GBASE®

南大通用数据中台

技术白皮书

www.gbase.cn

南大通用数据中台技术白皮书，南大通用数据技术股份有限公司

通讯方式

南大通用数据技术股份有限公司

天津市西青区工华道 2 号天百中心 3 号楼 3 层

电话：400-013-9696　　邮箱：info@gbase.cn

目 录

1 数据中台概述

1.1 数据中台定义

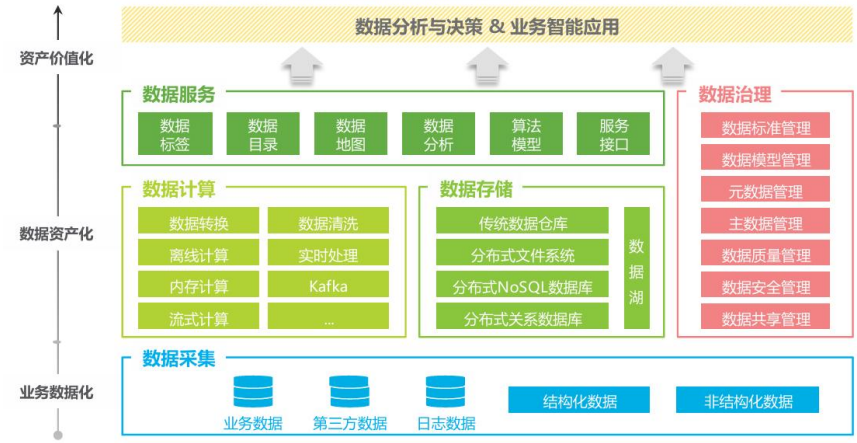
数据中台的定义自诞生以来经历了不断的发展演变。数据中台源于企业内部通过组织架构调整所形成的公共数据能力，通常通过将企业各部门和业务线所需的数据能力提炼并整合形成，是企业内部可复用的统一数据能力集合。

而随着相关理论和技术的持续发展，数据中台成为使企业综合数据能力建设的更好的一种形式。当前的数据中台可通过狭义与广义两种定义来进行描述。

狭义的数据中台指在企业内部通过对数据半成品、算法、模型、工具等能力的积累，支撑业务应用，为前台提供数据能力的企业级数据中枢平台。狭义数据中台聚焦在数据服务的创建和提供，并不包括数据本身的生产、加工、传输等基础性工作。

广义的数据中台是企业数据价值实现的能力框架，包括数据存储汇聚、数据开发、数据管理、数据服务、数据资产运营等能力。通常通过企业统一的一站式数据加工生产利用逻辑平台的形式具象化，是企业级数据价值生产的中枢平台。进一步的，在企业层面数据中台是企业业务数据化的承载体，是企业业务通过数据视角的一种呈现，担负了企业数字化所需的核心综合数据能力，是数据驱动企业的核心引擎。

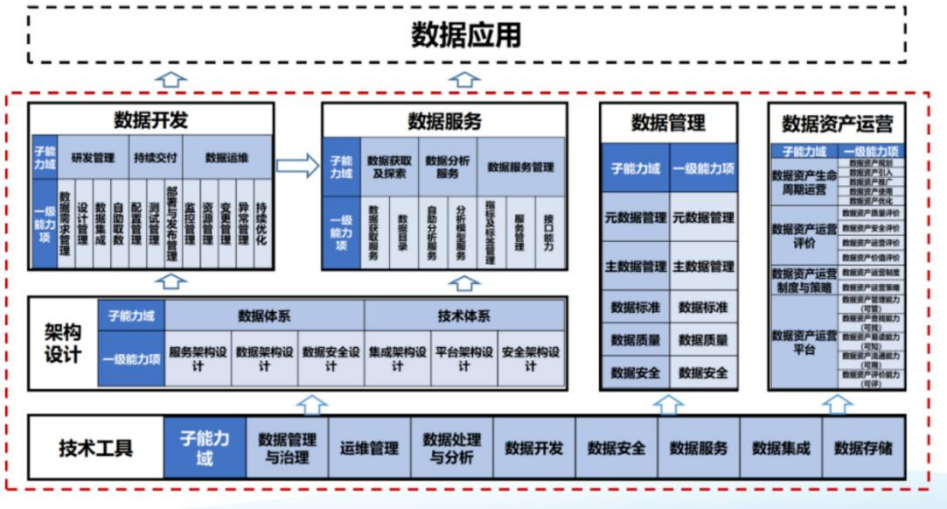
数据中台架构如下图所示：



南大通用能够基于在金融、电信、政企等行业积累多年的数据管理经验和技术优势，帮助客户搭建具有企业特色价值的数据中台，实现业务数据化、数据资产化以及资产价值化。

1.2 数据中台建设要求

根据中国信息通信研究院在 2023 年发布的《数据中台能力成熟度模型》框架，数据中台作为企业为了使数据产生业务价值所需要构建的能力集合，应当具备技术工具、架构设计、数据开发、数据服务、数据管理、数据资产运营共六大能力域。



技术工具是数据中台的物理基础设施。技术工具能力域从产品功能的角度集中体现了企业建设数据中台所需的全部技术工具能力集合。具体细分为数据管理与治理、运维管理、数据处理与分析、数据开发、数据安全、数据服务、数据集成、数据存储共 8 个子能力域。

架构设计完成了数据中台的整体设计图。架构设计是指基于企业数据中台建设及业务需求，参考已有战略和企业架构，结合组流程规范和技术工具等，对数据体系和技术体系的蓝图进行设计，最大限度满足业务和管理需求的过程。该部分重点关注湖仓设计、数据建模、平台工具的集成架构、安全架构等内容。

数据服务是数据中台面向用户的服务内容出口。数据服务部分规范了数据中台对外进行能力输出的方式，其中实现了对业务部门常规数据工作的下沉与复用，由此达到数据赋能业务的效果。具体分为数据获取及探索、数据分析服务、数据服务管理三个子能力域，该部分重点关注数据探索、数据获取、自助分析、模型管理、指标和标签管理等内容。

数据资产运营提供了促进使用数据中台的策略和措施。数据资产运营能力域的标准仍在编制过程中，当前包括数据资产生命周期运营、数据资产运营评价、数据资产运营制度与策略、数据资产运营平台四个子能力域，该部分重点关注资产生命周期、作为核心的评价体系，以及运营制度与策略。

数据开发是将数据加工形成可用半成品的生产过程，是面向数据全生命周期，聚焦于协同从数据需求输入到交付物输出的全过程。该部分标准仍在编制中，计划分为研发管理、持续交付、数据运维三个子能力域，分别对应企业数据开发过程中的研发、交付、运维三个主要环节。

数据管理提供了保障数据质量及安全的制度及措施。数据管理是包括提升整体数据质量、保障数据安全可用等在内的一系列工作的总和。计划分为元数据管理、主数据管理、数据标准、数据质量、数据安全五个能力域。

2 GBASE 数据中台价值

GBase 数据中台对于企业有五大核心价值：

- 湖仓一体，数据高效流转

GBase 数据中台依托湖仓一体架构作为其基础底座。能够高效整合并管理海量数据资源，有效打通数据仓库和数据湖，让数据和计算在湖仓之间自由流动，提升数据流转效率，促进数据协同共享，从而构建一个完整的有机大数据技术生态体系。湖仓一体不仅让数据有序、高效、高质、安全在湖仓间流转，降低运维成本，更助力实现数据的深度融合，为数据中台的建设提供重要的基础支撑。

