# TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

## VIỆN TRÍ TUỆ NHÂN TẠO

-----***-----

# BÁO CÁO MÔN HỌC KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

## ĐỀ TÀI

## Finding Average Temperature of Each Year using Hadoop-HDFS


**Giảng viên hướng dẫn:** TS. Trần Hồng Việt

Nhóm sinh viên thực hiện

1. Thái Nguyễn Hoàng Bách - 22022672
2. Quách Đắc Chính - 22022518
3. Vũ Minh Đức - 22022587
4. Đào Duy Hưng - 22022589


**HÀ NỘI, 12/2024**

# MỞ ĐẦU

Big Data ngày nay đã trở thành một phần không thể thiếu trong việc xử lý và phân tích dữ liệu lớn khi xu hướng sử dụng công nghệ ngày càng lớn đối với tất cả mọi người. Với sự phát triển của các công nghệ hỗ trợ như Hadoop, việc xử lý các tập dữ liệu lớn không chỉ trở nên khả thi mà còn hiệu quả. Điều đó sẽ được ứng dụng trong rất nhiều lĩnh vực như đời sống, giáo dục, …

Chúng em chọn đề tài **"Finding Average Temperature using Hadoop-HDFS"** nhằm mục đích áp dụng thực tiễn công nghệ MapReduce để phân tích nhiệt độ từ dữ liệu khí tượng và cung cấp thông tin về biến động thời tiết qua các năm.

Báo cáo gồm 4 chương:

**Chương 1:** Tổng quan về dữ liệu lớn và Hadoop MapReduce

**Chương 2:** Cách triển khai phân tích dữ liệu khí tượng.

**Chương 3:** Đánh giá kết quả và hiệu quả xử lý.

**Chương 4:** Kết luận và hướng phát triển.

**BÁO CÁO MÔN HỌC KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN** 1
Nhóm sinh viên thực hiện: 1
MỞ ĐẦU 1
**Chương 1: Tổng quan về dữ liệu lớn và Hadoop MapReduce** 3
1.1 Định nghĩa của Big Data 3
1.2 Đặc trưng của Big Data 3
1.3 Giới thiệu về Hadoop MapReduce 3
**Chương 2: Cách triển khai phân tích dữ liệu khí tượng** 5
2.1 Cấu trúc dữ liệu đầu vào 5
2.3 Quy trình xử lý dữ liệu với MapReduce 6
2.3.1 Mapper 6
2.3.2 Reducer 7
**Chương 3: Đánh giá kết quả** 9
3.1 Kết quả chạy chương trình 9
3.2 Đánh giá chương trình 9
**Chương 4: Kết luận và hướng phát triển** 10
4.1 Kết luận 10
4.2 Hướng phát triển 10

# Chương 1: Tổng quan về dữ liệu lớn và Hadoop MapReduce

## 1.1 Định nghĩa của Big Data

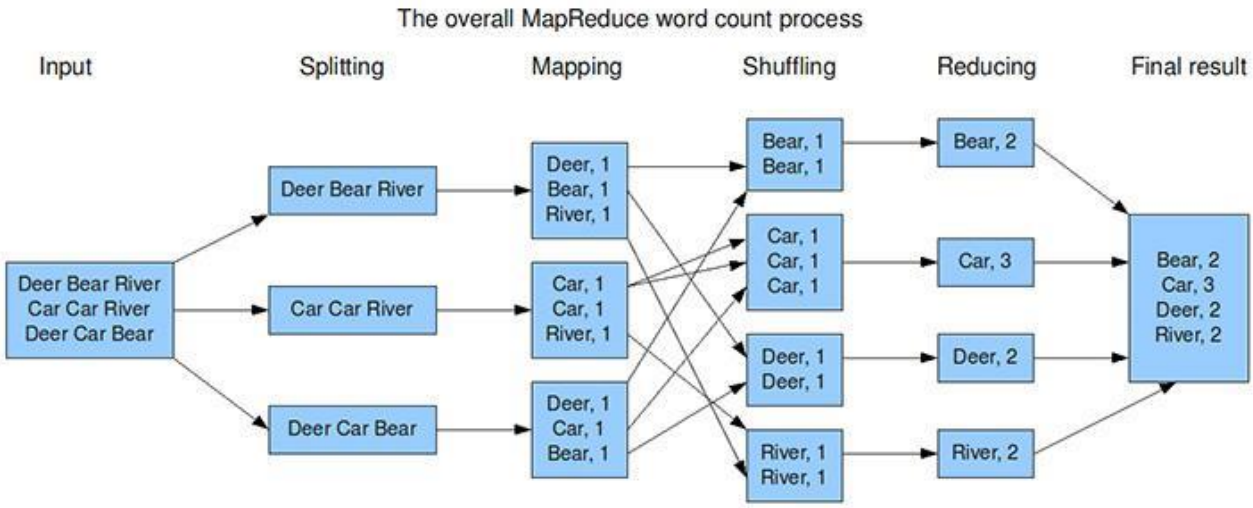
- Theo Wikipedia: Dữ liệu lớn thường bao gồm tập hợp dữ liệu với kích thước vượt xa khả năng của các công cụ phần mềm thông thường để thu thập, hiển thị, quản lý và xử lý dữ liệu trong một thời gian có thể chấp nhận được.
- Dữ liệu lớn bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư.

