关键/JIM GUAN

- **求职意向**：大数据相关
- **手机**：15624901008
- **邮箱**：guanwwjian@163.com

关于我

- 共3年工作经验
- 1年大数据工程师工作经验（数据中台）
- 2年大数据算法工程师工作经验（个性化推送相关）
- 善于接受新技术
- 创新能力强

教育工作经历

2019.4至今	App Annie	大数据工程师
2017.4~2019.3	搜狗公司	大数据算法工程师
2016.6 ~ 2016.12	SAP	Java开发实习生
2014.9 ~ 2017.4	北京理工大学	计算机科学与技术 硕士
2010.9 ~ 2014.7	同济大学	计算机科学与技术 本科

职业技能

- 编程语言：
  - 掌握：**Python**、**Linux Shell**
  - 熟悉：**Java**、**SQL**
- IDE：**vim**
- 大数据处理：**Spark**（数据中台、个性化推送模型预测流程）、**Hadoop Streaming**（日志清洗）、**Hive**（数据统计）、**Celery**
- 机器学习：**TensorFlow**（Wide&Deep CTR预估模型）、**Spark ML**
- 测试：熟悉**Python Unit Test**的开发，了解**ETL的数据质量QA**
- 工作流程：**Scrum**
- AWS: **S3**、**Cognito**、**ElasticSearch Service**、**API Gateway**
- 数据库：**Redis**、**ElasticSearch**、**Citus**、**MySQL**、**SQLite**
- 容器：**Docker**
- 代码版本管理：熟悉基于**git**的多人协作工作模式
- Restful接口服务: **Sanic**（个性化推送服务）、**Flask**
- Restful接口服务部署相关：**nginx**、**PM2**、**expect**
- 数据统计及可视化：**Kylin**、**Redash**
- **大学英语六级**
- 了解国内推送技术现状及部分细节

工作经历

参与Data Pipeline ETL框架项目

- Data Pipeline是一个App Annie自研的ETL框架项目用户可以在该框架的基础上使用Pyspark完成ETL流程的开发

- Data Pipeline定义了多个数据层，统一了数据的存储规范，抽取了ETL的共通逻辑，使用户通过PySpark建立ETL流程变得更简单
- Data Pipeline在公司计算平台的基础上，实现了ETL流程对EC2集群的随启随用，用完即停
- Data Pipeline针对Citus、ES等数据库建立了loader，使得将数据通过分布式的方式导入数据库更加简便快捷

参与Data Pipeline Ingest API项目

- Data Pipeline Ingest API为提供了一个Restful API接口, 数据发送方通过调用改接口可以向Data Pipeline提交数据或者数据的路径
- Data Pipeline Ingest API完全由AWS托管的服务构成（API Gateway、Lambda、Kinesis Data Firehose），服务稳定性由AWS保障，节省了MT导致的人力消耗
- Data Pipeline Ingest API使用IAM Role来进行鉴权，从而保证仅授权用户可以访问该接口

参与Auto Pipeline SQL ETL框架项目

- Auto Pipeline是一个App Annie自研的ETL框架项目，ETL开发人员可以使用这个框架完成ETL的编写工作
- Auto Pipeline对Pyspark的Spark SQL做了封装，完善了与ETL相关的功能，使用户（ETL的开发人员）只需编写SQL就可以完成对ETL流程的开发，大大缩短了ETL开发时间
- Auto Pipeline是在Data Pipeline的基础上开发完成的

参与了Data Pipeline Monitoring System项目

- Data Pipeline Monitoring System提供了一个带有UI的平台使用户可以方便的监控到所有的DataPipeline任务的运行状态
- 使用AWS ELasticSearch Service中的ElasticSearch存储任务状态，Kibana作为UI供用户访问
- 提供了基于IAM Role的鉴权机制，Kibana UI使用了基于Cognito的认证机制

