实时检测共享账户概要设计说明书

目录

[1引言 3](#_Toc529813899)

[1.1编写目的 3](#_Toc529813900)

[2背景 3](#_Toc529813901)

[3 项目架构 3](#_Toc529813902)

[4概要流程图 4](#_Toc529813903)

[5各个模块设计描述 5](#_Toc529813907)

[5.1数据来源模块描述 5](#_Toc529813908)

[5.2 数据存储模块描述 5](#_Toc529813909)

[5.3数据处理模块描述 6](#_Toc529813910)

[5.4数据展示模块描述 6](#_Toc529813911)

## **1引言**

### 1.1编写目的

利用大数据技术采集现网流量流向信息，根据这些原始信息运用计算框架判断此账户是否异常(超过一定阈值则触发报警机制，例如阈值为5时，同一个账号同时连接的终端数量不能超过该值，如果超过则报警)。

## 2背景

随着通信行业的普及，以及人们对网络的需求越来越大，采集现网流量的流向信息，可以通过实时监测数据来判断宽带账号的同时接入数量，把控共享账户的的阈值，如果超过规定数量会报警，这样有效控制了一个账号同时链接的终端数量，有效保护该账号安全，同时符合运营商的利益。

## 3 项目架构

|  |  |
| --- | --- |
| 技术 | 作用 |
| Hadoop | 分布式系统的基础架构 |
| Kafka | Kafka是一种高吞吐量的[分布式](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F/19276232)发布订阅消息系统，它可以处理消费者规模的网站中的所有动作流数据 |
| Storm | Storm是Twitter开源的、分布式的、高容错的实时计算系统 |
| 数据可视化 | 用前端技术，把统计出的数据展示出来 |

## 4概要流程图

根据前面的功能需求分析和系统整体的需求，将整个架构分为以下几层：数据采集层、数据处理层、数据存储层和数据展示层。

详细的流程图会在详细设计中给到。

数据可视化

数据存储

流式计算

数据采集

实时检测共享账户系统的总体设计是从运营商骨干网那里拿到数据信息进行数据分析，设计该系统的目标是将运营商提供的信息进行加工处理统计出异常账号，在总体设计时都需要考虑到这些问题，还要在进行详细设计时给出相关架构。

为了最后实现该系统，在总体设计中把握以下重点，然后在详细设计中可以分别对每个细节难点和重点一一解决。总体设计中包括以下核心的问题，在进行程序详细设计时需要针对这些个难点一一对应的进行分析：

(1)各大运营商每天产生的数据量是非常庞大的，要从各大运营商的数据库中获取所有的有效数据并且进行过滤、清洗是一项很复杂的工作，要实现数据的获取、转换以及加载的过程就需要把所有的可能有关的源数据和它们的数据的格式，以及Flume接收数据的格式，都清晰的整理出来才可以进行接下来的操作，所以Flume组件合理地设计和使用是获取和数据存储的关键。

(2)页面与数据库MySQL接口的对接问题。在页面开发人员对数据进行操作时，Web要与MySQL进行交互，它们两者的接口就需要互相统一，并且要有较高的效率。并且让其用可视化框架展示所统计出的结果，所以应该运用Ajax技术对页面的部分进行实时刷新，而且在编写MySQL接口的时候，需要考虑到它们的性能、效率、优化等问题，让接口符合系统的要求。

## 5各个模块设计描述

### 5.1数据来源模块描述

数据即金钱，数据是一个公司进行用户行为分析，数据统计，互联网金融等领域所必备的基础，那么获取数据以及数据的清洗汇聚是较为重要的，操作该项目的数据来源于运营商的业务系统传统的关系型数据库。

### 5.2 数据存储模块描述

本模块主要有两种方式，第一种是Flume获取数据源放入到Hadoop的HDFS分布式文件系统中，Storm再从HDFS中获取数据进行分析，第二种就是把Flume采集数据通过kafka的producer存储到缓存borker中，Storm从borker中获取数据，进行数据分析，将处理好的数据存储到MySQL数据库中。

但具体用哪一个还需要进行实际的运行，控制bug的频发，实现代码的优化，效率的提升。

### 5.3数据处理模块描述

该模块用到的技术是Hadoop生态圈中的Storm分布式计算框架。运用Storm中的Spout对HDFS或者kafka的borker中的文件进行处理，最后把处理好的数据存储到MySQL数据库中。

### 5.4数据展示模块描述

通过数据可视化技术将分析结果绘制图表展示到Web页面，为服务商的运营决策提供支持。

6.检测方法

5分钟内，同一个账号（UserAccount）下，只要满足下面的任意一个条件，表示出现共享账户问题：

* natIP去重求和数 > 5
* qqid去重求和数 > 20
* cookieValue + devName + osName去重求和数 > 5

7.统计需求

将1天中，每5分钟的异常账号总数绘制图表（只求异常账号总数）