Premier Reference Source

# Applied Big Data Analytics in Operations Management

Manish Kumar

IGI GLOBAL
DISSEMINATOR of KNOWLEDGE

# Applied Big Data Analytics in Operations Management

Manish Kumar
*Indian Institute of Information Technology, Allahabad, India*

A volume in the Advances in Business Information Systems and Analytics (ABISA) Book Series


www.igi-global.com

Published in the United States of America by
IGI Global
Business Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com



Library of Congress Cataloging-in-Publication Data

CIP Data Pending
ISBN: 978-1-5225-0886-1
eISBN: 978-1-5225-0887-8

This book is published in the IGI Global book series Advances in Business Information Systems and Analytics (ABISA) (ISSN: 2327-3275; eISSN: 2327-3283)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

# Advances in Business Information Systems and Analytics (ABISA) Book Series

ISSN: 2327-3275
EISSN: 2327-3283

## Mission

The successful development and management of information systems and business analytics is crucial to the success of an organization. New technological developments and methods for data analysis have allowed organizations to not only improve their processes and allow for greater productivity, but have also provided businesses with a venue through which to cut costs, plan for the future, and maintain competitive advantage in the information age.

The **Advances in Business Information Systems and Analytics (ABISA) Book Series** aims to present diverse and timely research in the development, deployment, and management of business information systems and business analytics for continued organizational development and improved business value.

## Coverage

- Business Models
- Business Process Management
- Forecasting
- Data Analytics
- Performance Metrics
- Data Strategy
- Business Intelligence
- Business Systems Engineering
- Business Information Security
- Information Logistics

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: http://www.igi-global.com/publish/.

The Advances in Business Information Systems and Analytics (ABISA) Book Series (ISSN 2327-3275) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit http://www.igi-global.com/book-series/advances-business-information-systems-analytics/37155. Postmaster: Send all address changes to above address. 

## Titles in this Series

*For a list of additional titles in this series, please visit: www.igi-global.com*

*Data Envelopment Analysis and Effective Performance Assessment*
Farhad Hossein Zadeh Lotfi (Islamic Azad University, Iran) Seyed Esmaeil Najafi (Islamic Azad University, Iran) and Hamed Nozari (Islamic Azad University, Iran)
Business Science Reference • copyright 2017 • 365pp • H/C (ISBN: 9781522505969) • US $160.00 (our price)

*Enterprise Big Data Engineering, Analytics, and Management*
Martin Atzmueller (University of Kassel, Germany) Samia Oussena (University of West London, UK) and Thomas Roth-Berghofer (University of West London, UK)
Business Science Reference • copyright 2016 • 272pp • H/C (ISBN: 9781522502937) • US $205.00 (our price)

*Automated Enterprise Systems for Maximizing Business Performance*
Petraq Papajorgji (Canadian Institute of Technology, Albania) François Pinet (National Research Institute of Science and Technology for Environment and Agriculture, France) Alaine Margarete Guimarães (State University of Ponta Grossa, Brazil) and Jason Papathanasiou (University of Macedonia, Greece)
Business Science Reference • copyright 2016 • 312pp • H/C (ISBN: 9781466688414) • US $200.00 (our price)

*Improving Organizational Effectiveness with Enterprise Information Systems*
João Eduardo Varajão (University of Minho, Portugal) Maria Manuela Cruz-Cunha (Polytechnic Institute of Cávado and Ave, Portugal) and Ricardo Martinho (Polytechnic Institute of Leiria, Portugal & CINTESIS - Center for Research in Health Technologies and Information Systems, Portugal)
Business Science Reference • copyright 2015 • 318pp • H/C (ISBN: 9781466683686) • US $195.00 (our price)

*Strategic Utilization of Information Systems in Small Business*
M. Gordon Hunter (The University of Lethbridge, Canada)
Business Science Reference • copyright 2015 • 418pp • H/C (ISBN: 9781466687080) • US $195.00 (our price)

*Enterprise Management Strategies in the Era of Cloud Computing*
N. Raghavendra Rao (FINAIT Consultancy Services, India)
Business Science Reference • copyright 2015 • 359pp • H/C (ISBN: 9781466683396) • US $210.00 (our price)


www.igi-global.com

701 E. Chocolate Ave., Hershey, PA 17033
Order online at www.igi-global.com or call 717-533-8845 x100
To place a standing order for titles released in this series,
contact: cust@igi-global.com
Mon-Fri 8:00 am - 5:00 pm (est) or fax 24 hours a day 717-533-8661

# Editorial Advisory Board

# Table of Contents

# Detailed Table of Contents

be too big and for the same the big data concept is of primarily utmost significance. The authors also proposed a framework which clearly depicts the usage of the opinion mining on operation management of various domains.

**Chapter 6**



The philosophical interpretation of operation management in the process of delivering a 360 degree digitized educational system requires a core approach of supervision, scheming and finally controlling the mechanism of delivery the educational service and its constitutes operations. In today's world of telecommunication, internetworking, cloud infrastructure and computational advancement, a collaborative mechanism of intelligent and reachable educational system is being conceptualized which provoke a notion of "one world one university". The chapter aim to illustrate different source and applications of BigData Generation in educational system operations, existing mechanism to store and process these generated data, limitations of existing mechanisms and conceptualization of a novel framework. Further it aims to discuss different intelligent analytics approach to improvise the teaching-learning philosophy.

**Chapter 7**



Management of data for an organization is crucial task but when data goes to its complex form then it becomes multifaceted as well as vital. In today era most of the organizations generating semi structured or unstructured data that requires special techniques to handle and manage. With the needs to handle unstructured data, semantic web technology provides a way to come up with the effective solution. In this chapter Synthetic Semantic Data Management (SSDM) is explained that is based semantic web technique and will helps to manage data of small and Midsized Enterprise (SME). SSDM provide the procedure to handle, store, manages and retrieval of semi structured data.

# Foreword

Operations Management is an area of management associated with control and design of production of goods and services. It guarantees business operations are efficient by a few resources needed, and are effective by meeting requirements from customers. As the availability of sensors and smart metering, data sets are growing rapidly in the field of operations management. Nevertheless, big data is more than high-volume data, it is even more complicated (both structured and unstructured). Challenges arise for the analytics for the big data in operations management, and this book aims to give a neat and creative solution to it.

My colleagues and I are current implementing researches over big data in medical image for a long period, and we have once read papers published by the authors. I am impressed by his achievements, which makes me accept the invitation to review this book. All chapters are peer-reviewed by at least two independent reviewers or editors. The authors reply to the comments promptly and accurately.

In total eight chapters are finally included in this book. Chapter 1 introduces a revolutionary concept of "big data analytics", which helps to predict behaviors of customer purchase. It can help build the customer-centric business. Chapter 2 presents an application of artificial neural networks in predicting the degradation of tram tracks using maintenance data. Chapter 3 develops a ZAMREN big data management to envisage efficiency and analytically manage IT resources. Chapter 4 offers a predictive analytics in operations management. Chapter 5 discusses the advantages and disadvantages from a big data perspective of applying opinion mining on operation management. Chapter 6 establishes a conceptual framework for educational system operation management synchronous with big data approach. Chapter 7 manages SME's semi-structured data via semantic technique. Finally, Chapter 8 gives the overview of big data security.

We hope this book offers a comprehensive and timely view of the area of emerging trends in big data analytics and its application in operations management, and that it will offer stimulation for further research.

*Yudong Zhang*
*Nanjing Normal University, China*

# Preface

Management, storage and information retrieval from huge amount of data require a paradigm shift in the way data is handled. This huge amount of structured, semi-structured or unstructured data can be termed as Big Data. Big Data can be characterized by 5V and these V's pose ultimate challenge in terms of storage, processing, computation, analysis, and management.

- **Volume:** This refers to the amount of data been generated from different sources such as data logs from twitter, click streams of web pages and mobile apps, sensor-enabled equipment capturing data, etc.
- **Velocity:** This refers to the rate at which data is generated and received. For example for an effective marketing offer to a consumer, ecommerce applications combine mobile GPS location and personal preferences.
- **Variety:** This refers to various types of structured, unstructured and semi-structured data types. Unstructured data consist of files such as audio and video. Unstructured data has many of the requirements similar to that of structured data, such as summarization, audit ability, and privacy.
- **Value:** This refers to the intrinsic value that the data may possess, and must be discovered. There are various techniques to derive value from data. The advancement in the recent years have led to exponential decrease in the cost of storage and processing of data, thus providing statistical analysis on the entire data possible, unlike the past where random samples were analyzed to draw inferences.
- **Veracity:** This refers to the abnormality in data. Veracity in data analysis is one of the biggest challenges. This is dealt with by properly defining the problem statement before analysis, finding relevant data and using proven techniques for analysis so that the result is trustworthy and useful. There are various tools and techniques in the market for big data analytics. Hadoop is Java-based programming framework that supports processing of large data sets. It was started out as a project by Yahoo to analyze its data.

Operations management can be termed as the area of management concerned with designing, controlling and supervising the process of production and reconstruction of business operation in the production of goods and services. The same management process is also applicable to the software development. The data generated is huge in volume and can be structured or unstructured and satisfies all the characteristics of big data. Organizations need big data analytics for better decision making and sustain competitive advantage in the market.

There is need of analyzing huge amount of data for better decision making. There can be predictive analysis, classification, clustering, or some other analytical techniques applied to this data. Big Data management requires a scalable architecture for performing the analytical tasks on it. There are some solutions available for management, storage and information retrieval from this huge amount of data like:

- Hadoop Distributed File System,
- MapReduce,
- Apache Spark, and
- HBase.

The objective of this book is to address different challenges in operation management and how to overcome these challenges using Big Data analytics. The application of the tools, techniques discussed in this book is related to the following areas:

- Software development process,
- Social networks analysis,
- Semantic analysis,
- Predictive analytics,
- Education System Analysis,
- Transport analysis, and
- Cloud computing.

The impact of this edited book is focused on big data analytics and its application in operation management. It is evident that data rate is very high in today's world and growing need of analysis in decision making is important area to focus upon. The content of this book will benefit the researchers in the area of information technology and operations management. The potential readers of this book are

1. The researchers, academician and engineers will be able to know the growing applications of big data analytics in operation management.
2. The industry persons dealing in data storage and its management will get to know how to analyze the big data.

3. The students of various universities studying information technology, computer science and management.

The intended audiences include:

1. Academician,
2. Researchers,
3. Engineers,
4. Industry professionals, and
5. Business persons.

Chapter 1 introduces big data and its 5V. It focuses on real time data collection, real time event processing, comprehensive data collection and deterministic data collection. The chapter discusses various approaches with use-cases to big data operations management. Authors present different approaches for big data operations management: SPLUNK approach, Reflex system approach, Cloud physics approach and XANGATI approach. These approaches perform the collection of heterogeneous data from various resources and perform Map reduce task on this data. SPLUNK has indexing, logs, drill down as its features. Reflex approach can be used for real time analysis by using virtualization query language.

Chapter 2 covers the application of analytical techniques on transport organization data. It discusses Artificial Neural Networks and a model to predict Gauge Widening. The chapter explains preprocessing of data and techniques for obtaining insight of various factors resulting track degradation and track maintenance. The problem of efficient maintenance of track is proposed to solve by a model which is required to predict the degradation of the tracks. The following is presented:

- Analyzing and trending the tram track attributes/variables over a period of time.
- The correlation between those variables and track deterioration will be identified.
- A model will be developed to predict the track deterioration/degradation based on tram track variables.
- The maintenance and replacement activities will be identified and a priority list for reparation of track rails will be prepared. This will lead to minimizing the maintenance costs and preventing unnecessary maintenance actions and therefore saving time

Chapter 3 introduces ZAMBiDM which envisaged to collectively bringing all the resources from the ZAMREN member institutions so that big data can be managed. It discusses various Big Data technologies employed in industry. It proposes a ZAMBiDM model and explains the functions of the system components. It also includes ZAMBiDM's operations and highlights the benefits of envisaging such a model. The ZAMBiDM is developed in-line with the following objectives:

- Design and build the ZAMBiDM architecture.
- Road map of ZAMBiDM.
- Acquire servers to accommodate large volumes of data.
- Collect and store large heterogeneous volume of data from academic and industry sectors including the NRENs worldwide.
- Manage the "V's" of data and these are volume, velocity, variety, value and veracity.
- Build operational processes.
- Develop relevant strategies for the respective operation nodes.
- Elevate ZAMBiDM data to executive level.

There are several challenges while applying analytical and mining technique. Implementation of a mining algorithm in parallel requires certain modification in the basic sequential approach. Authors discuss the challenges like:

- Division of a problem into the sub problems.
- Deployment of the problem on a grid or cluster.
- Load balancing.

Chapter 4 discusses predictive analytical techniques and their implementations. Authors present regression technique, Machine learning technique, random forest etc. Chapter provides a detailed explanation of linear regression model, Parallel Back propagation, parallel support Vector Machine.

Chapter 5 discusses the pros and cons of applying opinion mining in operations management from the big data perspective. Semantic analysis also known as opinion mining is used on text generated by systems for finding the opinion of the users. In this analysis of opinionated text that contains people's opinions toward entities such as products, organizations, individuals, and events. Opinion mining has been applied to a number of domains like hotels and restaurants, different products, movies and politics. Not only had this, but the ever growing growth of information on social media platforms influenced many companies to use this analysis in the operational management as well. Authors also include advantages and disadvantages of opinion mining in operational management.

Chapter 6 discusses the operational management in educational system and its benefits, it also explains some existing operational techniques to store and process big data. Authors propose a conceptualized framework of novel educational operational management. It is highly essential that operational management pertaining to educational system be studied very closely. The functions of operations may again include various essential activities like prediction, effective planning of capacity, data storage etc. Chapter explains the adaptation of information and communication technologies into the education system. Chapter also discusses complexities which exist in the education system in context of operation management. Chapter derives the relationship between operation management in education system and the big data analytics. It also provide information about the solving these problems using the big data analytical techniques.

Chapter 7 discusses a synthetic semantic data management approach for managing the data of small and medium size enterprises which can analyze customer and business intelligence. This chapter presents the survey of big data and semantic technology. It also discusses the semantic web languages, Synthetic Semantic Data Management (SSDM) and its implementation. This uncovers hidden facts, unknown relations, customer requirements and business needs. Small and medium size enterprise generates data which can be used for analysis so that better decision can be made. The chapter provides an explanation of using different scripting languages like OWL for managing small and medium size enterprise data.

Chapter 8 presents an overview of big data security referencing Hadoop framework. Authors discuss the possible solutions related to privacy and security of big data environment. Since security of data is main concern of every organization because of data sensitivity and its usefulness in decision making process. The hostile nature of digital data itself poses security challenges. This chapter tries to uncover big data security issues and to find the better solution for handling security challenges. The observation and analysis of different security mechanism are presented.

*Manish Kumar*
*Indian Institute of Information Technology, Allahabad, India*

# Acknowledgment

First of all, I would like to thank the Almighty.

Acknowledgement when honestly given is a reward without price. I am thankful to the publishing team at IGI Global accepting to publish this edited book. I thank Janine Haughton, development editor for helping me in bringing this project to a successful end. I would also thank Jan Travers, Kayla Wolfe, Meghan Lamb and Jacqueline Sternberg of IGI Global team for their various supports.

I must thank all the esteemed members of Editorial Advisory Board and all the reviewers whose valuable contribution is indispensible. I acknowledge and thank all the authors who have contributed in this book. I take this opportunity to extend my sincere gratitude to Prof. Yudong Zhang for accepting to write the foreword for this book and took his valuable time to go through the content of the book. I would like to thank Prof. Somenath Biswas and Prof. Shekhar Verma for their continuous encouragement and suggestions. I would like to acknowledge my research students for their support during entire process of this book.

I express my deep gratitude to my parents for their moral support, blessings and encouragement throughout my life. I would also like to thank my mother-in-law for all her blessings. I am greatly indebted to my wife and best friend for her constant support in every aspect of life and also responsible for encouraging me to finish this book. She has paid the real cost for this work by sacrificing many of her desires and carrying out our combined social responsibilities alone. I want to acknowledge her contributions and convey my sincere thankfulness. I am heartily thankful to Mahi and Gauri, my lovely daughters and inspiration, blessed by God. They always generate enthusiasm and spark in my work by their warm welcome and blossom smile at home.

The road towards completion of this edited book has been experienced as journey with responsibility, dedication and willing to give something valuable to the knowledge based society. I hereby acknowledge all the people who helped me and shared this memorable time. I would never have been able to complete this challenging task without their support and encouragement.

*Manish Kumar*
*Indian Institute of Information Technology, Allahabad, India*

# Chapter 1
# Big Data in Operation Management

**Arushi Jain**
*Ambedkar Institute of Advanced Communication Technology and Research, India*

**Vishal Bhatnagar**
*Ambedkar Institute of Advanced Communication Technologies and Research, India*

## ABSTRACT

*The word big data analytics have been increased substantially these days, one of the most prominent reasons is to predict the behavior of the customer purchase. This analysis helps to understand what customer wants to purchase, where they want to go, what they want to eat etc. So that valuable insights can be converted into actions. The knowledge thus gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them. This is the revolutionary change to build a customer-centric business. To build a customer centric business an organization must be observant about what customer is doing, must keep a record about what customer is purchasing and lastly should discover the insights to maximum the profit for customer. In this chapter we discussed about various approaches to big data management and the use cases where these approaches can be applied successfully.*

DOI: 10.4018/978-1-5225-0886-1.ch001

## INTRODUCTION

Big data is the revolutionary world in the field of information technology because of its enormous influence on different domain. Big data is the voluminous and complex collection of data that comes from different sources such as sensors, content posted on social media website, sale purchase transaction etc. Such voluminous data becomes tough to process using ancient processing application. By 2020, International data corporation (IDC) predicts the number will have reached 40 Zettabytes (ZB) that is the world will produce 50 times the amount of information. A huge surge in the amount of data being generated that needs to be stored and analyzed quickly has been witnessed in the recent years. For example walmart handles millions of sale purchase transactions per hour, Facebook handles 40 billion photos uploaded by its users each day. Organizations are using big data to analyze and find insights to build an optimal information system. Big Data can be defined using five V's. These are:

- **Volume:** This refers to the amount of data been generated from different sources such as data logs from twitter, click streams of web pages and mobile apps, sensor-enabled equipment capturing data, etc.
- **Velocity:** This refers to the rate at which data is generated and received. For example for an effective marketing offer to a consumer, ecommerce applications combine mobile GPS location and personal preferences.
- **Variety:** This refers to various types of structured, unstructured and semi-structured data types. Unstructured data consist of files such as audio and video. Unstructured data has many of the requirements similar to that of structured data, such as summarization, audit ability, and privacy.
- **Value:** This refers to the intrinsic value that the data may possess, and must be discovered. There is wide variety of techniques to derive value from data. The advancement in the recent years have led to exponential decrease in the cost of storage and processing of data, thus providing statistical analysis on the entire data possible, unlike the past where random samples were analyzed to draw inferences.
- **Veracity:** This refers to the abnormality in data. Veracity in data analysis is one of the biggest challenges. This is dealt with by properly defining the problem statement before analysis, finding relevant data and using proven techniques for analysis so that the result is trustworthy and useful. There are various tools and techniques in the market for big data analytics. Hadoop is Java-based programming framework that supports processing of large data sets. It was started out as a project by Yahoo to analyze its data but now it is part of the Apache project.

## LITERATURE SURVEY

To manage the growing demands, there is a need to increase the capacity and performance of tools and methods employed for analysis of data. Chen et al. (2014), in their work "Big data: A survey" focused on big data and reviewed related technologies and examined the application of big data in various fiels. Al-Jarrah et al. (2015), in their work "Efficient Machine Learning for Big Data: A Review" reviewed the data modeling in large scale data intensive field relating to model efficiency and new algorithm approaches. Hoffmann and Birnbrich (2012) to protect their customer from third party fraud proposed a conceptual link between retail bank activities in "The impact of fraud prevention on bank-customer relationships: An empirical investigation in retail banking". Srivastava and Gopalkrishnan (2015) revealed some of the best techniques which are used by the banks across the globe and can be used by the Indian banks to enhance their services offerings to the customers in "Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks". Azar and Hassanien (2014) for dimensionality reduction presented a linguistic hedges neuro-fuzzy classifier with selected features (LHNFCSF). In this paper author compared the new classifier with the other classifiers for various classification problems in "Dimensionality reduction of medical big data using neural-fuzzy classifier". Hassanien et al. (2015) focused on application, challenges and opportunities of big data in "Big Data in Complex Systems: Challenges and Opportunities". Wahi et al. (2014) proposed a social media and its implication on customer relationship management in "Social Media: The core of enterprise 2.0.". Shabeera and Madhu Kumar (2015), in their work "Optimizing virtual machine allocation in MapReduce cloud for improved data locality" focused on improving data locality by allocating virtual machines for executing map reduce jobs. Aloui and Touzi (2015) proposed a methodology for designing ontology on a new platform called "FO-FQ Tab plug-in" and then querying them smartly based on conceptual clustering and fuzzy logic in "A Fuzzy Ontology-Based Platform for Flexible Querying". Ghallab et al. (2014), in their work "Strictness petroleum prediction system based on fussy model" predicted the status of crude oil and then compared it with other petroleum values. Huang et al. (2015) summarized the latest application of big data in health science. The authors also reviewed the latest technologies of big data and discussed the future perspective of health sciences in "Promises and challenges of big data computing in health science". Jagadish (2015) in "Big Data and Science: Myths and Reality" explored myths about big data and exposed the underlying truth. Jin et al. (2015) introduced the concept of big data and described the challenges as well as solution to these challenges in "Significance and challenges of big data research". Ryan and Lee (2015) presented a Multi-tier resource allocation as resource management

technique for distributed systems in "Multi-tier resource allocation for data-intensive computing". Tiwari and Joshi (2015) in "Data security for software as a service" discussed security vulnerabilities of software as a service (SaaS) and its solution. Wahi et al. (2015) focused on whether the organization could able to address challenges posed by big data successfully or not. It also focused on the reasons why it is necessary to transit from the enterprise 1.0 stage to enterprise 2.0 stage in "Big Data: Enabler or Challenge for Enterprise 2.0. Deepak and John (2016) illustrated that information system is one of the most significant problem in fuzzy domain. Authors illustrated a case where hesitant membership value arrived from attribute value whose membership values are a family of set and also discusses the homomorphism between hesitant information systems in "Information Systems on Hesitant Fuzzy Sets". Bhanu and Tripathy (2016) in "Rough Set Based Similarity Measures for Data Analytics in Spatial Epidemiology" carried out epidemiological studies to understand a pattern and transmission of disease instances.

## RESEARCH ROADMAP

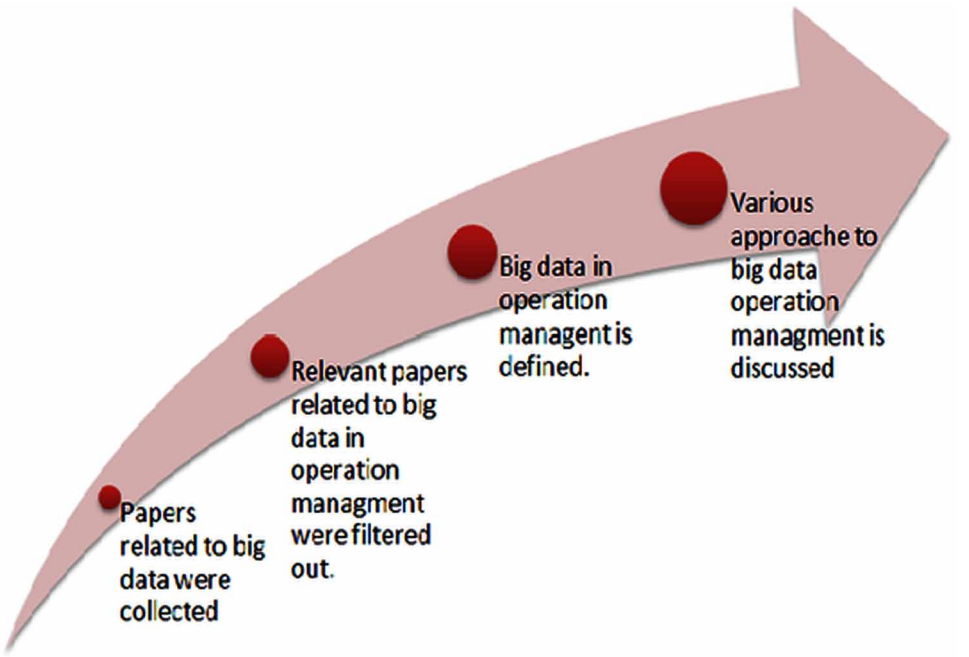
To manage the growing demands of this day and age, there is a need to increase the capacity and performance of tools and methods employed for analysis of data. Big Data requires new solutions to improve the capacity and to exploit the inherent value it brings with itself. Indeed, with the exponential growth of data, traditional data mining algorithms have been unable to meet important needs in terms of data processing, but it can be very costly and time taking for development of hardware which can handle such loads. Also, the hardware thus developed may become insufficient in a matter of months, given the rate at which the data is increasing and evolving. In order to exploit this voluminous data without much financial and time overhead, efficient processing model with a reasonable computational cost of this huge, complex, highly dynamic and heterogeneous data set is needed.

The above Figure 1 defines the methodology which was adapted to carry out the research work, Firstly the papers relevant to big data were collected, then papers related to big data operation management were filtered out followed by the brief discussion of big data in operation management and various approaches to big data operation management

## INTRODUCTION TO BIG DATA

Since the birth of the computers, improving the computing performance and the processing speed of the computers has been an omnipresent topic. Ever since the

*Figure 1. Research roadmap*




advent of big data, it has gained even more voice. Because of the limitation of physical devices, parallel computing was conceptualized and brought to life. It has gained a lot of fame since its advent and has now become a comprehensive field in computer science. The discipline includes the hardware concurrent designing and software concurrent designing. According to the hardware concurrent design for a computer, the different types of hardware architectures can be divided into single instruction stream and data stream (SISD); single instruction multiple data (SIMD) flows; multiple instruction stream data flow (MISD); the instruction stream and data stream (MIMD). The software concurrent design on the other hand includes the algorithms of concurrent design, high-speed internet clustering framework, and high-efficiency concurrent calculation model. Hardware concurrent design is skilled in numerical analysis and the floating point precision but it faces a technical bottleneck in case of big data, especially when the data is unstructured, while the software concurrent design, which includes parallel computing model and cluster system design, has been very successful in making up for the defects above.

For example, cluster system is constituted with the workstation or PC, which is linked via a high-speed network with a certain optimizing structure, and a unified dispatching visual human-computer interface, thus it is an efficient parallel processing system. These cluster systems have the good scalability and the system

development cycle is little short. With a good programming model, the system can be very successful in coping with big data.

Hadoop is one of the most popular cluster systems, which includes the high-performance calculation model Map Reduce and Spark. Hadoop is an excellent and robust analytics platform for Big Data which can process huge data sets at a really quick speed by providing scalability. It can manage all aspects of Big Data such as volume, velocity and variety by storing and processing the data over a cluster of nodes whose numbers can run into thousands. The best thing about it is that the nodes in the cluster need not possess complex and costly hardware but simply commodity hardware which neither costly nor difficult to procure. Hadoop has a proper architecture, which consists of two components, MapReduce and the Hadoop.

MapReduce is a programming paradigm which can run parallel processing tasks on nodes in a cluster simultaneously and is able to both process and generate really large data sets with ease. It takes input and gives the output in form of key-value pairs. MapReduce is able to achieve all this by simply dividing each problem into two broad phases, the map phase and the reduce phase. The Map phase processes each record sequentially and independently on each node and generates intermediate key-value pairs. The Reduce phase takes the output of the Map phase. It processes and merges all the intermediate values to give the final output, again in form of key-value pairs. The output gets sorted after each phase, thus providing the user with the aggregated output from all nodes in an orderly fashion. This can be easily used for clustering large data sets. MapReduce programs have to be written for the query and the classification takes place on a cluster of interconnected computers, whose numbers may run into thousands.

The Hadoop distributed file system (HDFS) is a highly scalable and distributed file system for Hadoop. HDFS is able to store file big in size across multiple nodes in a cluster. It is reliable as is replicates the data across the nodes, and hence theoretically does not require Redundant Array of Integrated Devices (RAID) storage. The default replication value is 3, that is, the data is stored on 3 nodes. Two of these are on the same rack and one is on a different rack. The jobs are done by two different tasks, the Job tracker (master) and the Task tracker (slave). The Job tracker schedules MapReduce jobs to the Task trackers which know the data location. The intercommunication among tasks and nodes is done via periodic messages called heartbeats. This reduces the amount of traffic flowing through the network. Hadoop can also be used with other file systems, but then these advantages are not available

With the introduction of YARN (Yet Another Resource Negotiator) in later releases, Hadoop was integrated with a number of wonderful projects which can be used for storing, processing and analyzing data a lot more efficiently, thus aiding

in exploration of data for undiscovered facts at a smooth pace. Hadoop is an open source project written in Java and is currently under the Apache Software Foundation and is free to use. The initial design idea is mainly used to deploy on the cheap hardware. The framework implements a distributed file system, referred to as "HDFS," which has high fault tolerance and high speed and scalability to store large amounts of data and implements a computation model of parallel processing large data sets; this model is of high speed computing in big data processing field. In addition to these advantages, the distributed system framework is open source software. With the advent of big data, it has become a good and reliable solution, which can store and process big data with ease. While in this day and year it has become a famous big data processing framework in the field of the big data analytics, at the same time, Hadoop ecosystem, with continuous improvement and optimization, has become better. With the advent of second generation, a lot more projects have been introduced into the hadoop ecosystem.

After collecting the data, it has to be processed so that meaningful information can be extracted out of it which can serve as decision support system. Therefore the analysts need to come up with a good technique for the same. One way of achieving this is MapReduce, it permits the filtering and aggregation of data stored in the HDFS so as to gain knowledge from the data. However, writing MapReduce requires basic knowledge of Java along with sound programming skills. Assuming one does possess these skills, even after writing the code, which is itself a labor intensive task, an additional time is required for the review of code and its quality assessment. But now, analysts have additional options of using the Pig Latin or Hive-QL. These are the respective scripting languages to construct MapReduce programs for two Apache projects which run on Hadoop, Pig and Hive. The benefit of using these is that there is a need to write much fewer lines of code which reduces overall development and testing time. These scripts take just about 5% time compared to writing MR programs. Although Pig and Hive scripts are 50% slower in execution compared to MR programs, they still are very effective in increasing productivity of data engineers and analysts by saving lots of time. Pig is a Hadoop project which was initially developed by Yahoo! and is now open source and free to use under Apache. Its scripting language is called Pig Latin. The script consists of data flow instructions which are converted to MapReduce instructions by its framework and used to process data. Hive is a Hadoop project which was initially developed by facebook and it is now an open source project under Apache. It has its own UI called Beeswaxand, its script is called Hive-QL which is very similar to the SQL being a declarative dialect. The script is then converted to MR instructions and the data is processed. Hive helps in querying and managing large dataset. Pig Latin is the

scripting language to construct MapReduce programs for an Apache project which runs on Hadoop. The benefit of using this is that there is a need to write much fewer lines of code which reduces overall development and testing time. However in order to improve the performance of the distributed system framework, many companies and supporters provide their first-class good components and high performance code for Hadoop, such as YARN, Hcatalog, Oozie, and Cassandra, which make the performance of Hadoop become more and more strong, and the application field of Hadoop is becoming more and more wide.

## Example

Finding maximum word count through Map reduce programming

Input: Text File
Output: Total word count
Mapper phase: (see Figure 2)
Reducer Phase: (see Figure 3)
Driver Class: (see Figure 4)
Output: (see Figure 5)

R language, as a data analysis and visualization tool, is widely used for statistical analysis, drawing, data mining, machine learning, and visualization analysis. Particularly, R tool has been built in a variety of statistical, digital analysis functions and drawing functions. R language is an open source software, and very useful

*Figure 2. Mapper class*

```
public class WordCount {

        public static class Map extends Mapper<LongWritable,Text,Text,IntWritable>{

                public void map(LongWritable key, Text value,
                                Context context)
                                throws IOException,InterruptedException {

                        String line = value.toString();
                        StringTokenizer tokenizer = new StringTokenizer(line);

                        while (tokenizer.hasMoreTokens()) {
                                value.set(tokenizer.nextToken());
                                context.write(value, new IntWritable(1));
                        }


                }
```

*Figure 3. Reducer class*

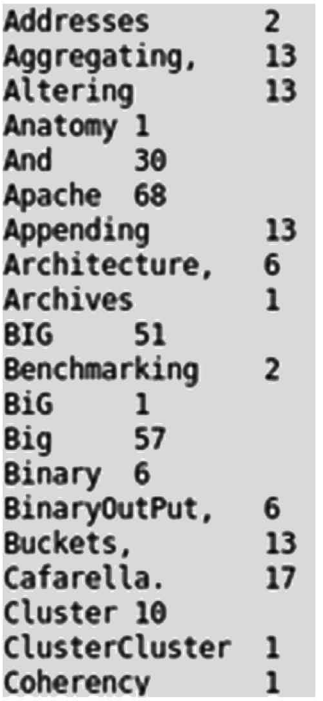
```
public static class Reduce extends Reducer<Text,IntWritable,Text,IntWritable>{

        public void reduce(Text key, Iterable<IntWritable> values,
                        Context context)
                        throws IOException,InterruptedException {
                int sum=0;

                for(IntWritable x: values)
                {
                        sum+=x.get();
                }
                context.write(key, new IntWritable(sum));

        }
}
```

*Figure 4. Driver class*

```
public static void main(String[] args) throws Exception {

        Configuration conf= new Configuration();
        Job job = new Job(conf,"mywc");

        job.setJarByClass(WordCount.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);


        Path outputPath = new Path(args[1]);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

        outputPath.getFileSystem(conf).delete(outputPath);

        System.exit(job.waitForCompletion(true) ? 0 : 1);
```

third-party packages which are written by the developers in it's community around the world can be downloaded and used for free. Therefore, its application field is wide, from statistical analysis, to applied mathematics, econometrics, financial analysis, financial analysis, human science, data mining, artificial intelligence, bioinformatics, biomedical, data visualization and many more (see Figures 2-5).

*Figure 5. Output*




## BIG DATA OPERATION MANAGEMENT

Big data operation management system is a virtualized ecosystem consisting of cloud and operational applications to understand what the growing concern concept is across IT systems.

Big data operation management = clustering analysis of big data + collaboration filtering of big data

For example, through recommendation system a user can choose item to purchase, subscribe, invest, or any other venture that the user may need suggestions for through building a recommendation system. Such systems may be personalized for the user to base recommendations based on data specific to the user. A great way of building a reliable recommender system for the users is collaborative filtering. Collaborative filtering is defined as a technique that filters the information sought by the user and patterns by collaborating multiple data sets such as viewpoints,

multiple agents and pre-existing data about the users' behavior stored in matrices. Collaborative filtering is especially required when a huge data set is present. The collaborative filtering methods are used to create recommender systems for a wide variety of fields with lots of data having varied formats such as sensing and monitoring of data in battlefields, line of controls and mineral exploration; financial data of institutions that provide financial services such as banks and stock markets; sensing of large geographical areas from which data is received from all kinds of sensors and activities; ecommerce and websites where the focus is to recommend products to users to increase sales, to name a few.

A definition of collaborative filtering which is somewhat newer and a bit narrow in sense states that it is a way of automating the process of making predictions, a process which is known as filtering, about the preferences and dislikes of a user by collecting data from as big a number of users as possible, a process which is known as collaborating, hence it is given the name collaborative filtering. The underlying assumption of the collaborative filtering approach is that if a person has the same opinion of an issue as a person B, A is more likely to have an opinion similar to B's opinion on a related but different issue. It is noteworthy that such predictions are specific to the user, but using data from a number of users forms them. The personal information of the user such as age, gender and location are generally not used in collaborative filtering (CF) but a partially observed matrix of ratings is used. The rating matrix may be binary or ordinal. The binary matrix contains the ratings by the users in columns in the form of likes or dislikes while the user' name or id is in the rows. The ordinal matrix contains ratings in form of a number of responses form the user such as excellent, very good, good, average, poor or simply in form of stars out of five or ten, a system that is used a lot in this day and age. The website's server, for example using click stream logging, can easily gather the rating matrix implicitly. Clicks on links to pages of goods or services being provided can be considered to be positive review of the user. While rating matrices can prove to be really handy, one major drawback they have is that they are extremely sparse, so it is very difficult of club similar users together in classes. This is so because each and every user does not give the reviews about each and every product. Thus collaborative filtering consists of storing this sparse data and analyzing it to create a recommendation system.

Cluster analysis or clustering is the exercise of taking a set of objects and dividing them into groups in such a way that the objects in the same groups are more similar to each other according to a certain set of parameters than to those in other groups. These groups are known as clusters. Cluster analysis is one of the main tasks in the field of data mining and is a commonly used technique for statistical analysis of

data. Cluster analysis does not refer to an algorithm but an exercise that has to be undertaken on the given data set. Various algorithms can be used for cluster analysis. The algorithms are divided into various categories and they differ significantly in their idea of what a cluster is constituted of and how the clusters are identified. The most popular ideas on the basis of which clusters are defined and identified include groups with small distances among the constituent members, areas of the data space which are highly dense, intervals or particular distributions. Clustering is a multi-objective problem that it is a mathematical optimization problem. A clustering algorithm consists of parameter settings such as a distance function, a density threshold (the number of clusters expected to be formed). Based on the available data set and the use of result as intended by the user, apt clustering algorithm may be used. For example, It is a perpetual task in search engine to group similar objects into distinct clusters and dissimilar object away from the clusters. Thus, clustering algorithms are an essential constituent for making a well performing search engine. The clustering is used to provide data to the users of the search engine as they post queries to search on various topics. The results are given to the user by the engine according to the similar objects in the particular user's cluster using previously gathered data about preferences of similar users. The better the performance of the clustering algorithm is for the users, the more the chances are that the users are able to find the thing they are looking for on the very first page itself and they don't have to spend time looking up further results. Therefore the definition based on which the algorithm forms clusters and defines objects has to be spot on to get the best results. The better this performance is, the more users are attracted to the search engine (see Figure 6).

Big data operation system focuses on collection of following data (Harzog, 2015):

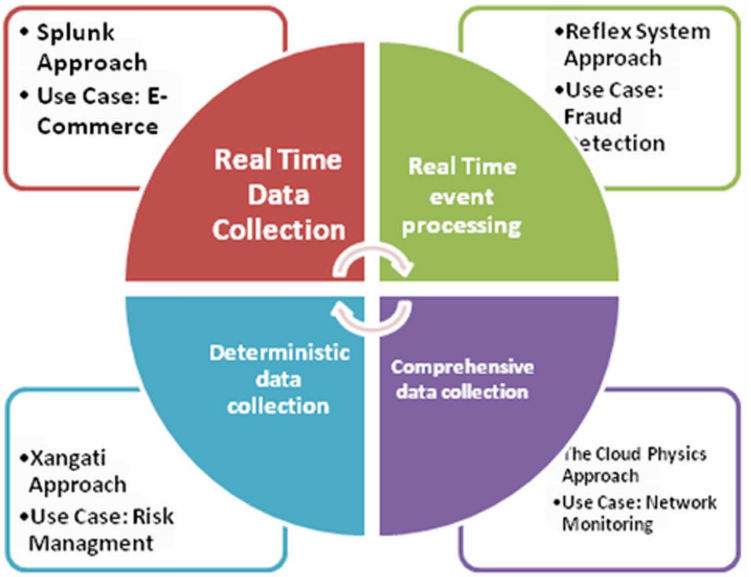
1. Real Time Data Collection
2. Real time event processing
3. Comprehensive data collection
4. Deterministic data collection

## Real Time Data Collection

It captures huge continuous data from remote sensors across the environment while allowing real time analytics by firing queries in the database. The two I/O bottlenecks which must be considered while designing network applications are (Harzog, 2015):

1. **Network I/O:** This is the resource allocation technique which is used to allocate bandwidth to different type of traffic. This feature also supports bandwidth

*Figure 6. Big data operation management*




prioritization depending upon the needs of the different network resources. Network I/O categorize the network traffic on the basis of already existing resource pools: Fault tolerance traffic, Iscsi traffic, Vmotion traffic, Management traffic, sphere Replication traffic, Nfs Traffic and virtual machine traffic.

2. **File system I/O:** In this incoming stream of data is organized, managed and scaled to meet the requirement.The results are then interpreted and evaluated to discern new patterns and trends. The knowledge which is gained can be used for making effective decision and to identify actionable patterns and trend.

## Real Time Event Processing

This includes event by event processing by analyzing continuous stream of real-time data with the help of continuous queries.The knowledge thus gained helps in understandingwhat can be used for making effective decision and to identify actionable patterns and trends(Harzog, 2015). This includes two important set:

1. **Aggregation-Oriented Processing:** In this online algorithm is created so that required processing of the data entering into the system can take place. It is the analytical process for analyzing and exploring large data sets to find hidden rules, associations, patterns among the parts of data, which help to formalize

and plan the future decision making process which is enabled by knowledge featureFor example, for an inbound event calculating the average of the data.

2. **Detection-oriented processing:** In this we look for the specific pattern and behavior of the events so that valuable insights can be converted into actions. This analysis helps to understand what customer wants to purchase, where they want to go on vacations, what they want to eat etc. So that valuable insights can be converted into actions. The knowledge thus gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them.

### Comprehensive Data Collection

Voluminous amount of data needs to be enterinto the computer system in the format which can be understood by the computer. For storing data into data warehouse it needs to follow certain rules while maintaining the consistency and integrity of the data. This helps in discovering, exploring and extracting unseen patterns and relationship among large data sets. Analyzing erroneous data leads to the discovery of wrong patterns and relationship. Thus false information will be gained resulting into the production of imprecise results (Harzog, 2015).

### Deterministic Data Collection

In this we focus on calculating as close value to the actual as possible and not an estimated value. This helps to understand the needs and behavior of growing IT industry in the better way because averaging value sometimes lead to enormous decision.

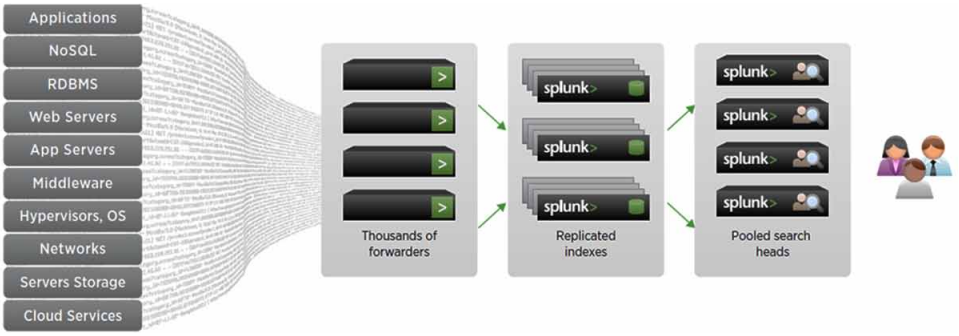
## VARIOUS APPROACHES AND USE CASES TO BIG DATA OPERATION MANAGEMENT

### Splunk Approach to Big Data Operations Management

Architecture of Splunk is based on the three basic principles (Harzog, 2015) (see Figure 7):

- Collection of voluminous data through various distributed sources. data is obtained from different sources having different data formats. It is important to transform the varying data format into the target one.

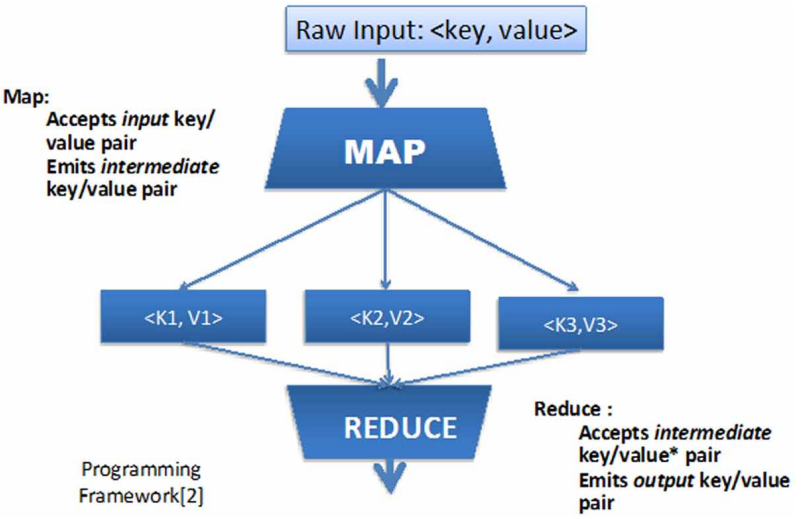
*Figure 7. Splunk approach*
*(Harzog, 2015)*




- Use of Map Reduce (MR) programming technique to process large data sets with a parallel and distributed algorithm. It consists of two phases:
  - **Map Phase:** It assigns the workload by dividing it into small parts and allocates tasks to Mapper, which handles each block of data.
  - **Reduce:** It combines the output of map phase to give the final output. MapReduce is a programming paradigm which can do parallel processing on nodes in a cluster. It takes input and gives the output in form of key-value pairs. MapReduce is able to achieve all this by simply dividing each problem into two broad phases, the map phase and the reduce phase. The Mapper processes each record sequentially and independently in parallel and generates intermediate key-value pairs.
    - **Map(k1, v1) → list(k2, v2)** The Reduce phase takes the output of the Map phase. It processes and merges all the intermediate values to give the final output, again in form of key-value pairs.
    - **Reduce(k2, list (v2)) → list(k3, v3)** The output gets sorted after each phase, thus providing the user with the aggregated output from all nodes in an orderly fashion (see Figure 8).
- Firing appropriate query according to our needs to gain insights .The knowledge which is gained can be used for making effective decision and to identify actionable patterns and trend.

Splunk has the ability of collecting large amount of real time data in a scale out manner which can be used for searching, monitoring and analyzing. It helps to understand the meaning of data facilititating future investigation. Splunk also helps to visualize the data through charts and reports for better understanding and makes the insights more clearer and refiner. The results are then interpreted and evaluated to discern new patterns and trends.

*Figure 8. Map reduce programming framework*




Key features of Splunk are:

- **Indexing:** In this, administrator collects maximum data from different sources and collects it at centralized location with indexes for making them searchable centrally.
- **Logs:** With the help of indexes, splunk can look out for the logs from different data centers and can indicate when a problem could occur.
- **Drill Down:** Splunk then uses drill down approach to determine the root cause of the problem and can generate alert for the future.

## Usecase

### *E-Commerce*

This approach helps to predict the behavior of the customer purchase. This analysis helps to understand what customer wants to purchase, where they want to go, what they want to eat etc. So that valuable insights can be converted into actions. The knowledge thus gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them. This is the revolutionary change to build a customer-centric business. To build a customer centric business an organization must be observant about what customer is doing, must keep a record about what customer is purchasing and lastly should discover the insights to maximum the profit for customer.

## Analysis

E-commerce generally requires:

1. Active participation from users
2. A simple way of representing users' interests to the software system
3. An algorithm to match people with similar interests.

Users with similar rating patterns are identified are clubbed together into classes, so that when one of them is in need of a recommendation, it can be provided using the data of the users in the class. The products matching the preferences of the user are then recommended, based on the data in the ratings matrix. This approach is used by algorithms, which use user-based collaborative filtering for building the recommendation system. A specific application of this is the user-based Nearest Neighbor algorithm.

Alternatively, there is item-based collaborative filtering, which identifies like-minded people from the list of costumers who avail a certain product. It uses an item-centric approach for collaborative filtering using the following basic steps:

1. An item-item matrix is built to determine the relationships between products on the basis of user buying patterns.
2. When the user searches a product, using the previously built matrix identifies the preferences and relevant recommendations are provided.

Relying on scoring or rating systems is not exactly ideal in tasks where there is a large variation in interests and a lot of users' interests may be idiosyncratic. This is so because rating which is averaged across all users ignores demands specific to a particular. This is particularly common in the music and movie industries. However, the developers may use alternative methods to combat information explosion, such as data clustering. Therefore, there is another form of collaborative filtering that does not rely on imposition of artificial behavior by a rating or binary system but uses implicit observations of the behavior of common user for filtering. Such systems log the user's behavior based on previous purchases to identify a pattern, which can be used to group that user with a class of users with similar behavior. This can then be used to provide recommendation, predict future purchases and can even be used to predict a user's behavior in a hypothetical situation, which can be essential for decision support systems. Analyzing each one on the basis of business logic, which helps to determine how a prediction might affect the business system,

then filters these predictions. For example, it is not useful to recommend a user a particular piece of art if they already have demonstrated that they own that artwork or a similar artwork.

## Reflex System Approach to Big Data Operation Management

This system has the ability of collecting large amount of streaming data from various sources. To scale up the vmware environment through the data collecting nodes and graph based mapping techniques this approach is widely used. Through this approach the real time data can be analysis by using virtualization query language (Harzog, 2015). The context can be then wrapped with the help of reflex software and can be presented to the user. This technology provides the framework for processing of distributed data sets across distributed computer clusters which handleactivities like distributed processing and parallel computation. It is cost-effective for handling large data sets with massive scalability and speed.It is an analytical process for analyzing and exploring large data sets to find hidden rules, associations, patterns among the parts of data, which help to formalize and plan the future decision making process which is enabled by knowledge feature (Harzog, 2015) (see Figure 9).

### Use Case

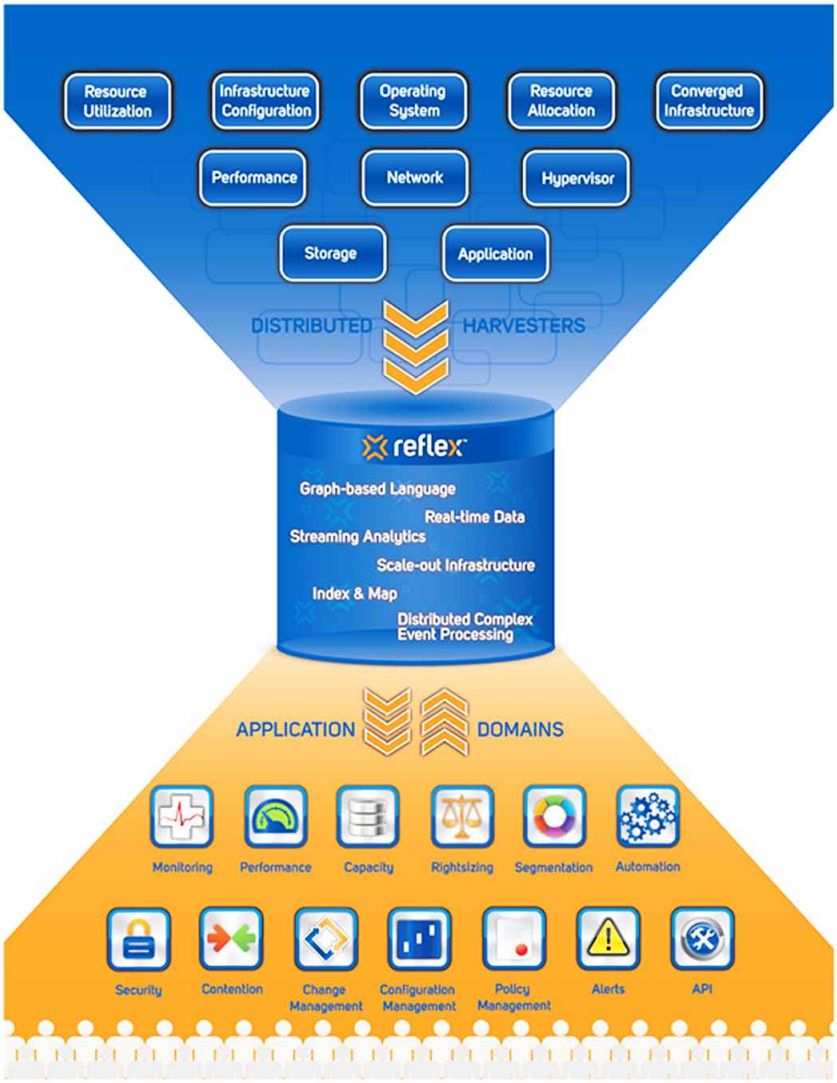
#### *Fraud Detection*

It provides the framework for crime and intelligence analysis and for national security. Fraudulent behavior of individuals and other wrong intentions of the people can be easily detected using these approaches and techniques. The knowledge which is gained can be used for making effective decision and to identify actionable patterns and trend. These approaches focus on extraction of named entities from reports, detecting criminal identities from existing crime datasets and Identifying criminal groups. This helps in discovering, exploring and extracting unseen patterns and relationship among large data sets.To safeguard against fraud and anti crime, organizations are liable to select the best methods to cater their needs.

#### *Analysis*

It tries to determine the amount of purchase, the time of purchase in case of credit card fraudulent. Affinity helps in determining the events which likely to happen with another event.Cluster analysis or clustering is the exercise of taking a set of objects and dividing them into groups in such a way that the objects in the same groups are more similar to each other according to a certain set of parameters than

*Figure 9. Reflex system approach*
*(Harzog, 2015)*




to those in other groups. These groups are known as clusters. Cluster analysis is one of the main tasks in the field of data mining and is a commonly used technique for statistical analysis of data. Cluster analysis does not refer to an algorithm but an exercise that has to be undertaken on the given data set. Various algorithms can be used for cluster analysis. The algorithms are divided into various categories and they differ significantly in their idea of what a cluster is constituted of and how the

clusters are identified. The most popular ideas on the basis of which clusters are defined and identified include groups with small distances among the constituent members, areas of the data space, which are highly dense, intervals or particular distributions. Clustering is a multi-objective problem that it is a mathematical optimization problem. A clustering algorithm consists of parameter settings such as a distance function, a density threshold (the number of clusters expected to be formed). Based on the available data set and the use of result as intended by the user, apt clustering algorithm may be used.

## The Cloud Physics Approach to Big Data Operation Management

This approach is based on hosting and analysis of the big data at the backend and thus facilitating secure hosting (Harzog, 2015). The knowledge gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them. This is the revolutionary change to build a customer-centric business. To build a customer centric business an organization must be observant about what customer is doing, must keep a record about what customer is purchasing and lastly should discover the insights to maximum the profit for customer. Voluminous amount of data enter into the computer system in the format which can be understood by the computer. This helps in discovering, exploring and extracting unseen patterns and relationship among large data sets (Harzog, 2015) (see Figure 10).

