عنوان مقاله: ETL چیست ؟

نویسنده مقاله: تیم فنی نیک‌آموز

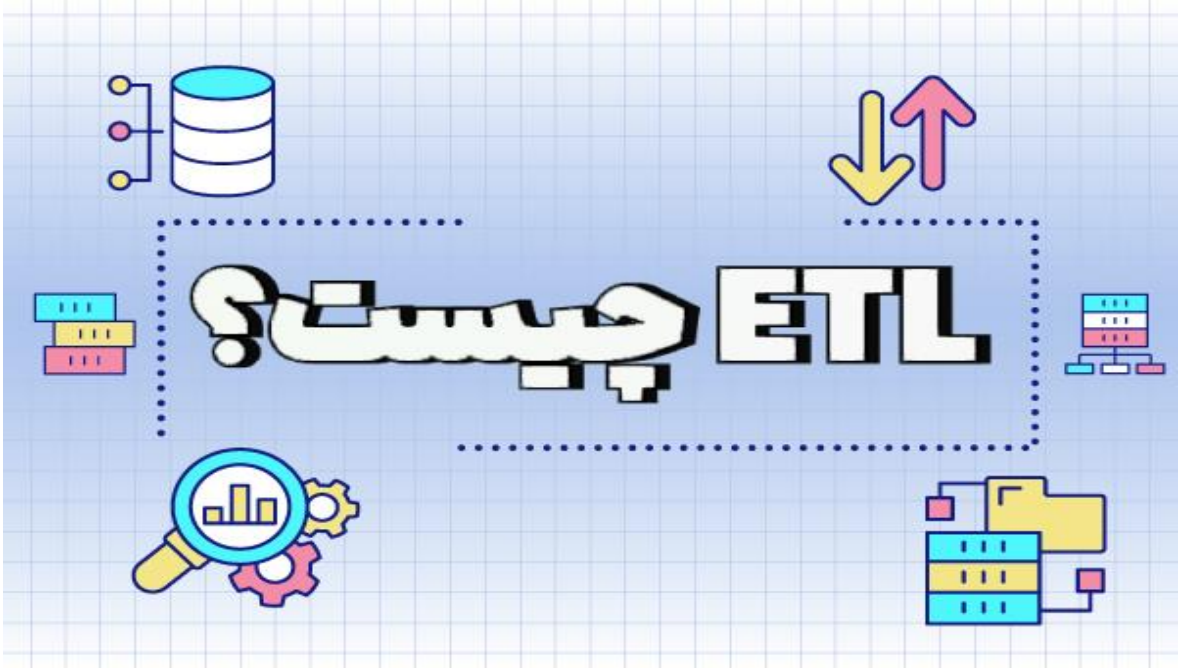
تاریخ انتشار: ۹ آبان ۱۴۰۲

منبع: https://nikamooz.com/what-is-etl

ETL به‌عنوان سنگ بنای پردازش داده‌ها، یک رویکرد ساختاریافته برای مدیریت داده‌ها ارائه می‌کند. در جهان کنونی که عصر کلان داده‌ها (Big Data) محسوب می‌شود، سازمان‌ها حجم وسیعی از data را به شکل مداوم از منابع مختلف دریافت می‌کنند. برای اخذ تصمیمات آگاهانه، تجزیه و تحلیل ترندها و دریافت بینش، لازم است سیستم سریع و قدرتمندی به‌منظور جمع‌آوری، پاک‌سازی و ذخیره‌سازی کارآمد داده‌ها وجود داشته باشد. در این شرایط، ETL به‌عنوان راه حل نجات‌دهنده سازمان‌ها مورد استفاده قرار می‌گیرد. در این مقاله، بررسی می‌شود که فرآیند ETL چیست و چه مزیتی‌هایی را به‌همراه دارد.

