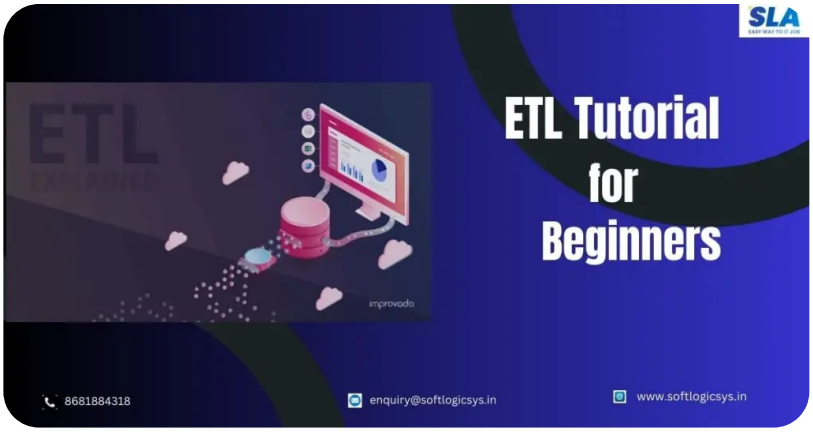

Share on your Social Media

ETL Tutorial for Beginners

Published On: September 14, 2024

ETL Tutorial for Beginners

Without extensive coding knowledge, you as an IT aspirant can shine in your data scientist career with ETL skills. Here is the comprehensive ETL tutorial that helps beginners understand the fundamental ETL concepts to kickstart their career in data warehousing jobs. Let's explore.

Download ETL Tutorial PDF

Introduction to ETL

The process of merging data from several sources into a sizable, central repository known as a **data warehouse** is called extract, transform, and load, or ETL. This ETL tutorial covers the following:

Featured Articles

Related Courses at SLA

- ETL Online Training
- ETL Training in OMR
- ETL Training in Chennai

Related Posts

- Overview of ETL
- How Does ETL Work?
- Fundamental ETL Components
- ETL Lookup Stage
- ETL Processes

Overview of ETL

Integrating data from several sources into a single data repository or warehouse is called extract, transform, and load, or ETL. There are three steps in the process:

- **Extract:** A source system's data is used.
- **Transform:** An analysis-ready format is created from the data. Filtering, sorting, combining, cleaning, deduplicating, and validating the data are some examples of what this can include.
- **Load:** The information is kept in a data warehouse or another type of system.

ETL Interview Questions and Answers

Applications of ETL

ETL can be applied in the following areas:

- **Develop data warehouses:** ETL is frequently used to provide a central repository for machine learning and data analytics.
- **Boost business intelligence:** ETL enables companies to present a unified picture of their data.
- **Enable data democracy and governance:** ETL can assist in making sure that data is accessible, safe, and useful while also granting everyone in the organization access to the information they require.

How Does ETL Work?

The ETL process includes three steps: extract, transform, and load.

Tableau Developer Salary in Chennai

Published On: October 12, 2024

Introduction A Tableau Developer designs, develops, and maintains dashboards and visualizations using Tableau software. Key...

ETL Project Ideas

Published On: October 12, 2024

Introduction An ETL Professional focuses on data integration by extracting data from various sources, transforming...

VMware Tutorial for Cloud Computing Aspirants

Published On: October 12, 2024

VMware Tutorial for Cloud Computing Aspirants VMware software allows you to run a virtual machine...

VBA Macros Tutorial for Beginners

Published On: October 10, 2024

VBA Macros Tutorial for Beginners VBA macros are programs that automate

repetitive operations in Microsoft...

Step 1: Extract

To extract data, it is necessary to copy or export it from several locations known as source sites. The raw data is then stored at a staging location until it is processed further.

Any kind of data can be found in source locations, such as flat files, emails, logs, web pages, CRM, ERP systems, spreadsheets, logs, and SQL or NSQL servers.

