



股票代码:300229

拓尔思数据要素白皮书

拓尔思数字经济研究院

2023 年 3 月

目 录

一、 我国数据要素市场发展概述.....	3
1. 数据量井喷式增长，数据要素市场规模将破千亿.....	4
2. 数据交易产业链形成，数据交易需求旺盛，盈利模式多样化.....	5
3. 我国数据交易机构建设迎来新浪潮，数据流通进入 2.0 时代.....	6
二、 拓尔思数据要素业务目标.....	7
三、 拓尔思数据要素业务体系.....	7
（一） “1+1=N” 的数据要素商业模式.....	7
（二） 数据要素治理体系.....	8
（三） 数据要素业务总体架构.....	10
（四） 数据要素供给业务视图.....	11
（五） 数据要素业务运营架构.....	12
四、 拓尔思数据要素业务介绍.....	15
（一） 云和数据服务业务发展之路.....	15
（二） 公司数据资产现状.....	17
1. 数据资产规模.....	18
2. 大数据底座产品清单.....	23
3. 数据安全产品清单.....	25
4. 云和数据服务产品清单.....	25
（三） 大数据底座和人工智能产品介绍.....	27
1. 网络大数据采集.....	27
2. 数据资源汇聚.....	30
3. 大数据管理.....	33
4. 数据中台.....	35
5. 数据质量管理.....	38
6. 人工智能技术平台.....	43
（四） 数据安全产品介绍.....	45
1. 网络安全市场.....	45

2. 产品服务模式.....	47
3. 市场规模测算.....	49
(五) 云和数据产品介绍.....	49
1. 政府.....	49
2. 金融.....	65
3. 媒体.....	69
4. 专利.....	71
5. 舆情.....	75
(六) 云和数据服务收入规模.....	79
(七) 小结.....	80
1. 拓尔思数据要素业务发展空间大，未来可期.....	80
2. 拓尔思有望成为国内另类数据服务龙头，加速数据要素价值释放	81
五、 拓尔思数据要素发展方向.....	83
(一) 强化数据要素底座基础.....	83
(二) 加强数据治理能力建设.....	83
(三) 提升数据要素变现能力.....	84
六、 拓尔思数据要素业务保障.....	84
(一) 组织保障.....	84
(二) 资金保障.....	85
(三) 人才保障.....	85

一、我国数据要素市场发展概述

数据作为新型生产要素，是数字化、网络化、智能化的基础，已快速融入生产、分配、流通、消费和社会服务管理等各环节，深刻改变着生产方式、生活方式和社会治理方式。数据要素已成为驱动中国式现代化的重要引擎。当前，经济全球化已进入一个由数据要素驱动的新时代。

2020年4月，中共中央、国务院发布的《关于构建更加完善的要素市场化配置体制机制的意见》中正式把数据作为生产要素单独列出，并提出了促进数据要素市场化配置的改革方向。后面，又陆续在《国务院办公厅关于印发要素市场化配置综合改革试点总体方案的通知》《中共中央国务院关于加快建设全国统一大市场的意见》《“十四五”数字经济发展规划》《国务院办公厅关于印发全国一体化政务大数据体系建设指南的通知》等政策文件以及中央深改委第二十六次会议上，进一步提出要促进数据要素资源在更大范围内畅通流动，加快构建数据基础制度体系，推动我国数据要素市场健康有序发展，加快构建以数据为关键要素的数字经济。2023年伊始，数据要素市场活跃。2022年底，《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》（简称“数据二十条”）对外发布，系统性布局了数据基础制度体系的“四梁八柱”，加速了数据流通交易和数据要素市场发展。各地正积极推进数据交易场所建设，优化交易环境，让选购数据产品像逛超市一样方便快捷。



图 1：我国数据要素市场培育的顶层机制建设

1. 数据量井喷式增长，数据要素市场规模将破千亿

数据为数字经济的发展提供了不可或缺的动力支持，近年来全球数据量呈井喷式发展。根据国际数据公司（IDC）预测，2025 年全球数据量将达到 163ZB。随着数据量指数级增长，数据分析算法和技术迭代更新，数据创新应用和产业优化升级，数据对社会变革的影响将更加深远。与此同时，随着数据要素相关政策文件推动数据要素市场深化改革，我国数据要素市场规模进入高速增长阶段。根据国家工业信息安全发展研究中心测算，2025 年将突破 1749 亿元，“十四五”期间我国数据要素市场规模复合增速将达到 26.3%，整体上进入高速发展阶段。

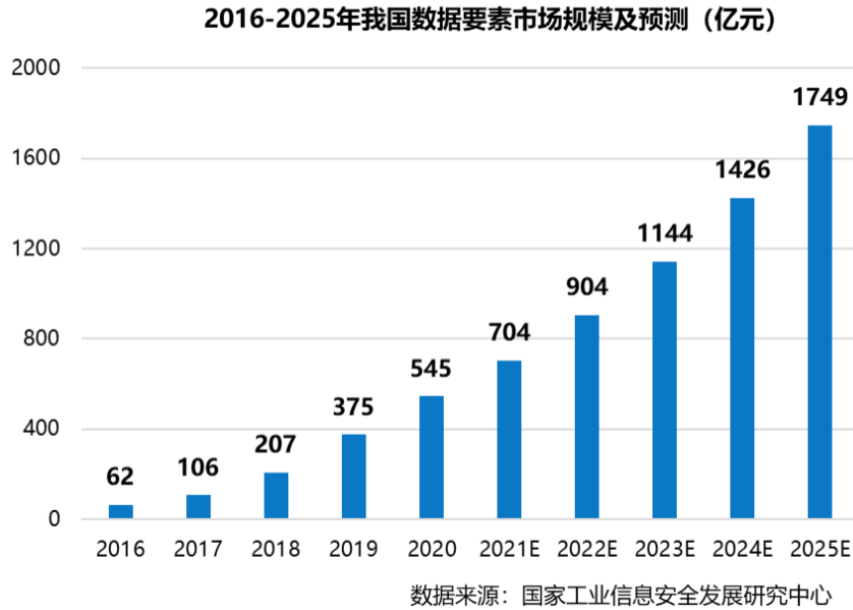


图 2：我国数据要素市场规模及预测

2. 数据交易产业链形成，数据交易需求旺盛，盈利模式多样化

数据流通发生在数据交易前、交易中、交易后和数据出入境等环节，目前已形成产业链生态雏形。在国家政策的推动鼓励下，数据产品交易从概念逐步落地，部分省市和相关企业在数据定价、交易标准等方面进行了有益探索。随着数据产品交易类型的日益丰富、交易环境的不断优化、交易规模的持续扩大，我国数据产品交易变现能力显著提高。根据国家工业信息安全发展研究中心等机构联合发布的《中国数据要素市场发展报告》，2021 年数据要素对我国 GDP 增长的贡献率和贡献度分别为 14.7%和 0.83 个百分点，呈现持续上升状态。市场空间方面，2021 年我国数据要素市场规模达到 815 亿元，2022 年市场规模接近千亿元，并且在“十四五”期间有望保持 25%的复合增速。目前数据交易所的盈利模式主要有三种：佣金模式、会员制模式、增值模式，但盈利模式多样不代表其盈利

能力强大，数据交易所作为“准公共服务机构”，主要目标是赋能市场，推动经济发展。

3. 我国数据交易机构建设迎来新浪潮，数据流通进入 2.0 时代

当前，数据要素市场发展势头强劲。随着数据要素相关政策的推动、数据安全和个人隐私法规的完善及数据流通技术的快速发展，数据产品交易迎来 2.0 时代。截至 2022 年底，全国数据交易所已近 50 家。除建立数据交易所外，各省市纷纷成立由各级政府主导的大数据集团，进一步推动数字化转型进程和数字经济发展。国家未来的目标是培育全国统一数据大市场，包括国家级数据交易所、区域数据交易中心、行业数据交易平台、数据商、数据供需方、第三方服务机构等，通过完整的生态推动数据产业蓬勃发展。

序号	成立时间	交易所	序号	成立时间	交易所	序号	成立时间	交易所
1	2014年	中关村数海大数据交易平台	19	2017年	青岛大数据交易中心	37	2021年	杭州国家数字交易中心
2		北京大数据交易服务平台	20		河南平原大数据交易中心	38		海南数据产品超市
3		香港大数据交易所	21		河南中原大数据交易平台	39		长三角数据要素流通服务平台
4	2015年	贵阳大数据交易所	22	2018年	东北亚大数据交易服务中心	40	2022年	海南大数据交易所
5		华东江苏大数据交易平台	23	2019年	山东数据交易平台	41		无锡大数据交易平台
6		武汉东湖大数据交易中心	24	2020年	安徽大数据交易中心	42		福建大数据交易所
7		武汉长江大数据交易中心	25		北部湾大数据交易中心	43		广州数据交易所
8		西咸新区大数据交易所	26		山西数据交易服务平台	44		青岛海洋数据交易平台
9		华中大数据交易平台	27		中关村医药健康大数据交易平台	45		郑州数据交易中心
10		河北京津冀大数据交易中心	28	2021年	北京国际大数据交易所	46		苏州大数据交易所
11		钱塘大数据交易中心	29		贵州省数据流通交易服务中心	47		辽宁工业大数据交易中心
12	2016年	交通大数据交易平台	30		北方大数据交易中心			
13		上海数据交易中心	31		上海数据交易所			
14		浙江大数据交易中心	32		华南国际数据交易公司			
15		哈尔滨数据交易中心	33		西部数据交易中心			
16		深圳南方大数据交易中心	34		深圳数据交易所			
17		亚欧大数据交易中心	35		合肥数据要素流通平台			
18		丝路辉煌大数据交易中心	36		德阳数据交易中心			

数据来源：拓尔思根据互联网公开数据整理

图 3：我国数据要交易所建设历程

二、拓尔思数据要素业务目标

拓尔思以完善自身数据要素资源体系、壮大数据要素应用体系及建立数据要素安全体系为发展目标，全面布局数据要素市场。

一是完善拓尔思数据要素资源体系，充分释放数据要素价值。强化拓尔思数据资源采集汇聚能力，推动数据高质量汇聚。提升公司在数据采集、标注、存储、传输、管理、应用等全生命周期管理的规范性，健全公司数据资源管理机制，优化数据质量评估方案，积极应用先进质量管理工具，形成数据质量管理闭环，从而进一步促进数据要素价值提升。

二是壮大拓尔思数据要素应用体系，深化数据驱动的行业应用。加强拓尔思在政府、金融、媒体、专利、产业大数据、开源情报等重点行业的数据应用模式探索，深化数据驱动的全流程应用，提升基于数据分析的服务并复制推广。

三是强化拓尔思数据要素安全体系，提升数据支撑和安全保障能力建设。加强公司数据安全管理体系建设，进一步明确安全主体责任和防护要求。围绕数据全生命周期的安全保护要求，强化数据安全监测、加密传输、访问控制、数据脱敏等环节的技术保障，稳步建立数据风险防控体系和预警机制。

三、拓尔思数据要素业务体系

（一）“1+1=N”的数据要素商业模式

拓尔思在经营过程中形成了一套 1+1=N 的数据要素商业模式。

“1+1”是指一个大数据底座和一个人工智能技术平台。

“N”是指行业 SaaS 服务集群，未来可实现 N 个云服务产品的拓展。

“1+1=N”的商业模式主要依托拓尔思自建的大数据中心，基于自研的大数据底座和人工智能技术平台，面向优势行业的业务应用，根据不同场景封装产品，基于公司成熟营销体系迅速推向市场。这种模式充分体现了拓尔思的数据要素具有可复制性强、迭代速度快、复用价值高以及无限供给等特点，实现了统一数据在产品化方面的裂变增长。



图 4：拓尔思数据要素商业模式：1+1=N

（二）数据要素治理体系

数据作为一种新的生产要素，在企业构筑竞争优势的过程中起着重要作用，企业应将数据作为一种战略资产进行管理。数据从业务中产生，在 IT 系统中承载，要对数据进行有效治理，需要业务充分参与，IT 系统确保遵从，这是一个非常复杂的系统工程。

拓尔思公司从 2010 年开始自建大数据中心，经过十多年的实践积累，建立了一套符合公司实际的数据治理体系。该体系从组织、制度、流程、标准、质控、执行、IT、数据产品等方面做出了明确规定，确保关键数据资产有清晰的业务管理责任，作业人员有规范的流程和指导，跨公司跨部门协作时有顺畅的工作机制，治理过程有充足的人才、组织、预算保障，数据产品创新有稳定的决策和优先流程。综合上述因素，最终建立有效的数据治理环境，数据的质量和安全得到保障，数据的价值才能真正发挥出来。



图 5：拓尔思数据要素治理框架

拓尔思在长期的业务经营过程中，形成了一套数据与知识融合的数据要素治理体系。



图 6：拓尔思数据要素治理体系

（三）数据要素业务总体架构

在规模持续高速增长的数据要素市场中，拓尔思拥有完整的大数据基础建设产品线、丰富的行业 SaaS 服务集群和可靠的安全一体化平台。拓尔思采集的数据均为公开的网络数据，基于自研的海贝大数据管理系统进行统一存储，通过 TRS 人工智能技术平台对源数据自动分类、自动抽取、自动标注、自动查重、质量校核等处理，使之成为数据要素。采用不同行业不同场景的业务模型、算法等，将数据要素进行不同维度的融合，包括数据级融合、行业级融合、场景级融合、决策级融合等，形成不同类型的数据产品，实现数据资源向数据资产转变。数据产品以 SAAS、DAAS、KAAS、MAAS 等不同方式，向政府、金融、媒体、专利、能源、企业等行业用户提供细分场景的在线服务，最终实现数据要素的价格化及商品化。拓尔思建立了一套数据安全管

理、数据运营及技术运维、数据要素生态管理等完整体系，为公司数据要素市场化过程提供保障。

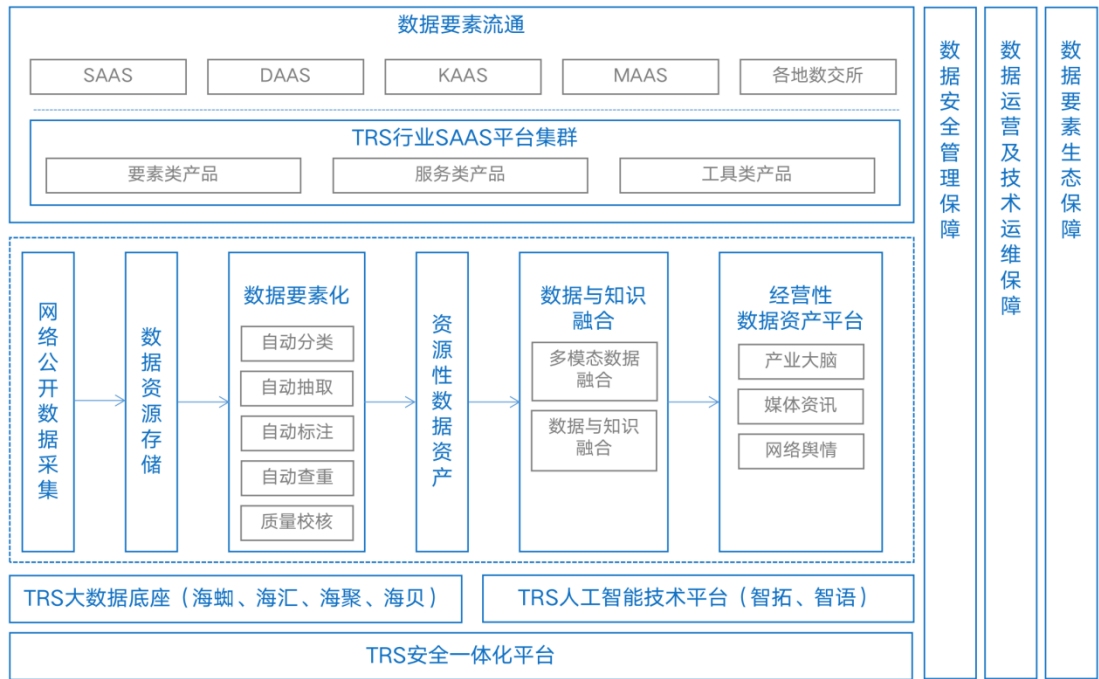


图 7：拓尔思数据要素业务总体框架

（四）数据要素供给业务视图

拓尔思采集的公开数据属于另类数据。目前，业界对另类数据暂没有统一定义。在当前数字时代的语境下，另类数据不同于传统的交易所、公司公告等披露的金融数据，它是由各类信息终端、平台或者系统产生的数据，这些数据能够用来帮助进行投资决策。之所以被称作“另类数据”，是因为当下数据的来源、形式和作用均有别于传统数据，让我们有了解事物外界变化的不同思维方式，所以称为“另类”。

