

星环一站式大数据平台 Transwarp Data Hub

2020年10月

星环信息科技（上海）有限公司

目录

CONTENTS

● 背景介绍

● 产品介绍

背景介绍

越来越复杂的数据环境

	传统数据	大数据
数据量	GB → TB	TB → PB以上
速度	数据量稳定、增长不快	实时产生处理、年增长率超60%
多样性	结构化数据	结构化、半结构化、非结构化数据
价值	统计报表	机器学习、深度学习

大数据是指超出传统数据库工具收集、存储、管理和分析能力的数据集。与此同时，及时采集、存储、聚合、管理数据，以及对数据深度分析的新技术和新能力，正在快速增长，就像预测计算芯片增长速度的摩尔定律一样。

— McKinsey Global Institute

- ✓ 数据规模巨大 (Volume)
- ✓ 数据类型多样 (Variety)
- ✓ 生成和处理速度极快 (Velocity)
- ✓ 价值巨大但密度较低 (Value)

越来越多样的业务需求

离线、在线与实时业务**并存**

检索型、分析型与智能型业务**并存**



结构化、半结构化与非结构化数据**并存**

对事务支持的**需求**

大数据环境下传统平台面临的挑战

存储管理能力不足

无法支撑海量多源异构数据的灵活高效存储
无法实现基于SQL的异构数据统一管理和访问

.....



综合搜索能力不足

无法实现PB级半/非结构化数据的组合、全文和语义搜索
无法实现千亿级数据搜索的秒级返回

.....



分析挖掘能力不足

计算任务井喷式增长，系统不堪重负
无法支撑PB级异构数据的快速分析和深度挖掘

.....



实时处理能力不足

无法实现流式数据的实时接入、复杂事件处理和机器学习
开发门槛高，不支持用SQL编写流应用

.....



产品介绍

星环一站式大数据平台 Transwarp Data Hub

Transwarp Studio



Waterdrop
SQL开发工具



Transporter
ETL/实时同步



Workflow
工作流引擎



Data Catalog
数据资产目录



Rubik
Data Cube设计



Pilot
BI/报表工具



Stream studio
流任务管理



Notebook
机器学习编程



Transwarp
Guardian

安全及
权限控制

Slipstream



事件驱动
SQL编程

实时流计算引擎
Realtime Streaming Engine

Inceptor



批处理
数据仓库

分析型数据库
Analytical Database

ArgoDB



数据仓库
数据集市

Hyperbase



NewSQL
在线数据库

操作型数据库
Operational Database

Search



搜索引擎
时空数据库

知识库
Knowledge Database

StellarDB



图数据库
图分析

Discover



数据挖掘
机器学习

数据科学平台
Data Science Platform



Transwarp
Manager

安装
监控
运维
管理

Transwarp Cloud Operating System (Embedded Edition)
容器化集群操作系统 (计算、内存、存储、网络资源调度)



