张德宇简历

姓名:张德宇

学历:本科

毕业院校:郑州成功财经学院

证书: 网络工程师

工作经验:3 年

专业:计算机科学与技术

基 本 信 息

电话:17611275382

邮箱:17611275382@163.com

工作性质: 全职

目标地点: 北京

技术岗位: 数据开发工程师

目标薪资: 面议

专 业 技 能

1. 熟练使用 Scala 对 Spark 程序的进行编写, 熟悉 Spark 运行模式的部署流程,能够使用 SparkSQL 完成数据查询分析与处理, 及 SparkStreaming + Kafka 实现数据的流式处理, 能够独立完成常见的 Spark 性能优化, 熟练使用 Spark 操作 Hive 对数据操作
2. 熟练掌握 Kafka API 完成 Spark 编程, 了解 Kafka 架构, 能够排查独立处理 Kafka 常见问题, 与 Zookeeper 之间协同工作的原理
3. 熟悉 Hadoop 底层, MapReduce 的原理、HDFS 存储、Hadoop 调优、Yarn 的调度机制, 能够根据业务需求完成 MapReduce 程序的编写
4. 熟悉 Hive SQL 语句, 能够熟练使用 Hive 对离线业务指标的进行分析, 及独立完成常见 Hive 优化
5. 了解 HBase 底层数据存储原理, 连接 Phoenix 的使用, 掌握 HBase 常用 API 完成对 HBase 基本操作
6. 深刻理解 Flume 采集模型, 具有使用 Flume 搭建大数据日志采集系统的经验
7. 了解 Sqoop, 能够使用 Sqoop 实现数据在多框架之间的迁移, 具有迁移 MySQL 数据的经验
8. 理解中间件 Canal 的工作原理, 监控 MySQL 数据库的 binary-log 的增量和订阅并发送到 Kafka
9. 熟悉 Jenkins 任务调度工具，能够独立完成调度任务脚本的编写
10. 熟练使用 Linux 操作命令, 独立编写常用的 Shell 脚本, 及 Python 脚本的编写, 具有相关需求开发经验, 熟悉 Python 高级特性、函数式编程、面向对象高级编程
11. 了解实时查询分析引擎 Impala 架构及其原理，熟悉其优缺点，及适用的场景
12. 熟悉 GreenPlum + RoaringBitmap 对多维交叉运算的运用，及 Bitmap 在大数据的应用
13. 熟练掌握 MySQL 的日常操作，索引的建立和优化策略, 能够使用 Explain 进行 MySQL 性能调优
14. 熟悉敏捷开发 Scrum 的具体开发流程, 熟悉敏捷开发工具 Confluence, Rally ,JIRA 的使用

工 作 经 历

公司名称: 北京腾云天下技术有限公司

职位名称: 大数据开发工程师

工作时间: 2016.6–至今

工作描述:

1.数据日常需求例行化加工，主要负责 Bitmap 数据加工例行化，技术架构选型，技术调研以及平台中技术难点的解决，持续对平台进行维护

2.使用 Spark、Hive 对海量数据进行离线分析，数据加工，用数据进行统计改变企业决策

项 目 经 验

项目一: TalkingData 数据应用服务项目

开发环境: **IntelliJ IDEA + JDK1.8 + Scala2.11 + Linux**

软件架构: **hadoop3.0 + Spark2.3 + Kafka0.8 + Hbase + GreenPlum + Hive + Jenkins**

项目描述:

- [移动观象台]数据收集通过开发 APP 嵌入 TD sdk,或服务用户调用服务交换服务进行数据的交换，或通过爬虫获取公开数据。然后将数据通过前置节点得到原始日志。然后将其发送到 Kafka 集群后，按照小时落入 HDFS 上。然后对人群洞悉数据(区域省份，喜好)做成移动观象台，同时将应用排名的中的数据，在移动观象台中进行展示
- [Bitmap 计算]通过 RoaringBitmap 对国家、省、市、设备、机型、系统、网络、安装、活跃，及客户定制(一二三城市)对数据进行加工，并将其导入到 GreenPlum 中用于后续的例行化查询，也可以对后续定制查询提供数据源
- [设备信息库]设备位置通过 wifi 定位库，基站(cell)定位库，ip 定位库，运用 Geohash 定位设备位置。计算出设备 Applist,应用标签，人口属性标签，及位置标签消费标签、常驻城市标签合并标签库，对标签进行 Bitmap 加工。并将设备信息库数据放入到 Hbase 中

责任描述:

- 项目中主要负责 Bitmap 相关例行化加工，导入到 GreenPlume 后进行 Bitmap 查询
- 设备信息库信息通过例行化加工程序，完善设备信息，提供用户查询数据服务的功能

技术描述:

- [Bitmap 计算]运用 Spark 从 kafka 不同 topic 中拉取数据,对相关数据进行聚合处理之后进行例行化加工，并对数据进行 bitmap 计算加工，将其导入到 GreenPlum 中。伴随着数据量的增加，例行化要想完成 T+1 的查询，天聚合库加工比较延后，导致后续任务比较延后，使用 SparkStreaming 实时处理天库所需要的字段将其放入到 Hbase 中，后续任务从 Hbase 中查询，保证任务 T+1。最后针对 Spark 程序进行基准测试，对其进行性能调优处理，最后配置 Jenkins 完成例行化调度
- [设备信息库]运用 Spark 将 HDFS 上的数据按照不同的定位库进行处理聚合加工，抽取出客户需要的标签，并将设备信息库的数据放入到 Hbase 中，方便后续查询

项目二：TalkingData 数据仓库系统

开发环境：**IntelliJ IDEA + JDK1.8 + Scala2.11 + Linux**

所用技术：**Spark + Hive + MySQL + Sqoop + Impala + jenkins**

项目描述：

- [数仓建设]将常用数据集进行分层整理，针对数据表进行定义-设计，内部表与外部表的规划，将之前 parquet，textFile 数据集进行分层的迁移，最终统一用 Hive 管理起来
- [数据迁移]将原始日志中的数据导入到数据仓库中，进行数据清洗，完成数据抽取，数据标准化，数据聚合，根据需求生成设备信息库宽表，天聚合库宽表等
- [数仓架构]主要分为四层 ODS、DWD、DWS、APP，ODS 原始数据层主要存放原始日志数据，对数据的备份，DWD 明细数据层去除空值，脏数据，异常数据，生成标准化库 DWS 服务数据层做一些聚合，宽表，生成天聚合库，设备信息库。APP 应用数据层，生成应用大库(应用信息，安装列表，活跃列表，分类表)，设备大库中的黑名单(android、idfa、imei、mac),标签基础数据(位置标签，行为标签)，设备位置库(原始位置，ip 定位库，wifi 定位库，连接小时)等生成完备位置库。及游戏深度库(付费基础数据，活跃、安装、游戏时间等标签)，人口属性库(男女，年龄，教师，快递员，司机)，及后续 Bitmap 加工后的数据

责任描述：

- 参与架构选型，分层设计，相关业务库表设计及使用到的技术进行调研
- 定义相关数据集 schema 的设计，建表语句编写，及文档的实时更新
- 完成对相关业务库的数据迁移，保证迁移后的数据能够正常运行，及相关问题排查

技术描述：

- ODS 层将原始日志数据进行存储不做任何的改变，将数据按天分区，压缩使用的 LZO，Parquet 存储格式，减少存储的空间，原有的 Parquet 数据直接使用 Hive 进行管理
- DWD 层对原始日志数据进行 ETL,使用 Hive 中的 UDF 与 UDTF 进行数据处理
- 使用 SparkSQL 对数据进行清洗，建立宽表，将 MySql 中的数据使用 Sqoop 使用 Hive 进行管理
- 使用 Impala 对 Hive 表进行即席查询，并使用 Zepplin 插件来访问服务
- 保证 MySQL 中的数据通过 Sqoop 能够导入到 Kafka 中，并解决一些导入导出中 Null 值的问题，并保证数据的一致性问题

项目三: 实时采集分析工具

开发环境：IntelliJ IDEA + JDK1.8 + Scala2.11 + Linux

软件架构：Spark + Kafka + Canal + Redis + ElasticSearch + Flume + SpringBoot + Kibana

项目描述：

- 设备信息库中过滤出设备数> 50 的 ardroidid,idfa,imei,mac，生成实时滚动信息做成设备信息库黑名单，采集 MySQL 数据黑名单进行实时合并展示

责任描述：

- 能够实时展示出当前产品线中设备黑名单的变化情况，每天正常运行

技术描述：

- 用 Flume 自定义拦截器对数据预处理，使用 TailDirSource 实现断点续传和监控多文件及使用 Canal 来监控 MySQL 数据新增变化将其导入到 Kafka 中进行缓存对接流式处理
- Kafka Direct + SparkStreaming + MySQL 手动维护 Kafka Offset 实现的精准一次消费
- 利用 Canal + Redis + SparkStreaming 进行流的合并，实现了两个流的连接操作，组成一张宽表，并保存数据到 ES 中，用于 Kibana 前端页面显示
- 使用 Flume 导出 Kafka 中数据到 HDFS，使用 File Channel 来保证数据的不丢失，Ganglia 做数据的监控，对 HDFS 端的小文件，添加参数解决小文件的问题

自 我 评 价

热爱技术, 喜欢钻研新的技术写一些功能之类的 Demo

工作之余，喜欢在社区阅读各类技术文章，对技术有自己的理解

热爱团体活动，喜欢羽毛球活动