awsmarketplace

Data Integration: **Data Transformation for Cloud Data Warehouses**

Challenge

How can data transformation be done at the scale of Cloud Data Warehouses?

Solution

Run ETL/ELT directly on your Cloud Data Warehouse with native cloud solutions.

The Importance of Accurate and Efficient Integration

Data integration is important when merging data sources and systems of two enterprises or consolidating applications within one company to provide an all-up view of the company's data. All this information is put into a data warehouse.

Data integration solutions allow businesses to create powerful data transformation jobs and orchestrate these processes in graphical tool. They include all the features businesses need to build and maintain enterprise data warehouses and analytical databases.

Successful integration is dependent on solutions that support the agile facilitation of the following phases and tasks:

Design

- Assess the requirements: why is the data transformation and integration is being done, and what are the objectives and deliverables are? From what systems will the data be sourced? Is all the necessary data available to fulfill the requirements?
- Analyze the source systems, that is, what are the options for extracting the data from the systems (update notification, incremental extracts, full extracts)? What is the required/available frequency of the extracts? In other words, what is the quality of the data?

Implementation

- Assess and determine the best tools to implement your data transformation system. Small companies and enterprises just starting with data warehousing should decide about the set of ETL solutions to consider for solution implementation. Using a new, better-suited platform or technology makes a system more effective, compared to staying with existing company practices. For example, this might include finding a tool that provides better scaling for future growth, a solution that lowers the implementation cost, and the decision to migrate the system to a new platform.

Testing

- Ensure that the unified data is correct and up-to-date with proper testing. Both IT and business leaders should be a part of the testing to help make sure that the results are as expected.
- The testing should include at least a Performance Stress test (PST), Technical Acceptance Testing (TAT) and User Acceptance Testing (UAT). Consider doing a bit of data integration, getting results (or failing) fast, then iterating.

ELT VS. ETL

ETL

IT pros are familiar with the acronym “ETL”—Extract, Transform, and Load. ETL is about taking data from a data source, applying any transformations that may be required, and then loading it into a data warehouse ready for running reports and accepting queries.

Another option is “ELT”—to perform the extraction, transformation, and loading of data in a different order. ELT can utilize the power of Amazon Redshift an example of columnar data storage technology, and the massively parallel processing capability that accompanies it. So with Amazon Redshift, using ELT, rather than ETL, is a more logical approach.

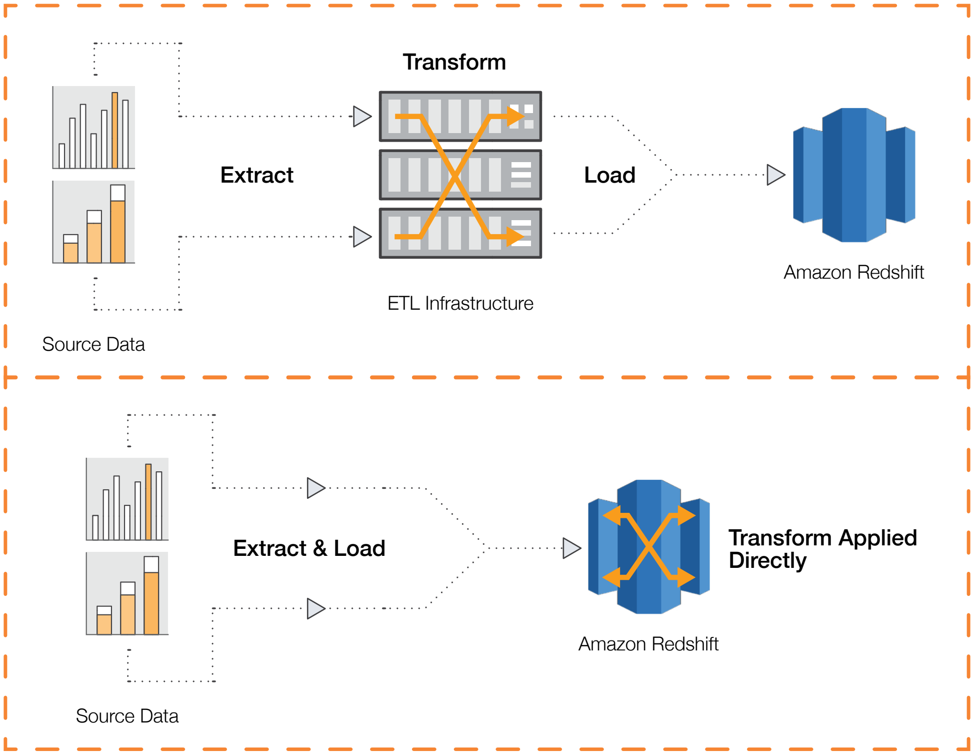
The “extract” part can be simple, if it’s a single data source like an ERP database. It can be a bit more complicated if the task is to extract data from an ERP database along with several line-of-business systems or third-party data sources. Either way, businesses can use the correct connectors and start the extract.

Likewise, with the “transform” activity, this can be a straightforward normalization of data contained in an ERP or line of business database, or something more involved like converting units of measure to a common basis.

Lastly, the “load” task sends the data into the data warehouse. Either way, ETL is characterized by a lot of string processing and variable transformation, and a lot of data parsing.

ELT

The ELT approach takes a compute-intensive activity and performs it where it makes most sense – in a powerful, cloud-based data store—rather than in an on-premises server that is perhaps already under pressure with its regular transaction-handling role. See graphic for a pictorial view:



The ETL vs. ELT approach explained

The "extract" activity is the same with ELT or ETL. The "load" activity is also the same, apart from the fact that what is being loaded is the un-transformed data. With ELT, the "transform" activity is different because it's taking place in the Cloud, inside Amazon Redshift for example.

Amazon Redshift, as mentioned previously is a columnar database, so index and record location operations are vastly quicker. It's also a massively parallel database, so the required transformations are carried out in parallel, not sequentially, with multiple nodes handling multiple transformations at the same time.

Managing and Selecting Right ETL Solution

The ETL process seems quite straightforward. However, like every application, there is the possibility that the process can fail. This can be caused by many situations such as missing extracts from one of the systems, missing values in one of the reference tables, or even a connection or power outage.

There are many ready-to-use ETL solutions available. The main benefit of using off-the-shelf ETL tools is the fact that they are optimized for the ETL process by providing connectors to common data sources like databases, flat files, mainframe systems, XML, and so on. They provide a means to implement data transformations easily and consistently across various data sources. This includes filtering, reformatting, sorting, joining, merging, aggregation and other operations ready to use. The tools also support transformation scheduling, version control, monitoring and unified metadata management. Some of the ETL tools are even integrated with business intelligence (BI) tools.

Conclusion

Many businesses are realizing that traditional on-premises data warehousing practices are characteristically slow due to the amount of time required in setting up these databases. Thus, more companies are moving large sets of information from on-premises databases to cloud-based data warehousing services. When moving to a cloud data warehousing solution, companies also want their integration solution to be agile, and one way this can be accomplished is with Amazon Redshift.

Visit aws.amazon.com/mp/etl to learn more about Data Integration Tools on AWS.