IE405.E31
Big Data Analytics Technology

Index Huynh

Lecturer: M.S. Nguyễn Hồ Duy Trí

HCMC, Ngày 10 tháng 8 năm 2025

Mục lục

Preface

Chương 1

Introduction

1.1 Syllabus

- **Prerequisites:** Discrete mathematics, statistical mathematics.
- **Topics Covered:** Big data foundation, models, processing tools.
- **Self-Taught Content**: Corporate culture.
- **Programming language:** Python

1.2 Objective

The objective of this course is to enable students to understand and articulate fundamental concepts related to big data. Students will learn to organize and store data using big data storage platforms, and will apply data mining, machine learning, and AI algorithms to practical tasks. Furthermore, students will be able to research and understand the systems and ecosystems of a chosen company.

1.3 Course Content

- **Overview**: The course begins with an **overview of big data**, focusing on the reasons for studying this topic.
- **MapReduce Model:** Students will learn about the **MapReduce model**, which is the foundational model and a prerequisite for further development in the field.
- **Apache Spark Framework**: This section covers **Apache Spark**, where students will acquire foundational knowledge of this essential big data processing tool.
- **Big Data Ecosystem**: The course concludes with a study of the **big data ecosystem**, which serves as both the final project and a seminar topic.

Chương 2

Overview

2.1 Introduction

2.1.1 Scenario

With the Samsung ecosystem, all activities are synchronized, helping users save a lot of time. This is achieved by applying numerous big data technologies. Some key examples include: parallel computing, massive data storage, data distribution, high-speed network connectivity, high-performance computing, task and workflow management, data analysis and mining, data collection, machine learning, and data visualization. These thechnologies are foundational and, in some ways, quite mature, which shows that learning cannot always keep pace with society's rapid development. The more details you can uncover.

2.2 Definition

Big Data is an information asset characterized by high-volume, high-velocity, and high-variety data. It requires new technologies to be processed in order to enable effective decision-making, discover hidden insights, and optimize data processing.

2.3 Big data processing workflow

Big data processing workflow is a critical sequence of steps that includes **collection**, **storage**, and **analysis**.

- **Data Collection**: In this initial phase, data is gathered from two main sources: **active** (using tools like web crawlers) and **passive** (from sources such as surveilliance video and clickstreams). The collected data is then transported to storage locations via high-speed connections, where it is integrated, cleaned, and de-duuplicated to ensure quality.

- **Data Storage**: For storage, it is essential to evaluate the infrastructure and budget to decide between **SSD** or **HDD** and to segment the data into smaller parts. To enable communication between these parts, selecting an appropriate network architecture (such as **DAS**, **NAS**, or **SAN**) is crucial. Next, a suitable database management system must be chosen (e.g., **HDFS**, **key-value**, column-oriented, or document-oriented databases). For efficient storage, programming models like **MapReduce**, **stream processing**, or **graph processing** are utilized based on specific objectives.

- **Data Analysis**: The final step is analysis. First, the specific problem and goals must be clearly defined. There is a wide range of analytical methods available, including **statistical analysis**, **data mining**, **text mining**, **network and graph data mining**, **clustering**,

classification and regression, and **association analysis**. Each diverse field requires its own customized analytical techniques.