Typical techniques for extracting data are:

- Partial extraction with notifications for updates
- Partial extraction with no notice of updates
- Full extraction

Step 2: Transform

The data processing step in the ETL process transforms the data in the staging area so that it can be used for analytics during the transformation stage. An integrated, useful data set is created from raw data.

The data is used for many tasks, including:

- Cleaning and Standardization
- Verification and Validation
- Filtering and Sorting
- De-duplication and data audits
- Calculations, Translations, and Formatting
- Data encryption and protection

Step 3: Load

The transformed data is put into its intended destination, which could be a straightforward database or perhaps a data warehouse, in this last stage of the ETL process.

The kind of destination is determined by the quantity and complexity of data as well as the unique requirements of the organization.

The procedure for loading can be:

- **Full Loading:** It only happens during initial data loading or disaster recovery.
- **Incremental Loading:** The process of importing updated data incrementally

Advantages of ETL

Since data is cleaned up throughout the ETL process before being uploaded to the final repository for additional analysis, data quality is improved.

To collect and format data, an automated data processing pipeline is offered, eliminating the need to delegate data transformation chores to other tools.

Limitations of ETL

Since ETL is a laborious batch process, it is best suited for creating smaller data warehouses that don't require frequent updates.

For integrating greater volumes of data that need to be updated in real time, further data integration methods like ELT, CDC, and data virtualization can be employed suitably.

Download ETL Syllabus PDF

Popular ETL Tools

Several well-known ETL software tools are as follows:

- Talend
- Azure Data Factory
- Oracle Data Integrator
- Amazon RedShift
- Integrate.io
- AWS Glue
- MarkLogic
- Matillion
- FlyData

Fundamental ETL Components

Some important ETL elements to think about are:

- **Managing Multiple Source Formats:** This allows different data formats to be handled.
- **Support for Change Data Capture (CDC):** This enables loading incrementally.
- **Auditing and Logging:** To ensure that data can be audited after loading and problems can be debugged, logging and auditing are necessary.
- **Fault Recovery:** The ability to smoothly resume operations if a data transfer issue arises
- **Notification Support:** Integrated alerts that notify the user when data is inaccurate
- **Scalability:** The capacity to grow to accommodate increasing volumes of data
- **Accuracy:** Every data point must be able to be verified at any time.

Importance of ETL

ETL contributes to better **data analytics** and data cleanliness. Additionally, ETL tools carry out other crucial business tasks, such as:

- Integrating disparate data formats to transfer information from antiquated systems to contemporary platforms.
- Synchronizing external data from suppliers, buyers, and vendors.
- Combining information from several systems that overlap.
- Combining transactional data in a way that makes sense for users.

ETL Lookup Stage

Since data can only be studied while it is in memory, the ETL search step cannot be utilized when dealing with large databases.

However, it does allow us to evaluate data with several options. Nonetheless, when compared to

the join and merge stages, it is the better option. Furthermore, condition-based data analysis is enabled by the lookup stage.

Features of the Lookup Stage

- It is a phase of processing.
- It uses a dataset to read data to operate on memory.
- Additionally, direct lookups on Oracle and DB2 are possible.
- Row validation is another lookup application. If a row doesn't have a matching entry, it is rejected.
- There can be one or more reference links in the lookup step, but only one input and one output link.

Three approaches can be used to work this stage:

- **Equality Match:** The standard appearance is another name for this. Here, the precise case-sensitive match in the data is examined.
- **Casesless Match:** This function searches for values that don't depend on the case.
- **Range Match:** A lookup stage can be set up to search for a range of values between two lookup columns with the use of the range function.

Compared to the join and merge stages, the lookup stage is better for handling smaller amounts of data because it processes data in memory.

The lookup step, however, cannot be used for very large amounts of data. A join stage or a merge stage is utilized for databases or datasets that include enormous quantities.

One of the join stage's shortcomings is that it cannot reject a row for which there is no corresponding entry; in contrast, the merge stage can assist us in rejecting such values.

