太平洋保险集团
“家园项目”大数据平台 DSG 应用
（oracle&kafka）

项目背景

根据太平洋保险集团的 IT 建设规划，在 2017 年年底，需要完成“一个太保，共同的家园”项目（简称家园项目），旨在给客户提供更加便携、全面的服务，通过一个家园平台，就能够完成所有的服务。

众所周知，太平洋保险的业务范围非常广泛，囊括了产险、寿险、车险等业务，同时，一个险种又由多个系统共同提供服务。现在要在一个平台上完成这些服务，数据的汇聚、集中、转换就成了整个项目的核心与难点。

项目需求

根据太保家园项目的最终目标，在一期建设中，需要将太平洋保险集团下属的寿险，产险，车险等 30 多个核心系统数据，通过实时同步复制的方式，统一集中到大数据平台。其中涉及数据的转换，标化，清洗，去重等一系列过程，具体需求如下：

1、 需要将核心数据从 30 多个系统同步到大数据平台的 kafka 组件。

2、 确保数据复制的实时性（秒级）和数据的准确性

3、 复制的数据需要加上时间，操作类型等标签，便于后端应用识别

4、 以生产环境的 dg 库作为数据的汇聚源端，减轻对生产库的影响

5、 入 kafka 的数据格式可灵活配置，以便更好的适配后端应用

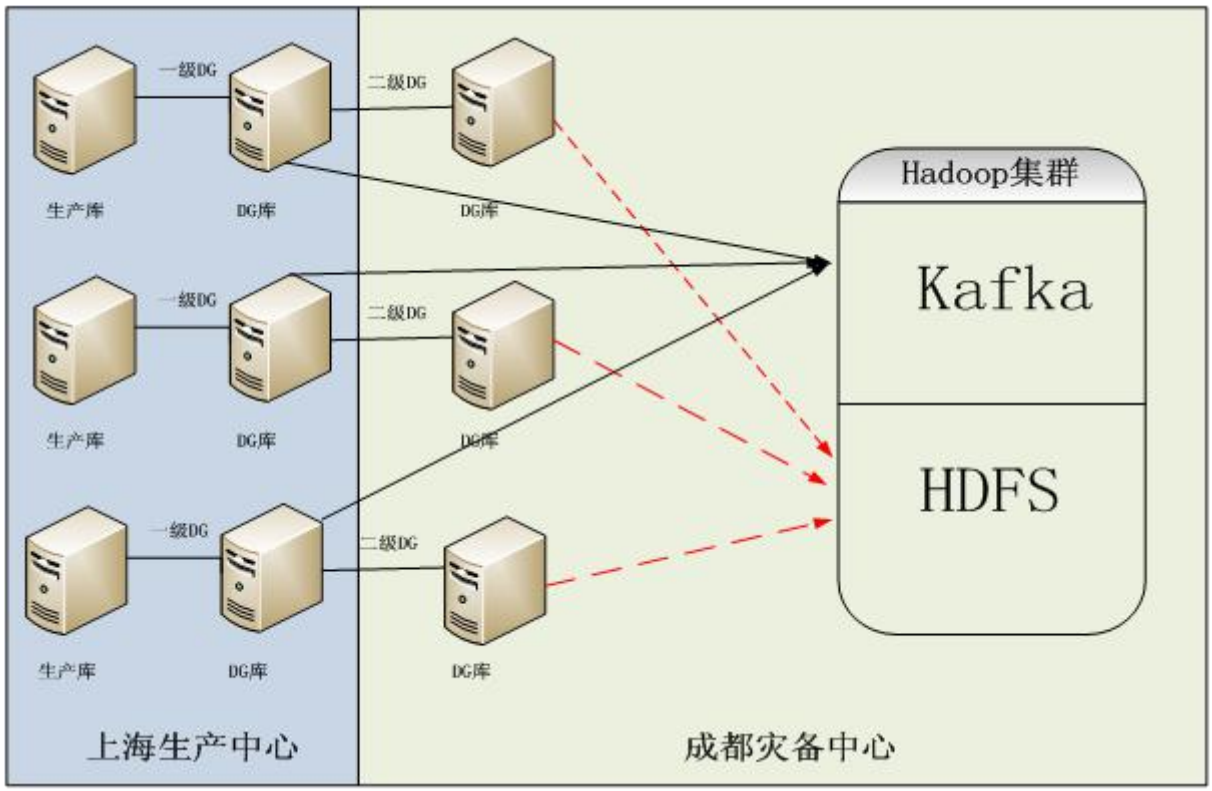
6、 需要具备数据操作统计和数据比对功能，便于核对数据的准确性

项目难点

在实现整个家园项目的数据汇聚中，根据项目需求和实际的生产环境情况，要完成整个数据同步，主要存在以下一些难点：

1. 涉及的业务系统众多。据初步规划，此平台需要接入的核心生产系统有30多个，既有oracle，也有mysql、db2等，每个系统的基础平台和数据格式千差万别，
2. 数据量大。目前整个平台需要的数据容量超过30T。并且源端业务系统是非常严格的7x24小时系统，这就给初始化带来很大的难度。
3. 网络带宽资源有限。生产环境数据都在上海数据中心，大数据平台在成都数据中心，中间的网络带宽是所有业务系统共用，因此不能过大占用带宽资源。
4. 业务量大。数据库每天的归档量均在800G以上，参与复制的核心表，每秒钟均有几百上千笔业务。
5. 延迟时间短。由于家园平台需要给客户提供实时的业务咨询与办理服务，复制的延迟不能超过10S，否则，用户的体验度大打折扣，违背家园项目建设的初衷。
6. 数据准确性要求高。家园平台承载着所有的查询、部分业务办理，如果数据不准确，必然引起业务逻辑混乱，无法为用户提供服务等问题。

解决方案



在此方案中，采用 DSG SuperSync 产品完成 oracle 到 kafka 的数据复制，方案架构如上图所示。在太保的系统架构中，生产中心位于上海，灾备中心位于成都。所有核心系统在本地生产中心均建有一级 DG 库，在成都灾备中心建有二级 DG 库。同时，此次项目的大数据中心也位于成都灾备中心。基于这种架构考虑，把数据量较大的全量同步放在成都的二级 DG 库上，这样可以节省上海到成都的带宽资源，同时提高同步效率。同时增量同步放在上海本地的一级 DG 库，以满足实时同步的要求。