- 全量入湖，打破数据孤岛

GBase 数据中台可提供业务系统数据全量入湖的能力，并实现贴源数据按照主题入湖、贴源数据标准化、历史数据归档存储等功能，保证数据的完整性、原始性和可访问性。全量入湖打破源系统数据孤岛，实现对数据的全面掌控和高效利用（统一存储、统一管理和集中服务）。同时建立起内外部数据融合机制，数据一点接入、多方共享，提供内外部数据的融合加工服务能力。

- 治理同步，提纯数据资产

GBase 数据中台可基于自身的数据质量管理能力帮助用户建立统一的治理框架，同步制定基础数据标准、元数据标准等，贯穿指导数据中台的整个实施工作，数据治理活动与数据中台的技术架构、数据管理流程、业务应用等各个方面紧密相连，形成一个协同运作的整体，过程中通过不断迭代和优化，使得数据资产逐渐变得清晰、规范、有序，进而为数字化转型和业务发展提供坚实的数据基础。

- 智能研发，统一规范，固化流程

GBase 数据中台通过可视化数据开发管理平台 DolphinScheduler，帮助用户搭建高效的“数据生产线”，该平台包括统一数据交换、统一数据开发、统一作业调度、统一数据服务、统一运维管理等功能模块，实现三个统一：统一开

发模式，提供一致视角的规范遵从、任务部署、任务驱动、任务管控；统一治理模式，提供一致标准的贯标管理、资产管理、质量管控；统一服务模式，提供一致模式的服务监控、服务量化。

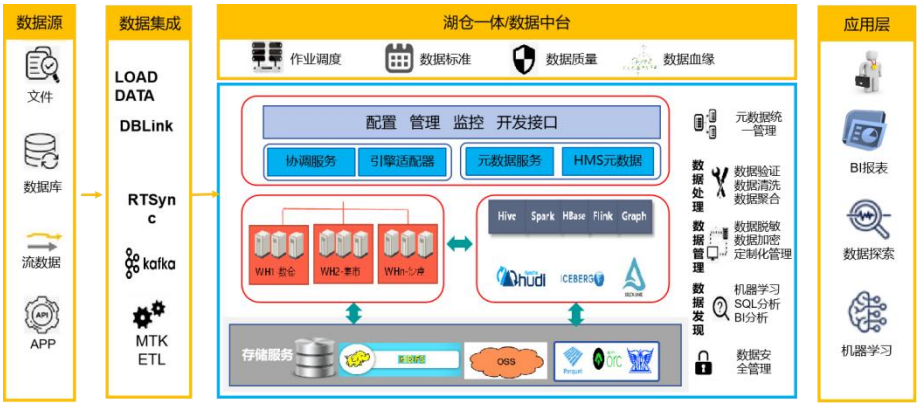
- 领域集市，让数据更贴近业务

基于 GBase 在金融、电信等行业的大规模应用实践经验，可以根据客户的业务管理领域，助力规划建设业务领域集市。各个集市遵循数据中台整体架构管控、遵从整体开发规范，基于数据开发管理平台相对独立地进行各领域的建设。通过数据集市建设，丰富了业务领域的相关数据，以维度建模方式建立易于业务理解、方便使用的宽表模型，从而降低数据使用的门槛，提高下游应用分析数据、开发应用效率的同时，又使得非技术人员也能更容易地进行数据分析和查询，满足多样化的用数需要。

3 GBASE 数据中台核心产品概述

3.1 GBASE 数据中台架构

南大通用数据中台以核心产品 GCDW 和 GBase HD 为基础，构建一个集数据湖与数据仓库为一体的数据基础平台，在具备了数据管理、数据处理与分析、数据安全、数据集成、数据存储等能力的同时，提供统一存储、统一元数据和统一的 SQL 任务调度服务接口，实现数据资产运营、数据开发管理等功能，最终形成一个完整的数据中台体系。



如上图，业务数据可以通过数据加载、kafka 实时同步、ETL 增量同步、DBLink 访问等方式流转到数据中台的存储服务中进行统一存储。GBase 数据中台提供统一的开发访问接口，调度 GCDW 计算引擎 WH 或者 Hadoop 生态计算引擎（Spark、Flink）对存储服务中的数据进行批量/流式计算，并将分析计算结果存储到存储服务中，供上层应用进行即席查询。此外，数据中台对存储服务中的所有业务数据提供数据安全管理、作业调度、数据标准管理、数据质量管理、数据血缘分析等数据管理与治理功能，助力用户发现数据价值，保障数据安全。

3.2 GCDW 云原生数据仓库

南大通用自主研发的南大通用云数据仓库 GBase Cloud Data Warehouse（简称 GCDW）是一款基于列存储的海量分布式大规模并行处理的弹性云数据仓库。南大通用云数据仓库计算和存储分离，适用于云上和云下环境，采用统一元数据服务架构，支持计算资源和存储资源的弹性扩展，适用于数据仓库、数据湖、大数据开发开放平台等应用场景，为用户提供海量数据的查询分析服务。

GBase Cloud Data Warehouse 支持公有云、私有云、虚拟机、物理机等部署环境，使用对象存储、HDFS 作为数据存储系统。GBase Cloud Data Warehouse 提供 SaaS 能力，为客户提供企业级弹性数据仓库系统，开箱即用，无需进行系统部署和优化；GCDW 解耦了计算资源和存储资源，实现了统一元数据管理和计算资源无状态，支持计算资源和存储资源独立部署、弹性扩展。用户通过 GBase Cloud Data Warehouse 提供统一服务管理接口进行数据库的计算资源、存储资

源管理。

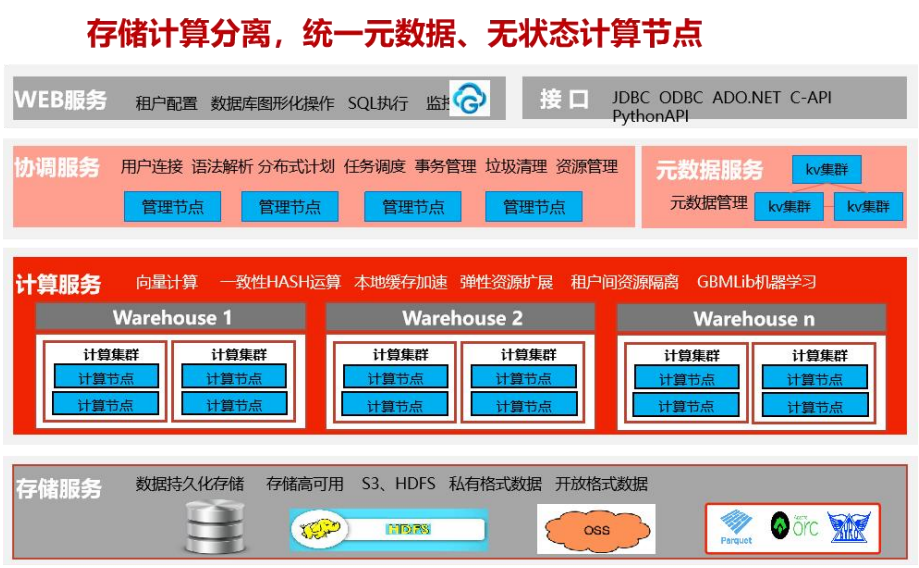
3.2.1 GCDW 云原生数据仓库应用场景

GBase Cloud Data Warehouse 在公有云上通过 SaaS 模式向用户提供企业级的数据仓库。用户仅需要具备云上账户，在云上根据自己的存储和计算需求选择某个规格的数据仓库，即可在其上加载用户自己的数据并进行数据查询和分析，用户之间资源完全隔离。同时，GBase Cloud Data Warehouse 具有弹性资源扩展能力，用户可以根据自身的需要随时弹性扩展计算资源或存储资源。

