数据分析平台二期需求分析说明书

项目编号：

涉及系统：（OBTM/CSR/DW/WF/SIEBLE/IVR）

需求接口人：

项目负责人：

密级：仅供内部访问

**Copyright © 2010中银金融商务有限公司 All Rights Reserved.**

**版本记录**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **版本号** | **版本日期** | **编写人** | **审核人** | **审批人** | **说明** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**目录**

[一、概述 1](#_Toc446406914)

[（一）编写目的 1](#_Toc446406915)

[（二）项目背景 1](#_Toc446406916)

[（三）基本定义 1](#_Toc446406917)

[二、 业务概述 5](#_Toc446406918)

[（一） 外呼业务概述 5](#_Toc446406919)

[（二）银行卡客服中心呼入数据分析需求概述 5](#_Toc446406920)

[（三） 综合客服中心呼入数据分析需求概述 6](#_Toc446406921)

[三、 功能描述 7](#_Toc446406922)

[（一）功能分类 7](#_Toc446406923)

[1.外呼营销中心数据分析新增功能 7](#_Toc446406924)

[2. 银行卡客服中心呼入数据分析优化功能 7](#_Toc446406925)

[3.综合客服中心呼入数据分析新增功能 7](#_Toc446406926)

[（二） 功能描述与需求分析 8](#_Toc446406927)

[1.响应率模型（含数据提取） 8](#_Toc446406928)

[2. Dashboard的可视化实现方案 16](#_Toc446406929)

[3. 综合客服中心的数据处理与数据分析 17](#_Toc446406930)

[3.建立导航体系 18](#_Toc446406931)

[4.用户/数据权限的实现方案 18](#_Toc446406932)

[5.操作行为关联分析 19](#_Toc446406933)

[6.特定业务场景可视化的实现方案 20](#_Toc446406934)

[7.投诉模型 21](#_Toc446406935)

[8.二次统计分析方案（即套娃功能） 22](#_Toc446406936)

[9.ElasticSearch搜索引擎优化 23](#_Toc446406937)

[10.提供专业数据分析工具的整合（H2O，Spark-R) 24](#_Toc446406938)

[11. IVR的可视化、算法、转移矩阵等的实现方案 24](#_Toc446406939)

[12. 重新清洗数据（含系统自动化运维、自动化部署） 26](#_Toc446406940)

[13. 特殊数据的标记及剔除统计 27](#_Toc446406941)

[四、 技术架构 29](#_Toc446406942)

[（一）数据源和架构图 29](#_Toc446406943)

[（二） 物理层 31](#_Toc446406944)

[（三） 数据层 33](#_Toc446406945)

[（四） 平台层 34](#_Toc446406946)

[（五） 应用层 35](#_Toc446406947)

[（六） 访问层 35](#_Toc446406948)

[（七） 运维体系 35](#_Toc446406949)

# 一、概述

## （一）编写目的

以我司数据分析平台一期建设成果为基础，提出外呼业务数据分析的总体需求方案及呼入业务二期优化需求，作为大数据分析平台外呼业务功能设计的依据及呼入业务功能优化依据，并且为下一阶段系统开发奠定基础。

## （二）项目背景

大数据营销对企业的业务流程理念、营销决策及消费者行为模式均产生巨大影响，伴随着消费者信息价值稀疏性不断扩大，客户知识维护成本日益增长。建立有效的数据分析、数据提取逻辑，实现精准化营销是外呼营销中心接下来的主要方向。

面对以上挑战，外呼业务管理部门亟需一个有效的名单数据分析及提取工具，用于辅助一线管理人员及坐席进行前端客户管理、机会管理和人员管理。

## （三）基本定义

|  |  |
| --- | --- |
| **专用名词** | **名词解释** |
| 名单利用率 | 下发名单量 /导入名单量 |
| 处理量 | 成功+拒绝+无法联系客户+回拨 |
| 有效名单量 | 成功+拒绝 |
| 触达量 | 成功+拒绝+回拨 |
| 成功量（笔数） | 成功分期笔数 |
| 成功量（户数） | 成功分期客户数 |
| 处理客户数量 | 当日处理的全部数据中所有有客户标记的客户数量 |
| 触达量 | 当日处理的全部数据中[成功+客户拒绝+回拨联系] |
| 首拨首次成功客户数量 | 首次接触的新名单（累计标记次数为1）中成功客户数量 |
| 首拨首次触达量 | 首次接触的新名单（累计标记次数为1）中[成功+客户拒绝+回拨联系] |
| 首拨首次客户拒绝客户数 | 首次接触的新名单（累计标记次数为1）中客户拒绝客户数 |
| 首拨首次未联系到客户数 | 首次接触的新名单（累计标记次数为1）中未联系到客户数 |
| 首拨首次回拨联系客户数 | 首次接触的新名单（累计标记次数为1）中回拨联系客户数 |
| 首拨首次放弃客户数 | 首次接触的新名单（累计标记次数为1）中放弃客户数 |
| 总分期交易金额 | 日分期总金额 |
| 笔均金额 | 日分期总金额/成功量（笔数） |
| 户均金额 | 日分期总金额/成功量（户数） |
| 消费数据日人均导入名单量 | 消费数据导入名单量/FTE |
| 消费数据日人均下发名单量 | 消费数据下发名单量/FTE |
| 账单数据日人均导入名单量 | 账单数据导入名单量/FTE |
| 账单数据日人均下发名单量 | 账单数据下发名单量/FTE |
| 日人均处理量 | 处理量/FTE |
| 日人均触达量 | 触达量/FTE |
| 日人均成功量（笔数） | 成功量（笔数）/FTE |
| 日人均成功量（户数） | 成功量（户数）/FTE |
| 日人均金额 | 总分期交易金额/FTE |
| 消费数据导入名单量 | 卡中心导入OBTM的量 |
| 消费数据下发名单量 | 每日下发至组长的量总和 |
| 账单数据导入名单量 | 卡中心导入OBTM的量 |
| 账单数据下发名单量 | 每日下发至组长的量总和 |
| 呼出ATT(s) | 通话时长/通话量 |
| 接通率 | 通话量/外拨量 |
| 通话比率 | 通话时长/登陆时长 |
| ACW比率 | ACW时长/登陆时长 |
| AUX比率 | AUX时长/登陆时长 |
| 置闲比率 | 置闲时长/登陆时长 |
| 振铃比率 | 振铃时长/登陆时长 |
| 其他比率 | 100%-（通话比率+ACW比率+AUX比率+置闲比率+振铃比率） |
| 非生产比率 | ACW比率+AUX比率+置闲比率 |
| A-ACW(s) | ACW时长/通话量 |
| A-AUX时间 | （用餐+工间休息+培训+开会）时长/（用餐+工间休息+培训+开会）次数 |
| A-AUX时间-用餐 | 用餐时长/用餐次数 |
| A-AUX时间-工间休息 | 工间休息时长/工间休息次数 |
| A-AUX时间-培训 | 培训时长/培训次数 |
| A-AUX时间-开会 | 开会时长/开会次数 |
| A-振铃(s) | 振铃时长/外拨量 |
| 疑似违规IB时长 | 使用硬电话通话时长 |
| A-内部通话时间 | 内部通话时长/内部通话量 |
| 咨询通话量 | 三方通话次数 |
| 咨询通话时长(s) | 三方通话时长 |
| AUX时长 | 置忙时长 |
| ACW时长 | 话后处理时长 |
| HOLD时长 | 保持时长+验密时长 |
| 在线成功率 | 成功量（户数）/（有效名单量+无效名单量） |
| 有效成功率 | 成功量（户数）/有效名单量 |
| 净成功率 | 成功量（户数）/下发名单量（消费+账单） |
| 首拨首次成功率 | 首拨首日成功客户数量/首拨首日新名单数量 |
| 首拨成功率 | 成功量（户数）/首拨新名单量 |
| 触达成功率 | 成功量（户数）/触达量 |
| 处理成功率 | 成功量（户数）/处理量 |

# 二、 业务概述

## （一） 外呼业务概述

外呼业务是呼叫中心的“利润中心”，需要充分挖掘客户价值，主动出击为企业创造利润。而面对庞大的客户资料数据库，现有的信息系统无法满足，仅依靠人工经验进行判断与分析，这种模式下工作效率低，成效低，严重阻碍了外呼业务的发展，而且随着人员流失，这种经验往往也无法很好的传承。

通过数据分析平台对卡中心名单数据的分析、筛选，从中找出成功率最高的客户群，一方面便于坐席找到目标客户，减少不必要的电话接通等待及沟通等待时间；另一方面，可便于业务管理人员针对名单数据的优劣情况制定针对性的营销策略和人力安排，建立数据库营销模式，最大化业绩产出。同时结合优秀员工的经验，不断完善数据分析平台的数据模型，将经验积累成企业的知识资产。

本次外呼需求主要针对外呼业务常用考核指标、外呼电话名单响应率预测模型、外呼考核中使用较为频繁的综合性展示分析，如员工综合表现展示与分析、变量相关性分析、数据下发量与坐席完成度的关系分析、预测反应率与实际反应率贴合度分析等。

## （二）银行卡客服中心呼入数据分析需求概述

本次呼入需求主要针对呼入电话行为的关联性分析进行优化、针对已完成计算或统计的结果集进行深度搜索的相关功能进行优化、组别信息的关联性分析、来电原因的综合性分析、重复来电多维度分析、IVR轨迹关联分析以及人工对比分析等。

## （三） 综合客服中心呼入数据分析需求概述

本次综合客服中心的数据涉及sieble数据、电子工单数据、客户满意度数据等，将以上数据通过excel导入的方式存入数据分析平台，建立数据间的多维度自定义检索与分析体系，如数据表明细查询、关键词检索、平均通话时长多维度统计分析、重复来电统计分析、月报统计报表等，并根据实际业务需求选择部分内容进行数据可视化的实现。

# 三、 功能描述

## （一）功能分类

本次需求可分为以下三个方面，第一是外呼营销中心数据分析功能的新增，第二是一期银行卡客服中心呼入数据分析部分功能的优化，第三是综合客服中心呼入数据分析功能的新增，详情如下：

### 1.外呼营销中心数据分析新增功能

1. 新增外呼常用考核指标的自定义统计及展现功能。
2. 新增外呼电话名单响应率预测模型的设计与开发。
3. 新增外呼电话综合性展示分析相关功能，如员工综合表现展示与分析、变量相关性分析、预测反应率与实际反应率贴合度分析等。

### 2. 银行卡客服中心呼入数据分析优化功能

1. 新增呼入电话批量检索功能批量处理后的深度搜索功能。
2. 优化呼入电话行为的关联性分析功能。
3. 优化组别信息的关联性分析功能。
4. 优化来电原因的综合性分析功能。
5. 优化重复来电多维度分析功能。
6. 优化IVR轨迹关联分析以及人工对比分析等功能。

### 3.综合客服中心呼入数据分析新增功能

1. 新增Sieble数据excel格式导入数据分析平台功能
2. 新增Sieble数据清洗、标签化等处理功能
3. 新增Sieble数据分布式存储功能
4. 新增Sieble数据表明细查询功能
5. 新增Sieble数据关键词检索功能
6. 新增Sieble数据检索结果二次查询功能
7. 新增Sieble数据中平均通话时长多维度统计分析功能
8. 新增Sieble数据中重复来电统计分析功能
9. 新增Sieble数据可视化图表功能
10. 新增Sieble数据中异常号码规则性剔除统计分析功能

## 功能描述与需求分析

### 1.响应率模型（含数据提取）

（1）功能简述

以客户响应率为中心指标，对有可能造成影响的变量进行大数据分析，找出与响应率关联性最强、共享度最高的变量值，并可对全体数据进行匹配标记与评分建模；通过评分建模的模式，运用统计学和概率论方法对不同评分阶段的名单进行响应率预测。

功能说明——变量需求

变量字段：客户群组标签、信用卡额度、信用卡账单欠款、还款情况、人行信用记录、OBTM交易记录等

获取路径：卡中心、TGW

需求目的：以客户响应率为关键指标，对变量字段内容进行针对性数据分析，将分析后的变量字段进行打分排序，找出与响应率指标贡献度最高的变量内容。

功能说明——评分建模及响应率预测需求

将已获取的原始变量内容通过组合分析、计算，生成所需数据源后进行响应率模式建立；

针对不同评分阶段的名单数据进行响应率预测。

功能说明——数据打标

将完成响应率预测的名单进行关键字段标注，方便进入OBTM系统后的数据筛选工作；

对于打标的字段内容仅作为OBTM后台数据筛选条件使用，不在坐席端进行显示；

名单数据需打标字段内容：预测响应率、本期账单可分期金额、本期消费最大可分期金额。

（2）功能细分

从用户交易记录中抽取新的参数，并计算新参数与响应率的相关性

训练决策树模型，展示参数与最终客户响应之间的逻辑关系

寻找对客户响应率有关系的参数

对参数进行卡方检验，按照卡方值的大小排序，列出和客户响应率最相关的参数

自动对新数据进行训练并生成新的模型

自动部署新的模型

（3）功能实现

首先需要和OBTM系统集成，OBTM发送每日需要拨出电话的用户至数据分析平台，数据分析平台预测拨出响应率后，将用户排序结果返回给OBTM。

集成信息抽取技术，对客户消费明细等记录进行深入分析，寻找并抽取与客户响应率有关的变量。例如：通过分析交易记录找出网购占比、家具家电消费占比、大宗交易占比、资金流入渠道、境外消费占比等变量。通过信息抽取技术的分析，将获得更多特征，对数据分析、建模和预测提供质量更高的数据。

运用统计学习方法对客户响应记录进行大数据分析，寻找影响客户响应率的随机变量，以及随机变量之间的关联分析。最后预估客户响应率。

利用机器学习技术设计出能够对客户进行评级打分的模型，对客户名单按照预测的分数进行排序。除了概率模型以外，再综合线性分类器、支持向量机、神经网络等多种算法对评级打分模型进行对比分析。

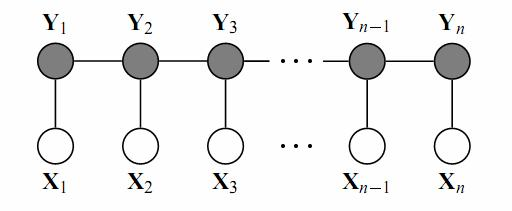
对成功交易的客户进行聚类分析，寻找成功交易的用户之间的共性。

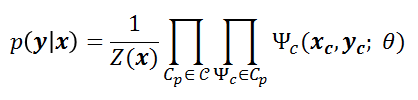
对成功交易的客户使用决策树等机器学习方法进行分析，寻找出成功交易的原因并将其可视化。

A.集成信息抽取技术

对客户消费明细等记录进行深入分析，寻找并抽取与客户响应率有关的变量。例如：通过分析交易记录找出网购占比、家具家电消费占比、大宗交易占比、资金流入渠道、境外消费占比等变量。通过信息抽取技术的分析，将获得更多特征，对数据分析、建模和预测提供质量更高的数据。

应用CRF算法对客户交易明细记录等短文本进行特征抽取和分析，例如标记出消费明细中出现的产品名称。CRF算法是一种马尔可夫随机场的概率模型，该算法在训练中将优化目标由全概率改为条件概率，该模型充分考虑到句子内上下文特征，同时又具备速度快，性能好等优点。被广泛应用来解决实体识别、词性标注等序列标记等问题。

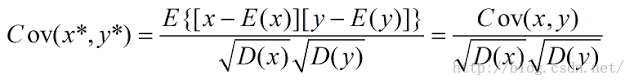




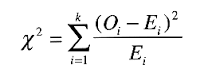
B.随机变量与客户响应率的统计分析

运用统计学习方法对客户响应记录进行大数据分析，寻找影响客户响应率的随机变量，以及随机变量之间的关联分析。最后预估客户响应率。

应用概率论中经典的相关性分析方法，对已有数据的输入变量（随机变量）计算协方差，进行相关性分析。

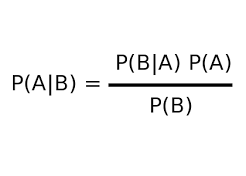


并同时根据客户是否响应的记录，对输入变量进行卡方统计，筛选特征与最终客户响应率之间的关系。



C.贝叶斯原理

预测客户响应率。

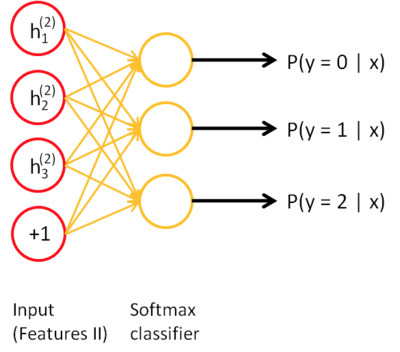


其中A代表客户最终响应与否，B代表客户信息

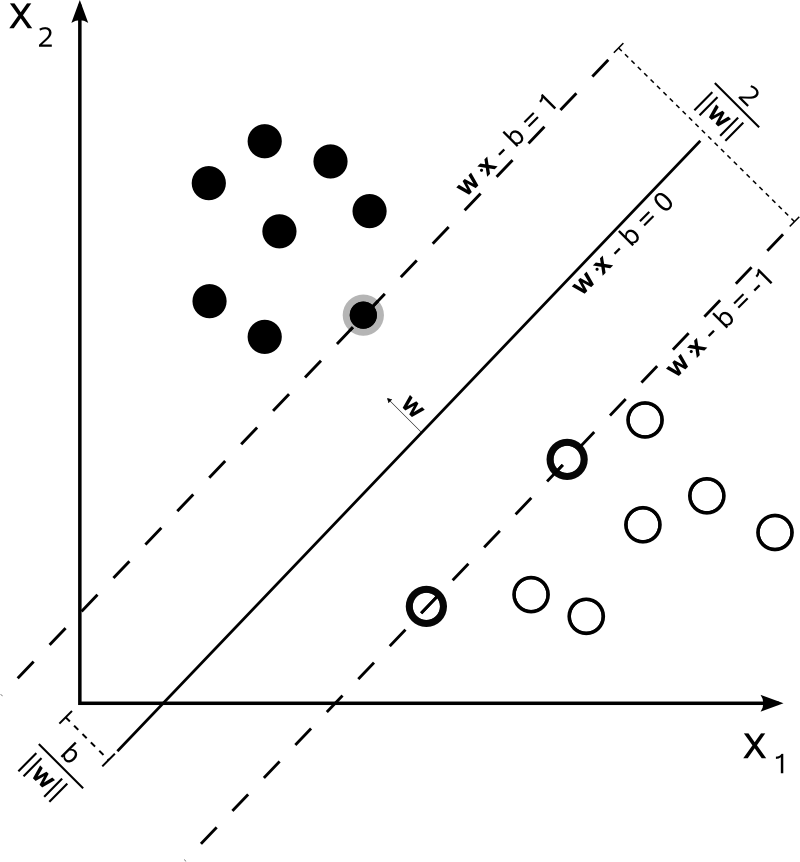
D.机器学习与评级打分

利用机器学习技术设计出能够对客户进行评级打分的模型，对客户名单按照预测的分数进行排序。除了概率模型以外，再综合线性分类器、支持向量机、神经网络等多种算法对评级打分模型进行对比分析。

除了应用贝叶斯原理来预测客户的响应率，还可以尝试使用神经网络来对客户响应率的概率模型建模。Softmax是一种使用神经网络来表达的概率模型的方法，如下所示，神经网络也可以应用来预测客户多大可能会完成交易。



支持向量机也是一种非常流行的概率机器学习算法，可以使用该方法训练机器对客户进行分类，对客户打分，判定客户多大可能会完成交易。支持向量机的原理是在数据集中寻找一个超平面，用来讲数据分为两类，新的数据的预测其实就是该数据距离此超平面的距离。



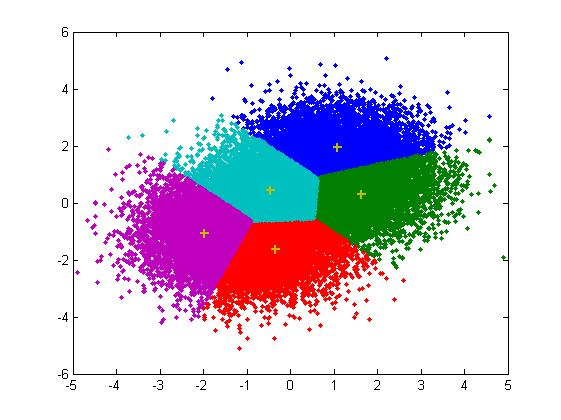
E.聚类分析寻找客户的共性

对成功交易的客户进行聚类分析，寻找成功交易的用户之间的共性。

完成交易客户之间有很多共性，可以对完成交易的数据进行聚类分析，然后探索这些客户在不同维度的分布，分析原因，为今后更好的优化流程提供科学的分析结果。聚类算法可以基于KMean和EM聚类来实现。

在同一类的用户中，在对用户的每个维度进行探索，发现隐藏的模式。

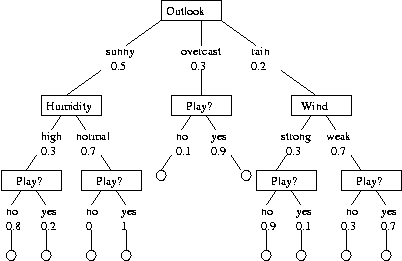
ThoughtWorks HD:Users:twer:Desktop:unnamed (1).png



F.决策树与数据可视化

对成功交易的客户使用决策树等机器学习方法进行分析，寻找出成功交易的原因并将其可视化。

虽然决策树的机器学习方法并不准确，但是决策树是机器学习模型中最容易被人类理解的模型。所以决策树虽然不能作为有效的分类算法，但往往成为可视化工具，并为业务决策提供强有力的数据支持。如下图所示：决策树通过统计历史数据，自动计算出特定的决策规则。



### 2. Dashboard的可视化实现方案

1. 功能简述

Dashboard可关联灵活配置报表、导航与用户/数据管理功能，是不同业务领域的用户方便查看与其相关的统计分析报表。

1. 功能细分
2. Dashboard展示页功能
3. Dashboard管理页功能
4. Dashboard与现有配置类统计分析页关联
5. Dashboard与导航关联
6. Dashboard与用户/数据管理关联
7. 功能实现

Dashboard需要在原有的自定义功能块的基础之上，将多个功能块的结果呈现在一个页面内。

其本质是在数据库中存储Dashboard和功能块的关联关系，以及显示的位置。在后台，为了能够快速展示，则需要提前设计定时任务，提前检索功能块的结果，在需要浏览的时候，最大限度的减小用户的等待时间。

### 3. 综合客服中心的数据处理与数据分析

1. 功能简述

将95566导出数据实现上传至数据分析平台，导入错误数据管理，数据清洗后自动存入数据分析平台，可自定义配置数据的统计分析。

1. 功能细分
2. 95566数据上传
3. 95566错误数据管理
4. 95566数据清洗
5. 95566用户/数据权限处理
6. 95566数据自定义统计分析
7. 95566工单数据、sieble数据、满意度等数据的关联
8. 功能实现

95566的数据也将被抽取至系统中，其中需要经过sqoop、hbase和pentaho的数据处理。同时原来的数据也将被搭上标签，标记出自己的数据来源。

前端的UI控件，在多处都需要增加对95566的支持。

### 3.建立导航体系

1. 功能简述

导航需根据不同用户显示不同的导航页，且管理员可配置导航与用户之间的关系

1. 功能细分
2. 导航管展现功能
3. 导航管理功能
4. 功能实现

管理员通过在后台配置，来配置基本用户的组别，以及每一个组别的显示模版，相同组别看到的页面一致，包括自定义的功能块以及Dashboard。

这些配置信息将存储在关系型数据库中，并和用户组别关联。

### 4.用户/数据权限的实现方案

1. 功能简述

针对敏感信息查看及数据下载进行权限管理，用户可自行配置各种权限组合。

1. 功能细分
2. 用户管理前端功能展现
3. 用户管理后台功能逻辑处理
4. 数据权限分类管理
5. 数据权限与呼入用户管理关联处理
6. 数据权限与呼出用户管理关联处理
7. 数据权限与95566用户管理关联处理
8. 功能实现

数据分析平台从用户及数据层面针对各级权限进行管理，开发配置工具用于进行权限配置。

同时对一期的用户权限管理进行重构，以适应新的Dashboard、用户导航、以及整合专业数据分析平台的需求。

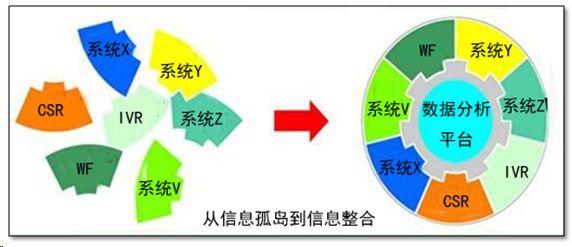
### 5.操作行为关联分析

1. 功能简述

将客户基础信息整合到现有数据中，实现客户基础信息、客户操作记录、客户办理业务信息等数据整合在一起的统计分析方案。

1. 功能细分
2. 客户基础信息数据处理
3. 三类数据整合后模型调整
4. 操作行为关联统计分析
5. 功能实现

根据CALLID整合从CSR、WF及IVR获取的相关数据形成数据链，根据业务需求进行各类统计计算，实现相关业务数据的关联分析，整合概念如下图所示：



### 6.特定业务场景可视化的实现方案

1. 功能简述

信用卡呼入业务、呼出业务以及95566呼入业务现有需求中无法通过自定义报表的方式实现的图表展示，如重复来电、IVR相关的图表，呼出业务中的气泡图等。

1. 功能细分
2. 重复来电定制化报表优化
3. IVR定制化报表优化
4. 呼出业务定制化报表
5. 95566定制化报表
6. 功能实现

结合不同业务特定的可视化需求与主流的可视化方法定制适合的可视化方案。

### 7.投诉模型

1. 功能简述

以客户投诉率为中心指标，对有可能造成影响的变量进行大数据分析，找出与投诉率关联性最强、共享度最高的变量值，并可对全体数据进行匹配标记与评分建模；通过评分建模的模式，运用统计学和概率论方法对不同评分阶段的来电进行投诉率预测。

1. 功能细分
2. 从投诉工单记录中抽取新的参数，并计算新参数与投诉率的相关性
3. 训练决策树模型，展示参数与最终客户投诉之间的逻辑关系
4. 寻找对客户投诉率有关系的参数
5. 对参数进行卡方检验，按照卡方值的大小排序，列出和客户投诉率最相关的参数
6. 自动对新数据进行训练并生成新的模型
7. 自动部署新的模型
8. 功能实现

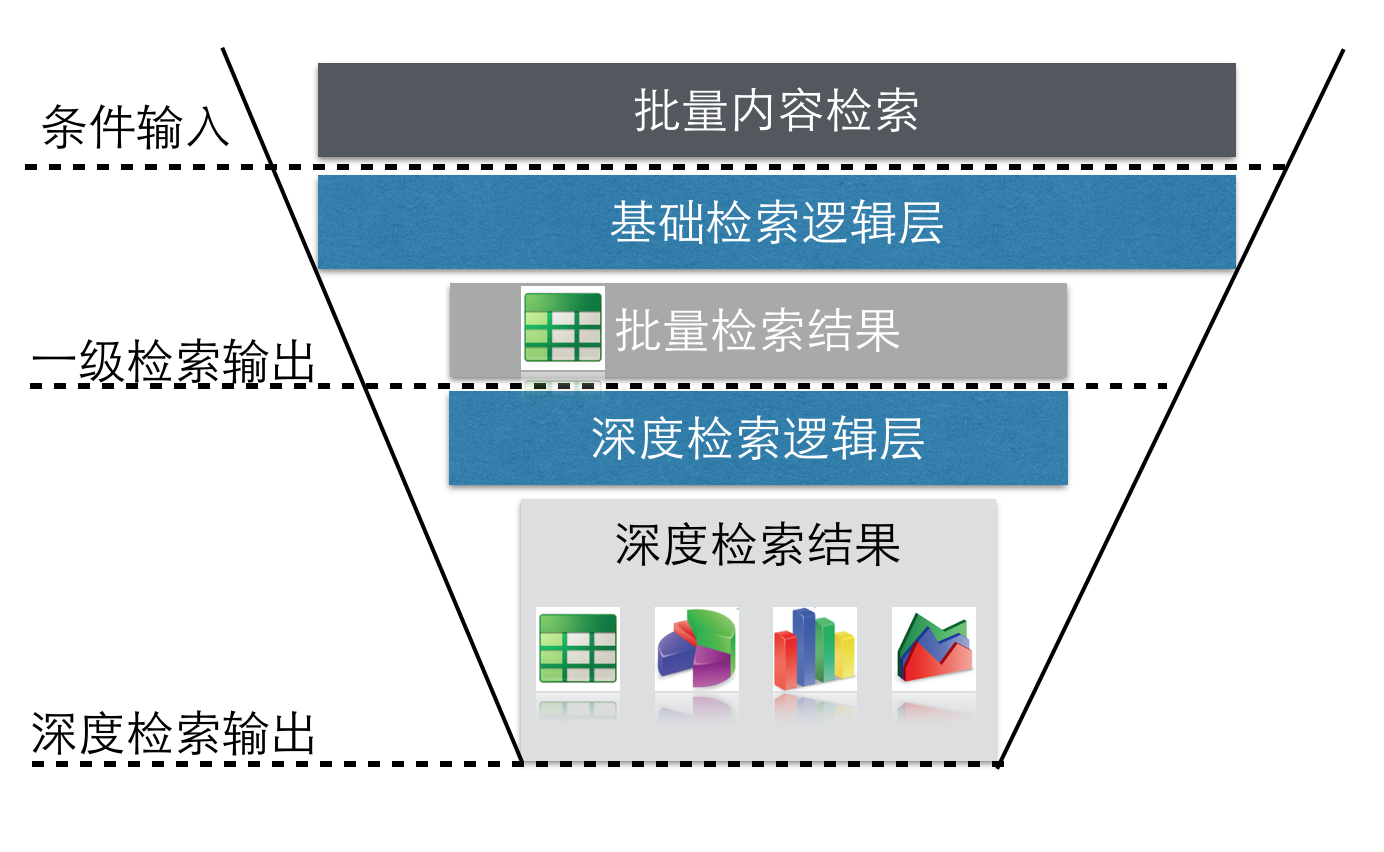
详见响应率模型实现方案。

### 8.二次统计分析方案（即套娃功能）

1. 功能简述

针对重复来电、平均通话时长等统计结果进行进一步的深度搜索，比如针对某种重复来电的搜索结果，自定义进一步查询或统计的条件（某组别、某交易）。

1. 功能细分
2. 批量检索后与检索条件关联
3. 检索结果二次统计分析前台展现
4. 重构后的检索与用户/数据权限的关联
5. 功能实现



深度统计的功能需要后台对数据检索条件的灵活支持，支持对搜索逻辑的增加，来实现深入的数据挖掘。

同时由于搜索条件的变化，需要通过良好的用户体验设计，在前端即时的展现深度搜索的结果。需要前端和后端的配合，同时在数据导出，图片下载等功能都要做出相应的修改。

### 9.ElasticSearch搜索引擎优化

1. 功能简述

优化现有搜索引擎的使用方式，充分利用分布式计算的能力，为数据探索和可视化提供实时的交互式体验。

主要涉及业务场景：呼出业务、呼入业务、95566业务中涉及检索的相关业务需求。

1. 功能细分
2. 自定义统计功能块与ES之间的接口重构
3. ES统计接口的开发
4. ES查询接口的开发
5. ES聚合函数接口的开发
6. 功能实现

原有系统因为重复来电的实现需要将大量数据从搜索引擎加载到应用中进行计算。为了满足程序设计的一致性，所以全部查询功能都需要加载大量的数据，导致程序稳定性受到影响。

此次将集成数据探索模块，需要重构原有的图表和表格数据检索功能。将原有自定义功能块翻译成能够执行的表达式，再由Envis数据探索模块发送查询请求至ElasticSearch，并充分利用ES的聚合统计功能。

### 10.提供专业数据分析工具的整合（H2O，Spark-R)

1. 功能简述

提供主流的自定义分析工具及接口，主要应用场景如下：响应率模型、投诉模型、首解率优化、年度呼叫中心白皮书等。

1. 功能细分
2. H2O portal的部署
3. H2O与Spark的结合
4. Spark-R接口提供
5. 功能实现

H2O和Spark-R是数据分析平台为数据科学家提供的数据挖掘和分析建模的工具和工作平台。

除了搭建H2O和Spark-R之外需要实现的是定制化的用户登录和权限管理，以及用户敏感信息的管控。

### 11. IVR的可视化、算法、转移矩阵等的实现方案

1. 功能简述

分析客服在IVR系统中的操作与转人工后进行的交易之间的关系，例如人工办理业务是否与之前的点选一致。

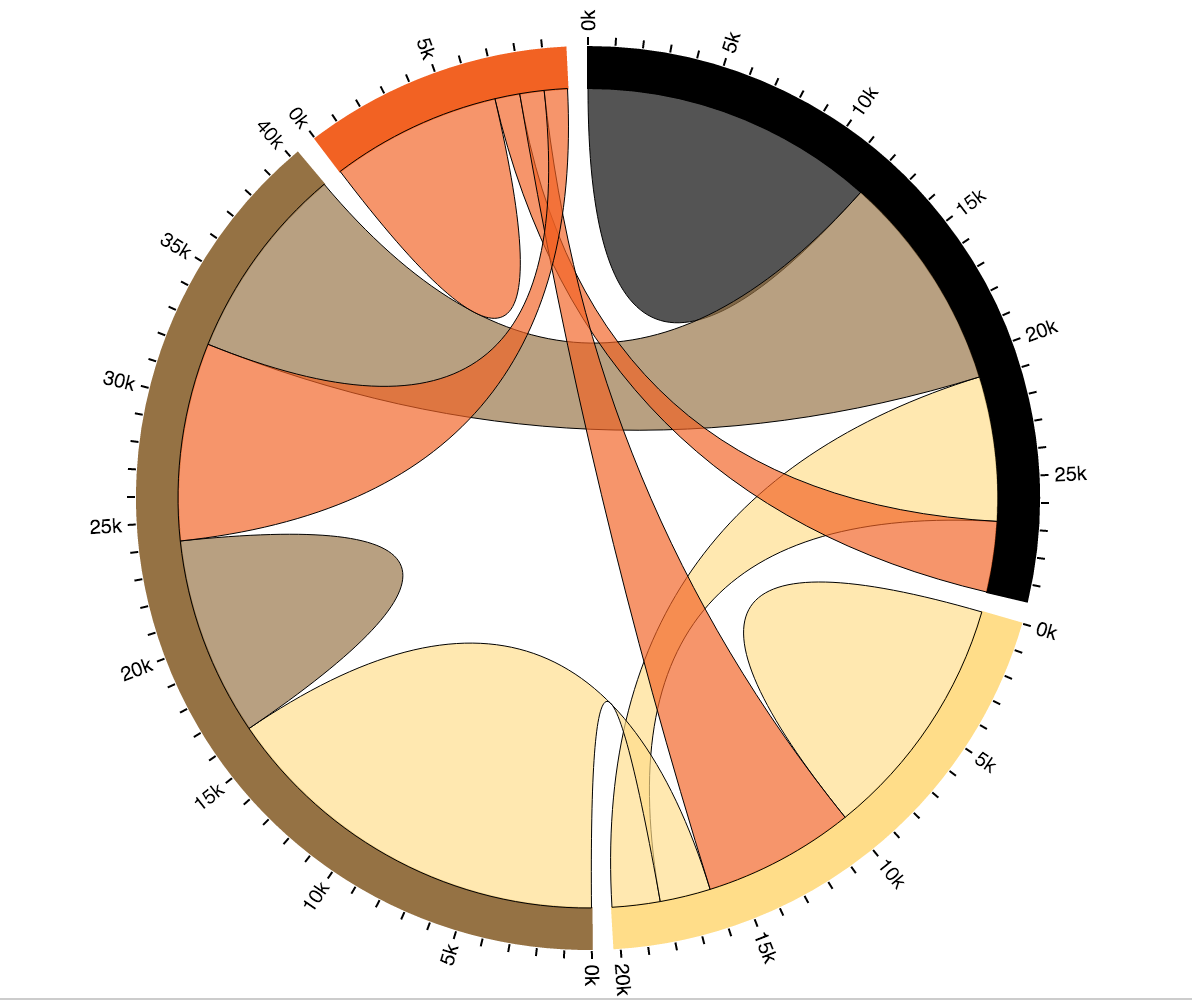
1. 功能细分
2. IVR路径规划的转移矩阵的可视化开发
3. IVR路径规划中转移矩阵的数据统计
4. IVR路径的模式抽取
5. IVR路径模式抽取的结果可视化
6. IVR路径展示-全结点前端展现
7. IVR路径展示-单结点前端展现
8. 功能实现

跟踪节点轨迹，整理并统计各种路径，针对全量路径数据进行关联及聚类等分析，包含轨迹的可视化。轨迹可视化方案可参看下图样式：



轨迹的集合就是一种树形结构，可以按照以上的方式进行展示。

转移矩阵的可视化方案可以参考：



越是相关的节点之间，他的转移概率越高，则连线越粗。

### 12. 重新清洗数据（含系统自动化运维、自动化部署）

1. 功能简述

数据分析平台的自动化部署、数据清洗、运维管理。

1. 功能细分
2. 自动化部署搜索引擎集群
3. 自动化部署pentaho
4. 自动化部署hadoop集群
5. 自动化部署spark集群
6. 自动化部署KDC
7. 自动化部署Nginx
8. 自动化部署web应用
9. 自动化部署sqoop
10. 自动化部署goserver
11. 功能实现

一旦出现增加标签的需求，所有已经进入搜索引擎的数据都需要重新打上标签，这时候就需要重新清洗数据。

清洗数据需要较长时间的准备和索引，整个过程极为复杂，所以对数据和系统的运行需要完全自动化，以避免人工操作产生的数据错误。

### 13. 特殊数据的标记及剔除统计

1. 功能简述

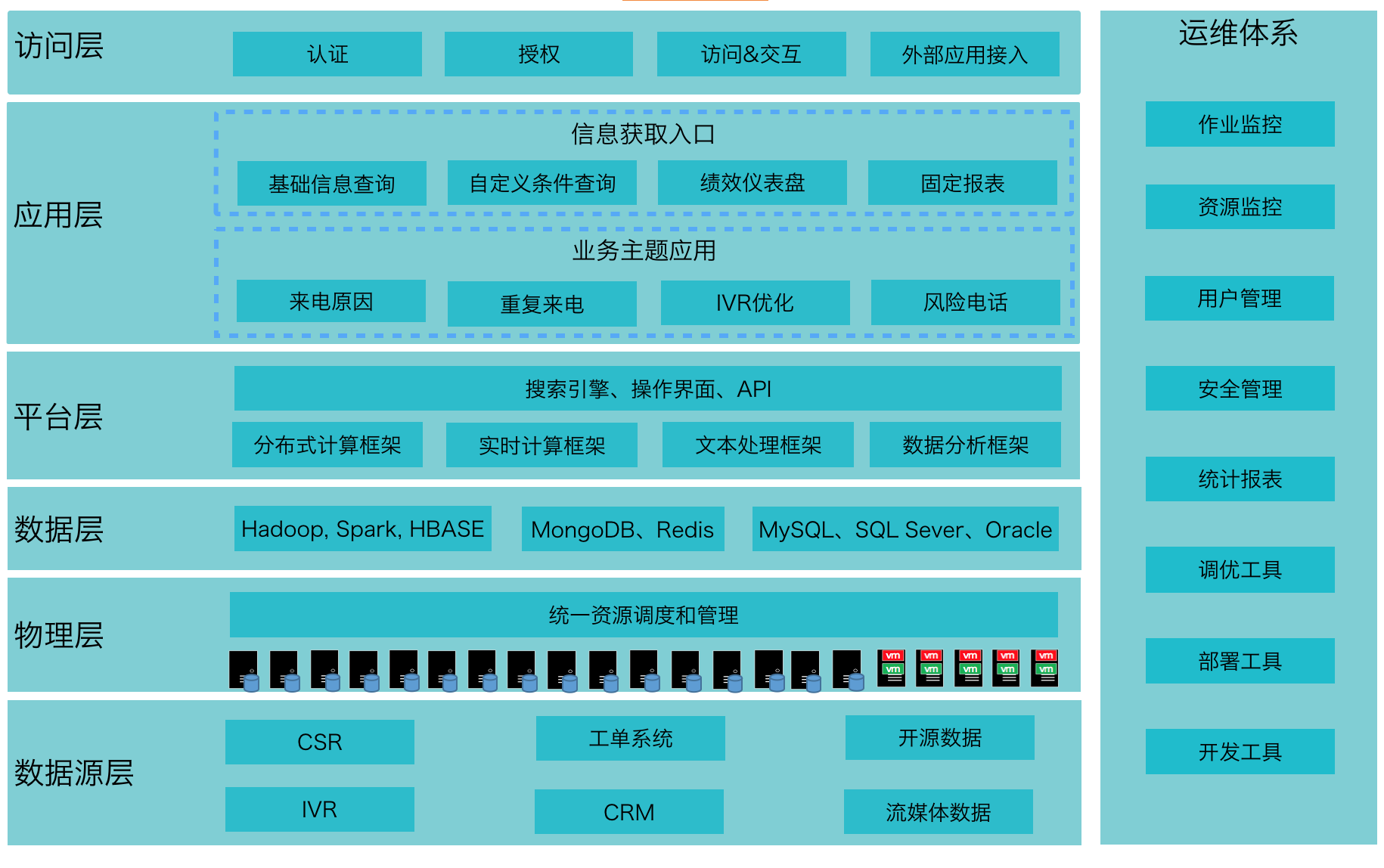
若电话符合某一规则，可被定义为干扰数据（如网络电话等），剔除重复来电的统计分析。

1. 功能细分
2. 配置pentaho对新数据的剔除逻辑
3. 重新索引所有在搜索引擎中的数据
4. 添加字段支持被删除数据的可视化分析
5. 功能实现

对所有来电的电话号码进行分析，找出网络电话、无号码的电话规则，并将这些信息加入到一个的字段中。在检索的时候，通过增加一个查询条件来过滤无效电话。

# 四、 技术架构

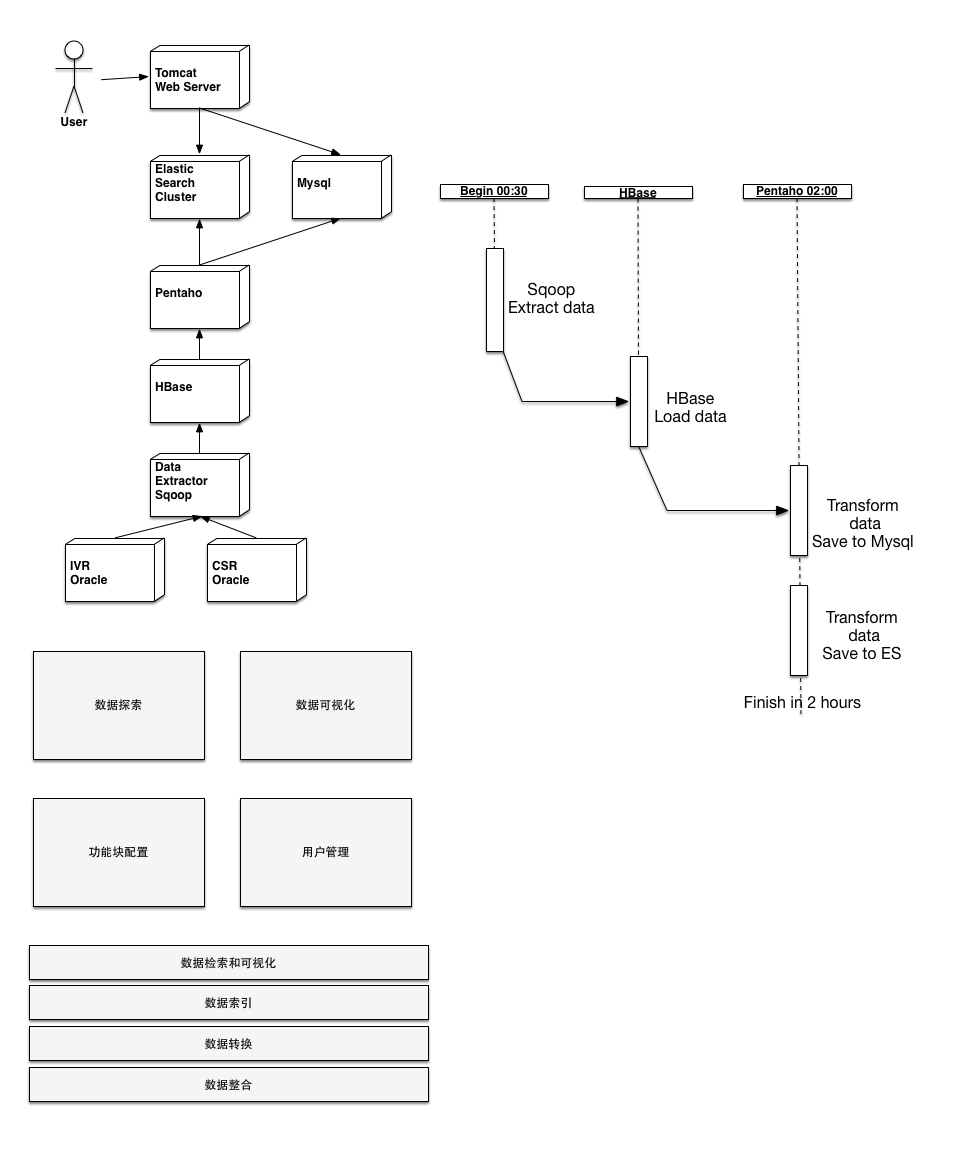
数据分析平台的核心是数据的采集、存储、分析、展现与预警。在第一期项目中，数据分析平台已经接入CSR、IVR和WF的部分数据，并搭建了以搜索引擎为核心的分析平台体系。下图是当前数据分析平台的整体技术框架。



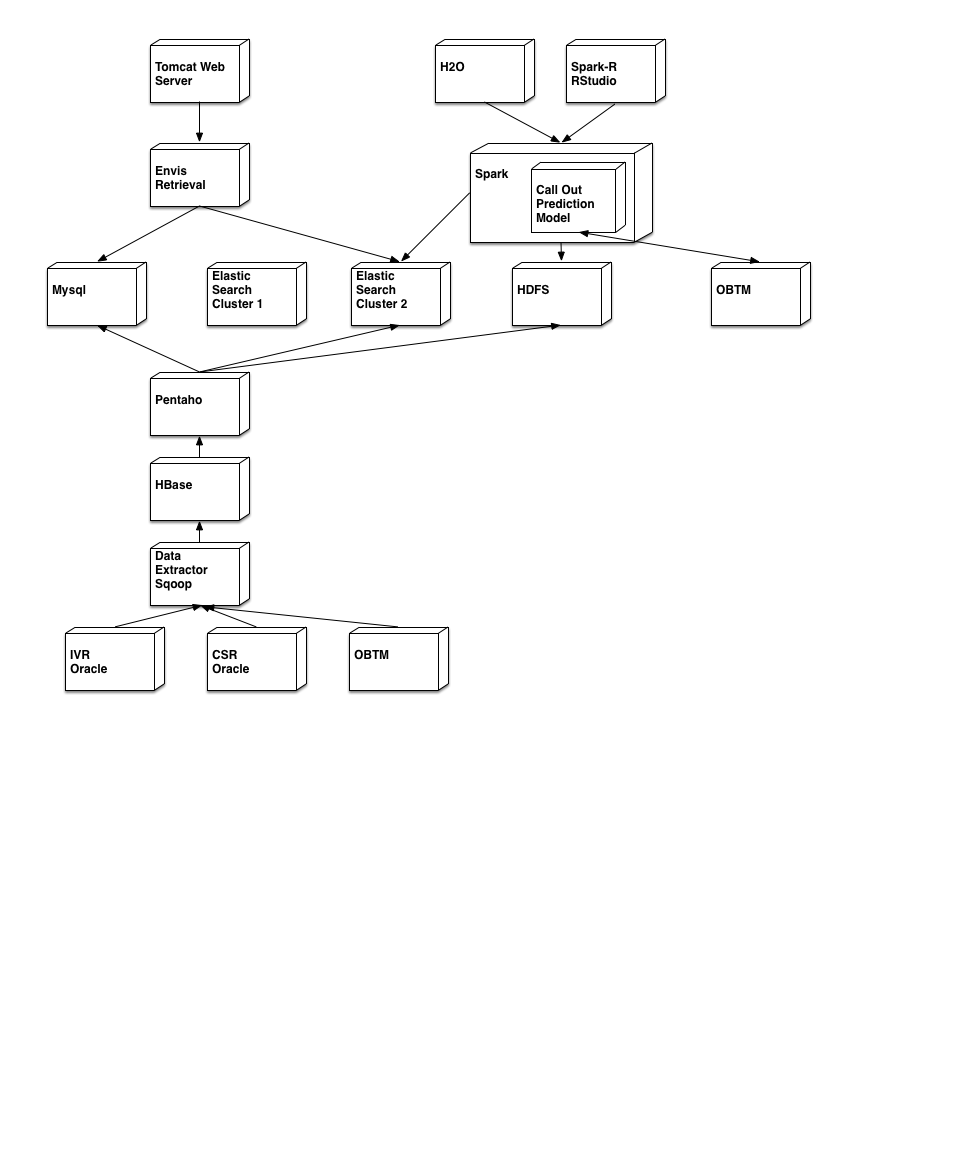
## （一）数据源和架构图

数据源层涵盖了二期项目需要纳入的新数据源，包括OBTM、DW的信用卡消费数据、CIF和CCMS。

平台一期的架构如下：

**

根据以上功能的增加，计划平台二期的架构为：



数据源增加了OBTM。同时为了实现响应率预测，增加了Spark计算呼出电话响应率。

Spark也为数据科学家提供了H2O和Spark-R等专业数据分析工具与平台。

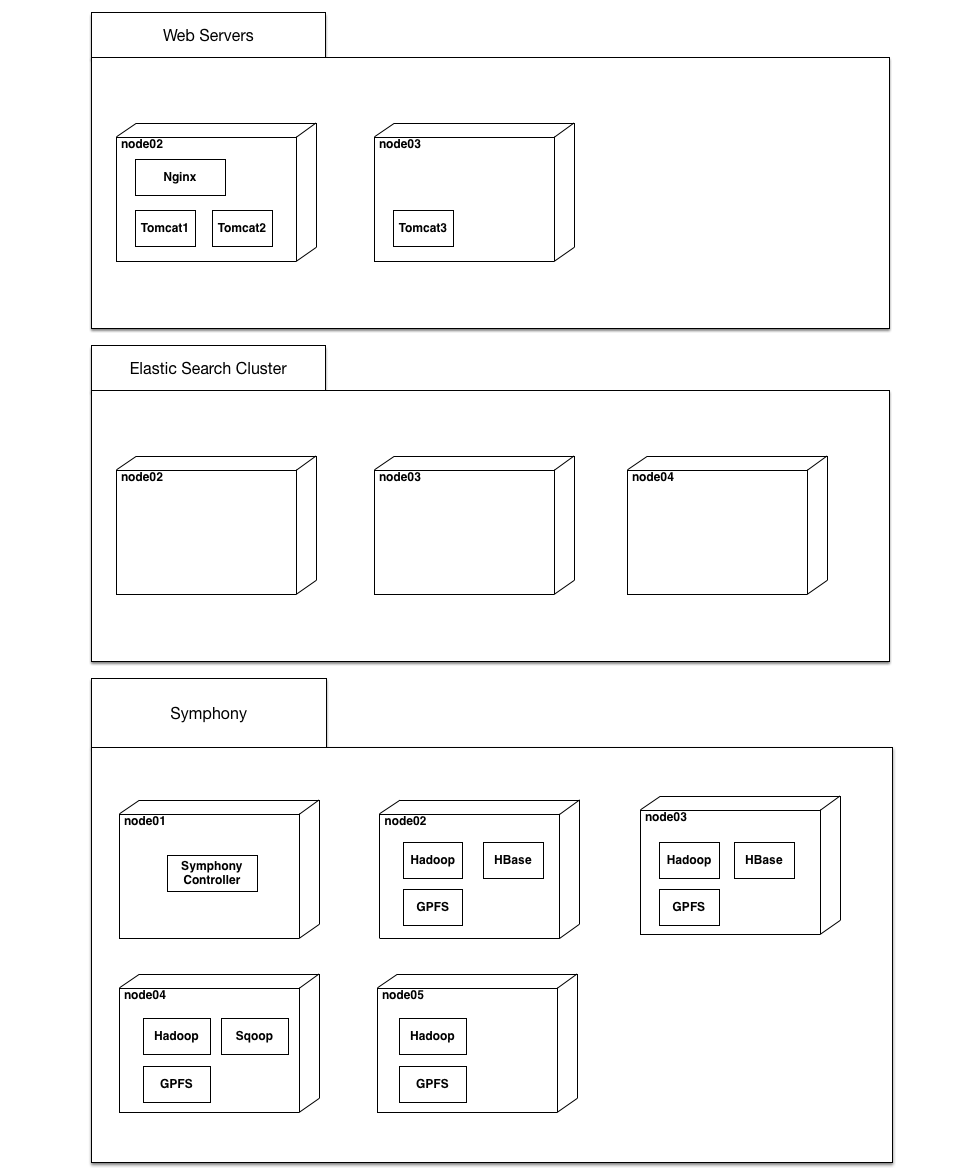
Pentaho在数据导出后再将数据存储到HDFS系统中，为数据科学家提供快捷的数据访问方式。

