中铝瑞闽智能制造新模式应用项目

技术方案规划书

2017-12-29

北京科技大学

目录

[1 项目介绍 4](#_Toc502267524)

[2 项目总体设计 7](#_Toc502267525)

[2.1 设计概述 7](#_Toc502267526)

[2.2 项目目标 9](#_Toc502267527)

[2.2.1 成本精益控制 9](#_Toc502267528)

[2.2.2 质量精益管控 9](#_Toc502267529)

[2.2.3 客户精益服务 10](#_Toc502267530)

[2.2.4 设备精益监控 10](#_Toc502267531)

[2.2.5 安全精益监督 11](#_Toc502267532)

[3 系统框架 12](#_Toc502267533)

[3.1 系统框架总体结构 12](#_Toc502267534)

[3.1.1 总体逻辑框架 12](#_Toc502267535)

[3.1.2 总体技术框架 13](#_Toc502267536)

[3.1.3 系统软硬件规划 15](#_Toc502267537)

[3.2 数据获取 15](#_Toc502267538)

[3.2.1 实时镜像备份 15](#_Toc502267539)

[3.2.2 网络爬虫scrapy 15](#_Toc502267540)

[3.3 数据仓库 17](#_Toc502267541)

[3.3.1 数据仓库设计 17](#_Toc502267542)

[3.3.2 元数据管理 18](#_Toc502267543)

[3.3.3 数据清洗 20](#_Toc502267544)

[3.3.4 ETL工具 21](#_Toc502267545)

[3.4 知识挖掘和模型构建 21](#_Toc502267546)

[3.5 应用服务与智能平台 22](#_Toc502267547)

[3.5.1 应用服务后台 22](#_Toc502267548)

[3.5.2 权限模块 23](#_Toc502267549)

[3.5.3 智能可视化平台 23](#_Toc502267550)

[3.6 数据服务平台 24](#_Toc502267551)

[3.7 大数据与系统集成 26](#_Toc502267552)

[4 成本精益控制 29](#_Toc502267553)

[4.1 基础成本指标 29](#_Toc502267554)

[4.1.1 原材料采购成本 29](#_Toc502267555)

[4.1.2 制造成本 29](#_Toc502267556)

[4.1.3 产品运输成本 30](#_Toc502267557)

[4.2 分析方法与模型 30](#_Toc502267558)

[4.2.1 原材料采购指标 30](#_Toc502267559)

[4.2.2 物料消耗指标 30](#_Toc502267560)

[4.2.3 能源消耗指标 31](#_Toc502267561)

[4.2.4 耗材消耗指标 32](#_Toc502267562)

[4.2.5 人力成本指标 32](#_Toc502267563)

[4.2.6 运输成本指标 33](#_Toc502267564)

[4.3 可视化展示 33](#_Toc502267565)

[4.3.1 原材料采购成本分析 33](#_Toc502267566)

[4.3.2 制造成本分析 34](#_Toc502267567)

[4.3.3 运输成本分析 34](#_Toc502267568)

[5 质量精益控制 36](#_Toc502267569)

[5.1 基础质量指标 36](#_Toc502267570)

[5.2 分析方法与模型 36](#_Toc502267571)

[5.2.1 铝溶液成分偏离度 36](#_Toc502267572)

[5.2.2 铝熔液成分波动率 38](#_Toc502267573)

[5.2.3 生产温度偏离和波动 39](#_Toc502267574)

[5.2.4 产品成材率和合格率 39](#_Toc502267575)

[5.3 可视化展示 40](#_Toc502267576)

[6 客户精益服务 43](#_Toc502267577)

[6.1 产品销售情况分析 43](#_Toc502267578)

[6.1.1 销售分析主要KPI指标： 43](#_Toc502267579)

[6.1.2 数据获取 43](#_Toc502267580)

[6.1.3 计算模型 44](#_Toc502267581)

[6.1.4 效果展示 46](#_Toc502267582)

[6.2 客户分析 47](#_Toc502267583)

[6.2.1 客户分析相关KPI 47](#_Toc502267584)

[6.2.2 数据获取 48](#_Toc502267585)

[6.2.3 客户评分评级计算模型 48](#_Toc502267586)

[6.2.4 效果展示 49](#_Toc502267587)

[6.3 服务质量评估 49](#_Toc502267588)

[6.3.1 服务质量评估主要KPI 49](#_Toc502267589)

[6.3.2 数据获取 50](#_Toc502267590)

[6.3.3 评估结果计算方法 50](#_Toc502267591)

[6.3.4 效果展示 51](#_Toc502267592)

[6.4 生产排程 52](#_Toc502267593)

[6.4.1 排程相关KPI指标 52](#_Toc502267594)

[6.4.2 辅助排程工具 52](#_Toc502267595)

[6.4.3 排程效果评估 53](#_Toc502267596)

[6.4.4 效果展示 55](#_Toc502267597)

[6.5 订单管理模块 56](#_Toc502267598)

[6.5.1 用户订单提交 56](#_Toc502267599)

[6.5.2 历史订单管理 57](#_Toc502267600)

[6.5.3 效果展示 57](#_Toc502267601)

[7 绩效智能决策 59](#_Toc502267602)

[7.1 KPI指标体系 59](#_Toc502267603)

[7.1.1 基础指标 60](#_Toc502267604)

[7.1.2 主题综合指标 61](#_Toc502267605)

[7.2 员工安全评估 61](#_Toc502267606)

[7.2.1 员工安全资质 62](#_Toc502267607)

[7.2.2 员工反馈指数 63](#_Toc502267608)

[7.3 自定义模板 65](#_Toc502267609)

[7.3.1 自定义KPI指标 65](#_Toc502267610)

[7.3.2 自定义计算方法 66](#_Toc502267611)

[7.4 可视化展示 68](#_Toc502267612)

