Hadoop数据分析平台 第11周

2013.01.08

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

完全分布式模式的安装和配置

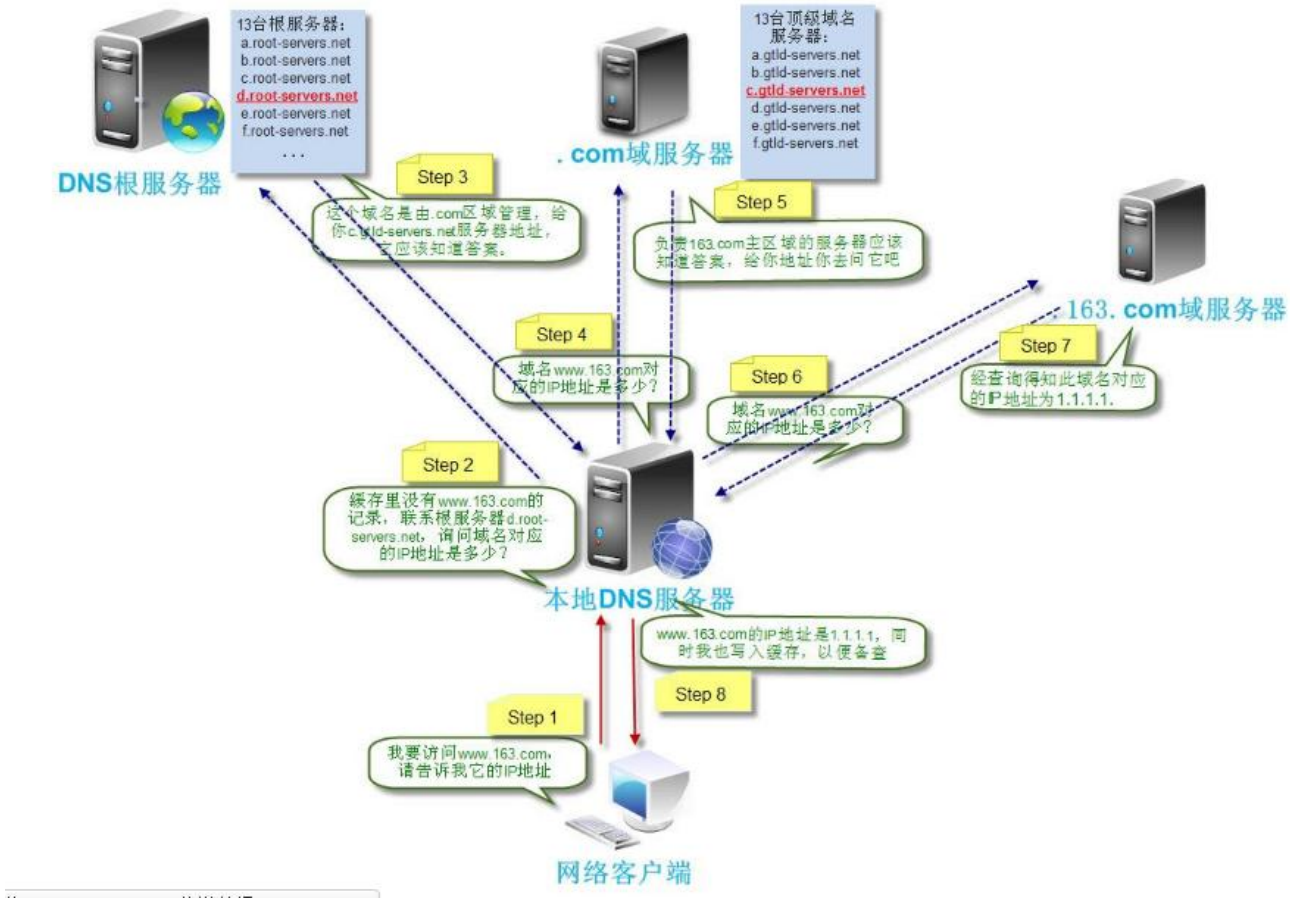
- 配置hosts文件
- 建立hadoop运行账号
- 配置ssh免密码连入
- 下载并解压hadoop安装包
- 配置namenode，修改site文件
- 配置hadoop-env.sh
- 配置masters和slaves文件
- 向各节点复制hadoop
- 格式化namenode
- 启动hadoop
- 用jps检验各后台进程是否成功启动

