百度智能云

“云智一体”技术与应用解析系列白皮书

智能大数据篇

目录 CONTENS

一、云智一体，促进数字化转型与智能化升级一步到位 …… 02

1.1 产业智能化，数字经济发展新阶段 …… 03
1.2 百度智能云战略升级 …… 03
1.3 云智一体的百度智能云架构2.0，加速产业智能化 …… 04

二、云智一体的智能大数据产品架构全景图 …… 05

2.1 数据要素成为数字经济时代的新战略资源 …… 06
2.2 大数据驱动千行百业数字化转型和业务创新 …… 07
2.3 大数据的未来发展趋势 …… 08
2.4 企业构建大数据平台的原则 …… 08
2.4.1 成效为先 …… 09
2.4.2 治理为要 …… 09
2.4.3 安全为本 …… 09
2.4.4 持续运营 …… 09
2.5 百度智能云大数据产品架构全景图 …… 10

三、云原生湖仓架构，构建企业大数据基础设施 …… 11

3.1 云原生 …… 12
3.1.1 按需创建 …… 12
3.1.2 存算分离 …… 12
3.1.3 弹性扩缩容 …… 13
3.1.4 运维托管 …… 14
3.2 数据湖架构 …… 14
3.2.1 BOS数据湖存储 …… 15
3.2.2 丰富的数据湖分析引擎 …… 16
3.3 一体化平台 …… 19
3.3.1 元数据统一管理 …… 20
3.3.2 自动数据集成 …… 20
3.3.3 全流程可视化数据治理开发 …… 21
3.3.4 数据湖分析 …… 21
3.4 成功实践 …… 22

四、智能数据挖掘，驱动数据资产价值最大化 …… 24

4.1 数据资产管理与运营平台DAMP …… 26
4.1.1 数据资产目录 …… 26
4.1.2 数据在线应用工具 …… 26
4.1.3 数据服务超市 …… 26
4.1.4 数据资产看板 …… 27
4.2 数据可视化Sugar BI …… 27

4.2.1 直连多数据源 27
4.2.2 丰富的图表和拖拽式编辑 28
4.2.3 炫酷大屏与智能交互 28
4.3 全功能AI开发平台BML 29
4.3.1 BML总体架构 29
4.3.2 BML为数据科学提供的核心功能 30
4.3.3 面向行业的智能数据挖掘解决方案 31
4.4 智能数据挖掘成功实践 32
4.4.1 邮储大脑机器学习平台 32
4.4.2 某集团研发中心设备健康管理系统 33
五、隐私计算铸就安全防护新体系 36
5.1 数据安全的落实规则 38
5.1.1 覆盖数据全生命周期的安全体系 38
5.1.2 隐私数据的安全保护闭环 39
5.1.3 安全合规的数据流通形式 39
5.1.4 主流的隐私计算核心引擎 40
5.2 百度全生命周期数据安全防护体系 40
5.2.1 资产安全 41
5.2.2 隐私保护 41
5.2.3 隐私计算 41
5.3 百度点石成功实践 43
六、企业数据资产构建与应用实践 45
6.1 数据治理 46
6.1.1 搭班子 46
6.1.2 做盘点 47
6.1.3 定标准 48
6.1.4 稳实施 48
6.2 资产管理与运营 49
6.2.1 资产管理 49
6.2.2 资产运营 49
6.3 业务应用 50
6.4 企业数据资产应用实践 51
6.4.1 徐工综合经营管理决策驾驶舱 51
6.4.2 陕西广电媒体融合数据资产建设 52
七、更多行业案例参考 56
八、关于本报告 57

云智一体，促进数字化转型与智能化升级一步到位 >>

1.1 产业智能化，数字经济发展新阶段

当前，我们正处在科技创新和产业发展最好的时代，以人工智能、大数据、云计算、5G等为代表的新一代信息技术蓬勃发展，驱动新一轮科技革命和产业变革。

在产业应用新兴技术和数据资源转型升级的过程中，数字化是基础；而随着人工智能等新一代信息技术的发展，以及数据的爆发式增长，产业加快应用智能技术，从海量数据中发现规律、训练模型、提炼知识，促进产出增加和效率提升，实现企业生产经营的智能化，进而实现整个产业的智能化升级。

因此，数字经济既包括数字化转型，也包括智能化升级，产业智能化是数字经济发展的新阶段。

1.2 百度智能云战略升级

百度是拥有强大互联网基础的领先AI公司。从成立的第一天起，百度搜索引擎就与人工智能、云计算等技术息息相关。经过多年的技术积累和产业实践，百度已形成了全面布局，从基础的算力和数据技术、深度学习算法及框架，到语音、视觉、自然语言处理等感知、认知技术，以及飞桨深度学习开源开放平台等，具备云智一体的独特优势。

基于百度的技术优势和对产业发展的洞察，经过跟合作伙伴的共同成长与产业实践，百度智能云的战略升级为：以“云计算为基础”支撑企业数字化转型，以“人工智能为引擎”加速产业智能化升级，云智一体“赋能千行百业”，促进经济高质量发展。

云智一体的“云”为数字化转型提供安全、稳定、灵活的数字化底座，“智能化引擎”为智能化升级提供领先的创新技术和平台。我们推荐客户数字化转型与智能化升级一起做，一步到位，这样效率更高，效果更好。

同时，如果客户只做数字化转型，我们有丰富的云计算产品可以支持；或者客户已经有了数字化基础，我们的智能化引擎也可以助力智能化应用。

百度智能云的“云”，不仅为数字化转型提供安全稳定、弹性灵活的云计算服务，同时也是适合跑AI的云，我们称之为“AI原生云”，继承云原生的优势，面向AI场景，提供极致弹性的高性能异构算力，打造简洁、高效的AI应用开发架构。

“云智一体”的智，是百度将深耕十余年的AI技术，以及AI赋能产业的实践经验，沉淀为助力产业智能化升级的技术、平台及方法论，是以加速产业智能化升级为己任、懂场景的AI。

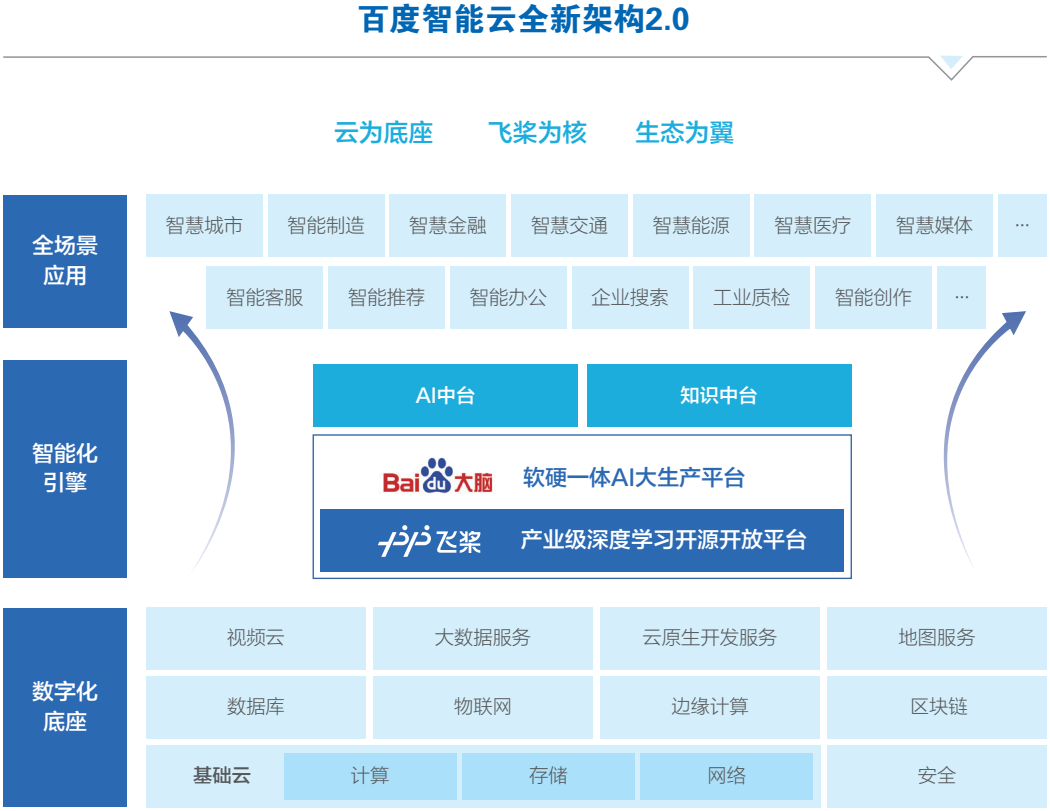
1.3 云智一体的百度智能云架构2.0，加速产业智能化

基于百度智能云的战略升级，云智一体的百度智能云架构2.0同时发布，包含数字化底座、智能化引擎和全场景应用（如图1.3－1所示）。

数字化底座，包括基础云、数据库、物联网、边缘计算、区块链等基础平台，以及视频云、大数据、云原生开发和地图服务等；同时还有安全模块，为技术创新和产业转型升级保驾护航。

在数字化底座之上是智能化引擎，百度自主研发的飞桨深度学习平台为核心，软硬一体AI大生产平台—百度大脑为支撑，包括AI中台和知识中台等。智能化引擎与行业深度融合，帮助企业建设自己的智能化中台，助力智能化升级。

基于数字化底座和智能化引擎，百度智能云深入行业场景，打造智能应用，同时与生态伙伴一起，在制造、能源、城市、金融、医疗、媒体等领域，助力数字化转型和智能化升级一步到位。



（图1.3-1）

我们希望以百度智能云为纽带，联合技术和产业生态伙伴，通过开源开放平台降低AI开发的门槛，加快人工智能融入传统产业，帮助越来越多的行业大步快跑，加入产业智能化的大潮。

云智一体的智能大数据产品架构全景图>>

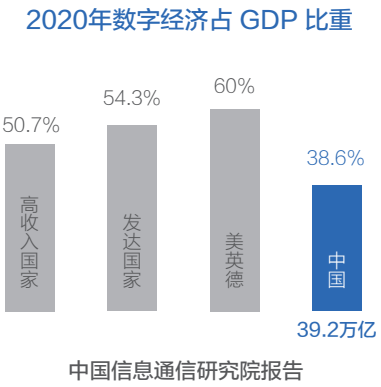
2.1 数据要素成为数字经济时代的新战略资源

随着新一轮科技和产业革命的浪潮席卷而来，特别是大数据、人工智能、移动互联网、云计算、5G等新一代信息技术的应用，人类进入数字经济时代。数字经济是以数字化的知识和信息作为关键生产要素，以数字技术为核心驱动力量，以现代信息网络为重要载体，通过数字技术与实体经济深度融合，不断提高经济社会的数字化、网络化、智能化水平，加速重构经济发展与治理模式的新型经济形态。

2019年，高收入国家的数字经济占GDP比重达到47.9%，其中发达国家这一比例高达51.3%，美国、德国等国家则超过60%。中国信通院发布的《中国数字经济发展白皮书(2021)》显示，2020年我国数字经济规模达到39.2万亿元，占GDP比重为38.6%，保持9.7%的高位增长速度，成为稳定经济增长的关键动力。

“十四五”规划纲要中将“数字经济核心产业增加值占GDP比重”作为创新驱动的关键指标之一，并预期“数字经济核心产业增加值占GDP比重”将从2020年的7.8%上升到2025年的10%，大数据将成为驱动未来数字经济发展的强大推动力。

数字经济成为“十四五”期间高质量发展新引擎



- 《上海加快发展数字经济推动实体经济高质量发展的实施意见》
- 《打造上海市数字化转型示范区全面推进杨浦区城市数字化转型行动方案（2021-2023年）》
- 《北京市关于加快建设全球数字经济标杆城市的实施方案》

……

（图2.1-1）

在数字经济中，数据要素是数字经济时代的新战略资源。2020年，国务院发布了关于要素市场化配置的指导意见，明确把“数据”与土地、劳动力、资本、技术等并列为五大生产要素。

和其他生产要素相比，数据生产要素有两个作用，即创造和放大。创造创新方面，数据作为一种新的生产能力，直接驱动了很多新模式、新经济形态、新产业。特别近两年疫情不断反复期间，很多产业实现了彻底的数字化和线上化，而今年大火的“元宇宙”概念，也是数字化产业发展的新方向。放大方面，数据要和资本、土地、劳动力、技术等其他生产要素协同，发挥乘数作用，解决供需优化、创新价值链流转方式等问题，放大其他要素价值。

数据要素成为数字经济时代战略资源

（图2.1-2）

2.2 大数据驱动千行百业数字化转型和业务创新

回顾近几年可以发现，大数据已经驱动零售、政务、城市治理、金融、创造业、教育、医疗等很多行业，进行了数字化转型或者业务创新的尝试与实践。

在零售行业，开店选址对最终业务成功非常关键，结合大数据技术和目标客群定位，实现精准智能选址，从源头降低经营的不确定性和风险；在制造业，以前设备交付给用户后，还面临着维护售后等运营问题，结合大数据和物联网技术，可以及时了解设备状态，实现预测性维护，大大提升了制造业的服务效率和用户体验；在视频行业，长、短视频等信息载体，已经实现了个性化推荐，这背后也是基于大数据技术，提升用户体验，创造新的增长空间。

通过这些例子可以看到，大数据在各行各业的数字化转型和业务创新里面有非常多的实践和案例，驱动着千行百业的数字化转型和业务创新。

大数据驱动千行百业数字化转型和业务创新

（图2.2 -1）

2.3 大数据的未来发展趋势

首先，4V特性更凸显（规模性- 多样性- 实时性- 价值型）。随着5G、IoT等技术的发展，以及企业更加重视数字化发展，数据总量成指数级增长，数据价值密度越来越低。同时，除传统结构化数据外，视频、语音、图像等海量非结构化数据不断积累，且随着企业数字化程度的提高，数据处理的实时性也越来越高。

其次，大数据和AI深度融合。AI也是一种数据处理技术，它与大数据的深度融合，让数据处理更高效，帮助企业从海量业务数据中挖掘更多价值。

第三，数据安全是生命线。关于数据安全，近几年国际和国内范围内都密集出台了多个数据安全类法规。从数据安全趋势来看，社会各方在数据流通与运营过程中，数据安全合规是基础生命线，必须满足。

综上，新型的数据基础设施、深度挖掘数据价值，以及保障数据安全，是构建大数据平台的关键。

大数据领域发展趋势

（图2.3 -1）

2.4 企业构建大数据平台的原则

2.4.1 成效为先

随着人工智能等新一代信息技术的发展，数据以指数级速度成倍增长，海量数据的存储和计算带来了巨大的IT成本和复杂性。云计算的发展有效实现了企业IT基础设施的弹性灵活性和成本低廉性；而当企业面对更大规模、类型更多样、更实时、价值密度低等数据挑战时，构建云原生湖仓数据基础设施就成为了一种新的有效应对方式。

该类新型数据基础设施优势：一是以云为基础，提供弹性低成本的数据存储、按需伸缩的计算资源；二是以湖仓引擎为架构，在低成本基础上保障各种数据处理场景中数据加工处理灵活性、数据分析高性能性、异构数据源融合分析等特性；三是提供一体化数据治理与开发平台，以统一元数据为抓手，支持数据集成、治理、开发、分析、服务等一站式数据服务。此外，大数据领域流传一句话即“数据是资产，也是负债”，衡量企业大数据落地效果，核心还是要看业务落地成效，即在繁杂的业务场景下能高效实现数据价值挖掘的能力。AI的大发展驱动了数据分析的智能化和数据价值的深度挖掘。

在百度智能云服务企业用户做数据应用落地过程中，我们发现，帮助企业构建敏捷智能BI、全功能AI开发平台，可以更高效地驱动数据应用落地。综上，新型的数据基础设施、深度挖掘数据价值，以及保障数据安全，是构建大数据平台的关键。

2.4.2 治理为要

在信息化时代，信息系统建设是面向具体业务的，例如人力管理建设-人力资源管理系统、设备管理建设-管理系统、项目管理建设-项目管理系统等。因系统间缺乏关联互助、建设标准不同，带来数据孤岛、标准不一致等问题；而数据的全面性、数据质量、数据及时性与一致性等直接影响数据的应用和效果。

因此，数据治理是企业大数据应用落地的必要环节。数据湖架构中的数据治理工具平台，可以提高数据治理的实施效率。在数据治理的实施过程中，搭建专业的实施班子，做好数据资源、IT架构、业务情况的盘点，定好数据分类、数据模型、数据质量、应用指标的标准，再实施数据采集、数据分层建设，可以有效保障数据治理实施效率和效果。

2.4.3 安全为本

“没有网络安全就没有国家安全”。《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》以及《关键信息基础设施安全保护条例》等多部法律法规及管理办法对行业提出了规范性的要求。例如《数据安全法》明确要求，“数据安全，是指通过采取必要措施，确保数据处于有效保护和合法利用的状态，以及具备保障持续安全状态的能力”。因此，大数据应用要以数据安全为本，从数据全生命周期来看，数据安全保障主要包括数据资产安全保障、数据隐私保护、数据流通安全等。

