**ETL Tools**

**ETL** is the traditional technique of extracting raw data, transforming it for the users as required and storing it in data warehouses. ELT was later developed, having ETL as its base. The three operations happening in ETL and ELT are the same except that their order of processing is slightly varied. This change in sequence was made to overcome some drawbacks.

* Extract - It is the process of extracting raw data from all available data sources such as databases, files, ERP, CRM or any other.
* Transform - The extracted data is immediately transformed as required by the user.
* Load - The transformed data is then loaded into the data warehouse from where the users can access it.

**BEST ETL TOOLS**

ETL tools can work in either cloud or on premises IT environments; they also come in either proprietary or open-source software. Here are some of the most popular ETL tools in those categories.

CLOUD ETL TOOLS

1. **AWS Glue**

AWS Glue is a nice fit for companies that use SQL databases, AWS and Amazon S3 storage services. AWS Glue enables you to clean, validate, organize and load data from disparate static or streaming data sources into a data warehouse or a data lake. It also allows you to process semi-structured data such as clickstream (e.g., website hyperlinks) and process logs. Its strength is in its ability to work with SQL, which many companies have competence in. On the programming side, AWS Glue executes jobs using either Scala or Python code.

With AWS Glue, you can schedule ETL jobs based on a schedule or event, or you can trigger jobs as soon as data becomes available. AWS Glue is an on-demand tool that automatically scales to accommodate the processing and storage resources that you need, and that gives you visibility of runtime metrics while it processes.

AWS Glue integrates well with other AWS systems and processes, so if AWS is your primary data repository and processor, AWS Glue works well. It also has APIs for third party JDBC (JAVA)-accessible databases like DB2, MySQL, Oracle, SyBase, Apache Kafka and MongoDB.

AWS offers free online courses. It also provides certification programs.

1. **Azure Data Factory**

Azure Data Factory is a pay-as-you-go cloud-based ETL tool that automatically scales processing and storage to meet your data and processing demands. Its strength is that it can be used by both IT professionals and end users. This is because the tool has both a no-code graphical user interface for end users and a code-based interface for IT. Both code and no-code interfaces feature data pulls from more than 90 connectors. Among these connectors are AWS, DB2, MongoDB, Oracle, MySQL, SQL, SyBase, Salesforce and SAP.

Azure Data factory is a nice choice for Microsoft shops, and for companies that want both their business end users and IT group to have access to ETL tools that enable them to pull data into data repositories.

Microsoft offers free online training. It also offers certifications for Azure Data Factory. Its standard technical support package provides 24×7 access to support engineers via email and phone, with a guaranteed response time that is within one hour.

1. **Google Cloud Dataflow**

Google Cloud Dataflow is part of the Google Cloud platform, and is well integrated with other Google services. Dataflow uses ApacheBeam open source technology to orchestrate the data pipelines that are used in DataFlow’s ETL operations. Google Cloud Dataflow requires IT expertise in SQL databases, and in the Java and Python programming languages.

This software can be deployed for both batch and real-time processing, and in either a scheduled or a real-time on demand mode. Because Google Cloud Dataflow is cloud-based, it can automatically scale to accommodate the processing and storage that you need for any ETL job. Google Cloud Dataflow is ideal for shops that heavily use the Google Cloud platform.

Through its Cloud Academy, Google offers a free online tutorial on Dataflow, offers hands-on training at $34/month and a Google certification program at $39/month.

ON PREMISES ETL TOOLS

1. **IBM InfoSphere DataStage**

InfoSphere DataStage is part of the IBM Information Server Platform. It uses a client/server design where jobs are created and administered via a Windows client against a central repository on a server. This server can be Intel-based, UNIX-based, LINUX-based or even an IBM mainframe. Regardless of platform, the IBM InfoSphere DataStage ETL software can integrate data on demand across multiple, high volumes of data sources and can target applications using a high performance parallel framework. InfoSphere DataStage also facilitates extended metadata management and enterprise connectivity.

InfoSphere DataStage is well suited for large enterprises that have mainframes or large servers, and high volume processing and data. These organizations tend to run on multiple clouds, and also in on premises data centers. The connecters supported by IBM InfoSphere DataStage range from AWS, Azure and Google, to SyBase, Hive, JSON, Kafka, Oracle, Salesforce, Snowflake, Teradata and others.

IBM InfoSphere DataStage is a robust ETL solution, and also a costly one. This tool is designed for IT professionals who have a sound understanding of SQL and also knowledge of the BASIC programming language, which InfoSphere DataStage uses.

IBM offers pay-for online and classroom training and certifications for DataStage. It also provides 24/7 technical support packages

1. **Oracle Data Integrator**

Oracle Data Integrator (ODI) is a strong platform for larger enterprises that run other Oracle applications such as Enterprise Resource Planning (ERP). ODI is designed to move data from point to point across an entire company’s business functions. Like ERP, it can support integrated workflows across entire organizations.