从数据源来看，另类数据的数据源种类主要可以分为以下三大类：

- 1. 个体相关数据：由个人在社会行为中所产生的相关数据，如社交媒体信息、产品评价、搜索记录、APP 使用、通话信息等；

2. 商业过程数据：由商业行为中的交互所产生的相关数据，如支付信息、产品销售数据、人流量数据、Web 数据、数据要素等；

3. 传感器数据：由卫星、5G 基站和传感器等其他设备所收集到的数据，例如车辆轨迹数据、农业种植数据、石油储量数据、生产相关数据等。

拓尔思主要采集 Web 数据、数据要素、社交媒体信息、新闻报道等，并将这些另类数据进行数据要素化处理，与业务场景进行紧密融合后，形成不同类型的数据产品，应用广泛且在不同行业中发挥着重要作用。拓尔思数据要素供给可满足政务、金融、媒体、专利、舆情等业务场景，具体业务视图如下所示：



图 8：拓尔思数据要素供给业务视图

（五）数据要素业务运营架构

拓尔思数据要素业务运营架构如下图所示：

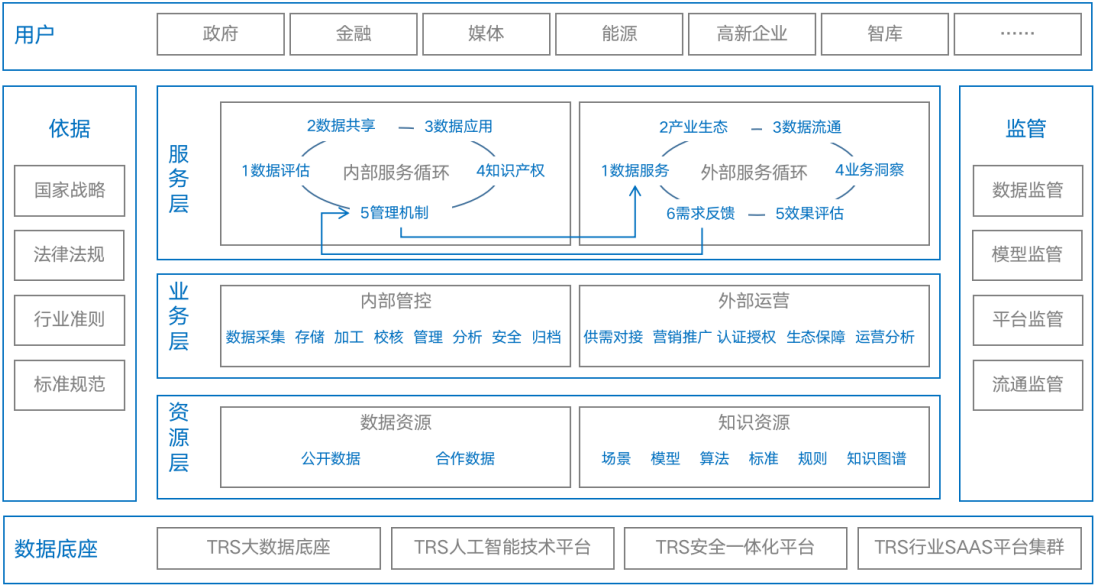


图 9：拓尔思数据要素业务运营架构

整个运营架构涉及运营依据、运营监管、产品运维、以及资源层、业务层、服务层等模块。

1. **运营依据是数据要素运营的导向。**运营依据是数据要素运营的导向，数据要素运营要以国家战略为指引，遵从法律法规的约束，符合地方政策、行业准则和标准规范的要求，并依此来开展数据要素运营活动，确保整个数据要素运营活动的合法合规，安全守正。

2. **运营监管是保障。**通过对数据要素运营过程中的核心能力要素监管，包括数据监管、模型监管、平台监管、流通监管等，确保数据要素运营和服务等活动开展的风险可控和数据产品的健康良性发展。

3. **大数据底座是基础。**由于数据要素运营关系到公司利益与品牌形象，因此要构建一个核心能力集中的、安全可控的、行为可追溯的大数据底座，承载基础数据要素的总体运营，包括统一的数据管理平台、人工智能平台、安全一体化平台和行业 SaaS 服务集群等。其既可以支撑面向公司部门之间数据共享应用、提升公司的治理和服务能

力的内部循环，又可以支撑面向社会主体开放应用、流通和释放数据要素价值的外部循环。

4. 资源层、业务层和服务层是数据要素运营的核心资源层。

资源层：数据要素运营的供给，包括数据资源和知识资源：数据资源主要是拓尔思的自采数据和生态数据；知识资源包括在数据要素运营和服务过程中所产生能够促进数据要素价值流通的应用场景、模型、算法标准和规则等价值资源。

业务层：主要包括内部管理和外部运营。内部管理主要是面向数据要素资源或资产的全生命周期管理，包括数据采集、存储、加工、校核、管理、分析、安全、归档等环节。外部运营先确认不同数据产品的运营主体，并明确其权责利。其次清晰运营模式，包括需求对接、合作方式、利益分配和监管机制等。然后封装或研发对外数据服务和产品，对外进行推销推广，推动数据要素价值流通。再次通过认证授权，确保生态相关参与方的数据访问范围和权限清晰、行为可追溯等。最后是生态保障和运营分析，使数据产品健康良好的持续发展。

服务层：包括面向公司部门数据共享的内部服务循环和面向社会的外部服务循环。内循环是确保安全可控前提下的数据共享应用和运营，重点关注数据评估、数据共享、数据应用、知识产权等环节的管理机制、路径和方法。外循环是开放的新兴数据要素市场形成和产业生态构建过程，重点关注数据服务、产业生态、数据流通、业务洞察、效果评估和需求反馈等环节。最后，内外服务循环通过稳定的管理机

制实现外部价值数据安全顺畅的回流反馈，内部数据优化服务及时有效的外推迭代，形成了一个大的业务服务闭环。

四、拓尔思数据要素业务介绍

（一）云和数据服务业务发展之路

拓尔思大数据中心自 2009 年开始筹备建设，2010 年正式成立，标志着公司开始布局数据服务。12 年来，拓尔思一直坚持探索“1+1=N”数据业务模式，持续研发了丰富的数据服务产品，目前已经形成了以服务政府、金融、媒体、新兴企业为主的行业 SaaS 产品集群。

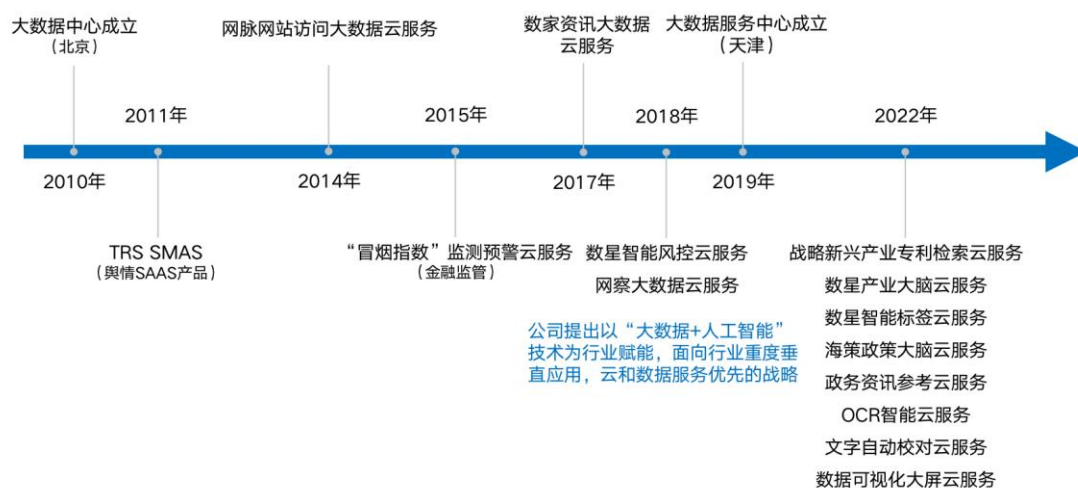


图 10：拓尔思数据业务发展之路

2011 年，拓尔思推出了第一款 SaaS 产品：TRS SMAS，这是一款面向政府、企事业单位、个人提供基于社交网络的舆情分析云服务。众所周知，2010 年是“微博元年”，人们更加熟悉的应用微博获取信息、沟通交流、休闲娱乐，政府部门也更趋向于通过微博发布政务信息、回应社会关切，微博对社会发展起到了越来越重要的作用。2011 年被人称为“微博红年”。以前铁道部部长刘志军涉嫌严重违纪被免职为标志性事件的微博反腐、以湖北官员冉建新非正常死亡为标志性

事件的微博维权、河南禹州天价过路费案引发的微博热议、湖北“死囚复生”谜案及洛阳李浩地窖囚禁6名女子案引发的微博追踪等都发生在2011年，均是通过微博在全国引发了巨大的网络舆情危机。拓尔思基于自研的网络信息采集产品和坚实的人工智能技术，迅速地研发了TRS SMAS推向市场。由于产品具备的社交数据更新及时、数据量覆盖较广、舆情分析功能较全、用户体验界面可视化可交互等特点，很快得到了高端用户的认可，如政府用户有外交部、质监总局、食药监局、最高法、卫生部、国家纪委等，企业用户有国家电网、移动、联通、佳能、神威药业、新华保险等。

2014年，拓尔思推出了第二款SaaS产品：TRS网脉大数据云服务，这是一款面向政府、媒体、企业等企业级客户提供专业的网站、APP、小程序分析服务，通过精准分析提高网站、APP、小程序友好性，帮助企业运营进而提升网络营销的有效性和便利性。该产品的推出，实现了存量的政府、媒体、企业用户的优势业务延伸，如外交部、国家气象局、全国总工会、陕西省政府、天津市政府、浙江日报、新疆日报、深圳报业、广发银行、邮储银行、海尔等都是网脉的核心用户。

2015年，拓尔思的全资子公司金信网银推出了“冒烟指数”监测预警平台。该平台是国内聚焦地方金融监管与金融风险防控，专门服务地方金融监管机构的大数据咨询平台，也是国内第一个非法集资监测预警平台。产品上线不久，在2015年年初公司监测预警了e租宝的风险，同时报送给相应的监管机构。受公安部的委托，公司的监测

预警平台也把风险企业报送给公安部。2016 年“冒烟指数”监测预警平台的业务推广开始走出北京，服务全国。

拓尔思 SaaS 产品每年保持健康有序的迭代，并随着公司对行业用户业务的深耕以及对国家政策信号的敏锐捕捉，几乎每隔一两年拓尔思都会推出新的数据服务产品。

拓尔思作为融媒体平台第一轮技术建设浪潮中的绝对主力，在 2016 年完成了以浙江日报报业集团“媒立方”为标杆的融媒体平台代表作后，精准地判断融媒体后时代的建设将从基础平台建设向媒体大数据服务转变，在 2017 年很快推出了数家资讯大数据平台。

2018 年初，公司提出坚持“大数据+人工智能”技术为行业赋能、“面向行业重度垂直应用”、“云和数据服务优先”的三大整体战略。同年推出了数星智能风控云服务、网察大数据云服务。

2019 年 4 月，公司成立了天津拓尔思数据服务中心，其定位是服务全集团，发展目标是成为集团核心的数据生产和服务基地，主要从事数据资产的采集、清洗、标引、加工和数据产品的研发，并为客户提供数据分析服务。

2022 年 9 月，拓尔思在秋季新品发布会上发布了 9 款 SaaS 服务产品，除虚拟人开放云服务外，其他 8 款产品均为数据产品。

（二）公司数据资产现状

拓尔思数据中心目前拥有 3 大专业 IDC 机房，上千台高端机架式服务器完成对互联网公开数据的 7*24 小时不间断采集，同时租用了

阿里云及微软云的服务器实现境外数据采集。数据中心监测的采集站点分类如下：

- ✧ 80 万+新闻网站
- ✧ 3000+资讯 APP
- ✧ 1200+国内报刊
- ✧ 20+三方平台
- ✧ 79 语种 50 万+境外站点

每日数据量高达 1 亿多条，几乎涵盖了全量的微博新增数据，网站覆盖度及数据存储量均明显高于业内水平。

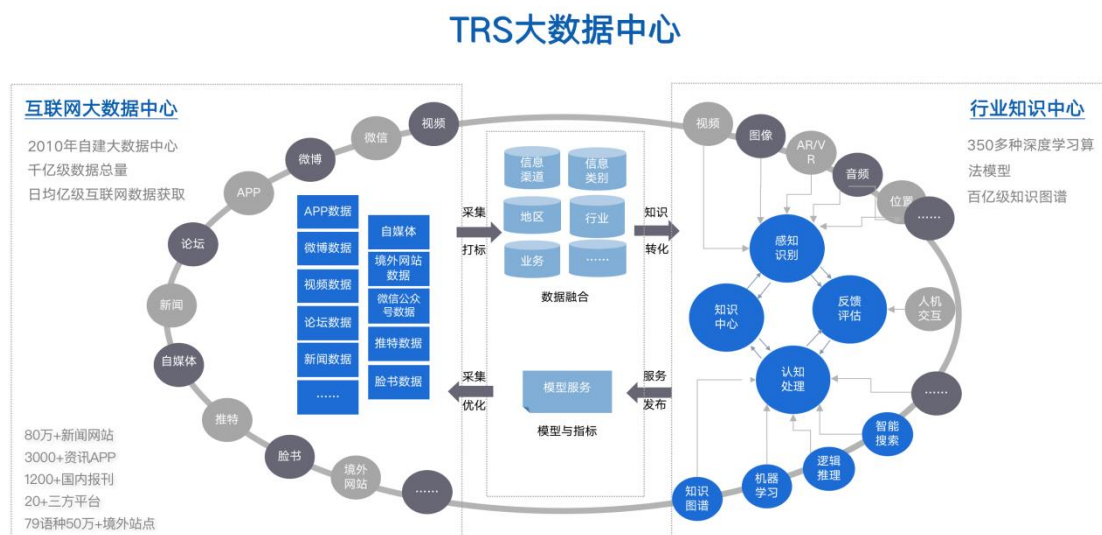


图 11：拓尔思数据中心总视图

1. 数据资产规模

（1）资源性数据资产（1300 亿）

拓尔思自建大数据中心，以长期服务多行业用户的开源数据应用持续累积为基础，拥有了规模及质量均位列业界前茅的、公开的资源性数据资产。目前，拓尔思拥有来自境内外、各行各业的互联网公开数据，总量超 1300 亿，具备数千亿数据量的数据索引、标记、查询、

拓尔思信息技术股份有限公司

挖掘分析能力,万亿级数据总量的秒级检索能力,日均亿级数据获取能力。拓尔思大数据中心不仅支撑公司多个专业化的数据智能服务,也具有对外的开放接口服务,可满足公司面向政府、媒体、金融、公安、商业等多行业用户的开源大数据云服务需求。

1300亿+公开类资源性数据资产



100亿+主题类资源性数据资产



图 12: 拓尔思资源性数据资产视图

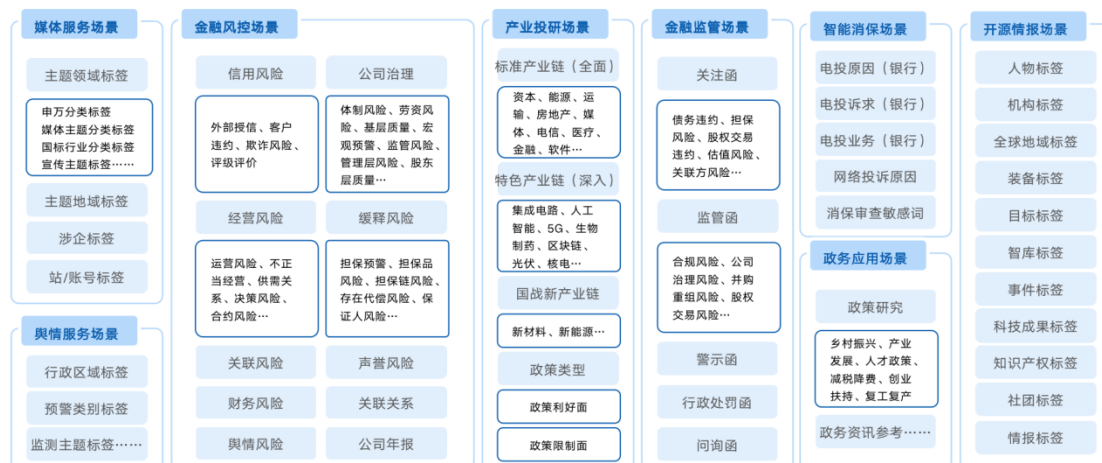
(2) 知识库 (30+大类)