Hadoop大集群实施

- 设备选型
- 是否使用虚拟机？
- 使用DNS代替hosts文件
- 使用NFS实现密钥共享
- 利用脚本复制hadoop——awk技巧

DNS

- Linux下使用bind



NFS

- 网络文件系统
- 《Hadoop权威指南》第266页

接下来，需确保公钥存放在用户打算连接的所有机器的*~/.ssh/authorized_keys* 文件中。如果 hadoop 用户的 home 目录在 NFS 文件系统中，则密钥可以通过键入以下指令在整个集群共享：

```
% cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

用awk生成脚本的技巧

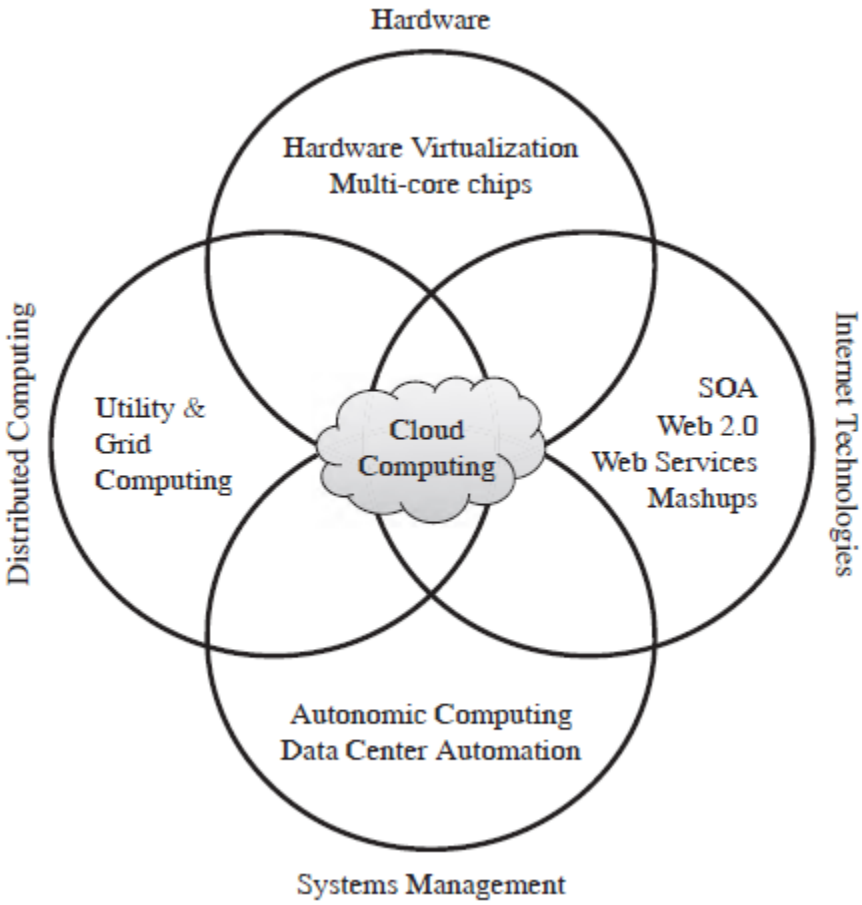
- 强大的武器：awk
- 怎样使用awk
- 生成脚本的技巧

云计算是什么？

- 相关软硬件厂商（争先恐后状）：我的产品就是云计算
- 不相关软硬件厂商（争风吃醋状）：云计算不就是一根网线加上计算机嘛
- 政府官员：云计算就是超级计算机
