TRANSWARP
DATA HUB

Transwarp Data Hub (TDH)
企业级一站式大数据综合平台
白皮书

星环信息科技（上海）有限公司

Transwarp Data Hub (TDH)
企业级一站式大数据综合平台

2006年Hadoop技术的出现标志着大数据技术时代的开始，经过10多年的蓬勃发展，大数据技术已经真正承托起一大批企业的数据基础架构。Transwarp Data Hub（简称TDH）是星环信息科技（上海）有限公司研发的企业级大数据平台，经过4年的快速演进，已成为国际一流的大数据平台。从2016年起，TDH正式成为Gartner认可的Hadoop国际主流发行版本。

TDH主要提供5款核心产品:Transwarp Inceptor是大数据分析数据库,Transwarp Slipstream是实时计算引擎,Transwarp Discover专注于利用机器学习从数据提中取价值内容,Transwarp Hyperbase用于处理非结构化数据,Transwarp Search用于构建企业搜索引擎。通过使用TDH，企业能够更有效的利用数据构建核心商业系统,加速商业创新。

TDH产品的主要技术优势包括以下几个方面：

极致的性能与可扩展性 TDH的批处理速度是开源Hadoop的10-100倍，是MPP的5-10倍，可以对从GB到PB级的数据量实现复杂的查询和分析。TDH具有高可扩展性，用户可以通过增加集群节点数量，线性提高系统的处理能力。

容器技术与大数据平台 TDH支持部署于TOS之上。TOS是为大数据应用量身订做的云操作系统，基于Docker和Kubernetes，支持一键部署TDH、扩容、缩容，同时支持基于优先级的抢占式资源调度和细粒度资源分配。

完整的SQL和ACID支持 Transwarp Inceptor是第一个实现完整SQL支持的Hadoop发行产品。它不仅支持SQL 2003，Oracle PL/SQL以及DB2 SQL PL，还实现了完整的ACID和CRUD功能。TDH提供JDBC和ODBC驱动连接，方便第三方工具运行于TDH之上。

低延迟的流处理 Transwarp Slipstream是同时支持事件驱动和微批处理的流处理引擎，计算延迟最低可至5ms。它提供标准的SQL编程接口，还支持高可用性（HA）和Exactly-Once的语义，从而支持7x24小时的生产业务。

丰富的机器学习和深度学习功能 Transwarp Discover支持用户通过R语言和Python开发机器学习项目,也可以用图形化的工具做分析。

大数据上的全文搜索 Transwarp Search支持通过SQL实现大数据上的秒级全文搜索，它利用层次化存储、堆外内存管理等创新性技术，极大的提高了系统的可用性。此外，Search还可以结合Inceptor提供较强的数据分析能力。

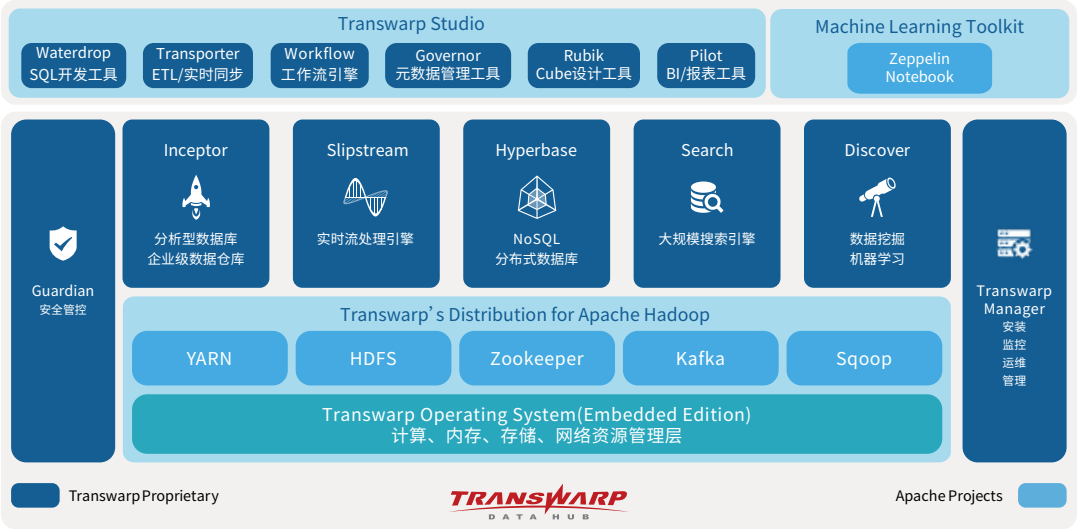
图形化的大数据开发工具套件 Transwarp Studio是TDH中的大数据开发工具集，包括元数据管理Governor、工作流Workflow、数据整合工具Transporter，Cube设计工具Rubik以及报表工具Pilot。用户可以使用这些图形化工具来提高大数据的开发效率，降低技术门槛。

