VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY (Autonomous)

Department of Computer Science and Engineering

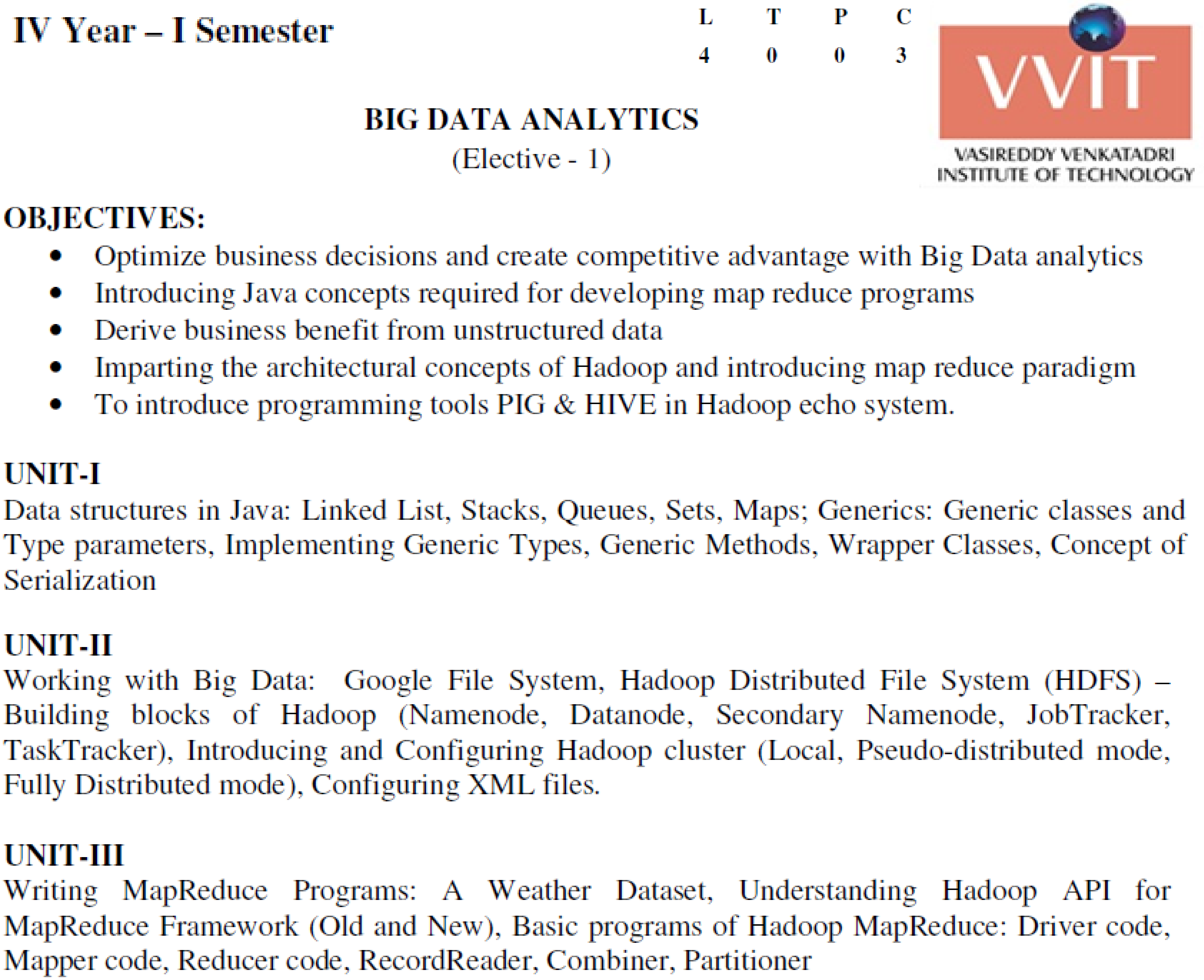
IV B.Tech – I Semester (Sections – C & D)

Big Data Analytics (Elective – I)

Table of Contents, Introduction & Overview of Subject

Outline of Today’s Session

- JNTUK Syllabus of Big Data Analytics
- What is Data?
- Data to Big Data Transformation
- What the need of Data Analytics?
- Data Analyst Vs. Data Scientist
- Popular Software Tools for Data Analytics
- Introduction to Big Data Analytics
- Revision of Java Programming Language and Discuss the need of Java for our subject.
- Discussion on Syllabus, Objectives & Outcomes.



IV Year – I Semester

L	T	P	C
4	0	0	3

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

BIG DATA ANALYTICS

(Elective - 1)

OBJECTIVES:

- Optimize business decisions and create competitive advantage with Big Data analytics
- Introducing Java concepts required for developing map reduce programs
- Derive business benefit from unstructured data
- Imparting the architectural concepts of Hadoop and introducing map reduce paradigm
- To introduce programming tools PIG & HIVE in Hadoop echo system.

