## 凤凰藏品数据库营销平台方案设计

### 概述

当今大部分企业决策者由于无法及时访问高质量、全面、可靠、个性化的运营信息，因而在制定关键企业决策时常常感到压力巨大。本方案考虑将凤凰藏品多个数据源的营销数据整合成一个统一的数据仓库，通过数据分析，将其作为指导营销策略制定、推动持续的业务流程改进的有效手段。

统一数据营销平台可满足企业范围内相关营销活动的商务智能需求，包括特定用户群的查询分析以及营销进度的跟踪管理。所有这些都通过一个个性化、交互式的智能信息显示界面来提供。

整个方案涉及到现有多数据源的信息抽取、转换清洗及加载（简称ETL）可以说是数据仓库最为复杂的过程。ETL引擎中，提供了对数据和元数据的数据质量、数据审计、完全集成关系和多维建模以及整个生命周期的管理。这里我们要根据系统特点建立ETL策略，如什么样的数据需要进行抽取、抽取完后如何进行审计和校正、数据的颗粒度有多高等。ETL 完成还要进行整个过程的监控及跟踪处理。

存放在数据库中的数据包括详细数据、汇总数据、计算数据、结果数据及元数据。数据存储是数据库系统的中心，取自多个数据源系统的明细数据，以及用于分析的集成汇总数据都存储在这个中心。随着数据库规模的增加，数据访问的性能依然要求保持较高水平，用户界面可以方便快捷地对数据进行查看、维护、管理。考虑到数据的安全性，要求定期对数据进行备份。

### 环境要求

Oracle 10g 数据库通过使用分区技术可以支持海量数据的存储，一个数据库最大数据量为8,000 PB（1PB=1024TB）。Oracle 10g 提供强大的并行处理能力。所以，考虑搭建一个满足Oracle 10g系统运行的物理环境，初始最大记录数在1亿，支持后续容量扩充。

|  |  |
| --- | --- |
| 服务器硬件要求 |  |
| CPU | Intel Pentium/Xeon系列处理器，主频>=2.8GHz，二级缓存>=1M  （或多核CPU具备同等性能） |
| 内存 | >=4GB |
| 硬盘 | 2\*1TB（考虑到数据安全性和备份，至少需要两个物理磁盘） |
| 光驱 | DVD-ROM |
| 网络 | 100Mbps以太网络 |

|  |  |
| --- | --- |
| 服务器软件要求 |  |
| 操作系统 | SUSE Linux Enterprise Server 11 |
| 数据库系统 | Oracle 10g R2 Enterprise |
| 工具 | PL/SQL数据库调试工具 |
| ETL运行环境 | Java SDK Version 6 |

### 关键技术

数据库平台的设计包含以下关键技术：多数据源识别、表/索引分区以及B/S架构展示界面。

#### 3.1 多数据源识别

多数据源识别要求从Excel和Access两种数据文件中提取数据，并整理为统一数据格式。

公司的数据分散在大量的EXCEL、ACCESS 和TXT文本文件中，现在为了统一管理，统一储存，能基本实现数据挖掘功能，需要将各种文件的数据提取出来，并把数据导入到ORACLE中。经过初步调研，公司目前的数据格式差别很大，各个文件的表头基本不同，如果利用目前市面上现存的ETL工具，无法正确抽取数据入库，需要自行研发ETL工具，来现实数据的抽取入库。



ETL 主要分为导出模块和导入模块，导出模块主要负责从三种数据源中提取数据出来组装成入库对象；导入模块主要负责将入库对象导入到ORACLE数据库中。导出模块主要从ACCESS ,EXCEL，TXT文件中导出数据，经过对公司各种文件分析，公司的各种文件表头字段差别很大，表现在以下方面：

1、表与表之间列的数目不同，有的表有很多列，包含的信息量大，有的表只有少量列，包含的信息少。

2、绝大多数表，列名不能描述该列存储的信息，如字段1，字段2

3、有些表，列名能基本表达该列存储的信息，但是对同一个信息，却有好几种不同的描述，如同是对姓名的描述，有 联系人 、name、等等描述

4、有些信息残缺：一行信息中，有些记录是空白的。

5、信息存储的多样化：如姓名，有些不光有姓名，后面还有某某经理；如手机号码：有的手机号码以138- 139- 等方式；有些表格中存有多个电话号码。

数据大而杂，给导出模块提出了很大的技术难度，对于此，提出两种解决方案：

**方案一**：由公司统一规范各个表的列名，ETL根据列名抽取数据组装成正确的入库对象。

该方案缺点：需要公司统一规范各个列的表名，耗费人力

优点：能提取的数据质量高

**方案二**：由ETL智能提取，每读一行，ETL来自动识别电话号码，手机号码，邮箱，邮编，地址，单位，姓名

该方案缺点：ETL对姓名，单位，地址的识别难度很高，目前业界也没有成熟的经验，属于人工智能的范畴，只能根据长度，以及关键字来模糊匹配,同时需要公司更多的数据来统计分析，提取出来的数据质量不能保证。

