BEYOND
TECHNOLOGY

SaCa® DataCompare

数据比对平台软件

产品白皮书

Neusoft

目录

一、概要

不断增长的结构化和非结构化数据增加了信息管理的复杂性，同时客户希望在跨异构环境中能更出色地管理数据，企业需要数据高度可用，需要能够不间断地访问数据，同时不会导致性能下降和服务中断，为此企业需要拥有冗余的分布式数据副本。然而，在当今复杂的 IT 环境中，在各个分布式数据副本之间保持数据一致性极具挑战，不幸的现实是可能出现数据差异。如果不良数据未被发现并解决，则可能导致错误的决策，最终出现运营、财务和法律风险。

在 SaCa DataCompare 的帮助下，企业可以信心十足地在其备份系统、报表/查询数据库、主从数据库和其他类型的冗余数据系统中实现数据一致性。在本文中，我们将这些系统统称为目标数据库。SaCa DataCompare 在源数据库与目标数据库之间执行定期检查，企业可根据需要设置检查频率，且无需使任一系统脱机。

SaCa DataCompare 为在对业务造成负面影响之前发现不同步的数据提供了一个易用且强大的解决方案。SaCa DataCompare 可与 SaCa CDC 实时数据复制产品，SaCa DataExchange 数据交换产品，SaCa DataTransform 数据转换与清洗产品一起部署，也可以单独部署，能够确保在各数据库之间保持数据一致性。

二、保持数据一致性面临的挑战

在我们讨论对帮助管理数据库间数据一致性的解决方案的需求之前，我们需要了解企业中出现数据不一致的常见原因。

当目标数据库中的数据偏离源数据库时就出现了数据差异。数据偏离的程度取决于各种因素，一些可能是有意而为之，一些可能是无意形成的。

即使使用能够可靠复制数据的产品，如 SaCa CDC 或 SaCa DataTransform，仍会存在可能导致出现数据差异的一些原因。如果目标数据库的目标是保持与源数据库严格一致，那么 IT 将需要实施相应的流程和策略以确保实现此目标。下面将描述可能导致出现数据差异的一些原因:

- **迁移错误**

在可以开始复制之前，要使用各种不同的迁移工具来帮助进行目标数据库的初始加载。迁移工具和复制产品中用于处理数据的配置上的差异可能会导致出现数据差异。

例如，如果某列值未知，迁移工具可能使用“?”，而复制产品可能使用“null”。当执行迁移时，可能存在待办事务未纳入其中，从而导致目标数据库上的数据缺失。

- **源数据库与目标数据库中的差异**

源数据库和目标数据库的配置差异，如编码、区域设置、字节顺序或数据库版本不同，可能导致在迁移和复制过程中出现细微差异。例如，不兼容的字符集或日期/时间格式和范围可能会导致目标数据库出现错误。

- **实例化错误**

在可以开始迁移或复制之前，将需要使用正确的模式和约束条件实例化目标数据库。如果这么做时出现故障，将导致源数据库和目标数据库不同步。

例如，如果无法设置主键/唯一键，可能会导致出现重复行。即使源数据库中无重复行也可能会创建重复行，因为无法保证完成迁移作业时不出现任何故障。其他实例化错误包括不

正确的迁移作业、脚本和触发器，这些可能导致不正确地修改数据。

- **配置错误**

复制产品的不当配置和意外配置可能导致出现差异。这种类型的差异不显示在复制日志中，因为从复制产品的角度来看，是按配置执行。这还可能让 QA 测试发现不了问题。

- **复制方面的漏洞**

虽然在源数据库与目标数据库之间启用了复制，并且复制运行完好，但也存在源数据库中插入的数据未被复制的情况。例如，当批量插入数据时，用户常常会使用数据库中的一些选项 (如 Oracle 数据库中的 NOLOGGING)，这些选项会导致复制系统避免捕获这些数据。

- **复制延迟**

使用异步复制，对源数据库进行更改与将这些更改提交到目标数据库之间将存在短暂的延迟。然而，如果不满足最大延迟要求，可能会违反服务级别协议或数据合规性要求。

- **基础架构故障**

系统故障、磁盘损坏和网络故障等基础架构错误可能导致源数据库与目标数据库之间出现数据差异。恢复有故障的系统后，一项主要任务是要确保源系统与目标系统之间的数据一致性毫发无损。

