AISWare DataInfrastructure产品

亚信科技大数据基础平台产品白皮书

AISWare DataInfrastructure是亚信科技全力打造的大数据基础平台产品，大数据平台作为企业数字化驱动和转型能力的基石，为了更好的支持业务，必须要向轻量化，微服务化和支持业务能力灵活构建的方面发展。在企业整体IT环境上云的背景下，继续为企业数字化转型提供大数据存储，计算，分析，管控能力，协助客户打造企业级的高效的，安全的一体化的大数据基础平台。

声明

任何情况下，与本软件产品及其衍生产品、以及与之相关的全部文件（包括本文件及其任何附件中的全部信息）相关的全部知识产权（包括但不限于著作权、商标和专利）以及技术秘密皆属于亚信科技（中国）有限公司(“亚信”)。

本文件中的信息是保密的，且仅供用户指定的接收人内部使用。未经亚信事先书面同意本文件的任何用户不得对本软件产品和本文件中的信息向任何第三方（包括但不限于用户指定接收人以外的管理人员、员工和关联公司）进行开发、升级、编译、反向编译、集成、销售、披露、出借、许可、转让、出售分发、传播或进行与本软件产品和本文件相关的任何其他处置，也不得使该等第三方以任何形式使用本软件产品和本文件中的信息。

未经亚信事先书面允许，不得为任何目的、以任何形式或任何方式对本文件进行复制、修改或分发。本文件的任何用户不得更改、移除或损害本文件所使用的任何商标。

本文件按“原样”提供，就本文件的正确性、准确性、可靠性或其他方面，亚信并不保证本文件的使用或使用后果。本文件中的全部信息皆可能在没有任何通知的情形下被进一步修改，亚信对本文件中可能出现的任何错误或不准确之处不承担任何责任。

在任何情况下，亚信均不对任何因使用本软件产品和本文件中的信息而引起的任何直接损失、间接损失、附带损失、特别损失或惩罚性损害赔偿（包括但不限于获得替代商品或服务、丧失使用权、数据或利润、业务中断），责任或侵权（包括过失或其他侵权）承担任何责任，即使亚信事先获知上述损失可能发生。

亚信产品可能加载第三方软件。详情请见第三方软件文件中的版权声明。

亚信科技控股有限公司（ 股票代码： 01675.HK ）

亚信科技创立于1993年，依托产品、服务、运营和集成能力，为电信运营商及其它大型企业客户提供业务转型及数字化的软件产品及相关服务，致力于成为大型企业数字化转型的使能者。

根据弗若斯特沙利文的资料，我们是中国电信行业最大的电信软件产品及相关服务供应商，按2017年收益 计，我们的市场份额为25.3%。根据同一资料来源，我们也是中国电信行业最大的BSS软件产品及相关服务供应 商，按2017年收益计，我们的市场份额为50.0%。我们是中国第一代电信软件的供应商，从20世纪90年代开始与 中国移动、中国联通和中国电信长期合作，支撑全国超过十亿用户。与电信运营商的长期合作关系让我们对电信 运营商的IT及网络环境以及业务运营需求有了深度理解，使我们能够开发出拥有500多种任务关键型电信级软件 的丰富的产品组合（软件产品主要面向电信运营商，对其业务运营至关重要），包括客户关系管理、计费账务、大 数据、物联网及网络智能化产品。截至2018年12月31日，我们有214家电信运营商客户，包括中国移动、中国联通 和中国电信的总部、省级公司、地市级公司、专业化公司和合营企业。

我们也正在积极拓展在中国非电信企业软件产品及相关服务市场的市场份额。凭借我们在电信软件产品及相关服务市场丰富的行业知识及专长及稳固的领导地位以及全方位、高度专业化的电信级产品图谱，我们相信我 们也已经就解决各类企业，尤其是大型企业在业务转型与数字化方面与电信运营商相类似的、最为根本的需求占 据了有利地位。截至2018年12月31日，我们有38家广电、邮政及金融、电网、汽车等行业的大型企业客户。通过资 源、管理、专业知识及技术专长的共享，我们能够同时服务电信和非电信企业市场，凭借协同效应赢取新业务并 保持竞争优势。

部分企业荣誉资质

ISO 9001质量管理体系认证

ISO 20000IT服务管理体系认证

信息系统集成及服务资质（ 一级 ）

CMMI 5级（ 能力成熟度模型集成5级 ）认证

国家规划布局内重点软件企业

2018年中国软件业务收入前百家企业前20强

2018年中国电子信息行业社会贡献500强

2018年中国电子信息研发创新能力50强企业

目录

一. 摘要 4
二. 缩略语与术语解释 5
三. 产品概述 6
3.1 趋势与挑战 6
3.2 产品定位 7
四. 技术介绍 8
4.1 亚信科技XX产品整体架构 8
4.2 产品功能架构 8
4.3 关键技术能力 9
4.3.1 XX关键能力1 9
4.3.2 XX关键能力2 9
五. 功能介绍 10
5.1 基础功能 10
5.2 特色功能 10
5.2.1 功能模块1 10
5.2.2 功能模块2 10
六. 场景应用方案 11
6.1 场景应用方案1 11
6.2 场景应用方案2 11
七. 带给客户的价值 12
八. 产品优势 13
九. 联系我们 14

返回目录

一. 摘要

从宏观形势来看，大数据在各行各业发挥了越来越重要的作用，已经逐步上升到国家战略层面。2020年10月29日，中国共产党第十九届中央委员会第五次全体会议审议通过了《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》。明确提出了“十四五”时期经济社会发展指导方针，这为做好未来五年经济社会发展工作指明了方向、提供了可遵循的原则。

数字化转型、智能化升级带来的需求，企业内部有降本增效的需求，企业向原生数字化企业发展，要实现业务数据化必须要规范建设大数据平台。商业生态环境无时无刻不在变化，企业也需要不断调整、扩展业务边界，加强生态合作，为数据集成、查询、实时分析等都带来需求。科技赋能5G，数字孪生，大数据，云计算，人工智能等，加速科技突破。大数据技术的不断突破创新，主要来源于厂商的主动性突破创新，例如湖仓一体、跨域跨源交互式查询等。

近年来大数据技术和应用迅猛发展，技术的不断演进，通过横向扩展，分布式集群部署方式比传统集中式架构性能更优，在数据平台架构云化重构、实时应用支撑、能力开放、智能运维等方面发挥了重要作用，为企业的大数据中心从大数据存储、计算、PaaS能力、运维、开放都有了飞跃式的发展。在此国家政策扶持、技术发展趋势良好的背景下，亚信紧跟开源大数据技术的发展，自研开发的大数据基础平台为电信行业，其它大企业提供完整的大数据平台解决方案，助力企业在大数据中心建设，发挥大数据价值方面保驾护航。

亚信AISWare DataInfrastructure是集大数据采集、数据转换、数据计算、数据存储、资源管控、运维一体化完整产品集和解决方案。

本白皮书将从产品概述、技术架构，主要功能、客户价值、产品优势等几个方面阐述亚信AISWare DataInfrastructure产品。

返回目录

二. 缩略语与术语解释

缩略语或术语	英文全称	解释
OSB	Open Service Broker	纳管组件的接入接口标准
DIF	DataInfrastructure	大数据基础设施平台
DP	Data Platform	大数据基础平台
CM	Cluster Manager	集群管控平台
CI	Cluster Insight	集群洞察平台
DLS	Data Lake Storage	数据湖存储
RCE	Real Compute Engine	实时计算引擎
SE	Search Engine	关联检索引擎
HQE	High Query Engine	高速查询引擎
GA	Graph Analyse Engine	图分析引擎

返回目录

三. 产品概述

AISWare DataInfrastructure帮助企业从存储，计算、分析，处理，管控、运维提供一体化的大数据基础平台，立足于开源，提供专业的大数据服务，运维，咨询，协助客户打造企业级的高效安全的大数据平台，并为客户提供大数据平台智能化的集群管控运维、集群洞察工具，提供可视化、实时的分布式流数据开发处理平台。

3.1. 趋势与挑战

国家大数据发展战略及政策为大数据发展保驾护航，大数据政策开始向各大行业和各细分应用领域延伸扩展，行业应用成为关注重点。受宏观政策环境，技术进步与数字应用普及渗透等多种因素影响。新基建政策提出加快第五代移动通信、工业互联网、大数据中心等建设。完善宏观经济治理政策指出提升大数据等现代技术手段辅助治理能力。发展战略性新兴产业政策明确推动互联网、大数据、人工智能等同各产业深度融合。

国家加强了对科技基础平台建设的支持，并且随着新型技术的发展，大数据平台迎来新的机遇；同时平台从百家争鸣进入到加剧竞争的阶段，市场对于产品化程度，技术深度，专业化都提出了新的要求。开源社区走向封闭：Cloudera和Hortonworks合并，Hadoop开源社区活跃度下降。Hadoop3.x新特性质量，稳定性下降，从开源到商用之间的鸿沟不断增大。

在技术层面，以开源为主导、自主可控要求提升，大数据平台建设从粗放走向精细，从小规模走向大规模，从少量使用走向多租户。对技术上的挑战主要体现在新技术特性的理解推广和使用，大数据平台规模上的增长，带来运维和性能调优的挑战。国产化，国产生态深度融合的要求逐步提升，带来对核心技术自主可控的强要求。

3.2. 产品定位

AISWare DataInfrastructure是亚信推出的大数据基础平台，作为企业数字化驱动和转型能力的基石，为了更好的支持业务，必须要向轻量化，微服务化和支持业务能力灵活构建的方面发展。在企业整体IT环境上云的背景下，继续为企业数字化转型提供大数据存储，计算，分析，管控能力，协助客户打造企业级的高效的，安全的一体化的大数据基础平台。

返回目录

四. 技术介绍

4.1. 亚信科技AISware BigData产品整体架构

亚信大数据域产品集包括：

AISWare DataDiscovery（AISW D2D）数据探索分析平台
AISWare Knowledge Graph（AISW KG）知识图谱工具
AISWare DataOS（AISW DataOS）数据中台操作系统
AISWare DataInfrastructure（AISW DIF）基础大数据平台
AISWare BigData产品体系中产品间关系及AISWare DataDiscovery位置如图1所示。

