex Data Types of Pig.
8. Pig is used in _____ process.

Answers:

1. Scripting
2. Pig Latin
3. Pig Engine
4. Local Mode, MapReduce Mode
5. Grunt
6. Fields and Aliases
7. Bag, Tuple, Map
8. ETL

B. Match Me

Column A	Column B
Map	Hadoop Cluster
Bag	An Ordered Collection of Fields
Local Mode	Collection of Tuples
Tuple	Key/Value Pair
MapReduce Mode	Local File System

Answers:

Column A	Column B
Map	Key/Value Pair
Bag	Collection of Tuples
Local Mode	Local File System
Tuple	An Ordered Collection of Fields
MapReduce Mode	Hadoop Cluster

C. True or False

1. PigStorage() function is case sensitive.
2. Local Mode is the default mode of Pig.
3. DISTINCT Keyword removes duplicate fields.
4. LIMIT keyword is used to display limited number of tuples in Pig.
5. ORDER BY is used for sorting.

Answers:

1. True
2. False
3. False
4. True
5. True

ASSIGNMENTS FOR HANDS-ON PRACTICE

ASSIGNMENT 1: SPLIT

Objective: To learn about SPLIT relational operator.

Problem Description:
Write a Pig Script to split customers for reward program based on their life time values.

Input:

Customers	Life Time Value
Jack	25000
Smith	8000
David	35000
John	15000
Scott	10000
Joshi	28000
Ajay	12000
Vinay	30000
Joseph	21000

- If Life Time Value is >1000 and <= 2000 → Sliver Program.
- If Life Time Value is >20000 → Gold Program.

ASSIGNMENT 2: GROUP

Objective: To learn about GROUP relational operator.

Problem Description:

Create a data file for below schemas:

- **Order:** CustomerId, ItemId, ItemName, OrderDate, DeliveryDate
- **Customer:** CustomerId, CustomerName, Address, City, State, Country

1. Load Order and Customer Data.
2. Write a Pig Latin Script to determine number of items bought by each customer.

ASSIGNMENT 3: COMPLEX DATA TYPE – BAG

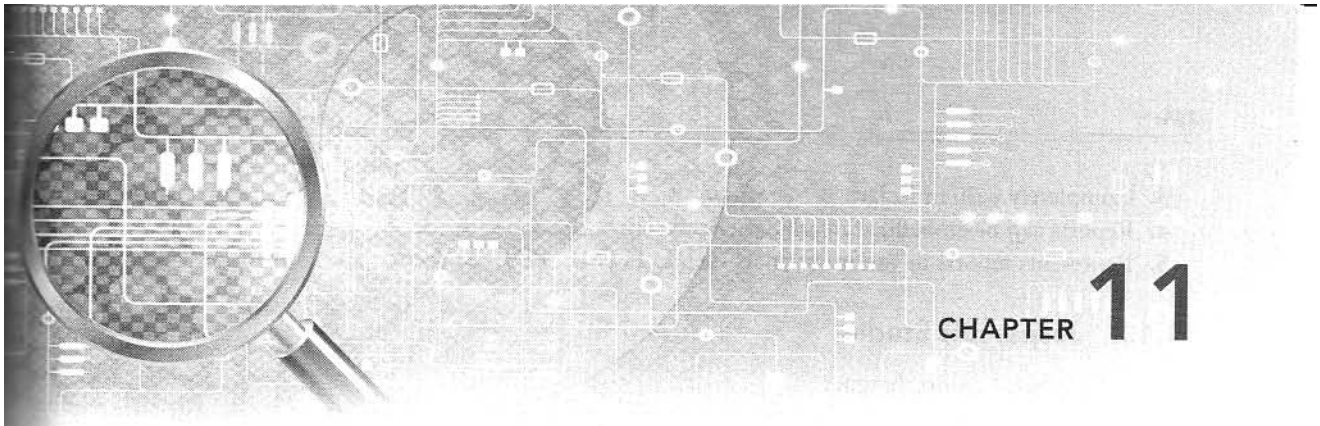
Objective: To learn complex data type – bag in Pig.

Problem Description:

1. Create a file which contains bag dataset as shown below.

User ID	From	To
user1001	user1001@sample.com	{(user003@sample.com),(user004@sample.com), (user006@sample.com)}
user1002	user1002@sample.com	{(user005@sample.com), (user006@sample.com)}
user1003	user1003@sample.com	{(user001@sample.com),(user005@sample.com)}

2. Write a Pig Latin statement to display the names of all users who have sent emails and also a list of all the people that they have sent the email to.
3. Store the result in a file.



CHAPTER 11

JasperReport using Jaspersoft

BRIEF CONTENTS

- What's in Store?
- Introduction to JasperReports
  - JasperReports
  - Jaspersoft Studio
- Connecting to MongoDB NoSQL Database
  - Syntax of Few MongoDB Query Language
  - Elements and Attributes
  - Creating Variables
  - Creating Report Parameters
- Connecting to Cassandra NoSQL Database

"The greatest value of a picture is when it forces us to notice what we never expected to see."

– John Tukey, American Mathematician

WHAT'S IN STORE?

We assume that you are familiar with the MongoDB and Cassandra NoSQL databases. The focus of this chapter will be to build on this knowledge to create JasperReports using Jaspersoft Studio. We will discuss few MongoDB Queries to retrieve data from MongoDB and place it on the report.

We suggest you refer to some of the learning resources provided at the end of this chapter for better learning.

11.1 INTRODUCTION TO JASPERREPORTS

11.1.1 JasperReports

1. Open-source reporting engine.
2. Helps to create page-oriented reports.

3. Completely written in Java.
4. Reports can be embedded in any Java Application.
5. Represents reports in various formats such as HTML, PDF, CSV, etc.

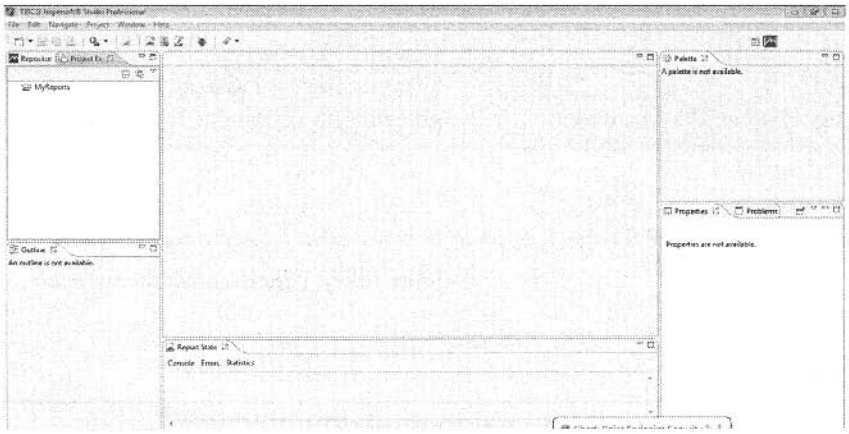
11.1.2 Jaspersoft Studio

1. Eclipse based report designer.
2. Available in two flavors: Eclipse plugin and as a standalone application.
3. Data can be accessed from multiple sources such as JDBC, XML, CSV, etc.
4. Supports big data components such as MongoDB, Cassandra, Hive, etc.
5. Provides sophisticated layouts such as charts, crosstabs, images, subreports, etc.

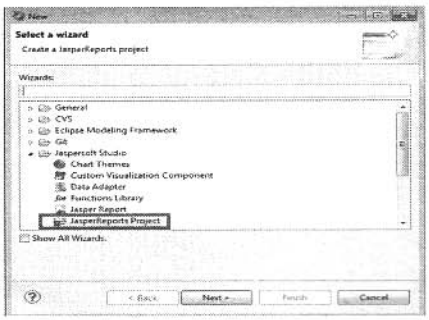
11.2 CONNECTING TO MONGODB NoSQL DATABASE

Jaspersoft uses **Jaspersoft MongoDB Query Language** to query the MongoDB databases. Jaspersoft MongoDB Query Language is a declarative language. It is used for stating which data to retrieve from which database. Connector converts this MongoDB query into the appropriate API calls. Then it makes use of MongoDB Java connector to query the MongoDB instance.

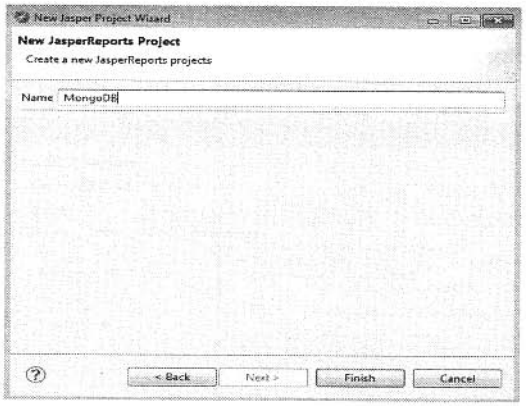
1. Double click on Jaspersoft Studio icon. You can see Jaspersoft Studio IDE as shown below.



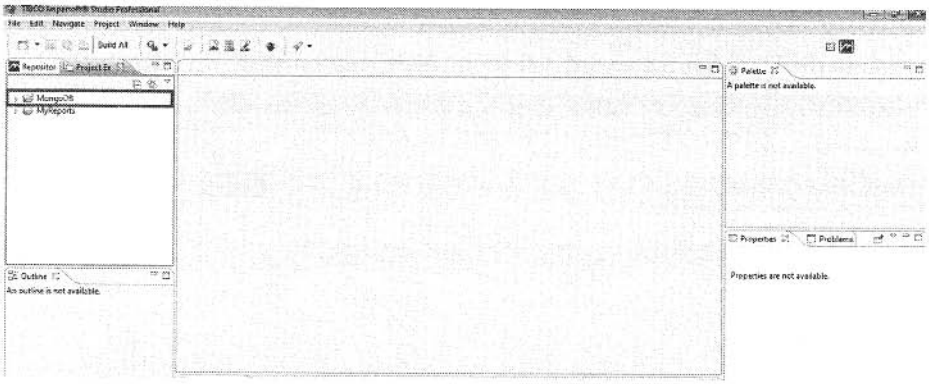
2. Select **File** → **New** → **Other** to create JasperReports Project. From the **New** wizard select **"JasperReports Project"** and click **"Next"**.



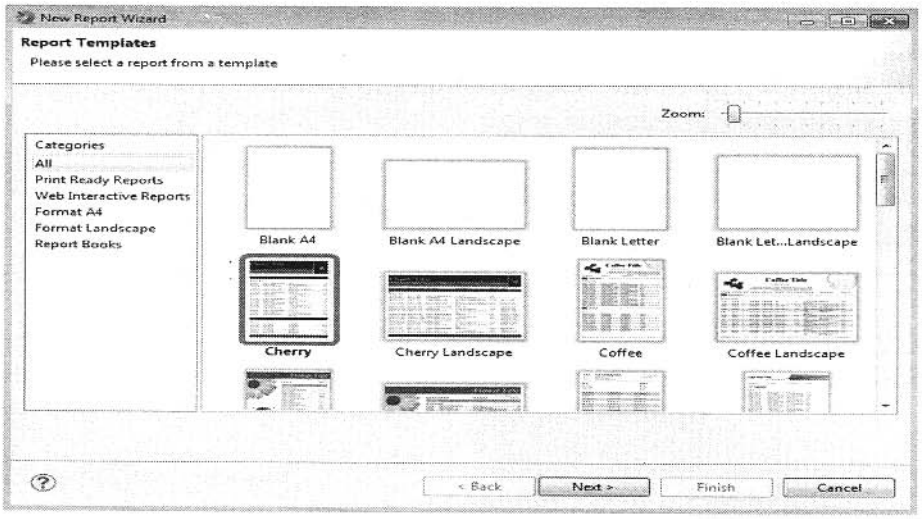
3. Mention project name as **"MongoDB"** and Click **"Finish"**.



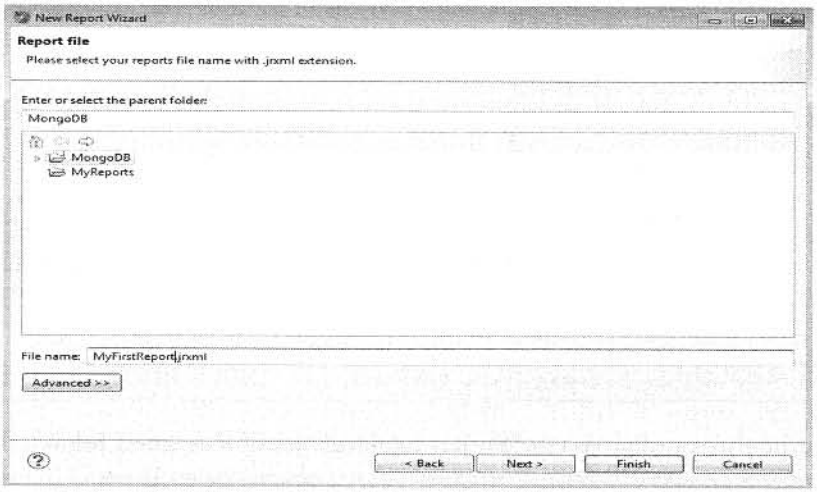
4. Now, you can see the newly created project in **"Project Explorer"** window as shown below.



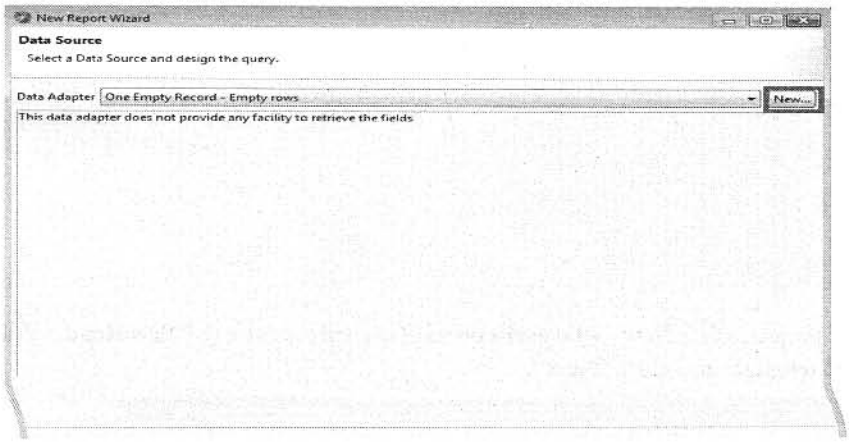
5. Right Click on the project, select **New → JasperReports**. This will take you to **"New Report Wizard"**. Select the required template and click **"Next"**.



6. Then specify the file name as shown below and click on **"Next"** button.

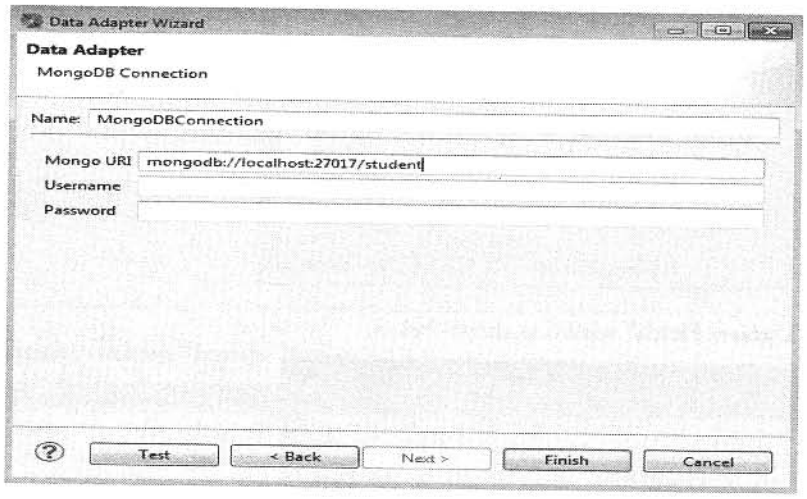


7. It will take you to the **Data Source** wizard. Click on **"New"** button as shown below.

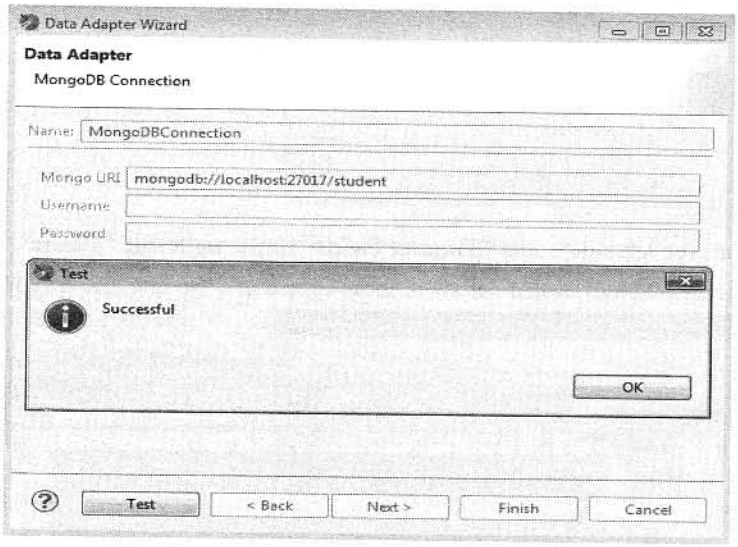


8. In the **"Data Adapter Wizard"**, select **"MongoDB Connection"** as shown below.

9. Specify the details as given below:
   Name: MongoDBConnection
   Mongo URI: mongodb://localhost:27017/student
   Note: Here, "student" is the name of the database.

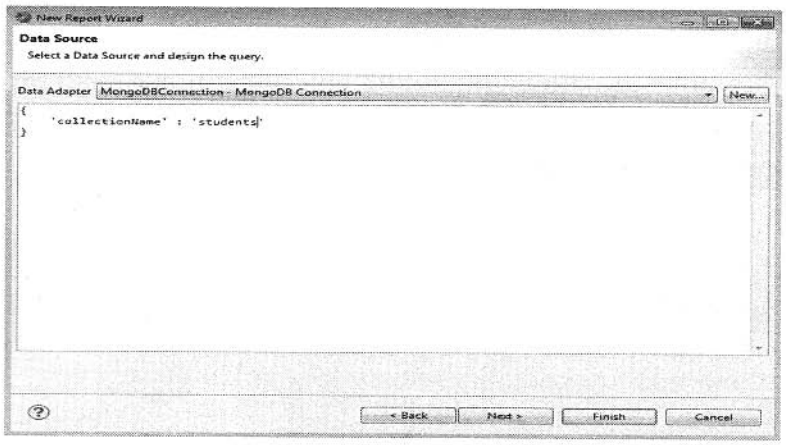


10. Click on **"Test"** button to test the connection. If the connection is properly set, you will get the message **"Successful"**.