多样化的数据处理功能 Transwarp Hyperbase用于存储和计算结构化或非结构化数据，包括日志记录、JSON/XML文件以及二进制数据（如图像和视频）。Hyperbase底层是KV的数据库，因此其非常适合高频次的数据入库、高并发精确检索等业务。

简易的操作和管理 Transwarp Manager是专门用于部署、管理和运维TDH集群的组件。它支持产品一键安装、一键升级和图形化运维，并提供了预警和健康检测功能，帮助用户简化运维过程。

统一的安全/多租户管理 Transwarp Guardian是TDH平台中实现安全控制和资源管理的中央服务平台，它支持Kerberos和LDAP认证，可以做细粒度的权限控制，并且提供租户管理功能。

Transwarp Data Hub 体系架构



Transwarp Data Hub由Apache Hadoop、5款核心产品、大数据开发工具集Studio、安全管控平台Guardian和管理服务Manager构成。

Transwarp 核心产品

产品名称及其主要特点和使用场景

Transwarp Inceptor

Inceptor是一款用于批量处理及分析的数据库。它支持SQL 2003标准、Oracle PL/SQL以及DB2 SQL PL，对Oracle、DB2以及Teradata都有很好的方言支持，是Hadoop领域对SQL标准支持最完善的产品。Inceptor的另一大优势是对ACID的支持，从而可以满足用户对数据处理中一致性和可靠性保障的需求。此外，Inceptor拥有极为优异的大数据分析性能，比Apache Hadoop处理速度快10倍以上，比MPP处理速度快5倍以上，在TPC-DS和TPC-H基准测试中也胜于其他Hadoop和MPP产品。目前，Inceptor被广泛地应用于数据仓库和数据集市的构建，在中国，已经有超过500家客户在Inceptor上创建了他们自己的商业应用。

Transwarp Slipstream

Slipstream是提供实时计算的产品，被广泛用于交通运输和物联网行业。和其他解决方案相比，Slipstream有几个突出的技术优势：完整的SQL支持使得实时业务开发过程更加简便；基于事件驱动的计算引擎可将延迟时间缩减到5毫秒，是Spark Streaming引擎的延时的1/100；此外Slipstream支持复杂事件处理能力（CEP），因此用户可以基于Slipstream用SQL语言开发比较复杂的在线流计算业务，如在线反欺诈应用等。Slipstream还提供完善的高可用性（HA）和Exactly-Once语义，而这些都是使实时应用稳定、可靠的保障。

Transwarp Discover

Discover是分布式机器学习平台，它包含了丰富的分布式算法库，还内置了多个行业应用模块，例如金融反欺诈、文本挖掘算法库等。Discover提供了R语言、Python和SQL接口，以帮助数据科学家开发自己的数据挖掘算法。通过内置Notebook工具Zeppelin，Discover可以非常灵活的支持数据工程师和科学家之间的团队协作。

Transwarp Hyperbase

Hyperbase是以Apache HBase为基础，融合了多项创新技术的NoSQL数据库：它采用了和Inceptor同样的SQL引擎，允许开发者们直接用SQL构建复杂应用；支持全局索引和次级索引，实现高速的非主键查询；提供原生的JSON/BSON格式支持以及对象存储（Object Store）技术，极大地简化了非结构化数据处理。

