个人信息

姓名: **电话:**

邮箱:工作年限:9年 **地址:学历:本科**

毕业院系:吉林大学

工作经历

xx有限公司

数据架构师 2017/04-至今

xx**数据科技有限公司**

数据架构师 2016/05-2017/03

xx**有限公司（teradata）**ETL 工程师

xx集团 2013/05-2016/04

java 工程师

2011/01-2013/05

项目经验

实时计算 2019/02-至今

团队人数:10

公司业务上会有许多实时计算的要求,需求繁多.非专业人员开发起来费时费力.部署时候总是遇到各种问题.而且运维程序难度很大.

基于以上种种问题,规划基于 flink 做实时平台.

项目 1:运营商抓取信息实时计算

公司采集运营商信息后,需要为特征平台,决策引擎,报告等提供数据指标,项目中解决了采集数据繁多,计算逻辑复杂,中间状态的保存,多条流关联等问题

将 400+的指标计算缩短到秒级别.保证线上稳定运行.

主要技术点: 1,flink Broadcast connect 多个流操作.

2,hbase 存储维度值.

3,中间状态的保存.

项目 2:日志分析,短信预警分析

这类需求比较多,但比较通用.都是些计算窗口内的触发数统计及触发预警类的需求.通过技术调研,采用 sql 接口的方式对外暴露平台,供业务方使用.使开发周期大大缩短

主要技术点: 1,sylph 的二次开发.丰富 source.sink.

2,时间窗口的划定

3,antlr4 解析自定义 sql 语法

项目 3:模型实时落地

贷后,催收等团队拥有大量的语音数据,市场的合规迫使语音质检场景的形成.团队主要负责于协助算法模型人员实时落地模型.缩短质检时长.

主要技术点: 1,语音数据如何形成实时流.目前使用 fastfs 存储切分后的音频,将音频的元数据信息存入 kafka

2,基于 docker 部署 tensorflow.实时消费 kafka.

遇到的问题:

1,第三方存储的介入,致使实时计算的延迟较长.将语音的切分点标注和切分的过程合并,切分后的音频发送到 kafka 中.(或者 hbase)目的是减少网络 io 磁盘 io.效率提升 30%+

2,基于 cpu 模型计算耗时过长,cpu 使用率过高.增加语音文件 key,将切分文件通过 kafka 打散,实现分布式.

数据中台搭建

2018/06-2019/01

团队人数:8

参与从0到1的数据平台建设,主导基础服务的落地,包括不限于元数据,数据质量,账户体系,计算引擎,仓库建设.并推动,指导业务线使用数据平台的服务.保证服务的正常.

项目1:元数据系统

基于 hive 仓库,提供自动化的表查询,血缘分析,实体分析,数据地图等功能.方便业务人员使用集团数据资产,挖掘数据的更高价值.截止目前使用人数400+.

主要技术点:

1,编写业务逻辑处理 hive 元数据变化

2,使用 es 存储检索索引.支持多种检索模式

3,neo4j 存储实体间关联关系

项目2:数据质量系统

和调度系统联动,及早的发现有问题的数据,解决资源浪费问题,帮助集团节省硬件的开销.

主要技术点:

1,spark 处理 mr 历史日志.分析出计算过程中的倾斜

2,分析 fsimage.分析表的生命周期,存储数据量.通过规则识别存储浪费情况

3,编写 etl 建立整体数据分析主题

4,数据压缩归档,

项目3:账户体系

数据平台的基础服务,提供认证和授权等重要功能.保证集团数据的安全.并提供审计等功能.

主要技术点:

1,设计整套权限认证机制.提出虚拟工作组,租户等概念,支持整体数据中心产品.

2,建立统一登录机制

3,采集 sql.解析 sql.

4,基于 ldap,sentry 实现大数据组件的权限打通.

