

ArteryBase 逻辑复制（ARRS）技术白皮书

1 概述

逻辑复制（Logical Replication）的含义为：从数据库中按照业务的要求，挑选一些数据复制。

ARRS 的全称为：ArteryBase Realtime Replication Server，即：ArteryBase 数据库实时复制服务器。

ARRS 实现 ArteryBase 数据库中的数据以毫秒级延迟将源库部分表、部分行、部分列的数据复制到 ArteryBase 或者 Sybase ASE 等异构数据库中；并具备可扩展性，通过实现 JAVA API 接口，可将数据复制到其他数据库、文件、大数据平台等多种目标中。

2 ARRS 的地位与作用

一、随着国家基础软件自主可控战略的推进，在数据库领域逐渐将 Oracle、Sqlserver、Sybase ASE、DB2 等替换为拥有自主知识产权的数据库。为避免替换工作对业务产生影响，绝大多数客户都采取了逐步替换的模式。

为避免替换为 ArteryBase 数据库给客户带来影响，我们开发了 ARRS，使在 ArteryBase 数据库中产生的数据可以以毫秒级延迟的速率复制到“旧”数据库中，使“新”“旧”数据库可以并行运行一段时间，为客户实现数据库的平稳替换提供保证。

二、客户在实际的业务应用中，存在较多需求各异的数据实时复制需求，如：一个数据中心可能部署了多个 ArteryBase 数据库，部分数据需要在多个数据库间共享；跨广域网的不同机房的 ArteryBase 数据库之间可能需要复制部分数据；高安全等级网络中的数据库需要向低安全等级的数据库中复制符合安全要求的部分数据等。ARRS 通过提取部分表、部分行、部分列的数据，实现以上的复制需求。

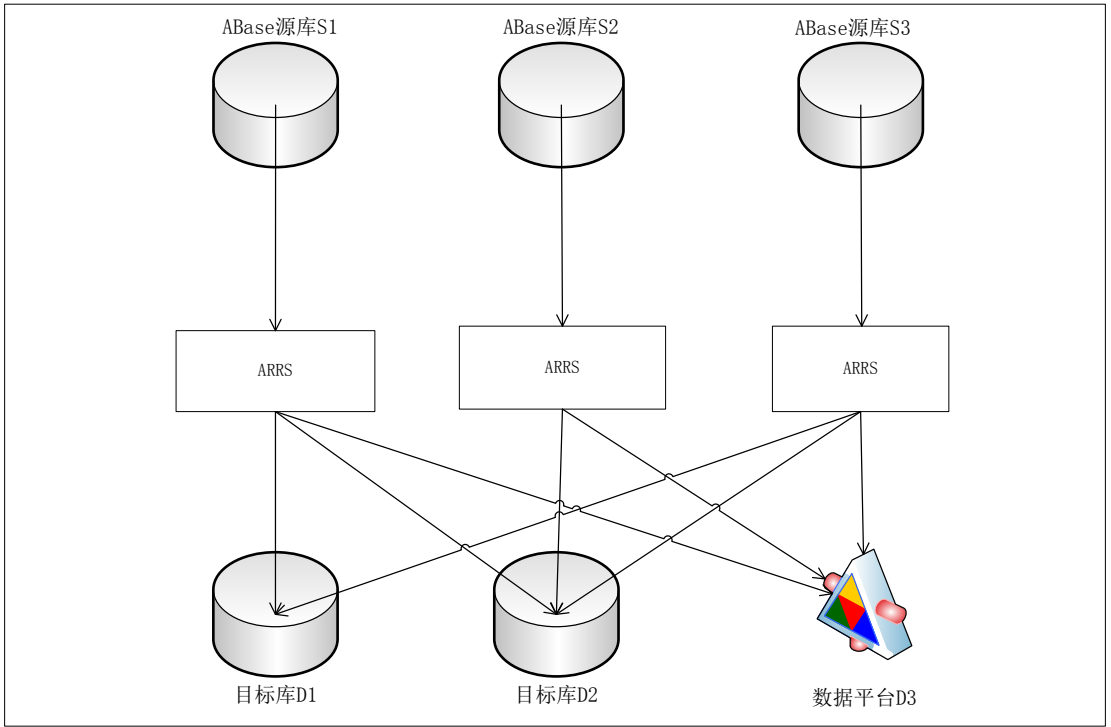
三、近年来数据挖掘、机器学习、AI、BI 等技术发展迅猛，客户期望能从多

年积累的历史数据中得到有价值的知识。研发人员通过实现 ARRS 提供的 JAVA API 接口，可将 ArteryBase 数据库中的数据按照不同需求复制到大数据平台、数据仓库、分布式文件系统。

