|  |
| --- |
|  |

|  |
| --- |
| **新统计hadoop数据处理业务** |
| 统计平台项目组 |
|  |
|  |
|  |
| **Kendy** |
| **2014-10-30** |
|  |

目录

[修改历史记录 1](#_Toc402447437)

[一、Hadoop集群共几台机器，ip都是什么？ 2](#_Toc402447438)

[二、数据从何处上传而来？上传逻辑是什么？ 2](#_Toc402447439)

[1、数据从何处上传而来？（六个server + 一个备机） 2](#_Toc402447440)

[2、上传逻辑是什么？ 2](#_Toc402447441)

[三、Hadoop源数据路径，路径格式是什么？数据格式是什么？ 3](#_Toc402447442)

[四、加工项脚本存放位置？程序存放位置？log存放位置？什么时候定时执行？ 3](#_Toc402447443)

[五、加工项处理都有哪些？ 3](#_Toc402447444)

[六、日报都计算了哪些数据？日报分哪几类？如何发日报？如何修日报？ 4](#_Toc402447445)

[七、页面上新统计数据在数据库中如何查询？ 4](#_Toc402447446)

[八、新统计数据库是如何分布的？配置库中有哪些比较重要的表？ 4](#_Toc402447447)

修改历史记录

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 修改人 | 颜色 | 备注 |
| 2014-10-30 | Kendy | 黑色 | 新建 |

**本文档主要记录了统计后台hadoop数据处理业务相关的内容，方便后面查询和参考。**

一、Hadoop集群共几台机器，ip都是什么？

|  |  |  |
| --- | --- | --- |
| Namenode | Secondary namenode | Datanone |
| 192.168.11.128 | 192.168.11.129 | 192.168.11.81 |
| 192.168.11.79 |
| 192.168.11.74 |
| 192.168.11.66 |
| 192.168.11.63 |
| 192.168.11.55 |
| 192.168.11.54 |
| 192.168.11.26 |
| 192.168.11.25 |
| 192.168.11.24 |
| 192.168.11.23 |
| 192.168.11.122 |
| 192.168.11.121 |
| 192.168.11.120 |
| 192.168.11.119 |
| 192.168.11.118 |
| 192.168.11.116 |
| 192.168.11.110 |
| 192.168.11.109 |
| 192.168.11.108 |
| 192.168.11.105 |
| 192.168.11.104 |
| 192.168.11.103 |
| 192.168.11.102 |
| 192.168.11.101 |

二、数据从何处上传而来？上传逻辑是什么？

# 1、数据从何处上传而来？（六个server + 一个备机）

|  |  |  |
| --- | --- | --- |
| Stat-server | 服务部署目录 | 数据文件存放目录 |
| 192.168.71.65 | /opt/taomee/stat/stat-server | /opt/taomee/stat/stat-server-data |
| 192.168.71.75 | /opt/taomee/stat/stat-server | /opt/taomee/stat/stat-server-data |
| 192.168.71.52 | /opt/taomee/stat/stat-overseas | /opt/taomee/stat/stat-overseas/stat-server-data |
| 192.168.1.72 | /opt/taomee/tongji/stat-server | /opt/taomee/tongji/stat-server-date |
| 192.168.1.177 | /opt/taomee/stat/stat-server | /opt/taomee/stat/stat-server-data |
| 192.168.1.88 | /opt/taomee/stat/stat-server | /opt/taomee/stat/stat-server-data |
| 192.168.71.43 | /opt/taomee/stat/stat-server | /opt/taomee/stat/stat-server-data |

# 2、上传逻辑是什么？

A. 登陆用户和密码：hadoop （hadoop@tm@0529）

B. Upload服务目录：/opt/taomee/hadoop/upload

C. 上传时间：每分钟



D. 源程序：com.taomee.bigdata.upload.Uploader

E. Jar包：stat-bigdata.jar

F. 本地数据：/opt/taomee/stat/stat-server-data/sent/

G. 上传到hadoop目录：/bigdata/input/

H. 本地备份目录：/opt/taomee/stat/stat-server-data/backup

I. 逻辑：本地数据先上传到hadoop的input目录中，然后备份到本地backup目录下，然后将已经放到备份目录的本地目录中数据删除。

三、Hadoop源数据路径，路径格式是什么？数据格式是什么？

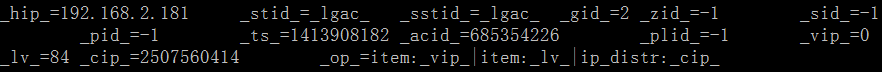
Basic数据：



Custom数据：



源数据格式：



服务端ip|统计项|子统计项|游戏id|区|服|平台|时间|账号|角色|vip|等级|客户端ip|op

四、加工项脚本存放位置？程序存放位置？log存放位置？什么时候定时执行？

1. 脚本目录：[hadoop@192.168.11.129:~/bigdata](mailto:hadoop@192.168.11.129:~/bigdata)
2. 程序存放目录：

<http://svn.taomee.com/data-center/server/trunk/stat-bigdata/src/com/taomee>

1. Log目录：[hadoop@192.168.11.129:~/bigdata/log](mailto:hadoop@192.168.11.129:~/bigdata/log)
2. 定时执行时间



凌晨两点会计算上一天所有数据



周一到周五，9点到18点，每隔两小时会计算新增等数据



每天下午两点会计算海外数据

五、加工项处理都有哪些？

1. 主要对以下几类数据进行加工

新增活跃(包括 留存、流失、回流) | 付费数据 | 分布数据 | 其他

1. 每天首先计算日报数据，日报数据完成后会计算后续页面基础项数据，页面基础项数据计算完，最后计算自定义数据。

六、日报都计算了哪些数据？日报分哪几类？如何发日报？如何修日报？

1. 日报主要分以下几类：

|  |  |  |
| --- | --- | --- |
| 日报分类 | 计算日报服务器 | 发送时间 |
| 页游日报（11封） | 192.128.11.129对应页面“核心数据” | 9点会自动检测发送，没发出的需要手动发送 |
| 注册转化日报 | 192.168.71.36 | 9:30自动测试发送 10:30自动正式发送 |
| 创想兵团日报 | 192.168.71.57 | 9:32自动发出 |
| 新增用户数，UV日报 | 192.168.11.128 | 10:13自动发出 |

1. 页游日报主要计算以下数据：

新增 | 活跃 | 付费（总付费，按条，包月）| 留存率（新增，活跃）|在线 |PCU|ARPPU

1. 日报有问题如何修（见文档《日报发送相关》 ）

七、页面上新统计数据在数据库中如何查询？

见文档《新统计数据库查找数据和处理数据》

八、新统计数据库是如何分布的？配置库中有哪些比较重要的表？

1. 100个库被分到三个服务器上，相应的库对应如下：

|  |  |  |
| --- | --- | --- |
| 192.168.71.72 | 192.168.71.43 | 192.168.71.76 |
| 0-29 | 30-59 | 60-99 |

2. 平时会经常查看的几个表

|  |  |
| --- | --- |
| 表名 | 用途 |
| t\_game\_info | 存放了游戏信息 |
| t\_report\_info | 存放report\_id 等信息 |
| t\_common\_task | 所有的加工项对应的task\_id |
| t\_common\_result | 加工结果表 |
| t\_data\_info | 数据项映射表 |
| t\_gpzs\_info | 存放每个游戏的gpzsid |