机器认知的核心能力之一是自然语言理解,而机器理解自然语言需要足够的知识库支撑。拓尔思基于自然语言处理技术、积累的各场景行业术语及主题数据形成的各行业知识图谱,对海量多源异构数据

拓尔思信息技术股份有限公司

进行融合、关联、标注以及知识化处理,构建形成大规模领域知识图谱。目前,公司拥有通用、行业/领域知识库 30+大类,标签规则 30000+。每类知识库具有规模大、语义关系丰富、结构友好、质量精良等特点。

30000+标签覆盖8大业务场景



30+专业领域知识资产



图 13: 拓尔思知识库视图

(3) 模型工厂 (350+)

拓尔思通过大规模自监督学习的多模态人工智能算法,融合自然语言理解,基于行业的千亿级图文大规模预训练,实现用更小的标注

样本数量,融合更多的模态信息,形成了 350+深度学习算法模型的模型工厂,可面向不同行业、不同落地场景对数据进行知识模型的加工。

350+深度学习算法模型



图 14：拓尔思模型工厂视图

（4）三大经营性数据资产平台

基于上述资源性数据资产、知识库、模型工厂的融合治理，拓尔思公司目前已形成了产业大脑、媒体资讯、网络舆情三大数据资产平台，为公司在网络信息内容治理、金融监管、政务资讯服务、金融机构风控管理、智能媒体服务、开源情报等领域的云和数据服务提供数智赋能。



2. 大数据底座产品清单

产品名称	适用行业	业务场景	业务简介
海蜘蛛大数据采集平台	通用	网络数据采集	根据互联网不同的数据类型与网站结构，采用分布式抓取技术，对指定网站定向数据采集、存储，然后进行分类、分词、查重、打标等数据加工，在专业知识库建立、企业竞争情报分析、报社媒体资讯获取、舆情监测等领域应用广泛。
海聚数据融合平台	通用	数据资源汇聚	采用分布式架构，以 flink 作为数据处理引擎，可拖拽式数据流程开发，提供几百种常用功能组件，支持对任务的进度实时监控和多维度统计分析，并提供了全功能 RESTful API。“海聚”能帮助企业打破数据孤岛、修正数据质量、汇集多数据源、转换数据形态，提升数据的利用价值。
海贝搜索数据库	通用	搜索引擎支撑	“海贝”是一款从内核到系统完全国产自研的搜索引擎数据库，其定位为大数据应用支撑软件，为大数据应用提供高效的数据存储、信息检索、统计分析等数据管理服务，支持几乎所有类型的数据，包括文本、数字、地理空间、图片等各种结构化、半结构化和非结构化数据。该系统吸收了拓尔思在信息检索和 NLP 领域多年的技术积累，不仅安全、高效、专业，还集成了众多开源检索系统不具备的企业级功能，能够有效降低用户的系统建设成本、开发成本和运维成本。
天骄数据中台	通用	大数据治理 数据质量管理	帮助大数据应用客户处理各种多源异构数据，完成数据治理，沉淀数据知识，控制数据质量，实现数据采集、数据接入、数据处理、数据存储、数据应用和数据智能，成为数据资产管理中枢，让客户从数据泥潭中解放出来，专注于业务应用和创新。

智拓人工智能技术平台	通用	AI 处理	“智拓”作为拓尔思新一代基于深度学习的语义智能技术平台，围绕自然语言理解（NLP）、知识图谱、图像检索三大核心技术，结合机器人流程自动化、数据中台等技术，面向媒体融合、智慧专利、智慧公安、智慧政务、智慧金融、开源情报分析等应用场景，提供文本、音视频、多模态等全栈 AI 服务能力，可以帮助用户快速构建自己的 AI 技术能力，有效提升行业语义智能的服务能力和效果。
------------	----	-------	---

表 1：拓尔思大数据底座产品清单

3. 数据安全产品清单

产品名称	适用行业	业务场景	业务简介
新一代安全一体化平台	通用	网络安全	提供数据双向隔离交换、数据单向隔离交换、音视频跨网调取、跨网接口调用、边界运维管理、边界态势感知、级联对接、异常流量分析、业务审计、访问控制、安全评估、安全咨询、安全运维等一系列具有自主特色的安全一体化能力平台，实现“全面防护，智能分析，自动响应”的防护效果。

表 2：拓尔思数据安全产品清单

4. 云和数据服务产品清单

产品名称	适用行业	业务场景	业务简介
数家资讯大数据云服务	通用	内容资讯	面向各行业用户提供个性化的全网资讯数据分析和知识服务。
网察大数据云服务	通用	舆情	从网络舆情线索发现、实时预警、分析研判、综合报告等环节为用户提供全面、及时、准确的服务。聚合网络新闻、微博、微信、境外媒体等多渠道海量数据，以信息监测、态势感知、关系挖掘、事件追踪、传播分析、智库共享等为目标，解决跨舆论场的网络信息洞察、分析、研判，实现从数据到智能决策的一体化大数据舆情服务。
海策政策大脑云服务	政府	政策分析	围绕政策制定、发布、执行、兑付、评价、监督等全生命周期流

产品名称	适用行业	业务场景	业务简介
			程，坚持问题导向、效果导向，通过运用大数据、知识图谱、语义智能、虚拟人等技术，构建了全国政策文件库，并提供一体化管理与智能分析的在线服务。
政务资讯参考云服务	政府	政务资讯	面向党政机关、企事业单位的秘书工作者提供全网政务资讯一站式云服务。旨在帮助秘书工作者及时了解最新政务动态、政策发布、行业趋势等，掌握一手前沿政务资讯。
“冒烟指数”监测预警平台	政府金融监管机构	金融监管	是国内聚焦地方金融监管与金融风险防控，专门服务地方金融监管机构的大数据咨询平台，也是国内第一个非法集资监测预警平台。
数星产业大脑云服务	政府、金融、产业园	产业招商区域产业分析	以产业数据中心为支撑，以智能数据标签引擎和全产业链知识引擎为核心底座，运用新一代信息技术，推动产业数字化、智能化升级，实现政策、空间、供应链、金融、科技、销售等产业要素与产业主体之间的高效协同，为企业生产经营提供数字化赋能，为产业生态建设提供数字化服务，为经济治理提供数字化手段。
数星智能风控云服务	金融	风险管理	提供智能多维标签、预警信号推送、风险事件跟踪、风险传导关联、综合异构图谱等贯穿多场景全流程风险管理功能，提升金融机构风控能力，在风控、合规、投研、监管等核心金融场景得到广泛应用。
战略新兴产业专利检索云服务	战略新兴产业	专利服务	与知识产权出版社联合开发，聚焦国家战略新兴产业发展需求，在强化搜索引擎技术与数据资源优势基础上，共同构建的具备自主可控核心能力的专业领域数据库。 一期全球基因专利检索数据库已上线。
网脉网站访问大数据云服务	通用	网站访问大数据分析	面向政府、媒体、行业企业级客户提供专业的网站、APP、小程序分析服务，通过精准分析提高网站、APP、小程序友好性，帮助企业运营进而提升网络营销的有效性和便

产品名称	适用行业	业务场景	业务简介
			利性。
OCR 智能云服务	金融、电力 政府、公安	文字识别	基于深度学习算法，专注复杂业务场景下将图片上的文字内容，智能识别成为可编辑的文本。支持通用文字识别、个人证照识别、票据凭证识别，金融与电力、政府、公安等行业特殊票据和办公文档的文字识别。支持提供定制化服务，可以有效地代替人工录入信息，满足各种客户的图片识别需求。
文字自动校对云服务	通用	文字校对	融合了拓尔思在自然语言处理和信息检索领域多年的技术积累，围绕深度学习、知识图谱等核心技术，为公文编辑、新闻出版、媒体内容审核等多种场景提供智能化、自动化的文本校对服务。目前，产品已经在公文校对、新闻检测、辅助写作、OCR 识别错误检测等场景广泛应用。
数据可视化大屏云服务	通用	数据可视化	提供零代码拖拽式的可视化大屏制作，以组件素材库、交互设置、数据接口等作为核心，聚焦多方面的数据图表应用，为管理者提供全面数据化、科学化的决策分析。目前，拓尔思数据可视化大屏云服务已在政务、教育、公检法等领域成功应用。
数星智能标签云服务	通用	数据标注	针对非结构化文本数据、面向多维度业务场景的知识挖掘引擎，以解决认知智能面向领域模型训练过程中样本数据积累难、技术门槛高、效率低等问题，敏捷化、轻量化服务于各垂直领域非技术型业务专家，快速构建领域分析模型。

表 3：拓尔思云和数据服务产品清单

（三）大数据底座和人工智能产品介绍

1. 网络大数据采集

网络大数据采集是指根据互联网不同的数据类型与网站结构，采用分布式抓取技术，对指定网站定向数据采集、存储，然后进行分类、

分词、查重、打标等数据加工，在专业知识库建立、企业竞争情报分析、报社媒体资讯获取、舆情监测等领域应用广泛。

拓尔思大数据中心的数据基于自研的互联网数据采集系统实现全网数据采集的。早在 2002 年，拓尔思就推出了第一代网络信息采集系统——TRS 网络信息雷达（以下简称“雷达”），已成功应用于数千家媒体、企业、金融和政府部门的数据中心或信息中心。随着互联网的发展，以及 web2.0 技术的兴起，各种风格迥异的网站、新闻 APP 等风云涌现，各种页面展现技术层出不穷，比如：异步加载、延时加载、下拉刷新、手机 APP、登录验证、参数时效性、访问屏蔽等等，“雷达”产品面临新技术的挑战。2020 年，拓尔思推出了 TRS 海蜘蛛大数据采集平台（以下简称“海蜘蛛”），这是基于“雷达”开发出来的新一代分布式大数据采集系统。“海蜘蛛”从采集点的维护到采集终端的部署、源码下载、内容解析，再到采集结果的存储都是集群式+分布式的，平台内嵌了雷达的内容和链接智能过滤模块、脚本引擎模块、内容模板和链接模板等等。把雷达采集流程按照模块进行拆分，细分为采集点管理、采集配置管理、采集属性管理、链接更新检查、源码下载（雷达下载、基于浏览器下载、手机模拟下载、插件下载、API 下载）、链接过滤和内容抽取、采集日志管理、解析结果分发几个独立的模块，任意模块均可集群式部署。

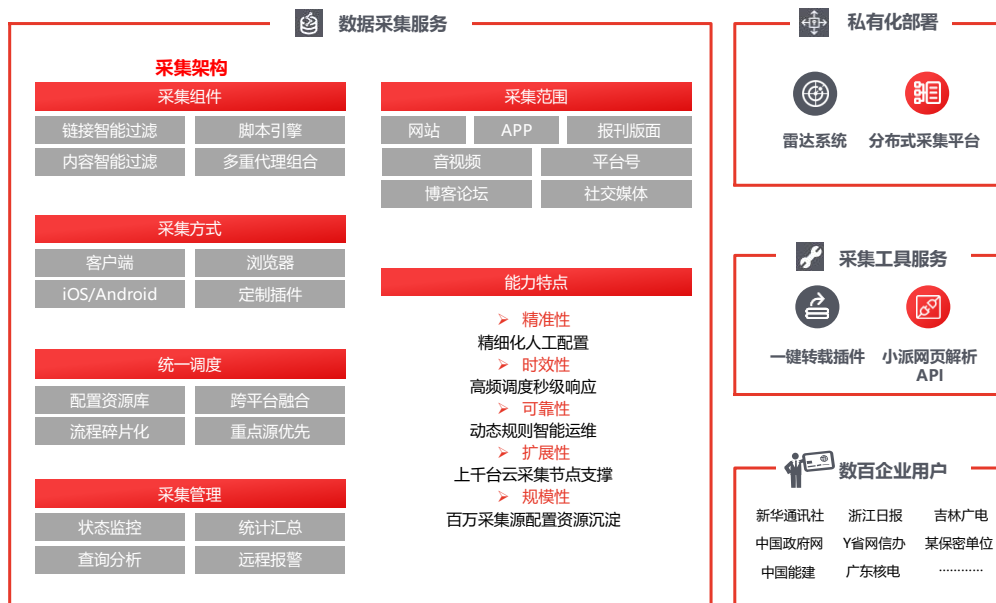


图 16 “海蜘蛛”总体架构

主流社交媒体和小视频相对传统网站，是舆论新阵地，信息密度高。由于官方网站提供的 API 存在限制数据采集速度和身份验证等问题，常规的网络爬虫存在一定的局限性，采集门槛更高。

针对这一新形势，拓尔思开发出了全新的海汇主流社交媒体和小视频采集平台（以下简称“海汇”），集成了拓尔思丰富的协议分析和分布式采集经验，实现了海量数据的稳定高效采集，覆盖了国内外常见的多种社交媒体和小视频站点。针对小视频，“海汇”还拥有文本快速检测和视频动态 OCR 等关键技术，可以实现视频内容的深入挖掘。“海汇”管理着数千个采集节点，通过对大规模采集过程中的任务调度、智能纠错、反爬对抗和验证码识别等技术的工程化应用，让海量公开数据的实时监控成为可能。“海汇”模拟自然人的行为特征、内容特征，进行拟人化帐号养殖。设计社交网络智能体时考虑包括访问、发布内容、评论内容、对话等行为的时间，基于智能写作和智能对话生成的内容特征，进行反检测以维持

社交网络智能体不被屏蔽，同时模仿社交网络智能体的行为特征进行采集任务配置与调度监控，以实现基于社会计算的数据采集技术。“海汇”的采集点全球各大洲都有覆盖，在阿里云、华为云、AWS、微软云、VPS 等供应商云主机验证通过，且对多语种都能够适配。

2. 数据资源汇聚

数据资源汇聚就是从源端经过抽取(extract)、转换(transform)、加载(load)至目的端的过程，它能够对各种分布的、异构的源数据(如关系数据)进行抽取，按照预先设计的规则将不完整数据、重复数据以及错误数据等“脏”数据内容进行清洗,得到符合要求的“干净”数据,并加载到数据仓库中进行存储,这些“干净”数据并加载到大数据管理平台中进行存储，这些“干净”数据就成为了数据分析、数据挖掘的基石。

海聚数据融合平台（以下简称“海聚”）是一款由拓尔思自主设计研发，具有强大数据集成能力的流批一体数据融合平台。“海聚”采用分布式架构，以 flink 作为数据处理引擎，可拖拽式数据流程开发，提供几百种常用功能组件，支持对任务的进度实时监控和多维度统计分析，并提供了全功能 RESTful API。“海聚”能帮助企业打破数据孤岛、修正数据质量、汇集多数据源、转换数据形态，提升数据的利用价值。

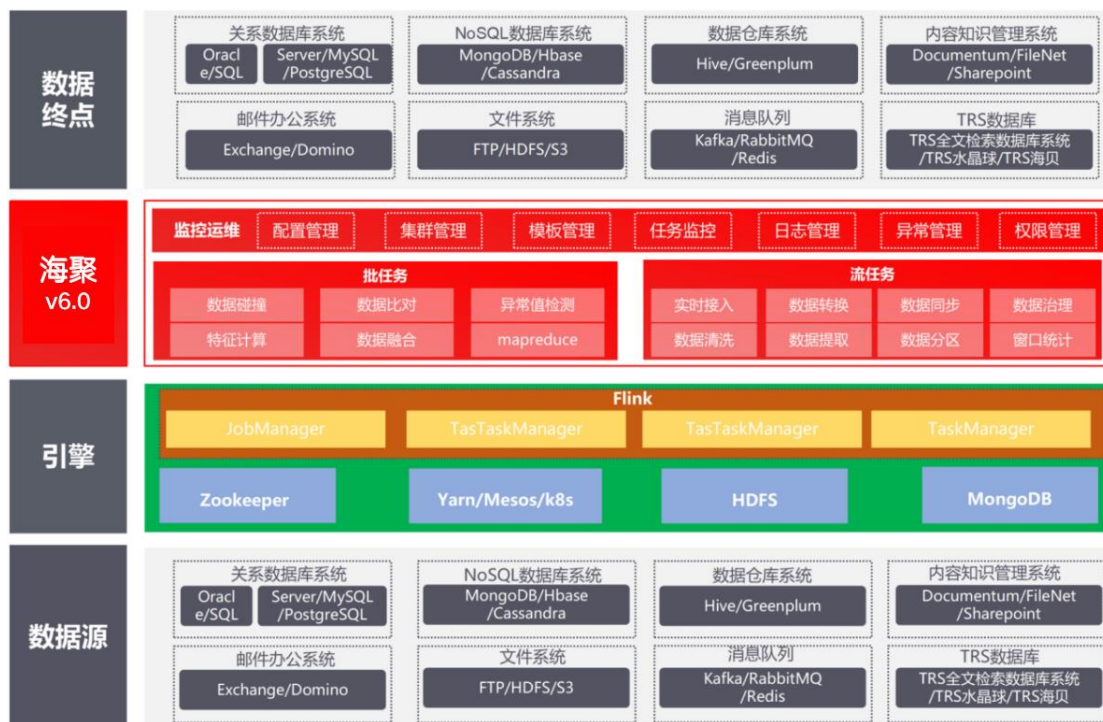


图 17 “海聚”总体架构

“海聚”支持多种任务类型，包括批任务、流任务、SQL 任务、JAR 任务。通过调度计划用于配置任务的定时运行策略，合理的给任务分配调度计划，可以使得多个任务协同运行并且合理使用服务器资源，避免出现过多任务争抢线程和数据库等资源以及大量任务排队等待等现象，为服务器稳定运行提供保障。“海聚”支持集中管理任务中使用的所有外部连接的资源，并生成相应的血缘图谱。“海聚”具体的应用场景如下：