- 广大围观者：云计算就是集群？Or Hadoop？Or Openstack？Or Vmware？Or …？网格和云计算有什么差别？

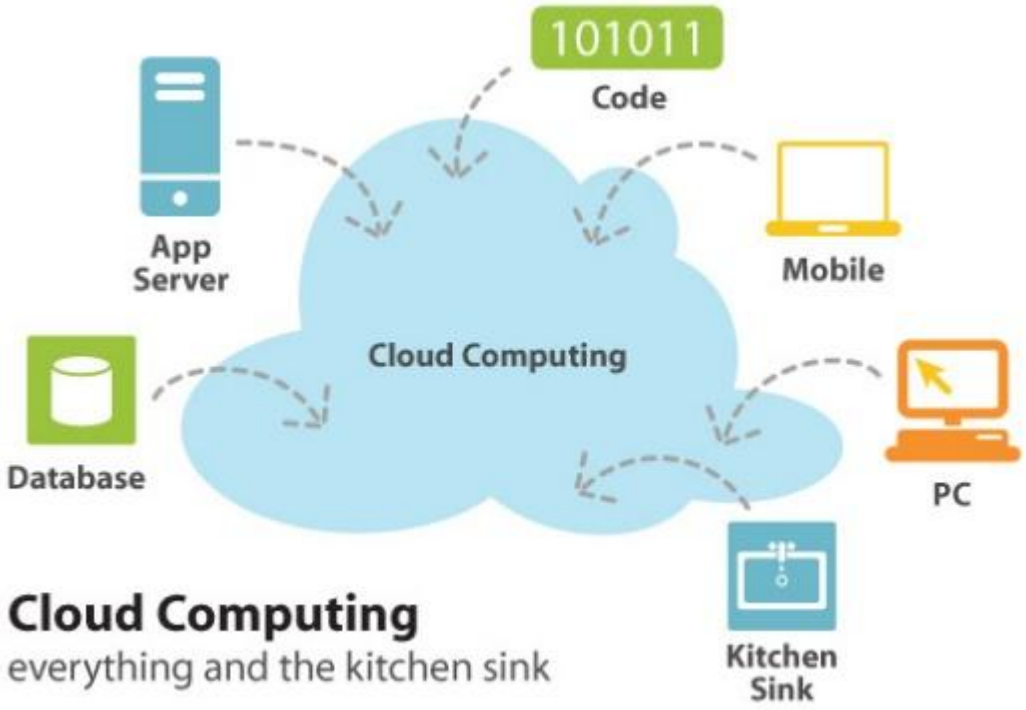
云计算是服务模式和拥有模式的革命

- **云计算是服务模式**：它不是新技术，更准确来说不应称之为技术，它是在一些关键技术日趋成熟后催生的一种新的服务模式
- 云计算通过**集中拥有**，使到用户能得到其本身无法得到的服务，或是以更低成本获得相同的服务，**降低拥有成本是云计算的核心价值**之一
- 云计算项目，**必先考虑服务模式和盈利模式的问题，其次才是投资和技术**



云计算的特征

- 自我服务
- 按使用量计费
- 弹性架构
- 可定制化



云计算怎样降低成本？

- 提高软硬件使用率
- 集中管理降低能耗
- 节约维护人员费用

能耗是日益严重的问题

- 2010年，美国计算机耗电量占总耗电量15%，预计到今年将翻一番
- 服务器在空转状态时的耗能，依然达到满载耗能的50%
- 现有关键计算硬件并非绿色设计，单位能源产生的计算能力成为重要指标。据某研究机构测试CPU降频5%，计算时间增加到原先1.04倍，但耗电降低50%