Transwarp Search

Search用于在企业内部构建大数据搜索引擎。它能够在PB数据量级上实现秒级延迟的搜索功能；在开发接口方面，Search提供了完整的SQL语法支持并提供了搜索语法SQL扩展，通过和Inceptor优化器有效结合，使开发者无需了解底层架构就可以开发出高效的搜索引擎。Search创新的使用了堆外内存管理技术来提高系统的健壮性，避免了GC问题对系统的影响；此外，Search还支持混合存储，通过将热数据存储在SSD上来提升查询速度。

系统平台和管理组件

产品及其主要特点

Apache Hadoop

Transwarp基于Apache Hadoop 2.7.2开发，以HDFS为文件系统，以YARN为资源管理平台。Transwarp对各组件性能进行了优化，提升安全性、稳定性，从而提供7×24小时的不间断服务。

Transwarp Operating System

TOS是为大数据应用量身订做的云操作系统。它基于Docker和Kubernetes，支持对TDH一键式部署、扩容、缩容，同时也允许其他服务和大数据服务共享集群，从而提高资源的使用率。TOS采用创新的抢占式资源调度模型，能在保障实时业务的同时，提高集群空闲时的资源占用，让批量作业和实时业务在互不干扰的情况下分时共享计算资源。

Transwarp Manager

Manager是负责配置、管理和运维TDH集群的图形工具。用户只需通过几个手动步骤，就可以在x86服务器上或基于Docker的云端平台上部署一个TDH集群。Manager的运维模块提供告警、健康检测、监控和度量这四项服务。用户可以轻松的浏览各服务的状态，并且在告警出现时采取恰当的措施以处理应对。此外，Manager还提供了一些便捷的运维功能，例如，磁盘管理、软件升级和服务迁移等。

Transwarp Guardian

Guardian为TDH提供集中的安全和资源管理服务。它支持LDAP和Kerberos，保护Hadoop集群免受恶意攻击和安全威胁，而且还可以对资源做细粒度的ACL控制。其多租户资源管理模块可以按照租户的方式管理资源，并通过一个图形化工具为用户提供权限配置以及资源配置接口。

Apache Kafka

Transwarp Kafka在Apache Kafka 1.0的基础上添加了大量安全特性：整合Kerberos以保护数据；允许Producer和Consumer使用不同的KDC来进行交叉认证；通过支持认证命令行以简化操作。

开发工具

组件及其主要特性

Transwarp Transporter

Transporter是一款用于设计和创建ETL任务的可视化工具。它支持从RDBMS到TDH的近实时数据同步功能，用户可以利用Transporter将数据从RDBMS迁移到Hadoop，再进行数据分析和挖掘工作。Transporter提供完整的数据整合功能，源系统支持多种格式的数据源，包括CSV、JDBC、XML、JSON以及关系数据库；支持多种常用的数据转换操作，例如，连接、聚合、清洗等。由于数据迁移过程中产生的数据处理任务都在Inceptor中完成，且受完整的ACID支持，因此用户不必为了ETL任务建立单独集群，也不用担心数据一致性问题。

Transwarp Workflow

Workflow是一个图形化的工作流设计、调试、调度和分析的服务平台，它支持Shell、SQL、JDBC、HTTP等任务类型，也可以写自定义Java任务。它还提供丰富的分析能力，如依赖关系、执行历史、甘特图等，可以帮助用户诊断工作流的执行状况。

Transwarp Rubik

Rubik是一款用于设计OLAP Cube的可视化工具，所建Cube可以实例化于HDFS或Holodesk。Rubik支持雪花模型和星形模型两种Cube设计模型，并支持多种格式的数据源(包括HDFS和远程RDBMS)。实验显示，在数据立方体的加速下，分析查询的速度可提高10倍。Rubik通过可视化方式提供服务，使数据分析师得到更友好的交互体验。

Transwarp Governor

Governor是TDH中的元数据管理和数据治理工具。用户可以用它来管理元数据(包括表和存储过程)，监控所有数据和程序的更改历史，进行数据血缘分析和影响分析。开发者可以利用Governor调试数据问题，追踪问题来源，并帮助数据管理者预测计划进行的元数据更改会造成哪些影响，因此Governor能够帮助用户提高大数据的数据质量。