## 1.2 Đặc trưng của Big Data

- Khối lượng lớn (Volume): Khối lượng dữ liệu rất lớn và đang ngày càng tăng lên, tính đến 2014 thì có thể trong khoảng vài trăm terabyte.
- Tốc độ (Velocity): Khối lượng dữ liệu gia tăng rất nhanh.
- Đa dạng (Variety): Ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc( tài liệu, blog, hình ảnh,...)
- Độ tin cậy/chính xác(Veracity): Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của bigdata.
- Giá trị(Value): Giá trị thông tin mang lại.

## 1.3 Giới thiệu về Hadoop MapReduce

- Hadoop là một framework mã nguồn mở dùng để lưu trữ và xử lý dữ liệu lớn trên các cụm máy tính thông thường.
- Hadoop bao gồm hai thành phần chính:
    + HDFS (Hadoop Distributed File System): Hệ thống lưu trữ phân tán, giúp lưu trữ dữ liệu lớn bằng cách phân mảnh dữ liệu trên nhiều nút.
    + MapReduce: Mô hình lập trình song song, xử lý dữ liệu qua các bước Map và Reduce.
- MapReduce là mô hình lập trình được thiết kế để xử lý dữ liệu song song trên các cụm máy tính phân tán. Mô hình hoạt động qua hai pha chính:
    + Map: Chuyển đổi dữ liệu đầu vào thành các cặp key-value trung gian.
    + Reduce: Tổng hợp và xử lý các cặp key-value để tạo ra kết quả cuối cùng.


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

# Chương 2: Cách triển khai phân tích dữ liệu khí tượng

2.1 Cấu trúc dữ liệu đầu vào

- Dữ liệu đầu vào là các bản ghi văn bản được mã hóa, mỗi dòng biểu diễn một quan sát từ trạm thời tiết. Ví dụ một dòng dữ liệu:

0067011990999991950051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9+00001+99999999999

- Ý nghĩa của các trường trong dữ liệu:

006701199099999:

- 006701: Mã nhận dạng của trạm quan trắc.
- 199099999: Mã bổ sung (có thể chỉ loại trạm hoặc chi tiết về thiết bị).

1950051507004:

- 1950: Năm quan trắc.
- 0515: Ngày tháng (15 tháng 5).
- 07004: Giờ quan trắc (07:00 UTC).

+68750+023550:

- +68750: Vĩ độ (+68.750 độ).
- +023550: Kinh độ (+23.550 độ).

FM-12+038299999:

- FM-12: Loại thông báo hoặc định dạng dữ liệu (thường là METAR hoặc SYNOP).
- +0382: Độ cao trạm quan trắc (382 mét trên mực nước biển).
- 99999: Mã chưa xác định hoặc không sử dụng.

V0203301:

- V020: Tốc độ gió hoặc thông tin bổ sung liên quan.
- 3301: Hướng gió.

N00671220001C:

- N0067: Mã quốc gia (67 là Nga).
- 1220001C: Dữ liệu bổ sung.

N9999999N9+00001+99999999999

- N9999999: Giá trị chưa xác định.
- N9+00001: Nhiệt độ thực tế (+0.001 độ C, có thể đã được nhân với 10).
- +99999999999: Dữ liệu bổ sung hoặc không xác định.

Các trường quan trọng được sử dụng trong phân tích:

- **Năm:** Lấy từ vị trí 15-19, biểu diễn năm quan sát (VD: "1950").
- **Nhiệt độ:** Lấy từ vị trí 87-92, là giá trị nhiệt độ (chia 10 để chuyển sang °C).
- **Chất lượng dữ liệu:** Lấy từ vị trí 92, kiểm tra giá trị để loại bỏ các dữ liệu không hợp lệ (VD: giá trị phải là "0, 1, 4, 5, 9").

Các dòng dữ liệu bị thiếu thông tin hoặc chứa giá trị nhiệt độ là +9999 (ký hiệu cho giá trị không xác định) sẽ bị loại bỏ.

## 2.3 Quy trình xử lý dữ liệu với MapReduce

### 2.3.1 Mapper

- Lớp AvgTemp_Mapper được thiết kế để xử lý các nhiệm vụ sau:

1. **Đọc từng dòng dữ liệu đầu vào:** Mỗi dòng là một quan sát thời tiết từ trạm.
2. **Trích xuất các trường dữ liệu cần thiết:** Gồm năm, nhiệt độ, và chất lượng.
3. **Phát các cặp key-value:** Với key là năm và value là nhiệt độ.