云计算模式也会增加成本

- 安全风险
- 可用性风险
- 绑架风险

盈利模式是云计算的核心问题

- 云计算领域的现状是项目找资金，资金找项目，折中点是有创意的盈利模式
- 互联网公司是云计算的先行者
- 技术相对于服务模式和盈利模式并不是门槛

云计算的形态

- 私有云
- 公有云
- 混合云

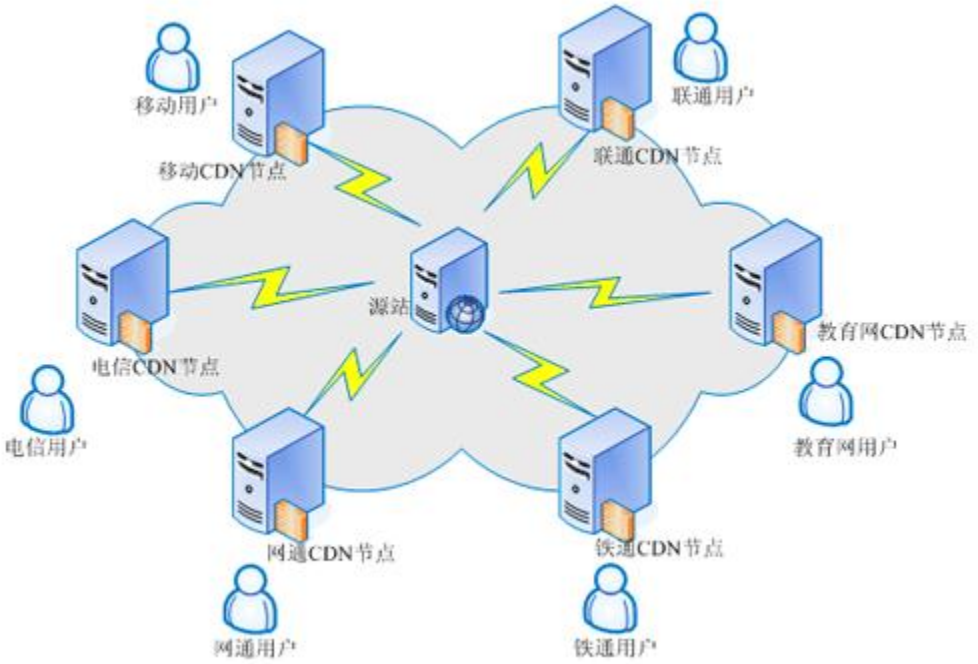
目前流行的开源云计算解决方案

- Hadoop
- Openstack

Hadoop在云计算中的用途

- 分布式文件系统提供的低单位成本的巨大的存储能力，高冗余度的可靠性
- Map-Reduce提供快速并行计算能力，这种能力可以随着节点数的增加线性递增

场景一：日志分析

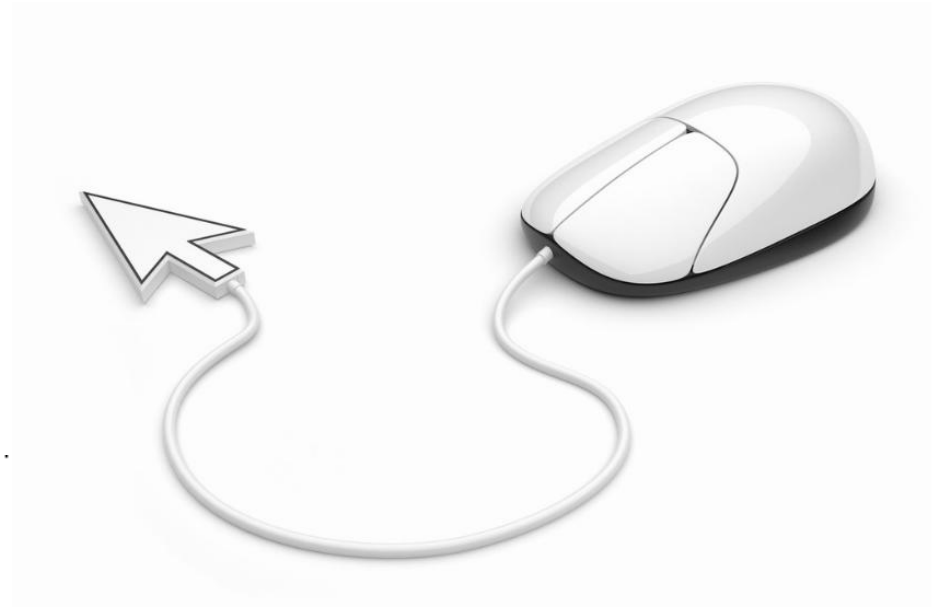

移动用户
移动CDN节点
联通用户
联通CDN节点
电信CDN节点
源站
教育网CDN节点
电信用户
教育网用户
网通CDN节点
铁通CDN节点
网通用户
铁通用户

探针设计