UNIT-I

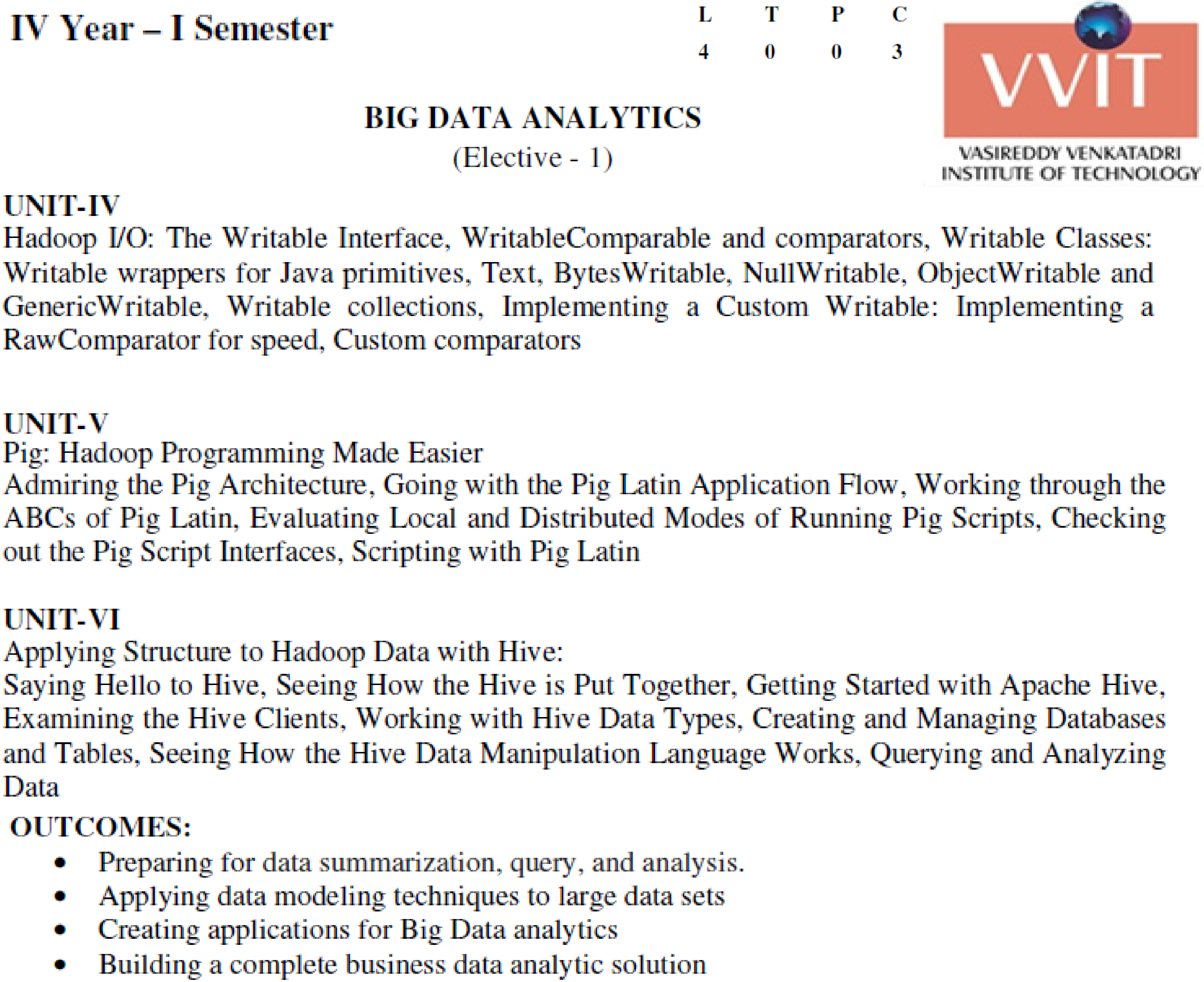
Data structures in Java: Linked List, Stacks, Queues, Sets, Maps; Generics: Generic classes and Type parameters, Implementing Generic Types, Generic Methods, Wrapper Classes, Concept of Serialization

UNIT-II

Working with Big Data: Google File System, Hadoop Distributed File System (HDFS) – Building blocks of Hadoop (Namenode, Datanode, Secondary Namenode, JobTracker, TaskTracker), Introducing and Configuring Hadoop cluster (Local, Pseudo-distributed mode, Fully Distributed mode), Configuring XML files.

