友信大脑系统概要

1 介绍

友信大脑，目标是整合友信所有系统的异构数据，提供数据计算与智能化分析功能，并且对外提供统一数据服务。

异构数据包括：数据库数据、日志数据、三方数据、外部爬取数据、报表、备注信息、图片附件、语音文件等。

在友信大脑中，一部分数据会以实体的方式保存在知识图谱图中、一部分数据会以时序数据保存在知识库中、一部分数据会经过特征计算保存在特征库中，数据通过不同方式发挥不同功能。

本平台整合已有资源（大数据框架、各种功能脚本等），对某些资源优化加固，某些资源微服务化，同时加入知识图谱、特征库等新资源。整合数据，对外提供统一数据服务。

2 功能模块

友信大脑包含如下功能：

- 知识图谱：以图的形式保存数据，提供反欺诈等应用；
- 数据整合：整合所有系统所有异构数据；
- 数据服务：对外提供统一数据服务；
- 特征计算：针对原始数据进行特征计算与整合，提供用户画像功能；
- 机器学习：在大数据环境下提供统一机器学习平台；
- 智能爬虫：爬取外部数据；
- 智能问答 ：对所有人提供统一数据搜索入口（带权限控制）；
- 基础服务：包括邮件服务、短信服务、调度服务、文件服务、日志服务、监控服务、安全服务、消息服务等；