Lookup Toolbar

These buttons are part of the ETL Lookup toolbar:

- **Stage properties:** This option aids in defining the stage name, link properties, and other attributes.
- **Conditions:** This button allows you to define all conditions.
- **Show all selected relations:** We can view all selected relations by using this button.
- **Cut, Copy, and Paste:** These include general functions like Cut, Copy, and Paste.
- **Load/Save column definitions:** This particular option allows you to load and store the data for the columns.
- **Find/Replace:** They do the standard Find/Replace tasks.
- **Column automatch:** Based on mapping, it assists in automatically matching columns.
- **Order of input/output link execution:** This one aids in rearranging by displaying the links that are used.

ETL Process

Batch processing is the foundation of the conventional ETL method. ETL processes handle massive amounts of data from source systems on a daily, weekly, or monthly basis. The ETL process's fundamentals are as follows:

- **Reference data:** Establish a set of guidelines outlining acceptable values.
- **Data extraction:** It is the process of taking data from sources and transforming it into a single, standardized format from a variety of forms, such as RDBMS, XML, JSON, and CSV.
- **Data validation:** It is necessary to make sure the data has the desired information in the right format; otherwise, the ETL processes that follow won't function.
- **Data transformation:** It is the process of cleansing, confirming the accuracy of the data, and organizing or grouping it to make analysis simpler.

- **Staging:** Data should be loaded into a staging environment for ease of rollback if something goes wrong.
- **Load to data warehouse:** If all checks out, push the data to the production data warehouse, where it will either replace previous data or be kept and its historical versions will be managed using timestamps.

Examples of ETL Process:

Here are some general examples of important ETL tasks.

Data Extraction

The initial stage of the ETL process is data extraction. There are numerous ways to extract data: it can be imported using APIs, streamed using tools like Kafka, or copied straight from storage devices.

Let's take a basic scenario where we have data files that need to be loaded into a target table in a data warehouse and uploaded to an FTP server. The following task employs a Type 4 Slowly Changing Dimension, where updated data is kept in a distinct history table with a timestamp for every previous iteration.

To manage earlier iterations of the data when extracting source data that has been transferred via FTP:

- **Manage two tables:** the history table, which contains older, timestamped versions of every data field, and the target data table.
- Make an automatic trigger that pulls files to the ETL machine when it finds a new file in a specified folder on the FTP server.
- Fill a temporary table with data loaded from the source file.
- The target table should already be loaded into a temporary lookup file.
- Perform the following actions for every record

in the source record:

- Check if the source data record passes validation, and then store it in a reject table.
- Compare the entry with the lookup table. Load a new record into the target table if it doesn't already exist.
- Save the updated value to the history table, load it into the target table, and overwrite the old value if the record does indeed exist in the lookup table and its value has changed. Do nothing if the value hasn't changed.

Surrogate Key Generation

To manage data coming in from many sources, ETL engineers construct a data field called a surrogate key. A distinct, numeric record identification known as the surrogate key is mapped to the original "natural keys" in the source data, such as transaction or customer IDs.

To create a surrogate key while loading data and overwriting existing data:

- From the source data, choose natural keys.
- Make a mapping table that transfers every value from the natural keys to the new surrogate key, which is a number. A value indicating the maximum key number of the most recent data field loaded ought to be present in the table.
- Carry out a loading procedure for every source file:
  - Verify that every value in the source data is accurately mapped by the mapping table.
  - Verify whether the surrogate key is already present in the target table for each data record. In that case, replace the current record.
  - Add a new record to the target table, add a new entry to the mapping table, and

increase the maximum key by one if the surrogate key is absent.

It is possible to create a somewhat more involved procedure that is comparable to loading data while preserving its historical version.

Etl Developer Salary

Header and Trailer Processing

A trailer is appended to each record, and a header with standard descriptive data is present in many data sources, such as network traffic data and legacy sources. The records are organized in blocks.