虽然 SaCa CDC 等复制或迁移产品通常具有检查点功能，但它们无法保证发生崩溃后对系统进行恢复时目标数据库中数据的质量。

- **用户错误**

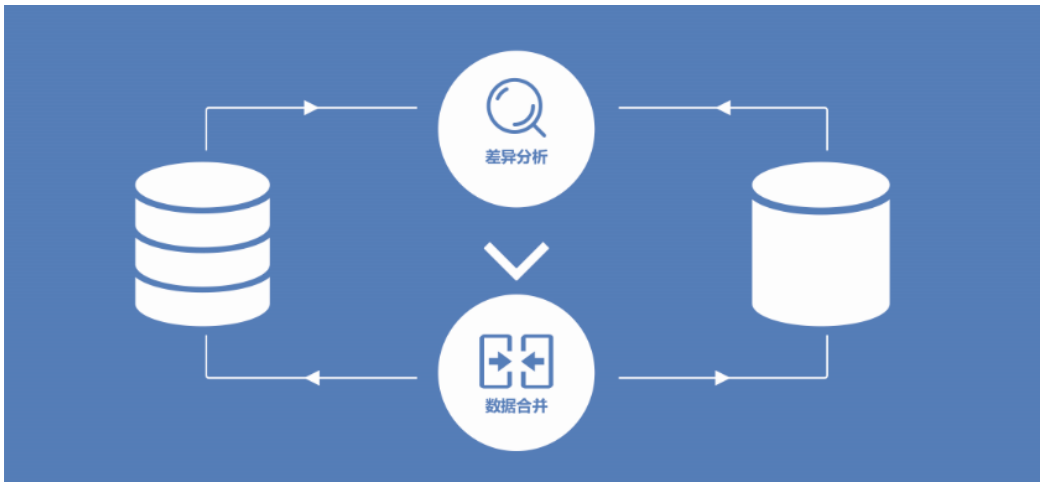
通常创建目标数据库的目的是分流源数据库的查询处理。这样可以在不影响源数据库上运行的应用程序的情况下生成丰富的运行报告。根据使用的技术，目标数据库不仅对于读操作是开放的而且对于写操作也可能是开放的。如果是这样，即使应用了 IT 策略，用户/DBA 也可能会无意或恶意修改数据。

- **应用程序错误**

使用目标数据库的应用程序可能因逻辑错误以及在应用程序升级期间更改数据。此外，即使目前复制运行正常且数据保持一致，IT 也可能会开发使用目标数据库的新应用程序，因此可能会在未来某个时候修改数据。

三、关于 SaCa DataCompare

SaCa DataCompare 是一款高性能且侵入性极低的数据比对工具，可帮助管理整个企业内的数据一致性，该产品可在多种情境下使用以确保数据一致性，如用来对关系型数据库、列存储数据库中的数据进行比对和同步。通过比对的方式分析两端数据库的差异性（数据结



构、数据量、数据项），并能够提供差异性报告，同时支持数据的合并。

SaCa DataCompare 可与 SaCa CDC 等产品以及其他类似产品无缝协作，它可对这些产品的功能加以补充并确保数据一致性。

现在我们已充分了解在整个企业内维护一致数据面临的挑战，下面我们将讨论 SaCa DataCompare 一些关键功能，这些功能不仅可帮助应对上述挑战，而且还能无缝融入 IT 组织中。