百度智能云提供全栈数据安全防护体系，全力保障数据资产不被恶意访问、数据隐私不被非法侵犯，通过业界领先的网络安全技术、信息安全技术、隐私计算技术等综合安全能力，实现数据要素的全方位保障。

2.4.4 持续运营

要保障数据应用价值的持续性，数据资产和应用都需要“持续运营”。大到产业、小到每个具体企业，都具有动态发展变化性，数据和业务必然也持续变化。所以，大数据的应用落地不是一次性项目，数据基础设施、数据治理、数据应用等实施完成后，需要持续的管理和运营。

结合产业实践不难发现，平台化的管理方式可以提高运营效率，比如建设数据资产管理和运营平台，有利于帮助企业实现数据“好管理”、“好找到”、“好理解”、“好应用”。

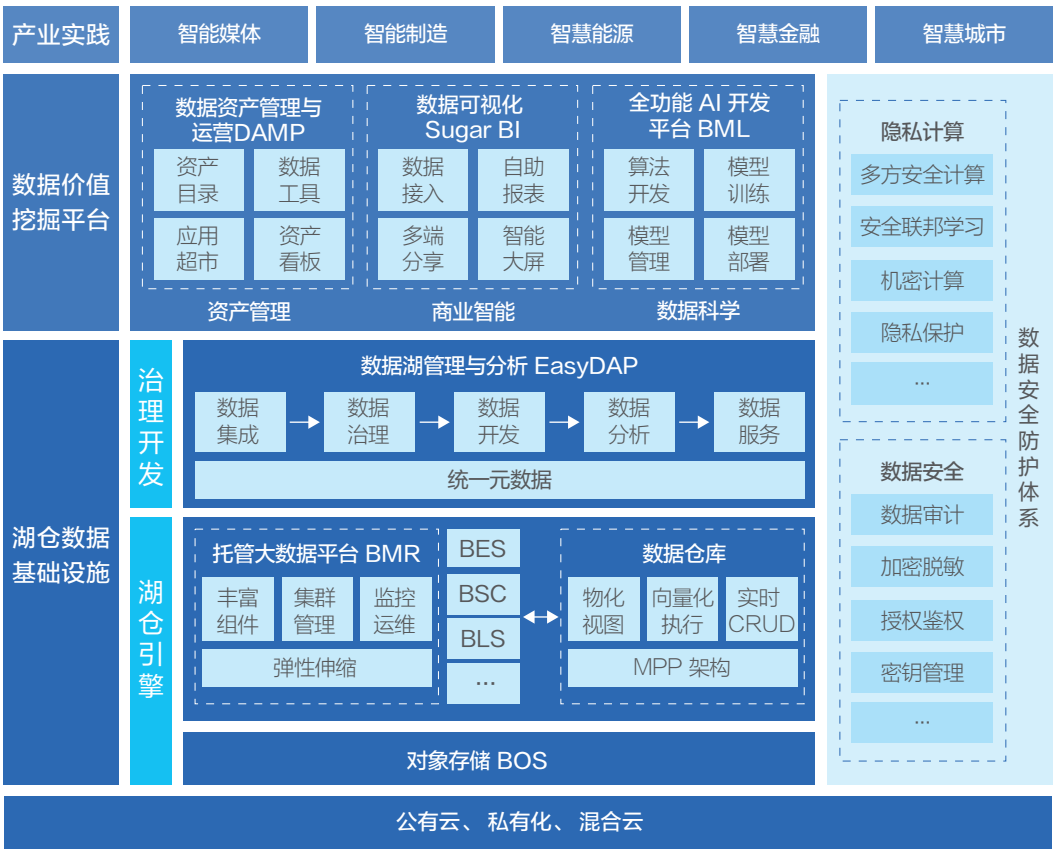
此外，运营过程配套建设运营体系比如数据覆盖、数据质量、数据成本、数据价值等，对于优化和指引未来规划价值也非常大。

2.5 百度智能云大数据产品架构全景图

基于以上原则，百度智能云以云和AI为依托，打造了云智一体的智能大数据产品架构全景图，注重实践经验积累，力争做到成效为先、治理为要、安全为本、持续运营，为更多企业构建云智一体的大数据技术与能力体系，实现对更大规模、更多类型、更多源数据进行处理和应用，全面助推企业数字化转型。

百度智能云大数据产品架构全景图共三层：底层，通过湖仓数据基础设施为企业提供数据存储、数据处理、数据开发等能力；中层，数据价值挖掘平台充分利用百度智能大数据技术，实现企业数据资产价值最大化；顶层，则基于底层和中层的技术，帮助各行各业落地大数据应用落地。以下篇章将做详细解读。

百度智能云大数据产品架构全景图

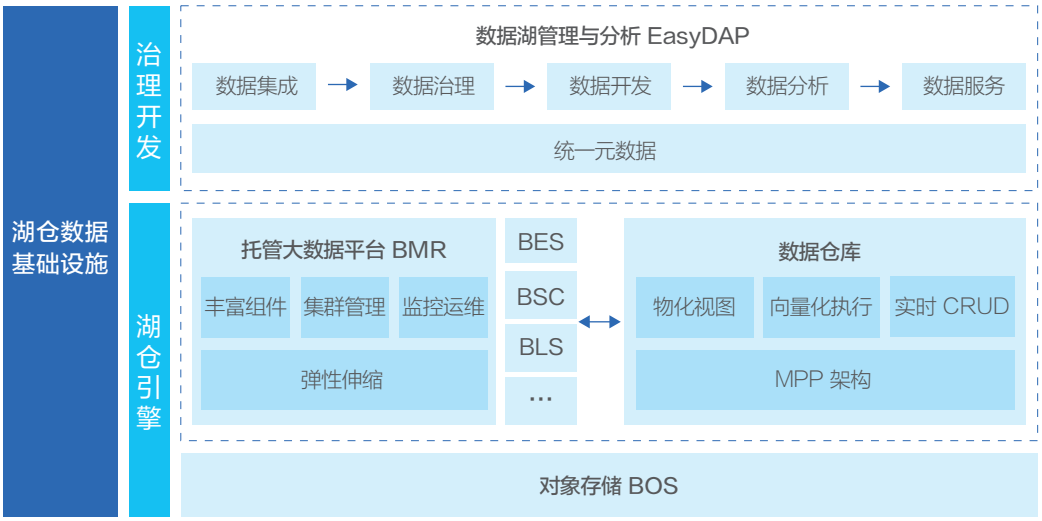


（图2.5 -1）

云原生湖仓架构，构建企业大数据基础设施 >>

随着企业数字化转型加速，企业日常运营中产生的数据量呈指数级增长，且数据类型更加多样化，数据的应用场景也日益繁杂，以及基于实时数据的快速决策越来越普及……因此，单一的数据仓库或者数据湖解决方案满足不了用户对数据挖掘和使用的需求，湖仓一体架构成为云原生时代数据架构演变的必然趋势。

云原生湖仓架构



（图3－1）

百度智能云湖仓一体架构的优势主要体现在：云原生、数据湖架构以及一体化平台，下面展开来讲。

3.1 云原生

大数据处理技术广泛应用于各个行业，为业务解决海量存储和海量分析需求，数据量的爆发式增长，对数据处理能力、基础设施成本、集群运维管理提出了更大的挑战。而云原生天生具有的高效部署、云计算资源成本和弹性扩展等优势，百度智能云大数据平台提供云原生架构的大数据集群服务，直接选型公有云大数据产品，即可享受按需创建、智能扩缩、运维托管等海量大数据处理能力。

3.1.1 按需创建

在百度智能云上，通过界面化点选的操作方式，即可根据业务场景可视化创建适合业务数据处理场景的BMR开源大数据处理集群、百度数据仓库、BES集群等。例如，在百度智能云创建BMR大数据处理集群，可按需选择组件如HDFS、Spark、Hive、Flink、Clickhouse等，同时提供不同组件版本的选择，平台自动进行组件参数推荐；按需配置集群套餐类型、集群节点规模、集群存储配置等，配置完成后，分钟级完成集群创建操作。

3.1.2 存算分离

数据总量指数级增长，相应地要求数据库有极致的弹性伸缩能力、持续服务能力和合理成本。传统的存算一体架

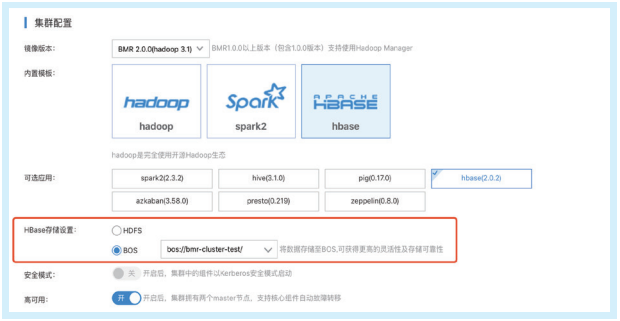
构弊端逐步显现，比如随着数据量和业务激增、资源调度互相影响、集群故障风险随着规模增大而剧增、不同集群数据无法共享等，导致系统架构的可靠性降低、运营困难和成本居高不下。

百度智能云云原生湖仓通过先进的计算存储分离架构，实现了计算资源和存储资源的灵活扩展，解决了数据同步的延时问题，并提升了计算横向扩展能力。表现在：集群因不可抗力出现异常时可快速恢复，可用性更高；数据存储更可靠，支持多组件数据挂载BOS对象存储，存储灵活，可弹性扩展，更方便，按需归档；集群节点灵活选配CDS云盘，多副本保证数据可靠不丢失。相比存算一体架构，这种计算存储分离架构更能兼顾性能、成本、灵活性等特点。

云原生 – 存算分离、冷热分离

存算分离架构，更高效、便宜

- 集群节点灵活选配 CDS 云盘
- 多组件数据存储支持挂载 BOS 对象存储
- Alluxio 等性能优化
- 结合 BOS 多层数据类型自动沉降



（图3.1.2 -1）

3.1.3 弹性扩缩容

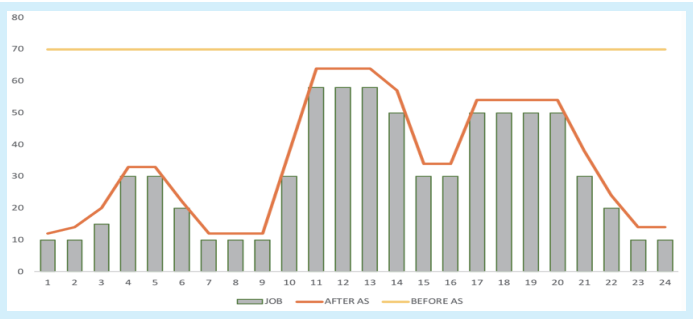
通过智能弹性伸缩，可以根据作业负载情况或业务周期时间，自动增减计算资源，不用为闲置的计算资源付费，降低用户成本。既可以基于场景按时间维度自由定制扩缩容规则，也可以根据CPU使用量、作业延迟度等自定义弹性伸缩规则。

云原生–弹性扩缩容

智能弹性伸缩，更高性价比

- 基于场景按时间自由定制扩缩容
  - √ 时间点
  - √ 周期
- 自定义规则，弹性扩缩容
  - √ CPU
  - √ 作业延迟等
- 智能分析，动态扩缩容
- 普通节点到竞价实例都支持

某公司上线 BMR 弹性伸缩后，成本下降40%



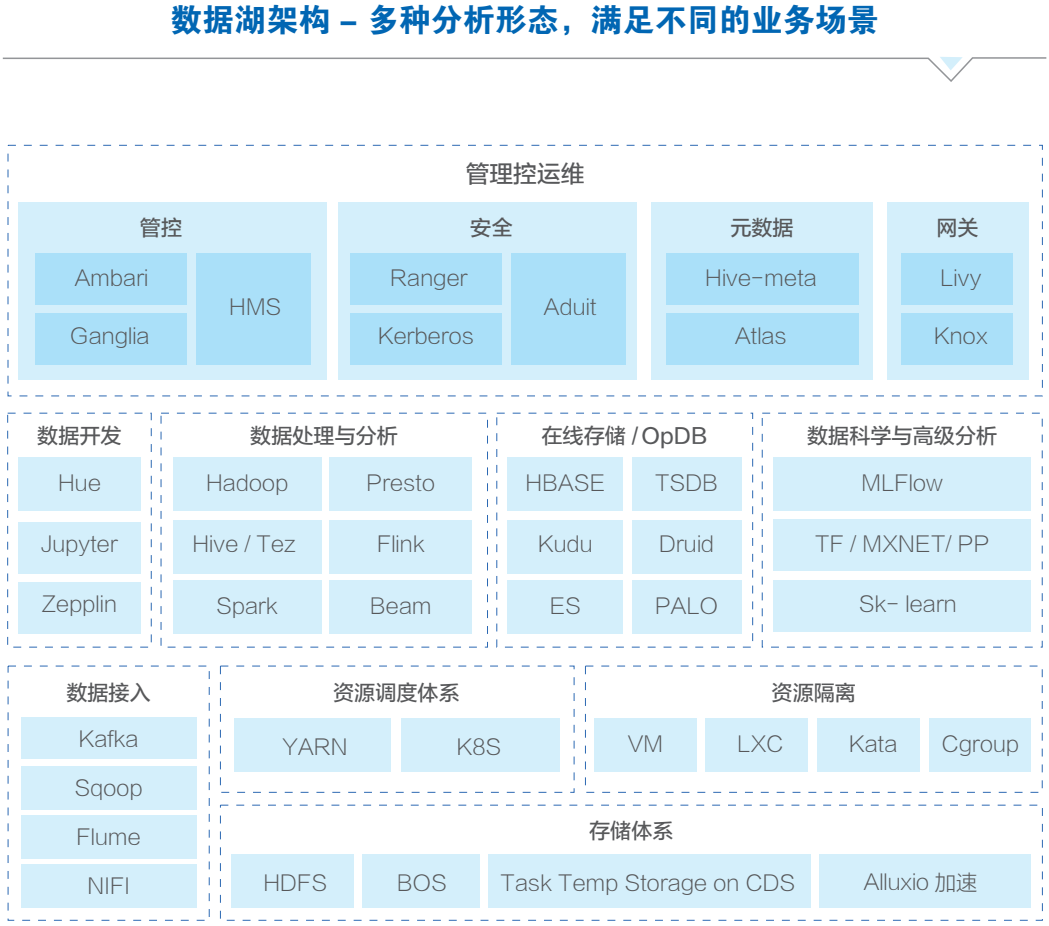
（图3.1.3 -1）

3.1.4 运维托管

提供全托管服务，用户只需专注于业务开发,无需关注部署运维,提供SLA保障、24小时技术专家支持、实时可视化监控运维告警大盘体系，更专业、更弹性、更可靠。

3.2 数据湖架构

数据在以指数级增长的同时，还逐渐呈现出类型更多样、更实时、价值密度低等特点，这些挑战加大企业IT设施的成本和复杂性，湖仓一体架构成为云原生时代数据架构演变的必然趋势。百度智能云大数据平台数据湖架通过BOS数据湖支持低成本存储任意规模结构化和非结构化数据；提供丰富的数据湖分析引擎，满足不同业务场景的数据分析需求，同时基于元数据的统一管理，支持数据湖分析、跨数据源联邦分析，为用户提供全面的数据分析能力。



（图3.2 -1）

3.2.1 BOS数据湖存储

百度智能云BOS，提供稳定、安全、高效、高可扩展的云存储服务。用户可以将任意数量和形式的非结构化数据存入BOS，并对数据进行管理和处置。BOS支持标准、低频、冷和归档等多种存储方式，满足多场景存储需求。具有如下特点：

海量规模。BOS提供超过EB级别物理空间，数万级别服务器，万亿级别文件规模，规模业内领先。帮助用户数据快速高效上云，既提供面向离线传输的磁盘传输方式，也提供面向在线的Cloud Flow产品，支持创建各种模型的在线传输任务。

高可靠性。BOS全面自研且自主可控的冗余存储架构，保障数据的持久存储能力，确保用户业务连续性。数据可靠性高达99.9999999999%（12个9），数据可用性高于99.95%。

极具性价比。BOS提供标准存储-多AZ、标准存储、低频存储-多AZ、低频存储、冷存储和归档存储等共 6级存储体系，成本逐级下降，用户可根据数据热度匹配合适的存储类型，合理控制成本。另外，BOS提供智能的生命周期管理和业界领先的读写性能，满足用户在大数据场景下的性价比需求。

行业首发智能存储。针对视频、图像、大容量文件等类型，百度智能云在BOS侧更好地封装集成了百度相关的AI能力，包括几十项图像的审核能力，以及图像增强和特效能力，这些特有能力的集成，能够轻松的降低编码复杂度。用户可以直接通过BOS的事件触发框架，在数据访问或者数据上传时，通过统一的接口拓展和调用这些智能处理能力、及时得到AI处理的效果，让数据分析和存储更近，降低管理、运维的复杂性。

（图3.2.1-1）

3.2.2 丰富的数据湖分析引擎

为满足数据湖数据的各种处理场景需求，百度智能云提供丰富的数据湖分析引擎，包括：兼容各类开源大数据处理分析组件的百度托管大数据分析平台BMR，提供了Spark、Flink、Hive、Hbase等多种数据分析处理引擎；百度数据仓库，专门应对高并发、低延时的PB级实时数据仓库分析场景；百度BES全文检索和分析引擎等。