Transwarp Proprietary

TRANSWARP
DATA HUB

Open Source

TDH产品定位



基于TDH构建产品解决方案

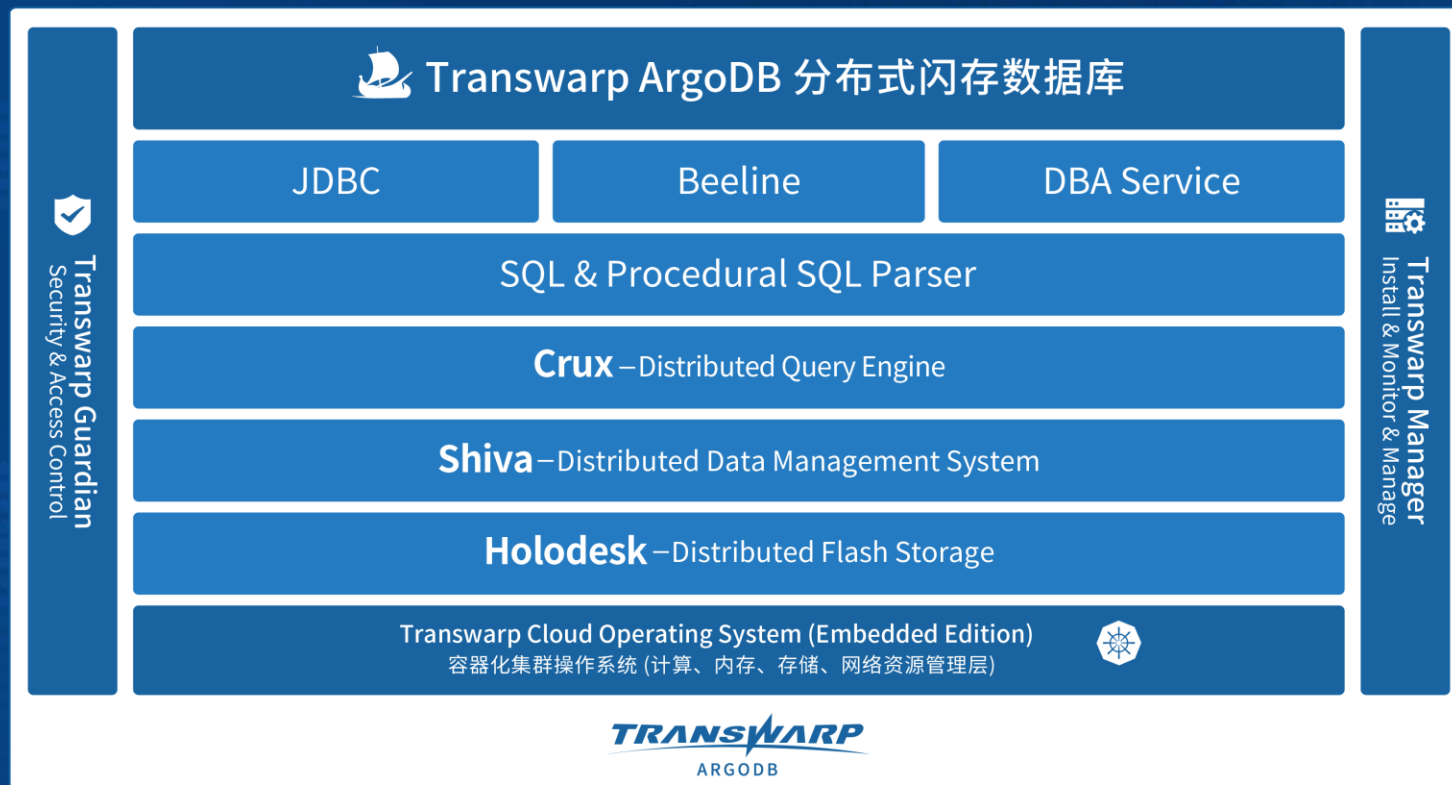



世界领先的高性能分析型数据库 & SQL引擎 Inceptor





- SQL支持最完整
 - ✓ 支持SQL 99/2003标准
 - ✓ 支持存储过程, 包括Oracle PL/SQL、DB2 SQL PL
 - ✓ 支持Oracle、DB2、Teradata等数据库方言
- 首个支持分布式事务处理
- 多种优化策略
- 混合负载均衡管理
- 细粒度调度算法SLA
- 数据联邦
- 数据稽核
- 多种数据格式
- 多种数据加载方式
- 海量多源异构数据的统一存储和管理