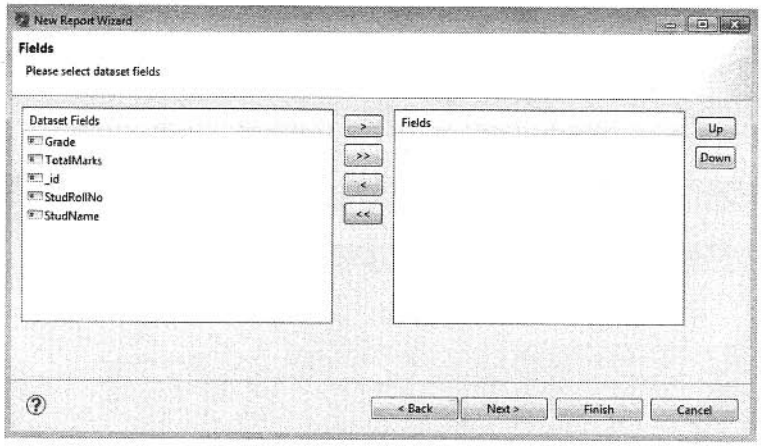


11. Let us write a MongoDB Query to retrieve data from the MongoDB database and click on the **"Next"** button. The bare minimal syntax (MongoDB Query Language) to retrieve all fields from the collection, "students" (the collection exists in the database, "student") is

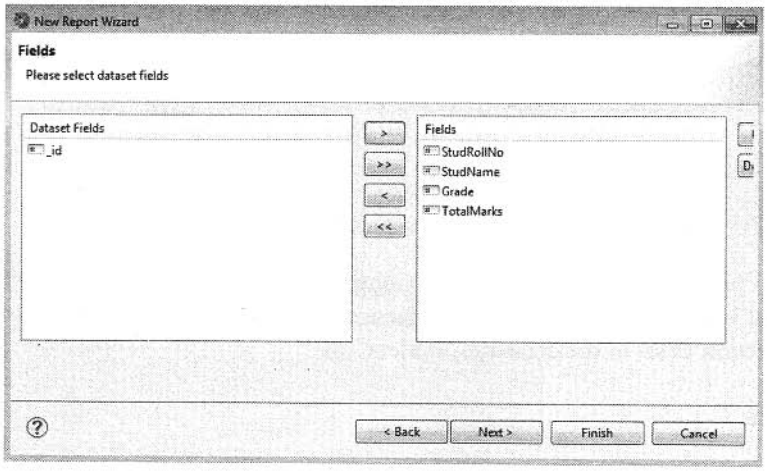
```
{
'collectionName': 'students'
}
```



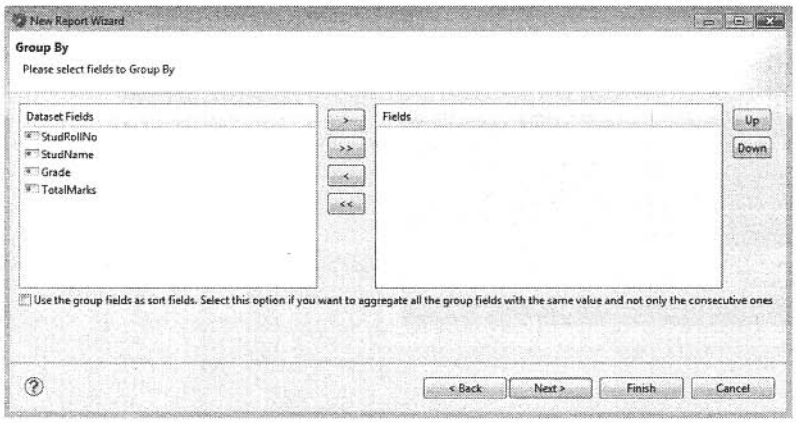
12. This will lead to the **"Dataset Fields"** wizard as shown below.



13. Select the required fields from the left side under **"Dataset Fields"** to the right side **"Fields"**. Click on **"Next"** button.

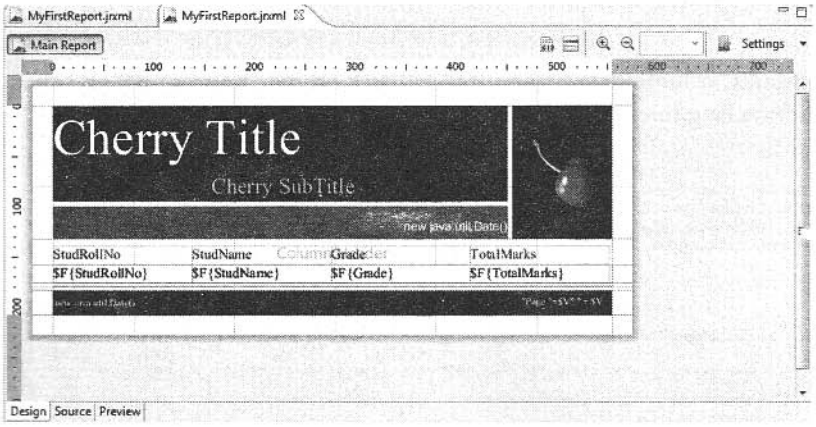


14. Ignore the "Group By" option and click on **"Next"**.

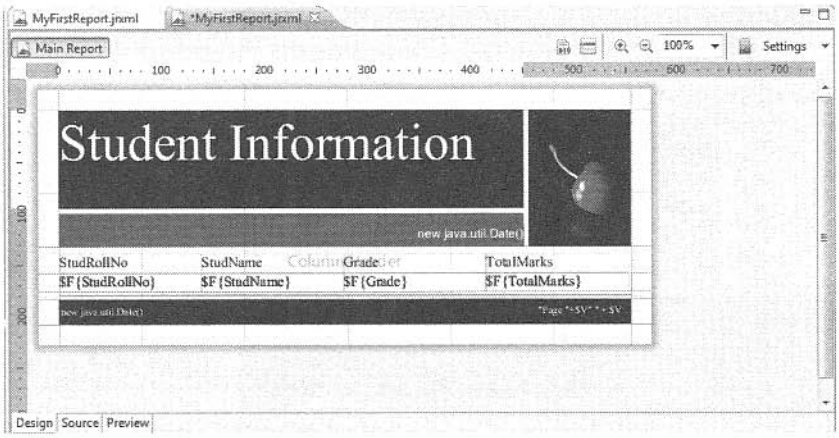


15. Finally, click on **"Finish"** button.

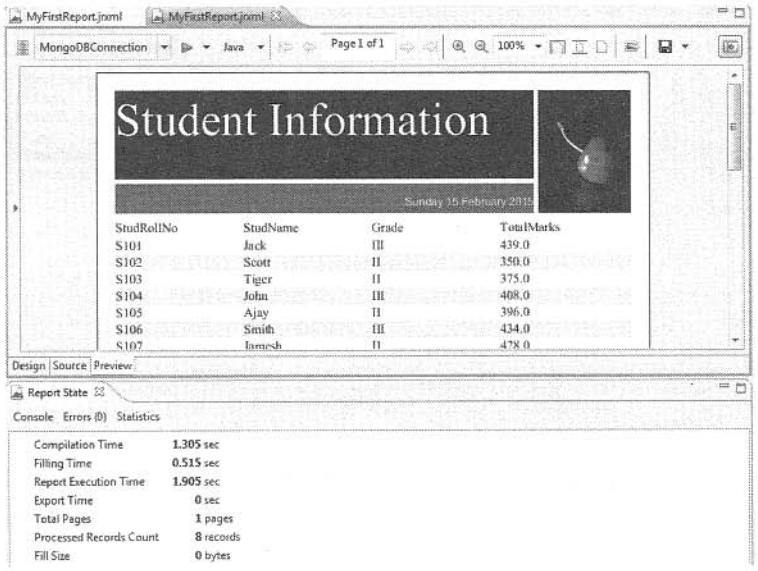
16. Your report will open in design mode as shown below. Double click on the text to change the title and delete the subtitle.



17. Specify the title name as "Student Information" as shown below.



18. Click on the **"Preview"** tab to preview the report.

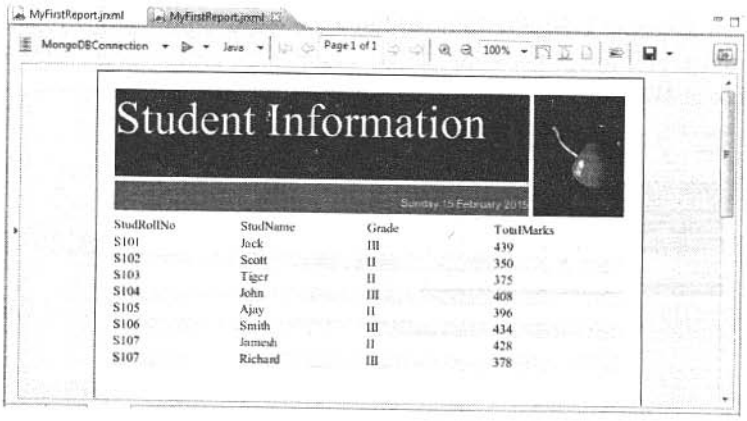


19. To change the "TotalMarks" column data type to Integer, click on the **"Source"** tab and change the "TotalMarks" class to "java.lang.Integer" as shown below.

```
19 }]]>
20     </queryString>
21     <field name="StudRollNo" class="java.lang.String"/>
22     <field name="StudName" class="java.lang.String"/>
23     <field name="Grade" class="java.lang.String"/>
24     <field name="TotalMarks" class="java.lang.Integer"/>
25     <background>
26         <band splitType="Stretch"/>
27     </background>
28     <title>
29         <band height="132" splitType="Stretch">
30             <image>
31                 <reportElement x="456" y="0" width="99" height="132" uuid="5514eb68-2f0b-4493-a789-9ac0d1
```

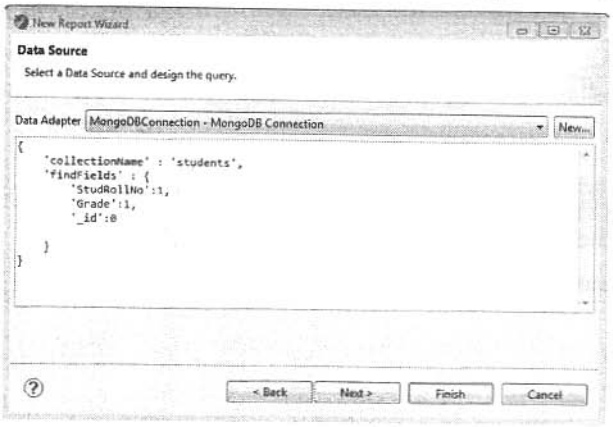
Design Source Preview

20. Click on the **"Preview"** tab to view the result.

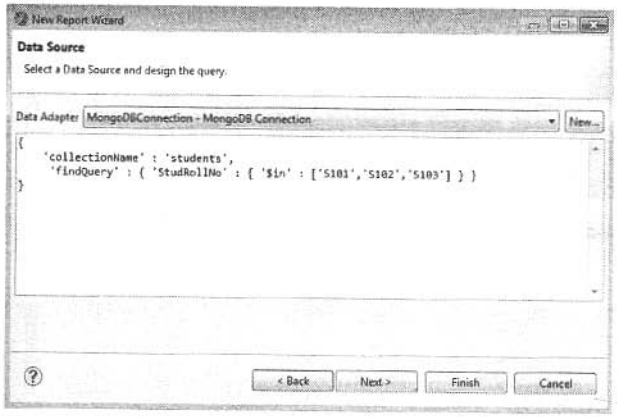


11.2.1 Syntax of Few MongoDB Query Language

1. To return only StudRollNo, Grade and suppress the _id field.

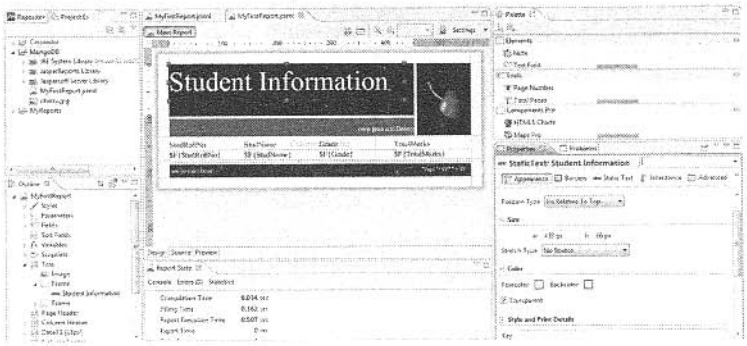


2. To select documents from a collection based on search criteria(s).



11.2.2 Elements and Attributes

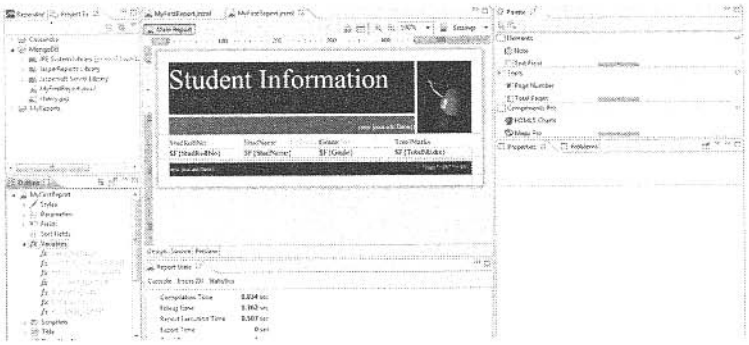
Attributes can be used to specify an element's behavior. The attributes are visible in the property tab. The elements such as title, column header, and image behavior can be specified by attributes available in the **"Properties Window" (Right Side of IDE)** as shown below.



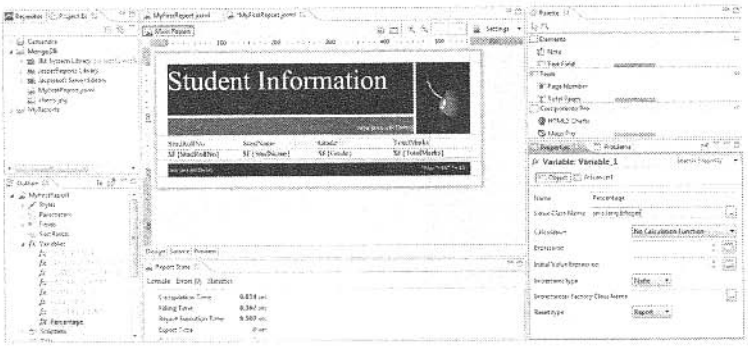
11.2.3 Creating Variables

Variables can be used to do complex calculations on the data extracted from the database. This can be stored and used later.

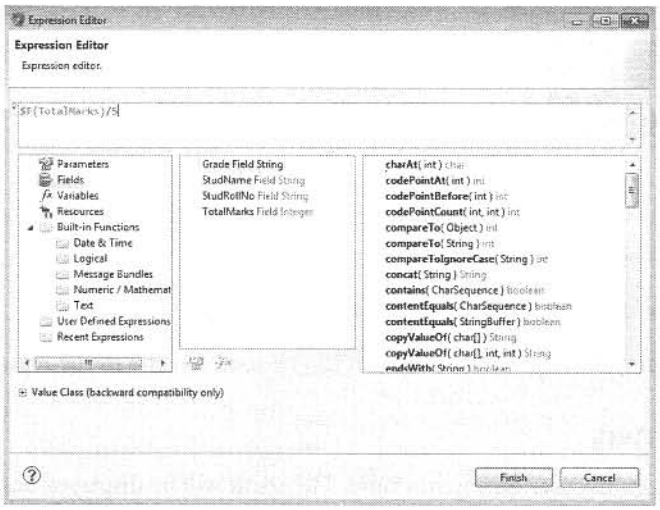
1. From the outline menu select **"Variables"** root node, Right click on it and select **"Create Variable"**.



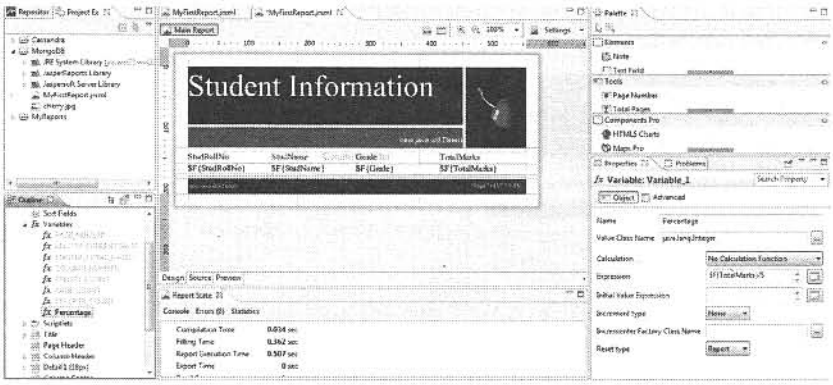
2. Once it is selected, its property will appear in the property window as shown below. Specify the variable name and the data type for the variable as shown below.



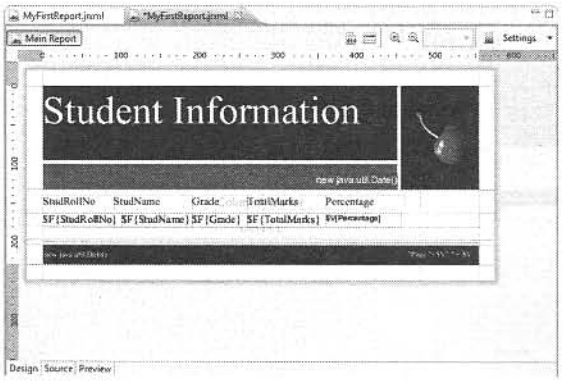
3. Click on the **"Expression"** tab to create an expression. Type an expression to calculate Percentage of Marks as shown below.