ETL یعنی چه ؟

ETL از سه واژه، استخراج (Extract)، تبدیل (Transform) و بارگذاری (Load) برگرفته شده است و فرآیند یکپارچه سازی داده ها (Data Integration) محسوب می‌شود. در این فرآیند یکپارچه‌سازی دیتا، ابتدا داده‌ها از منبع‌های مختلف جمع‌آوری می‌شوند، تبدیل و پالایش آن‌ها انجام می‌شود و درنهایت، به انبار داده (Data Warehouse) یا ریپازیتوری بارگذاری می‌شود. پس از بارگذاری داده‌ها به Data Warehouse، امکان تجزیه‌وتحلیل داده‌ها و گزارش‌گیری از آن به وجود می‌آید.



مؤلفه های اصلی ETL

ETL دارای ویژگی‌هایی است که در این بخش به آن‌ها می‌پردازیم.

- **تجمیع داده‌ها (Data Aggregation):** به واسطه ETL، می‌توان داده‌ها را از منابع مختلف جمع‌آوری و در یک نمای واحد و یکه تجمیع کرد. این کار در اخذ تصمیمات داده‌محور ضروری قلمداد می‌شود.
- **پاک‌سازی داده‌ها:** در ETL ، پاک‌سازی داده‌ها حائز اهمیت است و مواردی همچون تشخیص و رفع خطا، ناسازگاری‌ها و اشتباهات را در برمی‌گیرد.
- **تبدیل داده‌ها:** در فرآیند تبدیل / پالایش داده‌ها (Data Transformation)، اموری همچون تبدیل نوع های داده (Data Types)، ارزش بخشیدن به داده‌ها (ازطریق اطلاعات اضافه) و تجمیع داده‌ها در سطوح مختلف انجام می‌شوند. این مرحله برای اطمینان داشتن از سازگاری (Consistency) و فرمت داده‌ها لازم است.
- **بارگذاری داده‌ها:** ETL داده‌های پالایش‌شده را به یک انبار داده یا ریپازیتوری داده بارگذاری می‌کند. در این مرحله، داده‌ها برای امور تجزیه‌وتحلیل و گزارش‌گیری، بهینه‌سازی شده‌اند و آماده استفاده هستند.
- **مقیاس‌پذیری:** فرآیندهای ETL به گونه‌ای طراحی شده‌اند که امکان رسیدگی به حجم وسیعی از داده‌ها را به‌صورت کارآمد داشته و با رشد نیازمندی‌های دیتا، قابل Scale شدن باشند.
- **خودکارسازی:** شما می‌توانید فرآیندهای ETL را خودکارسازی کنید. با اتوماتیک شدن فرایندهای ETL، دیگر نیازی نیست که به‌صورت دستی بررسی کنید که آیا پردازش داده‌ها به‌درستی انجام می‌شوند یا خیر.

مزایای ELT چیست ؟

فرآیند های ETL به شما این تضمین را می‌دهند که داده‌ها به‌صورت تمیز (Clean) و قابل اکتفا هستند و آماده‌سازی آن‌ها برای تجزیه‌وتحلیل به‌درستی انجام شده است. سازمان‌ها فرآیند ETL را به‌منظور اخذ تصمیمات بهبودیافته و افزایش کارایی عملیاتی مجموعه خود به کار می‌برند.

عمده‌ترین مزیت های ETL عبارتند از:

- مجتمع‌سازی داده‌ها از منابع مختلف
- بهبود کیفیت کدها و رفع ناسازگاری‌های موجود در آن
- تبدیل داده‌ها به فرمت مناسب برای تجزیه‌وتحلیل
- نگهداری داده‌های تاریخی (Historical Data) و امکان پیگیری تغییرات و ترندها در طول زمان
- ارائه داده‌ها در قالب ساختاریافته (Structured) و تسهیل گزارش‌گیری و تحلیل آن‌ها
- بهبود رویکرد در تصمیم‌گیری‌های مبتنی‌بر داده‌ها

معایب ETL چیست ؟

هرچند وزن مزیت های ETL نسبت به چالش‌های آن سنگینی می‌کنند، اما کاستی‌هایی دارد که باعث می‌شوند ETL برای پروژه‌های بلندمدت و گسترده، انتخاب مناسبی قلمداد نشود. در ادامه به این موارد اشاره می‌شود.