◇ 全文检索场景

抽取、清洗、整合海量信息数据，为全文检索功能的重要组成部分——全文数据库提供优质数据。



图 18 “海聚”全文检索应用场景

◇ 数据治理场景

帮助组织建立数据质量管理体系，提升数据的完整性、规范性、及时性、一致性、逻辑性。



图 19 “海聚”数据治理应用场景

◇ 数据迁移场景

实现新旧系统与数据库切换的数据接入与数据格式匹配，确保应用系统顺利割接。



图 20 “海聚”数据迁移场景

◇ 数据整合场景

实现多源异构数据整合，加强数据管控能力，消除信息孤岛，更好的为上层应用提供服务。



图 21 “海聚”数据整合场景

◇ 数据共享交换场景

用于构建、运行和管理分布式应用系统间、云环境下应用系统之间等数据交换共享任务，以满足各种大型应用、各种复杂的网络环境下的业务需求，尤其适用于跨部门、跨地域、跨层级的数据交换共享应用。

3. 大数据管理

海贝搜索数据库（以下简称“海贝”）是一款从内核到系统完全国产自研的搜索引擎数据库，其定位为大数据应用支撑软件，为大数据应用提供高效的数据存储、信息检索、统计分析等数据管理服务，

支持几乎所有类型的数据，包括文本、数字、地理空间、图片等各种结构化、半结构化和非结构化数据。“海贝”不仅集成了众多开源检索系统不具备的企业级功能，还能够有效降低用户的系统建设成本、开发成本和运维成本。

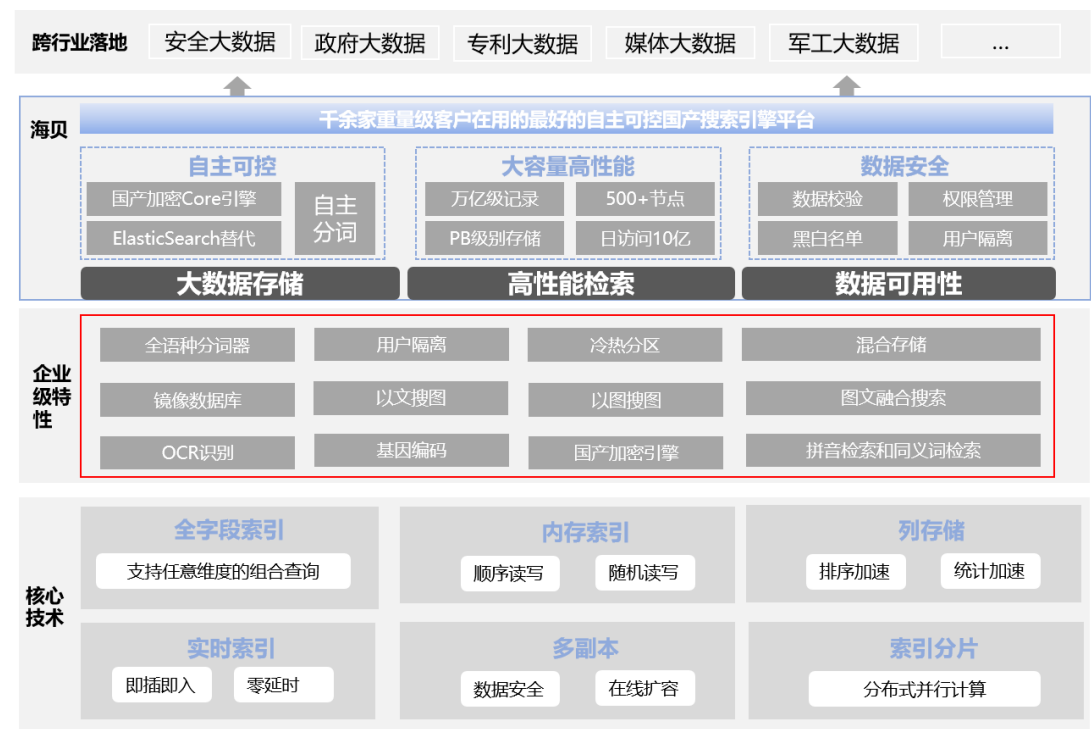


图 22 “海贝”总体架构

✧ 全语种分词器

“海贝”自带的 TRS 分词器，不仅可以支持中日韩等方块文字，也能处理英法德等拉丁语系，还可以处理藏文、蒙文、维文等少数民族语言，真正可以做到单一分词器处理全部语种，大大简化了系统的开发和运维工作。

✧ 用户隔离

“海贝”还具有其他同类产品不具备的用户隔离安全特性，系统支持用户数据的物理隔离和逻辑隔离，如图所示，模式 1 与模式 2 实现逻辑隔离，模式 1 与模式 3 实现物理隔离，这样既可以解决多租户

之间的访问安全，又可以防止用户之间的相互影响，提高系统的可用性。

◇ 冷热分区

海贝自带的分时归档视图，不仅可以实现冷热数据自动分区，同时还支持多种存储混合使用以提供高效的检索服务。

◇ 镜像数据库

用户可以通过简单的配置实现读写分离、大小库以及访问隔离等，大大降低了数据加工和查询之间的 CPU、内存、IO 等冲突，避免了数据入库导致的检索卡顿问题。

◇ 深度检索

随着互联网审查的愈发严格，有人经常将一些敏感词转成同音词或者同义词进行发布，或将非法文章通过图片的形式对外发布，企图躲避监管。针对这种情况，海贝集成了深度学习引擎，支持以文搜图、以图搜图、图文融合搜索，可以对图像中的文字进行 OCR 识别，也可以提取图像或者文本的特征数据，通过基因编码存储到海贝数据库里，成功实现图像相似性检索功能，并配合拼音检索、同音词检索等功能，让网络上的不法分子无处遁形。

4. 数据中台

随着互联网的高速发展，业务场景的不断增加，数据应用的需求越来越多，为快速响应业务的需求，很多企业都不同程度的存在烟囱式的开发模式，这种烟囱式的开发导致企业不同业务线的数据是割裂的，这就造成了数据的重复加工，导致研发效率、数据存储和计算资源的浪费，使大数据的应用成本越来越高，也带来指标口径不一致的

问题。产生这些问题的根源在于数据无法共享,为解决这一问题,2016年,阿里率先提出“数据中台”的口号。数据中台的核心是:避免数据的重复加工,通过数据服务化,提高数据的共享能力,赋能数据应用。

总的来说,数据中台具备异构数据统一计算、存储的能力,同时让分散杂乱的数据通过规范化的方式管理起来。数据中台借鉴了传统数仓面向主题域的数据组织方式,基于维度建模理论,构建统一的数据公共层和应用层。数据中台依赖于大数据平台完成数据研发全流程,同时增加了数据治理和数据服务化以及数据资产内容。数据中台是大数据时代的大背景下,为实现数据快(快速)、准(准确)、省(低成本)赋能业务发展的目标,将企业的数据统一整合起来,基于Onedata方法论借助大数据平台完成数据的统一加工处理,对外提供数据服务的一套机制。

拓尔思天骄数据中台(以下简称“天骄”)帮助大数据应用客户处理各种多源异构数据,完成数据治理,沉淀数据知识,控制数据质量,实现数据采集、数据接入、数据处理、数据存储、数据应用和数据智能,成为数据资产管理中枢,让客户从数据泥潭中解放出来,专注于业务应用和创新。



图 23 “天骄”总体架构

◇ 数据标准

可根据业务需要无限制制定领域内需要的数据标准，包括各种标准数据元，由一组属性规定其定义、标识、表示和允许值的数据单元。以及元数据，由数据元构成的数据对象，为数据交换，数据质量，数据管理和利用打下基础。

◇ 数据接入

支持各种多源异构数据接入到数据中台，数据可通过 HTTP/HTTPS 接口推送，或各种消息中间件推送，或各种关系型数据库，非关系性数据库接入数据，并完成数据质量检查，标准数据转换等工作。

◇ 数据处理

支持批流一体化对接入的数据和历史数据进行定制化处理，涵盖结构化和非结构化的数据处理；涉及常规的数据清洗，数据打标等传统数据处理能力，还支持自然语言处理和图像，音频，视频处理；除实时数据处理外，也支持历史数据处理；同时还支持自定义业务模型进行数据处理。

◇ 数据存储

支持原始数据和处理后的数据，直接存储到关系型数据库、非关系型数据库、全文检索数据库、OLAP、OLTP 数据库以及图数据库，并对整个存储进行实时监控和管理，严格控制数据访问安全。

◇ 数据应用

提供统一的数据应用管理功能，支持全库的检索和关联分析，支持基于应用系统定制化访问数据，支持字段级别的数据使用控制。通过可视化定制场景的方式配置数据应用，降低使用难度。

5. 数据质量管理

数据质量是数据治理的一项重要工作，数据质量低下会导致不正确的信息，不利于组织提升业务业绩。高质量的数据需要有效的数据质量管理工作提供保障。

数据质量管理工作包括校正数据，以及管理数据创建、数据转换和数据传输的整个生命周期。数据质量管理是一个持续的过程，要将数据质量管理和质量提升等流程制度化。同时，在数据处理过程中，要采取措施校正发现的问题，并消除问题的根源。数据质量管理包括数据质量分析、识别数据异常、定义业务需求及相关业务规则，在必要的时候对已经定义的数据质量规划进行合规性检查和监控，以及数据解析、标准化、清洗和整合。同时还包括问题跟踪，对已经定义的数据质量服务水平、合规性等进行监控。

拓尔思“天骄”提供一套完整的数据质量管理体系。

◇ 数据标准

数据标准是指保障数据的内外部使用和交换的一致性和准确性的规范性约束。其核心目的是通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，实现企业大数据平台数据的完整性、有效性、一致性、规范性、开放性和共享性管理，为数据资产管理活动提供参考依据。

数据标准是数据质量管理非常重要的一环，天骄数据中台涉及到的核心功能主要是定义数据元和元数据。

为了完整表示各种场景下的数据，在天骄数据中台中，可以定义六大类属性：标识类属性，定义类属性，表示类属性，管理类属性，关系类属性，融合类属性。

添加数据标准

标识类属性

* 中文名:

请输入

英文名:

请输入

中文全拼:

请输入

* 标识符:

请输入

语境:

请输入

同义名称:

输入后回车

定义类属性

* 说明:

请输入...

对象类词:

特性词:

应用约束:

请输入...

表示类属性

* 数据类型:

* 表示词:

值域:

请输入

归一化标识:

是

否

计量单位:

请输入

管理类属性

* 状态:

* 提交机构:

请输入

注册机构:

请输入

主要起草人:

请输入

批准日期:

请选择日期

关系类属性

分类方案值:

关系:

融合类属性

融合单位类型:

请输入

融合单位数据标准编码:

请输入

关系类属性

备注:

请输入...

图 24 “天骄”添加标准管理界面

元数据管理的范围将涵括数据产生、数据存储、数据加工和展现等各个环节的数据描述信息，帮助用户理解数据来龙去脉、关系及相关属性。按其描述对象的不同可以划分为三类元数据：业务元数据、技术元数据、和操作元数据。

在天骄数据中台中，除了可定义以上三种不同类型的元数据外，还可以根据需求自定义更多类型的元数据，例如：

The form contains the following fields and options:

- * 数据标准:** 数据记录敏感级别 (dropdown), 1.0 (dropdown)
- 限定词:** (dropdown)
- * 中文名:** 数据记录敏感级别
- 英文名:** sjjlmgb
- * 元数据标识符:** DD00000076
- * 数据项标识符:** DE90009997
- 是否原始字段:** ☒ 是 ☐ 否
- 是否查询条件:** ☒ 是 ☐ 否
- 是否订阅条件:** ☒ 是 ☐ 否
- 是否必填:** ☒ 是 ☐ 否
- 是否多值:** ☐ 是 ☒ 否
- 查询/订阅匹配方式:** 请选择查询/订阅匹配方式 (dropdown)
- ? 字段敏感度级别:** 98
- 字段分类:** (dropdown)
- 备注:** 请输入...
- Buttons:** 取消, 确定

图 25 “天骄”元数据自定义界面

◇ 数据源管理

除数据标准外，数据采集、数据接入、数据处理、数据存储等环节，对数据质量控制是至关重要。对不同数据来源进行严格的控制和管理，对原始数据标准化的定义和描述，以及原始数据约束和质量检查等，都属于数据源管理范畴。

多源数据接入：为了满足接入更多不同渠道，不同性质的数据，天骄数据中台支持多渠道，多文件格式的数据接入，包括文件，数据库，消息中间件。

数据规范化：不同数据源的同一类数据，接入平台后，可规范化为同一类数据，满足相同的数据标准，且支持不同数据源有不同的数据定义，并进行智能数据转换。

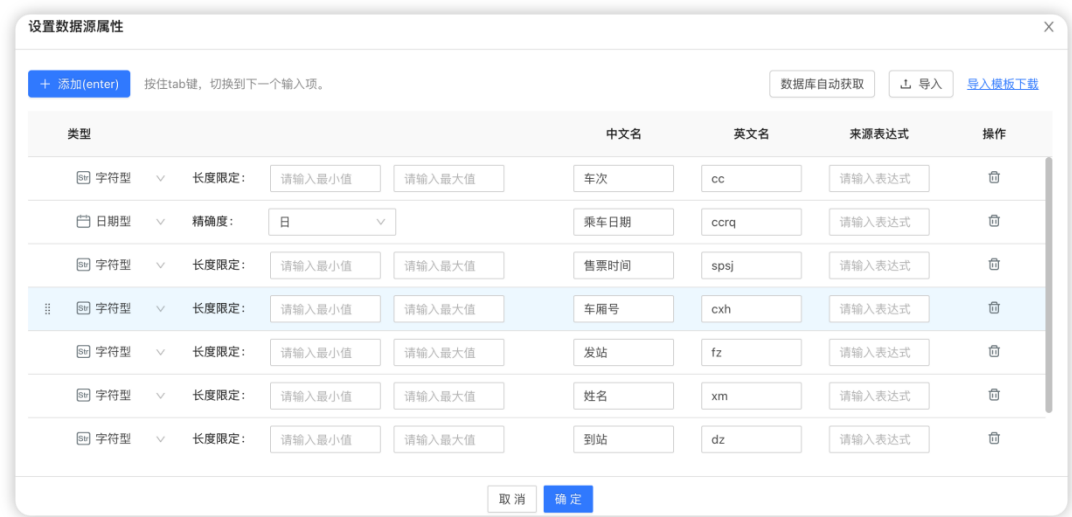


图 26 “天骄”数据规范化界面

数据对账：为了及时有效地发现问题，天骄数据中台还提供了数据对账功能，确保数据准确无误的接入到了数据中台中。

◇ 数据全流程监控

数据接入、数据处理、数据存储的每一个环节和细小步骤，都有细化到毛细血管级别的监控，包括每一步的输入和输出，有异常也会及时提醒和报警。

对任意数据的处理和质量问题，可点击查看全链路跟踪记录：



图 27 “天骄”全链路跟踪记录界面

◇ 数据血缘

天骄数据中台具备数据血缘功能。数据血缘是指数据产生的链路。直白点说，就是我们这个数据是怎么来的，经过了哪些过程和阶段。举个例子，比如在生产系统如淘宝网中，客户在淘宝网页中购买物品后，数据就被存到后台数据库表 A 中。当我们领导需要查看某个月卖的最火的是哪些物品时，我们需要对存入的这些数据进行加工汇总，形成一张新的表 B 来存储我们处理的数据，最后我们会根据 B 表进一步处理成我们前台展现使用的表 C。那么 A 表是 C 表数据最初的来源，是 C 表数据的祖先。从 A 表数据到 B 表数据再到 C 表数据，我们认为这条链路就是 C 表的数据血缘。

