



Επεξεργασία Δεδομένων - MapReduce

Msc In Information Systems (part time)

Μάθημα: Ειδικά θέματα σχεδίασης βάσεων δεδομένων

Διονύσης Κοροπούλης

MM4160008

Βασίλης Λαμπρακάκης

MM4160016

Περιγραφή σεναρίου

- Μετεωρολογικά δεδομένα μεγάλης κλίμακας
- Αφορούν τις θερμοκρασίες που έχουν καταγραφεί σε μία πόλη
- Θα εφαρμόσουμε το μοντέλο προγραμματισμού MapReduce σε δίκτυο 5 υπολογιστών (cluster)
- Εξαγωγή συμπερασμάτων για μέση μέγιστη θερμοκρασία των μηνών του έτους (από ανάλυση ημερήσιων θερμοκρασιών πολλών ετών)

Επεξήγηση του dataset

- Η χαμηλότερη και η υψηλότερη θερμοκρασία που καταγράφεται μέσα στην ημέρα
- Η κάθε γραμμή του dataset περιγράφει την χαμηλότερη και υψηλότερη τιμή θερμοκρασίας για μία συγκεκριμένη ημέρα
- Τα datasets περιλαμβάνουν όλες τις ημέρες από το έτος 1972 έως και το έτος 2013

Μορφή του dataset

Αρχεία τύπου csv

10052000	17.6	24.9
11052000	18.8	26
12052000	16	25
13052000	14.8	26.8
14052000	16	26.4