UNIT-III

Writing MapReduce Programs: A Weather Dataset, Understanding Hadoop API for MapReduce Framework (Old and New), Basic programs of Hadoop MapReduce: Driver code, Mapper code, Reducer code, RecordReader, Combiner, Partitioner



IV Year – I Semester

L	T	P	C
4	0	0	3

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

BIG DATA ANALYTICS

(Elective - 1)

UNIT-IV

Hadoop I/O: The Writable Interface, WritableComparable and comparators, Writable Classes: Writable wrappers for Java primitives, Text, BytesWritable, NullWritable, ObjectWritable and GenericWritable, Writable collections, Implementing a Custom Writable: Implementing a RawComparator for speed, Custom comparators

UNIT-V

Pig: Hadoop Programming Made Easier

Admiring the Pig Architecture, Going with the Pig Latin Application Flow, Working through the ABCs of Pig Latin, Evaluating Local and Distributed Modes of Running Pig Scripts, Checking out the Pig Script Interfaces, Scripting with Pig Latin

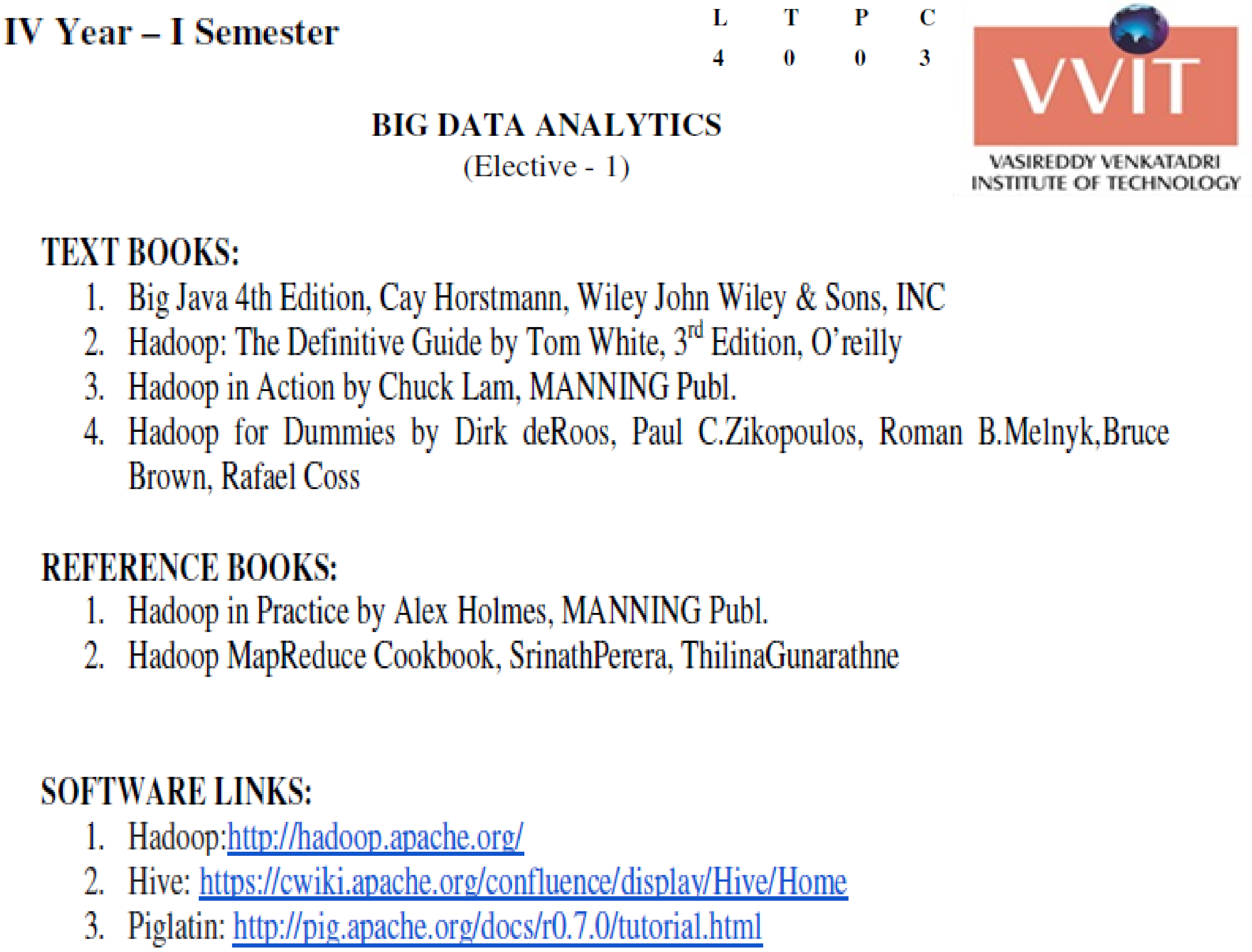
UNIT-VI

Applying Structure to Hadoop Data with Hive:

Saying Hello to Hive, Seeing How the Hive is Put Together, Getting Started with Apache Hive, Examining the Hive Clients, Working with Hive Data Types, Creating and Managing Databases and Tables, Seeing How the Hive Data Manipulation Language Works, Querying and Analyzing Data

OUTCOMES:

- Preparing for data summarization, query, and analysis.
- Applying data modeling techniques to large data sets
- Creating applications for Big Data analytics
- Building a complete business data analytic solution



IV Year – I Semester

L	T	P	C
4	0	0	3

VVIT
VASIREDDY VENKATADRI
INSTITUTE OF TECHNOLOGY

BIG DATA ANALYTICS

(Elective - 1)

TEXT BOOKS:

1. Big Java 4th Edition, Cay Horstmann, Wiley John Wiley & Sons, INC
2. Hadoop: The Definitive Guide by Tom White, 3rd Edition, O'reilly
3. Hadoop in Action by Chuck Lam, MANNING Publ.
4. Hadoop for Dummies by Dirk deRoos, Paul C.Zikopoulos, Roman B.Melnyk,Bruce Brown, Rafael Coss

REFERENCE BOOKS:

1. Hadoop in Practice by Alex Holmes, MANNING Publ.
2. Hadoop MapReduce Cookbook, SrinathPerera, ThilinaGunarathne

SOFTWARE LINKS:

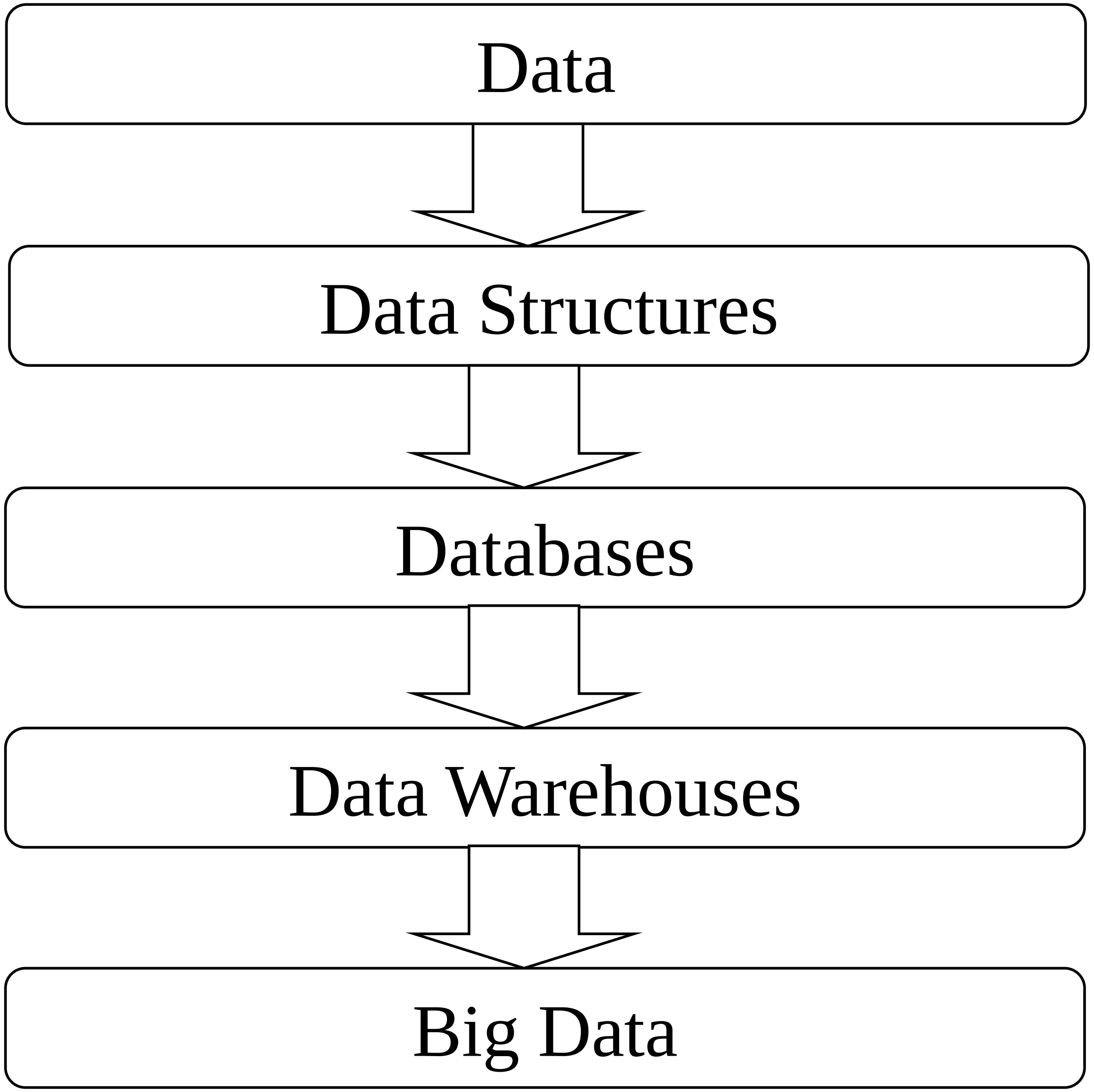
1. Hadoop:http://hadoop.apache.org/
2. Hive: https://cwiki.apache.org/confluence/display/Hive/Home
3. Piglatin: http://pig.apache.org/docs/r0.7.0/tutorial.html