在数据的处理过程中，从数据源头到最终的数据生成，每个环节都可能会导致我们出现数据质量的问题。比如我们数据源本身数据质量不高，在后续的处理环节中如果没有进行数据质量的检测和处理，那么这个数据信息最终流转到我们的目标表，它的数据质量也是不高的。也有可能在这个环节的数据处理中，我们对数据进行了一些不恰当的处理，导致后续环节的数据质量变得糟糕。因此，对于数据的血缘关系，我们要确保每个环节都要注意数据质量的检测和处理，那么我们后续数据才会有优良的基因，即有很高的数据质量。

6. 人工智能技术平台

从数据类型来分，大数据包括结构化和非结构化数据。非结构化数据特指公文、研究报告、音视频、设计图、工作照片等各种文件等。这些与数据库里的结构化数据不一样，这些数据的治理难度更大、挑战更多，如没有统一存储（各种附件，各种微信传输）、没有统一标准（都是各自写的文件）、数据种类特别多（除了结构化的，都是非结构化、半结构化的）、分散管理（各处室、各类文档未能统一管理，分散在各个不同应用和终端）。对于结构化数据，我们知道要盘点、要做标准、要整主数据、要梳理指标、要做质量控制，因为我们知道数据就在哪些目标库里。但是非结构化数据不一样，首先很难对数据全量进行盘点，因为这些文档分散在档案管理室、知识中心、各业务系统还算好的，无论是否内容和数量齐全，总归有个集中的地方。但是更多的都是各自散落存储在 OA、邮箱、云盘、个人存储等。所以非结构化数据治理的第一步是对非结构化数据的分布进行梳理，了解哪些是数据治理重点。第二步就是选定一个合适的场景，由应用场景牵引，跟数仓建设逻辑一样，自下而上建设见效快。然后再快速收集部分数据，用 NLP 等技术将非结构化数据结构化，然后再

拓尔思信息技术股份有限公司

利用数据库、大数据、图计算等技术处理数据，做出几个能看到效果的应用示范。所以在大数据治理中，非结构化数据的治理与 AI 技术紧密融合。

智拓人工智能技术平台(以下简称“智拓”)作为拓尔思新一代基于深度学习的语义智能技术平台，围绕自然语言理解(NLP)、知识图谱、图像检索三大核心技术，结合机器人流程自动化、数据中台等技术，面向媒体融合、智慧专利、智慧公安、智慧政务、智慧金融、开源情报分析等应用场景，提供文本、音视频、多模态等全栈 AI 服务能力，可以帮助用户快速构建自己的 AI 技术能力，有效提升行业语义智能的服务能力和效果。

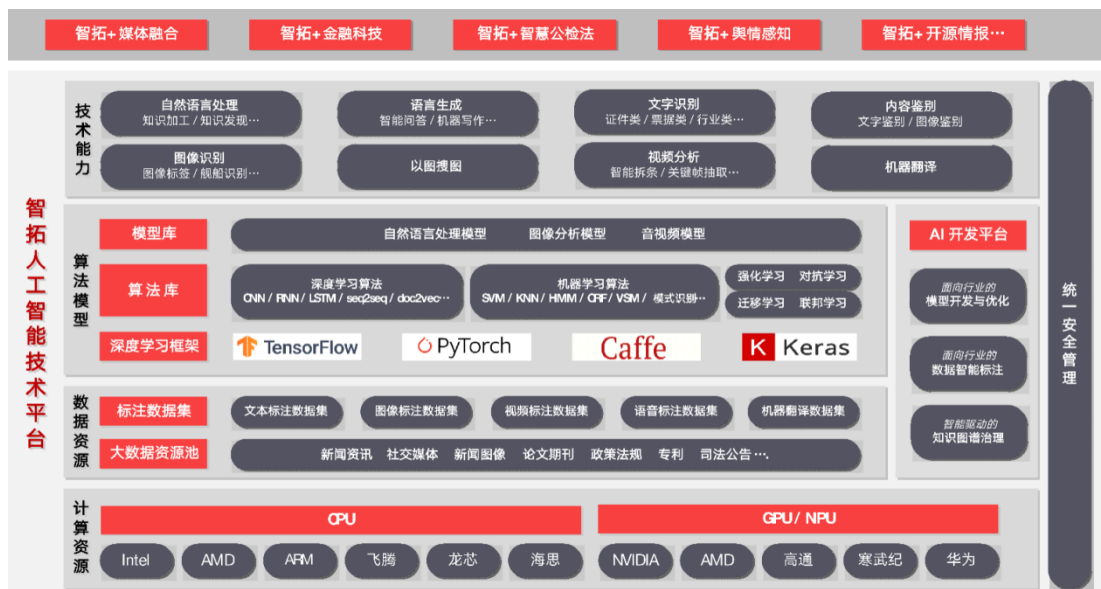


图 28 “智拓”总体架构

◇ 一站式 AI 开发

围绕行业应用构建一站式的 AI 能力平台，提供 AI 开发全生命周期管理功能，包括模型设计、训练、优化、评估、发布、下载等完整过程；提供全流程的图形化操作界面，支持可视化建模和 Notebook 建模，用户无编程或低代码编程即可完成算法模型训练及使用，降低

开发门槛、提升开发效率。生成的 AI 模型可以方便地发布到智语、智眼等产品中，有效地提升相关产品的行业应用效果。

✧ 智能数据标注

针对 AI 训练需要的大量标注数据的难题，面向行业的人工智能定制优化平台可以借助数据智能标注平台，向用户提供文本、图像、视频等数据的标注服务，包括机器预标注、可视化标注、标注效果评估等功能。用户可以构建自己的业务标注数据集，生成专属的行业 AI 模型。

✧ 模型优化

支持 AutoML，提供深度学习模型及传统机器学习的优化管理工作，包括参数调优设置、神经网络模型压缩、增量训练等功能。

✧ 知识驱动

深度学习的主要驱动力来自人工标注的大数据和高性能算力。知识图谱作为高质量的标注数据，对机器学习也是重要的信息源。将行业知识图谱等知识资源应用于深度学习中，将大数据驱动的深度学习神经网络加入知识驱动，通过多模型的集成学习，提升认知计算效果。另一方面，将深度学习成果用于知识发现，形成良性循环。

（四）数据安全产品介绍

1. 网络安全市场

IDC 预计，从行业维度来看，到 2026 年政府、金融、运营商仍将引领整体网络安全行业支出。除此之外，2021 年，中国“十四五”规划中明确指出建设制造强国的目标，在此背景下，中国工业互联网、

工控相关的市场开始进入快速增长期。未来五年，制造业（包含离散制造、流程制造）将成为中国网络安全支出增速最快的行业市场，五年复合增长率将达到 22.8%，到 2026 年，制造业 IT 安全支出预计将达到 22.2 亿美元。

积极寻求新机会、转危为机

2020 年以来，疫情肆虐全球，各国经济都遭受了不同程度的影响。与此同时，各地区冲突频发，地缘政治、供应链等问题层出不穷，大环境给予了网络安全厂商巨大的挑战。对于技术服务提供商来说，积极寻求新机会刻不容缓。例如，疫情催生了大批远程办公、远程教育等“远程+”场景的新需求，零信任网络访问在这个场景下迎来了市场的高速增长期。

软件化趋势增强，厂商云化、虚拟化、SaaS 化能力亟需提升

近几年，伴随云计算的快速发展，云基础设施部署越来越多，云上安全能力的建设需求成为了上云客户的刚需。公有云、专属云、边缘云、混合云上的安全一直以来都是最终用户云建设过程中最大的担忧。在此背景下，云安全市场将快速发展。厂商需要积极应对软件化趋势，提升其产品的虚拟化、云化、SaaS 化能力，从而抓住下一个五年安全市场的发展机遇。

聚焦重点行业，因地制宜

政府、金融和通信长期是网络安全支出的前三大行业，无论是出于合规还是业务要求，其对于网络安全的需求是持续且稳定的，厂商需要在这三大行业上进行稳定投入。除此之外，中国制造强国的目标

带动了制造业在安全领域的投入，“关基”条例的加持使得能源、公共事业等行业支出也在稳定增加。聚焦短期高增长行业和地区，因地制宜可以更好的发挥营销效果，实现增长。

深挖用户需求，持续提升产品和服务能力

近几年，网络攻击频发使得越来越多的用户采购网络安全产品和服务不仅是为了合规，更是需要其部署的产品和服务可以真正解决其所面临的安全挑战。用户对于安全产品和服务要求的进一步提高、新场景、新技术的持续发展都需要厂商不断提升其产品和服务能力。具体而言，提升产品性能、功能，增加自动化、智能化等能力，增强服务标准化、制度化、流程化、知识化建设等方面都需要厂商重点关注。

近几年，伴随中国网络安全相关政策监管体系的不断完善、企业业务稳定发展运行需求的不断增加，中国网络安全市场在疫情、地缘政治等因素的影响下仍实现了较为稳定的增长。未来，网络安全市场将向“软件化”、“服务化”进一步发展。

2. 产品服务模式

拓尔思全资子公司天行网安是国内最早从事网络安全和数据交换的企业，发明了国内第一台安全隔离网闸。经过二十多年的发展，已形成了以数据交换为核心的边界安全、物联网安全、大数据安全三大阵营产品线和解决方案。

虽然在传统安全建设思路的指导下，企事业单位或许已经购买了大量的安全设备、产品，但这些更多的只是为了满足合规的需求，如同孤岛般的安全设备无法满足企业真实的安全建设需要。只有将这些

孤岛一个个连接起来，以安全运营的理念去做好顶层向下的体系规划，对整个目标进行一体化安全管理，才能够真正建设好企事业单位坚实的城墙。针对安全管理者看不见、摸不清安全威胁的现状，天行网安未来将为用户提供一套一体化安全管理思路，逐步形成了数据双向隔离交换、数据单向隔离交换、音视频跨网调取、跨网接口调用、边界管理、边界态势感知、级联对接、异常流量分析、业务审计、访问控制、安全评估、安全咨询、安全运维等一系列具有自主特色的安全一体化能力平台，实现“全面防护，智能分析，自动响应”的防护效果。



图 29：拓尔思安全一体化平台

天行网安面向政府及公共安全市场，产品和安全解决方案应用于公安、检法、海关、金融等行业，头部用户有全国公民身份证号码查询服务中心、国家发改委、公安部、公安信息网、最高法、海关总署以及 17 个省级、直辖市公安厅（局）等，全国服务了 4000+用户，市场份额居于前列。

目前天行网安形成了以北京总部为中心，全国 31 个分支机构为基点，辐射全国的营销及服务体系。公司商业模式相对稳定，公司在行业内地位相对稳定且业务模式相对成熟，下游客户类型也相对稳定。

3. 市场规模测算

智通财经 APP 获悉，11 月 28 日，《IDCMarketForecast：中国网络安全市场预测，2022-2026》报告显示，2021 年，中国网络安全市场总投资规模为 122 亿美元，其中安全硬件产品投入达到 47 亿美元，占总体投入的 39%；安全软件产品投入达到 42.4 亿美元，占总体投入的 35%；安全服务产品投入达到 32.4 亿美元，占总体收入的 26%。IDC 预测，到 2026 年，中国 IT 安全市场投资规模将达到 319 亿美元，其中安全软件的市场占比将超过安全硬件，软件占比在 2026 年将达到 41%，即约 131 亿美元。

（五）云和数据产品介绍

1. 政府

（1）政务数据概述

一是依托数据采集打造政务数据资源库。当前，覆盖国家、省、市、县等层级的政务数据目录体系初步形成，数据生产部门按照政务数据目录和相关标准规范，通过人工或系统方式采集基础数据、主题数据、部门数据，各地区依托全国一体化政务服务平台向上级数据平台或数据中心归集，由数据主管部门按数据属性建立数据资源基础库和主题库。

二是政务数据融合分析以服务场景为牵引不断深入。各地区各部门依托政务大数据平台建立的政务数据仓库，围绕城市治理、环境保护、生态建设、交通运输、食品安全、金融服务、经济运行等应用场景开展数据分析应用，为多行业和多跨场景应用提供多样化共享服务。同时，围绕产业发展、市场监管、社会救助、公共卫生、应急处突等领域，推动开展政务大数据综合分析应用，为政府精准施策和科学指挥提供了重要支撑。

拓尔思的政务数据产品是面向政务数据融合分析以服务场景为牵引不断深入而服务的。拓尔思政务数据供给场景主要聚焦在 4 个方面：



图 30：拓尔思政务数据供给场景

① 产业大脑

国家“十四五”规划对推动产业数字化，提升产业链供应链现代化水平做出明确部署。产业大脑作为各级政府、园区和产业链链长提供产业选择的支撑、产业治理的工具和产业服务的平台，正有效帮助提升产业治理水平和协同能效。近年来国家密集出台相关政策文件，产业数字化迎来全新的发展局面。

当前，各政府部门间数据融合不足，产业的数字化程度不深，从“政务”到“产业”的数据平台和数据通道尚未形成，无法在常态数据监测、产业拐点研判、产业风险防范、产业阶段评估方面做到精准的产业管理，无法形成协同一致的“指挥棒”，极大地影响产业链上下游供需的匹配效率。从今年突如其来的新冠肺炎疫情可以看出，生产企业的产能、库存等信息未实现数字化，导致物资供需信息不对称，影响大规模医疗物资生产和调配的效率，无法满足国家从急到缓、从重到轻的精准化物资统筹管理需求。由此可见，和城市一样，产业也需要一颗“大脑”以加速产业的数字化进程。

全国范围内越来越多的城市，包括北京、上海、浙江、山东、广州、杭州等多地都走在新型产业治理的探索之路上。如浙江数字化改革方案明确提出以“产业大脑+未来工厂”“城市大脑+未来社区”为两大核心应用场景，第一次将产业大脑和城市大脑放到并重的位置。2022年6月，浙江省经济和信息化厅代拟起草《以“产业大脑+未来工厂”为引领加快推进制造业数字化转型行动方案（征求意见稿）》中明确提出，要提升产业大脑的应用赋能水平，有序推进企业大脑、行业大脑、产业大脑建设，形成互联互通、资源共享、高效协同的产业大脑体系。到2025年，建成行业产业大脑30个以上。山东省工信厅近日发布《关于深化改革创新促进数字经济高质量发展的若干措施》明确提出创建“产业大脑+晨星工厂”新模式，力争建成“产业大脑”50个以上。

② 政策大脑

当前国内外疫情防控与经济形势严峻复杂，稳增长、稳就业、稳物价面临新的挑战。各地政府纷纷提出以精准思维来高效统筹疫情防控和经济社会发展，紧盯发展需求，把准政策导向，积极对接国家政策体系，强化政策供给，全面提升支持政策的精准性和时效性。但在各地政府开展精准施策的工作中，存在一些普遍问题：

✧ 在改善优化营商环境时，各地政府都推出了大量的惠企政策。但是对企业而言，对政策缺乏了解，遇到事情不知道该找哪方面的政策文件；对政府部门的职能没有概念，找不准办理机关；办理状态不可知，过程性反馈少；缺乏政策申报指引，导致申请要素不齐全、信息描述不明确等。

✧ 政府职能部门在拟定政策时，缺乏跨层级、跨区域、跨部门的政策知识库支撑，如何避免高而浮和大而空的政策虚化使推进工作悬空？同时，还存在政策公开程序、答复内容不规范，政策内容合法性审查不严谨等现象。

✧ 政府管理部门在开展政策监管工作时，缺乏数字化监管工具，难以实现精准化、智能化的政务公开监管；缺乏面向社会公众的动态感知和历史数据的大数据分析，难以提供有效的决策参考等。

近年来各地出台政策的动作越来越频繁，涉及新任务、新产业的政策也越来越多，如惠企政策、科技创新政策、人才政策、复工复产政策、战略新兴产业政策等。针对以上精准施策难、如何消除推行政策盲点等问题，各地政府对政策大脑的需求呼之欲出。政策大脑不仅仅是一个大而全的政策文件库，主要需求还涉及多维的政策内容分析、