- ابزارهای ETL امکان ذخیره‌سازی داده‌ها را ندارند و باید آن‌ها را در یک Repository متمرکز، همچون انبار داده‌ها نگهداری کنید.
- به منظور مشاهده آخرین داده‌ها در ابزارهای BI و بصری‌سازی، لازم است آن‌ها را به‌صورت دستی به‌روزرسانی کنید.
- فرآیند ETL برای داده‌های گسترده و در حجم بالا، سربار زیادی به همراه دارد؛ زیرا این فرآیند، نیازمند به‌روزرسانی و نگهداری مداوم است و می‌تواند برای سازمان پرهزینه و زمان‌بر باشد.

شایان ذکر است که راهکارهایی برای مواجه و بهبود این چالش‌ها فراهم شده است.

ابزارهای ETL

ابزارهای ETL ، راه حل‌های نرم‌افزاری خاصی هستند که فرآیند استخراج داده‌ها از منابع مختلف، تبدیل آن‌ها به قالب‌های سازگار و بارگذاری این داده‌ها به Data Warehouse، **بازار داده** (Data Mart) و دریاچه داده (Data Lake) را تسهیل می‌بخشند. این ابزارها برای اموری همچون، یکپارچه‌سازی داده‌ها، **مدیریت کیفیت داده ها** (Data Quality Management) و پردازش داده‌ها در **هوش تجاری** (BI)، تجزیه‌وتحلیل و گزارش‌گیری ضروری به شمار می‌روند. انواع مختلفی از ابزارهای ETL در دسترس هستند که از لحاظ Feature ها، قابلیت‌ها و پشتیبانی آن‌ها از دیتاسورس‌ها با یکدیگر تفاوت دارند.

رایج ترین ابزار های ETL کدامند ؟

ابزار های ETL رایج به شرح زیر است:

۱- آپاچی نایفای (Apache NiFi)

یک ابزار متن‌باز (Open Source) برای یکپارچه‌سازی داده‌ها است که قابلیت دریافت و پالایش آن‌ها را دارد. در Apache NiFi، دیتاسورس‌های مختلفی پیشتیبانی می‌شوند و می‌توان آن را برای پردازش دسته ای (Batch Processing) و بلادرنگ (Realtime) داده‌ها به کار برد.

۲- Talend

به‌عنوان یک ابزار ETL متن‌باز که به‌صورت گسترده مورد استفاده قرار می‌گیرد، Talend دارای Feature های جامع و مختلفی برای دریافت و تبدیل داده‌ها است. رابط کاربرپسند Talend و امکان پشتیبانی آن از انواع دیتاسورس‌ها، دو مزیت مهم این ابزار ETL تلقی می‌شوند.

۳- Informatica PowerCenter

یک ابزار ETL قدرتمند و با قابلیت سازگاری بالا محسوب می‌شود و در آن، Feature های مختلفی برای یکپارچه‌سازی داده‌ها، کیفیت آن‌ها و حاکمیت داده‌ها (Data Governance) ارائه شده است. Informatica PowerCenter قابلیت پشتیبانی از تبدیلات پیچیده را دارد و Data Profiling و پاکسازی داده‌ها را به‌صورت قدرتمند انجام می‌دهد.

۴ -آپاچی اسپارک (Apache Spark)

هرچند در وهله اول اسپارک یک فریم ورک محاسباتی توزیع‌شده (Distributed) است، اما Feature هایی مانند Spark SQL و Spark Streaming دارد که می‌توان آن‌ها را برای تسک‌های ETL به کار برد. آپاچی اسپارک به‌دلیل قابلیت پردازش بلادرنگ داده‌ها مورد توجه و شناخت است.