Processing the payload and header:

- An ETL processing tool is the most convenient way to carry out this kind of processing. Create a separator inside the file to indicate the header, body, and trailer sections of the data after extracting it using the program of your choosing.
- Divide the data into the following three tables: headers, body, and trailers using the separators.
- See the header and trailer format documentation for guidance on converting into a usable format for the header and trailer tables.
- Ensure that you save the record ID that associates body data with headers and trailers.

Data Masking

Data masking, scrubbing, or anonymization is a standard need in data initiatives. This may be required:

- To avoid storing sensitive client data on non-production servers, during data testing or staging.

- It may be necessary to anonymize OLTP data while keeping all business-relevant information in each data record to avoid privacy and security concerns while transferring the data to a data warehouse.

Approaches for masking or concealing data:

- **Substitution:** It is the process of using fictitious data from a dictionary table to fill in each value in a sensitive data field.
- **Masking:** It is the process of replacing sensitive data with characters like *. For instance, displaying a 16-digit credit number with just the final 4 digits and 12 asterisks.
- **Hashing:** It is the process of changing a sensitive data value into a completely distinct value while maintaining the original data format and size via a one-way function.
- **Shuffling:** Data is shuffled or shifted at random between data records.
- **Randomization:** To replace the original sensitive data, randomization involves producing new data at random.

Data Quality and Data Cleansing

Validating the consistency and integrity of the data is essential for any ETL process, as is cleaning up inaccurate or non-standard data records. If done incorrectly, this important step can jeopardize all other processing processes.

Imagine a straightforward data quality procedure with two tests:

Syntax test: finds records with invalid characters, improper data types, incorrect data patterns, etc.

Reference test: identifies records with a correct data pattern that do not correspond with reference data that is known to exist. An example of this would be an order that contains a product that is not included in the products database.

To develop an automated data cleansing procedure, do the following things:

- Verify dates to make sure they follow business regulations and are formatted correctly.
- Verify that the IDs are within the permitted range of numbers or characters and contain the correct characters.
- Check the address's syntax and its constituent parts against a dictionary table, including the names of the nation, the city, and the street.
- Check phone numbers for format, including allowances for international number formats, then cross-reference country codes with a dictionary listing of recognized nations.
- For every other data field, do similar tests.
- To facilitate manual data correction and troubleshooting, save all data fields containing errors to a rejected file.
- If a data problem is discovered, report a warning if it is not a critical issue and save the data to the target table. If the problem is critical, report an error and refrain from saving the data to the target table.

ETL Training

Conclusion

ETL includes entering all data directly into the data warehouse and instantaneously transforming it at a later time to satisfy user requirements. We hope this basic ETL tutorial will help you gain a fundamental understanding of ETL jobs. Excel in your data warehouse career by learning the complete **ETL course in Chennai.**

Softlogic Academy

Softlogic Systems

KK Nagar [Corporate Office]

No.10, PT Rajan Salai, K.K. Nagar, Chennai – 600 078.
Landmark: Karnataka Bank Building
Phone: +91 86818 84318
Email: enquiry@softlogicsys.in
Map: Google Maps Link

OMR

No. E1-A10, RTS Food Street
92, Rajiv Gandhi Salai (OMR),
Navalur, Chennai - 600 130.
Landmark: Adj. to AGS Cinemas
Phone: +91 89256 88858
Email: info@softlogicsys.in
Map: Google Maps Link

Courses

- Python
- Software Testing
- Full Stack Developer
- Java
- Power BI
- Clinical SAS
- Data Science
- Embedded
- Cloud Computing
- Hardware and Networking
- VBA Macros

Navigation

- About Us
- Blog Posts
- Careers
- Contact
- Placement Training
- Corporate Training
- Hire With Us
- Job Seekers
- SLA's Recently Placed Students
- Reviews
- Sitemap

Important Links

- Disclaimer
- Privacy Policy
- Terms and Conditions

Social Media Links

Review Sources

- Google
- Trustpilot
- Glassdoor
- Mouthshut
- Sulekha
- Justdial
- Ambitionbox

Mobile App Development

DevOps

Indeed

Software Suggest

Sitejabber