优点：不需要公司投入人力。

导入模块主要将导出模块的入库对象导出到ORACLE中，由于一些信息的残缺，很难能保证入库对象的每一个字段都有信息，需要公司给出入库准则，以便程序检查入库，对不符合入库条件的记录，则不将其入库。

入库前需要检查是否有重复，这个也需要公司给出重复准则，入库前在oracle数据中查找看是否有重复，如果重复，则不新增oracle行，将信息补充到查找到的行中；如果没有重复，则直接入库。

#### 3.2 表/索引分区

Oracle 数据分区技术将大表和索引分成可以管理的小块，从而避免了对每个表作为一个大的、单独的对象进行管理。分区(partitioning)是一种“分而治之”的技术，它为海量数据管理提供了可扩展的性能。

分区通过将操作分配给更小的存储单元，减少了需要进行管理操作的时间。分区对应用是透明的，可以通过标准的SQL 语句对分区表进行操作。通过“分区忽略”技术，Oracle数据库的优化器在访问数据时会分析数据的分区情况，在进行查询时，那些不包含任何查询数据的分区将被忽略，从而大大提高系统的性能。

分区技术还提高了数据的可用性。当部分数据由于故障或其它原因不可用时，其它分区内的数据可以不受影响继续使用。Oracle 10g 数据库企业版支持多种数据分区方式，包括：范围分区(Range Partition)、列表分区(List Partition)、散列分区(Hash Partition)，以及以上各种分区组合的复合分区(Composite Partition)。在Oracle 10g 支持的所有操作系统版本中均支持以上各种数据分区方式。

**表分区有以下优点：**

1、数据查询：数据被存储到多个文件上，减少了I/O负载，查询速度提高。

2、数据修剪：保存历史数据非常的理想。

3、备份：将大表的数据分成多个文件，方便备份和恢复。

4、并行性：可以同时向表中进行DML操作，并行性性能提高。

**分区表设计原则：**

* 表大小超过2GB，或表记录超过1000万时，应该考虑对表进行分区
* 基于表的大部分查询应用，只访问表中少量的数据，对于这样的表进行分区，可利用分区排除无关数据查询的特性。
* 按时间段删除成批的数据，对于这样的表需要考虑进行分区，以满足维护需要
* 按时间周期进行表空间备份时，将分区与表空间建立对应关系
* 如果一个表中大部分数据都是只读，通过对表进行分区可将只读数据存储在只读表空间，对于数据的备份有利

**表分区种类：**

1）Range分区

最早、最经典的分区方法，通过对字段值的范围进行分区，适用于按时间周期进行数据的存储。数据管理能力强，范围分区的数据可能不均匀，与记录值有关，可维护性差。

举例：CREATE TABLE sales\_range(salesman\_id NUMBER(5), salesman\_name VARCHAR2(30), sales\_amount NUMBER(10), sales\_date DATE) COMPRESS PARTITION BY RANGE(sales\_date) (PARTITION sales\_jan2000 VALUES LESS THAN(TO\_DATE(‘02/01/2000’, ’DD/MM/YYYY’)), PATRIRION sales\_feb2000 VALUES LESS THAN(TO\_DATE(‘03/01/2000’, ‘DD/MM/YYYY’)));

2）HASH分区

基于分区字段的HASH值，自动将记录录入到相应的分区，分区一般是2的幂，易于实施，总体性能最佳，适合于静态的数据。HASH分区适合于数据的均匀存储。

举例：CREATE TABLE sales\_range(salesman\_id NUMBER(5), salesman\_name VARCHAR2(30), sales\_amount NUMBER(10))PARTITION BY HASH(salesman\_id) PARTITIONs 4;

3）LIST分区

通过对分区字段的离散值进行分区，分区之间没有关联，分区适合于对数据离散值进行控制，只支持单个字段，优缺点与RANGE类似。

举例：CREATE TABLE sales\_list(salesman\_id NUMBER(5), salesman\_name VARCHAR2(30), sales\_state VARCHAR2(20), sales\_amount NUMBER(10), sales\_date DATE)PARTITION BY LIST(sales\_state)(PARTITION sales\_west VALUES(‘California’, ‘Hawaii’),PARTITION sales\_east VALUES…