- **托管大数据分析平台BMR**

百度MapReduce（BMR），定位于组件完备、高性价比、开源开放的企业级托管大数据平台，包括丰富的hadoop生态组件，如Hadoop、Spark、Hive、Flume、Storm、Flink、Hbase、Druid等，用户可通过界面灵活选择所需组件和配置，一键创建所需集群，并且通过可视化界面、API方式灵活管理集群。此外，先进的存储计算分离技术和弹性伸缩技术，确保高可靠的同时，真正帮助用户做到用时高效获取资源、闲时释放资源，帮助用户用低成本获得最高计算性能。百度BMR特点如下：

完备大数据生态组件。包括：数据集成组件Sqoop、Kafka等，离线计算组件MapReduce、Tez、Spark等，流式计算组件Flink等，数据仓库组件Hive、Hbase、Druid等，数据分析组件Impala、Presto等，作业调度组件Azkaban、Ooize等，安全管理组件Kerberos、Ranger、LDAP等。

一键创建高效运维。根据模版可视化一键创建、开箱即用，支持按需灵活配置，如选组件、选版本、选套餐。此外，全方位可视运维，如平台运行状态、作业执行分析、资源使用监控、多租户配置管理、监控报警配置等。

智能弹性扩缩。通过智能弹性伸缩，可以根据作业负载情况或业务周期时间，自动增减计算资源，不会为闲置的计算资源付费，降低用户成本。支持基于场景按时间自由定制伸缩规则，如指定时间点、按天、周、月等周期配置。自定义规则，弹性伸缩如根据CPU负载指标、作业延迟指标、自定义指标依据等。

存算分离。通过先进的计算存储分离架构，集群不可抗力异常时可快速恢复，可用性更高；数据存储更可靠，存储弹性扩展，更方便。

多层级、全方位安全机制。BMR提供多层级的安全机制保障集群平台和数据的安全。基础架构上，用户基于BMR创建的大数据基础平台运行在独立的VPC网络环境下，从网络上进行隔离，并支持用户主机安全审计。提供Kerberos 和 Ranger组件，为用户提供基于用户身份、作业提交和资源使用，以及数据访问的认证和鉴权，细粒度保障平台和数据的安全。

- **百度数据仓库**

百度数据仓库是基于Apache Doris（百度捐赠）构建的企业级MPP数据仓库，专门应对高并发、低延时的PB级实时数据仓库使用场景，全面兼容MySQL协议，可以针对亿万级数据实现毫秒级多维分析透视和业务探查。

在架构上来看，百度数据仓库与常见的分布式存储系统的架构有些不同，主要有FE（Frontend）和BE（Backend）这两类系统进程，其中FE可以理解为百度数据仓库的管控节点，主要负责用户请求的介入、查询计划的解析、元数据的存储以及集群管理等工作，BE主要负责数据存储以及查询计划的执行，这两类系统进程都可以横向拓展，而不需要依赖任何第三方系统（如HDFS、ZooKeeper等），这样高度集成的架构设计也极大简化了一款分布式系统的运维成本。同时百度数据仓库在FE进程中实现了MySQL兼容协议层，这样用户通过标准MySQL客户端或其他各类工具即可便捷连接到百度数据仓库，并且还支持标准SQL语言，不论是简

单的单表聚合、排序过滤抑或复杂的多表关联、子查询、窗口函数、自定义函数等，都可以通过SQL快速完成，极大减少用户使用成本。

使用百度数据仓库时，可以从本地、RDS、BOS、百度智能云MapReduce等导入海量数据，进行大数据的多维分析。同时它还兼容主流BI工具，数据分析师可以通过可视化的方式分析和展示数据，快速获取洞察以辅助决策。此外，其还提供了全新UI支持，5分钟上手，即可轻松实现建库建表、数据导入、数据查询。百度数据仓库特点：

极致性能。采用了现代化MPP架构，使用高效列式存储引擎和向量化执行引擎，加以智能物化视图技术，可以实现极致性能，100台集群可达10wQPS，无并发瓶颈。

简单易用。提供标准SQL支持，完全兼容MySQL协议，同时，灵活的数据模型，Join表现优秀，并具有在线表结构变更、Bitmap索引精确去重等贴近业务场景的实用功能。

流批一体。可以支持批量和实时流式数据导入，行级别数据更新/删除，多版本机制解决读写冲突，导入事务支持，保证ACID，实现Exactly-Once语义。

极简运维。运维方面，百度数据仓库架构高度一体，无任何外部组件依赖，集群规模弹性伸缩，任何节点可线性拓展，无代码即可完成运维。

高可用性。对数据库而言，最核心的宗旨就是要稳定。百度数据仓库在稳定性方面，做到了主节点高可用，数据多副本存储，节点故障自动副本迁移，自动请求路由，数据分片自动均衡。

开放生态。百度数据仓库源自百度开源并捐赠的Apache Doris，核心代码全部对外开放，有近200名开发者曾为项目贡献代码；可以与主流大数据生态，比如数据源端的BOS/HDFS/Kafka等数据无缝导入、与Spark进行联邦数据分析、为ES提供分布式SQL查询，以及与主流BI工具进行适配。

（图3.2.2 -1）

• **百度Elasticsearch**

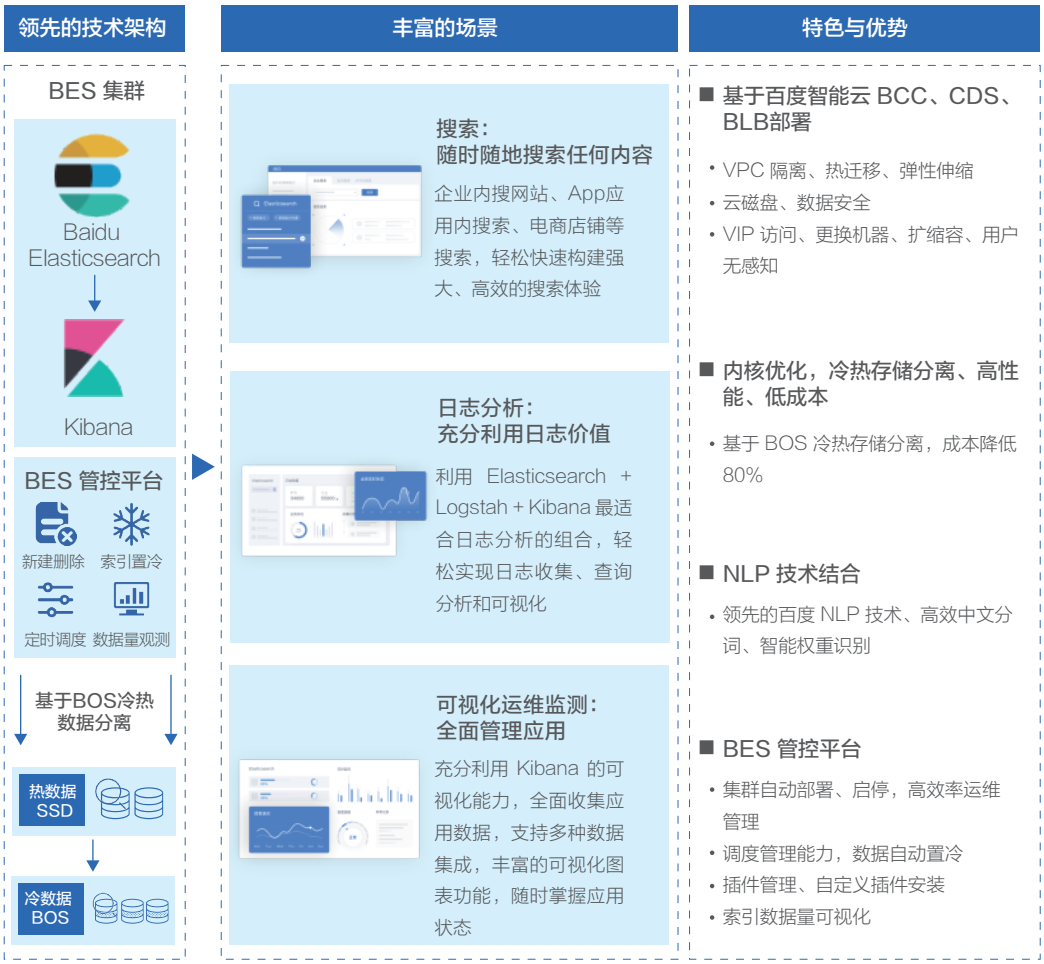
百度Elasticsearch（BES）是开源的全文检索和分析引擎Elasticsearch的托管服务，完全兼容开源Elasticsearch功能，同时内置了基于百度NLP技术的分词插件，以及即开即用、自动运维以及丰富的监控指标，减少用户运维开销，帮助用户快速启动业务分析。技术领先的冷热数据分离能力，帮助用户减少50%+的存储成本。同时具备冷数据可查询功能，用户使用流程无差异。BES应用场景包括：

日志分析。利用Elasticsearch+Logstah+Kibana最适合日志分析的组合，轻松实现日志收集、查询分析和可视化，充分挖掘和利用日志数据，辅助实现应用监控、设备运维等。

企业搜索。Elasticsearch作为搜索引擎，提供高性能且轻量的搜索能力，常用于企业内搜网站、App应用内搜索、电商店铺等搜索，轻松快速构建强大、高效的搜索体验。

可视化运维监控。Kibana作为Elastic生态一员，提供强大的可视化能力，全面收集应用数据，支持多种数据集成，丰富的可视化图表功能，随时掌握应用状态。

文本检索分析 – BES

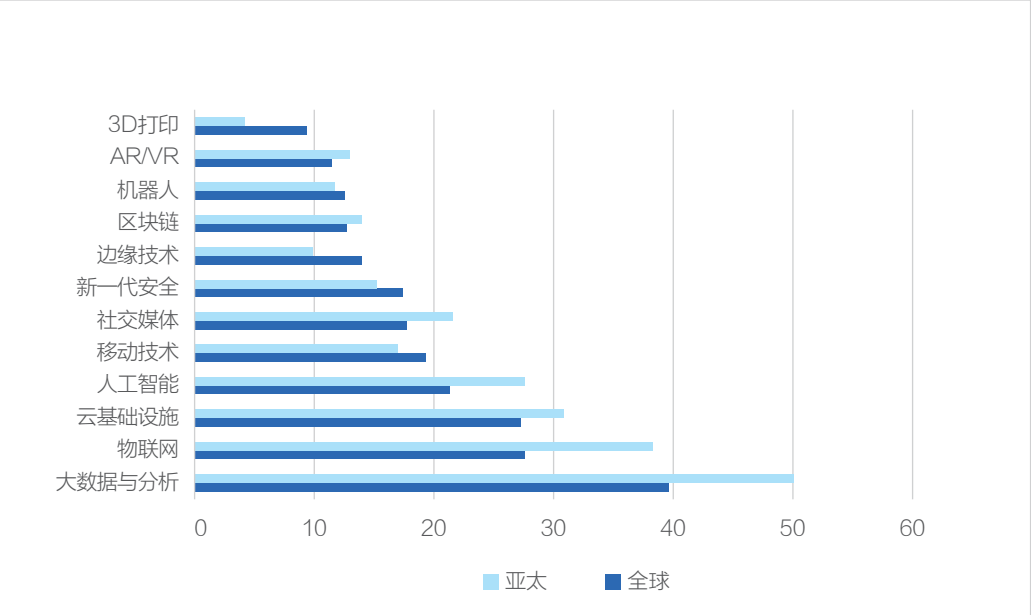


（图3.2.2-2）

3.3 一体化平台

根据相关机构的调研显示，大数据已成为数字化转型战略第一要务。企业要数字化转型，需要利用大数据技术来构建有价值的数据资产，并通过各种大数据分析技术，基于业务场景来进行数据分析，破解企业生产经营中的应用难题，实现业务的转型和创新。而大数据技术的复杂性、多样性、技术迭代等特性，为企业数字化转型中的大数据应用落地，带来了巨大挑战。

大数据为数字化转型战略中第一要务



数据来源：中国大数据平台市场研究报告 -2020，IDC

（图3.3 -1）

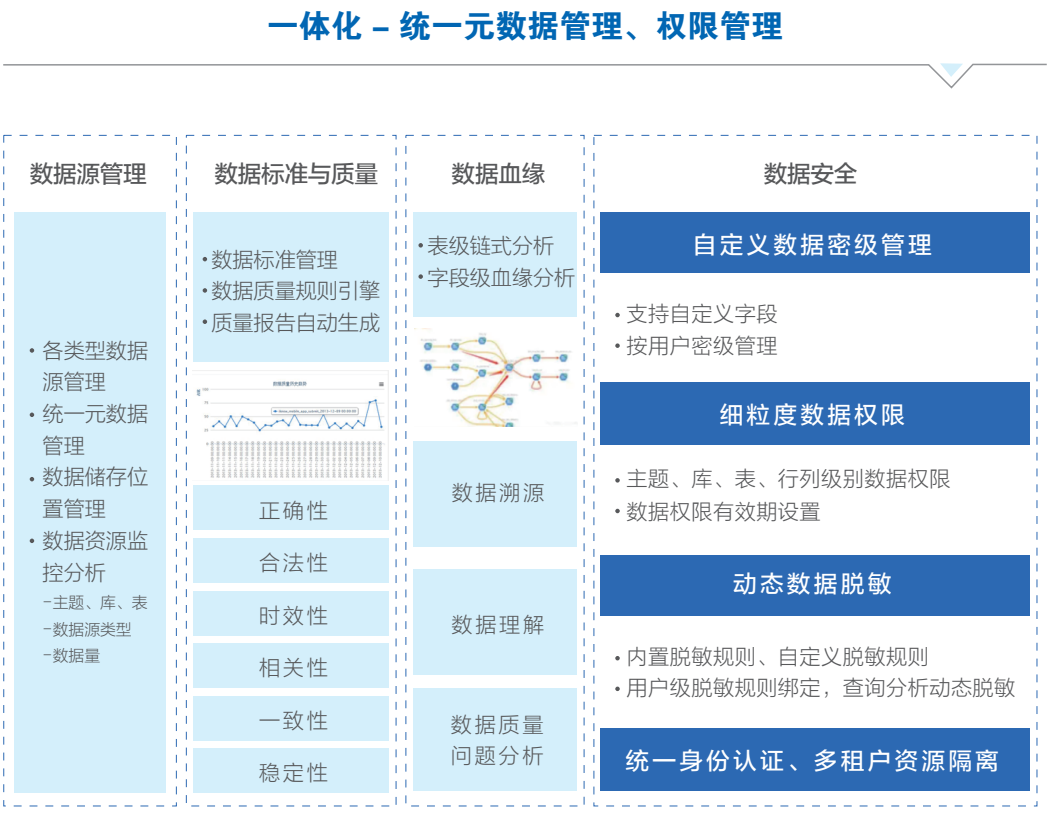
百度智能云大数据平台提供全场景、低门槛、开放兼容、安全可靠的一站式数据管理与分析平台EasyDAP。EasyDAP以数据流为核心，形成数据汇聚、开发、分析、管理、服务等一套完整产品体系，并提供简易的可视化开发操作界面，降低使用门槛。此外，其可以无缝对接主流大数据存储计算平台，兼容企业原有IT系统、技术栈等，并与百度智能云产品生态打通，降低数据中台的建设成本。同时，还具备资源隔离、权限控制、审计日志等安全能力，提高数据安全性。EasyDAP功能特性如下：

3.3.1 元数据统一管理

百度EasyDAP的元数据管理通过物理表、映射表的方式，将数据采集、开发、管理、分析所涉及到的企业各类数据源系统、数据仓库、数据湖中的相关元数据，纳入到EasyDAP平台进行元数据统一管理。实现全域数据的高效管理，以满足数据移动、数据权限管理、全域数据血缘分析、跨数据源联邦查询等需求。

百度EasyDAP的元数据管理，支持自定义主题，可灵活定义数据的组织结构，基于企业的组织结构、数据内容进行灵活的目录结构组织。结合元数据统一管理，在EasyDAP上可以实现数据库、表、行、列级别的权限控制。同时，EasyDAP提供MySQL等数据源的元数据自动同步。

为保证平台的开放性以及适配能力，元数据平台的核心能力可以通过API方式，对外开放服务接口，方便业务/作业人员进行数据的接入、管理和使用。



（图3.3.1 -1）

3.3.2 自动数据集成

在EasyDAP上可以通过两种方式实现数据的集成同步，即可视化ETL和整库整表数据同步。

可视化ETL，不需要写代码，通过可视拖拽式即可完成数据集成流水线创建，支持离线数据和实时流数据集成，离线数据同步。同时，支持数据库、MPP数仓、NoSQL、对象存储、HDFS分布式文件系统等几十种异构数据源的定时、周期性同步。支持http、kafka源数据实时流同步到DB、MPP数仓、NoSQL、对象存储、HDFS分布式文件系统等场景数据源。

