# 李宜润

18103966973 376942103@qq.com

## 基本信息

籍　　贯：河南省 驻马店市　　政治面貌：中共党员
英语水平：CET-6　　学历：工学硕士
目标岗位：大数据研发工程师 、spark 工程师、Hadoop 工程师

## 教育背景

| | | | |
|---|---|---|---|
| 2014.09-2017.07 | 中国石油大学（北京） | 地质资源与地质工程 | 工学硕士 |
| 2010.09-2014.07 | 长江大学 | 地质学 | 理学学士 |

## 科研成果

以第一作者发表中文核心（科学技术与工程）论文 1 篇，参与发表中文核心 1 篇（四作）。
以第一作者发表会议（第四届非常规油气地质评价学术研讨会）论文 1 篇。

## 工作经历

| | | |
|---|---|---|
| 2017.07-2017.12 | 华图教育 | 公共基础知识专职教师 |
| 2017.12-至今 | 培训 | 大数据 |

## 工作技能

理解 hdfs 分布式文件系统存储结构和高可用原理；
熟悉 Zookeeper 分布式服务框架，理解 HA 高可用集群；
掌握 hadoop mapreduce 计算框架编程，对 yarn 的资源调度，作业监控有一定认识；
熟悉 hive 数据仓库工具及 HQL 的书写，能对日志数据进行查询，统计等数据操作；
熟悉 linux 系统，了解常用的 linux 的 shell 命令，能在 linux 系统下搭建开发环境；
理解面向对象设计思想，熟练使用 Java 编程语言；
熟悉 kafka、flume 数据采集工具的使用，实现流式数据的过滤和分析；
能阅读英文技术文档。具备良好的文档写作能力；
理解 Hbase 的存储原理，Hbase 存储架构，实现数据的毫秒检索；
了解 Spark 相关组件，了解 Storm 运行流程；
熟悉 Python、Scala 编程语言，能运用 Scala 进行 spark RDD，spark streaming 编程。

## 项目经历

项目名称：Hive 项目-某视频网站运营指标分析
开发环境：eclipse+maven+jdk+linux
系统架构：hadoop+zookeeper+hive
需求描述：统计某视频网站的常规指标，各种 TopN 指标：
统计视频观看数 Top10；
统计视频类别热度 Top10；
统计视频观看数 Top20 所属类别包含这 Top20 视频的个数；
统计视频观看数 Top50 所关联视频的所属类别 Rank；
统计每个类别中的视频热度 Top10；
统计每个类别中视频流量 Top10；

*————态度决定人生，细节决定成败！*

统计上传视频最多的用户 Top10 以及他们上传的视频；
统计每个类别视频观看数 Top10。

项目描述：项目源数据是两个文件，一个是视频表，字段有视频的 ID 标识、视频上传者、视频的类别、视频的观看数、视频流量和视频相关视频的 ID 等。另一个表为用户表，字段有上传者的用户名，上传的视频数等。先使用 MapReduce 对视频表中的数据进行清洗，剔除不合要求的数据。再根据不同的需求，通过 Hive，使用 Hive sql 统计出各种 TopN 数据。

项目步骤：1、通过 MapReduce 对原始数据进行清洗，生成规范数据文件上传到 HDFS；
2、然后使用 Hive 对数据进行多维分析；
3、再把 Hive 分析结果使用 Sqoop 导出到 Mysql 中。

项目名称：HBase 项目-微博系统

开发环境：IDEA+maven+JDK+linux

软件架构：hadoop+zookeeper+ hbase

需求描述：用户发布微博内容；
用户社交体现：关注用户，取关用户；
拉取关注的人的微博内容。

项目描述：微博系统包括三张表，一张是微博内容表（RowKey：用户 ID_时间；Family：info；column：content；value：微博内容 String），一张是用户关系表（RowKey：用户 ID；Family：attends,fans；column：用户 ID；value：用户 ID；），一张是收件箱表（RowKey：用户 ID；Family：info；column：用户 ID；value：微博内容表的 RowKey)。当用户发布微博内容时，我们在微博内容表中添加相应的行。当有用户添加关注用户时，我们在该用户的用户关系表列簇（attends）中添加相应列，在被关注用户的用户关系表列簇（fans）中添加相应列，在收件箱表中添加相应列。收件箱表存放着每个用户及其关注用户的微博内容的 RowKey，收件箱表对所关注用户多个微博内容采用的是版本号的方法。当用户的关注用户发表微博内容时，在此用户的收件箱表中添加相应的版本号。