- **支持多种数据源**

支持主流的关系型数据库（Oracle、DB2、SQLServer、MySQL、以及达梦、南大、金仓等国产数据库），同时可灵活扩展新的数据源。

- **提供多种比对方式**

支持基于表结构比对识别结构化差异、数据量比对识别数量差异、数据项比对识别信息内容差异，支持表、视图、自定义 SQL 等方式。

- **适用于各种网络环境**

既可以在企业局域网内部使用，也可部署在跨网络、跨机房等环境中，即使是跨机房的方式也不会丢失其性能和配置的简便性。

- **Web 方式管理监控**

提供完全基于浏览器的方式配置比对模型，监控和执行差异分析流程，查看差异数据报告以及基于个性化配置的数据补齐过程。

- **快速安装部署**

提供多种平台产品安装包，解压即用，无需额外安装其他软件，而且对硬件的要求也不算高，绝大多数的计算机配置都可以很流畅地运行这个系统。

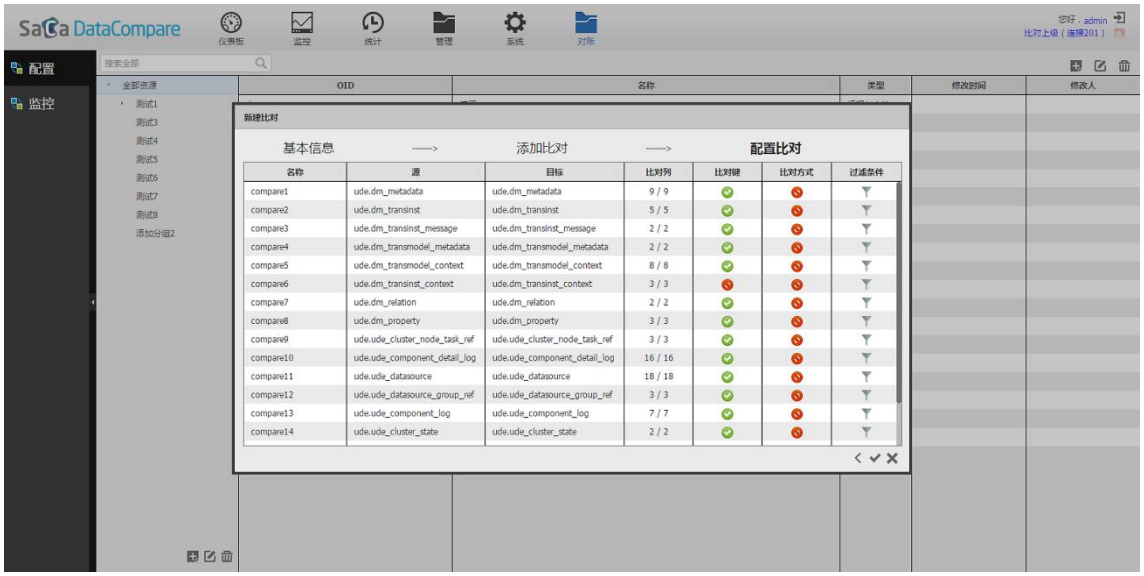
- **多样化调度方式**

支持调度，可以实时的不间断的调度、定时的基于某个固定的时间点或周期调度、手动执行或者基于 WebService 接口的触发调度。

四、工作原理

在探究 SaCa DataCompare 的架构、安装和配置细节之前，我们简单介绍一下 DataCompare 的工作原理。

与要么全有要么全无的方法不同，DataCompare 允许用户选择要比较的对象以及灵活地确定比较方式，以便只处理相关数据，只突显相关差异。



在初始比较（或行散列）步骤中，利用查询从源表和目标表检索行。如果源数据库与目标数据库属于不同的类型，列将转换为标准化的数据类型格式，以进行准确比较。

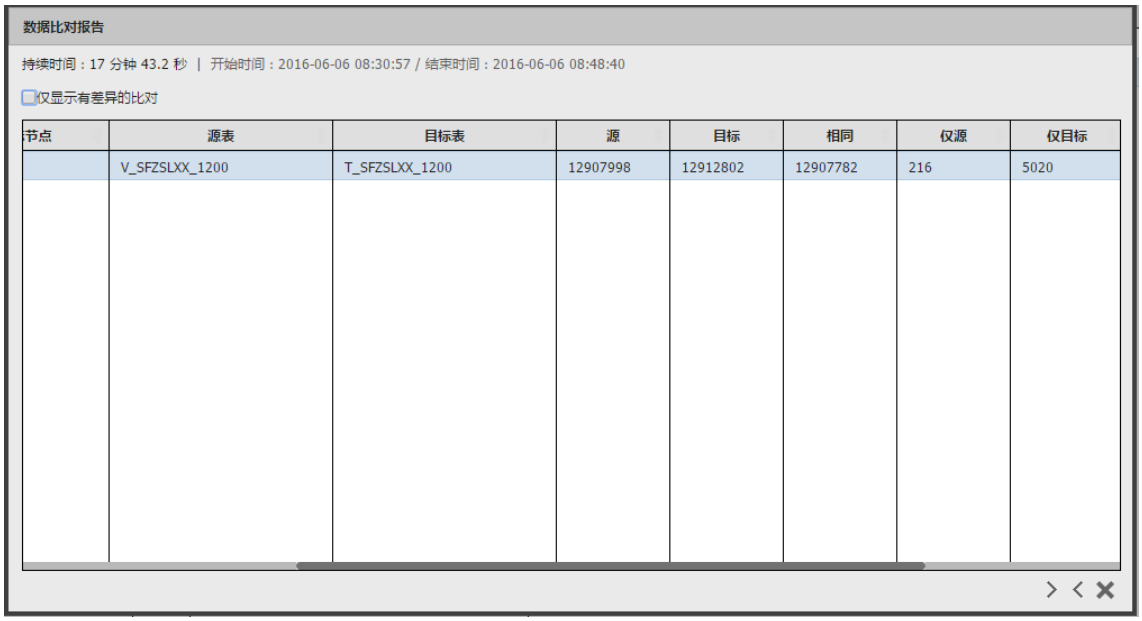
默认情况下，DataCompare 在比较行时，会以值一一对应的方式比较主键的所有列，而对所有非键列使用散列值。用于计算散列值的独特的数字签名缩小了通过网络传输以进行比较的数据，同时仍提供高效且高度可靠，但是并非绝对的机制来确定两行是包含相同还是不同的列值。

