Big Data Processing Technology

Introduction of Myself

CV

Introduction of the Course

- AI
- Machine Learning
- Big Data Processing Technology

Nowadays, **big data, machine learning, AI** are very sought-after IT skills that employers are pursuing. These are the top three skills in the field of information technology.

The relationships between big data processing technology, machine learning, and AI are as follows:
AI = **ML (learning algorithm)** + **Big Data (engineering)**; **ML** = **Traditional ML** + **DL** + **RL**;
In this class, we will focus on the engineering aspect and provide you with a broad set of tools.

My Goal for the Class

- 1. Inspire excitement about computer science engineering.
- 2. Be able to apply different tools or skills to big data processing problems, which will be **our final term project**.

Syllabus

Textbook: “Big Data Basic Programming, Experiments, and Case Tutorials (Version 2)” by Lin Ziyu, Tsinghua University Press

- Chapter One: Overview (2 theory classes)
- Chapter Two: Linux Basic Programming (2 theory + 4 practice)
- Chapter Three: Hadoop Introduction and Installation (4 theory + 8 practice)

- Chapter Four: HDFS Practice, Data Storage (2 theory + 4 practice)
- Chapter Five: HBase Practice, Database (8 classes)
- Chapter Six: MapReduce, Data Processing (4 classes)
- Chapter Seven: Hive, Data Warehouse, Data Processing (8 classes)
- Chapter Eight: Spark, Data Processing (12 classes)
- Chapter Nine: Flink + Kafka, Data Collecting (4 classes)
  Totally 62 classes

Assessment

1. Final Term Project (50%)

Execute a small project:

- Pass Criteria
  - Public competition or projects that demonstrate your efforts
  - Related to big data processing
  - In a competition, any submission will be considered as passing the exam. A good rank will also result in higher scores.
- Open Options for Projects
  - Kaggle
  - TianChi
  - Big Data Challenging
- A study group of up to three members
- Adherence to the honor code
- Brainstorming with me or Teacher Zhou is encouraged

2. Regular Performance (0%)

- Asking questions is strongly encouraged

3. In-Class Quizzes (50%)

- Quizzes will be randomly given during class

Student Background survey

- ? Python
- ? Linux
- ? Git
- ? Machine Learning
- ? Deep Learning
- ? Statistics Analysis or Linear Algebra

Questions

?