3 已知需求

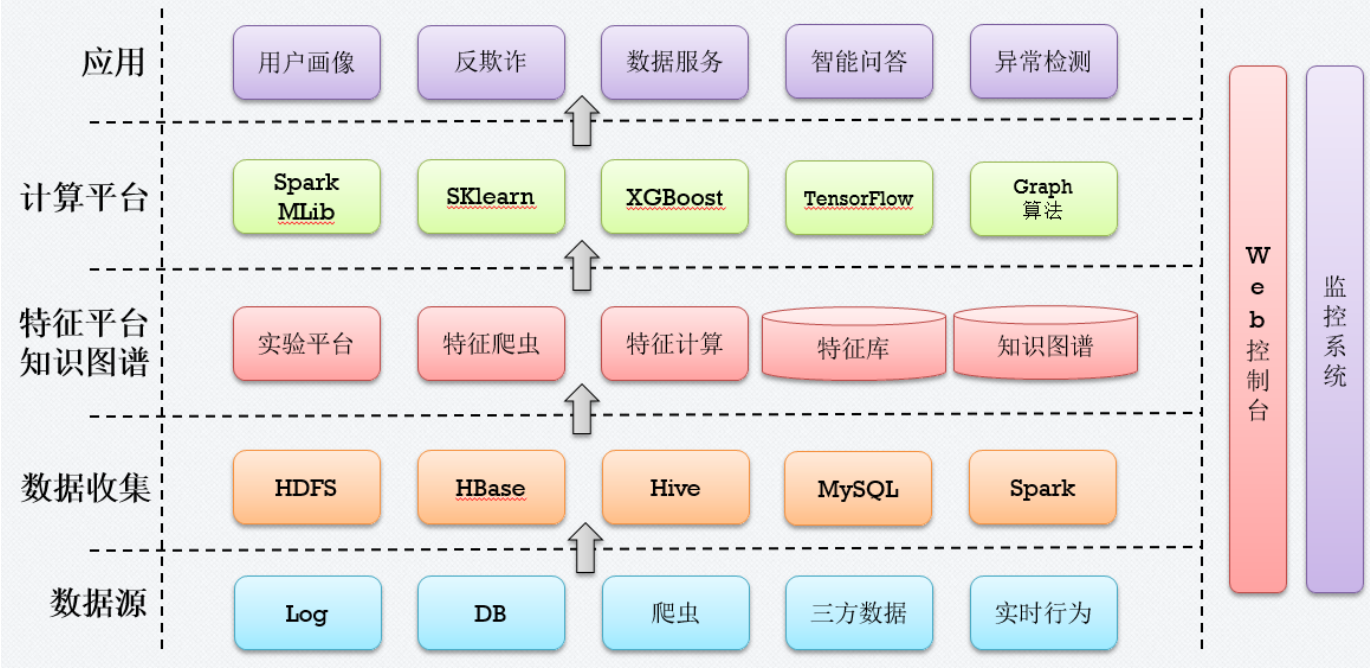
基于当前的需求，友信大脑在2019可能的落地需求如下：

- 基于知识图谱的黑名单服务（信审）；
- Yotta三方历史数据迁移与分析（政策、技术）；
- 客户App用户行为数据存储与分析服务（政策、技术）；
- 知识图谱数据服务，如反欺诈（多业务方）；
- 统一数据服务（多业务方）；
- 多系统元数据统一管理（技术）；
- 所有系统日志管理服务（技术）；
- 大数据量（亿级）数据查询服务（技术）；