为完全确保发现不同步的行，DataCompare 可配置为以列一一对应的方式比较非键行，而不使用散列。全列比较会降低处理性能且下降程度与列数成比例，还会增加网络使用率，因此不建议作为最佳实践。

在实际复制环境中，DataCompare 完成初始比较后，将看起来不同步的行存储在队列

中。不确定的原因是复制与比较同时执行，因此，差异可能是在源系统上捕获但尚未应用于目标系统的进行中事务引起的。

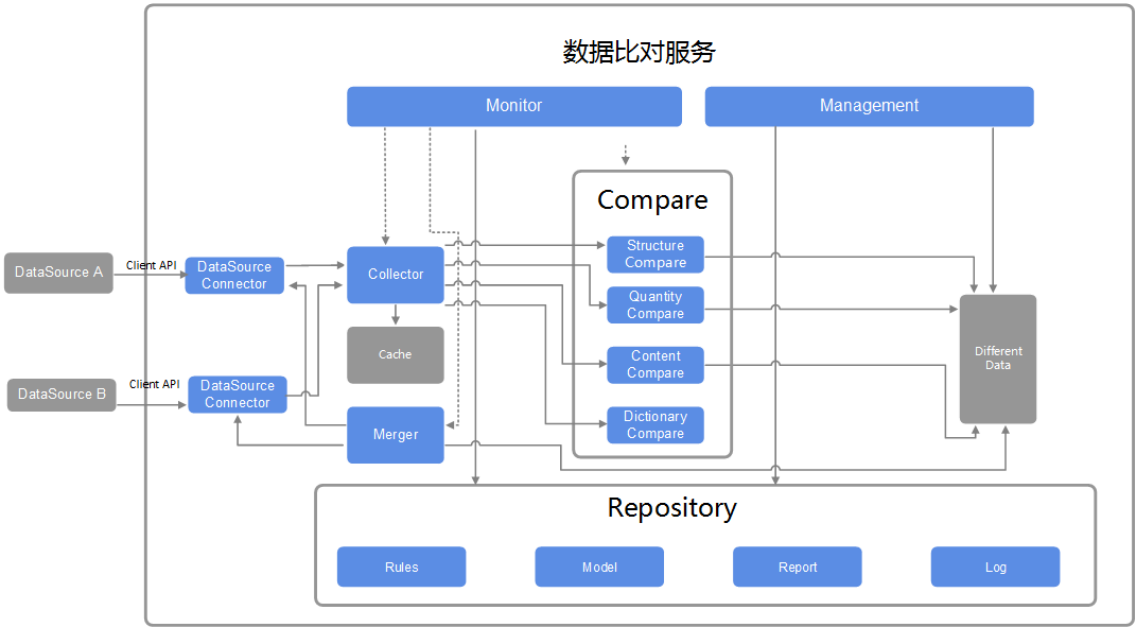
在确认步骤，也称为确认不同步步骤中，DataCompare 在不断变化的环境中确认行状态，从而确保结果准确。默认情况下，执行确认处理的线程与初始比较步骤并行执行，但每行的确认直到超出指定的复制延迟阈值后才执行。在这一步的最后，会给出对队列中的行的



评估结果，如下所示:

完成作业后，可以通过使用 SaCa DataCompare Web 界面或直接查看文件来查看比较报告和不同步报告。

五、技术架构



以上架构示意图展示了各个 SaCa DataCompare 组件的典型架构设置。箭头表示发起通信的起始位置。

Monitor&Management，使用 web 浏览器，用户可以连接到 DataCompare 服务器并配置 DataCompare 任务及其关联对象。在配置必需的 DataCompare 任务后，用户就可以开始进行比较以及查看报告了。

Compare 组件，是一个用于执行 DataCompare 任务的 Java 程序。DataCompare 使用 Repository 库中的配置信息来获得有关 DataCompare 任务的细节。用户可通过 Web 浏览器手动执行 DataCompare 任务。还可以使用 DataCompare 调度定期自动执行比较任务。DataCompare 服务器通常与源数据库和目标数据库安装在不同的机器上。

