本部分的题目来自Kimball的ETL Toolkit著作，原著未直接给出答案。这里的中文题目和答案是我参考其原著按自己的理解整理而来的，仅供参考。对于其中不确切的地方，欢迎大家一起沟通。有兴趣的朋友可以直接阅读原著。

分析

1. 什么是逻辑数据映射？它对ETL项目组的作用是什么？

逻辑数据映射（Logical Data Map）用来描述源系统的数据定义、目标数据仓库的模型以及将源系统的数据转换到数据仓库中需要做操作和处理方式的说明文档，通常以表格或Excel的格式保存信息.

逻辑数据映射应该贯穿数据迁移项目的始终，在其中说明了数据迁移中的ETL策略。在进行物理数据映射前进行逻辑数据映射对ETL项目组是重要的，它起着元数据的作用。项目中最好选择能生成逻辑数据映射的数据迁移工具。

2．在数据仓库项目中，数据探索阶段的主要目的是什么？

逻辑数据映射进行之前，需要首先对所有的源系统进行分析。对源系统的分析 通常包括两个阶段，一个是数据探索阶段（Data Discovery Phase），另一个是 异常数据检测阶段。 数据探索阶段包括以下内容： 1．收集所有的源系统的文档、数据字典等内容。 2．收集源系统的使用情况，如谁在用、每天多少人用、占多少存储空间等内容。 3．判断出数据的起始来源（System-of-Record）。 4．通过数据概况（Data Profiling）来对源系统的数据关系进行分析。 数据探索阶段的主要目的是理解源系统的情况， 为后续的数据建模和逻辑数据映 射打下坚实的基础。

1. 如何确定起始来源数据？

答：这个问题的关键是理解什么是System-of-Record。System-of-Record和数据仓库领域内的其他很多概念一样，不同的人对它有不同的定义。在Kimball的体系中，System-of-Record是指最初产生数据的地方，即数据的起始来源。在较大的企业内，数据会被冗余的保存在不同的地方，在数据的迁移过程中，会出现修改、清洗等操作，导致与数据的起始来源产生不同。起始来源数据对数据仓库的建立有着非常重要的作用，尤其是对产生一致性维度来说。我们从起始来源数据的越下游开始建立数据仓库，我们遇到垃圾数据的风险就会越大。

架构

4．在ETL过程中四个基本的过程分别是什么？

答： Kimball 数据仓库构建方法中，ETL 的过程和传统的实现方法有一些不同，主要 分为四个阶段，分别是抽取（extract）、清洗（clean）、一致性处理（comfo rm）和交付（delivery），简称为 ECCD。 1．抽取阶段的主要任务是： 读取源系统的数据模型。 连接并访问源系统的数据。 变化数据捕获。 抽取数据到数据准备区。 2．清洗阶段的主要任务是： 清洗并增补列的属性。 清洗并增补数据结构。 清洗并增补数据规则。 增补复杂的业务规则。 建立元数据库描述数据质量。 将清洗后的数据保存到数据准备区。 3．一致性处理阶段的主要任务是： 一致性处理业务标签，即维度表中的描述属性。 一致性处理业务度量及性能指标，通常是事实表中的事实。 去除重复数据。 国际化处理。 将一致性处理后的数据保存到数据准备区。 4．交付阶段的主要任务是： 加载星型的和经过雪花处理的维度表数据。 产生日期维度。 加载退化维度。 加载子维度。 加载 1、2、3 型的缓慢变化维度。 处理迟到的维度和迟到的事实。 加载多值维度。 加载有复杂层级结构的维度。 加载文本事实到维度表。 处理事实表的代理键。 加载三个基本类型的事实表数据。 加载和更新聚集。 将处理好的数据加载到数据仓库。 从这个任务列表中可以看出，ETL 的过程和数据仓库建模的过程结合的非常紧 密。换句话说，ETL 系统的设计应该和目标表的设计同时开始。通常来说，数据 仓库架构师和 ETL 系统设计师是同一个人。

5．在数据准备区中允许使用的数据结构有哪些？各有什么优缺点？

 1.在数据抽取阶段，我们需要连接到源系统，为了对源系统的影响尽量小，我们需要将抽取的数 据保存成文本文件或者放入数据准备区的表中，这样，当 ETL 过程出现错误而失败时，我们就可 以从这些文本文件开始 ETL，而不需要再次影响源系统。这种存储数据的方式叫数据准备区。 1) 固定格式的文本文件。Flat File 指的是一种保存在系统上的一种文本文件格式，它以类似数 据库的表的方式用行和列来保存数据。这种文件格式经常用来进行数据交换。用于保存数据 不太合适。 2) XML 数据集。多用于数据交换，用户保存数据不太合适。 3) 关系数据库的表。保存数据的较理想选择。 4) 独立的数据库表。独立的数据库表一般指建立的表和其他表没有外键约束关系。这样的表 多用于数据处理。 5) 三范式或者关系型模型。 6) 非关系型数据源。非关系型数据源一般包括 COBOL copy books、VSAM 文件、Flat 文件、S preadsheets 等。 7) 维度模型。 8) 原子事实表和聚集事实表。 9) 代理键查找表。

