



عنوان مقاله: تفاوت بین ETL و ELT چیست؟

نویسنده مقاله: تیم فنی نیک آموز

تاریخ انتشار: آذر ماه ۱۴۰۰

منبع: <https://nikamooz.com/difference-between-etl-and-elt>

آنچه در ان مقاله میخوانیم:

۱- ETL و ELT چیست؟

۲- مراحل هر فرآیند

۳- تفاوت آن ها

۴- نسل بعدی این فرآیندها

مقدمه

در سامانه های مدرن امروزی جابه جایی بین منابع مختلف جز دغدغه های مهندسين داده است. خیلی اوقات ما نیاز داریم بین منابع داخلی سازمان منابع را منتقل کنیم. مثلا داده هارا از دیتابیس های عملیاتی مانند SQL به دیتابیس های تحلیلی منتقل کنیم. ممکن است ما داده ها را از شبکه های اجتماعی مانند توئیتر، تلگرام و... بخوانیم و به دیتابیس های داخلی سازمان انتقال بدهیم. در تمامی این نکاتی که گفته شد ما به جابه جایی بین منابع مختلف نیاز داریم. ETL چیست؟ ما نیاز داریم که داده ها را از منابع مختلف بخوانیم و تغییراتی را روی آن اعمال کنیم و آن ها را در جایی ذخیره کنیم. به طي شدن این فرایند ETL. گفته می شود.

شامل سه بخش میباشد که این سه بخش عبارتند از ETL

- استخراج داده ها
- تبدیل داده ها
- ذخیره سازی داده ها



مراحل ETL:

استخراج داده ها: در حین استخراج، ETL داده ها را شناسایی می کند و از منابع خود کپی می کند، بنابراین می تواند داده ها را به دیتا بیس منتقل کند. داده ها می توانند از منابع ساختاریافته و بدون ساختار، از جمله اسناد، ایمیل ها، برنامه های کاربردی تجاری، پایگاه های داده، تجهیزات، حسگرها، اشخاص ثالث و غیره به دست آیند.

تبدیل داده ها: از آنجایی که داده های استخراج شده به شکل اصلی خود خام هستند، باید نقشه برداری و تبدیل شوند تا برای ذخیره داده نهایی آماده شوند. در فرآیند تبدیل، ETL داده ها را اعتبارسنجی، احراز هویت، کپی سازی و/یا جمع آوری می کند به گونه ای که داده های حاصل را قابل استفاده می کند.

ذخیره داده ها: ETL داده های تبدیل شده را به دیتا بیس مورد نظر منتقل می کند. در این مرحله می توانید داده های اولیه را بارگیری نمایید و تغییرات تدریجی را روی آن ها اعمال کنید. و در نهایت باید هر داده را در بسته بندی های مخصوص قرار دهید تا در زمان مناسب بشود از آنها استفاده کرد.

ELT چیست؟

ELT مخفف عبارت «extract, load, and transform» است - در این فرایند اول داده ها دریافت می شوند بعد ذخیره می شوند و در نهایت آن ها را برحسب نیاز سازمان داده ها را تبدیل میکنیم.

در این فرایند بخش ذخیره میانی که در فرایند ETL وجود داشتو کار ذخیره موقت را انجام می داد حذف شده است. ELT هم شامل سه بخش می باشد که این سه بخش عبارتند از:

- دریافت
- بارگذاری
- و تبدیل

مراحل ELT

استخراج: این مرحله اول شامل کپی کردن داده ها از سیستم منبع است.

بارگذاری: در طول مرحله بارگذاری، خط لوله داده ها را از منبع به سیستم هدف، که ممکن است یک انبار داده یا دریاچه داده باشد، تکرار می کند.

تبدیل: هنگامی که داده ها در سیستم هدف قرار می گیرند، سازمان ها می توانند هر تغییری را که نیاز دارند انجام دهند. اغلب سازمان ها داده های خام را به روش های مختلف برای استفاده با ابزارها یا فرآیندهای تجاری مختلف تغییر می دهند.