- 系统异常检测功能（合规、安全、技术）；
- 特征计算与用户画像功能（政策）；
- App UI用户反馈实验平台（企划、产品）；
- 大数据机器学习与深度学习平台（政策）；
- 智能爬虫服务（政策）；
- 智能报表，智能问答服务（所有人）；

4 架构设计

4.1、整体架构

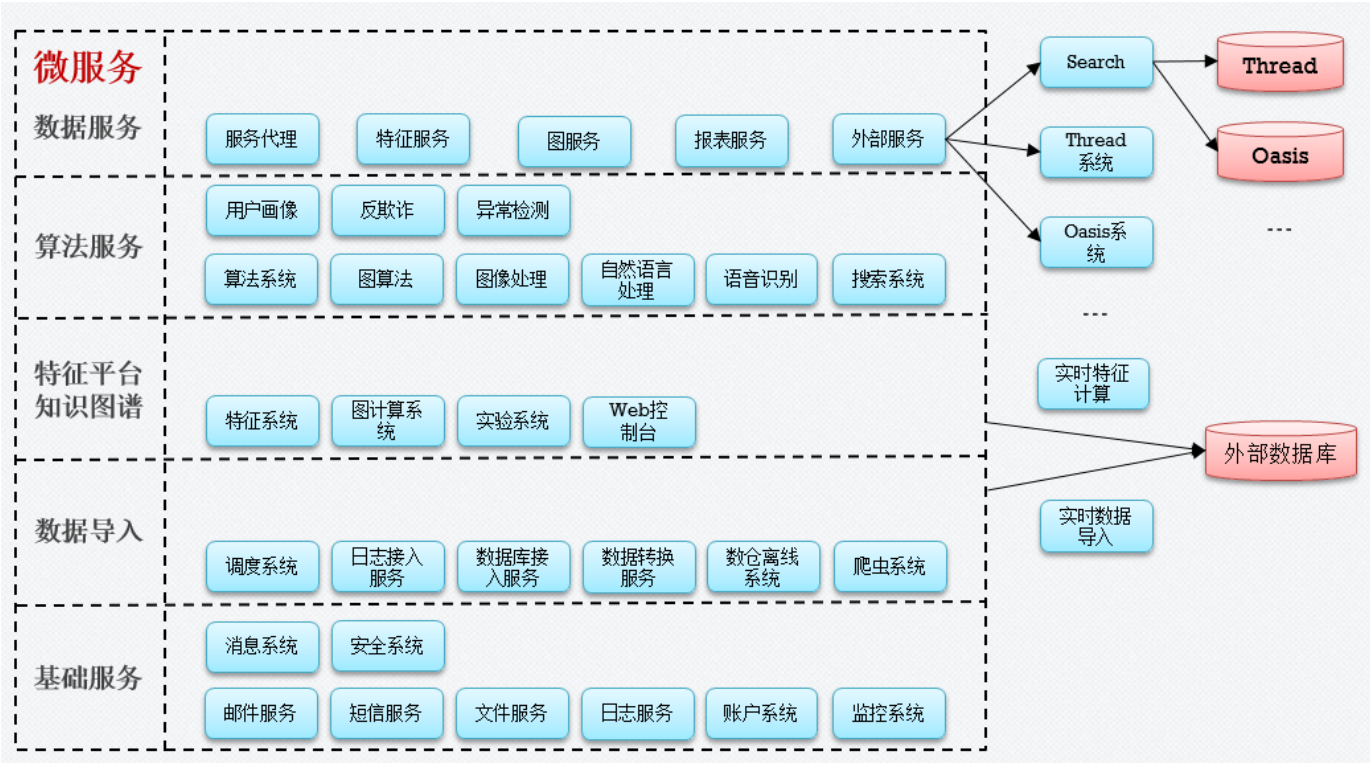
友信大脑从不同数据源收集数据，利用大数据框架存储数据，把部分数据存入知识图谱以及特征库，对外提供统一数据服务，或者通过计算平台提供反欺诈、用户画像、异常检测等智能化服务。