```
<script type="text/javascript">

  var _gaq = _gaq || [];
  _gaq.push(['_setAccount', 'UA-20237423-4']);
  _gaq.push(['_setDomainName', '.itpub.net']);
  _gaq.push(['_trackPageview']);

  (function() {
    var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
    ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-an
    var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
  })();

</script>
<div style="display:none">
<script type="text/javascript">
var _bdhmProtocol = (("https:" == document.location.protocol) ? " https://" : " http://");
document.write(unescape("%3Cscript src='" + _bdhmProtocol + "hm.baidu.com/h.js%3F5016281862f595e78
</script></div>
<!-- END STAT PV --></body>
</html>
```

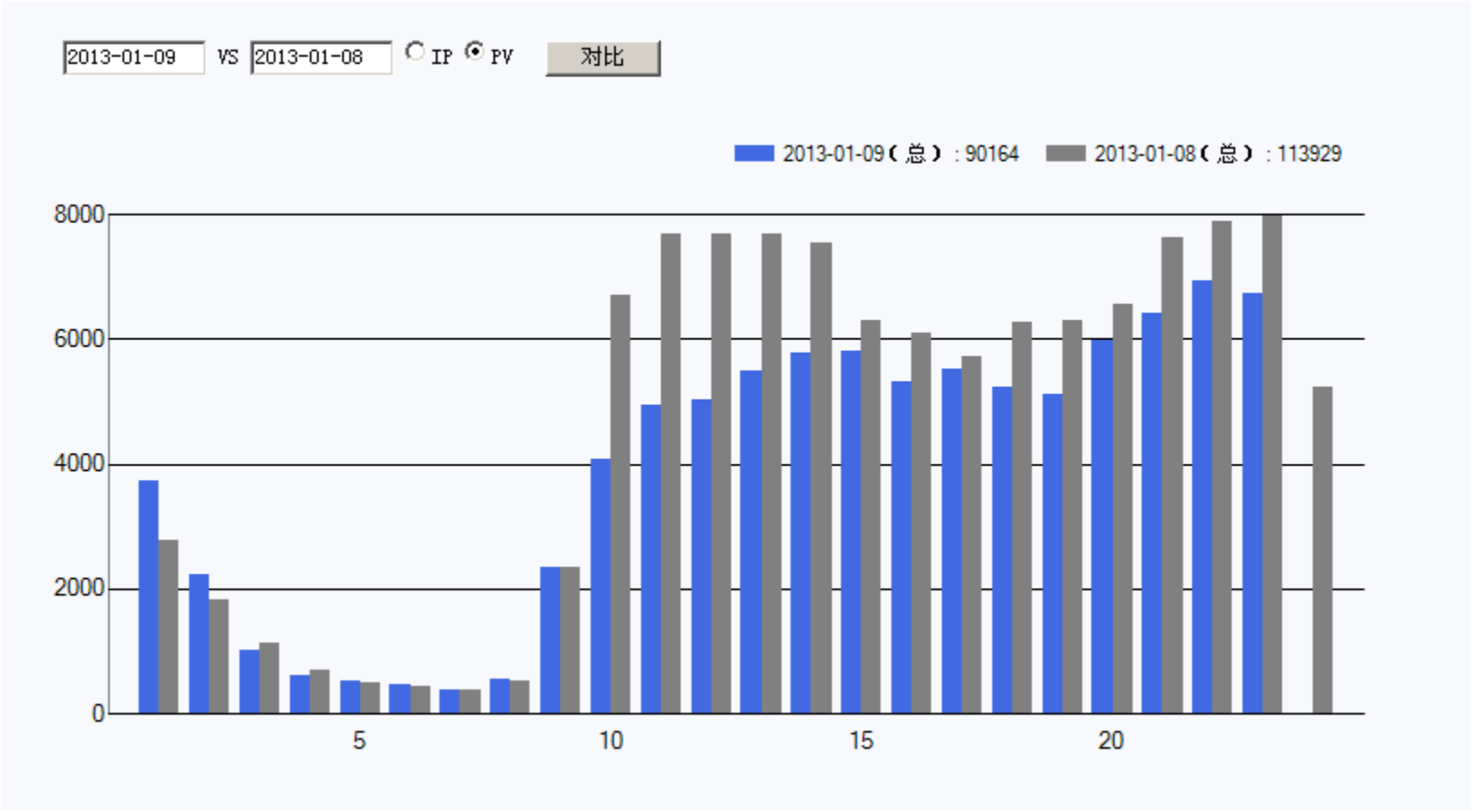
排除爬虫和程序点击，对抗作弊

- 用鼠标测动对抗爬虫
- 常用流量作弊手段
- 跟踪用户