世界领先的分布式闪存数据库 ArgoDB



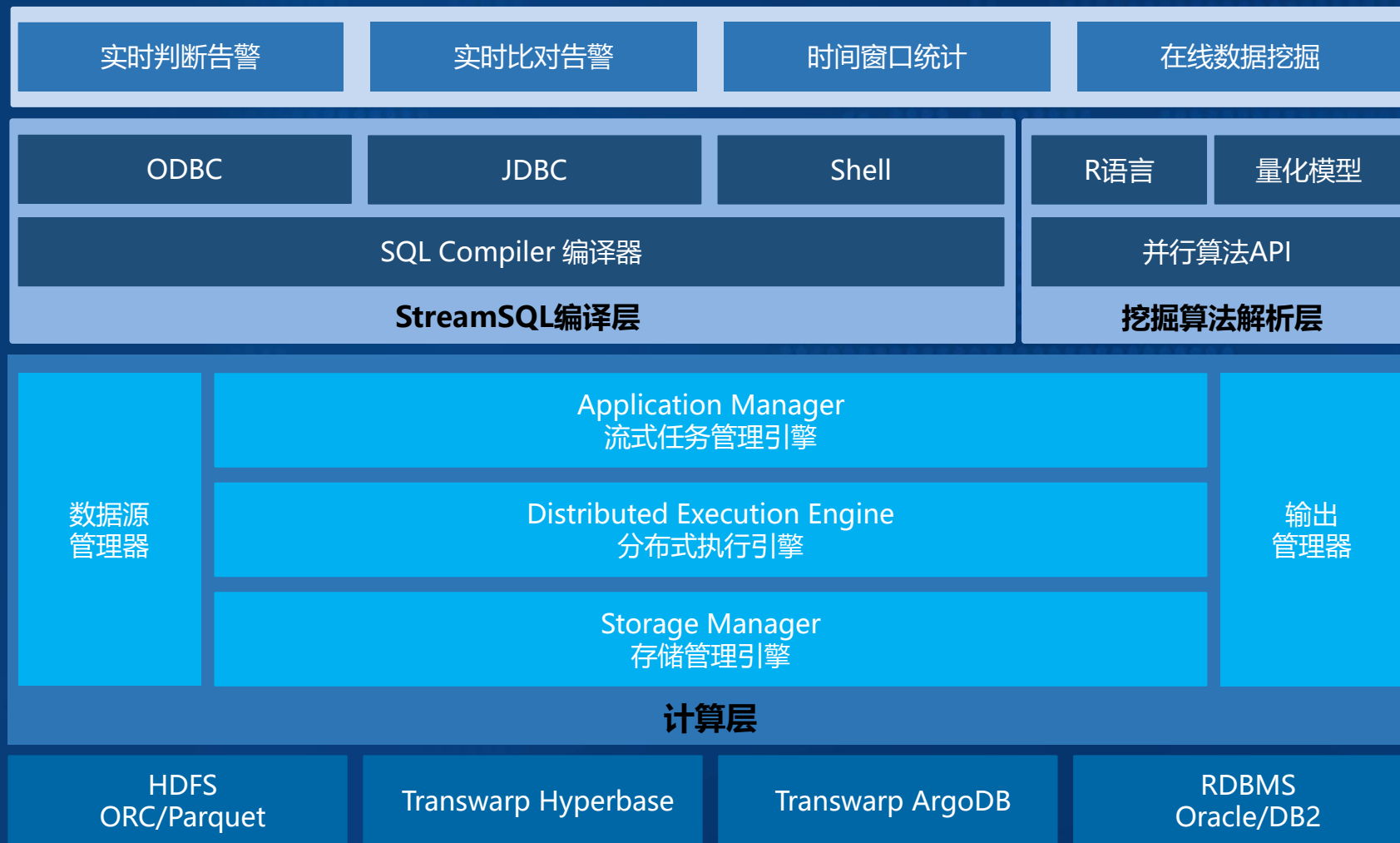
 **支持混合负载**
为高速硬件而生的存储格式

 **平台稳定可靠**
基于分布式一致性协议的存储引擎

 **运算能力强大**
专用的纯向量化计算引擎

「自主研发、自主创新、自主可控、国际领先的国产分布式数据库」

世界领先的实时流处理引擎 Slipstream



✓ 功能完善

- 完整的SQL支持
- 丰富的窗口功能
- 完善的复杂事件处理
- 流式规则引擎
- 流式立方体
- 流式微积分
- 流式机器学习
- 可视化设计与监控

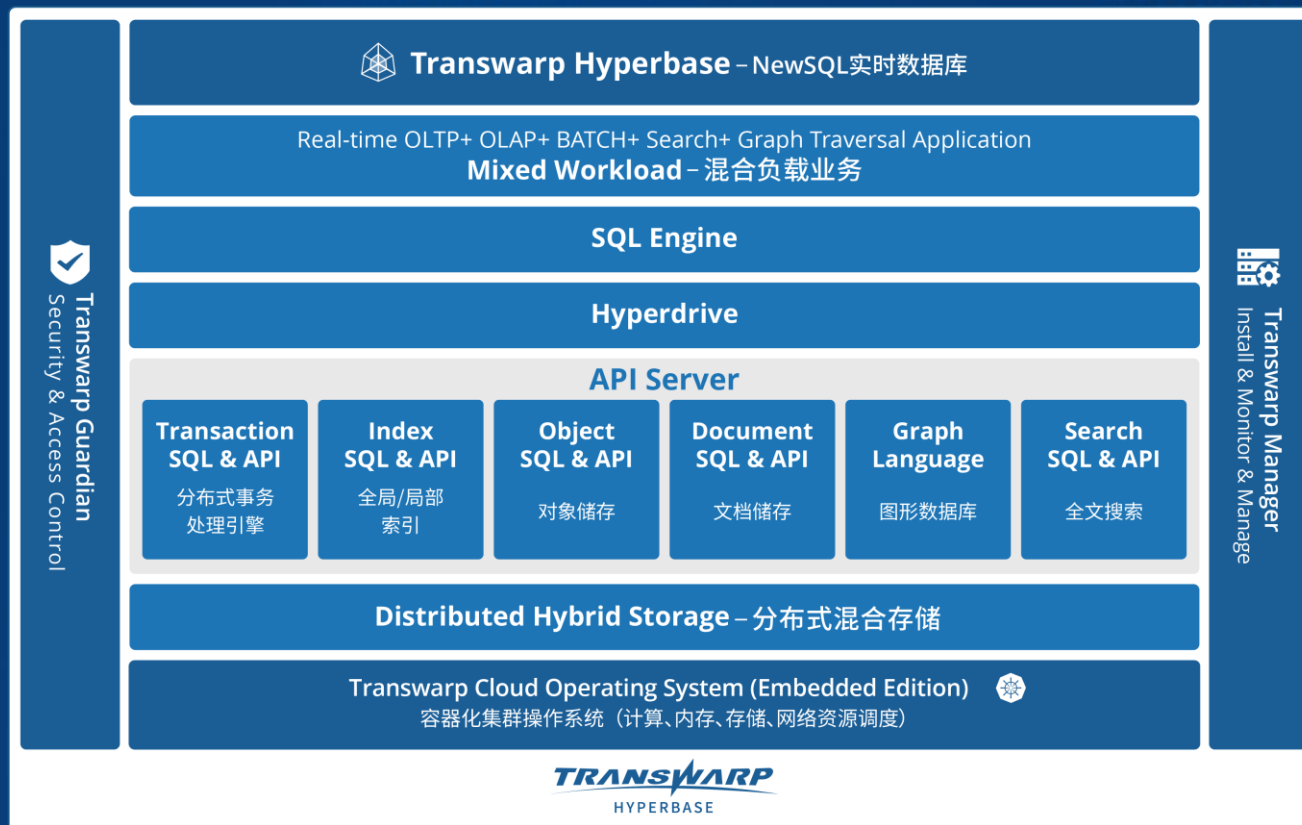
✓ 性能优异

- 低延时事件驱动模式
- 延迟低至<10ms
- 单节点百万级吞吐

✓ 安全可靠

- 安全认证和细粒度访问控制
- 数据不重不丢 (exactly-once)
- 多活的高可用保证
- 应用和资源隔离

更稳定、更健壮的操作型数据库 Hyperbase



- ✓ HBase升级到1.3，稳定性和健壮性大幅提高
- ✓ 开发多个运维工具，可维护性大幅提升
- ✓ 完善的SQL支持，方便应用开发
- ✓ 超高的并发访问支持，支撑高并发业务
- ✓ 支持多种索引 Global + Local + High-dimensional
- ✓ 结合Inceptor进行秒级高效分析
- ✓ 支持非结构化数据的存储和处理
- ✓ 支持海量数据的高速入库