۵- SSIS

یکی از ابزارهای داخلی Microsoft SQL Server به‌حساب می‌آید که در آن، ویژگی‌های مشخصی برای تبدیل و یکپارچگی داده‌ها عرضه شده است. Microsoft SQL Server Integration Services یا همان SSIS، برای سازمان‌هایی مناسب است که از تکنولوژی‌های مایکروسافت استفاده می‌کنند و قصد دارند به‌صورت محدود از این ابزار بهره ببرند.

۶- Apache Camel

Apache Camel یک فریم ورک یکپارچه‌سازی متن‌باز محسوب می‌شود که تمرکز آن، روی مسیریابی و قوانین میانجیگری برای انتقال بین دو نقطه نهایی است. معمولاً از این Framework به‌همراه سایر ابزارها استفاده می‌شود تا یک راه حل ETL ایجاد شود.

۷- CloverDX

CloverDX، یک پلتفرم ETL و یکپارچه‌سازی داده‌ها است که مواردی همچون، تبدیل داده‌ها، پاکسازی و Orchestration پشتیبانی می‌کند. لازم به ذکر است که در CloverDX، تمرکز برروی کیفیت داده‌ها و حاکمیت آن‌ها است.

۸- Pentaho Data Integration

Pentaho Data Integration، یک ابزار ETL متن‌باز است که در آن، قابلیت‌های گسترده‌ای برای دریافت و تبدیل داده‌ها ارائه می‌شود. این ابزار که با نام Kettle نیز شناخته می‌شود، به‌دلیل امکان Drag and Drop و سادگی استفاده از آن، مورد توجه قرار دارد.

۹- IBM InfoSphere DataStage

یکی از بخش‌های InfoSphere، قسمت DataStage است که شرکت IBM در آن، ویژگی‌های مختلفی برای ETL و یکپارچگی داده‌ها عرضه کرده است. IBM InfoSphere DataStage به شما گزینه‌های مختلفی برای برقرار ارتباط قوی ارائه کرده و همچنین از پردازش داده‌ها با مقیاس بالا پشتیبانی می‌کند.

۱۱- Oracle Data Integrator

Oracle Data Integrator یا همان ODI، یک ابزار ETL است که در آن، قابلیت‌های گوناگونی برای یکپارچگی داده‌ها، کیفیت داده‌ها و تبدیل آن‌ها وجود دارد. این ابزار ETL ، توسط شرکت اوراکل ارائه شده است و به‌طور خاص، برای دیتابیس‌های اوراکل مناسب است. مطالعه **مقاله انواع پایگاه داده** می‌تواند به‌عنوان مطلب مکمل برای شما مفید باشد.

۱۲- SAS

در SAS (سیستم تجزیه و تحلیل آماری | Statistical Analysis System)، ابزارهای مدیریت داده‌ها، ازجمله Feature های ETL، ارائه می‌شود که از آن‌ها برای یکپارچگی داده‌ها، کیفیت داده‌ها و تجزیه‌وتحلیل پیشرفته استفاده می‌شود.

۱۳- TIBCO Jaspersoft ETL

بخشی از TIBCO Jaspersoft، ابزار ETL است که به کارگیری آن می‌توانید از قابلیت‌های تبدیل و یکپارچگی داده‌ها بهره ببرید. این ابزار به‌طور خاص برای گزارش‌گیری و طراحی داشبورد کارایی دارد.