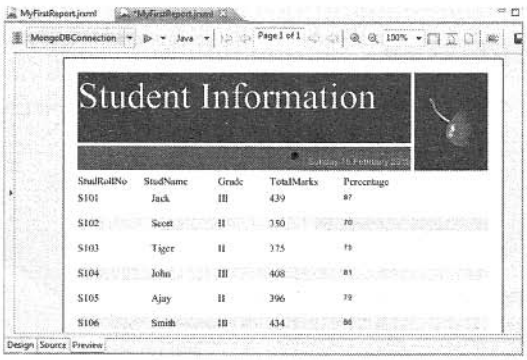
4. You can see the Percentage variable in outline window under the Variables item.



5. Now, drag and drop the variable into the detail band as shown below and add static label in the column header to display the "Percentage" column.



6. Click on the **"Preview"** tab to view the report.

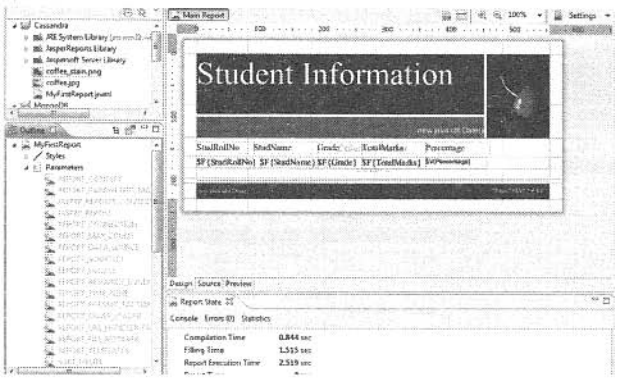


Note: Variable root node also contains built-in variables, which can be used directly in the report.

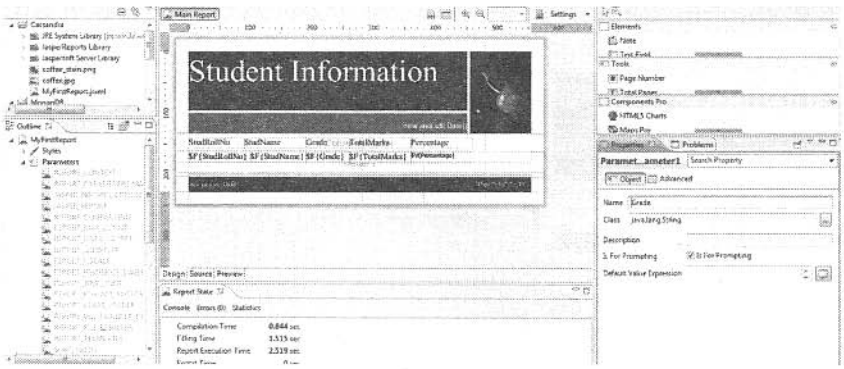
11.2.4 Creating Report Parameters

Report parameter is used to take input from the user during run time. The result will be displayed based on the provided input.

1. From the outline menu select the **"Parameters"** root node. Right click on it and select **"Create Parameter"**.



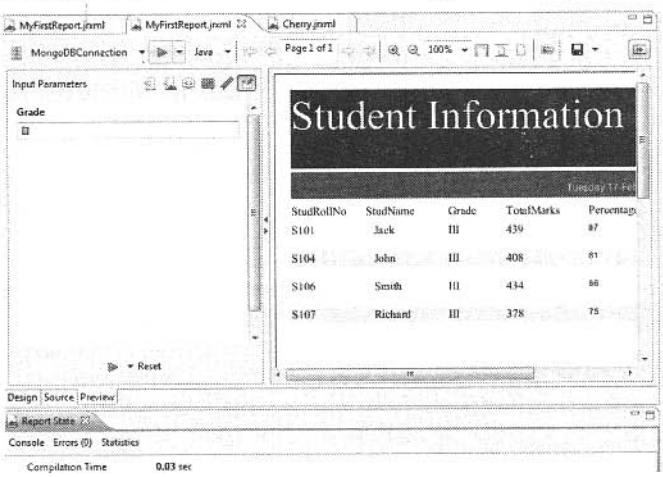
2. Once it is selected, its property will appear in the property window as shown below. Specify the name, data type, and description for the parameter and check **"Is For Prompting"** option.



3. Open the source tab and specify the parameter inside the queryString tag as shown below.

```
17    <queryString language="MongoDbQuery">
18        <![CDATA[{'collectionName' : 'students',
19                'findQuery' : { 'Grade' : $P{Grade} } }]]>
20    </queryString>
```

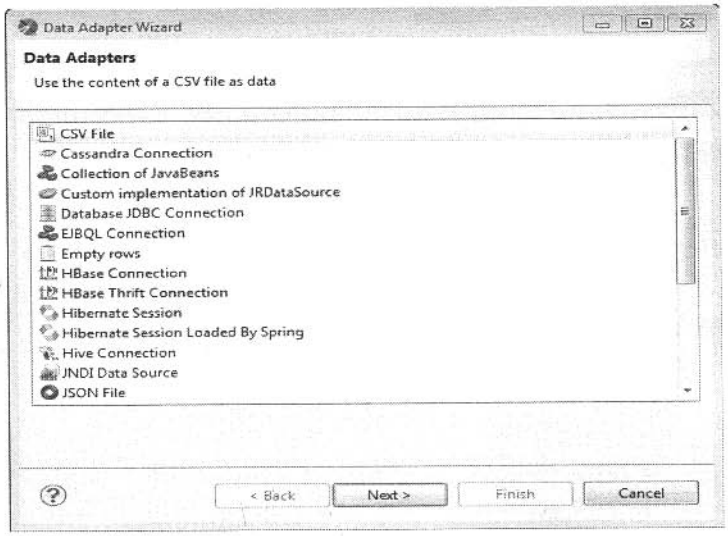
4. Click on the **"Preview"** tab. You will be prompted to enter the value for the "Grade" parameter as shown below. Click on the Run button (green arrow) to view the result.



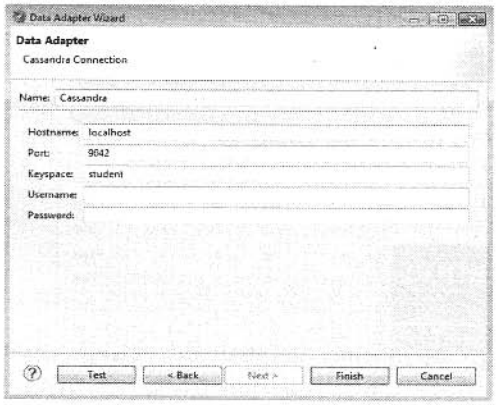
11.3 CONNECTING TO CASSANDRA NoSQL DATABASE

Jaspersoft studio uses Cassandra Query Language (CQL), which is similar to SQL, to retrieve data from Cassandra NoSQL database.

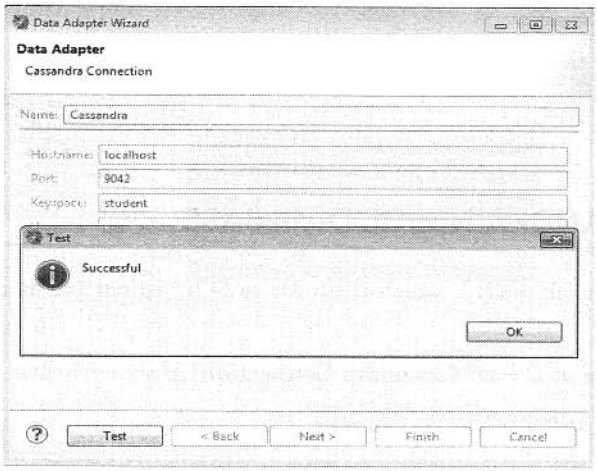
1. Create a JasperReports and state the data source as **"Cassandra Connection"** as shown below.



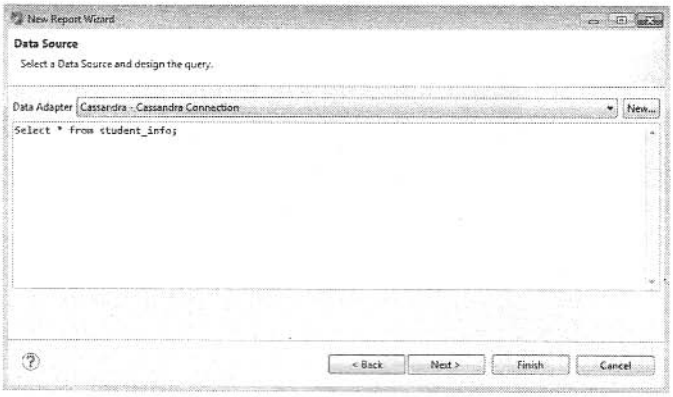
2. Click **"Next"** to get the Data Adapter Wizard. Mention the hostname, port, and keyspace as shown below.



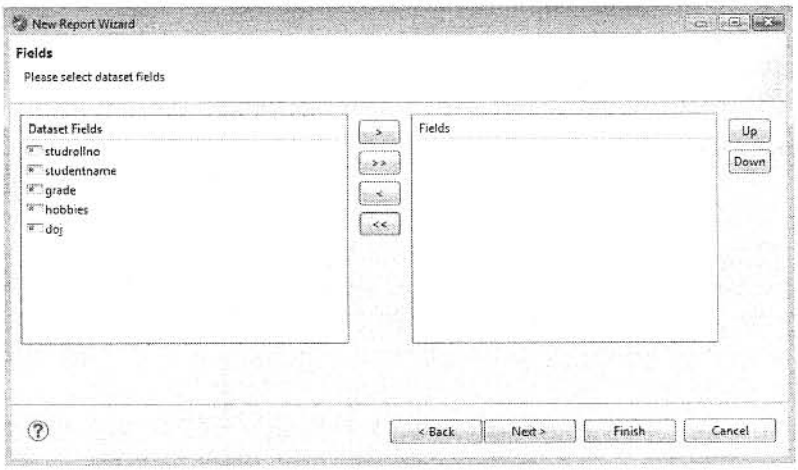
3. Click on the **"Test"** button to test the connection. If the connection is proper, you will get the **"Successful"** message as shown below.

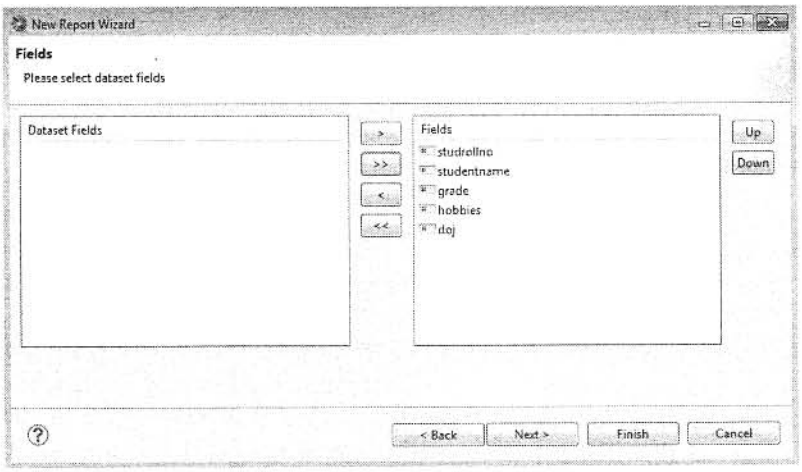


4. Write the CQL (Cassandra Query Language) query to retrieve data from the Cassandra database.



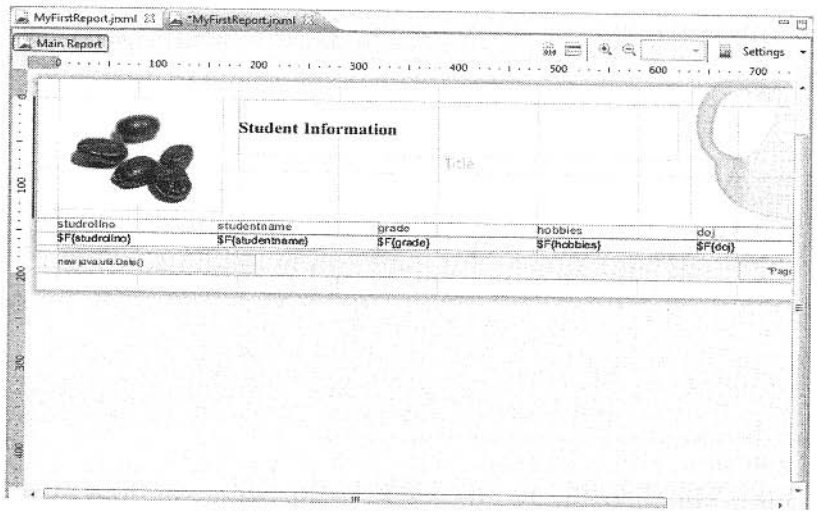
5. Select the required fields (under Dataset fields) for your report.

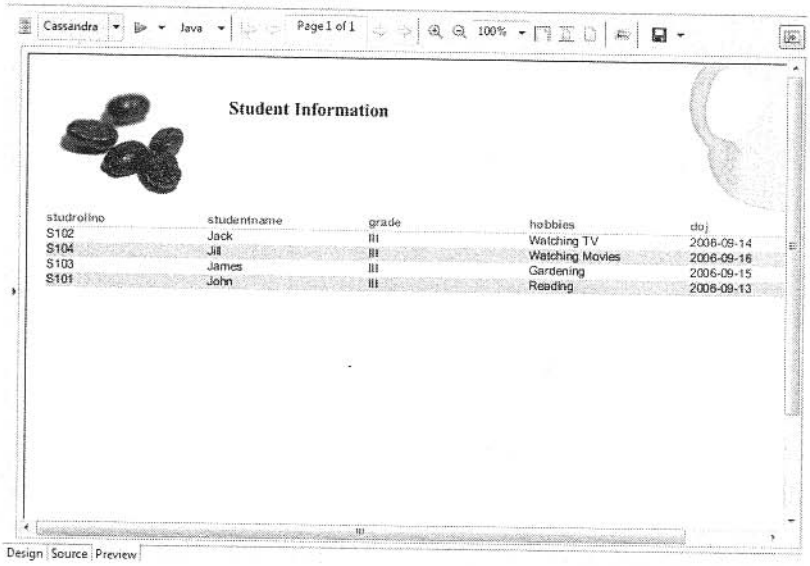




6. Click on **"Finish"** button to get the below message.

7. The report will open in design mode for editing. Create a report based on your requirement and view the report by clicking on the preview button.





REMIND ME

- Open-source reporting engine.
- Reports can be embedded in any Java Application.
- Jaspersoft studio is available in two flavors: Eclipse plugin and as a standalone application. Data can be accessed from multiple sources such as JDBC, XML, CSV, etc. Supports big data components such as MongoDB, Cassandra, Hive, etc.

POINT ME (BOOK)

- JasperReports for Java Developers by David R. Heffelfinger.

CONNECT ME (INTERNET RESOURCES)

- **https://community.jaspersoft.com/wiki/designing-report-jaspersoft-studio**
- **http://community.jaspersoft.com/wiki/jaspersoft-mongodb-query-language**

ASSIGNMENT FOR HANDS-ON PRACTICE

Background

You will be analyzing the scores and percentages of the trainees in various modules. You will be doing the analysis for scores and percentages in various assessments such as Test, Retest, Hands on, and/or Comprehensive Examination. Data for the assignment is available in JasperAssignment.accdb and TimeData.txt (see CD available with the book).

Database Section

Use either MongoDB or Cassandra.

Data for the assignment is provided on the CD (JasperAssignment.accdb and TimeDataForJasper Assignment.txt). Read the data from JasperAssignment.accdb into a .CSV or .TXT. Then from the .CSV or .TXT read the data into MongoDB or Cassandra.

Below are the table structures given in a standard RDBMS. Based on the requirements, create collections in MongoDB or tables in Cassandra.

Create a new database with name as 'JasperAssignment'. This database will include following tables. You are free to make names of table more meaningful.

1. Time (no need to create; load directly from the data provided on the CD).
2. Assessment (to be created).
3. Modules (to be created).
4. Trainees (to be created).
5. Score (to be created).

Table Details

Trainees Table:

Column Name	Data Type	Description
EmpKey	Int	Primary Key
EmpNumber	Int	Not Null
EmpName	Varchar(255)	Not Null

(*Continued*)

Column Name	Data Type	Description
BatchName	Varchar(50)	Not Null
Stream	Varchar(10)	Not Null
IBU	Varchar(4)	If Emp Number is less 100150 then assign SI else TRPU

Assessment Table:

Column Name	Data Type	Description
AssessmentKey	Int	Primary Key
AssessmentType	Varchar(100)	Not Null
DurationInMins	Int	Not Null

Modules Table:

Column Name	Data Type	Description
ModuleKey	Int	Primary Key
ModuleName	Varchar(100)	Not Null
ModuleCreditPoints	Int	Not Null

Score Table:

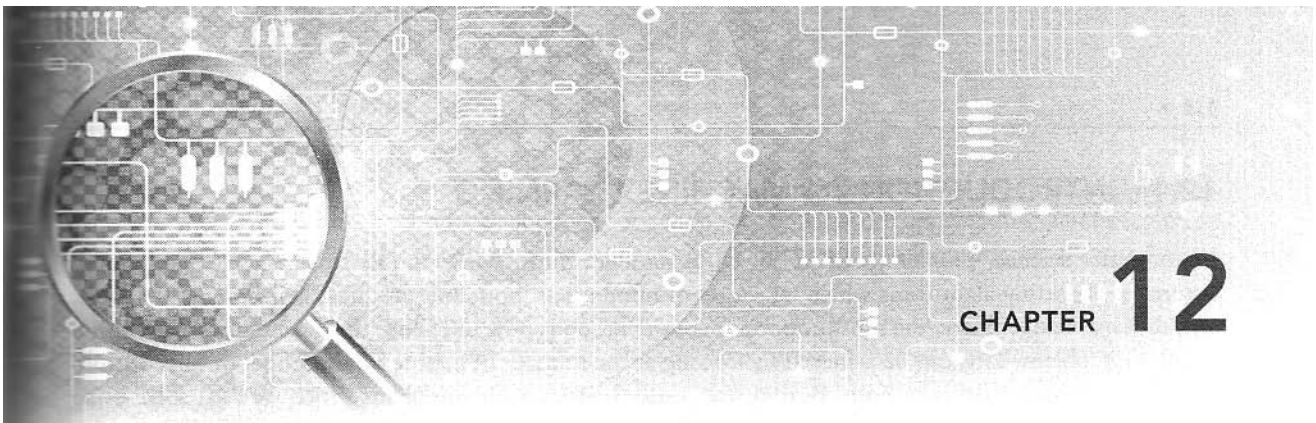
Column Name	Data Type	Description
ScoreKey	Int	Primary Key
AssessmentKey	Int	Not Null
ModuleKey	Int	Not Null
EmpKey	Int	Not Null
TimeKey	Int	Not Null
TotalScore	Int	Not Null
MaximumScore	Int	Not Null
Percentage	Int	Not Null

Reporting Section

Create the following report.