GBase Cloud Data Warehouse 是基于 GBase 8a MPP Cluster 专业打造的云上数据仓库管理系统，在 GBase 8a MPP Cluster 的功能基础上实现计算与存储分离，更加贴合云平台底层，让用户能够在云中更轻松地设置、操作，满足用户云环境下数据仓库需求。所以，GBase Cloud Data Warehouse 适用于分析类型的大数据平台、综合性 BI 系统、数据仓库和集市系统的云上系统。同时 GBase Cloud Data Warehouse 的算力弹性伸缩特性使其非常适用于业务规模（并发量、处理的数据量等）在不同时段变化较大的场景，能够在不同的时段根据业务需求自动匹配合适的资源，在节省用户成本的同时很好地保证用户业务的性能需求。GBase Cloud Data Warehouse 支持物理机部署，为已经购置了大量服务器的用户提供高性能的数据仓库系统，实现一份数据存储支撑用户所有的大数据分析应用。

3.2.2 GCDW 云原生数据仓库架构

GCDW 为存算分离设计架构的云原生数据仓库，在服务层分为存储服务、计算服务、协调服务、元数据服务、WEB 服务五层，并对外提供访问接口，架构图如下：



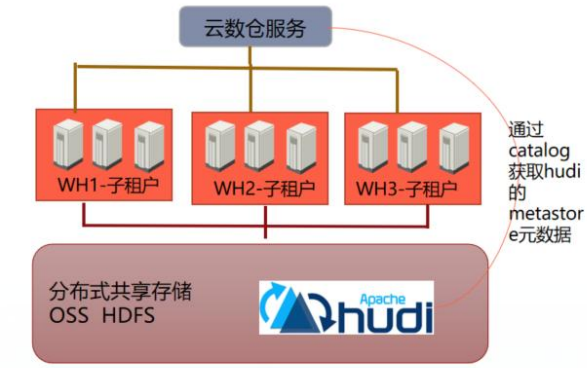
在存储服务层，GCDW 支持使用云端对象存储或者 HDFS 作为存储服务，以开放格式(ORC、Parquet)或者 GCDW 私有格式的形式将数据存储在存储服务中，实现完全的存储分离，用户的存储资源可以按需进行扩展。

在计算服务层，GCDW 将计算服务封装为多个 Warehouse 虚拟仓库，一个 WH 内可包含一个或多个计算集群，用户的 SQL 任务经协调服务解析后生成分布式执行计划下发给 WH 中的某个计算集群执行。可以理解为计算集群是执行 SQL 任务的最小单位。一个计算集群下可以包括多个计算节点，一个 WH 下可以包括多个计算集群，而一个 GCDW 实例又可以包括多个 WH。

在协调服务层，GCDW 通过多个管理节点来接收用户连接并进行语法解析、任务调度和分布式执行计划的生成，以及承担事务管理、垃圾清理、资源管理等功能。GCDW 的多个管理节点之间地位一致，无单点故障，一个管理节点在故障时，其任务可被其他管理节点接管。

在元数据服务层，GCDW 将管理节点和数据节点的元数据都单独使用一个完全解耦的 KV 集群进行管理，实现了管理节点和数据节点的无状态，为实现云原生特性提供了基础。此外，GCDW 提供了访问数据湖元数据的能力，能够通过 HMS 接入 Hudi/Hive 的 catalog，进而读取和访问数据湖中的表元数据，实现数据湖和数据仓库元数据的统一管理。用户可以通过 catalog 直接读取数据湖中的

表数据进行查询和关联分析，无需提前创建外部表。



在 WEB 服务层，GCDW 提供了可视化操作系统云际，用户通过访问云际可以便捷地进行租户配置、数据库管理、SQL 编辑与执行、GCDW 监控等操作，降低了 GCDW 操作的学习成本，提升了 GCDW 的易用性。

在对外接口上，GCDW 提供 ODBC、JDBC、ADO.NET、C API、Python API 和 TCL API 等通用接口，轻松访问和使用 GCDW。

3.2.3 GCDW 云原生数据仓库技术特点

GBase Cloud Data Warehouse 除了具备 GBase 8a MPP Cluster 的技术特点之外，同时还具有云上数据库高性能、高可用、弹性扩展、易管理等技术特点。

具体特点如下：

弹性扩展

GBase Cloud Data Warehouse 计算与存储分离，用户可以根据自身的需要随时弹性扩展计算单元或者存储单元。

- 云服务：支持主流云平台：阿里云、腾讯云、华为云等。
- 资源按需在线弹性扩展：
  - 云上存储资源对用户无容量限制，用户无需规划存储容量和扩容存储硬件，只需按实际需要存储数据即可；

- 计算资源无状态，秒级扩容，支持弹性扩展。计算集群根据任务负载情况（出现排队任务时），在最大计算集群数范围内自动启动同等规格的计算集群，承担新任务的执行；任务负载降低后，自动关闭扩展的计算资源，保留最小计算集群个数。用户按实际使用的资源付费，使用户在满足性能需求的同时高效使用所需资源。

多租户，资源管理

多租户，租户间资源隔离，租户内通过不同的 warehouse 计算资源负载不同的业务，支持按需申请计算资源，彻底解决传统数仓混合负载问题。

一份数据，数据无需冗余

用户数据以 GBase 自研格式存储在对象存储上，由第三方对象存储服务商提供存储服务，无需冗余存储。

元数据存储在 KV 数据库集群上，统一管理，无需协调服务在各协调节点上多份存储及同步管理。

数据集成，多源数据加载

数据加载具与集群高度集成，面向用户的 SQL 接口方式更符合用户的使用习惯。支持单表多数据源并行加载，支持多加载机对单表的并行加载，最大化提升加载性能。

多源数据加载支持从通用数据服务器拉取数据，支持 ftp、sftp、hdfs、S3、http、https 等多种文件传输协议；

高性能

- 数据文件多版本管理机制：数据文件的多版本管理机制使读事务不需要与写事务互斥就可以安全读取数据文件，读写事务并发时 MVCC 控制各事务对数据的版本可见性，有效提升了并发读写的性能。

- 多级缓存机制：直接从云存储读取数据受限于网络带宽，GCDW 计算服务节点设计了两级缓存预读数据，以提升数据读写速度。业务程序启动

时即将业务相关所有数据从云存储调入到二级缓存 SSD 中，将程序实时需要的数据装入一级缓存中。上层应用需要的数据查找顺序是：一级缓存、二级缓存、云存储。

高可用

- 数据高可用：GCDW 的数据保存在第三方云存储上，当前云存储普遍可提供 99.999%的数据可靠性，GCDW 由云存储保证数据的高可用和完整性。
- 服务高可用：GCDW 的所有服务都是多节点分布式部署运行，有节点服务故障时，会即时采用新的节点自动替换原故障节点并恢复到故障前的节点状态。