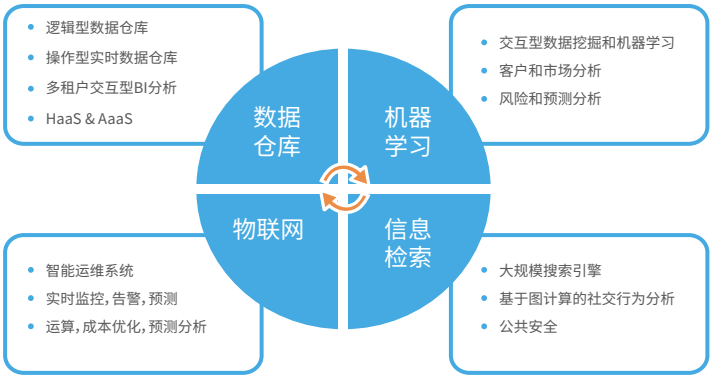
Transwarp Waterdrop

Waterdrop是TDH中一个SQL IDE工具。它包含的子模块有SQL编辑器、元数据管理器、SQL执行器、以及数据导入/导出。Waterdrop提供语法检测、SQL格式化和开发助手等功能，可帮助开发者极大地提高开发效率。

Transwarp Pilot

Pilot是基于Web的报表展现工具，轻量、灵活，可以快速部署。它支持多维度的分析和自助分析，提供数十种报表样式，对时序数据也有很好的展现。此外，Pilot还支持团队协作和共享，支持导入和导出报表。

目标市场



为什么选择TDH?

SQL 2003, PL/SQL & SQL PL以及SQL扩展支持

TDH是行业内第一个提供完整SQL 2003支持的SQL on Hadoop引擎，从2014年开始支持Oracle PL/SQL，从2015年起支持DB2 SQL PL。因此，用户可以很方便的把在Oracle或者DB2上创建的传统应用迁移到TDH，从而更好地利用大数据促进业务创新。为了适配各种数据库语言，TDH还允许用户设置数据库方言，目前可以很好的支持Oracle、DB2和Teradata。

为了降低开发流应用程序的难度，TDH还支持Stream SQL标准，其中包含流扩展后的SQL 99。因此，开发者可以在TDH上直接使用SQL而不是通过各种API来编写流计算程序，也不需要考虑任何打包或部署工作。为了更好的提供全文搜索服务，TDH中的Search也支持SQL的检索扩展语法（兼容Oracle标准）。由于支持标准的JDBC 4.0和ODBC 3.5，TDH可以兼容主流的数据库软件和中间件。

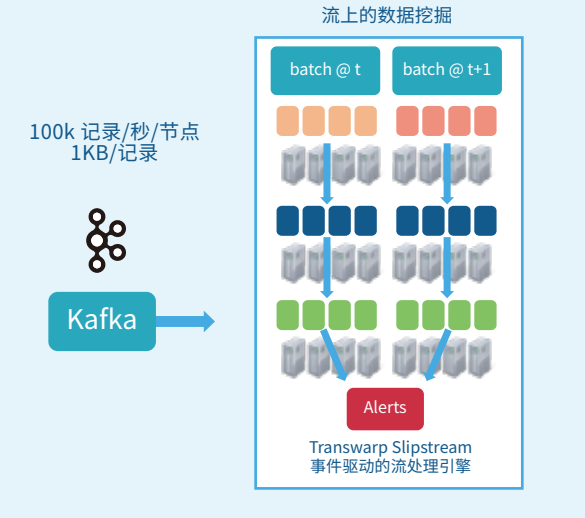
平台	SQL 92	SQL 99 & 2003	Oracle PL/SQL	DB2 SQL PL	Database 方言	DB-Link 扩展	Stream SQL	Search 扩展
TDH	是	是	是	是	是	是	是	是
Apache Hive	是	部分	部分	否	否	否	否	否
Apache Spark	是	是	否	否	否	否	否	否