基于特征平台 SNA 业务线上部署

2017/12-2018/05

风控引擎中有许多基于 neo4j 的规则，由于数据的变动，准确性不是很高。想通过模型将多种数据源拟合成一个指标，替换掉老的规则。

团队人数:3

难点：

将离线数据处理流程，离线训练的模型，部署到生产环境中，并实时打分提供指标

工作职责：

1，提供离线数据（个人信息，运营商信息，淘宝等信息），用于 ds 训练 dnn 模型

2，使用水滴将数据预处理，并形成 pipeline。

3，与业务部门沟通，梳理业务流程，明确线上数据的来源与数据频率。

4，使用 spark stream，hbase，mongo，kafka 等技术，模型部署上线

遇到的问题：

个别数据源的消息体过大，数据源的及时性及历史数据处理问题等

特征平台、模型平台开发（水滴） 2017/08-2018/03

项目介绍:

实现一个端到端的集特征处理，模型预测于一体的平台，方便 ma ，ds 探索新的数据源，快速验证业务模式。

团队人数：5

工作职责:

1,使用 spark 进行数据的预处理，流程化。主要有时间转换，类型转换，采样，连续变量离散化等功能

2,使用 docker 部署 GPU，提供 tensorflow 开发环境。

3,使平台产品化，提供功能界面，推动公司人员使用。

4,引入 h2o，部署 h2o 集群，使训练模型方便快捷，（修复 h2o 中文乱码，int96等 bug）

5,使用 spark ml，h2o 训练随机森林，逻辑回归等模型，验证特征的有效性

知识图谱 2017/04-2017/07

项目介绍:

通过知识图谱提供贷后失联修复功能，并在图结构上结合算法解决金融领域中风控问题，如反欺诈规则，组团欺诈的识别，信息不一致，关联关系风险识别等。

工作职责:

1,使用 spark stream 将通讯录，通话记录，淘宝等渠道的数据采集，清洗，加工成三元组存储在 neo4j 中

2,使用 spark graphx ,graphframes 进行标签传播，社区发现等图算法模型的验证

3,数据分层，在 hive 中将采集的数据，清洗，拍平形成中间表，方便针对不同场景快速建图，验证业务场景。

用户行为分析/用户画像 2016/05-2017/03

项目介绍:用户行为分析系统主要服务于银行行业,现已在中信,光大,华夏等银行开发部署,帮助企业发现问题所在. 改善其营销的策略.

本人职责:

1,与各银行技术部门进行架构评审，根据不同环境给出合理部署架构。

2,指导并参与平台的建设，积极与各部门沟通即时解决部署中的问题。

3,优化平台参数 yarn,hive 的配置项等。

4,数据模型的改造,老系统中不支持即时查询,需要根据客户需求做到即时查询的功能,将 hive 底层数据改造成宽表,星型模型等

山西移动 OPENAPI 2015/05-2016/04

项目介绍：使用 hive 构建一个数据集市.将移动数据汇总，提供接口供上层平台取数，做到数据的统一性，便于审计。

1，使用 tdch 将 teradata ， aster 的数据导入到 hadoop 上。
2，使用 spark 在 hive 上进行进行数据的清洗，关联。
3，将结果数据存在在 hbase 上，供前台日报的展示。

北京电信 CCR 2014/05-2015/04

项目介绍：基于用户的投诉工单，按照集团的规范划分种类，预测10000客服手动划分的正确性。及时的反馈新业务（4G）的客户意见。
1，北京电信每月都会有几万条客户投诉的工单，人工将上几个月的工单按照业务进行分类（python），输出模型，安装预定的模型进行下一个月的分类。
2，分类结果的展示即 CCR 界面。反应出最近几个月投诉的重点，以及具体的一个问题的跟踪.
3，将预测模型迁移到 spark 上。 使用 spark 做工单的拆分,使用 kmeans 聚类.
4，编写脚本用作数据提取,处理缓慢变化维度,配置到 automation(类似 kettle)中.

个人能力

专业能力
1,拥有5+年的大数据处理,数仓建设及实时计算相关经验.
2,主语言是 java,但能熟练使用 scala python 做相关数据分析,及工程实现.
3,拥有数据建模,业务建模能力.常见的算法模型能够灵活应用.
4,自驱动能力强,对于新技术有高涨的情绪研究,并能结合业务需求.加以利用.
组织能力
1,优秀的资源统筹,跨部门协作能力.
2,领导10+人的团队,完成集团数据中台核心功能.自学能力

1,博客地址:
2,github 地址: