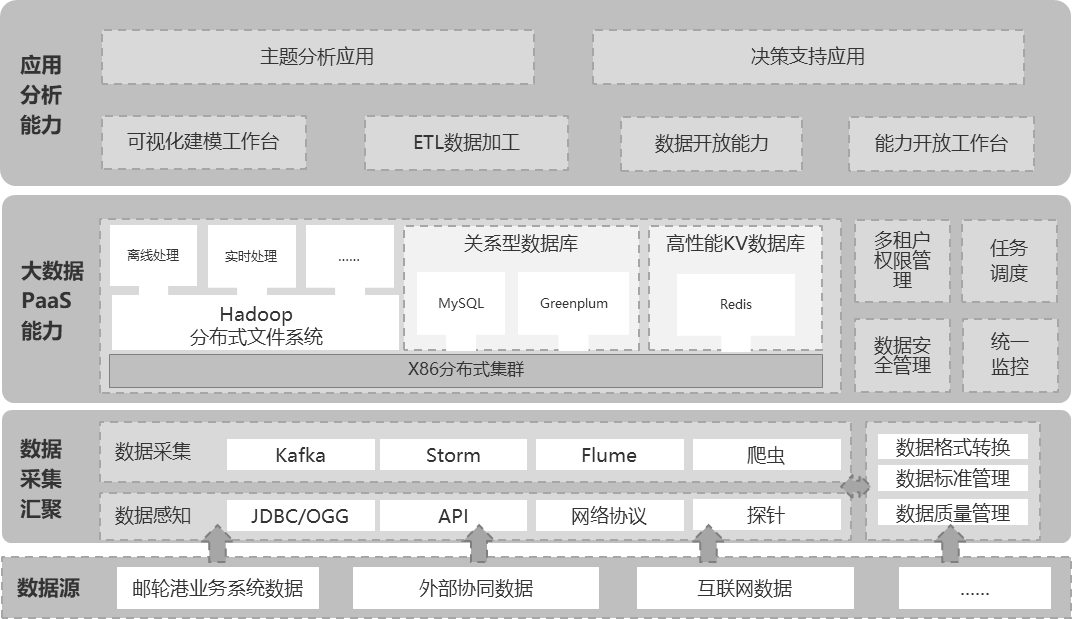
###### 大数据能力平台

平台整体架构



大数据能力平台提供基础的大数据存储、计算、分析及数据治理的平台和框架，是整个平台的核心，向上层应用提供基础能力。基于“集中、开放、安全、云化”的原则，构建三大能力：

1. 数据采集汇聚能力：将各种数据源接入到大数据平台或者其他数据仓库，为大数据计算和分析提供基础数据。
2. 大数据PaaS能力：提供了大数据处理过程中所需用到的各类工具及服务，覆盖了大数据全生命周期管理的各个环节，实现数据全流程管理，大数据量的批处理、实时处理和数据存储能力。
3. 应用分析能力：提供了多种数据转换和大数据分析挖掘算法，全面支持邮轮港的生产、销售、市场营销、财务管理、企业决策等各类应用。

数据采集汇聚层

数据源管理

数据源管理主要从数据源类型、数据源更新频率、数据结构三个维度进行分析及分类，根据不同的数据源、数据更新频率、数据结构，确认采用不同的技术框架实现数据感知与采集的功能需求。

按照数据源类型，可以将邮轮港大数据分为三类：

一、邮轮港各业务系统数据：

1. 游客基本资料（护照、身份证、近照、手机号码）等各类信息；
2. 面向游客的咨询、服务、向导等各类信息；
3. 邮轮公司、邮轮、码头、泊位、航线、航班、船舱等各类信息；
4. 票样、船票、订单、代理人、合同样式、合同、账务、行李、车辆、评价、满意度等各类信息；
5. 海关涉及到的报关单、违禁物品、违禁库存、凭证票样、行李单票样、违规信息、征税、缉私等各类信息；
6. 单一窗口涉及到的报关邮轮、报关游客、报关船员、报关航线、报关航班等各类信息；

二、外部协同数据：

1. 区港联动涉及到的公共交通、公安安全、安防点位、公共视频、智能人脸、游客分布、游客动线、联动协同、应急预案等各类信息；
2. 港区涉及到的智能感知设备（烟感、消防栓、消防闷盖、窨井盖、空气监测、水质监测等）、IBMS楼宇自控信息等各类信息；
3. 新技术应用涉及到的5G服务机器人、无人机、360全景信息、一体化自助机、手持终端等各类信息。
4. 电信运营商用户标签数据，如基础画像标签、兴趣标签、上网行为标签、位置标签等。

三、互联网数据：

1. 主要包括互联网上的社交数据、交易数据、信息发布数据等。
2. 基于爬虫抓取的各类信息数据。

数据源管理对各个服务系统的数据源进行管理，为数据的采集提供必要的信息，并和元数据信息进行关联，对于数据源上元数据的采集可以定制任务。数据源管理提供数据源的展现、新建、查询、编辑、删除功能。

(1)数据源信息

针对面向游客邮轮旅游全生命周期信息、面向内部管理及区港联动各类信息，提供数据源信息的展示，包括数据源的基本信息和关联的元数据信息，数据源的基本信息包括数据源类型、网络协议、网络端口、服务器名称、服务器IP等；展现元数据的基本信息，并提供链接展现元数据具体信息。

(2)数据源查询

针对面向游客邮轮旅游全生命周期信息、面向内部管理及区港联动各类信息，提供多选项关键字的模糊搜索功能。提供数据源的数据源类型、网络协议、网络端口、服务器名称、服务器IP等条件的搜索功能。

(3)数据源维护

针对面向游客邮轮旅游全生命周期信息、面向内部管理及区港联动各类信息，新建数据源应包括数据源基本信息的录入，并可添加任务，包括数据源类型、网络协议、网络端口、服务器名称、服务器IP等，并且新建时进行心跳检测，确保数据源的存在以及运行良好。

数据采集

国际邮轮产业上海数据中心目前数据主要来源于本期拟建的游客服务平台、口岸协同平台、票务平台、港口运营管理平台中各个应用系统。在本次建设中将对上述各个系统的数据规划数据采集功能，实现数据源管理、数据采集计划任务管理，从其他平台抽取各类数据，汇聚处理。其中的数据源管理，需要实现数据源的查看、新建、修改和删除等操作，和对数据源配置信息管理。计划任务管理功能支持不同的数据采集策略，包括周期性的数据采集，或者根据特定的预定义时间表的数据采集,或者使用事件管理器执行工作流进行数据采集，或响应突发事件的触发器执行数据采集。主要实现包括以下系统的数据采集：

* 单一口岸数据协同系统
* 游客离港系统
* 海关协同系统（海关旅检辅助应用系统）
* 公安协同系统
* 泊位预定系统
* 船供管理系统
* 综合可视化系统
* 信息发布系统
* 应急预案系统
* 船讯服务集成
* 港区作业管理系统
* 港区安全管理系统
* 港区设备管理系统
* 爱邮轮服务系统
* 爱邮轮后台系统
* 爱邮轮接口系统
* 渠道管理系统
* 交易管理系统
* 销售服务系统
* 票务接口系统

数据清洗

数据清洗是数据质量、数据应用重要一环，智慧邮轮港信息化项目涉及系统多，业务杂，数据模型多样，每天每刻都在产生大量数据，如游客服务数据，游客订单数据，游客值船数据，游客安检、登船数据、港区日常运营数据、港区设备运行状态数据等对数据清洗过虑，检测和去除数据集中的错误数据和无关数据，处理遗漏数据，去除空白数据域。

国际邮轮产业上海数据中心数据清洗采用两步策略开展：

第一步：根据邮轮港业务规则，预先定义好港区各类数据清洗算法，由计算机自动执行算法，对数据集进行清洗，然后产生清洗报告。

第二步：根据清洗报告，在邮轮业务专家的指导下，收集分析游客数据、邮轮数据、港口运营数据，除明显的边缘数据和重复记录，填补缺值数据等清洗动作。

清洗后，形成智慧邮轮港基础单元数据，用于数据挖掘和分析。

异构数据交换

不同平台不同业务系统存储数据的方式各不相同。平台需要提供灵活多样的系统接入和数据采集方式，支持多平台、复杂异构系统接入和数据采集的需要。

数据交换服务采用异构系统接入和数据采集处理方式，能够适配项目建设中诸多系统，数据信息内容、格式、存储方式等方面的差异（如不同数据库；不同库、表和字段定义方式；文件等），有效减少平台建设开发的工作量，提高平台建设实施效率。

提供对各种数据库系统的访问能力，以实现异构数据库的互联。

提供数据的过滤、填充功能，以实现抽取来源于各种系统的数据内容能够满足数据中心的要求。

提供数据库同步机制，以实现将各业务数据交换平台的中心数据库中的数据自动同步到数据中心的数据库中。

数据格式转换

数据交换提供数据整合所需的各种接口协议和数据格式支持，支持按照预定义的数据采集策略，根据设定的规则和条件，按需从各类数据源采集数据。保证各种格式的数据均能够按数据中心的规范格式存入数据库系统。 数据经格式转换、数据校验后等数据质量检查后写入数据存储库中。

数据标准管理

智慧邮轮港信息化建设应该是一个持续发展、面向政府、游客、产业的多元化开发的体系，因此从整体架构出发需建立一套严格的数据标准、统一规范数据模型、集中管理，做好数据标准架构、标准形式、使用标准，形成标准资产。建立数据标准，有利于后续各个系统协同实施，大量节约系统间开发成本和沟通成本；有利于保障项目实施工期按计划实施；同时数据挖掘分析之前，制定数据标准，规范元数据管理，通过数据标准化方案，对不同来源数据进行查找、检索和比对数据对象及其含义、特征和使用情况等方面的信息，帮助邮轮港将数据进行数据分类编码，落实信息化建设的“统一标准”原则，从根本上解决“信息孤岛”问题，为后续的邮轮信息化系统建设提供标准。

数据质量管理

针对面向游客邮轮旅游全生命周期信息、面向内部管理及区港联动各类信息，通过数据质量检核，暴露各系统数据质量问题，并提醒用户进行数据质量的改进。对不符合质量要求的数据进行清洗、筛选。质量规则包括规则名称、规则描述、规则标准（唯一、非空、字符大小等）。

* 数据质量规则管理

制定数据质量规则，录入规则的基本信息，包括规则名称、规则描述、规则标准。

提供多选项关键字的模糊搜索功能。提供质量规则的规则名称、规则描述、规则标准等条件的搜索功能，并可以查看绑定的元数据，进行页面的跳转，查看元数据的详细信息。

* 数据质量告警提示

当数据采集出现大量的不符合数据质量标准的数据时，对该数据对应的服务进行告警提示，通知维护人员进行排查，以保障数据和服务的安全性。

* 数据质量监控

监测从各个业务单位汇集的原始数据，通过自动/手动方式发现原始库中存在的数据质量问题。

系统提供唯一检查、非空检查、外键检查、代码检查、长度检查、值域检查、一致性检查、自定义检查等自动监测方式。

系统主要提供两个功能：

实体重复记录监测：列出所有表，对其中的重复记录进行比对去重。

参照关系监测：参照数据字典，提供标参数、列参数、库参数等监测。

该模块实现规则与参数模型、监测器（定时调用检查模型）之间的关联、还实现了方案和规则的管理功能，方便用户的维护。

* 数据质量报告

数据质量监控报告可以查看数据巡检任务的运行结果情况，比如被巡检数据源符合数据规则的数量，总的问题数据的数量，记录时间等，根据问题数量找到原始数据核对，保证数据的准确性。

大数据PaaS能力层

随着港城业务更进一步发展，信息浏览、搜索、关系交互传递，对于实时性的要求进一步提升，而信息的交互和沟通正在从点对点往信息链甚至信息网的方向发展，这样必然带来数据在各个维度的交叉关联，数据爆炸已不可避免。大数据PaaS能力层提供了大数据处理过程中所需用到的各类工具及服务，覆盖了大数据全生命周期管理的各个环节，实现数据全流程管理。具备以下四大特性：

* **高可靠，易扩展：**PaaS层基于Hadoop分布式系统基础框架，同时结合Docker容器技术，为用户提供高可靠，易扩展的架构
* **图形化界面：**工具组件皆采用图形化界面，UI设计友好，可拖拽式操作，降低用户的使用门槛
* **数据安全保障：**工具组件皆采用图形化界面，UI设计友好，可拖拽式操作，降低用户的使用门槛
* **微服务化部署：**各个工具组件采用微服务化部署模式，可简单插拔，便于用户使用

**数据存储**

平台将分布式文件存储作为平台最基础的存储架构，同时结合关系型数据库，内存数据库，实现分布式内存数据库+关系型数据库+HDFS的混合分布式存储框架实现对结构化数据、半结构化数据以及非结构化数据的存储管理，满足分布式分析和大数据挖掘的编程开发，为不同的大数据使用场景及其应用，提供最优的数据服务：

* **分布式文件系统：HDFS、Hive、Hbase**

提供强大且可靠的存储能力，可以存储海量的非结构化原始数据文件，所以邮轮港大数据中的，智能设备数据、数字设备数据、互联网数据，都将会纳入此平台存储管理。同时提供分布式分析和大数据挖掘的编程能力，支撑各类数据分析挖掘应用。

* **内存KV数据库：Redis**

具备高性能的数据读写操作能力，有效应对海量数据下的数据查询、分析挖掘场景，提供高并发低延时的数据查询服务。

* **关系型数据库：Mysql**

目前最流行的开源关系型数据库管理系统，它支持多线程高并发多用户，为各类应用服务提供基础的数据存储方式，支撑应用的数据查询，关联等场景。

数据存储包括结构化数据存储以及非结构化数据存储，为了支持后续的扩容，采用分布式数据库，结构化数据存储在数据库，非结构化数据存储在文件系统上。

分布式计算

数据离线处理

离线分析计算引擎是一个基于MapReduce编程框架的分布式计算平台，用于支撑海量大规模数据集的并行处理，其中包括对数据的加工、变换、传输等流程。上层应用通过提供数据处理的算法接口，将上层数据处理子系统有效的和海量数据计算进行整合，从而实现任务的集群化并行处理。

MapReduce处理的任务基本要求是：待处理的数据集可以分解成许多小的数据集，而且每一个小数据集都可以完全并行地进行处理。MapReduce计算模型的核心部分是map和reduce函数。用户自定义的规则，将输入的<key,value>转化成另外一个或一批<key,value>对输出即可。

数据实时处理

邮轮产业参与者每个人都是个数据源，每时每刻都在产生着数据与个性化、实时化的需求，在普遍认知中，数据的价值会随着时间的流逝而降低，所以在事件出现后必须尽快对它进行处理，最好是数据出现时便立刻对其进行处理，发生一个事件处理一次，而不是缓存起来进行批处理。实时分析处理是针对流式数据的实时处理（计算），它具有数据实时持续不断到达、到达次序独立、数据来源众多格式复杂、数据规模大且十分关注存储、注重数据的整体价值而不关注个别数据等特点。

实时分析处理一般有三个流程：数据实时采集，数据实时处理，实时查询服务。而这个过程也就对高性能、海量式、实时性、分布式以及可靠性提出了需求，并且，实时分析处理需要两个层：存储层和处理层。存储层需要支持记录定序和高度一致性，以便以快速、便宜且可重复的方式读取和写入大型数据流。处理层负责处理存储层中的数据，基于该数据运行计算，然后通知存储层删除不再需要的数据。除此之外，还必须为存储层和处理层制定可扩展性、数据持久性和容错规划。

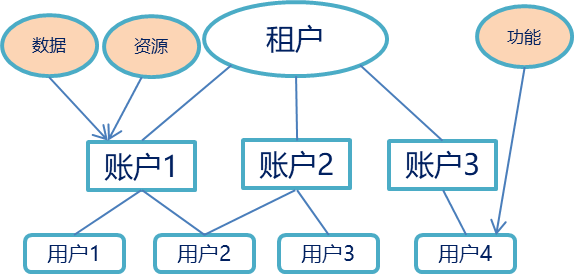
平台基于Flume实现批量数据采集、Kakfa消息队列、Storm流式处理以及Spark-Streaming实时计算方式，实现实时数据采集。



多租户权限管理

租户是PaaS平台中完整拥有资源管理权限的独立单位，租户之间资源彼此隔离。一个租户对应现实的公司、团队。每个租户可以包含一个或多个项目，也可以没有项目。租户内有一个最高权限帐号，称为“租户所有人”，就是创建租户的帐号，这个帐号可以执行租户级别的所有操作。

* **权限模型体系**



基于“租户-账户-用户”的权限模型体系，实现平台用户“一点登录、一点注册、一点鉴权、一点管控”。

* **资源权限**：CPU、内存、存储等资源映射于账户，通过账户区隔计算资源，存储资源
* **数据权限**：存于大数据平台的数据资源映射于账户，统一管控用户数据查询、数据使用
* **功能权限**：菜单权限，功能权限映射于用户，不同用户登录后对应于不同的菜单功能
* **KDC安全认证中心**

Hadoop本身不能提供完整的DBA控制功能，只有简单的目录权限控制等功能，不能满足大数据平台多租户权限管理要求。多租户权限管理系统基于Kerberos认证服务结合大数据权限模型，实现不同业务、不同用户的群组管理、数据表管理、字段管理、策略管理、访问记录管理等，真正实现了企业级别的大数据中心管理职能。

KDC（Key Distribution Center）是Kerberos建立的一个安全的、可信任的密钥分发中心。KDC在Kerberors中通常提供两种服务：

* Authentication Service(AS)：认证服务
* Ticket-Granting Service(TGS)：授予票据服务
* **RM资源管控**

大数据平台通过RM资源管控机制，可以为每个用户分配独立的数据空间和计算资源，以保证每个用户的业务应用过程能正常运行。

1、RourceManager资源管控中心

ResourceManager负责集群中所有资源的统一管理和分配，它接收来自各个节点（NodeManager）的资源汇报信息，并把这些信息按照一定的策略分配给各个应用程序（实际上是ApplicationManager）。

2、HDFS存储资源安全管控

ACL访问控制可以配置用户的HDFS信息。HDFS信息包括：

1. 文件数限额。
2. 可用文件数
3. 空间限额
4. 可用空间
5. 用户目录
6. 目录数
7. 文件数
8. 文件总大小

HDFS配置界面如下图所示：



图- HDFS安全配置界面

3、MapReduce计算资源安全管控

ACL访问控制可以增加、删除或修改队列信息。队列信息包括：

1. 队列名
2. 容量设置
3. 可用容量
4. 最大容量
5. 是否支持job优先级
6. 用户资源百分比限制
7. 用户可占队列容量系数
8. 队列中并发task上限值
9. 每个用户并发task上限值
10. 每个队列中可容纳job总数的系数
11. 初始化后并发执行的job数

MapReduce配置界面如下图所示：

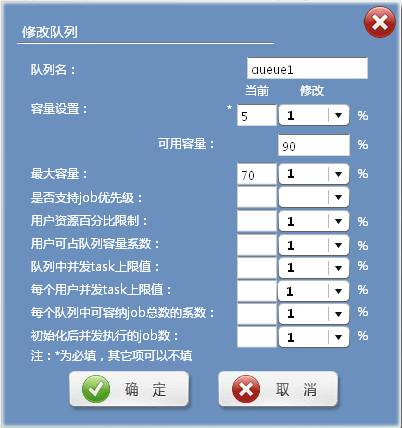


图- MR安全配置界面

* **资源审批分配**

建立租户资源申请流程，租户需求部门发起资源使用申请，包括计算资源和存储资源，以及使用期限，提交到大数据系统运营单位进行审批，系统运营单位根据租户的使用需求，评估租户所需的存储资源和计算资源，进行合理的分配。

存储资源包括：文件数限额、可用文件数、空间限额、可用空间、用户目录、目录数、文件数、文件总大小。

计算资源包括：队列名、容量设置、可用容量、最大容量、是否支持job优先级、用户资源百分比限制、用户可占队列容量系数、队列中并发task上限值、每个用户并发task上限值、每个队列中可容纳job总数的系数、初始化后并发执行的job数。

* **资源使用监控**

租户接入大数据系统采用统一帐号进行权限管理。

租户进行任何操作前应进行身份标识和鉴别，认证通过之后，只能操作所在租户权限范围内的数据和功能。

租户操作行为分为hiveweb、hdfs、hive访问三种类型。

对大数据系统的多个租户环境进行逻辑隔离，避免各租户环境互相影响。

每个租户都配置单独的队列以及资源限额，不会发生相互抢资源的情况。

为每个用户租户建立生命周期设置，设定租户的到期日期，在到期日前10天、5天、1天提醒大数据系统管理人员及对应租户。提醒方式支持短信以及邮件两种。到期日期可以配置。

* **资源回收**

PaaS平台根据租户的资源消耗情况分析，自动回收闲置的资源，包括：

* HDFS资源自动回收
* 内存资源自动回收
* CPU资源自动回收

待任务完成后，系统自动对租户分配的资源进行回收，重新初始化，以便为下个租户提供服务。

**数据安全管理**

满足在数据存储、传输过程中的安全保密性需求。在数据中心运营过程中涉及大量的敏感数据，在其处理过程中，特别是与各级单位数据交换过程中，要进行数据加密传输和存储，要保证数据的安全保密性。

数据安全管理主要负责对整个数据中心系统技术管理，包括人员层面，数据层面，服务层面，主要从以下几个方面入手。

数据加解密

* 数据加解密

密钥管理

实现加解密公钥密钥单独管理，与平台分离，存放于专线独立的服务器中，进行多重安全防护，如网络环境、硬件防火墙、操作系统防护，保证密钥安全；

密钥支持定期自动更换，防止长期公钥保持不变，被猜测破解。自动更新的新密钥，可以继续解密原有加密数，不是整体重新加密。可以配置密钥的自动更新周期。

多层加密：

相同数据，不同角色使用不同的加密算法或者密钥;

数据存储层、数据显示层、数据使用层可以支持不同的密钥;

数据加密支持随机加密策略，避免相同数据的加密结果相同。

敏感数据识别：

对于用户导入等新增的数据进行自动识别，数据进入Paas平台后，实现敏感字段自动识别。主要识别手机号、身份证号码等敏感信息。

自动定级：

使用敏感数据自动识别的结果对数据进行定级，并实现元数据的自动调整，定级规则同目前已实现的一级明文、二级脱敏、三级加密。

* 数据脱敏

**1.数据脱敏功能**

鉴于系统内的数据可能牵涉用户隐私以及生产数据，所以系统在数据使用以及对外提供时具备数据脱敏功能，同时系统也应具备脱敏恢复的功能和机制。系统的数据脱敏管理功能分为3个层次级别，分别为不脱敏、部分脱敏、完全脱敏。管理员可以根据安全要求设置数据的脱敏级别，系统应提供相应的管理界面供管理员进行设置和更改。

1)不脱敏

系统不对数据的使用以及对外提供数据进行脱敏处理，数据使用方和接收方可以看到原始的数据及用户敏感信息。

2)部分脱敏

系统对数据的使用以及对外提供数据进行部分脱敏处理，数据使用方和接收方通过数据无法完全获知敏感信息。例如：用户手机号经过部分脱敏处理之后，使用方只能看到尾号部分“ABCDEFG8888”。

3)完全脱敏

系统对数据的使用方以及对外提供数据进行完全脱敏处理，数据使用方和接收方通过数据无法获知敏感信息内容。例如：用户手机号经过部分脱敏处理之后，使用方只能看到无法识别的内容“ABCDEFGHIJK”。

**2.脱敏恢复功能**

系统需提供3种脱敏恢复机制，分别是自行恢复模式、界面操作模式和接口恢复模式。

自行恢复模式：是数据使用方通过系统提供的界面申请“脱敏恢复功能”，当申请通过之后由管理员分配相关的“脱敏恢复”密钥及算法。系统需支持多种加解密的方式以满足多用户使用的需求。

界面操作模式：是数据使用方通过系统提供的界面申请“脱敏恢复功能”，当申请通过之后由管理员分配相关的“脱敏恢复”界面使用权限，使用方通过系统提供的操作界面进行“脱敏恢复”使用。

接口恢复模式：是数据使用方首先通过系统提供的界面申请“脱敏恢复功能”，当申请通过之后有管理员分配相关的“脱敏恢复”密钥及数据连接信息配置以及SLA控制，使用方通过接口方式连接系统并进行“脱敏恢复”操作。系统需具备对接口脱敏恢复方式的SLA控制功能及安全性保证。

日志审计

用户行为分析：

使用户行为数据，通过模型训练和数据积累，最大限度预测非法操作、危险操作，形成用户安全画像。在非法操作实施前，进行预警，告知管理员进行相关研判和管控操作。

应用行为分析：

针对应用程序操作数据的日志记录信息，刻画应用程序行为画像。通过模型训练和数据积累，及时发现恶意程序和异常程序，提供应用程序衔接链的故障关联性分析。

审计规则管理：

目前审计预警规则，支持异地登录告警、密码猜测检测告警等。本期继续扩充审计预警规则库，加强审计风险识别能力。主要包括：任务调度异常监控告警、应用产品持续BUG监控告警。

实时审计：

为达到实时审计的目的，实时抽取、汇聚各类日志。非法操作实施后，可以立即被识别，并主动采取对应的封堵措施、关闭权限等。

区块链审计追踪：

引入区块链技术增强日志记录能力，利用区块链去中心化，不可伪造、销毁等特性，防止非法用户，越权对已经形成的日志文件恶意伪造、篡改、删除关键核心操作日志。真实追踪用户/系统的全部操作信息。

数据隔离与存储

隔离安全性：采用数据隔离手段，数据中心建立独立数据库，将数据中心数据与其他平台数据进行物理隔离。

存储保密性：智慧邮轮港数据存储区规划于私有云网内，该区域有大量的用于支撑业务的应用系统、应用服务器以及需要用到的信息资源库。存储加密基于密码管理局批准公安机关专用的密码算法实现数据加密，保证信息在存储、传输过程中的机密性和完整性。

数据可用性与隐私性管理

数据可用性：支持用户对数据的各项操作，如上传、修改、删除、查找进行控制。

数据隐私性：建立数据隐私保护策略，未经授权不得获取和查看数据中心各类数据，除监管部门审计外，对其他第三方引用数据进行授权控制。

数据加密传输

数据传输加密选择在链路层、网络层、传输层等层面实现，采用网络传输加密技术保证网络传输数据信息的机密性、完整性、可用性。对于管理信息加密传输，采用SSH、SSL等方式为维护管理提供数据加密通道，保障维护管理信息安全。用户数据加密传输，可采用IPSecVPN、SSL等VPN技术提高用户数据的网络传输安全性。

数据访问控制

在数据的访问控制方面，采用基于身份认证的权限控制方式，进行实时的身份监控、权限认证和证书检查，防止用户间的非法越权访问。采用有数据访问需求时才显性打开对应的端口或开启相关访问策略。虚拟应用环境下，设置虚拟环境下的逻辑边界安全访问控制策略，通过加载[虚拟防火墙](https://www.baidu.com/s?wd=%E8%99%9A%E6%8B%9F%E9%98%B2%E7%81%AB%E5%A2%99&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \t "_blank)等方式实现虚拟机间、虚拟机组内部精细化的数据访问控制策略。

数据发布接口与获取接口

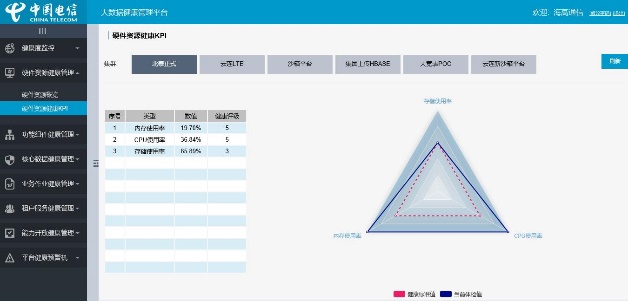
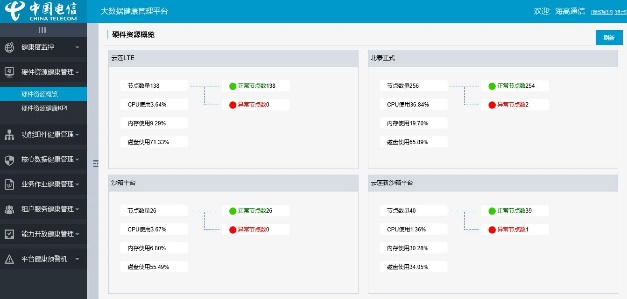
为各级领导决策、内部信息共享和信息发布等不同层次的信息服务需求，统一数据提供服务通过数据服务接口API为各系统及外部系统提供数据访问服务。数据共享服务接口采用统一标准，统一规范；数据获取接口基于第三方系统提供的接口进行开发。

统一数据提供服务一方面集成第三方提供的各类信息承载、专题数据分析及挖掘等服务组件，以统一接口方式提供给各类应用，服务组件可根据业务发展需要而动态调整。未来统一数据提供服务同时也提供开发接口服务，支持开发者或应用厂商调用平台提供的服务和自己的业务应用进行集成。

**统一监控系统**

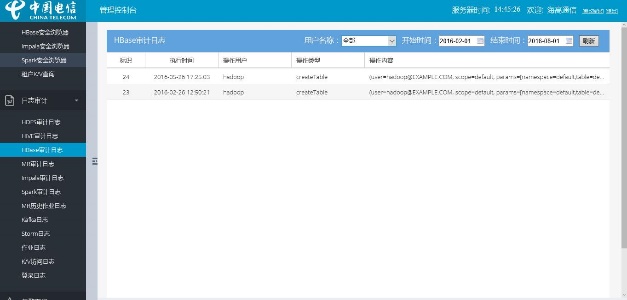
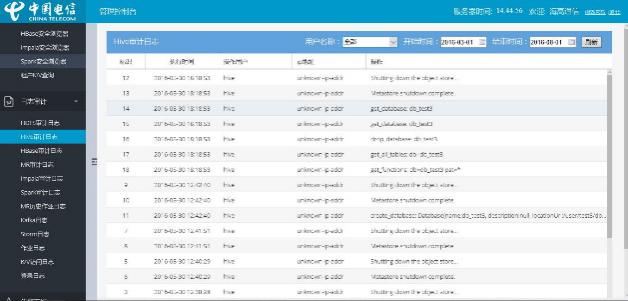
统一监控系统提供两种运行监控方式，包括：Web图形化界面监控方式、命令行监控方式。监控平台对以下各类资源进行监控，具体包括：

* 硬件资源监控：包括平台所属集群、节点数量（包括正常数量和异常数量）、CPU利用率及峰值、内存容量及利用率、物理机磁盘利用率、磁盘IO吞吐量及利用率、物理网络流量统计等。硬件资源监控概览图例如下所示：

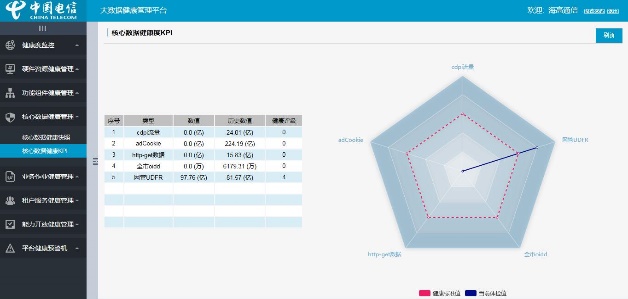


* 支撑组件监控：包含HDFS分散存储组件、MR并行计算组件、HIVE数据库、HBASE数据库、Impala交互式查询组件、Spark内存并行计算组件、Kfaka数据分发组件、Storm实时计算组件等。支撑组件监控概览图例如下所示：

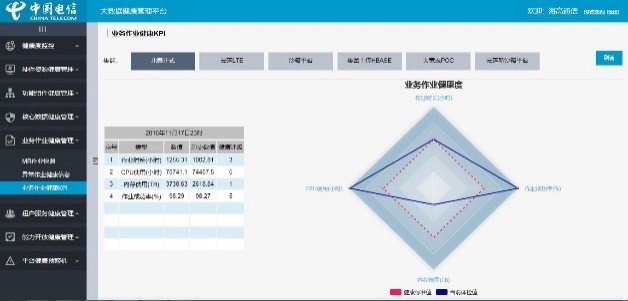
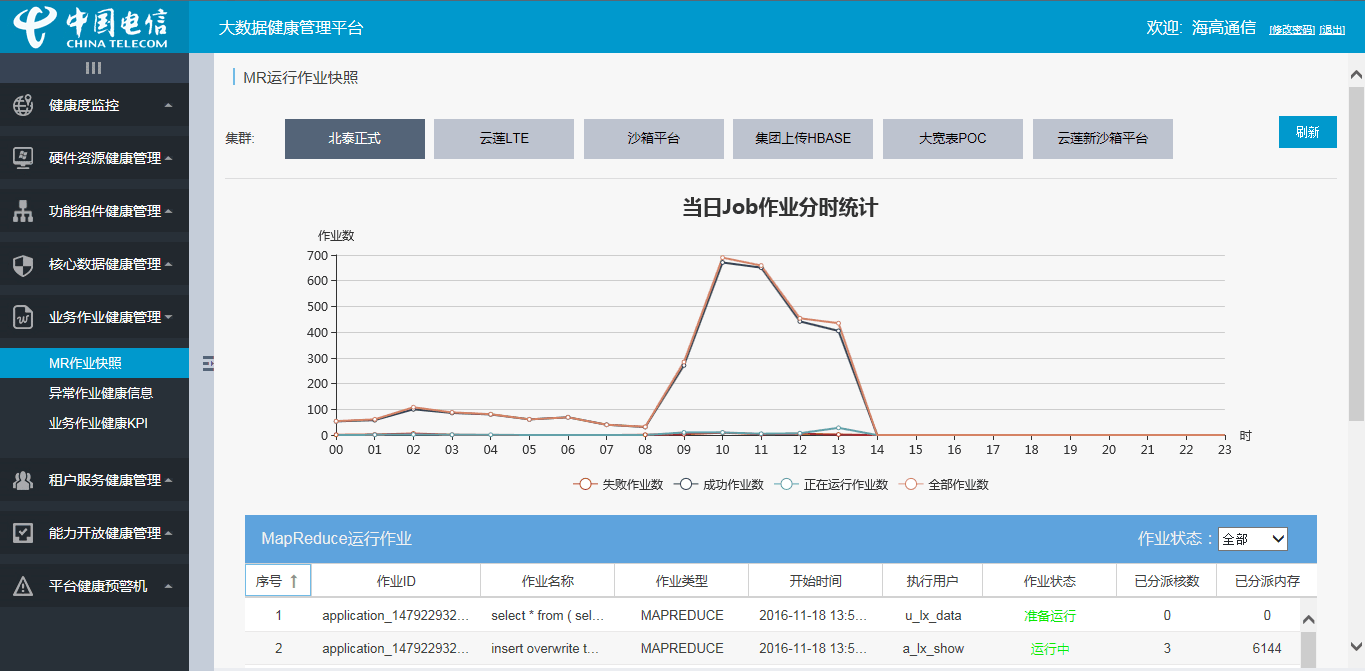




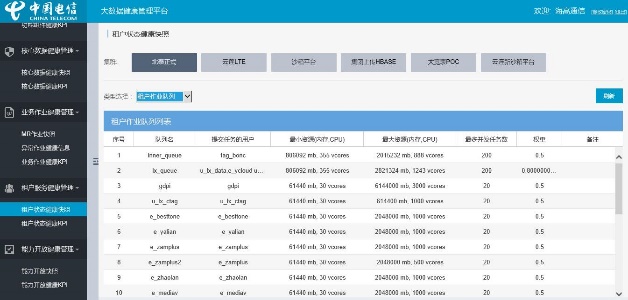
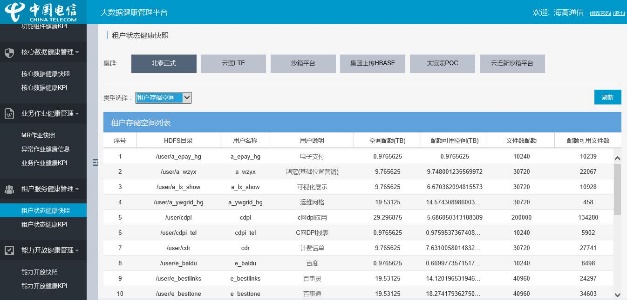
* 核心数据监控：监测平台核心数据的采集情况，发现核心数据潜在的隐患，核心数据监测包括：cdpi流量、adCookie、Http-get数据、全市OIDD、网管UDFR。 核心数据监控概览图例如下所示：



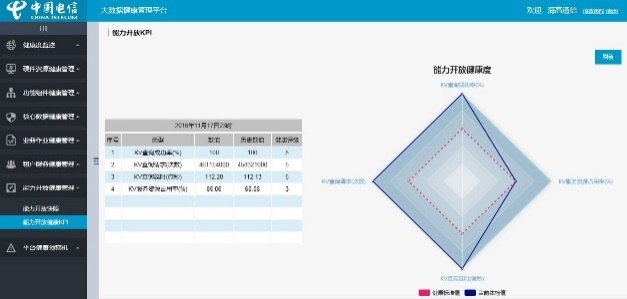
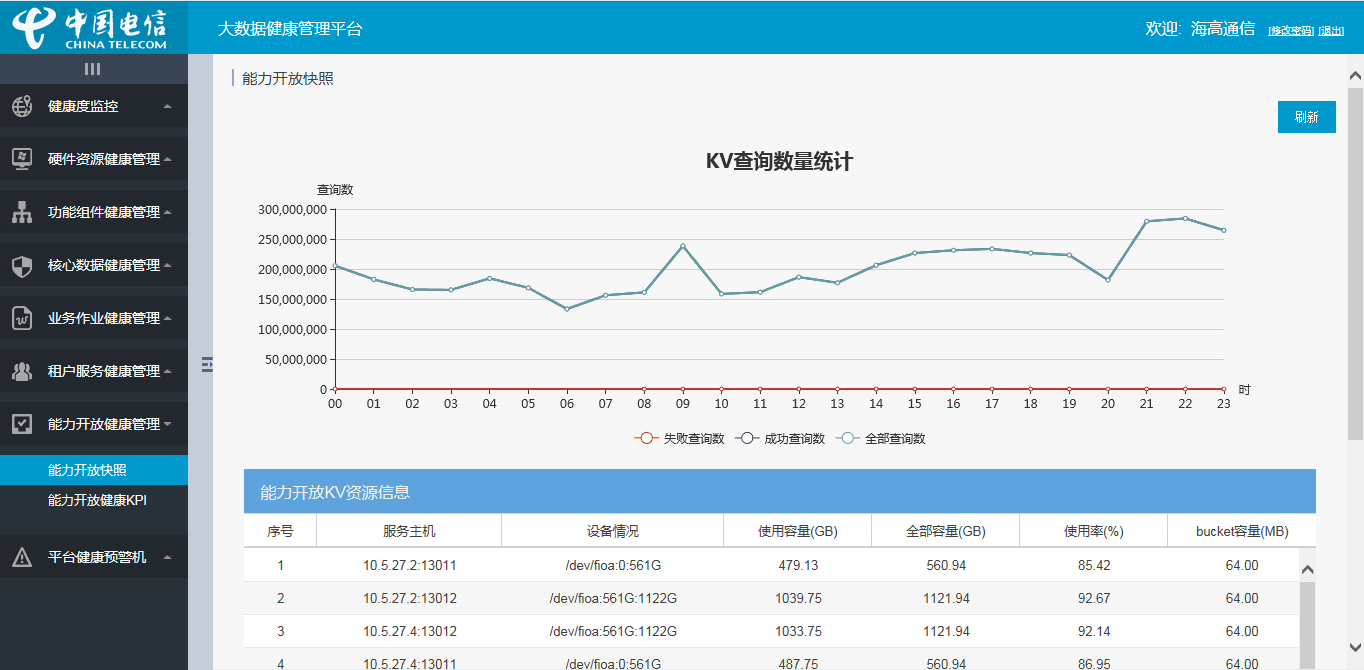
* 业务作业监控：平台可以监控任意状态下的作业运行情况，状态包括：全部、运行、结束、失败、中止。作业运行健康信息包括：作业ID、作业名称、开始时间、执行用户、作业状态、已分配CPU核数、已分配内存。业务作业监控概览图例如下所示：



* 租户服务监控：租户状态健康监测包含租户存储空间使用情况和租户作业队列执行情况。租户存储空间健康信息包括：HDFS目录、用户名称、用户说明、占用空间、空间配额、配额可用空间、文件数配额、配额可用文件数等。租户作业队列健康信息包括：队列名、提交任务的用户、最小资源（内存，CPU）、最大资源（内存，CPU）、最多同时可以执行的任务数、权重等。租户服务监控概览图例如下所示：



* 能力开放监控：详细记录外部用户的查询量，并生成能力开放统计值，用于监测能力开放业务是否运行正常。能力开放监测信息包括：KV查询数量统计、服务主机、设备情况、使用容量、全部容量、使用率、Bucket容量。 能力开放概览图例如下所示：



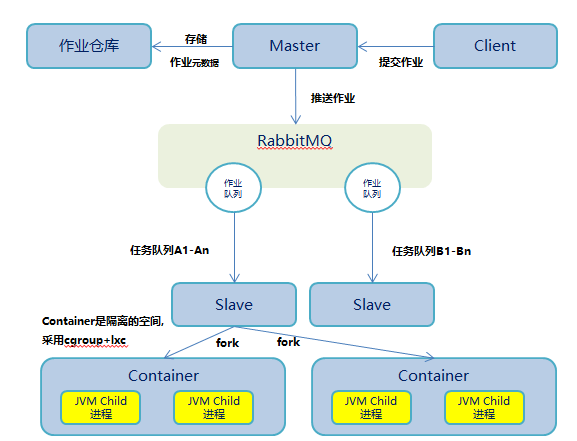
**高性能KV查询服务**

大部分应用场景都有实时性方面的需求，而要实现高性能查询服务，离不开支持高并发、低延时的K-V数据库。键值（Key-Value）存储数据库，这是一种NoSQL（非关系型数据库）模型，其数据按照键值对的形式进行组织、索引和存储。KV存储非常适合不涉及过多数据关系业务关系的业务数据，同时能有效减少读写磁盘的次数，比SQL数据库存储拥有更好的读写性能。

基于键值对数据存储模型的数据库管理系统（DBMS），数据存储元组没有统一的字段格式要求，数据的存储与访问均是以键值对作为标识进行的。基于内存数据库实现Key-Value键值对查询接口，适用于提供主键ID查询数据结果的情况。

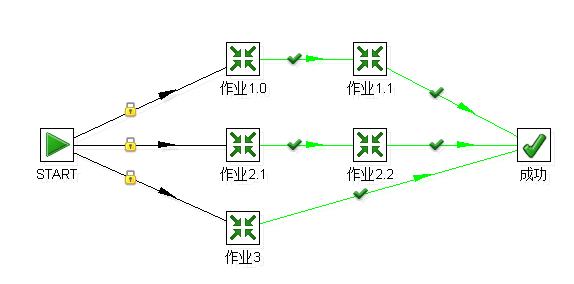
**任务调度**

利用好的资源管理与调度的策略和方法，使得整个集群的大量资源在能够实现更高资源利用率的同时加快所有计算任务的整体完成速度，这就是集群资源管理与调度系统的核心目标。



实现数据采集任务的统一调度功能，包括作业发布管理，作业审核管理，作业运行管理，作业进程管理等，通过可视化的界面，用户可以方便的实现采集作业的配置、发布管理等操作。

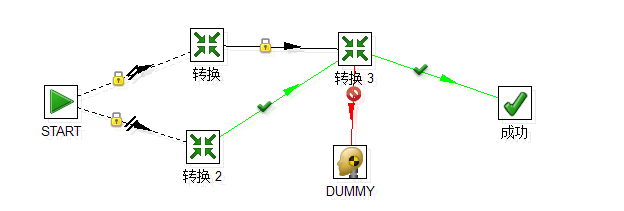
* **作业串行**



在有向无环的任务调度中可以设置各个节点之间的关系，可以是并行、串行，在串行关系中可以设置后续节点的执行条件，是在前置节点成功后执行，还是在前置节点失败后执行，亦或是无论前置节点成功还是失败都必须执行。

而对于任务而言，从START 到成功就是一个总的任务，而其中的作业1.0、作业1.1、作业2.1、作业2.2、作业３就是总任务下的每个分任务，每个分任务可以配置不同的任务属性和运行参数，执行不同的作业逻辑，例如：每个分作业分别汇集不同的数据源，这样就能按照设置条件自动按序执行任务。

* **作业并行**



对于需要并发的任务，调度系统也能以上图方式并发执行多任务，图中START到转换和转换２直接使用了并行设置，转换和转换2同时开始，转换到转换3之间是必定进行，无论转换是否成功，都会执行转换3，而转换2到转换3是只有当转换2成功时，才可以执行转换3，当转换3失败时，执行DUMMY，成功时执行成功节点。

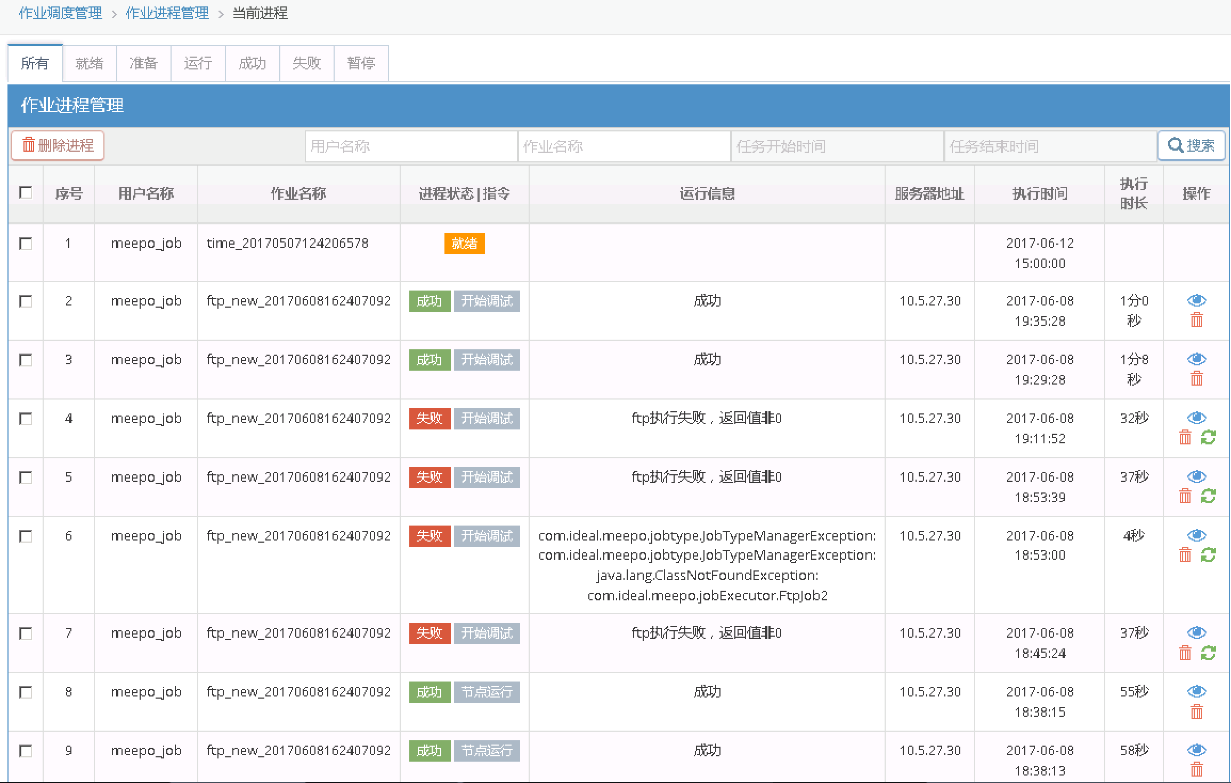
对于高负载的任务，调度系统在接受任务后，将自动向Yarn集群请求资源，Yarn集群分配资源，在资源充足的集群节点上运行该任务。

* **任务监控**

对于异常的作业，调度系统将采取如下的处理方式来对调度异常进行处理：

运行日志界面

系统维护人员可以通过调度管理工具的运行日志界面中看到三级日志界面，第一级是作业列表，通过失败任务选项卡筛选出所有执行失败的作业，第二级是作业执行的步骤，出错的步骤将以红色字体的方式提醒维护人员该作业已经在那个执行步骤出现了错误，第三级日志界面显示该异常的详细信息，以帮助维护人员定位错误原因。

****

1. 自定义异常数据记录

调度系统可以预先在系统中定义异常输出类型，例如身份证15或18位，在执行调度任务中，不满足15位或18位的数据将被判定为异常数据输出，并对异常的数据记录进行统计，以获得该汇聚任务的具体执行情况(成功数、异常数)等信息。

1. 作业异常后的恢复

　　对于出现异常的调度作业，调度系统将把通过“提醒”（短信、邮件）方式通知系统维护人员，在由系统维护人员处理完异常的作业后，在调度管理工具界面中的恢复功能，从而将任务回退到需要重新调度的步骤，进行恢复调度。

应用分析层

智慧邮轮港信息化将逐步深入，大量信息系统在邮轮港中广泛应用，物联网、云计算等技术与邮轮港经营生产紧密结合，港区设备运行、港区生产运营、游客服务、邮轮服务、边检海关协同办公等积累了大量的数据，包括邮轮票务销售数据、游客服务数据、游客行为数据、港区运营数据等，港区经营生产的各个阶段都可以被记录下来，游客服务、邮轮服务的各个环节也被记录下来，这些数据隐藏着大量的有价值的规律和信息，是智慧邮轮港的重要资产。

传统的数据报表、查询等简单分析手段不足以支撑邮轮港对于深层次信息的挖掘需求，数据融合、数据分析、数据挖掘等技术不断发展，漏斗分析、事件分析、行为分析、留存分析、属性分析等模型不断完善，神经网络、决策树、关联规则等挖掘算法不断成熟，基于Hadoop的分布式存储技术以及基于Storm、Spark、MapReduce等分布式计算技术迅猛发展，这些都为国际邮轮产业上海数据中心的数据分析处理及分析建模提供了坚实的技术支撑。

智慧邮轮港数据分析建模主要从以下两个方面进行考虑：

* 以需求为出发点，深化业务问题解决

建模结合智慧邮轮港业务发展目标，梳理港区经营生产过程中的瓶颈问题，以游客、邮轮为中心，以港区生产运营为导向，展开需求调研，清理业务流程和业务数据，研究数据来源、采集通道和映射关系，深入梳理数据基础，不断推进业务问题的解决并构建合理的业务模型。

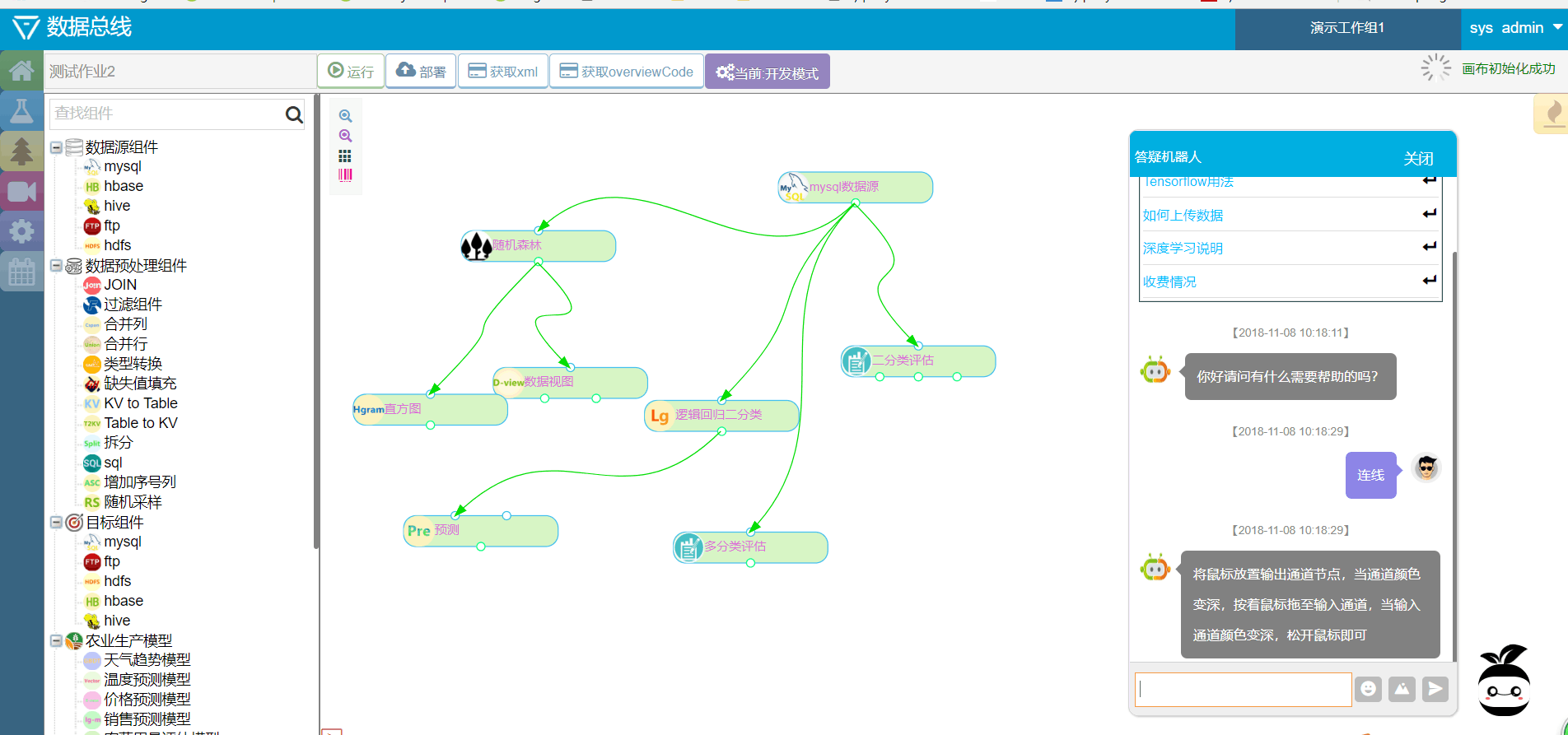
* 加大数据积累力度，夯实数据体系基础

系统建模围绕邮轮港智慧通关、智慧服务、智慧交通、智慧管理，提高数据采集效率，丰富数据积累力度；完善主数据、主题数据、数据建模等标准规范，构建包含业务、指标、报表等的数据体系，夯实数据体系基础

可视化建模工作台

大数据的数据分析与挖掘组件众多，使用者在使用大数据平台时不可避免的会经常性的调整所需的组件以及连接方式。因此，为了减少使用者的学习成本，以及更好的发挥大数据数据挖掘的实际作用，数据分析与挖掘模块提供可视化交换式的开发模式，依托大数据集群强大的计算能力，进一步提升租户对于海量数据的建模挖掘能力，提高开发效率。

* 交互式开发界面支持：全流程、图形化、组件拖拽式的交互式开发能力，快速便捷
* 完善的算法调试支持：提供小批量数据运行、断点调试、模型参数调试等
* 多租户管理支持：底层融合多租户权限体系，保证数据资源、存储资源以及计算资源的安全管控
* 数据挖掘建模支持：常用机器学习算法（随机森林，逻辑回归二分类，二分类评估，预测），统计分析（数据视图，数据直方图），模型管理及共享等



用户可以通过在“画布”中，自主选择需要的组件和其连接方式，从而开发自己的模型；对于已经开发好的模型，用户也可以再次进行编辑，随时调整组件的参数、布局等信息。系统提供了多种多样，适用于多数场景的数据分析与挖掘组件，以应对不同使用者的不同使用需求，数据分析与挖掘工具功能强大而且方便易用。

**1、分布式数据总线引擎**

* DSL

数据总线引擎中的DSL模块为总线节点对外提供可视化、组件式的拖拽界面提供了语义转换功能，将前端界面定义的组件、节点、连线、属性面板“翻译成”Spark Driver可以识别的语言。

* 交互式模式

为了提供用户一个敏捷性数据开发的环境（数据Schema动态解析、数据依赖执行、数据动态查询），总线引擎底层采用Spark Driver交互式程序，将DSL翻译后的语言输入Spark Driver，并获取响应结果，然后传递给总线引擎，最大可能的给用户的数据数据处理开发流程带来简便性的功能体验。

* 多租户资源管控

总线底层的作业交由Spark集群去完成，可以无缝集成现有的多租户权限、资源隔离等特性。

**2、组件配置项**

组件配置项包括：字段选择、参数设置、执行调优等，依据不同类别的组件，会有相应的不同的配置项条目。绝大多数的组件都将提供适用于自身的配置选项，在易用性与自由度之间形成平衡，帮助使用者在不失便捷的条件下最大限度的提升模型效果。

**3、组件功能**

组件功能项包括：单步执行、日志查看等，依据不同类别的组件，会有相应的不同的功能项条目。由于数据分析与挖掘的完整步骤通常会较为复杂，因此，为了方便使用者了解每一步骤的情况，该平台中的组件均提供了单步执行，以方便使用者及时进行调试验证。同时，组件也提供了日志查看功能，当执行异常时可利用日志进行问题排查，便于及时解决问题。

ETL数据加工

ETL用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。ETL是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去，成为联机分析处理、数据挖掘的基础。

* 数据源组件
* ftp数据源组件
* hdfs数据源组件
* kudu数据源组件
* oracle数据源组件
* mysql数据源组件
* gbase数据源组件
* 数据预处理组件
* Join数据预处理组件
* 过滤数据预处理组件
* 类型转换数据预处理组件
* 缺失值填充数据预处理组件
* SQL数据预处理组件
* 增加序号列数据预处理组件
* 数据目标组件
* ftp数据目标组件
* hdfs数据目标组件
* kudu数据目标组件
* oracle数据目标组件
* mysql数据目标组件
* gbase数据目标组件

**数据开放能力**

平台基于开放的数据服务体系，平台按照不同的服务开放场景，聚合数据服务能力的封装，数据服务与业务应用解耦合，完全可满足平台业务应用（3 - 5年以上）的持续扩展。平台已实现典型服务开放场景的数据能力封装，与具体业务应用解耦合，开发者也可在此基础上增加定义数据开放服务。目前平台已具备的典型服务开放场景数据能力封装包括：

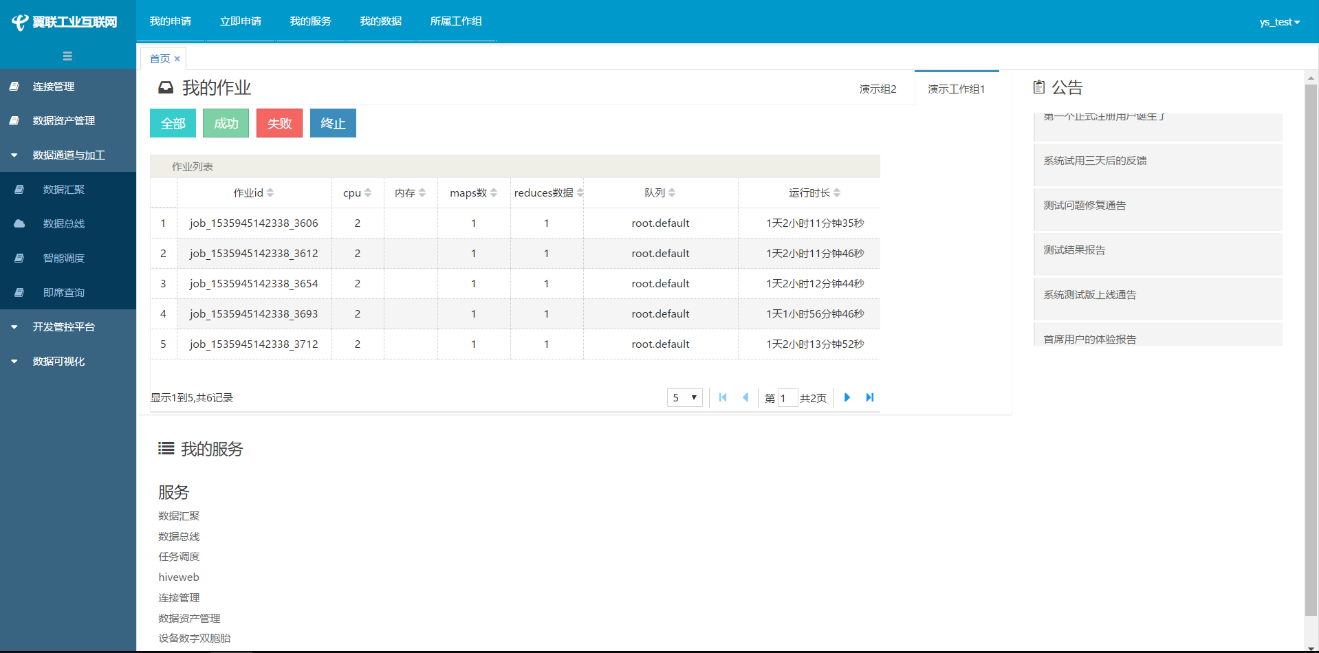
* 应用接入能力封装：实现第三方应用接入的统一标准化，定义了各类型应用接入的接口类型，如RestFul、Web Service、FTP、Socket等协议。平台提供界面化的服务基础信息、权限信息配置功能
* 查询分析能力封装：实现单次用户信息查询、批量用户信息查询的内部处理逻辑，将系统内部的数据根据查询请求组装、规整；平台支持界面模式与接口模式两种方式供外部数据需求方进行数据查询
* 数据挖掘能力封装：平台通过对挖掘工具的封装，提供统一的标准，对外提供数据挖掘能力。数据使用方通过系统提供的界面进行数据挖掘操作。目前主流的数据挖掘能力实现方案包括：MLlib，Weka，Mahout，本平台考虑更多需要对外提供挖掘能力API和与java语言、spark的集成成熟度，推荐选择MLlib挖掘组件

**能力开放工作台**

平台提供一站式集成化能力开放工作台，包含以下特征：

1. 针对服务对象的专属平台：根据用户角色的划分，可由用户自行订制，包括使用者，开发者，管理者，在工作台中提供有针对性的与之工作相关数据视图、图表、分析报告等。
2. 高集成度的工作平台：向用户提供一个一站式的工作平台。将用户所有的工作进行分类整理后，在工作台中提供统一的功能入口，满足用户所有的日常工作操作。
3. 符合用户操作习惯的个性平台：在工作台的标准定义下，用户可以根据操作习惯进行自定义配置。

工作台示意图，如下：



* 平台普通用户工作台的功能：
* 应用统一入口

提供用户所有订阅应用的统一入口。

* 服务申请

提供用户申请服务的功能

* 数据申请

提供用户查看本租户下所有表的数据字典及申请权限的功能

* 任务查看

提供用户任务运行情况的监控

* 租户管理员工作台的功能：
* 应用统一入口

提供用户所有订阅应用的统一入口。

* 资源申请

提供用户申请资源的功能

* 服务申请

提供用户申请服务的功能

* 数据申请

提供用户查看本租户下所有表的数据字典及申请权限的功能

* 工作组新建及分配资源人员

新建工作组，为工作组分配资源，及人员调配

* 数据资源管理

提供用户建立数据表，对数据结构的维护管理

* 管理者工作台的功能：
* 子系统管理

提供用户便捷的子系统接入，注销，使平台具备松耦合，热插拔的架构特点

* 租户管理

提供用户新建租户及对租户的资源，工作组的管理

* 用户管理

提供用户对平台用户的管理，包括账号，基础信息等

* 数据资源管理

提供用户建立数据表，对数据结构的维护管理，以及对工作组赋予或撤销表的权限

* 统一监控

系统重要指标实时动态显示，为管理者提供方便的监控手段，掌握系统实时的健康状态。

**主题分析应用**

票务经营分析

* **票务交易量分析**

(1)交易量分析

对平台上因船票交易产生的交易总量进行统计，分为平台交易额、平台交易票数及平台交易订单数量三个维度。

对平台上邮轮公司与代理商之间的交易产生的交易量进行统计，分为代理交易额、代理交易票数及代理交易订单数量三个维度。

对平台上代理商与代理商之间的交易产生的交易量进行统计，分为同业交易额、同业交易票数及同业交易订单数量三个维度。

以日为时间单位计算，计算交易量，可选择任意日期的交易总额、交易票数及交易订单数量，点击仪表盘可查看今日每小时的交易量。

以月为时间单位计算，计算交易量，可选择任意月份的交易总额、交易票数及交易订单数量，点击仪表盘可查看本月每日的交易量，如有需要可将时间周期进一步细化。

以年为时间单位计算，计算交易量，可选择任意年的交易总额、交易票数及交易订单数量，点击仪表盘可查看本年每月的交易量，如有需要可将时间周期进一步细化。

（2）交易偏差分析

统计船票交易平均量，查询日交易量，与平均量进行比较，算出偏差率。识别正负偏差率大的交易日，从航班数量、产品种类、价位等多角度分析偏差原因以数据指导船公司产品运营、指引代理商购物方向。

（3）交易变化量分析

从时间与用户两个维度对船票交易量的变化情况进行统计分析，主要包括一下几个方面的分析：

整体变化：对比各个时期的平台整体船票交易量，分析变化情况。

船公司：对比各个时期的单个船公司票交易量，分析变化情况。

代理商：对比各个时期的单个代理商交易量，分析变化情况。

同一产品：对比各个时期的同一产品交易量，分析变化情况。

* **库存分析**

（1）整体库存

以上架的总航班为主维度，多角度分析船票库存，主要分析内容如下：

统计所有入驻平台的船公司、分销商的上架的总航班的库存量，分为船公司库存量、分销商库存量、航班实际库存量

船公司库存量：统计所有上架的总航班，船公司手里面的船票库存量。

分销商库存量：统计所有上架的总航班，分销商手里面的船票库存量。

航班实际库存量：统计所有上架的总航班，实际的库存量。

（2）产品库存

以上架的同类型产品为主维度，多角度分析船票库存，主要分析内容如下：

统计所有入驻平台的船公司、分销商上架的同类型产品的库存总量，分为船公司库存量、分销商库存量、航班实际库存量

船公司库存量：统计所有上架的同类型产品，船公司手里面的船票库存量。

分销商库存量：统计所有上架的同类型产品，分销商手里面的船票库存量。

产品实际库存量：统计所有上架的同类型产品，实际的库存量。

为船公司未来产品运营决策提供数据支撑。

（3）航班库存

以单个上架的航班为主维度，多角度分析船票库存，主要分析内容如下：

统计所有入驻平台的船公司、分销商单个上架的航班的库存总量，分为船公司库存量、分销商库存量、航班实际库存量

船公司库存量：统计单个上架的航班，船公司手里面的船票库存量。

分销商库存量：统计单个上架的航班，分销商手里面的船票库存量。

产品实际库存量：统计单个上架的航班，实际的库存量。

* **购买力分析**

（1）购买力排名

以交易额、票量、交易占比等维度指标来评价代理商的购买能力，将代理商按照购买力进行排名，并可自定义排名时间周期，默认快捷查询周期为日、月、季度、年。

（2）购买力管理

对每个船公司下面的代理分销商的购买力进行管理，购买力的评价指标分别以交易额、票量、交易占比、购入产品类型分布等，可自定统计时间周期，默认快捷查询周期为日、月、季度、年。船公司可以直观看出与其发生交易的代理商的各项购买指标，为代理折扣、佣金比等确定提供数据支撑。为船公司与代理商的合作提供决策支撑。

（3）购买力趋势

对单个代理商的交易额、交易票量、购入产品类型分布进行统计分析，分别以日、月、季度和年为单位，行程购买力趋势图。可用来评价代理分销商的交易潜力。

* **吸引力分析**

分析产品、船公司对分销商的吸引力，也能侧面反应出产品对游客的吸引力，为船公司产品运营决策提供数据支撑。分析不同船公司同类型产品、同一公司不同类型产品、多有上架产品、不同船公司对分销商的吸引力。

（1）不同公司同类产品

以船票分销速度、库存剩余率及产品页面浏览量等指标为支撑，分析不同船公司同类型产品，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。

结合“价格、房型”等影响因素，如不同“价格”产品，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看价格在在吸引力方面的影响。

通过一定数据的积累，确定价格、船上项目的影响系数，将船票分销速度、库存剩余率及产品页面浏览量与指标系统通过算法，得出产品吸引力指数。

（2）同公司不同产品

以船票分销速度、库存剩余率及产品页面浏览量等指标为支撑，分为价格、船上项目等细化角度分析同一船公司不同类型产品对分销商的吸引力，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。

不同“价格”产品，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看价格在在吸引力方面的影响。

不同产品的船上项目，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看船上项目在吸引力方面的影响。

通过一定数据的积累，确定价格、船上项目的影响系数，将船票分销速度、库存剩余率及产品页面浏览量与指标系统通过算法，得出产品吸引力指数。

（3）所有产品

以船票分销速度、库存剩余率及产品页面浏览量等指标为支撑，分为价格、船上项目、房型等细化角度分析所有产品对分销商的吸引力，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。

不同“价格”产品，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看价格在在吸引力方面的影响。

不同产品的船上项目，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看船上项目在吸引力方面的影响。

不同产品的房型，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看船上项目在吸引力方面的影响。

通过一定数据的积累，确定价格、船上项目、房型的影响系数，将船票分销速度、库存剩余率及产品页面浏览量与指标系统通过算法，得出产品吸引力指数。

（4）品牌吸引力

以船票分销速度、库存剩余率及产品页面浏览量等指标为支撑，分为价格、船上项目、房型等细化角度分析不同船公司对分销商的吸引力，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。

不同“价格”产品，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看价格在在吸引力方面的影响。

不同产品的船上项目，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看船上项目在吸引力方面的影响。

不同产品的房型，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。以结果为导向，看船上项目在吸引力方面的影响。

通过一定数据的积累，确定价格、船上项目、房型的影响系数，将船票分销速度、库存剩余率及产品页面浏览量与指标系统通过算法，得出产品吸引力指数。

以船票分销速度、库存剩余率及产品页面浏览量等指标为支撑，分析船公司品牌对分销商的吸引力，分销速度越快、库存剩余率越小、网页浏览量越大的吸引力越强。

* **单产品分析**

(1)产品价位分析

将正在分销的产品及历史售卖的产品独立划分，为每个产品建立价格资料库，以线状图、堆积线图为主要前端样式组件，直观展示单产品价格变化趋势。

产品价位分析实现从时间周期、时间刻度高度可定义性，样式类型高度可配置，数据价值与美观高度最大化。

系统用户可以分别从时间主维度、价格主维度、分销主体主维度去查看产品的价格变化。

时间维度：若产品正在交易平台进行分销，系统默认展示本次分销期间的价格变化趋势，若产品暂未架上分销，系统默认展示最近一次分销的价格变化。展示周期为分销开始到分销结束，时间刻度为日。可调整查询时间周期，提供两种时间周期模式，一种是常规的时间周期，可以按日、月、季度及年来查询，一种是以每次分销期为时间周期。

价格维度：将查询产品的分销价位，按价格段化为档位，查看各个价格段发生的分销时间及出游时间。系统默认百元为分档单位，支持用户自定义配置分档标准。

分销主体：按船公司与代理商两个销售主体维度，统计产品价格变化，结合价格与时间维度，查看变化趋势。

(2)偏好分析

将正在分销的产品及历史售卖的产品独立划分，为每个产品建立代理商库，以排名列表、3D柱状图为主要前端样式组件，直观展示分销商对此产品的偏好程度。

统计每个产品在不同时期、不同定价情况下，购入此产品的分销商。可自定义统计时间。并可通过统一化归、去重等技术，分析出长期对此产品偏好的分销商，利用戴布拉图、蜘蛛网图，直观看出各个分销商对此产品分销情况的影响力。为船公司定价及选择合作方提供数据支撑。

(3)对比分析

利用图表联动技术，结合多样化前端组件，实现可视化图型与具体数据直观对比。按需对比分析不同影响因子，可对比分析内容举例如下：

船公司-代理商价位对比分析：分析对比同一产品船公司与代理商的分销价位情况及分销情况。

不同时期对比分析：按不同时间周期，对比产品价格情况、分销情况。如可利用堆积点图、堆积线图，对比不同月份产品的价格，下面配合数据表格。可直观看出不同月份的价位的差异情况，及具体价格数据。

价格-分销分析：统计不同价位的分销情况进行对比分析，利用帕累托图，直观展示分析结果。当用户需要对帕累托图的线进行大小或颜色等的调整时需要通过该线的右键菜单中相对应的属性或格式进行设定。

自定义对比指标：系统初始化默认价格、时间、船公司、代理商、船票分销情况，随着平台运营一段时间，一定数据的积累及对比，可自定义增加对比指标。

* **消费者分析**

通过获取游客的信息，分析游客消费特征，从地区、年龄特征出行方式度分析，用数据指导船公司及分销商运营管理。

(1)地域特征

获取游客所在地信息，以省、市为单位划分游客，进行多角度分析，如以下分析方向：

产品分布：分析各个类型产品的游客省市分布情况，可以在产品推新的时候，重点投向游客分布多的省市。

价位分布：各个省市购买船票的价位分布，可以在定价的时候，重点倾斜游客分布多的省市。

(2)年龄特征

获取游客的年龄，以各年龄段划分游客，进行多角度分析，如以下分析方向：

产品分布：分析各个类型产品的年段分布情况，可以在产品优化升级的时候，重点关注该产品购买力强的年龄段的需求。

价位分布：各个年龄段购买船票的价位分布，可以在定价的时候，重点倾斜该产品购买力强的年龄段

(3)出游方式分析

获取游客订单数据，以订单为单位，分析游客出行方式，统计闺蜜游、家庭游、情侣游、同学游、毕业游的分布情况。

产品分布：分析各个类型产品的游客出行方式，可以在产品优化升级的时候，重点关注该产品购买力强的出行方式游客的需求。

价位分布：各中出行方式的买船票的价位分布，可以在定价的时候，重倾斜该产品购买力强的出行方式的游客。

* **平台运营分析**

对平台的运营情况进行分析，包括入驻的船公司数量、代理商数量，平台浏览量PV等指标。

(1)入驻情况分析

统计入驻的船公司、代理商的数量，并形成变化趋势图。可以查询入驻数量的变化趋势。并计算入驻船公司、代理商数量的增量率，指标大小反映出票务交易服务平台的在票务交易活动中的影响力。

(2)平台使用分析

计算浏览量PV与方可访客量UV两个指标，反映了船公司、代理商是否使用邮轮票务交易服务平台。

访客数UV指一天之内到底有多少船公司、代理商访问了邮轮票务交易服务平台你的网站。

浏览量PV船公司、代理商访问票务平台时每打开一个页面，就记为1个PV。同一个页面被访问多次，浏览量也会累积。一个页面的浏览量越高，说明这个用实用。

(3)实用性分析

计算平均访问时长和平均访问页数，评价邮轮票务交易服务平台的实用性。

平均访问时长是用户访问网站的平均停留时间。平均访问时长=总访问时长/访问次数。如果船公司、分销商不喜欢网站的内容，可能稍微看一眼就关闭网页了，那么平均访问时长就很短;如果用户对网站的内容很感兴趣，在票务平台停留了很长时间，平均访问时长就很长。

平均访问页数是用户访问网站的平均浏览页数。平均访问页数=浏览量/访问次数。平均访问页数很少，说明访客进入票务平台后访问少数几个页面就离开了

* **分析决策报告**

(1)决策支撑看板

决策支撑看板实现从样式组件、展示数据、展示规则、关联规则、模型优化的自定义配置，最大程度贴合各种角色、岗位人员的分析决策需求。实现数据价值的最大化。

通过挖掘交易情况分析、库存分析、购买力分析、吸引力分析、单产品分析、消费者分析及平台运营分析几个维度的分析结果的关联关系，制定关联规则。通过多样化组件技术进行直观的分析结果数据，为管理者决策提供大量数据依据。

系统默认显示多种分析结果数据之间轻关联关系展示，每种分析结果提供多样化组件，可通过配置按需选择。并可自定义选择具体计算展示几种维度的分析结果，自定义展示关联。

通过一定数量数据的积累，挖掘关联数据，建立算法模型，制定关联规则，通过数据的不断校验，算法训练，优化规则。并可自定义算法：支持预置各类常用的数据挖掘算法，包括分类预测、回归预测、聚类分析、时序分析等。支持直接选用预置算法对数据进行分析。各类数据挖掘算法可以进行封装，使用户无需进行代码上的修改。提供友好方便的数据选择和参数调节功能，支持对参数的检验。支持参数调整后，即时展示分析结果，方便调整参数。支持一次选择多种数据挖掘方法对同一组数据进行分析，并一次给出各个方法的分析结果，以便节省逐个试用不同方法的对比时间。可方便、友好增加新的算法。

建立关联规则库，系统用户可按需选择几种维度分析结果之间关联规则，计算出最符合使用者的决策支撑数据。

(2)周期性报告

周期性报告实现从数据、时间、内容类型、报告类型均可配置，生成最大程度贴合各种角色、岗位人员需求的报告。避免冗余数据干扰决策分析，将数据的价值最大化。并利用大数据计算工具实现数据的高速计算，缩短计算响应时间，极速生成报告，提高工作效率。

数据类型可配置：将交易情况分析、库存分析、购买力分析、吸引力分析、单产品分析、消费者分析及平台运营分析内有的原始数据及分析数据组件化，建立原始数据组件库、分析数据组件库及关联数据组件库，系统用户按需选择数据类型，系统支撑组合数据类型。

时间可配置：系统实现时间周期、刻度可选择的最大自由化，系统用户选择日、周、旬、月、季度及年的时间周期，时间刻度可细化到小时、分钟。时间周期与时间刻度自由组合，可配置成日-小时、日-分钟、月-日-小时、月-日-分种等几十种组合数据。

内容类型可配置：将交易情况分析、库存分析、购买力分析、吸引力分析、单产品分析、消费者分析及平台运营分析几个维度的分析内容，分类成模块化系统用户自行选择内容类型，可复合选择。