# 项目介绍

高端铝合金功能材料智能制造新模式围绕中铝瑞闽高端铝合金功能材料的产品研发、生产制造、质量管控、仓储物流、优化决策等产品全生命周期的主要过程，建设以“集成化、精益化、数字化、互联化、智能化”为特征的高端铝合金功能材料的智能制造新模式，缩短新材料的研发周期、提高生产效率、提升产品质量、降低过程能耗。通过智能制造新模式的建设，提升行业智能化制造水平，增强企业核心竞争力，实现精益、高效、绿色、智慧生产。

实践表明，通过传统的技术与管理提升，已经难以有效解决制铝企业产品设计、生产制造、经营管理等多个生产与管理环节的全局协调优化问题，也难以解决产品质量、节能减排与生产效益的动态协调与管理控制存在的问题；更加难于满足制铝企业增效、降耗与转型升级的需求。先进的信息技术恰恰可以帮助解决铝企业特有的连续流程整体优化问题，具体包括提升生产效益、保障产品质量、节能降耗、提升生产计划兑现率、减缓机器设备损坏、缩短生产停车时间、减少大修次数等。

随着以社会化网络、移动互联网、云计算等信息技术的兴起和快速发展，数据已经成为社会化的战略资源。一个企业应有数据的规模和运用数据的能力正在成为综合实力和创新能力的重要组成部分。云计算、大数据技术，人工智能在促进工业化与信息化融合方面的重要作用和潜力已显现出来，已成为带动工业和社会发展的重要力量，已经成为驱动铝企业形成创新发展机制，突破增长极限，保障经济快速发展的主要动力。有效地组织和使用大数据将对企业发展与企业创新能力提升产生巨大的推动作用。近年来，云计算、大数据技术日趋成熟，大数据对企业生产、管理已经产生了重大影响；利用大数据技术，进一步提高产品质量及生产效率，降低生产成本及能源消耗，减少排放，实现绿色制造已成为铝企业发展的重要技术措施。大数据是一整套数据分析处理技术体系，更是一种复杂问题解决的思想方法；利用企业生产、经营管理业务数据、产品服务数据、宏观经济数据等构成的“企业大数据”，在企业经营管理决策中开始发挥重要作用。大数据对于企业管理模式创新具有重要影响，是改善企业生产管理能力、提升决策能力、形成企业管理创新的关键。

目前，中铝瑞闽生产包括熔铸、热轧、冷轧、退火及精整五大工序，生产过程涉及多个工序，每个工序均会产生大量的过程数据，包括合同订单信息、产品规范、工艺参数、生产消耗、实绩曲线等。对于这些数据的处理，目前多采用孤立的方式，即仅对单个产品的单个工序进行分析，且数据存储方式简单，保存期间短，没有有效利用这些数据，对生产过程的企业实现精细管理提供有力支持，主要表现在：

（1）不能在产品生产的各工序之间和工序内部实现质量信息的及时传递，出现质量异议，缺乏有效的过程溯源数据，难以准确定位出现问题的环节及快速找到出现问题的原因，导致批量的质量事故或长时间的生产停滞；

（2）由于产品生产过程非常复杂，多种因素耦合在一起，上游工序的生产结果会对下游工序产生遗传影响，简单的数据处理方式及数学模型无法满足高效、高精度的控制要求，也无法为工艺模型的优化提供有效支撑；

（3）产品的营销信息与生产过程数据没有有效关联和融合，不能为企业经营的科学决策提供支持，包括生产组织、资源分配等。

铝合金产品生命周期包括整个生产过程，涉及多个工序，每个工序均会产生大量的过程数据，涵盖铝合金产品生产过程的实时生产数据、产品信息与积累的经验知识，数据具有典型的大数据的“4V+1C”的特征（即Volume，Velocity，Variety，Value和Complexity），其突出的是生产数据以实时流数据为主体，数据量随时间持续快速增长，数据体内部蕴含复杂非线性关系，多源分布异构数据并存等特点；使数据的分析挖掘和应用的难度非常大。生产数据本身的多样性和复杂性及其所表征的铝合金产品生产流程的复杂性，使得企业大数据智能分析与决策技术研究具有典型意义。

因此，本项目研究中铝瑞闽生产制造数据、经营管理数据、营销数据等大数据的多元异构集成、可靠存储、可视化决策分析等关键技术；研究开发基于大数据的决策分析模型、算法，构建模型驱动的企业大数据智能分析与决策支撑平台。面对高端铝合金功能材料定制化、多样化需求，以及制造过程产品一次合格率低、质量异常频繁等问题，项目建设中将在目前企业ERP、APS、MES与PCS等信息与自动化系统架构基础上，通过对制造过程工业大数据的深度利用构建面向产业链内外协同的智能制造决策新模式，实现高端产品大规模个性化定制生产，有效降低制造成本，提升产品的价值链和精益服务能力，提升企业对于客户个性化需求的快速响应能力和核心竞争力。

# 项目总体设计

## 设计概述

针对中铝瑞闽生产数据拥有庞大的规模，不断变化的类型，不断演化的分析模式，存在信息量大，信息存储的格式繁多，信息较分散不集中，现有技术难以为铝制产品生产过程进行相应支持，难以对高端产品大规模个性化生产，多工序无法协同智能产生等具体问题，建立中铝瑞闽智能决策系统。

在智能分析与决策技术的基础理论、技术方法、模型算法与应用支撑平台等多个层面，构造综合集成的智能分析与决策技术体系，形成符合制铝生产大数据特点的数据整合与数据分析挖掘技术和智能分析与决策技术平台，建设横跨质量、成本、客户、设备和安全五个主题的KPI体系，以及基于自定义模版的绩效计算模型。

由于工业过程的复杂性及特点，将智能分析、决策应用与制铝生产系统控制理论相结合，围绕生产流程和工艺模型，研究模型驱动的适用于复杂工业系统特点的大数据整合（如何体现智能、流程、可视化等）、大数据融合、大数据分析挖掘技术。

（1）数据挖掘与分析方法库

对多对象、多变量、多路径、多证据、多目标的复杂优化问题，提出基于证据的动态规划方法；分别对成本、质量、客户、设备、安全、绩效等KPI指数进行在线监控与分析、及时发现管控中存在的问题、预测变化趋势、分析存在的风险、建立风险预警机制，最终实现成本精益控制、质量精益管控、客户精益服务、设备精益监控、安全精益监督、绩效精益管理。