针对场景的数据库、数据仓库到大数据平台的数据同步，EasyDAP提供整库整表的数据集成方式，提升数据入湖入仓效率。此外，数据库CDC实时同步能力，也在升级建设中。

3.3.3 全流程可视化数据治理开发

在企业数据资产建设和开发应用实践中，一般数据治理、开发、实施不分家。百度EasyDAP提供全流程可视化的数据治理开发能力，高效实现数据资产构建，具体包括：

数据标准。支持数据标准的创建、审核、发布，用于描述公司层面需共同遵守的数据含义和业务规则，同时支持码表管理，高效创建数据资产元数据、或者引入作为数据规范。

数据质量。内置常见数据质量规则引擎和模板，同时支持自定义数据质量规则引擎，可基于需求建设数据质量任务，自动生成数据质量报告。

数据血缘。支持自定义生成数据血缘，同时支持手动填报，保障数据血缘的完整性、准确性。

数据安全。支持库、表、行、列级别的数据权限管理，支持动态数据脱敏等。

离线开发。离线开发场景下，平台集成了MR、Spark、Hive离线数据处理引擎，提供SparkSQL、HSQL、Python、Scala等多语言支持。此外，可视化配置作业调度DAG，将多个作业组成工作流，按设计调度计划，保证平台中的业务数据流定期调度运行。

实时开发。实时流开发场景下，平台集成了Spark-streaming、Flink等多种流计算引擎，支持简易的在线SQL、自定义Jar程序等开发模式，以及高吞吐低延迟的实时计算、全方位多角度的监控告警机制等。

3.3.4 数据湖分析

提供数据湖分析能力，支持交互式分析多类型数据源数据、支持跨数据源联邦分析等能力，数据源类型包括但不限于HDFS，BOS对象存储，RDBMS，NoSQL，MPP等。具体特性如下：

易用高效。支持PB级数据高并发实时写入、实时查询，完全兼容PostgreSQL协议。

跨源联邦分析。直接分析RDS、PALO、BMR+Hbase、BOS、BMR+Hive、BMR+HDFS的数据。

场景丰富。支持快速点查询，支持复杂大数据分析。

优化分析。提供SQL耗时、扫描数据量等分析，指导优化。

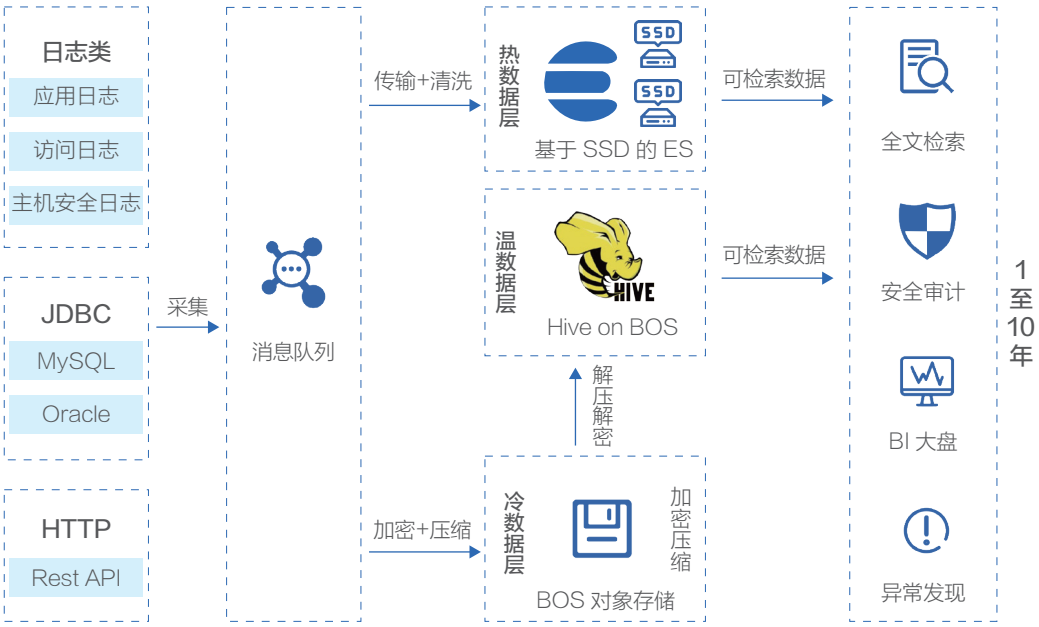
此外，还与大数据生态无缝打通，支持海量实时和离线数据的实时分析。

3.4 成功实践

度小满是一家成立于2015年的金融科技公司，其前身为百度金融，2018年完成拆分实现独立运营，在独立运营之初构建了基于IDC物理机房的日志系统。2020年，度小满开启上云进程，由百度智能云提供云服务，同时对之前的日志服务进行了技术改进。

在改进之前，度小满日志系统检索相关的架构，对高低频数据进行了分层存储，对于高频数据需求是基于SSD的ES落地的，对于低频的数据需求，则是基于对象存储的Hive来进行落地的。

基于 ES & Hive 的日志检索框架

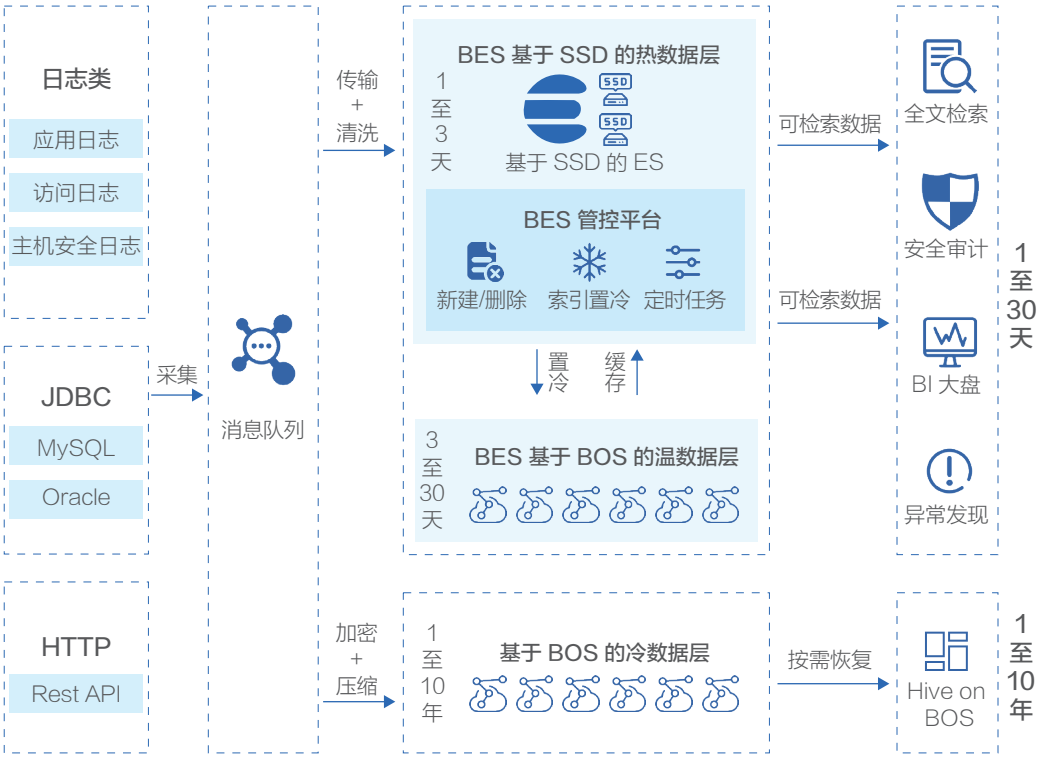


（图3.4 -1）

该架构痛点：一是随着数据量逐步上涨，基于SSD的ES存储成本升高，热数据只能保留7天；二是温数据层的hive检索框架，下游业务采用模糊查找检索，速度较慢，通常需要半小时甚至更长时间；三是数据加温过程中，恢复速度较慢，不能及时满足业务预期；最后，基于时长划分的日志存储架构，系统复杂度高，接口不统一。

针对以上痛点，百度智能云为度小满搭建了基于BES的日志检索框架。在数字采集、传输、清洗等日志场景基本保持不变的前提下，度小满基于BES的冷热分离架构，用BES热数据能力替代基于SSD的ES，用BES冷数据能力替代hive方案。

基于 BES 的日志检索框架



（图3.4 -2）

整体调优收益：

写查速度大幅提升。热数据写入SSD，导入性能不受影响；TB级索引，热数据秒级响应；温数据检索时效由半小时缩短为分钟级响应。

存储周期更长。可直接被检索数据，由7天全面升级为30天，最高可支持180天；得益于存储周期变长，可被直接检索数据达到PB级。

资源成本更低。BOS对象存储成本是SSD磁盘的13%，存储成本降低90%；ES资源需求量是之前的1/2；索引冷热属性设置方便，提升操作效率。

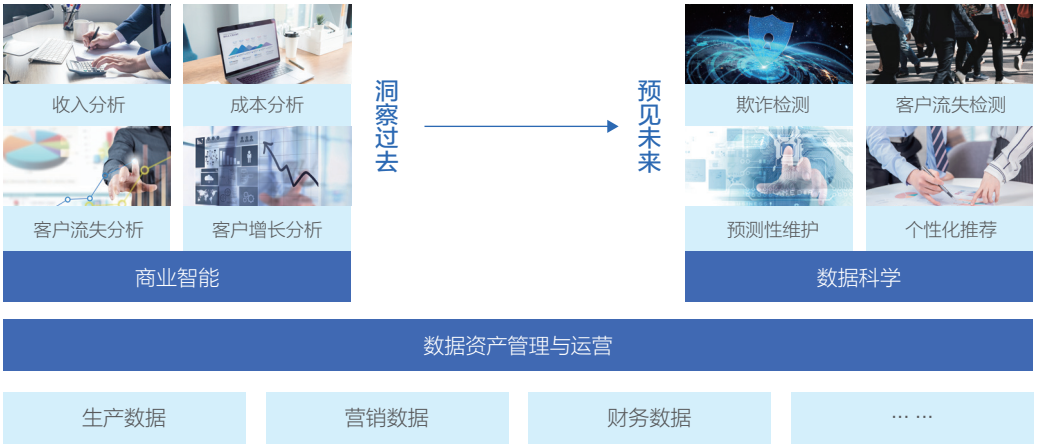
智能数据挖掘，驱动数据资产价值最大化>>

完成数据基础设施构建后，企业要实现数据资产价值最大化，需要进一步解决数据价值挖掘的问题。数据价值挖掘的第一步，需要把企业内生产、营销、财务、人力等各类原始数据，进行治理和挖掘形成数据资产，并管理起来。

数据资产的价值当前有两大类：一是商业智能，二是数据科学，前者帮助企业更好的洞察业务的过去，如各类报表、领导驾驶舱等都属于这一范畴，后者让企业预见未来，对业务进行预测性分析。

为应对业务发展的不确定性，企业需要利用机器学习、深度学习等技术，在数据科学领域预测未来趋势，如欺诈检测、客户流失检测、预测性维护以及个性化推荐等。通过智能技术的引入，让数据资产价值最大化。

智能数据挖掘，驱动数据资产价值最大化



（图4 -1）

在数据价值挖掘平台方面，百度智能云推出了三大产品，即数据资产管理与运营平台DAMP、数据可视化Sugar BI，以及全功能AI开发平台BML。

数据价值挖掘平台产品特性

数据资产管理与运营 DAMP

- 资产目录
- 应用超市

数据可视化 Sugar BI

- 智能图表生成
- 智能交互大屏

全功能AI开发平台 BML

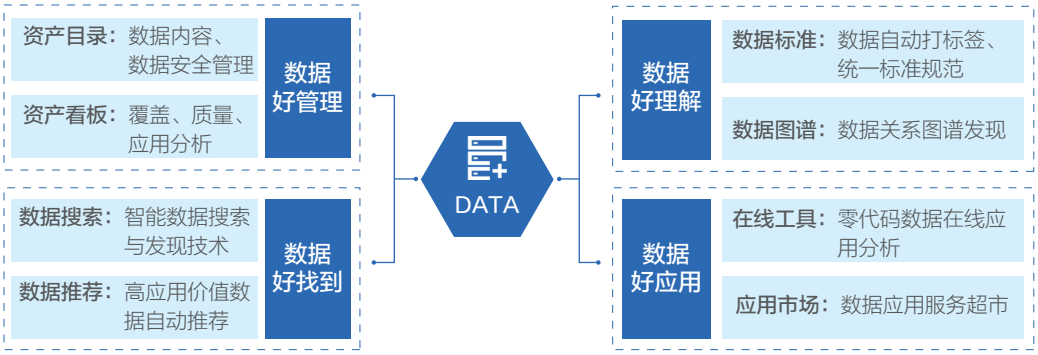
- 一站式、多开发场景
- AutoML& AutoDL

（图4 -2）

4.1 数据资产管理与运营平台DAMP

百度智能云通过数据资产管理与运营平台DAMP，将各类数据经过基础治理后形成的数据资产进行统一管理，以资产目录的形式让企业内部资产更清晰化，同时通过应用超市帮助企业更好的运营数据资产，实现数据资产要素好管理、好找到、好理解、好应用。

数据资产管理与运营平台DAMP特性



（图4.1-1）

百度智能云通过数据资产管理与运营平台DAMP的核心能力如下：

4.1.1 数据资产目录

对有价值、可分析和应用的数据做提炼，形成数据目录，并同步管理和展现元数据信息、数据质量、数据血缘、数据获取方式等数据详情。通过数据资产目录，方便数据分析和应用人员高效了解数据、找到数据、理解数据。

4.1.2 数据在线应用工具

基于整合多种数据源的数据资产目录，为分析师、产品经理、运营人员等提供在线数据融合分析功能，打破原有报表、工具的定制限制，支持零SQL基础的人员进行可视化拼接查询条件、或直接SQL查询，提供通用分析模板。

此外，数据基于业务化视角提供业务衍生标签的自定义功能，为上层应用提供统一的标签数据目录和标签调用接口，沉淀上层应用制作的模型标签，实现高价值标签共享复用。同时，提供标签圈群、群体分析等能力，帮助企业深入洞察圈定的目标群体、实现精准分析。

4.1.3 数据服务超市

以数据API形式将形成的数据资产目录、数据标签等对外开放，提高数据的应用和流通效率。同时，支持灵活创建、API注册，并配套提供激励机制，驱动数据服务API市场繁荣。

4.1.4 数据资产看板

基于数据资产统一管理和运营角度，直观分析、展示各类数据资产汇聚的覆盖率、数据资产综合质量、数据流通应用、数据资产关系图谱等情况，方便数据资产管理和运营者综合了解企业的数据资产情况，指导数据战略决策。

4.2 数据可视化Sugar BI

在商业智能方面，Sugar BI可以快速搭建数据可视化页面，帮助企业用户洞察过去。Sugar BI是百度自助BI报表分析和制作可视化数据大屏的工具，直连MySQL、本地Excel等各类数据源，通过丰富的图表和拖拽式编辑帮助用户5分钟即可生成可视化页面，并以炫酷大屏呈现，让数据信息更直观。同时，Sugar BI融合了百度语音、语义识别等多种AI技术，用户通过语音的方式就可以快速获取想要的数据。Sugar BI的核心特性包括直连多数据源，丰富的图表和拖拽式编辑方式、炫酷大屏，以及智能交互。

百度Sugar BI核心特性

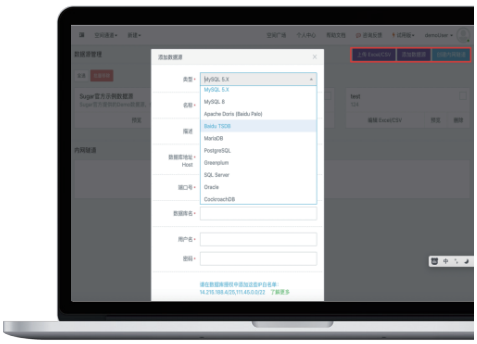
（图4.2 -1）

4.2.1 直连多数据源

Sugar BI支持MySQL、SQL Server、PostgreSQL等市场主流数据库都支持，同时支持本地Excel/CSV文件上传、API和静态JSON数据接入，并可通过提供的内网隧道功能，安全地访问企业内网数据库。还可支持来自同源数据库多张表交叉分析，以及跨数据库联合交叉分析。

直连多数据源，无缝打通云上数据

- 支持多种数据源，类型包括：MySQL、SQL Server、PostgreSQL、Oracle、IBM DB2、SAP Hana 等，并持续扩充
- 同时支持本地Excel/CSV文件上传、API 和静态 JSON 数据接入，并可通过提供的内网隧道功能，安全地访问您的内网数据库
- 支持同源异库交叉分析与跨源交叉分析