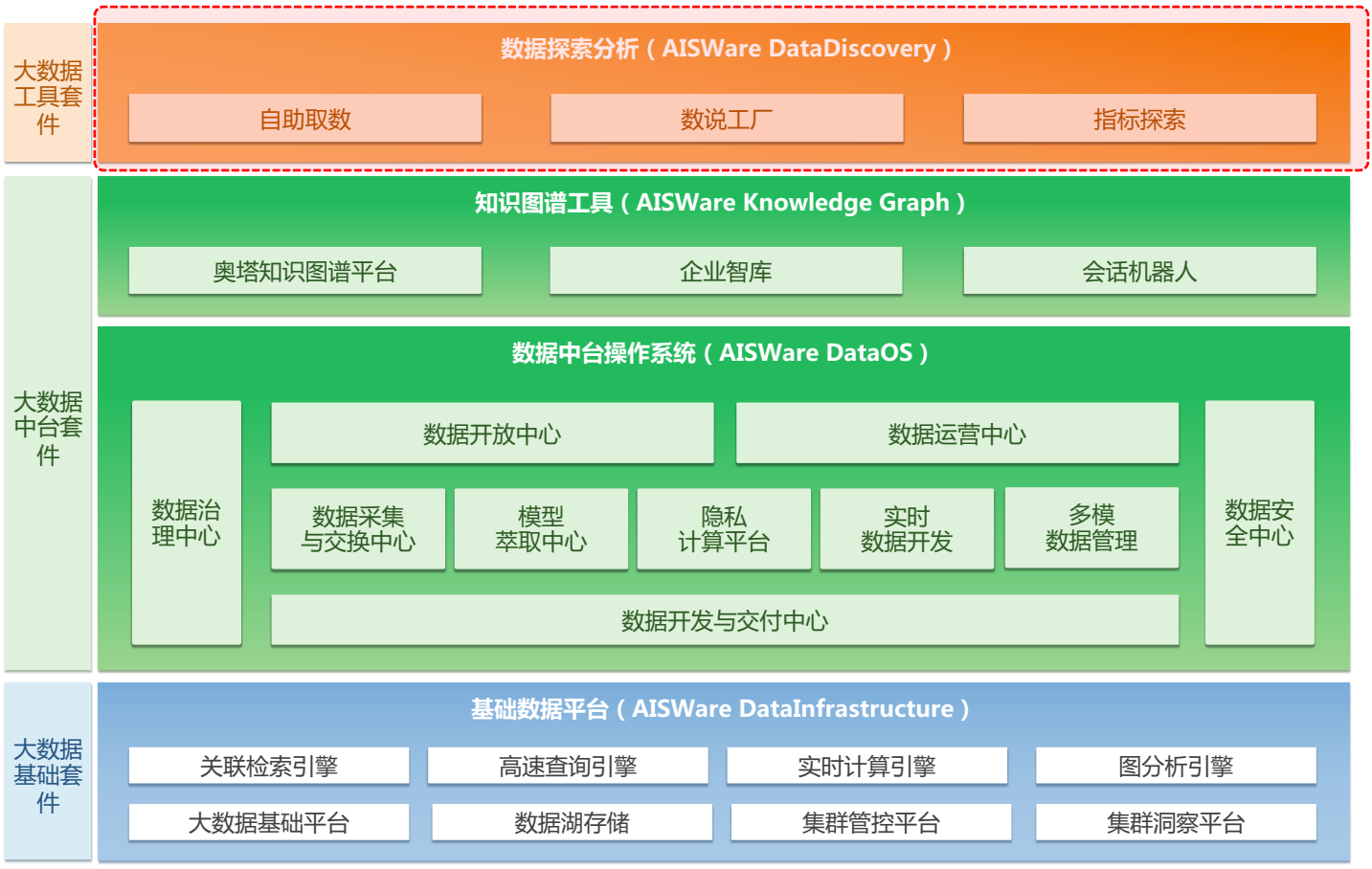


图1 亚信大数据域产品集总体架构

返回目录

4.2.亚信科技AISWare DataInfrastructure产品功能架构

亚信AISWare DataInfrastructurej基础数据平台整体技术架构如图所示，其中包括大数据基础平台、集群洞察、集群管控、数据湖存储、实时计算引擎、高速查询引擎、关联检索引擎、图分析引擎几部分。

AISW DIF-DP （大数据基础平台）是亚信推出的商业版本大数据平台，实现为租户提供统一的资源、运维管理及多种常用的容器化组件，并根据客户需求提供完整的计算、存储及组件使用环境，以支撑客户上层业务运营需求。

AISW DIF-CI （集群洞察）是大数据集群资源、性能、安全的深度洞察和智能规划，保障大数据集群的合理部署和不断优化，达到充分利用资源的目的。

AISW DIF-CM （集群管控)以多租户管理核心，面向企业实现大数据集群资源管控，实现大数据平台的租户能力开放管理能力。

AISW DIF-DLS （数据湖存储）实现结构化数据、半结构化数据及非结构化数据的统一存储，简化数据存储难度，提升对所有类型数据查询及分析的能力。

AISW DIF-RCE （实时计算引擎）实现多数据源的实时融合计算，提供实时业务所需要的计算场景。提升对于实时流数据的ETL、字段增强实时处理能力。

AISW DIF-SE （关联检索引擎）通过实现端到端的数据装载、构建索引、条件检索等处理过程，为业务系统提供快速查询能力，并针对非结构化数据提供存储检索和内容检索能力。

AISW DIF-HQE（高速查询引擎）应对大数据平台跨多种数据源的交互查询要求，屏蔽底层多类型跨数据源查询的复杂性，实现按租户权限为客户提供标准SQL的交互式高效查询场景。

AISW DIF-GA（图分析引擎）构建企业级图数据库，提供基于大规模图数据的图计算、图查询、图分析、图展现等能力。

返回目录

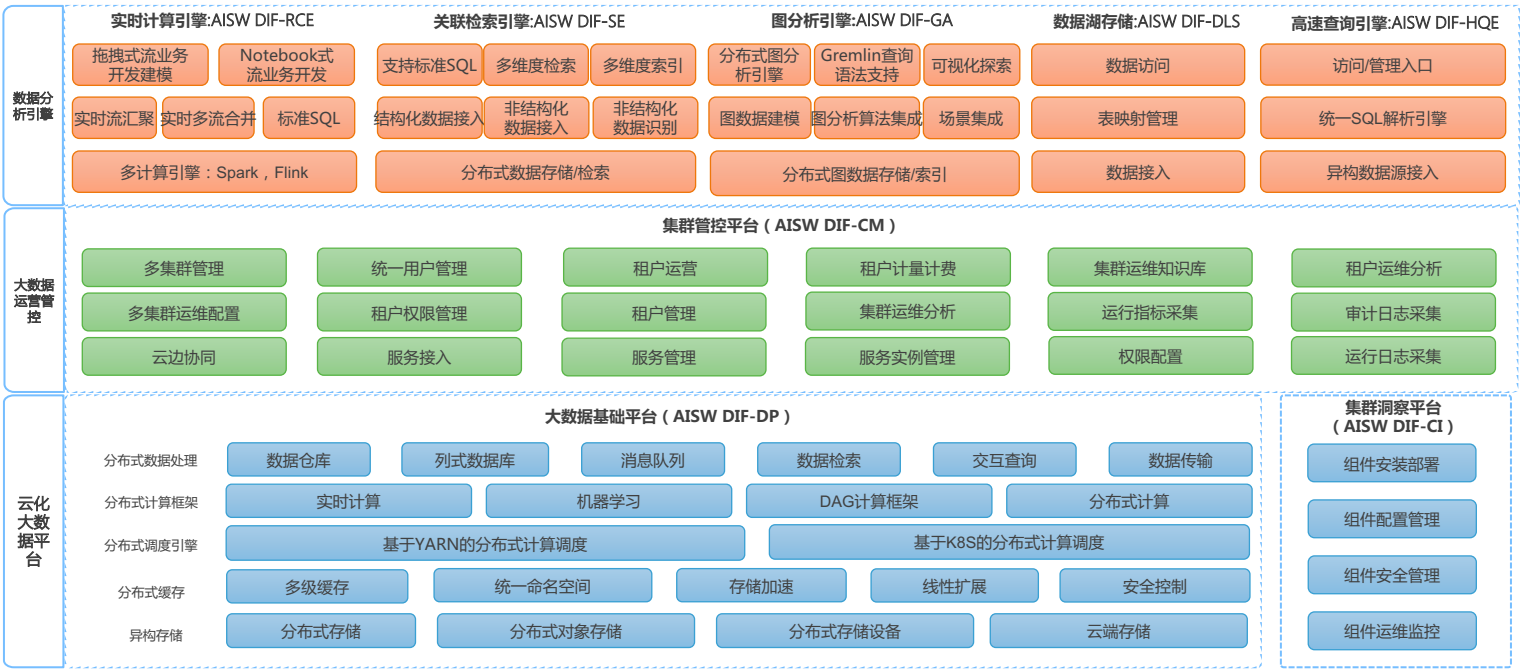


图2　AISWare DataInfrastructure产品架构

返回目录

4.2.1. AISW DIF-DP大数据基础平台产品架构

大数据基础平台提供一套完整的基于分布式文件系统的海量数据采集、存储、计算处理及运维的综合基础平台，采用分布式文件系统、列存储或混合存储、压缩、延迟加载等技术，只需要较为廉价的硬件设备投入即可提供对海量数据的存储能力。

采用分布式调度和资源管理技术，保证分布式并行运算的安全、高效和可靠；通过列存储引擎，提供key /value数据的实时存取，满足实时应用对大数据的读、写能力；采用交互式类SQL语句完成分析查询功能，提供快速、精准的数据多维分析功能，免去程序开发的复杂性，降低开发的难度，满足离线分析型应用对大数据的处理要求。提供丰富的组件包括分布式文件系统HDFS、资源管理与调度YARN、安全组件及能力Ranger、NoSQL数据库Hbase、数据加载处理Sqoop,Flume, Kafka、服务管理与YARN集成Slider、数据仓库Hive等。

用户可以快速的搭建起自身的企业级大数据基础平台，并在其上开展自身的数据分析业务。大数据基础平台作为开放的大数据基础平台产品，为用户提供了强大而丰富的平台能力，如数据存储能力、数据查询能力、数据计算和多维分析能力、索引分析能力、机器学习能力、图计算能力、资源管理调度能力、作业管控能力、平台运维管理能力。

图3　AISW DIF-DP产品架构

返回目录

4.2.1.1. 数据存储

HDFS（Hadoop Distributed File System）是Hadoop的分布式文件系统。HDFS集群主要由NameNode（管理者）和多个DataNode（工作者）组成。NameNode用来管理元数据；DataNode用来存储真实数据。用户可以通过NameNode与文件元数据建立联系或修改文件，并且通过DataNode直接访问实际文件的内容。

支持Federation ，在集群中将会有多个 namenode。这些 namenode 之间是联合的，也就是说，他们之间相互独立且不需要互相协调，各自分工，管理自己的区域。增加多备用NameNode支持。以此增强hadoop hdfs的高可用性，降低单个备用namenode瘫痪带来的集群管理风险。引入了纠删码技术（Erasure Coding），与三副本策略相比，提高50%以上的存储利用率。

4.2.1.2. 资源管理

在Hadoop中每个应用程序被表示成一个作业，每个作业又被分成多个任务。JobTracker是一个后台服务进程，启动之后，会一直监听并接收来自各个TaskTracker发送的心跳信息，包括资源使用情况和任务运行情况等信息。它的主要功能是：作业管理、状态监控和任务调度等。TaskTracker是JobTracker和Task之间的桥梁：一方面，从JobTracker接收并执行各种命令：运行任务、提交任务、杀死任务等；另一方面，将本地节点上各个任务的状态通过心跳周期性汇报给JobTracker。TaskTracker的主要功能是：汇报心跳和执行命令等。