What is Data?

Data are characteristics or information, usually numerical, that are collected through observation. In a more technical sense, data is a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum (singular of data) is a single value of a single variable.

Although the terms "data" and "information" are often used interchangeably, these terms have distinct meanings. In some popular publications, data is sometimes said to be transformed into information when it is viewed in context or in post-analysis.

Data to Big Data Transformation



Data

Data Structures

Databases

Data Warehouses

Big Data

What the need of Data Analytics?

Data Analytics is a new term for many people if you are also confused as to what is Data Analytics and what is it used for, then you’re at the right place.

As “Data Analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain.”

Data is extracted, acknowledged and bifurcated to identify and analyze behavioral data, techniques and patterns can be dynamic according to a particular business’s need or requirement.

Data Analytics is a broader term that has analysis as a subhead and analytics is basically the concepts used to do the analysis.

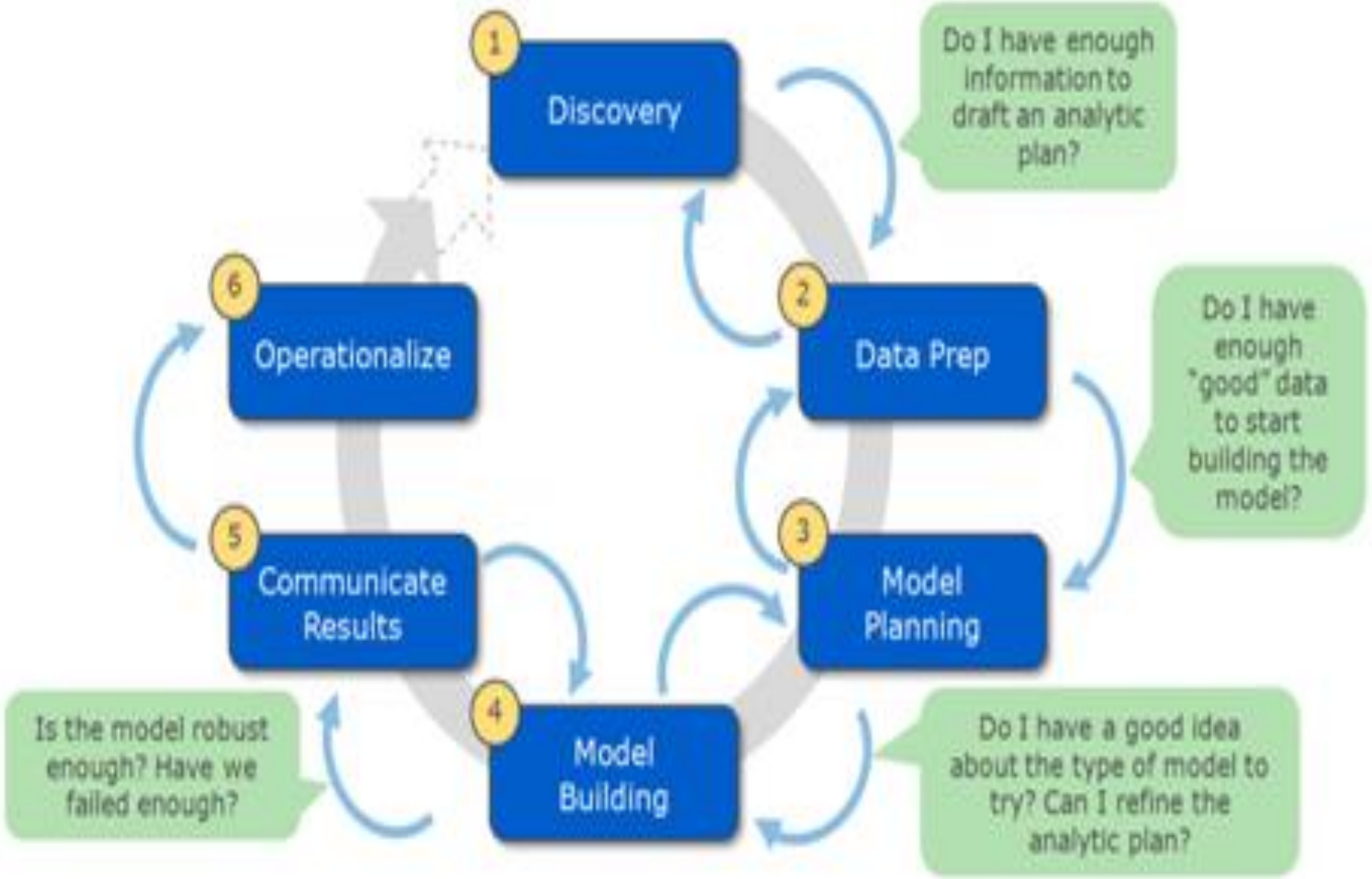
What the need of Data Analytics?

