1. HADOOP 背景介绍

1.1 什么是 HADOOP

1. HADOOP 是 apache 旗下的一套开源**软件平台**
2. HADOOP 提供的功能：利用服务器集群，根据用户的自定义业务逻辑，**对海量数据进行分布式处理**
3. HADOOP 的核心组件有
    A. HDFS（分布式文件系统）
    B. YARN（运算资源调度系统）
    C. MAPREDUCE（分布式运算编程框架）
4. 广义上来说,HADOOP 通常是指一个更广泛的概念——HADOOP 生态圈

1.2 HADOOP 产生背景

1. HADOOP **最早起源于 Nutch**。Nutch 的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，**遇到了严重的可扩展性问题**——如何解决数十亿网页的存储和索引问题。
2. 2003 年、2004 年**谷歌发表的两篇论文为该问题提供了可行的解决方案**。
    ——分布式文件系统（GFS），可用于处理海量网页的**存储**
    ——分布式计算框架 MAPREDUCE，可用于处理海量网页的**索引计**

算问题。

3. Nutch 的开发人员完成了相应的开源实现 HDFS 和 MAPREDUCE，并从 Nutch 中剥离成为独立项目 HADOOP，到 2008 年 1 月，HADOOP 成为 Apache 顶级项目，迎来了它的快速发展期。

2. HADOOP 学习参考资料

1. 知乎： https://www.zhihu.com/question/19795366
2. 官网文档：http://hadoop.apache.org/docs/r2.6.5/

3. 数据项目任务要求

1. 利用 Hadoop 的 mapreduce 架构处理给定的数据文件 city.txt

部分数据如下：

```
01  yaan   36 200
02 xuzhou 110    600
03 shuicheng  20 150
04 rikaze 60 300
01 nanchong   40 300
02 yancheng   90 200
03 chishui    50 200
01 mianzhu    70 500
01 chengdu    100    1200
02 nanjin 120    1000
03 guiyang    80 750
04 lasa   60 300
01 deyang 40 300
02 suzhou 120    800
03 zunyi  50 200
01 mianyang   80 500
01 guangan    47 100
```

（数据说明，第一列为省份编号，第二列为城市名称，第三列为 GDP，

第四列为人口）

2. 任务要求

利用 mapreduce 架构，通过自定义可序列化 bean，Partitioner，Comparator，Mapper，Reducer 等实现对 city.txt 中的内容进行分区排序。

3. 规则：

（1）奇数编号省份和偶数编号省份分开输出**（即有奇数编号省份的城市排序输出到一个文件，偶数编号省份的城市排序后输出到一个文件）**

（2）同一编号身份的城市，按照 GDP，人口，城市名称，从小到大排序**（即如果两城市 GDP 相同，则按照城市人口数量从小到大排序，如果城市人口数量相同，再按照城市名字符串排序）**

4. 示例：

数据输入：

```
01  yaan   36 200
02 xuzhou 110    600
03 shuicheng  20 150
04 rikaze 60 300
01 nanchong   40 300
02 yancheng   90 200
03 chishui    50 200
01 mianzhu    70 500
01 chengdu    100    1200
02 nanjin 120    1000
03 guiyang    80 750
04 lasa   60 300
01 deyang 40 300
02 suzhou 120    800
03 zunyi  50 200
```

```
01 mianyang   80 500
01 guangan    47 100
```

数据输出：

奇数编号省份输出结果：

```
1   yaan   36 200
1  deyang 40 300
1  nanchong   40 300
1  guangan    47 100
1  mianzhu    70 500
1  mianyang   80 500
1  chengdu    100    1200
3  shuicheng  20 150
3  chishui    50 200
3  zunyi  50 200
3  guiyang    80 750
```

偶数编号省份输出结果：

```
2   yancheng   90 200
2  xuzhou 110    600
2  suzhou 120    800
2  nanjin 120    1000
4  lasa   60 300
4  rikaze 60 300
```

5. 友情提示可定义以下类：

其中每类的定义如下：

```
public class CityBean implements
WritableComparable<CityBean>
```

```
public class CityComparator extends WritableComparator
public class CityGrouper extends WritableComparator
public class CityMapper extends Mapper<Object, Text,
CityBean, LongWritable>
public class CityPartitioner extends
Partitioner<CityBean, LongWritable>
public class CityReducer extends Reducer<CityBean,
LongWritable, CityBean, NullWritable>
public class CityRunner
```

（注意：以上仅供参考，可按照自己的思路实现）

6. 部分代码已在附件中给出，也可自行实现