**BD 2-1**

**1**

**Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering**

**School of Computer Science, in Beijing Institute of Technology, in this session, we will discuss Internal data acquisition.**

**2**

**The most often used internal data acquisition tool is ETL (Extract, Transform, load).**

**Whose Purpose is to Integrate all the relevant enterprise data for analysis.**

**With the help of ETL tools, structured data can be extracted from CRM, ERP, website traffic, and transformed, loaded into the data warehouse,**

**And the unstructured data like social media data, sensor logs, these can be offload to the big data cluster HDFS, Hadoop file system, or pig, and then using apache Sqoop, the bulk data can be integrated to data Warehouse for OLAP, online analysis processing, data mining, reporting and other Business intelligence analysis to get the business insight.**

**And data scientist can also directly use the unstructured data to conduct some analysis.**

**3 The first task of ETL tool is data extraction, because the new data is generated constantly as the business happens, Data extraction cannot be done just once, it needs to be executed periodically,**

**you can extract all the data every time, called full extraction, or you just extract the newly updated data part every time, incremental extraction**

**like showed in the diagram, in the 1st day, DB has 3 records 123, 2nd day, one record added, number 4, and 3rd day another record added, number 5,**

**Full extraction extracts all data in the entire source data storage every time, then Convert it into a format that can be recognized by the ETL tool and stored in the target data storage. Full extraction is simple and intuitive way, but rarely used, because full extraction produces a large amount of redundant data like the yellow part in the diagram, records number 123, and this reduces extraction efficiency;**

**in comparison, the incremental extraction** **Only extract data newly added or modified in source DB since the last extraction; like the right diagram first extraction extract records 123, second extraction just extract number 4 record, and the 3rd extraction just extract record number 5.**

**4**

**So most of data extraction is incremental data extraction.**

**incremental data extraction Extract new or modified data in the database since the last extraction, it normally would not have a big impact on the running business system.**

**“But the question” is how to accurately recognize the changing data part.**

**The Methods of capturing incremental data include Log comparison, Timestamp, Trigger and Full table comparison.**

**let’s learn these four methods one by one.**

**5**

**The first Data incremental extraction method- log comparison.**

**This method Evaluate the changed data through the Database's own log.**

**Oracle DBMS has a special component called changed data capture CDC,**

**Oracle CDC can identify the changed data since last extraction;**

**Use CDC to extract data while inserting, updating or deleting the source table, and the changed data is stored in the change table of the DB separately.**

**6**

**The second Data incremental extraction methods is Timestamp.**

**Add a timestamp field to the original table, and update the value of the timestamp field every time while modify the table data.**

**When extracting data, determine which data to extract by comparing the value of the last extraction system time and the timestamp field of each record.**

**7**

**The 3rd Data incremental extraction method is using Triggers.**

**Using this method, you can Create a trigger on the data table (for example, you can create three triggers for insert, modify and delete).**

**Whenever the source table data changes, the changed data is written to the temporary table through the corresponding trigger.**

**The extraction thread extracts data from the temporary table, Not the source table. The extracted data should be marked or deleted.**

**8**

**The 4th Data incremental extraction methods is full table comparison.**

**Full table comparison method Use MD5 check code:**

**The ETL tool creates an MD5 temporary table with a similar structure for the table to be extracted in advance.**

**The temporary table records the primary key of the source table and the MD5 check code calculated based on the data of all fields.**

**Each time of extracting data, the source table and the MD5 temporary table are compared with the MD5 check code, How? Using all the fields data to calculate the MD5 check code, if the MD5 code is as same as the MD5 code in the temporary table, which means the data in the original table has not been changes since last extraction, if the MD5 code is different, which means the data has been added, updated or deleted, and the MD5 check code is updated again.**

**9**

**In addition to relational databases, the data sources extracted by ETL could be files, such as TXT, EXCEL files, XML files, etc.**

**The extraction method of the file is generally full extraction, save the time stamp of the file or calculate the MD5 check code of the file before each extraction, and compare it during the next extraction. If it is the same, ignore this extraction**

**10**

**The data extracted from the data source may not fully meet the requirements of the target database, such as inconsistent data formats, data input errors, and incomplete data.**

**for example**

**The order of the last name and first name in the name field is different, which would result in inaccurate statistics, because Tom Jackson and Jackson Tom would be regarded as different person.**

**Or The different currency measurement unit, the global company summarizes the sales in currencies of various countries, Dollars, Pounds, RMB etc. These different currencies cannot be directly added up together, it must be transformed to unified currency unit.**

**ETL engines generally implement data transforming using different components.**

**As showed in diagram, some components in ETL engine include, field mapping, data filtering, data cleaning, data replacement, data calculation, data verification, Data encryption and decryption, Data consolidation, Data split and so on.**

**13**

**After data being transformed, they need to be loaded to the target data store , like data warehouse.**

**There are 2 common methods of data loading**

**1 Directly using SQL statements to insert, update and delete data in the target data store.**

**2** **Using bulk loading methods, such as BCP (Bulk Copy Program), relational DB-specific bulk reprint tools or API to insert, update and delete data in the target data store.**

**If you want to choose the ETL tool, Kettle is the most famous open source ETL tool.**

**14**

**Let’s make a Summary of internal data acquisition methods.**

**The mostly used internal Data acquisition tool is ETL, which means Extract, Transform, load.**

**In Extraction, there are 2 ways, Full extraction and Incremental extraction:**

**There are 4 Incremental extraction methods, which are Log comparison, time stamp, Trigger and full table comparison（using MD5 check code）.**

**The transform components include mapping, filtering, clarification, replacement, calculation, verification, encryption and decryption, merging, splitting and so on.**

**loading can be done by SQL statement loading; batch loading tool; or using API.**

**15**

**That is all for internal data acquisition, thank you and if you have any question, please feel free to contact me.**