Big Data

(Syllabus)

Instructor: Thanh Binh Nguyen

September 1st, 2019

S3Lab

Smart Software System Laboratory

> “Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”
>
> – Chris Lynch, Vertica Systems

General Information

- Instructor: Ph.D. Thanh Binh Nguyen
    - Email: binhnt@uit.edu.vn
- Credit: 4 (3 lectures + 1 lab)
- Course Code: IS6102
- Prerequisites: STAT 4033, CS 4323
- Language: Vietnamese.

Student Assessment

- Final Examination (QUIZ): 40% - 4 / 10
- Assignments, projects: 60% - 6 / 10
  - 1 Project (included presentations, 2 students per group)
- **Note**:
  - Do cheating during studying: 0 - Failed
  - Maximum absent time per semester is 6 hours

Contents

- Understanding Big data (1 week)
- Hadoop (2 weeks)
  - Hadoop? Bigdata and Hadoop.
  - Hadoop in Overview
  - Hadoop Distributed File System
  - Hadoop Yarn & MapReduce
  - Hadoop ecosystem

Contents

- Hadoop
    - Apache Flume & Sqoop (Data Loading tools)
    - Apache Pig
    - Apache Hive
    - Apache HBase
    - Apache Oozie
- Spark (1 week)
- NoSQL Databases (1 week)
- Streaming Analytics / Stream Processing (1 week)

Contents

- Big data analytics (4 weeks)
  - Understanding
  - Data Mining
  - Recommendation systems (2 weeks)
- Internet of Things (IoT) (1 week)
- Success stories
  - Netflix (1 week)
  - Uber (1 week)

Contents

- Projects Presentations (2 weeks)
  - Study a tool, framework (select one from Big data ecosystem diagram), or open source project (**Oryx.io**, **Raccoon**, …) which relate to big data.
  - Use them to solve a problem in practice.
  - Make a report, demo and do the presentation in 15’.
  - Register team, topic from 2nd week.
  - Time for presentation is the last 2 weeks of the course.
  - Student will receive bonuses for a completed system or solution.

Course’s Keywords

Big data

[Algorithm] [Analytics] [Descriptive Analytics] [Predictive Analytics] [Prescriptive Analytics] [Batch Processing] [Cassandra] [Cloud computing] [Cluster Computing] [Dark Data] [Data lake] [Data mining] [Data Scientist] [Distributed File System] [ETL] [Hadoop] [In-memory computing] [IoT] [Machine learning] [MapReduce] [NoSQL] [R] [Apache Spark] [Stream processing] [Structure & Unstructured Data] [BI] [Recommendation System]

Books & Materials

- Books
    - Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman. “Big Data For Dummies”. John Wiley & Sons, Inc.
    - Jeffery Aven. “Hadoop in 24 Hours.” SAMs, 2017.
    - Jiawei Han, Micheline Kamber, Jian Pe. ”Data Mining Concepts and Techniques”. Elsevier, 2012.
    - A. Maheshwari. “Data Analytics Made Accessible”. 2019.
- Blogs and Others references
    - https://github.com/samadhankadam/Hadoop-Ebook
    - https://data-flair.training/blogs
    - https/www.edureka.co/blog/
    - https://www.tutorialspoint.com/index.htm

Q & A

Cảm ơn đã theo dõi

Hy vọng cùng nhau đi đến thành công.