۱۴- Alteryx

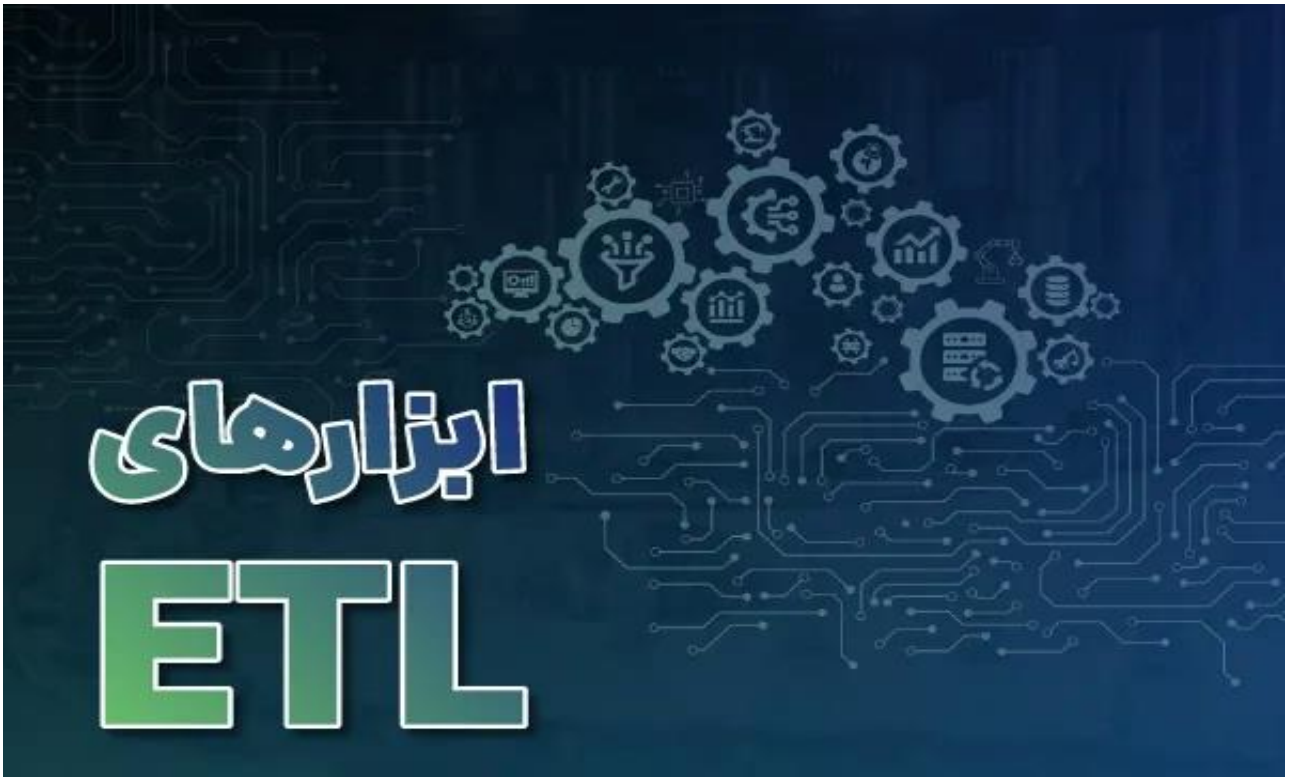
Alteryx یک پلتفرم تجزیه‌وتحلیل و آماده‌سازی داده‌ها محسوب می‌شود و در آن، قابلیت‌های ETL برای ترکیب داده‌ها (Data Blending)، پاک‌سازی داده‌ها و تجزیه‌وتحلیل پیشرفته فراهم شده است.

Matillion -۱۵

Matillion یک ابزار ETL ابر بومی (Cloud Native) است که به‌منظور یکپارچه‌سازی و پالایش داده‌ها در Data Warehouse های مبتنی‌بر فضای ابری طراحی شده است. به‌عنوان مثال، Amazon Redshift ، Google BigQuery و Snowflake، همگی انبارهای داده مبتنی‌بر فضای ابری به‌شمار می‌روند.

۱۶ -آپاچی کافکا (Apache Kafka)

هرچند Apache Kafka یک پلتفرم جریان توزیع‌شده است، اما می‌توان از آن برای دریافت بلادرنگ داده‌ها استفاده کرد و از آن به‌عنوان قسمتی از خط لوله (Pipeline) در ETL بهره برد.