（2）智能决策业务模型设计

围绕产品全生命周期的核心信息与数据，建立分析平台，构建顶层决策支持系统，结合统计分析，分类、聚类，回归预测，多目标优化等智能算法，针对生产调度、成本控制、质量管控、客户服务、设备监控、安全监督、绩效管理等主题业务，分别实现数据主题仓库构建，模型算法设计优化，实现产品全要素、全价值链、全流程、全生命周期的数据与信息集成。

（3）智能决策展示平台

构建整个高端铝功能材料制造流程的KPI导向图，实现数据虚拟分析与实际业务系统的互动，利用数据KPI导向图分析引擎及各数据分析支撑模型，将制造过程各类数据可视化、数字化和智能化，便于决策者和管理者对整个制造过程进行实时掌控，对存在的问题进行及时、准确的决策，提高企业的经济效益。

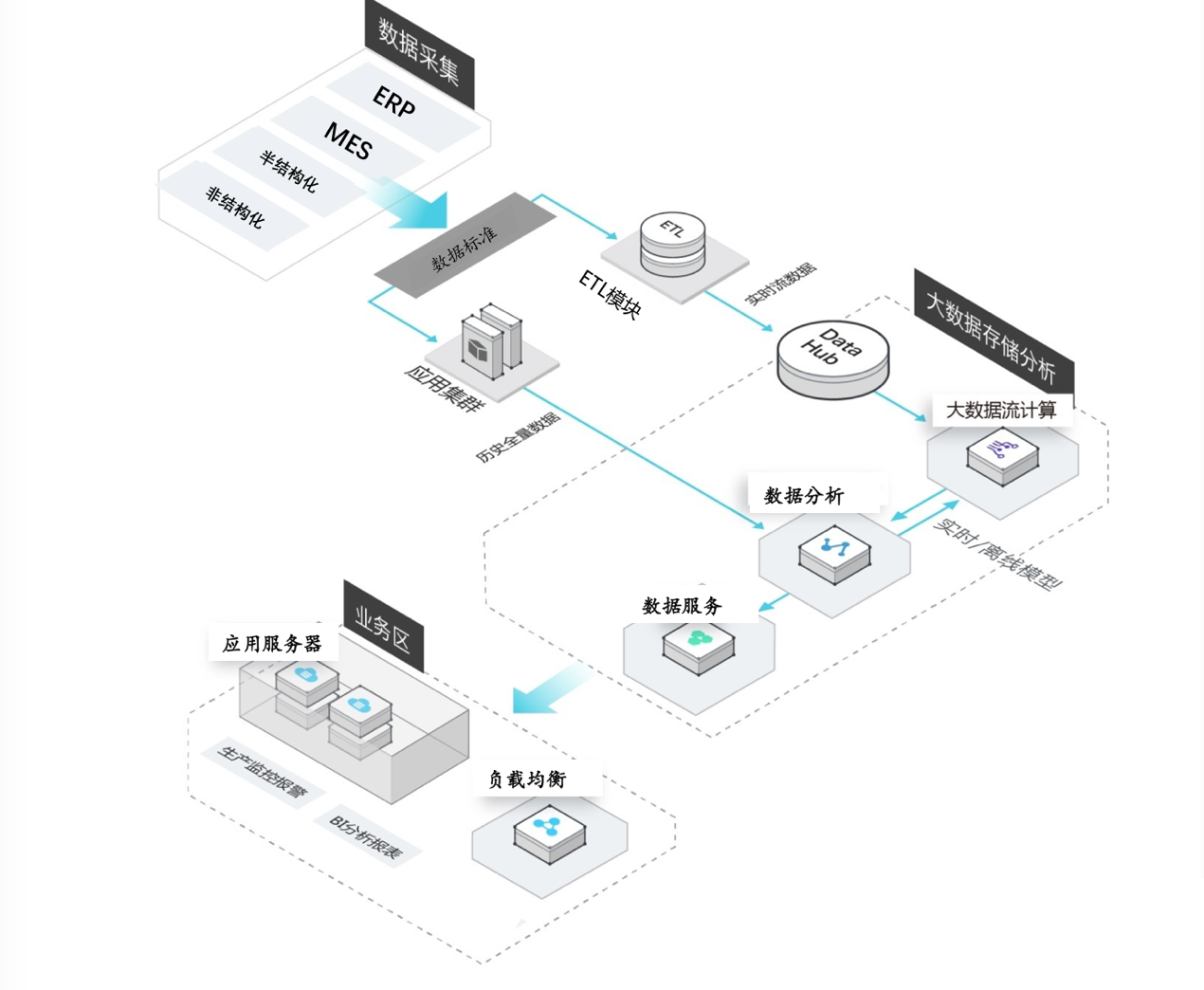


图 ‑1瑞闽智能决策支持平台

平台逻辑功能框架自底向上是分布式高维数据存储层、元数据服务层、复杂工业模型层、处理分析层以及制铝企业大数据分析处理应用层；同时，需要对分布式系统进行作业、资源调度、管理的协调与监控中间件的支持，支持工作流及其调度的设施。逻辑架构在逻辑上划分为实时数据流处理子系统与大数据智能分析与决策子系统。

实时数据流处理子系统接受实时传感器数据流，数据流元组记录随时间变化的空间信息，具有动态、海量、高维、时效、连续、多源、无限等特性。能够为分析调度、设计规划、异常预警等工业信息管理和决策提供支持，为工业提供更为全面和便捷的决策服务。

大数据智能分析子系统运用适于制铝领域内先进的数据处理技术，根据各工业子系统的需求和它们之间的内在联系，对来自多来源渠道、格式不一致的数据在综合生产信息的基础上进行抽取、集成，并进行深度智能分析与处理，获得可用于决策的模式、模型、规则和知识。

## 项目目标

在智能分析与决策技术的基础理论、技术方法、模型算法与应用支撑平台等多个层面，构造综合集成的智能分析与决策技术体系，形成符合制铝生产大数据特点的数据整合与数据分析挖掘技术和智能分析与决策技术平台，建设横跨质量、成本、客户、设备和安全五个主题的KPI体系，以及基于自定义模版的绩效计算模型。