6．简述ETL过程中哪个步骤应该出于安全的考虑将数据写到磁盘上？

Staging的意思就是将数据写到磁盘上。出于安全及ETL能方便重新开始在数据准备区Staging Area中的每个步骤中都应该将数据写到磁盘上即生成文本文件或者将建立关系表保存数据而不应该以数据不落地方式直接进行ETL。 例如在数据抽取阶段我们需要连接到源系统为了对源系统的影响尽量小我们需要将抽取的数据保存成文本文件或者放入数据准备区的表中这样当ETL过程出现错误而失败时我们就可以从这些文本文件开始ETL而不需要再次影响源系统。

抽取

7．简述异构数据源中的数据抽取技术。

答 在数据仓库项目中需要抽取的数据经常来自不同的数据源它们的逻辑结构和物理结构都可能不同即称之为异构数据源。 在对异构数据源进行整合抽取时我们需要做的事情依次是标识出所有的源系统对源系统进行概况分析定义数据匹配逻辑建立筛选规则生成一致性维度。 对于源数据的操作系统平台和数据平台各不相同的情况我们需要根据实际情况来确定如何进行数据抽取通常的方法有建立ODBC连接、定义接口文件、建立DBLINK等方法。

8．从ERP源系统中抽取数据最好的方法是什么？

 答ERP系统的产生是为了解决企业内异构数据的整合。这个问题也是数据仓库系统面临的主要问题。ERP的解决方案是将企业内的各个应用包括销售、会计、人力资源、库存和产品等建立在相同的平台和相同的应用框架下即在应用操作层将企业内的数据进行了一致性处理。而数据仓库是在应用操作层之上建立一致性的规则并进行一致性处理。目前比较流行的ERP系统有SAP、PeopleSoft、Oracle、Baan和J.D.EDwards大部分没接触过。 如果企业内只有一套ERP系统那么数据就已经是一致的了为数据抽取提供了方便。如果企业内除了ERP外还有其他系统则数据抽取会变得复杂。因为目前的ERP系统的数据模型都非常复杂可能有几百几千个表并且较难理解。直接在ERP系统上建立数据捕获和抽取是非常复杂的。最好的办法是购买能针对ERP系统数据抽取提供功能的ETL工具将ERP内部的复杂性留给ETL厂商处理。

9．简述直接连接数据库和使用ODBC连接数据库进行通讯的优缺点。

答 通常连接数据库的方式分为两类,一类是直接连接另一类是ODBC连接。 直接连接的方式主要是通过COBOL、PL/SQL、Transact-SQL等方式连接数据库。这种方式的优点是运行性能高可以使用DBMS提供的一些特殊功能。缺点是通用性差。 ODBC是为windows应用程序访问数据库提供的一组接口。ODBC的优点是灵活性通过改变驱动和连接方式可以使用不同的数据库。ODBC方式的缺点是性能差。使用ODBC连接方式实现ETL的话在ETL程序和至少要有两层分别是ODBC Manager层和ODBC Driver层。另外使用ODBC方式不能使用DBMS提供的一些特殊的功能。

10．简述出三种变化数据捕获技术及其优缺点。

 答 变化数据捕获CDC技术是ETL工作中的重点和难点通常需要在增量抽取时完成。实现变化数据捕获时最理想的是找到源系统的DBA。如果不能找到就需要ETL项目组自己进行检测数据的变化。下面是一些常用的技术。 A、采用审计列 审计列指表中如“添加日期”、“修改日期”、“修改人”等信息的字段。应用程序在对该表的数据进行操作时同时更新这些字段或者建立触发器来更新这些字段。采用这种方式进行变化数据捕获的优点是方便容易实现。缺点是如果操作型系统没有相应的审计字段需要改变已有的操作型系统的数据结构以保证获取过程涉及的每张表都有审计字段。 B、数据库日志 DBMS日志获取是一种通过DBMS提供的日志系统来获得变化的数据。它的优点是对数据库或访问数据库的操作系统的影响最小。缺点是要求DBMS支持并且对日志记录的格式非常了解。 C、全表扫描 全表扫描或者全表导出文件后进行扫描对比也可以进行变化数据捕获尤其是捕获删除的数据时。这种方法的优点是思路清晰适应面广缺点是效率比较差。