فرایند ETL چگونه است ؟

فرآیند ETL معمولاً مراحل متوالی و مشخصی است که با کمک آن‌ها، سازمان‌ها می‌توانند داده‌ها را از انواع دیتاسورس‌ها استخراج کنند و پس از پالایش و پاک‌سازی این داده‌ها، آن‌ها را برای تجزیه‌وتحلیل و گزارش‌گیری در سیستم هدف یا انبارهای داده بارگذاری کنند. با این دیدگاه، در ادامه این بخش از مطلب آموزش ETL ، فرآیند ETL را به‌صورت مرحله‌به‌مرحله بررسی خواهیم کرد.

۱- استخراج (E)

- **تعیین منبع‌های داده:** مرحله اول استخراج در فرایند ETL ، تعیین دیتاسورس‌ها است. در این گام، شما باید تعیین کنید داده‌ها باید در کجا قرار بگیرند. به‌عنوان مثال، ممکن است داده‌ها در فایل‌ها، API ها، وب سرویس‌ها، پایگاه‌های داده و سایر سیستم‌ها استقرار داشته باشند.
- **استخراج داده‌ها:** اکنون لازم است استخراج داده‌ها از دیتاسورس‌های موردنظر انجام شود. این عمل می‌تواند شامل کوئری‌نویسی پایگاه داده، اسکرپ کردن محتوای وب و خواندن فایل‌ها با فرمت‌های گوناگون باشد.

۲- تبدیل (T)

- **پاکسازی داده‌ها:** در این مرحله از تبدیل / پالایش در فرایند ETL ، لازم است خطاها، مقادیر جامانده یا ناسازگاری‌ها از داده‌های استخراج شده، حذف یا تصحیح شوند. این گام از ETL ، دقت و کیفیت داده‌ها را تضمین خواهد کرد.
- **تبدیل داده‌ها:** اکنون باید داده‌ها به یک قالب یا ساختار استاندارد شده تبدیل شوند. این فرآیند، مواردی همچون تغییر نوع‌های داده، ادغام داده‌ها از چند دیتاسورس و اجرای محاسبات و تجمیع را شامل می‌شود.
- **غنی‌سازی داده‌ها (Data Enrichment):** با استفاده از اطلاعات اضافی، ازجمله افزودن داده‌های جغرافیایی، ادغام داده‌های ارجاعی و تولید فیلدهای مشتق شده، می‌توانید به غنی‌سازی و بهبود داده‌ها بپردازید.

۳- بررسی کیفیت داده ها

با بررسی دقیق کیفیت داده‌ها، مطمئن می‌شوید که داده‌های پالایش شده دقیقاً مطابق با استانداردها و قوانین کسب‌وکار شما هستند. در غیر این صورت، احتمالاً نیاز باشد که داده‌ها پاکسازی و به مقدار بیشتری پالایش و تبدیل شوند.

۴- بارگذاری (L)

- **بارگذاری داده‌ها:** در این گام از فرایند ETL ، لازم است داده‌ها به سیستم هدف بارگذاری شوند. معمولاً سیستم هدف، محیط‌هایی مانند انبار داده، بازار داده یا پلتفرم‌های تجزیه‌وتحلیل داده هستند. داده‌ها به گونه‌ای ساختاریافته شده‌اند که کوئری‌نویسی و تجزیه و تحلیل روی آن‌ها به‌صورت کارآمد امکان‌پذیر است.
- **شاخص‌گذاری داده‌ها و Performance Tuning:** به منظور بهینه‌سازی کارایی کوئری‌ها و همچنین استخراج داده‌ها، می‌توان از شاخص‌گذاری (indexing) و اجرای Performance Tuning استفاده کرد. این مرحله باعث می‌شود شما خیالتان راحت شود که داده‌ها برای گزارش‌گیری و تجزیه‌وتحلیل آماده و قابل دسترس هستند.

