

**创新实验课程报告**

专 业 计算机科学与技术

学　　 号 1150310609

小 组 成 员 黄道龙（1150310613）

王陈阳（1150310609）

# 课题简介

## 1.1课题名称

分布式大规模网站访问日志管理

## 1.2课题任务

（1）假设一个网站有多个服务器，用户对网站的访问会随机分配到任意一台服务器。

（2）访问日志的格式 为 时间+用户IP的格式

（3）使用学过的数据结构和Maprduce架构，实现一个访问日志统计系统。

（4）实现日志汇聚，统计top k功能

（5）对极短时间内多次访问的用户，添加到临时黑名单结构中。

（6）自己构造足够量级的测试数据进行测试

# 运行环境

本次实验在WIndows 10系统下，利用Python 3.7语言进行开发测验。

# 数据生成

3.1数据存储格式

本实验的数据是模拟多个服务器被访问时产生的日志数据，所以生成的数据由多个文本文件保存，不同文本文件代表不同的服务器，并且每个文本文件的访问日志存储格式如下：

访问IP Access\_time

访问IP Access\_time

访问IP Access\_time

....................

其中访问IP的格式为标准的IPv4地址，Access\_time的格式是年-月-日-时-分-秒，即访问日志的时间的细粒度为秒。

3.2数据生成主要方法

3.2.1访问IP生成

通过随机生成在特定域的随机数，将其拼接为标准的IPv4地址作为本条日志的访问IP地址。

3.2.2访问时间的生成

先输入日志记录的开始时间及其日志记录范围，利用Unix存在的时间戳概念，将开始时间转换为时间戳。再随机生成日志记录范围内对应秒数的整数值，将两数相加，得到一个新的时间戳，该时间戳即为一个随机的访问时间，在将其转换为对应的访问时间的日志格式作为该条日志的访问时间。

3.3数据生成流程图

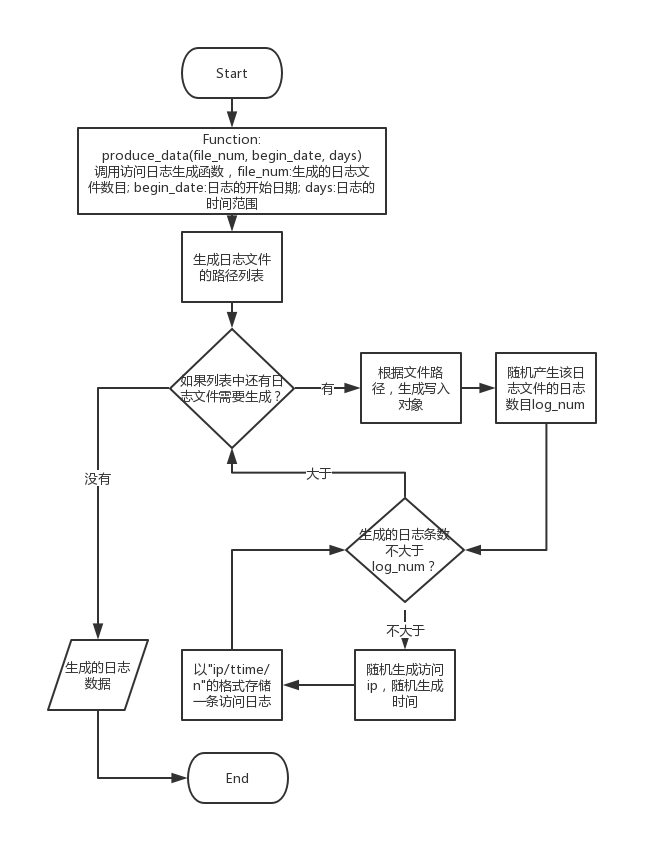


图1.数据生成流程图

3.4 数据的说明

本次实验中，生成的数据一共有100个文本数据文件，代表100个访问服务器的日志数据，日志数据一共有

# TopK

4.1主要方法

利用map-reduce的编程框架，对数据进行了两次map-reduce操作，将高频访问的IP按照访问次数排序输出。

4.2第一次map-reduce

4.2.1 map过程

读取生成的日志文本数据，将每一条日志文本数据：访问IP Access\_time以key:IP value：1的数据对存入程序中

4.2.2 reduce过程

将程序中存在的每一个key:IP相同的合并，其value为相同的IP的次数，即输出结果为：IP 访问次数

4.3 第二次map-reduce

4.3.1 map过程

将第一次map -reduce的结果以形式：key：访问次数 value：IP的数据对存储在程序中

4.3.2 reduce过程

利用map-reduce框架的自动排序，将访问次数高的数据对排序到前面，然后将其结果输入到一个文本文件中，该文件的内容即为所求TopK的次序

4.3程序流程图

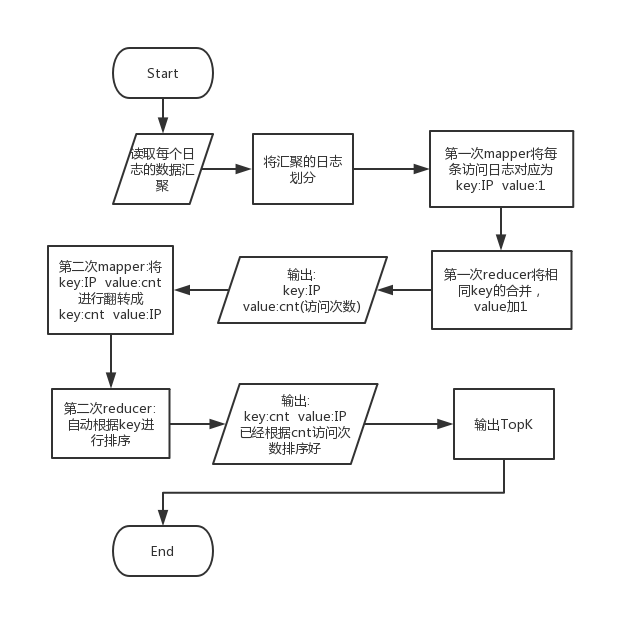


图2.TopK程序流程图