丰富的政策服务标签、全息的政策内容比对等，其核心是要构建政策算法化促进科学施策、精准施策和有效施策。从政策服务的流程环节来看，政策大脑涵盖政策制定、发布、执行、兑付、评价等全生命周期，涉及惠企政策服务、政策精准服务、政策综合服务、政策智能分析、政策兑付等相关应用场景。

③ 网络信息内容治理

国家“十四五规划”中明确了社会治理创新任务和 2035 年社会治理创新的远景目标。网络信息内容生态治理始终是管网治网的重要内容。近年来，在以习近平同志为核心的党中央高度重视和坚强领导下，中央网信办始终以提高人民群众在网络空间的获得感、幸福感、安全感为目标，扎实开展网络信息内容生态治理工作。从政策发布看，国家相关部门近年来发布了多个政策文件，如《关于促进平台经济规范健康发展的指导意见》、《网络信息内容生态治理规定》、《关于推进省级网信办网络安全应急指挥中心规范化建设的指导意见》等。从市场需求看，随着移动互联的日益发展，网络生态环境日趋复杂。“饭圈”乱象、网络水军、直播短视频、网络暴力、未成年人网络环境不规范等现象充斥着网络空间。政府急需科技赋能助力网络信息内容生态治理，加强日常监督，防范和治理各种网络生态问题，为广大网民营造风清气正、健康向上的网络空间。

网络信息内容生态治理已成为国家社会治理体系中必不可少的组成部分，也已成为网信部门的日常工作之一。未来网络空间的治理也将更加严格、全面，出于对监管手段的有效性和及时性的高需求，

必然会促进网络信息内容生态治理的市场规模扩大。由于该业务面向政府服务的特殊性，初期都以在电子政务内网建设软件平台与业务应用的模式为主。同时该类业务监测内容涉及互联网数据的实时更新，因此数据服务成为这类业务数据的核心来源方式。随着数据量的不断增长，用户方业务监测预警的要求越来越高，满足不同业务场景的应用服务对实用性、精准性的要求也将越来越高。从用户角度，国家、省、市、区（县）四级政府都有网信办的机构设置，政府每年有固定的信息化预算。该业务版块的市场需求整体将呈稳步增长之势。

④ 金融监管

金融科技的蓬勃发展在为金融业注入创新活力的同时，也使得金融风险构成的交叉性和复杂程度更加突出，从而给金融监管带来全新挑战。习近平总书记在 2021 年 8 月 17 日中央财经委员会第十次会议上明确要求金融领域“提高监管数字化智能化水平”。我国“十四五规划”第 21 章也提出“强化监管科技运用”。2021 年 12 月，央行出台《金融科技发展规划（2022-2025）》强调“加快监管科技的全方位应用，强化数字化监管能力建设”，监管科技的战略地位正逐步提升。加强金融监管和金融合规是维护社会稳定、金融稳定的基石，更是支持金融机构健康发展的加速器。

目前，非法集资、网络传销、非法放贷等各类非法金融活动涉及领域及企业数量众多，犯罪团伙通过专业组织分工，不断“创新”业务模式和手段，打着虚拟币、NFT 数字藏品、共享经济、国家扶持项目投资、以房养老等各种“合法”的旗号从事违法活动。相关监管部

门缺编制、缺数据、缺工具，导致金融风险防控工作面临发现难、预警难、研判难、认定难、处置难五大难题。因此需要借助监管科技手段，运用大数据思维方法，借助数字化平台工具，打破信息孤岛、打通业务协作流程，有效提升监管部门金融风险防控能力，守住风险底线，维护金融稳定。

（2）产品服务模式

① 数星产业大脑

数星产业大脑以产业数据中心为支撑，以智能数据标签引擎和全产业链知识引擎为核心底座，运用新一代信息技术，推动产业数字化、智能化升级，实现政策、空间、供应链、金融、科技、销售等产业要素与产业主体之间的高效协同，为企业生产经营提供数字化赋能，为产业生态建设提供数字化服务，为经济治理提供数字化手段。

数星产业大脑提供以下在线云服务：

✧ 宏观洞察：提供宏观经济运行指标的分析，包括但不限于 GDP，PMI，工业企业运行情况，通货膨胀，货币与信贷，以及监管政策等各类数据。利用 NLP 技术完成高精度数据指标抽取，通过对各类宏观指标的分析拆解，给出长期趋势和近期趋势的结果，帮助用户快速了解当前市场经济环境。

✧ 产业链分析：累计 5000+数星标准产业链、400+国家战略新兴产业链、2000+特色产业链，绘制 10 万+个上下游产业供需链图谱，明确供需关系及强度，令产业间关系得以量化。构建了全市场覆盖且

产业间可比的产业景气度追踪系统、集中度分析、产业风险传导、产业全景分析等模型。

✧ 产业地图：全方位、多维度、实时动态的产业数据展示和分析，呈现区域规划差异和发展成果，用更快捷的方式建立地区产业画像。

✧ 产业拓客：通过对企业和产业的评价，基于产业链上下游的建链、强链、补链的拓客引擎，精准获取高质量企业营销线索，实现优质服务支持、低效企业帮扶。

✧ 企业洞察：产业大脑汇聚多源数据，通过整合加工、标签化处理，构建“一企一档”，形成可检索的企业全生命周期数据资产，涵盖企业工商、经营、司法、知识产权、风险信息、所处产业链及具体环节及行业排名，助力地方政府精细化管理、培育企业。

✧ 事件中心：基于拓尔思特有标签引擎，超过上万个主题标签事件的智能搜索引擎（资讯、舆情、公告、政策、专利、投融资、境外等另类数据）

✧ 研究中心：汇集行业研报、个股研报、宏观研报等 40 多路研究报告，基于产业概述、市场规模等多类标签的提供研报图表搜索与定位、目录搜索与定位、内容搜索等核心功能。



图 31：数星产业大脑数据视图

在政府侧，产业大脑面向区域产业分析、区域精准招商、政府产业规划、强链补链研判等场景，为政府经济主管部门、产业园区管委会、研究机构提供区域产业发展深度分析、产业景气度集中度分析、宏观经济指标分析、区域政策比对等多种能力，服务区域产业转型升级，为区域经济的可持续发展提供分析决策支撑和全方位管理服务。

目前，数星产业大脑已成功服务了中关村、北京亦庄、辽宁营口、内蒙工商联、信用南京等客户，主要面向产业分析、政府招商场景，主要基于海量企业数据及多维度分析模型进行产业大脑构建，提供平台定制、系统开发、API 接口、数据运营、SaaS 应用等服务模式。上述项目中涉及数据服务采购规模平均在 100 万左右。

② 海策政策大脑

海策政策大脑围绕政策制定、发布、执行、兑付、评价等全生命周期流程，坚持问题导向、效果导向，通过运用大数据、知识图谱、语义智能、虚拟人等技术，构建了全国政策文件库，并提供一体化管理与智能分析的在线服务。



图 32：海策政策大脑政策分析示例

该产品帮助用户构建全国分类分级、集中统一、共享共用、动态更新的政策文件库，提供政策文件总览，可按层级、部门、区域、产业、主题等多维度自动打标分类，提供跨区域、跨产业、跨主题的政策比对分析，政策图谱分析，为政策制定提供高价值参考。基于细颗粒度的标签化服务，自动形成政策画像，对符合政策条件要求的企业进行精准的匹配推荐与政策推送，变“人找政策”为“政策找人”。可以按主题，以时间轴方式提供纲领性文件、指导性文件等政策执行脉络和统计分析。基于内容的自然语言处理技术，按主题、层级、部门等多维度提供政策执行工具分析。提供政策执行情况报告发布，帮

助政府部门一目了然地了解重大政策的执行情况。基于全网数据的传播，提供每条政策的社会评价分析，包括影响力指数、转载分析、互动分析等。为制定政策的政府部门更加科学全面的评价政策效果提供大数据支撑。

海策支持平台定制、系统开发、API 接口、数据运营、SaaS 应用等服务模式。目前已在中国政府门户、陕西省政府得到成功应用，每个项目的采购金额将近 300 万左右，其中每年的数据服务采购金额为 20 万左右。

③ 网察大数据云服务

网察大数据云服务秉承以用户需求为中心的设计理念，结合互联网舆情监测分析的实际业务流程，在功能设计层围绕“智汇、智库、智策”三大理念，依托人工智能、大数据、云计算等技术，实现了全舆论场的网络舆情洞察与研判。同时，顺应互联网视频内容快速发展等趋势，网察融合了视音频分析等语义智能技术最新成果，全面提升网络舆情分析的智能化、自动化水平，解决了网络舆情监测中短视频内容分析等新的痛点难点问题，实现了全舆论场、全媒体形式的网络舆情监测、分析、决策一站式服务。

在网络信息内容治理的应用场景下，网察聚焦网络低俗色情、饭圈乱象、网络暴力等网络生态问题持续整治，主要服务用户为全国各级网信办和网宣部。拓尔思是该领域中唯一一家具有平台+垂直场景应用+DaaS 服务+全国服务网络的综合服务厂商。

网络信息内容治理业务具有中央、省、市、区县多级联动的工作特点，垂直场景应用包括舆情/态势感知、统一指挥联动、数据中台、网宣等，能够满足网信用户实现网络信息内容治理的全栈业务需求。虽然这些应用都存在共性，但不同层级的政府部门在使用这些功能时仍存在很多联动机制、管理机制的差异化。拓尔思的用户分布在不同层级，垂直场景应用已孵化出省级、市级版本，成熟度高，普适性强。

拓尔思在网络信息内容治理领域成功服务了中央网信办，树立了行业标杆。通过综合实力与良好口碑在近两年内迅速覆盖了全国近50%的省级单位。几乎每个项目每年都有单独的数据服务采购需求。拓尔思在该领域竞争优势明显，是当之无愧的头部厂商。

④ 冒烟指数金融风控云服务

冒烟指数金融风控云服务聚焦地方金融风险防控，基于对企业大数据的采集分析、金融风险特征提取标注、量化预警模型搭建，及对“主动发现-量化监测-分级预警-核查化解-分类处置”全业务流程的梳理研究，形成金融风险防控全链条产品体系，致力于运用监管科技为地方金融监管局、公安经侦、行业监管等部门赋能，创新监管思维方法、提升数字化监管能力。



图 33：“冒烟指数”金融风控数据视图

针对辖区内财富管理、投资理财、私募、养老、房地产等重点领域企业，利用冒烟指数通过大数据的采集、融合、标注、分析，实现对非法集资、网络传销、非法放贷等金融风险的量化评估与分级预警，可视化呈现企业全息风险画像及地区整体风险态势，实现监管部门对金融风险的早发现、早预警、早处置。

在非法集资监测业务方面，“冒烟指数”成功服务了 400 多家政府监管单位，且续签率高，续签占比超 90%。公司深扎地方金融监管细分领域，在业内口碑及影响力位居前列。产品及服务嵌入地方金融监管的日常工作，成为防控金融风险的必备工具。另外多地用户在冒烟指数基础上结合本地特色特点进行定制化调整，形成“新冒烟指数”，包括“深圳海豚指数”、“重庆悟空指数”、“江西赣金鹰眼”、“温州追风捕影”等成功案例。其中，“重庆悟空指数”在全国率先推广“指尖打非”全流程 App, 实现了打非工作“移动化”落地，获得国家处非联办点名表扬的全新模式。“温州追风捕影”涉稳金融风险防范与处置应用被评为浙江省数字法治改革的经典案例，该应用已列入省拓尔思信息技术股份有限公司

“一本账 S1”，荣获浙江数字化改革第二批“最佳应用”、2021 年度浙江省改革突破奖。

在非法金融活动监测业务方面，“冒烟指数”成功服务了公安部，天津、南昌等几百个公安经侦部门，银保监、证监等金融监管部门。主要应用场景是针对相关行业领域建立专项预警模型，包括私募非集、养老诈骗、非法股票配资、非法放贷、网络传销、虚拟币传销等近十种重点非法金融活动，沉淀了数千项风险指标。

（3）市场规模测算

① 产业大脑

据中制智库统计，全国 31 个省市区共有 783 个产业集群。按每个产业集群构建一个产业大脑计算，全国共有 783 个产业大脑的空间。根据千里马招标平台公布的招标项目标的，一个市级产业大脑的招标金额在 300-600 万之间。取平均值测算，到 2025 年，全国产业大脑市场投资规模将在 30 亿元左右。

② 政策大脑

政策大脑的用户主要集中在省、市级政府，政策精准服务应用的用户主要集中在市、区级政府网站。基于千里马招标网近 3 年的公开招标信息统计分析，取平均值测算，到 2025 年，全国政策大脑市场投资规模将在 17 亿元左右。

业务类型	机构/网站数量 (个)	平均投入(万元)	市场规模(万元)
政策大脑	31(省)	300-600	9,300-18,300
	333(市级)	200-300	66,600-99,900

政策精准服务	948（市、区级政府网站）	50-100	47,400-94,800
合计			123,300-213,000

表 4：政策大脑市场规模测算

（行政区划数据来源于民政部网站公布数据，截至 2021 年 12 月 31 日。网站总数量的数据来源于观研报告网《中国数字政府市场行业现状深度研究与发展趋势分析报告》（2022-2029 年））

③ 网络信息内容治理

网络信息内容生态治理平台建设是根据国家每次十五规划的工作要求与部署，每五年做一次平台升级与改造。此类项目相对敏感，公开招标信息有限。基于拓尔思参与过投标项目的平均订单规模测算，网络信息内容治理业务 5 年内的市场总体规模将达 87 亿（ $558,100+63,140*5=873,800$ （万元））。

政府机构层级	机构数量（个）	平台建设投入（万元/5 年）	市场规模（万元/5 年）
中央	1	50,000	50,000
省级	31	5,000	155,000
市级	333	500	166,500
县级	1866	100	186,600
合计			558,100

表 5：网络信息内容生态治理业务平台建设市场规模（5 年）

平台运维保障和 DaaS 服务是每年购买服务，每年市场规模估算如下：

政府机构层级	机构数量（个）	每年运维和 DaaS 投入（万元/年）	市场规模（万元/年）
中央	1	3,000	3,000
省级	31	500	15,500

市级	333	50	16,650
县级	1866	15	27,990
合计			63,140

表 6：网络信息内容生态治理业务运维及 DaaS 服务每年市场规模

（行政区划数据来源于民政部网站公布数据，截至 2021 年 12 月 31 日）

④ 金融监管

拓尔思的金融监管业务主要在非法集资监测和非法金融活动监测两大版块。据测算，到 2025 年，金融监管的市场规模投入将达 16.5 亿。

非法集资监测的服务用户主要是各地的金融监督管理局。根据千里马招标平台公布的招标项目，取平均值测算，每年非法集资监测预警的数据服务市场规模在 3 亿元。

政府机构层级	机构数量 (个)	每年 DaaS 投入 (万元/ 年)	市场规模 (万元/年)
省级	31	50	1,550
市级	333	30	9,990
县级	1866	10	18,660
合计			30,200

表 7：非法集资数据服务市场规模测算

非法金融活动监测服务用户主要是公安部、各地公安经侦部门、银保监、证监等金融监管部门等，主要应用场景包括私募非集、养老诈骗、非法股票配资、非法放贷、网络传销、虚拟币传销等。根据千里马招标平台公布的招标项目，取平均值测算，每年非法金融活动监测的数据服务市场规模在 2.5 亿元左右。

政府机构层级	机构数量 (个)	每年 DaaS 投入 (万元/ 年)	市场规模 (万元/ 年)
省级	31	20	620
市级	333	20	6,660
县级	1866	10	18,660
合计			25,940

表 8：非法金融活动监测市场规模测算

综上所述，拓尔思的数据产品在政务数据要素市场的产业大脑、政策大脑、网络信息内容治理、金融监管 4 个应用场景的总体规模在 2025 年将达 96.5 亿。

2. 金融

(1) 金融数据概述

一是依托业务流程采集汇聚海量数据。传统的金融数据是指金融机构在其服务的全流程直接或间接从个人金融信息主体，以及企业客户、外部数据供应方等外部机构采集数据。采集方式可以分为传感器等边端设备采集、人工采集或系统采集和网络采集。

二商业智能、风险管理等继续呈现出旺盛的需求。从市场需求看，在 IT 架构转型、自主创新、金融科技技术升级、严监管以及中台等驱动因素的共同推动下，银行 IT 解决方案市场需求呈现出旺盛增长态势。其中，信贷管理、商业智能、风险管理等持续成为市场热点。银行业通过多种方式在信贷、理财、投行等多业务条线全流程采集海量企业金融数据、个人金融数据和外部数据，在中后台归集包括财务、审计等在内的银行核心数据。