### 成本精益控制

成本精益控制包括对熔铸工序、热轧工序、冷轧工序等各工序的核心成本KPI指标进行分析。核心成本KPI指标包括熔炼时的各类物料消耗偏离度和波动率信息，热轧时的耗材消耗、燃料消耗、热轧板成材率及冷轧时的耗材消耗、产品废品率等。

进行基础数据整合、相关数据搜集，数据预处理功能用以对原始数据进行处理，并重整，建立相应的数据之间关系，并按存储模型要求进行存储。

根据目标KPI进行历史数据分析，包括数理统计分析、成本问题挖掘、问题原因追溯等。基于历史数据，统计成本KPI指标统计期望、标准差等特征参数，分析当前指标状态，发现现存成本问题，利用Pearson相关性计算方法，对工序输入、控制、输出参数相关性进行分析，综合使用特征选择、多元回归分析等方法实现问题原因追溯。

### 质量精益管控

质量精益管控包括对熔铸工序、热轧工序、冷轧工序等各工序的核心质量KPI指标进行分析。核心质量KPI指标包括熔铸时熔液成分偏离度、波动率以及熔炼温度，热轧时轧制温度和热轧板质量，冷轧时产品质量等。

实施基础数据整合、相关数据搜集，数据预处理与数据存储、统一数据访问接口。数据采集自熔铸至成品的整个制造流程关键工艺装备、系统中的工艺参数、控制设定参数和物料、质量参数。

根据目标KPI进行历史数据分析，包括数理统计分析、控制问题挖掘、问题原因追溯等。通过对各产线的产品和过程实施过程质量信息采集，统计分析各指标当前状态，与历史数据进行对比，挖掘质量控制问题，基于Pearson相关性计算方法，分析工序输入、控制、输出参数相关性，综合使用特征选择、多元回归分析等方法实现问题原因追溯。

### 客户精益服务

客户精益服务分别包含产品导向、时间导向、客户导向的销售分析。

进行基础数据整合、数据预处理等步骤用以对原始数据进行处理，根据各个分析主题将数据按照不同关键字索引进行归类划分，建立相应的数据之间关系，并按存储模型要求进行存储。

采用不同模型对不同类型的数据进行分析，利用均值、同比变化率、加权得分等方式实现对产品、效益、客户的定量分析，并且针对数据的不同特点，使用折线图、柱状图等方式实现可视化展示，兼具实用性与美观性。对客户进行了评级与打分，让销售部门可以根据不同类型的客户群体提供不同的服务，做到客户的精益化服务。

### 设备精益监控

设备精益监控包括设备性能维护和发挥类指标计算和基于指标的厂区警戒划分。

针对铝生产工序中的多种设备，收集设备维护和生产的设备基本信息，使用领域相关公式完成设备完好率、设备利用率和维修费用率的计算，从多个角度反应设备管理工作的情况。还通过设备的故障信息完成设备工序状况系数计算，依据工序状况系数应用聚类方法实现对厂区设备不同警戒程度的划分。

### 安全精益监督

安全精益监督主要包含员工安全资历计算、员工反馈指数计算和安全反馈表单智能筛选系统。

员工安全资历计算主要基于人力资源系统记录的员工档案数据，通过无监督聚类算法对员工进行聚类，并根据其与中心点的偏移量化得到安全资历指数，将其作为员工的基本信息进行存储。员工反馈指数主要基于安全反馈系统收集的统计记录，通过公式计算得出有效反馈率再结合相应的系数，最终得到员工的有效反馈程度的评价指标。安全反馈表单筛选系统通过关键词模糊分类，基于安全资历和反馈指数的优先级划分以及审核人员的重要性评价完成对整体表单的筛选处理，降低审核人员的工作量。

# 系统框架

## 系统框架总体结构

### 总体逻辑框架

本项目是在中铝瑞闽生产流程范围内，面向生产制造和经营管理相关的分析评价、预测与决策等数据分析挖掘需求，以产品质量管理为主线，针对生产工序间质量分析、生产数据挖掘为目标的节能减排、营销决策支持等具体需求，深度融合实时生产数据、生产管理信息以及工艺知识，研发基于云计算与大数据技术的数据分析挖掘、深度学习、决策分析模型，利用已有成果和开源技术研发模型驱动的可视化、系统化的中铝瑞闽智能分析与决策应用支撑平台。