可靠的数据安全

- 完善的用户认证及权限管理：提供完善的用户、角色和账号控制策略，保证集群数据库访问的安全性。
- 丰富的加密函数：支持多种加密函数，如 AES_ENCRYPT、ENCRYPT、MD5、SHA1、SHA、SHA256、SM4 等。
- 动态数据脱敏：支持默认脱敏、随机脱敏、自定义脱敏、哈希脱敏和指定位置脱敏五种数据脱敏函数。

完备的 SQL 标准化支持

- 支持 SQL92 ANSI/ISO、SQL99、SQL2003 标准；
- 支持 ODBC、JDBC、ADO.NET、C API、Python API 和 TCL API 等接口；

易管理

GBase Cloud Data Warehouse 的云上服务不需要用户安装部署，GCDW 提供了云际（GCDW 客户端）可视化操作界面，只需要在云际选择需要的规模就可以获取相应的服务。

湖仓融合

- 统一存储：支持和数据湖共用一个存储底座（S3、HDFS），支持配置使用多个存储系统实现存储资源的隔离和扩展，支持快速读写数据湖上的开放格式（ORC、Parquet）数据。

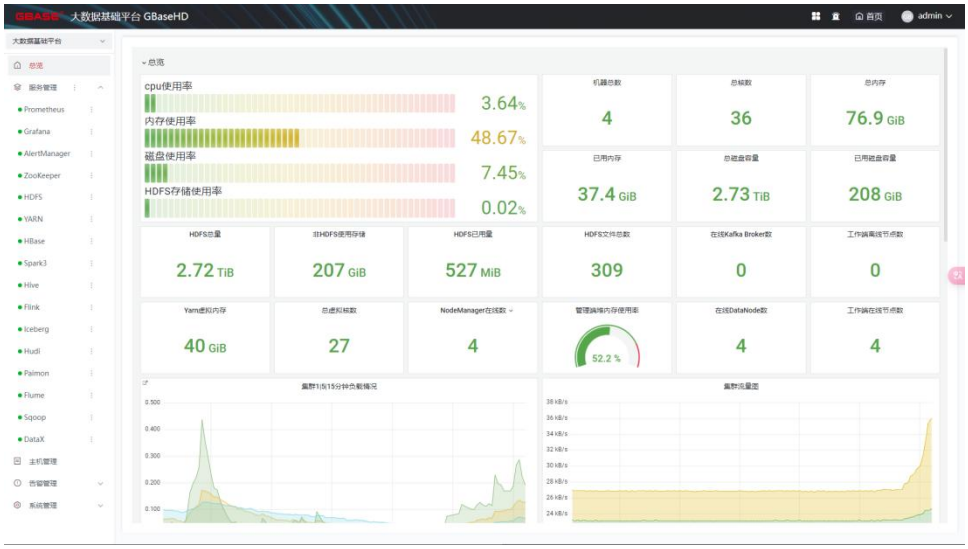
- 统一元数据：支持通过 catalog 对接数据湖的元数据，将其作为 GCDW 元数据服务的一个扩展。用户通过 GCDW 即可访问和统一管理数据湖及数据仓库 GCDW 的元数据，同时依赖数据库本身的 ACID 特性实现事务管理。

- 统一计算：对用户提供统一的任务调度，通过库内的统一元数据视图可以提供一致性事务保障，封装不同引擎的调度执行方法对用户屏蔽引擎差异。提供针对不同计算引擎的方言转换器，用户通过 SQL 语句下发任务，操作不同的计算引擎执行；SQL 方式简化了业务，屏蔽底层架构差异。复杂 NoSQL 计算，无法在业务方面使用 SQL 完全表达，直接方言转发给对应计算引擎执行。

- 统一接口及权限：使用 GCDW 的开发接口、用户权限体系作为仓湖的接口和权限管理。对用户提供统一的开发接口，内部使用各计算引擎的接口进行引擎对接；使用 GCDW 的 RBAC 权限体系进行仓湖的权限控制。

3.3 GBase HD 数据湖

GBaseHD 是南大通用基于开源 Hadoop 构建的一款强大的大数据处理平台，完美地融入了 Hadoop 生态系统，能够高效地完成大规模异构数据的处理和分析。无论是实现数据的快速入湖建仓，还是进行深入的数据分析建模，抑或是构建宽表以满足数据市场访问需求，GBaseHD 均能提供全方位的支持。它同时支持离线和实时两种数据处理模式，不仅可以帮助用户建立高效的离线数仓，还能助力构建响应迅速的实时数仓，从而满足各种数据应用的需求。



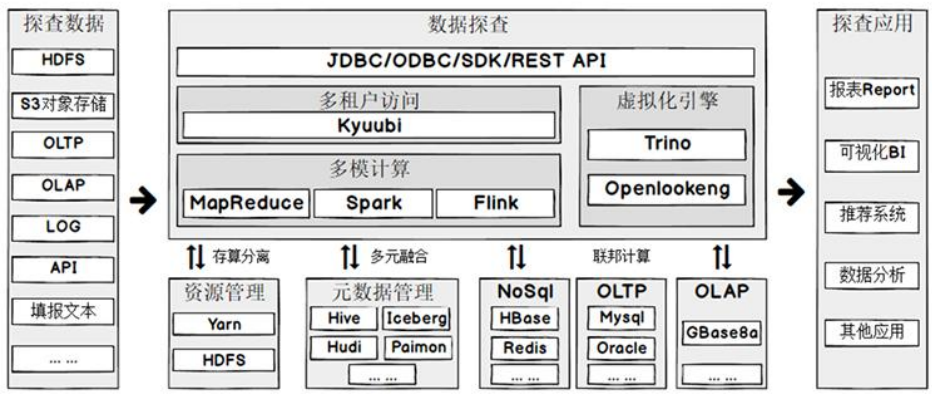
GBase HD 作为南大通用数据中台中湖组件，为 GCDW 提供 HDFS 存储服务和非结构化数据计算服务。

3.3.1 GBase HD 应用场景

GBase HD 基于数据湖强大的生态能力，具有多样的应用场景：

- 数据湖探查，存算分离、多模计算、跨源分析

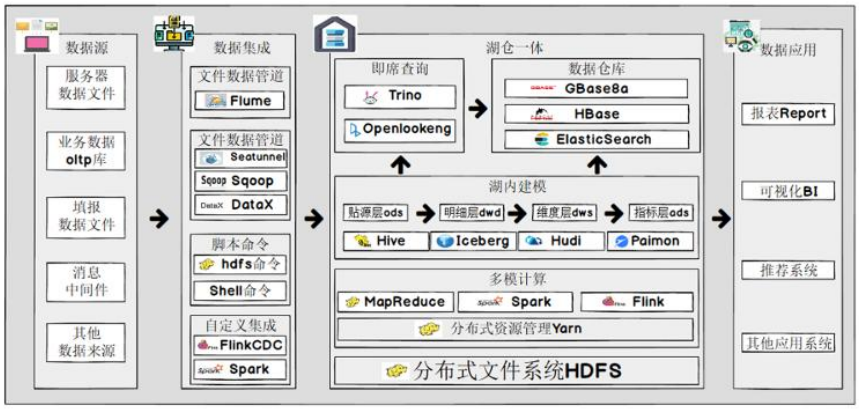
GBaseHD 依托多元计算引擎 Hadoop-MapReduce、Spark、Flink、可选的虚拟化计算引擎 Trino/Openlookeng（Base Presto），提供一站式的批计算、流计算、联邦查询等多模计算、融合分析的数据湖探查。