数据质量

11．数据质量检查的四大类是什么？为每类提供一种实现技术。

答数据质量检查是ETL工作中非常重要的一步主要关注一下四个方面。 1正确性检查Corret 检查数据值及其描述是否真实的反映了客观事务。例如地址的描述是否完全。 2明确性检查Unambiguous 检查数据值及其描述是否只有一个意思或者只有一个解释。例如地名相同的两个县需要加区分方法。 3一致性检查Consistent 检查数据值及其描述是否统一的采用固定的约定符号来表示。例如币别中人民币用CNY。 4完全性检查Complete 完全性有两个需要检查的地方一个是检查字段的数据值及其描述是否完全。例如检查是否有空值。另一个是检查记录的合计值是否完全有没有遗忘某些条件。

12．简述应该在ETL的哪个步骤来实现概况分析？

答数据概况分析是对源数据内容的概况进行分析应该在项目的开始后尽早完成它会对设计和实现有很大的影响。在完成需求收集后就应该立即开始数据概况分析。 数据概况分析不光是对源系统的数据概况的定量描述而且为ETL系统中需要建立的错误事件事实表Error Event Table和审计维度表Audit Dimension打下基础为其提供数据。

13．ETL项目中的数据质量部分核心的交付物有那些？

答ETL项目中数据质量部分的核心的交付物主要有下面三个 1数据概况分析结果 数据概况分析结果是对源系统的数据状况的分析产物包括如源系统中有多少个表每个表有多少字段其中多少为空表间的外键关系是否存在等反映源系统数据质量的内容。这些内容用来决定数据迁移的设计和实现并提供给错误事件事实表和审计维度表需要的相关数据。 2错误事件事实表 错误事件事实表及相关的一系列维度表是数据质量检查部分的一个主要交付物。粒度是每一次数据质量检查中的错误信息。相关维度包括日期维度表、迁移信息维度表、错误事件信息维度表其中错误事件信息维度表中检查的类型、源系统的信息、涉及的表信息、检查使用的SQL等内容。错误事件事实表不提供给前台用户。 3审计维度表 审计维度表是给最终用户提供数据质量说明的一个维度表。它描述了用户使用的事实表的数据来源数据质量情况等内容。

14．如何来量化数据仓库中的数据质量？

 答在数据仓库项目中通常通过不规则数据的检测工作Anomaly Detection来量化源系统的数据质量。除非成立专门的数据质量调查项目组否则这个工作应该由ETL项目组完成。通常可以采用分组SQL来检查数据是否符合域的定义规则。 对于数据量小的表可以直接使用类似下面的SQL完成。 select state count from order\_detail group by state 对于数据量大的表一般通过采样技术来减少数据量然后进行不规则数据检测。类似SQL如下。 select a. from employee a select rownum counter a. from employee a B where a.emp\_id b.emp\_id and modb.counter truncselect count from employee/10000 0 如果可以采用专门的数据概况分析工具进行的话可以减少很大的工作量。

建立映射

15．什么是代理键？简述代理键替换管道如何工作。

在维度表的迁移过程中，有一种处理方式是使用无意义的整型值分配给维度记录并作为维度记录的主键，这些作为主键的整型值称为代理键（Surrogate Key）。使用代理键有很多好处，如隔离数据仓库与操作环境，历史记录的保存，查询速度快等。

同时，在事实表的迁移过程中，为了保证参照完整性也需要进行代理键的替换工作。为了代理键替换的效率高一些，我们通常在数据准备区中建立代理键查找表（Surrogate Mapping Table or Lookup Table）。代理键查找表中保存最新的代理键和自然键的对应关系。在对事实表进行代理键替换时，为了保证效率高，需要把代理键查找表中的数据加载到内存中，并可以开多线程依次替换同一记录的中的不同代理键，使一条事实记录在所有的代理键都替换完后再写如磁盘中，这样的替换过程称为代理键替换管道（Surrogate Key Pipeline）。

16．为什么在ETL的过程中需要对日期进行特殊处理？

答：在数据仓库的项目中，分析是主导需求，而基于日期和时间的分析更是占了很大的比重。而在操作型源系统中，日期通常都是SQL的DATETIME型的。如果在分析时，使用SQL对这种类型的字段临时处理会出现一些问题，如效率很差，不同的用户会采用不同的格式化方法导致报表不统一。所以，在数据仓库的建模时都会建立日期维度表和时间维度表，将用到的和日期相关的描述都冗余到该表中。