4）Composite分区

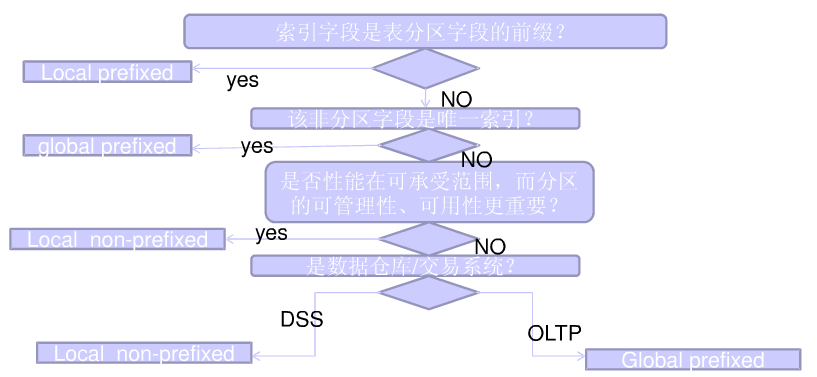
支持range-hash，Range-list分区，既适合于历史数据，又适合于数据均匀分布，粒度更细，支持local indexes不支持global indexes。

举例：PARTION BY RANGE(txn\_date) SUBPARTITION BY LIST(state)

**索引分区：**

Local index（同表分区），Global index（与表分区不同），Non partition index。Local index提供了更好的可用性，大数据量时，能充分利用local index的并行查询能力，在性能、高可用和可管理之间平衡。

索引分区选择策略



#### 3.3 B/S架构展示界面

基于B/S架构可以实现瘦客户端的设计，服务器运行Web Server，客户端向Web Server发送HTTP请求、接收HTTP响应，完成数据的查询、维护过程。

待补充。

### 实施方案

数据库完整网络架构图如下。Oracle服务器管理数据库；ETL客户端从Excel和Access文件中提取用户信息，并通过Oracle客户端与Oracle服务器进行交互，完成用户信息的入库；本地业务客户端通过本地以太网与Oracle服务器的Web服务器通信，完成数据库的查询、维护操作；远程业务客户端通过Internet与Oracle服务器的Web服务器通信，完成数据库的查询、维护操作。

图. 数据库网络架构设计

#### 4.1 环境搭建

环境搭建的主要工作是软硬件平台的安装。

硬件服务器环境必须已经具备。如果硬盘条件不满足（2\*1TB），则需要额外安装硬盘。如果网络条件不满足（100MBps），则需要额外安装局域网络。为了方便Linux操作系统的安装，服务器需要具备DVD-ROM。

软件安装主要是Linux和Oracle。Suse Linux 11g可以从Internet上下载刻录成DVD，然后在服务器上安装。Oracle安装文件也可以从Internet上下载，可以用U盘安装，安装过程中务必选择企业版。

安装前准备如下：

1. 建立用户

groupadd dba –g 1000

useradd oracle –g dba –u 1000 –d /oracle –c “Oracle”

passwd oracle

mkdir –p /oracle/products/10g

mkdir –p /oracle /orains/patch

cd /

chown –R oracle:dba oracle

1. 上传安装文件

安装文件上传到/oracle/orains，patch文件上传到/oracle/orains/patch

1. Oracle 用户解压

gunzip ..

cpio –idmv < ..

cd patch

unzip ..

1. Root登录安装

xhost +

xdpyinfo|grep display

su – oracle

DISPLAY=:0.0

export DISPLAY

cd /orains/database/install

vi oraparam.ini 修改使之支持suse11

cd..