تفاوت بین ETL و ELT

به طور معمولی ETL برای اشاره به هر خط لوله داده استفاده می شود که در آن داده ها از منبع استخراج و تبدیل می شوند و برای استفاده توسط کاربر نهایی در جدول نهایی بارگذاری می شوند. تبدیل می تواند در پایتون، اسپارک، اسکالا، SQL در انبار داده و غیره اتفاق بی افتد.

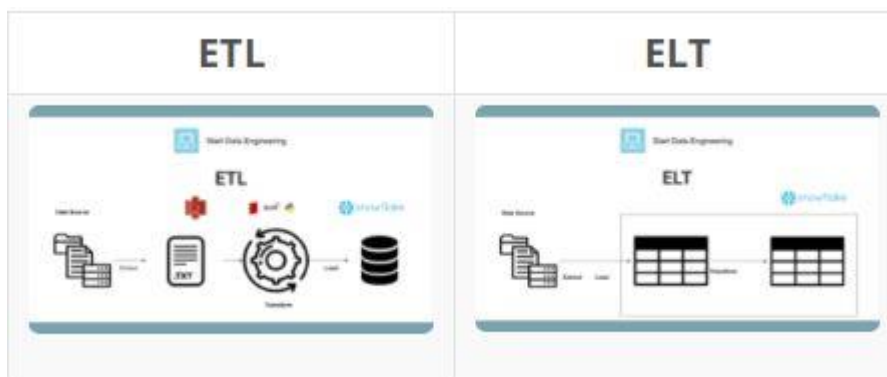
اخیراً از ELT برای اشاره به خطوط لوله داده استفاده می شود که در آن داده ها در انبار داده تبدیل می شوند. این می تواند گیج کننده باشد زیرا، طبق تعریف، این وظیفه خط لوله ETL است که داده ها را در SQL و یا انبارهای داده ذخیره کند.

منظور کاربران از ETL و ELT چیست؟

ETL: داده های خام در برخی از حافظه های ذخیره سازی فایل ذخیره می شوند، با یک پایتون/spark/scala یا سایر زبان های غیر sql تبدیل می شوند و در جداول بارگذاری می شوند تا توسط کاربر نهایی استفاده شوند.

مرحله تبدیل تا حد زیادی پیچیده ترین مرحله در فرآیند ETL است.

ELT: داده های خام در انبار داده بارگذاری می شوند و با استفاده از SQL به جدول نهایی تبدیل می شوند تا توسط کاربر نهایی استفاده شود.



در قسمت بعدی قصد داریم به طور مفصل فرایند های ETL و ELT را مقایسه کنیم.

شاخص	نکات	ELT	ETL
هزینه جاری	بسته به ابزار داده شما، انجام تبدیل توابع بدون ابزار داده میتواند درمقایسه با تبدیل در ابزار داده ها بسیار ارزان تر باشد.		✓
تخصص مهندسی	هر دو فرایند درک خوبی از سیستم های توزیع شده، کدنویسی، نگهداری، اشکال زدایی و SQL دارند.	✓	✓
زمان توسعه	بسته به ابزارهای موجود، ممکن است متفاوت باشد. اگر ابزارهایی مانند Fivetran یا dbt دارید، ELT مناسب تر است.	✓	
قابلیت های دگرگونی	SQL آنقدر قدرتمند نیست و زبان های برنامه نویسی/فریمورک هایی مانند پایتون، اسکالا و اسپارک تبدیل های پیچیده را انجام میدهند.		✓
تأخیر بین تولید داده و در دسترس بودن کاربر نهایی	به طور کلی تأخیر در ETL بیشتر است زیرا SQL سرعت کمتری نسبت به پایتون و دیگر زبان ها و فریمورک ها دارد.		✓
سرعت	ELT به دلیل در دسترس بودن ابزارهای (Airbyte، Fivetran و غیره) امکان تحویل سریع تر با ویژگی بالاتر برای داده ها را فراهم می کند.	✓	

توجه داشته باشید که مقایسه بالا برای یک خط جریان داده ساده است و بسته به خط جریان داده شما این ویژگی ها می تواند تغییر کند. اکثر شرکت ها از ترکیب رویکردهای ETL و ELT استفاده می کنند. ممکن است خطوط جریان داده ایی را پیدا کنید که در آن داده ها با استفاده از چارچوبی مانند Apache Spark تبدیل شده، و در جدول ابزار بارگذاری می شوند و با استفاده از SQL در ابزار داده شما تغییر شکل می دهند.