Yarn的主要思想是把jobtracker的任务分为两个基本的功能：一个是资源管理，一个是任务监控，这两个任务分别用不同的进程来运行。在Yarn中，有一个全局的资源管理器（ResourceManager）和每个应用程序的应用程序管理器（ApplicationMaster）。ResourceManager和每个节点（NodeManager）组成了处理数据的框架，ResourceManager是整个系统资源的最终决策者。每个应用程序的ApplicationMaster是框架具体的Lib，它的任务是从ResourceManager出获得资源，并在NodeManager上执行和监控任务。支持开源新版本YARN的优化或新增功能。

返回目录

4.2.1.3. 批量处理

Hive是建立在 Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 定义了简单的类 SQL 查询语言，称为 HQL，它允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉 MapReduce 开发者的开发自定义的 mapper 和 reducer 来处理内建的 mapper 和 reducer 无法完成的复杂的分析工作。

Spark是一个大数据分布式编程框架，它使用函数式编程范式扩展了MapReduce模型以支持更多计算类型——不仅实现了MapReduce的算子map 函数和reduce函数及计算模型，还提供更为丰富的算子，如filter、join、groupByKey等；可以涵盖广泛的工作流。Spark使用内存缓存来提升性能，因此进行交互式分析也足够快速(就如同使用Python解释器，与集群进行交互一样)。缓存同时提升了迭代算法的性能，这使得Spark非常适合数据理论任务，特别是机器学习。

4.2.1.4. 实时处理

SparkStreaming是Spark核心API的一个扩展，可以实现高吞吐量的，具备容错机制的实时流数据处理。Spark Streaming将接收到的实时流数据，按照一定时间间隔，对数据进行拆分，交给Spark Engine引擎，最终得到一批批的结果。

Apache Flink是一个框架和分布式处理引擎，用于对无界和有界数据流进行有状态计算。Flink设计为在所有常见的集群环境中运行，以内存速度和任何规模执行计算。

Kafka是一个低延迟高吞吐的分布式消息队列，适用于离线和在线消息消费，用于低延迟地收集和发送大量的事件和日志数据 。Kafka通过副本来实现消息的可靠存储，同时消息间通过Ack来确认消息的落地，避免单机故障造成服务中断。同时副本也可以增加扇出带宽，支持更多的下游消费者订阅。

返回目录

4.2.1.5 列式数据库

HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PC Server上搭建起大规模结构化存储集群。在Hadoop生态系统中，Hadoop HDFS为HBase提供了高可靠性的底层存储支持，Hadoop MapReduce为HBase提供了高性能的计算能力，Zookeeper为HBase提供了稳定服务和failover机制。

4.2.1.6 运维管理

基于Apache Ambari开发的大数据集群管控工具实现对Hadoop平台组件的可视化一键部署，实现集群配置的统一管理与维护，实现对Hadoop平台告警及通知。

实现节点监控，监控集群中的每一个节点的状态的监控以及节点之间的通信状态的监控。单个节点的监控包括节点内存队列的大小，进程的活动状态，节点的IO,CPU和内存的情况，需要在页面中查看到。节点之间的监控是指集群中如果发现出现故障的节点，及时的进行主备切换或者剔除集群。

返回目录

4.2.2. AISW DIF-CI集群洞察产品架构

集群性能洞察：利用运维专家在多个生产Hadoop集群环境的性能评判经验，沉淀出多种性能算法模型，对集群的多方面关键指标进行计算、判断和展现.

集群负载分析：计算资源、存储资源、负载情况的采集及分析，可以回溯自定义时间段的资源使用情况、指导集群、租户的资源分配、合理有效配置集群资源.

集群安全洞察：通过对权限、数据操作，登录操作的安全审计，寻找安全漏洞并告警

集群运维工具:提供日志中心、运维知识库、和租户分权限的运维视图.为集群管理员、租户管理员提供有效的运维工具。

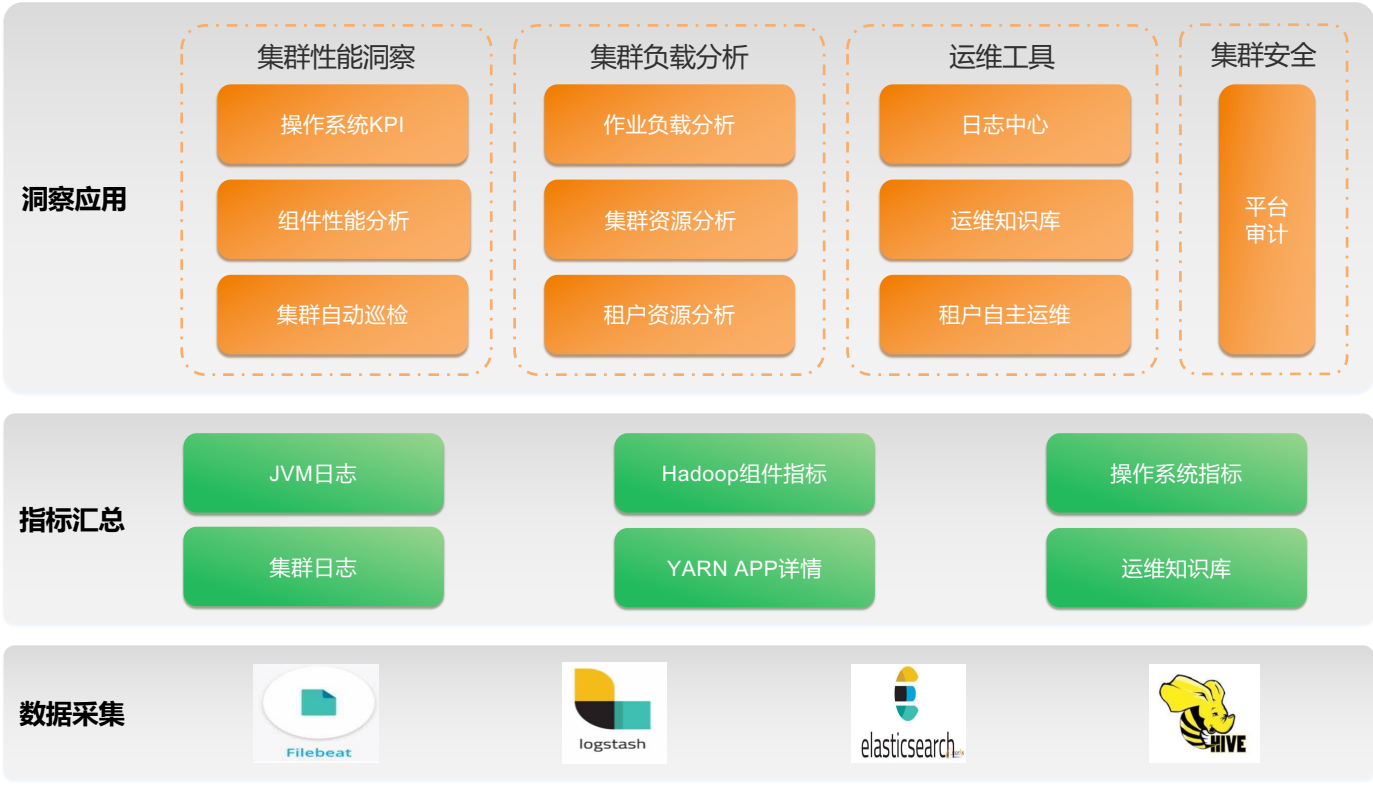


图4　AISW DIF-CI产品架构

返回目录

4.2.2.1. 集群性能洞察

对集群主机、网络、HDFS、Yarn、ZK、Hive、Hbase、Spark、等Hadoop服务的关键指标进行自动检查 ,并生成集群巡检报告.降低了运维知识门槛,对集群的健康度和性能指标进行快速概览。

通过对HDFS image Metadata的分析，对小文件的数量和占比进行统计. 可以辅助判定集群NameNode的效能, 分析出小文件的归属用户, 通知用户整改数据导入、计算的代码. 采集数据节点日志,将读写速度慢的节点统计出来, 展示这些慢速节点的IP和读写速度。

4.2.2.2. 集群负载分析

提供可视化的运维界面，展示集群、队列资源的利用率、消耗资源最多TOP10的作业，作业耗时的分布图、作业完成情况,等待情况等；该功能也为整个集群的资源评估提供依据。

4.2.2.3. 集群安全审计

审计功能可以保证Hadoop中用户数据安全，检测非法入侵和违反安全规则的行为，实现基于策略的实时检测和预警，实现基于用户行为模式的异常数据行为检测。

4.2.2.4. 集群运维工具

提供日志中心、运维知识库、和租户分权限的运维视图.为集群管理员、租户管理员提供有效的运维工具。

返回目录

4.2.3. AISW DIF-CM集群管控产品架构

集群管控运营在架构上主要分为接入层和功能层两层，其中功能层主要是提供管理、运营及分析能力，接入层主要实现大数据组件的统一接入。

功能层包括多主管理、资源分配、细粒度权限管理、运营分析及服务接入管理6个主要功能，多租户管理主要正常租户的入驻及管控；服务资源管理实现按租户的需求分配资源和实例，支撑变更和删除，并支持显示的申请及审批功能；大数据组件的细粒度管控功能，满足组件细粒度的授权，提升开放的安全能力；租户运营分析提供以租户维度的分析及实际用量提醒功能；支撑细粒度运营；提供服务接入能力，保障基于OSB的自动注册及接入，并形成统一的服务目录。