（2）产品服务模式

数星智能风控云服务（以下简称“数星风控”）是一款面向风控、合规、投研、监管等核心金融场景的，围绕海量异构数据，利用人工智能、大数据、自然语言处理、知识图谱技术，全面监控海量金融实体多维风险，提升金融机构风控能力。

“数星风控”帮助金融机构提供以下在线服务：

① 企业风险画像：实现了全国 8000+万家企业全面覆盖，支持舆情、司法、经营多维度分析企业风险与价值，实现风险指数量化计算，全面掌握企业实时风险。

② 预警信号推送：支持实体信号归一化处理，重大事件分钟级响应，信号准确率超 98%，NLP 风险识别准确率超 85%，风险信号实时触达客户经理，实时提示重大风险。

③ 全网事件监控：覆盖境内外全渠道异构数据，支持企业、行业、区域等多实体监控，异动事件实时展现，涉事主体关联风险回溯，实时监测异动热点风险事件。

④ 风险异构图谱：多维度挖掘潜在风险传导与疑似关系：股东、投资、董监高；担保、被担保；客户、供应商；关联交易、潜在关系、政府关系；债权人、项目、融资。实现异构风险关联与传导。

⑤ 风险组合透视：实现区域+行业、区域+政策风险叠加透视，区域、行业风险监控及竞品对比，支持自定义组合风险透视。实现区域行业全企业风险穿透。

⑥ 集团系传导分析：基于股权关系、业务积累、舆情挖掘构建集团系智能图谱，分析成员层级关系、风险穿透、关系拓客等。

目前，“数星风控”已服务包括中国银行、农业银行、国家开发银行、进出口银行、平安银行、中国人保等高端客户，提供数据订阅（按年）+软件+本地化定制的产品服务组合。数星智能风控大数据平台获得 2020 年中关村大数据优秀软件，服务客户国家开发银行国际信贷大脑获得“人民银行 2020 科技发展进步二等奖”，平安银行智慧风控获得“Gartner2020 年金融创新奖亚太区冠军”，结合金融行业对于数据安全与数据合规的特点，未来混合型服务模式将是符合行业特点的服务模式。拓尔思以提供数据服务+NLP+本地化实施的高壁垒型服务体系，以及在头部客户的丰富经验，建立金融行业综合壁垒。

在商业智能与风险管理领域，拓尔思目前收入来源主要来自于高端客户相关案例，未来竞争策略主要基于现有头部银行的成功案例，逐步下沉形成标准化产品+定制化服务的组合，逐步往腰部以下商业银行进行复制。

（3）市场规模测算

根据工信部赛迪研究院《2021 年度中国银行业 IT 解决方案市场份额分析报告》的统计，2021 年中国银行业管理与监管类解决方案的市场规模达到 160.2 亿元，同比增长了 20.6%，其中，风险管理市场继续保持稳定增长态势，继续位居管理类第二大子市场，2021 年市场规模达到 32.05 亿元，占管理与监管类解决方案市场的 20.0%。金融风险数据服务是风险管理市场中重要的一个子市场。

根据银保监会网站公布的最新《银行业金融机构法人名单》显示，截至 2021 年末，国内共有 4602 家银行业金融机构。据统计，在 4602 家金融机构中，具体包括：开发性金融机构 1 家；政策性银行 2 家；国有大型商业银行 5 家；国有控股大型商业银行 1 家；股份制商业银行 12 家；城市商业银行 128 家；民营银行 19 家；外资法人银行 41 家；住房储蓄银行 1 家；农村商业银行 1596 家；农村合作银行 23 家；农村信用社 577 家；村镇银行 1651 家；农村资金互助社 39 家；贷款公司 13 家；信托公司 68 家；金融资产管理公司 5 家；金融租赁公司 71 家；企业集团财务公司 255 家；汽车金融公司 25 家；消费金融公司 30 家；货币经纪公司 6 家；其他金融机构 33 家。

全国政策银行、商业银行、股份制商业银行、城商行、外资银行、民营银行共计 216 家。按每家银行每年在金融风险数据服务 200 万的投入计算，整体规模将达 4.3 亿元。

全国农村中小银行共计 3902 家。中小银行平均按一个客户监测 1 万家发放贷款或投资并购的交易对手或标的资产，收费标准为 100 元/家/年时。近 4000 家中小银行每年在金融风险数据服务的投入规模将达到 40 亿元。

全国其他信托、资产管理、信贷等金融机构共计 484 家。按每家机构每年在金融风险数据服务 100 万的投入计算，整体规模将达 4.8 亿元。

综合上述三类金融机构每年在金融风控数据服务方面的投入规模，拓尔思在该领域的总体市场规模大约为 49.1 亿元/年。

3. 媒体

(1) 媒体数据概述

一是依托媒资管理平台采集汇聚海量数据。对媒体机构内的稿件数据、产品数据、运营数据、行为数据，以及外部互联网资源等各类数据资源有效汇聚与管理，通过建设统一的数据标准与数据资源管理，实现统一的基础数据服务给业务方。

二是以服务场景牵引生产的智能服务数据。基于机器学习、深度学习及迁移学习等人工智能技术，提供自然语言处理、图片识别、OCR识别和视频分析等基础能力。基于上述基础能力进行数据的深层次加工和业务组织封装，实现业务智能化，包括智能推荐、用户画像、内容标引、专题分析、内容审校和智能专题等。

拓尔思媒体数据要素完全覆盖上述需求，包括融媒体生产平台全链条数据、互联网公开采集的数据，以及基于此类数据的标引和精加工数据分析及场景服务。

(2) 产品服务模式

数家资讯大数据云服务（以下简称“数家”）汇聚了全网新闻资讯类的内容数据，以“多维知识库+智能语义”的方式对内容进行结构化标引，融合多领域对资讯的应用需求，为用户提供高信源、精加工、全链路、富场景的资讯内容服务，助力传媒企业依托大数据能力实现媒体融合发展的深化和转型。

“数家”提供的数据服务包括：媒资管理、互联网素材供给、媒体大数据（线索发现、热点分析、专题跟踪、传播效果、榜单定制）、

OpenAPI 开放平台、用户资产数据（用户数据、行为数据、智能推荐）等，可全面满足传媒行业用户对报纸库、期刊库、网站库、APP 库、微信库、微博库、图片库、视频库、线索库、报料库、文献资料库、人物报道库、地域报道库、机构报道库等媒资全库数据服务的需求。



图 34：数家资讯大数据云服务业务视图

“数家”已实现了产品的全面 SaaS 化，支持数据服务和开放接口服务，提供 200 余种 API 接口，涵盖基础数据服务、增值场景服务、智能分析服务、传播分析服务、专题分析服务、定向采集服务，全方位赋能多元内容业务场景，为用户提供丰富的选择。目前“数家”的 OpenAPI 开放平台已申请在北京数据交易所挂牌。

未来“数家”将向智能化服务深化发展，加大用户数据、行为数据、智能推荐等数据治理与服务优化。敏捷开拓数据服务灵活、开放、便捷可定制的新模式，加大发展合作伙伴创造数据服务和数据交易的生态和市场。

拓尔思服务了数百家主流媒体，参与建设了众多示范性项目，用户涵盖 72% 中央媒体、61% 省级媒体、40% 行业媒体，具有良好坚实的

用户基础。其中新华社、教育报、浙报传媒等已成为“数家”连续签约的数据服务客户，单个用户每年的 SaaS 服务贡献达数百万元。

（3）市场规模测算

据中国记协统计，截至 2022 年 9 月，中央主要新闻单位 18 家，全国性行业类媒体 107 家，全国性新闻期刊 28 家。按每个省、市基本都有 1-2 家主流地方媒体计算，省级新闻媒体有 60+，地市级新闻媒体有 600+。2022 年 8 月，中宣部发布全国 2585 个县级融媒体中心已建成运行。全国所有主流新闻媒体(互联网新闻平台未计)共计 3398 家。全国性媒体（153 家）每年购买数据服务的投入按 100 万计算，省、市地方性媒体（660）每年购买数据服务的投入按 50 万计算，县级融媒体（2585）每年购买数据服务的投入按 10 万计算，“数家”在媒体数据服务市场每年的总体规模将达 7.4 亿。

4. 专利

（1）专利数据概述

自进入互联网时代以来，我国发明专利、注册商标、著作权登记等知识产权数量迅速增长，知识产权密集型行业快速发展，知识产权保护服务市场的需求不断增加，同时中国经济在世界经济中的重要地位吸引了越来越多的外资企业进入中国。市场主体和保护标的增长为知识产权保护服务行业创造了更多的市场需求和发展机会。中国国家知识产权局知识产权运用促进司司长雷筱云表示，目前，中国知识产权服务机构超过 7.3 万家，从业人员超过 86.5 万人。随着近年各国高度重视知识产权保护与驱动科技创新发展经济，中国企业对创新和

知识产权服务的需求正在迅速增长，专利服务产业正迎来新的黄金发展期。

专利是同时具备技术、法律和商业三方面信息的综合体，具有天然的情报属性。因此，以专利这一综合体为基础，对专利大数据的利用，已成为当前技术创新活动中的重要一环。科技型企业需要借助专利大数据为生产经营活动提供技术支撑和研发指引；投资机构需要借助专利大数据对拟投项目进行技术实力评估和技术价值判断；政府部门更需要依托专利大数据对产业规划、科研项目立项、创新资源投入等做出科学决策。

专利数据按照专利种类分为发明专利、外观设计和实用新型专利三类数据库。专利除了是一种知识成果的所有权凭证之外，它本身也是一个大数据系统，专利大数据经年积累、对外公开、不可篡改，还经过了政府知识产权部门的审核把关，因此具有相当的客观性、准确性。一般专利服务是通过专利代理机构办理的，可以接受委托人的委托，在委托权限范围内以委托人的名义办理专利申请或其他专利事务的服务机构信息服务，如专利检索、专利数据分析等。

（2）产品服务模式

战略新兴产业专利检索云服务是拓尔思信息技术股份有限公司与知识产权出版社联合开发，基于国家战略产业布局和新兴产业发展的需要，在强化搜索引擎技术与数据资源优势基础上，共同构建的具备自主可控核心能力的专业领域数据库，以在线检索方式提供云服务。该产品是以十个新兴产业专利检索矩阵组成，包括新一代信息技术、

生物技术、新能源、新材料、高端装备、新能源汽车、绿色环保、航空航天、海洋装备、数字创意。

目前一期全球基因专利检索云服务已上线，涵盖全球 40 多个国家和组织的生物序列相关专利数据及 32 个公共资源库，通过对区域创新趋势和区域头部专利申请人的分析，可以帮助用户识别潜在的创新者和未来可能合作的重要伙伴，为我国生物技术领域的发展和创新提供强有力的“智力支撑”。支持核苷酸序列检索核苷酸数据库、蛋白质序列检索蛋白质数据库、翻译核苷酸检索蛋白质数据库、蛋白质序列检索翻译核苷酸数据库，共五种检索策略。用户可以通过输入一条或多条裸序列、fasta 格式序列进行检索；也支持上传序列文件（支持.txt 和.FASTA 格式文件）进行检索，目前最多可支持 200 条序列同时进行检索。通过调整检索条件、序列条件、专利条件和算法条件，实现精准检索；多维度筛选序列检索结果，为不同研究目的获取目标序列。



图 35：拓尔思全球基因专利数据库

专利代理机构和专利申请单位可以利用上述专利数据库，进行专利服务全生命周期工作开展：

- ✧ 检索、筛选、处理和分析知识产权信息；
- ✧ 进行产业专利导航分析，规划创新发展方向、路径；
- ✧ 挖掘、分析知识产权信息披露的法律、技术、经济等价值；
- ✧ 绘制专利地图，分析技术领域发展和竞争态势，判断、预警专利风险；
- ✧ 分析、评价知识产权的使用价值和市场作用。

（3）市场规模测算

专利服务主要面向两大群体服务：一是专利代理机构，二是专利申请单位。

一般专利服务是通过专利代理机构办理的，代理机构也是经省专利管理局审核，国家知识产权局批准设立。国家知识产权局近日发布的《全国专利代理行业发展状况（2021 年）》中显示，截至 2021 年底，专利代理机构达到 3934 家（大陆）。根据市场情况，每家专利代理机构平均每年在单个行业专利库投入专利检索服务的规模为 10-50 万元计算，则单个行业专利库每年专利检索服务的市场规模将达 4~20 亿元。拓尔思计划推出十个战略新兴产业的专利检索服务，那每年专利服务的总体规模将达 40~200 亿元，取平均值测算为 120 亿元。

另外，根据 2022 年 1 月 12 日，国新办举行 2021 年知识产权相关工作统计数据发布会的数据显示，截至 2021 年底，我国国内拥有有效发明专利的企业达到 29.8 万家，较上年增长 5.2 万家。国内企

业拥有有效发明专利 190.8 万件，同比增长 22.6%。体现出我国市场主体创新创造能力不断增强。按每家大中型企业每年 10 万专利服务投入计算，我国大中型工业企业每年专利服务的市场规模将达 48 亿。

综上所述，专利代理机构和大中型工业企业每年在战略新兴行业专利服务投入的市场规模将达 168 亿元。

5. 舆情

（1）拓尔思舆情数据

一是舆情已形成四大细分市场的格局。舆情研究和应用的主要推动力来自于网络空间治理、风险和危机管理、开源情报挖掘三大方面的市场需求。目前舆情产业已经形成了政府、媒体、教育科研和商业等四大细分市场。政府开展舆情工作，既有网络信息内容治理的工作任务要求，又可以汲取社情民意，为决策提供参考，具有很强的政策性和现实意义。媒体开展舆情工作，一是有自身融媒体业务发展的需求，二是媒体对业界政务舆情研究的成果大量涌现，可转化为党政部门、企业和社会团体提供舆情报告的增值服务。教育科研开展舆情工作既有相关课题研究的需要，也可以实现科技开源情报服务的获取。企业开展舆情工作，主要可实现企业口碑监测、风险与危机管理和竞品动态监测等。

二是网络空间快速发展带来新的风险和挑战。首先，社交媒体、自媒体的迅速崛起和发展，呈现多元的网络内容生产，内容生产主体的泛化，消解了主流价值的传播与影响。其次，微信、微博、短视频等新的传播形式不断升级迭代，使网络传播形式呈现多模式，使舆情

场更加复杂。最后，网络平台化社效趋势，信息传播“蝴蝶效应”凸显。

三是全舆论场舆情态势感知关键技术日趋复杂。随着新闻推送服务、短视频、区块链等个性化、多媒体的发展，网络舆情监测更加强调对全舆论场的舆情态势感知。网络舆情态势感知需要实现三大能力，即对网络信息热度、敏感度和突发事件的“发现力”；对跨舆论场的多维度量化分析的“研判力”；以及对整体态势的综合“评估力”。要做到全天候全方位感知网络安全态势，就要对全网舆情的全量数据进行精准挖掘，对敏感信息进行分类，并精准定位热度信息；对多舆论场的内容进行智能分析，能够做到全面把握，并进行舆情分析研判；对整体舆情态势进行展示和预测，辅助相关机构制定和发布决策。

（2）产品服务模式

网察大数据云服务（以下简称“网察”）融合了视音频分析等语义智能技术最新成果，全面提升网络舆情分析的智能化、自动化水平，解决了网络舆情监测中短视频内容分析等新的痛点难点问题，实现了全舆论场、全媒体形式的网络舆情监测、分析、决策一站式服务。

“网察”聚合了网络新闻、微博、微信、论坛、博客、APP、电商、电子报、境外媒体等多渠道海量数据，以信息监测、态势感知、关系挖掘、事件追踪、传播分析、智库共享等为目标，依托基于深度学习的自然语言处理引擎，结合互联网舆情监测分析的实际业务流程，解决跨舆论场的网络信息洞察、分析、研判。从网络舆情线索发现、实

时预警、分析研判、综合报告等各个环节为用户提供全面、及时、准确的 SaaS 服务。

网察完成了新一代网络舆情分析 8 大关键技术的实现：

- ① 从定性分析到定量分析演变。通过指标体系量化业务知识，实现各类指标的自动计算。
- ② 从单舆论场分析到多舆论场融合分析。构建了一套综合量化指标体系，可全面评估网络舆情状态。
- ③ 从关键词分析到构建知识图谱演进，不断深入挖掘人物、事件之间的关系。
- ④ 从文本分析到跨媒体分析转变。
- ⑤ 从单语种分析到跨语种分析变化。
- ⑥ 从正负面情感分析到全面情感分析细化。
- ⑦ 从内容分析到“内容+群体人物+行为+时间+空间/地域”分析转变。
- ⑧ 从内容监测到事件分析变化。

