Hadoop应用开发实战案例(第二周)书面作业

使用Map-Reduce编程（或结合其它方法）计算：

1 根据上述日志文件，计算该天的独立ip数，pv数（注意要筛选日志，并非每条记录都要统计），被传输页面的总字节数

2 统计来源网站，列出域名及带来的独立ip数

以下题目为选做（可以不做）

3 统计用户使用的浏览器种类，计算出各种浏览器占的百分比

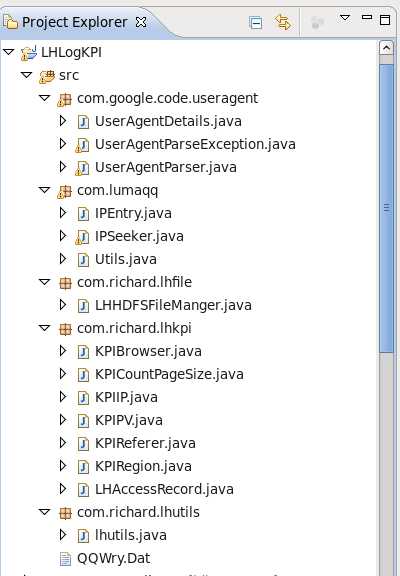
4 统计用户来自的地域（各省、直辖市、自治区，国外），计算各地域访问占的百分比

# 项目准备

1. Accesslog数据上传到hdfs:如下图



1. 项目创建：(编辑格式为 utf-8)项目结构：如下图



1. 项目文件介绍：

com.google.code.useragent包名下的文件是

开源的用于解析useragent,本项目中用到了

getBrowserName获取浏览器名称：

由于源码托管采用hg管理，所以我把下载的代码上传到如下链接：

<http://download.csdn.net/detail/lihuinihao/6707101>

com.lumaqq包名下的文件及QQWry.Dat用于传入ip地址来获取ip地址的地域，下载地址：

<http://download.csdn.net/detail/lihuinihao/6702099>

com.richard.lhfile包下是我封装的hdfs操作接口

可以进行文件上传下载，本项目主要用到文件检测及删除，因为在hadoop1.x中 输出文件夹如果存在会报错，所以在运行前先检测是否存在 存在则删除：格式如下使用：

if（LHHDFSFileManger.fileExists(output, conf)）

{

LHHDFSFileManger.deleteFile(output, conf);

}

com.richard.lhkpi包下为本次作业的相关代码：

KPIBrowser.java //浏览器统计

KPIIP.java //独立ip统计

KPIPV.java //pv统计

KPICountPageSize.java //页面传送总字节数

KPIReferer.java //网站来源统计

KPIRegion.java //ip地域统计

LHAccessRecord.java //单行数据解析保存

以单列模式实现

Com.richard.lhutils.java

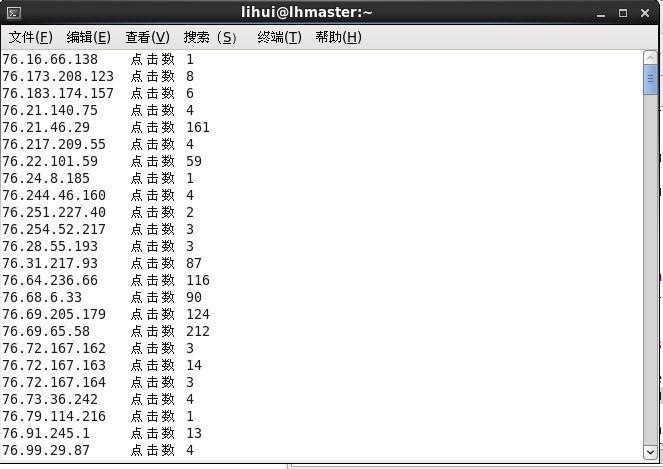
//目前实现两个函数

getPercent 转换百分比

ipRegion 获取地域信息

1. 6个作业执行结果截图如下：

独立ip执行结果截图：



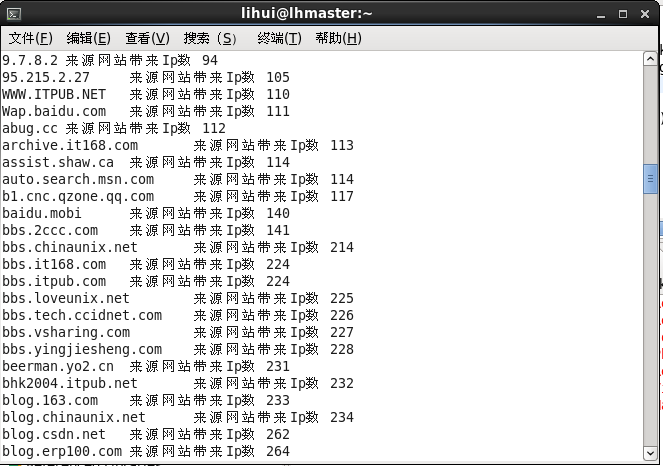
pv执行结果截图：



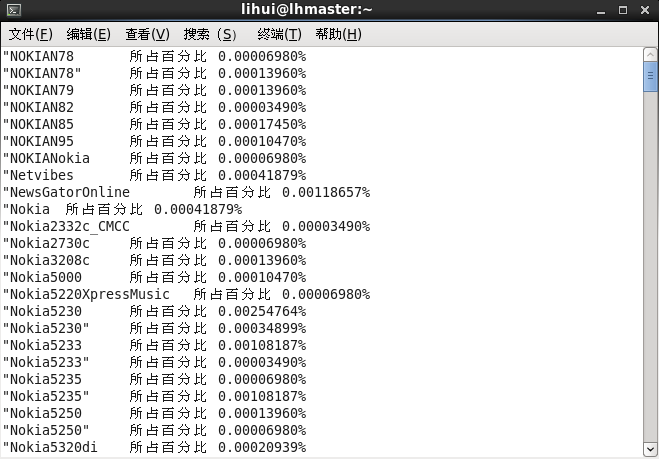
页面字节数截图如下：



网站来源执行结果截图：



浏览器统计执行结果截图：



ip地域统计 执行结果截图：