از دیگر تفاوت های این دو فرایند میتوان به موارد زیر اشاره کرد:

در فرآیند ELT، تبدیل داده ها بر اساس نیاز در خود سیستم هدف انجام می شود. در نتیجه، مرحله تبدیل زمان کمی می برد، اما اگر قدرت پردازش کافی وجود نداشته باشد، می تواند فرآیندهای پرس و جو و تجزیه و تحلیل را کند کند. فرآیند ETL برای مجموعه داده های کوچکی که نیاز به تبدیل های پیچیده دارند مناسب است. فرآیند ELT برای مجموعه داده های بزرگتر، ساختاریافته و بدون ساختار مناسب تر است.

در فرآیند ETL، تبدیل داده ها در یک منطقه خارج از انبار داده انجام می شود و کل داده ها باید قبل از بارگذاری تبدیل شوند. در نتیجه، تبدیل مجموعه داده های بزرگ تر می تواند زمان زیادی را در پیش بگیرد، اما تجزیه و تحلیل می تواند بلافاصله پس از تکمیل فرآیند ETL انجام شود.

ELT - نسل بعدی ETL:

ELT یک تغییر مدرن در فرآیند قدیمی تر استخراج، تبدیل و بارگذاری (ETL) است که در آن تبدیل ها قبل از بارگیری داده ها انجام می شود. اجرای تبدیل ها قبل از فاز بارگیری منجر به پیچیده شدن این فرآیند می شود. ابزارهای ETL به جریان های پردازشی برای تبدیل داده ها قبل از بارگیری داده ها در مقصد نیاز دارند. از سوی دیگر، با ETL، کسب و کارها از جریان های پردازش در مقصد استفاده می کنند تا داده ها را به طور مؤثر در خود سیستم منبع تبدیل کنند. این حذف یک مرحله میانی، فرآیند بارگذاری داده ها را ساده می کند.

از آنجایی که ETL داده ها را قبل از مرحله بارگیری تغییر می دهد، زمانی که یک مقصد به یک منبع داده خاص نیاز دارد، این فرآیند ایده آل است. این می تواند شامل مواردی باشد که در انواع داده های پشتیبانی شده بین منبع و مقصد ناهمگنی وجود دارد، توانایی محدود برای مقیاس بندی سریع پردازش در مقصد، یا محدودیت های امنیتی که ذخیره داده های خام را در مقصد غیرممکن می کند.

با این حال، وقتی مقصد یک انبار داده مانند Amazon Redshift، Google BigQuery، Snowflake و Microsoft Azure SQL Data Warehouse باشد، ELT عملکرد بهتری دارد. سازمان ها می توانند داده های خام خود را در هر زمان، و در هر شرایطی در خط لوله تغییر دهند.

نتیجه گیری:

امیدواریم این مقاله منظور کاربران را وقتی که ETL یا ELT می گویند را روشن کند. دفعه بعد که با این اصطلاح برخورد کردید، به یاد داشته باشید که ELT برای اشاره به خط لوله داده استفاده می شود که در آن داده ها با استفاده از SQL در انبار داده شما تبدیل می شوند. ETL به هر خط لوله داده ای اشاره دارد که شامل انتقال داده از یک سیستم به سیستم دیگر است. و در آخر هنگام طراحی خط لوله داده ELT، به ویژگی های زیر توجه کنید:

۱. هزینه توسعه
۲. هزینه جاری
۳. تخصص مهندسی
۴. الزامات تحول
۵. الزامات تاخیر

منبع: <https://www.startdataengineering.com/post/elt-vs-etl/>