2-2-cn
1
大家好，我是北京理工大学计算机学院数据科学与知识工程研究所的车海莺，本节我们讨论组织内部数据获取
2
最常用的内部数据采集工具是 ETL（Extract、Transform、load）。
可支持数据提取、数据转换与处理、数据加载；其目的是整合所有相关的企业数据以进行分析和处理以获取完整的全貌，
而不是片面的局部。左边 CRM、ERP，借助 ETL 工具，网站流量这些结构化数据可以被提取、转换并加载到数据仓库中，
而社交媒体数据、传感器日志，这些非结构化数据可以卸载到大数据集群 HDFS、Hadoop 文件系统，使用 Flume、pig 等工具，
然后使用 apache Sqoop，可以将批量数据从 Apache Hadoop 集成到结构化数据存储，例如关系数据库或数据仓库。（
Apache Flume 是一种分布式、可靠、可用的服务，用于高效收集、聚合和移动大量日志数据。
它具有基于流式数据流的简单灵活的架构。Apache Pig 是一个用于分析大型数据集的平台，
它由用于表达数据分析程序的高级语言以及用于评估这些程序的基础设施组成。
Apache Sqoop(TM) 是一种工具，旨在在 Apache Hadoop 非结构化数据存储和结构化数据存储（例如关系数据库）之间高效传输批量数据。）
在右侧，将相关的结构化和非结构化数据整合到数据仓库后，分析可以进行 OLAP、在线分析处理、数据挖掘、报表等商业智能分析。
而数据科学家也可以直接使用非结构化数据进行一些分析。
3
ETL 工具的首要任务是数据抽取，数据抽取不能只做一次，需要周期性的重复执行，因为新的数据是随着业务的发生而不断产生的。
提取数据时，可以每次都提取所有数据，也可以每次只提取新更新的数据部分，即全量提取和增量提取。
完全提取是一种简单直观的方式，如左图所示，每次都提取整个源数据存储中的所有数据，然后将其转换为 ETL 工具可以识别的格式并存储在目标数据存储中。
相比之下，增量提取仅提取自上次提取以来在源数据库中新添加或修改的数据； 同时不应对现有业务系统造成太大压力；
例如，它不应该减慢正在运行的业务 DBMS。 让我们详细了解每个。

4
所以大部分数据抽取都是增量数据抽取。
增量数据抽取 抽取自上次抽取后数据库中新增或修改的数据，同时一般不会对正在运行的业务系统造成太大影响。
但问题是如何准确识别变化的数据部分。
捕获增量数据的方法包括日志对比、时间戳、触发器和全表比较。
让我们一一学习这四种方法。
5
第一种数据增量提取方法——日志比较。
此方法通过数据库自己的日志评估更改的数据。
Oracle DBMS 有一个特殊的组件更改数据捕获 CDC，
Oracle 组件 CDC 可以识别自上次提取以来发生变化的数据；
使用 CDC 在插入、更新或删除源表的同时提取数据，改变的数据单独存储在 DB 的 change table 中。
6

第二种数据增量提取方法是时间戳。
在原表中添加时间戳字段，每次更新表数据的同时修改时间戳字段的值。
提取数据时，通过比较上次提取系统时间的值和每条记录的时间戳字段的值来确定要提取哪些数据。
对于支持时间戳自动更新的数据库，当数据库表中的其他字段发生变化时，系统会自动更新时间戳的值。
如果不支持自动时间戳的更新，需要手动更新时间戳。
7
第三种数据增量提取方法是使用触发器。
使用这种方法，您可以在数据表上创建一个触发器（例如，您可以创建插入、修改和删除三个触发器）。
每当源表数据发生变化时，都会通过相应的触发器将变化的数据写入临时表。
提取线程从临时表中提取数据，而不是从源表中提取数据。
提取的数据被标记或删除。
优势是更好的数据提取性能
缺点是为业务建立触发器，对业务系统产生影响。
8
第四种数据增量提取方法是全表比较。全表比较法使用 MD5 校验码：
ETL 工具为要提前提取的表创建一个结构相似的 MD5 临时表。临时表记录了源表的主键和根据所有字段的数据计算的 MD5 校验码。
每次提取数据时，源表和 MD5 临时表都用 MD5 校验码进行比较，以确定是否添加，更新或删除了原始表中的数据，并且同时更新了 MD5 校验码
如何？使用所有字段数据计算 MD5 校验码，如果 MD5 码与临时表中的 MD5 码相同，则说明原表中的数据自上次提取后没有变化，
如果 MD5 码不同，表示数据已添加、更新或删除，再次更新 MD5 校验码。
这种方法的缺点是被动比较整个表的数据；
限制是当表中没有主键或唯一列并且包含重复记录时，MD5 方法的准确性较差。
9
其他数据源 数据提取除了关系型数据库，ETL 提取的数据源还可以是文件，
如 TXT、EXCEL 文件、XML 文件等。
文件的提取一般为全量提取，每次提取前保存文件的时间戳或计算文件的 MD5 校验码，下次提取时进行比较。 如果相同，则忽略此提取
10
从数据源提取的数据可能不完全满足目标数据库的要求，如数据格式不一致、数据输入错误、数据不完整等。
例如姓名字段中姓氏和名字的顺序不同，这将导致统计不准确，因为汤姆杰克逊和杰克逊汤姆将被视为不同的人。
或者不同的货币计量单位，全球公司以各个国家的货币，美元，英镑，人民币等来汇总销售额。这些不同的货币不能直接加在一起，必须转换为统一的货币单位。
ETL 引擎通常使用不同的组件实现数据转换。如图所示，ETL 引擎中的一些组件包括：字段映射、数据过滤、数据清洗、数据替换、数据计算、数据校验、数据加解密、数据合并、数据拆分等。
从数据源中抽取的数据不一定会完全满足目的数据库的要求，例如数据格式不一致，数据输入错误，数据不完整等。
ETL 引擎中一般以组件化的范式实现数据转换
比如
1 姓名字段中的姓和名的顺序不同导致统计数据不准确
2 计量单位，全球化公司对各国货币表示的销售额进行汇总

数据转换后，需要加载到目标数据存储，如数据仓库。
常见的数据加载方法有 2 种
1 直接使用 SQL 语句在目标数据存储中插入、更新和删除数据。
2 使用批量加载方法，例如 BCP（批量复制程序）、关系数据库特定的批量重印工具或 API 来插入、更新和删除目标数据存储中的数据。
如果要选择 ETL 工具，Kettle 是最著名的开源 ETL 工具

12
让我们总结一下内部数据采集方法。
最常用的内部数据采集工具是 ETL、Extract、Transform、load。
在提取中，有两种方法，全量提取和增量提取：
有 4 种增量提取方法，它们是日志对比，时间戳，触发器，全表比较（使用 MD5 校验码）。
变换包括映射、过滤、说明、替换、计算、验证、加解密、合并、拆分等，通常使用组件程序来实现转换任务。
加载：可以通过 SQL 语句加载完成；
批量加载工具； 和 API

13
在我们学习了大数据资源、内部和外部数据。
基于多维度的数据，我们可以全面了解组织或业务。
感谢您的关注，如果您有任何问题，请随时与我联系。