统一事件驱动和批量处理的流处理引擎

Transwarp Slipstream是统一支持微批处理和事件驱动的混合流计算引擎，可以同时支持低延时和高吞吐的实时计算场景。在事件驱动的模式下，数据触发的计算任务延迟可以低至5毫秒。这样一来，用户就可以利用Slipstream来开发对延迟时间敏感度较高的应用，比如在线反欺诈应用。而在微批处理的模式下，Slipstream能够提供极高的吞吐，适合运用在某些对吞吐量要求较高的特殊行业，例如交通的视频检测。

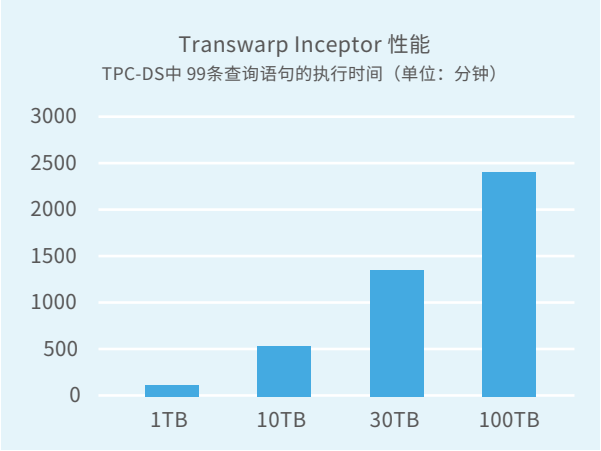
为了满足用户在流上构建复杂应用的需求，Slipstream还提供了CEP(复杂事件处理）的引擎，用户可以基于SQL写出复杂的数据处理规则和逻辑，这也是Slipstream的另一个独特的技术创新。

Slipstream支持高可用性和Exactly-Once语义，不需要开发者亲自编写数据处理或重复执行的逻辑。另外，机器学习应用也可以运行在事件驱动的流引擎上。

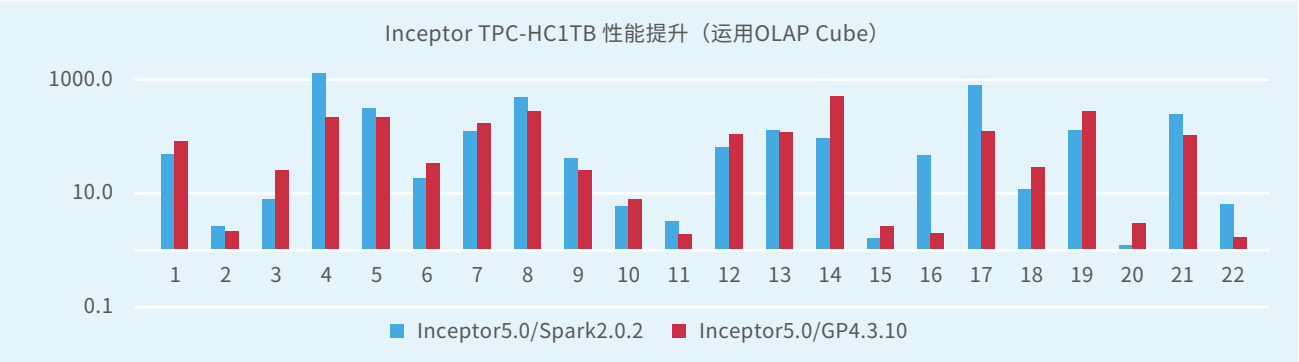


极致的性能

Transwarp Inceptor作为大数据分析工具有着极致的执行性能和扩展能力。星环深度优化了Inceptor的分布式计算引擎，并使其具有灵活的扩展能力，同时它会对数据重分布和广播的逻辑进行调试优化以达到更好性能。Inceptor Holodesk是一个基于SSD或内存的列式存储引擎，能提供非常高的数据读取速度。此外，Inceptor还采用了基于成本的优化和基于规则的优化来为执行任务选择最佳执行计划。所有这些特性都有助于Inceptor提高批量处理过程的效率和扩展性，在TPC-DS各个数据量级别的测试中，Inceptor都有非常好的表现，如右图所示。