Ημέρα

Μήνας

Έτος

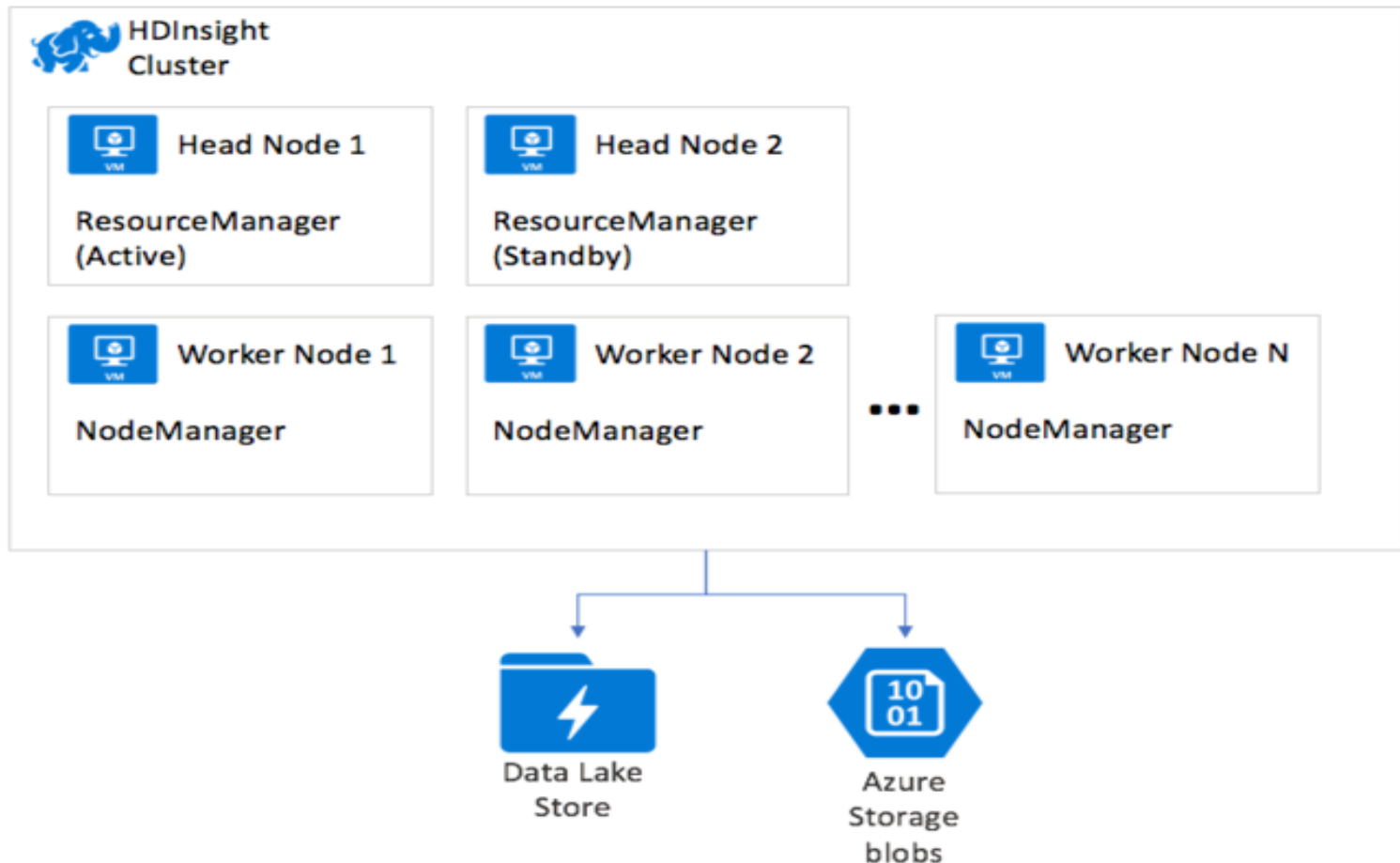
Χαμηλότερη
θερμοκρασία

Υψηλότερη
θερμοκρασία

Περιβάλλον δοκιμής σεναρίου

Microsoft Azure

Azure HDInsight



Δημιουργία δικτύου υπολογιστών (cluster) στο HDInsight - Προδιαγραφές

○ 5 εικονικές μηχανές

Εκ των οποίων :

- 2 Head Nodes (Master) – για περιπτώσεις αστοχίας του πρώτου Head Node


4 Cores, 7GB RAM / node

- 3 Worker Nodes (Slaves)

4 Cores, 7GB RAM / node

Cluster type : Hadoop 2.7 on Linux

Δημιουργία δικτύου υπολογιστών (cluster) στο HDInsight – Προδιαγραφές (2)

 Vassilis
HDInsight cluster

[Cluster Dashboard](#) [Secure Shell \(SSH\)](#) [Scale cluster](#) [Move](#) [Delete](#)

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

SETTINGS


Locks

Automation script

GETTING STARTED

Quick start

Cluster nodes

5 nodes 

TYPE	NODE SIZE	CORES	NODES
Head	A3	8	2
Worker	A3	12	3

Applications

Script actions

Cores in West Europe for subscription

Ανάλυση Κώδικα

Κλάσεις και μέθοδοι:

- SumCount – Βοηθητική κλάση
- Mapper Input
- Mapper
- Υπολογισμός μέσης τιμής θερμοκρασιών
- Cleanup()
- Παράδειγμα map() – cleanup()
- MeanReducer
- Cleanup()
- Παράδειγμα reduce() – cleanup()
- Driver

SumCount – Βοηθητική κλάση

- Δημιουργία της βοηθητικής κλάσης SumCount που υλοποιεί το WritableComparable interface
- Σκοπός: καταγραφή σε ένα αντικείμενο SumCount το άθροισμα θερμοκρασιών και το πλήθος τους

Code Snippet:

```
public class SumCount implements WritableComparable<SumCount> {  
    DoubleWritable sum;  
    IntWritable count;  
  
    public void addSumCount(SumCount sumCount) {  
        set(new DoubleWritable(this.sum.get() + sumCount.getSum().get()), new  
IntWritable(this.count.get() + sumCount.getCount().get()));  
    }  
}
```

Mapper Input

- .csv Dataset

01012000,-4.0,5.0

02012000,-5.0,5.1

03012000,-5.0,7.7

...

Code Snippet:

```
public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
```

```
    // gets the fields of the CSV line
```

```
    String[] values = value.toString().split(",");
```

```
    // defensive check
```

```
    if (values.length != 3) {
```

```
        return;
```

```
    }
```

Mapper

Code Snippet:

```
// gets date and max temperature
String date = values[DATE];
Text month = new Text(date.substring(2));
Double max = Double.parseDouble(values[MAX]);
Double min = Double.parseDouble(values[MIN]);
// if not present, put this month into the map
if (!maxMap.containsKey(month)) {
    maxMap.put(month, new ArrayList<Double>());
}
// adds the max temperature for this day to the list of temperatures
maxMap.get(month).add(max);
```

- Παράδειγμα:

012014, [5.0]

012014, [5.1]

012014, [2.3]



012014, [5.0 , 5.1 , 2.3]

Υπολογισμός μέσης τιμής θερμοκρασιών

Sample input data:

01012000, 0.0, 10.0

02012000, 0.0, 20.0

03012000, 0.0, 2.0

04012000, 0.0, 4.0

05012000, 0.0, 3.0

Mapper #1: lines 1, 2

Mapper #2: lines 3, 4, 5

Mapper#1: mean = $(10.0 + 20.0) / 2 = 15.0$

Mapper#2: mean = $(2.0 + 4.0 + 3.0) / 3 = 3.0$

Reducer mean = $(15.0 + 3.0) / 2 = 9.0$

Λάθος υπολογισμός της μέσης τιμής!!

Η μέση τιμή υπολογίζεται ως εξής:
 $(10.0 + 20.0 + 2.0 + 4.0 + 3.0) / 5 = 7.8$

Σωστός τρόπος!!