- 离线数仓，湖仓一体，离线分析

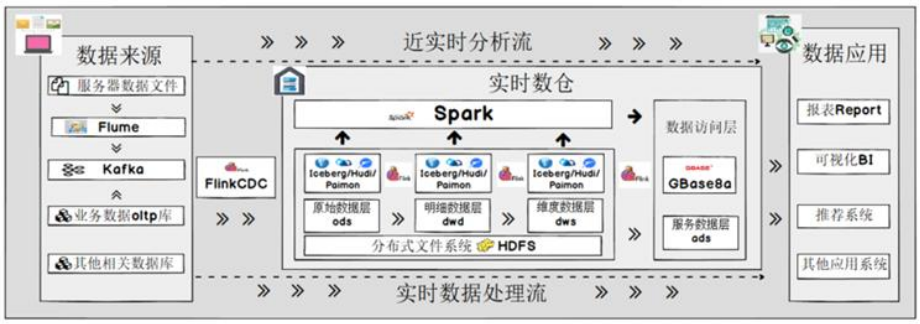
离线使用场景：T+1 离线数据分析场景：面向大批量离散化数据，分析建模提取构建关键指标数据，服务面向报表、BI、数据 API、数据挖掘等。

分析数据来源：OLTP 业务库（GBase8s、Mysql、Oracle 等）+ 服务器数据文件 + 填报数据文件（excel、txt、csv 等）+ 消息中间件 。



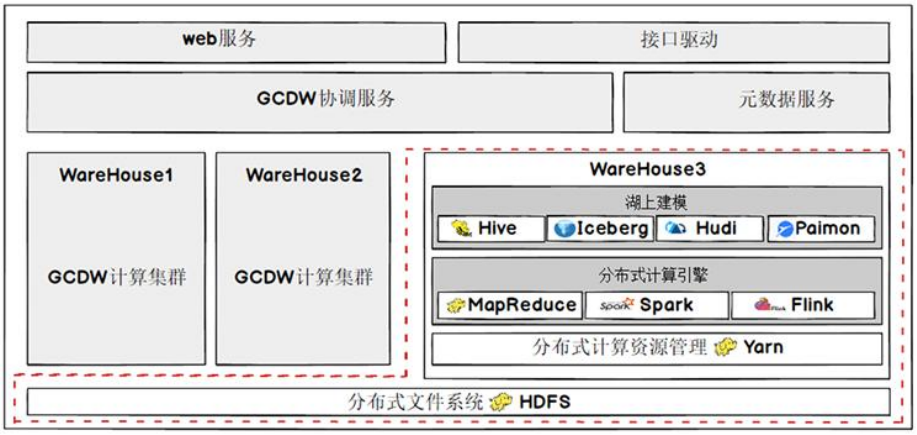
- 实时数仓，流批一体，实时计算

使用场景，实时/近实时数据分析场景，对数据分析结果时效性要求高的场景可以使用，比如实时推荐系统、实时数据 BI 看板，实时报表查看等场景



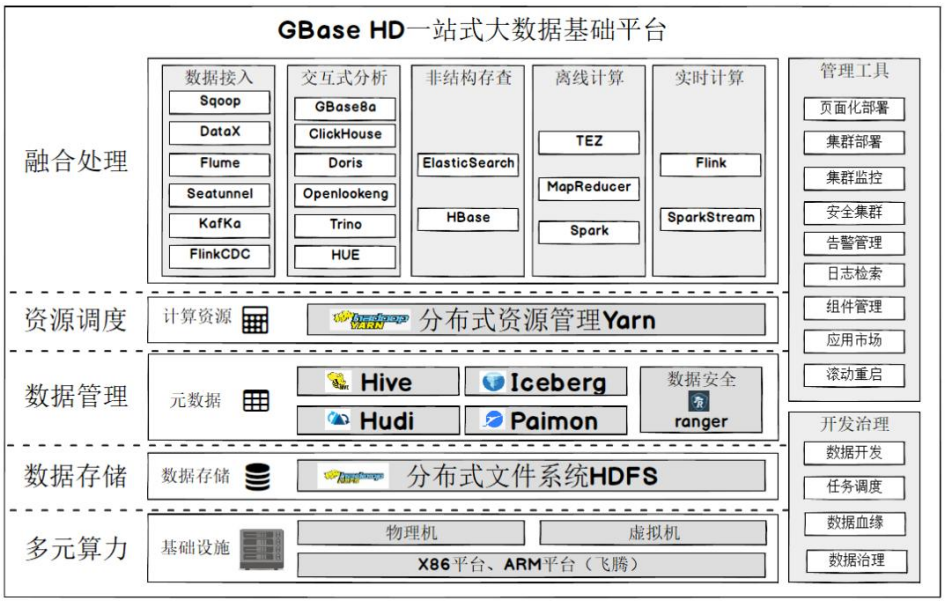
- 仓湖一体，仓下拓湖，简化操作，高效分析

仓湖一体：将 Hadoop 集群的数据处理能力，通过虚拟化技术，映射到 GCDW 上。在 GCDW 设计架构下，Hadoop 数据湖中的数据模式由 GCDW 负责全面管理。通过利用 GCDW 提供的强大 OLAP 分析执行客户端，实现了对数据湖和数据仓的统一联邦分析能力，极大地增强了数据分析的灵活性和效率。



3.3.2 GBase HD 架构

GBase HD 具备适配各种信创基础设施、统一存储、多元计算引擎等特性，架构图如下：



GBase HD 一站式大数据基础平台涵盖了多个维度的能力，包括数据存储、数据管理、资源调度、融合处理和平台管理等。在融合了 Hadoop 生态存储和计算引擎的基础上，提供了数据存、算、管、用的一系列能力，并能对数据进行分析治理，同时通过平台管理工具为 GBase HD 稳定平稳运行提供了有力保障。

3.3.3 GBase HD 技术特点

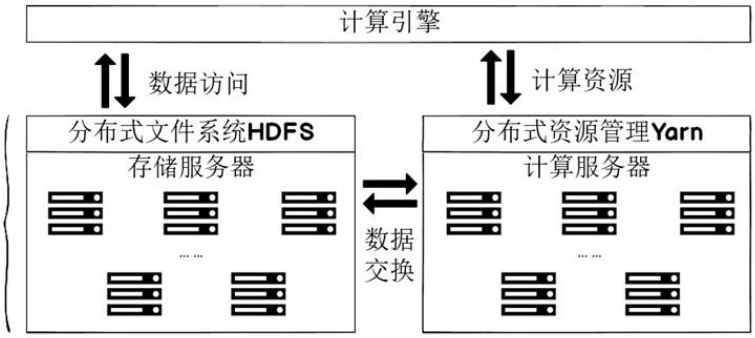
GBase HD 具备存算分离、流批一体处理、资源隔离、统一调度等多种技术特点。具体特点如下：