（图4.2.1-1）

4.2.2 丰富的图表和拖拽式编辑

基于拖拽式编辑来实现丰富的图表和大屏内容，无需代码，降低使用BI分析门槛。

（图4.2.2 -1）

4.2.3 炫酷大屏与智能交互

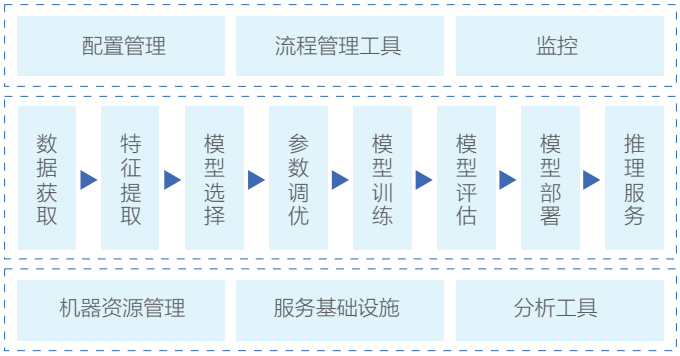
Sugar BI基于领先的人工智能交互技术、实现大屏语言交互功能，处于国内领先地位。

（图4.2.3 -1）

4.3 全功能AI开发平台BML

在数据科学方面，百度智能云通过全功能AI开发平台BML为数据科学场景提供全流程开发支持，帮助用户预测未来。BML整合了大数据和百度AI技术，可以实现从数据源管理、数据清洗与扩充、数据标注、数据预处理，到模型构建、模型管理与优化、预测服务部署、服务管理与监控等全流程能力支撑，降低企业应用数据技术的门槛。

BML为数据科学提供全流程开发支持



（图4.3 -1）

4.3.1 BML总体架构

BML可以训练图像、视频、语音等非结构化数据进行深度学习，还可接入各种结构化数据。在数据准备阶段提供数据分析和模型构建，在模型输出阶段提供模型管理、模型部署和应用构建。

全功能AI开发平台BML总体架构

（图4.3.1 -1）

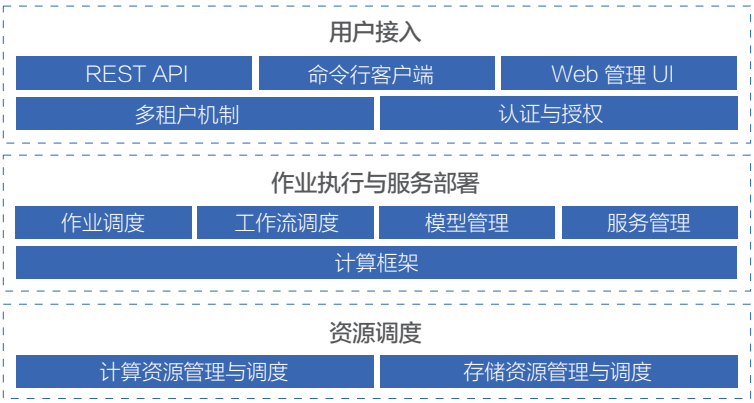
4.3.2 BML为数据科学提供的核心功能

BML为数据科学提供高效的算力管理和调度、高性能数据科学引擎、自动机器学习、丰富的建模方式四大核心功能。

在算力管理和调度方面，BML提供计算资源、存储资源的管理和调度。在这之上，提供一套作业执行与调度机制，帮助用户实现模型与服务管理。

高效算力管理与调度

- 异构算力资源池化管理
- 作业及工作流调度系统
- 模型管理与服务管理
- 丰富的计算框架支持

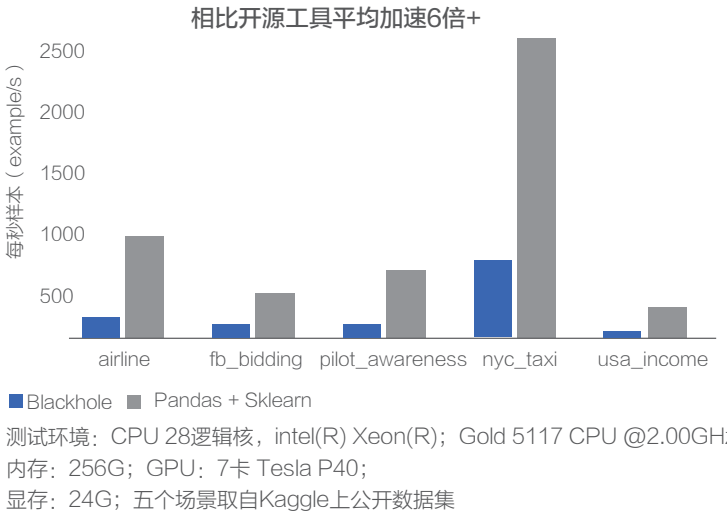


（图4.3.2 -1）

在高性能数据科学引擎方面，在数据科学场景下，从业者更倾向与机器学习强相关的Pandas/Sk-learn，因此BML提供高度兼容Pandas/Sk-learn接口，面向单机的数据分析和机器学习，提供5-10倍的开源工具的数据处理能力。

高性能数据科学引擎

- 面向单机数据分析、机器学习
- 5-10倍于开源工具的数据处理能力
- 高度兼容 Pandas/Sk-learn/SQL 接口



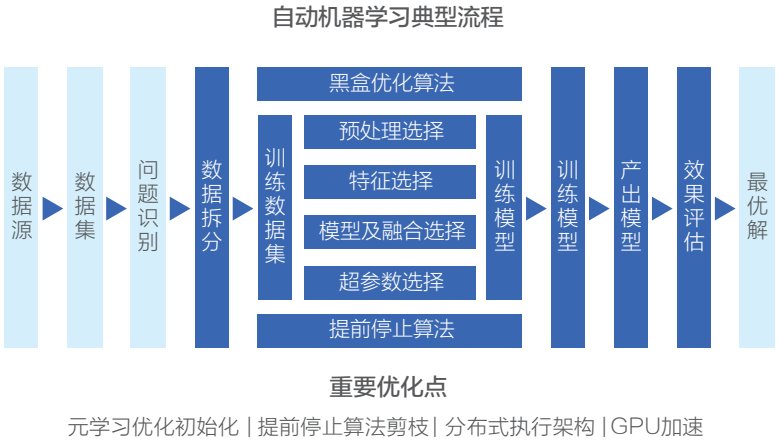
测试环境：CPU 28逻辑核，intel(R) Xeon(R)；Gold 5117 CPU @2.00GHz；
内存：256G；GPU：7卡 Tesla P40；
显存：24G；五个场景取自Kaggle上公开数据集

（图4.3.2 -2）

在自动机器学习方面，BML提供自动建模工具，实现从数据拆分、训练数据集、黑盒优化算法、模型训练、效果评估等全流程的自动化。

自动机器学习

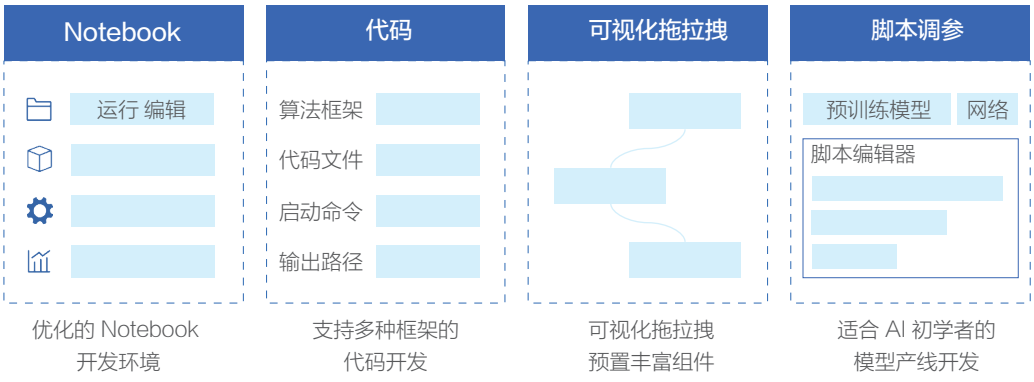
- 全流程自动化
- 组件可插拔可扩展
- 多种特征组合选择
- 结果方便人工精调



（图4.3.2 -3）

在丰富的建模方式方面，BML提供丰富的交互界面、文本编辑器、可视化的拖拉拽、脚本调参等工具，覆盖多种不同的建模场景。

丰富的建模方式



（图4.3.2 -4）

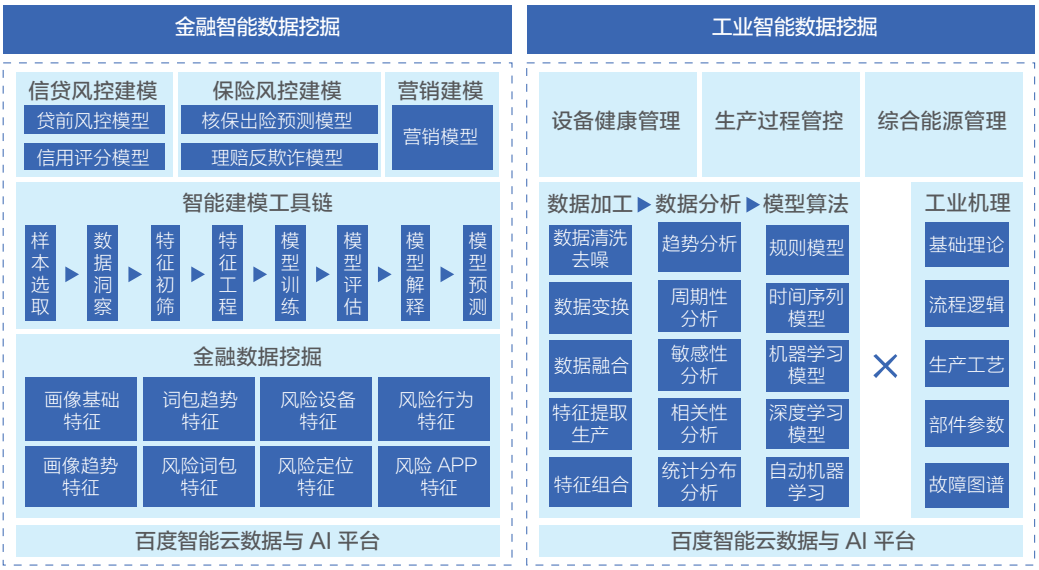
4.3.3 面向行业的智能数据挖掘解决方案

面对各行业的不同需求，百度智能云打造各种场景化解决方案，以金融和工业解决方案为例：

金融智能数据挖掘解决方案分为三层，底层是金融数据挖掘，中层是智能建模工具链，其基于BML开发平台，上层是各类风控建模。

工业智能数据挖掘解决方案，基于百度BML为数据加工、数据分析、模型算法、工业机理等流程提供设备健康管理、生产过程管控和综合能源管理等功能保障。

面向行业的智能数据挖掘解决方案举例



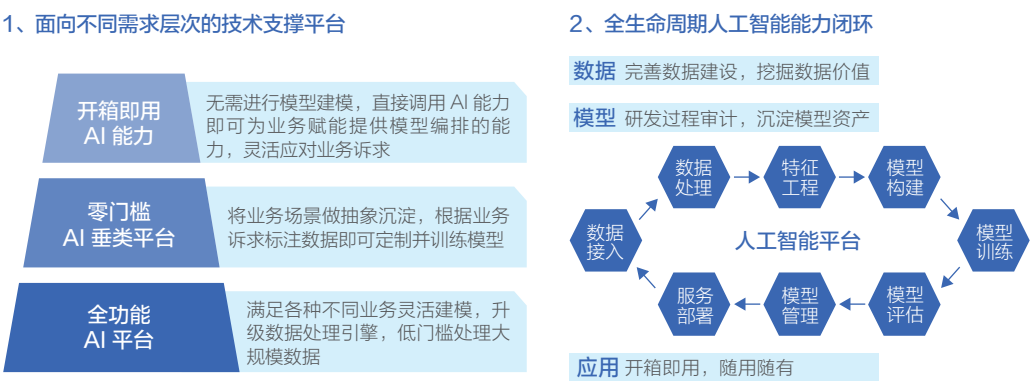
（图4.3.3 -1）

4.4 智能数据挖掘成功实践

4.4.1 邮储大脑机器学习平台

中国邮政储蓄可追溯至1919年开办的邮政储金业务，至今已有百年历史。2007年，中国邮政储蓄银行有限责任公司正式挂牌成立，2012年整体改制为股份有限公司。邮储银行在人工智能方向有三大目标：一是做一套全层次支撑平台，包括全功能底层、垂类AI场景以及AI能力；二是形成数据建设、构建模型、模型应用和应用数据回流等数据智能闭环。

邮储银行人工智能平台目标

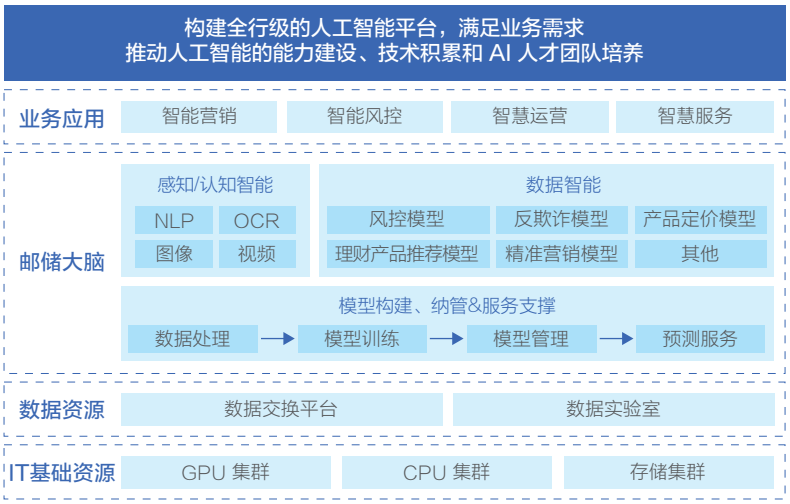


（图4.4.1-1）

百度智能云联合邮储银行，构建全行级人工智能平台，完善整体能力建设、技术积累和AI人才培养。在赋能效率方面，每天可运行数百个模型训练与预估服务，可实现14亿行、80多个维度的大规模数据训练，更新周期从月级缩短到小时级，图像模型能力扩容从天级缩短到秒级。

邮储大脑机器学习平台方案

- 全行统一的人工智能平台
- 完善的安全机制
- 科学的管理体系
- 实现全生命周期支持
- 满足合规管理要求
- 提升智能化赋能效率
  - √ 数百模型训练任务、预测服务同时运行及部署
  - √14亿行、800多维大规模数据训练
  - √ 相同模型训练更新周期从月级缩短到小时级
  - √ 图像模型部署扩容从天级缩短至秒级



（图4.4.1-2）

4.4.2 某集团研发中心设备健康管理系统

某集团研发中心属于轨道交通领域的一家车辆主机厂，基本业务包括整车以及关键部件智能运维检测服务，支持城轨列车、市域车的整车研发设计和制造工作。

2019年中央发布了交通强国的建设纲要，2020年中国城市轨道交通协会也发布了智慧城轨发展纲要，这两项政策性的纲要都提到要尽快引入人工智能等先进技术，加快推动产业的升级，这为该研发中心的业务发展提供了政策性引导和方向。

截止2022年4月，国内已有45个城市200多条共计7500多公里的地铁线路；城铁快速发展背后是传统运维模式所带来的矛盾日益凸显，带来运营运维成本日益增大。相关机构预测“需要在未来的50年投入等同于现有资产3倍价值的运维费用，才能保持现有的运营表现”。这是很大运维压力，也同样是一个重大机会，基于现有环境和压力情况，很多地铁业主在探索检修模式的转变，比较典型的就是从传统的计划修、故障修、定期维修，转变为预测修，并把这种转变列入中远期的目标和规划。

为了响应这种转变需求，该研发中心提出了包括状态检测、健康评估、寿命预测及延寿评估等业务，覆盖设计、运营和报废等阶段的全生命周期智能运维技术服务。这种运维服务横跨整车多个维度，将整车业务需求逐层分解，从整车分解到关键系统，再到关键部件和节点，逐步突破，化整为零，降低运营故障率和运维成本，提高乘客服务质量。

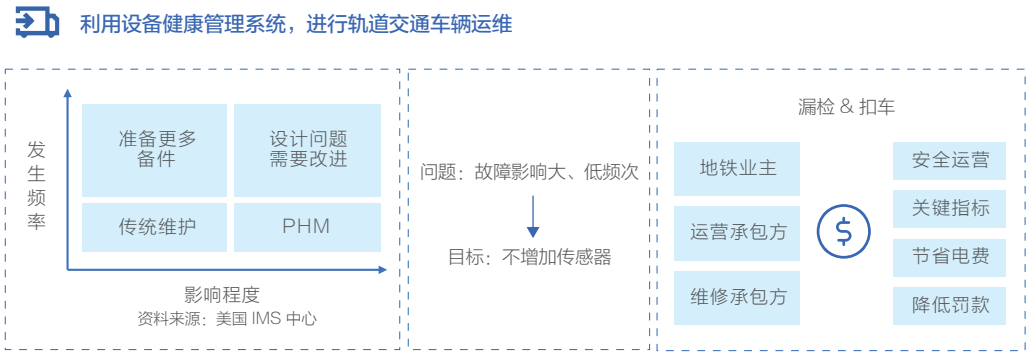
业务需求

（图4.4.2 -1）

业务痛点：场景应用方面，面对庞大数据压力情况下的数据分析能力不足。当前每列车每天可产生2G左右的数据，像上海有7000+辆车（1100多列，按6节编组估算）、杭州有5000+辆车（800多列，按6节编组估算），数据压力非常大，传统数据分析能力无法满足要求；再加上需要做预测模型，需求的算法模型非常复杂。