Inceptor可以很好地适配各种交互式数据分析和OLAP场景。Holodesk提供了索引支持，并且可以有效的利用SSD来加速扫描，因此对交互式分析场景的业务可实现多倍的提速。对于模式较为固定的数据报表业务，用户可以采用OLAP Cube技术使分析性能提高10-100倍。我们以1TB的数据集为对象进行测试，发现在OLAP Cube的加速下，TPC-H在Inceptor中的运行速度比SparkSQL和Greenplum快近100倍。



友好的开发体验

大数据的技术门槛很高，通常需要开发者充分了解系统的底层架构，掌握集群部署、操作和应用程序开发等各种能力，对软件工程的扩展构成阻碍。为了解决上述问题，星环力争简化TDH的使用过程，并在系统部署和管理等方面提供友好的用户体验。

SQL通常是开发者用于实现OLAP业务、批量处理、流处理、搜索以及其他应用程序的首选工具，因此为了方便程序开发，TDH对SQL提供了全面的支持。所以在写代码之前，开发者既不需要研究目标引擎的实现细节，也不需要深入了解API或考虑在API发生变化时应如何正确的维护应用程序。此外，由于TDH对事务的支持，开发者无需再为应用程序编写复杂的逻辑以处理事务异常，这将极大的提高他们的工作效率。总的来说，TDH可以实现在降低技术门槛的前提下，保障极高的开发效率，是使TDH成为一个成功大数据产品的关键性优势。

ACID支持

ACID对于大数据的数据处理和数据清洗过程至关重要。如果没有ACID，数据的插入修改过程将存在各种潜在问题，终端用户需要亲自探究事务操作的失败原因并找出避免和解决问题的方法，这使用户应用的过程变得复杂，甚至根本不可行。更糟糕的是，如果没有ACID，当两个应用程序向同一个数据块中写入数据时，会很容易出错。

TDH是第一个提供完整ACID支持的Hadoop商业化产品。Transwarp Inceptor实现了串行化的事务隔离，并通过两阶段锁和MVCC协议保证数据的一致性。

	Transwarp Inceptor	Hive	Impala	Oracle
CRUD 支持	支持	支持	不支持	支持
事务类型	事务+自治事务	自动提交事务	不支持	事务+自治事务
隔离级别	可串行	不支持	不支持	只读+可串行+读取提交
事务错误处理	事务+ PL/SQL + SQL PL	不支持	不支持	事务+PL/SQL
数据一致性	支持	不支持	不支持	支持

容器技术和Kubernetes资源管理

TDH中的组件都针对Docker作了优化，计算引擎也可以使用Kubernetes进行资源管理，得以使TDH以较低的成本部署在公共云或者私有云上。星环还充分利用Docker和Kubernetes的资源隔离能力和对资源调度的QoS支持，提供了弹性的资源共享，保障数据、资源、应用之间的隔离，实现了更好地多租户管理，以支持各种不同的业务需求。

容器技术给TDH的部署和维护带来了非常显著的提升，它支持动态扩容、缩容，支持灰度升级，可以实现在不停服的情况下对系统进行升级。

丰富的机器学习框架

Transwarp Discover为终端用户提供了R语言接口用于进行数据挖掘，并实现了超过60种分布式机器学习算法和多种行业模型，包括金融行业的交易反欺诈模型、文本挖掘模型等，从而加快机器学习在这些行业的落地商用。

统一的安全和多租户管理功能

Hadoop的安全问题对于将产品投入实际生产而言是一个极大的挑战。从2017年年初开始，国际上发生了一系列的HDFS入侵事件，Hadoop集群被攻击的案件不断被报道出来。为了保障Hadoop的安全，更好的认证和授权服务成为刚需。因此，我们设计并实现了统一的安全和资源管理服务Guardian来保护TDH。在Guardian的保护下，所有的应用服务都可以借助Kerberos实现数据加密，或者通过LDAP实现身份验证。Guardian还实现了租户级别的资源管理，提供对HDFS和所有数据库对象的细粒度访问控制。

支持SQL的搜索引擎

Transwarp Search主要用于构建企业内部搜索引擎，它能够支持PB级别的高速全文检索。Search支持使用SQL做数据检索。和API编程相比，SQL的好处在于，它不仅可以结合SQL编译器提供更优秀的性能，而且可以避免底层存储升级引入的兼容性问题。Transwarp Search与SQL引擎优化器深度整合，优化器会根据数据特点优化数据搜索的执行过程，例如进行下推聚合运算，以提供更好的性能。