接入层主要是提供OSB接口的标准接入，并提供接入代理共，保障异构多集群、云边集群的标准化接入，并具备灵活的扩展能力。

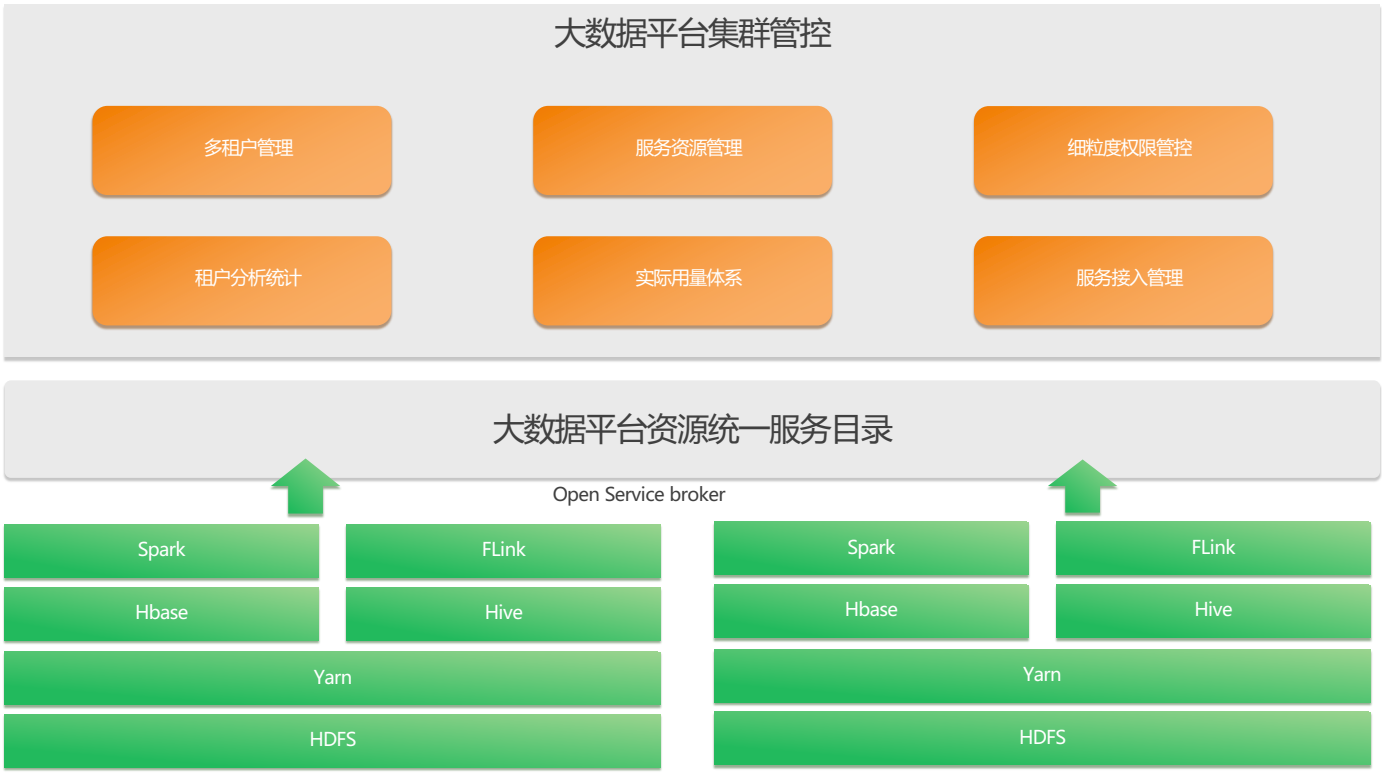


图5　AISW DIF-CM产品架构

返回目录

4.2.3.1. 多租户管理

构建树形多租户模型，逐级管理适配不同的管理模式。支持租户的生命周期管理，实现创建、修改、删除功能。提供租户的时间到期配置，在租户到期后，租户不可访问，租户需求后，租户提供访问能力 。

实现租户成员功能，实现RBAC（Role Based Access Control，基于角色的访问控制）的方式来实现用户的权限管控。当一个用户登录后，通过给用户在制定租户上分配一个角色，可以实现用户在租户上的授权操作 。

4.2.3.2. 服务资源管理

服务管理功能为集群管控的核心功能，实现大数据平台的资源分配及线上申请两种模式。租户管理员可以可根据子租户租户需求分配资源。子租户也可根据自身的需求提出申请，管理员审批后资源生效。对于租户资源使用不均衡，可以根据使用动态调整。当不需要该服务时，可以删除该服务。

在租户资源分配后，租户管理员可根据需求创建服务实例，服务实例的资源配额不能超过该租户的资源额度。租户管理员可以直接创建该租户的服务实例，租户成员创建实例需要首先提出申请。服务实例创建后，根据需求需要变更实例。租户成员需要提出申请，租户管理员审批通过后，实例变更生效。租户管理员或者成员均可删除该实例。删除实例时要做风险提示，删除后该实例物理删除，需要慎重操作 。

返回目录

4.2.3.3. 细粒度权限管控

实现大数据组件HDFS、Hive、HBase、MapReduce2、Spark、Kafka的细粒度权限管控功能，满足细分权限的分配，满足租户对权限控制的需求。对于每个服务实例均支持多条权限策略。支持对多个用户的授权功能。

服务名称	权限控制粒度	权限
HDFS	文件目录、文件数	支持对文件路径的权限管理，包括文件的读/写/执行
Hive	数据库，表	支持数据库的细粒度权限管理，实现表、列的授权，权限包括查询、更新、创建、删除、修改
MR2	队列	支持对队列的权限管理，包括：队列管理，队列查询
Kafka	主题	支持对Kafka实例主题的权限管理，实现细粒度的权限控制，权限包括：发布、消费、配置、创建、删除等
HBase	命名空间	支持对Hbase的命名空间的细粒度权限管理，支持表、列簇、列的授权，权限包括：查询、更新、创建、删除等
Spark	文字内容队列	支持对队列的权限管理，包括：队列管理，队列查询

4.2.3.4. 运营支撑

租户维度的资源视图，按照租户显示租户数量、服务数及用户数，显示按照服务配额参数显示占用情况，并呈现服务实例的实际使用量并显示明细情况。使平台管理员及租户管理员能及时集群资源使用情况，对集群资源动态调整，提升集群资源利用率；

租户的实际用量提供查看功能，并且可以配置门限，形成资源使用告警，已保障业务不会因为资源资源使用情况导致宕机。户管理员提前扩容，用以保障业务的连续性。

支持大数据组件精细化运营能力，使平台管理员直观看到平台的收入，支持大数据平台运营能力。提供服务定价、服务订购、服务计量、服务账单四个核心功能。

4.2.3.5. 服务接入管理

基于Open Service Broker标准并且实现服务的自动接入、自动注册，并支持订购界面的自动生成，实现云边、异构、不同版本的大数据平台资源管控。并规范统一的接入标准，满足物理多租及逻辑多租组件的接入管理。

返回目录

4.2.4. AISW DIF-DLS数据湖存储产品架构

实现结构化数据、半结构化数据及非结构化数据的流批统一存储底座，打通数据通道构建数据T+0的实时数据业务场景，满足多种业务分析的诉求，提供多种计算引擎和存储引擎支撑等。

数据湖存储：实现数据统一存储，解决数据存储多份的问题，批流数据统一存储，支持基于HDFS，Ozone，FastDFS，AWS S3和Aliyun OSS 等存储引擎，数据缓存层支持 Alluxio和JindoFS。

数据湖映射：提供数据格式演变，支持添加，删除，更新或重命名，提供隐藏分区支撑，提供分区布局演变，可以随着数据量或查询模式的变化而更新表的布局，快照控制可实现使用完全相同的表快照的可重复查询，或者使用户轻松检查更改，数据修剪优化使用表元数据使用分区和列级统计信息修剪数据文件，支持事务支持读取操作永远不会看到部分更改或未提交的更改。

计算引擎：根据不同数据应用的计算场景要求，实现批量计算、实时计算、批流混合计算的能力，支持Spark、Flink、Hive、Presto、Hive MR等。

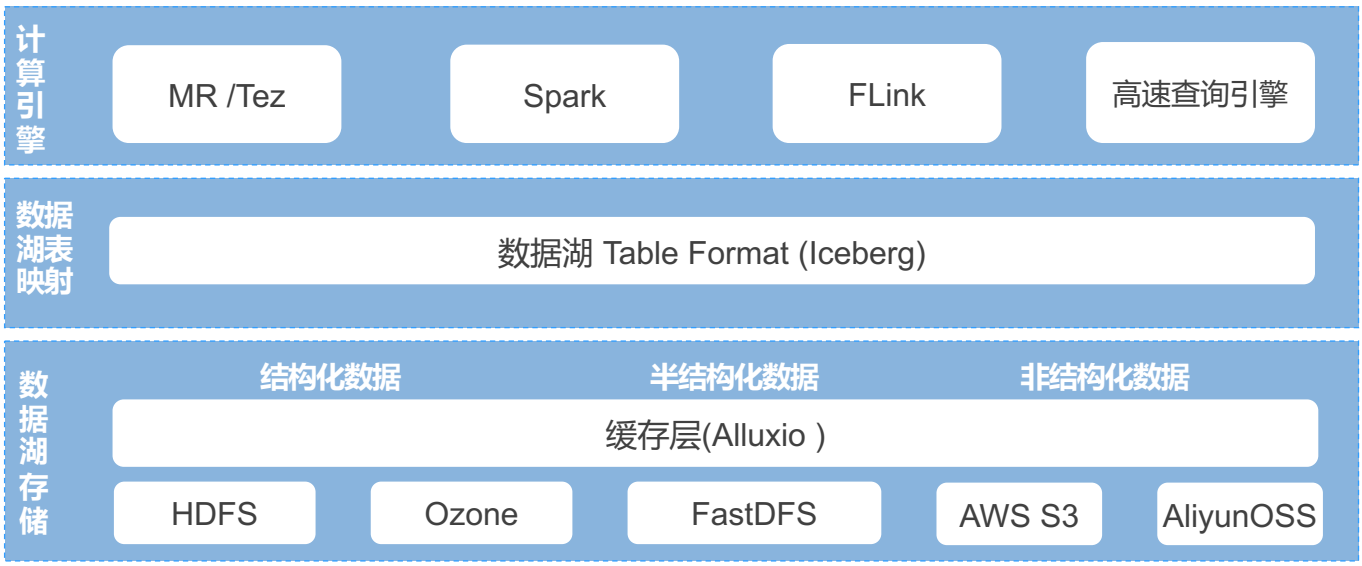


图6 AISW DIF-DLS产品架构

返回目录

4.2.4.1. 结构化数据统一存储

实现结构化数据统一存储，解决Lamda架构带来的数据存储多份的问题，批流数据统一存储，数据只保存一份，实现融合的交互式分析及批量方式。

1. 数据全量入数据湖，即：全量数据批量入湖，增量数据T+0更新入湖；

2. 支持ACID事务，多方可并发读写数据，相互不被干扰及数据入库的一致性；

3. 大数据数据仓库实现T+0数据分析，满足实时分析场景的时效性要求；

4.2.4.2. 非结构化数据存储

提供对非结构化数据（文本、图片、视频等）进行数据装载、查询检索、标准化服务为一体的非结构化数据存储检索系统，同时与大数据平台Hadoop、HBase、Elasticsearch紧密集成，实现详单查询、文本检索、图片查询等查询场景。