Cleanup

```
protected void cleanup(Context context) throws IOException,
InterruptedException {

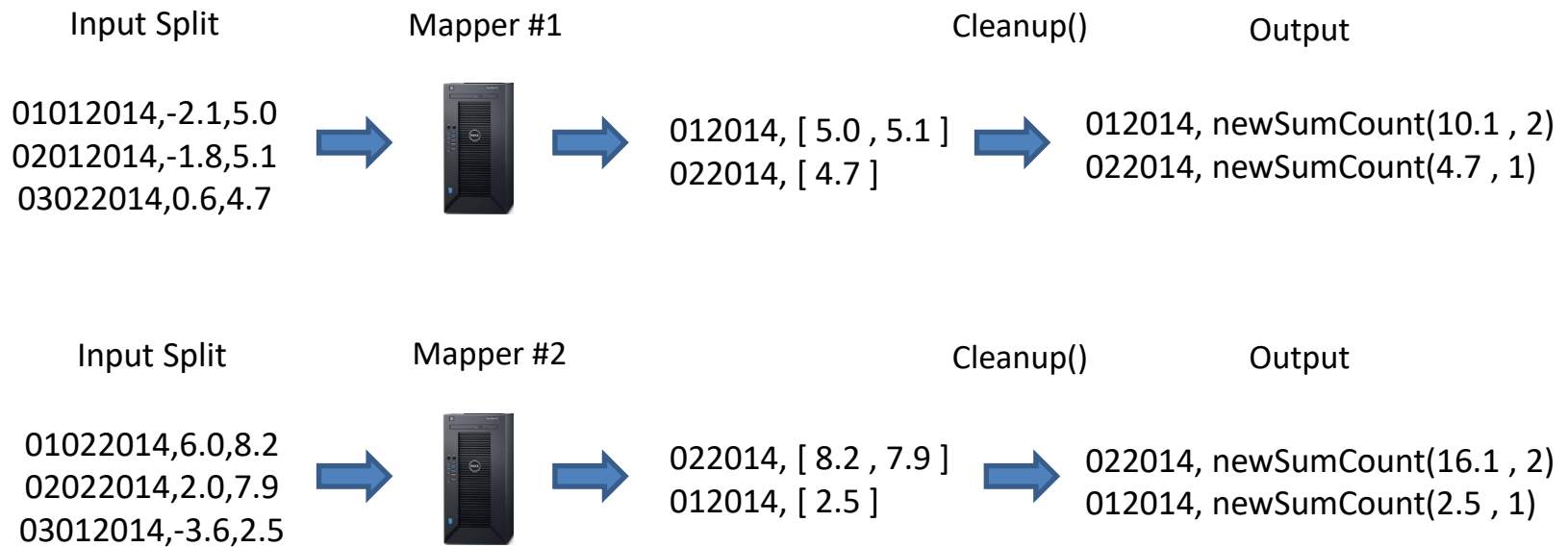
    // loops over the months collected in the map() method
    for (Text month: maxMap.keySet()) {

        List<Double> temperatures = maxMap.get(month);

        // computes the sum of the max temperatures for this
        month
        Double sum = 0d;
        for (Double max: temperatures) {
            sum += max;
        }

        // emits the month as the key and a SumCount as the
        value
        context.write(month, new SumCount(sum,
            temperatures.size()));
    }
}
```

Παράδειγμα map() – cleanup()



MeanReducer

```
public void reduce(Text key, Iterable<SumCount> values, Context
context) throws IOException, InterruptedException {

    SumCount totalSumCount = new SumCount();

    // loops over all the SumCount objects received for this
    month (the "key" param)
    for (SumCount sumCount : values) {

        // sums all of them
        totalSumCount.addSumCount(sumCount);
    }

    // puts the resulting SumCount into a map
    sumCountMap.put(new Text(key), totalSumCount);
}
```

Cleanup

```
protected void cleanup(Context context) throws
IOException, InterruptedException {

    // loops over the months collected in the reduce() method
    for (Text month: sumCountMap.keySet()) {

        double sum = sumCountMap.get(month).getSum().get();
        int count = sumCountMap.get(month).getCount().get();

        // emits the month and the mean of the max temperatures
        // for the month
        context.write(month, new DoubleWritable(sum/count));
    }
}
```


Παράδειγμα reduce() – cleanup()

Output – Mapper #1

012014, newSumCount(10.1 , 2)

022014, newSumCount(4.7 , 1)

012014, [newSumCount(10.1 , 2) , newSumCount(2.5 , 1)]

022014, [newSumCount(4.7 , 1) , newSumCount(16.1 , 2)]

Output – Mapper #2

022014, newSumCount(16.1 , 2)

012014, newSumCount(2.5 , 1)

Reducer



= 12.6/3 = 4.2

= 20.8/3 = 6.93

Cleanup()