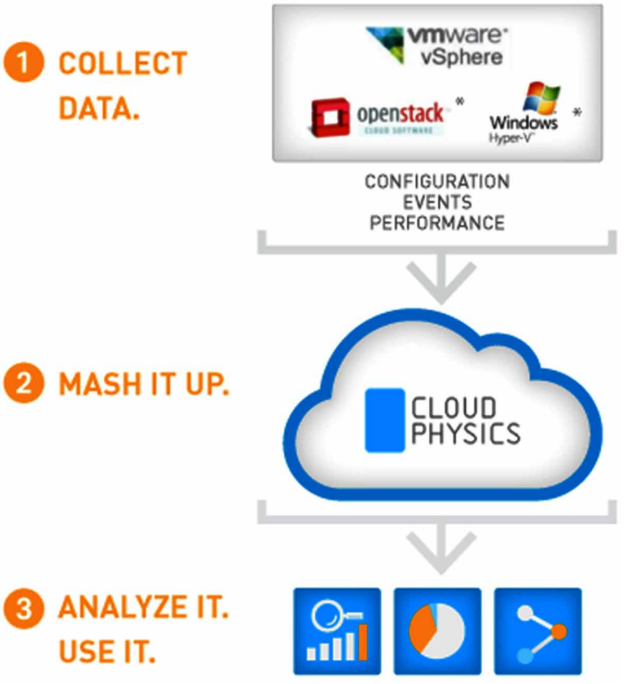
### Use Case

#### *Network Monitoring*

Big data operation management can be used to achieve the following:

1. **Load Balancing:** It helps in managing and analyzing input and output stream for better throughput.
2. **Data Filtering:** It helps in analyzing the activity of specific packet.
3. **Real Time Analytics:** It helps in analyzing and exploring large data sets to find hidden rules, associations, patterns among the parts of data, which help to formalize and plan the future decision making process which is enabled by knowledge feature.

*Figure 10. Cloud Physics approach*
*(Harzog, 2015)*




## Analysis

Cloud physics approach can be vital in analyzing data from wireless sensor networks. One such application can be detection of landmines hidden beneath the soil, especially in civilian regions the algorithms can find clusters of electromagnetic activity in areas where such spikes in radiations are not expected by analyzing satellite images, which contain electromagnetic field sensing data. This can be useful in saving scores of civilian lives by identifying hidden explosives before they can cause any harm.

## Xangati Approach to Big Data Operation Management

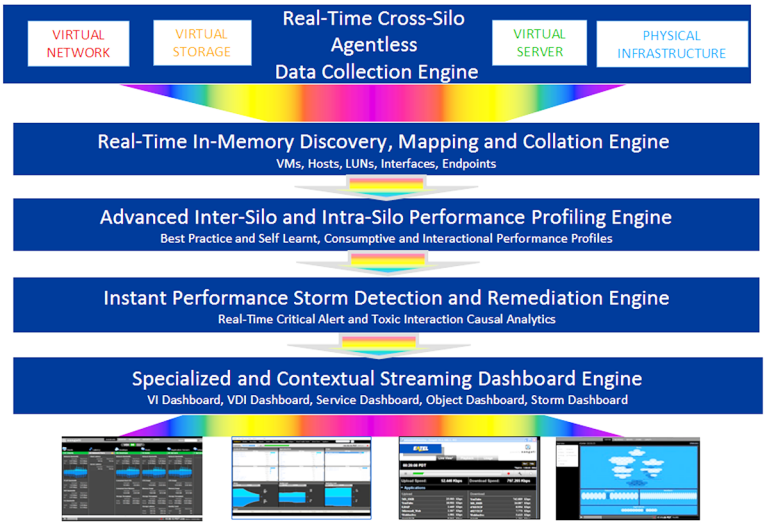
In this approach voluminous amount of data is collected through physical and virtual network (Harzog, 2015). This helps in analyzing in- memory database and thus facilitating in discovering, exploring and extracting unseen patterns and relationship among large data sets. It captures huge continuous data from remote sensors

across the environment while allowing real time analytics by firing queries in the database. This incoming stream of data is then organized, managed and scaled to meet the requirement. The results are then interpreted and evaluated to discern new patterns and trends. The knowledge, which is gained, can be used for making effective decision and to identify actionable patterns and trend.

## Use Case

For identification of cancerous tissue, using cancerous and non-cancerous data set samples does this. These data sets are labeled as samples data sets. Then, both samples are randomly mixed. Then these are applied with different clustering algorithms and accordingly the results are checked for the correct results (since these are known samples and results are known beforehand) and hence the percentage of correct results obtained is calculated. This is the learning phase of clustering mechanism. Now, if for some arbitrary sample data the same algorithm is applied, the result received after analysis of the data set can be expected by the user to have same percentage of accuracy in diagnosing the data set as the accuracy the algorithm achieved in the learning phase on the samples data sets. This approach can be used to find the most suitable algorithm for this analysis. The suspicious data sets from patients can now be analyzed by computers to identify if the disease is present or

*Figure 11. Xangati approach*
*(Harzog, 2015)*



not. Experimentation has revealed that data sets for cancerous tissue can be most accurately analyzed using non-linear clustering algorithms. This can be used to infer that these kinds of data sets are non-linear in nature.

## COMPARATIVE ANALYSES OF VARIOUS APPROACHES

Table 1 describes the comparative analysis of various big data operation management.

*Table 1. Comparative analyses*

| S.no | Approach | Description | Example | Analysis |
|---|---|---|---|---|
| 1 | **Splunk Approach** | Indexing: In this, administrator collects maximum data from different sources and collects it at centralized location with indexes for making them searchable centrally. Logs: With the help of indexes, splunk can look out for the logs from different data centers and can indicate when a problem could occur. Drill Down: Splunk then uses drill down approach to determine the root cause of the problem and can generate alert for the future. | This approach helps to predict the behavior of the customer purchase. This analysis helps to understand what customer wants to purchase, where they want to go, what they want to eat etc. | An item-item matrix is built to determine the relationships between products on the basis of user buying patterns. When the user searches a product, using the previously built matrix identifies the preferences and relevant recommendations are provided |
| 2 | **Reflex System Approach** | This system has the ability of collecting large amount of streaming data from various sources. To scale up the VMware environment through the data collecting nodes and graph based mapping techniques this approach is widely used. Through this approach the real time data can be analysis by using virtualization query language | It provides the framework for crime and intelligence analysis and for national security. Fraudulent behavior of individuals and other wrong intentions of the people can be easily detected using these approaches and techniques. | It tries to determine the amount of purchase, the time of purchase in case of credit card fraudulent. Affinity helps in determining the events which likely to happen with another event.Cluster analysis or clustering is the exercise of taking a set of objects and dividing them into groups in such a way that the objects in the same groups are more similar to each other according to a certain set of parameters than to those in other groups. |

*continued on next page*

*Table 1. Continued*

| S.no | Approach | Description | Example | Analysis |
|---|---|---|---|---|
| **3** | **The Cloud Physics Approach** | This approach is based on hosting and analysis of the big data at the backend and thus facilitating secure hosting | 1) Load Balancing: It helps in managing and analyzing input and output stream for better throughput.<br>2) Data Filtering: It helps in analyzing the activity of specific packet.<br>3) Real Time Analytics: It helps in analyzing and exploring large data sets to find hidden rules, associations, patterns among the parts of data, which help to formalize and plan the future decision making process which is enabled by knowledge feature | Cloud physics approach can be vital in analyzing data from wireless sensor networks. One such application can be detection of landmines hidden beneath the soil, especially in civilian regions the algorithms can find clusters of electromagnetic activity in areas where such spikes in radiations are not expected by analyzing satellite images, which contain electromagnetic field sensing data. This can be useful in saving scores of civilian lives by identifying hidden explosives before they can cause any harm. |
| **4** | **Xangati Approach** | In this approach voluminous amount of data is collected through physical and virtual network.This helps in analyzing in- memory database and thus facilitating in discovering, exploring and extracting unseen patterns and relationship among large data sets | For identification of cancerous tissue, using cancerous and non-cancerous data set samples | Both samples are randomly mixed. Then these are applied with different clustering algorithms and accordingly the results are checked for the correct results (since these are known samples and results are known beforehand) and hence the percentage of correct results obtained is calculated. |

(Harzog, 2015)

Big data operation management system is a virtualized ecosystem consisting of cloud and operational applications to understand what the growing concern concept is across IT systems. This approach helps to predict the behavior of the customer purchase. This analysis helps to understand what customer wants to purchase, where they want to go, what they want to eat etc. So that valuable insights can be converted into actions. The knowledge thus gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them. This is the revolutionary change to build a customer-centric business. To build a

customer centric business an organization must be observant about what customer is doing, must keep a record about what customer is purchasing and lastly should discover the insights to maximum the profit for customer.

## CONCLUSION

Big Data Operation Management has become a revolutionary word which focuses on creating a customer-centric business by collecting and analyzing voluminous amount of data from distributed data sources.The knowledge thus gained helps in understanding the needs of every customer individually so that it becomes easier to do the business with them. It also focuses on understanding the various customers purchasing behavior and trends so that valuable insights can be converted into action. This is the revolutionary change to build a customer-centric business. It also helps in analyzing and exploring large data sets to find hidden rules, associations, patterns among the parts of data, which help to formalize and plan the future decision making process which is enabled by knowledge feature.

## REFERENCES

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*, *2*(3), 87–93. doi:10.1016/j.bdr.2015.04.001

Aloui, A., & Touzi, A. G. (2015). A Fuzzy Ontology-Based Platform for Flexible Querying. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(3), 12–26. doi:10.4018/IJSSMET.2015070102

Azar, A.T., & Hassanien, A.E. (2014). Dimensionality reduction of medical big data using neural-fuzzy classifier. *Soft computing, 19*(4), 1115-1127. DOI 10.1007/s00500-014-1327-4

Bhanu, S. K., & Tripathy, B. K. (2016). Rough Set Based Similarity Measures for Data Analytics in Spatial Epidemiology. *International Journal of Rough Sets and Data Analysis*, *3*(1), 114–123. doi:10.4018/IJRSDA.2016010107

Chen, M., Mao, S., & Liu, Y. (2009). Big Data: A Survey. *Springer-. Mobile Networks and Applications*, *19*(2), 171–209. doi:10.1007/s11036-013-0489-0

Deepak, D., & John, S. J. (2016). Information Systems on Hesitant Fuzzy Sets. *International Journal of Rough Sets and Data Analysis*, *3*(1), 55–70.

Ghallab, S. A., Badr, N. L., Salem, A. B., & Tolba, M. F. (2014). Strictness petroleum prediction system based on fussy model. *International Journal of Service Science, Management, Engineering, and Technology*, *5*(4), 44–65. doi:10.4018/ijssmet.2014100104

Harzog, B. (2015). *Big data operations management*. Retrieved from https://www.virtualizationpractice.com/big-data-operations-management-19219/

Hassanien, A. E., Azar, A. T., Snasel, V., Kacprzyk, J., & Abawajy, J. H. (2015). *Big Data in Complex Systems: Challenges and Opportunities*. Springer.

Hoffmann, A. O. I., & Birnbrich, C. (2012). The impact of fraud prevention on bank-customer relationships. *International Journal of Bank Marketing*, *30*(5), 390–394. doi:10.1108/02652321211247435

Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and challenges of big data computing in health science. *Big Data Research.*, *2*(1), 2–11.

Jagadish, H. V. (2015). Big Data and Science: Myths and Reality. *Big Data Research*, *2*(2), 49–52. doi:10.1016/j.bdr.2015.01.005

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, *2*(2), 59–64. doi:10.1016/j.bdr.2015.01.006

Ryan, T., & Lee, Y. C. (2015). Multi-tier resource allocation for data-intensive computing. *Big Data Research*, *2*(3), 110–116. doi:10.1016/j.bdr.2015.03.001

Shabeera, T. P., & Madhu Kumar, S. D. (2015). Optimizing virtual machine allocation in MapReduce cloud for improved data locality. *International Journal of Big Data Intelligence.*, *2*(1), 2–8. doi:10.1504/IJBDI.2015.067563

Srivastava, U., & Gopalkrishnan, S. (2015). Impact of Big Data Analytics on Banking Sector: Learning for Indian Bank. *Big Data. Cloud and Computing Challenges.*, *50*, 643–652.

Tiwari, P. K., & Joshi, S. (2015). Data security for software as a service. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(3), 47–63. doi:10.4018/IJSSMET.2015070104

Wahi, A. K., Medury, Y., & Misra, R. K. (2014). Social Media: The core of enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *5*(3), 1–15. doi:10.4018/ijssmet.2014070101

Wahi, A. K., Medury, Y., & Misra, R. K. (2015). Big Data: Enabler or Challenge for Enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(2), 1–17. doi:10.4018/ijssmet.2015040101

## ADDITIONAL READING

Shabeera, T. P., & Madhu Kumar, S. D. (2015). Optimizing virtual machine allocation in MapReduce cloud for improved data locality. *International Journal of Big Data Intelligence.*, *2*(1), 2–8. doi:10.1504/IJBDI.2015.067563

Srivastava, U., & Gopalkrishnan, S. (2015). Impact of Big Data Analytics on Banking Sector: Learning for Indian Bank. *Big Data. Cloud and Computing Challenges.*, *50*, 643–652.

Tiwari, P. K., & Joshi, S. (2015). Data security for software as a service. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(3), 47–63.

Wahi, A. K., Medury, Y., & Misra, R. K. (2014). Social Media: The core of enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *5*(3), 1–15. doi:10.4018/ijssmet.2014070101

Wahi, A. K., Medury, Y., & Misra, R. K. (2015). Big Data: Enabler or Challenge for Enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(2), 1–17. doi:10.4018/ijssmet.2015040101

## KEY TERMS AND DEFINITIONS

**Big Data:** is the voluminous and complex collection of data that comes from different sources such as sensors, content posted on social media website, sale purchase transaction etc. Such voluminous data becomes tough to process using ancient processing applications.

**Value:** This refers to the intrinsic value that the data may possess, and must be discovered.

**Variety:** This refers to various types of structured, unstructured and semi- structured data types.

**Velocity:** This refers to the rate at which data is generated and received.

**Volume:** This refers to the amount of data been generated from different sources such as data logs from twitter, click streams of web pages and mobile apps, sensor-enabled equipment capturing data, etc.

# Chapter 2
# Application of Artificial Neural Networks in Predicting the Degradation of Tram Tracks Using Maintenance Data

**Sara Moridpour**
*RMIT University, Australia*

**Ehsan Mazloumi**
*RMIT University, Australia*

**Reyhaneh Hesami**
*Yarra Trams, Australia*

## ABSTRACT

*The increase in number of passengers and tramcars will wear down existing rail structures faster. This is forcing the rail infrastructure asset owners to incorporate asset management strategies to reduce total operating cost of maintenance whilst improving safety and performance. Analysing track geometry defects is critical to plan a proactive maintenance strategy in short and long term. Repairing and maintaining the correctly selected tram tracks can effectively reduce the cost of maintenance operations. The main contribution of this chapter is to explore the factors influencing the degradation of tram tracks (light rail tracks) using existing geometric data, inspection data, load data and repair data. This chapter also presents an Artificial Neural Networks (ANN) model to predict the degradation of tram tracks. Predicting the degradation of tram tracks will assist in understanding the maintenance needs of tram system and reduce the operating costs of the system.*

DOI: 10.4018/978-1-5225-0886-1.ch002

# 1. INTRODUCTION

## 1.1. Background

Transport organisations have traditionally focused on construction and expansion of transport infrastructure. After completing the expansion of transport networks, the emphasis has shifted from developing new infrastructure to intelligently maintaining the existing ones (Hensher, 2000; El-sibaie & Zhang, 2004; Andrade & Teixeira, 2013). In recent years, economical constraints have influenced budget allocation to transport sectors (Alfelor, Carr, & Fateh, 2001; Jovanovic, 2004; Larsson, 2004; Andrade & Teixeira, 2011). This resulted in highlighting the development of maintenance management systems in transport sectors particularly in transport infrastructure. Maintenance management systems assist organisations in deciding when and how to maintain transport infrastructure facilities to minimise the maintenance cost/time and diminish replacement activities (Jovanovic, 2004; Larsson, 2004; Zhao, Chan, Stirling, & Madelin, 2006; Ahmad & Kamaruddin, 2012; Yaghini, Khoshraftar, & Seyedabadi, 2013).

## 1.2. Statement of Problem

Melbourne tram network is the largest urban tram network in the world which consists of 250 kilometres of track that runs 31,500 scheduled tram services every week (Yarra Trams 2015). Many parameters involve ensuring that Melbourne tram system is operating to its safe and best practice standards. Track infrastructure is one of the fundamental elements of the tram system. The condition of the track infrastructure can influence the system operation either directly or indirectly. To keep the track infrastructure in a reasonable condition and to obtain the most benefit out of its life cycle, an optimised maintenance and renewal regime is required. Providing a maintenance plan to recover the serviceability of tram tracks from damages and preventing further wear-out is essential for such a large network. Currently, tram track maintenance activities are achieved by manual inspections across the network. Yarra Trams has a fixed number of maintenance teams who are responsible to manually inspect the status of the tram tracks and identify whether tracks need maintenance. They approximately estimate a time frame for the scheduled maintenance. Since, the inspections are done manually, the human error is unavoidable. Mistakes in inspection and detection of track faults as well as inaccurate predic-

tion of maintenance time frame are the challenges in current maintenance system. In addition, prioritising the maintenance projects is often a significant challenge. Poorly planned maintenance schedules results in high maintenance and operation costs. This is due to repairing or replacing tram tracks at very early or late stage. Unnecessary maintenance and replacement of tram tracks or maintenance at the very late stages of the damage is very costly.

## 1.3. Research Objectives

To solve this problem, a model is required to predict the degradation of tram tracks. In this chapter, this will be done by analysing and trending the tram track attributes/variables over a period of time. Afterwards, the correlation between those variables and track deterioration will be identified and a model will be developed to predict the track deterioration/degradation based on tram track variables. The results of this chapter can be used to identify the maintenance and replacement activities and prepare a priority list for reparation of track rails. This will assist in minimising the maintenance costs and preventing unnecessary maintenance actions and therefore, saving time (Jovanovic, 2004; Larsson, 2004; Budai, Huisman, & Dekker, 2005; Guler, Jovanovic, & Evren, 2011; Caetano & Teixeira, 2013; Peng & Ouyang, 2013). Moreover, by preventing unnecessary maintenance actions, disturbance/interruption of traffic will be reduced and the delay time experienced by private car drivers and tram passengers will be decreased (Shafahi & Hakhamaneshi, 2009; Liu, Xu, & Wang, 2010; Sadeghi, 2010; Sadeghi & Askarinejad, 2010; Peng & Ouyang, 2012; Wu, Flintsch, Ferreira, & Picado-santos, 2012; Zakeri & Shahriari, 2012; Yaghini et al., 2013; Tang, Boyles, & Jiang, 2015). This chapter aims to gain deep insights into the tram track (light rail) maintenance dataset, examine the relationship between various variables and understand how different factors might impact degradation of tram tracks for the first time in the literature. This chapter also develops a time dependent Artificial Neural Networks (ANN) model to predict degradation of tram tracks for the first time in the literature as a function of various external factors over time.

This chapter begins by explaining the data used in this study. Then, the variables influencing the degradation of tram tracks will be identified. It is followed by presenting an ANN model to predict the degradation of tram tracks. The final section summarises the insight from this study and identifies recommendations and directions for future research.

## 2. TRAJECTORY DATA

The data used to analyse and predict the degradation of tram tracks is divided into three main categories including: inspection data, load data and repair data. These three main categories are explained in the following sections.

### 2.1. Inspection Data

The degradation of rail track geometry is usually quantified in terms of track geometry defects and rail surface faults. Track geometry defects include: longitudinal level defects, horizontal alignment defects, gauge deviations and track twist (Quiroga & Schnieder, 2013). According to the literature, gauge deviation has been considered as a popular indicator of track geometry defect (El-sibaie & Zhang, 2004; Shafahi & Hakhamaneshi, 2009; Sadeghi, 2010). Consistent with the literature, gauge deviation can be used to measure the track geometry defects. Gauge deviation is provided by Yarra Trams inspections of tram tracks.

The inspection data used in this study has been collected for Melbourne tram network between December 2009 and December 2013 over various segment types of several routes. The segment types include crossover, curve, H-crossing, and straight sections. According to the literature (Yousefikia, Moridpour, Setunge, & Mazloumi, 2014; Moridpour & Hesami, 2015), despite forming almost 10% of total length of routes under study, curves are more likely to experience exceeding the failure limits, in comparison to other segment types such as the H-crossing, Crossover and Straight segments. Therefore, for the purposes of this analysis, the focus is on curve segments. The inspection data includes the gauge values at curve segments at every 10 cm.

### 2.2. Load Data

Load data is another set of data that should be used to predict the degradation of tram tracks. The load data consists of Million Gross Tonne (MGT) without passengers and the number of trips per day. Gross Tonne is the product of total weight including the weight of locomotives and wagons as well as the weight of the average annual daily traffic volume passing the tram tracks. The load data used in this study is available from 2009 to 2013.

### 2.3. Impact of Repair

Track repair data should also be supplied for the track maintenance analysis. The repair data includes the physical address of tracks which were repaired, the time

when repairs were made as well as the repair type. Similarly, the repair data from 2009 to 2013 is used for predicting the degradation of tram tracks.

Table 1 shows the summary of the prepared data used in the degradation analysis. As presented in this table, the data relates to 4 routes and in total, there are 12,745 records related to 8,158 inspection points. This means that on average there are about two observations/inspections relating to a certain inspection point.

## 3. DATA ANALYSIS

The first stage in tram track maintenance studies is gaining insights into the available dataset and understanding how different factors might impact degradation of tram tracks (i.e. gauge widening). In this section, impact of different factors on tram track degradation is analysed.

### 3.1. Track Deterioration Profile Over Time

For unrepaired inspection points, track deterioration over time is worked out based on the available dataset. This is carried out with respect to the number of inspections available in the dataset for each inspection point. Figures 1 and 2 show the "Average Gauge Widening" and "Average Increase in Gauge Widening since the First Inspection" versus "Average Number of Months since the First Inspection", respectively.

As shown in Figures 1 and 2, the deterioration trend for inspection points for which data from at least 4 inspections is available is different to the other two lines. For instance, it has a smaller slope between the first and the second inspections. The reason could be because the first and second inspections for the other two groups ("2 Inspections" and "3 Inspections") were made at a later stage in a track life cycle (when tracks were in bad conditions). Another insight from these two figures is that

*Table 1. Summary of the prepared data used in the degradation analysis*

| Route Number | Number of Inspection Points | Number of Records | Minimum Inspection Date | Maximum Inspection Date |
|---|---|---|---|---|
| 57 | 847 | 1,329 | 12/12/2009 | 25/11/2013 |
| 59 | 3,170 | 5,131 | 12/12/2009 | 25/11/2013 |
| 70 | 2,082 | 3,337 | 13/12/2009 | 29/11/2013 |
| 75 | 2,059 | 2,948 | 13/12/2009 | 29/11/2013 |
| Total | 8,158 | 12,745 | 12/12/2009 | 29/11/2013 |

*Figure 1. Deterioration trend for sections with different inspection numbers*

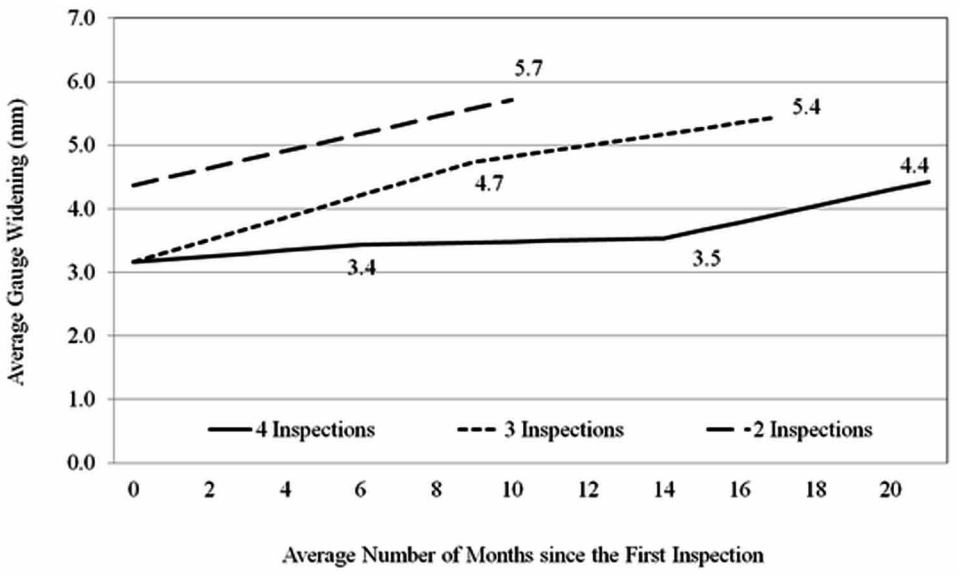
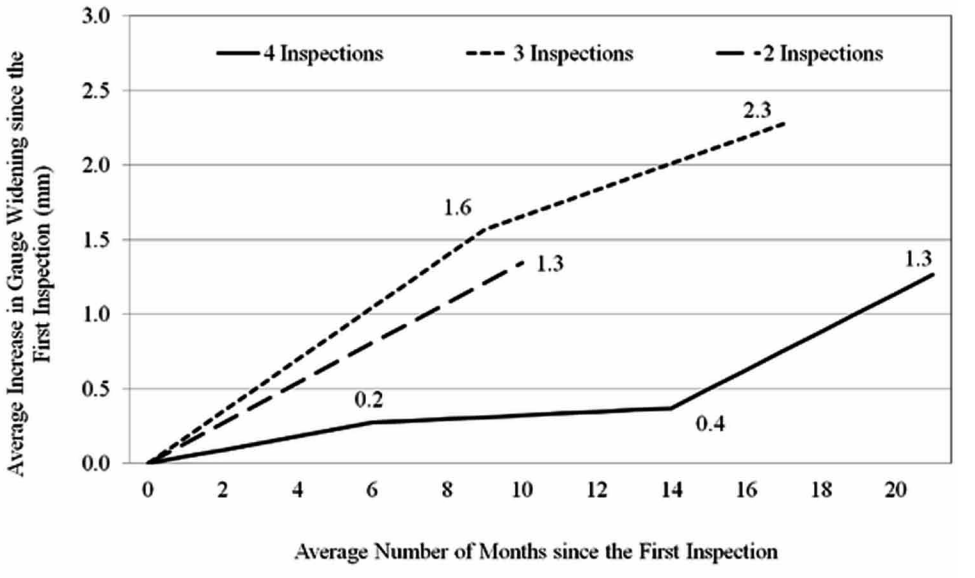



*Figure 2. Deterioration rate trend for sections with different inspection numbers*



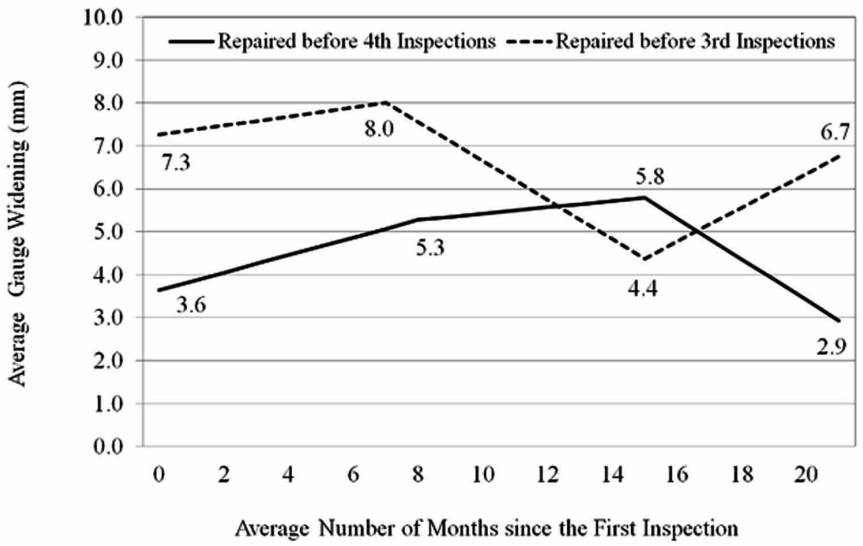
the average number of months for the second inspection ("4 Inspections") is around 6 months, whereas in the other two lines ("2 Inspections" and "3 Inspections"), it's around 9-10 months.

## 3.2. Impact of Repair on Deterioration Profile

For sections that have data from at least 4 inspections and have been repaired, the impact of repair on deterioration profile is investigated. As seen in Figure 3, this analysis is carried out with respect to when the repair was done, i.e. before the third inspections or the fourth inspection. As an example, the solid line shows the average gauge for inspection points that were repaired before the forth inspection. According to this line, due to a repair, the average gauge for this group of curves drops from 5.8 mm in the third inspection to 2.9 mm in the fourth inspection. This yields an almost 50% improvement, which is consistent with the "Repaired before third inspection" group (drop from 8.0 mm to 4.4 mm which is also around 50% improvement).

Another insight from the graph below is that post-repair degradation rate is greater than pre-repair degradation rate. According to the "Repaired before Third Inspection" line, the degradation rate is around 0.1 mm per month (from 7.3 mm

*Figure 3. Impact of repair on degradation of curves: Relationship between "Gauge Value Change per Month" and "Rail Profile"*



to 8.0 mm in 7 months) whereas it is around 0.3 mm per month (from 4.4 mm to 6.7 mm in 7 months) after repair which is three times higher than the pre-repair rate.

## 3.3. Impact of Other Factors on Deterioration

This section aims to examine the relationship between various variables (for which data is available for the study) and track deterioration. The available variables can be categorised into two main groups (based on their types) including continuous and categorical variables. This section adopts different approaches to examine the relationship each variable has with track deterioration. For the purpose of the analysis in this section, "Gauge Value Change per Month" is calculated for inspection points that were not repaired and causal relationships are investigated in relation to "Gauge Value Change per Month".

### 3.3.1. Continuous Variables

If a variable can take on any value between two specified values, it is called a continuous variable. In our dataset, the following variables are continuous.

- "MGT" as the million gross tonne value for the track where the inspection point is located.
- "Trips" which is the number of trips for the track where the inspection point is located.
- "Curve Radius" as the radius of curve for the track on which the inspection point is located.
- "Years since Instalment" which is the number of years since tracks were originally installed.

To examine the relationship between available continuous variables and "Gauge Value Change per Month", correlation analysis is adopted. Table 2 shows the correlation values and their associated p-values. Higher correlation values indicate higher association between the variable and "Gauge Value Change per Month". P-values also show if the associated correlation value is statistically significant or not. If a p-value is greater than 0.05 (95% accuracy level), it means that the correlation value is not statistically significant. As seen in Table 2, MGT is positively correlated with "Gauge Value Change per Month", and the p-value is 0 that means it is statistically correlated. "Trips" has a p-value greater than 0.05 so it is not significantly correlated with "Gauge Value Change per Month". Both "Curve Radius" and "Years since Instalment" variables are significantly and negatively correlated

*Table 2. Correlation analysis between "Gauge Value Change per Month" and continuous variables*

| Variable | Correlation Value | P-Value |
|---|---|---|
| MGT | 0.09 | 0.00 |
| Trips | 0.02 | 0.37 |
| Curve Radius | -0.15 | 0.00 |
| Years since Instalment | -0.14 | 0.00 |

with "Gauge Value Change per Month". Their negative values indicate that as they increase, "Gauge Value Change per Month" decreases.

To obtain further insights into the impact of various variables on deterioration rate (i.e. "Gauge Value Change per Month") Table 3 shows the average of various factors in different quartiles of "Gauge Value Change per Month".

Also, Table 4 shows the values of different variables in different quartiles of "Curve Radius". As seen, the greatest average "Gauge Value Change per Month"

*Table 3. Average of various factors in different quartiles of "Gauge Value Change per Month"*

| Quartiles based on "Gauge Value Change per Month" | "Gauge Value Change per Month" | Trips | MGT | Years since Instalment | Curve Radius |
|---|---|---|---|---|---|
| 1 (0.00 - 0.07) | 0.04 | 52,384 | 1,486,005 | 24 | 142 |
| 2 (0.08 - 0.15) | 0.11 | 49,226 | 1,410,567 | 27 | 163 |
| 3 (0.16 - 0.30) | 0.22 | 47,030 | 1,371,649 | 28 | 183 |
| 4 (0.30 - 7.27) | 0.63 | 50,201 | 1,544,150 | 20 | 132 |
| Total | 0.25 | 49,678 | 1,452,837 | 25 | 155 |

*Table 4. Average of various factors in different quartiles of "Curve Radius"*

| Quartiles based on "Curve Radius" | Curve Radius | "Gauge Value Change per Month" | Trips | MGT | Years since Instalment |
|---|---|---|---|---|---|
| 1 | 18 | 0.19 | 73,343 | 2,096,975 | 9 |
| 2 | 61 | 0.41 | 50,457 | 1,625,048 | 12 |
| 3 | 229 | 0.17 | 33,430 | 804,620 | 53 |
| 4 | 268 | 0.20 | 44,220 | 1,289,104 | 30 |
| Total | 155 | 0.25 | 49,678 | 1,452,837 | 25 |

occurs in the second quartile where the average radius is around 61m. This group consists of tracks that are relatively new (aged 12 years that is almost half of the average age of 25 years) experiencing almost average number of "Trips" and "MGT" compared to the overall "Trips" and "MGT" values across the entire sample. This contradicts our expectations, and therefore we hypothesise that there are other factors causing degradation in this group of tracks.

### 3.3.2. Categorical Variables

A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, thus assigning each individual to a particular group or "category." In order to examine the relationship between each categorical variable and "Gauge Value Change per Month", this study adopts an ANalysis Of VAriance (ANOVA) which statistically compares the average "Gauge Value Change per Month" in different categories of each variable. In our dataset, the following variables are categorical variables:

- Rail Profile which is associated with the type of rail profile at the inspection point (i.e. 41 kg, 57 kg, 60 kg, 80 lb, 96 lb).
- Rail Type which is associated with the inspection point (i.e. GP, TT, G or T).
- Route which is a categorical variable to indicate the route number (i.e. 57, 59, 70, or 75). In this study four different tram routes are used.

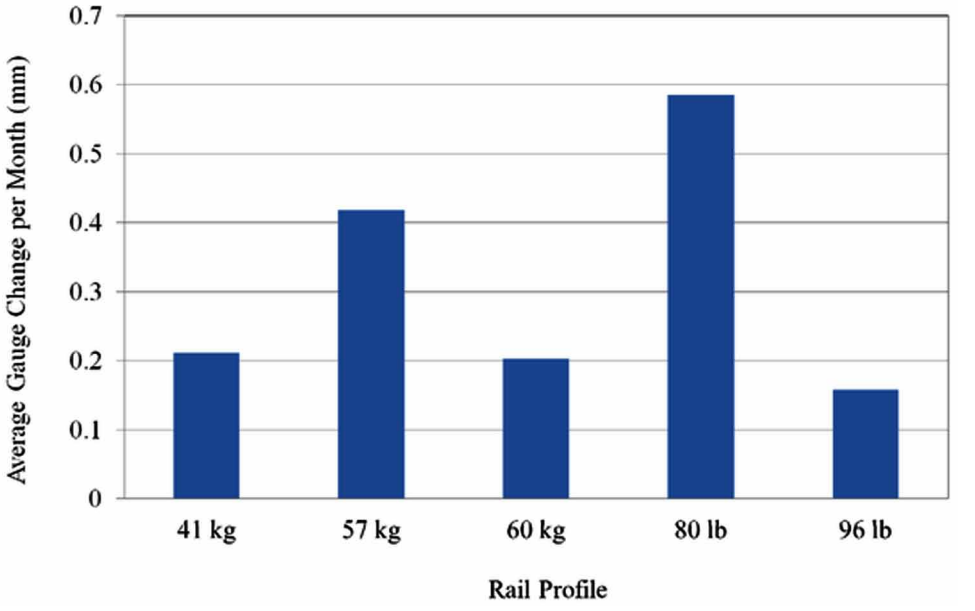
Table 5 shows the results of the ANOVA, including F-Value and significance value. Higher values of F-value indicate greater difference between average values of "Gauge Value Change per Month" in different categories of variables. Significance value (if less than 0.05) suggests the differences are statistically significant. According to Table 5, all categorical variables are important for modelling "Gauge Value Change per Month". However, based on F-Values, "Rail Profile" has the greatest impact.

To visually inspect the impact of "Rail Profile" on "Gauge Value Change per Month", Figure 4 shows average "Gauge Value Change per Month" in different

*Table 5. ANOVA between "Gauge Value Change per Month" and categorical variables*

| Variable | F-Value | Significance |
|---|---|---|
| Rail Profile | 84.3 | 0 |
| Rail Type | 51.5 | 0 |
| Route | 55.3 | 0 |

*Figure 4. Impact of repair on degradation of curves: Relationship between "Gauge Value Change per Month" and "Rail Type"*




categories of rail profile. Accordingly, "Gauge Value Change per Month" varies noticeably in various rail profile types. This shows that type of rail profile is an influencing factor on degradation of tram tracks and significant in rail maintenance. Figure 5 presents the relationship between different categories of "Rail Type" and average "Gauge Value Change per Month". Figure 6 also depicts average "Gauge Value Change per Month" in different tram routes. Accordingly, route 59 has the highest "Gauge Value Change per Month" and route 70 has the lowest. This shows that for valid analysis of the degradation and maintenance of tram tracks, each route has to be individually assessed.

## 4. RESEARCH METHODOLOGY

After identifying the variables influencing the gauge values, this section aims to develop a model to predict gauge widening as a function of various external factors. In this study, ANN are used to predict gauge widening.

*Figure 5. Impact of repair on degradation of curves: Relationship between "Gauge Value Change per Month" and "Route"*

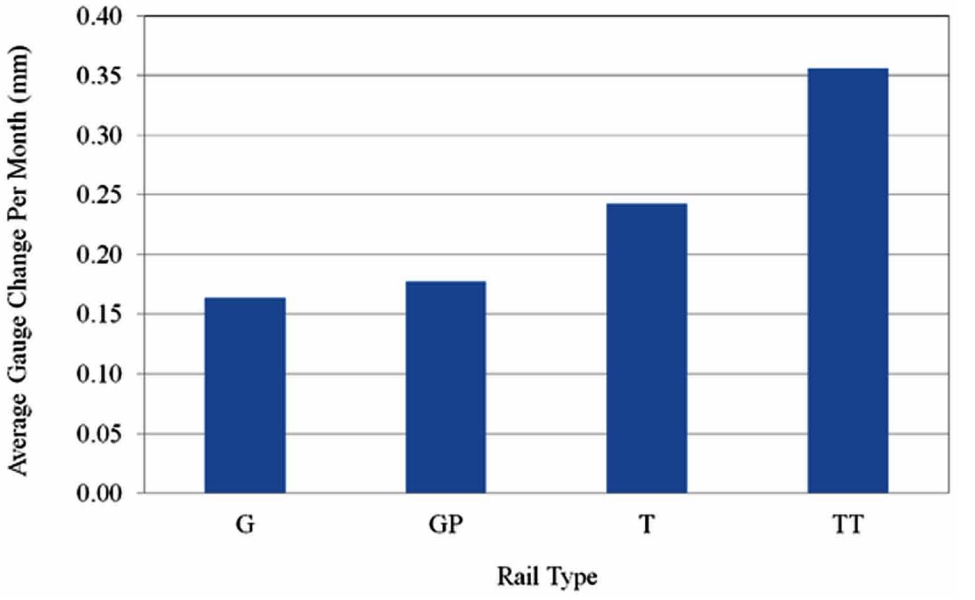
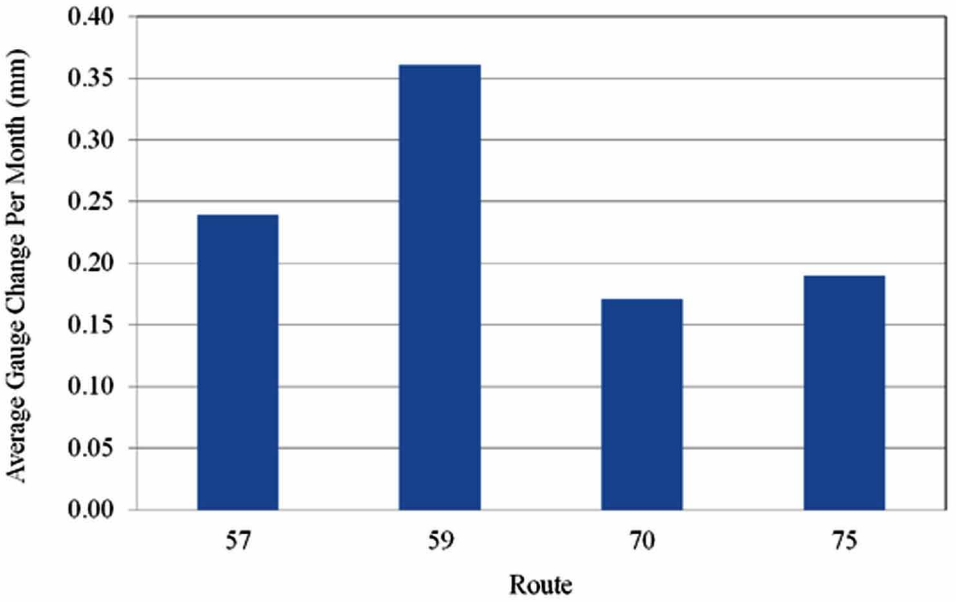



*Figure 6. The relationship between different variables and "Gauge Value Change per Month"*



## 4.1. Artificial Neural Networks

Artificial neural networks, commonly termed as neural networks or neural models, are a class of information processing systems which consist of a set of inter-related processing elements called neurons (Mazloumi, Rose, Currie, and Moridpour, 2011; Mazloumi, Moridpour. Currie, and Rose, 2012). Each neuron in the network carries out simple computation tasks such as receiving inputs from other neurons and computing an output sent to other neurons. A group of neurons constitutes a layer. The layer that receives the input data from external sources is called the "input layer", and the one which outputs the final results of the network is termed the "output layer". Often, between the input and output layers, there is one or more hidden layers, where the main data processing task is conducted.

The number of layers in a neural network determines if it is a single, two, or multi-layer neural network. When connections only exist in one direction between adjacent layers, i.e. they do not form a feedback loop, the network is defined as having feed-forward architecture. The most common structure of neural networks is the multi-layer feed-forward neural network. Figure 5 shows the typical topology of a three layer feed forward neural network. Each inter-connection in the network has an associated weight (e.g. $W_{ij}$ or $W_{jk}$) and each neuron has an assigned threshold (e.g. $\theta_j$ is the threshold associated with neuron $j$). These weights and thresholds are the model parameters calibrated through a "training procedure".

In the training mechanism, a number of input-output pairs (training examples) are introduced to the model, and the model then understands the relationship between the inputs and corresponding outputs through adjusting its parameters. Training involves finding the optimal values for model parameters. In this study, a "Back Propagation Training Algorithm" has been adopted.

An important factor that influences the neural network outcomes is the initial point to start searching for the optimal answer. Therefore, different initial model parameter may lead to a different optimal point. One way to alleviate this problem and to increase the likelihood of obtaining near-optimum local minima is to train several neural networks with a random set of initial weights and to make predictions based on an ensemble of neural networks.

In this study, the idea of ensemble prediction is adopted, and each neural network one hidden layer and is trained 10 times. The number of neurons in the hidden layer is calculated automatically by the model to achieve the best prediction accuracy. The results are then averaged to calculate the final prediction values corresponding to a certain set of input values. However, for ease of referencing, the neural network ensemble is referred as a neural network model, and results that are reported are the average for the ensemble. In addition, for the first time in light rail maintenance

prediction, the prediction model is time dependant and degradation of tram tracks is predicted over time. It should be noted that in the existing models, the degradation predictions is based on the mass passing through tracks. Understanding the influence of different values of mass on degradation of tracks may be difficult, even for experts in this area. However, understanding the rate of degradation over time is easy for everyone to understand and interpret.

## 4.2. A Model to Predict Gauge Widening

Assume that we are interested to know $G_{n+1}$ that is the value of Gauge in the $(n+1)^{th}$ inspection for a certain inspection point. If we know the "Change in Gauge per Month between $n^{th}$ and $n+1^{th}$", we can then use the following formula to work out $G_{n+1}$.

$$G_{n+1} = G_n + DM_{n+1,n} \times M_{n+1,n} \tag{1}$$

Where:

$G_{n+1}$: Gauge value in the $(n+1)^{th}$ inspection for a certain inspection point,
$G_n$: Gauge value in the $n^{th}$ inspection for the same inspection point,
$DM_{n+1,\,n}$: Change in Gauge between the $n^{th}$ and $n+1^{th}$ inspections (per month),
$M_{n+1,\,n}$: Number of months between the $n^{th}$ and $n+1^{th}$ inspections.

Therefore, in this study, the variable that is predicted is $DM_{n+1,\,n}$. An advantage of using neural networks is that it is not required to find significant predictors before building the model. All potential predictors can be fed into the model, and the model itself determines the importance of predictors (in predicting the output variable), and uses them accordingly. Therefore, in this study, we will use all available predictors and let the model decide on the importance of predictors. The following predictors are then used as inputs to the model:

- **Repaired:** A flag variable indicating if a repair has been done since the last inspection.
- **Rail Profile:** Rail profile associated with the inspection point.
- **Rail Type:** Rail type associated with the inspection point.
- **Curve Radius:** the curve radius for the track on which the inspection point is located.
- **Years Since Instalment:** number of years since the track was originally installed.

- **Months Since Last Inspection:** number of months since the point was last inspected.
- **Gauge Last Inspection:** the gauge value in the last inspection.
- **MGT:** the total weight from 2009 to 2013.
- **Trips:** the number of trips per day from 2009 to 2013.
- **Route:** A categorical variable to indicate the route number.

## 5. RESULTS

To develop the model, the entire dataset was randomly split into a Training (80%) and a Testing (20%) subset. The model is trained based on the training subset and is tested against the data examples in the Testing subset. The Testing subset was kept away from the model so the data examples in the Testing subset remain "unseen" to the model.

To evaluate the performance of the model, prediction error for the data example *i (i= 1,…,m)* is calculated as $e_i = G_i - \bar{G}_i$ where $G_i$ is the observed Gauge value, and $\bar{G}_i$ is the corresponding predicted value. The following performance measures can be adopted to evaluate the performance of the model:

$$\text{Mean Error (ME)} = \frac{\sum_{i=1}^{m} e_i}{m} \tag{2}$$

$$\text{Mean Relative Error (MRE)} = \frac{\sum_{i=1}^{m} \frac{e_i}{G_i}}{m} \tag{3}$$

$$\text{Root Mean Square Error (RMSE)} = \frac{\sqrt{\sum_{i=1}^{m} e_i^2}}{m} \tag{4}$$

Table 6 summarises a set of details about the structure of the model that gave rise to the best prediction accuracy, and the training process used to train the model. As

*Table 6. The structure of the best model*

| **Model Structure** | |
|---|---|
| Neural Network Type | Feed Forward |
| Number of layers | 3 layers (1 Input, 1 Hidden and 1 Output layer) |
| Number of neurons in the hidden layer | 6 |
| Training Method | An ensemble of 10 neural networks |
| Training / Testing data split | 80% training and 20% testing dataset |

presented, the final/best model is an ensemble of 10 neural networks each having 1 input layer, 1 hidden layer and 1 output layer. The hidden layer also has 6 neurons.

Figure 6 presents the importance of each variable in predicting "Gauge Value Change per Month" when the best model is used. As expected, "Repaired" and "Gauge Last Inspection" are the most important, and "Curve Radius", "Months since Last Inspection", "Years since Instalment" and "Trips" are the least important factors in predicting "Gauge Value Change per Month". It is worth noting that the importance values are outputs of the training process, and as such are related to the training dataset.

After the model is trained, its performance is tested based on data examples in the testing dataset. These examples were held out, and have not been used in training the model. Therefore, it can be expected that performance measures derived from the testing dataset are similar to the model's future performance. Table 7 shows the prediction performance of the developed model based on the testing dataset. As seen in the table, the average error (ME) is -0.13 mm which means on average the model over-predicts the Gauge values by 0.13 mm which yields an error of around 14 percent (MRE). Overall, RMSE of 1.58 mm indicates satisfactory prediction accuracy.

Figure 7 also presents the predicted values plotted against the observed Gauge values. As presented in this figure, the data points are generally spread along the 45 degree line which confirms the acceptable prediction accuracy (see Figures 8 and 9).

*Table 7. Prediction performance of the degradation model for curves*

| **Performance Measure** | | |
|---|---|---|
| ME (mm) | MRE (percent) | RMSE (mm) |
| -0.13 | -14% | 1.58 |

*Figure 7. Architecture of a three layer feed-forward neural network*

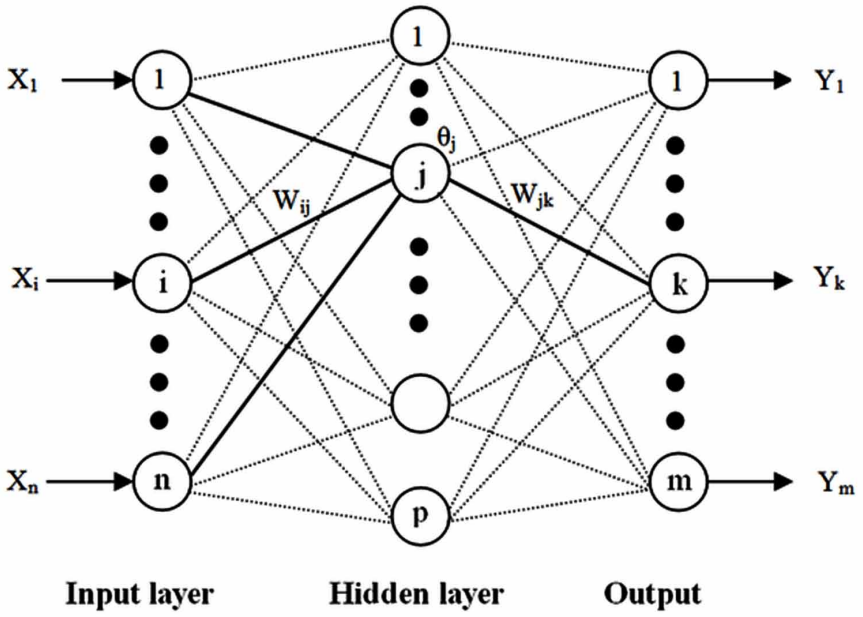



*Figure 8. Importance of various variables in prediction of "Gauge Value Change per Month"*

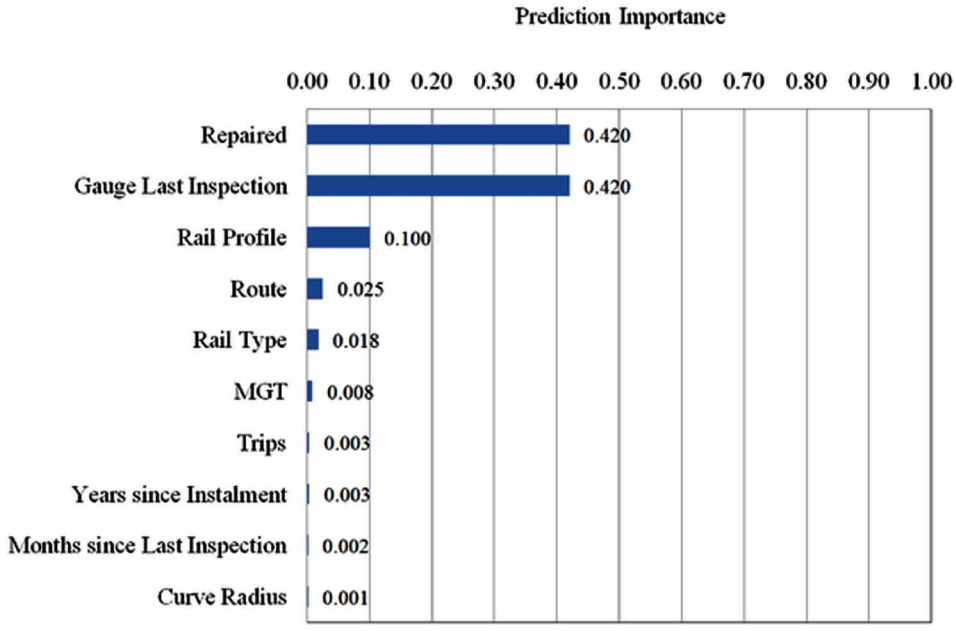
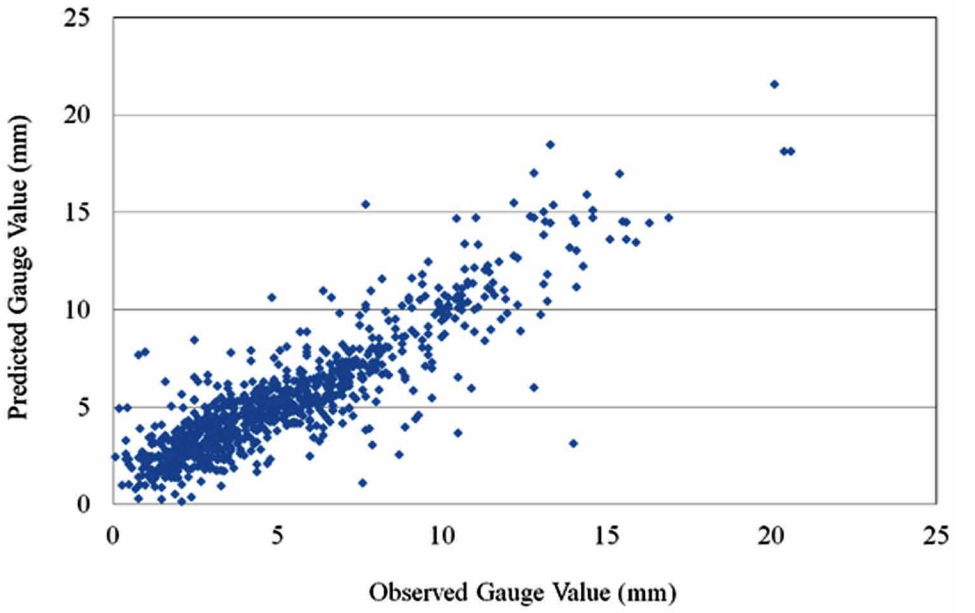


*Figure 9. Prediction vs. observed Gauge values*




## 6. CONCLUSION AND RECOMMENDATIONS

The analysis presented in this chapter intended to obtain insights into the impact of various external factors on track degradation and track maintenance based on the available dataset. Only curve segments were the focus of the study.

The preliminary analysis to explore the available data revealed that repairing a track has a great impact on the value of gauge widening. On average, a repair improves the gauge widening by around 50%. Degradation rate after a repair can be as three times higher than a pre-repair rate. Continuous variables were undergone a correlation analysis. The outcome suggests that "MGT" is positively and "Curve Radius" and "Years since Instalment" are negatively correlated with "Gauge Value Change per Month". Available categorical variables were tested using an Analysis of Variance (ANOVA) to understand how they relate to "Gauge Value Change per Month". ANOVA tests whether the average "Gauge Value Change per Month" varies significantly across various categories of a categorical variable. Accordingly, various categories across each categorical variable (i.e. Rail Profile, Rail Type, and Route) have statistically different average "Gauge Value Change per Month". Therefore, it is important to consider all these variables when modelling track gauge widening.

An ensemble of neural network models was also used to predict "Gauge Value Change per Month". The outcomes of the prediction process suggest that the most

important factors in predicting "Gauge Value Change per Month" are "Repaired", "Gauge Last Inspection", and "Rail Profile" and the least important factors are "Curve Radius", "Months since Last Inspection", "Years since Instalment" and "Trips" are the least important factors in predicting "Gauge Value Change per Month". The developed model presented a reasonably good prediction accuracy of 86% (error of 14%).