1.提供非结构化数据的采集及加载，实现批量加载、数据压缩及清洗转换等功能。

2.提供混合存储引擎，支持对不同大小的文件匹配到不同的存储引擎。

3. 提供非结构化数据的查询检索引擎，支持全文检索、分词、索引创建等能力

4.提供对外的标准服务，支持内容检索API、数据聚合API等

4.2.4.3. 缓存层提升查询性能

基于Alluxio构建数据湖加速层，使用多级存储(MEM、SSD及HDD)功能，实现数据湖存储的缓存，提升数据的访问效率。

1.性能加速：屏蔽存储计算分离后系统IO，网络性能瓶颈。

2.多级缓存：充分利用不同硬件设备性能特性，满足数据的冷热访问要求。

3.统一命名空间：屏蔽不同存储系统的差异性，对上提供统一的访问方式。

返回目录

4.2.5. AISW DIF-RCE实时计算引擎产品架构

为企业级用户搭建统一的分布式流式数据处理平台，实现统一的实时数据接入、处理、订阅，全面保障实时的业务场景开发。

实时开发管理：一站式完成流作业管理能力，根据流数据的特征，进行统一建模管理；

实时数据服务：依据实时业务场景的特点，提供个性化数据订阅和数据推送能力；

实时分析处理：主要对流式数据进行业务逻辑运算，包括：字段计算、多流合并、维度汇总和复杂事件处理；

实时数据交换：主要完成B域订购数据、缴费数据、消费数据等的采集、清洗转换、分发，同时支持O域数据的采集；

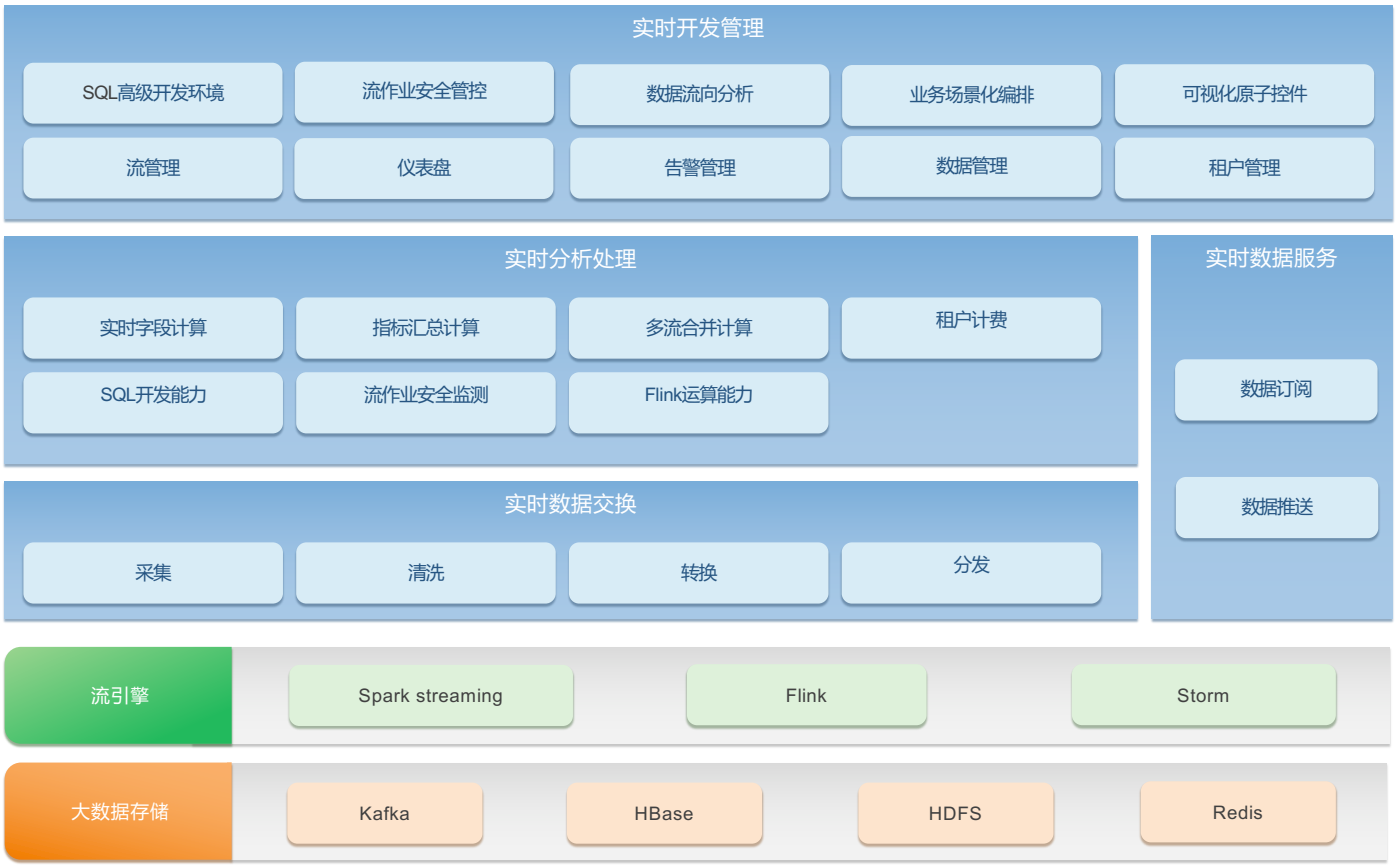


图7 AISW DIF-RCE产品架构

返回目录

4.2.5.1. 实时流数据交换

采用Master-Slave的分布式架构，针对不同系统的多种数据源，提供一站式实时采集、预处理和分发的功能，全界面化数据流采集配置和管理，摆脱单调的自定义脚本和手动流程管理数据流。

- 异构系统间统一调度处理：支持异构系统、平台、数据库间数据调度流程的编排、调度、处理和监控；
- 全界面化操作能力：丰富的图形化操作界面，控件式无编码开发功能，开发0门槛。
- 分布式线性动态扩展：实现节点动态线性扩展，从而满足高性能要求。
- 第三方软件集成能力：提供插件式开发，将对外服务、功能封装成API供其他软件调用；

4.2.5.2. 实时流数据分析处理

通过高速分布式缓存Redis Cluster 完成流数据和批数据的关联运算，满足多维度指标的分组统计运算、实时汇总计算、多流合并计算。

实时字段计算：通过高速分布式缓存Redis Cluster 完成上网类、位置类、订单缴费类等流数据计算，运算速度快，高并发，高吞吐，并为用户提供托拉拽控件的方式，完成SQL即可完成标签，易用性强。

实时汇总计算：实时增量/全量数据汇总分析，支持多指标多维度并行计算，汇总结果直接输出给外部系统使用，提高效率，支持SQL语法，便捷、易用。

多流合并计算：解决多种流数据合并处理，例如：位置流+内容流的实时join场景，完全基于Spark、Flink内存机制，而非与外部组件交互，提供双时间窗口设定机制，规避时序性、延迟性。

4.2.5.3. 实时数据开发

提供开发者基于控件模式的流数据开发编排能力，屏蔽了复杂的底层开发过程，降低开发门槛。提供向导式开发过程，简单易用，大幅提升客户感知。

返回目录

4.2.5.4. 实时数据服务

依据实时业务场景的特点，提供个性化数据订阅和数据推送能力；对实时的数据模型，实现租户间的实时发布共享，为用户提供更加便捷的实时数据服务。

租户进行实时数据的个性化订阅：

- 表字段级别的细粒度订阅；
- 提供订阅周期：随机，每天，每周，每月和每批次；

租户实时数据发布共享：

- 统一化的流数据模型定义、共享、审批、订阅的全周期管理。
- 租户管理员对实时作业、实时数据模型进行审核。

4.2.5.5. 实时任务监控及告警

全面展示作业运行的健康状态，包括运行时长、任务的运行情况。

提供实时和历史的性能指标分析和展示，同时提供性能优化的参数设定，即时生效。

Kafka核心指标实时监控，同时提供告警项和阈值设定，实时分级展示告警信息。

返回目录

4.2.6. AISW DIF-SE关联检索引擎产品架构

通过提供数据的存储、建立丰富的索引，多样化的查询接口，支持各种结构化业务数据解析，能够为更多的用户，丰富的数据类型，为多样化的业务提供通用的查询能力。

高效的查询性能：通过对不同业务场景建立索引，实现对流数据和批量数据的高效查询。

灵活的查询接口：提供可视化查询界面和API查询接口，通过定义丰富的查询参数支撑灵活的数据查询。

便捷的聚合查询：通过定义预定义函数，实现SQL的聚合查询，屏蔽底层查询的复杂性。

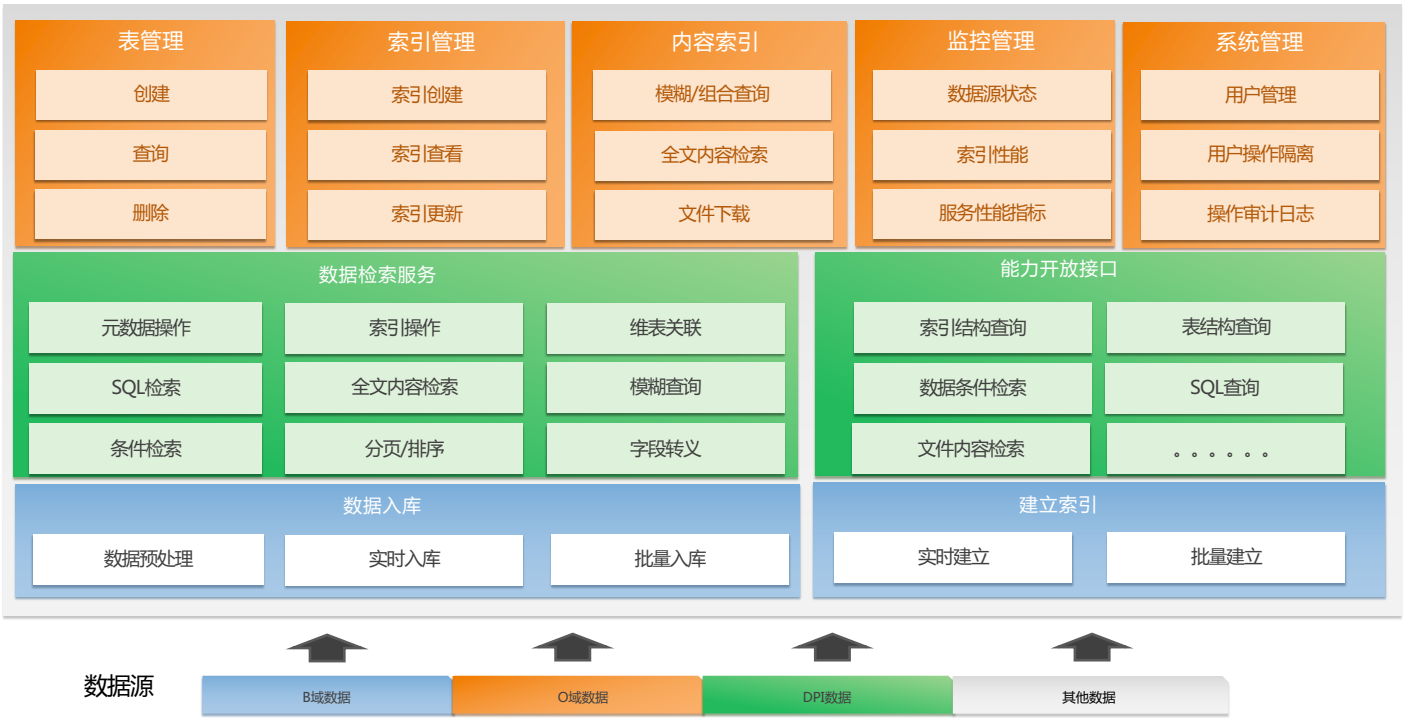


图8　AISW DIF-SE产品架构

返回目录

4.2.6.1. 实时数据入库

将结构化的数据文件实时读入flume source中，通过flume sink实时地将结构数据入库到HBase中。在数据实时接入的过程中，会根据数据内容以及数据的格式动态计算出该条数据对应的数据表，实现不同数据入到不同的表中。实时入库功能实现数据在入数据库的同时进行字段解析、建立索引等工作。