Data Analytics is needed in Business to Consumer applications (B2C).

Organizations collect data that they have gathered from customers, businesses, economy and practical experience.

Data is then processed after gathering and is categorized as per the requirement and analysis is done to study purchase patterns and etc.

Data Analytics Life Cycle


1
Discovery
Do I have enough information to draft an analytic plan?
2
Data Prep
Do I have enough "good" data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed enough?
5
Communicate Results
6
Operationalize

Data Analyst Vs. Data Scientist

If you are new to the field it can be difficult to know the difference between both roles.

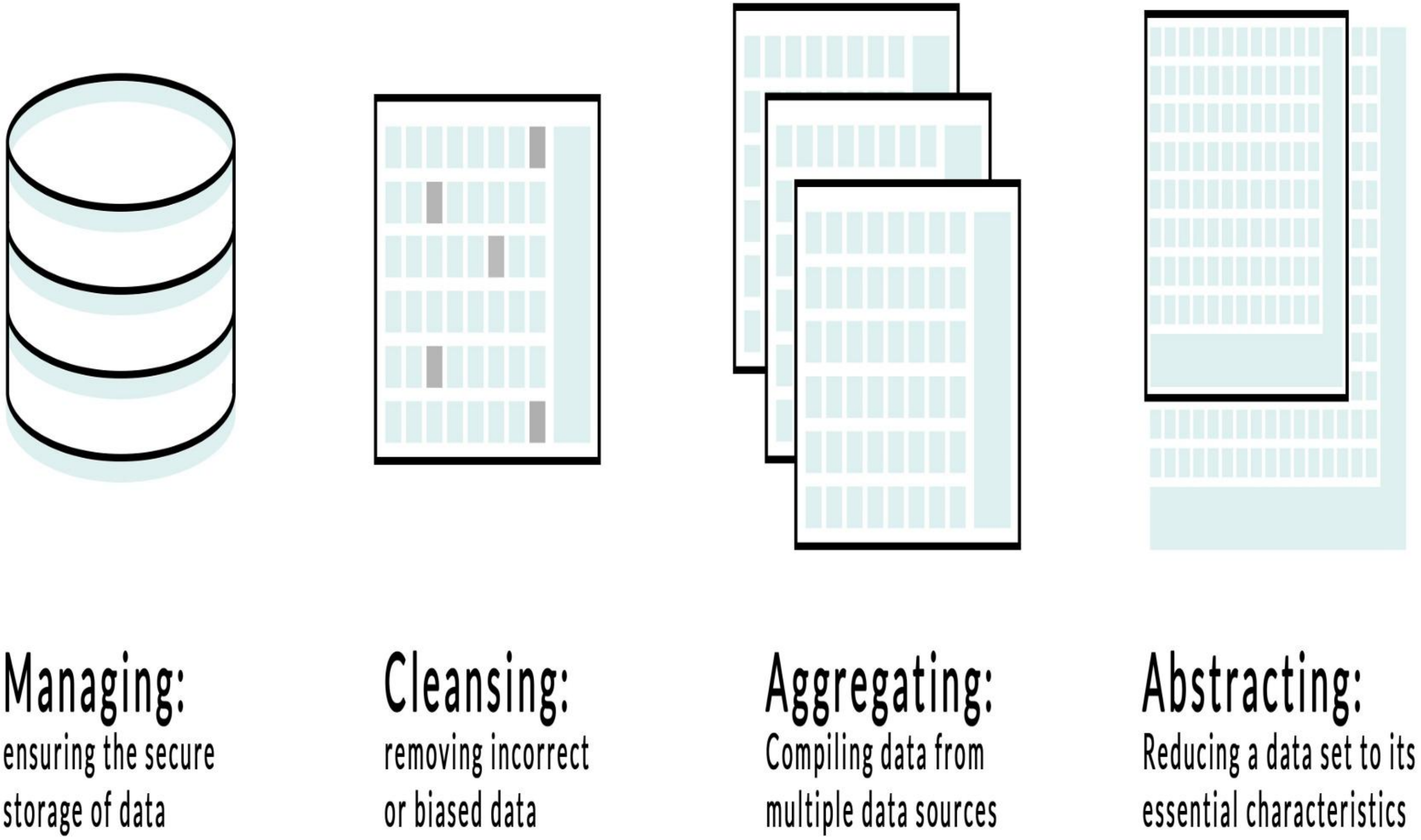
The good news is that there are clear industry standards published by the Institute of Apprenticeships (IfA) outlining the required skills for a **Data Analyst** and **Data Scientist**.