## 7. FUTURE RESEARCH DIRECTIONS

The analysis presented in this report can be extended in several ways. Prediction accuracy can be improved in several ways including:

- The model presented in this report is based on data from all routes and hence can be used to model degradation in all routes. If enough data becomes available, a separate model for each route can yield a higher accuracy and robustness. Also, if the data is available, the influence of wheels on track degradation can be analysed and modelled.
- The model presented in this report only focuses on Curves. A separate model to predict degradation for other track components (e.g. straight, H-crossing, crossover) can help prioritise the maintenance across the entire network.
- Similar procedure can be used to analyse the deterioration data and model the rate of deterioration/defection of different assets of tram system (e.g. overheads).

## REFERENCES

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, *63*(1), 135–149. doi:10.1016/j.cie.2012.02.002

Alfelor, R., Carr, G., & Fateh, M. (2001). Track degradation assessment using gage restraint measurements. *Transportation Research Record: Journal of the Transportation Research Board*, *1742*, 68–77. doi:10.3141/1742-09

Andrade, A. R., & Teixeira, P. F. (2011). Uncertainty in rail-track geometry degradation: Lisbon-Oporto line case study. *Journal of Transportation Engineering*, *137*(3), 193–200. doi:10.1061/(ASCE)TE.1943-5436.0000206

Andrade, A. R., & Teixeira, P. F. (2013). Unplanned-maintenance needs related to rail track geometry.*Proceedings of the ICE - Transport*.

Budai, G., Huisman, D., & Dekker, R. (2005). Scheduling preventive railway maintenance activities. *The Journal of the Operational Research Society*, *57*(9), 1035–1044. doi:10.1057/palgrave.jors.2602085

Caetano, L., & Teixeira, P. F. (2013). Availability approach to optimizing railway track renewal operations. *Journal of Transportation Engineering*, *139*(9), 941–948. doi:10.1061/(ASCE)TE.1943-5436.0000575

El-sibaie, M., & Zhang, Y. J. (2004). Objective track quality indices. *Transportation Research Record: Journal of the Transportation Research Board*, *1863*, 81–87. doi:10.3141/1863-11

Gazder, U. & Ratrout, N. T. (2015). A new logit-artificial neural network ensemble for mode choice modeling: a case study for border transport. *Journal of Advanced Transportation.* DOI: 10.1002/atr.1306

Guler, H., Jovanovic, S., & Evren, G. (2011). Modelling railway track geometry deterioration.*Proceedings of the ICE - Transport*.

Hensher, D. A. (2000). Transport economics: A personal view. *Journal of Advanced Transportation*, *34*(1), 65–106. doi:10.1002/atr.5670340105

Jovanovic, S. (2004). Railway track quality assessment and related decision making. *IEEE International Conference on Systems, Man and Cybernetics*. doi:10.1109/ICSMC.2004.1400992

Larsson, D. (2004). *A study of the track degradation process related to changes in railway traffic.* (PhD thesis). Lulea University of Technology.

Liu, R., Xu, P., & Wang, F. (2010). Research on a short-range prediction model for track irregularity over small track lengths. *Journal of Transportation Engineering*, *136*(12), 1085–1091. doi:10.1061/(ASCE)TE.1943-5436.0000192

Mazloumi, E., Moridpour, S., Currie, G., & Rose, G. (2012). Exploring the Value of Traffic Flow Data in Bus Travel Time Prediction. *Journal of Transportation Engineering*, *138*(4), 436–446.

Mazloumi, E., Rose, G., Currie, G., & Moridpour, S. (2011). Prediction Intervals to Account for Uncertainties in Neural Network Predictions: Methodology and Application in Bus Travel Time Prediction. *Engineering Applications of Artificial Intelligence*, *24*(3), 534–542.

Moridpour, S., & Hesami, R. (2015). Degradation and performance specification of Melbourne tram tracks. *The 3rd International Conference on Transportation Information and Safety (ICTIS)*.

Peng, F., & Ouyang, Y. (2012). Track maintenance production team scheduling in railroad networks. *Transportation Research Part B: Methodological*, *46*(10), 1474–1488. doi:10.1016/j.trb.2012.07.004

Peng, F., & Ouyang, Y. (2013). Optimal clustering of railroad track maintenance jobs. *Computer-Aided Civil and Infrastructure Engineering*, *29*(4), 235–247. doi:10.1111/mice.12036

Quiroga, L., & Schnieder, E. (2013). Railway systems. In H. Czichos (Ed.), *Handbook of Technical Diagnostics*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-25850-3_26

Sadeghi, J. (2010). Development of railway track geometry indexes based on statistical distribution of geometry data. *Journal of Transportation Engineering*, *136*(8), 693–700. doi:10.1061/(ASCE)0733-947X(2010)136:8(693)

Sadeghi, J., & Askarinejad, H. (2010). Development of improved railway track degradation models. *Structure and Infrastructure Engineering*, *6*(6), 675–688. doi:10.1080/15732470801902436

Shafahi, Y., & Hakhamaneshi, R. (2009). Application of maintenance management model for Iranian railways based on the Markov chain probabilistic dynamic modeling. *Scientia Iranica*, *16*, 87–97.

Tang, S., Boyles, S. D., & Jiang, N. (2015). High-speed rail cost recovery time based on an integer optimization model. *Journal of Advanced Transportation*, *49*(5), 634–647. doi:10.1002/atr.1294

Wu, Z., Flintsch, G., Ferreira, A., & Picado-santos, L. (2012). Framework for multiobjective optimization of physical highway assets investments. *Journal of Transportation Engineering*, *138*(12), 1411–1421. doi:10.1061/(ASCE)TE.1943-5436.0000458

Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, *47*(3), 355–368. doi:10.1002/atr.193

Yarra Trams. (2015). *Facts and Figures, Yarra Trams.* Retrieved from http://www.yarratrams.com.au/about-us/who-we-are/facts-figures/

Yousefikia, M., Moridpour, S., Setunge, S., & Mazloumi, E. (2014). Modeling degradation of tracks for maintenance planning on a tram line. *Journal of Traffic and Logistics Engineering.*, *2*(2), 86–91. doi:10.12720/jtle.2.2.86-91

Zakeri, J. A., & Shahriari, S. (2012). Developing a deterioration probabilistic model for rail wear. *International Journal of Traffic and Transportation Engineering*, *1*, 13–18.

Zhao, J., Chan, A., Stirling, A., & Madelin, K. (2006). Optimizing policies of railway ballast tamping and renewal. *Transportation Research Record: Journal of the Transportation Research Board*, *1943*, 50–56. doi:10.3141/1943-07

## ADDITIONAL READING

Ahmed, N. (1983). Management Science Department, Ball State University, Muncie, USA. An analytical decision model for resource allocation in highway maintenance management. *Transportation Research Part A, General*, *17*(2), 133–138. doi:10.1016/0191-2607(83)90067-5

Andersen, T. M., & Rasmussen, B. S. (1999). Effort, Taxation and Unemployment. *Economics Letters*, *62*(1), 97–103. doi:10.1016/S0165-1765(98)00216-X

Andrews, J. (2013). A modelling approach to railway track asset management. *Proceedings of the Institution of Mechanical Engineers. Part F, Journal of Rail and Rapid Transit*, *227*(1), 56–73. doi:10.1177/0954409712452235

Campos, J. (2009). Development in the application of ICT in condition monitoring and maintenance. *Computers in Industry*, *60*(1), 1–20. doi:10.1016/j.compind.2008.09.007

Chang, H., Liu, R. & Wang, W. (2010). Multistage linear prediction model of track quality index. *Traffic and Transportation Studies: 1183-1192*.

Chrismer, S. M. (1994). *Mechanics-based model to predict ballast-related maintenance timing and costs. PhD*. University of Massachusetts Amherst.

Currie, G., & Burke, M. (2013). Light Rail in Australia-Performance and Prospects. *The 36th Australasian Transport Research Forum*, 2-4 October, Brisbane, Australia.

Demharter, K. (1982). Setzungsverhalten des gleisrostes unter vertikaler Lasteinwirkung. *Minchen, Deutschland: der Technischen Univerität München*.

Esveld, C. (2001). Modern railway track, The Netherlands. *MRT Productions, Zaltbommel*.

Heckl, M. A., & Abrahams, I. D. (2000). Curve squeal of train wheels, part 1: Mathematical model for its generation. *Journal of Sound and Vibration*, *229*(3), 669–693. doi:10.1006/jsvi.1999.2510

Hokstad, P., Langseth, H., Lindqvist, B. H., & Vatn, J. (2005). Failure modeling and maintenance optimization for a railway line. *International Journal of Performance Engineering*, *1*, 51–64.

Hummitszch, R. (2005). *Calculation schemes for MDZ and "modified standard deviation"*. Technical University of Graz.

Iwnicki, S. D., Grassie, S., & Kik, W. (2000). Track settlement prediction using computer simulation tools. *Vehicle System Dynamics*, *33*, 37–46.

Larsson, D. (2004). *A study of the track degradation process related to changes in railway traffic. PhD*. Luleå University of Technology.

Lyngby, N., Hokstad, P., & Vatn, J. (2008). RAMS Management of Railway Tracks. In K. Misra (Ed.), *Handbook of Performability Engineering*. Springer London. doi:10.1007/978-1-84800-131-2_68

Mazzuchi, T. A., Van Noortwijk, J. M., & Kallen, M. J. (2008). *Maintenance optimization. Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, Ltd.

Meier-hirmer, C., Riboulet, G., Sourget, F., & Roussignol, M. (2009). Maintenance optimization for a system with a gamma deterioration process and intervention delay: Application to track maintenance. *Proceedings of the Institution of Mechanical Engineers O. Journal of Risk and Reliability*, *223*, 189–198.

Pham, H., & Wang, H. (1996). Imperfect maintenance. *European Journal of Operational Research*, *94*(3), 425–438. doi:10.1016/S0377-2217(96)00099-9

Pham, H., & Wang, H. (1999). Some maintenance models and availability with imperfect maintenance in production systems. *Analysis of Operations Research*, *91*, 305–318. doi:10.1023/A:1018910109348

Podofillini, L., Zio, E., & Vatn, J. (2006). Risk-informed optimisation of railway tracks inspection and maintenance procedures. *Reliability Engineering & System Safety*, *91*(1), 20–35. doi:10.1016/j.ress.2004.11.009

Sato, Y. (1995). Japanese studies on deterioration of ballasted track. *Vehicle System Dynamics*, *24*(sup1), 197–208. doi:10.1080/00423119508969625

Shyr, F., & Ben-akiva, M. (1996). Modeling rail fatigue behavior with multiple hazards. *Journal of Infrastructure Systems*, *2*(2), 73–82. doi:10.1061/(ASCE)1076-0342(1996)2:2(73)

Veit, P. (2007). Track quality-Luxury or necessity? *RTR special: Maintenance & renewal, Railway Technical Review: 8-12*.

Zhang, T., Andrews, J., & Wang, R. (2013). Optimal scheduling of track maintenance on a railway network. *Quality and Reliability Engineering International*, *29*(2), 285–297. doi:10.1002/qre.1381

Zhang, Y. J., Murray, M., & Ferreira, L. (2000). Modelling rail track performance: An integrated approach. *Proceedings of the Institution of Civil Engineers: Transport*, *141*, 187–194.

Zio, E., Marella, M., & Podofillini, L. (2007). Importance measures-based prioritization for improving the performance of multi-state systems: Application to the railway industry. *Reliability Engineering & System Safety*, *92*(10), 1303–1314. doi:10.1016/j.ress.2006.07.010

## KEY TERMS AND DEFINITIONS

**Artificial Neural Networks:** Artificial Neural Networks are a group of models inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown.

**Gauge:** Track gauge is the spacing of the rails on a railway track and is measured between the inner faces of the load-bearing rails.

**Prediction:** Prediction is a statement about an uncertain event and it is often based on experience or knowledge.

**Rail Degradation:** Degradation is the wearing down of rail.

**Track:** The track on a railway or railroad is the structure consisting of the rails, fasteners, railroad sleepers and ballast (or slab track), as well as the underlying subgrade.

**Tram:** A tram is a rail vehicle which runs on tracks along public urban streets, and also sometimes on a segregated right of way.

**Twist:** Track twist is used to describe cant gradient which may be expressed in percentage of cant change per unit of length.

# Chapter 3
# ZAMREN Big Data Management (ZAMBiDM) Envisaging Efficiency and Analytically Manage IT Resources

**Jameson Mbale**
*Copperbelt University, Zambia*

## ABSTRACT

*The ZAMREN member institutions deal with heterogeneous teaching and research materials drawn from all walks-of-life such as industry, and NRENs world over. To deal with such huge data that is in terabits for academic and economic gain becomes a mammoth task to manipulate, process, store and analyse. It is in view of that the ZAMREN Big Data and Data Management, in this work abbreviated as ZAMBiDM, is envisaged to collectively gather relevant heterogeneous large volumes of a wide variety of data from all sectors of economy. The data would be analytically managed in storage, processing and obtaining actionable insight real-time as a way to solve high-value skilled academic and industrial business problems, in order to prepare graduates for competitive future workforce. The data would be collected from all line-ministries of Zambia such as education, agriculture, health, mining, lands, communications, commerce, including industries and NRENs worldwide and be analytically analysed to exploit strategic actions that would enhance decision making in executing relevant tasks.*

DOI: 10.4018/978-1-5225-0886-1.ch003

## INTRODUCTION

The ZAMREN Big Data and Data Management, in this work abbreviated as ZAMBiDM, is envisaged with the capacity to handle the heterogeneous large volumes of data drawn from all sectors of economy, including the NRENs worldwide. The proposed ZAMBiDM is demonstrated in Figure 1 as model scale-out architecture with capacity to handle the four "V's", these are volume, variety, velocity, variability and complexity. The model has strategically built-in components that would accelerate processing of ZAMREN Big Data. These components include the ZAMREN Big Data and Management which would deal with the institutions Big Data process it to achieve business and technological goals. Surrounding it is the virtualisation which whittles down Big Data and reduces its sprawling bulk to manageable size. This is followed by a series of strategy components that are linked to the operational nodes.

### Statement of the Problem

In this era, as technology evolves, the demands on the use of large volumes of data escalate. Managing such information in form of Big Data has become a challenge. In order to overcome such challenges, the industry, particularly in the developed world has engaged various technologies to manage the Big Data storage and processing. Some of these technologies and methods employed are resource virtualisation, analytic of data to predict the output, utilisation of analytic tools, Big Data management, and management of the four (4) V's. However, none or very little has been done to tackle the Big Data challenges in the developing world. It is in view of that the ZAMBiDM was envisaged to collectively bring all the resources from the ZAMREN member institutions. Using the collective resources, the ZAMREN would build the ZAMBiDM with the capacity to manage the four V's, these are volume, variety, velocity and variability of data.

### Objective of the ZAMBiDM

The ZAMBiDM is developed in-line with the following objectives:

- Design and build the ZAMBiDM architecture.
- Write the ZAMBiDM road map.
- Employ highly qualified data analytics, and data-were housing personnel.
- Develop data management, analytic and quality tools.
- Acquire servers to accommodate large volumes of data.

- Collect and store large heterogeneous volume of data from academic and industry sectors including the NRENs worldwide.
- Virtualise all the data, software and hardware resources.
- Manage the four (4) "V's" of data and these are volume, variety, velocity and variability.
- Build operational processes or nodes.
- Develop relevant strategies for the respective operation nodes.
- Elevate ZAMBiDM data to executive level.

## Organisation of the Chapter

The Chapter is organised in the following parts: the first section introduces the ZAMBiDM and highlights the statement of the problem. The Literature Review section discusses various Big Data technologies employed in industries. The third section introduces the proposed ZAMBiDM model and explains the functions of the system components. The fourth section discusses the implantation of the model. The last section concludes the ZAMBiDM's operations and highlights the benefits of envisaging such a model.

# LITERATURE REVIEW

The emerging, handling and benefit of Big Data had been discussed by other scholars. In this work the Big Data had been looked at dwelling on the four following areas: management, strategies, analytics and virtualisation.

## Big Data

Philip Russom (2013) characterised Big Data as a very large data sets in form of structured which included relational data; unstructured, which involved human language text; semi-structured which comprised RFID and XML; and streaming, which included machines, sensors, Web applications, and social media. Philip further, sited some examples, such as the insurance companies that processed unstructured Big Data used technologies for Natural Language Processing in the form of text analytics, where the output would feed into older applications for risk and fraud or actuarial calculations that benefited from large data samples. Second example was bout the Big Data that streams from sensors enabled companies manage mobile assets, deliver products, identify noncompliant operations, spot vehicles that need maintenance, quality assurance in manufacturing industry, but to site a few.

According to Oracle White Paper (2014) defined Big Data as a concept applied by people to describe data sets whose size was beyond the capability of commonly used software tools to capture, manage, and process the data. The Report stressed that the sheer size of the data, combined with complexity of analysis and commercial imperative to create value from it, had led to a new class of technologies and tools to tackle it. The Report further narrated that the concept tended to be used in multiple ways, often referring to both the type of data being managed as well as the technology used to store and process it. The Report sited some companies where such technologies originated from such as Google, Amazon, Facebook and Linked-ln, to name a few, where they were developed to analyse the massive amounts of social media data they were dealing with. The Report summed up the description of Big Data by expressing that the concept was being described by the four (4) V's and these were: first, the volume, which classified the size of the data in larger quantities you have never encountered in your organisation; second the velocity, the rate at which the data was being processed; third, variety, expressed the syntax and semantics that determined the extent to which data could be reliably structured into a relational database and content exposed for analysis; and fourth, value, was the extent to which the commercial value of the data could be predicted ahead of time so that ROI could be calculated and project the budget acquired by the company.

## Data Management

Philip Russom (2013) defined Data Management as a broad practice that encompasses a number of disciplines, including data warehousing, data integration, data quality, data governance, content management, event processing, database administration to name a few.

## Big Data Management

Having defined Big Data and Data Management separately, Philip Russom (2013) combined the two to define Big Data Management (BDM) that it was about these concepts and on how they work together to achieve business and technology goals. He highlighted that preparing Big Data for use in different departments and applications required a combination of Data Management tools, techniques and teams. He also emphasised that the collaboration of multiple Database Management and the integration of their tools and platforms were critical success factors for Big Data Management. He further stated the benefits that included accurate business insights, greater business value from Big Data, and business optimization.

## Big Data Strategies

Philip Russom (2013) expressed the importance of strategies that they needed to spell out a wide range of road maps, standards, preferences, and stewardship guidelines, depending on what the organisations needed and had culture for. He cited some strategies that the company could lay out a road map for maturing from structured to semi-structured to unstructured data. He also emphasised that since Big Data comes in different forms, the companies or organisations needed to determine the preferred platforms and interfaces for capturing, storing, and processing for each form. He further noted that companies should design a workflow for managing Big Data forms in their original state, plus processing that into other states for customer intelligence, data warehousing, discovery analytics, and others.

According to Issues Paper (2011) expressed that Big Data Strategy was intended for agency senior executive and business programme managers. It stressed that Big Data Strategy was designed to highlight key opportunities and challenges that Big Data would bring to Government agencies. It further explained that strategy aimed at assisting agencies to take advantage of such opportunities and realise the potential benefits of the new technologies. It emphasised that strategy would deliver on elements of the ICT Strategy and that Big Data allowed more focused and evidenced-based policy design and service implementation that in turn allowed citizens to interact with the Government in a personalised and seamless way.

## Big Data Analytics

Oracle White Paper (2014) described data analytic as the systematic computational analysis of data or statistics, by extracting knowledge that can commercially be exploited in some form. Whereas, Issues Paper (2013) said that Big Data Analysis provided profound insight into a number of key areas of society including health care, medical and other sciences, transport and infrastructure, education, communication, meteorology and social sciences. The Paper also highlighted the key areas that Big Data Analytic may influence included the data management, personalisation of services, problem solving and predictive analytics, productivity and efficiency. The Paper further stressed that a successful Big Data Strategy was expected to assist in realising each of the priority areas observed in the ICT Strategy such as the delivery of better services, improved efficiency of Government operations, and open engagement. Wayne, R. (2011) explained that organisations embarking on an enterprise data strategy needed a broad-based portfolio data management tools to support various data integration and quality tasks and automate processes where

possible. Actually, the Wayne, R. (2011) emphasised that executive must recruit business managers and analysts to formulate and oversee a data strategy, define rules, policies and procedures to maintain the accuracy, completeness and timeliness of critical data elements, and partner with IT to execute the programme.

Actually, Peason,T., at el. (2013) narrated that early adopters of Big Data Analytics had gained a significant lead over the rest of the corporate world. He reported that more than four hundred (400) large companies examined were found that those with the most advanced analytics capabilities were outperforming competitors by wider margins. He vividly outlined the four (4) areas where analytics could be relevant and these included: improving existing products and services, improving internal processes, building new products or service offerings, and transforming business models. He also expressed that Big Data leaders work on developing a horizontal analytics capability by learning how to overcome internal resistance, and create both the will and the skill to use data throughout the organisation. He further cited some prosperous firms such as the Nordstorm which elevated responsibility for analytics to a higher management level in its organisation, pushed to make analytical tools and insights more widely available and embedded analytics-driven goals into its most important strategic initiatives. The other example is about the electronic company which added incentives for senior executives to tap Big Data capabilities, and firm's leadership had reinforced that approach with a steady drumbeat of references to the importance of analytics in delivering business results.

In view of the above, Carl, W. O. at el., (2012) argued that the Big Data Analytics was not new, but what has changed was the awareness that analytics could create competitive advantage if the right information was provided to the right decision makers at the right time using the right tools.

## Big Data Virtualisation

Ashutosh (2012) expressed that the next frontier was learning how to manage Big Data throughout its entire lifecycle and that virtualization was the secret weapon that organizations could wield to battle the Big Data management and the challenge. He emphasized that virtualization was indeed the "hero" when it came to managing Big Data. Ixia Report (2014) pointed out that data virtualisation was the easiest, fastest, and most agile data access approach for analytics. The Report stressed that serving data from across big data, enterprise and cloud sources, analytics solutions are more complete and accurate with data virtualisation. The Report further, highlighted that virtualisation was emerging as an effective way to simplify the deployment and provisioning of Hadoop, SQL and NoSQL databases. The Report gave an instance

where a company's Big Data application could be virtualised in the cloud, without a need to purchase big servers. It went on to clarify that a company could acquire processing power on demand without having to build a data centre for data analysis or they could even lease servers by the hour. The Report stressed that virtualisation raised efficiency dramatically and offered the flexibility to scale and meet changing needs. Also, the Report emphasised that only virtualisation could whittle down Big Data and reduce its sprawling bulk to manageable size. In addition, the Report explained that organisations virtualise their data sets so that multiple applications could reuse the same data footprint, plus a smaller data footprint could be stored on vendor-independent devices.

### Zambia Research Education Network (ZAMREN)

According to NUANCE (2008), Mbale (2006) narrated that Zambia Research Education Network (ZAMREN) was a specialised internet service provider dedicated to supporting the needs of the research and education communities in Zambia. They also pointed out that its membership was open to Universities, Research Institutions, Colleges, Communications utility companies and various Government Departments. They further alluded that ZAMREN was an association for tertiary level research and education institutions, which collectively intend to: secure low-cost broadband connectivity to all its member institutions; share its education resources via the dedicated infrastructure; and provided Advanced ICT services to its member institutions.

## THE PROPOSED ZAMBiDM MODEL

The ZAMBiDM model demonstrated in Figure 1 would be composed of four functional components these include: ZAMREN Big Data and Data Management as the core component, virtualisation which is the outer of the core, a series of operational Strategies connected to processes in the form of units or nodes.

### The ZAMREN Big Data and Data Management Component

The ZAMREN Big Data and Data Management would consist of heterogeneous data repository from Zambian public and private sectors and world-wide NRENs. The data could be structured, semi-structured, non-structured, etc. Such data would cross-cut all sectors of work. In fact, ZAMREN as an UbuntuNet Alliance member is well connected to the NRENs world over. Such a link would give ZAMREN an opportunity to access a wider range of academic and research materials from

*Figure 1. The ZAMBiDM Model*


the NRENs across the world. The majority of the resource would come from the academic sectors covering teaching, research, library, weather, sport materials, and news. The teaching materials would include: notes, slides, electronic books, but to name a few. Whereas, the research would comprise of prescribed Conference and Journal materials.

## ZAMBiDM Virtualisation

Surrounding it is the multiple resources, in this work classified as ZAMBiDM sever, network, operating system-level and application virtualisation mechanism. As discussed above, the ZAMBiDM would draw huge heterogeneous data in terabits from all walks-of-life. To manage such voluminous data it is an enormous and complicated task. To operate such system, multiple applications and resources such as operating systems, need to be run and processed preferably on a single hardware installation through a virtualisation mechanism. Such a mechanism has the capacity to configure multiple resources to function in a single hardware. That's, virtualisation would scale down the huge data to a manageable operational size of the resource. In

addition, the huge resources would be virtualised to make the multiple applications be reused the same resource footprint. In fact, ZAMBiDM would virtualise its Big Data to be operated in the cloud computing environment, where it would not require large servers and other relevant equipment. In fact virtualising the ZAMBiDM would yield a lot of benefits to the system. Actually, the Red Had Summit (2014) described such benefits as: easy access to Big Data; seamless integration of Big Data and existing data sets; sharing of integration specifications; collaborative development on Big Data; any reporting or analytical tool can be used; enterprise democratization of Big Data; fine-grained of security Big Data; and increased time-to-market of reports on Big Data. Such attributes benefits would make the ZAMBiDM very agile in accessing the huge data for analytics processing.

## ZAMBiDM Strategies and Operational Processes/Nodes

Around the virtualisation mechanism follows a series of dynamic strategies of actions that are connected and drives the ZAMBiDM functional nodes. Depending on the revolutionised technologies, these series of strategies automatically are transformed to fit the new wave of environmental methods. Each set is developed with a numerous of strategies that are operational of that particular process. The integration and synergies of these numerous strategies and process produce desired functions.

The ZAMBiDM dynamic strategies are connected to Operational Processes or Nodes these include: analytic tools, data quality tools, data accessibility, update data definitions, access data needs, elevate data to executive level, road map manage Big Data, and many more.

## Analytic Tools

The ZAMBiDM system Engineers, would develop the dynamic analytic tools that are adaptive and robust to both current and future technological trends. In that way, the tools would produce effective data analytical applications with the capability to provide intended solutions. These tools would support and accommodate all kinds of data analysis processing. Stephen at el. (2009) listed and discussed the characteristics of such analytic tools as: provision of interaction and function that might be needed by those who used it for their analytical tasks; supported efficiency, seamless transition from one step to the next of the analytical process; grounds the entire analytical experience in a single, central workplace, with all displays, interactions, and functions within easy reach from there; and doesn't require a lot of fiddling with things to whip them into shape to support your analytical needs.

## Data Quality Tools

Iwing, at el. (2014) defined data quality as the data which has fitness for use. Hence, the ZAMBiDM would deal with data quality which is fit for use in all walks-of-life. Actually, the ZAMBiDM needs to set up the principles of data quality tools these include, the institution's vision, policy, strategies, data cleaning, prioritising data quality procedures, and dimension of data quality. The actual data quality tools are given by Andreescu at el., (2014) such as the Centrus Merge/Purge, ChoiceMaker, ETLQ, ETI*Data Cleanser, Entity Search Server, FirstLogic, Hummingbird ETL, Integrater, Information ETL, Identity Search Server, DataBlade, DataStage, DeDupe, dfPower, DoubleTake, MatchIT, Migration Archtect, NaDIS, Sanopsis, Sagent, SQL Server 2000 DTS, QickAddress Batch, WizeSame, WizSame, WizeWhy, but to name a few. However, under Data Quality, it is recommended that there is no need to re-envent the wheel, as such, ZAMBiDM would use the listed data quality tools.

## Elevate Data to Executive Level

The collected heterogeneous data should be analytically transformed to executive consumption. This is where the data could be used to predict the performance of the output. The information provided could help the managers have an insight of production. In that way, they will be able to make tangible plans and decisions on the management of the success of their ZAMREN member institutions. The managers would heavily strategize on the accessing and utilising of huge, in terabits of relevant teaching and research data. As the saying goes that: "information is power", this would avail the necessary data for institution's development. Adequate and viable information is useful in a number of ways such as reduction of escalating cost of information gathering, measure the effectiveness of individual processes, raise enterprise visibility, and make the correct decision at right time.

## Road Map Manage Big Data

The ZAMBiDM development team would need to reposition in setting up a number of strategic preparation that includes: human capacity development, adequate data storage, produce tools for effective data accessibility, and data definitions.

# IMPLEMENTATION OF ZAMBiDM

The ZAMBiDM recruits highly qualified personnel with skills in data analytics, resource virtualisation, data warehousing, network and system administration. These

specialised personnel are employed to manage the ZAMREN Big Data. The System Administrators and other Engineers collect heterogeneous data that includes teaching, research and industrial resources from all walks-of-life and store it in the ZAMBiDM system. Then, these experts manage the ZAMREN Big Data by formatting it to fit into the system. Thereafter, they carry out virtualisation to scale down voluminous and heterogeneous data to a level where they can manage such resources. Having virtualised the resources, the ZAMBiDM experts develop multiples of the strategies that drive the functionality of the connected processes as demonstrated in Figure 1. Also the experts design and develop the processes based on the user's requirements. As new requirements arise, the System Administrators, Network Engineers and others would quickly design and develop new processes and their strategies.

## CONCLUSION

The ZAMBiDM is envisaged as a tool that would enable ZAMREN member institutions to adapt to the dynamic and revolutionised technological trends that demand the usage of high volume of data classified as Big Data. The success in managing Big Data by the ZAMBiDM would revolutionise the education and research in the ZAMREN community. This is supported by McKinsey Report (2011) which stated that Big Data had the potential to revolutionize not just research, but also education. The establishment of this envisaged tool should extract value from the ZAMBiDM data which would enable the institutions operate at a business level for their economic gain and sustain their academic programmes. This is also supported by the IDC Report which described Big Data technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis. In view of this, ZAMREN as a consortium composed of various academic institutions, require the use and deployment of state-of-art technologies to effectively process the heterogeneous data drawn from all sectors of economy. This implies that academic institutions should not only operate as non-profit generation entities, but they should also run like industries that sustain their programmes. For instance, tertiary institutions in the sub-Saharan region rely on government funding or support. Hence, this work is envisaged to equip the tertiary institutions with the tools that would allow them manage the operations of their programmes at business level. That's the ZAMREN member institutions be in a position to implement the data analytic which would help them predict the following: academic performance output, and identifies operational constraints. In that way, the ZAMREN member

institutions would be able to map up strategies for improvement of student's passing rate, generation of funds to sustain the institution programmes, but to mention a few. In fact, the ZAMREN institutions should be in a position to utilise the Big Data management in order independently provide solutions to their existing problems or challenges. This is emphasised by Rock (2011) who pointed out that to acquire solutions for corporate goals, requires the use of business intelligence and data management that would exploit and manipulate strategic actions, organisational capabilities, and enabling technologies. In this way, ZAMREN institutions require the manipulation of the ZAMBiDM Big Data into a business venture for economical gain and yield high student passing rate.

## REFERENCES

Ashutosh, A. (2012). *Website.* Retrieved from http://www.forbes.com/sites/ciocentral/2012/07/05/best-practices-for-managing-big-data/

*Big Data Strategy - Issues Paper.* (2013). Commonwealth of Australia.

Carl, W. O., & Dan, V. (2012). *White Paper, Big Data: trends, strategies, and SAP technology*. IDC.

Ixia Report. (2014). *Designing network for the big data superstorm*. Author.

Jameson, M. (2006). *Zambia Research and Education Networks (ZAMREN) Formation as an Emerging Academic Networking Strategy*. Author.

McKinsey, R. (2011). *Advancing Personalized Education*. Computing Community Consortium.

*NUANCE*. (2008). Retrieved from http://www.ubuntunet.net/zamren

Oracle White Paper. (2014). *Information Management and Big Data – A Reference Architecture*. Author.

Peason, T., & Wegener, R. (2013). *Big Data: The organizational Challenge*. Bian and Company.

Russom, P. (2013). *TDWI Best Practices Report: Managing Big Data*. Cloudera, Dell Software, Oracle, Pentaho, SAP.

IDC Report. (2011). *Extracting Value from Chaos*. Retrieved from http://idcdocserv.com/1142

Rock, R. (2011). *Data Management for BI: Big Data, Bigger Insight, Superior Performance*. Aberdeen Group, a Harte-Hanks Company.

Wayne, R. (2011). *Creating an enterprise data strategy: managing data as corporate asset*. BeyeNetwork.

# Chapter 4
# Predictive Analytics in Operations Management

**Harsh Jain**
*Indian Institute of Information Technology, Allahabad, India*

**Amrit Pal**
*Indian Institute of Information Technology, Allahabad, India*

**Manish Kumar**
*Indian Institute of Information Technology, Allahabad, India*

## ABSTRACT

*Operations management is a field of management which emphasizes on managing the day to day operations of business organizations. These organizations possess a huge amount of data which needs to be analysed for proper functioning of business. This large amount of data keeps some useful information hidden inside it, which needs to be uncovered. This information can be retrieved using predictive analytics techniques, which predict the patterns hidden inside the data. This data is heterogeneous, processing of such huge amount of data creates challenges for the existing technologies. MapReduce is very efficient in processing this huge amount of data. In the field of operation management, data needs to be processed efficiently, so it is highly required to process data using parallel computing framework due to its large size. This chapter covers different techniques of predictive analytics based on MapReduce framework which helps in implementing the techniques on a parallel framework.*

DOI: 10.4018/978-1-5225-0886-1.ch004

## PREDICTIVE ANALYTICS

Predictive Analytics comes under the field of data mining which attempts to analyse the data and extract information out of it ("Big data Analytics and Predictive Analytics", 2015). Predictive analytics is very helpful in the field of operations management as it helps in predicting the behaviour of certain operations. Information extracted out of raw form of data can be used to present trends and behaviours that are hidden inside the data. Predictive Analytics is applied to any event whether from present, past, or future. For example, identifying any fraudulent event in context of credit cards or identifying suspects involved in a crime. Predictive Analytics refer to applying several techniques on historical and past data to visualize future outcomes ("What is Predictive Analytics?", 2015).

Predictive Analytics compute probabilities for each and every possible outcome, and perform prediction at detailed level of granularity. Prediction differs from forecasting in a way that it is a technology which learns from experience to predict the future trends to deduce better conclusions.

Predictive Analytics is a technique which seeks to uncover hidden patterns and relationships in data. These techniques can be classified based on different parameters ("Predictive Analytics", 2015):

1. Based on underlying methodology:
    a. Regression technique
    b. Machine learning technique
2. Based on type of outcome variables:
    a. linear regression address continuous outcome variables
    b. others such as Random Forest

Predictive Analytics, a statistical and data mining technique that can be used on any kind of data, structured or unstructured, is certainly not a new technology (Halper, 2014). In fact, it is in use for decades. However, market adoption and visibility of the technology is increasing for a number of reasons:

1. **Computing Power Increases:** In past it used to take hours or days to get the output of a predictive model which now takes minutes. In early days, it was rather difficult to afford the computing power needed to analyse data that changes regularly in real time environment. With the rise in computing power it is now possible for the organizations to use predictive analytics to analyse data and predict future for their business (Halper, 2014).

2. **Value is Better Understood:** Almost every organization wants to take Business Intelligence to next level to unfold the regularities and irregularities hidden inside the data related to their business. These organizations are interested in knowing how their customers will react to the given scenario based on past experiences. They understood the value of predictive analytics (Halper, 2014).
3. **Economic Consideration:** The recession has affected every business to greater extent. Organizations have realized the importance of data, that it can be very useful to understand market and its trends. Adopters realize that it is very important to gain insight of every aspect related to data. To be successful in a competitive environment, companies must utilize data and analytics to its fullest advantage (Halper, 2014).

## TYPE

Generally, predictive modelling refers to "scoring" data with predictive analytics and forecasting techniques. However, the term "predictive analytics" is used to refer to various disciplines/models, such as descriptive models, predictive models or decision models ("Predictive Analytics", 2015).

## 1. Predictive Models

Predictive models analyse the relation between one or more attributes of the input data ("Predictive Analytics", 2015). These models are used to estimate the likelihood that a similar unit will exhibit the same performance in different sample. This category focusses on modelling in many fields, such as operational management, where they understand the complex data patterns revealing the customer performance and identifying fraudulent activities in a business. Predictive models perform calculation over live transactions, in order to solve many issues related to operational management, such as to evaluate the risk to the organization.

The sample units whose attributes and performances are known, is known as "training samples". While other sample units whose attributes are known but performance is unknown, is called "out of sample" units. The "out of sample" units do not have any relation with "training sample" units. For example, during a simulated crime scene, the blood splatter pattern is an example of "out of sample" units which needs to be analysed.

## 2. Descriptive Models

Descriptive Models establish relationships in data which is used to classify customers and products into groups ("Predictive Analytics", 2015). Descriptive modelling differs from predictive modelling in a way that it identifies many different relationships between customers or products, whereas predictive modelling attempts on predicting single customer behaviour. Descriptive models differ from predictive models in a way as they do not rank the customers on the basis of their behaviour to a particular action. Descriptive models are used to classify the customers based on their product selection.

## 3. Decision Models

Decision models establish relationship between all the elements of decision, the known data attributes, the decisions, the results forecasted, in order to predict decision that involves several variables ("Predictive Analytics", 2015). These models are used in optimizing the favourable outcomes, i.e., maximizing the interesting ones and minimizing the uninteresting ones. Decision models are used to produce set of logical decisions and business rules that will consider every action of customers.

# PREDICTION IN OPERATIONS MANAGEMENT

Operations management deals with various activities of an organization. Operations management includes supply chain management, optimization, production process improvement, analysis of complex system. Operations management deals with the

*Table 1. Distinction between different models of predictive analytics*

| Predictive Models | Descriptive Models | Decision Models |
|---|---|---|
| It attempts to uncover relation and patterns between dependent and independent variables. | It attempts to find cluster of elements that possess similar behaviour/ characteristics. | It attempts to find optimal outcome for a certain decision. |
| Consider only specific variables | Consider as many variables as possible | Consider only specific decisions |
| Examples: next customer preference, fraud, credit worthiness, system failure | Examples: customer segmentation based on sociodemographic | Examples: critical path, network planning, scheduling, resource optimization, simulation, stochastic modeling |

(Buytendijk & Trepanier, 2010)

problems of various sectors such as transport, defence, logistics, health, manufacturing and retail sectors ("Operations Management", 2015). In the field of operations management, prediction can be made in several aspects for e.g., prediction of demand of a particular product, cost or profit estimate and schedule estimation etc. Prediction helps managerial authorities and organizations develop organizational plans and reduce the probability of risks in future. Managers can manage the demand of a product according to its supply and present trend in the market. Prediction in the field of operations management can be done at two levels: quantitative level and qualitative level.

3 kinds of prediction techniques that are used for operations management:

1. **Judgemental Forecasts:** These forecast make predictions on subjective inputs that are coming from various sources ("Operations Management", 2015). This type of forecasting technique is used when the data is outdated, or there is very short time to collect it. Judgemental forecast includes several types of surveys such as consumer surveys, sales-force opinions etc.
2. **Time-Series Forecasts:** A time-series is a sequence of observations that are made at regular interval of times and the sequence is ordered according to time ("Operations Management", 2015). It uncovers the patterns that are hidden in these observations. Time-series forecast works on quantitative level of prediction and helps predict the future events
3. **Associative Models:** Associative models deal with quantitative prediction in business organizations ("Operations Management", 2015). For e.g. it can be used for predicting demand or estimating price of a particular product. It includes a predictor variable and develops a mathematical equation that predicts the value of that predictor variable. The mathematical equation takes some features into account and tries predicting the value of some unknown feature. This predictor variable predicts the value of our interest.

In context of this chapter, some different methods, for making prediction in the field of operations management, for e.g. regression methods and machine learning methods are explained. These methods lie in the category of quantitative predictions. They can be used in several aspects such as predicting demand based on historical data, estimating price of a product based on several features like fixed and variable costs etc., determining the risky events that might occur in future based on past trends etc. All the underlying methods need is some data which is attributed by relevant features. The data will be manipulated and analysed by the techniques, and the result might help the organization solve many of its problems.

## RELATED WORK

In the past years, there has been a trend of handling management activities with data mining techniques, especially if those management activities are concerned with large amount of data. Many researchers have suggested several data mining techniques that help manage the operations in the organization. These organizations vary in range such as corporate sector, IT companies, health organizations, financial organizations and academic institutions etc.

The work of Song et al. (2013) presented that storage management in IT industries can be handled with data mining techniques. There is a large amount of data generated in IT companies on daily basis. So this paper suggests that big data analytics such as predictive analytics on a parallel scalable framework can be used to handle the large amount of data generated by different operations in IT sector.

The work of Islek and Oguducu (2015) suggested that demand of several products in distribution warehouses can be predicted with data mining techniques. The paper suggested some of the machine learning techniques such as Bipartite graph clustering, bayesian Network algorithm and artificial neural network learning which are used for forecasting purpose. The methodology used the datasets of a nuts and fruits company from Turkey. Attributes related to warehouse and products have been used to describe data which is used in training and testing algorithms. This paper concluded that data mining techniques are reliable approaches for predicting demand. Similarly data mining techniques can be used for other operations management activities also.

## ANALYTICAL TECHNIQUES

Predictive analytics techniques and approaches can be broadly categorized into two groups:

- Regression techniques, and
- Machine learning techniques.

### 1. Regression Techniques

Regression techniques behold the foundation of predictive analytics techniques. These techniques derive a mathematical equation. The equation derived can be used as a model to represent relationship between all the features or attributes under consideration. Depending upon the application, there are a plethora of techniques that lie under this category. Some of the techniques are:

- Linear Regression Model, and
- Logistic Regression etc.

## 2. Machine Learning Techniques

Machine Learning, comes under the field of artificial intelligence, came into existence to help computers to learn and understand. Today, it consists of several techniques which find application in wide variety of fields such as fraud detection, medical diagnosis, analysis of stock market, and weather forecasting. Sometimes it becomes very easy to predict the value of unknown dependent variable without analysing the relationship between other variables which are known and independent. However, in some cases, where it is very complex to deduce relationships between variables, machine learning techniques come to the rescue. Some of them are given below:

- Back Propagation Network,
- Support Vector Machine, and
- Naïve Bayes, etc.

## MAPREDUCE

MapReduce is a programming model which is used for processing and generating large data sets in form of <*key,value*> pair. There are two kinds of functions used in MapReduce. A map function is specified which processes data in form of <*key,value*> pair and produces a set of all intermediate <*key,value*> pairs. The job of reduce function is to combine all the intermediate values which are related to same intermediate key after the processing of map function (Dean & Ghemawat, 2008) and (Lammel, 2008).

MapReduce is a program that consists *of map()* and *reduce()* function. *map()* function performs filtering and sorting over data, for example, sorting the names of students into queues. *reduce()* function carries the summary operation over the resultant data, for example, counting the number of students that are present in each row (Lammel, 2008). The "MapReduce" system gives its best performance over a distributed network, performing several tasks in parallel, managing data and control transfer between different units of system and providing fault tolerance.

There are many languages which provides libraries for mapreduce framework. Apache Hadoop provides an open source support for the implementation of MapReduce program over distributed network. "MapReduce" is a term which was originally a proprietary of Google technology, but later it has been generacized. MapReduce

is a framework which is capable of processing parallelizable problems which deals with huge datasets, with the help of large number of computers, referred to as cluster or grid. MapReduce can process data in any form, whether structured (database) or unstructured (file system). It takes the advantage of locality of data (Lammel, 2008).

- **Map Step:** Local data is processed by *map()* function which is applied on each worker node. If same input data is present at two instances, then that data is processed only once.
- **Shuffle Step:** Data belonging to one key is redistributed to one worker node, such that each worker node contains data related to one key only.
- **Reduce Step:** Each worker node now processes data belonging to same key in parallel.

Map and reduce operations are performed in distributed environment. In *map()* construct, mapping operations do not have any relation with each other, i.e., they are mutually independent. Being independent of each other, they can be parallelized. Hence, *map()* operations can be performed simultaneously. However parallel computation depends on the number of independent data sources or the number of CPUs available for processing. Similarly, reducers perform their function efficiently, only if output of the mapper function that belongs to the same intermediate key is provided to the same reducer node. MapReduce can be applied to a significantly large amount of data, for example, MapReduce framework can be used to sort terabyte/petabyte of data in ascending/descending order in few hours. The parallelism of framework also provides fault tolerance and reliability in processing: if one mapper or reducer goes out of proper functioning, processing can still be continued by rescheduling the work.

MapReduce can also be understood as a 5-step parallel and distributed computation (Lammel, 2008):

1. **Prepare the *map()* Input:** There are several map processors which are assigned input key value *K1* and all the data associated to that key value, that each map processor is going to work on.
2. **Run the User Provided *map()* Code:** User provides the code for *map()* function depending on the type of problem statement. Map executes only once for each key value *K1* and generates intermediate <*key,value*> pairs which is organized by key values *K2* at reduce processor.
3. **Shuffle the Output of *map()* Function to Reduce Processor:** The output data produced by *map()* processors is redistributed to reducers so that each reducer contains data belonging to the same key.

4. **Run the *reduce()* Code Which is Provided by User Based on the Problem Statement:** User provides the code for *reduce()* function based on the problem statement. Reduce runs exactly once for each key value *K2*.
5. **Produce the Final Output:** Last step is to collect the output of each reduce processor and sort it by the key value *K2*.

The five steps described above are performed in sequence. i.e., next step cannot be started until and unless previous step completes. However, some intermediate steps can be interspersed only if final result remains the same.

In some cases, it is often found that input data is already partitioned among different servers, so step 1 is sped up because it is not needed to prepare the input, as map processor can process the data which is available on its own site. Similarly step 3 can also be performed efficiently by allocating reduce function to the data closely available to it.

Map and reduce function of MapReduce framework are defined in terms of <*key,value*> pair. Map accepts <*key,value*> pair of data related to a type of domain as input and returns a list of intermediate <*key,value*> pairs related to entirely different domain.

$$\text{Map}(k1,v1) \rightarrow \text{list}(k2,v2)$$

A Mapper function performs its task simultaneously with other mapper functions. It processes input <*key,value*> pair and produces a list of intermediate <*key,value*> pairs as result. Once all mapper functions complete their tasks, it is the task of reducer functions to combine all the values belonging to same key and producing one group for each key representing all the values for that key.

Groups are now processed by different reducer function, which was given as output by map function, simultaneously. Reduce function produces a set of values that belong to same domain (key).

$$\text{Reduce}(k2, \text{list}(v2)) \rightarrow \text{list}(v3)$$

In the above system all the calls can return more than one value, whereas reduce functions return either value v3 or empty set. Finally, the result of all the reduce functions are combined to produce overall result. Therefore the MapReduce framework generates a list of <*key,value*> pairs from a list of values.

# REGRESSION TECHNIQUES

## Linear Regression Model

Linear Regression model is a regression technique of predictive analytics (Elkan, 2013). Let x be a vector of real numbers, of fixed length *p*, which is called dimension or dimensionality of *x*. Let *y* be its real valued label.

Linear Regression model is given by:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots. + b_px_p$$

In the above function, right hand side is called linear function of *x*, which is defined by coefficients $b_0$ to $b_p$. These coefficients are given as output by data mining algorithms.

In the above model, if $x_i=0$ for all i=1, 2, …, p, then $y=b_0$ which is called intercept. However it is not possible for *x* to have all its values as 0. Coefficient $b_i$ represents the amount by which *y* increases if $x_i$ increases by 1 and the values of all other features remain unchanged.

$$\hat{y}_i = f\left(x_i; b\right) = b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

The objective of this model is to find optimal values of coefficients $b_0$ to $b_p$. By optimality, it is meant to say that sum of squared errors on the training set must be minimized, where the squared error on training example *i* is $\left(y_i - \hat{y}_i\right)^2$. This algorithm finds the following:

$$\hat{b} = argmin_b \sum_{i=1}^{n} \left(f\left(x_i; b\right) - y_i\right)^2$$

The objective function $\sum_{i=1}^{n} \left(f\left(x_i; b\right) - y_i\right)^2$ is called sum of squared error or SSE in short. If the number of training tuples n is less than number of features p then the optimal values of coefficients are not unique. Even if n is greater than p, the optimal coefficients have multiple equivalent values, if some of the input features are linearly related. Here, "equivalent" means that the different sets of coefficients achieve the same minimum SSE.

Now the regression equation are best suited for prediction. There are two types of predictions:

1. Predicting the value of *y,* which is called a response variable. The response variable corresponds to value of predictor variable, *x.*
2. Computation of mean response $\mu_0$ when $x=x_0$. For this case, the mean response $\mu_0$ is estimated by $\hat{\mu}_0 = \hat{b}_0 + \hat{b}_1 x_0$

There are various other techniques that comes under the category of Regression model, for e.g. Logistic Regression Model, Discrete Choice Model, Probit Regression etc.

## MACHINE LEARNING TECHNIQUES

### Parallel Backpropagation

Backpropagation is method to train artificial network which is very useful in prediction ("Backpropagation", 2015) and (Han & Kamber, 2001). Though simple backpropagation can be used to train the network, we would like to discuss parallel backpropagation technique because predictive analytics deal with really huge dataset. Backpropagation method focusses on computing the gradient of a loss function with respect to every weight in the neural network. The gradient is then given back to the network as input and weights of the network are updated using the gradient value so that the loss function gets minimized (Butler, 2014).

Backpropagation is a supervised learning algorithm as it knows the desired output for each input value, therefore making it easy to calculate the gradient of loss function. Backpropagation is applied over a "multilayer feed forward network"- a network with multiple layers, one input layer, one or more hidden layers and an output layer, and contains no cycle. The activation function used by artificial neurons of the network should be differentiable.

Huge amount of data are produced continuously from different sources. It is highly tedious task to analyse the pattern hidden inside the data. Parallel version of backpropagation along with mapreduce framework works efficiently to analyse the large dataset. This method is known as MBBN (MapReduce based Backpropagation Neural Network) (Liu et al., 2010).

## Implementation

Simple backpropagation algorithm serves as a basis for the implementation of MBBN. To implement this method, mapper functions and reducer functions are described considering the basic backpropagation algorithm. Programmer gives the key/value pair as input to the mapper function which generates a set of intermediate key/value pairs. Once the mapper's job is done, reducer job is to merge/combine all the values which are related to same intermediate key. Once reducer is also done, final result is obtained (Liu et al., 2010).

The procedure is described below:

1. The system is given several items of the training dataset.
2. All mapper tasks are executed simultaneously. Each mapper function is given one training item as an input which calculates the values of the weights to be updated for that training item.
3. Now several reducers are executed. It collects all the update-values for one weight, using it as an intermediate key, and calculates the average of all the values for that weight.
4. All the weights in network are updated.
5. Repeat step 2, 3, and 4 until expected result is achieved

Backpropagation *mapper()* function is shown in Figure 1, backpropagation *reducer()* function is shown in Figure 2 and backpropagation *main()* function is shown in Figure 3.

*Figure 1. Parallel Backpropagation mapper() function*
*(Liu et al., 2010)*

**Procedure 1 Backpropagation Mapper**

| | |
|---|---|
| 1 | Input key/value pair < *key*, *value* = inputItem >. |
| 2 | Compute the update-value for all the weights using *value* as the input of the network |
| 3 | For every weight *w*, get its weight *w* |
| *4* | Emit intermediate key/value pair <*w*, *w*> |

*Figure 2. Parallel Backpropagation reducer( ) function*
*(Liu et al., 2010)*

**Procedure 2 Backpropagation Reducer**

1 Input key/value pair < *key* = *w*, *value* = *w* >.
2 *sum* ←0, *count* ← 0
3 *sum* ← *sum* + *value*
4 *count* ← *count* + 1
5 **If** more pairs <*key*, *value*> are collected, go to step 1 **Else** output key/value pair <*key*, *sum/count*>

*Figure 3. Parallel Backpropagation main( ) function*
*(Liu et al., 2010)*

**Procedure 3 Backpropagation Main Function**

1 Run the job whose mapper and reducer functions are the backpropagation mapper and reducer in procedure 1 and 2.
2 Compute the output value of each reducer and do a batch update on weights of the network
3 **If** the precision of the network on the training set is not better than the expected precision, go to step 1 **Else** procedure end.

Time complexity of this method is described below.

From the above given description of implementation, it is clearly noted that mapper tasks and training set items are equal in number, reducer tasks and weights of network are equal in number (Liu et al., 2010).

The arcs of the neural network can be computed using the give relation:
Where:

- v= number of attributes, number of inputs in neural network
- o= number of classes for classification, number of outputs of neural network
- L= number of middle layers
- $K_i$= number of nodes at $i^{th}$ middle layer, i=1,2,…, L

Let,

- number of items in training set is n
- number of mapper nodes is m
- number of reducer nodes r.

Complexities:

1. Mapper is assigned the task to compute update-values of all weights, so complexity of mapper task is *O(a).* Overall complexity is *O(an/m).*
2. Reducer is assigned the task to compute average of all update-values, so complexity of reduce task is *O(n).* Overall complexity is *O(an/r).*
3. So the complexity of each iteration of this method is *O(an/m)+O(an/r).*
4. Let there are N number of iterations, so overall complexity of MBBN is:

O(anN/m)+O(anN/r)

## Parallel Support Vector Machine

There is an enormous amount of data available due to the advancements in electronics and computer technology. It has become an important issue for the organizations to process and analyse a huge amount of data. There is an increase in demand of techniques to process large amount of data in almost every field, such as, medical, bioinformatics, business, web, banking, organization etc. Parallel algorithms are the solution to the problem of analysing such huge amount of data.

Support Vector Machine comes under the category of supervised learning, because the classes of classification are already known in this method ("Support Vector Machine", 2015) and (Han & Kamber, 2001). SVM can be used for both classification and regression. It emphasizes on mapping the input training data to

higher dimension. SVM computes a linear optimal separating hyperplane, i,e,, a plane/boundary that separates tuples of one class from another. The main objective of SVM is to find hyperplane using support vector and margins (Butler, 2014).

Since there is a very huge amount of data present, storage requirements increases rapidly. Parallelization is a solution proposed for this problem, which suggests that the problem should be split in subsets and training tuples should be assigned to different subsets. Parallel SVM can be realized with the help of MapReduce framework.

The parallel SVM can be seen as being implemented on cascade model. The SVM network is trained through the solutions of partial *subSVMs*, where each *subSVM* serves as a filter. Solutions obtained from partial *subSVMs* help us reach the global optimum. In this model optimization problem of large scale data is divided into independent and smaller optimization problem. The support vector of previous *subSVM* is fed as input to the next *subSVM* (Sun & Fox, 2012).

In the architecture given below it is clearly visible that the support vector from two *subSVMs* are combined and given as input to *subSVM* of next level. The process is carried out until only one SVM process is left. Ultimately on the last level, the output of the overall system is obtained. In this architecture each *subSVM* has to deal with a smaller set of input training data. Therefore, set of training data for each SVM is smaller than the training data set of whole problem.

Parallel SVM architecture is shown in Figure 4.

## Algorithm

Preparation

- Configure computation environment.
- Data is partitioned and distributed to different computation nodes.
- Create a partition file.

Main class for parallel SVM is shown in Figure 5 and Map class & Reduce class is shown in Figure 6.