国际领先的大规模统计和搜索融合引擎 Search



Transwarp Inceptor SQL Engine

Search-driver

API server

内存管理

存储管理

搜索引擎

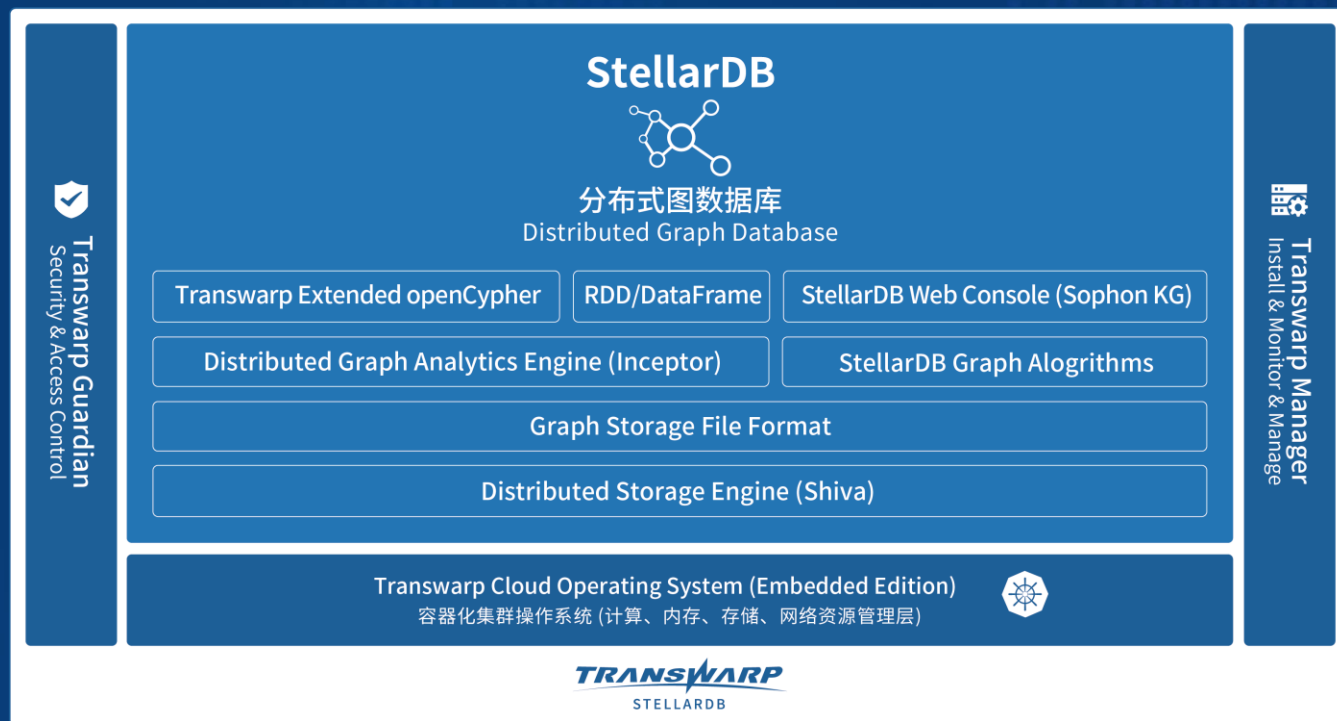
聚合计算

Lucene索引

「公安/监管领域的首选搜索引擎」

- ✓ 支持SQL 2003和全文检索SQL扩展，高性能检索和分析
 - ✓ 支持读写分离技术，服务稳定性增强
 - ✓ 支持堆外的内存管理技术
 - ✓ 可扩展性为PB级别，单节点可支持30TB
 - ✓ 支持对冷热数据不同的处理方式（有效降低10%~20%的内存空间）
 - ✓ 提供压缩速度更快、压缩率更高的存储方式（提高15%~25%的性能）
 - ✓ 支持时空地理信息的高效处理
- 硬件成本降到1/3
 - 千亿级数据搜索秒级返回
 - PB级数据分析能力
 - PB级非结构化数据的存储和检索
 - 时空分析秒级响应

企业级分布式图数据库 StellarDB



「在银监的担保链、公安的人物/团伙关系、审计的项目/企业/法人关系等拓扑分析领域得到广泛应用」

机器学习开发工具 Discover



Discover Spark-Shell, Pyspark, SparkR...

```
spark-shell
// 设置SparkContext
val sc = SparkContext("local[*]", "Discover")
// 加载数据
val data = sc.textFile("hdfs://...")
// 转换操作
val lines = data.mapPartitions(new IteratorFactory() {
  def newIterator(): Iterator[String] = {
    // 从文件中读取一行
    val line = ...
    // 处理数据
    ...
  }
})
// 输出结果
lines.saveAsTextFile("hdfs://...")
```



Discover TensorFlow教程

```
tf.nn.conv2d(x, w, [1, 1, 1, 1], [1, 1, 1, 1], name="conv2d")
tf.nn.conv2d(x, w, [1, 1, 1, 1], [1, 1, 1, 1], name="conv2d")
tf.nn.conv2d(x, w, [1, 1, 1, 1], [1, 1, 1, 1], name="conv2d")
```

Discover Inceptor SQL教程

sepal_length	sepal_width	petal_length	petal_width
5.1	3.5	5.4	0.2
4.9	3.0	5.4	0.2
4.7	3.2	5.3	0.2
4.6	3.1	5.5	0.2
5.0	3.6	5.4	0.2
5.4	3.9	5.7	0.4
4.8	3.4	5.4	0.3
5.0	3.4	5.5	0.2
4.4	3.0	5.4	0.2