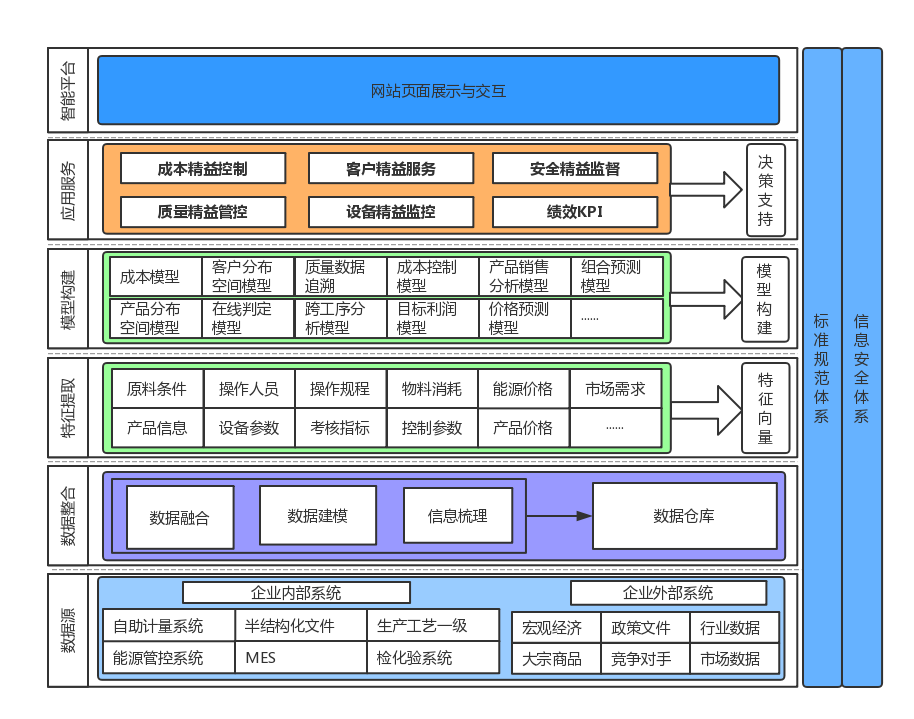


图 ‑1瑞闽智能决策平台逻辑框架

### 总体技术框架

在本项目中，数据来源可以分为产品数据、运营数据、价值链数据、外部数据这四个方面，不仅数据规模较大，且来源不一。所以最终采用了Hadoop的HDFS和spark的并行计算框架来作为数据处理、存储和分析的平台，构建了包含了数据源、数据整合、模式识别、模型构建、应用服务、智能平台共六个层次的智能决策系统。下面由下至上依次介绍每个层次的实施技术方案。

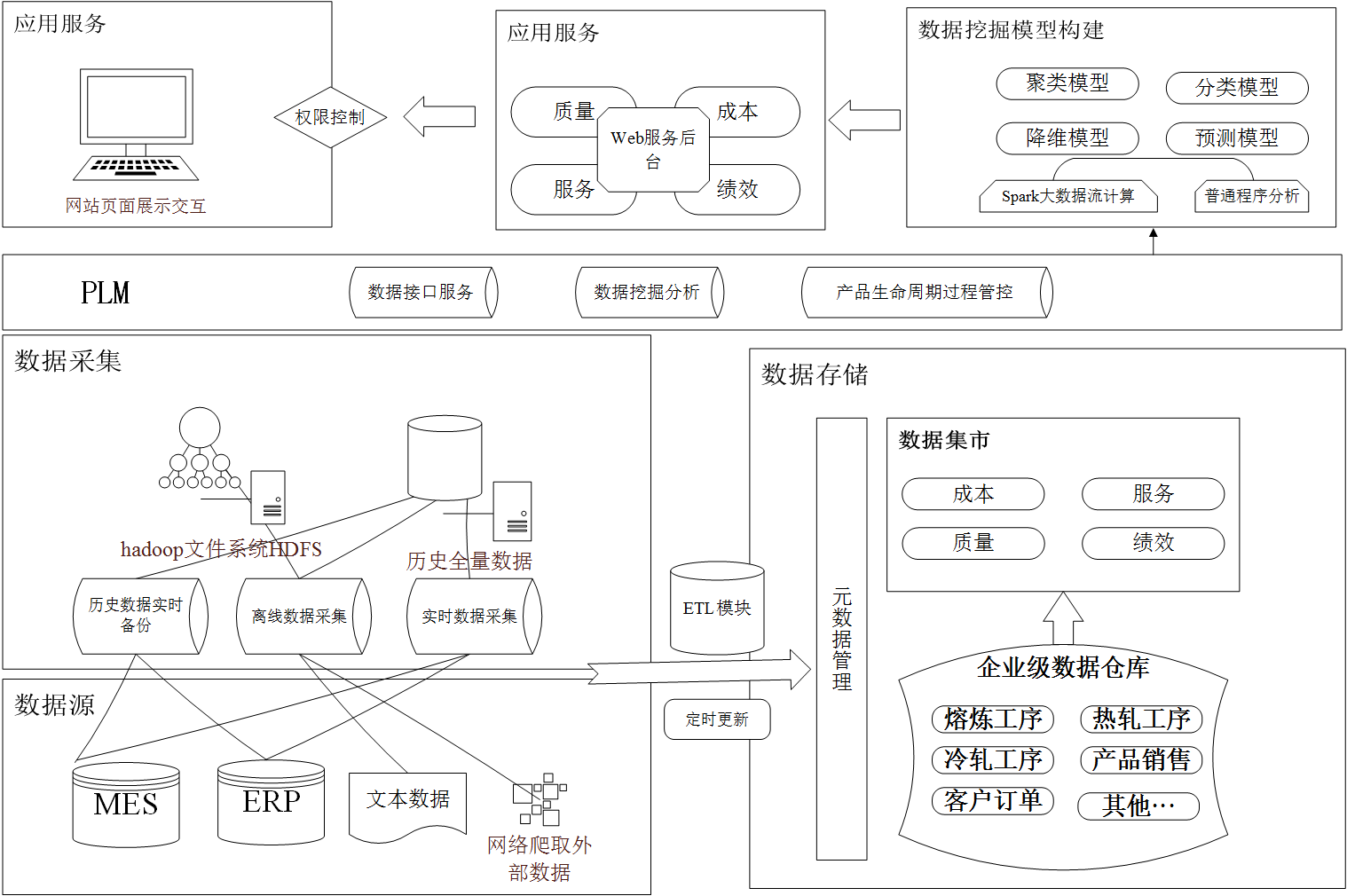


图3-2瑞闽智能决策平台技术框架

1. 数据源

数据源包括企业内部系统和企业外部系统，对于企业内部系统，进行实时备份镜像，主题数据仓库从备份镜像数据库中抽取数据；外部数据包括宏观经济、行业数据、市场数据等，采用网络爬虫从多个网络开放数据源进行爬取。

1. 数据整合

数据整合的核心步骤就是构建数据仓库，为后面的知识挖掘和模型构建等数据分析操作提供数据准备。

传统的OLTP（在线事务处理系统）使用了所有的访问数据的操作 CRUD（创建、读取、更新、删除），数据仓库则是完全不同的另一种应用程序，它是用来分析数据并从数据中发现新的价值，它并不是解决所有问题的通用结构，而聚焦于某一类问题，例如质量、成本等。

1. 特征提取

根据质量、成本、客户、设备、安全等主题的KPI，从原始数据或数据仓库中针对问题提取出有效特征用于进行分析，提高分析预测过程中数据处理的效率，增强分析及预测结果的准确性及可靠性。