The program is explained as follows: The large dataset D is partitioned into smaller size subsets $D_1$, $D_2$, …, $D_n$ and given as input to different nodes. Then a partition file is created according to twister command, which will be used in Twister configuration.

*jobConf* is a command which is used to configure the parameters such as Map, Reduce parameters and class names. While help in computation. *TwisterDriver* is used to initiate MapReduce task. In each node, Map task is performed. Partitioned training data from local file system is loaded into the first layer. After that support

*Figure 4. Parallel SVM architecture based on cascade model*
*(Sun & Fox, 2012)*

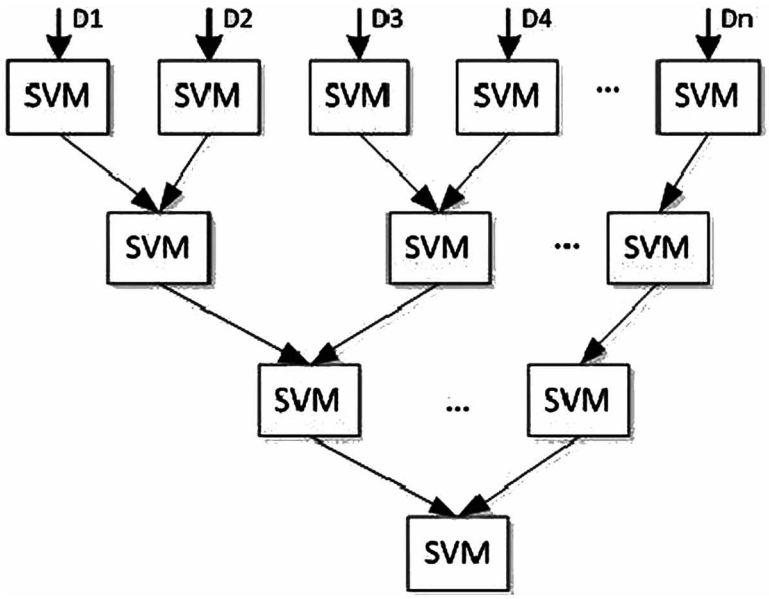



*Figure 5. Parallel SVM main class*
*(Sun & Fox, 2012)*

**Main class**
**JobConf**; //configure the MapReduce parameters and classnames
**TwisterDriver;** //to initiate the MapReduce tasks
While(condition) //not combined to one SVM
**JobConf;** //reconfigure the MapReduce parameters;
**TwisterDriver;** // initiate new MapReduce tasks, Broadcast combined support vectors to each computation node;
Get feedback results;
If(condition) break; // if one SVM obtained, program finished
**End main class**

vectors of previous layer are combined in pair and given as input to the nodes of next layer. Trained support vectors are given to the Reduce jobs as input. Reducer job is to collect all support vectors from all mapper jobs and to send to the client node. This iterative training process will stop executing when support vector of all *subSVM* are combined to one SVM (Sun & Fox, 2012).

*Figure 6. Parallel SVM map class & reduce class*
*(Sun & Fox, 2012)*

```
Map class
    If(the first layer SVM)
      Load data from local file system;
    else
      Read data broadcasted by Main
    class End if
    Svm_train(); //the parameters of the SVM model are transformed through
    jobConf.
    Collector; //sent the training result to Reduce job through message.
End Map class
Reduce class
    Read data transformed from Map job;
    Combine support vectors of each two subSVM into one sample set.
    Collect; //feedback all the trained support vectors
End Reduce class
```

## CASE STUDY

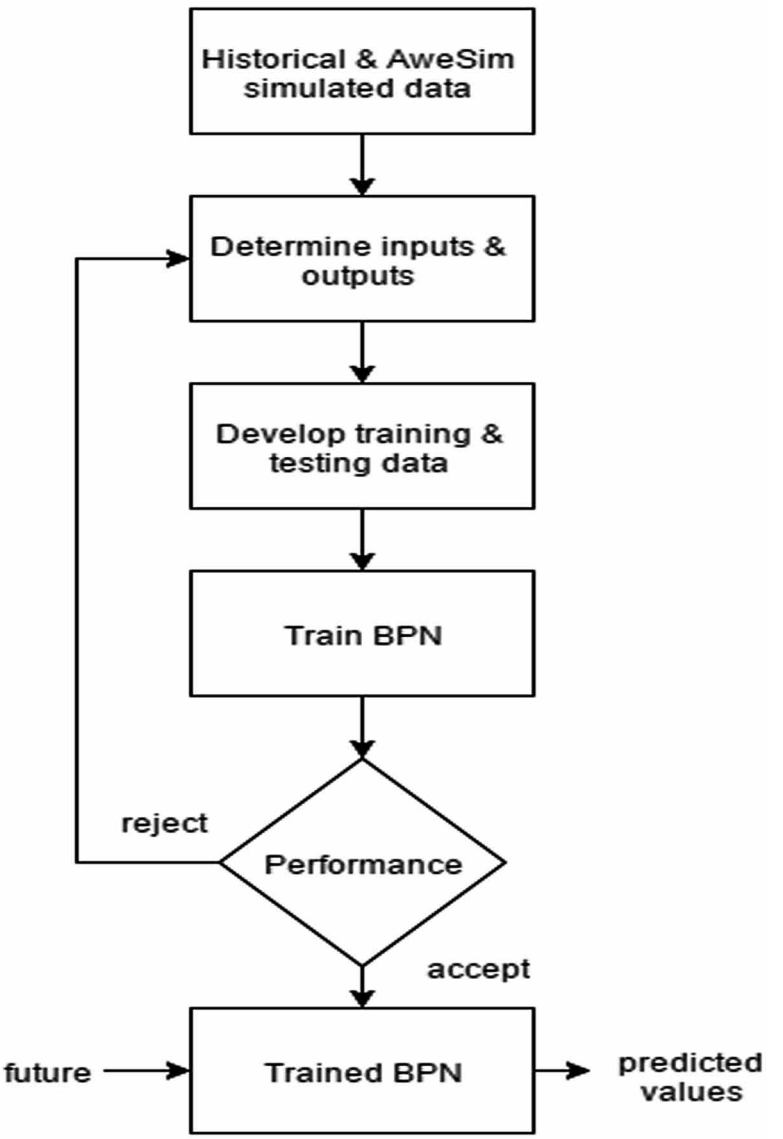
### Implementation of Back-Propagation Neural Network for Demand Forecasting in Supply Chain

Demand forecasting is the most important element of supply chain management in any organization. Forecasting technique must be efficient otherwise it would lead to cost increment, lose customers etc. This case study which is presented by Cheng et al. (2006), trains neural network using backpropagation algorithm. The neural network is trained to make predictions for a corporation in Taiwan which produces electrical connectors. In this study a tool called AweSim was used to accumulate the orders of different types of connectors. This set of data was used to train neural network to give better prediction results.

This study presents the feasibility of a backpropagation neural network in predicting demand for a simulated set of data. In this approach simple backpropagation neural network was used. However to implement the forecasting approach in parallel manner, the above explained Parallel backpropagation on MapReduce can be used (see Figure 7).

According to the study, five variables have been defined, that would define the data and help train the neural network. The five variables are as follows:

*Figure 7. Flow chart for the process of demand forecasting using backpropagation neural network*
*(Cheng, Hai-Wei & Chen, 2006)*



1. T: time interval between two arrival orders
2. Q: quantity of a connector
3. $P_A$: operator time
4. $P_B$: machining capacity
5. $P_C$: inventory level

Data of orders of electrical connectors, provided by the corporation, was not sufficient to train a neural network. So a simulation tool called AweSim was used to simulate large amount of data based on the records provided by the company. The data which is generated by AweSim is normalized, so that the network does not become saturated and indifferent to a new set of records.

Network architecture for demand forecasting consists of three layers namely as input layer, hidden layer, output layer. Input layer consists of five neurons, each corresponding to one of the five variables. Number of neurons in output layer were also five. A gradient descent backpropagation approach is used to train the network. A professional software called Neural Network Professional II/Plus was used to train the network. Once the network has been trained on the training data, it is the time to test it on testing data. Testing data a set of records that has not been used in training process.

The study shows that backpropagation neural network is capable of predicting demands based on the data of electrical connectors. Similarly, the neural network can be trained for predicting the demand of several other products. Neural network has been proved to be a powerful tool in different processes of operations management such as demand forecasting, cost estimation etc.

## RECENT RESEARCHES

There have been several researches done in the field of prediction in operations management. This section describes one of the recent researches that has been carried out by researchers. The work of Sarhani and El Alfia (2014) presented an intelligent system for supply chain demand demand forecasting, which is based on support vector machine. As it is seen earlier in this chapter that several machine learning methods serve the purpose of making predictions in the area of operations management, this paper also presents one of the methods that helps predict the demand in the field of supply chain management.

This paper suggests a hybrid method which includes support vector machine and particle swarm optimization algorithm. Support vector machine is a machine learning method which is used in classification and prediction whereas particle swarm optimization is an optimization algorithm which is used to optimize the parameters of support vector machine in this hybrid approach. This approach uses a modified form of support vector machine which is support vector regression so that it can be used for regression models.

The fundamental concept of SVR is to nonlinearly map input training data into a feature space of higher dimensions. This space may have infinite number of dimensions i.e., $R^{nh.}$.

SVR function is given by the following equation

$$f\left(x\right)=\omega\varphi\left(x\right)+b$$

where,

- $f(x)$: forecasting value
- $\omega$ and $\varphi$: coefficients to be optimized

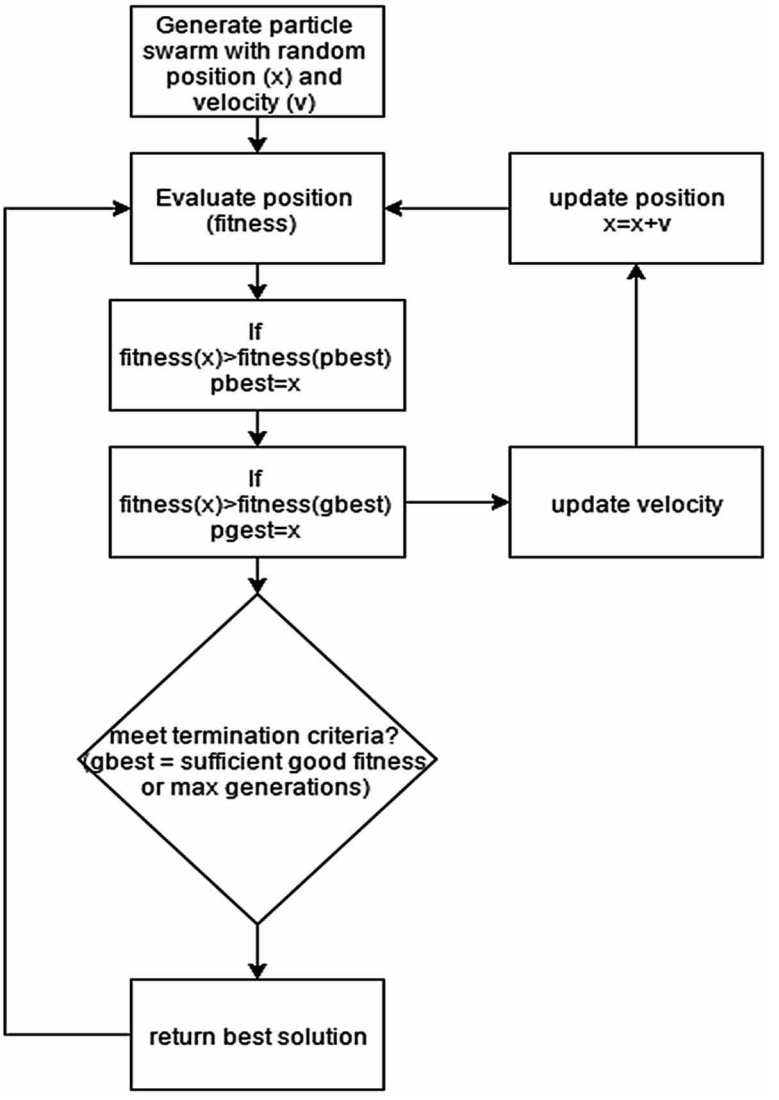
Particle Swarm optimization is an optimization algorithm which is used to optimize the coefficients ω and φ, along with one more parameter which is called penalty parameter, C. It discovers for an optimal combination of these three parameters. In context of this research paper, the focus has been given to optimize ω only, while other parameters are manually provided (see Figure 8).

There are two kinds of data available, training data and testing data. The accuracy of prediction technique is measured by mean absolute percentage error (MAPE) on testing data.

$$MAPE = 100 * \frac{\sum\left(prediction - real\right) / real}{n}$$

This paper compares the results of this hybrid approach with another hybrid approach of same style i.e., SVR coupled with genetic algorithm. The paper concludes that SVR with PSO gives better results than with GA in case of demand prediction.

*Figure 8. Flow chart for particle swarm optimization*
*(Sarhani & El Alfia, 2014)*



## CONCLUSION

To perform prediction analysis over large scale data, coming from operation management process, different prediction analytics techniques can be used. MapReduce is resulting as a great solution for processing the large volume of data using a parallel framework. There is a need to modify the conventional techniques for extracting information from this type of data using parallel framework of MapRecue. These modified parallel techniques have been discussed in this chapter. In order to achieve efficiency and scalability MapReduce framework based methods are used. These techniques helps in predicting different behaviours of operations. There are several other techniques which can be used for similar purpose.

## REFERENCES

Halper, F. (2014). *Predictive Analytics for Business Advantage*. TDWI Research.

Butler, M. (2013). *Predictive Analytics in Business*. Butler Analytics.

Buytendijk, F., & Trepanier, L. (2010). *Predictive Analytics: Bringing the tools to the data*. Redwood Shores, CA: Oracle Corporation.

Song, Y., Alatorre, G., Mandagere, N., & Singh, A. (2013, June). Storage mining: where IT management meets big data analytics. In *Big Data (BigData Congress), 2013 IEEE International Congress on* (pp. 421-422). IEEE. doi:10.1109/BigData.Congress.2013.66

Islek, I., & Oguducu, S. G. (2015, June). A retail demand forecasting model based on data mining techniques. In *Industrial Electronics (ISIE), 2015 IEEE 24th International Symposium on* (pp. 55-60). IEEE doi:10.1109/ISIE.2015.7281443

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. doi:10.1145/1327452.1327492

Lämmel, R. (2008). Google's MapReduce programming model. *Science of Computer Programming*, *70*(1), 1–30. doi:10.1016/j.scico.2007.07.001

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kauffman.

Elkan, C. (2013). *Predictive analytics and data mining*. Retrieved from http://www.cseweb.ucsd.edu

Liu, Z., Li, H., & Miao, G. (2010, August). MapReduce-based backpropagation neural network over large scale mobile data. In *Natural Computation (ICNC), 2010 Sixth International Conference on* (Vol. 4, pp. 1726-1730). IEEE. doi:10.1109/ICNC.2010.5584323

Sun, Z., & Fox, G. (2012, January). Study on parallel SVM based on MapReduce. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*.

Cheng, Y. H., Hai-Wei, L., & Chen, Y. S. (2006, June). Implementation of a back-propagation neural network for demand forecasting in a supply chain-A practical case study. In *Service Operations and Logistics, and Informatics, 2006. SOLI'06. IEEE International Conference on* (pp. 1036-1041). IEEE.

Sarhani, M., & El Afia, A. (2014, November). Intelligent system based support vector regression for supply chain demand forecasting. In *Complex Systems (WCCS), 2014 Second World Conference on* (pp. 79-83). IEEE. doi:10.1109/ICoCS.2014.7060941

Predictive Analytics. (2015). In *Wikipedia*. Retrieved August 27, 2015, from https://en.wikipedia.org/wiki/Predictive_analytics

*What is Predictive Analytics ?* (2015). Retrieved August 27, 2015, from, http://www.predictiveanalyticsworld.com/predictive_analytics.php

*Big Data Analytics and Predictive Analytics*. (2015). Retrieved August 30, 2015, from, http://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics

Backpropagation. (2015). In *Wikipedia.* Retrieved September 9, 2015, from https://en.wikipedia.org/wiki/Backpropagation

Support Vector Machine. (2015). In *Wikipedia.* Retrieved September 15, 2015, from https://en.wikipedia.org/wiki/Support_vector_machine

*Operations Management*. (2015). Retrieved November 18, 2015, from https://ids355.wikispaces.com/Ch.+3+Forecasting

## KEY TERMS AND DEFINITIONS

**Backpropagation:** Backpropagation is an algorithm to train the multilayer feed foreword neural network. The key concept of backpropagation algorithm is based on gradient descent method, which attempts to reduce error using gradient approach.

**Linear Regression Model:** Linear regression model builds a model that defines a linear relationship between dependent variable and one or more independent variables.

**Machine Learning:** Machine learning deals with the study of algorithms that manipulates the data and help the machine learn and make predictions. Machine learning includes several techniques such as associative modelling, clustering, classification etc.

**MapReduce:** MapReduce refers to a programming model which manipulates the data in form of <key,value> pair. MapReduce consists of two key processes: map is a process to analyse the data in <key,value> pair and generates a set of intermediate pair and reduce is a process which combines all the values corresponding to same intermediate key.

**Operations Management:** Operations management includes management techniques for several operations in an organization such as demand forecasting of a product, price estimation, storage management activities etc. It deals with designing, developing and managing the operations in business organization.

**Predictive Analytics:** Predictive analytics refers to the data mining techniques which analyse and manipulate the data to predict future trends and uncover the hidden patterns. It includes various methods such as regression methods, machine learning methods etc.

**Regression Analysis:** Regression analysis is a statistical technique which serves the purpose of establishing the relationship between known and unknown variables. It focusses on computing the value of dependent variable based on the analysis of one or more independent variables.

**Support Vector Machine:** It is a machine learning model which is used in classification and prediction problems. It comes under the category of supervised learning as the class of every record in the dataset is known beforehand. It builds a model which basically assigns the records of the dataset to their respective classes.

# Chapter 5
# Pros and Cons of Applying Opinion Mining on Operation Management: A Big Data Perspective

**Mahima Goyal**
*Ambedkar Institute of Advanced Communication Technologies and Research, India*

**Vishal Bhatnagar**
*Ambedkar Institute of Advanced Communication Technologies and Research, India*

## ABSTRACT

*The web data is growing at an immense pace. This is true for the social networks also. The data in the form of opinion of an individual is gathered to find the nuggets out of the same. The development in the application of opinion mining is rapidly growing due to various social sites which prompted us to pursue exhaustive literature survey in the field of opinion mining application in operation management and to classify the existing literature in this field. In this context the authors had identified the pros and cons of applying the opinion mining on operation management from the perspective of big data. The authors had considered the amount of data involved to be too big and for the same the big data concept is of primarily utmost significance. The authors also proposed a framework which clearly depicts the usage of the opinion mining on operation management of various domains.*

DOI: 10.4018/978-1-5225-0886-1.ch005

## INTRODUCTION

Data collected from the social media channels is huge and unstructured. It must be made ready to use. Here comes the role of big data to store and analyze these large volumes of data in an effective and efficient way. 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. Big Data analytics are the ways to convert the unstructured data available in the form of the social media into the structured data. Different tools for big data analytics must be developed for performing this task.

Sentiment analysis (opinion mining) techniques analyze opinionated text, which contains people's opinions toward entities such as products, organizations, individuals, and events. Businesses are increasingly capturing more data about their customers' sentiments that has led to the proliferation of sentiment analysis. Marketing, finance, and the political and social sciences are the major application areas of sentiment analysis. Sentiment analysis techniques are further divided into three sub-groups, namely document-level, sentence-level, and aspect-based.

Opinion mining has been applied to a number of domains like hotels and restaurants, different products, movies and politics. Not only this, but the ever growing growth of information on social media platforms has influenced many companies to use this analysis in the operational management as well. Operational management in the field of marketing can be used to predict the sentiment of customers for any new product launched.

An another application of it would be identifying trends and events that act as warning signs for supply-chain risks. Social media channels would be used for monitoring critical keywords or phrases associated with the suppliers' names or specific sourcing markets. Such a system can help continuously update suppliers' and sourcing markets' risk profiles and even trigger contingency plans in case of, e.g., bankruptcies or natural disasters .

As opinion mining has its advantages, it has its disadvantages too. In some of the recent reports, it is being seen that people are being paid to write fake reviews of particular products which can hamper the growth of the product. It is being seen that fake users tend to write good reviews about the bad products and bad reviews about the good products. This term is called opinion fraud.

The chapter will be further divided into various subtopics. The sections include an introduction, Research Review, a framework for pros and cons of opinion mining in operation management based on the big data perspective. The framework for research includes the research methodologies used in writing the review. The analysis for opinion mining includes the steps performed while analyzing the classification framework.

## Literature Review

*Table 1. Literature review*

| Name of the Authors | Name of the Article | Features of article |
|---|---|---|
| 1.Wood et al. (2013 a) | Using Sentiment Analysis to Improve Decisions in Demand-Driven Supply Chains | Sentiment Analysis has the potential be used in supply management decisions. |
| 2.Wood et al.(2013b) | Expanding Sales and Operations Planning using Sentiment Analysis: Demand and Sales Clarity from Social Media | Sentiment Analysis can be used as a tool in sales and operation planning(S &OP) |
| 3.Gandomi A. and Haider M. | Beyond the Hype: Big Data Concepts, Methods, and Analytics | Big data can be used in social media analytics to convert to handle high volumes of user generated content. |
| 4. Wood et al.(2014 c) | Sentiment Analysis in Supply Chain Management | Sentiment analysis can be used by distant suppliers in retrieving the information on consumer demands. |
| 5. Hu M. and Liu B. (2004) | Mining and Summarizing Customer Reviews | A feature based summary of the customer reviews using data mining techniques |
| 6. Marrese-Taylor et. al (2014) | A Novel Deterministic Approach for Aspect-Based Opinion Mining in Tourism Products Reviews | An aspect based approach is shown on tourism products. |
| 7. Ding et al.(2008) | A Holistic Lexicon Approach to Opinion Mining | This paper shows an overall approach of opinion mining by finding the opinion words which are context dependant. |
| 8. Liu B. (2012) | Sentiment Analysis and Opinion Mining | The definitions, types and examples of opinion mining are explained. |
| 9. Bai X.(2012) | Predicting Consumer Sentiments from Online Text | It provides the vocabulary for extracting the sentiments by detecting the dependencies among the words. |
| 10. Kwon et al.(2014) | Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics | Analytics of big data in the form of data quality management is depicted using empirical study. |

## PROS AND CONS FRAMEWORK

Opinion forms the basis for people, organizations and social communities to take accurate and effective decisions. People mostly ask their friends, peers and knowledgeable persons about their opinion on an entity while taking the decision, since, they believe the experiences, observations, concepts and beliefs of other individuals will help them in boosting the decisiveness of that entity. With the recent growth of the social media on the web in the form of reviews, blogs, Twitter, forum discussions and social networks, the inception of expressing the opinions and views on these portals has become a huge trend. Due to this large volume of opinionated data is being loaded for analysis. Thus, opinion mining plays a pivotal role to focus on the opinions which express or imply positive or negative sentiments. It involves both Natural language Processing(NLP) and machine learning techniques. Opinion Mining has various applications. It can be applied to different consumer products and services. It can be used for twitter analysis and measuring sales performance of an organization.

It can be categorized into three levels-Document level, Sentence level and Aspect level. The Document level classifies the review into either positive or negative as it assumes that whole review is about the single entity. The assumption considered in this level is that it is being assumed that the whole document consists of an opinion about the same or single entity. For example-a review about a particular product whether it is a camera, phone etc.. In such types, opinion about the particular product is calculated in terms of either positive or negative as a whole .However, it will not be applying for a blog post because in such posts the opinion holder compares different products with each other.

The Sentence level describes the sentence into subjective or objective by using subjectivity analysis and then the opinion orientation of each sentence is found. This level is closely related to Subjectivity Analysis by(Tsytsarau et al., 2011). In this, the first task is subjectivity classification where a sentence is categorized into objective or subjective sentence. Objective sentences are those sentences which contain facts or no sentiments explicitly. On the other hand, subjective sentences are those which contain opinion or sentiments . However, it must be noted that the objective sentences may also contain opinions implicitly. For example- consider the sentence“The battery backup of the camera is very great”. This sentence has an opinion defined explicitly so it is a subjective sentence. Now consider “The screen broke in two days”. It appears that the sentence is objective but actually it provides an important opinion about the screen of the phone. Thus, subjectivity classification poses as a challenge task. The second task is the sentiment classification which aims at finding the opinion orientation of each sentence by classifying them into positive, negative or neutral sentiment.

The aspect level is quite popular among the researchers. It determines the most important features of an entity by extracting the aspects which describe the entity. The different approaches of opinion mining are Sentiment Classification, Subjectivity Analysis, Lexicon based, Statistical based, Dictionary based and Semantic based. For example-"This room is very large". This sentence has an aspect "room" of the entity "hotel". The orientation of the opinion is positive . An another example can be cited from the movie domain as" I didn't like the storyline yet the movie was scintillating". In this, although the orientation of "storyline" is" negative" but the "movie" has an orientation "positive". Feature level allows to classify different features into different polarities-mainly positive and negative.

The different approaches of opinion mining are Sentiment Classification, Subjectivity Analysis, Lexicon based, Statistical based, Dictionary based and Semantic based. Sentiment classification is the technique of classifying the polarity with supervised learning.

The assumption considered in this level is that it is being assumed that the whole document consists of an opinion about the same or single entity. For example-a review about a particular product whether it is a camera, phone etc.. In such types, opinion about the particular product is calculated in terms of either positive or negative as a whole .However, it will not be applying for a blog post because in such posts the opinion holder compares different products with each other.

Sentiment Analysis or opinion mining is the field of analyzing social media text using machine learning or NLP techniques. With the growth of Web 2.0, the data on social media platforms is increasing in huge amounts. This data is in the form of text and is highly unstructured. There comes the importance of big data. This huge volume of data, when analyzed, can be used by supply chain managers in forecasting the sales of the products and creating marketing strategies. It can be done by analyzing the social media postings of different customers about a product. The different pros and cons of the application of opinion mining in operation management is described in the following sections.

The Figure 1 describes the different advantages of applying opinion mining in operation management.

## Advantages of Opinion Mining in Operation Management

The advantages of the application of opinion mining in operation management are listed below:

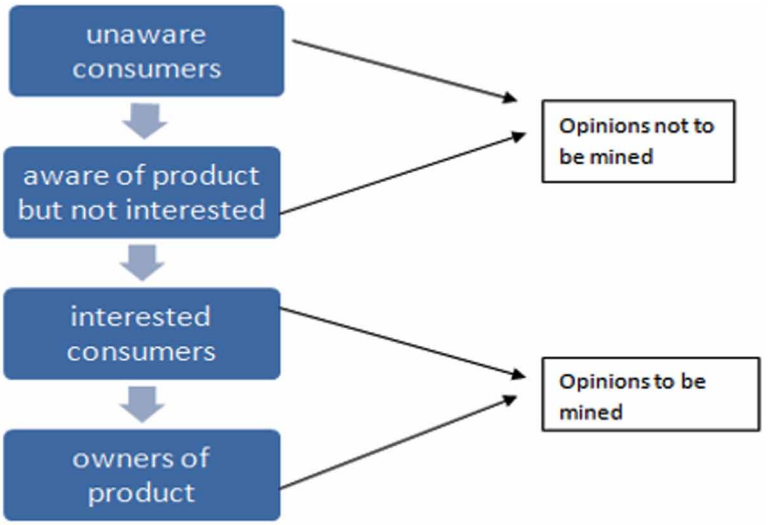
*Figure 1. Pros of opinion mining in operation management*


- **Forecast Actual Sales:**Research and validations show that the sales of an organization can be predicted by the analyzing the social media behavior. (Abbasi et al. 2012).It has been investigated that there is some sort of relationship between the products reviewed online, the release of the products and the actual sales of products . This relationship can provide an aid in predicting the sales of the products.(X. Yu, Liu, Huang, & An, 2012) .This prediction is important in the overall growth of the organization. The sales of the product is related to the overall revenue of the organization.
- **Identify Different Volumes of Customers**: The sentiment analysis, based on the underlying social media reviews about different entities can be used to provide an easy method for an organization to search different customers in each segment. Customers at different levels have unique things to say about the product. This pipeline is described with the fact that the customers who are interested and have potential may ask about the attribute of the product .The owners of the product can make comments about the product

either positive or negative. There are some 'unaware customers' who do not know anything about the product. There is another level of consumers who are aware of product but not interested. These type of customers ask general questions about the product. The Figure 2 shows the above concept in a pictorial manner.

- **Appreciation of consumer is evaluated while new initiatives are formed:** A consumer response to supply chain sustainability initiatives(SCSI) can help a firm to gain customer response using opinion mining(Wood 2014b). Firms can get the idea from the opinion mining whether the changes made in products lead to more revenue or not. If consumers are incurious about the new initiatives then organizations must not introduce new SCSI.
- **Profitable for Distant Firms**: Opinion Mining helps the distant organizations to take better decisions in understanding the consumers . It opens the avenues for these firms located far away from their customers to cater their need for better and quality products. Opinion mining combined with operation management can deal with different volumes of consumers in different streams. The comments of the consumers are elucidated to indicate the decisions that are needed to be taken by an organization.

*Figure 2. Pipeline of consumers*
*(Warren 2008)*



## Disadvantages of Opinion Mining in Operation Management

There are certain limitations of applying opinion mining in sentiment analysis which are illustrated below in Figure 3.

### Miscategorised Post Dealing with Sarcasm

A major limitation of opinion mining is the text to be dealt with sarcasm and irony. Not much attention is given to the posts that contains sarcasm as these types of posts generate a wrong value of sentiment. The incorrect value of the sentiment can lead to inaccurate information about the customers' sentiments of the products in an organization. This misleading and miscategorised post can hamper the growth by decreasing the sales and revenue of a firm.

### Fake Reviews, Leading to Opinion Frauds

In the recent times, it is being observed that a lot of fraud reviews are being created by fraudsters to distort the quality of a product. Such fake reviews can lead to the exponential decrease in the revenue of a firm if all the decisions of identifying the customers' sentiments are wholly made using the sentiment analysis. Moreover, identification of such fraudulent reviews is an another tedious task which is still not been automated.

*Figure 3. Cons of sentiment analysis in operation management*

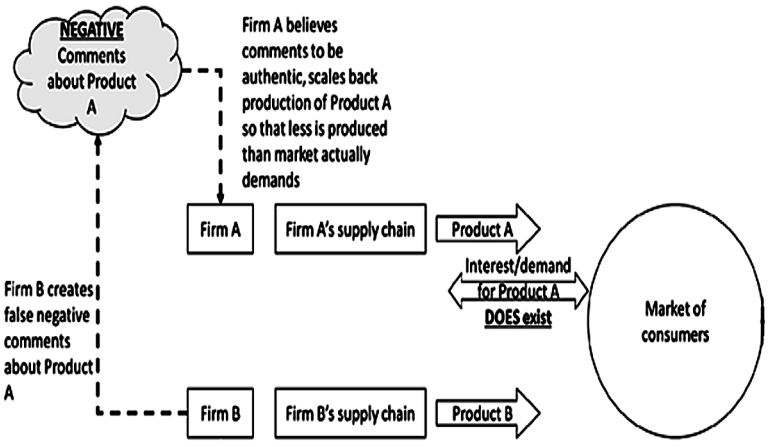
## Shortage of Skilled Staff of NLP

The task of sentiment analysis covers machine learning and NLP techniques. Both these techniques are current and new which has not been catered by number of people in the recent time. Due to the shortage of skilled staff in this area, automation of social media posts is difficult in its own terms. A large number of organizations won't be applying sentiment analysis in operation management due to the added overhead of training people in the new field which would add to the overall cost of a firm.

## Threats from Illegitimate Data

A new term 'opinion warfare' is introduced in which firms create fraudulent reviews to mislead their competent firms. As in Figure 4, consider two firms-Firm A and Firm B competing in the market. It is being assumed that the Firm A takes its operation management decisions based on the sentiment analysis. Suppose Firm B gets to know about this and it deliberately create negative reviews about the product A of Firm A. Due to the negative comments about the product A, the Firm A will decrease the procurement of the Product A which will ultimately lead to the fall in the sales of the Firm A for product A. This type of situation can create a lot of confusion for the organizations and can lead to bad decision making. Detecting of such misleading posts can be a bit tricky as automated tools are required to identify such malicious attacks.

*Figure 4. Firm B creates a false-positive outcome that misdirects the efforts of Firm A (Wood et. al 2013a)*



## SENTIMENT ANALYSIS IN OPERATION MANAGEMENT USING BIG DATA

Organizations are very keen in knowing the popularity of the brand. They use sentiment analysis for this purpose to analyze the information in an organization. As the information is quite large in number, analyzing with the conventional relational database management systems is quite cumbersome and slow. Therefore, Big data comes into play . It is a cost effective solution for analyzing this high volume data. Hadoop is a popular tool for implementing the Big data.It is an open source tool which is available in different versions. The first step in analyzing the firm information is to collect the data from the different available sources. The second step is to load the data in hadoop using tools like Apache Flume. After the data is loaded in the hadoop, data transformation is performed. In this step, Mapreduce programs are used for this purpose. Mapper and reducer function are written in different programming languages like Java, Python, .Net, Ruby. Mapreduce transforms (map) it to useful information, and then combine (reduce) for reporting. The final step is to load it into the Business Intelligence(BI) platform and analyze this meaningful data by different tools like R programming language.

The features that must be looked upon in big data are-heterogeneity and noise accumulation. Big data is termed as heterogeneous since data is accumulated from a large number of sources. A lot of unstructured and heterogeneous data are available in the form of big data. Predictive models can be estimated based on several parameters. Predictive analysis with big data perspective can also be applied to a large number of applications. It involves predicting the future based on current or historic data when the volume of the data is huge. Big Data analytics can also be applied to predictive analysis in social media by different ways. This big data is a powerful trend in identifying trends from the customers' sentiments. Social media channels associated with the different suppliers and markets are analyzed to take accurate decisions for the organization. It can help in updating the market risk profiles in case of any natural disasters.

## CONCLUSION AND FUTURE RESEARCH

This chapter widely shows the pros and cons of applying opinion mining in operation management from the big data perspective. A lot of advantages like actual forecasting of sales, developing customers through marketing, identifying different volumes of customers, distant firms profited and consumer appreciation in new initiatives is

being described. Some disadvantages also show that opinion mining in operation management may lead to added difficulties for the organization. It includes Fake reviews from competitors, Threats from the illegitimate data and miscategorised post dealing with sarcasm. We have tried to explain the core concepts of opinion mining, their different approaches and types. We have also touched the concept of big data analytics in different fields.

The applications of opinion mining must be empirically validated in the future to show the real existence. Opinion frauds and fraudulent must be identified by the automated tools to prevent any mishappening in the future.

## REFERENCES

Bai, X. (2012). Predicting consumer sentiments from online text. *Decision Support Systems*, *50*(4), 732–742. doi:10.1016/j.dss.2010.08.024

Ding, X., Liu, B., & Philip, S. Y. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08)*. doi:10.1145/1341531.1341561

Ding, X., Liu, B., & Yu, P. (2008). A holistic lexicon approach to opinion mining. In *Proceedings of the Internationl Conference on Web Search and Web Data Mining*.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*.

Lincoln, C., Reiners, T., Srivastava, H. S., & Linh, D. N. K. (2013). *Using sentiment analysis to improve decisions in demand-driven supply chains*. Anzam.

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, *41*(17), 7764–7775. doi:10.1016/j.eswa.2014.05.045

Wang, X., Petkova, B., & Wood, L. C. (2014). Do suppliers benefit from supply chain sustainability programs? The case of Wal-Mart. In *Proceedings of the 12th ANZAM Operations, Supply Chain and Services Management Symposium*. The University of Auckland Business School.

Warren, K. (2008). *Strategic management dynamics*. Hoboken, NJ: Wiley.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2013). *Expanding sales and operations planning using sentiment analysis: Demand and sales clarity from social media*. Anzam.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2014). *Sentiment analysis in supply chain management. In Encyclopedia of Business Analytics and Optimization*. Hershey, PA: IGI Global.

Woodward, M. (2012). 5 Supply chain predictions for 2013. *Supply Chain Management Review*.

Yu, S., & Kak, S. (2012). *A survey of prediction using social media*. arXiv:1203.1647

Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720–734. doi:10.1109/TKDE.2010.269

Zitnik, M. (2012). Using sentiment analysis to improve business operations. *XRDS: Crossroads*, *18*(4), 42–43. doi:10.1145/2173637.2173655

## ADDITIONAL READING

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. doi:10.1109/MIS.2013.30

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *Management Information Systems Quarterly*, *36*(4), 1165–1188.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *Management Information Systems Quarterly*, *36*(4), 1165–1188.

Cunha, J., Silva, C., & Antunes, M. (2015). Health Twitter Big Bata Management with Hadoop Framework. *Procedia Computer Science*, *64*, 425–431. doi:10.1016/j.procs.2015.08.536

Lohr, S. (2012). The age of big data. *New York Times, 11*.

## KEY TERMS AND DEFINITIONS

**Machine-Learning:** It is a field of computer science that deals with the learning theory in artificial intelligence and construction of algorithms for making predictions on data.

**Natural Language Processing:** It is a field of computer science that deals with the interaction between human language and computer.

**Operation Management:** It deals with the process of production and business operations.

**Opinion Mining:** It is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions of different entities.

**Sentiment Classification:** It is the technique of classifying the polarity with supervised learning. Example-Naïve Bayes classification algorithm, Support Vector Machines etc.

**Subjectivity Analysis:** It is the technique of classifying the sentence into subjective and objective sentences.

# Chapter 6
# A Conceptual Framework for Educational System Operation Management Synchronous with Big Data Approach

**Ganeshayya Ishwarayya Shidaganti**
*M. S. Ramaiah Institute of Technology, India*

**Prakash S**
*Dayanad Sagar University, India*

## ABSTRACT

*The philosophical interpretation of operation management in the process of delivering a 360 degree digitized educational system requires a core approach of supervision, scheming and finally controlling the mechanism of delivery the educational service and its constitutes operations. In today's world of telecommunication, internetworking, cloud infrastructure and computational advancement, a collaborative mechanism of intelligent and reachable educational system is being conceptualized which provoke a notion of "one world one university". The chapter aim to illustrate different source and applications of BigData Generation in educational system operations, existing mechanism to store and process these generated data, limitations of existing mechanisms and conceptualization of a novel framework. Further it aims to discuss different intelligent analytics approach to improvise the teaching-learning philosophy.*

DOI: 10.4018/978-1-5225-0886-1.ch006

## INTRODUCTION

The educational systems at the modern times have witnessed higher dimensionality of the transformation as compared to the conventional educational system that existed 15 years back. The modern day education system is highly influenced by cut-edge technologies in order to furnish a quality oriented knowledge sharing process. Be it a school or college or any advanced higher educational institutions, the impact of ICT can be seen worldwide. Right from primary levels in schools to peak level post graduation, all the actors involved in educational system are regularly using World Wide Web along with various sophisticated and interactive educational-related tools to gain more access to knowledge. Hence, in order to understand such positive transformation, it is highly essential that operational management pertaining to educational system be studied very closely. Operational management relates to process planning, formulating designs, and managing all the critical acts responsible for delivering the products or services (Panneerselvam, 2012). The actors in the operational management of educational institutions are engaged in designing various forms of policies (administrative policies, recruitment policies, courseware design policies, student assessment policies, faculty's skill development policies facilities maintenance and upgradation policies, ICT maintenance policies, etc.) (Harrison et al., 2012; Brown, 2014) (UNESCO, 2014). The functions of operations may again include various essential activities e.g.

- **Prediction:** The existing operational management supported by ICT can be utilized to predict the performance of the institution or even to understand success rate of various strategies for enhancing the quality of education.
- **Effective Planning of Capacity (Intake):** An educational institution can review the past data and can make a decision about the increasing or decreasing an intake of students.
- **Data Storage:** Various forms of information in daily operations (administrative data, student's information, or data pertaining to faculties) in the campus will be required to be stored in either in local server (physical) or in cloud (virtual).
- **Quality Assurance:** Reviewing the operations and analyzing its success rate can be highly useful in maintaining the educational quality.

This chapter discusses about the significance of operational management in education system and discusses the adoption of ICT that significantly generates a Big Data. The chapter also discusses about the existing techniques from the research-based literatures of operational management to process such complex educational data and finally it proposes a conceptualized framework.

## OPERATIONAL MANAGEMENT IN EDUCATIONAL SYSTEM

At present, the educational system in majority of the country have started adopted various web-based application that supports majority of the client-based system. Applications e.g. Entrar, iCampus, collegeSAAS, etc. Such application acts as a bridge between the teaching communities with the learner's community most effectively. They are not only used for updating attendance or act like mere e-notice board, but they allow authenticated access of many educational resources, ranking the faculty performance from students and vice-versa, etc. There are also availability of various mobile applications which uses tablet computers and smartphones for providing educational services. Hence, there is a generation of massive volume of data over months and months.

At present, there is a big proliferation of k-12 classroom by various types of worldwide learners. A K-12 classroom grades their students from the public before even joining the actual educational establishment in reality. Supported by the potential of cloud computing, it is most widely practiced in Canada, United States of America, and various other nations. Interestingly, adoption of k-12 classes significantly saves a lot of expenditure, space, and time of learners, parents, and instructors. The benefits of the existing operational management in the educational system are as follows:

- **Data / Service Availability:** With the use of cloud-based services, the existing system renders better data / service availability.
    - **Benefits:** Owing to the adoption of cloud computing, the existing operational tools allows the availability of the data as well as various resources to its users (learners, parents, and instructors). It acts as a terminal where all sorts of data requirement can be met. Although, such facility is a great boon for users at present.
    - **Limitation:** It equally posses a challenges for the same user in future. Educational data keeps on collecting in one storage system, but it is of no value if it doesn't provide any value added services or knowledge.
- **Efficient Communication:** With the availability of online classes e.g. Eureka, Coursera, Edline etc., there is an efficient communication system with existing operational tools.
    - **Benefits:** Sitting at comfort of house or along with existing commitment, it becomes possible to get knowledge along with certification for the learners, which is highly productive and develops good level of competence.

  - **Limitations:** Such online classes uses video streams, which can be live or offline. Offline streaming is not a problem, however, online streaming cannot ensure better Quality-of-Service for the user with low-configuration resources.
- **Reliable Data Recovery:** Recovery of data from the conventional servers are sometimes not possible or is highly expensive in nature. However, existing operational tools used in educational system ensure proper recovery of data during system / server crash condition.
  - **Benefits:** With the adoption of distributed clusters and virtual machines, the replications of the data over distributed environment ensures that even in crash condition from one distributed node, service is instantly resumed from another functional node.
  - **Limitation:** Owing to data replication, the issue of data integrity surfaces. Another biggest limitation is accessibility of the data (in encrypted form) by the service providers. It poses a slight threat to the genuine users. Hence, there is a greater deal of hesitation for adopting cloud based system although it claims of high fault tolerance.

Therefore, the operational management system practiced for the existing educational system allows a good level of participation from maximum numbers of users. The advantage of existing operational management system is its potential to provide data availability, a giant storage system and its limiting factors will be its less applicability of heterogeneous data. Educational data are highly heterogeneous in nature which may be either unstructured or structured in nature. Owing to such data charecteristics problems, various data analytical theory and applications are not practical. The upcoming sections will give more clear visualization of the problems.

## ADOPTION OF ICT OPERATIONS IN EDUCATIONAL SYSTEM

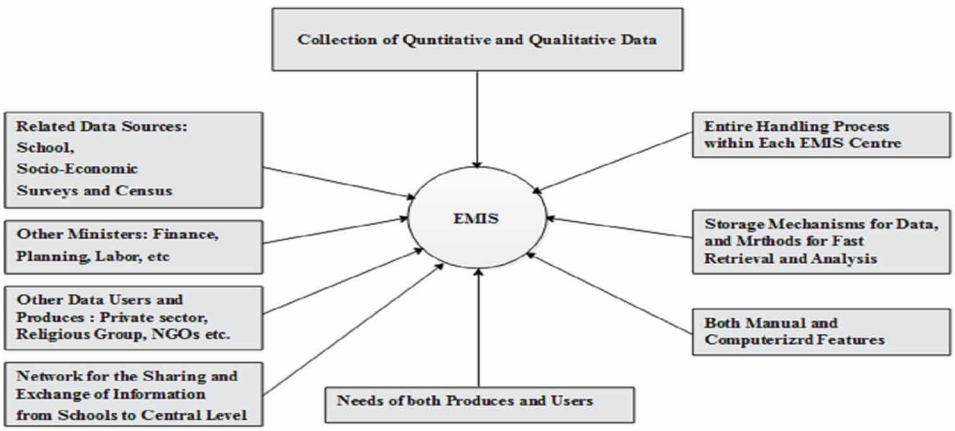
From the modern educational viewpoint, there are two significant forms of ICT operations management being standardized for educational system e.g. Educational Management Information System (EMIS) and Learning Management System (LMS) (Nolan, 1997; Kats, 2013).

## Operations Based on Educational Management Information System

Basically, EMIS is a significant part of operational management that collaborates various process, people, and advance computing process for furnishing precise information required to upgrade an effective operational management for educational system. There are two more contrast of EMIS operational technique e.g. Decision Support Information System (DSIS) (Aqdam & Bonab, 2014) and Statistical Information System (SIS) (Ellison, 2004). DSIS is responsible for carrying out data analysis of the stored data of operational management for any specific institution. However, SIS performs low-level statistical analysis on heuristic data. The standard components involved in the EMIS from the UNESCO report (2014) is shown in Figure 1.

The tools used for operating EMIS are cellular phones, tablet computers, Google Earth, Business Intelligence, Open Source Softwares, and internet (Bernbaum & Moses, 2011). Unfortunately, EMIS is still a hypothetical standard and is highly centralized in nature. The existing service delivery for educational resource require an operational process to be highly decentralized in nature and it is highly dependent on EMIS data to be i) highly reliable, ii) timely, iii) accurate, and iv) understandable. Majority of the existing techniques of operational management compromises the above four points as – Reliability factor is highly influenced by human resources and funding. Time factor can be tradeoff with reliability and accuracy. Moreover with the evolution of the various technologies e.g. clouds computing for decentralize data storage and optical network for high speed data rate, the operational management system for EMIS will need still more years to reformulate itself.

*Figure 1. Essential components of EMIS*

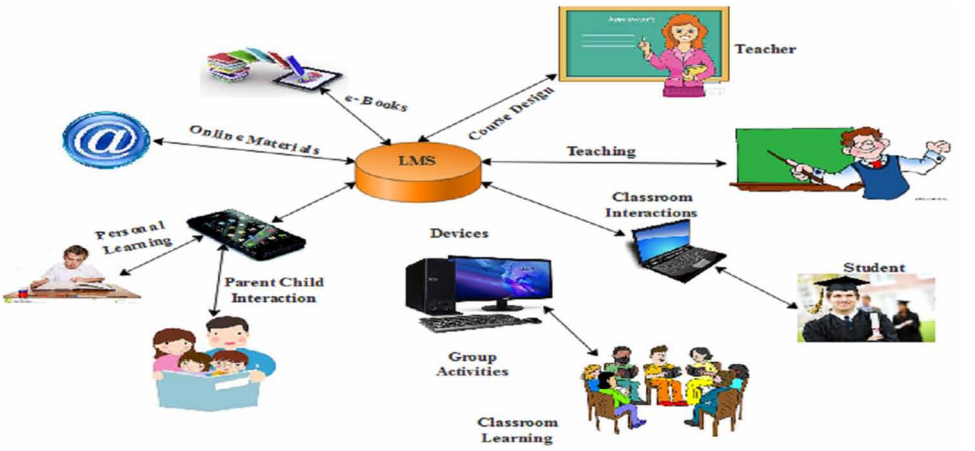


## Operations Based on Learning Management System

LMS or Learning Management System can be said to be set of software application that assists in the service delivery process of educational system. The significant operations involved in the process of LMS are i) maintaining archives of documentations, ii) tracking the contents, iii) monitoring the administration, and iv) knowledge transfer using online learning tools. This is one of the significant standards of imparting education that manages all the operational aspects of the learning (Alam et al., 2010). Critical operations for the course registration process are handled in this type of operational management along with analysis of skill gap, progress monitoring system, as well as comprehensive reporting system. Majority of the existing operational tools used by the LSM system are web-based for better access to the educational resources as well as effective administration. Adoption of LMS services mostly sought by regulated industries, self-service of student, scheduling e-learning system. It is also used currently for collaborative learning system along with operational management of structuring essential training resources (see Figure 2).

The operational management involved in use of LSM calls for i) designing an interactive website where teachers can communicate with students online and ii) effective database management. However, things have changed in past 5-6 years, where the area of LMS has witnessed the evolution of MOOC (Massive Open Online Courses). It is basically an online course where there is no conventional registration process based on eligibility criteria of learners just like what exists in conventional educational establishments. It is also an open access and supports enrollment of infinite number of participants. MOOC have drastically changed the

*Figure 2. Essential components of LMS*



definition of online learning in the present day. The year 2012 have witnessed various MOOC providers e.g. Coursera, edX, Udacity, etc. Table 1 highlights the lists of essential MOOC providers. Conventional LMS system works on video-based learning materials still now, but MOOC uses various forms of pedagogy in the form of different study materials for enriching the educational quality. Some of the ground-breaking advantages of adoption of MOOC is that it provides secure and reliable assessment with globally acclaimed certification, which the learners can gain along with their present commitments.

## COMPLEXITIES IN EXISTING EDUCATIONAL OPERATIONAL MANAGEMENT

The prior section has discussed about the existing operational management technique adopted in LMS resulting in generation of massive transactional data of various forms e.g. word, PDF, PowerPoint, image, audio, video, and various other types of files formats. However, such form of educational and knowledge sharing process often adopts the operations of collaborative learning more than the conventional LMS. Collaborative learning process consists of participation of a group of student for a same topic. The system is however more interactive and performs a massive exchange of data. It should be understood that data exchanged in this process are archived in the core web server of the operational system and hence, it is essential to understand the sorts of complexities framing up the small and simple data to large and complex data owing to existing operational management in education system.

*Table 1. List of non-profit and commercial MOOC providers*

| Types | MOOC Providers | Institutional Participants |
|---|---|---|
| Non-Profit | Stanford Online | Stanford University |
| | Khan Academy | |
| | edX | Havard University, MIT, University of Queensland, IIT Bombay, UC Berkeley, Australian National University etc |
| | MOOEC | Griffith University, University of Queensland, |
| Commercial | WizIQ | IIT Delhi |
| | Canvas Network | University of Utah, Santa Clara University |
| | Udacity | Cloudera, Cadence, Nvidia, San Jose State University, Autodesk |
| | Coursera | University of Maryland, University of Tokyo, Stanford University. |
| | NovoEd | Stanford University, Princeton, Carnegie Foundation. |

This section will mainly discuss about the complexities that arises from the existing operational management techniques.

## Complexities in Data Management

An effective data management calls for clear visualization of the types of data. Various RDBMS process already exists to manage such generated data, but in the most recent days, the complexities of the data being generated are increasing. Imagine of any educational section is maintaining the digital records of their daily activities; there are various forms of heterogeneous data being massively generated and stored in either local server or in cloud. However, the extent of data being generated from the LMS (collaborative learning) is exponentially high as the streams of data are generated round the clock. Owing to adoption of cloud resources by existing online learning tools (e.g. Cloudera). However, there are specifically three types of data being found to be generated from educational sector as structured data, unstructured data, and semi-structured data.

- **Structured Data:** Structured data are those data that can be stored in rows and columns of a table using SQL. Structural data is characterized with relational key and quite adaptable for mapping with pre-designed fields. It should be noted that 5-10% of the structured data can only represent all informatics data.
- **Unstructured Data:** Unstructured data mainly consists of multimedia and text-related information and constitutes about 80% of the overall informatics data. Majority of the educational data are unstructured in nature as they consists of audio, video, images, word format documents, email message, PowerPoint, etc. Although such forms of data too posses their own structure but they can never be called as structured as their data cannot suitably fit in the existing database neatly.
- **Semi-Structured Data:** Semi-structured data posses the interesting charecteristics. Although, such forms of data cannot be resided in a relational database management system, but they have some unique features of organization that makes them more suitable to perform data analysis. Example of semi-structured data is text in the forms of XML, JSON, etc doesn't have formal structure but it possess specific levels of tags to assist in data management. Another good example of semi-structured data is weblog data, healthcare data, supply chain data, and sensor data.

Both unstructured and Semi-Structured data could be generated by machine or human. Apart from educational system, unstructured data is also found generated from machines e.g. satellite images, surveillance feed, scientific data, and sonar/radar data. Similarly, generation of unstructured data from human are data generated from social networking sites, website content, mobile data, internal communication system (corporate emails, logs etc). Hence, the complexities surfaced from existing operation function in educational data management are generation of unstructured and semi-structured data, where the corporate spends a lot for storage but doesn't emphasize on its utility. At present, Hadoop and MapReduce are considered as standards for managing such forms of unstructured and semi-structured data. Hadoop and MapReduce are found compliant of the standard termed as Unstructured Information Management Architecture.

## EXISTING OPERATIONAL TECHNIQUES TO STORE/PROCESS BIGDATA

The role of cloud computing towards educational excellence has been discussed by various prominent authors e.g. (Wang & Xing, 2011; Sqalli et al., 2012), most recently. Various prominent educational institutions around the world have started with adoption for public cloud and at present some of them are also moving to the private cloud. 2-3 years back the educational data uploaded on clouds were usually course materials or some other resources with less complex problems pertaining to type of data. But at present, even the educational data have become much complex and posses all the 5V characteristics of Big Data (Holley et al., 2014). The high usage of ETL scripts as well as presence of numerous layers of integrations does not posses common semantics that increases the problems of data integration to manifold. Usages of ETL scripts are only found theoretically sound but in reality it is an expensive affair. There are various literatures in the existing system where researchers have identified the potential advantage of using advanced datamining technique over educational data. The study conducted by Ali (2013) have discussed that datamining techniques can be used for prediction of enrollment for prospect student, forecasting profiling of students, developing curriculum, performance of students, and many more. Study of data mining pertaining to educational data from online learning tool was seen in the work of Fernandez et al. (2014) and Thakar et al. (2015). The author have discussed about the challenges and benefits of online learning in cloud environment. Papamitsiou and Economides (2014) have adopted qualitative method to study the effectiveness of prior educational datamining tech-

nique. Another researcher Ramasubramanium et al. (2009) have presented a model for datamining using set theory considering educational data. The authors used SQL and normal query language.

BigData is the advancement of conventional datamining technique. The author e.g. Daniel (2014) have investigated the impact of BigData for analyzing educational data. The study found that there are positive opportunity of applying BigData approach over i) administrator (like allocation of resources, support system, academic programming), ii) student (planning of learning, feedback system, assessment system), and lectures (enhance teaching, assist student in academic distress). Similar direction of review-based study was also carried out by Drigas and Leliopoulos (2014) and Prakash et al. (2014). The study conducted by Johnson (2014) have discussed about the standard ethics that should be considered while analyzing datamining process over educational big data.