1 实时入库：入库过程数据不落地，极大提升入库时效性；

2 杜绝数据丢失：采用新的架构，各方面确保数据可靠性和准确性；

3 跨月数据分拣入库机制；

4.2.6.2. 实时数据索引

将写入hbase中的数据，按照自定义字段实时建立索引到solr中，也支持全字段建立索引，这样更好满足业务的随机字段检索需求。

整个实时数据索引功能中，使用了业界主流的三个组件：hbase、solr、hbase index，来完成整个过程。

4.2.6.3. 高效且多维的数据查询

提供按照业务查询需求建立的多字段索引，实现并发的多维数据查询能力。

1 实时建立全字段索引：不仅按照rowkey规则建立，同时可以完成其他字段的索引创建；

2 实时多样查询：不仅按照rowkey查询数据，同时非rowkey同样查询；

3 采用新的核心架构，支撑多场景下的并发条件查询；

4.2.6.4. 检索服务调用

可以作为统一的查询检索服务对外提供API调用，外部系统可以通过接口调用来完成业务数据的检索功能。

返回目录

4.2.7. AISW DIF-HQE高速查询引擎产品架构

基于开源分布式SQL查询引擎Presto构建产品。支持从多种数据源获取数据，一条标准SQL查询可以将多个数据源的数据进行合并查询分析，屏蔽底层多类型组件语言的差异性，实现跨数据源数据分析。

实现跨数据源的大规模的交互式查询分析，主要异地多集群的融合查询。支持标准SQL查询，对于不同的数据源基于标准SQL即可查询， 通过标准 SQL可以实现两个数据源的融合查询 。并实现集群的多租户管理，提供多类型的对外访问接口。

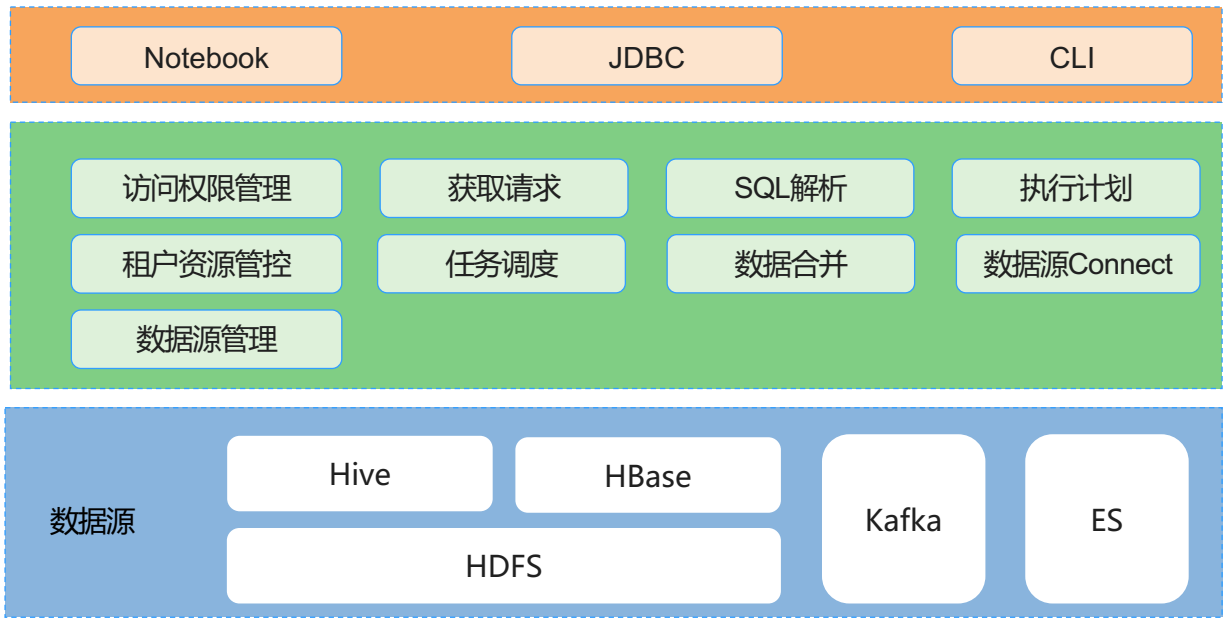


图9　AISW DIF-HQE产品架构

返回目录

4.2.7.1. SQL解析功能

支持标准SQL的执行，实现标准SQL的词法、语法分析及对应的语义分析，并支持对应的执行计划，并可对执行计划进行分段，对分段计划进行调度能力，分配到对应的执行节点，并连接数据源，执行对应的计划，获取执行结果，并根据分段执行的接口进行合并。

4.2.7.2. 权限管理

用户权限管理，实现与权限组件SSO的用户同步，并支持用户的变更及删除操作。

数据源权限管理，支持数据源的权限、Schema的权限及表全的管理，并支持安全策略的管理，实现策略的创建、变更及删除。

4.2.7.3. 数据源支持

支持Hive、Mysql、Kafka、Elasticsearch、PostgreSQL、Redis等常用数据源。

支持数据源连接的添加，即可在线添加支持数据源的连接，并支持数据原的管理功能。

4.2.7.4. 访问方式支持

提供JDBC、CLI及Notebook三种访问方式。

Notebook访问可以支持数据源的管理，支持SQL语句的输入及执行，并显示执行的记录，对于小数据量可支持查看及预览，大数据量的可以提供下载功能。

返回目录

4.2.8. AISW DIF-GA图分析引擎产品架构

支持大规模的超大图计算和查询，兼顾图计算和图查询的高并发、低延时要求。针对复杂数据关系分析的独特场景，其性能远胜于传统数据库技术。未来还会在大量单场景的基础上向知识图谱不断演化。

图计算能力：实现原始数据预处理转换、图数据顶点过滤、图属性计算、图数据加载入库等数据接入及预处理、加载工作。

服务提供：提供固有的图算法库，针对不同图场景实现算法支持；建立实时、批量的图数据模型库并对图模型进行有效管理。

服务开放：通过标准接口，提供实时图数据的实时图数据交互查询，图数据的全量展现等。

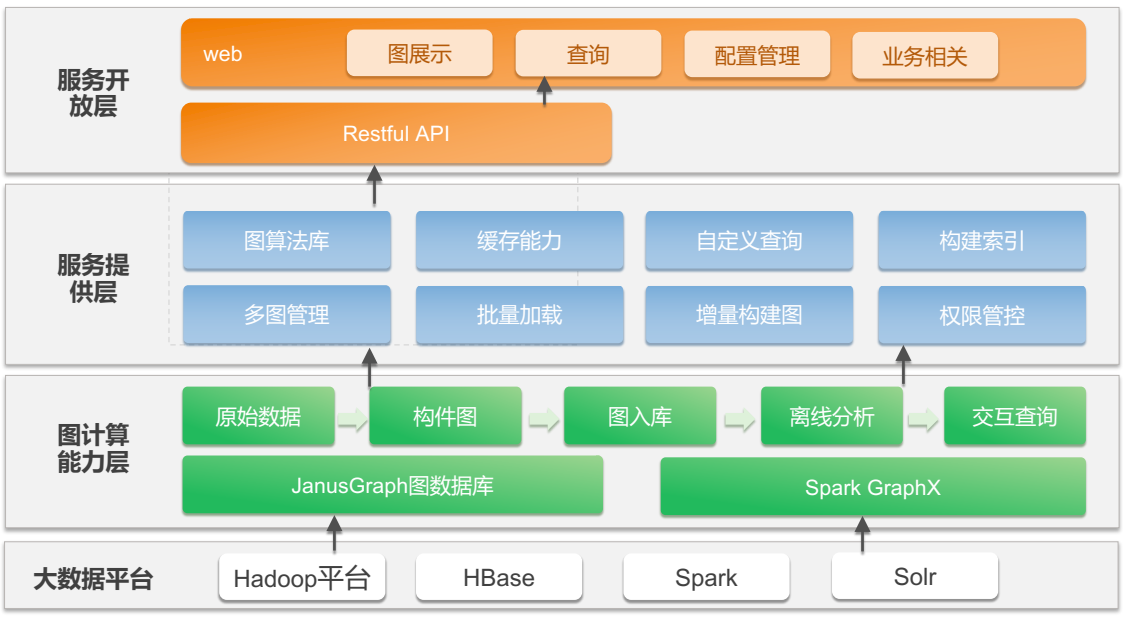


图10　AISW DIF-GA产品架构

返回目录

4.2.8.1.海量图数据存储查询

统筹每一份图数据的从原始数据到最终数据加载的完整流程，并管理这些数据处理流程，以保证图数据正确完整地被固化到图数据库或分布式文件系统中。实现图数据的存储，实现横向扩展，存储量可达千亿边级别。支持大量的并发事务和操作图处理。在千亿边的数据规模下，提供毫秒级的实时点、边查询。

向导式的操作界面和图数据展示，支持点边复杂条件的查询，提供图查询、图管理、gremlin查询等各类API；可以直观的展现查询结果。

4.2.8.2. 图模型库

模型库提供一系列的图分析模型和运行这些模型的参数和方式。这些图分析模型可以是与JanusGraph图数据库交互的本地分析模型，也可以是运行在Spark上的SparkGraphX全局图分析模型。模型库提供预设的图分析模型以外，还预留了自定义模型接口，用户可以根据业务需求开发上述任意类型的图分析模型，并将自定义模型注册加载到模型库中。

4.2.8.3. 图数据可视化

提供可视化的图数据检索分析工具，内置图检索工具、基本图查询接口，实现用户对图的关键实体挖掘和实体间关系的探索。进行准实时地响应节点搜索、多跳查询、最短路径分析等在线查询分析操作，直观展现全量图数据的复杂关系，满足客户在各个场景的定制化需求。

返回目录

4.3. 关键技术能力

介绍大数据基础平台产品关键技术能力，所包含的各个模块的技术能力

4.3.1. 大规模分布式文件存储

HDFS是Hadoop的分布式文件系统，实现大规模数据可靠的分布式读写。HDFS针对的使用场景是数据读写具有“一次写，多次读”的特征，而数据“写”操作是顺序写，也就是在文件创建时的写入或者在现有文件之后的添加操作。HDFS保证一个文件在一个时刻只被一个调用者执行写操作，而可以被多个调用者执行读操作。HDFS提供通用存储能力，可以存储结构化、半结构化、非结构等不同格式的数据。具有分布式线性扩展能力，可提供EB级数据存储能力。

4.3.2. 多源异构跨域查询引擎

多源异构数据查询引擎基于SQL的分布式查询引擎，与Hadoop生态无缝结合，实现海量数据秒级查询，支持多源异构系统，支持一站式SQL融合分析。跨源融合查询，实现跨数据源的大规模的交互式查询分析，主要异地多集群的融合查询。标准SQL支持，支持标准SQL查询，对于不同的数据源基于标准SQL即可查询， 通过标准SQL可以实现两个数据源的融合查询 。多租户隔离，实现多租户隔离机制，使租户间的查询不受影响。支持Hive、HBase、ES、Kafka、Mysql数据源。

4.3.1. 云边协同

大数据管控平台由目前的多集群资源管控向云边协同演进，重点实现云边集群的资源管控、云边集群协同及云边任务调度系统能力。云边集群统一集中运营，实现大数据资源的按需分配，支撑租户需求。云边集群资源统一调度，建立统一的资源调度中心。云边集群跨源查询，提供融合查询能力，支持跨集群的数据查询。