- 存算分离，统一存储，统一计算资源

存算分离：存储机器高存储高吞吐，计算机器高内存高带宽 ，存储节点和计算节点分离部署；

统一存储：提供分布式文件系统 HDFS，高可用，多副本；

统一计算资源：分布式资源管理 Yarn，高可用，多队列，多租户。



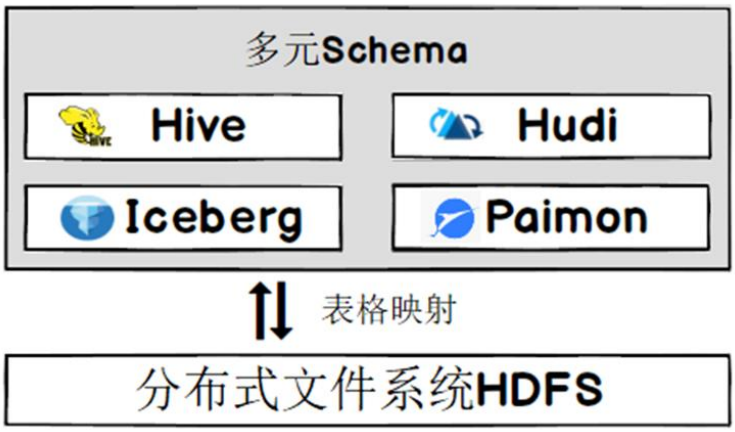
- 多模计算，流批一体，联邦计算

多模计算，流批一体：提供基于磁盘的离线计算引擎 MapReduce，基于内存微批次计算引擎 spark，基于内存的实时流计算引擎 flink。

联邦查询：提供基于 Presto，深度开发的联邦查询引擎 Trino、Openlookeng。

- 湖上表格，多元格式，场景丰富

湖上 schema 提供了面向传统稳定数仓的 hive、面向湖格式支持 ACID 和多版本控制的 Iceberg、面向低延时，近实时的 Hudi、同时提供了增加数据湖+LSM 的 Paimon。 提供了四种湖上表格范式，支撑传统湖仓、实时湖仓等方案的实施。



- 租户管理，存储多租户、资源多租户，计算多租户

  存储多租户：

  - 存储资源是分布式文件存储服务 HDFS 中可分配的数据存储空间资源。
  - 目录是 HDFS 存储资源分配的基本单位，租户通过指定 HDFS 文件系统的目录来获取存储资源。

  资源多租户：

  - 计算资源是分布式资源管理服务 Yarn 中可分配的计算资源空间。
  - 执行队列是 Yarn 计算资源分配的基本单位，租户通过定义不同的队列来分割租户资源。

  计算任务多租户：

  - 计算任务针对 spark、flink 计算引擎之上提供 kyuubi 服务，提供 JDBC、ODBC 访问方式。
  - 实现对 spark、flink 计算引擎提供多租户并行访问，租户间保证资源隔离

- 统一开发，统一编排，统一调度

数据开发调度平台：为 GBaseHD 提供的大数据生态系统增添了强大的数据调度和开发功能。通过提供直观的界面化和低代码开发方式，该平台极大地简化了数据开发流程，使得大数据的管理和应用更加高效、便捷。

3.4 GBASE 数据中台工具

南大通用数据中台提供标准开发接口，支持与 ETL、BI、数据建模等各类第三方工具的对接集成。此外，南大通用提供集群运维管理平台、数据同步工具、任务编排调度工具等。

3.4.1 GDOM 运维管理平台

南大通用 GBase 数据库运维管理系统（简称：GDOM），它是 GBase 自主研发的、专门为 GBase 8a 系列产品量身打造的企业级运维管理平台，旨在为 GBase 8a 系列产品提供全生命周期的运维保障，在提供可视化监控的同时，通过集群管理、主机管理、健康检查、告警等一系列功能，降低客户运维成本，提高客户运维效率，实时保障集群 7*24 正常运行。

GDOM 具备以下技术特点：

- 灵活易用的集群管理能力：支持根据任务配置向导可视化地对集群进行管理操作，包括集群安装、集群扩容、集群升级、节点替换等常用操作
- 丰富的监控指标：内置 140 余项告警指标，全面实时掌握集群的运行状态，在指标异常时生成告警信息，及时感知集群异常

- 深度健康检查：内置涉及 6 个维度的健康检查项，可自由配置健康检查任务并进行定时调度，生成健康检查报告，全面掌握集群、运行环境以及节点情况，保障集群正常运行
- 用户体验良好：采用业界流行的 Vue 框架，前端控件样式精美，交互设计贴近用户体验，并内置操作反馈，帮助用户低学习成本使用各项功能
- 高可用：支持全面高可用部署，包括后端服务高可用、资源库高可用以及 Agent 高可用，保障 GDOM7*24 小时稳定运行
- 快速响应：底层双资源库架构并引入 Redis 缓存组件，显著提升前端页面响应速度，统计性图表、列表快速生成无需等待

3.4.2 RTSync 数据同步工具

南大通用实时同步系统，简称：GBase RTSync，它是一款自主研发的异构及同构数据库增量数据实时同步产品，具备实时性、一致性、精准性、易扩展性和可集成特性，适用于 OLTP 数据库与 OLAP 数据库联动向应用系统提供数据管理和数据分析功能的业务场景，可以实现将 OLTP 数据库的数据实时同步到 OLAP 数据库，从而使得 OLAP 数据库具备了实时数据分析的基础，解决数据增量同步问题，能够有效提升数据仓库系统、BI 系统和决策支持系统的数据分析效率和及时性

GBase RTSync 具备以下技术特点：

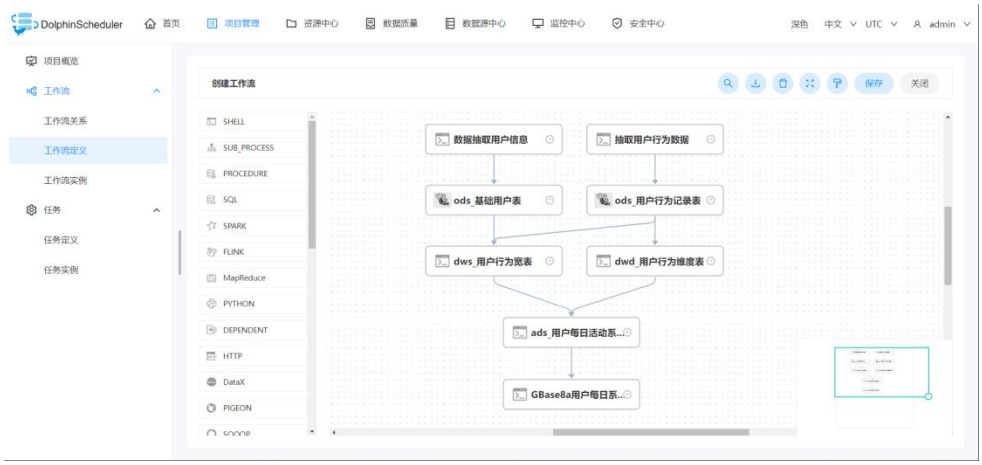
- 实时性：GBase RTSync 通过流模式实现源数据库到目标数据库的实时数据同步，确保上层业务系统能够通过目标数据库获取实时业务数据；
- 灵活性：既支持独立部署，完成从源数据库到目标数据库的数据同步； 也支持部署部分组件，完成从源数据库捕获增量数据到消息中间件的功能，并实现与第三方同步系统的集成；同时，可根据业务需要，既可以实现整库级别的增量数据同步，也可以实现表级增量数据同步，甚至可以实现字段级数据同步；
- 支持断点续传：GBase RTSync 支持断点续传功能，当出现网络异常或者是程序异常的情况下，在网络恢复以后或者程序重新启动以后， 能够捕获到出现异常时的断点，并从该点继续执行数据同步，从而确保同步到目标数据库的数据能够保持与源数据库一致，避免了目标数据库中数据重复或数据丢失情况的出现；

