**IUNI大数据分析平台方案**

V1.0.0

## 目录

[一．总体架构图 3](#_Toc391883981)

[二．日志收集 3](#_Toc391883982)

[2.1 收集步骤 3](#_Toc391883983)

[2.2 清洗规则 3](#_Toc391883984)

[2.3 处理流程 4](#_Toc391883985)

[2.4 示例配置文件附件 5](#_Toc391883986)

[2.5 包含模块 5](#_Toc391883987)

[三．数据仓库 5](#_Toc391883988)

[3.1 功能描述 5](#_Toc391883989)

[3.2 表设计 5](#_Toc391883990)

[四．数据分析/挖掘 6](#_Toc391883991)

[4.1 功能描述 6](#_Toc391883992)

[4.2 需求描述 6](#_Toc391883993)

[五．数据接口 6](#_Toc391883994)

[5.1 数据库 6](#_Toc391883995)

[5.2 消息队列 6](#_Toc391883996)

[5.3 Web service 6](#_Toc391883997)

[六．部署方案 7](#_Toc391883998)

[七．附录 8](#_Toc391883999)

[八．修订历史 8](#_Toc391884000)

## 一．总体架构图



## 二．日志收集

### 2.1 收集步骤

1. 在各业务服务器上部署一个flume agent，实时监控业务服务器日志文件收集日志，验证日志格式是否正确，写入日志产生时间（时间戳形式）、服务器IP地址、响应时间（若日志记录为0，则写为0.001，表示1ms），访问速度（消息大小 / 响应时间），ad\_id，用户ID，清洗日志。
2. Guard守护本地的flume agent以保证其一直运行，退出时则启动本地的flume agent。
3. Collector收到日志后将日志分别写入hdfs和hbase。

### 2.2 清洗规则

1. 通用规则：日志格式需匹配正则表达式，正则表达式由配置文件配置；
2. 存入HDFS规则：

2.1 将匹配正则表达式配置的日志存入usefull文件，不匹配的日志保存到useless文件。

eg：将status值为2开头的文件存入usefull文件，配置正则表达式为：^2\\d+\\d+$。

2.2 文件需要保存为text格式。

2.3 可配置为按时间（单位秒，一天即为86400秒），生成日志文件。

2.4 可配置文件名前缀后缀，以便按照日志内容保存到不同的文件。

eg：按服务器名保存 filePrefix = test-%{hostname}。

* 1. 日志文件行数限制，当日志文件达到配置的数量时，便产生新日志文件。

1. 存入HBASE规则

3.1 status为2开头的，同2.1；

3.2 URL请求非静态资源，由配置项staticRes指定静态资源，多种静态资源类型由空格分开；

eg: staticRes = css js jpg jpeg png gif ico img bmp min small，表示url为css js jpg jpeg png gif ico img bmp min small的静态资源请求都不存入hbase。

3.3 配置文件中指定的特殊的URL的静态资源请求，

eg: staticRes.jpg.specialUrl = specialUrl，表示url中包含specialUrl的jpg请求需要存入hbase。

### 2.3 处理流程



### 2.4 示例配置文件附件



### 2.5 包含模块

1. Guard
   1. 守护本地agent，可指定进程名称，默认为flume-agent（参数1）；
   2. 发送日志文件，将指定目录下的日志文件重命名，并发送到collector，若发送成功，则将其MV到arch目录（参数2）；
   3. 提供删除指定天数以前的文件（参数3）；
2. Flume iterator

验证日志格式是否正确，若正确，写入如下信息：日志产生时间（时间戳形式），服务器IP地址，域名，响应时间（若该字段值为0 ，则写为1，表示1ms），访问速度（消息大小 / 响应时间）, ad\_id；

1. Flume file-roll

将日志文件按条件写成两份，一份为完整的日志，后缀为full，另一份为有需要的日志，后缀为use，（后缀可配置）；

1. Flume hbase-serializer

按分隔符取成N列，可配置各列名；

1. Hive loader

将现有历史日志导入HIVE。

## 三．数据仓库

### 3.1 功能描述

日志记录写入HBASE中保存，并使用HIVE创建关联表与HBASE中的日志记录表关联，以方便管理和查询。

### 3.2 表设计

HIVE-HBASE关联表建表语句示例：

HBASE表设计：

1. 开启SNAPPY压缩；
2. 使用bloomfilter；
3. 预分配region，数量为100；
4. Rowkey设置为Long.toHexString(current timestamp). Reverse() + log timestmap + auto increment id；
5. 使用二级索引，索引表设计:region start key + index name + index column value + user table rowkey；

## 四．数据分析/挖掘

### 4.1 功能描述

使用MapReduce、UDF、UDAF、Impala、Mahout等工具进行数据分析与挖掘。

### 4.2 需求描述

1. 业务需求
   * 1. 网站访问指标：

PV：访问量统计；

UV-IP：独立IP统计；

UV-VK：独立cookie统计；

Source：用户来源域名统计；

Channel：用户访问渠道；

Browser：用户访问设备统计；

Ip-Count：Ip访问总次数；

Ip-Time：Ip访问总时间；

1. 运维需求

## 五．数据接口

### 5.1 数据库

将已经处理好的数据写入关系型数据库（Mysql/Oracle），开放关系型数据库提供给业务系统查询。

### 5.2 消息队列

定时将处理好的数据写入消息队列（ActiveMQ/RabbitMQ/jms），供业务系统查询。

### 5.3 Web service

开放数据处理服务接口供其他系统调用，详见附录[《IUNI大数据分析接口规范》](#_7.1_IUNI大数据分析接口规范)。

## 六．部署方案

详见附录[《IUNI大数据分析平台部署方案》](#_7.2_IUNI大数据分析平台部署方案)。

## 七．附录

### 7.1 IUNI大数据分析接口规范

### 7.2 IUNI大数据分析平台部署方案

## 八．修订历史

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 修订人 | 版本号 | 备注 |
| 2014-06-30 | 陈智伟 | V1.0.0 | 初步文档制定 |