Chart Report:

- This Report has to be built by extracting data from 'JasperAssignment' database.
- Module Names on X-axis, Percentage Scored in the various modules on Y-axis. Use Bar chart.
- Take 'Emp Name' as input parameter. Take values of parameter from drop down list.
- Take 'Assessment Type' as Input Parameter with Available values as 'Test' and 'Retest' in drop down List.



CHAPTER 12

Introduction to Machine Learning

BRIEF CONTENTS

- What's in Store?
- Introduction to Machine Learning
  - Machine Learning Definition
- Machine Learning Algorithms
  - Regression Model – Linear Regression
    - Methodology
    - Implementation of Regression using R
  - Clustering
    - K-Means
    - K-Means implementation using R
  - Collaborative Filtering
    - History of Collaborative Filtering
    - Collaborative Filtering Algorithms
      - Euclidean Distance
      - Manhattan Distance
      - Pearson–Correlation Co-efficient
  - Association Rule Mining
    - Binary Representation
    - Item Set and Support Count
    - Why should you consider support and confidence?
  - Decision Tree
    - What are the uncertainties?
    - What is a decision tree?
    - Where it is used?
    - Advantages of Decision Tree
    - Disadvantages of Decision Tree

"If the statistics are boring, you have got the wrong numbers."

– Edward Tufte

WHAT'S IN STORE?

The focus of this chapter is to build knowledge about Machine Learning Algorithms. We will discuss supervised, unsupervised learning and few learning algorithms of these categories. We will also discuss implementation of Regression Model and K-means algorithms using **R** statistical tool.

We suggest you refer to some of the learning resources provided at the end of this chapter for better learning.

12.1 INTRODUCTION TO MACHINE LEARNING

In computer science, problems can be solved using algorithms. For example, to sort a set of numbers, a number of sorting algorithms which take a set of numbers as input and produce their ordered list as the output are available. Here, the most important consideration is how to choose the most efficient algorithm to solve a problem. This can be achieved by looking at the number of instructions essential to solve a problem, the memory available or both. But, in real world there are some problems which we can't solve using these kinds of algorithms. One such problem is filtering emails. In this case, the input is a file of characters and the output is to ascertain whether the email is spam mail or not.

To resolve this sort of problem, we can use machine learning algorithms. In learning algorithms, we utilize sample data, feed it to system and then train the system to produce the approximate model. For example, for email classification, we can compile thousands of example messages for spam and we can then make the system to learn to distinguish between a mail that qualifies as a spam message and the email that is NOT spam. In this situation, we may not be able to identify the process completely, but we can construct an approximate model. This approximate model may NOT be able to explain everything, but may be able to account for some part of it. This helps us to create certain patterns, which can then be used to understand the process or to make predictions. This is known as "**Machine Learning**". Machine Learning is one of the branches of Artificial Intelligence.

When we apply machine learning methods to large databases, it is known as **Data Mining**. This isn't just a database problem; it is also a part of Artificial Intelligence. The system should be intelligent enough to study the data in different environments.

Real-Time Example for Machine Learning

1. **Google Search Engine:** It works well because of the learning algorithm implemented by Google. This learning algorithm has learned how to rank web pages.
2. **Facebook:** When you use phototyping application, it is able to recognize your friend's photos because of machine learning algorithm.

12.1.1 Machine Learning Definition

Arthur Samuel (1959), Machine Learning: It is a field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998), Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Example

Tom's definition is the latest Machine Learning Definition. According to Tom's definition, we will try to identify E, P, T in Email Spam Classification Problem. In this problem,

Task – Classifying email as spam or not.
Experience – Labeling the email as spam or not.
Performance – Number of emails correctly classified as spam or not.

12.2 Machine Learning Algorithms

Machine Learning Algorithms can be classified into two categories.

1. **Supervised Learning:** In supervised learning, the classifier undergoes a process of training based on known classifications and through supervision it attempts to learn the information contained in the training dataset. Example: Predicting the selling price of the house based on pricing of other houses for sale in the neighborhood.
2. **Unsupervised Learning:** In unsupervised learning, the classifier tries to find some structure from the given data set without any known classification. That structure is known as Cluster. Example: Google News collects tens of thousands of news stories and automatically clusters them together. So that news stories that have the same content are displayed together.

12.2.1 Regression Model – Linear Regression

Regression Model is used to predict numbers. With the help of regression you can predict profit, sales, house values, etc. Regression Model serves as a good example for classification (supervised learning). Here, we will discuss *Linear Regression*.

Linear Regression is used to predict the relationship between two variables. The variable that needs to be predicted is known as the dependent variable and the variables that are used to predict the value of the dependent variable are known as independent variables.

12.2.1.1 Methodology

The general equation for Regression Model is as follows:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

Here, Y is the dependent variable; x_1, x_2, x_3 are the independent variables (these variables are used to predict the value of Y); $b_1, b_2, \ldots, b_n$ are the co-efficient of the respective independent variables (these values are determined from the input data); a is the intercept. Slope and intercept can be calculated using the below expression.

$$\text{Slope } (b) = [N\Sigma XY - (\Sigma X)(\Sigma Y)]/[N\Sigma X^2 - (\Sigma X)^2]$$
$$\text{Intercept } (a) = [\Sigma Y - b(\Sigma X)]/N$$

Example: Assume you wish to predict the housing price in Washington. Further assume that you have a dataset and you can plot the dataset as shown in Figure 12.1. Here, the horizontal axis indicates the sizes of different houses with the area specified in square feet and the vertical axis indicates the price of the houses

Figure 12.1 Housing price prediction.

in thousands of dollar. Let us say that you want to predict the price of house spread over 750 square feet. In this case, you can use learning algorithm to place a straight line through the data and based on that you can predict that you can get the house for about $150,000.

12.2.1.2 Implementation of Regression Model using R

R is a programming language used for data analysis. The following steps describe how to implement **"Housing Price Prediction"** using **RStudio**.

1. You can import data into the Environment as shown below. The name of the file is Housing.txt.

The following lines read data from the file "D:/Housing.txt" and prints the data on the console. Here "Housing" is a variable.

```
>Housing = read.table("D:/Housing.txt",header=TRUE)
>Housing
```

2. Let us consider all the attributes for predicting the price of a house. Construct a data frame as shown below. In **R**, the data frame is an array-like structure.

```
> myhouse = c("Bedrooms","Bathrooms","Kitchen","Size","Price")
> Housing1 = Housing[myhouse]
> Housing1
  Bedrooms Bathrooms Kitchen Size   Price
1        2         2       1 1250 5300000
2        3         2       2 1750 8400000
3        1         1       1  750 3200000
4        2         1       1 1250 4700000
5        1         1       1 1500 5500000
> |
```

3. To construct a multiple linear regression model with "houseprice" as the response variable and all the other attributes as the explanatory variables, use lm (linear model) command:

```
> houseprice = lm(Price ~ Bedrooms + Bathrooms + Kitchen + Size, da
ta=Housing)
> houseprice

Call:
lm(formula = Price ~ Bedrooms + Bathrooms + Kitchen + Size, data =
Housing)

Coefficients:
(Intercept)     Bedrooms    Bathrooms      Kitchen
   -1266667       -33333       600000      1600000
       Size
       3067

> |
```

4. To predict housing price, use predict command.

```
> predict(houseprice,data.frame(Size=1500, Bathrooms=2,Bedrooms=2,K
itchen=1))
       1
6066667
> |
```

The result displays the housing price prediction for a house whose size is 1500 square feet, 2 bedrooms, 2 bathrooms, and 1 kitchen.

Note: Price is in lakhs and Size is in square feet.

12.2.2 Clustering

Clustering is the process of grouping similar objects together. Clustering is an example of unsupervised learning. One can use clustering algorithms to segment data as in classification algorithms. However, classification models are used to segment data based on previously defined classes that are mentioned in the target, whereas clustering models do not use any target.

Clustering can be used to group items in a supermarket. For example, butter, cheese and milk can be placed in the "dairy products" group.

You can use clustering when you want to explore data. Clustering algorithms are mainly used for natural groupings. There are different categories of clustering. Refer Figure 12.2 for Clustering categories.

1. **Hierarchical:** Hierarchical cluster identifies the cluster within the cluster. A news article group can further have other groups such as business, politics, and sports in which each group can still have subgroups. For example, inside sports news there could be news on baseball sport, news on basketball sport, and so on.
2. **Partitional:** Partitional creates a fixed number of clusters. The K-means clustering algorithm belongs to this category. Let us study the K-mean clustering algorithm in detail.

12.2.2.1 K-Means

The steps involved in K-means algorithm are as follows:

1. Choose the final required number of clusters.
2. Examine each element in the population and assign it to one of the clusters depending on the minimum distance.
3. Each time a new element is added to the cluster, the centroid's position is recalculated. This process is performed until all the elements are grouped into the required number of clusters.

Centroid: It is a point whose parameter values are the mean of the parameter values of all the points in the cluster.

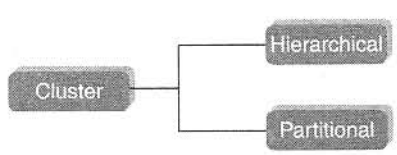


Figure 12.2 Categories of clustering.

12.2.2.2 K-Means Algorithm Implementation using R

Let us discuss implementation of K-means clustering using R.

1. You can import data into the Environment as shown below. The name of the file is Cars.txt. This file contains entry for Petrol cars and their corresponding mileage in kilometers.

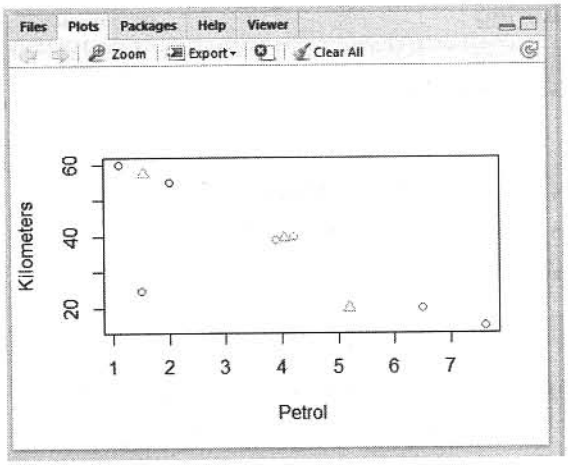
```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> cars = read.table("D:/Cars.txt",header=TRUE)
> cars
  Petrol Kilometers
1    1.1         60
2    6.5         20
3    4.2         40
4    1.5         25
5    7.6         15
6    2.0         55
7    3.9         39
>
```

2. Apply K-means algorithm as shown below. The data set is split into 3 clusters and the maximum iteration is 10.

```
> cars3 = kmeans(cars, centers=3, iter.max=10)
> cars3
K-means clustering with 3 clusters of sizes 3, 2, 2

Cluster means:
  Petrol Kilometers
1   5.20       20.0
2   1.55       57.5
3   4.05       39.5

Clustering vector:
[1] 2 1 3 1 1 2 3

Within cluster sum of squares by cluster:
[1] 71.140 12.905  0.545
 (between_SS / total_SS =  95.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

3. Next, you can plot clusters as shown below:

```
> plot(cars[cars3$cluster ==1, ], col = "red", xlim=c(min(cars[,1
]),max(cars[,1])), ylim=c(min(cars[,2]),max(cars[,2])))
>
> points(cars[cars3$cluster ==2, ], col ="blue")
>
> points(cars[cars3$cluster ==3, ], col ="green")
>
> points(cars3$centers,pch=2,col="orange")
>
```



12.2.3 Collaborative Filtering

Collaborative filtering is a technique used for recommendation. Let us assume there are two people A and B. A likes Apples and B also likes Apples. In this case, we can assume that B has similar liking as A. So we can go ahead and recommend options for A to B as well.

12.2.3.1 History of Collaborative Filtering

The history of collaborative filtering started with Information Retrieval and Information Filtering.

Information Retrieval

The information retrieval era was from 1960s to 1980s. It is about retrieving information based on the queries/questions and these contents are mostly static in nature. Indexes, if built, help with the retrieval of information. For example, a collection of books can be indexed by title, author, and summary specified for the book. However, information requirements do not stay the same and change from time to time. This kind of information is known as dynamic content. An example of dynamic content is Google Search Engine, which provides us with dynamic content based on our search criteria.

Information Filtering

The exponential growth of Web has led to the explosion of information. So we require some technique to reduce the information overload and sieve information that is relevant to the users. This is then utilized to build long-term profile of the users' needs. Email Filtering (Filtering Spam Messages) is an example for information filtering.

Collaborative Filtering

Collaborative filtering is a category of information filtering. It is nothing but predicting user preferences based on the preferences of a group of users. You can use collaborative filtering when information needs are more complex than keywords or topics. Collaborative filtering is concerned with quality and taste.

Collaborative filtering can be defined as **Social Navigation**. We say that human beings are social animals owing to their tendency to follow other people's advice or judgment when looking for information or buying products. This is in a way similar to an ant looking for food trudging behind other ants.

12.2.3.2 Algorithms of Collaborative Filtering

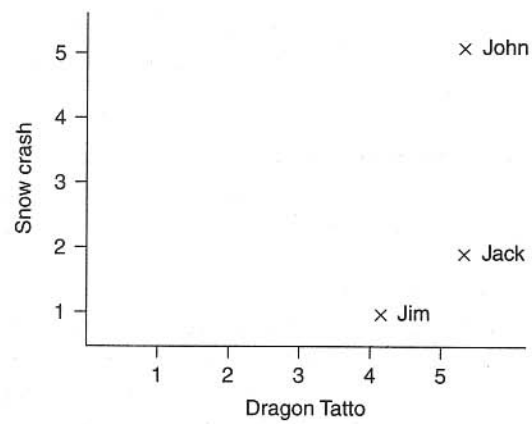
Let us discuss few of the collaborative filtering algorithms. In collaborative filtering, the important concern is "How to find someone who is similar?" It involves two steps:

1. **Collecting Preferences**

	Snow Crash	Dragon Tattoo
John	5★	5★
Jack	2★	5★
Jim	1★	4★

2. **Finding Similar Users:** Any of the following algorithms can be used to find similar users.
   - Euclidean Distance Score
   - Manhattan Distance or Cab Driver
   - Pearson–Correlation Co-efficient

Start by plotting the data as shown below. Here, X represents the Dragon Tatto and Y represents the Snow Crash.



***Euclidean Distance*:** The source of Euclidean distance is Pythagorean Theorem:

$$c = \sqrt{a^2 + b^2}$$

The distance between two points in the plane with coordinates (x, y) and (a, b) is given by

$$\text{dist}((x, y), (a, b)) = \sqrt{(x-a)^2 + (y-b)^2}$$

Example:

$$\text{dist}\ ((2,-1),\ (-2,\ 2)) = \sqrt{[2-(-2)]^2 + [(-1)-2]^2} = \sqrt{(2+2)^2 + (-1-2)^2}$$
$$= \sqrt{(4)^2 + (-3)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

The distance between Ms. Y (a person) and John is given in Figure 12.3.

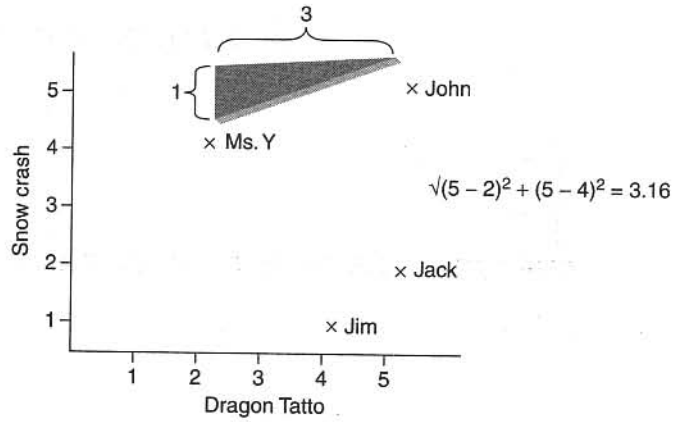


Figure 12.3 Euclidean distance between Ms. Y and John.

The distance between Ms. Y and all the three people is given below.

	Distance from Ms. Y
John	3.16
Jack	3.61
Jim	3.61

Note: For *N*-dimensional thinking refer the book specified in the Point Me Section.

Manhattan Distance or Cab Driver Distance: Here each person is represented by x and y.

$$(x_1, y_1) \Rightarrow \textbf{John and } (x_2, y_2) \Rightarrow \textbf{Ms. Y}$$

Formula to calculate Manhattan Distance for 2D is

$$|x_1 - x_2| + |y_1 - y_2|$$

Manhattan distance for John and Ms. Y is 4 (Figure 12.4).

So the distances between Ms. Y and all three people are as below:

	Distance from Ms. Y
John	4
Jack	5
Jim	5

John is the closest match. Manhattan distance is fast to compute. It is suitable for applications like Facebook to find another almost similar user amongst the millions of users.

Note: For *N*-dimensional thinking refer the book specified in the Point Me Section.

But the users have different behavior when it comes to their rating.

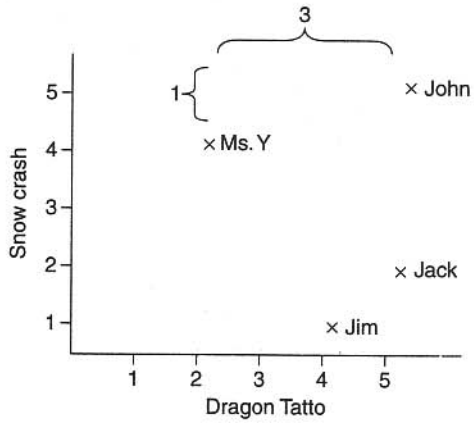


Figure 12.4 Manhattan Distance between Ms. Y and John.

Bill's ratings is between 2 and 4

Hailey's ratings is between 1 and 4

	Angelica	Bill	Chan	Dan	Hailey	Jordyn	Sam	Veronica
Blues Traveler	3.5	2	5	3	–	–	5	3
Broken bells	2	3.5	1	4	4	4.5	2	–
Deadmau5	–	4	1	4.5	1	4	–	–
Norah Jones	4.5	–	3	–	4	5	3	5
Phoenix	5	2	5	3	–	5	5	4
Slightly Stoopid	1.5	3.5	1	4.5	–	4.5	4	2.5
The Strokes	2.5	–	–	4	4	4	5	3
Vampire Weekend	2	3	–	2	1	4	–	–

Problem: How do we compare the similarities?

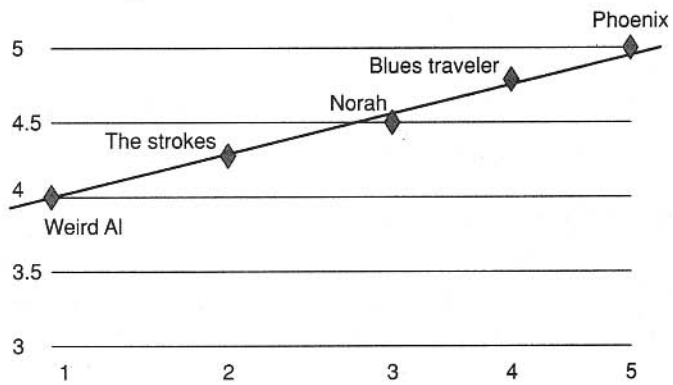
For example, does Hailey's "4" mean the same as Jordan's "4" or Jordan's "5"? The variability in this can create problems in the recommendation system.

The solution to this is **Pearson–Correlation Co-efficient**.

Pearson–Correlation Co-efficient

	Blues Traveler	Norah Jones	Phoenix	The Strokes	Weird Al
Clara	4.75	4.5	5	4.25	4
Robert	4	3	5	2	1

When we plot the chart, the data appears as below:



Here, the straight line indicates perfect agreement.

The formula to calculate PCC (Pearson–Correlation Co-efficient) is as stated below:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

Return ranges are between 1 and −1. 1 means perfect agreement and −1 implies disagreement. Let's first compute the expression in the numerator:

$$(4.75 \times 4) + (4.5 \times 3) + (5 \times 5) + (4.25 \times 2) + (4 \times 1) = 19 + 13.5 + 25 + 8.5 + 4 = 70$$

Proceeding further, let's compute the remaining part of the numerator:

Sum of Clara's ratings is 22.5.

Sum of Robert's ratings is 15.

They rated 5 bands, therefore

$$22.5 \times 15/5 = 67.5$$

Numerator value is: 70 − 67.5 = 2.5

Let us further compute the denominator:

Step 1:

$$(4.75)2 + (4.5)2 + (5)2 + (4.25)2 + (4)2 = 101.875$$

Step 2: Sum of Clara's ratings is 22.5. When we square that, we get 506.25. Divide this number by the number of co-rated bands (5) and we get 101.25. Putting it all together:

$$\sqrt{101.875 - 101.25} = \sqrt{0.625} = 0.79057$$

Similarly compute it for Robert:

$$\sqrt{55 - 45} = 3.162277$$

Putting it all together into the PCC equation gives us the following:

$$r = 2.5/0.79057\ (3.162277) = 2.5/2.5 = 1.00$$

So there is a perfect match between Clara and Robert.

Which Similarity Measure to use when?

1. **Pearson:** Use Pearson when different users use different scales.
2. **Euclidean or Manhattan:** Use Euclidean or Manhattan if you have values for all the attributes.

Note: You can use R to implement Collaborative Filtering Algorithms.

12.2.4 Association Rule Mining

Association rule mining is also referred to as market basket analysis. Few also prefer to call it as affinity analysis. It is a data analysis and data mining technique. It is used to determine co-occurrence relationship among activities performed by individuals and groups.

Examples:

1. It is widely used in retail wherein the retailer seeks to understand the buying behavior of customer. This insight is then used to cross-sell or up-sell to the customers.
2. If you have ever bought a book from Amazon, this should sound familiar to you. The moment you are done selecting and placing the desired book in the shopping cart, pop comes the recommendation stating that customers who bought book "A" also bought book "B".
3. Who can forget the urban legend, the very famous beer and diapers example. The legend goes... there was a retail firm wherein it was observed that when diapers were purchased beer was purchased as well by the customer. The retailer cashed in on this opportunity by stocking beer coolers close to the shelves that housed the diaper. This just to make it convenient for the customers to easily pick both the products.

An association rule has two parts (a) an antecedent (if) and (b) a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

PICTURE THIS...

Retailer "BigDailies" wants to cash in on their customers buying patterns. They want to be able to enact targeted marketing campaigns for specific segments of customers. They wish to have a good inventory management system in place. They wish to learn about which items/products should be stocked together to provide ease of buying to customers, in other words, enhance customer satisfaction.

Where should they start? They have had some internal discussions with their sales and IT staff. The IT staff has been instructed to design an application that can house each customer's transaction data. They wish to have it recorded every single day for every single customer and for every transaction made. They decide to meet after a quarter (3 months) to see if there is some buying pattern.

Table 12.1 illustrates a subset of the transaction data collected over a period of three months.

The table presents an interesting methodology called association analysis to discover interesting relationship in large datasets. The unveiled relationship can be presented in the form of association rules or sets of frequent items. For example, the following rule can be extracted from the above dataset:

$$\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$$

It is pretty obvious from the above rule that a strong relationship exists between the sale of diapers and beer. Customers who pick up a pack or two of diapers also happen to pick a few cans of beers. Retailers can leverage these sort of rules to partake of the opportunity to cross-sell products to their customers.

Table 12.1 Sample transactional dataset

Transaction ID	Transaction details
1	{bread, milk}
2	{bread, milk, eggs, diapers, beer}
3	{bread, milk, beer, diapers}
4	{diapers, beer}
5	{milk, bread, diapers, eggs}
6	{milk, bread, diapers, beer}

Challenges that need to be addressed while progressing with association rule mining are as follows:

1. The larger the dataset, the better would be the analysis results. However, working with large transactional datasets can be and is usually computationally expensive.
2. Sometimes few of the discovered patterns could be spurious or misleading as it could have happened purely by chance or fluke.

12.2.4.1 Binary Representation

Let us look at how we can represent the sample dataset in Table 12.1 in binary format (Table 12.2).

Explanation of the binary representation in Table 12.2: Each row represents a transaction identified by a "Transaction ID". An item (such as bread, milk, eggs, diapers, and beer) is represented by a binary variable. A value of 1 denotes the presence of the item for the said transaction. A value of 0 denotes the absence of the item from the said transaction. Example: for transaction ID = 1, bread and milk are present and are depicted by 1. Eggs, diapers, and beer are absent from the transaction and therefore denoted by zero. The presence of the item is more important than its absence, and for the same reason an item is called as an asymmetric variable.

12.2.4.2 Itemset and Support Count

Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be the set of all items in the market basket dataset.
Let $T = \{t_1, t_2, t_3, \ldots, t_n\}$ be the set of all transactions.

Itemset: Each transaction t_i contains a subset of items from set I. A collection of zero or more items is called an itemset. If an itemset contains k elements, it is called a k-item itemset. Example: the itemset {Bread, Milk, Diapers, Beer} is called a 4 itemset.

Table 12.2

Transaction ID	Bread	Milk	Eggs	Diapers	Beer
1	1	1	0	0	0
2	1	1	1	1	1
3	1	1	0	1	1
4	0	0	0	1	1
5	1	1	1	1	0
6	1	1	0	1	1

Transaction width: Transaction width is defined as the number of items present in the transaction. A transaction t_j contains an itemset X if X is a subset of t_j. Example transaction t_6 contains the itemset {bread, diapers} but does not contain the itemset {bread, eggs}.

Item support count: Support is an indication of how frequently the items appear in the dataset. Item support count is defined by the number of transactions that contain a particular itemset. Item support count can be expressed as follows: *Number of transactions that contain a particular itemset.*

Example: Support Count for {Diapers, Beer} is 4.

Mathematically, for an item set X the support count $\sigma(X)$ can be expressed as

$$\sigma(X) = \left|\{t_i\} \middle| X \subseteq t_i,\ t_i \in T\right|$$

The symbol | - | denotes the number of elements in the set.

Association rule: It is an implication rule of the form $X \rightarrow Y$ where X and Y are disjoint items, that is, $X \cap Y = \phi$. Support and confidence are the two factors that are utilized to get to the strength of the association rule mining.

The Support for an itemset is defined as follows:

Support $(x_1, x_2, \ldots)$ = Number of transactions containing $(x_1, x_2, \ldots)$/Total number of transactions (n)

Support for X → Y = Number of transactions containing $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ $/n$ (total number of transactions)

Example:

Support for {Milk, Diapers} → {Beer} as per the dataset in Table 12.1 is as follows:

Support for {Milk, Diapers} → {Beer} = 3/6 = 0.5

Confidence of the rule is

Confidence of $(x_1, x_2, \ldots)$ implies $(y_1, y_2, \ldots)$ = Support for $(x_1, x_2, \ldots)$ implies $(y_1, y_2, \ldots)$/Support for $(x_1, x_2, \ldots)$

Confidence of {Milk, Diapers} → {Beer} = Support for {Milk, Diapers} → {Beer}/Support for {Milk, Diapers}

Substituting, the actual values, we get

Confidence of {Milk, Diapers} → {Beer} = 0.5/Support for {Milk, Diapers} = 0.5/0.67 = 0.7462

Why should you consider support and confidence?

It is vital to consider support and confidence owing to the following reasons: A rule which has low support may occur simply by chance. To place big bets on it may prove futile. If you deliberate from the business perspective, it may prove to be rather unexciting and non-lucrative purely due to the fact that it does not make sense to promote items that customers seldom buy together. Support is used to chuck off uninteresting rules.

Confidence is a measure of reliability of inference of an association rule. For a given rule $X \rightarrow Y$, the higher the confidence the more likely it is for Y to be present in transactions that contain X. Confidence of a rule can also be used to provide an estimate of the conditional probability of Y given X.

The results of association rule analysis should be considered astutely and judiciously. It does not necessarily imply causality. For causality to be in effect, it requires knowledge about the causal and effect attributes

in the data. It requires the relationship to be observed, recorded, and studied over a period of time. For example: Ozone depletion leads to global warming. Think association rule mining, and think establishing co-occurrence relationship between items in the antecedent and consequent of the rule.

Note: You can use R to implement Association Mining Algorithm.

12.2.5 Decision Tree

PICTURE THIS...

It is that time of the year again. The college fest is going to be next week. It is going to be a week long affair. Your friends have started planning on the kind of stalls that they will put up. You too want to try out this stall thing. However, you are yet to decide on what stall you should go for. You have to communicate your decision to the organizing committee in a day's time. The time is short. The decision has to be made quickly. You do not want to end up with a wrong decision. It is your first time at putting up a stall. You want to go for maximum profit.

You have zeroed down your choice to either an ice-cream stall or a burger stall from a gamut of choices available. How about using a decision tree to decide on the same? Let us look at how we can go about creating a decision tree.

Decision to be made: Either an ice-cream stall or a burger stall.

Payoff: 3500 INR in profit if you put up a burger stall and a 4000 INR in profit if you put up an ice-cream stall.

12.2.5.1 What are the uncertainties?

There is a 50% chance of you succeeding to make profit with a burger stall and a 50% chance of you failing at it.

As per the weather forecast, it will be downcast sky and may drizzle or pour slightly throughout the week. Keeping this into consideration, there is a 40% chance of success and 60% chance of failure with an ice-cream stall.

Let us look at the cost of the raw materials:

For Burger: 700 INR for the burger buns, the fillings, and a microwave oven to keep it warm.

For Ice-creams: 1000 INR for the cone, the ice-cream, and a freezer to keep it cold. Refer Figure 12.5.

Let us compute the effective value as per the below formula:

$$\text{Expected value for burger} = 0.5 \times 3500 \text{ INR} - 0.5 \times 700 \text{ INR} = 1400 \text{ INR}$$

$$\text{Expected value for ice-cream} = 0.4 \times 4000 \text{ INR} - 0.6 \times 1000 \text{ INR} = 1000 \text{ INR}$$

The choice is obvious. Refer Figure 12.6. Going by the expected value, you will gain by putting up a burger stall.

The expected value does not imply that you will make a profit of 1400 INR. Nevertheless, this amount is useful for decision-making, as it will maximize your expected returns in the long run if you continue to use this approach.

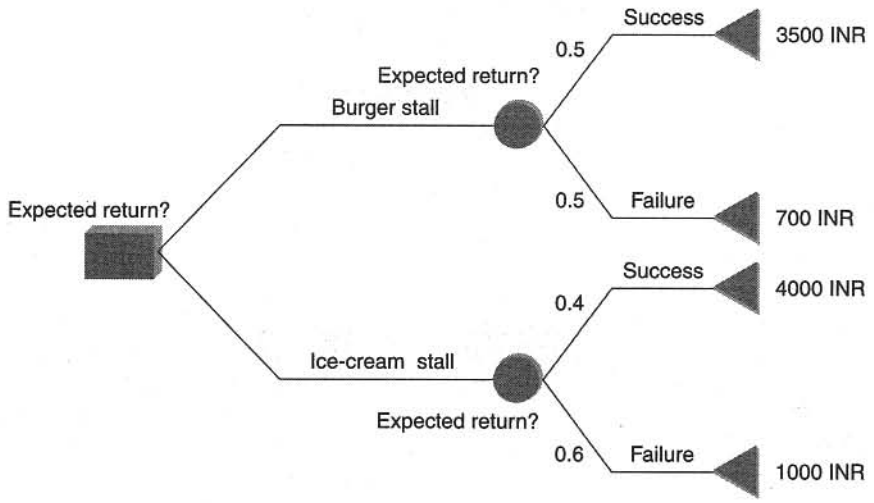


Figure 12.5 A sample decision tree with expected return value yet to be arrived at.

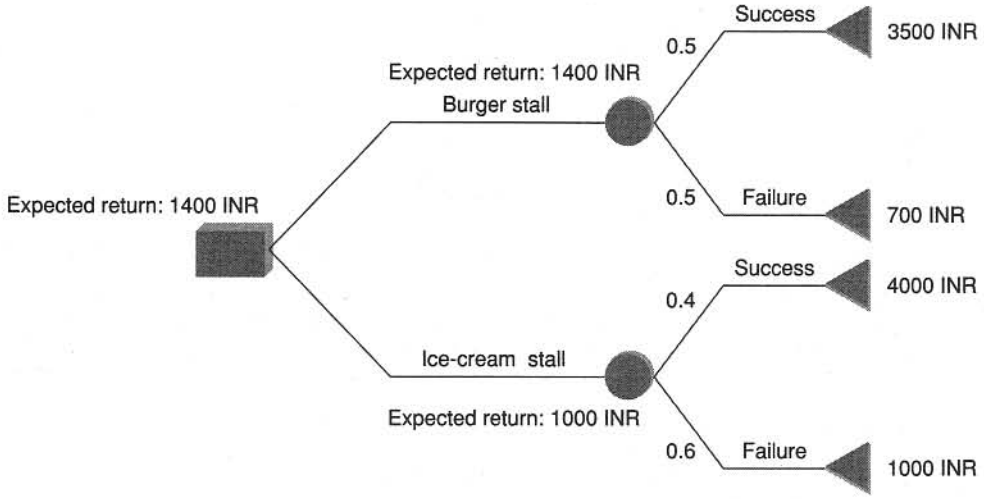


Figure 12.6 A sample decision tree with computed expected return.

PICTURE THIS...

You have just completed writing the script of a romantic story. There are two takers for it. (a) The television network: they are interested in making a daily soap of it that will be telecast on prime time. (b) XYZ Movie Company: they have also shown interest. You are confused. Should you sell the rights to the TV Network or XYZ Movie Company?

The TV network payout will be a flat 500,000 INR. XYZ Movie Company will pay in accordance to the audience response to the movie.

Payouts and Probabilities

TV Network payout:
Flat Rate: 500,000 INR

XYZ Movie Company Payout:

Small Box Office: 250,000 INR
Medium Box Office: 600,000 INR
Large Box Office: 800,000 INR

Probabilities:

P (Small Box Office): 0.3
P (Medium Box Office): 0.5
P (Large Box Office): 0.2

For greater understanding, let us create a payoff table:

Decisions	Small Box Office	Medium Box Office	Large Box Office
Sign up with TV Network	500,000 INR	500,000 INR	500,000 INR
Sign up with XYZ Movie Company	250,000 INR	600,000 INR	800,000 INR
Probabilities	0.3	0.5	0.2

Refer Figure 12.7.

Let us compute the effective value as per the below formula:

$$\text{Expected value for TV Network} = 0.3 \times 500{,}000 + 0.5 \times 500{,}000 + 0.2 \times 500{,}000 = 500{,}000 \text{ INR}$$

$$\text{Expected value for XYZ Movie Company} = 0.3 \times 250{,}000 + 0.5 \times 600{,}000 + 0.2 \times 800{,}000 = 535{,}000 \text{ INR}$$

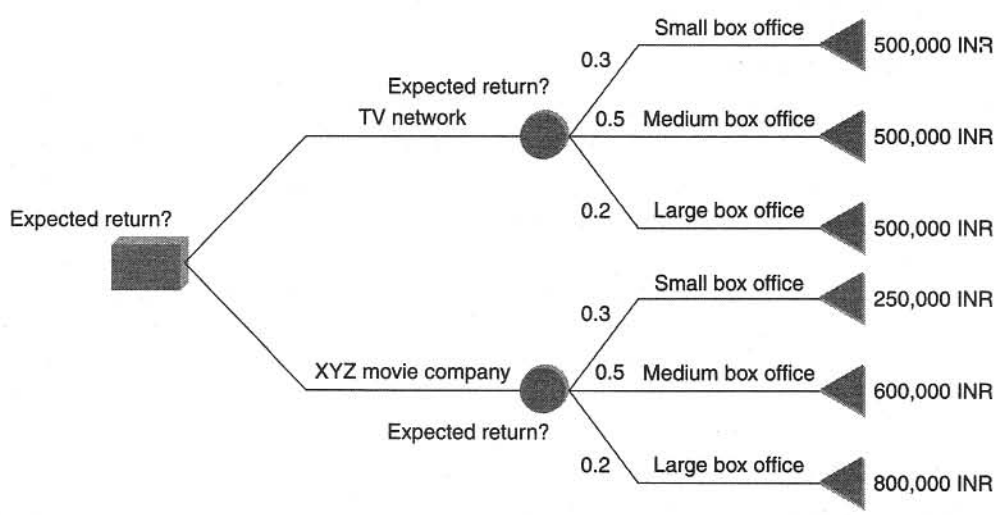


Figure 12.7 A sample decision tree with expected return value yet to be arrived at.

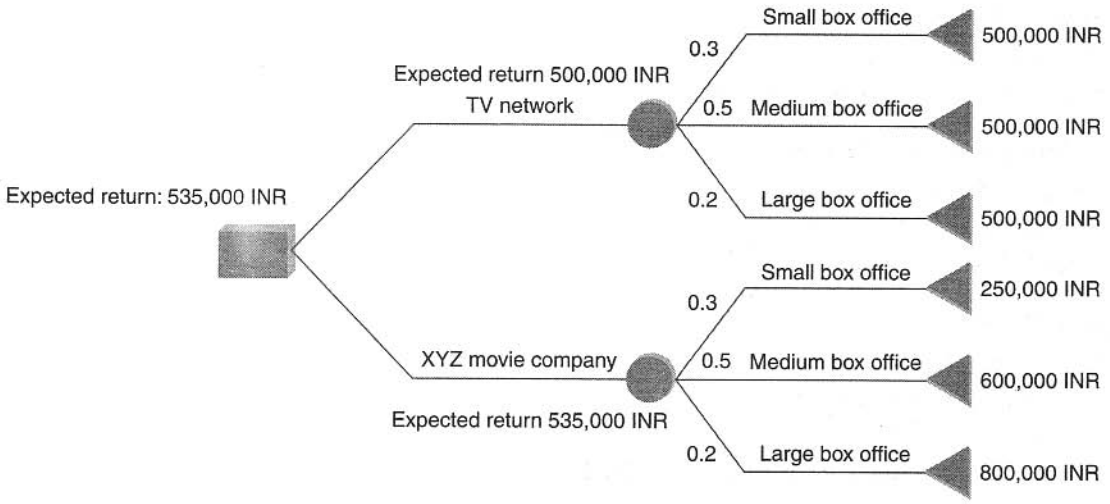


Figure 12.8 A sample decision tree with computed expected return.

The choice is obvious. Refer Figure 12.8. Going by the expected value, you will gain by selling the rights of your script to XYZ Movie Company. The expected value does not imply that you will make a profit of 535,000 INR.

12.2.5.2 What is a Decision Tree?

A decision tree is a decision support tool. It uses a tree-like graph to depict decision and their consequences. The following are the three constituents of a decision tree:

1. **Decision nodes:** Commonly represented by squares
2. **Chance nodes:** Represented by circles
3. **End nodes:** Represented by triangles

12.2.5.3 Where is it used?

Decision trees are commonly used in operations research, specifically in decision analysis. It is used to zero down on a strategy that is most likely to reach its goals. It can also be used to compute conditional probabilities.

12.2.5.4 Advantages of using a Decision Tree

1. Easy to interpret.
2. Easy to plot even when there is little hard data. If one is aware of little data such as alternatives, probabilities, and costs, it can be plotted and can lead to useful insights.
3. Can be easily coupled with other decision techniques.
4. Helps in determining the best, worst, and expected value for a given scenario or scenarios.

12.2.5.5 Disadvantages of Decision Trees

1. **Requires experience:** Business owners and managers should have a certain level of experience to complete the decision tree. It also calls for an understanding of quantitative and statistical analytical techniques.

2. **Incomplete information:** It is difficult to plot a decision tree without having complete information of the business and its operating environment.
3. **Too much information:** Too much information can be over-whelming and lead to what is called as the "paralysis of analysis".

REMIND ME

There are two types of Learning Algorithms:

- **Supervised Learning:** In this learning, classifier undergoes a process of training based on known classifications, and through supervision it attempts to learn the information contained in the training dataset. Regression Model is an example for supervised learning.
- **Unsupervised Learning:** In this learning, the classifier tries to find some structure from the given dataset without any known classification. That structure is known as Cluster. Clustering is an example for unsupervised learning.
- Collaborative filtering is an example for user based recommendation.
- Association rule mining is an example for item-based recommendation.
- Decision tree is a decision support system and used in operations research.

POINT ME (BOOK)

- A Programmer's Guide to Data Mining by Ron Zacharski.

CONNECT ME (INTERNET RESOURCES)

- https://www.coursera.org/course/ml: Coursera "Machine Learning" course from Stanford
- http://www.rstudio.com/resources/training/online-learning/

TEST ME

A. Fill Me

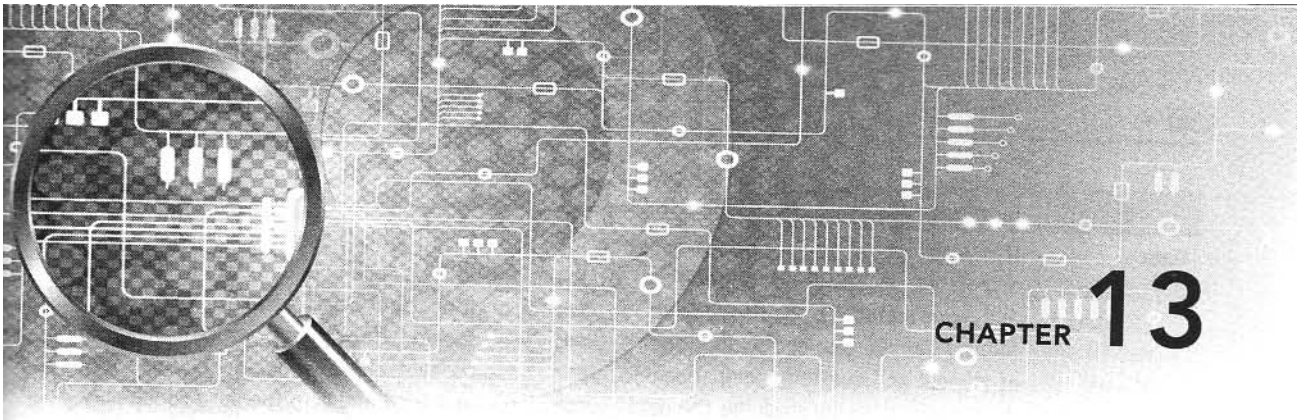
1. Decision tree is a _____ method.
2. R is _____ tool.
3. _____ is a user-based recommendation algorithm.
4. K-mean splits dataset in to _____ of Clusters.
5. Regression Model deals with _____.

Answers:

1. Supervised Learning
2. Statistical
3. Collaborative Filtering
4. Fixed number
5. Numeric Values

ASSIGNMENT FOR HANDS-ON PRACTICE

Write an R Program to implement Association Mining for frequently used item set. [Hint: You can construct your own dataset]



CHAPTER 13

Few Interesting Differences

BRIEF CONTENTS

- Difference between Data Warehouse and Data Lakes.
- Difference between RDBMS and HDFS.
- Difference between HDFS and HBase.
- Difference between MapReduce and Pig.
- Difference between MapReduce and Spark.
- Difference between Pig and Hive.

"Data is the new science. Big Data holds the answers."

– Pat Gelsinger

WHAT'S IN STORE?

The focus of this chapter is to bring out the differences between various Hadoop ecosystem components for easy lookup and remembrance. It will be good to read this chapter sequentially for a better absorption. Starting with data warehouse versus data lakes, it builds further to explain the differences between HDFS and RDBMS, then goes on to highlight differences of MapReduce with Pig, Spark, etc.

13.1 DIFFERENCE BETWEEN DATA WAREHOUSE AND DATA LAKE

Who coined the term data lake? It was Pentaho CTO, James Dixon. He described data mart (a subset of data warehouse) as akin to a bottle of water "cleansed, packaged and structured for easy consumption" while a data lake is more like a body of water in its natural state. Data flows from the streams (the source systems) to the lake.

	Data Warehouse	Data Lake
Data	It is a structured data extracted from multiple disparate data sources, integrated, transformed and made suitable to guide management decisions (allows the creation of trending reports such as manual and quarterly comparisons). Due diligence is done by studying the data sources, understanding business processes and profiling data before data is stored.	Data lake is a data storage repository to store huge amount of data in its original format until it is needed. All the data is stored here. The data which is needed today, the data which may be needed and used and also the data which may never be used basis it may be used someday.
Schema	Schema on write. Before data is written into the data warehouse, schema has to be compulsorily defined. In fact, data is not loaded into the warehouse until the use for it has been defined.	Schema on read. It stores raw data (the data that has not yet been processed for a purpose) in its original format. It is only at the time of reading that shape and structure need to be defined for the data.
Data types	It is structured data from transactional systems.	This includes data from web logs, sensor data, social media data, text and images.
Purpose	In-use clearly defined purpose.	Undetermined.
Cost associated with storing data	High cost storage.	Commodity, off-the-shelf servers with low-cost storage are used and therefore it helps to quickly scale from terabytes to petabytes to exabytes.
Agility	Data warehouses are less agile and have somewhat a fixed configuration.	Since data lakes lack a structure, it is easy to configure, reconfigure data models, queries and applications.
Data Security	Data warehouses have been around for decades and that has led to gaining maturity when it comes to securing data.	Data lakes are usually open-source, low-cost storage. Although significant strides are being made to secure data, it will still be some time before maturity sets in.
Users	The data in data warehouse is highly structured, purpose built and organized well therefore it is easy to use and comprehend. About 80% of the users in the organization are able to use the data to meet their operational needs like pull up reports, track their key performance metrics, etc. Few users (about 10%) need the data for more analysis. And the last 10% of the users – the data scientist tribe – will perform deeper analysis on the data. They usually tend to visit the original data sources (also a few new ones) in addition to the D/W.	Usually, data scientists who can play around with the data from varied sources.
Accessibility	More complicated and costly to make any changes to the structure, etc.	Highly accessible, quick to update.

13.2 DIFFERENCE BETWEEN RDBMS AND HDFS

	RDBMS	HDFS
Data Model	It is a relational database. Data are stored in tables comprising rows and columns.	It is not a database. It is a distributed clustered file system with support for parallelism and redundancy.
Data	Essentially used for structured data. Supports semi-structured data to a limited extent.	It is used for all varieties of data – structured, semi-structured and unstructured. It supports serialization and various file formats such as text, JSON, XML, Avro, etc.
Data Storage	Usually used for datasets (sizes in GBs, TBs).	Used for large datasets (sizes ranging in terabytes, petabytes, exabytes, etc.).
Querying	SQL (Structured Query Language).	Declarative language – HiveQL.
Schema	"Schema on Write". Mandates that the schema be pre-defined before placing data in it.	"Schema on Read". Stores data in its native format. It is only to be formatted at the time of reading data from it.
Speed	Reads are fast.	Both writes and reads are fast.
Cost	Available both as open-source (MySQL, PostgreSQL, etc.) as well as licensed ones such as Oracle, IBM-DB2, MS SQL Server, Teradata, etc.	Open-source and free (Apache Hadoop).
Usage	OLTP (Online Transaction Processing).	Data discovery and data analytics.
Throughput (throughput refers to the amount of data processed in a period of time)	Low	High
Scalability	Vertical (scale up). Scales by increasing the horsepower of the machine (CPU, RAM, Hard disk capacity, etc.).	Horizontal. Also called as linear scalability or scaling out. Scales by adding nodes to the cluster.
Hardware	High-end servers.	Commodity hardware.
Integrity	High. This is owing to strict adherence to the ACID (Atomicity, Consistency, Isolation/Integrity and Durability) properties of transactions.	Low. It supports BASE (Basically Available Soft state Eventual consistency).

13.3 DIFFERENCE BETWEEN HDFS AND HBASE

	Hadoop HDFS	HBase
What is it?	Hadoop Distributed File System (HDFS) stores huge amounts and types of data in a distributed (provides faster read/write access) and redundant (provides better availability and fault tolerance) manner on industry standard commodity hardware.	It is an open-source, distributed, non-relational scalable NoSQL database that runs on top of the Hadoop cluster and provides a random real-time read–write access to the data. It is like a layer on top of HDFS. Using APIs, it is possible to write NoSQL queries and get the results. It is modeled after Google's BigTable.
Storage	Datasets are divided into smaller subsets called chunks/blocks and stored across clusters.	Data is stored in key–value pairs in column-oriented database that uses column families to group similar or frequently accessed data together.
Flexibility to read-write	Write once, Read many times.	Multiple read–write of data stored in HDFS.