该研发中心与百度智能云结合业务多轮沟通之后，共同确定了解决方案，通过探索构建设备健康管理系统支持轨道交通车辆运维服务，并参照美国IMS中心提出的运维方式选择模型，选择故障影响大且频次低的项点来研究。同时，该方案还要重点考虑，尽可能充分利用现有车辆的数据资源，减少对车辆整改。

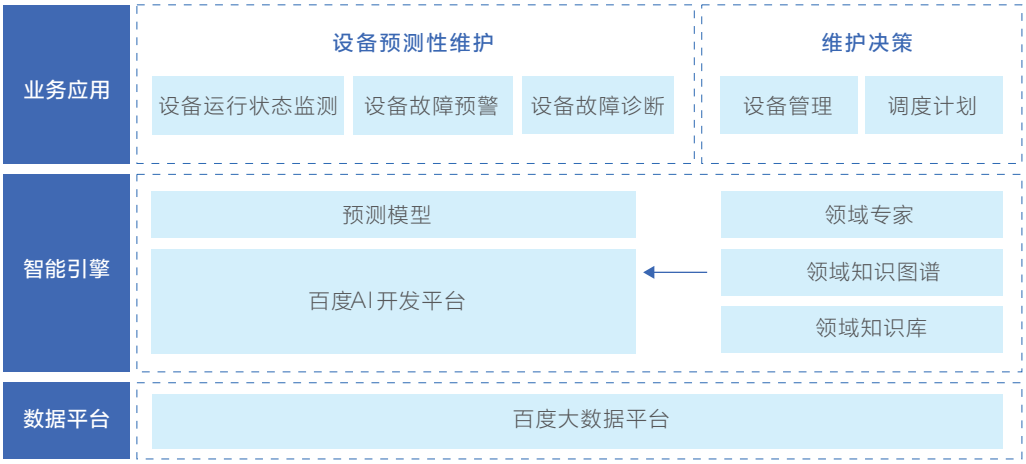
解决方案总体思路



（图4.4.2 -2）

解决方案的架构主要分为三层：一是底层数据平台，二是结合行业知识和百度AI算法平台形成的智能引擎层，用于输出各项预测模型，第三层包括设备预测性维护和维护决策。

设备健康管理系统架构

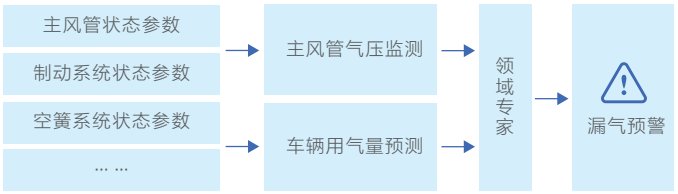


（图4.4.2 -3）

简单的举一个研发模型应用的例子，“制动系统主风管漏气预测”模型。轨道交通车辆在运营过程中制动装置、空簧等装置都需要大量用气，主风管漏气会严重影响列车的运营，轻度漏气会导致车辆运营的能耗增加，重度漏气会引发车辆救援事故。模型通过AI算法识别压力变化估算车辆用气的整体情况，建立智能决策模型，预测列车是否漏气，以尽早对车辆进行检修，减少车辆运营能耗，防止事故的发生。

应用场景示意 – 制动系统主风管漏气预测

轨道交通车辆在运营过程中制动装置、空簧等装置都需要大量用气，主风管漏气会严重影响列车的运营，轻度漏气会导致车辆运营的能耗增加，重度漏气会引发车辆救援事故



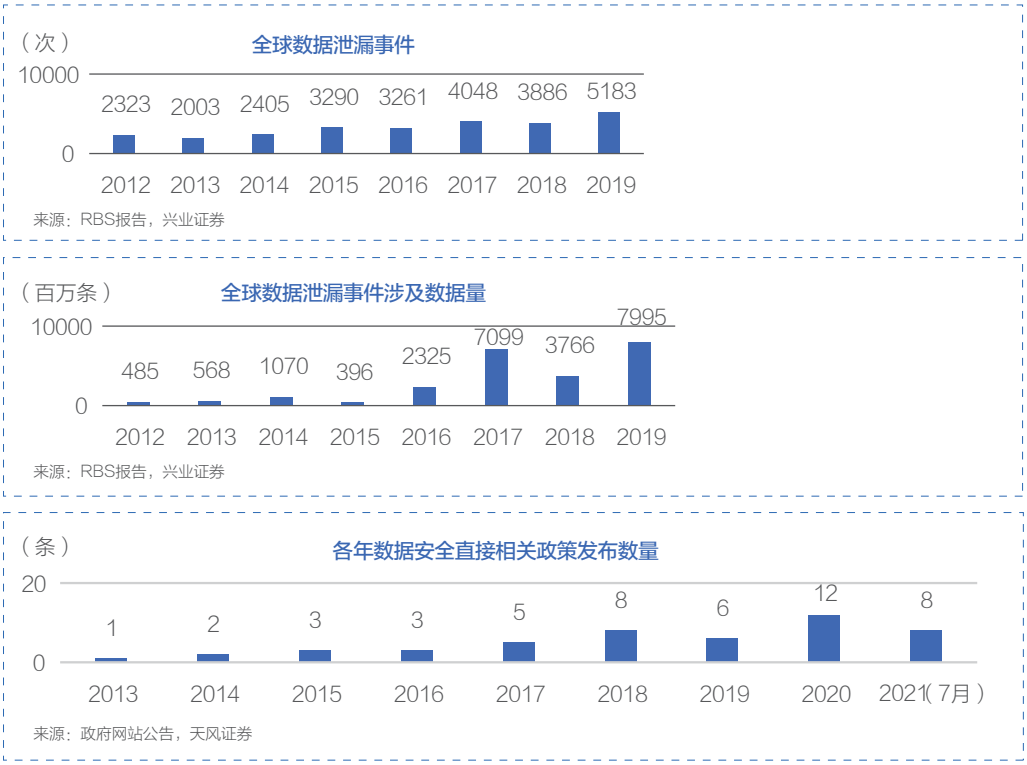
- 通过 AI 算法识别压力变化估算车辆用气的整体情况，建立智能决策模型，预测列车是否漏气
- 尽早对车辆进行检修，减少车辆运营能耗，防止事故的发生

（图4.4.2 -4）

隐私计算
铸就安全防护新体系>>

根据兴业证券统计，2019年全球有近6000件数据泄漏事件的发生。一方面数据安全事件不断上升，另一方面整个社会都在数字化，泄漏数据的影响也越来越大，数据安全事件驱动数据安全修改的法律法规与监管政策持续完善。

数据安全事件驱动数据安全政策持续加码



（图5-1）

以最新颁布的《中华人民共和国数据安全法》为例，其关键引导有：

保障数据安全。通过采取必要措施，确保数据处于有效保护和合法利用的状态，以及具备保障持续安全状态的能力。坚持总体国家安全观，建立健全数据安全治理体系，提高数据安全保障能力。

促进数据开发利用。国家支持数据开发利用和数据安全技术研究，鼓励数据开发利用和数据安全等领域的技术推广和商业创新，培育、发展数据开发利用和数据安全产品、产业体系。

维护国家利益。维护国家主权、安全和发展利益，国家对与维护国家安全和利益、履行国际义务相关的属于管制物项的数据依法实施出口管制。

保护个人/组织权益。国家保护个人、组织与数据有关的权益，鼓励数据依法合理有效利用，保障数据依法有序自由流动，促进以数据为关键要素的数字经济发展。

其中，公共数字资产开放对政府工作意义重大包括：

提升政务数字化能力。数据开放能极大的提升政务数字化的能力，提升公共服务的能力和质量。

提升执政透明度。公共数字资产开放，能极大的推动在数字化时代的政务透明度建设，提升互信。

赋能企业及公民。数字资产开放创新，有利于赋能传统行业企业数字化转型，提升传统行业的竞争力。

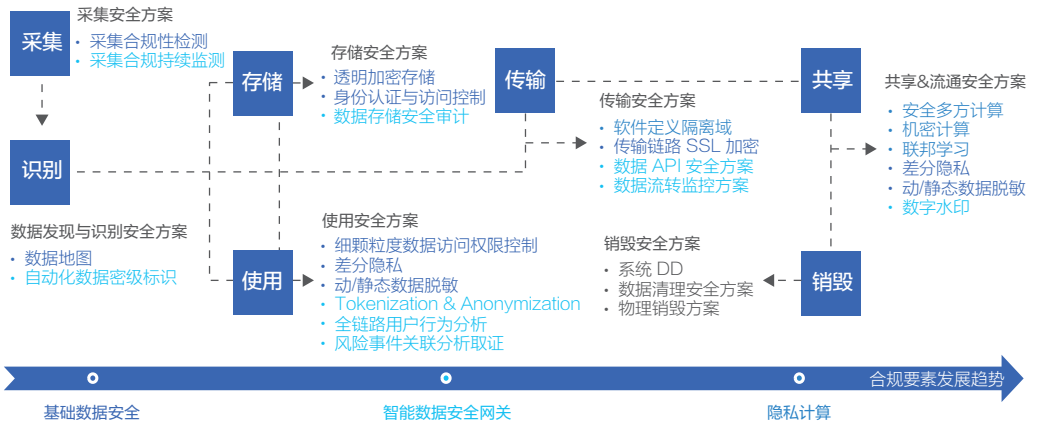
培育增长新动能。利用以数据驱动创新为特征的前沿技术，推动数字科技、人工智能、智能制造融合，培育增长新动能。

5.1 数据安全的落实规则

5.1.1 覆盖数据全生命周期的安全体系

数据全生命周期包括采集合规性检测和合规持续监测，数据发现与识别安全方案（数据地图和自动化数据密级标识）；存储安全方案（加密存储、身份认证与访问控制和数据存储安全审计）；数据使用安全方案（细颗粒度数据访问权限控制、差分隐私、动/静态数据脱敏、Tokenization & Anonymization、全链路用户行为分析和风险事件关联分析取证），传输安全方案（软件定义隔离域、传输链路SSL加密、数据API安全方案和数据流转监控方案）；共享&流通安全方案（安全多方计算、机密计算、联邦学习、差分隐私、动/静态数据脱敏和数字水印），以及销毁安全方案（系统 DD、数据清理安全方案和物理销毁方案）等。数据合规要素发展涵盖网络安全、信息安全以及隐私计算等各个方面。

数据生命安全周期的安全体系



（图5.1.1 -1）

5.1.2 隐私数据的安全保护闭环

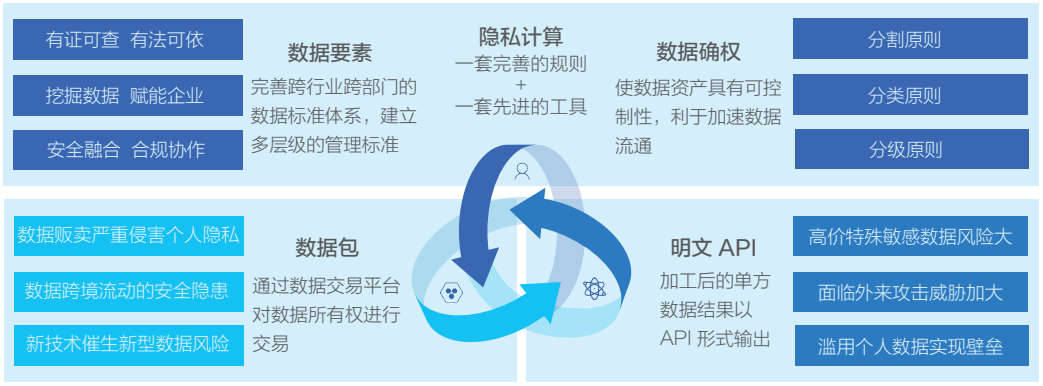
隐私数据的安全措施要形成闭环，包括资产集中管控、全面应用分析、敏感数据资产感知、敏感数据防护、泄露风险拦截和事件审计溯源等。做好安全保护闭环管理，可以避免流程环节中的安全漏洞，提高安全保护效率和质量，有利于数据安全机制的有效执行。

（图5.1.2 -1）

5.1.3 安全合规的数据流通形式

传统的数据流通，加工后的单方数据结果以数据包或明文API形式输出，存在高价特殊敏感数据风险大、面临外来攻击威胁加大以及滥用个人数据等问题，不能有效的保护个人隐私以及商业的机密。

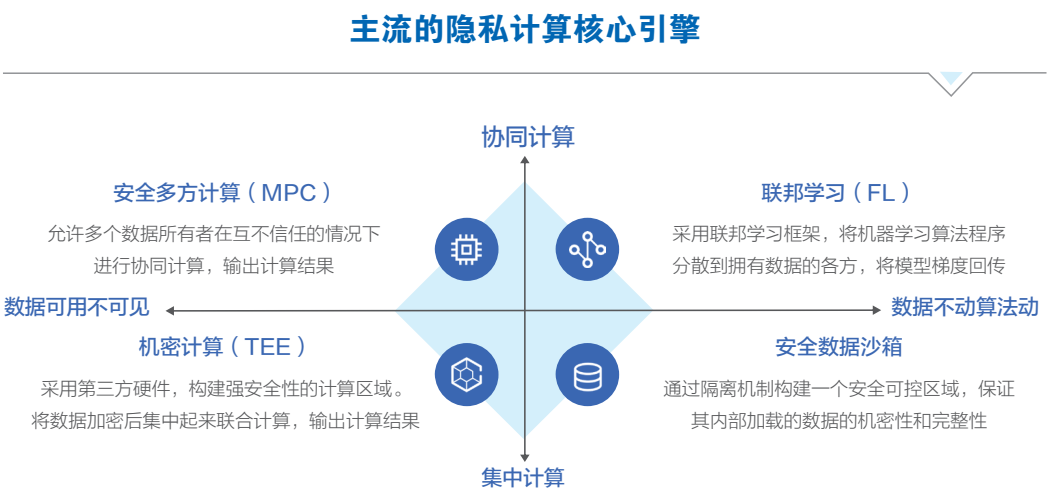
隐私计算是一套完善的规则和一套先进的工具，从数据要素角度完善跨行业跨部门的数据标准体系，建立多层级的管理标准；从数据确权角度使数据资产具有可控制性，利于加速数据流通，保护个人隐私和商业机密。



（图5.1.3 -1）

5.1.4 主流的隐私计算核心引擎

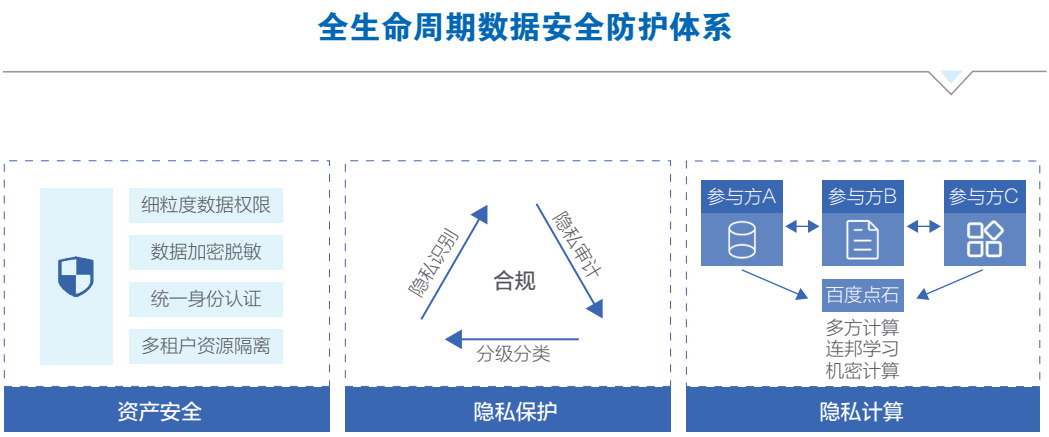
主流的隐私计算核心引擎可以分为四类：在数据可用不可见的基础上，安全多方计算（MPC）用了大量数学方法做密态计算，允许多个数据所有者在互不信任的情况下进行协同计算，输出计算结果。在数据可用不可见的集中计算领域，机密计算（TEE）是最流行的处理机密和隐私的技术手段，采用第三方硬件，构建强安全性的计算区域，将数据加密后集中起来联合计算，输出计算结果。在数据不动算法动方向，联邦学习（FL），采用联邦学习框架，将机器学习算法程序分散到拥有数据的各方，将模型梯度回传。还有数据不动算法动的集中计算的式，安全数据沙箱，通过软硬件结合，使用隔离机制构建一个安全可控区域，保证其内部加载数据的机密性和完整性同时，又不失可用性。



（图5.1.4 -1）

5.2 百度全生命周期数据安全防护体系

百度数据安全体系贯穿了大数据基础设施构建、数据价值挖掘的全过程，覆盖了数据全生命周期，从多个维度保护企业数据安全。



（图5.2 -1）

5.2.1 资产安全

在资产安全方面，百度数据安全体系提供细粒度数据权限、数据加密脱敏、统一身份认证、多租户资源隔离等技术，确保资产生命周期过程中的安全性，以及数据在企业内外部应用过程中的安全性。