1. 模型构建

根据不同的业务流程及KPI指标，需要构建相应的模型进行处理。目前行业内广泛使用的数据分析语言是python，而基于python构建的scikit-learn机器学习库提供了大量机器学习算法的实现，能快速实现为对数据的回归、分类、聚类及降维等操作。

1. 应用服务

采用MVC三层结构，构建应用服务系统后台。MVC包括三个层次内容，controller用于控制请求，model利用之前构建的主题模型，根据当前业务需求进行后台服务编程，view层用于客户端显示界面的控制。

应用服务系统需要根据主题及绩效KPI编写成本、质量、客户、安全、设备及绩效KPI六个业务处理模块，模块具有低耦合高内聚的特性。

结合上面提到的需求及实施条件，采用Django这种基于MVC模式的Web框架来构建智能决策系统应用服务后台。Django采用python语言进行开发，可以有效的结合python广泛而有效的社区，及其开发的一系列高效的第三方库资源开展包括数据分析，模型构建，逻辑业务处理等一列列工作。

1. 智能平台

智能决策展示平台使用B/S模式，即服务端和浏览器端，拟采用两种方式进行与决策人员的交互，包括网页浏览器端展示及微信平台的消息推送。

使用浏览器作为访问端增强了系统的灵活性，能够便捷的在不同设备，不同操作系统，不同浏览器设备进行访问，微信平台的信息推送则可以增强信息传递的即时性。

相应的开发任务包括网页开发和移动端开发两部分，使用基于网页的开发语言html、css、javascript，以及借助构建起上的前端开发框架jquery、vue.js等提高开发效率。界面的可视化展示则使用echarts，echarts是网页图表展示组件，提供数据接口，将数据处理后传入相应组件即可获得预期的可视展示效果。

### 系统软硬件规划

表3-1 软硬件需求表

|  |  |
| --- | --- |
| 层级 | 实现方案 |
| 数据管理 | 关系型数据-Oracle、非结构化数据-HDFS |
| 开发语言 | Python3.5 |
| 模型构建 | python工具模块、spark计算集群 |
| 服务后台搭建 | web服务框架Django1.9 |
| 客户端 | 网站页面，echarts(3.0)进行网页内嵌图表展示 |
| 硬件需求 | 1台应用服务器，1台数据库服务器，部分数据存入大数据平台 |

## 数据获取

### 实时镜像备份

采用ADG进行数据库的实时备份，ADG包括的是当前或接近当前的数据，ADG反映的是当前业务条件的状态，ADG的设计与用户或业务的需要是有关联的，ADG的更新是根据业务的需要进行操作的，而没有必要立即更新，因此它需要一种实时或近实时的更新机制。

### 网络爬虫scrapy

对于宏观经济、行业数据、市场数据等外部数据需要从不同的有效网络数据源进行爬取，如何高效统一管理不同类型数据的网络爬虫是需要解决的问题，scrapy具有爬去速度快，爬取功能强大，使用简单的特点，一个典型的scrapy结构如下图所示：



图 ‑3 scrapy体系结构

Scrapy主要包括了以下组件：

引擎(Scrapy): 用来处理整个系统的数据流处理, 触发事务(框架核心)

调度器(Scheduler): 用来接受引擎发过来的请求, 压入队列中, 并在引擎再次请求的时候返回. 可以想像成一个URL（抓取网页的网址或者说是链接）的优先队列, 由它来决定下一个要抓取的网址是什么, 同时去除重复的网址

下载器(Downloader): 用于下载网页内容, 并将网页内容返回给蜘蛛(Scrapy下载器是建立在twisted这个高效的异步模型上的)

爬虫(Spiders): 爬虫是主要干活的, 用于从特定的网页中提取自己需要的信息, 即所谓的实体(Item)。用户也可以从中提取出链接,让Scrapy继续抓取下一个页面

项目管道(Pipeline): 负责处理爬虫从网页中抽取的实体，主要的功能是持久化实体、验证实体的有效性、清除不需要的信息。当页面被爬虫解析后，将被发送到项目管道，并经过几个特定的次序处理数据。

下载器中间件(Downloader Middlewares): 位于Scrapy引擎和下载器之间的框架，主要是处理Scrapy引擎与下载器之间的请求及响应。

爬虫中间件(Spider Middlewares): 介于Scrapy引擎和爬虫之间的框架，主要工作是处理蜘蛛的响应输入和请求输出。

调度中间件(Scheduler Middewares): 介于Scrapy引擎和调度之间的中间件，从Scrapy引擎发送到调度的请求和响应。

Scrapy运行流程大概如下：首先，引擎从调度器中取出一个链接(URL)用于接下来的抓取；引擎把URL封装成一个请求(Request)传给下载器，下载器把资源下载下来，并封装成应答包(Response)；然后，爬虫解析Response；若是解析出实体（Item）,则交给实体管道进行进一步的处理；若是解析出的是链接（URL）,则把URL交给Scheduler等待抓取。

## 数据仓库

建立数据仓库的主要任务是建立中铝瑞闽企业全局性、基础性的数据中心。本研究以云平台、Hadoop为技术依托，进行了中铝瑞闽基础数据接入、数据清洗算法、统计口径整理、数据编码中铝瑞闽指标体系设计及实现，最终使中铝瑞闽数据达到指标规范、口径一致、数据字典标准。

研究为中铝瑞闽企业提供面向中铝瑞闽资源开发业务对象，支持数据采集、数据存储、分析处理和决策支持的数据管理体系结构和方法。通过中铝瑞闽智能决策系统数据仓库的建设，实现中铝瑞闽企业数据的统一集中管理、数据共享和数据交流，为中铝瑞闽企业提供数据分析应用环境，使中铝瑞闽管理者全面、及时、准确地掌握企业信息，为中铝瑞闽生产经营管理提供可靠依据，实现生产经营决策的科学性和及时性；促进和规范中铝瑞闽企业乃至整个制铝行业的信息标准化建设。

### 数据仓库设计

数据仓库是一个用于支持管理决策的面向主题的（Subject Oriented）、相对稳定的（Non-Volatile）、集成的（Integrate）、反映历史变化（Time Variant）的数据集合。