4.2、服务架构

友信大脑开发过程中会按照微服务的方式来开发，每个服务统一以Spring Boot + Spring Cloud为开发框架，统一前端样式。

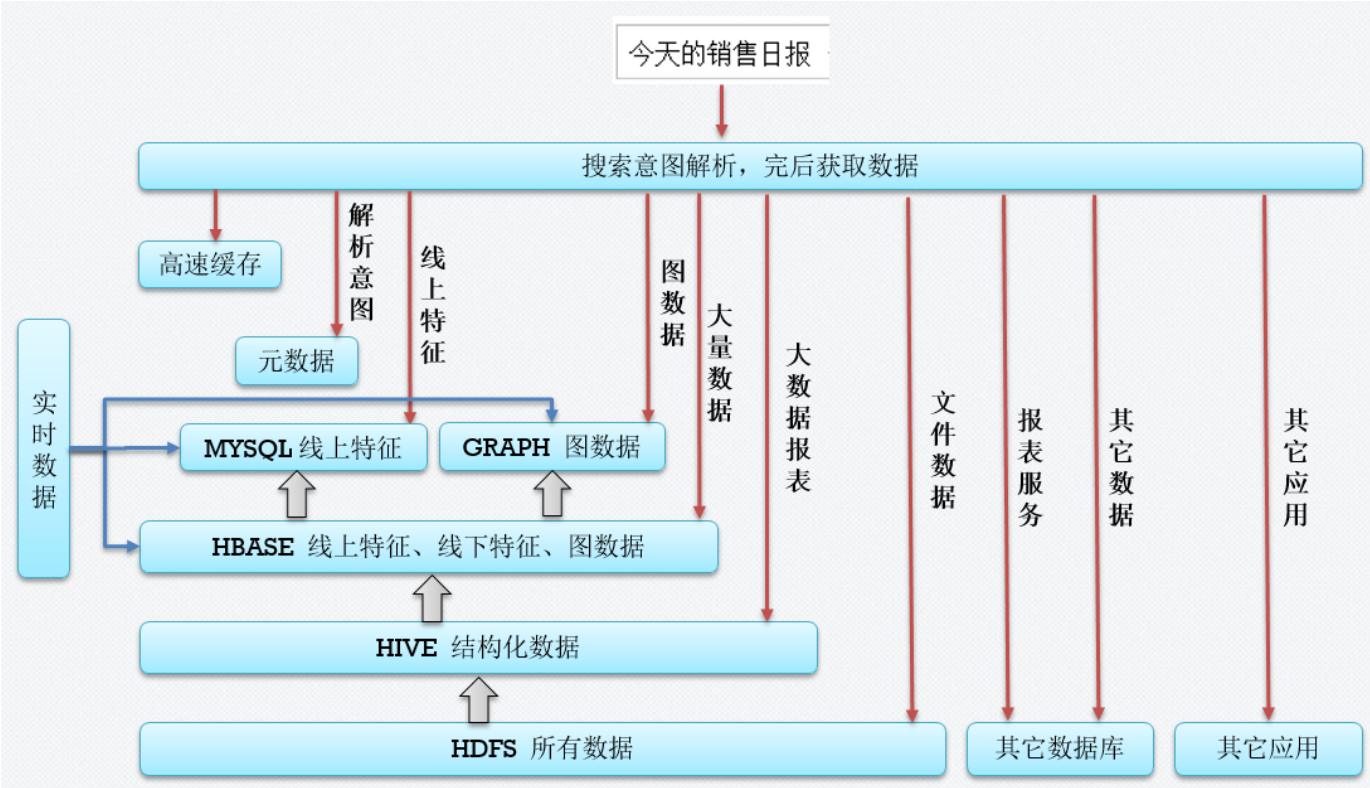
一些基础服务及数据服务开发完成后，可以作为通用服务提供出来。



4.3、数据服务

友信大脑作为公司内部的一个数据搜索平台，2019年的最终愿景之一是给公司所有人员提供数据查询服务（带权限认证）。

用户在查询一个数据时，不再关心数据来源于哪个系统，友信大脑会自动从特征数据库、知识图谱、大数据存储、其它系统数据库获取结果，甚至直接调用报表服务或者其它系统服务。