MacHardy and Pardos (2015) have adopted Bayesian Knowledge Tracing for extracting the knowledge from educational video contents. Bayesian Knowledge Tracing is basically an approach of assessing effectiveness of knowledge mastered by instructor. The study outcome was assessed using data from various online videos with respect to mean Root Mean Square Error and significance in probability. The presented model was used for understanding behaviour of student.

In order to design an efficient operational management system for storing and processing educational BigData, it is highly essential that both the semi-structured as well as unstructured data from multiple formats and sources be emphasized. As the operational management of conventional education system highly differs from the existing educational system, therefore, development and management of any novel BigData projects toward ensuring an optimal management will required highly professional and skilled developers. It is highly important that system for operational techniques for educational data management should be user-friendly, which is not the case in the existing system (Zhao, 2015). Adoption of Big Data approach on educational system can highly enrich the quality of knowledge dissemination process and furnish a competitive learning culture. One of the core motives of any higher technical (or non-technical) education system is to use the multiple heterogeneous disciplinary knowledge to enhance the educational contents and knowledge sharing system that surfaces a pleasant environment incorporating higher dimensionality of competence and analytical skills among the learners. There are various forms of underlying and hidden patterns which are quite difficult to be explored in complex concept e.g. education; however, adoption of BigData can bring a better form of integration and associate various forms of conventional and novel data sources from multiple source points of data.

## About BigData Storage

The phenomenon of storage of big data is very different from the conventional data storage system owing to its logistical and technical problems (Vaitheeswaran & Arockiam, 2015). There is an increasing interest among the organizations for exploiting the resources of transformed big data from their data storage system. The storage location of big data is not only a location to reposit the data but it strongly represents it as a platform of data. Hence, this section will showcase an essential factor that characterizes the storage system of BigData.

- **Potential to Truncate Migration of Data:** The adoption of data migration is very much large and high in the present system owing to the need of migrating the present data from existing storage system to the upcoming storage system. The process of migration is expensive as the organization will be required to spend money on both existing system and futuristic system of storage as well as it can be also time consuming. Hence, BigData storage system is highly anticipated to possess the characteristics of eliminating data migration problems.
- **Scalable Capacity:** The need of data capacity cannot be predicted by any organization as business requirements changes with evolution of customer's demands. Hence, a BigData storage system should be scalable to extend its capacity in most cost effective manner in case of increasing size of the data suddenly. BigData approach controls cost of data storage by accomplishing the scale by extracting storage devices and commodity servers of industry standards. However, the task is not that simple, as the operation of scaling up of the capacity should take place without affecting the performance of applications that uses existing storage locations in BigData.
- **Pervasiveness of Data:** Design of BigData storage system must ensure the data accessibility worldwide. It is highly essential that such BigData storage must support distributed data storage system as BigData is usually a bigger stream of massive volume of data with 5V problem.
- **Supportability of Legacy Silos:** In order to maintain higher accessibility and availability of the critical data (defined by urgent requirements of access by clients anytime), the present organizations are creating a novel storage instances for meeting the storage requirement of their growing data. Therefore, an efficient storage of BigData should have better accessibility without any dependency on adhoc interventions. BigData storage should have higher degree of supportability for legacy silos of storage.

Hence, from the above discussion, it can be seen that BigData storage system posses some charecteristics which differs them from the conventional data storage system. The conventional data storage system posses the static capacity but BigData storage system posses the capability to increase the capacity in most cost effective manner without any possibility of negative impact performance of data storage in existing storage location. The practice of adopting HyperScale Computing Environment is seen to be used by Facebook, Google, and Apple etc. These organization uses large number of commodity servers with storage system. In case of condition of service outage, the servers are instantly replaced by mirrors. At present, operational tools like Hadoop, MapReduce, Cassandra, NoSQL etc are used a software framework for BigData storage system.

## Benefits of BigData Storage

The learning as well as teaching methodology of the existing educational system has completely revolutionized. The educational system of almost every part of the work is now connected with usage of ICT as the operational tools. Hence, the operational management of present educational era calls for higher degree of supportability of increasing mobility of student and adoption to changing mode of educational delivery process. The present situation of education system also has higher adoption of Massive Open Online Courses, where large number of learners and instructors collaborate to redefine the educational system. Practice of such platforms will result in increase of a giant volume of users online and generation of double of that volume the size of contents being exchange by the users. Hence, BigData storage system can be adopted by the online user of MOOC community in order to revolutionize the quality of educational delivery as well as enhancing the teaching-learning experiences. The concept of pervasiveness in BigData storage will allows the data and service availability for both students and instructors. It is also possible for BigData storage system to encounter the need of data migration from MOOC; however, it will be eliminated if it is using BigData approach for storing and processing the data. The charecteristics of personalization will allow the user to enhance the study materials and mode of knowledge delivery system.

## Obstructions towards BigData Usage

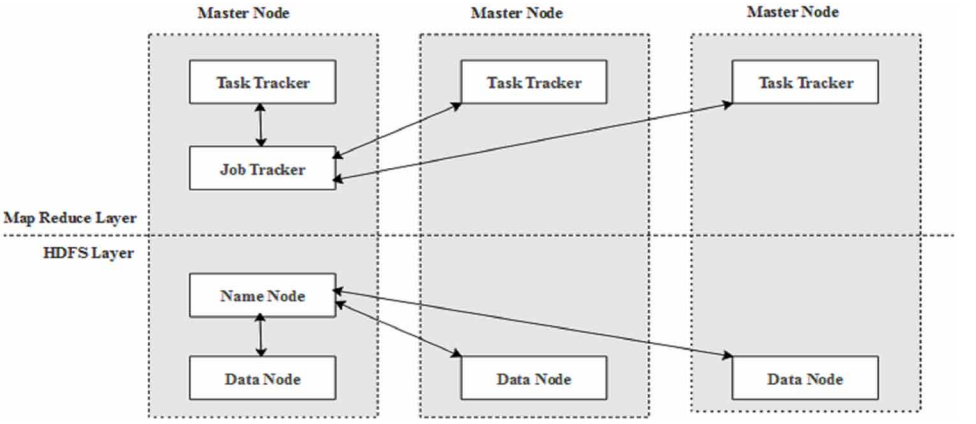
Although BigData approach has some of the potential advantages, but adoption of BigData is just a beginning. One of the biggest factors of impediment to use BigData approach is the security problems associated with it. Normally, the data used the educational institutions have some of the private information that have a very restricted access. However, certain challenges are encountered by the service

provider when the informed consent from the educational establishment is not received properly owing to the massive size of the data. Hence, it is great threat to the privacy as protocols cannot be set properly discretizing the accessibility of the data according to the level of privacy. The soft computing approaches that are in practice uses the conventional security techniques that doesn't even have the capabilities to identify the upcoming Trojans. Moreover the amount of awareness for the usage of BigData is very less. The amount of the transformation in the services in the BigData is quite faster as well as highly reliable that enables the system to formulate the policy implications. Dependency of the highly skilled software developer, lack of quality data, lack of common language across the platform are and lack of user friendly interface are also some of the critical obstructive factors for adoption of BigData. The system will also limit the availability of the resources to the learners owing to the complex learning technology. At present, the IBM data analytics can only process the data originated in last 2 years and existing BigData approaches limits analyzing historical data.

## Existing Operational Tools in Educational BigData

Basically Hadoop can be said to be a software framework that is used for repositing the massive stream of data over cloud and is responsible for running the BigData applications on the distributed clusters. Hadoop is much seeked software for the researchers and beginners as it has excellent capability for processing and furnishes storage of any forms of complex and heterogeneous data. The fundamental architecture of Hadoop can be seen in Figure 3.

*Figure 3. Fundamental architecture of Hadoop*



The processing part in Hadoop is carried out by MapReduce that is responsible for generation of large datasets on distributed clusters. MapReduce consists of two fundamental procedure called as Map and Reduce. The first one is used for filtering followed by sorting. The processing supported for Hadoop is accompanied by master and slave node and it is the layered on top of HDFS (Hadoop Distributed File System) layer. HDFS is designed on Java that reposits the data on multiple computing terminals. YARN is another important component of the Hadoop that is responsible for scheduling and handling resources from distributed environment. Other applications that uses Hadoop framework are Hive, Pig, Hbase, HCatalog, Cassandra, Ambari, Flume, Chukwa, Oozie, Sqoop, Solr, Zookeeper, etc. At present, Hadoop is one of the most frequently adopted operational tool for educational BigData as it has the capability of repositing and processing massive amount of streaming (online video classes) of big data in much faster rate. Some of the other essential charecteristics of Hadoop are:

- **Better Flexibility:** Hadoop doesn't request the massive stream of data to first store and then process it that happens in the existing system. It allows the applications to store massive amount of data and process it. Hadoop also offers flexibility to process unstructured data.
- **Maximized Computing Power:** Hadoop can offer computing capabilities for processing massive size of data. Adoption of more nodes with computational capability will offer more processing capabilities.
- **Fault Tolerance:** Owing to the master and slave approach in high level architecture of Hadoop, the work never stops or fails on any node. In case of node-disfunctionality, the redirection occurs from older node to other functional node and hence avoids the service outage.
- **Scalability:** Hadoop allows the system to enhance by appending more number of computing nodes with minimal maintenance on administration.
- **Cost-Effectiveness:** Hadoop is based on Java and thereby it has higher supportability to the open source applications.

## SIGNIFICANT ISSUES IN EXISTING OPERATIONAL TECHNIQUES

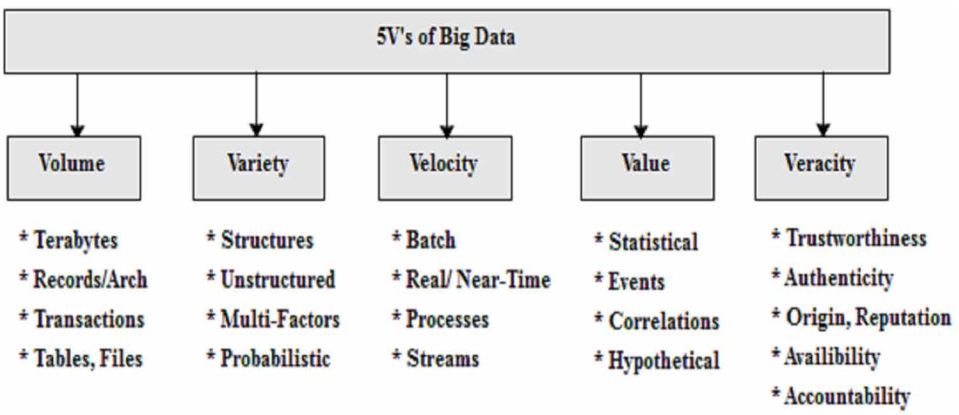
A clear-cut definition of the BigData is quite scarce to be found in any standard literatures. By the name "*Big*" and "*Data*" associated with the term BigData, it gives us an impression of massiveness of the data size. However, size is not the only one factor that has imposed an impediment towards processing BigData. First of all, BigData doesn't directly represent massiveness of data, but it represents a

technique that should be adopted when the data size is growing abnormally large and highly difficult to be stored in conventional RDBMS source or managed using conventional data mining tool (Sun & Han, 2012).

Some of the core problems that posses as an impediment to implement a successful operational management system for massively growing data are as follows (refer Figure 4):

1. **Data Volume:** The size of the data has increased from gigabytes to petabytes and zettabytes at present. Various organizational applications (e.g. hospitals, e-educational system, e-governance) are generating more than terabytes of data almost every day, which are quite hard to be stored on conventional RDBMS system. The size of such data is growing in every 1 second.
2. **Data Variety:** Owing to different nature of the data origination point, it is quite general that there are generation of multiple formats and types of data. Owing to multiple formats, it is sometime highly difficult to even access the data structure using conventional RDBMS tools. Hence, it significant affect realizing multiple dimensional data (structured, unstructured, and semi-structured).
3. **Data Velocity:** Data is also generated at a very high speed. Normally, the data generated from the sensor networks, financial stock exchange points, and social networks keeps on generating about massive terabytes of information within a fraction of seconds on the similar internet which also gives services to trillions of online customer at same time. Hence, speed of incoming massive data really poses a bigger threat to the proper storage system.
4. **Data Veracity:** Owing to critical issues of data volume, variety, and velocity, the extent of uncertainty of the data is quite high too. Users are not sure about

*Figure 4. 5Vs of Big Data*



the preciseness and amount of the non-redundant data in their availability. A fraction level of existence of data uncertainty will lead to incorrect computation or further data processing

5. **Data Value:** Basically, the problem of data value is directly linked with a reliable process of knowledge extraction. With the emergence of cloud storage and all sorts of advance storage system to process the conversion problems of structuredness of data, storage will be never a big issue for such massively growing data. But, even the storage of data is quite expensive one and return-of-investment can only happen when certain datamining process can be applied over such data. Unfortunately, due to the problem of unstructured and semi-structured data, existing mining techniques cannot be directly applied on massively growing data posing a critical impediment in operational management.

The identified problems of the proposed study are as follows:

- The multi-dimensional problems of the big data results in data heterogeneity problem that represents the extents of different forms of data and its associated meta-data information.
- Such data results in maximum pre-processing time.
- Cost maximum resources to pre-process it.
- Doesn't ensure data reliability.
- The uncertainty nature of the massively growing data leads to occurrence of false positives / errors in analysis of data.
- Applicability of existing datamining approaches are less in big data over cloud and hence there is need of novel data mining technique to extract knowledge.

The problems with the existing operational management of educational BigData are many. Although usage of operational tools e.g. Hadoop and MapReduce make the storage and processing efficient, but still it is accompanies by significant problems. It is already known that majority of the research work done most recently e.g. (Sivaraman, 2014; Seay, 2015) have implemented Hadoop considering the educational data. However, all such research works posse's limitations. Hadoop has two main file systems e.g. HDFS and GFS that are developed exclusively for computer racks over the datacenters. The localization of the original block and replica of the files are done by Hadoop over the clusters and rack servers, which doesn't go well with loosely coupled distributed environments. This problem will render the non-applicability of implementing datamining problems too. Hence, there is a need of a novel framework that can be adopted for mining the data from the cloud. It is extremely essential that a novel design should address such problem and furnish a

long-term and persistent educational BigData storage. Another significant issues in usage of Hadoop is that it uses association rule mining along with apriori approach. Such approach very often uses Boolean association rule for frequent itemsets. Unfortunately, adoption of such technique can generate large number of opinions thereby minimizing the efficiency of the association rule mining. This problem will also lead to minimization of validity for association rule mining too. Usage of apriori algorithms can be also used to explore more number of frequent item sets but it consumes a maximum time and well as memory while reading the rows and columns of database. This fact significantly affects parallelization.

## CONCEPTUALIZED FRAMEWORK OF NOVEL EDUCATIONAL OPERATIONAL MANAGEMENT

As discussed in the prior sections that education is one of the most complex domains to extract the patterns and therefore evolving up with a novel idea to harness the power of educational BigData. The prime goal of the proposed study is to design a framework that can perform effective search and extraction of precise knowledge of massive streams of repositories in the form of BigData in cloud storage. Following are the research objectives to accomplish the above mentioned research aim:

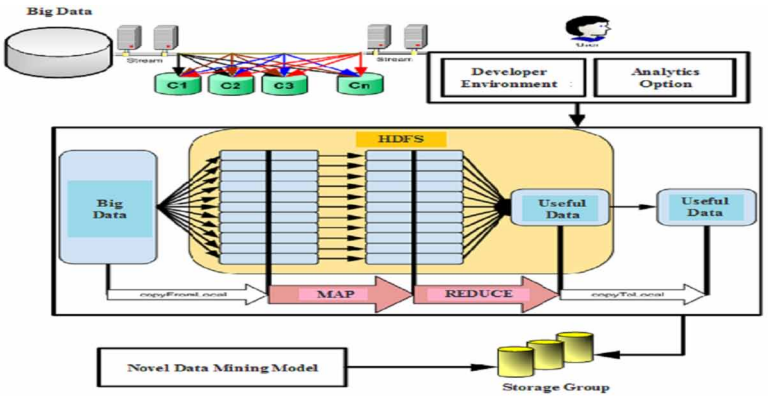
- To perform an extensive and in-depth investigation of the previous datamining techniques on big data in general as well as in cloud scenario and extracts the research gap along with open issues.
- To develop modelling of distributed clusters for BigData storage in elastic cloud framework.
- To design a novel framework integrated with the data mining for performing knowledge discovery process in BigData.
- To design a sophisticated and efficient search engine for extracting precise knowledge using data mining techniques in stream of big data.
- To accomplish standard benchmarking of the proposed framework.

### Proposed Methodology

An analytical model will be frame up that can effectively perform knowledge discovery. In order to do so, the analytical model will be designed using open source framework for storage in Hadoop and programming model MapReduce virtually. The proposed work will create a Knowledge repository and development of search engine that will also enhance the capability of search and sharing applications that will be developed on the top of BigData.

The proposed study develops a virtual platform that enables domain applications to not only access live streams of data. A large number of streams of data arriving continuously upon a data-center from hundreds and thousands of clusters located in distributed manner can swamp even the most robust clusters in cloud. In order to support the historical data feeds, the framework is also required to store data from feeds in a format easily compatible in cloud environment. Storing continuous feeds along with queries from large numbers of clusters requires highly scalable and efficient storage and retrieval mechanisms. Therefore, the next phase of the study focuses on designing a technique for Streaming data to Cloud and store as HDFS format for fulfilling the goal of the study. It is suggested that the cloud is organized in the Hadoop way. Both of Hadoop's storage system (Hadoop Distributed File System, HDFS) and data processing system (Map-Reduce framework) have a master/slave architecture, in which the master is responsible for storing file meta data or scheduling jobs, and slaves are responsible for storing file content or task execution (task is one piece of split job). The two masters could be deployed either in the same physical node or in different physical nodes. In its storage system, when a client accesses a file, it firstly contacts the master to retrieve the file's meta-data, then it directly retrieves the content from the slaves which store a part data of the file. In its data processing system, the file used is submitted to the master, and then the master distributes the file to the slaves which store the processed data, or to those slaves near by the slaves storing the processed data. As a result, the data is processed almost locally and the data movement (thus bandwidth requirement) is reduced. Hadoop supports data compression mechanism to reduce bandwidth cost as much as possible. The schematic architecture of the proposed system is highlighted in Figure 5.

*Figure 5. Schematic architecture of the proposed study*



Development of efficient software module using Hadoop and MapReduce framework to categorize the material being uploaded to be stored based on text, words in audio & visual text in video. To achieve this purpose, Hadoop (HDFS) will be designed for long term persistent storage to the datasets that are managed as indexed files. MapReduce will be enhanced to replicate the data for ensuring its longevity, to decrease the latency when retrieving it and to provide opportunities for better parallelism.

The proposed study accomplishes this parallelism in two ways e.g. i) the individual data files can be processed in parallel and ii) MapReduce can be typically configured to create replicas of files for safe cloud-storage purposes. These replicas can also be processed in parallel. MapReduce will also cache data connections. Therefore, frequent data transfers between the same pair of nodes do not need to set up a data connection every time. Then Hadoop will be designed to execute the user defined functions in parallel using the stream processing pattern for data managed by the MapReduce. The same user defined function will be applied to every data record in a data set managed by MapReduce and also to each segment of the dataset independently. An important advantage provided by a system such as using Hadoop is that often data can be processed in place, without moving it.

A sophisticated search engine will be developed using data mining techniques in Hadoop / MapReduce framework that will address all the needed requirements such as dealing with large amount of data sets and to increase the user interaction performance of data mining which will go behind simply locating documents containing certain keywords, but will help users find material on the basis of related ideas, cross-referenced research, spoken or displayed words and diagrams in audio and video material. Finding all frequent item sets is the core of association data mining and it has maximum calculating workload, then according to the minimum confidence threshold the effective data mining rules will be constructed from the frequent item sets. To achieve this, the new enhanced algorithm for data mining will be implemented in Hadoop/ MapReduce framework on cloud environment. Therefore, the proposed study discusses about a software analytical framework that can use BigData and can perform an efficient knowledge discovery process using storage and efficient search engine.

## Possible Advantages of Conceptual Framework

The expected outcomes of the study are as follows:

- **Efficient Framework:** The proposed study introduces a completely novel framework that can effectively perform knowledge discovery from Big Data considering cloud environment.

- **Cost Effective Knowledge Discovery:** It is anticipated that the study will put forward an application that considers real time dataset and is compliant of time and space complexity.
- **Superlative Semantics:** The proposed system attempts to make use enriched ontological representation in the object properties in the context of cloud SaaS service discovery on big data.
- **Computational Effective Data Mining:** The proposed system performs modelling of the big data based on dependent and independent constraint, where in all the study phases, the proposed system performs filtering of ir-relevant contextual based data that reduces computational complexity.

## Research Implication

The proposed system offers customizable analytical features on educational data using BigData approach that offers a highly enriched knowledge delivery process in any forms of educational system. The technical feasibility of the proposed system is quite cost effective owing to adoption of open-source tools e.g. Hadoop and MapReduce software framework. The economic feasibility of the conceptual framework is also quite high as hardware / infrastructure dependency becomes negligible owing to adoption of cloud services. Usage of novel semantics along with data mining algorithms allows the conceptual framework to extract maximum knowledge from the highly challenging unstructured data. The implication of the proposed research work is quite high and is not restricted to a specific case study. It can be applied on school-level educational system to advanced-level educational system on any discipline. Owing to the adoption of open source framework, it can take the input of high-dimensional data that can be either semi-structure or unstructured.

## Benefits and Disadvantage of Conceptual Framework in Operation Management

The operation management in the education system can be greatly automated for performing complicated analytical operation by the adoption of proposed conceptual framework. The direct advantage of the proposed study is to enhance the teaching-learning experience. The adoption of digitization and automation in educational institution leads to generation of massive set of education data. Such data may include information related to administrative operation and technical operation related to the higher education subjects. However, the data just increase its size of storage which doesn't lead to generation of much value added information for

future. The conceptual framework has the potential to capture the spontaneously growing educational data that allows the stakeholders to extract the highly valuable information e.g.

1. The proposed idea of conceptual framework can identify the problems of dimensionality in the massive and unstructured/semi-structure educational data using open source and highly customizable software framework. It not only enhances industry standard data storage over clusters but also performs a highly customizable data retrieval process.
2. The presented framework can assess the performance of faculties as well as student based on varying parameters which are very different than conventional predefined feedback-based parameters. A stakeholder can define its own parameter to judge the performance of both students and faculties. This feature will also encourage stakeholders of operation management to adopt various industry-based quality standards to further improve the process delivery.
3. It can identify the critical factors responsible for poor academic performance. Based on wide scope of data, the proposed system applies semantics and data mining that can extract all the hidden subspaces in high-dimensional data thereby discovering the latent data from massive data. Such features lacks in existing analytical applications.
4. It can predict essential parameters that can leverage the skills of actors involved in education industry. The proposed system can forecast the precise remedy that is essential for actors to fulfill any existing gap of a particular skill.
5. The proposed system will also involve a module for risk analysis that includes amount of costing and probability of return of investment to be assessed on multi-dimensional performance parameter. It can also enable the stakeholder to make decision making towards futuristic education-related ventures. It will also assists in evaluating the impact of any unforeseen risk towards educational quality.

The above mentioned points are the direct reflection of the fact that the proposed conceptual framework can offer a potential analytical power in the operation management of the educational section by enhancing the teaching-learning experience, minimizing the administrative task, decision-making in better strategic planning.

Although the proposed system is designed to overcome the problems pertaining to non-usage of educational data, it still has some disadvantage. The proposed system doesn't emphasize on security towards accessibility features owing to its focus on developing analytics on education data. Due to adoption of open-source framework,

it may give added advantage to hackers, who use Java extensively to access any open-source framework. This disadvantage can be overcome by developing a novel cryptographic algorithm in future.

## CONCLUSION

This chapter has discusses about the significance of the BigData especially from the context of the educational section. The discussion says that modern education systems are increasingly adopting cloud-services that lead to enormous rise of educational data. Such data are characterized by massiveness in size, high dimensionality, and 5V of BigData. Owing to such problems, the educational data are not utilized after it is being stored. Although there are some of the research attempts towards educational datamining and BigData analytics, but still simplicity, preciseness, and validation of such framework is missing. Hence, the chapter introduces a simple operational management framework that can efficiently use the educational based BigData.

## REFERENCES

Alam, G. M., Hoque, K. E., & Oke, O. K. (2010). Quest for a better operation system in education: Privatization and teacher educationalization or voucherilization glimpsing from consumer and product perspectives. *African Journal of Business Management*, *4*(6), 1202–1214.

Ali, M. M. (2013). Role of data mining in education sector. *International Journal of Computer Science and Mobile Computing*, *2*(4), 374–383.

Aqdam, B. A., & Bonab, M. B. (2014). Possible Evaluation Replacement of Supporting Decision Systems in Administrative Education System in Sarab Town. Academic Press.

Bernbaum, M., & Moses, K. (2011). *Education Management Information Systems*. Education Policy and Data Center.

Brown, C. A. (2014). *Globalization, International Education Policy and Local Policy Formation: Voices from the Developing World*. Springer.

Daniel, B. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*. doi:10.1111/bjet.12230

Drigas, A.S., & Leliopoulos, P. (2014). The Use of Big Data in Education. *International Journal of Computer Science Issues*, *11*(5).

Ellison, R. (2004). *Scoping exercise for possible work by DFID on Education Management Information Systems.* Academic Press.

Fernández, A., Peralta, D., Benitez, J. M., & Herrera, F. (2014). E-learning and educational data mining in cloud computing: An overview. *International Journal of Learning Technology*, *9*(1), 25. doi:10.1504/IJLT.2014.062447

Harrison, G. (2012). Outdoor Program Administration: Principles and Practices. *Business Economics (Cleveland, Ohio)*, 406.

Holley, K., Sivakumar, G., & Kannan, K. (2014). Enrichment Patterns for Big Data. *IEEE International Congress on Big Data*.

Johnson, J.A. (2014). The Ethics of Big Data in Higher Education. *International Review of Information Ethics*.

Kats, Y. (2013). Learning Management Systems and Instructional Design: Best Practices in Online Education. *Education*, 367.

MacHardy, Z., & Pardos, Z. A. (2015). Evaluating The Relevance of Educational Videos using BKT and Big Data.*Proceedings of the 46th ACM Technical Symposium on Computer Science Education*.

Nolan, C. J. P., Ayres, D. A., & Mckinnon, D. H. (1997). *An operational model for the implementation of computerised school information systems*. Massey University. doi:10.1007/978-0-387-35090-5_12

Panneerselvam, R. (2012). Production and Operations Management. *Business Economics (Cleveland, Ohio)*, 720.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational data mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Prakash, B. R., Hanumanthappa, M., & Kavitha, V. (2014). Big Data in Educational Data Mining and Learning Analytics. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(2).

Ramasubramanian, P., Iyakutti, K., & Thangavelu, P. (2009). Enhanced data mining analysis in higher educational system using rough set theory. *African Journal of Mathematics and Computer Science Research*, *2*(9), 184–188.

Seay, C. (2015). Using Hadoop on the Mainframe: A Big Solution for the Challenges of Big Data. IEEE International Information Technology, 765-769.

Sivaraman, E. (2014). High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing Using Hadoop.*IEEE International Conference on Intelligent Computing Application*. doi:10.1109/ICICA.2014.16

Sqalli, M. H., Al-Saeedi, M., Binbeshr, F., & Siddiqui, M. (2012). UCloud: A simulated Hybrid Cloud for a university environment.*IEEE International Conference on Cloud Computing*. doi:10.1109/CloudNet.2012.6483678

Sun, Y., & Han, J. (2012). Mining Heterogeneous Information Networks: A Structural Analysis Approach. *ACM SIGKDD Exploration Newsletter.*, *14*(2), 20–28. doi:10.1145/2481244.2481248

Thakar, P., & Mehta, A., Manisha. (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *International Journal of Computers and Applications*, *110*(15).

UNESCO. (2011). Transforming Education: The Power of ICT Policies. *Education*, 237.

Wang, B., & Xing, H. Y. (2011). *The application of cloud computing in education informatization. In Computer Science and Service System* (pp. 2673–2676). CSSS.

Zhao, Z. (2015). *Visualization of HashStash with Qt*. (Masters Thesis). Brown University.

## KEY TERMS AND DEFINITIONS

**Cassandra:** It is a distributed database management system developed for managing massive data over multiple commodity servers. It is an open source application.

**CollegeSAAS:** Supported by Software-as-a-Service, it is web-based college management tool used to manage various administrative task along with supportability of various client-specific applications. It is already in use by Universities of India.

**Entrar:** Entrar is the school management application of Delhi Public School, Bangalore. This simple web-based portal shares heterogeneous forms of data giving rise to increase of educational and information data.

**ETL Scripts:** A programming script that is adopted in conventional datamining technique for Extracting data, Transforming data to appropriate format, and Loading data to targeted node.

**Hadoop:** It is a software framework used for storing the BigData in the rack servers or the clusters. Hadoop is an open source and has capability of scaling up one server to thousand computing nodes.

**HDFS:** It stands for Hadoop Distributed File System and is a java based file system for Bigdata.

**iCampus:** It is a collaborative educational tool designed by Massachusetts Institute of Technology (MIT) along with Microsoft.

**MapReduce:** It is a processing framework of larger dataset and it also poses capability to generate larger dataset.

**NoSQL:** It is a type of database that provides storage and retrieval mechanism used for BigData as well as various real-time

**Silo:** Silo is a termed used with database isolation that directly represents a group of fixed data that is not required by the organization in regular operations on database management system.

# Chapter 7
# Management of SME's Semi Structured Data Using Semantic Technique

**Saravjeet Singh**
*Chitkara University, India*

**Jaiteg Singh**
*Chitkara University, India*

## ABSTRACT

*Management of data for an organization is crucial task but when data goes to its complex form then it becomes multifaceted as well as vital. In today era most of the organizations generating semi structured or unstructured data that requires special techniques to handle and manage. With the needs to handle unstructured data, semantic web technology provides a way to come up with the effective solution. In this chapter Synthetic Semantic Data Management (SSDM) is explained that is based semantic web technique and will helps to manage data of small and Midsized Enterprise (SME). SSDM provide the procedure to handle, store, manages and retrieval of semi structured data.*

DOI: 10.4018/978-1-5225-0886-1.ch007

## INTRODUCTION

Big data analytics is process to analyze organizations data that may be of structure, unstructured or semi structured type and leads to meaningful insights and information. This analysis process uncovers hidden facts, unknown relations, customer requirements and business needs. Result of analysis provides effective customer services and edge over competitive environment. Data analysis done by traditional tool but big data analysis require special tool that can handle volume, verity and velocity. These may be of Hadoop related or NoSQL database type.

Semantic technologies can be effectively works with Variety, a third V of big data. Semantic technology is a type of graph model that deals with RDF data. This graph model helps to store data and analyze it using semantic query tool. Semantic technique provides easy and cost effective solution to unstructured data analysis. As in the current world wide web environment complexity and amount of data increase at unexpected rate that demands new storage techniques that can handle data semantically.

Traditional data storage methods are based on relational techniques that are not effective to handle unstructured data so unstructured nature of data motivates to use semantic technology. Facebook and Google like big data generator uses semantic graph for its data storage and handling. Semantic technology uses RDF as a flexible data model and ontology to represent linked data. Ontologies are use to represent semi-structured data and provides means to store and extract knowledge. Current relational model and XML model are widely used to represent structured and semi structured data with limited functionality. XML model provides syntax based data representation but weak semantic support whereas entity relationship model (relational model) can deal with semantic at non user end but create difficulty for end user to understand when it is transformed to physical data model.OWL and RDFS ontology provide best means of storage of semantic and easy query based data retrieval.

With increasing market antagonism and complex business performance management, it is significantly important for contemporary organizations to maintain a single group of core entities across many systems within an enterprise to improve business efficiency and customer satisfaction. There arises high demand to analyze data, which refers to core business entities a company uses repeatedly across many business processes, such as lists or hierarchies of customers, suppliers, accounts, or organizational units.

In this chapter, we propose SSDM (Synthetic semantic data management) approach. SSDM provide data management solution for a Small and Medium size Enterprise (SME) that will be helpful for analyzing customer behavior and business intelligence. Ontologies are used for data representation data management and semantic queries are used to retrieve facts for stored data. In this chapter we manage data of food chain enterprise (SME) using OWL script. For storing data we use propose algorithm (SSDM) that implemented on RDOTE tool and data is retrieved using SPARQL queries and DL inference. This paper covers following topics after the introduction section:

- Survey of Big data and semantic technology,
- Semantic Web Languages,
- Synthetic Semantic Data Management (SSDM),
- Steps of Implementation,
- Tool used for SSDM,
- Conclusion, and
- References.

As an outcome of this chapter, we will be able to store and manage data using semantic technique and also able to handle variety of data. Chapter gives the of following:

- Introduction to semantic technique,
- Different type semantic scripts and their importance for big data management,
- Difference between XML, RDF, RDFs and OWL scripts,
- SSDM approach for data management,
- SPARQL query execution,
- DL inference for collecting facts from defined rule, and
- RDOTE and Protégé tool for data management and inference.

## BACKGROUND

Semantic technology presents a way to connect data so the knowledge can be retrieved from them. Semantic technology is also known as linked data technique because it enables to connect different type of data in graph like connected structure and meaningful way that makes computers to take decision in own way. Semantic technology gives storage method of knowledge instead of data, and user can understand data and reason data at execution time. Concept of semantic web was given by Tim Berner Lee in 2001 as extension of web with the use of World Wide Web consortium. WWW at the time of semantic web invention just provide understanding to the user but semantic web intends to enhance the capabilities of WWW and to overcome the limitations of WWW. Semantic web provides way that enables user information to be understood by machines. Semantic web provides a uniform structure to present information of WWW, databases and other type of structured or semi structure data sources (Lu et al., 2007)

Semantic web works to provide consistent data with less volume to be readable by machines. Relational databases like MySQL and Oracle from flexibility to store data. These databases capture data with little meaning so less integrity with more raw data. Semantic data model overcomes this limitation by providing meaning based data storage methods. Semantic data models are based on the concept of subject, predicate, object and hieratically relationship.

Lu et al. (2007) present and demonstrate Scalable Ontology Repository abbreviated as SOR. SOR is used for the storage, inference, reasoning and inference of ontological data. It acts as basic and practical model for ontology management system. Apart for management SOR also acts as a complete framework for the generation and querying the semantic data. With the guidelines of SOR one can inference the semantic owl data and also fetch hidden fact in it using SPARQL like semantic query language. This paper explains the semantic web languages and their relationship among each other SOR ontology act as basic for semantic web and with advancement of technology it act as supporting technique for world wide web consortium (W3C). W3C supports all the semantic technique. It starts with HTML and continues with non relation technology like XML, RDF, and OWL etc.

Wang et al. (2009) X Sun, F Cao, N Kanellos, K Zhang, Y Pan and Y Yu in this paper gives a proposal to implement Semantic MDM (SMDM). SMDM technique uses relational table as a RDF sets and semantic query is applied on it. This approach actually uses Semantic graph called Ontology graph or ontologies on which queries are applied. Core MDM ontology and external imported domain ontologies

are the two type of ontologies that are used in this technique. The SMDM engine use relation MDM database as a RDF store that links Core MDM ontology to open data ontology. SPARQL queries are used on ontologies or RDF set. SPARQL is converted to SQL when applied on relational data.

This research gives two aspect of SMDM that are:

- SMDM enables discovery of hidden relationship and extension of SPARQL to rule base SPARQL.
- SMDM provides SPARQL to SQL translation.

Features of SMDM system are:

- Cross-domain linking provides linkage between different relational databases as virtual RDF set.
- Dynamic concept extension means user can extend MDM concept by user defined rule or OWL.
- Semantic query and implicit relationship discovery enables semantic query execution using rule base technique and discovery of hidden relationship in different RDF set, it also drives new information.

In this architecture SPARQL query act as input to SMDM system, and input is translated into SQL query using SPARQL to SQL translator. D2RQ act as mapping logic for RDB to ontology translation. Dialog rule engine is use for ontology reasoning and reason for user defined rule.

Big Data: Big data is a data that is big in the sense of storage; needs more processing capacity of conventional database system and moves faster than normal data. Big data is little nebulous term due to diverse technology coverage. Input for big data system could be data from social network, web server, audio streaming, broadcast data, satellite content, mobile data, cellular data, web page content, medical images, geographical data, automobile telemetry, stock market data and goes on. The values of big data for an organization depend on either its analytical use or new product of organization. For example Facebook uses persons profile information \ for analysis and adds new options or promotional ads according to person's activity. Due to big data characteristics it is very difficult to perform analysis on Big Data using conventional techniques (Sagiroglu & Sinanc, 2013).

## Characteristics of Big Data

Big data can be defined by its characteristics, which are defined by 5 V's.

- **Velocity:** Big Data is characterized by velocity due to its varying speed of incoming and flow. Velocity of big data is define by the rate at which new data is generated and speed at which it is propagated from one location to another, these location can be their origin or where it is use. Velocity of big data can be understand by rate at which social site messages are propagated and get viral in second. Transaction of bank, stock market data, trading activities related data movement is hasty.
- **Varity:** Concept of Varity in big data deals with heterogeneity in data, it can be structured or semi structured. In past generated data very really found to be heterogeneous, due to same type of data we can easily put that data into tables or relations. Now a day's data generated is of different type and according to statistics eighty percent data is unstructured that forces to generate new solution to store data. Unstructured data includes photos, videos, text, graphics etc.
- **Volume:** Size of Big Data is very large than conventional data. The term "Big" for big data is coupled due to its volume that is in the range of Petta byte and can reach to Zeta in future. Volume of big data is amount of data generated per second and according to facebook, total ten billion messages are generated per day, four billion 'like' clicks and four hundred million updates are done by the users. Except facebook there are many other social sites that are less or more equals to facebook and they have approximate same data generation amount. Data with this volume cannot be stored and analyze with traditional system so it requires new techniques for data handling.
- **Variability:** Variability deals with inconsistency in speed of data loading in database and challenging when data related to social sites with its Peak size. Dependability or disorderliness refers to the variability of data. Social site content's reliability, hash tags; abbreviations etc are related to variability of data.
- **Value:** Big Data is valuable because it is use for analysis and result of analysis is use for important Business decision. Value is not a comment on big data but a type of complement that makes other factor's solution necessary and effective. Big data is valuable for business and analysis activities, as big data source is social sites and now a day's social sites are heart of business so their data is as valuable.

Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior (Wigan & Clarke 2013). Thus main challenge of Big Data:

- Designing of such system that can efficiently and effectively handle Big Data.
- Filtering and analysis of Big Data for business values.

Application of Big Data is for cloud computing, Business decision for customer retention, Social Sites like Facebook, Google etc, and Business extension.

Semantic technology comes to existence in around 1998 and becomes the future internet technology. Semantic technology enables computer system to work based on meaning instead of syntax. Jeopardy of big data ended with interception of big data and semantic technology. Now a day's big data techniques and technologies are effectual due to semantics.

Roughly all big data solution by now uses lots of the components that are dependent fortification of semantics. Analytics of sophisticated text, large volume of data and unstructured data are comes under semantic buttressing. Today's era big data and semantic techniques commonly exist in following way:

- Generation of intelligent responses,
- Processing of context and semantic data,
- Development of proactive system based on aggregated knowledge,
- To enhance response time, and
- Increased accuracy based on context based understanding.

Steiner el al.(2012) present a new approach used by Google called knowledge graph; a graph that represents real world object and their relationship. Knowledge graph includes entity likes person, movies, sports team, their features and their relationship between other entities. Knowledge graph makes searching faster in four ways:

- By disambiguation of search queries,
- By search log based summarization of keys,
- By explorative search suggestion, and
- By adding real world entities or through real coverage.

With knowledge graph now search based on things not strings and user can sense words without any ambiguity (Singhal, 2012). This work demonstrates how static, structured world facts can be combined for profit with dynamic unstructured social network fact (Steiner el al., 2012).

Cuesta et al. (2013) outline the foundation of architecture used to handle big data that uses No-SQL technology. The architecture is named as SOLID (Service Online Index Data architecture) disintegrates big data complexities from its real time consumers and producers. This architecture provides easy data handling approach by using RDF concept. SOLID provides storage efficient solution by compressing data storage space and indexes. It also tackles separation of real time data generation complexities from big data semantic consumptions. Moller and Wandelt(2013) explain the concept, to solve ABDEO (Accessing Big Data with Expressive Ontologies. This solution is based on DL SHI: A concept based instance retrieval on big data store.

Calvanise and Haase (2013) are founder of Optique, a solution of big data handling using OBDA (Ontology Based Data Access) approach. ODBA approach deals with big data dimensions that include variety, volume and velocity and reduce the data handling complexity and cost. By using ontology based end user query interface for heterogeneous data OBDA system reduced the big data complexity. Ontology in this model is used to handle Variety attribute of big data whereas Velocity and volume are managed by data integration and query evaluation techniques. In short Optique provides a way to handle big data integration using semantic technology.

Qu (2013) proposed a method to process Big data using RDF technique. According to this technique data is converted to semantic graph and RDF data set. Then queries are applied on RDF data using semantic techniques that were matched with graph data and gives result in semantic form. Semantic technique focuses on semantic or meaning of data instead of structure. It deals with knowledge retrieved from data instead of raw data. Following Figure 1 depicts the main focus of semantic and non semantic technology relatives to data-knowledge hierarchy.

## SEMANTIC WEB LANGUAGES

Ontologies are backbone of current semantic web. Ontology provides precise and shared linked data that can be understand by machines. This linked data gives concept and relationship. Figure 2 shows language hierarchy.

*Figure 1. Data knowledge relationship*

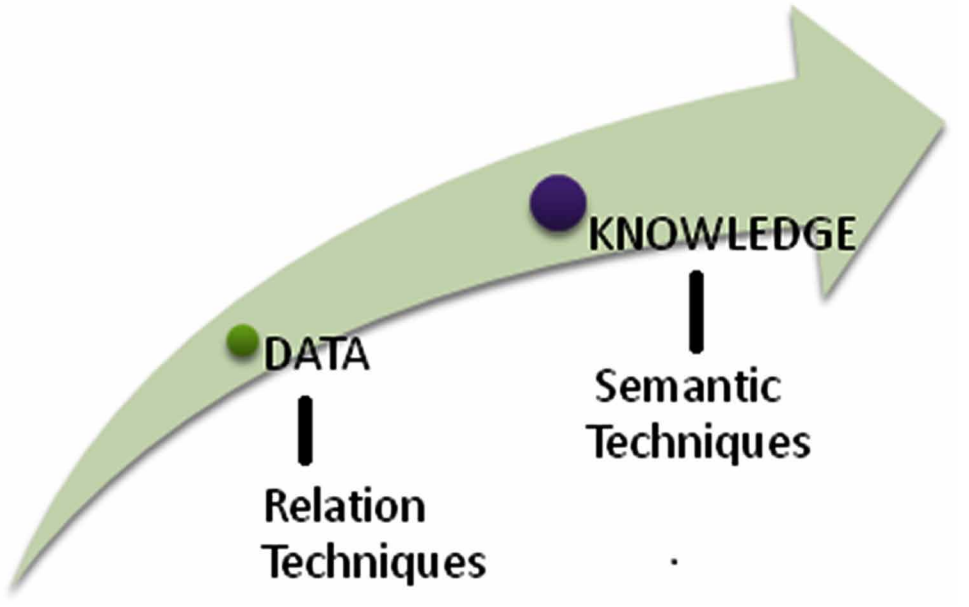
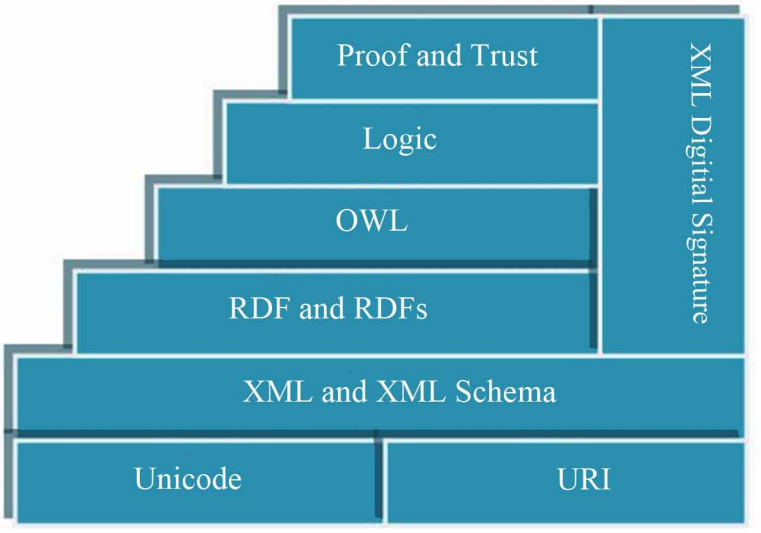



*Figure 2. Semantic web layer architecture*
*(Gil & Ratnakar 2002)*



Ontologies can describe semi-structured, heterogeneous and distributed information due to support to define complex object semantics. Languages used for ontology design are XML, RDF(s), and OWL. These languages are defined in layer architecture that dreams to drive semantic web to its peak potential. The layer architecture provides the base for semantic web (Gil & Ratnakar 2002).

## a. XML

XML stands for eXtendible Markup Language is extension of HTML. It is tag based language that provides facility to define use defined tag and compatibility with HTML tags. It provides linear structure and is web based language

## b. RDF

RDF stands for Resource Descriptor Framework. RDF is extension of XML to provide understanding to machine. It defines information structure and provides understanding to machine that is not provided by XML

RDF provides definition of data model that consist resource, properties and statements. Resources are not directly accessible from web and describe by URI. Properties are attributes of resource and statements are combination of resource and property value.RDF defines data in triplet that include subject predicate and object, where as predicate is property and object is property value. Expressive power and reasoning capabilities of RDF is limited. Interoperability of RDF is partial and its syntax is XML based. It support natural language processes and act as super set of HTML based languages. There exists two term RDF and RDFs both are interchangeable but following is deference between RDF and RDFs (Resource Descriptor framework [RDF], 2014)

- RDF is a conceptual data model and exists using triplets or rdf/xml formats.
- RDF includes statement having subject, predicate and object relationship.
- RDF is a abstract model of data.
- RDF models data and provide base for graph representation of data.
    - RDFS stand for RDF Schema.
    - It is a language to define RDF words.
    - RDFs are a building block of ontology. RDFs are similar to column in relational table.
    - It gives schema level information. It is a terminology.
    - For example cow and animal are RDF set and relation between cow and animal is RDFs. For example cow is a sub class of animal is RDFs.

## c. OWL

OWL is Ontology Web Language defined to overcome the RDF short comes and to provide more effective model to semantic web. OWL is developed in three stages named OIL, DAML and OWL. OWL consist three main components that are ontology, axioms and facts. Fact is a triplet or class that includes individual, property and value. Axioms provides link between classes with its complete or partial information to give logical meaning of class. Ontologies are set of axioms, fact and relationship between other ontology. Expressive power and reasoning capabilities of owl is highest. It provides built-in versioning capabilities. It is scalable and support natural language processing. OWL-Lite, OWL-Full and OWL-DL are three type of owl. It's still under development so developer needs to update with its updated features [13]. Following terms are used to distinguish XML, RDF and OWL and Table 1 summarize all these terms.

1. **Context:** Context defines the environment in which item can be used. Context for language can vary from interpreted term. XML uses namespaces to define context. Namespace is limited form of context. <*xmlns:*label*=*"URI"*> is used to specify the namespace. In this syntax label is prefix that specify the particular element of schema if it missing then default namespace is referred. RDF uses XML namespace to identify schema. The only difference is, in case of RDF namespace also define the location but in case of XML <*xsi: schemaLocation*> is use to specify particular location of schema OWL uses XML namespace to refer ontology but to use ontology definition import tag is used to import particular ontology.
2. **Subclasses and Properties:** Subclasses defines the relationship between different classes on the bases of relationship and common criteria. This relationship can be denoted by is-a relation. Relationship between two classes is decided by properties or we can say properties define the type of relation between different classes. Combination of class and properties completely defines entity. XML does provide class and property term instead of these, it uses element and sub elements. Lack of class and properties term XML mapping to semantic model gives ambiguous result. RDF provides class and properties. Class in RDF is define by <*rdfs: class*> and subclass is define by <*rdfs:subclassOf*>. Similarly we can define properties. In owl classes are define by <*CLASS IRI="#classname*> tag and properties are define by <*objectProperty*> or <*dataProperty*> tag.

*Table 1. Comparison of XML, RDF and OWL*

| Comparison Term | XML | RDF | OWL |
|---|---|---|---|
| **Different Context for environment** | YES | YES | YES |
| **Class and property** | Element and attributes are used in place of class and property (no defined semantics | Individual are act as class and predicates are properties <rdf:class> <rdf:property> | Same as RDF with different level of properties |
| **Inheritance** | No, but can be extends using addition feature | Yes using subclassOf and subpropertyOf | Yes using subclassOf and subpropertyOf |
| **Data Type** | Basic data type and enumeration | Only literals using HTML links | Only literals using HTML links |
| **Property Range** | Local and global | Global only P rdfs:range C Prdfs:domain C | Global same as rdf |
| **Property Domain** | Implicitly in which element are defined | Global | Global |
| **Binary operation** | N0 | Yes | Yes |
| **Property Type** | Yes | Yes <rdf:id> | Same as rdf |
| **Cardinality** | Based on restriction of data type value | Based on class | Based on class and property |
| **Instance** | | | |
| **Structure** | Tree based(hierarchical model) | Triplet based(syntax based model) | Fact based(syntax based model) |
| **Terminological term** | Schema | Ontology | ontology |

3. **Data Types:** It includes type of data that can be primitive or complex type. Primitives include numeric and string whereas combination of these comes under complex data type. XML uses rich range of data types these include string, numeric and complex. OWL and RDF uses only literal and linked construct of HTML.

Example of XML data types

```
<xsd:example ...>
    <xsd:first id="Age" name="john">
      <xsd:restriction base="integer">
       <xsd:minInclusive value="1">
      </xsd:restriction>
    </xsd:first>
    ...
  </xsd:example>
```

similarly base can be float, string etc.

Data type in rdf is define as HTML links that are in the form of literals following is example of rdf data type definaton

IRI http://www.w3.org/2015/06/22-rdf-datatype-ns#HTML

Or

rdf:HTML

similarly owl uses rdf based data types.

## d. Instances

Instances are object of classes described by specified properties of class.

## e. Range, Domain, and Constraints

Range decides the class that can assign value. Domains describe the property holding classes. Constraints are defined to restrict and relate the properties of a class.

Cardinality constraints define the range limit of class. XML uses type attribute to limit the range of element. maxInclusive and minInclusive are used to specify maximum and minimum range of an element. Cardinality of XML element can be global if no parent else local. RDF provide global range and domain on the bases of property. OWL provides local and global range by using restrict attribute.

Below is example of XML domain.

```
<? xml version = "1.0" encoding="utf-8"?>
<Dexample>
       xsi:type = "typens:Dexample">
  <DomainName>First</DomainName>
  <FieldType>Firstfield</FieldType>
 <Description>Domain creation example</Description>
  <Owner>saravjeet</Owner>
  <MaxValue >5000</MaxValue>
  <MinValue >50</MinValue>
</Dexample>
```

RDF Domain and range example

```
   <--http://www.semanticweb.org/laptop/ontologies/2015/5/unti-
tled-ontology-101#Height -->
    <owl:ObjectProperty rdf:about="http://www.semanticweb.org/
laptop/ontologies/2015/5/untitled-ontology-101#Height">
        <rdfs:domain rdf:resource="http://www.semanticweb.org/
laptop/ontologies/2015/5/untitled-ontology-101#child"/>
        <rdfs:range rdf:resource="&owl;Thing"/>
        <rdfs:subPropertyOf rdf:resource="&owl;topObjectProper
ty"/>
    </owl:ObjectProperty>

OWL domain and range example
<Declaration>
        <Class IRI="#child"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#height"/>
    </Declaration>
    <SubClassOf>
        <Class IRI="#child"/>
        <Class IRI="#child"/>
    </SubClassOf>
    <SubObjectPropertyOf>
        <ObjectProperty IRI="#height"/>
        <ObjectProperty abbreviatedIRI="owl:topObjectProper
```

```
ty"/>
    </SubObjectPropertyOf>
    <ObjectPropertyDomain>
        <ObjectProperty IRI="#height"/>
        <Class IRI="#child"/>
    </ObjectPropertyDomain>
    <ObjectPropertyRange>
        <ObjectProperty IRI="#height"/>
        <Class abbreviatedIRI="owl:Thing"/>
    </ObjectPropertyRange>
```

Following is example of XML constraint.

```
<xsd:example2 ...>
     <xsd:person id="young" name="john">
       <xsd:restriction base="integer">
        <xsd:minInclusive value="18">
        <xsd:maxInclusive value="40">
       </xsd:restriction>
     </xsd:person>
   </xsd:example>
```

## f. Inheritance

Inheritance defines the parent child relationship. According to it valid constraint and properties for parent are also valid for child. Single, multiple, multilevel are the type of inheritance. XML does not directly provides inheritance extends element and data types where as RDF and OWL supports inheritance using subclassOf and subpropertyOf tags.

```
RDF inheritance
<rdfs:subClassOf rdf:resource="http://www.semanticweb.org/lap-
top/ontologies#child"/>
http://www.semanticweb.org/laptop/ontologies#Height -->
    <rdfs:ObjectProperty rdf:about="http://www.semanticweb.org/
laptop/ontologies/2015/5/untitled-ontology-101#Height">
        <rdfs:subPropertyOf rdf:resource="&owl;topObjectProper
```

```
ty"/>
    </rdfs:ObjectProperty>
```

Inheritance in owl

```
<Declaration>
     <Class IRI="#child"/>
    </Declaration>
    <SubClassOf>
        <Class IRI="#child"/>
        <Class IRI="#child"/>
   <SubObjectPropertyOf>
        <ObjectProperty IRI="#height"/>
        <ObjectProperty abbreviatedIRI="owl:topObjectProper
ty"/>
    </SubObjectPropertyOf>
```

## g. Ontology a Terminological Knowledge

Ontology is base of semantic model and defines core term of semantic web. XML defines data in the form of schema that is similar to ontology with limited expressiveness. It define element of in built type and user define type having rich capability of data types. RDF defines ontology using rdfs language. Ontology in rdf consists subject predicate and object. OWL defines ontology in triplet form. OWL provides rich support for ontology design. It provides complex class and axioms and facilitate effective query processing (Bizer Heath & Berners 2009).

## h. Structure

XML uses tree structure to represent data. It uses hierarchal data model. RDF uses triplet data model that includes individual, property and statement. It is syntax based approach and uses binary relation between data- data and relation – relation. OWL is class, property and axiom based language. Due to big data characteristics, special architecture is requires to store and handle it. Semantic technologies are very helpful for implementing big data application. Advancement in semantic technology and its effectiveness toward data storage gives hike to semantic technology.

## SYNTHETIC SEMANTIC DATA MANAGEMENT

Data management reasserts IT role for managing critical information that is business data, its quality and reliability standards. Data present in various shapes, sizes and forms, represents valuable information that needs to be standardized and stored in a meaningful way within a central repository accessible by all business units. These shapes can be of type unstructured, structured or semi structured. SSDM provides a way to manage and store semi unstructured data in meaning full way. Many big enterprises, production houses and service companies monitor social media sources like Facebook, Twitter and Google+ etc. to get customer reviews and feedback. Technology giants companies rely on semantic techniques to manage their data in the expectancy of knowledge. SSDM try to simulate semantic architectures on relatively limited number of data.