返回目录

五. 功能介绍

5.1. 基础功能

功能点	功能点描述
分布式数据存储	支持分布式数据存储的各种操作，包括： 1、分布式数据存储，数据一致性； 2、分布式文件访问，文件读取； 3、分布式文件系统高可用； 4、数据存储策略； 5、数据均衡； 6、压缩存储；
分布式资源管理	支持分布式资源管理，包括： 1、集群资源管理； 2、集群资源调度； 3、YARN服务框架； 4、应用优化调度与处理；
集成Hive	提供数据仓库Hive能力，包含: 1、提供Hive+Tez+LLAP； 2、支持DDL、DML、DQL操作 3、Hive Warehouse Connector支持Spark 4、元数据管理 5、支持ACIDv2； 6、权限管理和配置
集成Hbase	支持Nosql数据库，包含: 1、HMaster管理，异步RPC机制 2、RegionServer管理、负载均衡，逻辑分组 3、数据库接口、API、Web Service 4、Region任务管理 5、数据查询读写，Hfile操作
集成Phoenix	提供SQL层访问NoSQL数据，包括： 1、数据查询操作界面、命令; 2、数据查询操作转换； 3、列编码与日志； 4、支持Hive、Python;

返回目录

功能点	功能点描述
集成Spark	支持spark的各种操作，包括： 1、缓存持久化; 2、元数据检查点、元数据管理; 3、数据检查点，数据批处理，时间片管理 4、DAG、JOB、TASK管理 5、流输出、监控、监听接口
集成Kafka	提供Kafka分布式消息系统，包含： 1、分布式和分区; 2、数据副本; 3、数据生产； 4、数据消费; 5、消息传送机制
集成Ambari	使用Ambari运维管理集群,包含: 1、支持中文界面，使用全新的UI界面； 2、支持批量增删节点、组件；密码、数据库、用户能够集中配置； 3、增加操作提示功能； 4、集群监控； 5、主机配置; 6、服务管理
底层基础服务	提供底层基础能力,包含: 1、提供TEZ引擎; 2、提供ZK基础服务; 3、集成Ranger，支持认证、授权、审计等 4、组件支持操作系统的Kerberos认证

返回目录

功能点	功能点描述
集群负载分析	1、集群CPU、内存、存储的监控。2、集群作业的监控。3、集群数据节点DataNode\NodeManager\Region Server\ 管理节点的情况NameNode\ResouceManager\HMaster4、集群负载分析。
集群自动巡检	对集群主机、网络、HDFS、Yarn、ZK、Hive、Hbase、Spark、等Hadoop服务的关键指标进行自动检查
租户自主运维	1、租户作业的appid,类型、开始时间、完成时间、当前状态、内存、CPU使用情况。2、队列CPU和内存。3、租户可以杀死自己的作业、租户调取失败作业的部分container 日志（来定位作业失败的原因。4、租户负载分析5、Hive数据库使用趋势6、Hbase的region读写次数分析
HDFS性能分析	支持集群慢速节点画像、RPC画像、name node JVM监控、HDFS小文件分析
YARN性能分析	通过对集群的内存利用率和任务等待个数分析。如果等待个数多，利用率高。会建议用户扩容。如果等待的个数多，利用率少，那么需要对集群的队列资源进行重新配置。详细的建议参见队列分析部分。如果等待任务为0，内存利用率较高，则该集群的得分很高，暂时不需要重新配置。
Hbase性能分析	对Region的读写请求次数进行分析，发现存在数据倾斜的Region.以及Region倾斜度的排名。
运维监控权限改造	各个租户登录系统之后，只可以看到自己权限以内的报表
运维知识库	支撑现场运维，积累运维经验。
hadoop审计	审计用户对数据的操作.自定义告警策略，实时捕获不规范操作。对审计和告警的结果进行多维度的分析。审计和告警全部记录，支持关键字搜索。
日志中心	支持第三方日志通过kafka接入。收集hadoop集群组件日志，展示日志，支持关键字搜索
集群负载分析	1、集群CPU、内存、存储的监控。2、集群作业的监控。3、集群数据节点DataNode\NodeManager\Region Server\ 管理节点的情况NameNode\ResouceManager\HMaster4、集群负载分析。

返回目录

功能点	功能点描述	
多租户管理	租户模型管理，及支持租户的入驻，包括： 1、租户的生命周期管理，包括创建、变更及删除； 2、租户到期时间配置； 3、租户成员管理，支持三种角色；	4、树形多租户模型； 5、租户简略信息统计
服务资源配额	面向租户按需分配资源，包括： 1、资源配额或申请； 2、资源配额变更； 3、资源配额删除；	
服务实例	租户根据应用需求创建所需实例，包括： 1、服务实例创建或申请； 2、服务实例变更； 3、服务实例删除；	
我的待办	在服务配额及服务实例的申请、变更时，需要提交申请工单，申请信息会以代办方式，显示我的待办功能中，包含: 1、待办工单 2、已处理工单	
细粒度 权限管控	支持大数据组件的细粒度权限管控，包含: 1、HDFS细粒度权限 2、Hive细粒度权限 3、HBase细粒度权限	4、Spark细粒度权限 5、MR2细粒度权限 6、Kafka细粒度权限
大数据平台 多租户管理	基于OSB实现大数据平台组件的多租户功能，包括： 1、HDFS基于Namespace实现多租户管理 2、Hive基于Database实现多租户管理 3、MR基于Yarn queue实现多租户管理	4、HBase基于Namespace实现多租户管理 5、Spark基于Yarn queue实现多租户管理 6、Kafka基于Topic实现多租户管理
服务及工具 接入管理	开放的服务接入框架，支持基于OSB组件及工具的接入，包括： 1、服务接入; 2、服务注册;	3.、服务接入变更 4、服务下线 5、服务日志
运营支撑	支持租户分析及实际用量提醒,包含: 1、按照租户提供基础统计、服务已分配量及服务实际使用量的分析功能，提供图表显示，并可按照租户角色查看; 2、对服务实例的实际使用量提醒，并可提供自定义门限; 3、组件的计量计费能力，提供定价、订购及账单等	

返回目录

功能点	功能点描述
流数据管理	支持sparkstreaming引擎作业的管理 支持Flink引擎作业的管理 支持Storm，Jstorm自定义作业的管理 支持流数据的断点续读功能，基于offset进行断点读取功能
租户管理	支持与多租户系统集成，从多租户系统同步租户和资源信息 多租户下的数据权限隔离 多租户下的流作业和操作隔离 支持成员在归属的不同租户间，自有切换功能
图形化控件开发	预处理类控件、字段增强类控件、实时汇总分析类控件、多流关联类控件、数据输出控件
基础配置管理	包括对计算引擎配置、数据存储配置、缓存信息配置、审计日志管理
索引建立	提供对于实时入库的数据进行实时建立索引 提供对已经导入hbase的数据批量建立索引。
数据服务	支持多条件的数据检索功能，可根据返回字段设置需要返回的结果数据 支持通过录入文本关键字进行全文内容检索，可以设置根据不同字段的权重进行检索 支持利用通配符进行模糊查询 支持按照RowKey字段条件，查询数据明细记录。支持精确查询和范围查询 支持按照SQL语句进行查询和汇总。 支持多条件检索，全文检索，模糊查询，都可以按照索引字段排序
实时入库	支持结构化的csv数据格式的实时入库， 支持实时接入结构化数据过程中动态获取非结构化数据转存入HBase中。 数据可以根据设定的规则入到不同的表中 按照Schema的字段信息对原始数据进行过滤筛选以及字段类型转换 提供数据过滤规则，不满足规则的异常数据直接过滤掉 按照预先设定好的规则，组装生成对应的rowkey值
批量入库	按照Schema的字段信息对原始数据进行字段类型转换 提供数据过滤规则，不满足规则的异常数据直接过滤掉 按照预先设定好的规则，组装生成对应的rowkey值 提供批量数据导入作业状态查询功能

返回目录

5.2. 特色功能

5.2.1. 大数据多租户能力模型

大数据平台包括组件众多，面向租户资源分配不统一，创建操作比较繁琐，需要功能通过CLI、管理UI、Ranger来实现，随之租户数量的增多，此管理方式成为租户运营的障碍。

抽象大数据HDFS、Hive、HBase、MapReduce2、Spark、Kafka组件的多租户模型，通过资源管理能力面向租户开放。

服务名称	租户模型	配额参数
HDFS	基于Namespace实现多租户管理	最大文件数量，最大存储容量
Hive	基于Database实现多租户管理	数据库大小
MR2	基于Yarn queue实现多租户管理	内存大小
Kafka	基于Topic实现多租户管理	Topic存活时间，分区大小，分区数
HBase	基于Namespace实现多租户管理	最大表数量，最大的region数目
Spark	基于Yarn queue实现多租户管理	内存大小

5.2.2. 面向垂直行业客户的运维和体验管理

采用行业通用的Open Service broker（OSB）协议实现大数据平台组件的纳管，基于该标准协议集群管控平台可以实现对于不同厂商、不同版本的接入，并且可以支持的云边集群资源的统一管控。提供OSB接口的接入代理，实现OSB的自动接入，并且可以在集群管控中自动生成订购界面。组件只需要提供OSB7个接口定义即可，无需其他开发工作。通用的OSB协议，不但可以接入大数据平台组件，也可以实现其他第三方组件及工具接入。

返回目录

5.2.3. 跨中心，异构集群云化

大数据云化平台管理，主要解决云化平台上的基于多中心，多云平台上的多集群管理和运营。未来的企业云架构会是一个物理云，虚拟云和容器云混搭的架构，并且基于5G的云边架构，形成云端，边缘短的分级云平台，云化大数据平台适应这种基础设施的变化，也会形成跨中心，多平台，多集群的多级部署，多种大数据平台的能力需要统一整合， 统一管理和统一运营。

5.2.4. 增强实时数据场景开发

为企业级用户搭建统一的分布式实时计算平台，实现统一的实时数据接入、处理、订阅，全面保障实时的业务场景。

- 增强Flink引擎的支持，按需选择引擎，灵活切换。
- 租户计费功能，面对租户进行流数据的计量和算费。
- 流作业安全管控，使得各租户后台执行的程序包更加有序和清晰，防范数据的违规操作。
- 面向业务人员，将提供业务场景化编排功能，提供场景化的配置。
- 数据流向分析，对于流数据模型之间的转变过程，进行分析展示。
- 流原子控件能力，提供原子功能的控件，提升编排开发的灵活度，更好的满足需求 。

返回目录

5.2.5. 提升检索效率和易用性

通过提供数据的存储、建立丰富的索引，多样化的查询接口，支持多种非结构化数据解析，能够为更多的用户，丰富的数据类型，多样化的业务提供通用的查询能力。

- 自然语言检索：利用自然语言解析能力，使对非结构化数据的搜索结果更为精确。
- 非结构化数据查询：实现对非文本、语音数据的解析和建立索引，支持非结构化数据检索查询。
- 支持统一SQL方式的检索和查询，提升其易用性。

5.2.6. 智能租户资源分配

集群资源利用打分:大数据集群上的作业运行在不同的队列上，可能会出现有的队列忙，有的队列闲的情况，这就是资源分配不合理，如果能对它们整体合理调度，可以提升集群的资源利用率，同时保证SLA。该功能综合集群的资源使用率、任务完成率、等待率进行建模,最后的给集群资源利用打分.