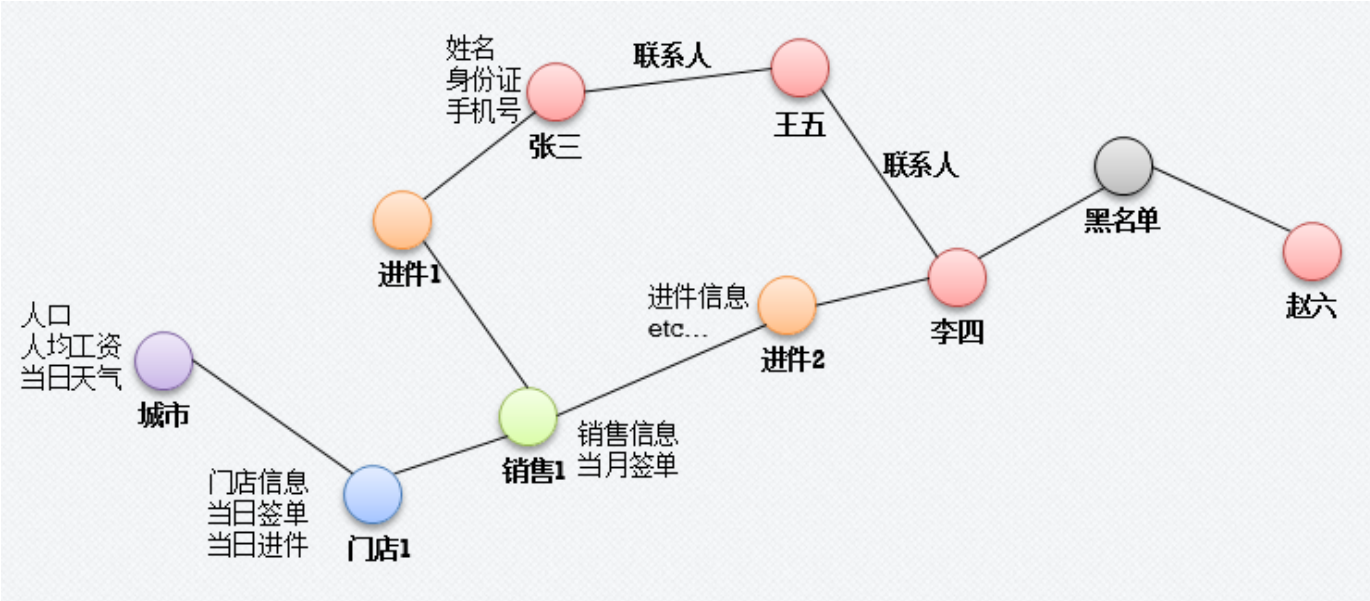


5 功能说明

5.1、知识图谱

友信大脑的核心是知识图谱，把各个系统的数据以实体的方式联系起来，包括自然人、进件、门店、销售、信审人员、城市、公司，以及黑名单、失信记录等。

基于知识图谱可以提供各种图运算，实现反欺诈、黑名单等多种应用。



5.2、数据平台

经过多年的发展，我们已经建立并使用了基于Hadoop和Spark的大数据平台。

友信大脑对于数据量的要求跟现在的量比会急剧上升，对于整个平台的稳定性及性能要求都会更高。

友信大脑使用到的相关框架如下：

- HDFS：保存所有原始数据，以及图片、日志文件、语音等非结构化数据；
- HIVE：保存离线数据与报表所用到的各个系统所有的结构化数据；
- HBASE：保存所有线上特征数据、线下实验用的特征数据、知识图谱的原始数据、以及某些有查询需求的大批量数据（亿级）；
- GRAPH：知识图谱所用的图数据库，把数据以实体的方式保存下来；
- MYSQL：保存线上特征数据及平台元数据；
- ES：保存需要创建索引的搜索类数据；
- KAFKA：实时消息平台；
- CANAL：MYSQL实时BINLOG同步框架；
- SPARK：实时处理框架；

对于基础数据存储框架，所有数据都保存到HDFS中，HIVE和HBASE的数据都可以从HDFS恢复，MYSQL和GRAPH的数据都可以从HBASE恢复。

5.3、数据服务

收集到多个系统的数据，整理计算出其它特征数据，以及把数据存储到知识图谱中后，友信大脑将可以对外提供统一数据服务与管理，包括：

- 提供线上特征数据；
- 提供基于图的知识图谱数据；
- 提供亿级别的大数据量的查询功能；
- 通过SQL调用其它系统数据库提供数据；
- 调用报表系统提供报表数据；
- 调用其它系统服务提供数据；
- 提供文件类数据；

友信大脑会有一套完整的账户系统与安全系统，可以把每个员工的权限管理到每条数据，并且保证系统的整体安全。

5.4、特征平台

基于原始数据进行特征计算，是数据分析必不可少的步骤，友信大脑对外提供统一的特征计算平台，包括功能如下：

- 对原始数据进行数据处理（异常值过滤、数据异构转同构）、数据加工（统计、平滑、归一等）等操作，实现界面可视化配置；
- 提供线下建模实验平台，包含批量特征计算及机器学习平台，以满足大数据量下的数据建模工作；
- 提供Web管理界面，对特征的上下线进行统一化管理，并且可视化展示特征的各种统计指标（最小值、最大值、均值、方差等）；
- 分析人员可以在友信大脑上计算出相关特征，完后导出到自己的MYSQL数据库中进行后续处理；
- 特征平台通过私钥及其它安全配置，可以保证特征计算逻辑、模型配置及超参数等信息的绝对保密；
- 线上特征与线下特征分开存储，保证线下特征实验不会影响到线上特征库的稳定性与性能；

5.5、计算平台

随着友信数据量越来越大，很多数据如Yotta三方数据，以及即将接入的客户App用户行为数据，要在单机上分析这些数据已经越来越不现实。

友信大脑会提供一套大数据计算平台，供相关数据分析人员使用，预期的平台如下：

- Spark Mlib（机器学习）
- Sklearn（机器学习）
- XGBoost（机器学习）
- TensorFlow（深度学习）
- Pytorch（深度学习）
- Spark Graph（图计算）

特征平台通过私钥及其它安全配置，可以保证特征计算逻辑、模型配置及超参数等信息的绝对保密。

分析人员可以在友信大脑上计算出相关特征，完后导出到自己的MYSQL数据库中进行后续处理。

5.6、智能爬虫

提供智能爬虫系统，爬取外部数据。

5.7、智能问答

提供基于搜索框的智能问答服务，不同用户（领导、销售、信审）可以直接在搜索框中输入所要数据。

系统智能解析其想要的数据，通过知识图谱、特征库、其它系统数据库、其它系统数据服务获取到结果并返回，权限不够的用户不会看到相应数据。

根据结果数据的类型自动切换展示方式，包括单条结果、多条表格、某个报表、图形化展示、甚至下载某个文件。

5.8、智能应用

基于友信大脑提供的知识图谱、特征计算、机器学习与深度学习平台，用户可以基于友信大脑开发多种智能类应用，以下列举了一些可能的需求：

- 反欺诈服务；
- 黑名单服务；
- 用户画像服务；
- 异常检测模型；
- 图像处理服务；
- 自然语言处理服务；
- 语音识别服务；
- 搜索服务；