5.2.2 隐私保护

在隐私保护方面，百度数据安全体系实现了事前安全隐患发现、事中敏感数据保护、事后精准溯源的安全保护闭环，为用户提供安全合规的数据应用能力。

5.2.3 隐私计算

在隐私计算方面，百度智能云基于百度内部数据安全及隐私保护业务场景以及合作伙伴业务实践，整合数据脱敏技术、安全多方计算技术（MPC）、机密计算（TEE）、联邦学习（FL）、安全数据沙箱等技术，沉淀形成的数据安全及隐私保护解决方案——百度点石隐私计算平台，实现“数据可用不可见”与“数据不动算法动”。

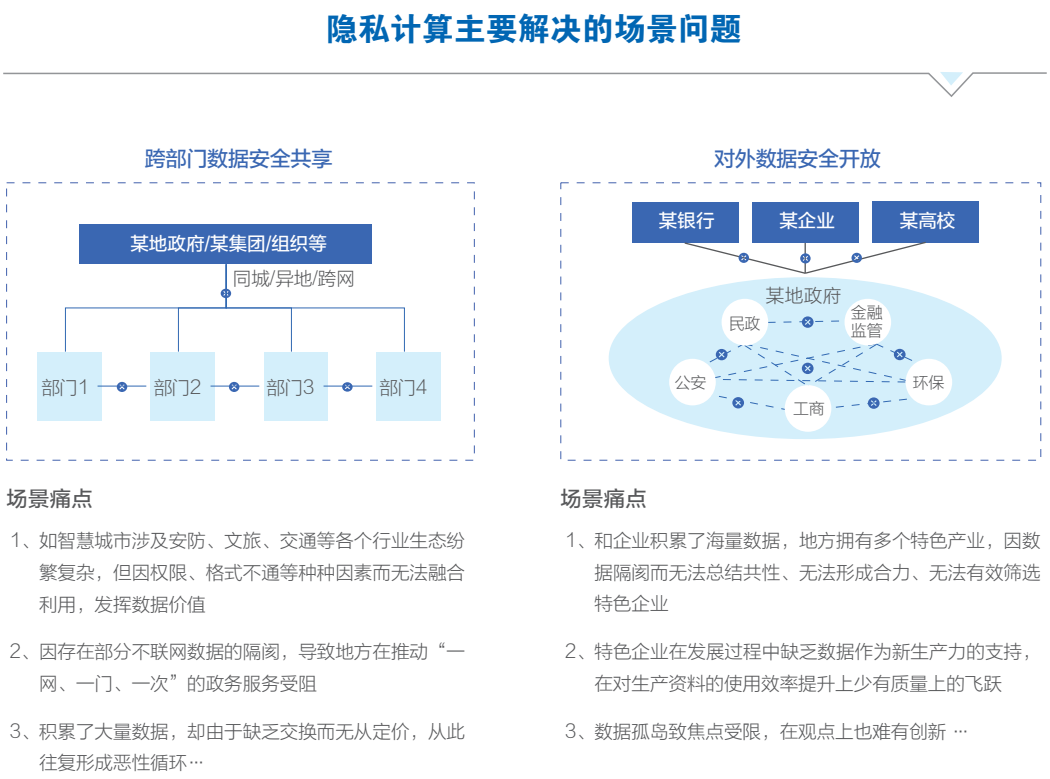
百度点石隐私计算平台架构

（图5.2.3 -1）

隐私计算主要解决的场景问题

智慧政务中跨部门数据安全共享：如智慧城市涉及安防、文旅、交通等各个行业生态纷繁复杂，存在不同的数据隔阂，导致地方在推动“一网、一门、一次”的政务服务受阻；积累了大量数据，却由于缺乏应用而无从定价，从此往复形成恶性循环。

政务数据赋能产业：在信息化浪潮中积累了海量数据，地方拥有多个特色产业，因数据隔阂而无法总结共性、无法形成合力、无法有效筛选特色企业；特色企业在发展过程中缺乏数据作为新生产力的支持，在对生产资料的使用效率提升上少有质量上的飞跃；数据孤岛导致治理认知受限，在观点上也难有创新。



（图5.2.3 -2）

基于场景的解决方案

百度数据安全及隐私保护解决方案整合了：安全数据沙箱、联邦学习、多方计算、机密计算四大隐私计算引擎，分别在广告营销、金融风控、智慧政务等多领域提供完善的解决方案，企业用户无需关注复杂的隐私计算技术，即可轻便享用世界领先的隐私计算技术带来的便利。

安全、合规是百度智能云服务客户的基础。目前，百度智能云共获取了40+项国家、国际机构认可的资质认证，包括SOC1 Type2、 SOC2 Type1 、SOC2 Type2、SOC3等多项SOC安全审计，以及MTCS最高安全评级等国内外安全权威机构认证。同时，百度智能云是国内首家通过ISO 27032、ISO 29151、ISO 27081、ISO 27017、BS 10012认证的云服务商。

安全合规是百度智能云服务客户的基础

百度智能云获取了40+项国家与国际机构认可的资质认证

（图5.2.3 -3）

5.3 百度点石成功实践

目前，百度点石数据安全及隐私保护方案已在政务、金融、医疗、电商、教育、媒体等多个领域成功落地。

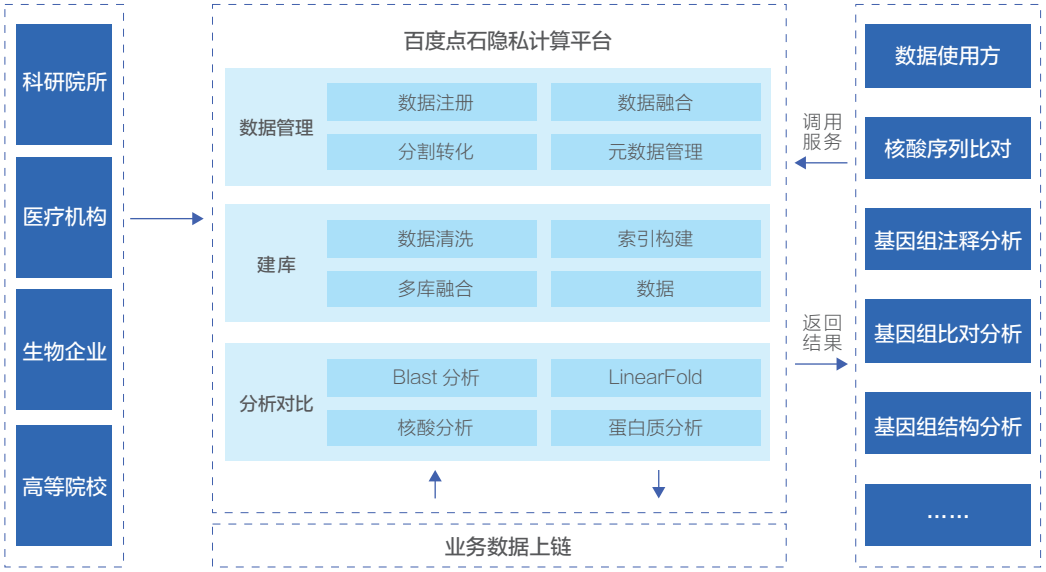
百度点石案例覆盖诸多行业

服务过的客户涵盖政府、金融、医疗、电商、教育、新闻体育等诸多领域

（图5.3 -1）

以某基因研究所为例，该研究所拥有大量中国公民基因信息，在医疗机构、生物制药等领域有重要价值，一旦基因数据泄漏将十分严峻。百度智能云针对该场景，量身打造一整套数据安全方案，全力保障科研院研究时、在医疗机构和高校使用数据时保证不会涉及个人隐私、敏感数据安全问题。

医疗案例：打破数据孤岛释放数据价值



（图5.3 -2）

企业数据资产构建与应用实践 >>

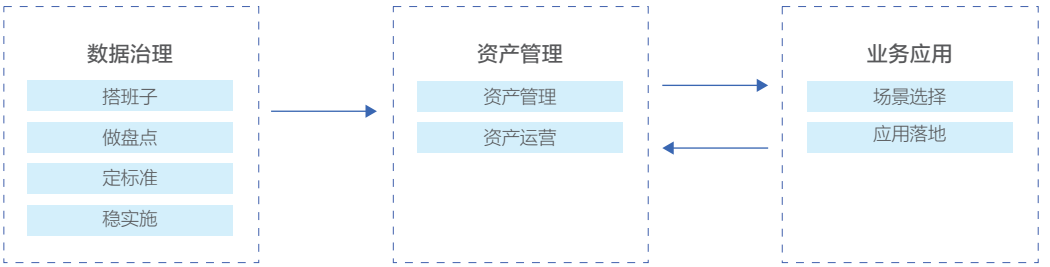
45

在信息化时代，企业构建的信息系统是面向具体业务场景的，传统的烟囱式建设模式，会带来数据孤岛和治理困难；数据生产要素自身又具有变化性、流通安全性，这些特性又让管理效能成为问题。从数据应用角度来看，行业和场景不同、企业发展阶段不同，数字资产应用选择点也会不一样。企业在大数据应用落地过程中会面临数据治理、数据孤岛和治理困难、资产管理效能低下，以及业务场景繁杂等问题。

百度从做搜索业务起，就开始运用大数据技术进行互联网海量数据的采集、加工、深度分析挖掘，匹配信息与用户需求。到近些年大力发展AI和智能应用，是进一步对更大规模、更多类型、更多源数据的治理、管理和应用过程。同时，百度智能云在工业、能源、金融、政务等重点行业也积累了丰富的数据资产建设经验。

数据资产建设与应用过程，可以抽象为三步：数据治理，形成资产；数据资产管理；数据反哺到业务中产生价值。

企业数据资产构建与应用过程



（图6 -1）

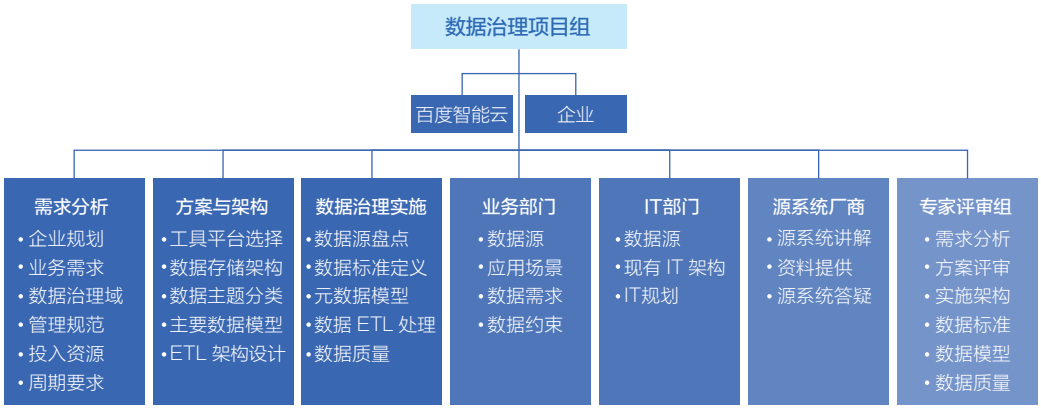
6.1 数据治理

数据治理，要做到稳健有序。在整个过程中，数据治理有四个关键点，即搭班子、做盘点、定标准和稳实施。

6.1.1 搭班子

数据治理实施过程是一个跨部门、跨角色协同的过程，所以搭班子十分重要。跟企业用户做数据治理时，首先要跟企业的业务部门、IT部门甚至企业在建系统厂商等建立专业的数据治理实施项目组，搭建稳健班子来确保数据治理有序性和高效性。

搭班子，有效协同



（图6.1.1 -1）

6.1.2 做盘点

做盘点，是数据治理的基础。除了盘点数据源、数据类型等，还需要盘点业务和IT架构。盘点业务，数据治理实施部门或者牵头部门往往既不产生数据，也不应用数据。因此需要跟业务部门协同，沟通业务场景痛点、数据需求，有助于确保实施过程中的流畅性，形成有效的数据结果和资产。盘点IT架构，数据治理要盘点数据源、数据类型、网络架构及服务器等资源配置，盘点清楚IT架构，对实施过程中的高效性非常重要，保证数据实施过程的高效性和有序性。

做盘点，覆盖全面

（图6.1.2 -1）

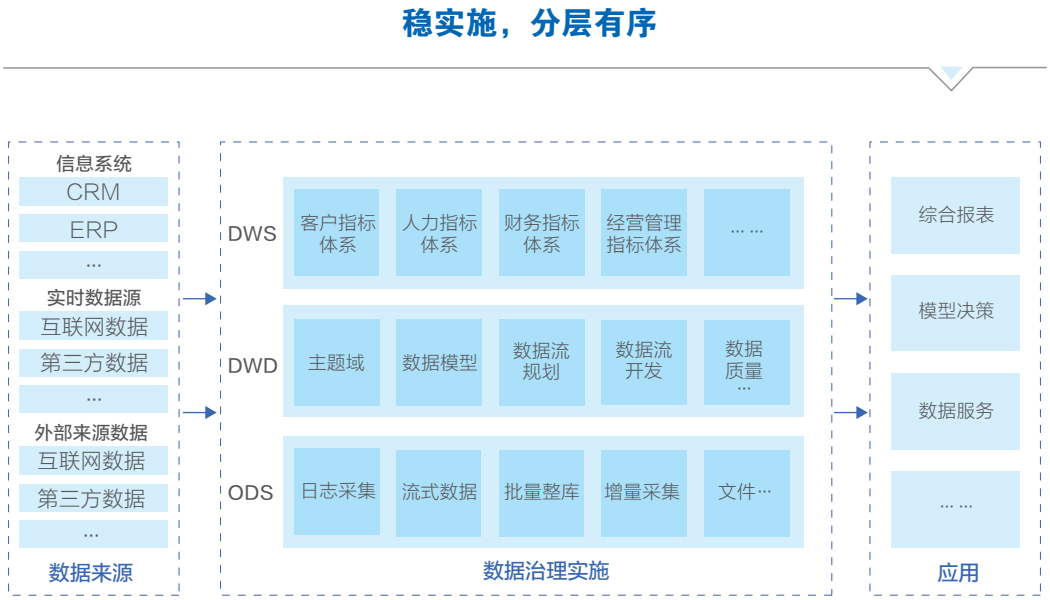
6.1.3 定标准

数据搭班子保证数据治理的过程有序性，做盘点为数据治理实施打下基础，定标准则保证数据治理结果的正确性，保证符合预期。定标准一般包括数据分类标准、数据建模标准、数据质量标准和数据应用标准。

（图6.1.3 -1）

6.1.4 稳实施

做好准备工作后，开始数据实施工作。首先要通过大数据平台对涉及数据源如信息系统的结构化数据、第三方数据等进行数据采集，形成ODS运营数据层；再从主题域、数据模型、数据流规划等维度定好模型标准，形成DWD数据层；最后通过分析业务的客户指标、人力和财务指标等，定义DWS数据模型，再基于数据开发平台进行数据实施，形成企业数据资产。



（图6.1.4 -1）

6.2 资产管理与运营

数据作为一种新生产要素和资产，需要持续管理和运营，以发挥最大业务应用价值。

6.2.1 资产管理

构建数据资产管理平台，可以提高企业数据管理和运营效率，让数据资产实现好管理、好找到、好理解和好应用。

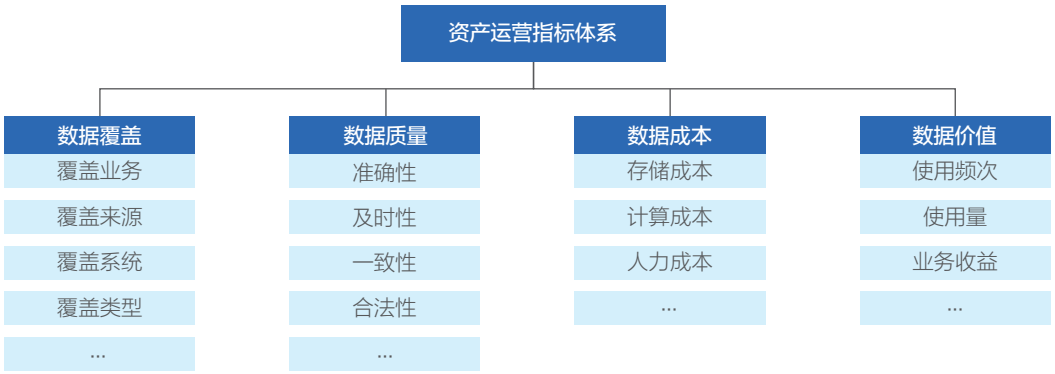
对于数据管理人员，可以通过资产目录管理企业数据资产的分类、目录和主题等，通过资产看板对数据资产形成更宏观的认知。数据资产管理平台提供的数据搜索和推荐功能，可以帮助业务方数据应用者更高效的找到数据。数据资产管理平台的数据标准和数据图谱功能，可以帮助数据应用者找到数据后，更好的理解数据来龙去脉和业务含义。数据资产管理平台还可以提供数据在线查询分析和数据自动生成标签的能力，让数据应用者在平台上找到并理解数据后，帮助他们高效应用数据。

