

# 需求分析

项目名称： 互联网搜索日志数据挖掘

项目类别：  
☐ 电子商务  
☐ 移动终端应用  
☒ 大数据分析  
☐ 物联网应用  
☐ 人机交互应用  
☐ 其他( )

命题企业： 北京瑞德云网科技有限公司

咨询邮箱： zhangguangjun@ict.ac.cn

2017 年 12 月 1 日

# 项目需求分析

## 一、引言

### 1.1 项目背景

随着经济的发展，互联网已经成为人们生活的一部分，方便了生活的方方面面。互联网的普及也引发了一系列的产业的发展，网购就是其中最重要的标识，同时，为了达到更加精准的商品推荐，更加高效的广告投放成了互联网公司的主要发展方向。大量的网民上网日志来分析用户的一些喜好，分析上网时间、频率、浏览的内容等等进行多维度的分析，进而进行精准的广告投放与营销。

### 1.2 项目目的

通过真实的互联网搜索日志应用的开发案例，能够帮助学生快速的掌握大数据应用开发流程，掌握大数据应用的核心技术，能在以后的工作和学习中做到有的放矢，更加的游刃有余。本案例是企业真实的大数据应用的简缩版，主要在数据量和数据脱敏方面做了处理，旨在帮助同学掌握真实的大数据应用实现的案例。

## 二、项目需求

### 2.1 功能需求

#### 2.1.1 数据录入

互联网搜索日志的录入，搜索日志可以通过搜索系统提取得到，搜索日志可以是实时收集流入大数据集群也可以是定时的同步到大数据集群。本项目采用的是已经收集好的日志数据文本，对整个大数据处理流程做分析的项目，数据已经准备好。上传到大数据集群的 HDFS 文件系统中即可完成数据的录入。

#### 2.1.2 数据处理

对现有的网络搜索日志进行清洗处理，最终处理成结构化的数据储存到 HDFS 文件系统中，提供数据分析提取分析。

#### 2.1.3 数据分析

对清洗过后的搜索日志数据进行分析，根据现有的数据字段属性情况，可以对互联网搜索日志数据进行统计分析，能够快速统计出当天或者是实时的热词词频，统计实时的关键词的 TOP10 报表，以及搜索词频上升和下降的情况，绘制出搜索关键词的词云图。

#### 2.1.4 数据展现

使用统计的结果搜索的热词进行报表的绘制，根据词频情况绘制

出词云图。

## 2.2 性能需求

### 2.2.2 可扩展性

大数据集群可以快速无缝的横向扩展，数据达到一定的规模之后不需要担心数据承载出现问题，加存储或者服务器即可完成集群规模的横向扩展。数据会根据集群的情况合理调整调度，尽可能的合理使用资源。

### 2.2.3 稳定性

采用的是大数据领域主流的组件进行开发，每个技术都是非常领先。大数据处理平台采用的是高可用，数据副本机制保障数据的安全。并且在行业已经运用到各行各业中，数据的规模也是在 PB 级别，社区活跃度非常高，在运维和开发成本上相对较低，可持续性较强。

## 2.3 任务要求

### 2.3.1 大数据平台搭建

能通过安装文档个人完整的把 CRH 技术平台搭建完成，并且正常运行，为后续的案例开发实验做基础。

### 2.3.2 数据接入

独立把数据上传至 HDFS 分布式文件系统，并且存储在 Hive 数据

仓库中，提供给 Zeppelin 查询展现

### 2.3.3 数据展现

使用 Zeppelin 通过 JDBC 连接 Hive，写一下 Sql 把 Hive 中的数据提取出来做报表展现

### 2.3.4 数据分析

使用 Dataiku 直接读取 HDFS 文件系统中的数据，抽取数据做数据的分析统计，统计搜索词频绘制词云图。

## 三、运行环境

### 3.1 软件环境

服务器操作系统：RedHat 或者 CentOS（英文版）

服务器操作系统版本：RedHat7 或者 CentOS7

JDK 版本：Oracle1.8

CRH 版本：CRH5.1

分析工具：Dataiku

### 3.2 硬件环境

推荐测试环境：

内存：8G

存储：100G

CPU：双核处理器

推荐生产环境：

内存：128G 或者 256G

存储：服务器满配每块盘 3T 或者 4T

CPU：48 核

### 3.2 网络环境

每台服务器或者操作系统之间能相互连通，有时间同步服务器，终端能连接上即可。

## 四、实现过程

### 4.1 实现思路

- 搭建大数据处理集群
- 安装数据分析工具
- 准备数据
- 数据接入存储为原始数据
- 数据清洗为结果数据
- 结果数据提供数据分析使用
- 对结果数据进行分析统计
- 对统计结果进行展现或者生产报表

## 4.2 实现技术

### 4.2.1 HDFS

大数据分布式文件存储，实现数据的存储保证数据的安全，HDFS 文件系统的容量可以横向的扩展。

### 4.2.2 HIVE

Hive 是基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的 sql 查询功能，可以将 sql 语句转换为 MapReduce 或 Spark 等任务进行运行。

### 4.2.3 ZEPPELIN

提供了 web 版的类似 ipython 的 notebook，用于做数据分析和可视化。背后可以接入不同的数据处理引擎，包括 spark, hive, tajo 等，原生支持 scala, java, shell, markdown 等

### 4.2.4 DATAIKU

企业级客户提供基于云技术的大数据服务分析平台，数据分析工程师可以很简单的完成数据的收集，分析，展现。

## 4.3 实现计划

### 4.3.1 CRH 环境搭建

使用 CRH5.1 搭建大数据基础平台，搭建过程详见 CRH 安装手册

### 4.3.2 数据录入

把本地测试数据 PUT 到 CRH 大数据集群的 HDFS 文件系统中。

把数据录入到 HIVE 数据仓库中，提供 Zeppelin 展现使用。

### 4.3.3 数据分析

使用 Dataiku 连接 CRH 大数据平台 HDFS 文件系统中，抽取数据，对互联网搜索日志数据做统计分析，计算词频。

### 4.3.4 数据展现

对统计的关键词词频统计结果做报表展示。使用 Zeppelin 连接 HIVE 做报表展示。