۵- اعتبارسنجی داده ها

علاوه‌بر مراحل استخراج، تبدیل و بارگذاری در ETL ، شما باید به اعتبارسنجی داده‌هایی که بارگذاری شده‌اند، توجه کنید. با این اعتبارسنجی، مطمئن خواهید شد که هیچ داده‌ای در طول فرایند ETL ، گم یا تخریب نشده است.

۶- زمان بندی و خودکارسازی

شما می‌توانید فرایند ETL را به گونه‌ای خودکارسازی کنید که برمبنای یک زمان‌بندی مستمر (شبانه یا روزانه) اجرا شود. این کار باعث می‌شود سیستم هدفتان مطابق با آخرین داده‌های دریافتی از دیتاسورس‌ها، به‌روزشده باقی بماند.

۷- ورود (Logging) و نظارت (Monitoring)

شما می‌توانید با کمک لاگین و مانیتورینگ، کارایی فرایند ETL را پیگیری کنید و امکان تشخیص و توجه به خطاها را داشته باشید. لازم به ذکر است که روال‌هایی (Procedures) برای رسیدگی به خطاهای احتمالی در فرایند ETL توسعه یافته‌اند.

۸- نگهداری و تکرار فرآیند

نگهداری (Maintenance) و تکرار (Iteration) در فرایند ETL اهمیت زیادی دارد؛ زیرا شما باید به‌صورت مداوم به حفظ و به‌روزرسانی فرایند ETL بپردازید تا آن را مطابق با منبع‌های داده‌ها، نیازمندی‌های کسب‌وکار و مدل‌های داده‌ها هماهنگ کنید.

۹- مصرف داده ها

زمانی که داده‌ها به سیستم هدف شما بارگذاری شوند، این امکان فراهم می‌شود که این Data را برای امور مختلفی، از جمله هوش تجاری، گزارش‌گیری، تجزیه‌وتحلیل و تصمیم‌گیری به کار ببرید. شایان ذکر است که بهتر است تمام فرآیند ETL و جزئیات مراحل مختلف آن را مستندسازی کنید.

آینده ETL چیست؟

آینده ETL ، به سمت‌وسوی عملیات بلادرنگ، ابر بومی و مبتنی‌بر داده‌ها سوق داده شده است. فرایند های ETL به گونه‌ای درحال تکامل هستند که از تجزیه‌وتحلیل داده‌های جریانی پشتیبانی کنند. این موضوع به سازمان‌ها، امکان تصمیم‌گیری سریع و آنی براساس داده‌ها را خواهد داد. در طول زمان، راه حل‌های ETL ابر بومی به شهرت فراوانی رسیده‌اند و به همین دلیل، ایجاد Data Pipeline هایی که منعطف و مقیاس‌پذیر باشند، تسهیل پیدا کرده‌اند.

به‌صورت کلی، آینده ETL در جهت سرعت، خودکارسازی و هوشمندی است و با کمک آن، سازمان‌ها امکان بهره‌وری مطلوب از داده‌ها به‌منظور دریافت بینش و اخذ تصمیمات آگاهانه را خواهند داشت. در این بخش، به این سؤال پاسخ داده شد که چرا سازمان ها به ETL نیاز دارند. در بخش بعد، اهمیت و نقش ETL در هوش تجاری را مورد بررسی قرار می‌دهیم.

نقش ETL در هوش تجاری

همانطور که در **مقاله نقش ETL در هوش تجاری** بررسی کردیم، ETL در اکوسیستم هوش تجاری، یک بخش جدایی‌ناپذیر تلقی می‌شود و به‌عنوان لایه یکپارچه‌سازی و آماده‌سازی داده‌ها عمل می‌کند. به‌واسطه فرایند ETL در هوش تجاری، این اطمینان حاصل می‌شود که داده‌ها قابل اکتفا، سازگار و آماده برای تجزیه و تحلیل در ابزارهای BI هستند. شما می‌توانید برای یادگیری نحوه ساخت داشبوردها و گزارش‌گیری ازطریق ابزارهای هوش تجاری، **دوره آموزشی طراحی داشبورد با Power BI Desktop** را مشاهده کنید.