- R和Python语言支持
- 100+分布式算法实现，部分性能较开源3~10倍提升
- 提供行业应用模板，如客户画像、客户分群、客户流失、担保圈分析等

大数据开发工具套件 Transwarp Studio

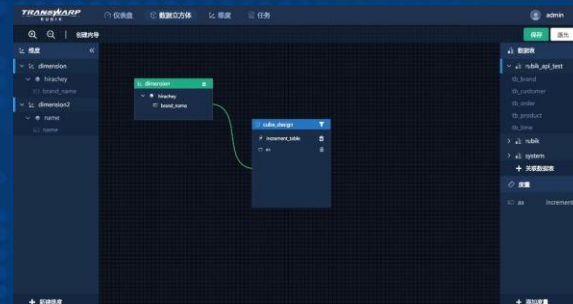
全面提升大数据开发和使用体验



ETL/实时同步工具 - Transporter



工作流引擎 - Workflow



Data Cube设计工具 - Rubik

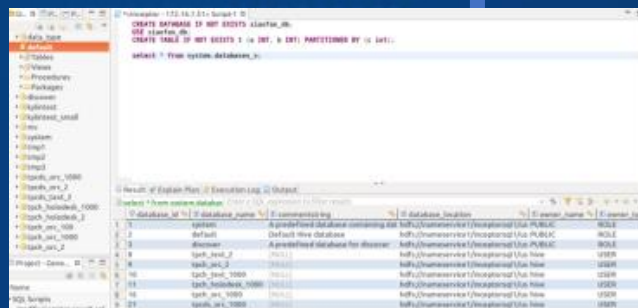
数据集成

数据转换

数据治理

数据建模

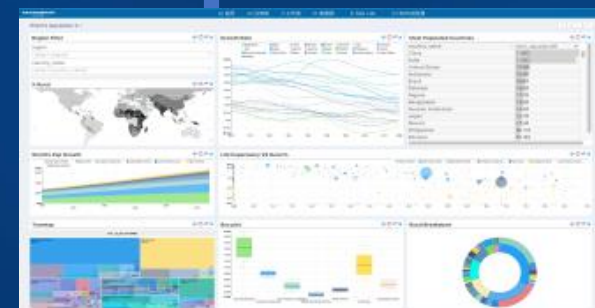
数据展现



SQL开发工具 - Waterdrop



数据资产目录 - Data Catalog



BI报表工具 - Pilot



完整SQL支持

- 支持 SQL2003、ACID/ 分布式事务、Oracle PL/SQL、主流数据库方言等
- 实现全产品线的SQL统一访问和开发，用SQL编写全场景应用



逻辑数据仓库

- 将平台中的数据库系统融合为一个逻辑上完整统一的大数据数仓
- 实现GB~PB级多源异构数据的高效存储和统一管理



低延事件驱动流处理

- 低延时高吞吐，延时<10ms，单节点百万级吞吐
- 支持复杂事件处理和流式机器学习
- 统一事件驱动模式和微批处理模式



大规模综合搜索

- 支持组合、全文、语义等多种搜索方式
- 支持PB级非结构化数据的存储与检索
- 支持高并发访问
- 千亿级数据搜索秒级返回



全面的机器学习开发支持

- 支持R和Python语言
- 100+分布式算法实现，部分性能较开源提升3~10倍
- 多种行业模板，如客户画像、反欺诈等



图形化开发工具套件

- 八个图形化的大数据开发工具
- 覆盖数据集成、转换、治理、建模和展现等全部大数据开发环节



友好的开发体验

- 全场景覆盖，用SQL编写全场景应用