方案优势

该方案具有以下优势：

1. 从架构层面，依赖于 DSG 产品对异构平台的完美支持，将全量数据同步到集群的 hdfs，增量数据同步到 kafka，很好地解决了两个数据中心的网络带宽资源有限的问题。

2. 为减轻生产库的压力，支持以生产库的 DG 库作为源端进行数据复制
3. 通过 cjson 模板，可高度自定义入 kafka 的数据格式
4. 可自定义输出数据内容，针对采集的数据可进行增删改操作后，投递到 kafka 中
5. 数据可校验。投递入 kafka 的数据，操作数据会通过明细，定时统计，累计统计三个维度进行记录，并把该记录定时存放在指定位置，例如数据库中，hdfs 中或者文件系统中，以便后续业务进行数据操作的回查，实现数据校验的功能。
6. DSG SuperSync 软件支持不同平台上的 Oracle 数据库之间的快速同步，包括首次数据同步和增量数据复制。DSG SuperSync 采用完全逻辑的方式进行数据同步，可以跨越不同平台；并且在数据同步过程中，采用了 DSG 独有的 XF1 文件格式、数据流压缩技术和快速数据抽取和装载技术。在配置多个同步通道的情况下，可以快速将现有数据库内的数据同步到目标数据库，并在其后将同步期间的增量数据一并复制到目标数据库实现数据追平。目前 DSG SuperSync 支持主流平台(HP/IBM/SUN/Comppaq/PC)上的 Oracle 各版本(Oracle8i - 10g)之间的数据复制。
7. DSG SuperSync 产品的数据复制效率，在该领域中是最高的。在 kafka 的投递端，可以采用多线程、多并发等方式进行加速投递，现场效率可以达到每秒 2 万条的

DSG 简介

DSG 是领先的致力于数据存储管理的专业厂商，提供优秀的大数据管理软件和数据安全、灾难恢复、数据抽取共享、数据归档检索和一体化管理平台在内的解决方案，产品包括：备份、容灾、数据同步复制/抽取/共享、数据归档、数据稽核等，在国内得到了广泛的应用。目前公司拥有员工近 300 余人，全国设有 3 个研发中心、20 多个办事处和分支机构，服务网点覆盖全国，在中国市场拥有数百家电信、金融和政府行业的高端用户。

SuperSync 数据同步复制软件应用，（国内 800 余家客户，在原有强大的 Oracle 的实时同步复制/灾备外，还可以支持 Mysql/Sql/DB2/PostgreSql/Hana/Qcubic/Redis/Teradata/浪潮 K-DB/达梦/南大 Gbase 等国内外各类数据库与 Hadoop、HBase、Phoneix、Storm、Flume、Spark、Kafka、tibc、阿里云间的实时同步复制，可根据 kafka 等格式需求定制（添加字段/数据转换/分类等），应用在大数据共享、读写分离和实时灾备等方面。