```
t="sso_username_utf8";
break}})();
A(e,"mouseover",F,false);
A(e,"mousemove",F,false);
z=function(){var a,c;
if(e.body){a=e.body.clientWidth|e.documentElement.clientWidth;
c=e.body.clientHeight|e.documentElement.clientHeight}else{a=e.docu
```

需要的统计图表


2013-01-09
VS
2013-01-08
IP
PV
对比
2013-01-09(总) : 90164
2013-01-08(总) : 113929
8000
6000
4000
2000
0
5
10
15
20

遇到的问题

- 日志的保存需要大量的空间
- 日志的备份成本
- 统计时滞明显，不能满足业务要求

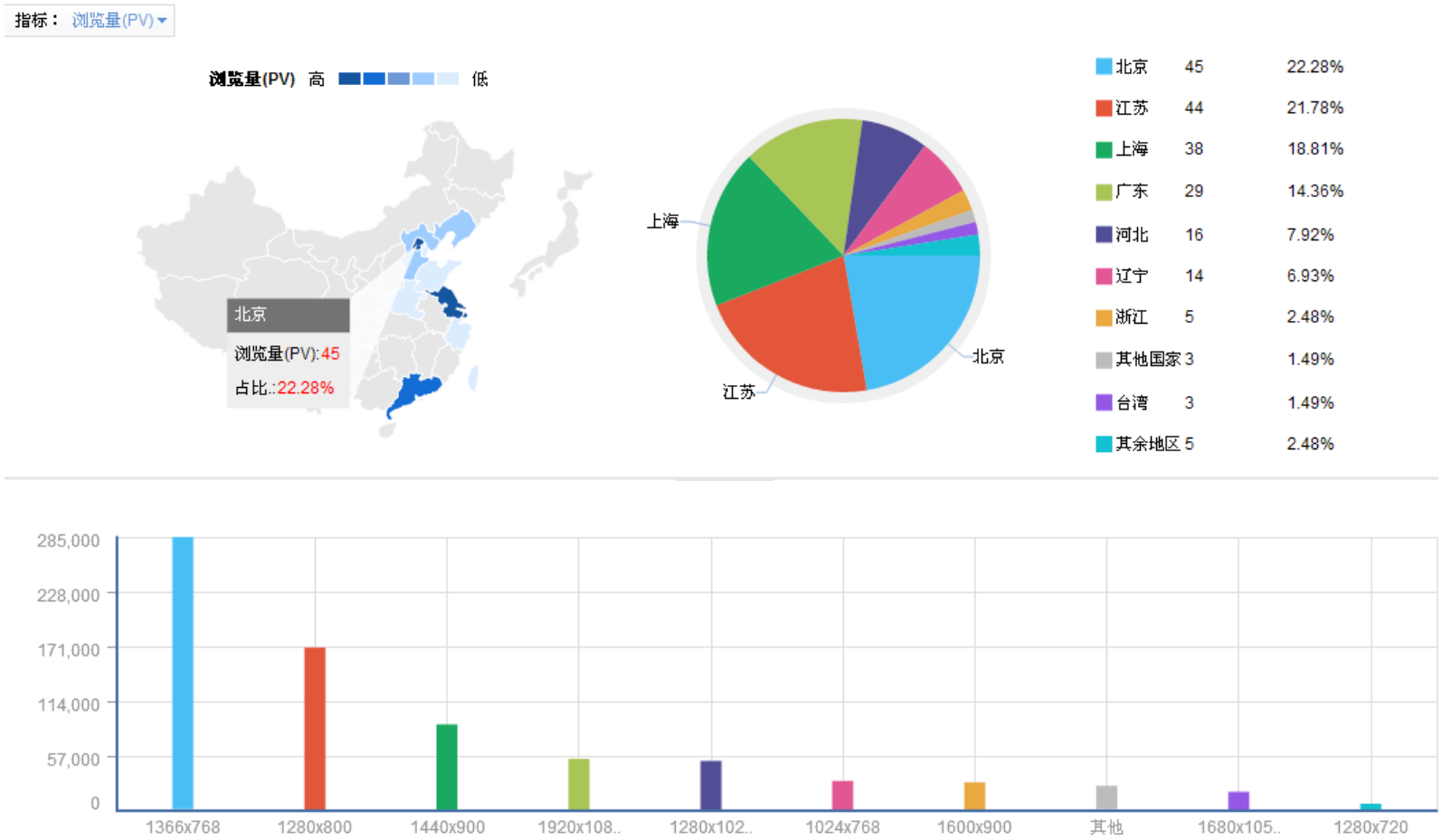
Hadoop方案

- 部署多个节点的Hadoop集群
- 探针激活java程序，在内存保存一定数量的日志信息后，利用API集中写入到HDFS
- HDFS既能保存日志，同时也提供了备份功能
- 用定时脚本清除过期的日志
- 用定时脚本激活pig进行统计，统计结果回写到输出文件
- 应用通过API读取输出文件里的数据，再展示给用户

Hadoop+Hbase方案

- 部署Hadoop+Hbase集群
- 探针激活java程序，程序把每条日志利用API集中写入到HBase（也考虑过批量入库）
- Hbase保存数据，它基于HDFS提供了冗余备份
- 利用时间戳和生存期自动清除过期日志