Scalability	Multiple nodes can be added to the cluster and therefore hugely and infinitely scalable.	Can store huge amounts of data.
Analytics	Supports batch analytics. Batch analytics is high latency analytics whereby a huge volume of data is processed. There is latency as the data is collected, stored and then processed for analysis and reporting.	Supports batch and real-time analytics. Real-time analytics is low latency as the data from multiple disparate data sources and in any data format is filtered, aggregated, enriched and analyzed to help identify simple patterns, urgent situations, automate immediate actions etc.
Access to data	Only sequential access to data.	Random access to data.
Write Pattern	Append only.	Random write, bulk incremental.
Read Pattern	Full table scan.	Random read, small range scan or table scan.
Dynamic changes	Rigid architecture that does not allow changes.	Allows for dynamic changes and can be utilized for standalone applications.
Latency (how long does it take a file system/database to respond to a single request)	High latency operations.	Low latency access to small amounts of data from within a larger dataset.
Accessibility	Primarily accessed through MapReduce jobs.	Can be accessed through shell commands, client APIs in Java, REST, Avro or thrift.

13.4 HADOOP MapReduce VERSUS PIG

	Hadoop MapReduce	Pig
What is it?	• It is a low-level language. • It leads to a huge amount of custom user code that is difficult to maintain and reuse.	• It is a high-level scripting language. • It is a data flow language. • Pig is an application that runs atop MapReduce, YARN or Tez. • It is written in Java. • It converts (compiles) Pig Latin scripts into MapReduce jobs. • Approx. 10 lines of Pig code are equivalent to 200 lines of Java code. • Pig Latin scripts are easier to read for someone without a Java background.
Joins	Performing datasets joins is very difficult.	Pig helps easily with sort, parse, join, etc.
Extensibility	Not easy to extend. Functions need to be written from scratch.	Easily extendable using UDFs. The UDFs are fairly reusable.
Development cycle	It is fairly long comprising writing mappers, reducers – compiling, packaging the code, submitting the jobs, etc.	No need of compiling and packaging. The Pig operators are converted to MapReduce tasks internally.
Code Portability	May not be supported with all versions of Hadoop.	Works with any version of Hadoop.

13.5 DIFFERENCE BETWEEN HADOOP MapReduce AND SPARK

Spark can run standalone or on top of Hadoop YARN (Yet Another Resource Negotiator). In other words, Spark is not bound to Hadoop, although both Spark and Hadoop MapReduce are included in the distributions by Hortonworks and Cloudera.

	Hadoop MapReduce	Spark
Data storage	Stores data in local disk. This implies that data is written to and read from hard disk.	Stores data in memory.
Speed	Slow speed. However, MapReduce has given good results for ETL style jobs and transformations where it needs to pass through a piece of data only once.	Fast speed. This is primarily because data is stored in memory for processing. However, there could be major performance issues if the size of the program is too big to fit into memory. The performance issue is compounded if Spark is running on top of YARN along with other resource demanding services. Spark has been known to perform well for iterative computations that go over the same data many times.

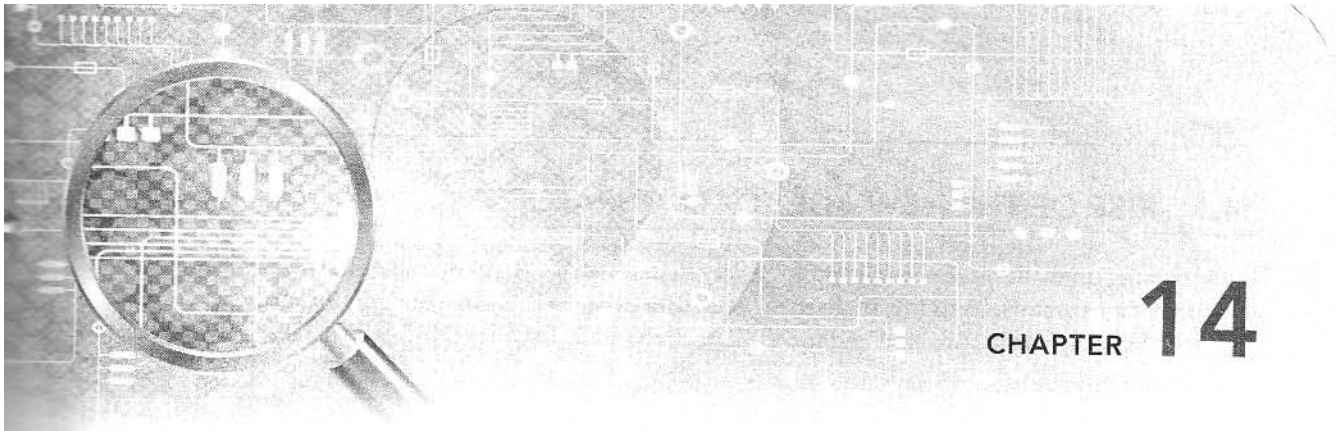
	Hadoop MapReduce	Spark
Real Time	Suitable for batch processing.	Suitable for batch and real-time processing. Spark can process and analyze data the moment it is captured and insights fed back to the user for appropriate action.
Scheduler	External schedulers such as Oozie are required.	Schedules tasks itself.
Latency	High latency.	Low latency.
Interactivity	No in-built interactive mode although Hive has a command line interface.	Has interactive mode that helps to run commands with immediate feedback.
Cost	Less expensive hardware.	Lots of RAM required for in-memory processing; increasing it in the cluster gradually increases its cost.
Difficult level	Difficult to program. We need to code/handle each process.	Fairly easy to program with the availability of RDD (Resilient Distributed Dataset). RDDs provide fault tolerance to Spark.
Platform developed on	Developed using Java.	Has APIs for Python, Java Scala and Spark SQL.
SQL support	MapReduce runs queries using Hive Query Language.	Spark has its own query language called Spark SQL.
Machine Learning	Supports Apache Mahout tool for machine learning.	Has its own machine learning library called MLib.
Caching	MapReduce does not cache in memory data so it is not as fast as Spark.	Spark caches in memory data, therefore it is very apt for iterative analytics.
Language Supported	MapReduce basically supports C, C++, Ruby, Groovy, Perl, and Python.	Spark supports Scala, Java, Python, R, and SQL.
Security	Hadoop supports Kerberos and LDAP for authentication. It also has support for traditional file permissions as well as ACLs (Access Control Lists).	Spark security is not as mature as Hadoop. However, Spark can integrate with HDFS and use HDFS ACLs and file level permissions.

Components in Hadoop Ecosystem and Their Equivalent in Spark

Hadoop	Spark
Hive	Spark SQL
Apache Graph	GraphX
Impala	Spark SQL
Apache Storm	Spark Streaming
Apache Mahout	MLib

13.6 DIFFERENCE BETWEEN PIG AND HIVE

	Pig	Hive
Language used	Pig has a procedural scripting language called Pig Latin for describing operations like reading, writing, filtering, transforming, merging and joining data. These are largely the operations that were performed using MapReduce. The difference being now it can be done without having to write thousands of lines of Java code. Pig is "Scripting for Hadoop".	Hive has a declarative language called HiveQL. Hive is "SQL for Hadoop". It basically is used for data summarizations, ad-hoc querying and analysis of large datasets.
Data	Works with both structured and semi-structured data.	Essentially works with structured data.
Who uses?	Researchers and Programmers	Data Analysts
Operates on	Client side of the cluster.	Server side of the cluster.
Support for Avro file format	Supports Avro file format.	Does not have support for Avro file format. However, with the evolution of Serge, support for Avro file format can be provided.
Developed at	Yahoo	Facebook
Partitioning and Bucketing	Does not support partitioning and bucketing.	Makes use of partitioning and bucketing.
Speed of loading and execution	Faster than Hive.	Slower than Pig.
Usage	It can be suitably used for prototyping and rapidly developing MapReduce jobs.	It can be used to perform ad hoc queries or generate report data spread across multiple nodes in the cluster.
Schema	Hive defines tables schema + stores schema information in database.	Pig does not have dedicated metadata of database.



CHAPTER 14

Big Data Trends in 2019 and Beyond

BRIEF CONTENTS

- What's in Store?
- Rise of the New Age "Data Curators"
- CDOs are Stepping up
- Dark Data in the Cloud
- Streaming the IOT for Machine Learning
- Edge Computing
- Open Source
- Hadoop is Fundamental
- Chat bots will get Smarter
- Container(ed) Revolution
- Visualization Commoditization

"Information is the oil of the 21st century, and analytics is the combustion engine."

– Peter Sondergaard, Senior Vice President, Gartner

WHAT'S IN STORE?

This chapter discusses the big data trends in 2019 and beyond. The years ahead will see an increase in the adoption of open-source technologies. Hadoop is and will remain fundamental although there will be increased usage of the in-memory Spark. The years ahead will also awake the container(ed) revolution. The last half-a-decade has been witness to the commoditization of visualization. The rising wave of IoT (Internet of Things) will lead to processing being done on the edge of the network before moving it to the central data center in the cloud. The world will witness the power of empowered computing – edge and quantum. It is time to utilize and draw value/insight from the abundant dark data. Also bots will mature and get smarter in the coming years.

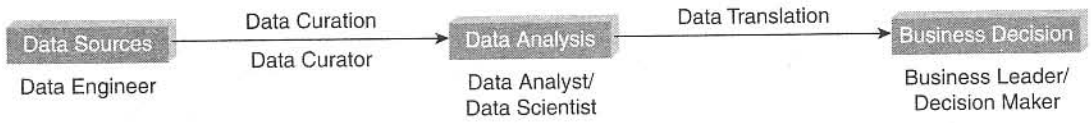
14.1 RISE OF THE NEW AGE "DATA CURATORS"

In the coming times, data curator will become a very significant role. To understand the role and responsibilities of the data curators, let us first understand what data curation is. In simple terms, data curation implies ensuring that people can easily find and use the data now and in the future. It can be detailed down to:

1. Identifying the most relevant data sources.
2. Sourcing (getting) the data from data owners.
3. Fixing any issues with the data such as missing values, unexpected values, and mysterious variables. Also preparing the data if not suitable for analysis; for example, if the data has too much details or is too granular to be picked up for analysis then maybe summarization or aggregation might help in the analysis.
4. Using annotations, tags, appropriate and crisp documentation, etc. to help make the data easy to find, use, reuse and present.
5. Preserving the data such that it is available for use, reuse, etc.

Data curator is a role that sits between data engineers and data consumers (data analysts, data scientists, etc.).

Let us discuss how data was made available to data consumers when there was no role as data curator. Data consumers, basis their analytics task, would raise requests for data. They would provide specifications/details such as the data sources or datasets required for the job, the format in which data should be provided, the tools that they will use to run analysis on the data, how frequently data should be updated, and some details about the analysis task on hand. Data engineers would then run the errand of getting the data from the identified sources, ask for more details if required, blend the data from two or more disparate sources, transform it as per requirements, secure access, etc. The data engineers have a complete understanding of the infrastructure and the formatting of the data; however, they may or may not understand the data completely. The data consumers, on the other hand, have a fairly good understanding of the data but usually do not have much idea of the systems, processes and tools to bring data to the present form.



Enter the role of the data curator to bridge this gap between IT – data engineers and data consumers (data analysts, scientists, etc.). They work closely with both the data engineers as well as data consumers. On one hand, they have up-to-date knowledge about datasets, their provenance (origin), and what data curation is needed; on the other hand, they also understand the different types of analysis to be performed on specific datasets as well as the expectations of latency and availability set by diverse business users.

14.2 CDOS ARE STEPPING UP

With data becoming the new oil, the clear mandate is to create more and more value from the organization's data. Enter the role of CDO (Chief Data Officer) to help with data leveraging (use the existing data assets in the best possible way), data enrichment (augment the value of data by blending, bringing together internal and external data), data monetization (exploring newer avenues of earnings and revenues), data upkeep (ensuring proper data quality and governance), data protection (ensuring security and adequate protection of data), etc.

Let us look at few statistics:

- Estimated number of CDOs globally as per Gartner:
  2010: 15
  2014: 400
  2018: 4000+
- Percentage of large companies with a CDO in place:

	CDO	
Year	Yes	No
2012	12%	88%
2017	56%	44%
2018	63%	375

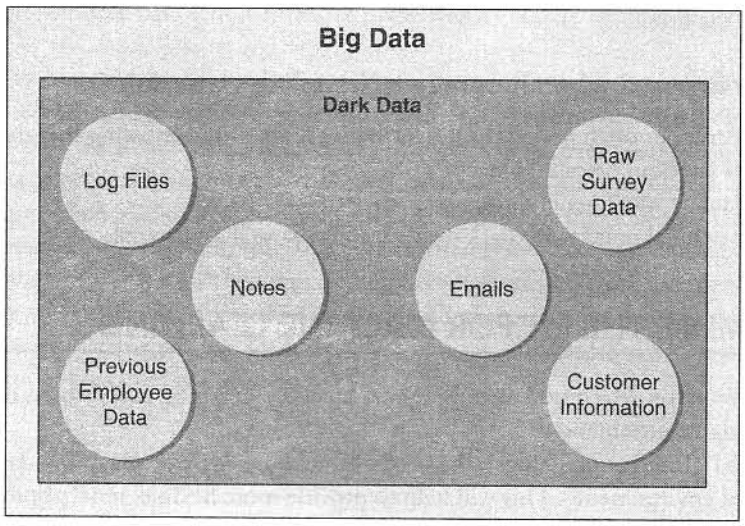
14.3 DARK DATA IN THE CLOUD

What is Dark Data?

According to Gartner, dark data is "the information assets organization's collect, process and store during regular business activities, but generally fail to use for other purposes".

We all know about big data. Data that is big in volume, variety and velocity. Not all the big data that is collected by the organizations are processed, analyzed or used. This data that is collected and stored by the organizations thinking that it will be used for some analysis sometime in future but is never put to any use is "dark data". Dark data is a subset of big data. It also happens to constitute the biggest portion of the large volume of big data that organizations collect and store.

As per IDC (International Data corporation), 90% of unstructured data is "dark data".



Why is it important then to consider dark data? Primarily because this dark data could mean opportunity or opportunities lost for an organization. It may have untapped, undiscovered useful insights which could spell success for the organization. Another reason why it becomes important is if the dark unanalyzed data is not handled well, it can result in a lot of problems such as legal and security problems.

Why then the Dark Data has not been handled the way it should have been?

There are several reasons for this. Few are listed below:

1. **Not knowing what to do with the data:**
   Picture this:
   A bank receives online applications for credit cards. Their focus is mainly on collecting and analyzing customer details and eligibility. They attach little or no priority to finding out how the customer came to the application page. If this data was collected and analyzed, it could have provided useful insights about the usage of the bank website or could have helped improve the application.
2. **Disconnect among departments:** Typically in large organizations, departments operate in silos. They have their own data collection and storage processes. They may or may not reveal these processes to other departments. So, there is a good chance of data lying unused even though the possibility is that some or all of this data may be relevant/useful to some other department.
3. **Technology and tool constraint:** Again, if we take the case of a large organization, there is a high possibility that all applications do not use the same technology and tools for data collection, storage, etc. Sometime the integration of all this data becomes a problem and at times impossible. There could be integration issue, data quality issues, data governance issues, data ownership issues, etc.

What problems can dark data cause?

1. **Opportunity lost:** This is the data that is as yet untapped. It may have useful insights which can help the company surge ahead of competition.
2. **Legal and regulatory issues:** A lot of this data is lying around sometimes secured, sometimes unguarded. Any inadvertent disclosures could lead to intelligence risk, legal liabilities and even loss of reputation for the firm.

What is the way forward when it comes to handling dark data?

1. Properly structuring or categorizing the data will go a long way in ensuring that the process of searching for the data later can be eased out.
2. Securing the data by way of encryption, etc.
3. Having clearly defined policies for dark data retention as well as disposal.

14.4 STREAMING THE IoT FOR MACHINE LEARNING

Let us quickly do a recap on Machine Learning (ML). Machine Learning uses "stored data" for training in a "controlled" learning environment.

With the rise of IoT (Internet of Things), the need of the hour is to use "streaming data in real time" in a "much less controlled environment". This will help to provide more flexible, more appropriate responses to a variety of situations including communicating with humans.

Gone are the days of operating in silos. Today, IoT data, streaming analytics, machine learning, and distributed computing have come together to offer a very powerful, yet an inexpensive proposition to store and analyze big volume and varied types of digital data.

Some examples of IoT, Big Data, and Machine Learning working together include:

1. **Healthcare:** Continuous monitoring of chronic diseases.
2. **Smart Cities:** Traffic patterns and congestion management.
3. **Transportation:** Optimizing routes and fuel consumption.
4. **Automobile:** Smart cars.
5. **Retail:** Location-based advertising.

14.5 EDGE COMPUTING

Edge computing allows data from IoT devices to be analyzed at the edge of the network before being sent to a data center or cloud.

It takes advantage of microservices architectures to allow some portion of computing to be moved to the edge of the network which reduces network traffic and processing time.

Benefits that edge computing provides are as follows:

1. **Cost-effective data processing solution:** Edge computing lowers IoT costs by locally performing vital data computing. Businesses can then decide which services to run locally and which will reside in the Cloud. This greatly optimizes IoT solution costs, by augmenting with cellular-based technologies at a lower-cost, and improves return on investment.
2. **Faster response times:** Data latency (the time the request is submitted to the time the response/outcome is available) is reduced by cutting out round trips to the Cloud. This delivers faster responses. This in turn ensures that critical operations function smoothly without breaking down or incidents occurring.
3. **Security and compliance:** Edge computing reduces or avoids general data transfer between devices and the Cloud. Data can be filtered basis its sensitivity. Sensitive information can be processed locally with non-sensitive data sent for further processing to adhere to strict security and compliance frameworks.
4. **Interoperability between legacy and new devices:** Enterprises typically will have legacy as well as modern smart devices. Edge computing acts as an interpreter between legacy and modern devices. Legacy devices have their own communication protocols. Edge computing helps convert the same for easy consumption and comprehension by new smart devices and the Cloud. This enables legacy devices to be connected to the latest IoT platforms, extending the life of older IT architecture.
5. **Dependable operations with sporadic Internet connectivity:** Edge computing enables equipment that is remotely placed or has intermittent Internet connectivity to work seamlessly. Equipment can function offline without any disruption providing an ideal scenario for fast analysis of data in distant or remote locations such as oil rigs, solar farms, rural areas, etc. It also can detect equipment failure even in instances of restricted connectivity.

14.6 OPEN SOURCE

The future of open source is looking brighter than ever with corporate buy-ins. Two major acquisitions which made headline in 2018 were that of GitHub by Microsoft for an estimated 7.5 billion USD and Red Hat by IBM for a whopping 34 billion USD.

The coming years will see a widespread adoption of open-source components/tools in the software development lifecycle as also the tools that can help developers manage open source to avoid late stage security and compliance issues.

14.7 HADOOP IS FUNDAMENTAL AND WILL REMAIN SO!

The last decade witnessed the failure of quite a few big data projects. While there were reasons galore, two prominent ones were:

1. Spark displacing Hadoop as a stand-alone installation.
2. Data lakes becoming popular in Cloud storage layers.

However, Hadoop is fundamental and is likely to remain so with various Hadoop ecosystem components being used to analyze data. It is yet again touted that Hadoop together with Spark, Business Intelligence tools for integration and visualization will rule the data analytics markets.

14.8 CHATBOTS WILL GET SMARTER

"By 2020, over 50% of medium to large enterprises will have deployed product chatbots," said Van Baker, Research Vice President at Gartner at the Gartner Application Architecture, Development & Integration Summit, held March 12–13, 2018 in Mumbai.

We have already come a long way from the subpar interactions with bots like Siri, Alexa, Google Assistant and have seen them mature over a period of time. Yet there is still room for improvement.

Some of the business use cases for chatbots are:

1. Use your phones or mobile apps to place orders.
2. Handle customer functions: order status, order cancellations, return instructions, tracking numbers, order modifications, account balance, etc.
3. Act as personal digital assistants that help employees do basic tasks: reserving conference rooms, registering mileage, recording expenses, etc.
4. Provide automated support responses to customer inquiries .
5. Navigate through customer services, such as government or administrative services.

Chatbots with their ability to use natural language processing to map spoken or written input to intent, are increasingly entering the workspace. This technology is here to stay.

14.9 CONTAINER(ED) REVOLUTION

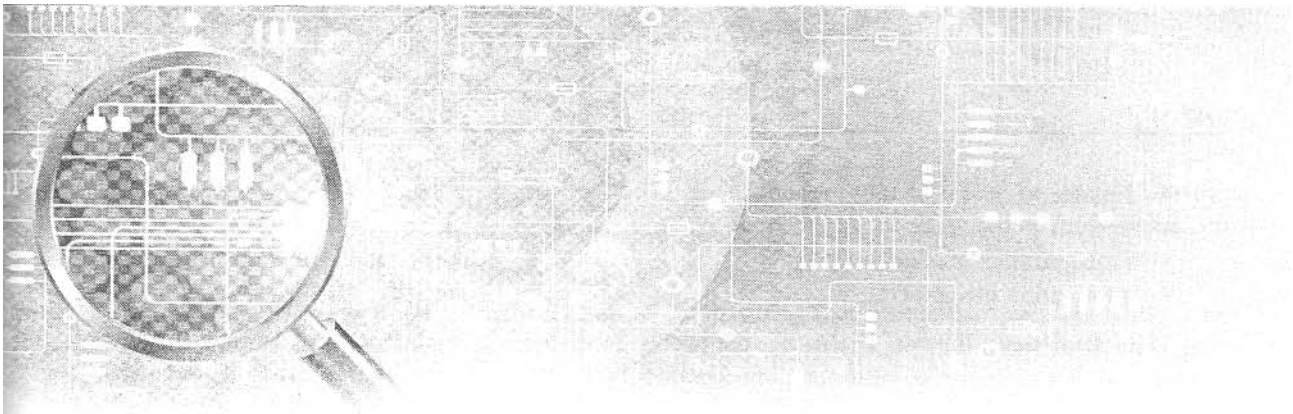
Docker-based container technology is becoming popular by the day. The cloud providers are taking advantage of container technology to make the provisioning of nodes faster and to facilitate greater resource sharing – allowing ephemeral clusters to seem persistent.

Hadoop itself, which recently hit its 3.0 release, will soon support the ability for code deployed to YARN to run in the context of a Docker container, thus allowing Hadoop job code dependencies to differ from what may be installed on each node in the cluster.

14.10 COMMODITIZATION OF VISUALIZATION

Visualization is now commoditized. Now visualization ceases to provide competitive advantage. Good analytics and visualization are available for free for most part. The constraint no longer seems to be the monetary investment but rather the constraint and limitation are more in terms of the available time to perform analysis and reporting.

Just as an example, one of the market leaders, Microstrategy, now provides connectors to Tableau, QlikView and PowerBI – the self-service BI tools and the three market leading visualization tools (Tableau, QlikView and PowerBI) between them together with Plot.ly, D3 ecosystem and few open-source geospatial tools provides entry level analytics for free.



Glossary

A

Ad-Hoc Query Any query that cannot be determined prior to the moment the query is issued. A query that consists of dynamically constructed SQL, which is usually constructed by desk-top resident query tools.

Ad-Hoc Query Tool An end-user tool that accepts an English-like or point-and-click request for data and constructs an ad-hoc query to retrieve the desired result.

Affinity Analysis Refer to Association rule mining.

Ambari Component of Hadoop Ecosystem. It is a web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.

Apache Spark It is a cluster computing framework which is open source and distributed. It is fast, in-memory data processing engine. It can be a standalone installation or can sit atop the Hadoop cluster.

Association Rule Mining It is a well-researched method of finding relations between variables in large databases.

Availability In distributed system, availability implies that a read/write request from a client will always be catered to.

B

Big Data Big data is data that is big in Volume, Variety and Velocity. It is difficult to store and process big data using traditional database and software techniques. Big data has challenges with the following:

- Capture
- Curation
- Storage
- Search

- Transfer
- Analysis
- Visualization
- Information privacy, etc.

Big Data Analytics Big data analytics is the process of examining big data to uncover patterns, unearth trends, and find unknown correlations and other useful information to make faster and better decisions.

BigTable It is a compressed, proprietary data storage system built on Google File System.

BSON Binary JSON.

Business Data Information about people, places, things, business rules, and events, which is used to operate the business. It is not metadata. (Metadata defines and describes business data.)

C

Cassandra Born at FaceBook, Cassandra is an open source, NoSQL database, and a wide column data store. It supports tunable consistency. It has support for MapReduce.

Central Warehouse A database created from operational extracts that adheres to a single, consistent, enterprise data model to ensure consistency of decision-support data across the corporation. A style of computing where all the information systems are located and managed from a single physical location.

Clustering Clustering is the process of grouping similar objects together. The objects in the same group are more similar (called a cluster) to each other than to those in other groups (clusters).

Chatbot It is an Artificial Intelligence (AI) program designed to simulate conversations with humans. It is also called Chatterbot/Smartbot/Talkbot, etc. Example: Siri, Alexa, Google Assistant, etc.

Chukwa Component of Hadoop Ecosystem. It is a data collection system for managing large distributed systems.

Collaborative Filtering It is used by several recommender systems. Example: it is about making predictions about the interests or preferences of a user by collecting preferences/interests or taste information from many users.

Column Database Each storage block has data from only one column. For example: Cassandra, HBase, etc.

Consistency In distributed system, consistency implies that a read request from a client will always fetch the most recent write.

CouchDB CouchDB is a document-oriented database. It has a JavaScript interface. It uses a multi-version concurrency version. It uses JSON to store data.

D

Dark Data According to Gartner, dark data is "the information assests organizations collect, process and store during regular business activities, but generally fails to use for other purposes".

Data Items representing facts, text, graphics, bit-mapped images, sound, analog or digital live-video segments. Data is the raw material of a system supplied by data producers and is used by information consumers to create information.

Data Curation In simple terms, data curation implies ensuring that people can easily find and use the data now and in the future.

Data Curator Data curator is a role that sits between data engineers and data consumers (data analysts, data scientists, etc.). They work closely with both the data engineers as well as data consumers. On one hand, they have up-to-date knowledge about datasets, their provenance (origin), and what data curation is needed and on the other hand, they also understand the different types of analysis to be performed on specific datasets as well as the expectations of latency and availability set by diverse business users.

Database Schema The logical and physical definition of a database structure. Example: the schema of an "Employee" table.

Employee Table:

ColumnName	Data Type and Length	Constraints
EmpID	Varchar(6)	Primary Key
EmpName	Varchar(30)	
EmpDesg	Varchar(20)	Not Null
EmpUnit	Varchar(10)	Not Null

Data Integration This concept describes extraction of business data from various disparate sources to produce a unified view for the end user.

Data Loading The process of populating the data warehouse. Data loading is provided by DBMS-specific load processes, DBMS insert processes, and independent fastload processes.

Data Mart A data mart is usually a subset of the data warehouse and is oriented to the needs of a specific business line or team.

Data Mining A technique using software tools, geared for the user who typically does not know exactly what he's searching for, but is looking for particular patterns or trends. Data mining is the process of sifting through large amounts of data to produce data content relationships. This is also known as data surfing.

Data Model A logical map that represents the inherent properties of the data independent of software, hardware or machine performance considerations. The model shows data elements grouped into records, as well as the association around those records.

Data Modeling A method used to define and analyze data requirements needed to support the business functions of an enterprise. These data requirements are recorded as a conceptual data model with associated data definitions. Data modeling defines the relationships between data elements and structures.

Data Store A place where data is stored; data at rest. A generic term that includes databases and flat files.

Data Warehouse An implementation of an informational database used to store sharable data sourced from an operational database-of-record. It is typically a subject database that allows users to tap into a company's vast store of operational data to track and respond to business trends and facilitate forecasting and planning efforts. Ralph Kimball uses the following terms to describe a data warehouse:

- Subject-oriented
- Integrated
- Non-volatile
- Time-variant

Data Visualization A graphic representation of data.

DBA Database Administrator.

DCL Data Control Language statements. DCL statements are as listed below:

- GRANT
- REVOKE

DDL Data Definition Language statements. A few DDL statements are as follows:

- CREATE TABLE
- ALTER TABLE
- DROP TABLE
- CREATE INDEX
- DROP INDEX
- TRUNCATE TABLE

Decision trees It is a decision support tool. It uses a tree-like graph or model of decisions and their possible consequences.

Descriptive Analytics Descriptive analytics helps to answer the following questions:

- What happened?
- Why did it happen?, etc.

It is reporting on events, occurrences of the past.

DML Data Manipulation Language statements. The DML statements are the following:

- SELECT
- INSERT
- UPDATE
- DELETE

Document Database It maintains data in collections constituted of documents. For example, MongoDB, Apache CouchDB, Couchbase, MarkLogic, etc.

Sample Document in Document Database

