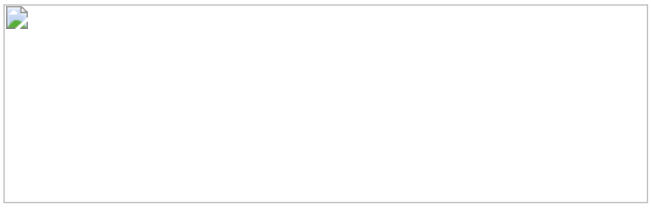
# 重画GoogleClusterTrace数据

2014-11-17 (2014-11-17 07:52:00) 128次阅读 google (/tags/google) cluster (/tags/cluster) trace (/tags/trace) 数据 (/tags/%E6%95%B0%E6%8D%AE)

由于项目计划书写作需要，重画了Qi Zhang, Mohamed Faten Zhani, Raouf Boutaba, Joseph L. Hellerstein, Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud. IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 2, NO. 1, JANUARY-MARCH 2014.中的TaskEvent分布统计图。原图更跟重画图如下：

原图：

重画图：



数据来源：

介绍：

https://code.google.com/p/googleclusterdata/wiki/ClusterData2011_1

所有文件列表及校验和：

https://commondatastorage.googleapis.com/clusterdata-2011-1/SHA256SUM

格式说明：

https://commondatastorage.googleapis.com/clusterdata-2011-1/schema.csv

数据文件示例连接：

https://commondatastorage.googleapis.com/clusterdata-2011-1/job_events/part-00017-of-00500.csv.gz

重画的步骤如下。

1 由于数据存放在https://commondatastorage.googleapis.com/clusterdata-2011-1/

需要FQ才能访问，故所有数据处理都是在墙外的位于东亚的azure服务器完成的。故首先建一个云服务器，并完成环境配置。

（主要是装个python）

2 下载数据文件（数据总量较大，1.51G）

```
import urllib2

url = 'https://commondatastorage.googleapis.com/clusterdata-2011-1/'
f = open('C:\\SHA256SUM')
l = f.readlines()
f.close()
for i in l:
    if i.count('task_events')>0:
        fileAddr = i.split()[1][1:]
        fileName = fileAddr.split('/')[1]
        print 'downloading', fileName
        data = urllib2.urlopen(url+fileAddr).read()
        print 'saving', fileName
        fileDown = open('C:\\task_events\\'+fileName, 'wb')
        fileDown.write(data)
        fileDown.close()
```

注意：

(1) 执行脚本前要将所有文件列表及校验和文件SHA256SUM

（https://commondatastorage.googleapis.com/clusterdata-2011-1/SHA256SUM）

放到C盘根目录下，它负责生成其他文件的下载链接。

(2) 这里只下载了task_events，如果要分析其他数据的话，参考前文提到的格式说明及介绍修改要下载的文件部分。

3 生成要处理的文件名

```
f = open('C:\\SHA256SUM')
l = f.readlines()
f.close()
fName = open('C:\\task_events_file_name.txt', 'w')
for i in l:
    if i.count('task_events')>0:
        fileAddr = i.split()[1][1:]
        fileName = fileAddr.split('/')[1]
        fName.write(fileName+'\r\n')
fName.close()
```

4 统计

```
import gzip

fName = open('C:\\task_events_file_name.txt')
fileNames = fName.readlines()
fName.close()
cntMapGratis = {}
cntMapProduction = {}
cntMapOthers = {}
#fileNames = ['part-00000-of-00500.csv.gz']
for l in fileNames:
    print 'now at: '+ l.strip()
    f = gzip.open('C:\\task_events\\'+l.strip())
    for log in f.readlines():
        log = log.split(',')
        if log[9]!='' and log[10]!='':
            index = log[9]+' '+log[10]
            priority = int(log[8])
            if priority <= 1: #Gratis Task
                cntMap = cntMapGratis
            elif priority >= 9 and priority <= 11:
                cntMap = cntMapProduction
            else:
                cntMap = cntMapOthers
            if not index in cntMap:
                cntMap[index]=1
            else:
                cntMap[index]+=1
    f.close()
fReasult = open('C:\\CPUandMEMuseGratis.txt', 'w')
for i in cntMapGratis:
    fReasult.write(i+' '+str(cntMapGratis[i])+"\r\n")
fReasult.close()

fReasult = open('C:\\CPUandMEMuseProduction.txt', 'w')
for i in cntMapProduction:
    fReasult.write(i+' '+str(cntMapProduction[i])+"\r\n")
fReasult.close()

fReasult = open('C:\\CPUandMEMuseOthers.txt', 'w')
for i in cntMapOthers:
    fReasult.write(i+' '+str(cntMapOthers[i])+"\r\n")
fReasult.close()
```

5 使用matlab绘制

clear all

close all

%load('D:\\CPUandMEMuseGratis.txt')

%load('D:\\CPUandMEMuseProduction.txt')

load('D:\\CPUandMEMuseOther.txt')

%CPUandMEMuse = CPUandMEMuseGratis;

%CPUandMEMuse = CPUandMEMuseProduction;