数据仓库与大数据本质上没有必然联系，但随着数据量的急剧增加，TB、PB  
级的数据越来越常见，传统的技术已无法满足如此大量数据的存储与分析要求，因而数据仓库需要引入大数据技术，作为数据仓库的一环，解决大数量下的存储与分析问题，让数据仓库在大数据量下仍然高效。

构建一个数据仓库系统主要涉及以下几个步骤：收集商业需求、进行数据建模、数据加工与处理、数据查询与分析。

1. 收集商业需求

前面我们说到，数据仓库的目的是为了支持管理决策，所以我们在构建数据仓库的时候首先需要考虑的做的是收集商业需求，深入理解企业的各个业务流程（Business Process），抓住核心需求，为数据仓库的建立提供指导方向，这样才能反过来为业务提供更好的决策数据支撑，让数据仓库价值的最大化。

1. 进行数据建模

目前中铝瑞闽公司的业务要求，我们提出了成本、质量、服务及绩效四个主题的内容，针对产线的各个工序，都需要从该工序中选取合适的数据、合理的维度（时间和空间就是两个不同的维度）建立数据模型。因为分析需求总是多种多样、猝不及防的，只有建立合适的数据模型，才能更加从容的面对未来多变的分析需求，建模的重要性可见一斑。

1. 数据加工处理

完成数据建模之后就可以进行数据的加工和处理了，数据仓库里我们称之为ETL（Extraction, Transformation, and Loading）流程，主要的工作是从各个业务系统或者日志系统提取数据，并做必要的转换，保证数据的正确性，然后加载到数据仓库系统中，以便之后的查询和分析。

### 元数据管理

元数据管理任何文件系统中的数据分为数据和元数据。数据是指实际的数据，就是我们能看到的一条条记录。而元数据指用来描述一个表的特征的系统数据，比如表的字段信息、访问权限、拥有者以及数据块的分布信息等等。

数据仓库的元数据主要包括三类数据库管理系统的数据字典、ETL处理流程产生的日志、BI建模和分析中工具或文档中记录的信息。

1. DBMS数据字典

数据库管理系统（DBMS）中的元数据一般在所有的数据仓库都会包含，因为数据仓库一般都是基于数据库搭建的，而数据库本身的管理系统就会自动维护一套数据字典供用户查询。这些信息一般包括：

* 数据库的关系模型，包含的对象及对象的描述；
* 数据库的表结构、字段信息及描述；
* 表和字段中的主外键、索引、约束等信息；
* 各对象的存储位置和操作权限等。

1. ETL处理日志

ETL是数据仓库管理和维护的基础，就像是数据仓库的血液维系着整个数据的新陈代谢。我们需要时刻关注血液的循环是否正常，它是保证数据完整性、一致性、准确性和及时性的重要参考依据，所以我们需要记录ETL任务的处理日志，我一般会记录以下几类信息：

* 任务信息、调用的程序或脚本、前置任务；
* 数据来源、加载目标、转化规则或计算公式；
* 数据的刷新类型、刷新频率，任务调度信息；
* 每次运行的起始时间、结束时间、操作记录数、任务状态及出错信息。

1. BI分析模型

BI分析模型主要有两类，一类是数据仓库常见的多维模型，另一类是根据具体业务构建的商业分析模型。无论是哪类模型，其实都已经在分析的层面上，所以都有必要记录以下几类信息：

* 分析模型的设计和结构；
* 模型的分析应用和商业价值；
* 模型中指标的定义、计算方法；
* 模型的展现和效果。

　　模型一般由分析师设计和构建，所以这类信息一般会以文档的形式记录下面，或者制作成相应的PPT进行展示。这里必须注意的是分析模型在构建之初就必须明确应用的环境、体现的价值或可能实现的预期，明确这些是为了更好地应用到实践中，如果只是单纯为了实现这样的模型或者基于相应算法的实现，那么很有可能最终模型会变成一种摆设；再有一点就是模型的展现，模型需要优化其在可视化层面的效果，也就是要让其他人能够更好地理解模型的使用和价值，一切底层的算法和数据的处理只是为了让模型在最终的展现上更加有效。

### 数据清洗

建立中铝瑞闽智能决策系统的数据仓库目的主要是为联机分析和决策支持等相关应用业务服务，是为了优化企业资源调配、提升企业的协同和资源共享，从而提高企业的竞争力和创新力。高质量的数据能够促进和提高企业决策的正确性，而质量差的数据则起到反面的作用。数据质量的控制成为数据仓库建设发展过程中越来越引起重视的突出问题之一。

与此同时，由于生产环境中的各种原因，导致中铝瑞闽数据系统中，现有数据库存在这样或那样的数据质量问题，主要表现为：不正确的属性值、重复记录、拼写问题、不合法值、空值、不一致值、缩写词不同、不遵循完整性规则、时间序列流数据缺失问题等。此外，从多数据源中抽取数据时，由于各数据源的数据库表结构的设计可能不相同，当完成从多数据源到数据仓库的数据迁移后，同样会产生一些错误或冗余等信息。上述这些问题构成了所谓的数据质量，故需要进行系统的数据预处理。要对中铝瑞闽智能决策系统建设数据仓库时所面临的复杂数据预处理问题，提出公司数据预处理流程。

数据预处理操作包括数据清洗、数据维度转换、数据一致性处理等。

数据清洗方法包括：删除重复记录、补充缺失值、剔除极端错误值、设置数据范围、五数概括法等。

数据维度转换包括：a) 在数据维度过高时使用相关性分析、主成分分析、随机森林等方法进行降维；b) 在数据维度过低或缺少数据维度时使用各种离散化、聚类、自定义分组等方法扩充数据维度。

数据一致性处理包括：a) 使用标准化、归一化的方式对不同的指标度量体系进行统一，例如对单位的统一，对小数位数的定标，对原始数据的最小-最大标准化等；b) 在数据具有多渠道来源时设定不同数据来源的优先级别，以保证数据的完整性。