（图6.2.1 -1）

6.2.2 资产运营

数据资产运营的目标主要分两类，一是通过运营驱动企业核心资产要素的建设完善，二是通过运营驱动企业核心要素发挥最大价值，提升ROI。围绕这两个目标，可以确定数据资产运营体系的关键指标，即数据覆盖、数据质量、数据成本以及数据价值。通过建设一系列指标体系，来提高数据资产运营的ROI。

资产运营，以指标体系辅助持续运营

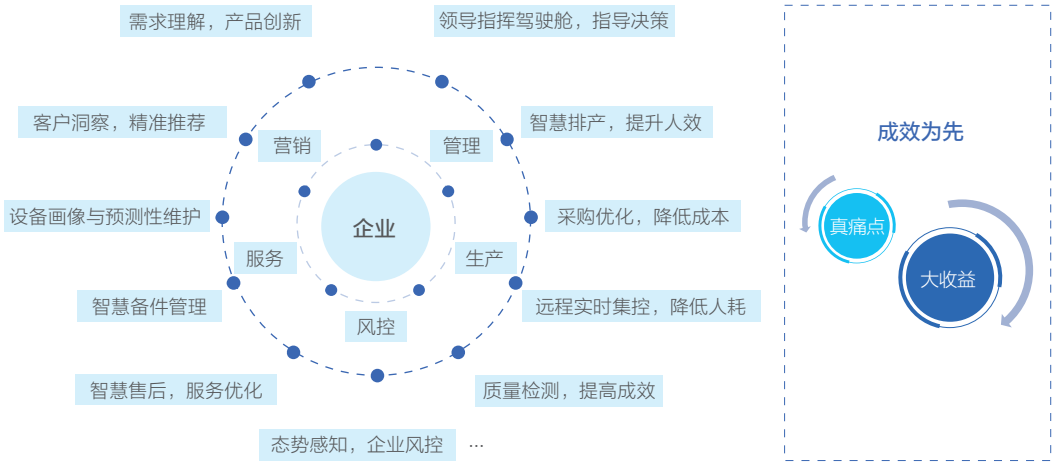


（图6.2.2 -1）

6.3 业务应用

行业有一句话“数据是资产也是负债”，数据不仅会产生价值也会带来存储和运营等成本，因此要成效为先，通过数据资产解决企业真实业务痛点，帮助企业获取最大化收益。一个企业的核心发展有几个方向，即管理、生产、营销、服务和风控，每个维度又包括具体应用场景。百度智能云在领导指挥驾驶舱、智能排产、结构优化、产品创新和预测性维护等具体应用场景方面，帮助不同行业的众多企业做了应用实施落地工作，积累了丰富的行业实践经验。

基于痛点和收益驱动场景选择



（图6.3 -1）

6.4 企业数据资产应用实践

6.4.1 徐工综合经营管理决策驾驶舱

徐工集团工程机械有限公司，是我国工程机械行业规模最大、产品品种与系列齐全、极具竞争力、影响力和国家战略地位的千亿级龙头企业。其核心痛点：一是部门繁杂，跨部门协作难度大；二是挖掘机售卖后维修及时性和满意度不理想；三是挖掘机单价高，很多客户是通过贷款购买，回款压力和难度大。

百度智能云为其量身打造的企业综合经营管理决策驾驶舱解决方案，实施效果和收益包括：

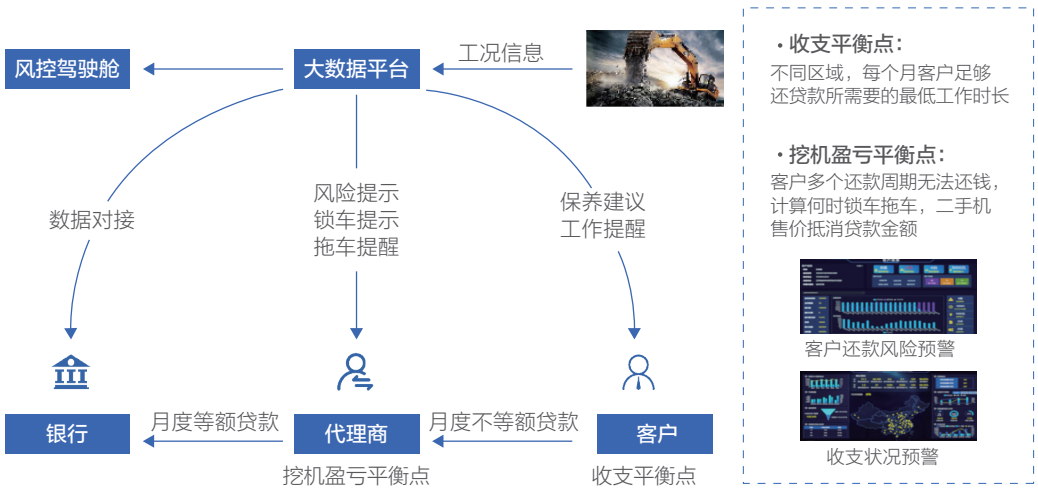
一是通过大数据平台建设，实现数据互通互联，打通产品线、部门藩篱，研、产、供、销、服一张图，管理者全局实时掌控，精准决策。

（图6.4.1-1）

二是通过把挖掘机的物联网日志+采购数据+维修数据打通，结合百度地图数据，让挖掘机售后做到就近服务，按资源配置地图分配服务，就近选择资源最丰富、最匹配的维修商，大幅提升维修效率。此外，通过物联网数据、CRM中部件更换数据等形成的专业知识库，对部件进行预测性维护，从设备类型、工作地区、海拔湿度、工况等维度预测部件磨损，维修人员还可通过挖掘机轨迹回溯，提前形成故障预判，大幅提升维修及时性和满意度。

（图6.4.1-2）

三是评估不同区域、每月客户足够还贷款所需要的最低工作时长，评估挖机盈亏平衡点，即如果客户多个还款周期无法还钱，计算何时锁车拖车，二手机售价刚好能还贷款。



（图6.4.1-3）

6.4.2 陕西广电媒体融合数据资产建设

陕西广电融媒体集团是整合原陕西广播电视台和陕西广播电视集团而成，是大型现代主流媒体集团，拥有新闻广播、陕广新闻等广播节目和陕西卫视等电视频道。

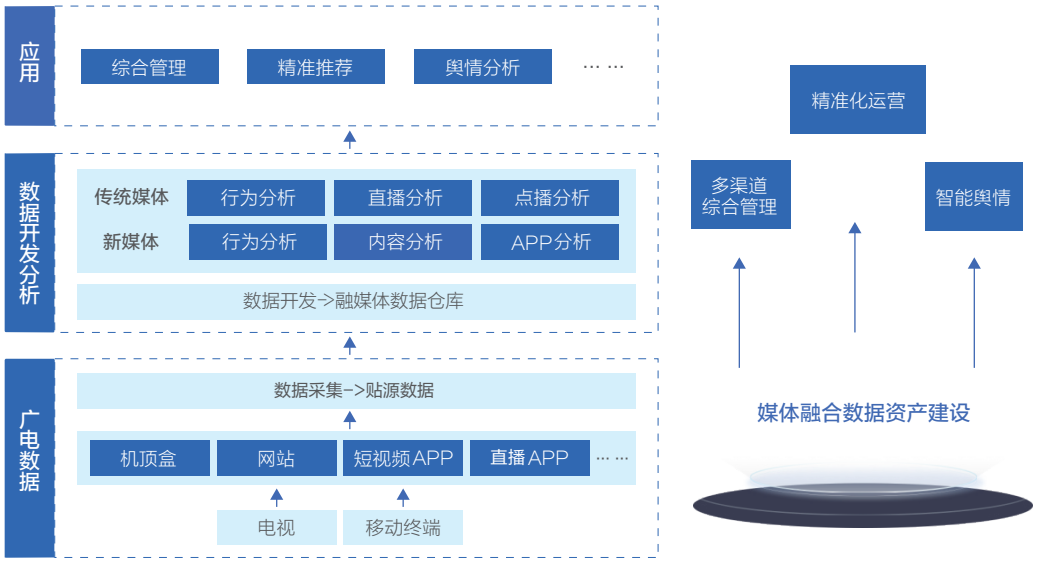
随着中共中央办公厅国务院办公厅发布的《关于加快推进媒体深度融合发展的意见》媒体融合进入新阶段，媒体融合和创新发展向纵深发展。陕西广电融媒体集团作为广电行业的融合创新发展的排头兵，为广电融合模式的探索技术创新还有一些项目孵化主要当参谋做创新。

从纸媒到PC互联网再到移动互联网，媒体积累的数据越来越多，渠道也越来越多，内容和形式也更加的多样和复杂。因此，媒体融合创新面临着 “数据缺乏治理和应用”，“多渠道客群运营复杂”，“内容和运营深度融合困难”等多重挑战。

（图6.4.2-1）

基于百度智能云云智一体的智能大数据平台，陕西广电建设了媒体融合数据资产平台，在数据资产建设的基础上，进一步进行数据分析和价值深度挖掘，实现了实时的多渠道用户综合管理、精准化用户运营，以及舆情智能感知。

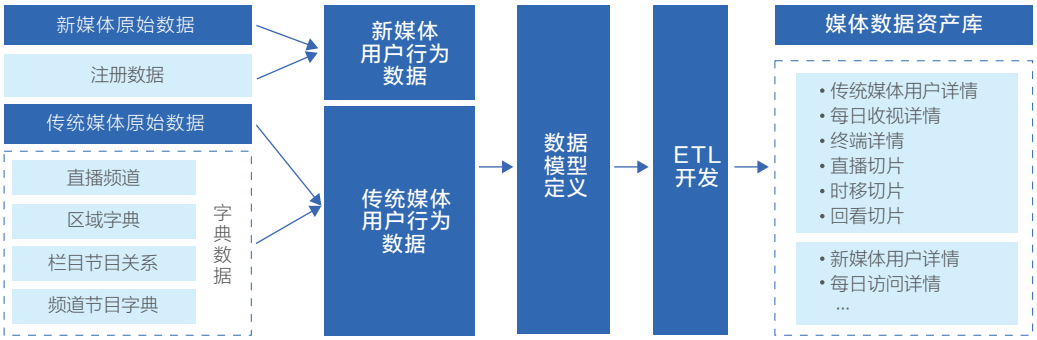
百度智能大数据助力陕西广电媒体融合创新



（图6.4.2 -2）

媒体融合数据资产建设。基于百度智能大数据平台完成新媒体原始数据、传统媒体原始数据、以及字典数据采集，形成ODS层数据资产，包括新媒体用户行为数据仓库、传统媒体用户行为数据仓库。接着，通过分析媒体融合业务应用场景对于数据的需求，定义了媒体数据资产数据模型，形成DWD和DWS媒体数据资产集市层，包括传统媒体、新媒体用户详情、直播回看等。百度智能大数据平台支持的数据开发处理形态非常丰富，有免代码拖拽式的可视化开发Studio、离线脚本开发、实时流开发、跨源联邦分析等，所以整个数据治理、数据资产建设的过程比较高效。

媒体融合数据资产建设



（图6.4.2 -3）

数据应用创新环节。基于数据分析驱动实现了多渠道用户的综合管理，根据管理和运营需求定义好运营指标体系和计算公式等，再通过百度智能大数据平台进行数据开发。百度智能大数据平台除了支持丰富的数据开发形态，也支持可视化作业编排调度，使用非常灵活、方便。管理运营指标开发计算好后，写入实时数据仓库，再配置成可视化图表或者大屏，管理或运营人员即可高效通过可视化的报表或者大屏进行多渠道综合管理。最后，实现了可视化综合管理一张图、多渠道多维度分析报表、实时趋势变化分析等可视化报表。

数据驱动多渠道综合管理

可视化综合运营管理决策

（图6.4.2 -4）

数据驱动精准化运营。精准化运营优化的维度一般包括营销增强、需求洞察、渠道优化、产品提升等，首先需要针对每个优化方向，定义出了运营数据指标体系，例如用户群的基本属性、用户偏好、用户潜在需求等，这些数据指标有的是事实标签，通过数据统计即可获取，有的是模型标签、预测标签，需要通过多数据源的交叉融合分析和算法开发进行预测，都可以在百度智能大数据平台上完成。基于丰富的运营指标体系，可通过地域、需求、偏好、关系等多维度运营指标数据，有效的实现用户圈生命周期精准运营，提高各类转化率。

数据驱动精准化运营

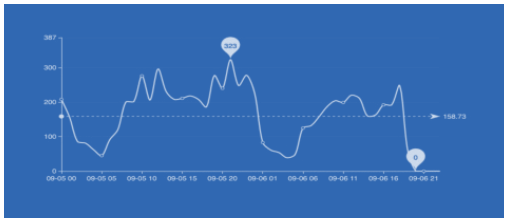
（图6.4.2 -5）

智能舆情分析。实时舆情监测，可以洞察网民群众对某事件所持有的观点态度分布，分析统计持续关注某事件的群体画像特征分布，感知网民群众日常表现的关注内容及话题分布等。同时基于大数据平台实施最终事件发展每一个关键点，刻画事实的脉络轨迹，还可以构建事件与事件的关系，包括事件间的因果关系和顺承关系，为媒体的深度洞察提供数据参考。

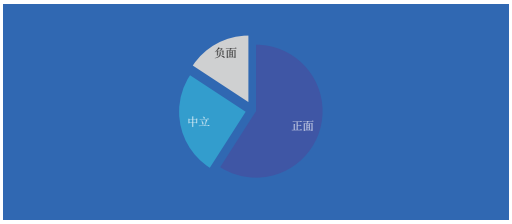
此外，还可以通过综合媒体传播声量走势，动态统计热点舆情事件的传播量，描绘声量曲线，掌握热点舆情的整体传播态势；通过对舆情文本深度分析，了解舆情事件的情感属性偏向，实时监测舆论走势。

数据驱动智能舆情

事件传播影响分析



趋势热度



情感分布

（图6.4.2-6）

百度从做搜索业务起，就开始运用大数据技术进行互联网海量数据的采集、加工，并深度分析挖掘，匹配信息与用户需求，让人们最平等便捷地获取信息、找到所求；再到大力发展AI和智能应用，是更进一步的对更大规模、更多类型、更多源的数据进行处理和应用，是以实际行动在践行“用科技让复杂的世界更简单”。

更多行业案例参考

智慧水务	泉州水务集团有限公司	构建大数据平台，实施数据治理、落地水泵精准调压、供水综合调度、管网漏损预警和定位、排水调度优化、污水处理监管等应用。
城市数据治理	沛县	基于数据湖治理开发分析工具，高效实现27个部门的数据整合治理和打通，改变了数据孤岛情况、实现数据开放、共享、互联互通。
城市大脑	智慧丽江	构建城市数据综合管理平台，打通数据壁垒、实现数据资源共享。落地城市综合指挥调度驾驶舱、智慧党建、智慧文旅、数字化城市治理等应用。
智能制造	XCMG 徐工集团	构建大数据平台，实施数据治理、落地领导综合指挥驾驶舱、客户画像、设备画像、故障预测、智慧售后等应用。
金融大数据	百信银行 AIBANK	构建托管金融大数据平台，服务业务数据处理和分析及智能化创新。
融媒体大数据	陕西广电网络 SHANXI BC&TV NETWORK	构建融媒体大数据平台和融媒体数据资源库，落地多渠道用户综合管理与分析、用户全生命周期精准运营、舆情舆论态势实时感知等应用。
气象大数据	重庆市气象局	构建气象数据综合分析挖掘平台，建设气象数据资源库，气象智能探测系统、智能预报系统、气象数据服务系统、智慧防灾预警等系统。
智慧交通	广州黄埔	通过“车城网”平台的打造，实现道路、智能网联汽车、城市市政基础设施等城市要素数字化，形成全区公共数字底座，为各种应用提供支撑。
智慧教育	北京大学 PEKING UNIVERSITY	支撑北大图书馆数字化转型，提供整体的数据平台和数据可视化分析方案，实现面向用户、面向各业务中心、面向管理决策等的综合数据分析应用，优化了布局、优化服务效能。
互娱分析	iQIYI 爱奇艺	提供全托管、弹性可扩展、高性价比的云上大数据分析服务，按需创建、智能弹性伸缩、现金的存算分离架构大大降低的TCO成本。
能源数据综合分析	中国石油	数据分析驱动广域网升级，落地链路画像、用户画像、应用画像、流量回溯分析、基础服务与情报联动等数据分析与应用。

关于本报告

本报告内容来源于百度智能云重磅打造的“云智技术论坛”系列活动第四期，主题为《云智一体，让大数据发挥大价值》。百度多位资深技术专家与行业伙伴共同探讨大数据发展新形势下，企业如何运用云智技术打造安全可靠的数据基础设施与智能挖掘平台，最大化发挥数字资产的核心价值。

《云智一体 让大数据发挥大价值》

精彩回顾

百度智能云
微信公众号

百度 AI
微信公众号