- 定时执行一java程序从hbase读出数据统计，结果写入mysql
- 应用直接从mysql中读出结果展示
- 本方案的优点是可以统计更为复杂的数据

复杂的统计图表



2013.01.08

复杂的统计图表

新访客

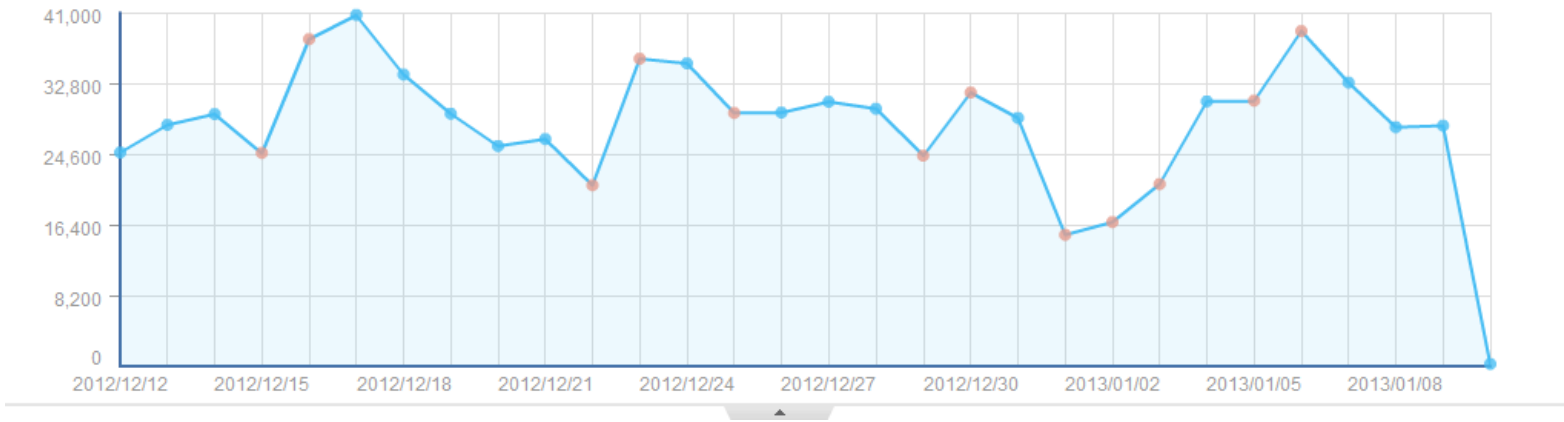
51.57%

浏览量：	134299
访客数：	29008
跳出率：	74.94%
平均访问时长：	00:07:10
平均访问页数：	3.6

老访客

48.43%

浏览量：	704296
访客数：	27243
跳出率：	32.59%
平均访问时长：	00:25:54
平均访问页数：	13.8



2013.01.08

场景二：某运营商数据分析实例

- 运营商网分程序：
  - 网分预处理程序
  - 网分位置统计程序
- 位置更新表

编号	字段名称	字段类型	中文名称	备注
1	CDRID	INTEGER	主键	
2	IMSI	VARCHAR2(50)	IMSI	
3	CGI	VARCHAR2(50)	位置信息	位置信息
4	STARTTIME	VARCHAR2(50)	当前时间	位置登记的时间
5	IMEI	VARCHAR2(50)		
6	UPDATETYPE	VARCHAR2(50)	更新类型	状态类型（开关机）
7	RELEASETIME	VARCHAR2(50)		此字段网分项目不需要。
8	OLDLAC	VARCHAR2(50)		
9	LCTUPDATEREJCAUSE	VARCHAR2(50)		
10	INSTIME	VARCHAR2(50)		

运营商Hadoop集群数据分析实例

- 网分预处理程序
- 输入：网分数据