As methodology provides theoretical techniques and step to be performed in implementing the idea so SSDM is divided into following steps that are developing core ontologies on identified RDF data sets, enrich the core ontology by adding semi structured or structure data, add rules and relations and finally executes semantic query techniques. To implement this work we used semantic based knowledge management technique (OWL). As with the advances of semantic web and semantic technology there is need to store and manage data in semantic form. Semantic technology gives storage method of knowledge instead of data, and user can understand data and reason data at execution time.

### Ontology

Ontology is a type of linked data structure and act as building block of semantic web. Ontology can be defined as "human intelligible and system understandable data storage method in the form of subject, relationship and object form. RDFS (Resource Descriptor Framework Schema) is first language that is recommended by w3c in 2004 for developing ontology. RDFS was built on RDF and extends features of RDF like class and properties. OWL (Ontology Web Language) is latest recommendation of ontology development. Ontology structure can be defined with quintuple as follow:

O = (CL, I, CD, P, AD) where:

- Represents ontology.
- CL represents set of classes or concept.
- I represent instance or individual of ontology.
- CD is concept of ontology and can be super or equivalent concept.
- P is properties of an ontology it can be class property or data property.
- AD act as definition of property and it includes domain value and range value.

Ontology can be visualized using graph structure where each class is shown by a node and their relationship is indicated by a directed line. Steps to develop ontology are:

- Define concepts called entities,
- Finding relation between entities,
- Organize entities, and
- Define the attributes, values they can take and their properties and define instances.

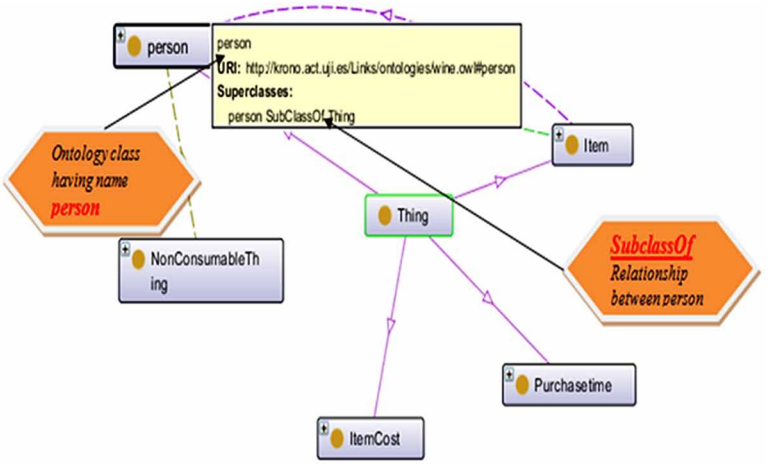
An example ontology graph is shown in figure below where node person, item, thing etc are classes and directed line represent relationship. In the Figure 3 a rectangular box explains person is class, its super class is Item and person and item have subclass relationship that is denoted by person *SubclassOf* Thing. These classes may have much relationship with other classes or with subclasses. Ontology has *rdf:* class and *rdf: subclassOf* default relationship for classes

OWL provides features like cardinality constraints, disjoint classes, equality, reasoning support etc. OWL-Lite, OWL-DL and OWL-Full are type of OWL out of these three only OWL-DL is widely. SPARQL, RDQL etc are well known query languages for ontology.

## 1. Properties

Properties are creates for classes and object. Properties basically expand the knowledge base and act as user derive relationship between classes and object. Class properties are Boolean expression between classes or regular expression; similarly object properties are regular expression of in built data type and object value or constant value. An object property between two classes can be formed as below

*Figure 3. Ontology*




Item *hascost* cost (i.e. milk *hascost* 40)

Where item and cost are two different classes and *hascost* is user defined object property, *hascost* maps two different classes. Owl script for this expression is as

```
http://www.semanticweb.org/laptop/food#hascost -->
<owl:ObjectProperty rdf:about="http://www.semanticweb.org/lap-
top/food#hascost"/>
        <rdfs:domain rdf:resource="http http://www.semanticweb.
org/laptop/food.owl#Item"/>
        <rdfs:range rdf:resource=" http://www.semanticweb.org/
laptop/food.owl#ItemCost"/>
    </owl:ObjectProperty>
```

## 2. Inference Rule

Inference is a kind of technique that is used for expansion and acquisition of knowledge. Inference is used to add new information using rules, previous data and

metadata. OWL-DL is a type of inference method and reasoning technique that is based on Hemit, Fact++ and Pellet like reasoner. These inferences method provides a way to classify, realize and verify semantic data. These rules includes reflective, symmetric, transitive etc. A very simple semantic rule of transitive type is

(a isbrotherof c) and (a isbrotherof c) - > (b issibblingof c)

In above expression left hand side data is actual data in ontology whereas right hand side is derived data using inference method. This rule provides expansion of knowledge and act as DL query to get knowledge from ontology. DL inference will also works on graphical data and gives output in graph forms.

## 3. SPARQL

SPARQL is recommended by W3C in 2008 as semantic query language. SPARQL is similar to RDFs syntax and have conjunction and disjunction of data. SPARQL is widely used for semantic query. A simple SPARQL query to search all subject and object is

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?subject ?object
        WHERE { ?cost rdfs:subClassOf ?object }
```

In the above code first four lines are used to specify the ontology and other namespace related information and vary from tool to tool. Whereas last two lines are used to fetch data from owl script and it is similar to SQL query structure with little difference in syntax. Select clause always contains subject and object name or its variable or aliases name. Condition is specified using Where clause that contains subject predicate and object. It can also have some expression using logical statements. For example

```
SELECT ?itemname ?price ?day
WHERE { ?itemname a:price ?x .
?x rdf:value ?price .
?x p:day ?day. }
```

The above query returns *itemname,* its price and purchase day associated with each items in owl script. We can also put more constraint on query like

```
SELECT ?itemname ?price ?day
WHERE { ?itemname a:price ?x .
?x rdf:value ?price .
?x p:day ?day.
FILTER(?day ==:mon)
 }
```

Now this query returns all item purchased on Monday with its purchase cost. Same way can extract data from owl script.

## SSDM Implementation Step

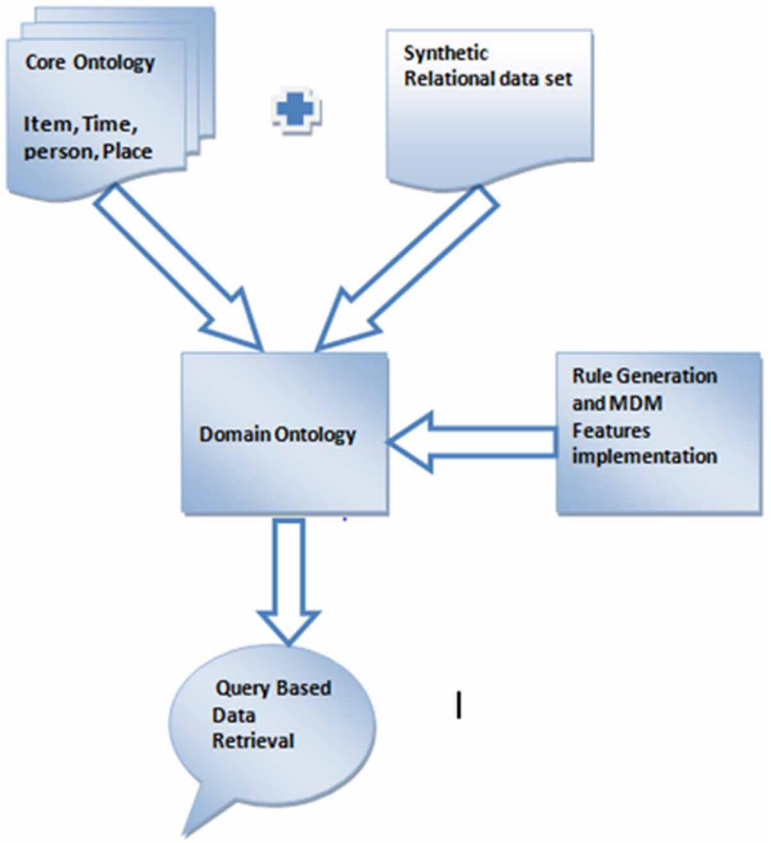
Within the scope of this work a semantic technique is proposed that act as a solution to data management process. For this process food purchase data of an organization is collected and processed using SSMD technique. The SSDM includes logic interface, reasoner capability, SPARQL query execution and expansion of knowledge base. The overall process and associated steps that are shown in Figure 4. This process has three basic step that are explain below

### 1. Development of Core Ontology

In this step core ontology is created that is based on key attributes of focused data. . In core ontology classes are created by identified entities of domain for which solution is designing. In this work identified area for data management is SME's (Small and Medium size Enterprise) that is food purchase data of an organization. Identified entities of Food Chain organization are food item name, person detail, food item cost, purchase time. Sample script for created core ontology for limited data is

```
<?xml version="1.0"?>
<!DOCTYPE Ontology [
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY xml "http://www.w3.org/XML/1998/namespace" >
    <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
    <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
```

*Figure 4. Work flow of SSDM*




```
>
]>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
     xml:base="http://www.semanticweb.org/laptop/ontolo-
gies/2015/7/untitled-ontology-105"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     ontologyIRI="http://www.semanticweb.org/laptop/ontolo-
gies/2015/7/untitled-ontology-105">
    <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-
```

```
syntax-ns#"/>
    <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-
schema#"/>
    <Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSche-
ma#"/>
    <Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#"/>
    <Declaration>
        <Class IRI="#cost"/>
    </Declaration>
    <Declaration>
        <Class IRI="#item"/>
    </Declaration>
    <Declaration>
        <Class IRI="#perchasetime"/>
    </Declaration>
    <Declaration>
        <Class IRI="#person"/>
    </Declaration>
</Ontology>
<!-- Generated by the OWL API (version 3.4.2) http://owlapi.
sourceforge.net -->
```

## 2. Finding of Synthesis Data and Convert to RDB

In this step domain related data set is detected. This identified data set contains information of customer, purchased item and other relative information. This data set gives information of customer purchase behavior and can be in any format and we need to convert into standard common format. To convert to common format detected data set is converted to relational database by using SQL * loader. SQL * loader is a feature of SQL that helps to convert bulk that into table. In this conversion each sheet of input is converted to single table of oracle. Following steps are performed for this conversion.

- Creation of target container.
- Generation of control file contain information for loader.
- Execution of sql loader by specifying require parameters.

Control file contains detail of source file and destination file. It contains a sql loader command that loads data from source to a sql table by giving the detail of input file format and columns. It has structure like follow:

Load data infile "path of input file" into target target_name filld1, field2……fieldn;

load and infile are sql loader keywords.

## 3. Development of Domain Ontology

In this data collected in above step is converted to ontology by using RDOTE tool. Relational data is stored in oracle database version depends on type of data. RDOTE provides relational data to ontology conversion and uses either MYSQL or Oracle as relational data storage space. Already created Core ontology is uploaded to RDOTE first and then connectivity is made between RDOTE and Oracle. For connectivity, tool required database username, password, hostname, Sid, database and Port number. Database can be mysql or oracle whereas sid vary from database to database. After connecting to database data and initial core ontology both are available in RDOTE. According to the requirement and core ontology properties we can enrich core ontology query data. Queries are same as oracle queries except some semantics of RDOTE requires.

### *a. Mapping Relational Schema to Ontology*

After connecting to database we need to extract all tables from oracle database and find out system table and user defined table. In this step we also check table is master data table or not. These are as below

```
Start
Create connection to the database
Check user defined data base from all table
List master data table
For each elected master table
{
If (table need association with other table)
{
Use foreign key constraint
```

```
Map primary key and foreign key and create ontology
}
else
{
Get single table and create maping for same table primary key
Create ontology
}
End
```

### *b. Mapping Attributes*

Now we get tables and attributes of that table, next is to add data properties and object properties for ontology classes. If attributes are mapping two classes then we create object property and if attribute are designated for constant value than we create data property. Each object property has range and domain, whereas data property have only domain and some predefined data type. Simple or non foreign key attributes are map to object property where as foreign key attributes map to data property. Below step explain this concept

```
Create connection to the database
For each selected table
{
For each required column
{
If(column act as foreign key)
{
Create object property in ontology
Define range value and domain value of object property
}
Else if         (column is normal attribute)
          {
Create data property for ontology
Set domain value
         }
}}
```

### *c. Mapping Constraints*

After adding concept and properties to ontology, next we need to add some addition or optional constraint to property. These constraints include minimum cardinality, maximum cardinality, exact cardinality and unique cardinality.

### *d. Data Extraction*

In this step we extract record from database and add to the ontology. For this extraction we need to deploy query, create instance and create property. These are as follow:

Mapping between data extracted from database and ontology I done through instance. Instance use to maps data extracted using query and classes then used to deepen the ontology. Instance creation requires class, queries, merge string, actual data and rename option. Class and queries should be created before instance creation. Ontology creation totally depends on instance and instances are compiled before conversion and return zero on failure and n rows on successful. If instance have error then ontology is created with zero field. Properties combine two instances with a object or data relationship. Relationship should be created be forehead property creation, and it basically connect two classes. Final ontology is stored in OWL file format. Below is subset of generated ontology.

```
http http://www.semanticweb.org/laptop/food.owl#Purchased_lunch
-->

    <owl:ObjectProperty rdf:about=" http http://www.semantic-
web.org/laptop/food.owl#Purchased_lunch">
        <rdfs:domain rdf:resource="&j.0;ConsumableThing"/>
        <rdfs:range rdf:resource="http http://www.semanticweb.
org/laptop/food.owl#Time"/>
    </owl:ObjectProperty>

    <!-- http http://www.semanticweb.org/laptop/food.owl#drinks
-->
    <owl:ObjectProperty rdf:about="http://krono.act.uji.es/
Links/ontologies/food.owl#drinks">
        <rdfs:range rdf:resource="&j.0;PotableLiquid"/>
        <owl:propertyDisjointWith rdf:resource="http http://
www.semanticweb.org/laptop/food.owl#notdrinks"/>
        <rdfs:domain rdf:resource="http http://www.semanticweb.
```