### ETL工具

ETL，是英文 Extract-Transform-Load 的缩写，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。

DataX 是一个异构数据源离线同步工具，致力于实现包括关系型数据库(MySQL、Oracle等)、HDFS、Hive、MaxCompute(原ODPS)、HBase、FTP等各种异构数据源之间稳定高效的数据同步功能。

## 知识挖掘和模型构建

python的scikit-learn内置丰富的算法和分析模型，同时结合google开源的tensorflow，可以快速高效的实现神经网络算法。

scikit包含丰富的模型，包括贝叶斯、随机森林算法、线性鉴别分析（LDA）、稀疏回归、逻辑回归、稀疏线性、特征建模、决策树分裂、邻接点分析等，模型概览及选择可参考下图：

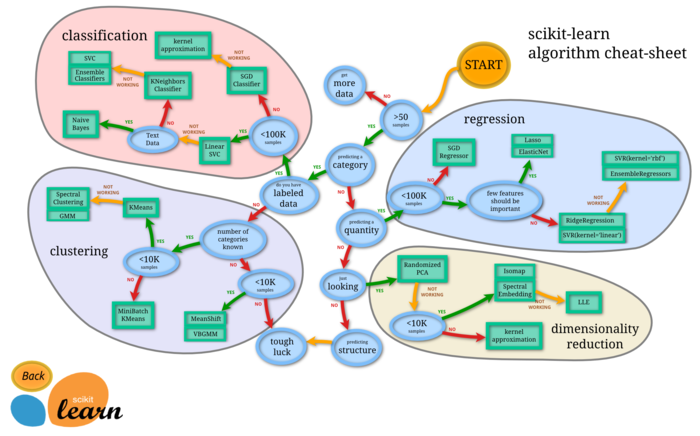


图 ‑4 scikit-learn算法参考

同时可以使用tensorflow构建神经网络模型，一个完整的tenflow构建图计算如图所示：



图 ‑5 tensorflow图计算

## 应用服务与智能平台

### 应用服务后台

智能决策应用服务后台由MVC模式的Web框架Django实现。

Django 是一个高级的 Python 网络框架，可以快速开发安全和可维护的网站。由经验丰富的开发者构建，因此可以专注于编写应用程序，而无需重新开发。Django是免费开源的，有一个繁荣昌盛而积极的社区，在应用程序开发过程中遇到的问题都可以去社区中讨论以进行解决。Django总体结构如下图所示：

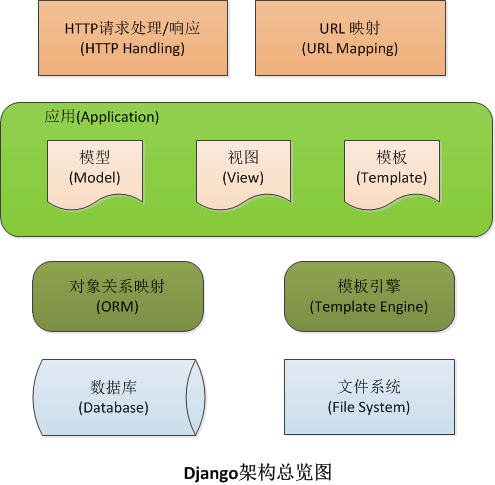


图 ‑6 Django结构概览

### 权限模块

针对中铝瑞闽不同层级的管理人员，在应用服务系统中开发权限系统，基础指标包括成本、质量、服务三个主题的KPI指标，汇总自每个主题的分析结果，面向部门级管理人员，直观展现企业底层业务运行现状；主题综合指标是指针成本、质量、服务三个主题的总体评价性指标，面向集团管理人员，基于系统构建的数据仓库，利用特定的计算方式获得个主题综合指标数值，以供管理人员对企业整体状况进行评估。

### 智能可视化平台

平台可视化展示使用网页形式展现，图表展示使用echarts组件库。echarts是一款开源免费的web组件，它的功能丰富，涵盖各行业图表，满足各种需求，同时它也具有活跃的社区，在应用开发过程中遇到的问题都可以得到有效的解答。

ECharts 提供了常规的折线图，柱状图，散点图，饼图，K线图，用于统计的盒形图，用于地理数据可视化的地图，热力图，线图，用于关系数据可视化的关系图，treemap，多维数据可视化的平行坐标，还有用于 BI 的漏斗图，仪表盘，并且支持图与图之间的混搭。

可以在下载界面下载包含所有图表的构建文件，如果只是需要其中一两个图表，又嫌包含所有图表的构建文件太大，也可以在在线构建中选择需要的图表类型后自定义构建。常用的echarts图表如下图所示：



图 ‑7 echarts常见图表展示

## 大数据与系统集成

### 产品生命周期管理

由于现有的信息化系统之间不兼容和数据的异构问题， 造成了严重的“信息孤岛”，海量数据没有挖掘它的潜在价值，因而现有的系统还缺乏产品与服务价值链上的智能决策能力。为了更好的解决系统之间数据不兼容、异构和不完整等问题，在实施智能制造过程中采用产品生命周期管理（Product Lifecycle Management PLM）架构对重要基础数据实行标准化、规范化的统一管理。

确立以PLM为核心数据平台的智能决策层，可以实现产品与服务数据管理的一体化和规范化，从而形成统一、规范的数据流，为大数据分析提供数据支撑平台；可以实现产品制造流程的标准化和智能化，从而实现产品全流程制造各环节的高度协同和整体优化；可以实现系统间信息交换的标准化和模块化，从而实现各子系统之间的无缝连接，确保信息交换和传递过程的畅通、准确。

### 大数据计算

针对中铝瑞闽在生产过程及企业内部管理过程中产生的海量、多源、异构数据，在数据分析过程可能会出现的单一主机无法进行有效计算的情况，需要结合大数据平台，利用大数据分布式平台的低成本（大量普通PC机即可）、高效性（并发处理）、可靠性（分布式存储，自动存储多分副本）及扩展性（可以扩展至更多集群节点），将部分主题数据存入大数据平台中进行调度计算（包括直接存入csv文件及mongodb中）。