CPUandMEMuse = CPUandMEMuseOther;

x=CPUandMEMuse(:,1);

```
y= CPUandMEMuse(:,2);
s = CPUandMEMuse(:,3)/10000000;
s = log(s);

%max_r = 0.002; %for production and gratis
max_r = 0.001; %for other only
s = s/max(s)*max_r;

for i=1:size(x)
if x(i) == 0 || y(i) == 0
s(i)=0;
end
end

t= 0:pi/10:2*pi;
figure();
grid on
for i=1:size(x)
if x(i)~=0 && y(i)~=0
pb=patch((s(i)*sin(t)*0.5+ x(i)),(s(i)*cos(t)+y(i)),'b','edgecolor','k');
alpha(pb,.3);
end
end
axis([0 0.5 0 1]);
xlabel('CPU size');
ylabel('Memory size');
set(gca,'FontSize',25);
set(get(gca,'XLabel'),'FontSize',30);
set(get(gca,'YLabel'),'FontSize',30);

%saveas(gcf,'D:\\CPUandMEMuseGratis.jpg')
%saveas(gcf,'D:\\CPUandMEMuseProduction.jpg')
saveas(gcf,'D:\\CPUandMEMDemandOther.jpg')
```

附注：

1. Task通过优先级划分类别的

0-1 是Gratis

9-11 是Production

其他（2-8）是Other

2. 画图的时候，圆的半径表示数量的对数（log）

相关文章


- 全国最全的省，市，县，电话号前缀，邮编数据 (/posts/yLGv728)
- 从天气网提取气象预报数据 (/posts/lpbf20)
- ThinkPHP 3.2.3 数据缓存与静态缓存 (/posts/ECV656e)
- 两个步骤修改Mac、windows、类linux系统（ubuntu系统为例）的hosts文件来访问Google (/posts/EDQo0e5)
- 同步的数据过大，导致shareplex超时，并自动kill掉了同步会话 (/posts/ETalbeb)
- jQuery validate 表单验证，涵盖各种类型数据 (/posts/yZD7cc4)
- 向sqlserver插入二进制数据(比如图片) (/posts/oQh0d5)
- Android学习之Adapter(数据适配器) (/posts/yXzZe73)
- 8.数据库函数 (/posts/yBavaa4)
- 远程查询批量导入数据 (/posts/sf988e)

含 GOOGLE 标签的热门文章

- 两个步骤修改Mac、windows、类linux系统（ubuntu系统为例）的hosts文件来访问Google (/posts/EDQo0e5)
- Google的云存储技术：Google Storage的开通试用及其API的简单应用 (/posts/y3xB898)
- Google、Oracle CEO就Android专利问题共同出庭 (/posts/ydM6235)
- Google中国业务可能缺乏许可证 (/posts/4vo1ec)
- PDA开发系列：Google地图接口 (/posts/30a638)
- Google Gson解析Json数据应用实例 (/posts/yFjoec8)
- 如何提高网站的Google RP值 (/posts/ynzMdf3)
- CwCity.de – 德国老牌无限容量免费PHP空间 Google翻译 – 最好的免费在线翻译网站+为你的网站建立多语言版本 (/posts/Echp987)
- Google笔试题_改进 (/posts/yU4G002)
- Google与微软为jQuery等类库提供的CDN服务 (/posts/EJyw580)

含 CLUSTER 标签的热门文章

- Cluster (/posts/LpG2b6)
- 记录：A Cluster-Based Resampling Method for Pseudo-Relevance Feedback (/posts/EHP8760)
- 转【实战体验几种MySQLCluster方案】 (/posts/NVsd3f)
- Hadoop建立Cluster实例 (/posts/y53je34)
- 安装配置OPENCMS的Replication cluster(从)详细过程 (/posts/MUdcf6)
- helios架构详解（二）客户端架构和cluster (/posts/WyH70d2)
- redis-集群（cluster）扫盲篇（一） (/posts/W9qLed0)
- ML_Clustering & Retrieval（一） (/posts/AYyr33c)
- K-means:如何选择K(cluster的数目) (/posts/WIC2937)
- MapReduce:Simplified Data Processing on Large Clusters(中文翻译3) (/posts/yRSF84d)

含 TRACE 标签的热门文章

- .NET中的跟踪与调试(Trace&Debug) (/posts/K1wd17)
- enable Assembly Load Trace (/posts/Q6U22d)
- 工具04_SQL Trace/DBMS_SYSTEM (/posts/yn8y4f1)
- 将FlashPlayerDebugger的trace()功能输出到日志 (/posts/y4v0726)
- Linux网络应用编程之集线器（Packet Tracer仿真） (/posts/W3U559d)
- php实现网页trace方法 (/posts/yjPQefb)
- .Net 中的反射(动态创建类型实例) - Part.4（转自http://www.tracefact.net/CLR-and-Framework/Reflection-Part4.aspx） (/posts/JDqf00)
- MFC中TRACE (/posts/Wi7f850)
- 使用Debug和Trace (/posts/3dz19f)
- tracert命令 (/posts/aBKac2)

---

上一篇：使用Python爬取mobi格式电纸书 (/posts/EnDT3db)
下一篇：Python的并行求和例子 (/posts/EnDh962)

相关文章

- 全国最全的省，市，县，电话号前缀，邮编数据 (/posts/yLGv728)
- 从天气网提取气象预报数据 (/posts/lpbf20)