多年来，“网察”服务了大量的政府、金融、媒体、企业用户。

政府：外交部、工信部、公安部、交通运输部、教育部、发改委、最高法、最高检、监察委、市场监督管理总局、海关总署等部委，以及数百个地方政府单位都是“网察”的用户。很多用户与拓尔思都是多年的合作，续签率非常高。如“网察”为国家发改委提供日常监测、智能预警和舆情汇编等服务，通过对国民经济和发展规划、国外经济形势及发展情况，宏观经济预测等实时监测和分析，帮助用户及时掌

握互联网舆情，第一时间获知相关负面舆情。根据网察态势评估和分析，对网络热点事件进行多维数据分析和舆情趋势研判，避免事件持续发酵恶化，为事件处置和方针决策提供数据依据。双方连续合作4年以上，获得用户高度认可，并主动把“网察”引荐给其他政府单位。

企业：中石油、中国移动、中国电信、中国联通、海尔、佳能、索尼、三星、浪潮、光大银行、上汽通用汽车、宝马等著名企业都是“网察”的用户。百胜中国是全球最大的餐饮集团，“网察”为其提供“定向监测”、“实时预警”到“研判建议”的一站式服务。百胜通过“网察”从全网采集海量信息，收集消费者对产品、服务、活动的满意度和反馈意见，以及行业政策、同行业企业动态信息等内容，并自动生成个性化报告辅助市场决策。同时，根据预警提示及时处理各类问题，满足客户不同时期的不同需求，提高产品质量与消费者满意度，构建与维护品牌与消费者之间的良好关系，持续提升品牌价值与影响力。至今“网察”为其服务6年，为用户预警拦截上万起潜在舆情风险，客户满意度高。

（3）市场规模测算

据共研行业调研所统计，近年来我国舆情大数据市场规模增速有不断加快的趋势，2021年中国舆情大数据市场规模达134.38亿元，同比增长13.26%，这也得益于近年来我国5G、云计算等相关技术的兴起和发展，在“十四五”等相关政策支持下，我国舆情大数据市场未来可期。2022年中国舆情大数据市场规模达134.38亿元，同比增长13.26%。

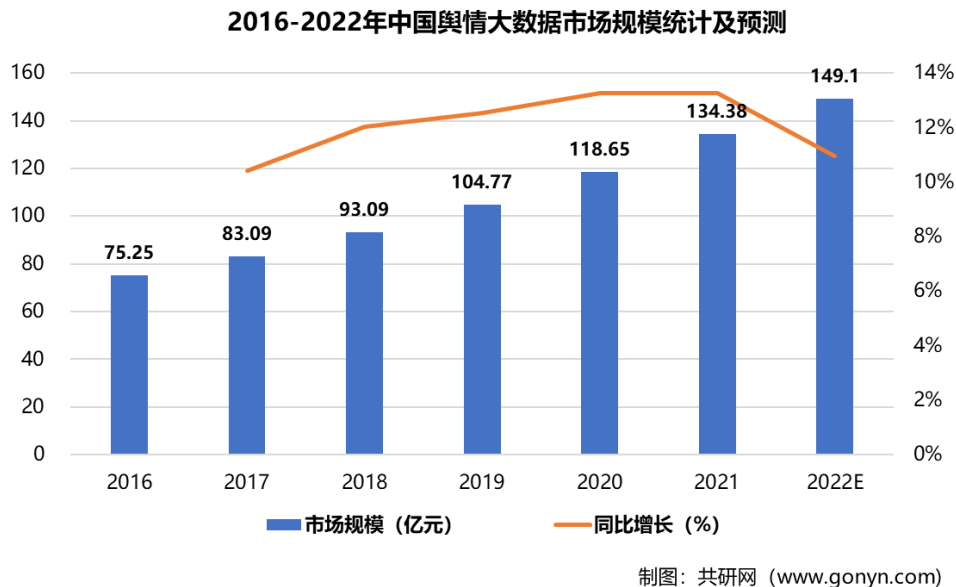


图 36：2016-2022 年中国舆情大数据市场规模测算

（六）云和数据服务收入规模

拓尔思于 2018 年提出大力发展 SaaS 服务战略，公司 2018~2021 年云和数据业务收入高速增长，2021 年相关收入达 3.78 亿元，占营业收入比例达到 36.84%。在 2022 年经营计划中，公司明确提出支持加快发展云和数据智能服务战略，预计未来将持续保持较快增长态势。

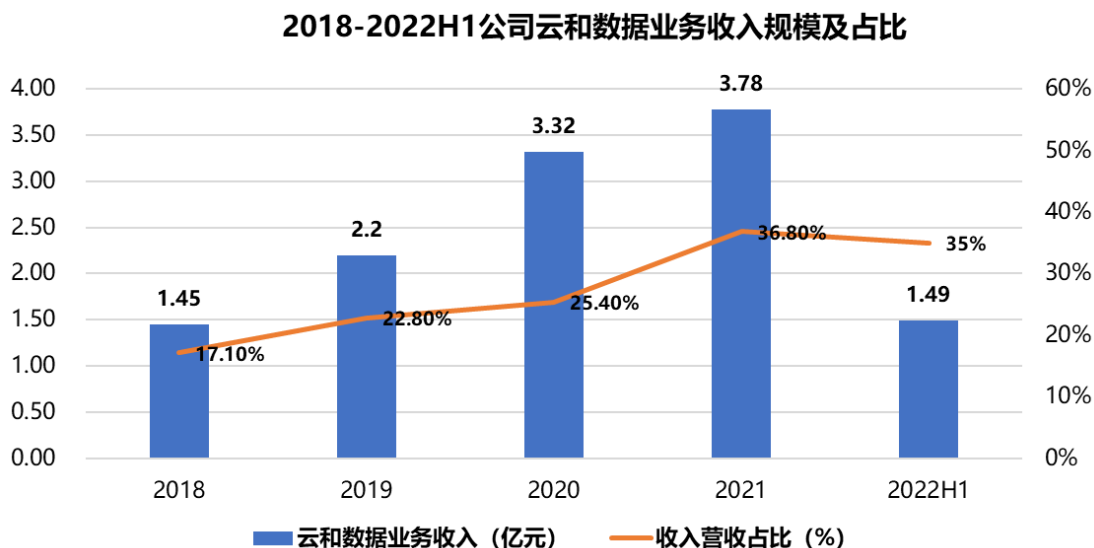


图 37：拓尔思云和数据业务收入规模及占比

SaaS 商业模式具备多重优势，公司加速转型有望进一步提升核心竞争力。公司加速业务云转型，整个商业模式正由销售解决方案、软

件产品全面转向 SaaS 服务模式。传统的数字政府、融媒体等行业解决方案的生产由重建设转向重运营、重增值服务，通过叠加专业领域的数据服务，进一步深化服务能力。未来公司在专利、虚拟人、机器人等新领域业务开拓中也将采用 SaaS 服务模式，提升公司 SaaS 营收占比的同时，进一步完善业务生态，提升行业竞争力。

（七）小结

1. 拓尔思数据要素业务发展空间大，未来可期

综上所述，至 2025 年，拓尔思的数据产品对应行业应用的市场规模总计为 1175 亿。具体市场规模测算分布如下：

政务数据要素市场的产业大脑、政策大脑、网络信息内容治理、金融监管 4 个应用场景的总体规模将达 96.5 亿。

金融风控数据服务市场的总体规模将达 150 亿（49.1 亿元/年*3 年）。

媒体数据服务市场的总体规模将达 22 亿（7.4 亿/年*3 年）。

专利检索服务市场的总体规模将达 504 亿元（168 亿/年*3 年）。

舆情大数据市场的总体规模将达 403 亿元（134.38 亿/年*3 年）。

拓尔思全资子公司天行网安的安全一体化业务，根据 IDC 预测，到 2026 年，中国 IT 安全市场投资规模将达到 319 亿美元。

在上述优势领域中，拓尔思均和各赛道的头部用户建立了长期良好的合作，用户黏性强，复购率高，因此能够保持现有业绩稳定。基于现有头部用户的成功案例与标杆示范，拓尔思将加大各行业/区域的生态网络建设，制定灵活的合作模式和数据服务模式，逐步往腰部

及以下的用户进行覆盖。未来三年，拓尔思的行业 SaaS 集群业务业绩将迎来大幅增长，未来可期。

2. 拓尔思有望成为国内另类数据服务龙头，加速数据要素价值释放

拓尔思采集的是网络公开数据，包括网页、新闻、公开数据、社交数据、百科等，已实现从数据采集、存储到数据的加工处理，是另类数据市场上的轻处理数据提供者。此外，另类数据本质上是在投资决策、发展战略决策、趋势预测、风险预警等方面，追求极致的量化理念。展望未来，另类数据有望成为地方政府、头部投资机构信息优势的重要来源，在应用范围、应用深度和应用频率上持续、显著提升。

在非法集资监管领域，拓尔思的“冒烟指数”已成为了一套事实标准。以前地方政府金融监督管理局和公安经济侦查部门所认定的企业风险高低是两套不同的语系。目前，拓尔思通过“冒烟指数”让相关部门有统一衡量标准，让监管部门和公安部门在同样的标准上看待一个“冒烟”企业。北京、重庆、深圳、温州等很多地方都在引用“冒烟指数”标准，这成为了该领域一套自下而上的事实标准，而不是国家定的标准。“冒烟指数”目前应用在 400 多家金融监管单位，助力行业规范发展。

“冒烟指数”是拓尔思另类数据在金融监管方面应用成功的典型。首先，公司围绕被监管对象收集各类大数据资源，采集了网络舆情数据、工商数据、法院数据、招聘数据、投诉数据、公司官网数据，还

有一些监管数据，形成了拓尔思金融风险大数据中心。其次，通过 NLP 技术对企业名称、经营范围、经营活动、相关描述进行自动识别、打标、抽取。再次，从风险分析和所掌握数据，构建“冒烟指数”模型。针对非法集资特征，形成了合规性、收益率、投诉率、传播力、特征性等一级指标，还包括二级指标、三级指标。此外，还针对网络借贷、私募股权、小贷公司单独设计了指标体系。最后，通过冒烟指数金融风控云服务将“冒烟指数”分析结果和基础数据可视化呈现出来。放眼全国，全国有 200 多万类金融机构，拓尔思监测的将近有 150 万家企业，对地方的金融监管确实起到了很重要的作用。预警过有冒烟信号的企业 6 万多家，成功预警的有近 2000 家。

拓尔思的另类数据通过与 NLP 技术、深度学习模型融合，专注于风险预警、智能投研、企业经营情况分析、行业景气度预测、智能招商评估等场景，在网络信息内容治理、非法集资监测、互联网金融风控、金融监管、产业大脑、媒体大数据等领域得到了非常成功的商业应用。并且拓尔思的多个指标已成为业内事实标准。[拓尔思从轻数据处理提供者成为了信号提供者。](#)

拓尔思在数据要素市场探索出来的“1+1=N”模式，在统一数据中心的基础上，发展了数款分别面向政府、金融、媒体、专利、舆情等 SaaS 产品，这类数据和云服务在近三年内的业绩得到了较快增长。这也体现了数据要素具有可复制性强、迭代速度快、复用价值高以及无限供给等特点，使另类数据的价值得到充分释放。“1+1=N”的商业

模式也得到了成功验证，拓尔思在全面加速数据+云服务转型的战略指导下，将迎来更多的数据产品创新，拓展更多的市场机会。

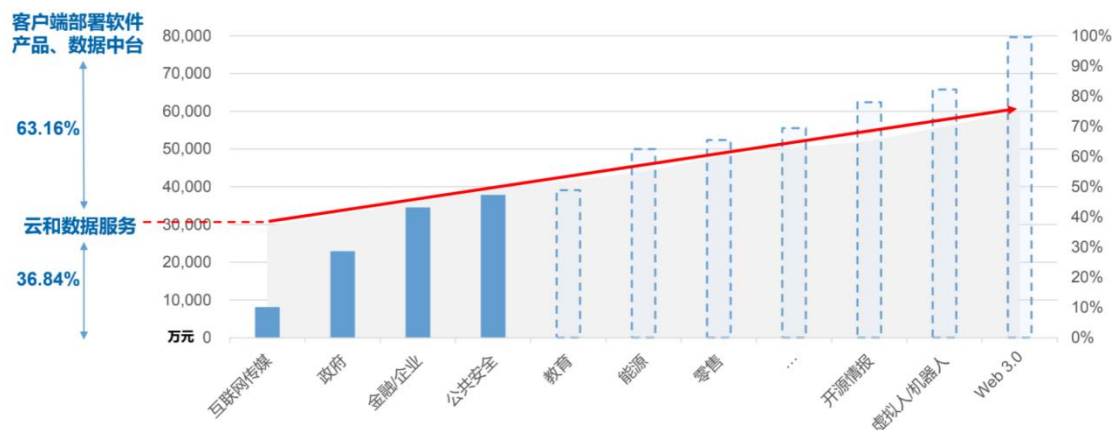


图 38：拓尔思云和数据服务业绩增长曲线

五、拓尔思数据要素发展方向

（一）强化数据要素底座基础

内容方面，围绕着数字经济的新产业、新业态、新模式扩大构建农业、工业、交通、教育、安防、城市管理、公共数据交易等方面的数据资源开发。

采集方面，进一步扩大网络公开数据采集范围，加大多模态数据采集力度，继续拓展生态数据汇聚融合。

技术方面，提升多模态内容分析能力，加大语义检索、稠密向量、多模态检索等方向的产品研发。

（二）加强数据治理能力建设

一是优化数据资产的分类分级。在现有基础上，将全资子公司的所有数据资产纳入统一管理，对拓尔思全量数据进行全面盘点、分类定级。优化数据管理机制，实现数据台账动态维护。通过台账数据治理，达到纲举目张，构建全公司数据资产目录体系。

二是加强数据资产的质量校核。明确数据质量度量规则 and 标准，健全天津数据服务中心与各行业 SaaS 集群产品部门之间的数据服务反馈机制，及时对问题数据进行核查，自动探查数据结构和数据内容、数据纠正。

三是加大领域业务专家的引入。根据数据产品的实际应用情况，为了进一步提升业务场景的预警、预测指标的精准度，加大业务专家的引入，一起设计针对性的预训练任务、算法、模型等，让拓尔思数据产品在各自垂直领域的应用效果得到显著提升。

（三）提升数据要素变现能力

一是专注优势产品，做精做细，做优做强。警惕新产品盲目扩张，要“小步快跑”，采取“滚雪球”的发展方式。

二是建立专注行业 SaaS 产品的营销团队，不断丰富生态布局，进一步市场扩张。

三是加快推进拓尔思数据产品在各地数据交易所挂牌，丰富数据要素的流通与变现渠道。

六、拓尔思数据要素业务保障

（一）组织保障

拓尔思作为数据要素市场主体开展数据服务活动，将积极落实数据管理主体责任，实施“管理层重视、一把手负责、全员参与”的数据要素管理模式。建立健全公司自身的数据管理组织架构、管理制度和自我评估机制，加强对公司现有数据资产的分类分级管理及合规性建设，加强数据质量管理，确保数据的真实性、准确性、完整性、时

效性，加强数据安全保障，提升边界安全传输、数据安全治理等数据安全技术能力。

同时，建立公司内部的数据要素统计核算制度，明确公司数据资源的统计范围、统计指标和统计方法，清晰摸底、准确掌握公司数据资产的价值。

（二）资金保障

统筹运用公司财政资金预算，加大对数据要素领域关键核心技术研发、创新产品及应用孵化等方面的资金支持力度，积极引入风险投资，鼓励公司内部开展数据要素领域的产品和服务创新。开展对外投资，加强与垂直领域数据服务商技术与产品等方面的合作与共享。

保障公司开展数据收集、归集、治理、共享、开放及其质量和安全管理等工作涉及的经费，保障必要数据采购经费。加强数据治理人才引进和储备。

（三）人才保障

建立公司数据要素虚拟工作专班，建立首席数据官、首席安全官方等主体责任制，开展公司数据要素领域的产品融合与业务规划统筹工作，提出数据要素领域的新技术研究及新产品研发的建议与指导，建立完善的数据安全技术体系、建设数据安全生态，确保公司数据产品安全等。

加强公司业务骨干培训，分层次、分类别组织开展首席数据官和数据要素市场化配置等专题培训，打造具有良好数据素养的人才队伍。创新公司数据人才评价与激励机制，健全数据人才服务和保障机制。