3 架构

整体架构

逻辑复制包括 ArteryBase 源库、ARRS、目标库（大数据平台）。
一个 ArteryBase 源库需要一个 ARRS 实例，一个 ARRS 实例可以对应多个目标库。

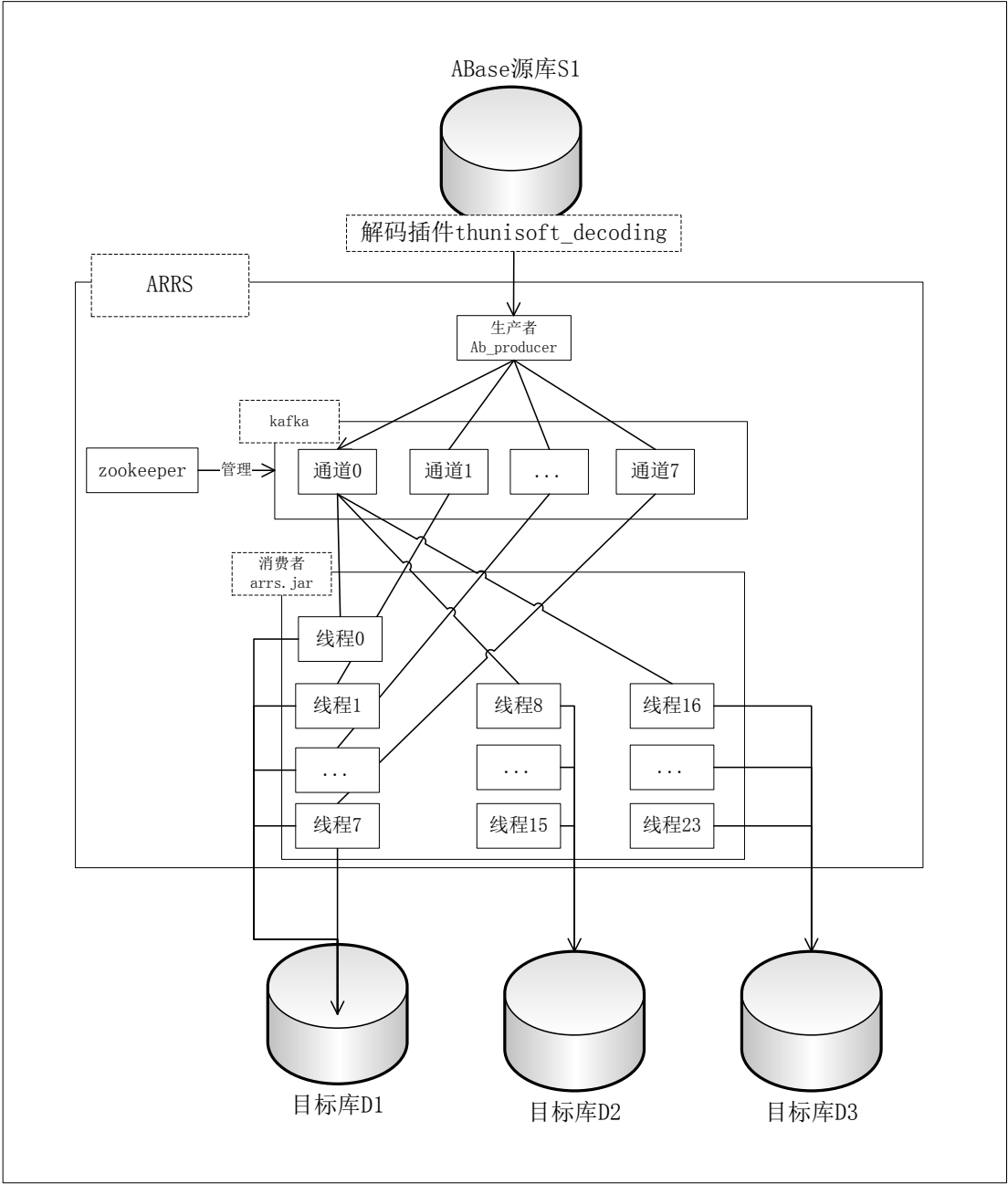


组件

ARRS 包括 6 个组件，分别是：解码插件、生产者、zookeeper、kafka、消费者、复制通道，除解码插件需要在源库部署外，其余 5 个组件作为一个整体部署在独立的服务器上。