Data Analyst Vs. Data Scientist

The Data Analyst

In essence, the primary role of a Data Analyst is to collect, organize and study data to provide business insight. As stated in the IfA Data Analyst Standards *“Data Analysts are typically involved with managing, cleansing, abstracting and aggregating data, and conducting a range of analytical studies on that data.”*

Data Analyst Vs. Data Scientist



Managing:
ensuring the secure storage of data

Cleansing:
removing incorrect or biased data

Aggregating:
Compiling data from multiple data sources

Abstracting:
Reducing a data set to its essential characteristics

Data Analyst Vs. Data Scientist

Managing, Cleansing, Abstracting and Aggregating Data: A Definition

Managing: involves planning, executing and maintaining data processes for the secure storage of data and information assets.

Cleansing: the process of checking data quality and accuracy by recognizing then removing incorrect or biased data from a database.

Abstracting: the process of removing characteristics from a dataset to reduce it to a set of essential characteristics for more efficient data processing.

Aggregating: the process of compiling information from multiple data sources to prepare combined datasets for data processing.

Data Analyst Vs. Data Scientist

The Data Scientist

Data Scientists build upon the core competencies of a Data Analyst with additional Machine Learning and Software Engineering skills. The IfA Data Scientist Standards explains; *“Data Scientists are dynamic and adaptable, addressing varied problems with varied techniques. They actively explore innovative ways to use existing and new statistical, algorithmic, predictive, machine learning and artificial intelligence tools and techniques, to find significant and valuable patterns in data and transform these into information for their organization.”*