3.4.3 DolphinScheduler 任务编排调度工具

DolphinScheduler 是一个分布式易扩展的可视化 DAG 工作流任务调度系统，适用于企业级场景，提供了一个可视化操作任务，工作流和全生命周期数

据处理过程的解决方案，为 GBaseHD 提供的大数据生态系统增添了强大的数据调度和开发功能。

通过提供直观的界面化和低代码开发方式，该平台极大地简化了数据开发流程，使得大数据的管理和应用更加高效、便捷。



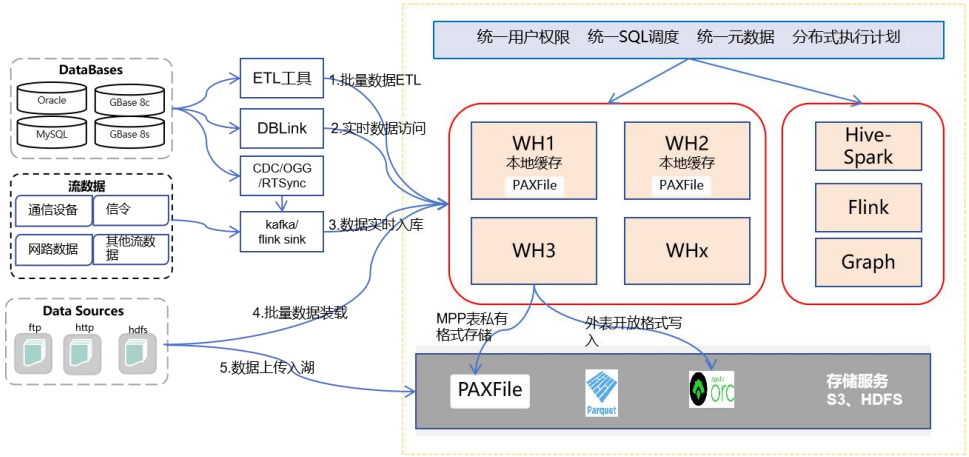
DolphinScheduler 具备以下技术特点：

- 高可用性：去中心化的多 Master 和多 Worker 服务对等架构设计，避免单 Master 压力过大，同时采用任务缓冲队列避免过载，确保稳定性
- 易用性：DAG 监控页面，所有流程定义都是可视化，通过拖拽任务完成定制 DAG，通过 API 方式与第三方系统集成，一键部署
- 支持多种任务类型：支持 Shell、MR、Spark、Hive、Python 等近 20 种任务类型
- 高扩展性：支持自定义任务类型，调度器使用分布式调度，调度能力随集群线性增长，Master 和 Worker 支持动态上下线

4 GBASE 数据中台建设能力

4.1 数据集成能力

GBase 数据中台支持流式数据加载、批量数据加载、实时数据入库等多种数据接入方式，支持通过外部表功能将数据存储到开放格式数据文件中，统一使用 GCDW 接口实现全业务数据的接入。



- ETL 工具批量迁移其他数据库中数据
- DBLINK 实时访问其他数据库，进行数据的交换
- CDC 软件实时解析 OLTP 数据库的事务日志获取增量数据，之后通过 Kafka 或 Flink 入库到 GCDW 中
- 流式数据写入到 Kaka/Flink 中，再实时入库到 GCDW
- 数据文件批量加载入库
- 复用数据湖的数据集成能力

4.2 数据存储与管理能力

GBase 数据中台支持使用对象存储、HDFS 等数据湖存储作为存储底座，即可存储结构化数据文件，也可以存储非结构化数据（图片、语言、视频等），实现企业内所有数据存储的统一管理。

在进行存储时，支持将数据存储为数据湖生态的开放格式数据（ORC、Parquet），也支持将数据存储为 GCDW 私有格式数据进而提升性能。在使用 GCDW 私有格式存储数据时，支持设置高效压缩比以节省存储资源。

GBase 数据中台将元数据单独存放在 K/V 集群中进行管理，实现了元数据与数据的解耦，从而实现在 K8s 云平台上进行基础组件的部署和弹性伸缩。

4.3 数据处理与分析能力

GBase 数据中台在数据处理和分析能力上同时具备数据湖和数据仓库的特

性，融合了 GBase 的 MPP 计算引擎和 Hadoop 生态计算引擎（MR、Spark、Flink、Presto 等），用户可按需调度进行计算。

- 离线计算：对于结构化数据，可调用 GCDW 计算引擎使用 WH 直接对数据湖中的开发格式数据进行高性能计算，基于 MPP 引擎，相比于 Hadoop 生态计算引擎可大幅提升计算性能，提高计算效率。
- 实时计算：提供统一的访问调度接口，调度流式计算引擎 Spark streaming/Flink 对实时数据进行高速计算，满足业务分钟级甚至秒级的计算需求。
- 机器学习：融合 Spark 机器学习算法、MPP GBMLlib 实现 In-Database Analysis
- 多级缓存：对象存储扩展性强、成本低、吞吐量高，但是相对于本地存储对象存储也有网络访问延时高的问题,GCDW 采用了二级缓存和数据预读的技术，抵消了网络访问的延时，大幅提升了数据访问的性能。
  - 计算节点接收任务包后，解析出需要读写的库表列表在计算节点本地构建虚拟表，同时后台异步从对象存储预读数据到计算节点。计算节点从对象存储以 PCFile 文件为单位预先读取到本地后进行两级缓存保存。从对象存储预读的数据先写入一级缓存，一级缓存中的冷数据下沉到二级缓存（通常为 SSD 介质）中存储。读文件时判断二级缓存使用率，超过 70%后台会自动开启清理工作。

4.4 数据安全能力

GBase 数据中台提供完善可靠的数据安全机制。

- 完善的用户认证及权限管理:提供基于 RBAC 的用户、角色和账号控制策略，保证数据访问的安全性。
  - 支持用户密码安全策略管控及用户黑白名单配置
- 丰富的加密函数：支持多种加密函数，如 AES_ENCRYPT、ENCRYPT、MD5、SHA1、SHA、SHA256、SM4 等。
- 支持动态数据脱敏：支持默认脱敏、随机脱敏、自定义脱敏、哈希脱敏和

指定位置脱敏五种数据脱敏函数。能够在访问表数据时对数据进行脱敏。

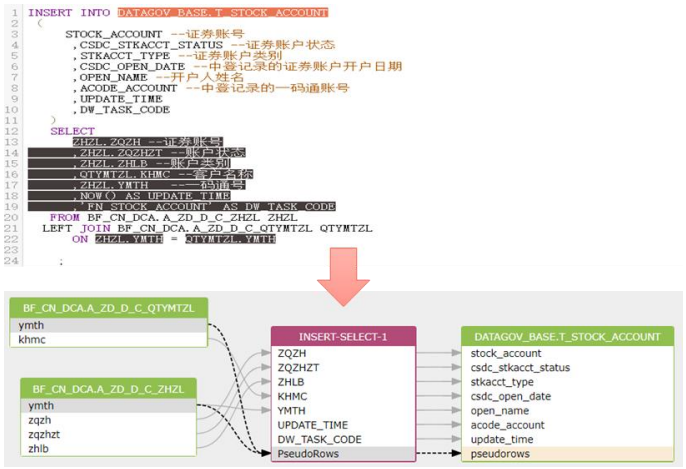
- 高效透明的数据存储加密：支持将数据加密存储落盘到文件中，在访问数据时自动进行加密/解密，同时加解密负载对整体性能影响小于 5%
- 审计策略及审计日志：支持记录 DDL、DML、DQL，持久化审计日志用于审计数据安全。
- 支持高性能的在线备份恢复，为数据文件提供全方位的安全保障