Repository 元数据库，用于存储 DataCompare 任务，调度等一系列配置信息，报告信息以及运行日志信息。

六、产品优势

- 完全基于 Web 方式配置、管理和监控
- 高速度、侵入性极低

- 支持异构数据库
- 高效处理大量数据
- 支持跨网络跨机房的比对
- 支持数据不断变化的实时数据库
- 源系统和目标系统无需停机
- 对硬件和网络资源的影响很低
- 提供有关不一致数据的详细、可指导行动的报告

七、客户示例

全国人口基本信息资源库，是在全国人口信息管理系统建设基础上，充分利用公安综合业务通信网络资源，建立部级人口信息管理系统和全国人口基本信息资源库，集中存储全国常住人口的文字和照片信息，为全国各级公安机关广大民警提供人口信息快速查询查证服务。

挑战

公安部信息部门在数据管理方面面临着多项挑战，其中包括确保人口基本信息数据的完整性。

公安部人口信息库的数据来源于所有的 31 个省数据中心，这就需要一种解决方案来发现公安部与省级的数据差异并在不停机的情况下在各个数据库之间重新同步数据，以确保一致，并避免因服务级别降低而导致收入损失的情况。

这些不一致的数据包括：

不同的记录，显示目标中键与源相同但数据与源不同的记录。

只在源中，显示目标中未出现的源中的记录。

只在目标中，显示源中未出现的目标中的记录。

相同的记录，显示目标中键和数据都与源相同的记录的。

解决方案

公安部数据中心机房部署了 SaCa DataCompare 产品，因而在数据不同步时能够快速、自动发现这种情况，且不会中断数据库可用性。不再需要运行复杂的查询并投入大量的人力资源以发现受影响的数据，在某些情况下可节省数周的时间。

东软的 IT 服务团队创建了一系列 SaCa DataCompare 任务，用以处理 31 个省/直辖市差异数据，定位差异的数据，以及自动修复数据环境，这进一步节省了时间和资源，从而扩展了 SaCa DataCompare 的价值。

优势

通过使用 SaCa DataCompare，公安部无需手动将数据从一个省级数据中心复制到公安部数据中心站点，也无需使某数据中心长时间地停机以从另一个数据中心重新初始化化。这样，就提高了系统可用性，降低了错误风险。

当从省级数据中心将数据迁移到部数据中心时，利用该解决方案高效地检测可能的错误和数据遗漏，从而简化了流程，避免了可能的数据丢失。

利用扩展解决方案的高级同步功能，无需重新实例化数据库，每小时可处理 1 亿多行的数据。

八、总结

在当今复杂的 IT 环境中，一个不幸的现实是可能会出现数据差异。如果不一致的数据未被发现并解决，则可能导致错误的决策、不满足服务级别协议要求，最终出现运营、财务和法律风险。

SaCa DataCompare 为在对业务造成负面影响之前发现不同步的数据提供了一个易用且强大的解决方案。SaCa DataCompare 与 SaCa CDC，SaCa DataExchange 等系列产品一起，提供了确保数据一致性的实时数据集成和持续可用性解决方案。

九、关于东软

东软创立于 1991 年，是中国领先的 IT 解决方案和服务提供商。公司主营业务包括：行业解决方案、产品工程解决方案及相关软件产品、平台及服务等。目前，拥有员工 20000 余名，在中国建立了 6 个软件研发基地，8 个区域总部，在 40 多个城市建立营销与服务网络；在大连、南海、成都和沈阳分别建立 3 所东软信息学院和 1 所生物医学与信息工程学院；在美国、日本、欧洲、中东设有子公司。

东软是中国第一家上市的软件企业，是第一家通过 CMM5 和 CMMI（V1.2）5 级认证的软件企业，是中国最大的离岸软件外包提供商。

东软将"超越技术"作为公司的经营思想和品牌承诺。作为一家以软件技术为核心的公司，东软通过开放式创新、卓越运营管理、人力资源发展等战略的实施，全面构造公司的核心竞争力，创造客户和社会的价值，从而实现技术的价值。

东软致力于成为最受社会、客户、投资者和员工尊敬的公司，并通过组织与过程的持续改进，领导力与员工竞争力的发展，联盟与开放式创新，使公司成为全球优秀的 IT 解决方案和服务提供商。

产品网站：http://platform.neusoft.com

技术社区：http://plus.neusoft.com

电话：400 655 6789

邮箱：platform@neusoft.com

微信：东软平台产品