012014, [newSumCount(12.6 , 3)]

022014, [newSumCount(20.8 , 3)]

Driver

```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf,
args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Usage: Mean <in> <out>");
        System.exit(2);
    }
    Job job = Job.getInstance(conf);
    job.setJobName("Mean");
    job.setJarByClass(Mean.class);
    job.setMapperClass(MeanMapper.class);
    job.setReducerClass(MeanReducer.class);
    job.setMapOutputKeyClass(Text.class);
    job.setMapOutputValueClass(SumCount.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

Εντολές για την εκτέλεση του κώδικα

// Ανέβασμα του αρχείου JAR και των datasets από το τοπικό filesystem στον Headnode του Cluster

```
C:\>scp C:/data/Meantest2.jar ssh sshuser@Vassilis-ssh.azurehdinsight.net:
```

```
C:\>scp C:/data/milano_temps.csv ssh sshuser@Vassilis-ssh.azurehdinsight.net:
```

```
C:\>scp C:/data/milano2.csv ssh sshuser@Vassilis-ssh.azurehdinsight.net:
```

```
C:\>scp C:/data/milano3_high.csv ssh sshuser@Vassilis-ssh.azurehdinsight.net:
```

// Remote σύνδεση στο cluster που έχουμε σηκώσει στο AZURE

```
C:\>ssh sshuser@Vassilis-ssh.azurehdinsight.net
```

// Δημιουργία φακέλου στο HDFS όπου θα αποθηκευθούν τα δεδομένα μας

```
sshuser@hn0-Vassil:~$ hdfs dfs -mkdir /project_v_d
```

// Αποθήκευση των αρχείων από τον Headnode σε φάκελο που δημιουργήσαμε στο HDFS

```
sshuser@hn0-Vassil:~$ hadoop fs -copyFromLocal milano_temps.csv /project_v_d/milano_temps.csv
```

```
sshuser@hn0-Vassil:~$ hadoop fs -copyFromLocal milano2.csv /project_v_d/milano2.csv
```

```
sshuser@hn0-Vassil:~$ hadoop fs -copyFromLocal milano3_high.csv /project_v_d/milano3_high.csv
```

// Εκτέλεση του κώδικα στα data του input directory και εξαγωγή αποτελεσμάτων στον φάκελο out3

```
sshuser@hn0-Vassil:~$ hadoop jar Meantest2.jar Mean /project_v_d /example/data/out3
```

Counters

Map-Reduce Framework

```
Map input records=15257  
Map output records=784  
Map output bytes=14616  
Map output materialized bytes=16196  
Input split bytes=332  
Combine input records=0  
Combine output records=0  
Reduce input groups=784  
Reduce shuffle bytes=16196  
Reduce input records=784  
Reduce output records=784  
Spilled Records=1568  
Shuffled Maps =2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=624  
CPU time spent (ms)=8230  
Physical memory (bytes) snapshot=1014059008  
Virtual memory (bytes) snapshot=7325364224  
Total committed heap usage (bytes)=802160640
```

```
Mean$MeanMapper$Temperature
```

Counters (με ακραίες τιμες θερμοκρασίας)

```
Map-Reduce Framework
  Map input records=10227
  Map output records=616
  Map output bytes=11424
  Map output materialized bytes=12662
  Input split bytes=103
  Combine input records=0
  Combine output records=0
  Reduce input groups=616
  Reduce shuffle bytes=12662
  Reduce input records=616
  Reduce output records=616
  Spilled Records=1232
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=150
  CPU time spent (ms)=2888
  Physical memory (bytes) snapshot=372637696
  Virtual memory (bytes) snapshot=473862144
  Total committed heap usage (bytes)=236978176

Mean$MeanMapper$Temperature
  HIGH_TEMPERATURE=4
  LOW_TEMPERATURE=3

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
```

Αποτελέσματα

- Map tasks = 2
- Reduce tasks =1
- Map input records=15257
- Map output records=784
- Reduce input groups=784
- Reduce output records=784
- Shuffled Maps =2
- Merged Map outputs=2

```
012010 5.481481481481482
012008 8.538709677419355
012009 4.32
012000 8.629032258064516
012001 7.283870967741937
051972 18.296129032258058
012002 7.3419354838709685
051973 20.95870967741935
012003 7.358064516129032
051974 16.896129032258067
012004 7.838709677419355
051975 21.257741935483875
012005 8.016129032258064
051976 20.351935483870967
012006 6.716129032258063
051977 20.60741935483871
012007 9.738709677419358
051978 16.385483870967743
051980 21.285161290322584
```

ΤΕΛΟΣ ΠΑΡΟΥΣΙΑΣΗΣ

ΕΥΧΑΡΙΣΤΟΥΜΕ!!