```
org/laptop/food.owl#person"/>
    </owl:ObjectProperty>

    <!-- http http://www.semanticweb.org/laptop/food.owl#eatby
-->
   <owl:ObjectProperty rdf:about="http http://www.semanticweb.
org/laptop/food.owl#eatby">
        <rdf:type rdf:resource="&owl;InverseFunctionalProper
ty"/>
        <rdfs:domain rdf:resource="http http://www.semanticweb.
org/laptop/food.owl#Item"/>
        <owl:inverseOf rdf:resource="http http://www.semantic-
web.org/laptop/food.owl#eats"/>
        <rdfs:range rdf:resource=" http http://www.semanticweb.
org/laptop/food.owl#person"/>
    </owl:ObjectProperty>
```

## SOLUTION

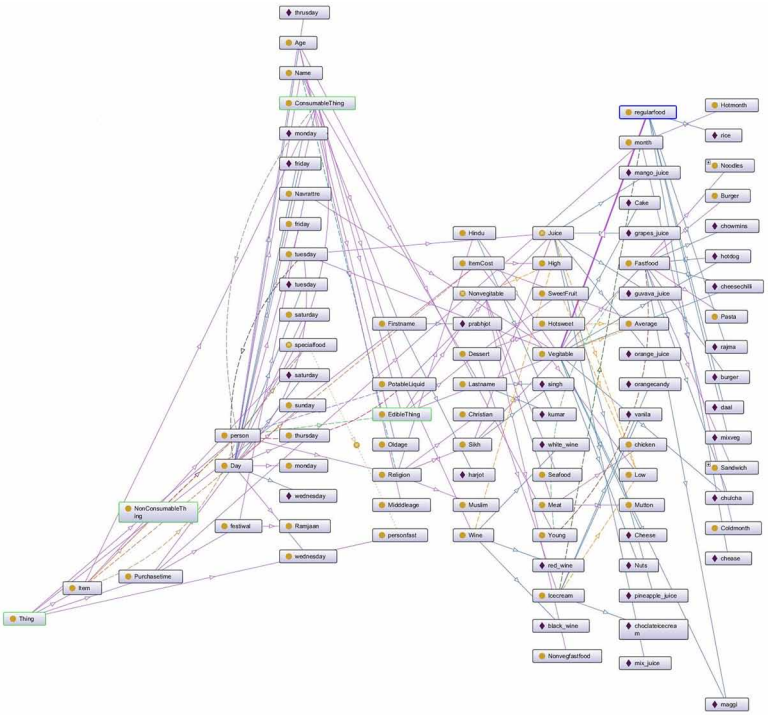
Final data management solution can be visualized by using any ontology visualization tool like Protégé. In this data is shown in the form of graph node where each node corresponds to a class and link between nodes shows relationship between classes. After creating ontology some addition rules or constraints are implied on the generated solution. These rules are based on additional requirement of particular data element e.g. rule based on person liking and purchased item image etc. Figure 5 shows the graphical view of domain ontology an output of SSDM.

## Tool

### 1. Protégé

Protégé is an open source Owl ontology and terminology development tool. Protégé provides several features like graphical user interface for creating and managing ontologies, visualization components for understanding and displaying complex relationship using graph structure, and a programming interface. Protégé community provides a facility of Plug-in architecture which helps to upgrade system and make system having rich libraries. Protégé provides SPARQL query execution environment and additional support to run logical queries by using DL Query engine. It

*Figure 5. Domain ontology an output of SSDM*




is domain-independent, freeware, platform-independent technology for developing and managing terminologies, ontologies, and knowledge bases many application domains. In protégé terminology and ontology are represented by class, slots and facts. Classes are entities, slots are attributes and properties whereas characteristics of slot are describe by facts. Knowledge base in protégé includes classes, slots, facts, axioms and instances. Protégé provides following features

- Creation of ontology in rdf or owl format.
- Creation of classes and properties.
- SPARQL and DL query support.
- OWLViz plug-in for graph based data visualization.
- FaCT and Hermit reasoned for compiling ontology.
- Merging of two ontology.

- To find to ontology difference.
- Creation of java code of ontology.
- Conversion of ontology to image file.

Protégé is platform independent tool so it is available for different operating systems, for 64 bit windows 8 it requires 500KB of RAM and java run time environment for execution.

## 2. RDOTE

RDOTE (Relational Data to Ontology Transformation Engine) is a tool used to convert relational data to ontology. RDOTE converts table data to a ontology and it can integrate data from more than one data source to a single ontology. *RDOTE* is available under GNU/GPL license and it provides easy graphical user interface. Through GUI user can import and export ontology, map data to ontology and mapping from table to class and column to properties. RDOTE provides following features:

- Import and export ontology.
- Connect to relational database.
- Provide connectivity to Oracle and MYSQL database.
- Connectivity with D2RQ.
- R2RML support.
- Execution of SQL queries.
- Creation of queries, instances, properties and merging of strings.
- It requires initial ontology having classes and properties.

RDOTE can easily enrich ontology with relational data to do this it requires already created initial ontology. It doesn't create new ontology but train created ontology. RDOTE requires primary key for each relational table if we want to integrate different table to ontology or want to map single table data to ontology. RDOTE is java based tool that requires Java run time environment to execute and work in windows operating system. Size of RDF dump is depended on JAVA heap size whereas it can create two million RDF dump in five minute with one GB heap size. RDOTE requires 14 MB RAM space and can connect only with oracle and MYSQL database.

## CONCLUSION

This study will be helpful for small and medium size enterprise to take their business decision. To start any SME (Small and Mid-sized Enterprise) identification of probable customers and their preferences are crucial because the area of operation for any SME is quite limited in comparison to any big enterprise. As SME cannot spent too money on the survey as compare to big enterprise to find the customer behavior about the product, so this research will be helpful for the SME and new organization to take sound business decision. As in current era customer detection and attraction can be done through social media that requires on relational techniques to operate and handle data. Facebook and google are the best way to publicize once idea and to know the current trends. So these social sites demands to have graph based technique to work with them. This study will help to manage semi structure data of an organization. Different semantic techniques are explain in this chapter will be helpful while selecting appropriate technique. SSDM technique is explained using food purchase habit and this technique will be helpful for the management of semantic data and for the conversion of relational data. Apart from this user can directly insert unstructured or semi unstructured data into the derive solution. SSDM approach allows unstructured data to be the part of domain ontology. AS a final output of SSDM we got a semantic database on which we can query the data using either SPARQL queries or using direct inference. Obtained domain ontology can be explore using data logics or graphviz (owlviz). This work can be used in food chain, transport system, sale and purchase organizations. In last we can conclude that this chapter will be help full for the management of SME's semi structured data using semantic techniques and act as easy way for detection of probable customer.

## REFERENCES

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Calvanese, D., Giese, M., Haase, P., Horrocks, I., Hubauer, T., Ioannidis, Y., . . . Koubarakis, M. (2013). Optique: OBDA solution for big data. In The Semantic Web: ESWC 2013 Satellite Events (pp. 293-295). Springer Berlin Heidelberg.

Cuesta, C. E., Martínez-Prieto, M. A., & Fernández, J. D. (2013). Towards an architecture for managing big semantic data in real-time. In *Software Architecture* (pp. 45–53). Springer Berlin Heidelberg.

Gil, Y., & Ratnakar, V. (2002, May). A Comparison of (Semantic) Markup Languages. In *FLAIRS Conference* (pp. 413-418).

Lu, J., Ma, L., Zhang, L., Brunner, J. S., Wang, C., Pan, Y., & Yu, Y. (2007, September). SOR: a practical system for ontology storage, reasoning and search. In Proceedings of the 33rd international conference on Very large data bases (pp. 1402-1405). VLDB Endowment.

Möller, R., Neuenstadt, C., Özçep, Ö. L., & Wandelt, S. (2013). Advances in accessing big data with expressive ontologies. In KI 2013: Advances in Artificial Intelligence (pp. 118-129). Springer Berlin Heidelberg.

Pérez, J., Arenas, M., & Gutierrez, C. (2006, November). Semantics and Complexity of SPARQL. In *International semantic web conference* (Vol. 4273, pp. 30-43).

Qu, Z. (2011). Semantic processing on big data. Advances in Multimedia, Software. *Engineering and Computing*, *2*, 43–48.

Ranganathan, A., & Liu, Z. (2006, November). Information retrieval from relational databases using semantic queries. In *Proceedings of the 15th ACM international conference on Information and knowledge management*(pp. 820-821). ACM. doi:10.1145/1183614.1183747

RDF Working Group. (2014). *Resource Description Framework (RDF)*. Retrieved from www.w3.org/RDF/

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE. doi:10.1109/CTS.2013.6567202

Singhal, A. (2012, May 16). Introducing the knowledge graph: things, not strings. *Official Google Blog*. Message posted to https://googleblog.blogspot.in/2012/05/introducing-knowledge-graph-things-not.html

Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., & Van de Walle, R. (2012, November). Adding Realtime Coverage to the Google Knowledge Graph. In *11th International Semantic Web Conference (ISWC 2012)*.

Wang, X., Sun, X., Cao, F., Ma, L., Kanellos, N., Zhang, K., & Yu, Y. et al. (2009). SMDM: Enhancing enterprise-wide master data management using semantic web technologies. *Proceedings of the VLDB Endowment*, *2*(2), 1594–1597. doi:10.14778/1687553.1687600

Wigan, M. R., & Clarke, R. (2013). Big data's big unintended consequences. *Computer*, *46*(6), 46–53. doi:10.1109/MC.2013.195

World Wide Web Consortium. (2014). *Extensible Markup Language (XML)*. Retrieved from www.w3.org/XML

# Chapter 8
# An Overview of Big Data Security with Hadoop Framework

**Jaya Singh**
*Indian Institute of Information Technology, Allahabad, India*

**Ashish Maruti Gimekar**
*Indian Institute of Information Technology, Allahabad, India*

**S. Venkatesan**
*Indian Institute of Information Technology, Allahabad, India*

## ABSTRACT

*Big Data is a very huge volume of data which is beyond the storage capacity and processing capability of traditional system. The volume of data is increasing at exponential rate. Therefore, there is the need of such mechanism to store and process such high volume of data. The impressiveness of the Big data lies with its major applicability to almost all industries. Therefore, it represents both, the tremendous opportunities and complex challenges. Such omnipotent eminence leads to the privacy and security related challenges to the big data. Nowadays, security of big data is mainly focused by every organization because it contains a lot of sensitive data and useful information for taking decisions. The hostile nature of digital data itself has certain inherited security challenges. The aim of Big data security is to identify security issues and to find the better solution for handling security challenges. The observation and analysis of different security mechanism related to the issues of big data and their solutions are focused in this chapter.*

DOI: 10.4018/978-1-5225-0886-1.ch008

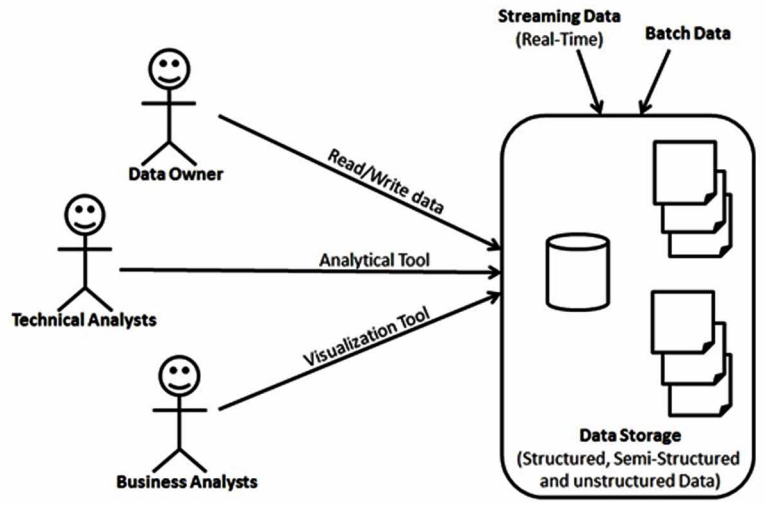
## 1. INTRODUCTION

Big data is a phenomenon that is defined by very rapid expansion of raw data. It refers to the large volume of data which is more than the storage capacity and requires more processing power than the traditional systems. Currently we are living in the world where data is the most valuable thing. So, the important thing is how to store, process and analyse the data, to get more knowledge from it. This large volume of data comes from many applications like sensors, social networks, online shopping portals and Government agencies. Storing and processing such data is a challenging task.

Big data is distributed everywhere across the multiple machines. It is a massive or vast collection of not only great quantity of data but also various kinds of complex data which previously never would have been considered together and it exceeds the processing capacity of conventional database system to capture, store, manage and analyse. Figure 1 shows the framework of Big Data through two data sources (real-time streaming data & batch data) and three data analysts (Data owner, technical analysts & business analysts) along with data storage infrastructure.

There are mainly three categories of data: structured data, semi-structured data and unstructured data (Bill Vohries, 2013). Structured data are highly organized data which have a pre-defined schema like relational database management system. Semi-structured data are those data which cannot be stored in rows and tables in a typical database. They have inconsistent structure like logs, tweets, sensor feeds.

*Figure 1. Big Data architecture*



Unstructured data lack structure or are not structured like free form text, reports, and customer feedback forms. Big data is the combination of all the three types of data. It has to face three important challenges (B. Gerhardt et al., 2012):

- **Volume:** The volume of data is very large and cannot be processed on a single system. Its size may be in Terabytes, Petabytes and so on.
- **Velocity:** We need to fetch and process that data again and again. So we need to access it several times. So velocity is the speed to fetch data stored on particular node and the speed of the data coming in from various sources.
- **Variety:** It consists of structured, unstructured and semi-structured data. Hence managing different types of data is the main challenge.

In addition to the 3 V's there are some other challenges of big data that are presented below:

- **Veracity:** It is the quality of captured data, which can change dynamically. Veracity of data affects the accuracy of data analysis results.
- **Value:** It is the knowledge that can be extracted from huge amount of data by performing data analysis. This value is very important aspect of data from business point of view.

Advancement in digital devices and sensors, communication, calculation and storage has generated vast assemblage of data. Collecting and managing data for industries and enterprises, scientific research, government and civilization was the main challenge. Since Big data is a modern forthcoming technology in new era which can bring great profit to business enterprises and also a lot of issues like data management and sharing, understanding the data, addressing data quality and handling errors, displaying meaningful result and the most important are privacy, security and trust of data. The appropriate solutions to these challenges are necessary to help the commercial organizations to move towards this technology to increase the value of enterprises, science and society. Big data typically covers huge volume of information related to personal identities and therefore security and privacy of the user becomes a big concern. Without the correct security in right place, Big data can be at high risk. Business organizations must have to confirm that they have the perfect balance between utility and privacy of data, when they are gathering it. Before the data is stored, it should be made appropriately anonymous, by removing any identity related information of any specific user. This in itself may be a security challenge as eliminating identity related information might not be sufficient

to ensure that the data will remain unidentified. There are two main things which should be considered from a security point of view, the one is securing enterprises and its employee's information in a Big data perspective and the other is using Big data technology to analyse, and even predict security incidents.

## 2. BIG DATA TECHNOLOGIES

For processing the huge amount of data from different sources, a new framework was introduced named Hadoop. In fact, the idea of Hadoop came from the concept published by Google. A paper on MapReduce was published in 2004 by Google. They solved this problem by providing parallel processing mode. MapReduce is divided into two steps: Map step and Reduce step. In Map step, queries are split and spread across parallel nodes and follows parallel processing. In Reduce step, the results of Map step are collected and transmitted. This framework became very successful. Therefore, Apache Software Foundation adopted this framework and published the open-source implementation which is known as Hadoop. Hadoop framework is a collection of many tools such as HBase, Hive, Mahout, AVRO, Flume, Spark, Pig, Zookeeper, SQOOP, Couchdb, Storm, Ambari, Oozie etc. The description of some of the tools are as follows:

- **HBase:** It is a database management system implemented in java in which data is stored in column oriented form. It is not a relational database. It is well suited for sparse datasets and provides linear and modular scalability.
- **Hive:** It provides the facility of ad hoc querying, summarization and managing large datasets residing in distributed storage on top of Hadoop.
- **Pig:** Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs and it is compliant to parallel processing.
- **Sqoop:** It is used to import/export data between Hadoop and relational database. This tool helps users to import large data into existing relational database within Hadoop clusters. It is also useful in controlling the import process and parallelism.
- **Zookeeper:** It is a high-performance centralized service for configuring distributed synchronization, naming and providing group services.
- **Mahout:** It is a machine learning library to bring implementation of large number of algorithms for data analysis, clustering, classification and collaborative filtering in Hadoop clusters.

- **Avro:** It is a serialization system, which provides a suitable way to present complex data structure within Hadoop. It is flexible and adaptable with XML configuration.
- **Oozie:** It is a workflow manager based on time and data triggers which combines multiple jobs sequentially into a logical unit of work.
- **Flume:** It is a distributed, reliable and available service for collecting, aggregating and moving huge number of log files. It is robust and fault tolerant because of having simple and flexible architecture which is based on streaming data flows.
- **Spark:** It is an open source data analytics cluster computing framework to run algorithms even faster on Hadoop data.
- **Ambari:** It is a software package to make Hadoop management simpler for managing, provisioning and monitoring Hadoop clusters.

Basically, Hadoop framework is prepared by combining two technologies namely Hadoop Distributed File System(HDFS) and Map Reduce. The clusters of Hadoop follow the master slave architecture. HDFS is a file system which is responsible for the storage of data in distributed manner. HDFS improves the availability of data by replicating the blocks of data across multiple nodes to solve the issue of node failure. MapReduce is powerful technology for parallel processing of large volume of distributed data.

Originally Hadoop was developed without keeping security in mind, so anybody could submit arbitrary code to be executed. After sometime, authentication and authorization was added by including Kerberos to Hadoop, but that was having some flaws. Now, Hadoop community support some security features through the current version of Hadoop and Kerberos implementation with use of firewalls, basic HDFS permissions and ACLs.

## 3. WORKING OF HADOOP FRAMEWORK

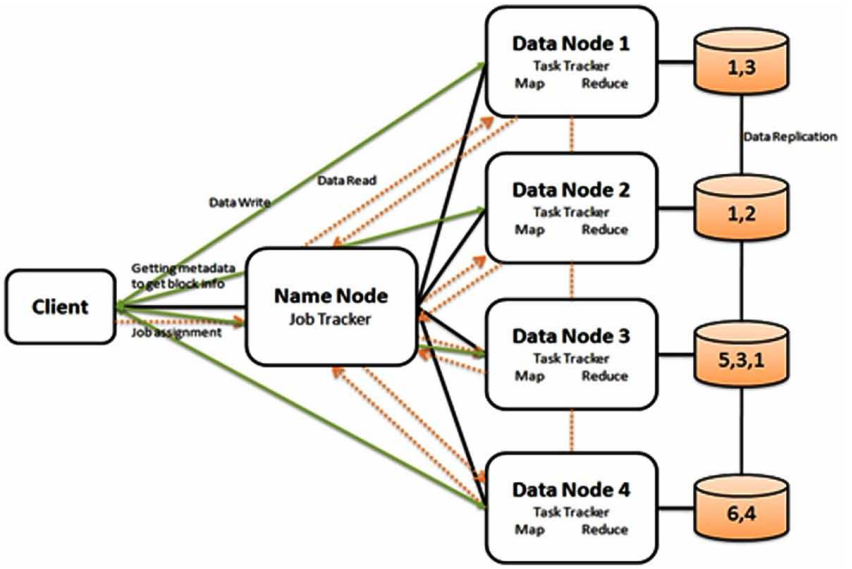
Hadoop framework follows the Master-Slave architecture. With the help of Hadoop, large volume of data can be stored and processed in parallel manner. HDFS allows the fast data transmission rates and also allows the system to complete its all operation even in case of any node failures. It combines the file systems on local nodes to create a single huge virtual file system. For Processing, Hadoop uses MapReduce. Initially, data is divided into chunks which are then parallel processed by Map step. The results of the maps are then sorted and given as input to the Reduce step.

Normally, both input and output are stored in the file system. This framework also takes care of scheduling of multiple jobs, monitoring of currently running tasks and re-execution of failed tasks.

Figure 2 shows the framework of Hadoop. In HDFS data is divided into blocks and then these blocks are stored separately. When client wants to store the file in HDFS then first it sends request to the namenode. Namenode will then send with the reply that how and where to store the blocks of this file. Client then divide the file into blocks of fixed size(128 MB default size) and then store those on slaves accordingly. When the block is sent to the slave then slave makes multiple (by default 3) replicas of that block to achieve availability. After storing the replicas on different slaves it will notify the namenode about the storage and then namenode updates the metadata. In this way HDFS stores the files in distributed manner.

However Hadoop framework was considered to work in a precise isolated environment. But moving it to the public cloud, many challenges arose related to the security mechanisms of Hadoop. Traditional security mechanisms, which were made for securing small static data, are inadequate. Nowadays most of the technologies are being dependent on Big data and therefore, security of Big data is an important issue.

*Figure 2. Hadoop framework*



## 4. PAST RESEARCHES ON BIG DATA SECURITY

In Big data, everyone can provide different and multiple security issues according to their opinion and the environment. However main security issues are similar to other domains that is Authentication, Authorization, Confidentiality (data), Integrity and Accounting. In addition, verifying the integrity of the data before using it for taking decision, preserving privacy of the data while processing and confidentiality of the data while transmission are other common security issues to be considered in Big data. Even though security issues in the Big data are common with some other domains, this section gives a look on security problems that have been found so far in the Big data environment by various research.

M. R. Jam et. al. (2014) discussed about the Security mechanisms of Hadoop. They have discussed about the layers of defence for Hadoop nodes. These layers are Perimeter Level Security, Authentication, Authorization, OS security and the data protection. There is a very strong security at the file system level in Hadoop, but the granular support is necessary to entirely secure access to data by users and Organization Intelligence applications, which is absent in it.

P. Adluru et. al. (2015) have proposed their own security mechanism for Hadoop. The user must be authenticated for access of distributed data and its storage on the cloud to ensure security and privacy of Big data and with the implementation of secure random encryption mechanism a more secure infrastructure can be formed. Their main goal was to present a Hadoop framework which sustains the privacy and security of the data stored on the cloud environment.

M. Kaushik and A. Jain (2014) have discussed the challenges of security and privacy in Big data. They have classified the big data security challenges into four categories such as Infrastructure Security for secure data in distributed programming environment, Data Privacy preservation in analytics and data mining with granular access control, Data Management to secure the stored data and log files with granular auditing and Integrity and Reactive Security for input validation at end-point with real-time security observations. Depending on the data and security policies of the organization, the security is applied to different security levels of data.

A. Katal et. al. (2013) have discussed about the issues, challenges and good practices of big data security. They have discussed challenges such as Privacy of big data, data access and sharing of data in secure manner along with storage and processing issues. They have also mentioned some technical challenges such as fault tolerance, scalability, maintaining quality of data and processing heterogeneous data.

Cloud Computing is used to store and process big data so they have categorized the big data security in four different categories which are network level, user authentication level, data level, and generic issues (V. Narasimha et. al., 2014). Further issues and challenges in the environment are Internode Communication,

Data Protection, Administrative rights for the clusters, Authentication of applications and nodes, and log maintenance. As a solution of these issues, Narasimha et. al. (2014) proposed layered framework for assuring cloud. They have suggested file and network encryption, logging, introduction of honeypot nodes, access control and third party secure data publication for cloud.

## 5. SECURITY ISSUES AND CHALLENGES RELATED TO BIG DATA

Initially, Hadoop framework was developed without keeping security in mind. But later it is observed that without security, data is at high risk. The trustworthiness of data sources is necessary to be proved. Technologies must be explored for identifying security threats in Big data. There have been a lot of issues and challenges (Jaseena et. al. 2014; CSA, 2012) related to the security in the Big data and Hadoop framework which are as follows:

- Unauthorised access of information in Hadoop environment.
- Execution of arbitrary code by attackers.
- Illegal read/write of a data block of any file in Hadoop.
- Eavesdropping or sniffing to data packets being sent to users.
- Unauthorised access of privileges to change priority of job, delete, modify or submission of any malicious job to a queue.

There are some conceptual points which focus on the security requirements of Big data. These points must be understood by the enterprises and organizations to implement some efficient technologies for creating a secure Big data environment.

- Big data must not compromise its essential characteristics i.e. volume, velocity, variety.
- It should not compromise the elementary technology and functionality of the Hadoop clusters.
- It must alert about security threat to Hadoop environment or data stored within the blocks.
- It should address the three kinds of security violations: unauthorised release of information, variation of information and repudiation of resources.
- Secure computation is the main requirement in distributed programming framework for Big data.

- It is important to secure data storage and transaction logs with real-time security monitoring and end-point input validation.
- Automated data transmission needs extra security, which are probably not present in Hadoop framework.
- It is also required to validate the trustworthiness and accuracy of information on receiving large quantity of data by a system.

## 6. SECURITY SOLUTIONS NEEDED FOR BIG DATA

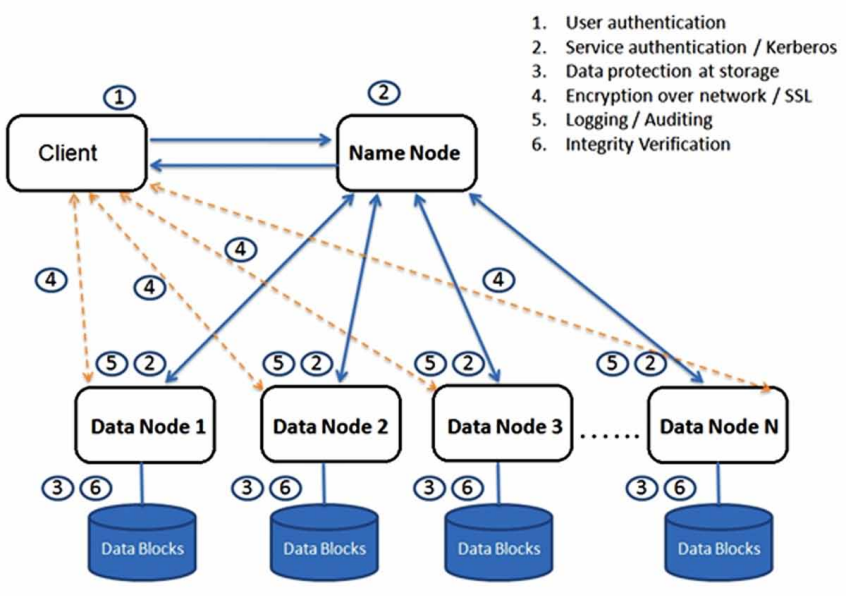
Security is a form of protection that makes separation between resources and the threat. Big data is a new technology for most of the enterprises and organizations. Any type of technology which is not well understood can always introduce new susceptibilities. Major issue involved with Big data is the collection and processing of a great amount of sensitive data from different sources. Hence, data protection is very important. The implementation of Hadoop typically contains open source code, which is potential for default credentials and unrecognized back doors.

If we consider Hadoop framework for Big data, by default security is disable in it. If we need to enable the secure Hadoop framework then we can enable by configuring it. The process of configuration is complex. In the current Hadoop framework, master node has the metadata and slave nodes have the actual data which is fragmented. Even though slave authentication is done but still data is stored in plain text form and slaves have the administrative rights of their own node. This leads to the confidentiality and integrity issue in the Hadoop framework.

- **Confidentiality:** If the slave is compromised or slave itself is malicious then all the data stored in it is at risk. Malicious slave can then easily read that data and then privacy breach will happen. If attacker got the physical access to the slave then he can directly read the data and in this way confidentiality of the data will be lost.
- **Integrity:** Malicious slave can modify the data and then there will be no use of processing it because it will not give the accurate results.

The security requirements related to Big data is mainly divided into five categories. Figure 3 shows the security requirements for the Hadoop framework based on the five categories. Numberings given in the Figure 3 shows the requirement of the respective security component in the entities of the framework.

*Figure 3. Hadoop framework with security requirements*




## a. Authentication

It is the beginning step of security process. It is specific to the security objectives which guarantees that entity are who they claim to be. Access-control, non-repudiation, entity authentication, key authentication, message authentication etc. are the examples of specific objectives. Service authentication in Hadoop defines the services which are needed to preserve the user's choice (wish) when selecting one of many Hadoop services. Figure 3 shows the location where user authentication and service authentication is required in Hadoop framework.

The main purpose of authentication in Hadoop is to verify the identity of users and system. In Hadoop, authentication is provided by Kerberos as in other environment. After getting Ticket Granting Ticket (TGT), the client uses this Ticket to initiate a request for Service Token for accessing Hadoop services from Hadoop Key Distribution Center (H-KDC). This service token is then used for client authentication in Hadoop environment to access Hadoop services.

Authentication process is strong in Kerberos, but it also consist some risks which are dangerous for security. The drawbacks of Kerberos are as follows:

- Kerberos uses third party authentication.
- Requirement of continuous availability of a central server is responsible for single point of failure.
- Strict time requirement for sessions.
- In Kerberos, no authentication is required for requesting ticket, hence attacker can gather equivalent of /etc/passwd by requesting many tickets.
- Other security vulnerabilities related to Kerberos in the different version of the Hadoop is given in the common vulnerabilities database (Apache Hadoop Vulnerabilities, CVSS-CVE). It shows that vulnerabilities lead to authentication, confidentiality and integrity issues in Hadoop.

Problems related to Kerberos should be solved in efficient manner to reduce the security risks in Hadoop. It is also necessary to control user authentication when user approaches from dynamic locations.

## b. Authorization

This is the process of specifying access rights to a user to do or have something and also controlling access on particular. Hadoop framework uses service level authorization. This is primarily managed by file permission in HDFS. It is an initial mechanism to confirm that user connecting to a particular Hadoop services are authorized to access the provided services. It also specifies that what a user can and cannot do within the Hadoop clusters.

Authorization defines the access policies of Hadoop clusters. Kerberos is responsible for providing Service-level authorization in Hadoop because authorization arises in the perspective of authentication. After successful authentication, user may be authorized for accessing different types of services. For enabling it, some configuration has to be done in core Hadoop as well as in Kerberos. It also provides Access Control Lists (ACLs) for every Hadoop services.

## c. Accountability

Accounting or Auditing is used to measure the consumption of resources by user during access. It is also an important key aspect of security in Hadoop framework because of providing report on data (like, from where that data came? what is the use of that data? how it is being used? etc.). The accounting can also include the system time for a particular user during a session. There is a secondary name node,

which is the solution to single point of failure of Hadoop namenode (master node). The task of the secondary namenode is to generate and maintain the log files of the system, so in case of failure of master namenode, we can retrieve the metadata back. The log files collected by this secondary namenode can also be used for auditing the system. Having a robust audit data is always helpful for organizations to check and manage their regulatory compliance with controlling the situation proactively. Figure 3 shows that at every slave node logs must be maintained for auditing purpose.

## d. Data Protection

Data protection is used to defend data at rest as well as in motion. The security of Big data is incomplete without the protection of sensitive data in it. Data protection consist of encryption and key management technique to guarantee that data is fully protected in and around the Hadoop clusters.

Encryption provides an added security layer for the protection of data during its transmission or when it is kept. But, in Hadoop, data protection is lacking because there is absence of encryption at the storage level. Hadoop provides data encryption only on RPC, block data transfer and over HTTP. Hadoop framework is using SSL for secure transfer of data over network but while storing data in HDFS, the data is stored in plain text format only because HDFS doesn't involve encryption engine.

For the total security of Hadoop, end-to-end encryption of data is needed. But its essential requirements are efficiency and availability. Once the data is stored in encrypted form, we have to decrypt it every time before processing. This increases computation overhead. There are encryption techniques available like homomorphism based encryption which can be used to process data without decrypting it. However, there are some limitations to this technique with respect to computational time, etc. Next, efficient key management techniques among the slaves are important in this scenario. Keys must be kept secure and not to be shared with unauthorized users.

## e. Integrity Verification

This is the procedure of manually administrating the verification of information for core consistency. The current version of Hadoop is providing integrity verification at block level using checksum. In Hadoop, to increase the availability of data, every block is replicated multiple times and stored on different nodes. When these blocks are replicated, client must know that the blocks sent by them are stored as it is on the slave node without any modification. To achieve this, every time the block is replicated, the hash is calculated and it is checked with the hash of original block. In this way integrity of the data is verified in Hadoop.

There are many organizations such as Cloudera, Hortonworks, Vormetric etc. are currently working on Hadoop security to protect Big data from malicious activities. They have developed many tools to enhance the security features of Hadoop. Some important projects for Hadoop Security are as follows:

- **Apache Knox Gateway:** Provides authentication and authorization to secure REST API gateway for Hadoop.
- **Apache Sentry:** Provides authorization and access control mechanism for Hadoop data and metadata.
- **Apache Ranger:** Also known as Apache Argus. It is similar to Apache Sentry as it deals with authorization and access control.
- **Cassandra:** It removes the problem of single point of failure in Hadoop by providing high scalability and availability.
- **Project Rhino:** The aim of this project is to add all the security features which are missing in Hadoop such as support for encryption and key management, single sign-on, common authorization, and to improve audit logging.

## 7. SUMMARY

Big data is a huge volume of data, which cannot be stored and processed on a traditional system. To store and process this vast amount of data, a new technology named Hadoop has been introduced. Hadoop is created by the combination of HDFS and MapReduce. It was developed without keeping security in mind. Later, it is observed that Big data needs a very strong security because it contains a large volume of sensitive data from various industries and organizations. Some security features like Kerberos, ACLs has been added to this framework, and still they could not provide an effective security infrastructure for Big data.

This chapter has discussed about security challenges and issues of Big data in Hadoop framework with the possible security solutions, which are needed to secure the Hadoop environment from malicious users and threats. We have gone through five important security requirements of Big data: Authentication, Authorization, Accountability, Data Protection and Integrity Verification. These security requirements must be fulfilled to completely secure the Hadoop infrastructure for Big data.

## 8. RESEARCH SCOPE

Even after providing lot of solution to security issues, Hadoop is still vulnerable to few attacks. It is observed that Hadoop uses Kerberos to provide Authentica-

tion services. Kerberos is the third party based authentication scheme and there is a common Key Distribution Center. Therefore, there are possibilities that it may get compromised under some circumstances. To avoid such risks, Authentication without third party can be introduced in place of Kerberos.

Another risk in existing Hadoop framework is, it does not provide encryption of data at storage level. If the node where data is stored wants to access the data or it gets compromised by the attacker then it can easily read the data, because data is stored in plain text form. Due to this, data stored at that node will get lost or misused. As a solution to this, HDFS must provide encryption engine, which will encrypt the data before storing it on data node and decrypt the data before processing it or process the encrypted data. It may increase the processing time of Big data but it will keep the data secure. Hence there is a need for the light weight cryptographic algorithms and while processing the data, privacy should be preserved.

Security always comes at a cost. Thus, after providing security solutions to the Big data, its efficiency may get affected. The main challenge in securing Big data is to maintaining the efficiency and performance of the system, which ensures security.

## REFERENCES

Adluru, P., Datla, S. S., & Zhang, X. (2015). *Hadoop Eco System for Big Data Security and Privacy*. Long Island Systems, Applications and Technology Conference (LISAT). doi:10.1109/LISAT.2015.7160211

*Apache Hadoop Vulnerabilities*. Retrieved from https://www.cvedetails.com/vulnerability-list/vendor_id-45/product_id-22215/Apache-Hadoop.html

CSA. (2012). *Top Ten Big Data Security and Privacy Challenges*. Retrieved from http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf

Gerhardt, Griffin, & Klemann. (2012). *Unlocking Values in the Fragmented World of Big Data Analytics*. Cisco Internet Business Solutions Group.

Inukollu, V. N., Arsi, S., & Ravuri, S. R. (2014). Security Issues Associated with Big Data in Cloud Computing. *International Journal of Network Security & Its Applications*, *6*(3), 45–56. doi:10.5121/ijnsa.2014.6304

Jam, Khanli, Akbari, & Javan. (2014). *A Survey on Security of Hadoop*. 4th International eConference on Computer and Knowledge Engineering (ICCKE).

Jaseena, K.U., & David, J. M. (2014). Issues, Challenges and Solutions: Big Data Mining. *CS & IT-CSCP, 4*(13), 131-140.

Katal, A., Wazid, M., & Goudar, R. H. (2013). *Big Data: Issues, Challenges, Tools and Good Practices*. Sixth International Conference on Contemporary Computing (IC3). doi:10.1109/IC3.2013.6612229

Kaushik, M., & Jain, A. (2014). Challenges to Big Data Security and Privacy. *International Journal of Computer Science and Information Technologies*, *5*(3), 3042–3043.

Vohries. (2013). *The Big Deal about Big Data: What's Inside – Structured, Unstructured, and Semi-Structured Data.* [Web blog post]. Data Magnum.

## ADDITIONAL READING

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big Data: Related Technologies, Challenges and Future Prospects*. Springer Briefs in Computer Science. doi:10.1007/978-3-319-06245-7

Liu, C., Yang, C., Zhang, X., & Chen, J. (2015, August). External integrity verification for outsourced big data in cloud and IoT: A big picture. *Future Generation Computer Systems*, *49*, 58–67. doi:10.1016/j.future.2014.08.007

## KEY TERMS AND DEFINITIONS

**Big Data:** It is a large volume of data which is beyond the storage capacity and processing power of traditional systems.

**Big Data Security:** It is to achieve the confidentiality, authenticity and integrity of the Big data collected, stored and processed.

**Hadoop:** It is the combination of HDFS and MapReduce.

**Hadoop Distributed File System (HDFS):** It is used to store data in distributed manner.

**MapReduce:** It is a process of splitting and executing queries across multiple nodes and combining the results from individual nodes to generate final output.

# Related References

To continue our tradition of advancing information science and technology research, we have compiled a list of recommended IGI Global readings. These references will provide additional information and guidance to further enrich your knowledge and assist you with your own research and future publications.

Aalmink, J., von der Dovenmühle, T., & Gómez, J. M. (2013). Enterprise tomography: Maintenance and root-cause-analysis of federated erp in enterprise clouds. In P. Ordóñez de Pablos, H. Nigro, R. Tennyson, S. Gonzalez Cisaro, & W. Karwowski (Eds.), *Advancing information management through semantic web concepts and ontologies* (pp. 133–153). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2494-8.ch007

Abu, S. T., & Tsuji, M. (2011). The development of ICT for envisioning cloud computing and innovation in South Asia.[IJIDE]. *International Journal of Innovation in the Digital Economy*, *2*(1), 61–72. doi:10.4018/jide.2011010105

Abu, S. T., & Tsuji, M. (2012). The development of ICT for envisioning cloud computing and innovation in South Asia. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 453–465). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch207

Abu, S. T., & Tsuji, M. (2013). The development of ICT for envisioning cloud computing and innovation in South Asia. In I. Oncioiu (Ed.), *Business innovation, development, and advancement in the digital economy* (pp. 35–47). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2934-9.ch003

Adams, R. (2013). The emergence of cloud storage and the need for a new digital forensic process model. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 79–104). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch004

Adeyeye, M. (2013). Provisioning converged applications and services via the cloud. In D. Kanellopoulos (Ed.), *Intelligent multimedia technologies for networking applications: Techniques and tools* (pp. 248–269). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2833-5.ch010

Aggarwal, A. (2013). A systems approach to cloud computing services. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 124–136). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch006

Ahmed, K., Hussain, A., & Gregory, M. A. (2013). An efficient, robust, and secure SSO architecture for cloud computing implemented in a service oriented architecture. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 259–282). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch011

Ahuja, S. P., & Mani, S. (2013). Empirical performance analysis of HPC benchmarks across variations in cloud computing.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 13–26. doi:10.4018/ijcac.2013010102

Ahuja, S. P., & Rolli, A. C. (2011). Survey of the state-of-the-art of cloud computing.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(4), 34–43. doi:10.4018/ijcac.2011100103

Ahuja, S. P., & Rolli, A. C. (2013). Survey of the state-of-the-art of cloud computing. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 252–262). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch018

Ahuja, S. P., & Sridharan, S. (2012). Performance evaluation of hypervisors for cloud computing.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(3), 26–67. doi:10.4018/ijcac.2012070102

Akyuz, G. A., & Rehan, M. (2013). A generic, cloud-based representation for supply chains (SC's).[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(2), 12–20. doi:10.4018/ijcac.2013040102

Al-Aqrabi, H., & Liu, L. (2013). IT security and governance compliant service oriented computing in cloud computing environments. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 143–163). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch006

Al-Zoube, M., & Wyne, M. F. (2012). Building integrated e-learning environment using cloud services and social networking sites. In Q. Jin (Ed.), *Intelligent learning systems and advancements in computer-aided instruction: Emerging studies* (pp. 214–233). Hershey, PA: Information Science Reference; doi:10.4018/978-1-61350-483-3.ch013

Alam, N., & Karmakar, R. (2014). Cloud computing and its application to information centre. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 63–76). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch004

Alhaj, A., Aljawarneh, S., Masadeh, S., & Abu-Taieh, E. (2013). A secure data transmission mechanism for cloud outsourced data.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 34–43. doi:10.4018/ijcac.2013010104

Alharbi, S. T. (2012). Users' acceptance of cloud computing in Saudi Arabia: An extension of technology acceptance model.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(2), 1–11. doi:10.4018/ijcac.2012040101

Ali, S. S., & Khan, M. N. (2013). ICT infrastructure framework for microfinance institutions and banks in Pakistan: An optimized approach.[IJOM]. *International Journal of Online Marketing*, *3*(2), 75–86. doi:10.4018/ijom.2013040105

Aljawarneh, S. (2011). Cloud security engineering: Avoiding security threats the right way.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(2), 64–70. doi:10.4018/ijcac.2011040105

Aljawarneh, S. (2013). Cloud security engineering: Avoiding security threats the right way. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 147–153). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch010

Alshattnawi, S. (2013). Utilizing cloud computing in developing a mobile location-aware tourist guide system.[IJAPUC]. *International Journal of Advanced Pervasive and Ubiquitous Computing*, *5*(2), 9–18. doi:10.4018/japuc.2013040102

Alsmadi, I. (2013). Software development methodologies for cloud computing. In K. Buragga & N. Zaman (Eds.), *Software development techniques for constructive information systems design* (pp. 110–117). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3679-8.ch006

Anand, V. (2013). Survivable mapping of virtual networks onto a shared substrate network. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 325–343). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch014

Antonova, A. (2013). Green, sustainable, or clean: What type of IT/IS technologies will we need in the future? In P. Ordóñez de Pablos (Ed.), *Green technologies and business practices: An IT approach* (pp. 151–162). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1972-2.ch008

Ardissono, L., Bosio, G., Goy, A., Petrone, G., Segnan, M., & Torretta, F. (2011). Collaboration support for activity management in a personal cloud environment. [IJDST]. *International Journal of Distributed Systems and Technologies*, *2*(4), 30–43. doi:10.4018/jdst.2011100103

Ardissono, L., Bosio, G., Goy, A., Petrone, G., Segnan, M., & Torretta, F. (2013). Collaboration support for activity management in a personal cloud environment. In N. Bessis (Ed.), *Development of distributed systems from design to application and maintenance* (pp. 199–212). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2647-8.ch012

Argiolas, M., Atzori, M., Dessì, N., & Pes, B. (2012). Dataspaces enhancing decision support systems in clouds.[IJWP]. *International Journal of Web Portals*, *4*(2), 35–55. doi:10.4018/jwp.2012040103

Arinze, B., & Anandarajan, M. (2012). Factors that determine the adoption of cloud computing: A global perspective. In M. Tavana (Ed.), *Enterprise Information Systems and Advancing Business Solutions: Emerging Models* (pp. 210–223). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1761-2.ch012

Arinze, B., & Sylla, C. (2012). Conducting research in the cloud. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 50–63). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch004

Arshad, J., Townend, P., & Xu, J. (2011). An abstract model for integrated intrusion detection and severity analysis for clouds.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(1), 1–16. doi:10.4018/ijcac.2011010101

Arshad, J., Townend, P., & Xu, J. (2013). An abstract model for integrated intrusion detection and severity analysis for clouds. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 1–17). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch001

Arshad, J., Townend, P., Xu, J., & Jie, W. (2012). Cloud computing security: Opportunities and pitfalls.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(1), 52–66. doi:10.4018/jghpc.2012010104

Baars, T., & Spruit, M. (2012). Designing a secure cloud architecture: The SeCA model.[IJISP]. *International Journal of Information Security and Privacy*, *6*(1), 14–32. doi:10.4018/jisp.2012010102

Bai, X., Gao, J. Z., & Tsai, W. (2013). Cloud scalability measurement and testing. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 356–381). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch017

Baldini, G., & Stirparo, P. (2014). A cognitive access framework for security and privacy protection in mobile cloud computing. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 92–117). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch006

Balduf, S., Balke, T., & Eymann, T. (2012). Cultural differences in managing cloud computing service level agreements. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1237–1263). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch512

Banerjee, S., Sing, T. Y., Chowdhury, A. R., & Anwar, H. (2013). Motivations to adopt green ICT: A tale of two organizations.[IJGC]. *International Journal of Green Computing*, *4*(2), 1–11. doi:10.4018/jgc.2013070101

Barreto, J., Di Sanzo, P., Palmieri, R., & Romano, P. (2013). Cloud-TM: An elastic, self-tuning transactional store for the cloud. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 192–224). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch013

Belalem, G., & Limam, S. (2011). Fault tolerant architecture to cloud computing using adaptive checkpoint.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(4), 60–69. doi:10.4018/ijcac.2011100105

Belalem, G., & Limam, S. (2013). Fault tolerant architecture to cloud computing using adaptive checkpoint. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 280–289). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch020

Ben Belgacem, M., Abdennadher, N., & Niinimaki, M. (2012). Virtual EZ grid: A volunteer computing infrastructure for scientific medical applications.[IJHCR]. *International Journal of Handheld Computing Research*, *3*(1), 74–85. doi:10.4018/jhcr.2012010105

Bhatt, S., Chaudhary, S., & Bhise, M. (2013). Migration of data between cloud and non-cloud datastores. In A. Ionita, M. Litoiu, & G. Lewis (Eds.), *Migrating legacy applications: Challenges in service oriented architecture and cloud computing environments* (pp. 206–225). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2488-7.ch009

Biancofiore, G., & Leone, S. (2014). Google apps as a cloud computing solution in Italian municipalities: Technological features and implications. In S. Leone (Ed.), *Synergic integration of formal and informal e-learning environments for adult lifelong learners* (pp. 244–274). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4655-1.ch012

Bibi, S., Katsaros, D., & Bozanis, P. (2012). How to choose the right cloud. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1530–1552). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch701

Bibi, S., Katsaros, D., & Bozanis, P. (2012). How to choose the right cloud. In X. Liu & Y. Li (Eds.), *Advanced design approaches to emerging software systems: Principles, methodologies and tools* (pp. 219–240). Hershey, PA: Information Science Reference; doi:10.4018/978-1-60960-735-7.ch010

Bitam, S., Batouche, M., & Talbi, E. (2012). A bees life algorithm for cloud computing services selection. In S. Ali, N. Abbadeni, & M. Batouche (Eds.), *Multidisciplinary computational intelligence techniques: Applications in business, engineering, and medicine* (pp. 31–46). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1830-5.ch003

Bittencourt, L. F., Madeira, E. R., & da Fonseca, N. L. (2014). Communication aspects of resource management in hybrid clouds. In H. Mouftah & B. Kantarci (Eds.), *Communication infrastructures for cloud computing* (pp. 409–433). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4522-6.ch018

Bonelli, L., Giudicianni, L., Immediata, A., & Luzzi, A. (2013). Compliance in the cloud. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 109–131). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch008

Boniface, M., Nasser, B., Surridge, M., & Oliveros, E. (2012). Securing real-time interactive applications in federated clouds. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1822–1835). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch806

Boukhobza, J. (2013). Flashing in the cloud: Shedding some light on NAND flash memory storage systems. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 241–266). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch015

Bracci, F., Corradi, A., & Foschini, L. (2014). Cloud standards: Security and interoperability issues. In H. Mouftah & B. Kantarci (Eds.), *Communication infrastructures for cloud computing* (pp. 465–495). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4522-6.ch020

Brown, A. W. (2013). Experiences with cloud technology to realize software testing factories. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 1–27). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch001

Calcavecchia, N. M., Celesti, A., & Di Nitto, E. (2012). Understanding decentralized and dynamic brokerage in federated cloud environments. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 36–56). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch003

Calero, J. M., König, B., & Kirschnick, J. (2012). Cross-layer monitoring in cloud computing. In H. Rashvand & Y. Kavian (Eds.), *Using cross-layer techniques for communication systems* (pp. 328–348). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0960-0.ch014

Cardellini, V., Casalicchio, E., & Silvestri, L. (2012). Service level provisioning for cloud-based applications service level provisioning for cloud-based applications. In A. Pathan, M. Pathan, & H. Lee (Eds.), *Advancements in distributed computing and internet technologies: Trends and issues* (pp. 363–385). Hershey, PA: Information Science Publishing; doi:10.4018/978-1-61350-110-8.ch017

Cardellini, V., Casalicchio, E., & Silvestri, L. (2012). Service level provisioning for cloud-based applications service level provisioning for cloud-based applications. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1479–1500). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch611

Carlin, S., & Curran, K. (2013). Cloud computing security. In K. Curran (Ed.), *Pervasive and ubiquitous technology innovations for ambient intelligence environments* (pp. 12–17). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2041-4.ch002

Carlton, G. H., & Zhou, H. (2011). A survey of cloud computing challenges from a digital forensics perspective.[IJITN]. *International Journal of Interdisciplinary Telecommunications and Networking*, *3*(4), 1–16. doi:10.4018/jitn.2011100101

Carlton, G. H., & Zhou, H. (2012). A survey of cloud computing challenges from a digital forensics perspective. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1221–1236). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch511

Carlton, G. H., & Zhou, H. (2013). A survey of cloud computing challenges from a digital forensics perspective. In M. Bartolacci & S. Powell (Eds.), *Advancements and innovations in wireless communications and network technologies* (pp. 213–228). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2154-1.ch016

Carpen-Amarie, A., Costan, A., Leordeanu, C., Basescu, C., & Antoniu, G. (2012). Towards a generic security framework for cloud data management environments. [IJDST]. *International Journal of Distributed Systems and Technologies*, *3*(1), 17–34. doi:10.4018/jdst.2012010102

Casola, V., Cuomo, A., Villano, U., & Rak, M. (2012). Access control in federated clouds: The cloudgrid case study. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving Federated and Self-Manageable Cloud Infrastructures: Theory and Practice* (pp. 395–417). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch020

Casola, V., Cuomo, A., Villano, U., & Rak, M. (2013). Access control in federated clouds: The cloudgrid case study. In *IT policy and ethics: Concepts, methodologies, tools, and applications* (pp. 148–169). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2919-6.ch008

Celesti, A., Tusa, F., & Villari, M. (2012). Toward cloud federation: Concepts and challenges. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 1–17). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch001

Chaka, C. (2013). Virtualization and cloud computing: Business models in the virtual cloud. In A. Loo (Ed.), *Distributed computing innovations for business, engineering, and science* (pp. 176–190). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2533-4.ch009

Chang, J. (2011). A framework for analysing the impact of cloud computing on local government in the UK.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(4), 25–33. doi:10.4018/ijcac.2011100102

Chang, J. (2013). A framework for analysing the impact of cloud computing on local government in the UK. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 243–251). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch017

Chang, J., & Johnston, M. (2012). Cloud computing in local government: From the perspective of four London borough councils.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(4), 1–15. doi:10.4018/ijcac.2012100101

Chang, K., & Wang, K. (2012). Efficient support of streaming videos through patching proxies in the cloud.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(4), 22–36. doi:10.4018/jghpc.2012100102

Chang, R., Liao, C., & Liu, C. (2013). Choosing clouds for an enterprise: Modeling and evaluation.[IJEEI]. *International Journal of E-Entrepreneurship and Innovation*, *4*(2), 38–53. doi:10.4018/ijeei.2013040103

Chang, V., De Roure, D., Wills, G., & Walters, R. J. (2011). Case studies and organisational sustainability modelling presented by cloud computing business framework.[IJWSR]. *International Journal of Web Services Research*, *8*(3), 26–53. doi:10.4018/JWSR.2011070102

Chang, V., Li, C., De Roure, D., Wills, G., Walters, R. J., & Chee, C. (2011). The financial clouds review.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(2), 41–63. doi:10.4018/ijcac.2011040104

Chang, V., Li, C., De Roure, D., Wills, G., Walters, R. J., & Chee, C. (2013). The financial clouds review. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 125–146). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch009

Chang, V., Walters, R. J., & Wills, G. (2012). Business integration as a service. [IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(1), 16–40. doi:10.4018/ijcac.2012010102

Chang, V., & Wills, G. (2013). A University of Greenwich case study of cloud computing: Education as a service. In D. Graham, I. Manikas, & D. Folinas (Eds.), *E-logistics and e-supply chain management: Applications for evolving business* (pp. 232–253). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3914-0.ch013

Chang, V., Wills, G., Walters, R. J., & Currie, W. (2012). Towards a structured cloud ROI: The University of Southampton cost-saving and user satisfaction case studies. In W. Hu & N. Kaabouch (Eds.), *Sustainable ICTs and management systems for green computing* (pp. 179–200). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1839-8.ch008

Chang, Y., Lee, Y., Juang, T., & Yen, J. (2013). Cost evaluation on building and operating cloud platform.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *5*(2), 43–53. doi:10.4018/jghpc.2013040103

Chao, L. (2012). Cloud computing solution for internet based teaching and learning. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 210–235). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch015

Chao, L. (2012). Overview of cloud computing and its application in e-learning. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 1–16). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch001

Chauhan, S., Raman, A., & Singh, N. (2013). A comparative cost analysis of on premises IT infrastructure and cloud-based email services in an Indian business school.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(2), 21–34. doi:10.4018/ijcac.2013040103

Chen, C., Chao, H., Wu, T., Fan, C., Chen, J., Chen, Y., & Hsu, J. (2011). IoT-IMS communication platform for future internet.[IJARAS]. *International Journal of Adaptive, Resilient and Autonomic Systems*, *2*(4), 74–94. doi:10.4018/jaras.2011100105

Chen, C., Chao, H., Wu, T., Fan, C., Chen, J., Chen, Y., & Hsu, J. (2013). IoT-IMS communication platform for future internet. In V. De Florio (Ed.), *Innovations and approaches for resilient and adaptive systems* (pp. 68–86). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2056-8.ch004

Chen, C. C. (2013). Cloud computing in case-based pedagogy: An information systems success perspective.[IJDTIS]. *International Journal of Dependable and Trustworthy Information Systems*, *2*(3), 1–16. doi:10.4018/jdtis.2011070101

Cheney, A. W., Riedl, R. E., Sanders, R., & Tashner, J. H. (2012). The new company water cooler: Use of 3D virtual immersive worlds to promote networking and professional learning in organizations. In Organizational learning and knowledge: Concepts, methodologies, tools and applications (pp. 2848-2861). Hershey, PA: Business Science Reference. doi:10.4018/978-1-60960-783-8.ch801

Chiang, C., & Yu, S. (2013). Cloud-enabled software testing based on program understanding. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 54–67). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch003

Chou, Y., & Oetting, J. (2011). Risk assessment for cloud-based IT systems. [IJGHPC]. *International Journal of Grid and High Performance Computing*, *3*(2), 1–13. doi:10.4018/jghpc.2011040101

Chou, Y., & Oetting, J. (2012). Risk assessment for cloud-based IT systems. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 272–285). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch113

Chou, Y., & Oetting, J. (2013). Risk assessment for cloud-based IT systems. In E. Udoh (Ed.), *Applications and developments in grid, cloud, and high performance computing* (pp. 1–14). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2065-0.ch001

Cohen, F. (2013). Challenges to digital forensic evidence in the cloud. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 59–78). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch003

Cossu, R., Di Giulio, C., Brito, F., & Petcu, D. (2013). Cloud computing for earth observation. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 166–191). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch012

Costa, J. E., & Rodrigues, J. J. (2014). Mobile cloud computing: Technologies, services, and applications. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 1–17). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch001

Creaner, G., & Pahl, C. (2013). Flexible coordination techniques for dynamic cloud service collaboration. In G. Ortiz & J. Cubo (Eds.), *Adaptive web services for modular and reusable software development: Tactics and solutions* (pp. 239–252). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2089-6.ch009

Crosbie, M. (2013). Hack the cloud: Ethical hacking and cloud forensics. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 42–58). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch002

Curran, K., Carlin, S., & Adams, M. (2012). Security issues in cloud computing. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 200–208). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch014

Dahbur, K., & Mohammad, B. (2011). Toward understanding the challenges and countermeasures in computer anti-forensics.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(3), 22–35. doi:10.4018/ijcac.2011070103

Dahbur, K., Mohammad, B., & Tarakji, A. B. (2011). Security issues in cloud computing: A survey of risks, threats and vulnerabilities.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(3), 1–11. doi:10.4018/ijcac.2011070101

Dahbur, K., Mohammad, B., & Tarakji, A. B. (2012). Security issues in cloud computing: A survey of risks, threats and vulnerabilities. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1644–1655). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch707

Dahbur, K., Mohammad, B., & Tarakji, A. B. (2013). Security issues in cloud computing: A survey of risks, threats and vulnerabilities. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 154–165). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch011

Daim, T., Britton, M., Subramanian, G., Brenden, R., & Intarode, N. (2012). Adopting and integrating cloud computing. In E. Eyob & E. Tetteh (Eds.), *Customer-oriented global supply chains: Concepts for effective management* (pp. 175–197). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0246-5.ch011

Davis, M., & Sedsman, A. (2012). Grey areas: The legal dimensions of cloud computing. In C. Li & A. Ho (Eds.), *Crime prevention technologies and applications for advancing criminal investigation* (pp. 263–273). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1758-2.ch017

De Coster, R., & Albesher, A. (2013). The development of mobile service applications for consumers and intelligent networks. In I. Lee (Ed.), *Mobile services industries, technologies, and applications in the global economy* (pp. 273–289). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1981-4.ch017

De Filippi, P. (2014). Ubiquitous computing in the cloud: User empowerment vs. user obsequity. In J. Pelet & P. Papadopoulou (Eds.), *User behavior in ubiquitous online environments* (pp. 44–63). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4566-0.ch003

De Silva, S. (2013). Key legal issues with cloud computing: A UK law perspective. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 242–256). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch013

Deed, C., & Cragg, P. (2013). Business impacts of cloud computing. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 274–288). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch015

Deng, M., Petkovic, M., Nalin, M., & Baroni, I. (2013). Home healthcare in cloud computing. In M. Cruz-Cunha, I. Miranda, & P. Gonçalves (Eds.), *Handbook of research on ICTs and management systems for improving efficiency in healthcare and social care* (pp. 614–634). Hershey, PA: Medical Information Science Reference; doi:10.4018/978-1-4666-3990-4.ch032

Desai, A. M., & Mock, K. (2013). Security in cloud computing. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 208–221). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch011

Deshpande, R. M., Patle, B. V., & Bhoskar, R. D. (2014). Planning and implementation of cloud computing in NIT's in India: Special reference to VNIT. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 90–106). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch006

Dhamdhere, S. N., & Lihitkar, R. (2014). The university cloud library model and the role of the cloud librarian. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 150–161). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch009

Di Martino, S., Ferrucci, F., Maggio, V., & Sarro, F. (2013). Towards migrating genetic algorithms for test data generation to the cloud. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 113–135). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch006

Di Sano, M., Di Stefano, A., Morana, G., & Zito, D. (2013). FSaaS: Configuring policies for managing shared files among cooperating, distributed applications.[IJWP]. *International Journal of Web Portals*, *5*(1), 1–14. doi:10.4018/jwp.2013010101

Dippl, S., Jaeger, M. C., Luhn, A., Shulman-Peleg, A., & Vernik, G. (2013). Towards federation and interoperability of cloud storage systems. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 60–71). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch005

Distefano, S., & Puliafito, A. (2012). The cloud@home volunteer and interoperable cloud through the future internet. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 79–96). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch005

Djoleto, W. (2013). Cloud computing and ecommerce or ebusiness: "The now it way" – An overview. In *Electronic commerce and organizational leadership: perspectives and methodologies* (pp. 239–254). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2982-0.ch010

Dollmann, T. J., Loos, P., Fellmann, M., Thomas, O., Hoheisel, A., Katranuschkov, P., & Scherer, R. (2011). Design and usage of a process-centric collaboration methodology for virtual organizations in hybrid environments.[IJIIT]. *International Journal of Intelligent Information Technologies*, *7*(1), 45–64. doi:10.4018/jiit.2011010104

Dollmann, T. J., Loos, P., Fellmann, M., Thomas, O., Hoheisel, A., Katranuschkov, P., & Scherer, R. (2013). Design and usage of a process-centric collaboration methodology for virtual organizations in hybrid environments. In V. Sugumaran (Ed.), *Organizational efficiency through intelligent information technologies* (pp. 45–64). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2047-6.ch004

Dreher, P., & Vouk, M. (2012). Utilizing open source cloud computing environments to provide cost effective support for university education and research. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 32–49). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch003

Drum, D., Becker, D., & Fish, M. (2013). Technology adoption in troubled times: A cloud computing case study.[JCIT]. *Journal of Cases on Information Technology*, *15*(2), 57–71. doi:10.4018/jcit.2013040104

Dunaway, D. M. (2013). Creating virtual collaborative learning experiences for aspiring teachers. In R. Hartshorne, T. Heafner, & T. Petty (Eds.), *Teacher education programs and online learning tools: Innovations in teacher preparation* (pp. 167–180). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1906-7.ch009

Dykstra, J. (2013). Seizing electronic evidence from cloud computing environments. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 156–185). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch007

El-Refaey, M., & Rimal, B. P. (2012). Grid, SOA and cloud computing: On-demand computing models. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 12–51). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch102

El-Refaey, M., & Rimal, B. P. (2012). Grid, SOA and cloud computing: On-demand computing models. In N. Preve (Ed.), *Computational and data grids: Principles, applications and design* (pp. 45–85). Hershey, PA: Information Science Reference; doi:10.4018/978-1-61350-113-9.ch003

Elnaffar, S., Maamar, Z., & Sheng, Q. Z. (2013). When clouds start socializing: The sky model.[IJEBR]. *International Journal of E-Business Research*, *9*(2), 1–7. doi:10.4018/jebr.2013040101

Elwood, S., & Keengwe, J. (2012). Microbursts: A design format for mobile cloud computing.[IJICTE]. *International Journal of Information and Communication Technology Education*, *8*(2), 102–110. doi:10.4018/jicte.2012040109

Emeakaroha, V. C., Netto, M. A., Calheiros, R. N., & De Rose, C. A. (2012). Achieving flexible SLA and resource management in clouds. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 266–287). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch014

Etro, F. (2013). The economics of cloud computing. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 296–309). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch017

Ezugwu, A. E., Buhari, S. M., & Junaidu, S. B. (2013). Virtual machine allocation in cloud computing environment.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(2), 47–60. doi:10.4018/ijcac.2013040105

Fauzi, A. H., & Taylor, H. (2013). Secure community trust stores for peer-to-peer e-commerce applications using cloud services.[IJEEI]. *International Journal of E-Entrepreneurship and Innovation*, *4*(1), 1–15. doi:10.4018/jeei.2013010101

Ferguson-Boucher, K., & Endicott-Popovsky, B. (2013). Forensic readiness in the cloud (FRC): Integrating records management and digital forensics. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 105–128). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch005

Ferraro de Souza, R., Westphall, C. B., dos Santos, D. R., & Westphall, C. M. (2013). A review of PACS on cloud for archiving secure medical images.[IJPHIM]. *International Journal of Privacy and Health Information Management*, *1*(1), 53–62. doi:10.4018/ijphim.2013010104

Firdhous, M., Hassan, S., & Ghazali, O. (2013). Statistically enhanced multi-dimensional trust computing mechanism for cloud computing.[IJMCMC]. *International Journal of Mobile Computing and Multimedia Communications*, *5*(2), 1–17. doi:10.4018/jmcmc.2013040101

Formisano, C., Bonelli, L., Balraj, K. R., & Shulman-Peleg, A. (2013). Cloud access control mechanisms. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 94–108). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch007

Frank, H., & Mesentean, S. (2012). Efficient communication interfaces for distributed energy resources. In E. Udoh (Ed.), *Evolving developments in grid and cloud computing: Advancing research* (pp. 185–196). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0056-0.ch013

Gallina, B., & Guelfi, N. (2012). Reusing transaction models for dependable cloud computing. In H. Yang & X. Liu (Eds.), *Software reuse in the emerging cloud computing era* (pp. 248–277). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0897-9.ch011

Garofalo, D. A. (2013). Empires of the future: Libraries, technology, and the academic environment. In E. Iglesias (Ed.), *Robots in academic libraries: Advancements in library automation* (pp. 180–206). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3938-6.ch010

Gebremeskel, G. B., He, Z., & Jing, X. (2013). Semantic integrating for intelligent cloud data mining platform and cloud based business intelligence for optimization of mobile social networks. In V. Bhatnagar (Ed.), *Data mining in dynamic social networks and fuzzy systems* (pp. 173–211). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4213-3.ch009

Gentleman, W. M. (2013). Using the cloud for testing NOT adjunct to development. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 216–230). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch010

Ghafoor, K. Z., Mohammed, M. A., Abu Bakar, K., Sadiq, A. S., & Lloret, J. (2014). Vehicular cloud computing: Trends and challenges. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 262–274). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch014

Giannakaki, M. (2012). The "right to be forgotten" in the era of social media and cloud computing. In C. Akrivopoulou & N. Garipidis (Eds.), *Human rights and risks in the digital era: Globalization and the effects of information technologies* (pp. 10–24). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0891-7.ch002

Gillam, L., Li, B., & O'Loughlin, J. (2012). Teaching clouds: Lessons taught and lessons learnt. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 82–94). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch006

Gonsowski, D. (2013). Compliance in the cloud and the implications on electronic discovery. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 230–250). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch009

Gonzalez-Sanchez, J., Conley, Q., Chavez-Echeagaray, M., & Atkinson, R. K. (2012). Supporting the assembly process by leveraging augmented reality, cloud computing, and mobile devices.[IJCBPL]. *International Journal of Cyber Behavior, Psychology and Learning*, *2*(3), 86–102. doi:10.4018/ijcbpl.2012070107

Gopinath, R., & Geetha, B. (2013). An e-learning system based on secure data storage services in cloud computing.[IJITWE]. *International Journal of Information Technology and Web Engineering*, *8*(2), 1–17. doi:10.4018/jitwe.2013040101

Gossin, P. C., & LaBrie, R. C. (2013). Data center waste management. In P. Ordóñez de Pablos (Ed.), *Green technologies and business practices: An IT approach* (pp. 226–235). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1972-2.ch014

Goswami, V., Patra, S. S., & Mund, G. B. (2012). Performance analysis of cloud computing centers for bulk services.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(4), 53–65. doi:10.4018/ijcac.2012100104

Goswami, V., & Sahoo, C. N. (2013). Optimal resource usage in multi-cloud computing environment.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 44–57. doi:10.4018/ijcac.2013010105

Gräuler, M., Teuteberg, F., Mahmoud, T., & Gómez, J. M. (2013). Requirements prioritization and design considerations for the next generation of corporate environmental management information systems: A foundation for innovation.[IJITSA]. *International Journal of Information Technologies and Systems Approach*, *6*(1), 98–116. doi:10.4018/jitsa.2013010106

Grieve, G. P., & Heston, K. (2012). Finding liquid salvation: Using the cardean ethnographic method to document second life residents and religious cloud communities. In N. Zagalo, L. Morgado, & A. Boa-Ventura (Eds.), *Virtual worlds and metaverse platforms: New communication and identity paradigms* (pp. 288–305). Hershey, PA: Information Science Reference; doi:10.4018/978-1-60960-854-5.ch019

Grispos, G., Storer, T., & Glisson, W. B. (2012). Calm before the storm: The challenges of cloud computing in digital forensics.[IJDCF]. *International Journal of Digital Crime and Forensics*, *4*(2), 28–48. doi:10.4018/jdcf.2012040103

Grispos, G., Storer, T., & Glisson, W. B. (2013). Calm before the storm: The challenges of cloud computing in digital forensics. In C. Li (Ed.), *Emerging digital forensics applications for crime detection, prevention, and security* (pp. 211–233). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4006-1.ch015

Guster, D., & Lee, O. F. (2011). Enhancing the disaster recovery plan through virtualization.[JITR]. *Journal of Information Technology Research*, *4*(4), 18–40. doi:10.4018/jitr.2011100102

Hanawa, T., & Sato, M. (2013). D-Cloud: Software testing environment for dependable distributed systems using cloud computing technology. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 340–355). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch016

Hardy, J., Liu, L., Lei, C., & Li, J. (2013). Internet-based virtual computing infrastructure for cloud computing. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 371–389). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch016

Hashizume, K., Yoshioka, N., & Fernandez, E. B. (2013). Three misuse patterns for cloud computing. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 36–53). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch003

Hassan, Q. F., Riad, A. M., & Hassan, A. E. (2012). Understanding cloud computing. In H. Yang & X. Liu (Eds.), *Software reuse in the emerging cloud computing era* (pp. 204–227). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0897-9.ch009

Hasselmeyer, P., Katsaros, G., Koller, B., & Wieder, P. (2012). Cloud monitoring. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 97–116). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch006

Hertzler, B. T., Frost, E., Bressler, G. H., & Goehring, C. (2011). Experience report: Using a cloud computing environment during Haiti and Exercise24.[IJISCRAM]. *International Journal of Information Systems for Crisis Response and Management*, *3*(1), 50–64. doi:10.4018/jiscrm.2011010104

Hertzler, B. T., Frost, E., Bressler, G. H., & Goehring, C. (2013). Experience report: Using a cloud computing environment during Haiti and Exercise24. In M. Jennex (Ed.), *Using social and information technologies for disaster and crisis management* (pp. 52–66). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2788-8.ch004

Ho, R. (2013). Cloud computing and enterprise migration strategies. In A. Loo (Ed.), *Distributed computing innovations for business, engineering, and science* (pp. 156–175). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2533-4.ch008

Hobona, G., Jackson, M., & Anand, S. (2012). Implementing geospatial web services for cloud computing. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 615–636). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch305

Hochstein, L., Schott, B., & Graybill, R. B. (2011). Computational engineering in the cloud: Benefits and challenges.[JOEUC]. *Journal of Organizational and End User Computing*, *23*(4), 31–50. doi:10.4018/joeuc.2011100103

Hochstein, L., Schott, B., & Graybill, R. B. (2013). Computational engineering in the cloud: Benefits and challenges. In A. Dwivedi & S. Clarke (Eds.), *Innovative strategies and approaches for end-user computing advancements* (pp. 314–332). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2059-9.ch017

Honarvar, A. R. (2013). Developing an elastic cloud computing application through multi-agent systems.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 58–64. doi:10.4018/ijcac.2013010106

Hossain, S. (2013). Cloud computing terms, definitions, and taxonomy. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 1–25). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch001

Hudzia, B., Sinclair, J., & Lindner, M. (2013). Deploying and running enterprise grade applications in a federated cloud. In *Supply chain management: Concepts, methodologies, tools, and applications* (pp. 1350–1370). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2625-6.ch080

Hung, S., Shieh, J., & Lee, C. (2011). Migrating android applications to the cloud. [IJGHPC]. *International Journal of Grid and High Performance Computing*, *3*(2), 14–28. doi:10.4018/jghpc.2011040102

Hung, S., Shieh, J., & Lee, C. (2013). Migrating android applications to the cloud. In E. Udoh (Ed.), *Applications and developments in grid, cloud, and high performance computing* (pp. 307–322). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2065-0.ch020

Islam, S., Mouratidis, H., & Weippl, E. R. (2013). A goal-driven risk management approach to support security and privacy analysis of cloud-based system. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 97–122). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch006

Itani, W., Kayssi, A., & Chehab, A. (2013). Hardware-based security for ensuring data privacy in the cloud. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 147–170). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch008

Jackson, A., & Weiland, M. (2013). Cloud computing for scientific simulation and high performance computing. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 51–70). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch003

Jaeger, M. C., & Hohenstein, U. (2013). Content centric storage and current storage systems. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 27–46). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch003

James, J. I., Shosha, A. F., & Gladyshev, P. (2013). Digital forensic investigation and cloud computing. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 1–41). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch001

Jena, R. K. (2013). Green computing to green business. In P. Ordóñez de Pablos (Ed.), *Green technologies and business practices: An IT approach* (pp. 138–150). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1972-2.ch007

Jeyarani, R., & Nagaveni, N. (2012). A heuristic meta scheduler for optimal resource utilization and improved QoS in cloud computing environment.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(1), 41–52. doi:10.4018/ijcac.2012010103

Jeyarani, R., Nagaveni, N., & Ram, R. V. (2011). Self adaptive particle swarm optimization for efficient virtual machine provisioning in cloud.[IJIIT]. *International Journal of Intelligent Information Technologies*, *7*(2), 25–44. doi:10.4018/jiit.2011040102

Jeyarani, R., Nagaveni, N., & Ram, R. V. (2013). Self adaptive particle swarm optimization for efficient virtual machine provisioning in cloud. In V. Sugumaran (Ed.), *Organizational efficiency through intelligent information technologies* (pp. 88–107). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2047-6.ch006

Jeyarani, R., Nagaveni, N., Sadasivam, S. K., & Rajarathinam, V. R. (2011). Power aware meta scheduler for adaptive VM provisioning in IaaS cloud.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(3), 36–51. doi:10.4018/ijcac.2011070104

Jeyarani, R., Nagaveni, N., Sadasivam, S. K., & Rajarathinam, V. R. (2013). Power aware meta scheduler for adaptive VM provisioning in IaaS cloud. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 190–204). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch014

Jiang, J., Huang, X., Wu, Y., & Yang, G. (2013). Campus cloud storage and preservation: From distributed file system to data sharing service. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 284–301). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch012

Jing, S. (2012). The application exploration of cloud computing in information technology teaching.[IJAPUC]. *International Journal of Advanced Pervasive and Ubiquitous Computing*, *4*(4), 23–27. doi:10.4018/japuc.2012100104