报告类型可配置：系统建立报告类型模型，分为内容形式类型及报告格式类型。系统提供数据表格、多样化图型的内容形式类型，如仪表盘、帕累托等。默认提供excel、PDF、word、txt、图片等传统的报告格式类型，excel、word支持兼容模式、图片支持PNG、JPG、GIS。

报告用途可配置：通过四个配置，生成最终使用者所需内容报告，并可进行二次筛选、淘汰数据。可加入报告背景字段，logo等信息，系统支持在线上传、编辑报告名称、封面，生成更为专业的数据支撑报，系统支持在线导出、在线打印。

游客服务画像分析

游客画像是建立在一系列真实数据之上的针对游客的用户模型，即根据用户的属性及行为特征，抽象出相应的标签，拟合而成的虚拟的形象，主要包含基本属性、社会属性、行为属性及心理属性。需要注意的是，游客画像是将一类有共同特征的游客聚类分析后得出的，因而并非针对某个具像的特定个人。

根据大数据针对游客的属性及行为建模，统计分析并预测游客行为模式。

* **标签取数**

标签和游客画像的数据整合中心，可支撑业务部门的日常取数需求，也可作为其他系统的数据支撑来源。

* **营销推荐**

精准营销推荐是标签和游客画像重要的数据服务功能之一，通过不同的标签组合，可以圈定一批目标用户进行精准营销，数据可推送至短信平台、促销平台等进行短信、邮件、赠品的下发。

* **多维分析**

基于标签的游客多维分析是另一项非常重要的数据服务功能，标签可以更好的连接数据和业务人员，降低数据获取和操作的难度，赋能业务人员参与到数据分析之中。

游客服务满意度分析

游客满意度分析是涉及游客消费行为和景区服务管理的重要指标。游客满意度包括总体满意度和单项满意度，其测度涉及指标体系和测度模型，其中指标体系的设定建立在对游客满意影响因素及其特征分析的基础上。由于影响游客满意因素的复杂性，游客满意度是多维度、动态和难测度的。

* **游客满意度的指标体系**

游客满意度 作为游客满意的定量表述，是衡量旅游景区经营绩 效（经济和社会效益）的综合性指标，是旅游相关产 业者考察旅游产业发展情况的标准之一，对于旅游 业的发展有着举足轻重的作用，为旅游业的发展提 供了管理指导，通过建立结构方程模型，考察出游期望对 满意度的影响，其中，对出游期望赋予4个评价因子：形象、过去经历、交流、有形性。

* **游客满意度的测度模型**

结合大数据建模基础，通过建立结构方模型，将重游意向分为短期、中期、长期，并考察了满意度和新奇度分别对不同时间尺度重游意向的影响。满意度对短期重游意向有显著影响，新奇度 对中、长期重游意向有显著影响；在对旅游期望的定义、影响因素及特点研究的基础 上，提出了旅游体验质量的交互模型。顾客满意度指 数理论和旅游业的“食、住、行、游、娱、购”六大要素特点构建了旅游地顾客满意度指数的测评模型，建立了测评标准；在旅游六大要素的基础上加上了社区和人员服务两大要素，构建了旅游目的地服务质量评价模型。

港城协同联动分析

基于大数据建模模型，实现港城数据联动，以港城业务为对象的管理协同、以港区事件联动为重点的可视化管理、以融合通讯为工具手段的协同指挥，通过数据关联融合形成港城关键体征运行监控管理，全面呈现港城运行综合态势，感知港区运行风险和发展趋势，便于领导及各综合部门调整相应处置方法和政策。

港城运营总体预测

实现指标的分析和监测，包括重点指标情况、各区域指标总体评价等，支持向下钻取，支持按不同业务需求进行多级数据分析工作。

* **平台支持对指标总体运营监测、分析。**
* **对区域的运营情况进行评分**

总体运营得分及不达标指标情况，根据优劣情况在予以颜色标注。

* **各区域总体运营排名**

分析所有区域的总体运营情况，并根据评分进行排名。

港城人流监控与分析

* **人群密度分析**

通过建模分析，使用不同颜色的区块叠加来实时描述人群分布、密度和变化趋势。可视化引擎采用Echarts实现，可通过配置灵活调控。

* **区域特点**

根据区域的划分，使每个区域TOP3的标签构成。区域可以是行政区、商务区，也可以是自定义的区域。

* **区域管理**

默认区域为行政区，通过框选形式，对区域进行增删改查的管理。

港城车流监控与分析



* **车辆查询**

根据查询条件可经过卡口、对机动车辆、异常牌照车辆、违法车辆以及红名单车辆进行查询。以给人更直观的感觉、更丰富的图片信息展示，在提供基本过车信息的同时，还能展示路面状况、车辆形状、驾驶员和副驾驶员的面部特征，可切换到以记录的方式进行展示。

* **过车查看**

可以对卡口的实时过车、历史过车进行查看。

* **车辆分析**

可通过多种算法，对过车记录进行数据处理和分析，从而筛选出满足提交的过车信息和车辆信息。例如：行车轨迹分析、落脚点分、同行车辆分析、连续违法分析、频繁过车分析、车流量统计、初次入港分析、车流量分析、行车规律分析、多车同行、夜间行车、车流量预测等。

* **以车搜车**

通过对过车图片进行模型分析，获得分析结果，然后根据分析结果进行结构化搜索，从而快速准确的获得类似过车信息。

输入一张车辆照片，可在海量卡口图片中根据外形特征检索出与其最相似的车辆，然后根据车辆的某一明显特征（如：车灯坏了，凹痕，车内饰物等），通过对输入照片框选感兴趣区域，进行精细检索，检索出与该部分特征最相似的车辆。

* **违章停车监测**

通过对各个路段的监控和车辆过卡口的预测，可以监测和分析出哪些车辆在哪些路段上进行了违章停车，系统对违章停车的车辆给予提醒，让管理人员可以迅速去处理违章停车的车辆，保证港区路段畅通安全。

港区能耗监测与分析

平台对港区的能耗数据实时收集并做监测。

* **能耗区域显示**

对整个虚拟港区建立各个能耗区域，用户可通过选择不同的能耗区域，分层级寻找各个能耗设备组或者设备的状态情况。

* **能耗设备组显示**

通过能耗设备组可现实组系列的设备能耗状态，例如：通过查看二楼大厅空调组的能耗情况，可以清楚了解当前二楼空调设备的能耗状态是否达到要求及可控。

* **单设备能耗情况**

通过查看单个设备的能耗情况，来了解设备的能耗状态及组成。

* **能耗分析及预测**

平台对各个设备的能耗数据收集，可做各类数据分析及预测模型，帮助港区合理使用能耗、节约使用能耗。

* **能耗报警**

通过接入能耗数据，平台可以随时收集、计算、预警能耗有问题的设备，辅助用户做好能耗管理工作。

港区设备监测与分析

港区合并监控主要分为三类应用场景，即设备波动时提醒、数据恶化时预警、风险超标时督办。对应到系统监控手段上，当采集的设备数据异常时，可通过平台的监测手段具体处理各个设备的故障和问题，通过对应不同操作，实现早发现、早介入、早解决。具体如下：

* **提醒通知**

实时干预，当设备数据出现明显波动时，通知相关业务、管理人员。

* **设备预警**

当监控的设备指标开始持续恶化，平台预警标记该设备。

* **故障处理**

当设备出现问题，协助用户处理督办该设备故障。

数据统计与报表

用户可以对不同的业务需要进行统计，报表可按时间、区域、指标自由组合成多种方式，供用户选择统计。

* 支持报表模板个性化配置。
* 支持报表的自动生成、导出功能。
* 当展示时既有报表又有图形的时候，按上图、下表展示。
* 当点击报表中某个指标，能够展示该指标的趋势图。当指标为月指标时，展示从本年1 月到上月末的趋势图；当指标为日指标时，展示从本月1 日到前1 日的趋势图。
* 支持可自定义查询条件。

决策支持应用

决策支撑看板

决策支撑看板实现从样式组件、展示数据、展示规则、关联规则、模型优化的进行配置，最大程度贴合各种角色、岗位人员的分析决策需求。实现数据价值的最大化。

通过票务经营分析、游客满意度分析、港城联动分析、港城运营总体分析、港城人流车流分析、港城设备能耗分析、数据统计与报表等多平台多维度分析结果制定关联规则。通过多样化组件技术进行直观的分析结果数据，为管理者决策提供大量数据依据。

通过一定数量数据的积累，挖掘关联数据，建立算法模型，制定关联规则，通过数据的不断校验，算法训练，优化规则。支持预置各类常用的数据挖掘算法，包括分类预测、回归预测、聚类分析、时序分析等。支持直接选用预置算法对数据进行分析。各类数据挖掘算法可以进行封装，使用户无需进行代码上的修改。提供友好方便的数据选择和参数调节功能，支持对参数的检验。支持参数调整后，即时展示分析结果，方便调整参数。支持一次选择多种数据挖掘方法对同一组数据进行分析，并一次给出各个方法的分析结果，以便节省逐个试用不同方法的对比时间。可方便、友好增加新的算法。

建立关联规则库，系统用户可按需选择几种维度分析结果之间关联规则，计算出最符合使用者的决策支撑数据

周期性报告

周期性报告实现从数据、时间、内容类型、报告类型均可配置，生成最大程度贴合各种角色、岗位人员需求的报告。避免冗余数据干扰决策分析，将数据的价值最大化。并利用大数据计算工具实现数据的高速计算，缩短计算响应时间，极速生成报告，提高工作效率。

数据类型可配置：将交易情况分析、库存分析、购买力分析、吸引力分析、单产品分析、消费者分析及平台运营分析内有的原始数据及分析数据组件化，建立原始数据组件库、分析数据组件库及关联数据组件库，系统用户按需选择数据类型，系统支撑组合数据类型。

时间可配置：系统实现时间周期、刻度可选择的最大自由化，系统用户选择日、周、旬、月、季度及年的时间周期，时间刻度可细化到小时、分钟。时间周期与时间刻度自由组合，可配置成日-小时、日-分钟、月-日-小时、月-日-分种等几十种组合数据。

内容类型可配置：将交易情况分析、库存分析、购买力分析、吸引力分析、单产品分析、消费者分析及平台运营分析几个维度的分析内容，分类成模块化系统用户自行选择内容类型，可复合选择。

报告类型可配置：系统建立报告类型模型，分为内容形式类型及报告格式类型。系统提供数据表格、多样化图型的内容形式类型，如仪表盘、帕累托等。

报告用途可配置：通过四个配置，生成最终使用者所需内容报告，并可进行二次筛选、淘汰数据。可加入报告背景字段，logo等信息，系统支持在线上传、编辑报告名称、封面，生成更为专业的数据支撑报，系统支持在线导出、在线打印。

决策支持

决策支持系统和数据仓库执行的都是决策和趋势分析类的应用。决策支持系统中的一些技术可以很好地集成到数据仓库中，使数据仓库的分析能力更加强大。例如：决策支持系统中的传统统计分析模型可以帮助用户对数据仓库中的数据进行更加有效、更加深入的分析，从而更好地掌握和利用信息。

数据仓库是面向决策分析的，具有从各种数据源抽取数据，并对数据进行清洗、聚集和转换等各种处理能力。从这个意义上说，数据仓库既是一个信息资源的存储机制，也是一个信息资源分析、处理的支持决策的系统。政府组织的决策行为依赖于信息资源的良好组织，因而基于数据仓库的政府信息资源存储必然成为主流的政府信息资源存储方式。