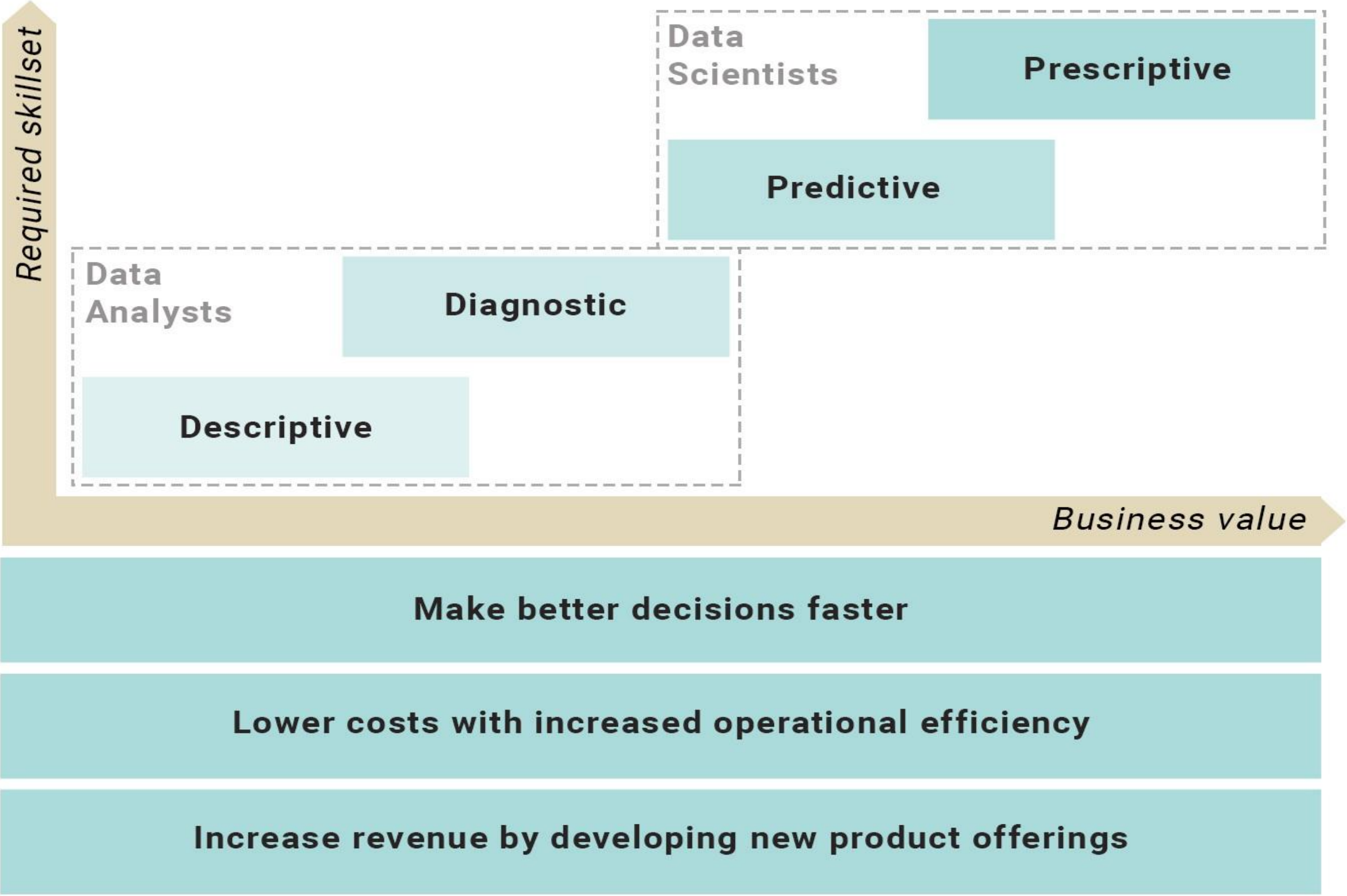
Data Analyst Vs. Data Scientist

What are the typical problems a Data Analyst and a Data Scientist might work on?

Different types of analytics can be categorized into “The Four Analytic Capabilities” — a widely used framework put forward by Gartner Research.

These approaches increase in complexity, from Description and Diagnostic (more traditional techniques), to Predictive and Prescriptive (more sophisticated techniques), providing a useful way to demonstrate the progression from Data Analytics to Data Science.

Data Analyst Vs. Data Scientist


Required skillset
Data Scientists
Prescriptive
Predictive
Data Analysts
Diagnostic
Descriptive
Business value
Make better decisions faster
Lower costs with increased operational efficiency
Increase revenue by developing new product offerings

Data Analyst Vs. Data Scientist

The Four Analytic Capabilities: A Definition

Descriptive: What happened? Example: What is the turnover this month?

Diagnostic: Why did it happen? Example: In your monthly report, you can see that last month's sales performance declined. What caused this?

Predictive: What will happen? Example: Imagine you are a retailer and you want to maximize product sales while minimizing waste. How can you accurately forecast how much stock you need?

Prescriptive: What should I do? Example: Based on the traffic predictions, what are the best marketing initiatives you can put in place to maximize the prospects-to-lead ratio?

Data Analyst Vs. Data Scientist

While each company will face different data challenges and business problems, a Data Analyst will often be tasked with performing descriptive and diagnostic analysis to provide business insights.

In contrast, a Data Scientist would be expected to apply predictive and prescriptive analytics to develop business solutions. Their work requires strong programming skills and deep theoretical knowledge, combined with the dedication and curiosity to keep up with the latest tools and techniques.

Popular Software Tools for Data Analytics

https://www.softwaretestinghelp.com/big-data-tools/

Introduction to Big Data Analytics

https://www.guru99.com/what-is-big-data.html

Revision of Java Language and Discuss the need of Java for our subject.

Why Learning Java is a Starting Point For Big Data Developers Of The Future?

Java is a big friend to Big Data scientists and developers.

Here I am going to tell you why is that so

https://towardsdatascience.com/why-learning-java-is-a-starting-point-for-big-data-developers-of-the-future-9a9b6d240dea

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Discussion on Syllabus, Objectives & Outcomes

OBJECTIVES:

- Optimize business decisions and create competitive advantage with Big Data analytics.
- Introducing Java concepts required for developing map reduce programs.
- Derive business benefit from unstructured data.
- Imparting the architectural concepts of Hadoop and introducing map reduce paradigm.
- To introduce programming tools PIG & HIVE in Hadoop echo system.

Discussion on Syllabus, Objectives & Outcomes

OUTCOMES:

- Preparing for data summarization, query, and analysis.
- Applying data modeling techniques to large data sets.
- Creating applications for Big Data analytics.
- Building a complete business data analytic solution.