تفاوت فرایند ETL و ETL

فرایند ETL و ELT ، دو رویکرد یکپارچه‌سازی داده‌ها محسوب می‌شوند که ترتیب پردازش داده‌ها تفاوت دارند. در ETL، ابتدا داده‌ها از دیتاسورس‌ها استخراج می‌شدند و پس از آن، به گونه‌ای تبدیل و پالایش می‌شدند که مناسب سیستم هدف (معمولاً انبار داده) باشند. در نقطه مقابل، در فرایند ELT ، داده‌ها به‌صورت مستقیم به سیستم هدف منتقل می‌شوند و فرآیند پالایش آن‌ها، بعد از دریافت، یا همان Ingestion داده، رخ می‌دهد. معمولاً سیستم هدف در ELT، دریاچه داده یا انبار داده است.

به‌صورت کلی، ETL برای داده‌های ساختاریافته در Data Warehouse ها به کار می‌رود، در حالی که فرایند ELT برای پلتفرم‌های مبتنی‌بر فضای ابری و مدرن داده‌ها مناسب هستند؛ زیرا این پلتفرم‌ها، امکان رسیدگی به داده‌های نیمه‌ساختاریافته (Semi-Structured) یا خام (Raw) را دارند و می‌توانند در تجزیه‌وتحلیل و پردازش داده‌های در مسیر مقصد، انعطاف‌پذیر عمل کنند. برای مطالعه تفاوت‌های این دو فرایند، پیشنهاد می‌کنیم **مقاله تفاوت فرایند ETL و ETL** را مطالعه کنید.

انتخاب بین ETL وELT

اگر بخواهید میان فرایند ETL و ELT یکی را انتخاب کنید، موارد زیر در تصمیم‌گیری شما حائز اهمیت خواهند بود:

- توجه به ساختار داده‌ها
- اهمیت حجم داده‌ها
- توجه به ریپازیتوری مورد استفاده (اینکه Data Lake یا Data Warehouse است.)
- سرعت پردازش داده‌ها
- مقیاس‌پذیری و مقدار هزینه
- توجه به پیچیدگی‌های تبدیل داده‌ها
- بررسی مورد استفاده و نیازمندی‌های سازمان

مزایای ELT چیست ؟

هرچند فرایند ETL فواید خاص خود را دارد، اما بهتر است با مزیت‌های فرآیند ELT آشنا شوید. مزیت‌های ELT عبارتند از:

- سازگاری با دریاچه داده‌ها
- مقیاس‌پذیری و سرعت مناسب
- امکان ذخیره‌سازی داده‌های خام
- کاهش هزینه ذخیره‌سازی و جلوگیری از تکرار داده‌ها

جمع بندی ETL: چیست و چه کاربردی دارد ؟

فرایند ETL یکی از بخش‌های بنیادی از پایپ لاین داده‌ها محسوب می‌شود و نقش کلیدی و پراهمیتی در اطمینان از قابل اکتفا بودن داده‌ها و مناسب بودن‌شان برای گزارش‌گیری و تجزیه‌وتحلیل ایفا می‌کند. در این مطلب آموزش ETL، ابتدا به این سؤال پاسخ داده شد که مفهوم ETL چیست و سپس ابزارهای کاربردی آن، به‌همراه چگونگی عملکرد این فرآیند به‌طور مفصل و با جزئیات شرح داده شدند. به‌طور کلی، ETL در مدیریت کارآمد داده‌ها، بهبود کیفیت آن‌ها و تسهیل فرایند تصمیم‌گیری سازمان‌ها کاربردی است و با کمک آن، می‌توان به بینش عمیقی از داده‌ها رسید و کسب و کار را در مسیر موفقیت و پیشرفت هدایت کرد.