./runInstaller

选择Advanded installation，/oracle/oraInventory dba,dba, Custom, OUIHome, /oracle/products/10g, Software only

1. 用户登录

服务器上启动监听服务，客户端上安装Oracle客户端，可以访问数据库。

#### 4.2 数据源统计与分析

统计所有文件的表头、格式、内容，确认是否满足入库条件。设计智能匹配算法，在没有表头的前提下，也可以提取用户名称、电话号码、用户地址、邮箱等关键信息用于入库。

提取规则需要在分析过程中不断完善，直至所有数据源文件分析完成。数据源文件提取是否完整，需要人工验证，对于错误提取的信息或者遗漏的信息，通过完善规则的方式修正。最终保证所有数据源可以在统一的提取规则下完整地、正确地导入。

#### 4.3 建立数据库

表1 客户信息表customer\_profile，表中的字段设计如下。考虑到节省存储资源，字段尽量设计为变长；客户号为10位数字，支持用户数100亿；数据来源表示导入的数据文件ID，不允许为空。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 字段类型 | 是否自增长 | 是否主键 | 是否可以为空 | 说明 |
| customer\_id | number(10) | 是 | 是 | 否 | 客户号 |
| customer\_name | Varchar2(20) | 否 | 否 | 是 | 客户姓名 |
| customer\_addr | Varchar2(200) | 否 | 否 | 是 | 客户地址 |
| customer\_phone | Varchar2(20) | 否 | 否 | 是 | 客户电话 |
| customer\_mail | Varchar2(50) | 否 | 否 | 是 | 客户邮箱 |
| customer\_gender | Char(2) | 否 | 否 | 是 | 客户性别 |
| customer\_birth | Char(10) | 否 | 否 | 是 | 客户生日 |
| customer\_age | number(3) | 否 | 否 | 是 | 客户年龄 |
| cutomer\_job | Varchar2(50) | 否 | 否 | 是 | 客户单位 |
| customer\_idcard | Varchar2(20) | 否 | 否 | 是 | 身份证 |
| customer\_source | number(10) | 否 | 否 | 否 | 数据来源 |
| customer\_info | Varchar2(200) | 否 | 否 | 是 | 备注信息 |

表2 通讯详情表 session\_content，业务发生次数最多支持100亿，状态标记0无效，1有效。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 字段类型 | 是否自增长 | 是否主键 | 是否可以为空 | 说明 |
| session\_id | number(10) | 是 | 否 | 否 | 业务号 |
| customer\_id | number(10) | 否 | 否 | 否 | 客户号 |
| ack\_flag | number(1) | 否 | 否 | 否 | 响应状态 |
| deal\_flag | number(1) | 否 | 否 | 否 | 定购状态 |

表3 通讯统计表 session\_statistics，业务发生次数最多支持100亿，短信内容最多支持200个中文字符。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 字段类型 | 是否自增长 | 是否主键 | 是否可以为空 | 说明 |
| session\_id | number(10) | 是 | 是 | 否 | 业务号 |
| session\_date | Date | 否 | 否 | 否 | 业务日期 |
| msg\_info | Varchar2(400) | 否 | 否 | 是 | 短信内容 |
| send\_num | number(10) | 否 | 否 | 否 | 通信量 |
| ack\_num | number(10) | 否 | 否 | 否 | 响应人数 |
| Deal\_num | number(10) | 否 | 否 | 否 | 订购人数 |

表4 ETL导入状态表etl\_manger，文件个数支持100亿，文件类型1-excel，2-access，3-txt，文件状态1-未导入，2-已导入，3-部分导入，4-验证不通过，5-验证通过。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 字段类型 | 是否自增长 | 是否主键 | 是否可以为空 | 说明 |
| customer\_source | number(10) | 是 | 是 | 否 | 数据来源 |
| file\_name | Varchar2(100) | 否 | 否 | 否 | 文件名 |
| file\_type | Number(1) | 否 | 否 | 否 | 文件类型 |
| file\_state | Number(1) | 否 | 否 | 否 | 文件状态 |
| customer\_num | Number(10) | 否 | 否 | 否 | 客户数目 |
| import\_time | Date | 否 | 否 | 否 | 导入时间 |

#### 4.4 数据源导入与验证

数据源完成导入后需要从两方面进行数据验证：

1. 软件验证：新增的客户记录数目等于数据源中包含的记录数目，记录内容匹配完全一致。验证工作可以定时、定量进行，保证数据库的数据是准确的。
2. 人工验证：未防止软件Bug和程序运行异常给数据库带来破坏，人工验证也是很有必要的，但数据量较大，人工验证可以采用随机抽取方式。人工验证的准确性也可以在一定统计意义上代表数据库内数据的正确性。

#### 4.5 数据库性能调优

1）采用建立索引的方式，优化关键字的查询性能。

2）调整Oracle的SGA参数，达到数据库访问性能的最佳方式。

3）采用多磁盘，并行I/O方式，减少数据交互时的瓶颈。

#### 4.6 数据备份

数据库需要定时（每周）导出一次到备用磁盘上，防止磁盘故障带来的数据丢失。

#### 4.6 界面设计

待补充。