4.5 数据服务能力

GBase 数据中台对内对外均提供良好的数据服务能力。对于私有格式数据，提供租户间访问数据能力。各租户间通过为需要访问的数据创建 sharedata 来实现数据共享。对于开放格式数据，其他业务系统数据可以直接进行读写操作。

在外部业务系统访问数据上，GBase 数据中台支持以下特性：

- 支持通过 gccli 访问，也支持通过 ODBC/JDBC/ADO.NET/C API/Python API 各种接口进行访问
- 支持统一的访问接口和调度，在访问调度 Hadoop 生态引擎时，通过 SQL 方言转换为其他计算引擎下发任务
- 支持统一元数据管理和访问能力。能够通过 catalog 对接数据湖的元数据并将其缓存到 GCDW 中，用户可以直接通过 GCDW 查询数据湖中的库表结构并进行读写操作。
- 支持读写高并发及高并发扩展能力。
  - 数据文件通过 MVCC 多版本进行管理，实现读写并发
  - 自由配置负载策略，当并发业务数量达到一定数量级后，弹性增加计算集群承载业务访问。
- 管理节点服务高可用，多个管理节点提供访问服务，且可配置负载均衡访问，内部通过 failover 机制时刻保证对外访问服务的高可用。
- 内置血缘分析功能，基于 AST 抽象语法树实现，描述数据的来源和去向，以图的形式展现血缘关系



4.6 资源隔离能力

GBase 数据中台具备良好的存储、计算资源物理隔离能力。

- 存储资源隔离：支持各租户使用不同的存储系统进行存储资源隔离；支持单租户内为不同的业务库设置不同的存储系统以实现存储资源隔离
- 计算资源隔离：支持为各租户分配不同的互相完全隔离的计算资源，以计算节点为最小单位进行分配。计算节点间的资源隔离由 Paas 平台实现。

5 GBASE 数据中台建设优势

5.1 技术优势

GBase 基于自身在大数据行业多年的技术积累所研发的数据中台具备以下技术优势：

- 存算分离，存储、计算资源独立扩展，避免资源浪费的同时能够利用 HDFS、S3 等服务统一存储格式化、非格式化数据。
- 性能强劲，采用多级缓存预读机制和优化的私有存储格式，基于 MPP 引擎和 GBase 8a 在各行业优化改进多年的分布式执行计划，以高性能完成大数据量的跑批计算和即席查询
- 湖仓兼容，既支持通过 catalog 直接访问数据湖的元数据，也支持通过 SQL 方言转换等方式调度 Hadoop 生态的其他计算引擎对数据湖中的数据进行计算
- 高可用，无论是管理节点还是数据节点，基于 GBase 8a 在各行业多年的生产实践，提供了强有力的高可用特性，内部通过 failover 机制结合 K8s 平台的调度能力，保障系统 7*24 小时稳定运行
- 资源弹性伸缩，利用 K8s 平台基础能力，将计算节点资源封装为 pod，实现计算资源的快速弹性伸缩，满足业务潮汐变化需求。

5.2 行业经验优势

GBase 在金融、电信、政企、安全等行业具有广泛的实践案例，部署节点数>96000 个，总数据量>500PB，覆盖各行各业核心业务系统，积累了大量的行业经验。

基于行业经验，GBase 能帮助用户挖掘数据中台建设的核心需求，结合行业内数据中台的建设案例，综合考量建设数据中台的必要性和建设价值，并帮助企业按照需求进行统筹规划和考虑，分多阶段逐步完成各项能力建设。

此外，GBase 拥有强大的大数据技术团队，技术骨干均拥有 10 年以上数据库产品经验，核心研发成员拥有 5 年以上数据库产品开发经验，在全国各省市均配备本地化服务团队和原厂技术服务支持，7*24 小时响应客户需求，极大保

障系统的稳定运行。

5.3 信创兼容优势

GBase 数据中台全面兼容信创生态，从国产芯片、操作系统、服务器再到数据上下游应用工具和开发平台，GBase 均进行了充足和完善的适配工作，支持用户打造全栈化国产的数据中台体系。

GBase 生态支持情况：

6 GBASE 数据中台运行环境和技术指标

6.1 运行环境要求

GBase 数据中台对云上资源要求如下：

- Kubernetes 集群硬件配置要求：

硬件	配置要求
CPU	64 核以上
内存	128G 以上
磁盘	1T SSD/Nvme 以上
网络	容器网络使用内部数据传输的万兆网络

- Kubernetes 集群软件配置要求：

软件	配置要求
版本	建议 Kubernetes1.20 及以上版本
权限	使用 Kubernetes 管理员权限的账户进行安装 （必须满足权限要求）
基础功能要求	支持 kubectl 命令操作 kubernetes 的资源 支持 helm 模板安装（helm V3）

GBase 数据中台对物理机软、硬件环境要求如下：

- 硬件环境：
  - 服务器平台：x86_64 的标准 PC 服务器、PowerLinux 服务器、浪潮 K1、华为泰山、中科曙光、海光等；
  - 存储：本地存储（SATA、SAS、SSD 等）、阵列存储（SAN、NAS）、SSD、Flash 卡；
  - 网络环境：千兆、万兆、InfiniBand
- 操作系统和平台：
  - CPU：Intel、AMD、申威、龙芯 3B、飞腾、X86、Power、鲲鹏、海光等；

- 操作系统：CentOS、Red Hat、SUSE、中标麒麟、PowerLinux、深度、银河麒麟、凝思、中科方德、普华等 64 位操作系统；
- 基于 x86 及 Power 的虚拟机，如 VMware ESX 及 KVM、OpenStack、docker 等虚拟化技术。

6.2 技术指标

技术指标	描　述
数据容量	无上限
数字精度	65
表的个数	无限制；单租户支持的表数量>1000 万
每个表中列的个数	2000
每个表中行的个数	140737488355328
表中一行的内部长度	65534000 字节
一个 INTEGER 类型列的长度	4 字节
日期类型列中表示年的位数	4 位
用户名包含字符的个数	16 字符
CHAR 类型列的长度	255 字符
BLOB 列的长度	32K 字节
VARCHAR 类型列长度	随字符集而不同，UTF8MB4、GB18030 是 8192，GBK、UTF-8 为 10922 字符
数据库名长度	48 字符
表名长度	56 字符
列名长度	64 字符
索引名长度	64 字符
别名长度	255 字符
编码格式	UTF-8、UTF8MB4、GBK 、GB18030
Warehouse 规格	选项如下 X-Small:1 Small:2 Medium:4 Large:8 X-Large:16

	2X-Large:32 3X-Large:64 4X-Large:128
单节点配置	由用户自行定义，推荐最小配置为 8core、16GB、100GB 本地磁盘作为二级缓存
单 warehouse 中计算集群个数	1~10 个
单个计算集群支持的并发数	200，可配置，通过扩展新的计算集群提升并发量；