ElasticSearch增加了一个Cluster，做为清洗数据的备用Cluster。

## 物理层

平台目前物理层包括了当前数据分析平台所要用到三件环境。当前架构包括5台IBM Power 8服务器。

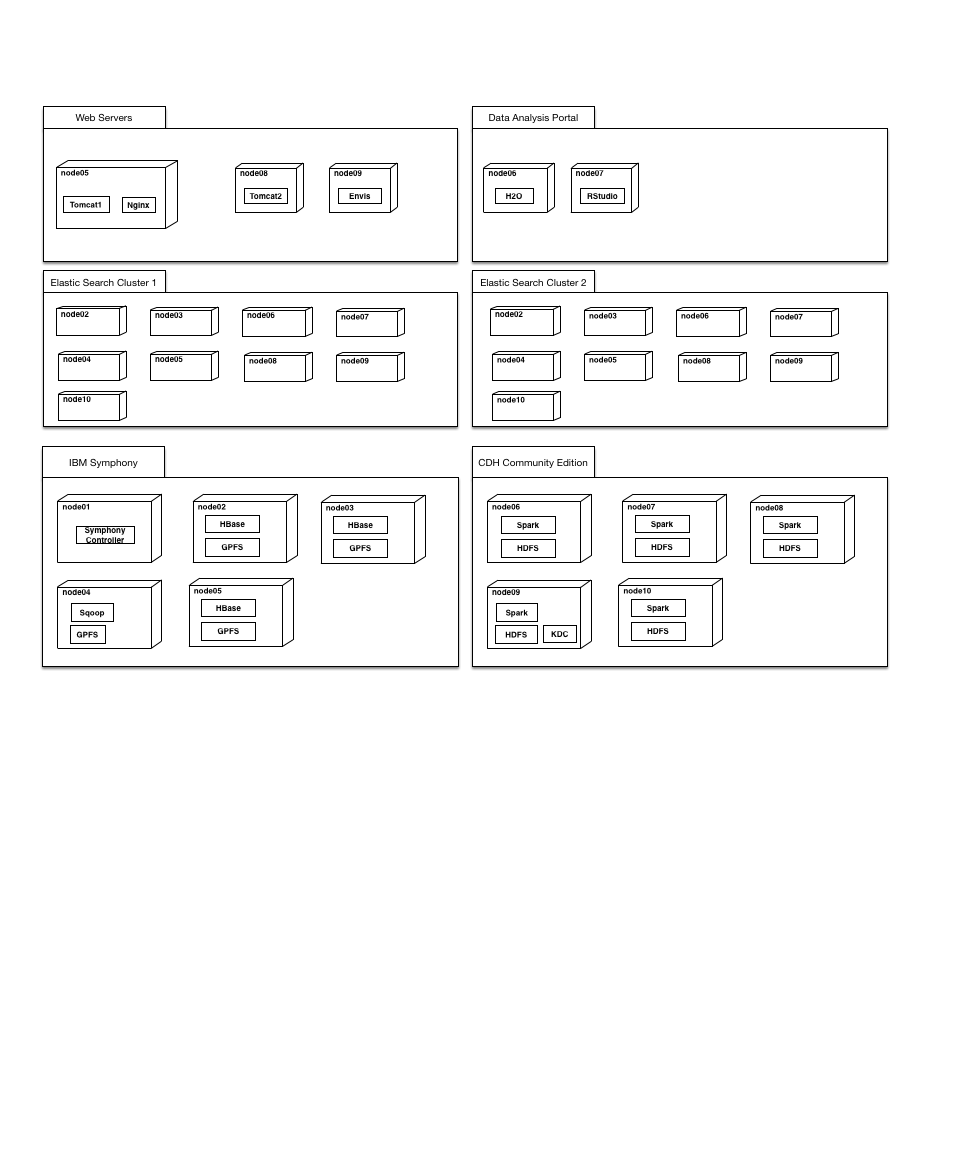
为了保证底层资源能够被合理分配和充分调动，包含一个统一的资源管理和调度平台-- Symphony。



根据二期新增数据量及处理性能要求，同时考虑平台的扩容性，评估需增加5台x86架构服务器：

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | CPU | 内存 | 硬盘 |
| 1 | 16核 | 64G | 5~10T |
| 2 | 16核 | 64G | 5~10T |
| 3 | 16核 | 64G | 5~10T |
| 4 | 16核 | 64G | 5~10T |
| 5 | 16核 | 64G | 5~10T |

增加的5台x86架构服务需部署了新的spark、elasticsearch、h2o、spark-r、KDC等。这些节点的作用已在架构图中描述。



## 数据层

物理层之上是数据层，或者称之为数据存储层。本方案将数据存储分为了两个部分：GPFS（包含开源Hadoop）和MySQL数据库。

Hadoop主要用于原始数据和明细数据的存储，数据存储于GPFS之上，并利用NoSQL数据库HBASE来访问。Hadoop是当前业内最为主流的分布式系统基础架构，既可以满足任意类型数据的存储需求，也可以通过简单增加机器以实现性能和存储空间的快速扩展。

在当前方案中，RDBMS仍然是系统的重要组成部分。因为RDBMS在结构化数据检索方面有着自己得天独厚的优势。Hadoop更适合大量数据的批量处理；而RDBMS则适宜于单条结构化数据的处理。但RDBMS受限于部署方式和成本，当前方案只会要求将部分汇总的数据和基础维度数据存储于RDBMS之上，这类数据的特点是高访问频度，但数量很小。当前采用的RDBMS为MySQL。

## 平台层

平台层主要作用是进行数据处理。分布式计算框架、数据分析框架以及实时计算框架都在这一层加以实现。

在各个计算框架之上，方案添加了企业级搜索引擎Elastic Search，用于数据的检索。换句话说，搜索引擎是系统所有数据的出口。

采用搜索引擎的原因有两点：一，搜索引擎借助于其强大的索引等功能，可以满足从海量数据中迅速获取所需数据的需求；二，在搜索引擎之上，开发一个统一访问语言，用于与搜索引擎进行对接；只要上层应用的数据输入输出接口，按照该统一访问语言进行设计，即可迅速地完成平台层与应用层的连接。

## 应用层

平台层之上就是应用层。如果说平台层及之下的各层，还与系统的实际使用人员相隔较远，应用层则会直接而对系统的使用人员。

应用层包括两个部分：业务应用主题应用和信息获取入口。

一期业务应用主题就是来电原因分析、重复来电识别、IVR图形展示等。在第二期项目中，还将引入外呼业务分析、外呼名单呼应率预测、IVR轨迹分析、重复来电深入分析等。

信息获取入口，则是指数据的各种展现形式，包括基础数据查询、Dashboard、固定报表、预警报表、动态多维分析报表等。

## 访问层

访问层是包含了一些对外的入口，所以认证、授权、单点登录SSO及与外部系统对接，都将在这一层来完成。

## 运维体系

运维体系是整个系统正常运转的保障，作业监控、资源监控、用户管理、权限配置（功能及数据权限）、运维统计报表、安全管理、调优工具、开发工具，以及持续集成、部署工具等，都将包含在整个运维体系之中。