但是，并不是所有的日期都被转化为日期维度表的外键。日期维度表中的记录是有限的，有些日期如生日等可能会比日期维度表中记录的最小日期还要早，这类字段可以直接在数据仓库中保存SQL的DATETIME型。而像购买日期等与分析的业务紧密相关的通常都需要转化为日期维度表的外键，可以用日期维度表中统一的描述信息进行分析。

17．简述对一致性维度的三种基本的交付步骤。

答：数据整合的关键就是生成一致性维度，再通过一致性维度将来自不同数据源的事实数据合并到一起，供分析使用。通常来说，生成一致性维度有如下三个步骤：

1.标准化（Standardizing）

标准化的目的是使不同数据源的数据编码方式，数据格式等相同，为下一步数据匹配打下基础。

2.匹配（Matching and Deduplication）

数据匹配的工作有两种，一种是将不同数据源的标识同一事物的不同属性匹配到一起，是数据更完善；另一种是将不同数据源的相同数据标识成重复，为下一步的筛选打下基础。

3.筛选（Surviving）

数据筛选的主要目的是选定一致性维度作为主数据（Master Data），也就是最终交付的一致性维度数据。

18．简述三种基本事实表，并说明ETL的过程中如何处理它们。

19．简述桥接表是如何将维度表和事实表进行关联的？

20．迟到的数据对事实表和维度表有什么影响？怎样来处理这个问题？

元数据

21．举例说明各种ETL过程中的元数据。

答：元数据是 ETL 项目组面对的一个非常重要的主题，对于整个数据仓库项目也 是非常重要的一部分。对于元数据的分类和使用没有很确定的定义。 通常来说，我们可以把元数据分为三类，分别为业务元数据（Business Metada ta），技术元数据（Technical Metadata）和过程处理元数据（Process Execu tion Metadata）。 业务元数据，是从业务的角度对数据的描述。通常是用来给报表工具和前端用户 对数据进行分析和使用提供帮助。 技术元数据，是从技术的角度对数据的描述。通常包括数据的一些属性，如数据 类型、长度、或者数据概况分析后一些结果。 过程处理元数据，是 ETL 处理过程中的一些统计数据，通常包括有多少条记录被 加载，多少条记录被拒绝接受等数据

22．简述获取操作型元数据的方法。

23．简述共享业务元数据和技术元数据的方法。

优化/操作

24．简述数据仓库中的表的基本类型，以及为了保证引用完整性该以什么样的顺序对它们进行加载。

25．简述ETL技术支持工作的四个级别的特点。

26．如果ETL进程运行较慢，需要分哪几步去找到ETL系统的瓶颈问题。

27．简述如何评估大型ETL数据加载时间。

实时ETL

28．简述在架构实时ETL时的可以选择的架构部件。

1．微批处理（microbatch ETL，MB-ETL）微批处理的方式和我们通常的ETL处理方式很相似，但是处理的时间间隔要短，例如间隔一个小时处理一次。2．企业应用集成（Enterprise Application Integration，EAI）EAI也称为功能整合，通常由中间件来完成数据的交互。而通常的ETL称为数据整合。对实时性要求非常高的系统，可以考虑使用EAI作为ETL的一个工具，可以提供快捷的数据交互。不过在数据量大时采用EAI工具效率比较差，而且实现起来相对复杂3．CTF（Capture, Transform and Flow）CTF是一类比较新的数据整合工具。它采用的是直接的数据库对数据库的连接方式，可以提供秒级的数据。CTF的缺点是只能进行轻量级的数据整合。通常的处理方式是建立数据准备区，采用CTF工具在源数据库和数据准备区的数据库之间相连接。数据进入数据准备区后再经过其他处理后迁移入数据仓库。4．EII（Enterprise Information Integration）EII是另一类比较新的数据整合软件，可以给企业提供实时报表。EII的处理方式和CTF很相似，但是它不将数据迁移入数据准备区或者数据仓库，而是在抽取转换后直接加载到报表中。

29．简述几种不同的实时ETL实现方法以及它们的适用范围。

1触发器在源系统建立触发器 建议源系统业务简单负载小

2日志部分源系统数据库可以基于日志实时抽取建议源系统的数据库支持日志变化捕捉

3.ESB企业信息总线直接通过应用层取数据建议数据量小而且源系统应用层提供接口

4.通过ODBC或JDBC直接取源系统数据建议使用源系统的数据安全性低业务非核心 基于日志应当是比较好的一种方式。

30．简述实时ETL的一些难点及其实现方法。

1 实时抽取会造成对源系统的影响如何使得影响最小 2 源系统实时更新变化数据仓库要求相对稳定如何不被影响或影响最小 实现方式 关键哪些需求需要实时把这部分需求剥离出来但是想办法实现我碰到过一些几乎都能迂回满足需求。