```
6579585409657406184|460000722940589|460-00-10193-7513|2012-03-15 23:59:35|0126
```

- 输出：网分基础表

序号	字段	字段描述
1	IMSI	
2	IMEI	
3	CGI	
4	TIME	
5	CALLDUR	当是语音记录时，该字段表示通话时长。
6	UPDATETYPE	
7	MTCALLNUM	
8	MOCALLEDNUM	
9	RESOUCETYPE	数据来源表示：1、语音；2、短信、3、位置

运营商Hadoop集群数据分析实例

- 网分预处理程序

输入与输出格式必须上下文一致

```
public static class Map extends Mapper<LongWritable, Text, NullWritable, Text>
{
    public void map ( LongWritable key, Text value, Context context )
                                                    throws IOException, InterruptedException
    {
        String line = value.toString();
        TableLine tableline = new TableLine();

        try
        {
            tableline.fromLoc(line);
        }
        catch ( ArrayIndexOutOfBoundsException e)
        {
            context.getCounter(Counter.LINESKIP).increment(1);
            return;
        }

        context.write( NullWritable.get(), tableline.outText());
    }
}
```

输入

输出

把输入的行转换为String

//如果字段数不足则计数增加

输出Key和Value

运营商Hadoop集群数据分析实例

- 网分预处理程序

```
public int run(String[] args) throws Exception
{
    Configuration conf = getConf();

    Job job = new Job(conf, "WangFenPreprocess");
    job.setJarByClass(WangFenPreprocess.class);

    FileInputFormat.addInputPath( job, new Path(args[0]) );
    FileOutputFormat.setOutputPath( job, new Path(args[1]) );

    job.setMapperClass( Map.class );
    job.setOutputFormatClass( TextOutputFormat.class );
    job.setOutputKeyClass( NullWritable.class );
    job.setOutputValueClass( Text.class );

    job.waitForCompletion(true);
```

在网页显示
必须与类名一致

与程序输出格式一致

运营商Hadoop集群数据分析实例

- 网分位置统计程序
- 输入：网分预处理程序结果
- 输出：网分位置表

序号	字段	字段描述
1	IMSI	
2	CGI	
3	STAY_TIME	
4	UPCOUNT	上行指令次数
5	TIMFLAG	

运营商Hadoop集群数据分析实例

- 网分位置统计程序

上下文对应

```
public static class Map extends Mapper<LongWritable, Text, Text, Text>
{
    String date;
    String [] timepoint;

    /**
    public void setup ( Context context )
    {
    public void map ( LongWritable key, Text value, Context context ) throws IOException, I
    {
        String line = value.toString();
        TableLine tableline = new TableLine();


        context.write( tableline.outKey(), tableline.outValue() );
    }
```

先于所有的Map程序

运营商Hadoop集群数据分析实例

- 网分位置统计程序

```
public void reduce ( Text key, Iterable<Text> values, Context context ) throws IOExceptio
{
    String imsi = key.toString().split("\\|")[0];
    String timeflag = key.toString().split("\\|")[1];


    //用一个TreeMap记录时间
    TreeMap<Long, String> timeToCGI = new TreeMap<Long, String>();
    String valueString;

    for ( Text value : values )
    {
        valueString = value.toString();
        try
        {
            timeToCGI.put( Long.valueOf( valueString.split("\\|")[1] ), valueString.split
        }
        catch ( NumberFormatException e )
        {
            context.getCounter(Counter.TIMESKIP).increment(1);
            continue;
        }
    }
```

使用迭代获取所有Value

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间