设置多个复制通道的目的是为了提升数据在目标库的执行速率，多个复制通道可以并行执行数据库更新操作。

各组件的逻辑关系如下：

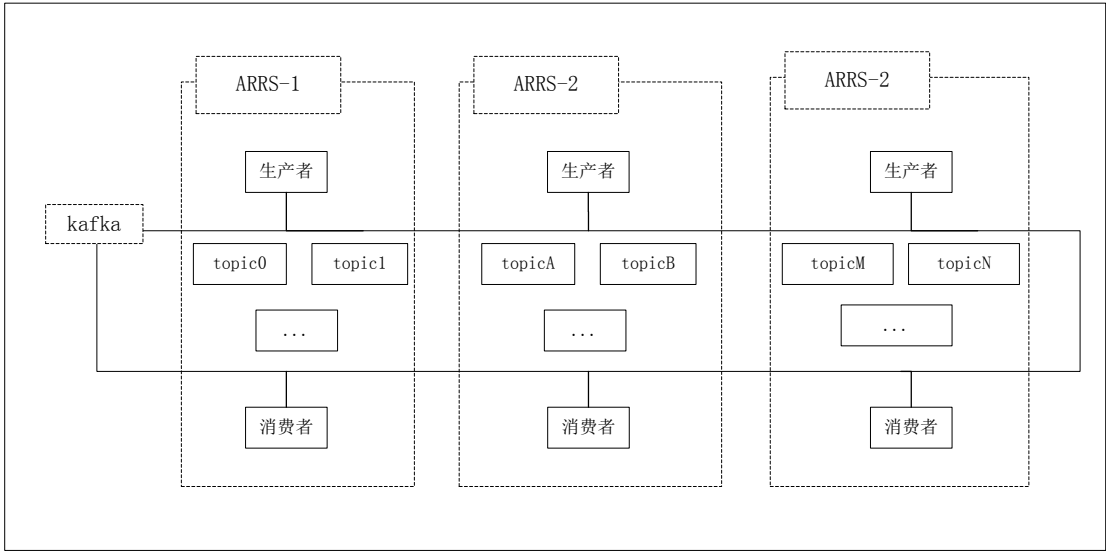


复制流程

当复制启动后，ArteryBase 数据库的解码插件从源库的事务日志中提取数据，由生产者接收处理后并推送到 kafka，消费者将 kafka 中的数据获得后，发送到目标端数据库并执行。

kafka 在 ARRS 中的地位

kafka 作为 ARRS 的消息队列，生产者将产生的消息推送到 kafka 的不同的 topic 中（每个复制通道都对应 kafka 一个单独的 topic），消费者从 kafka 对应的 topic 中取数据。部署在同一台服务器上的多个 ARRS 实例可以复用同一个 kafka。



4 复制速率

源库、目标库是普通的实体机（16 核+64GB 内存+RAID 10 机械硬盘），ARRS 8 个复制通道全开的情况下，复制速率可达 15 万个事务/分钟。每天可复制 2 亿个事务。如果需要更高的复制速率，源库、目标库应升级为固态硬盘。

源库每秒更新 750 条数据，当开启 4 个复制通道时，可实现毫秒级的复制延迟。（99.9%延迟在 10 毫秒以内，100%延迟在 100 毫秒以内。）

5 ArteryBase 源库限制

操作系统：	centos 6.5 64bit 及以上、windows 2008 R2 及以上
硬盘：	剩余磁盘空间 50G 以上
数据库版本	ArteryBase 3.6.1、ArteryBase 3.6.2
网络带宽：	千兆以上

6 ARRS 环境限制

操作系统:	centos 6.5 64bit 及以上
CPU:	4 核以上
内存:	剩余内存 8G 以上
硬盘:	剩余磁盘空间 100G 以上
网络带宽:	千兆以上