```
{
"Book Name":          "Fundamentals of Business Analytics",
"Publisher":          "Wiley India",
"Year of Publication": "2011"
}
```

DW Data Warehouse. Refer to Data Warehouse.

DWH Data Warehouse. Refer to Data Warehouse.

E

Edge Computing It allows data from Internet of things devices to be analyzed at the edge of the network before being send to a data center or cloud.

EDW Enterprise Data Warehouse. Refer to Data Warehouse.

EIS Executive Information System.

ELT Extract, Load and Transform. It is essentially the same as ETL.

Enterprise A complete business consisting of functions, divisions, or other components used to accomplish specific objectives and defined goals.

Enterprise Data Data that is defined for use across a corporate environment.

ETL Extract Transform Load. ETL is a process of extracting data from a multitude of disparate sources, transforming it and loading it into the enterprise data warehouse or data marts.

G

Graph Database They are also called network database. A graph stores data in nodes. For example, Neo4J, HyperGraphDB, etc.

Greenplum Database from EMC.

H

Hadoop Hadoop is an open-source project of the Apache foundation. It is a framework written in Java, originally developed by Doug Cutting in 2005 who named it after his son's toy elephant.

HBase Hadoop Database.

HDFS Hadoop Distributed File System.

Hive Component of Hadoop Ecosystem. It enables analysis of large data sets using a language very similar to standard ANSI SQL.

Horizontal Scaling A data processing system such as Hadoop is designed to run on clusters of low-cost commodity hardware. It is easy to scale such an arrangement by adding more nodes to the cluster as and when there is need for more storage and processing speed.

I

Impala It is an open-source, distributed, SQL query engine, and analytic database from Cloudera architected to leverage the flexibility and scalability of Hadoop.

Information Data that has been processed in such a way that it can increase the knowledge of the person who receives it. Information is the output or "finished goods" of information systems. Information is also what individuals start with before it is fed into a Data Capture transaction processing system.

Internet of Things (IoT) IoT is sometimes referred to as Internet of Everything. IoT refers to all the web-enabled devices that collect, send and act on data they acquire from their surrounding environments using embedded sensors, processors, and communication hardware. Examples: Smart TVs, wearables, smart appliances, etc. Another example of IoT is the use of smart city technologies to monitor traffic, weather conditions, etc.

J

JSON Java Script Object Notation. MongoDB, a NoSQL uses JSON documents in its collection.

K

Key Performance Indicator (KPI) KPIs are measures to gauge the progress of an Enterprise. It is a means to gauge the progress (or lack of it) in realizing the firm's objectives or strategic plans.

Key–Value Database It maintains a big hash table of keys and values. For example, Dynamo, Redis, Riak, etc.

L

Latency Time It is defined as the time interval between the stimulation and the response. In simple words, it is the time elapsed since the input is fed into a system till the desired outcome is made available.

Linear Regression Linear Regression is used to predict the relationship between two variables, a scalar dependent variable "Y" and one or more explanatory variables denoted by "X".

M

Machine Learning A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Mahout Component of Hadoop Ecosystem. It is a scalable machine learning and data mining library.

MapReduce MapReduce is an algorithm design pattern. It essentially consists of three steps:

- *Step 1 – Mapper function*:
  It takes your input data and outputs a series of keys and values to use in calculating the results.
- *Step 2 – Shuffle and Sort:*
  The output from "Step 1 – Mapper function" which is typically the unordered list of keys and values is then put through a shuffle and sort step to ensure that all the fragments that have the same key are placed next to one another in the file.
- *Step 3 – Reducer function:*
  The reducer function works on the sorted output to reduce/aggregate the input to a single output result per key.

Market Basket Analysis Refer to association rule mining.

Metadata Metadata is data about data. Examples of metadata include data element descriptions, data type descriptions, attribute/property descriptions, range/domain descriptions, and process/method descriptions. The repository environment encompasses all corporate metadata resources: database catalogs, data dictionaries, and navigation services. Metadata includes things like the name, length, valid values, and description of a data element. Metadata is stored in a data dictionary and repository. It insulates the data warehouse from changes in the schema of operational systems.

MPP Massive Parallel Processing. The "shared nothing" approach of parallel computing.

MongoDB MongoDB, the term has come from the word "humongous". It is a NoSQL database. It is a document-oriented database with records stored as JSON (Java Script Object Notation) documents. It supports automatic sharding and MapReduce operations.

Example:

```
{
BookTitle: "Big data and Analytics",
BookAuthor1: "Seema Acharya",
BookAuthor2: "Subhashini Chellapan",
Publisher: "Wiley India"
}
```

N

NewSQL NewSQL is a new modern RDBMS. It supports relational data model and uses SQL as their primary interface. It is a database that has the same scalable performance of NoSQL systems for On Line Transaction Processing (OLTP) while still maintaining the ACID guarantees of a traditional database.

Normalization The process of reducing a complex data structure into its simplest, most stable structure. In general, the process entails the removal of redundant attributes, keys, and relationships from a conceptual data model.

NoSQL (Not only SQL) A NoSQL database provides a storage and retrieval mechanism that allows data to be modeled in forms other than relational or tabular forms of conventional databases. Data can be stored in any of the following forms:

- Key–Value pairs: *Example:* Amazon DB
- Document-oriented *Example:* MongoDB
- Column-oriented *Example:* Cassandra
- Graph-based *Example:* Neo4j

The key features of a NoSQL database are its ability to horizontally scale and it's adherence to CAP (Consistency, Availability, and Partition tolerance) theorem.

NUMA Non-Uniform Memory Access. Here, the memory access time depends on the memory location relative to the processor.

O

OBIEE Oracle Business Intelligence Enterprise Edition (formerly Siebel Analytics).

ODBC Open Database Connectivity. A standard for database access co-opted by Microsoft from the SQL Access Group consortium.

ODI Oracle Data Integrator, an ETL tool.

OLAP On-Line Analytical Processing.

OLTP On-Line Transaction Processing. OLTP describes the requirements for a system that is used in an operational environment.

Oozie Component of Hadoop Ecosystem. It is a workflow scheduler system to manage Apache Hadoop jobs.

Operational Database The database-of-record, consisting of system-specific reference data and event data belonging to a transaction-update system. It may also contain system control data such as indicators, flags, and counters. The operational database is the source of data for the data warehouse. It contains detailed data used to run the day-to-day operations of the business. The data continually changes as updates are made, and reflects the current value of the last transaction.

Operational Data Store (ODS) An ODS is an integrated database of operational data. Its sources include legacy systems and it contains current or near term data. An ODS may contain 30 to 60 days of information, while a data warehouse typically contains years of data.

P

Partition Tolerance In distributed system, partition tolerance implies that the system will continue to function even when network partition occurs.

Pig Component of Hadoop Ecosystem. Pig is an easy to understand data flow language.

Predictive Analytics Predictive Analytics helps to answer the following questions:

- What will happen?
- Why will it happen?

It uses data from the past to make predictions for the future.

Prescriptive Analytics The key questions that prescriptive analytics helps to answer are as follows:

- What will happen?
- When will it happen?
- Why will it happen?
- What should be the action taken to take advantage of what will happen?

Q

Query A (usually) complex SELECT statement for decision support. See Ad-Hoc Query or Ad-Hoc Query Software.

Query Response Time The time it takes for the warehouse engine to process a complex query across a large volume of data and return the results to the requester.

Query Tools Software that allows a user to create and direct specific questions to a data base. These tools provide the means for pulling the desired information from a database. They are typically SQL-based tools and allow a user to define data in end-user language.

R

RDBMS It expands to Relational Database Management System. In RDBMS, data is neatly organized into rows and columns. It conforms to the relational data model.

Examples: Oracle Corporation – Oracle, Microsoft – Microsoft SQL Server, IBM – DB2, Teradata Corporation – Teradata, EMC – Greenplum, etc. Amongst the open sources, the popular RDBMS are MySQL and PostgreSQL, etc.

Replication The word "replication" means duplication of data. Replication in database parlance means the electronic copying of data from a database on one computer or server to a database in another computer or server. This ensures that all users share the same level of information.

Replication Factor Replication Factor connotes the number of data copies of a given data item/data block stored across the network.

Regression Analysis It is a statistical process for estimating the relationships among variables.

S

SaaS Software as a Service. It is a software distribution model. Here, the application or applications are hosted by vendors or service providers and made available to the customers over a network.

Scalability The ability to scale to support larger or smaller volumes of data and more or less users. The ability to increase or decrease size or capability in cost-effective increments with minimal impact on the unit cost of business and the procurement of additional services.

Schema The logical and physical definition of data elements, physical characteristics and inter-relationships.

Scorecard The concept of Balanced Scorecard was given by Kaplan and Norton in 1992. It is a measurement system built on integrated data and helps an organization view business performance.

Securability The ability to provide differing access to individuals according to the classification of data and the user's business function, regardless of the variations.

SELECT An SQL statement (command) that specifies data retrieval operations for rows of data in a relational database.

Semantic Mapping The mapping of the meaning of a piece of data.

Semi-structured Data Semi-structured data is a cross between the structured and unstructured data. It is a type of structured data, but does not have the strict data model structure. Semi-structured data is also referred to as self-describing data. It makes use of tags or other types of markers to identify certain elements within the data; however, the data doesn't have a rigid structure.

Example of semi-structured data: XML (EXtensible Markup Language), JSON (Java Script Object Notation), etc.

Sharding The word "shard" means part of a whole. Sharding in database parlance means a horizontal partitioning of the data in the database into smaller, faster and more easily managed shards.

Shared Disk Architecture It is a distributed computing architecture. Here, all disks are accessible from all cluster nodes.

Shared Memory Architecture It is a multiprocessing design where numerous processors access a globally shared memory.

Shared Nothing Architecture In shared nothing architecture, neither the memory nor the disk is shared. Each node is independent and self-sufficient. There is no single point of contention across the system.

Sqoop Component of Hadoop Ecosystem. It is used to transfer bulk data between Hadoop and structured data stores such as relational databases.

Structured Data Structured data is data that has strict adherence to a data model or pre-defined schema. Before we store the data, it requires defining what fields of data will be stored, what will be the data types (integer, date, character, variable length character string, etc.) of each of the data field, what will be the restrictions/constraints (PRIMARY KEY, NOT NULL, UNIQUE, etc.) on these data fields, etc.

Structured data is usually managed using SQL (Structured Query Language).

Example of structured data: Any Relational Database Management System (RDBMS) such as Oracle, DB2, MS SQL Server, MySQL, PostgreSQL, etc.

Symmetric Multiprocessing System (SMP) In SMP (Symmetric Multiprocessing System), two or more identical processors share a common main memory. The processors have complete access to the IO devices and are controlled by a single operating system instance. SMP are tightly coupled multi-processor systems.

T

Throughput Value It is the rate of production or the rate at which something can be processed. In other words, it is the productivity of a system, procedure, process or machine over a unit period.

U

UMA Uniform Memory Access. Here, all processors access the physical memory uniformly.

Unstructured Data Unstructured data cannot be easily classified and fitted into neat boxes. Unstructured data does not have conformance to a pre-defined data model. It is typically text-heavy.

Example of unstructured data: Photos, videos, chat conversations, short text messages (SMS), email, free-form text, audios, blog entries, wikis, etc.

V

Vertical Scaling Traditional database architectures (RDBMS) are designed to run well on a single machine. In order to handle larger volumes of operations, the approach usually employed is to upgrade the machine with a faster processor or more memory.

X

XML EXtensible Markup Language. It was designed to describe data.

Y

YARN Yet Another Resource Negotiator. The fundamental idea of MapReduce version 2.0 (MRv2) is to split up two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons. We now have a global ResourceManager (RM) and a per-application ApplicationMaster (AM).

Z

ZooKeeper Component of Hadoop Ecosystem. It is a coordination service for distributed applications.

About the Book

BIG DATA is a term used for massive mounds of structured, semi-structured and unstructured data that has the potential to be mined for information. The real power lies not just in having colossal data but in what insights can be drawn from this data to facilitate better and faster decisions.

This book *Big Data and Analytics* is a comprehensive coverage on the concepts and practice of Big Data, Hadoop and Analytics. From the *Do It Yourself* steps and guidelines to set up a Hadoop Cluster to the deeper understanding of concepts and ample time-tested hands-on practice exercises on the concepts learned, this ONE book has it all!

Salient Features of the Book

- The book is an exhaustive introduction to Big Data Technology Landscape.
- The concepts are presented in simple language for easy comprehension by beginners.
- The concepts are explained with the help of illustrations and real-life industrial strength application examples.
- The book is accompanied with STEP-BY-STEP Hands-On chapters on
  - NoSQL databases such as MongoDB and Cassandra
  - MapReduce Programming
  - Pig – the data flow language
  - Hive – data warehouse built on top of Hadoop
  - Jasper Soft Studio to design report by pulling the data from MongoDB and Cassandra
- For easy lookup and remembrance, the book carries comparison of various Hadoop components such as HDFS vs. HBase, MapReduce vs. Pig, Pig vs. Hive, Pig vs. Spark, etc.
- For better learning, comprehension and retention, the book is supported by the following:
  - Various types of self-assessment such as "Fill in the Blanks", "Crossword", "Match the Columns", "Hands on Assessment", etc. The solutions to these are also provided.
  - References/web links/bibliography at the end of every chapter.
- **Glossary** of terms frequently used in the big data and analytics at the end of the book.

Resources available on www.wileyindia.com

- Installation guidelines.
- Dataset for JasperReports Assignment.
- Dataset to practice Import from CSV in Cassandra.
- Dataset to practice MongoDB_Import.

follow us on
facebook.com/wileyindia twitter.com/wileyindiapl linkedin.com/in/wileyindia

Wiley India Pvt. Ltd.
Customer Care +91 120 6291100
csupport@wiley.com
www.wileyindia.com
www.wiley.com

WILEY

ISBN 978-81-265-7951-8
9 788126 579518