项目步骤：1、创建命名空间以及表名的定义；
2、创建微博内容表、用户关系表、用户微博内容接收邮件箱表；
3、发布微博内容；
4、添加关注用户、移除（取关）用户；
5、获取关注的人的微博内容；
6、测试。

项目名称：Spark Streaming 实时流处理日志项目

开发环境：IDEA+maven+JDK+linux

软件架构：hadoop+zookeeper+flume+ kafka+ spark+hbase

需求描述：实时（到现在为止）的日志访问统计操作

项目描述：项目数据来源的日志为 Python 脚本产生的，通过 crontab 定时执行 Python 脚本模仿服务器日志的产生，日志包括 ip、time、url、status、referer。然后使用 flume 采集产生的日志数据并 sink 到 Kafka 消息队列中，然后将日志信息传给 Spark Streaming 进行实时数据处理。最后将计算结果写入到 Hbase 上。

项目步骤：1、通过 Python 脚本模仿日志的产生；
2、Flume 的选型，在本例中设为 exec-memory-kafka；

3、打开 kafka 一个消费者，再启动 flume 读取日志生成器中的 log 文件，可看到 kafka 中成功读取到日志产生器的实时数据；

4、让 Kafka 接收到的数据传输到 Spark Streaming 当中，这样就可以在 Spark 对实时接收到的数据进行操作了；

5、Spark 中对实时数据的操作分为数据清洗过程、统计功能实现过程两个步骤。其中统计功能的实现基本上和 Spark SQL 中的操作一致，体现了 Spark 的代码复用性，即能通用于多个框架中；

6、计算结果写入到 Hbase。

项目名称：手机通信话单分析项目

开发环境：IDEA+maven+JDK+linux

软件架构：hadoop+zookeeper+flume+ kafka +hbase

需求描述：通信运营商每时每刻会产生大量的通信数据，需要定时定期的对已有数据进行离线的分析处理。例如，当日话单，月度话单，季度话单，年度话单，通话次数，通话总时长等等。项目需求就是要满足用户对通信话单的实时查询和展示。

项目描述：项目一共分为三块，第一部分为数据生产，在这部分要清楚项目数据的结构和内容，预判可能出现的问题并进行数据清洗，将数据写入到日志文件中；第二部分是将生产的数据落地到 HBase 中，首先是用 Flume 监控日志文件，采集实时产生的数据到 kafka 集群，再调用 Kafka 和 HBase 的 API，将数据写入到 HBase 中。第三部分是对 HBase 中采集到的数据进行分析，统计出我们想要的结果，将统计结果写入到 MySQL 中让用户查询。

项目步骤：1、数据生产：a）创建 Java 集合类存放模拟的电话号码和联系人；b）随机选取两个手机号码当做“主叫”与“被叫”，产出 call1 与 call2 字段数据；c）创建随机生成通话建立时间的方法，产出 date_time 字段数据；d）随机一个通话时长，单位：秒，产出 duration 字段数据；e）将产出的一条数据拼接封装到一个字符串中；f)、将通话数据写入到本地文件中；

2、数据消费：a) 编写 kafka 消费者，读取 kafka 集群中缓存的消息，并打印到控制台以观察是否成功；b)编写调用 HBaseAPI 相关方法，将从 Kafka 中读取出来的数据写入到 HBase；

3、数据分析：a) 按照时间维度来统计通话，比如统计 2017 年所有月份所有日子的通话记录，那这个维度我们大概可以表述为 2017 年*月*日。b) 通过 Mapper 将数据按照不同维度聚合给 Reducer。c) 通过 Reducer 拿到按照各个维度聚合过来的数据，进行汇总，输出。d) 根据业务需求，将 Reducer 的输出通过 Outputformat 把数据输出到 MySQL 中。

## 自我评价

- 乐于沟通，能快速融入团队，具备团队合作精神；
- 逻辑思维能力强，思路清楚，学习能力强，对新技术有着强烈的好奇心；
- 对工作尽职尽责，乐于从事有挑战性的工作；
- 具有良好的英语阅读能力，能阅读英文资料、技术文档等；

## 个人主页

个人主页：https://larry-arun.github.io/

在线简历：https://larry-arun.github.io/resume/

*————态度决定人生，细节决定成败！*