ODI can process data integration requests that range from high-volume batch loads to service-oriented architecture (SOA) data services that enable software components to be called and reused in new processes. ODI also supports parallel task execution for faster data processing and offers built-in integrations with other Oracle tools, such as Oracle GoldenGate and Oracle Warehouse Builder.

ODI ETL software supports data integration for both structured and unstructured data. It supports relational databases, and has a library of APIs for third party data and applications. On the big data side, ODI also supports Spark Streaming, Hive, Kafka, Cassandra, HBase, Sqoop and Pig. ODI is a sophisticated and proprietary tool that requires IT expertise and experience in Java programming.

On a subscription basis, Oracle offers access to online training and certifications for ODI.

Technical support is available, and will be added to licensing fees.

1. **Informatica PowerCenter Mapping Designer**

Informatica PowerCenter is an enterprise-strength ETL tool that is best utilized by large organizations with the need to move data across many different business functions. PowerCenter extracts, transforms and loads data from a variety of different structured and unstructured data sources that span internal and external (cloud-based) enterprise applications. PowerCenter has many APIs to variety of different third party applications and data.

Common data formats that PowerCenter works with include JSON, XML, PDF and Internet of Things (IoT) machine data. PowerCenter can work with many different third party databases, such as SQL and Oracle database. PowerCenter will transform data based upon the transformation rules that are defined by IT.

Informatica PowerCenter furnishes a user-friendly graphical interface that is designed for the use of business users, but the tool is best used by IT, as it is highly sophisticated. PowerCenter can automatically scale to meet processing and data needs at the same time that it works to optimize performance.

Although PowerCenter is a proprietary ETL tool, it can work in both cloud and on premises environments.

Informatica offers PowerCenter online training subscriptions and provides learning paths for developers, administrators and data integrators through its Informatica University.

It also offers technical support options that companies can subscribe to.

**ELT Tools**

**ELT** is the technique of extracting raw data from the source and storing it in data warehouse of the target server and preparing it for endstream users. ELT comprises of 3 different operations performed on the data:

* Extract - Extracting data is the technique of identifying data from one or more sources. The sources may be databases, files, ERP, CRM or any other useful source of data.
* Load - Loading is the process of storing the extracted raw data in data warehouse or data lakes.
* Transform - Data transformation is the process in which the raw data source is transformed to the target format required for analysis.

**BEST ELT TOOLS**

1. **Hevo**

Hevo allows you to replicate data in near real-time from 150+ sources to the destination of your choice including Snowflake, BigQuery, Redshift, Databricks, and Firebolt. Without writing a single line of code. Finding patterns and opportunities is easier when you don’t have to worry about maintaining the pipelines. So, with Hevo as your data pipeline platform, maintenance is one less thing to worry about.

For the rare times things do go wrong, Hevo ensures zero data loss. To find the root cause of an issue, Hevo also lets you monitor your workflow so that you can address the issue before it derails the entire workflow. Add 24\*7 customer support to the list, and you get a reliable tool that puts you at the wheel with greater visibility. Check Hevo’s in-depth documentation to learn more.

If you don’t want SaaS tools with unclear pricing that burn a hole in your pocket, opt for a tool that offers a simple, transparent pricing model. Hevo has 3 usage-based pricing plans starting with a free tier, where you can ingest upto 1 million records.

1. **Luigi**

Luigi is a Python library that provides a framework for building complex data pipelines. Luigi was built at Spotify. The purpose of Luigi is to allow you to automate and chain batch processes.

Features:

* Supports dumping data to and from databases
* Supports running machine learning algorithms.
* Luigi features reliable throughput with real-time elasticity, scalable to millions of events per month.
* Ability to build up long-running pipelines that comprise thousands of tasks.
* Support for running Python MapReduce jobs in Hadoop, Hive, as well as Pig.
* Ships with file system abstractions for HDFS and local files to ensure that the system can handle failures and that your data pipeline will not crash in a state containing partial data.
* The server comes with a Web UI for workflow management and visualization of the dependency graph of the workflow.
* It handles dependency resolution
* Command-line integration.
* Stores the state of the ELT pipeline in Elasticsearch.

Best-suited use case:

Luigi is best suited for organizations that run thousands of tasks every day, and they need to organize it in complex dependency graphs. It’s especially suited for building complex and ever-changing ELT pipelines.

Drawbacks:

* Steep learning curve. You would need you to invest heavily in engineering resources that can build and maintain this infrastructure.
* Hard to test tasks using the API.
* If the scheduler is busy or other concurrent users are using the UI, the UI suffers from disappointingly sluggish performance.

1. **Blendo**

Blendo is known as one of the best ELT tools enabling customers to centralize all of their different datasets and data sources into a central location. They are in the business of building new connectors and maintaining the ones already created.

Over the years, they’ve grown to over 40+ integrations. They provide a fast way to replicate your application, databases, events, and files into the fully managed and elastic cloud warehouses such as BigQuery, and Redshift.

Features:

* Fully managed data pipelines as a service
* Limited maintenance and configuration
* Automated schema migrations
* 40+ connectors and counting

Best-suited use case:

It is a good choice for businesses that want to move data from Facebook Ads, Google Ads, Google Analytics, Hubspot, LinkedIn Ads, Mailchimp, MySQL, Salesforce, Shopify, and Stripe to Amazon Redshift, Google BigQuery, Microsoft SQL Server, Snowflake, PostgreSQL, and Panoply.

Drawbacks:

* Blendo works as a extract and load kind of a set up. They do not provide a way to transform the data before or after loading to the warehouse. This becomes limiting when your ETL use cases start to evolve.

1. **Matillion**

Matillion is one of the best ELT Tools that is built specifically for Amazon Redshift, Azure Synapse, Google BigQuery, and Snowflake. Matillion has an ELT architecture. It sits between your raw data sources (internal, external, and third-party data) and your BI and Analytics tools.

Matillion ELT takes away the compute-intensive activity of loading data from your on-premise server that is perhaps already under pressure with its regular transaction-handling role and instead leaves this process to the data warehouses that tend to have an infinite amount of parallel processing resources.

Features:

* Pay-as-you-go model with no long term financial commitments.
* Scalable built to take advantage of the power and features of your data warehouse.
* Makes complex tasks simple with an intuitive UI and approach to data transformation.
* Automated data workflows.
* Drag-and-drop browser-based UI so you can build your ELT jobs in minutes.

Best-suited use case:

If you’re using Amazon Simple Storage Service (S3), Amazon Redshift, Azure Synapse, Google BigQuery, or Snowflake for your data Warehousing needs, then Matillion is a good choice for your use case. However, keep in mind that Matillion doesn’t support ELT load jobs to other Data Warehouses—it is designed specifically for those solutions.

Drawbacks:

* Learning curve – understanding and implementing complex features becomes challenging for new development teams.
* You can sometimes encounter validation failure in scheduled jobs for no discernable reason.
* Clustering is not supported which means large load jobs can take up a long time to process or even lead to OOM errors.
* Integration with version control systems is a complex undertaking.

1. **Talend**

The Talend cloud data integration tool is known as one of the best ELT tools. It is a modern big data and cloud integration software to connect, extract, and transform any data across the cloud and on-premises. They are enabling companies to harness the power of their enterprise information and to turn that data into insights so that they can get ahead.

Talend provides a data integration platform natively designed for the new Big Data and the cloud-centric world that empowers companies to immediately turn data into business insights.

Features:

* A subscription-based data management platform.
* Variety of connectors to various data sources.
* Management and monitoring capabilities.
* Log collection and display.
* Easily deployable in a cloud environment.
* Data can be loaded into your data lakes and warehouses without formatting which makes the ingestion speed much quicker.
* A healthy online community that can assist you with any technical support issue.
* Connectors for Snowflake, Amazon Redshift, Azure Data Lake Storage Gen2, Azure SQL Data Warehouse, Databricks Delta Lake, Google BigQuery, Oracle, Teradata, Microsoft SQL Server, SaaS, Packaged Apps, SMTP, FTP/SFTP, LDAP, and more.

Best-suited use case:

If you have your data in on-premises data warehouses, Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, SAP, Salesforce, Oracle — Talend connectors support all these use cases and more. Talend is full of features and built-in components.

Drawbacks:

* The job editor is quite heavy and it can stall during heavy tasks.

1. **Streamsets**

Streamsets is a cloud-first enterprise ELT tool for extracting data from SaaS applications and databases into data warehouses and data lakes for last-mile analysis. Enterprises use Streamsets to consolidate dozens of data sources for analysis.

Features:

* Easy self-serve model for replicating data from more than 100 applications and databases.
* Highly extensible – customers can add new data sources.
* Available in the AWS store.
* Ability to replicate, merge, as well as segment and route data.
* HIPAA, GDPR, SOC 2 compliant.

Best-suited use case:

If you have IoT edge devices, and you need a lightweight execution agent that runs pipelines on edge devices then you will find that SteamSets is an ideal ELT for this use case. StreamSets is one of the many ELT Tools that integrate seamlessly with the old Java platform.

Drawbacks:

* It is not a good fit for very low latency use cases as it is much better suited for batch data processing.
* It has fewer connectors compared to most tools in this list although the community regularly updates it with new connectors.