AI模型计算资源分配额度: 按照流计算、高优先级、一般队列进行分类，参考资源使用情况、作业等待情况等指标进行优化。该功能需要收集前期作业的运行情况，通过人工智能的算法，不断学习迭代，给出最优化的租户资源分配方案，供管理员选择。

5.2.7.多层关系数据的图存储分析

提供从数据采集、分析、构建、可视化展现的一站式开发服务，同时提供安装部署、监控等配套工具。通过标准REST API接口提供各类图数据的检索和分析服务，包括图管理、图检索、图深度分析等。

通过图构建、解析、索引、入库加载，快速构建图数据，同时提供高效遍历检索图数据能力。

返回目录

六. 场景应用方案

6.1. 技术领先的大数据集群

使用DP5.3搭建大数据集群，数据存储在三副本基础上，新增了纠删码策略，可提高存储利用率；使用HDFS联邦技术，两个NameNode分别提供给两部委、业务中心单独使用，水平扩展了NameNode的性能；每个NameNode又支持2个备用的SecondNamenode，提高了高可用能力，目前集群900个节点，数据容量50PB。

返回目录

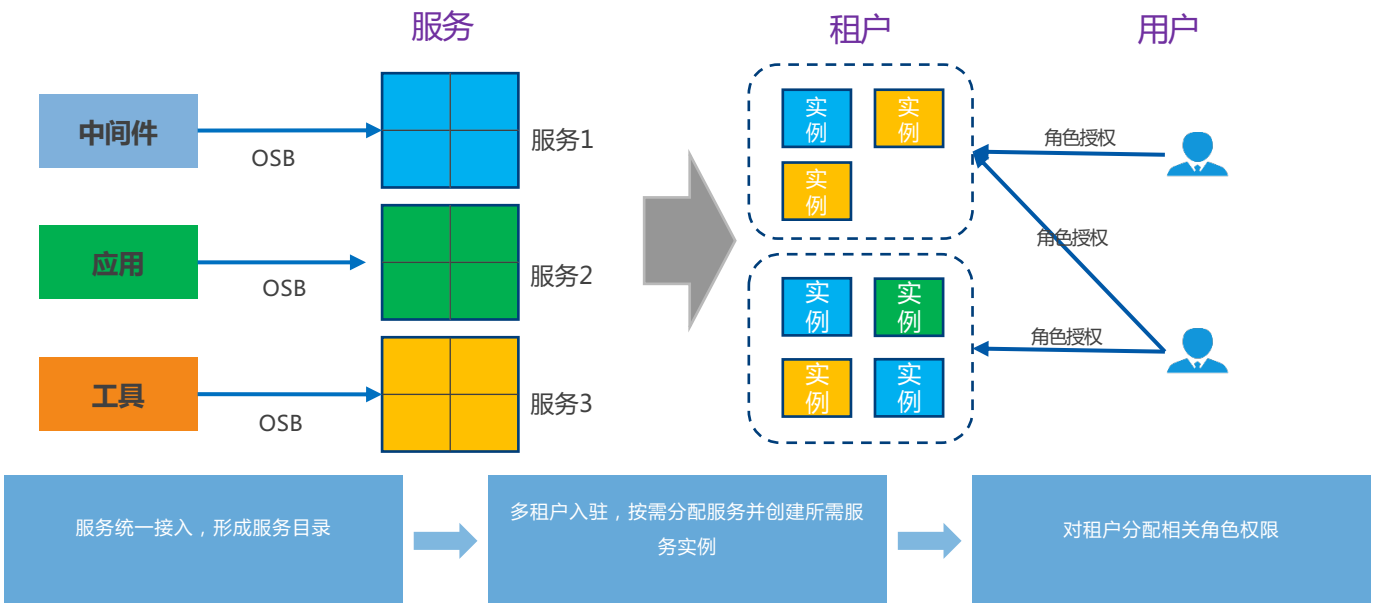
6.2. 大数据平台多租户管理

目前大数据领组件及工具众多，有些组件不支持租户管理，有些组件租户管理自成体系，无租户统一模型，无法统一管理。 实现大数据平台的多租户管理，以租户为对象开放大数据平台能力，并提供监、分析等手段。

在大数据平台多租户管理的场景中，提供的能力主要有：

1、通过租户、用户、服务、实例、工具及组件概念，将大数据生态系统的上下游所有基础技术组件、开发中间件进行有效整合，实现工具、组件的统一接入管理，；

2、通过服务资源管理功能，满足租户“按需分配、即分即用、弹性扩容”的诉求。有助于提升资源的有效利用，节省企业投资。



返回目录

6.3. 支撑大数据平台运营

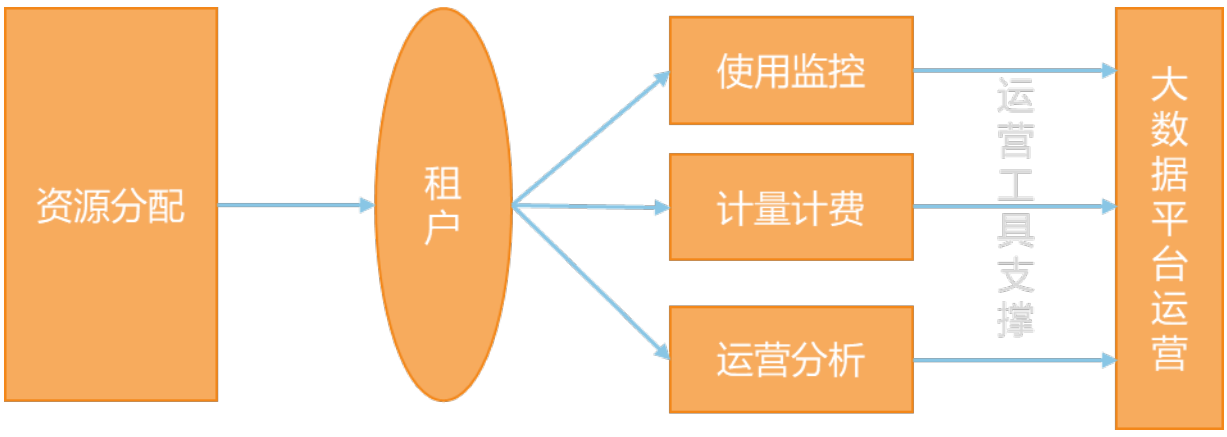
随之5G、AI的到来，数据量增大，对数据的诉求日益增多，应用也百花齐放，平台的计算及存储资源总是有限，对大数据平台的运营尤为重要。

在大数据平台运营的场景中，提供的能力主要有：

1、资源监控能力，监控租户分配资源及实际使用资源，根据监控情况，可对资源进行动态的调整。

2、计量计费能力，提供包年包月及按量计费两种计费模式，以租户维度精细掌握资源消费情况，使平台投入成本及费用精细化管理；

3、运营分析能力，提供租户分析能力，根据资源的使用情况和趋势，可以直观的配置租户资源。并且可按照租户提供实际用量的提醒功能，保障不会由于资源不足出现中断的事件。



返回目录

6.4. 日常智能运维

租户和管理员可以通过可视化的界面，对集群的资源、集群访问情况、集群运行情况进行监控，及时发现日常运维的问题，并通过不断积累形成知识库，支撑日常运维工作。

收集集群、组件的基础指标和日志，分类处理、多维度分析，快速从集群视角、租户视角中提取关键运维指标。提供可视化的操作界面，实时监控集群日常运行情况、租户所使用的队列资源用量。

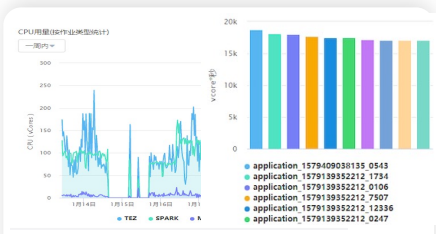
监控HDFS性能分析专家系统模型的结果，通过查看RPC调用分析，监控Namenode的RPC调用平均时长和等待时长，并可以监控RPC调用最多的用户TOP10. 以及用户发起的RPC指令，监控集群的负载情况。

通过查看访问告警的图表汇总，发现/apps/hbase目录的数据访问产生的告警信息，包括用户、操作、目录的信息等。

按照自定义的时间范围查看租户所使用的队列资源用量.由此判定租户的资源平均利用率、峰值.对指导资源分配、扩缩容有指导意义。

资源调度监控

展示作业详情列表, 包含APPID 、用户、时间、任务类型、进度、状态，通过查看作业调度日志，监控调度异常，并定位异常调度的原因。



集群运行洞察

监控租户的CPU和内存用量，MR、SPARK、TEZ使用资源的情况，便于租户管理资源、监控异常，判断是否可以运行更多的作业。

数据访问监控

监控Hive数据库空间趋势图，租户Hive表详情列，格式、大小，数据表的存储， 监控租户Hbase表的region读写次数。

返回目录

七. 带给客户的价值

- **低成本、高效的计算存储：**提供多种分布式计算、存储技术支撑海量数据低成本存储，多种批量、实时计算技术支撑高效的计算。
- **简单、易用的资源服务：**屏蔽底层复杂的资源调用、资源隔离的管理，为客户提供可量化、可伸缩、透明的各类资源服务。
- **降低开发门槛，提高开发效率：**为客户提供基于多种场景化的开发工具，通过可视化配置完成开发，降低技术使用门槛，提升开发效率。
- **低成本自动化运维：**为客户提供自动化部署，缩短建设周期；引入人工智能技术，实现智能化运维，降低后期维护成本。

八. 产品优势

亚信AISWare DataInfrastructure产品的优势集中体现在：

多种数据存储方式：提供如HDFS、Hive、Hbase等多类型存储方式，HDFS使用纠错码节省数据备份数量，大大节省存储空间。

多种计算融合：实现多种实时数据处理技术，批流数据处理技术及计算向边缘延伸，大大提高计算效率。

多集群管理：实现跨域、多集群，边云协同的统一资源调度，大大扩展集群节点数量，实现大规模的集群管理。

定制化多种开发查询工具：预置场景化实时处理工具、交互式数据检索工具及图分析工具，为客户提供便捷的开发查询需求。

全方位洞察，智能运维：通过资源、性能、安全的深度洞察和智能规划，保障大数据集群的合理部署和不断优化，达到充分利用资源的目的。

返回目录

九. 联系我们

亚信科技（中国）有限公司

地址：北京市海淀区中关村软件园二期西北旺东路10号院东区亚信大厦

邮编：100193

传真：010-82166699

电话：010-82166688

Email：5G@asiainfo.com

网址：www.asiainfo.com

Thank you

亚信科技依托产品、服务、运营、集成能力助力企业数字化，持续创造新价值。