Johansson, D., & Wiberg, M. (2012). Conceptually advancing "application mobility" towards design: Applying a concept-driven approach to the design of mobile IT for home care service groups.[IJACI]. *International Journal of Ambient Computing and Intelligence*, *4*(3), 20–32. doi:10.4018/jaci.2012070102

Jorda, J., & M'zoughi, A. (2013). Securing cloud storage. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 171–190). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch009

Juiz, C., & Alexander de Pous, V. (2014). Cloud computing: IT governance, legal, and public policy aspects. In I. Portela & F. Almeida (Eds.), *Organizational, legal, and technological dimensions of information system administration* (pp. 139–166). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4526-4.ch009

Kaisler, S. H., Money, W., & Cohen, S. J. (2013). Cloud computing: A decision framework for small businesses. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 151–172). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch008

Kanamori, Y., & Yen, M. Y. (2013). Cloud computing security and risk management. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 222–240). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch012

Karadsheh, L., & Alhawari, S. (2011). Applying security policies in small business utilizing cloud computing technologies.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(2), 29–40. doi:10.4018/ijcac.2011040103

Karadsheh, L., & Alhawari, S. (2013). Applying security policies in small business utilizing cloud computing technologies. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 112–124). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch008

Kaupins, G. (2012). Laws associated with mobile computing in the cloud.[IJWNBT]. *International Journal of Wireless Networks and Broadband Technologies*, *2*(3), 1–9. doi:10.4018/ijwnbt.2012070101

Kemp, M. L., Robb, S., & Deans, P. C. (2013). The legal implications of cloud computing. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 257–272). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch014

Khan, N., Ahmad, N., Herawan, T., & Inayat, Z. (2012). Cloud computing: Locally sub-clouds instead of globally one cloud.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(3), 68–85. doi:10.4018/ijcac.2012070103

Khan, N., Noraziah, A., Ismail, E. I., Deris, M. M., & Herawan, T. (2012). Cloud computing: Analysis of various platforms.[IJEEI]. *International Journal of E-Entrepreneurship and Innovation*, *3*(2), 51–59. doi:10.4018/jeei.2012040104

Khansa, L., Forcade, J., Nambari, G., Parasuraman, S., & Cox, P. (2012). Proposing an intelligent cloud-based electronic health record system.[IJBDCN]. *International Journal of Business Data Communications and Networking*, *8*(3), 57–71. doi:10.4018/jbdcn.2012070104

Kierkegaard, S. (2012). Not every cloud brings rain: Legal risks on the horizon. In M. Gupta, J. Walp, & R. Sharman (Eds.), *Strategic and practical approaches for information security governance: Technologies and applied solutions* (pp. 181–194). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0197-0.ch011

Kifayat, K., Shamsa, T. B., Mackay, M., Merabti, M., & Shi, Q. (2013). Real time risk management in cloud computation. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 123–145). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch007

King, T. M., Ganti, A. S., & Froslie, D. (2013). Towards improving the testability of cloud application services. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 322–339). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch015

Kipp, A., Schneider, R., & Schubert, L. (2013). Encapsulation of complex HPC services. In C. Rückemann (Ed.), *Integrated information and computing systems for natural, spatial, and social sciences* (pp. 153–176). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2190-9.ch008

Kldiashvili, E. (2012). The cloud computing as the tool for implementation of virtual organization technology for ehealth.[JITR]. *Journal of Information Technology Research*, *5*(1), 18–34. doi:10.4018/jitr.2012010102

Kldiashvili, E. (2013). Implementation of telecytology in georgia for quality assurance programs.[JITR]. *Journal of Information Technology Research*, *6*(2), 24–45. doi:10.4018/jitr.2013040102

Kosmatov, N. (2013). Concolic test generation and the cloud: deployment and verification perspectives. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 231–251). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch011

Kotamarti, R. M., Thornton, M. A., & Dunham, M. H. (2012). Quantum computing approach for alignment-free sequence search and classification. In S. Ali, N. Abbadeni, & M. Batouche (Eds.), *Multidisciplinary computational intelligence techniques: Applications in business, engineering, and medicine* (pp. 279–300). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1830-5.ch017

Kremmydas, D., Petsakos, A., & Rozakis, S. (2012). Parametric optimization of linear and non-linear models via parallel computing to enhance web-spatial DSS interactivity.[IJDSST]. *International Journal of Decision Support System Technology*, *4*(1), 14–29. doi:10.4018/jdsst.2012010102

Krishnadas, N., & Pillai, R. R. (2013). Cloud computing diagnosis: A comprehensive study. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 1–18). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch001

Kübert, R., & Katsaros, G. (2011). Using free software for elastic web hosting on a private cloud.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(2), 14–28. doi:10.4018/ijcac.2011040102

Kübert, R., & Katsaros, G. (2013). Using free software for elastic web hosting on a private cloud. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 97–111). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch007

Kumar, P. S., Ashok, M. S., & Subramanian, R. (2012). A publicly verifiable dynamic secret sharing protocol for secure and dependable data storage in cloud computing. [IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(3), 1–25. doi:10.4018/ijcac.2012070101

Lasluisa, S., Rodero, I., & Parashar, M. (2013). Software design for passing sarbanes-oxley in cloud computing. In C. Rückemann (Ed.), *Integrated information and computing systems for natural, spatial, and social sciences* (pp. 27–42). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2190-9.ch002

Lasluisa, S., Rodero, I., & Parashar, M. (2014). Software design for passing sarbanes-oxley in cloud computing. In *Software design and development: Concepts, methodologies, tools, and applications* (pp. 1659–1674). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4301-7.ch080

Lee, W. N. (2013). An economic analysis of cloud: "Software as a service" (saas) computing and "virtual desktop infrastructure" (VDI) models. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 289–295). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch016

Levine, K., & White, B. A. (2011). A crisis at hafford furniture: Cloud computing case study.[JCIT]. *Journal of Cases on Information Technology*, *13*(1), 57–71. doi:10.4018/jcit.2011010104

Levine, K., & White, B. A. (2013). A crisis at Hafford furniture: Cloud computing case study. In M. Khosrow-Pour (Ed.), *Cases on emerging information technology research and applications* (pp. 70–87). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3619-4.ch004

Li, J., Meng, L., Zhu, Z., Li, X., Huai, J., & Liu, L. (2013). CloudRank: A cloud service ranking method based on both user feedback and service testing. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 230–258). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch010

Liang, T., Lu, F., & Chiu, J. (2012). A hybrid resource reservation method for workflows in clouds.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(4), 1–21. doi:10.4018/jghpc.2012100101

Lorenz, M., Rath-Wiggins, L., Runde, W., Messina, A., Sunna, P., Dimino, G., & Borgotallo, R. et al. (2013). Media convergence and cloud technologies: Smart storage, better workflows. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 132–144). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch009

M., S. G., & G., S. K. (2012). An enterprise mashup integration service framework for clouds. *International Journal of Cloud Applications and Computing (IJCAC), 2*(2), 31-40. doi:10.4018/ijcac.2012040103

Maharana, S. K., Mali, P. B., Prabhakar, G. J. S., & Kumar, V. (2011). Cloud computing applied for numerical study of thermal characteristics of SIP.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(3), 12–21. doi:10.4018/ijcac.2011070102

Maharana, S. K., Mali, P. B., Prabhakar, G. J. S., & Kumar, V. (2013). Cloud computing applied for numerical study of thermal characteristics of SIP. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 166–175). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch012

Maharana, S. K., P, G. P., & Bhati, A. (2012). A study of cloud computing for retinal image processing through MATLAB.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(2), 59–69. doi:10.4018/ijcac.2012040106

Maharana, S. K., Prabhakar, P. G., & Bhati, A. (2013). A study of cloud computing for retinal image processing through MATLAB. In *Image processing: Concepts, methodologies, tools, and applications* (pp. 101–111). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3994-2.ch006

Mahesh, S., Landry, B. J., Sridhar, T., & Walsh, K. R. (2011). A decision table for the cloud computing decision in small business.[IRMJ]. *Information Resources Management Journal*, *24*(3), 9–25. doi:10.4018/irmj.2011070102

Mahesh, S., Landry, B. J., Sridhar, T., & Walsh, K. R. (2013). A decision table for the cloud computing decision in small business. In M. Khosrow-Pour (Ed.), *Managing information resources and technology: Emerging Applications and theories* (pp. 159–176). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3616-3.ch012

Marquezan, C. C., Metzger, A., Pohl, K., Engen, V., Boniface, M., Phillips, S. C., & Zlatev, Z. (2013). Adaptive future internet applications: Opportunities and challenges for adaptive web services technology. In G. Ortiz & J. Cubo (Eds.), *Adaptive web services for modular and reusable software development: Tactics and solutions* (pp. 333–353). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2089-6.ch014

Marshall, P. J. (2012). Cloud computing: Next generation education. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 180–185). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch012

Martinez-Ortiz, A. (2012). Open cloud technologies. In L. Vaquero, J. Cáceres, & J. Hierro (Eds.), *Open source cloud computing systems: Practices and paradigms* (pp. 1–17). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0098-0.ch001

Massonet, P., Michot, A., Naqvi, S., Villari, M., & Latanicki, J. (2013). Securing the external interfaces of a federated infrastructure cloud. In *IT policy and ethics: Concepts, methodologies, tools, and applications* (pp. 1876–1903). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2919-6.ch082

Mavrogeorgi, N., Gogouvitis, S. V., Voulodimos, A., & Alexandrou, V. (2013). SLA management in storage clouds. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 72–93). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch006

Mehta, H. K. (2013). Cloud selection for e-business a parameter based solution. In K. Tarnay, S. Imre, & L. Xu (Eds.), *Research and development in e-business through service-oriented solutions* (pp. 199–207). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-4181-5.ch009

Mehta, H. K., & Gupta, E. (2013). Economy based resource allocation in IaaS cloud. [IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(2), 1–11. doi:10.4018/ijcac.2013040101

Miah, S. J. (2012). Cloud-based intelligent DSS design for emergency professionals. In S. Ali, N. Abbadeni, & M. Batouche (Eds.), *Multidisciplinary computational intelligence techniques: Applications in business, engineering, and medicine* (pp. 47–60). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1830-5.ch004

Miah, S. J. (2013). Cloud-based intelligent DSS design for emergency professionals. In *Data mining: Concepts, methodologies, tools, and applications* (pp. 991–1003). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2455-9.ch050

Mikkilineni, R. (2012). Architectural resiliency in distributed computing.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(4), 37–51. doi:10.4018/jghpc.2012100103

Millham, R. (2012). Software asset re-use: Migration of data-intensive legacy system to the cloud computing paradigm. In H. Yang & X. Liu (Eds.), *Software reuse in the emerging cloud computing era* (pp. 1–27). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0897-9.ch001

Mircea, M. (2011). Building the agile enterprise with service-oriented architecture, business process management and decision management.[IJEEI]. *International Journal of E-Entrepreneurship and Innovation*, *2*(4), 32–48. doi:10.4018/jeei.2011100103

Modares, H., Lloret, J., Moravejosharieh, A., & Salleh, R. (2014). Security in mobile cloud computing. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 79–91). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch005

Moedjiono, S., & Mas'at, A. (2012). Cloud computing implementation strategy for information dissemination on meteorology, climatology, air quality, and geophysics (MKKuG).[JITR]. *Journal of Information Technology Research*, *5*(3), 71–84. doi:10.4018/jitr.2012070104

Moiny, J. (2012). Cloud based social network sites: Under whose control? In A. Dudley, J. Braman, & G. Vincenti (Eds.), *Investigating cyber law and cyber ethics: Issues, impacts and practices* (pp. 147–219). Hershey, PA: Information Science Reference; doi:10.4018/978-1-61350-132-0.ch009

Moreno, I. S., & Xu, J. (2011). Energy-efficiency in cloud computing environments: Towards energy savings without performance degradation.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(1), 17–33. doi:10.4018/ijcac.2011010102

Moreno, I. S., & Xu, J. (2013). Energy-efficiency in cloud computing environments: Towards energy savings without performance degradation. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 18–36). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch002

Muñoz, A., Maña, A., & González, J. (2013). Dynamic security properties monitoring architecture for cloud computing. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 1–18). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch001

Mvelase, P., Dlodlo, N., Williams, Q., & Adigun, M. O. (2011). Custom-made cloud enterprise architecture for small medium and micro enterprises.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(3), 52–63. doi:10.4018/ijcac.2011070105

Mvelase, P., Dlodlo, N., Williams, Q., & Adigun, M. O. (2012). Custom-made cloud enterprise architecture for small medium and micro enterprises. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 589–601). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch303

Mvelase, P., Dlodlo, N., Williams, Q., & Adigun, M. O. (2013). Custom-made cloud enterprise architecture for small medium and micro enterprises. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 205–217). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch015

Naeem, M. A., Dobbie, G., & Weber, G. (2014). Big data management in the context of real-time data warehousing. In W. Hu & N. Kaabouch (Eds.), *Big data management, technologies, and applications* (pp. 150–176). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4699-5.ch007

Ofosu, W. K., & Saliah-Hassane, H. (2013). Cloud computing in the education environment for developing nations.[IJITN]. *International Journal of Interdisciplinary Telecommunications and Networking*, *5*(3), 54–62. doi:10.4018/jitn.2013070106

Oliveros, E., Cucinotta, T., Phillips, S. C., Yang, X., Middleton, S., & Voith, T. (2012). Monitoring and metering in the cloud. In D. Kyriazis, T. Varvarigou, & K. Konstanteli (Eds.), *Achieving real-time in distributed computing: From grids to clouds* (pp. 94–114). Hershey, PA: Information Science Reference; doi:10.4018/978-1-60960-827-9.ch006

Orton, I., Alva, A., & Endicott-Popovsky, B. (2013). Legal process and requirements for cloud forensic investigations. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 186–229). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch008

Pakhira, A., & Andras, P. (2013). Leveraging the cloud for large-scale software testing – A case study: Google Chrome on Amazon. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 252–279). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch012

Pal, K., & Karakostas, B. (2013). The use of cloud computing in shipping logistics. In D. Graham, I. Manikas, & D. Folinas (Eds.), *E-logistics and e-supply chain management: Applications for evolving business* (pp. 104–124). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3914-0.ch006

Pal, S. (2013). Cloud computing: Security concerns and issues. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 191–207). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch010

Pal, S. (2013). Storage security and technical challenges of cloud computing. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 225–240). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch014

Palanivel, K., & Kuppuswami, S. (2014). A cloud-oriented reference architecture to digital library systems. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 230–254). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch014

Paletta, M. (2012). Intelligent clouds: By means of using multi-agent systems environments. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 254–279). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch017

Pallot, M., Le Marc, C., Richir, S., Schmidt, C., & Mathieu, J. (2012). Innovation gaming: An immersive experience environment enabling co-creation. In M. Cruz-Cunha (Ed.), *Handbook of research on serious games as educational, business and research tools* (pp. 1–24). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0149-9.ch001

Pankowska, M. (2011). Information technology resources virtualization for sustainable development.[IJAL]. *International Journal of Applied Logistics*, *2*(2), 35–48. doi:10.4018/jal.2011040103

Pankowska, M. (2013). Information technology resources virtualization for sustainable development. In Z. Luo (Ed.), *Technological solutions for modern logistics and supply chain management* (pp. 248–262). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2773-4.ch016

Parappallil, J. J., Zarvic, N., & Thomas, O. (2012). A context and content reflection on business-IT alignment research.[IJITBAG]. *International Journal of IT/Business Alignment and Governance*, *3*(2), 21–37. doi:10.4018/jitbag.2012070102

Parashar, V., Vishwakarma, M. L., & Parashar, R. (2014). A new framework for building academic library through cloud computing. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 107–123). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch007

Pendyala, V. S., & Holliday, J. (2012). Cloud as a computer. In X. Liu & Y. Li (Eds.), *Advanced design approaches to emerging software systems: Principles, methodologies and tools* (pp. 241–249). Hershey, PA: Information Science Reference; doi:10.4018/978-1-60960-735-7.ch011

Petruch, K., Tamm, G., & Stantchev, V. (2012). Deriving in-depth knowledge from IT-performance data simulations.[IJKSR]. *International Journal of Knowledge Society Research*, *3*(2), 13–29. doi:10.4018/jksr.2012040102

Philipson, G. (2011). A framework for green computing.[IJGC]. *International Journal of Green Computing*, *2*(1), 12–26. doi:10.4018/jgc.2011010102

Philipson, G. (2013). A framework for green computing. In K. Ganesh & S. Anbuudayasankar (Eds.), *International and interdisciplinary studies in green computing* (pp. 12–26). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2646-1.ch002

Phythian, M. (2013). The 'cloud' of unknowing – What a government cloud may and may not offer: A practitioner perspective.[IJT]. *International Journal of Technoethics*, *4*(1), 1–10. doi:10.4018/jte.2013010101

Pym, D., & Sadler, M. (2012). Information stewardship in cloud computing. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 185–202). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch109

Pym, D., & Sadler, M. (2012). Information stewardship in cloud computing. In S. Galup (Ed.), *Technological applications and advancements in service science, management, and engineering* (pp. 52–69). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1583-0.ch004

Qiu, J., Ekanayake, J., Gunarathne, T., Choi, J. Y., Bae, S., & Ruan, Y. … Tang, H. (2013). Data intensive computing for bioinformatics. In Bioinformatics: Concepts, methodologies, tools, and applications (pp. 287-321). Hershey, PA: Medical Information Science Reference. doi:10.4018/978-1-4666-3604-0.ch016

Rabaey, M. (2012). A public economics approach to enabling enterprise architecture with the government cloud in Belgium. In P. Saha (Ed.), *Enterprise architecture for connected e-government: Practices and innovations* (pp. 467–493). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1824-4.ch020

Rabaey, M. (2013). A complex adaptive system thinking approach of government e-procurement in a cloud computing environment. In P. Ordóñez de Pablos, J. Lovelle, J. Gayo, & R. Tennyson (Eds.), *E-procurement management for successful electronic government systems* (pp. 193–219). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2119-0.ch013

Rabaey, M. (2013). Holistic investment framework for cloud computing: A management-philosophical approach based on complex adaptive systems. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 94–122). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch005

Rak, M., Ficco, M., Luna, J., Ghani, H., Suri, N., Panica, S., & Petcu, D. (2012). Security issues in cloud federations. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 176–194). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch010

Ramanathan, R. (2013). Extending service-driven architectural approaches to the cloud. In R. Ramanathan & K. Raja (Eds.), *Service-driven approaches to architecture and enterprise integration* (pp. 334–359). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4193-8.ch013

Ramírez, M., Gutiérrez, A., Monguet, J. M., & Muñoz, C. (2012). An internet cost model, assignment of costs based on actual network use.[IJWP]. *International Journal of Web Portals*, *4*(4), 19–34. doi:10.4018/jwp.2012100102

Rashid, A., Wang, W. Y., & Tan, F. B. (2013). Value co-creation in cloud services. In A. Lin, J. Foster, & P. Scifleet (Eds.), *Consumer information systems and relationship management: Design, implementation, and use* (pp. 74–91). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-4082-5.ch005

Ratten, V. (2012). Cloud computing services: Theoretical foundations of ethical and entrepreneurial adoption behaviour.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(2), 48–58. doi:10.4018/ijcac.2012040105

Ratten, V. (2013). Exploring behaviors and perceptions affecting the adoption of cloud computing.[IJIDE]. *International Journal of Innovation in the Digital Economy*, *4*(3), 51–68. doi:10.4018/jide.2013070104

Ravi, V. (2012). Cloud computing paradigm for indian education sector.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(2), 41–47. doi:10.4018/ijcac.2012040104

Rawat, A., Kapoor, P., & Sushil, R. (2014). Application of cloud computing in library information service sector. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 77–89). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch005

Reich, C., Hübner, S., & Kuijs, H. (2012). Cloud computing for on-demand virtual desktops and labs. In L. Chao (Ed.), *Cloud computing for teaching and learning: strategies for design and implementation* (pp. 111–125). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch008

Rice, R. W. (2013). Testing in the cloud: Balancing the value and risks of cloud computing. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 404–416). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch019

Ruan, K. (2013). Designing a forensic-enabling cloud ecosystem. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 331–344). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch014

Sabetzadeh, F., & Tsui, E. (2011). Delivering knowledge services in the cloud. [IJKSS]. *International Journal of Knowledge and Systems Science*, *2*(4), 14–20. doi:10.4018/jkss.2011100102

Sabetzadeh, F., & Tsui, E. (2013). Delivering knowledge services in the cloud. In G. Yang (Ed.), *Multidisciplinary studies in knowledge and systems science* (pp. 247–254). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-3998-0.ch017

Saedi, A., & Iahad, N. A. (2013). Future research on cloud computing adoption by small and medium-sized enterprises: A critical analysis of relevant theories.[IJANTTI]. *International Journal of Actor-Network Theory and Technological Innovation*, *5*(2), 1–16. doi:10.4018/jantti.2013040101

Saha, D., & Sridhar, V. (2011). Emerging areas of research in business data communications.[IJBDCN]. *International Journal of Business Data Communications and Networking*, *7*(4), 52–59. doi:10.4018/IJBDCN.2011100104

Saha, D., & Sridhar, V. (2013). Platform on platform (PoP) model for meta-networking: A new paradigm for networks of the future.[IJBDCN]. *International Journal of Business Data Communications and Networking*, *9*(1), 1–10. doi:10.4018/jbdcn.2013010101

Sahlin, J. P. (2013). Cloud computing: Past, present, and future. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 19–50). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch002

Salama, M., & Shawish, A. (2012). Libraries: From the classical to cloud-based era. [IJDLS]. *International Journal of Digital Library Systems*, *3*(3), 14–32. doi:10.4018/jdls.2012070102

Sánchez, C. M., Molina, D., Vozmediano, R. M., Montero, R. S., & Llorente, I. M. (2012). On the use of the hybrid cloud computing paradigm. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 196–218). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch011

Sasikala, P. (2011). Architectural strategies for green cloud computing: Environments, infrastructure and resources.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(4), 1–24. doi:10.4018/ijcac.2011100101

Sasikala, P. (2011). Cloud computing in higher education: Opportunities and issues. [IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(2), 1–13. doi:10.4018/ijcac.2011040101

Sasikala, P. (2011). Cloud computing towards technological convergence.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(4), 44–59. doi:10.4018/ijcac.2011100104

Sasikala, P. (2012). Cloud computing and e-governance: Advances, opportunities and challenges.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(4), 32–52. doi:10.4018/ijcac.2012100103

Sasikala, P. (2012). Cloud computing in higher education: Opportunities and issues. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1672–1685). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch709

Sasikala, P. (2012). Cloud computing towards technological convergence. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1576–1592). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch703

Sasikala, P. (2013). Architectural strategies for green cloud computing: Environments, infrastructure and resources. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 218–242). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch016

Sasikala, P. (2013). Cloud computing in higher education: Opportunities and issues. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 83–96). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch006

Sasikala, P. (2013). Cloud computing towards technological convergence. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 263–279). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch019

Sasikala, P. (2013). New media cloud computing: Opportunities and challenges. [IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(2), 61–72. doi:10.4018/ijcac.2013040106

Schrödl, H., & Wind, S. (2013). Requirements engineering for cloud application development. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 137–150). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch007

Sclater, N. (2012). Legal and contractual issues of cloud computing for educational institutions. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 186–199). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch013

Sen, J. (2014). Security and privacy issues in cloud computing. In A. Ruiz-Martinez, R. Marin-Lopez, & F. Pereniguez-Garcia (Eds.), *Architectures and protocols for secure information technology infrastructures* (pp. 1–45). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4514-1.ch001

Shah, B. (2013). Cloud environment controls assessment framework. In *IT policy and ethics: Concepts, methodologies, tools, and applications* (pp. 1822–1847). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2919-6.ch080

Shah, B. (2013). Cloud environment controls assessment framework. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 28–53). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch002

Shang, X., Zhang, R., & Chen, Y. (2012). Internet of things (IoT) service architecture and its application in e-commerce.[JECO]. *Journal of Electronic Commerce in Organizations*, *10*(3), 44–55. doi:10.4018/jeco.2012070104

Shankararaman, V., & Kit, L. E. (2013). Integrating the cloud scenarios and solutions. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 173–189). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch009

Sharma, A., & Maurer, F. (2013). A roadmap for software engineering for the cloud: Results of a systematic review. In X. Wang, N. Ali, I. Ramos, & R. Vidgen (Eds.), *Agile and lean service-oriented development: Foundations, theory, and practice* (pp. 48–63). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2503-7.ch003

Sharma, A., & Maurer, F. (2014). A roadmap for software engineering for the cloud: Results of a systematic review. In *Software design and development: Concepts, methodologies, tools, and applications* (pp. 1–16). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4301-7.ch001

Sharma, S. C., & Bagoria, H. (2014). Libraries and cloud computing models: A changing paradigm. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 124–149). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch008

Shawish, A., & Salama, M. (2013). Cloud computing in academia, governments, and industry. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 71–114). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch004

Shebanow, A., Perez, R., & Howard, C. (2012). The effect of firewall testing types on cloud security policies.[IJSITA]. *International Journal of Strategic Information Technology and Applications*, *3*(3), 60–68. doi:10.4018/jsita.2012070105

Sheikhalishahi, M., Devare, M., Grandinetti, L., & Incutti, M. C. (2012). A complementary approach to grid and cloud distributed computing paradigms. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1929–1942). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch811

Sheikhalishahi, M., Devare, M., Grandinetti, L., & Incutti, M. C. (2012). A complementary approach to grid and cloud distributed computing paradigms. In N. Preve (Ed.), *Computational and data grids: Principles, applications and design* (pp. 31–44). Hershey, PA: Information Science Reference; doi:10.4018/978-1-61350-113-9.ch002

Shen, Y., Li, Y., Wu, L., Liu, S., & Wen, Q. (2014). Cloud computing overview. In Y. Shen, Y. Li, L. Wu, S. Liu, & Q. Wen (Eds.), *Enabling the new era of cloud computing: Data security, transfer, and management* (pp. 1–24). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4801-2.ch001

Shen, Y., Li, Y., Wu, L., Liu, S., & Wen, Q. (2014). Main components of cloud computing. In Y. Shen, Y. Li, L. Wu, S. Liu, & Q. Wen (Eds.), *Enabling the new era of cloud computing: Data security, transfer, and management* (pp. 25–50). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4801-2.ch002

Shen, Y., Yang, J., & Keskin, T. (2014). Impact of cultural differences on the cloud computing ecosystems in the USA and China. In Y. Shen, Y. Li, L. Wu, S. Liu, & Q. Wen (Eds.), *Enabling the new era of cloud computing: Data security, transfer, and management* (pp. 269–283). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4801-2.ch014

Shetty, S., & Rawat, D. B. (2013). Cloud computing based cognitive radio networking. In N. Meghanathan & Y. Reddy (Eds.), *Cognitive radio technology applications for wireless and mobile ad hoc networks* (pp. 153–164). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4221-8.ch008

Shi, Z., & Beard, C. (2014). QoS in the mobile cloud computing environment. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 200–217). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch011

Shuster, L. (2013). Enterprise integration: Challenges and solution architecture. In R. Ramanathan & K. Raja (Eds.), *Service-driven approaches to architecture and enterprise integration* (pp. 43–66). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4193-8.ch002

Siahos, Y., Papanagiotou, I., Georgopoulos, A., Tsamis, F., & Papaioannou, I. (2012). An architecture paradigm for providing cloud services in school labs based on open source software to enhance ICT in education. [IJCEE]. *International Journal of Cyber Ethics in Education*, *2*(1), 44–57. doi:10.4018/ijcee.2012010105

Simon, E., & Estublier, J. (2013). Model driven integration of heterogeneous software artifacts in service oriented computing. In A. Ionita, M. Litoiu, & G. Lewis (Eds.), *Migrating legacy applications: Challenges in service oriented architecture and cloud computing environments* (pp. 332–360). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2488-7.ch014

Singh, J., & Kumar, V. (2013). Compliance and regulatory standards for cloud computing. In R. Khurana & R. Aggarwal (Eds.), *Interdisciplinary perspectives on business convergence, computing, and legality* (pp. 54–64). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-4209-6.ch006

Singh, V. V. (2012). Software development using service syndication based on API handshake approach between cloud-based and SOA-based reusable services. In H. Yang & X. Liu (Eds.), *Software reuse in the emerging cloud computing era* (pp. 136–157). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0897-9.ch006

Smeitink, M., & Spruit, M. (2013). Maturity for sustainability in IT: Introducing the MITS.[IJITSA]. *International Journal of Information Technologies and Systems Approach*, *6*(1), 39–56. doi:10.4018/jitsa.2013010103

Smith, P. A., & Cockburn, T. (2013). Socio-digital technologies. In *Dynamic leadership models for global business: Enhancing digitally connected environments* (pp. 142–168). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2836-6.ch006

Sneed, H. M. (2013). Testing web services in the cloud. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 136–173). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch007

Solomon, B., Ionescu, D., Gadea, C., & Litoiu, M. (2013). Geographically distributed cloud-based collaborative application. In A. Ionita, M. Litoiu, & G. Lewis (Eds.), *Migrating legacy applications: Challenges in service oriented architecture and cloud computing environments* (pp. 248–274). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2488-7.ch011

Song, W., & Xiao, Z. (2013). An infrastructure-as-a-service cloud: On-demand resource provisioning. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 302–324). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch013

Sood, S. K. (2013). A value based dynamic resource provisioning model in cloud. [IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 1–12. doi:10.4018/ijcac.2013010101

Sotiriadis, S., Bessis, N., & Antonopoulos, N. (2012). Exploring inter-cloud load balancing by utilizing historical service submission records.[IJDST]. *International Journal of Distributed Systems and Technologies*, *3*(3), 72–81. doi:10.4018/jdst.2012070106

Soyata, T., Ba, H., Heinzelman, W., Kwon, M., & Shi, J. (2014). Accelerating mobile-cloud computing: A survey. In H. Mouftah & B. Kantarci (Eds.), *Communication infrastructures for cloud computing* (pp. 175–197). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4522-6.ch008

Spyridopoulos, T., & Katos, V. (2011). Requirements for a forensically ready cloud storage service.[IJDCF]. *International Journal of Digital Crime and Forensics*, *3*(3), 19–36. doi:10.4018/jdcf.2011070102

Spyridopoulos, T., & Katos, V. (2013). Data recovery strategies for cloud environments. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 251–265). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2662-1.ch010

Srinivasa, K. G., S, H. R. C., H, M. K. S., & Venkatesh, N. (2012). MeghaOS: A framework for scalable, interoperable cloud based operating system.[IJCAC]. *International Journal of Cloud Applications and Computing*, *2*(1), 53–70. doi:10.4018/ijcac.2012010104

Stantchev, V., & Stantcheva, L. (2012). Extending traditional IT-governance knowledge towards SOA and cloud governance.[IJKSR]. *International Journal of Knowledge Society Research*, *3*(2), 30–43. doi:10.4018/jksr.2012040103

Stantchev, V., & Tamm, G. (2012). Reducing information asymmetry in cloud marketplaces.[IJHCITP]. *International Journal of Human Capital and Information Technology Professionals*, *3*(4), 1–10. doi:10.4018/jhcitp.2012100101

Steinbuß, S., & Weißenberg, N. (2013). Service design and process design for the logistics mall cloud. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 186–206). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch008

Stender, J., Berlin, M., & Reinefeld, A. (2013). XtreemFS: A file system for the cloud. In D. Kyriazis, A. Voulodimos, S. Gogouvitis, & T. Varvarigou (Eds.), *Data intensive storage services for cloud environments* (pp. 267–285). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3934-8.ch016

Sticklen, D. J., & Issa, T. (2011). An initial examination of free and proprietary software-selection in organizations.[IJWP]. *International Journal of Web Portals*, *3*(4), 27–43. doi:10.4018/jwp.2011100103

Sun, Y., White, J., Gray, J., & Gokhale, A. (2012). Model-driven automated error recovery in cloud computing. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 680–700). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch308

Sun, Z., Yang, Y., Zhou, Y., & Cruickshank, H. (2014). Agent-based resource management for mobile cloud. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 118–134). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch007

Sutherland, S. (2013). Convergence of interoperability of cloud computing, service oriented architecture and enterprise architecture.[IJEEI]. *International Journal of E-Entrepreneurship and Innovation*, *4*(1), 43–51. doi:10.4018/jeei.2013010104

Takabi, H., & Joshi, J. B. (2013). Policy management in cloud: Challenges and approaches. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 191–211). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch010

Takabi, H., & Joshi, J. B. (2013). Policy management in cloud: Challenges and approaches. In *IT policy and ethics: Concepts, methodologies, tools, and applications* (pp. 814–834). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2919-6.ch037

Takabi, H., Joshi, J. B., & Ahn, G. (2013). Security and privacy in cloud computing: Towards a comprehensive framework. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 164–184). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch007

Takabi, H., Zargar, S. T., & Joshi, J. B. (2014). Mobile cloud computing and its security and privacy challenges. In D. Rawat, B. Bista, & G. Yan (Eds.), *Security, privacy, trust, and resource management in mobile and wireless communications* (pp. 384–407). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4691-9.ch016

Teixeira, C., Pinto, J. S., Ferreira, F., Oliveira, A., Teixeira, A., & Pereira, C. (2013). Cloud computing enhanced service development architecture for the living usability lab. In R. Martinho, R. Rijo, M. Cruz-Cunha, & J. Varajão (Eds.), *Information systems and technologies for enhancing health and social care* (pp. 33–53). Hershey, PA: Medical Information Science Reference; doi:10.4018/978-1-4666-3667-5.ch003

Thimm, H. (2012). Cloud-based collaborative decision making: Design considerations and architecture of the GRUPO-MOD system.[IJDSST]. *International Journal of Decision Support System Technology*, *4*(4), 39–59. doi:10.4018/jdsst.2012100103

Thomas, P. (2012). Harnessing the potential of cloud computing to transform higher education. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 147–158). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch010

T.M. K., & Gopalakrishnan, S. (2014). Green economic and secure libraries on cloud. In S. Dhamdhere (Ed.), Cloud computing and virtualization technologies in libraries (pp. 297-315). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-4631-5.ch017

Toka, A., Aivazidou, E., Antoniou, A., & Arvanitopoulos-Darginis, K. (2013). Cloud computing in supply chain management: An overview. In D. Graham, I. Manikas, & D. Folinas (Eds.), *E-logistics and e-supply chain management: Applications for evolving business* (pp. 218–231). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-3914-0.ch012

Torrealba, S. M., Morales, P. M., Campos, J. M., & Meza, S. M. (2013). A software tool to support risks analysis about what should or should not go to the cloud. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 72–96). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch005

Trivedi, M., & Suthar, V. (2013). Cloud computing: A feasible platform for ICT enabled health science libraries in India.[IJUDH]. *International Journal of User-Driven Healthcare*, *3*(2), 69–77. doi:10.4018/ijudh.2013040108

Truong, H., Pham, T., Thoai, N., & Dustdar, S. (2012). Cloud computing for education and research in developing countries. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 64–80). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch005

Tsirmpas, C., Giokas, K., Iliopoulou, D., & Koutsouris, D. (2012). Magnetic resonance imaging and magnetic resonance spectroscopy cloud computing framework. [IJRQEH]. *International Journal of Reliable and Quality E-Healthcare*, *1*(4), 1–12. doi:10.4018/ijrqeh.2012100101

Turner, H., White, J., Reed, J., Galindo, J., Porter, A., Marathe, M., & Gokhale, A. et al. (2013). Building a cloud-based mobile application testbed. In *IT policy and ethics: Concepts, methodologies, tools, and applications* (pp. 879–899). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2919-6.ch040

Turner, H., White, J., Reed, J., Galindo, J., Porter, A., Marathe, M., & Gokhale, A. et al. (2013). Building a cloud-based mobile application testbed. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 382–403). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch018

Tusa, F., Paone, M., & Villari, M. (2012). CLEVER: A cloud middleware beyond the federation. In M. Villari, I. Brandic, & F. Tusa (Eds.), *Achieving federated and self-manageable cloud infrastructures: Theory and practice* (pp. 219–241). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-1631-8.ch012

Udoh, E. (2012). Technology acceptance model applied to the adoption of grid and cloud technology.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(1), 1–20. doi:10.4018/jghpc.2012010101

Vannoy, S. A. (2011). A structured content analytic assessment of business services advertisements in the cloud-based web services marketplace.[IJDTIS]. *International Journal of Dependable and Trustworthy Information Systems*, *2*(1), 18–49. doi:10.4018/jdtis.2011010102

Vaquero, L. M., Cáceres, J., & Morán, D. (2011). The challenge of service level scalability for the cloud.[IJCAC]. *International Journal of Cloud Applications and Computing*, *1*(1), 34–44. doi:10.4018/ijcac.2011010103

Vaquero, L. M., Cáceres, J., & Morán, D. (2013). The challenge of service level scalability for the cloud. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 37–48). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch003

Venkatraman, R., Venkatraman, S., & Asaithambi, S. P. (2013). A practical cloud services implementation framework for e-businesses. In K. Tarnay, S. Imre, & L. Xu (Eds.), *Research and development in e-business through service-oriented solutions* (pp. 167–198). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-4181-5.ch008

Venkatraman, S. (2013). Software engineering research gaps in the cloud.[JITR]. *Journal of Information Technology Research*, *6*(1), 1–19. doi:10.4018/jitr.2013010101

Vijaykumar, S., Rajkarthick, K. S., & Priya, J. (2012). Innovative business opportunities and smart business management techniques from green cloud TPS.[IJABIM]. *International Journal of Asian Business and Information Management*, *3*(4), 62–72. doi:10.4018/jabim.2012100107

Wang, C., Lam, K. T., & Kui Ma, R. K. (2012). A computation migration approach to elasticity of cloud computing. In J. Abawajy, M. Pathan, M. Rahman, A. Pathan, & M. Deris (Eds.), *Network and traffic engineering in emerging distributed computing applications* (pp. 145–178). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1888-6.ch007

Wang, D., & Wu, J. (2014). Carrier-grade distributed cloud computing: Demands, challenges, designs, and future perspectives. In H. Mouftah & B. Kantarci (Eds.), *Communication infrastructures for cloud computing* (pp. 264–281). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4522-6.ch012

Wang, H., & Philips, D. (2012). Implement virtual programming lab with cloud computing for web-based distance education. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 95–110). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch007

Warneke, D. (2013). Ad-hoc parallel data processing on pay-as-you-go clouds with nephele. In A. Loo (Ed.), *Distributed computing innovations for business, engineering, and science* (pp. 191–218). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2533-4.ch010

Wei, Y., & Blake, M. B. (2013). Adaptive web services monitoring in cloud environments.[IJWP]. *International Journal of Web Portals*, *5*(1), 15–27. doi:10.4018/jwp.2013010102

White, S. C., Sedigh, S., & Hurson, A. R. (2013). Security concepts for cloud computing. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 116–142). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch005

Williams, A. J. (2013). The role of emerging technologies in developing and sustaining diverse suppliers in competitive markets. In *Enterprise resource planning: Concepts, methodologies, tools, and applications* (pp. 1550–1560). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-4153-2.ch082

Williams, A. J. (2013). The role of emerging technologies in developing and sustaining diverse suppliers in competitive markets. In J. Lewis, A. Green, & D. Surry (Eds.), *Technology as a tool for diversity leadership: Implementation and future implications* (pp. 95–105). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2668-3.ch007

Wilson, L., Goh, T. T., & Wang, W. Y. (2012). Big data management challenges in a meteorological organisation.[IJEA]. *International Journal of E-Adoption*, *4*(2), 1–14. doi:10.4018/jea.2012040101

Wu, R., Ahn, G., & Hu, H. (2012). Towards HIPAA-compliant healthcare systems in cloud computing.[IJCMAM]. *International Journal of Computational Models and Algorithms in Medicine*, *3*(2), 1–22. doi:10.4018/jcmam.2012040101

Xiao, J., Wang, M., Wang, L., & Zhu, X. (2013). Design and implementation of C-iLearning: A cloud-based intelligent learning system.[IJDET]. *International Journal of Distance Education Technologies*, *11*(3), 79–97. doi:10.4018/jdet.2013070106

Xing, R., Wang, Z., & Peterson, R. L. (2011). Redefining the information technology in the 21st century.[IJSITA]. *International Journal of Strategic Information Technology and Applications*, *2*(1), 1–10. doi:10.4018/jsita.2011010101

Xu, L., Huang, D., Tsai, W., & Atkinson, R. K. (2012). V-lab: A mobile, cloud-based virtual laboratory platform for hands-on networking courses.[IJCBPL]. *International Journal of Cyber Behavior, Psychology and Learning*, *2*(3), 73–85. doi:10.4018/ijcbpl.2012070106

Xu, Y., & Mao, S. (2014). Mobile cloud media: State of the art and outlook. In J. Rodrigues, K. Lin, & J. Lloret (Eds.), *Mobile networks and cloud computing convergence for progressive services and applications* (pp. 18–38). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4781-7.ch002

Xu, Z., Yan, B., & Zou, Y. (2013). Beyond hadoop: Recent directions in data computing for internet services. In S. Aljawarneh (Ed.), *Cloud computing advancements in design, implementation, and technologies* (pp. 49–66). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1879-4.ch004

Yan, Z. (2014). Trust management in mobile cloud computing. In *Trust management in mobile environments: Autonomic and usable models* (pp. 54–93). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4765-7.ch004

Yang, D. X. (2012). QoS-oriented service computing: Bringing SOA into cloud environment. In X. Liu & Y. Li (Eds.), *Advanced design approaches to emerging software systems: Principles, methodologies and tools* (pp. 274–296). Hershey, PA: Information Science Reference; doi:10.4018/978-1-60960-735-7.ch013

Yang, H., Huff, S. L., & Tate, M. (2013). Managing the cloud for information systems agility. In A. Bento & A. Aggarwal (Eds.), *Cloud computing service and deployment models: Layers and management* (pp. 70–93). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2187-9.ch004

Yang, M., Kuo, C., & Yeh, Y. (2011). Dynamic rightsizing with quality-controlled algorithms in virtualization environments.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *3*(2), 29–43. doi:10.4018/jghpc.2011040103

Yang, X. (2012). QoS-oriented service computing: Bringing SOA into cloud environment. In *Grid and cloud computing: Concepts, methodologies, tools and applications* (pp. 1621–1643). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0879-5.ch706

Yang, Y., Chen, J., & Hu, H. (2012). The convergence between cloud computing and cable TV.[IJTD]. *International Journal of Technology Diffusion*, *3*(2), 1–11. doi:10.4018/jtd.2012040101

Yassein, M. O., Khamayseh, Y. M., & Hatamleh, A. M. (2013). Intelligent randomize round robin for cloud computing.[IJCAC]. *International Journal of Cloud Applications and Computing*, *3*(1), 27–33. doi:10.4018/ijcac.2013010103

Yau, S. S., An, H. G., & Buduru, A. B. (2012). An approach to data confidentiality protection in cloud environments.[IJWSR]. *International Journal of Web Services Research*, *9*(3), 67–83. doi:10.4018/jwsr.2012070104

Yu, W. D., Adiga, A. S., Rao, S., & Panakkel, M. J. (2012). A SOA based system development methodology for cloud computing environment: Using uhealthcare as practice.[IJEHMC]. *International Journal of E-Health and Medical Communications*, *3*(4), 42–63. doi:10.4018/jehmc.2012100104

Yu, W. D., & Bhagwat, R. (2011). Modeling emergency and telemedicine heath support system: A service oriented architecture approach using cloud computing. [IJEHMC]. *International Journal of E-Health and Medical Communications*, *2*(3), 63–88. doi:10.4018/jehmc.2011070104

Yu, W. D., & Bhagwat, R. (2013). Modeling emergency and telemedicine health support system: A service oriented architecture approach using cloud computing. In J. Rodrigues (Ed.), *Digital advances in medicine, e-health, and communication technologies* (pp. 187–213). Hershey, PA: Medical Information Science Reference; doi:10.4018/978-1-4666-2794-9.ch011

Yuan, D., Lewandowski, C., & Zhong, J. (2012). Developing a private cloud based IP telephony laboratory and curriculum. In L. Chao (Ed.), *Cloud computing for teaching and learning: Strategies for design and implementation* (pp. 126–145). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-0957-0.ch009

Yuvaraj, M. (2014). Cloud libraries: Issues and challenges. In S. Dhamdhere (Ed.), *Cloud computing and virtualization technologies in libraries* (pp. 316–338). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4631-5.ch018

Zaman, M., Simmers, C. A., & Anandarajan, M. (2013). Using an ethical framework to examine linkages between "going green" in research practices and information and communication technologies. In B. Medlin (Ed.), *Integrations of technology utilization and social dynamics in organizations* (pp. 243–262). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-1948-7.ch015

Zapata, B. C., & Alemán, J. L. (2013). Security risks in cloud computing: An analysis of the main vulnerabilities. In D. Rosado, D. Mellado, E. Fernandez-Medina, & M. Piattini (Eds.), *Security engineering for cloud computing: Approaches and tools* (pp. 55–71). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2125-1.ch004

Zapata, B. C., & Alemán, J. L. (2014). Security risks in cloud computing: An analysis of the main vulnerabilities. In *Software design and development: Concepts, methodologies, tools, and applications* (pp. 936–952). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-4301-7.ch045

Zardari, S., Faniyi, F., & Bahsoon, R. (2013). Using obstacles for systematically modeling, analysing, and mitigating risks in cloud adoption. In I. Mistrik, A. Tang, R. Bahsoon, & J. Stafford (Eds.), *Aligning enterprise, system, and software architectures* (pp. 275–296). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2199-2.ch014

Zech, P., Kalb, P., Felderer, M., & Breu, R. (2013). Threatening the cloud: Securing services and data by continuous, model-driven negative security testing. In S. Tilley & T. Parveen (Eds.), *Software testing in the cloud: Perspectives on an emerging discipline* (pp. 280–304). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2536-5.ch013

Zhang, F., Cao, J., Cai, H., & Wu, C. (2011). Provisioning virtual resources adaptively in elastic compute cloud platforms.[IJWSR]. *International Journal of Web Services Research*, *8*(3), 54–69. doi:10.4018/jwsr.2011070103

Zhang, G., Li, C., Xue, S., Liu, Y., Zhang, Y., & Xing, C. (2012). A new electronic commerce architecture in the cloud.[JECO]. *Journal of Electronic Commerce in Organizations*, *10*(4), 42–56. doi:10.4018/jeco.2012100104

Zhang, J., Yao, J., Chen, S., & Levy, D. (2011). Facilitating biodefense research with mobile-cloud computing.[IJSSOE]. *International Journal of Systems and Service-Oriented Engineering*, *2*(3), 18–31. doi:10.4018/jssoe.2011070102

Zhang, J., Yao, J., Chen, S., & Levy, D. (2013). Facilitating biodefense research with mobile-cloud computing. In D. Chiu (Ed.), *Mobile and web innovations in systems and service-oriented engineering* (pp. 318–332). Hershey, PA: Information Science Reference; doi:10.4018/978-1-4666-2470-2.ch017

Zheng, S., Chen, F., Yang, H., & Li, J. (2013). An approach to evolving legacy software system into cloud computing environment. In X. Yang & L. Liu (Eds.), *Principles, methodologies, and service-oriented approaches for cloud computing* (pp. 207–229). Hershey, PA: Business Science Reference; doi:10.4018/978-1-4666-2854-0.ch009

Zhou, J., Athukorala, K., Gilman, E., Riekki, J., & Ylianttila, M. (2012). Cloud architecture for dynamic service composition.[IJGHPC]. *International Journal of Grid and High Performance Computing*, *4*(2), 17–31. doi:10.4018/jghpc.2012040102

# Compilation of References

Adluru, P., Datla, S. S., & Zhang, X. (2015). *Hadoop Eco System for Big Data Security and Privacy*. Long Island Systems, Applications and Technology Conference (LISAT). doi:10.1109/LISAT.2015.7160211

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, *63*(1), 135–149. doi:10.1016/j.cie.2012.02.002

Alam, G. M., Hoque, K. E., & Oke, O. K. (2010). Quest for a better operation system in education: Privatization and teacher educationalization or voucherilization glimpsing from consumer and product perspectives. *African Journal of Business Management*, *4*(6), 1202–1214.

Alfelor, R., Carr, G., & Fateh, M. (2001). Track degradation assessment using gage restraint measurements. *Transportation Research Record: Journal of the Transportation Research Board*, *1742*, 68–77. doi:10.3141/1742-09

Ali, M. M. (2013). Role of data mining in education sector. *International Journal of Computer Science and Mobile Computing*, *2*(4), 374–383.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*, *2*(3), 87–93. doi:10.1016/j.bdr.2015.04.001

Aloui, A., & Touzi, A. G. (2015). A Fuzzy Ontology-Based Platform for Flexible Querying. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(3), 12–26. doi:10.4018/IJSSMET.2015070102

Andrade, A. R., & Teixeira, P. F. (2011). Uncertainty in rail-track geometry degradation: Lisbon-Oporto line case study. *Journal of Transportation Engineering*, *137*(3), 193–200. doi:10.1061/(ASCE)TE.1943-5436.0000206

Andrade, A. R., & Teixeira, P. F. (2013). Unplanned-maintenance needs related to rail track geometry.*Proceedings of the ICE - Transport*.

*Apache Hadoop Vulnerabilities*. Retrieved from https://www.cvedetails.com/vulnerability-list/vendor_id-45/product_id-22215/Apache-Hadoop.html

Aqdam, B. A., & Bonab, M. B. (2014). Possible Evaluation Replacement of Supporting Decision Systems in Administrative Education System in Sarab Town. Academic Press.

Ashutosh, A. (2012). *Website.* Retrieved from http://www.forbes.com/sites/ciocentral/2012/07/05/best-practices-for-managing-big-data/

Azar, A.T., & Hassanien, A.E. (2014). Dimensionality reduction of medical big data using neural-fuzzy classifier. *Soft computing, 19*(4), 1115-1127. DOI 10.1007/s00500-014-1327-4

Backpropagation. (2015). In *Wikipedia.* Retrieved September 9, 2015, from https://en.wikipedia.org/wiki/Backpropagation

Bai, X. (2012). Predicting consumer sentiments from online text. *Decision Support Systems*, *50*(4), 732–742. doi:10.1016/j.dss.2010.08.024

Bernbaum, M., & Moses, K. (2011). *Education Management Information Systems*. Education Policy and Data Center.

Bhanu, S. K., & Tripathy, B. K. (2016). Rough Set Based Similarity Measures for Data Analytics in Spatial Epidemiology. *International Journal of Rough Sets and Data Analysis*, *3*(1), 114–123. doi:10.4018/IJRSDA.2016010107

*Big Data Analytics and Predictive Analytics*. (2015). Retrieved August 30, 2015, from, http://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics

*Big Data Strategy - Issues Paper.* (2013). Commonwealth of Australia.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Brown, C. A. (2014). *Globalization, International Education Policy and Local Policy Formation: Voices from the Developing World*. Springer.

Budai, G., Huisman, D., & Dekker, R. (2005). Scheduling preventive railway maintenance activities. *The Journal of the Operational Research Society*, *57*(9), 1035–1044. doi:10.1057/palgrave.jors.2602085

Butler, M. (2013). *Predictive Analytics in Business*. Butler Analytics.

Buytendijk, F., & Trepanier, L. (2010). *Predictive Analytics: Bringing the tools to the data*. Redwood Shores, CA: Oracle Corporation.

Caetano, L., & Teixeira, P. F. (2013). Availability approach to optimizing railway track renewal operations. *Journal of Transportation Engineering*, *139*(9), 941–948. doi:10.1061/(ASCE)TE.1943-5436.0000575

Calvanese, D., Giese, M., Haase, P., Horrocks, I., Hubauer, T., Ioannidis, Y., . . . Koubarakis, M. (2013). Optique: OBDA solution for big data. In The Semantic Web: ESWC 2013 Satellite Events (pp. 293-295). Springer Berlin Heidelberg.

Carl, W. O., & Dan, V. (2012). *White Paper, Big Data: trends, strategies, and SAP technology*. IDC.

Cheng, Y. H., Hai-Wei, L., & Chen, Y. S. (2006, June). Implementation of a back-propagation neural network for demand forecasting in a supply chain-A practical case study. In *Service Operations and Logistics, and Informatics, 2006. SOLI'06. IEEE International Conference on* (pp. 1036-1041). IEEE.

Chen, M., Mao, S., & Liu, Y. (2009). Big Data: A Survey. *Springer-. Mobile Networks and Applications*, *19*(2), 171–209. doi:10.1007/s11036-013-0489-0

CSA. (2012). *Top Ten Big Data Security and Privacy Challenges*. Retrieved from http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf

Cuesta, C. E., Martínez-Prieto, M. A., & Fernández, J. D. (2013). Towards an architecture for managing big semantic data in real-time. In *Software Architecture* (pp. 45–53). Springer Berlin Heidelberg.

Daniel, B. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*. doi:10.1111/bjet.12230

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. doi:10.1145/1327452.1327492

Deepak, D., & John, S. J. (2016). Information Systems on Hesitant Fuzzy Sets. *International Journal of Rough Sets and Data Analysis*, *3*(1), 55–70.

Ding, X., Liu, B., & Yu, P. (2008). A holistic lexicon approach to opinion mining. In *Proceedings of the Internationl Conference on Web Search and Web Data Mining*.

Ding, X., Liu, B., & Philip, S. Y. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08)*. doi:10.1145/1341531.1341561

Drigas, A.S., & Leliopoulos, P. (2014). The Use of Big Data in Education. *International Journal of Computer Science Issues*, *11*(5).

Elkan, C. (2013). *Predictive analytics and data mining*. Retrieved from http://www.cseweb.ucsd.edu

Ellison, R. (2004). *Scoping exercise for possible work by DFID on Education Management Information Systems.* Academic Press.

El-sibaie, M., & Zhang, Y. J. (2004). Objective track quality indices. *Transportation Research Record: Journal of the Transportation Research Board*, *1863*, 81–87. doi:10.3141/1863-11

Fernández, A., Peralta, D., Benitez, J. M., & Herrera, F. (2014). E-learning and educational data mining in cloud computing: An overview. *International Journal of Learning Technology*, *9*(1), 25. doi:10.1504/IJLT.2014.062447

Gazder, U. & Ratrout, N. T. (2015). A new logit-artificial neural network ensemble for mode choice modeling: a case study for border transport. *Journal of Advanced Transportation.* DOI: 10.1002/atr.1306

Gerhardt, Griffin, & Klemann. (2012). *Unlocking Values in the Fragmented World of Big Data Analytics*. Cisco Internet Business Solutions Group.

Ghallab, S. A., Badr, N. L., Salem, A. B., & Tolba, M. F. (2014). Strictness petroleum prediction system based on fussy model. *International Journal of Service Science, Management, Engineering, and Technology*, *5*(4), 44–65. doi:10.4018/ijssmet.2014100104

Gil, Y., & Ratnakar, V. (2002, May). A Comparison of (Semantic) Markup Languages. In *FLAIRS Conference* (pp. 413-418).

Guler, H., Jovanovic, S., & Evren, G. (2011). Modelling railway track geometry deterioration.*Proceedings of the ICE - Transport*.

Halper, F. (2014). *Predictive Analytics for Business Advantage*. TDWI Research.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kauffman.

Harrison, G. (2012). Outdoor Program Administration: Principles and Practices. *Business Economics (Cleveland, Ohio)*, 406.

Harzog, B. (2015). *Big data operations management*. Retrieved from https://www.virtualizationpractice.com/big-data-operations-management-19219/

Hassanien, A. E., Azar, A. T., Snasel, V., Kacprzyk, J., & Abawajy, J. H. (2015). *Big Data in Complex Systems: Challenges and Opportunities*. Springer.

Hensher, D. A. (2000). Transport economics: A personal view. *Journal of Advanced Transportation*, *34*(1), 65–106. doi:10.1002/atr.5670340105

Hoffmann, A. O. I., & Birnbrich, C. (2012). The impact of fraud prevention on bank-customer relationships. *International Journal of Bank Marketing*, *30*(5), 390–394. doi:10.1108/02652321211247435

Holley, K., Sivakumar, G., & Kannan, K. (2014). Enrichment Patterns for Big Data. *IEEE International Congress on Big Data*.

Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and challenges of big data computing in health science. *Big Data Research.*, *2*(1), 2–11.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*.

IDC Report. (2011). *Extracting Value from Chaos*. Retrieved from http://idcdocserv.com/1142

Inukollu, V. N., Arsi, S., & Ravuri, S. R. (2014). Security Issues Associated with Big Data in Cloud Computing. *International Journal of Network Security & Its Applications*, *6*(3), 45–56. doi:10.5121/ijnsa.2014.6304

Islek, I., & Oguducu, S. G. (2015, June). A retail demand forecasting model based on data mining techniques. In *Industrial Electronics (ISIE), 2015 IEEE 24th International Symposium on* (pp. 55-60). IEEE doi:10.1109/ISIE.2015.7281443

Ixia Report. (2014). *Designing network for the big data superstorm*. Author.

Jagadish, H. V. (2015). Big Data and Science: Myths and Reality. *Big Data Research*, *2*(2), 49–52. doi:10.1016/j.bdr.2015.01.005

Jam, Khanli, Akbari, & Javan. (2014). *A Survey on Security of Hadoop*. 4th International eConference on Computer and Knowledge Engineering (ICCKE).

Jameson, M. (2006). *Zambia Research and Education Networks (ZAMREN) Formation as an Emerging Academic Networking Strategy*. Author.

Jaseena, K.U., & David, J. M. (2014). Issues, Challenges and Solutions: Big Data Mining. *CS & IT-CSCP, 4*(13), 131-140.

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, *2*(2), 59–64. doi:10.1016/j.bdr.2015.01.006

Johnson, J.A. (2014). The Ethics of Big Data in Higher Education. *International Review of Information Ethics*.

Jovanovic, S. (2004). Railway track quality assessment and related decision making. *IEEE International Conference on Systems, Man and Cybernetics*. doi:10.1109/ICSMC.2004.1400992

Katal, A., Wazid, M., & Goudar, R. H. (2013). *Big Data: Issues, Challenges, Tools and Good Practices*. Sixth International Conference on Contemporary Computing (IC3). doi:10.1109/IC3.2013.6612229

Kats, Y. (2013). Learning Management Systems and Instructional Design: Best Practices in Online Education. *Education*, 367.

Kaushik, M., & Jain, A. (2014). Challenges to Big Data Security and Privacy. *International Journal of Computer Science and Information Technologies*, *5*(3), 3042–3043.

Lämmel, R. (2008). Google's MapReduce programming model. *Science of Computer Programming*, *70*(1), 1–30. doi:10.1016/j.scico.2007.07.001

Larsson, D. (2004). *A study of the track degradation process related to changes in railway traffic.* (PhD thesis). Lulea University of Technology.

Lincoln, C., Reiners, T., Srivastava, H. S., & Linh, D. N. K. (2013). *Using sentiment analysis to improve decisions in demand-driven supply chains*. Anzam.

Liu, Z., Li, H., & Miao, G. (2010, August). MapReduce-based backpropagation neural network over large scale mobile data. In *Natural Computation (ICNC), 2010 Sixth International Conference on* (Vol. 4, pp. 1726-1730). IEEE. doi:10.1109/ICNC.2010.5584323

Liu, R., Xu, P., & Wang, F. (2010). Research on a short-range prediction model for track irregularity over small track lengths. *Journal of Transportation Engineering*, *136*(12), 1085–1091. doi:10.1061/(ASCE)TE.1943-5436.0000192

Lu, J., Ma, L., Zhang, L., Brunner, J. S., Wang, C., Pan, Y., & Yu, Y. (2007, September). SOR: a practical system for ontology storage, reasoning and search. In Proceedings of the 33rd international conference on Very large data bases (pp. 1402-1405). VLDB Endowment.

MacHardy, Z., & Pardos, Z. A. (2015). Evaluating The Relevance of Educational Videos using BKT and Big Data.*Proceedings of the 46th ACM Technical Symposium on Computer Science Education*.

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, *41*(17), 7764–7775. doi:10.1016/j.eswa.2014.05.045

Mazloumi, E., Moridpour, S., Currie, G., & Rose, G. (2012). Exploring the Value of Traffic Flow Data in Bus Travel Time Prediction. *Journal of Transportation Engineering*, *138*(4), 436–446.

Mazloumi, E., Rose, G., Currie, G., & Moridpour, S. (2011). Prediction Intervals to Account for Uncertainties in Neural Network Predictions: Methodology and Application in Bus Travel Time Prediction. *Engineering Applications of Artificial Intelligence*, *24*(3), 534–542.

McKinsey, R. (2011). *Advancing Personalized Education*. Computing Community Consortium.

Möller, R., Neuenstadt, C., Özçep, Ö. L., & Wandelt, S. (2013). Advances in accessing big data with expressive ontologies. In KI 2013: Advances in Artificial Intelligence (pp. 118-129). Springer Berlin Heidelberg.

Moridpour, S., & Hesami, R. (2015). Degradation and performance specification of Melbourne tram tracks. *The 3rd International Conference on Transportation Information and Safety (ICTIS)*.

Nolan, C. J. P., Ayres, D. A., & Mckinnon, D. H. (1997). *An operational model for the implementation of computerised school information systems*. Massey University. doi:10.1007/978-0-387-35090-5_12

*NUANCE*. (2008). Retrieved from http://www.ubuntunet.net/zamren

*Operations Management*. (2015). Retrieved November 18, 2015, from https://ids355.wikispaces.com/Ch.+3+Forecasting

Oracle White Paper. (2014). *Information Management and Big Data – A Reference Architecture*. Author.

Panneerselvam, R. (2012). Production and Operations Management. *Business Economics (Cleveland, Ohio)*, 720.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational data mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Peason, T., & Wegener, R. (2013). *Big Data: The organizational Challenge*. Bian and Company.

Peng, F., & Ouyang, Y. (2012). Track maintenance production team scheduling in railroad networks. *Transportation Research Part B: Methodological*, *46*(10), 1474–1488. doi:10.1016/j.trb.2012.07.004

Peng, F., & Ouyang, Y. (2013). Optimal clustering of railroad track maintenance jobs. *Computer-Aided Civil and Infrastructure Engineering*, *29*(4), 235–247. doi:10.1111/mice.12036

Pérez, J., Arenas, M., & Gutierrez, C. (2006, November). Semantics and Complexity of SPARQL. In *International semantic web conference* (Vol. 4273, pp. 30-43).

Prakash, B. R., Hanumanthappa, M., & Kavitha, V. (2014). Big Data in Educational Data Mining and Learning Analytics. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(2).

Predictive Analytics. (2015). In *Wikipedia*. Retrieved August 27, 2015, from https://en.wikipedia.org/wiki/Predictive_analytics

Quiroga, L., & Schnieder, E. (2013). Railway systems. In H. Czichos (Ed.), *Handbook of Technical Diagnostics*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-25850-3_26

Qu, Z. (2011). Semantic processing on big data. Advances in Multimedia, Software. *Engineering and Computing*, *2*, 43–48.

Ramasubramanian, P., Iyakutti, K., & Thangavelu, P. (2009). Enhanced data mining analysis in higher educational system using rough set theory. *African Journal of Mathematics and Computer Science Research*, *2*(9), 184–188.

Ranganathan, A., & Liu, Z. (2006, November). Information retrieval from relational databases using semantic queries. In *Proceedings of the 15th ACM international conference on Information and knowledge management*(pp. 820-821). ACM. doi:10.1145/1183614.1183747

RDF Working Group. (2014). *Resource Description Framework (RDF)*. Retrieved from www.w3.org/RDF/

Rock, R. (2011). *Data Management for BI: Big Data, Bigger Insight, Superior Performance*. Aberdeen Group, a Harte-Hanks Company.

Russom, P. (2013). *TDWI Best Practices Report: Managing Big Data*. Cloudera, Dell Software, Oracle, Pentaho, SAP.

Ryan, T., & Lee, Y. C. (2015). Multi-tier resource allocation for data-intensive computing. *Big Data Research*, *2*(3), 110–116. doi:10.1016/j.bdr.2015.03.001

Sadeghi, J. (2010). Development of railway track geometry indexes based on statistical distribution of geometry data. *Journal of Transportation Engineering*, *136*(8), 693–700. doi:10.1061/(ASCE)0733-947X(2010)136:8(693)

Sadeghi, J., & Askarinejad, H. (2010). Development of improved railway track degradation models. *Structure and Infrastructure Engineering*, *6*(6), 675–688. doi:10.1080/15732470801902436

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE. doi:10.1109/CTS.2013.6567202

Sarhani, M., & El Afia, A. (2014, November). Intelligent system based support vector regression for supply chain demand forecasting. In *Complex Systems (WCCS), 2014 Second World Conference on* (pp. 79-83). IEEE. doi:10.1109/ICoCS.2014.7060941

Seay, C. (2015). Using Hadoop on the Mainframe: A Big Solution for the Challenges of Big Data. IEEE International Information Technology, 765-769.

Shabeera, T. P., & Madhu Kumar, S. D. (2015). Optimizing virtual machine allocation in MapReduce cloud for improved data locality. *International Journal of Big Data Intelligence.*, *2*(1), 2–8. doi:10.1504/IJBDI.2015.067563

Shafahi, Y., & Hakhamaneshi, R. (2009). Application of maintenance management model for Iranian railways based on the Markov chain probabilistic dynamic modeling. *Scientia Iranica*, *16*, 87–97.

Singhal, A. (2012, May 16). Introducing the knowledge graph: things, not strings. *Official Google Blog*. Message posted to https://googleblog.blogspot.in/2012/05/introducing-knowledge-graph-things-not.html

Sivaraman, E. (2014). High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing Using Hadoop.*IEEE International Conference on Intelligent Computing Application*. doi:10.1109/ICICA.2014.16

Song, Y., Alatorre, G., Mandagere, N., & Singh, A. (2013, June). Storage mining: where IT management meets big data analytics. In *Big Data (BigData Congress), 2013 IEEE International Congress on* (pp. 421-422). IEEE. doi:10.1109/BigData.Congress.2013.66

Sqalli, M. H., Al-Saeedi, M., Binbeshr, F., & Siddiqui, M. (2012). UCloud: A simulated Hybrid Cloud for a university environment.*IEEE International Conference on Cloud Computing*. doi:10.1109/CloudNet.2012.6483678

Srivastava, U., & Gopalkrishnan, S. (2015). Impact of Big Data Analytics on Banking Sector: Learning for Indian Bank. *Big Data. Cloud and Computing Challenges.*, *50*, 643–652.

Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., & Van de Walle, R. (2012, November). Adding Realtime Coverage to the Google Knowledge Graph. In *11th International Semantic Web Conference (ISWC 2012)*.

Sun, Y., & Han, J. (2012). Mining Heterogeneous Information Networks: A Structural Analysis Approach. *ACM SIGKDD Exploration Newsletter.*, *14*(2), 20–28. doi:10.1145/2481244.2481248

Sun, Z., & Fox, G. (2012, January). Study on parallel SVM based on MapReduce. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*.

Support Vector Machine. (2015). In *Wikipedia.* Retrieved September 15, 2015, from https://en.wikipedia.org/wiki/Support_vector_machine

Tang, S., Boyles, S. D., & Jiang, N. (2015). High-speed rail cost recovery time based on an integer optimization model. *Journal of Advanced Transportation*, *49*(5), 634–647. doi:10.1002/atr.1294

Thakar, P., & Mehta, A., Manisha. (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *International Journal of Computers and Applications*, *110*(15).

Tiwari, P. K., & Joshi, S. (2015). Data security for software as a service. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(3), 47–63. doi:10.4018/IJSSMET.2015070104

UNESCO. (2011). Transforming Education: The Power of ICT Policies. *Education*, 237.

Vohries. (2013). *The Big Deal about Big Data: What's Inside – Structured, Unstructured, and Semi-Structured Data.* [Web blog post]. Data Magnum.

Wahi, A. K., Medury, Y., & Misra, R. K. (2014). Social Media: The core of enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *5*(3), 1–15. doi:10.4018/ijssmet.2014070101

Wahi, A. K., Medury, Y., & Misra, R. K. (2015). Big Data: Enabler or Challenge for Enterprise 2.0. *International Journal of Service Science, Management, Engineering, and Technology*, *6*(2), 1–17. doi:10.4018/ijssmet.2015040101

Wang, X., Petkova, B., & Wood, L. C. (2014). Do suppliers benefit from supply chain sustainability programs? The case of Wal-Mart. In *Proceedings of the 12th ANZAM Operations, Supply Chain and Services Management Symposium*. The University of Auckland Business School.

Wang, B., & Xing, H. Y. (2011). *The application of cloud computing in education informatization. In Computer Science and Service System* (pp. 2673–2676). CSSS.

Wang, X., Sun, X., Cao, F., Ma, L., Kanellos, N., Zhang, K., & Yu, Y. et al. (2009). SMDM: Enhancing enterprise-wide master data management using semantic web technologies. *Proceedings of the VLDB Endowment*, *2*(2), 1594–1597. doi:10.14778/1687553.1687600

Warren, K. (2008). *Strategic management dynamics*. Hoboken, NJ: Wiley.

Wayne, R. (2011). *Creating an enterprise data strategy: managing data as corporate asset*. BeyeNetwork.

*What is Predictive Analytics ?* (2015). Retrieved August 27, 2015, from, http://www.predictiveanalyticsworld.com/predictive_analytics.php

Wigan, M. R., & Clarke, R. (2013). Big data's big unintended consequences. *Computer*, *46*(6), 46–53. doi:10.1109/MC.2013.195

Wood, L. C., Reiners, T., & Srivastava, H. S. (2013). *Expanding sales and operations planning using sentiment analysis: Demand and sales clarity from social media*. Anzam.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2014). *Sentiment analysis in supply chain management. In Encyclopedia of Business Analytics and Optimization*. Hershey, PA: IGI Global.

Woodward, M. (2012). 5 Supply chain predictions for 2013. *Supply Chain Management Review*.

World Wide Web Consortium. (2014). *Extensible Markup Language (XML)*. Retrieved from www.w3.org/XML

Wu, Z., Flintsch, G., Ferreira, A., & Picado-santos, L. (2012). Framework for multiobjective optimization of physical highway assets investments. *Journal of Transportation Engineering*, *138*(12), 1411–1421. doi:10.1061/(ASCE)TE.1943-5436.0000458

Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, *47*(3), 355–368. doi:10.1002/atr.193

Yarra Trams. (2015). *Facts and Figures, Yarra Trams.* Retrieved from http://www.yarratrams.com.au/about-us/who-we-are/facts-figures/

Yousefikia, M., Moridpour, S., Setunge, S., & Mazloumi, E. (2014). Modeling degradation of tracks for maintenance planning on a tram line. *Journal of Traffic and Logistics Engineering.*, *2*(2), 86–91. doi:10.12720/jtle.2.2.86-91

Yu, S., & Kak, S. (2012). *A survey of prediction using social media*. arXiv:1203.1647

Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720–734. doi:10.1109/TKDE.2010.269

Zakeri, J. A., & Shahriari, S. (2012). Developing a deterioration probabilistic model for rail wear. *International Journal of Traffic and Transportation Engineering*, *1*, 13–18.

Zhao, Z. (2015). *Visualization of HashStash with Qt*. (Masters Thesis). Brown University.

Zhao, J., Chan, A., Stirling, A., & Madelin, K. (2006). Optimizing policies of railway ballast tamping and renewal. *Transportation Research Record: Journal of the Transportation Research Board*, *1943*, 50–56. doi:10.3141/1943-07

Zitnik, M. (2012). Using sentiment analysis to improve business operations. *XRDS: Crossroads*, *18*(4), 42–43. doi:10.1145/2173637.2173655

# About the Contributors

**Manish Kumar** received his PhD degree on Data Management in Wireless Sensor Networks from Indian Institute of Information Technology, Allahabad India in 2011. He received his M.Tech in Computer Science from Birla Institute of Technology, Mesra (Ranchi) India. He is a professional member of IEEE and ACM. Currently he is working as an Assistant Professor in Department of Information Technology at Indian Institute of Information Technology, Allahabad India. His research interest includes data mining and warehousing, data management in wireless sensor networks and big data analytics. He has contributed in a number of books in the same areas and has many national and international publications in renowned journals and conferences.

* * *

**Vishal Bhatnagar** holds BTech, MTech and PhD in the engineering field. He has more than 13 years of teaching experience in various technical institutions. He is currently working as an Associate Professor in Computer Science in Engineering Department at Ambedkar Institute of Advanced Communication Technologies and Research (Government of Delhi), GGSIPU, Delhi, India. His research interests include database, advance database, data warehouse, data-mining, social network analysis and big data analytics. He has to his credit more than 50 research papers in various international/national journals and conferences.

**Ashish Maruti Gimekar** received his B.E. degree in Computer Engineering from Mumbai University. Currently, he is pursuing M.Tech with the specialization in Cyber Law and Information Security at Indian Institute of Information Technology, Allahabad, India. His major research interests include big data security and network security.

**Mahima Goyal** is pursuing her MTech in Information Security at GGSIPU, Delhi, India. Her research interests include database, data warehouse, data mining, business analysis and big data analysis.

**Reyhaneh Hesami** received her PhD in 2008. she is a skillful professional with diverse range of experience in academia and multidisciplinary industries including rail, road and steel. Reyhaneh has proven ability to combine, adapt and transfer knowledge, techniques and experiences from various disciplines to deliver robust, innovative and holistic solutions to the stakeholders.

**Arushi Jain** is doing her M.tech in Information Security at Ambedkar Institute of Advanced Communication Technologies and Research,New Delhi,India. Her research interest include database, data warehouse, data mining, big data, virtual reality.

**Harsh Jain** was born in Uttar Pradesh, India, in 1992. He received B.Tech degree in computer science and engineering from Uttar Pradesh Technical University, Uttar Pradesh, India, in 2014. He is currently pursuing M.Tech in software engineering from Indian Institute of Information Technology Allahabad. His current research work includes prediction of software error proneness using neural network and evolutionary algorithms. His areas of interest are data structure, algorithms, data mining and machine learning.

**Ehsan Mazloumi** received his PhD in Traffic and Transportation Engineering from Monash University. With more than 10 years industry experience, he has implemented many analytical solutions in various industries. Ehsan also pursues his research interests through collaborations with RMIT university.

**Jameson Mbale** is a Professor in the Department of Computer Science at the Copperbelt University, in Zambia. He is a Director of Research and Innovations. He teaches the following courses: Multimedia Computing and Web Service, Advanced Mobile Computing and Networking, and Telecommunications.

**Sara Moridpour** holds a Bachelor of Civil Engineering and Masters degree in Transportation Planning and Engineering from Sharif University of Technology, Iran. She received her PhD degree from Monash University. She has 9 years of work and research experience in the field of traffic and transport. Her main research interests

include on driving behaviour modelling and analysis, traffic policy analysis and management, transport network modelling and optimization, transport infrastructure asset management, traffic safety and application of artificial intelligence in traffic/ transport. She has been working in the School of Civil, Environmental and Chemical Engineering, RMIT University, from 2010.

**Prakash S** has 19 years of experience in Training & development, Education management, IT Infrastructure Management, both in academia and in IT Training industry. Currently he is with Dayananada Sagar University as Professor in Computer science & Engineering Department. He heads "Industry Integrated Learning Programs" and VMware Instructor Training Centre at the School of Engineering, Dayananada Sagar University, Bengaluru.His research area is on Mobile communications, cloud Computing and Big Data.

**Ganeshayya I. Shidaganti** is Assistant Professor in Dept. CSE, M.S.Ramaiah Institute of Technology. He has completed his Under Graduate (at BVBCET) and Masters (at MSRIT) in Computer Science and Engineering .He has 4 years of Teaching Experience. His main area of interest include studies related to Big data and its Applications. He is currently pursuing research in the area of Big Data and Analytics in Education System.

**Jaiteg Singh** is Associate Professor in School of Computer Sciences at Chitkara University, in Punjab (India). He has done his PhD in computer science. He has published 40 research paper in various national and international journals of repute included 6 books.

**Jaya Singh** received her B.Tech degree in Information Technology from Uttar Pradesh Technical University, Lucknow, India. She received her M.Tech degree in Computer Science and Engineering from SHIATS (Formerly Allahabad Agriculture Institute), Allahabad, India. Currently, she is a full-time Research Scholar at Indian Institute of Information Technology, Allahabad, India. Her major research interests include Network and Data security, Cryptography and Big Data security.

**Saravjeet Singh** is an Assistant Professor in the School of Computer Science at the Chitkara University, in Punjab(India). He has done his B tech and M tech in computer Science. He has 4.5 years of teaching experience. He teaches the follow-

ing courses: Programming Languages, Data Structures, Algorithm Design and Data Base Management systems. His area of research is semantic and spatial databases. He has published 3 research paper and one book on semantic databases.

**S. Venkatesan** is an Assistant Professor at the Indian Institute of Information Technology Allahabad. His current research interests include Mobile Agent Security, Cloud Computing Security & Privacy and Big Data Security.

# Index

## I



## L



## M



## N



## O



## P



## R



## S



## T

## V



## W



## X



## Z