**HƯỚNG DẪN LÀM BÀI TẬP CUỐI KỲ MÔN NHẬP MÔN DỮ LIỆU LỚN**

**HKII 2024 – 2025**

**Ngày báo cáo: tuần 17**

**File nộp: báo cáo word (.docx) + project + script + database**

Mỗi nhóm (2-4 SV) thực hiện các chức năng sau:

# **Bắt buộc**

1. Cài đặt Ubuntu Server 22.04 LTS.
2. Cài đặt cụm Hadoop nhiều node.
3. Các phần mềm, framework nên sử dụng phiên bản mới nhất.
4. Cài đặt framework Apache Spark trên Hadoop Cluster.
5. Viết lại và thực thi các chương trình MapReduce: WordCount, WordMean, WordMedian, WordStandardDeviation bằng PySpark trên Hadoop Cluster.

# **Tùy chọn**

1. Có thể chạy thêm các chương trình khác bằng PySpark trên Hadoop Cluster.

# **Lưu ý:**

# các nhóm tạo user, đặt tên theo quy định (VD: An-master, Binh-slave, Trung-slave) và cấp các quyền cần thiết.

# File báo cáo phải có **Bảng phân công công việc** (tất cả thành viên đều thực hiện cài đặt, viết và thực thi chương trình) và **Tài liệu tham khảo**.