- 开发门槛低，全产品线支持SQL编程
- 开发体验好，全套图形化大数据开发工具



业务平滑迁移

- 业界SQL支持最完善
- 实现Oracle、DB2、Teradata等传统平台应用的平滑迁移
- 中国邮政集团：全球首例替换Teradata



容器化集群操作系统

- 采用容器编排技术进行资源管理
- 支持TDH的一键部署、升级和扩缩容
- 支持基于优先级的抢占式资源调度和细粒度资源分配



极致的性能体验

- TDH 的批处理速度是开源 Hadoop 的 10~100倍，是MPP的5~10倍
- 实现GB~PB级多源异构数据的高性能复杂查询和挖掘分析



高可用 / 弹性扩展

- 平台中的所有集群均实现数据与服务的高可用，以确保平台健康稳定运行
- 不停服情况下，实现集群的动态扩容和缩容



简易部署 / 一键升级

- 一站式/可视化部署大数据集群，时间从几周减少到几分钟
- 不停服情况下，实现平台组件的一键升级



一站式集群管理

- 对平台组件和服务进行一站式/可视化监控、管理和优化
- 提供监控、度量、告警、健康检测、磁盘管理、软件升级、服务迁移等功能



统一的安全多租户管理

- Guardian负责平台的统一安全控制和资源管理，支持Kerberos和LDAP认证
- 实现租户级资源管理，提供对HDFS和所有数据库对象的细粒度访问控制



系统科学的认证培训体系

- 系统科学的大数据与人工智能认证培训课程体系
- 颁发工信部急需紧缺人才培养证书
- 经验丰富的培训讲师团队



完备规范的客户服务体系

- 标准的客户服务流程
- 严格的服务考核标准
- 专业的技术服务团队



Thanks

www.transwarp.io

星环信息科技（上海）有限公司 版权所有

公司地址 / Our Office

上海：徐汇区虹漕路88号越虹广场B座11F&12F

北京：海淀区西直门北大街甲43号金运大厦B座1101室

广州：天河区体育东路140-148号南方证券大厦1015-1016室

联系电话：4007-676-098