Ví dụ đầu ra từ Mapper:

1950 +01

1950 +22

1949 -11

AvgTemp_Mapper.java

```java
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class AvgTemp_Mapper extends Mapper<Object, Text, Text, Text> {
    private Text yearKey = new Text();
    private Text tempValue = new Text();

    @Override
    protected void map(Object key, Text value, Context context) throws IOException, Int
        String line = value.toString();
        // Tách các trường từ dữ liệu
        String year = line.substring(15, 19); // Năm
        String temp = line.substring(87, 92); // Nhiệt độ
        String quality = line.substring(92, 93); // Chất lượng dữ liệu
        if (!temp.trim().equals("+9999") && quality.matches("[01459]")) {
            yearKey.set(year);
            tempValue.set(temp);
            context.write(yearKey, tempValue);
        }
    }
}
```

2.3.2 Reducer

- Lớp AvgTemp_Reducer thực hiện các nhiệm vụ quan trọng nhằm tổng hợp và tính toán các thông tin từ dữ liệu nhiệt độ trung gian do Mapper phát ra:

1. **Nhận các cặp key-value từ Mapper:**
    - Các giá trị nhiệt độ được nhóm theo từng năm (key) do Mapper phát ra.
    - Ví dụ: Các nhiệt độ của năm "1950" sẽ được nhóm chung để xử lý.
2. **Tính toán các thông số nhiệt độ:**
    - **Nhiệt độ trung bình:** Tổng các giá trị nhiệt độ chia cho số lượng bản ghi.
    - **Nhiệt độ cao nhất:** Duyệt qua các giá trị và lấy nhiệt độ lớn nhất.
    - **Nhiệt độ thấp nhất:** Duyệt qua các giá trị và lấy nhiệt độ nhỏ nhất.
3. **Xuất kết quả:**
    - Kết quả cho mỗi năm (key) được định dạng thành chuỗi bao gồm thông tin số lượng bản ghi, nhiệt độ trung bình, cao nhất và thấp nhất.

**Chèn ảnh kết quả

- Lớp này đóng vai trò quan trọng trong việc tổng hợp dữ liệu, đảm bảo rằng các kết quả phân tích là chính xác và đầy đủ để phục vụ các mục tiêu nghiên cứu khí hậu.

AvgTemp_Reducer.java

```java
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class AvgTemp_Reducer extends Reducer<Text, Text, Text, Text> {
    @Override
    protected void reduce(Text key, Iterable<Text> values, Context context) throws IOEx
        int count = 0;
        int sum = 0;
        int maxTemp = Integer.MIN_VALUE;
        int minTemp = Integer.MAX_VALUE;

        for (Text value : values) {
            int temp = Integer.parseInt(value.toString());
            sum += temp;
            count++;
            maxTemp = Math.max(maxTemp, temp);
            minTemp = Math.min(minTemp, temp);
        }

        double average = (double) sum / count;
        String result = String.format("Số lượng: %d, TB: %.2f°C, Cao nhất: %.2f°C, Thấ
                count, average / 10, maxTemp / 10.0, minTemp / 10.0);

        context.write(key, new Text(result));
    }
}
```

# Chương 3: Đánh giá kết quả

3.1 Kết quả chạy chương trình

- Sau khi triển khai chương trình trên tập dữ liệu mẫu, kết quả đầu ra như sau:

***Chèn ảnh kết quả

3.2 Đánh giá chương trình

**Ưu điểm:**

- Chương trình xử lý chính xác các dữ liệu nhiệt độ lớn và phức tạp.
- Khả năng mở rộng dễ dàng nhờ sử dụng Hadoop.
- Xử lý hiệu quả dữ liệu không hợp lệ thông qua bộ lọc Mapper.

**Hạn chế:**

- Phụ thuộc vào định dạng dữ liệu đầu vào; dữ liệu không chuẩn hóa cần được xử lý trước.
- Kết quả đầu ra chưa được tích hợp visualization để dễ hiểu hơn.

# Chương 4: Kết luận và hướng phát triển

4.1 Kết luận

- Đề tài "Phân tích dữ liệu khí tượng bằng Hadoop MapReduce" đã hoàn thành các mục tiêu đề ra, bao gồm:
- Triển khai thành công hệ thống phân tích nhiệt độ từ dữ liệu lớn.
- Xử lý hiệu quả dữ liệu không hợp lệ và cung cấp các thông số quan trọng như nhiệt độ trung bình, cao nhất và thấp nhất theo từng năm.
- Chương trình cho thấy tiềm năng ứng dụng lớn của Hadoop trong các bài toán xử lý dữ liệu khí tượng và khoa học môi trường.

4.2 Hướng phát triển

- **Tích hợp thêm yếu tố phân tích:** Bao gồm độ ẩm, lượng mưa, tốc độ gió để mở rộng phân tích khí hậu.
- **Visualization:** Sử dụng các công cụ như Tableau hoặc Power BI để trực quan hóa kết quả.
- **Tối ưu hóa hiệu suất:** Sử dụng các kỹ thuật caching và nén dữ liệu để giảm thời gian xử lý.
- **Ứng dụng thực tế:** Mở rộng hệ thống để xử lý dữ liệu khí tượng theo thời gian thực, hỗ trợ các cơ quan nghiên cứu khí hậu và dự báo thời tiết.