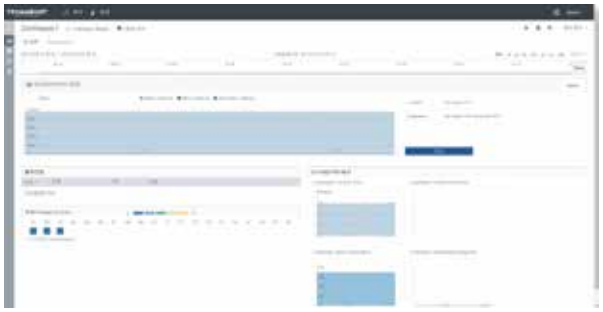
多样的开发工具以及便利的访问接口

TDH提供了大数据开发工具套件Transwarp Studio解决开发者在做大数据应用的痛点和效率问题，其中：Transporter用于设计和创建ETL任务，提供各数据源到大数据平台的数据迁移；Workflow提供可视化的工作流设计，以及对工作流的调试、调度，对工作流任务具有丰富的分析功能；Rubik用于实现OLAP Cube的设计以及实例化，以加速交互式多维度分析；Governor，负责提供对元数据的访问、历史跟踪、状态统计等功能，为用户提供数据治理服务；Waterdrop，一种SQL IDE，提供SQL编辑、执行以及数据导入导出等功能，协助用户提高SQL开发效率；Pilot，轻量的在线报表工具，支持多种报表样式，提供丰富美观的报表展现。

Transwarp Studio几乎涵盖了数仓开发的各个环节，包括数据入库，数据血缘管理，数据质量管理，交互式分析加速，调度流，报表展现等。它可以帮助用户解决数据分析中复杂的问题，降低大数据应用的构建成本，使技术人员能够快速的在Hadoop平台上构建正确的应用，提炼数据的价值。

易用的管理

Transwarp Manager提供了可视化的界面管理以控制Hadoop集群，以及一站式的集群部署、操作和预警功能。通过Manager，运维人员可以轻松部署和集中操作整个TDH集群，部署时间可从几周减少到几分钟，而且仅依靠若干次点击就能完成一些复杂的硬件磁盘管理。Manager还通过度量标准收集和展示框架，采用扁平化风格和可视化界面来展示TDH集群服务状态各个指标，制定配置的更改和全范围的报告和诊断工具来帮助用户监管集群性能，优化性能和利用率。



关于我们：

星环科技是全球领先的大数据与人工智能基础平台供应商，专注于提供企业级容器云计算、大数据和人工智能核心平台的研发和服务，打造大数据和人工智能生态的“中国心”。公司以上海为总部，以北京、广州、多伦多为区域总部，并在南京、郑州、成都设有支持中心，同时在深圳、天津、武汉等地设有办事机构。经过多年自主研发，星环科技建立了四条产品线：一站式大数据平台Transwarp Data Hub(TDH)、基于容器的云操作系统Transwarp Operating System(TOS)、人工智能平台Transwarp Sophon和超融合大数据一体机TxData Appliance，并拥有多项专利技术。2016年被Gartner评为全球最具有前瞻性的数据仓库及数据管理解决方案厂商，2017年被IDC评为中国大数据市场领导者。公司产品已经在十多个行业应用落地，是国内落地案例最多的大数据和人工智能平台供应商。目前星环科技已完成C轮融资，由腾讯领投。

核心技术：

基于Hadoop构建企业级数据仓库和数据集市

高性能、扩展的分布式数据库

融合低延时的事件驱动机制和复杂批处理编程模型的流处理引擎

具备统计、机器学习和深度学习的人工智能平台

应用行业：

金融、电信、交通、物流、政府、公共安全、媒体、电力、能源、零售、制造业、医疗、教育等。

部分用户：

电话：4008-079-976　北京分公司：010-62682815

网址：www.transwarp.io

咨询：sales@transwarp.io　技术支持：support@transwarp.io

上海：徐汇区虹漕路88号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

20180207