参与了多个ETL流程的开发及日常维护工作

- 参与了多个ETL流程的开发
- ETL Task不稳定时对Task进行处理
- 多次参与大规模的数据迁移

“今日十大热点”pingback回执日志通用统计流程

- 功能：根据配置，对“今日十大热点”pingback日志进行分钟级PV和天级UV统计，该流程支持按多个不同字段维度对日志进行筛选和分类
- 使用编程语言：**Python、Shell、SQL**
- 优势：
  - 配置简单：配置文件中每种统计维度组合只需要一行配置项
  - 配置项无需频繁修改：pingback日志中的已有字段新增字段值无需作出任何修改，新字段值的相关统计结果将自动列在统计表中
  - 维护简单：当日志中出现新的字段，只需对hive视图作出修改，无需建新的Hive表
- 架构：
  - 使用**Hadoop Streaming**对数据进行清洗
  - 使用**Hive**加载清洗好的数据
  - 根据配置文件生成SQL语句并提交至Hive
  - 将Hive的统计结果插入**Mysql**
  - 使用**Redash**对Mysql中的统计结果进行可视化

搜狗“今日十大热点”通知栏推送流程技术侧负责人

- 通知栏推送简介：通知栏推送是APP提升用户留存的一项重要机制，对于已经安装APP的用户，有一定概率可以在APP进程被杀死的情况下展示通知栏消息，一旦用户点击通知栏消息，APP就会被激活至前台，从而引导用户使用APP
- 用户规模 DAU：50万 推送目标用户总量500万

- 制定了客户端的推送Payload接口格式
- 制定了推送相关Pingback回执格式
- 根据pingback回执日志，进行Push相关的统计工作
- 根据pingback回执日志，收集并更新客户端Pushid集合
- 推送相关的客户端、前端、后端、运营的协调以及问题排查工作
- 参与推送通道调研及推送整体架构设计

“今日十大热点”个性化推送**CTR**预估模型

- 功能：计算用户-文章对的CTR预估评分
- 使用编程语言：**Python**、**Shell**
- 使用基于**Tensorflow**构建的**Wide&Deep**模型
- 使用**Spark**进行训练数据的分布式预处理，完成特征提取并将特征处理成**TFRecord**格式
- 使用GPU训练模型，训练过程使用了动态学习率
- 使用Spark加载Tensorflow模型实现分布式CTR评分预测流程，当Tensorflow模型结构发生变化时，无需修改预测流程代码

“今日十大热点”个性化推送服务

- 功能：收到推送指令后，为每一个用户选择目前CTR预估评分最高的文章，根据文章聚合用户，并把每一篇文章的推送用户集合提交至推送平台进行推送
- 使用编程语言：**Python3**
- 工作流程：为每个用户维护一个优先队列，接收CTR预估模型计算的文章评分，并将该文章评分插入到对应用户的优先队列中，评分最高的文章在队首。每当收到推送指令时，取出每个用户的队首文章，按文章聚合提交推送
- 优先队列使用**Redis**的**Sorted Set**结构实现
- Restful接口服务使用Python3下的**Sanic**框架
- 为了保证推送速度，使用**Celery**框架提供的生产者-消费者模式来分布式获取每个用户的队首元素，使发送速度达到了100万用户/分钟

“搜狗搜索”文字搜索个性化“猜你想搜”个性化后端流程维护

- 简介：负责维护已有的“猜你想搜”服务，猜你想搜就是对用户的搜索词进行个性化推荐，以提升用户使用体验
- 使用编程语言：**C++**、**Python**、**Shell**
- 线下流程根据用户搜索pingback日志使用**协同过滤**算法计算“搜索历史-推荐词”映照表
- 搜索历史-推荐词映照表以及用户搜索历史存储在**Redis**当中
- 线上Restful接口使用C++开发，负责接收pingback记录用户搜索历史，以及根据搜索历史查询“搜索历史-推荐词”映照表，并把对应的推荐词返回给用户客户端
- 使用**AC自动机**算法配合运营黑名单进行黑名单模糊匹配