ETL Developement

Tushar B. Kute,
http://tusharkute.com

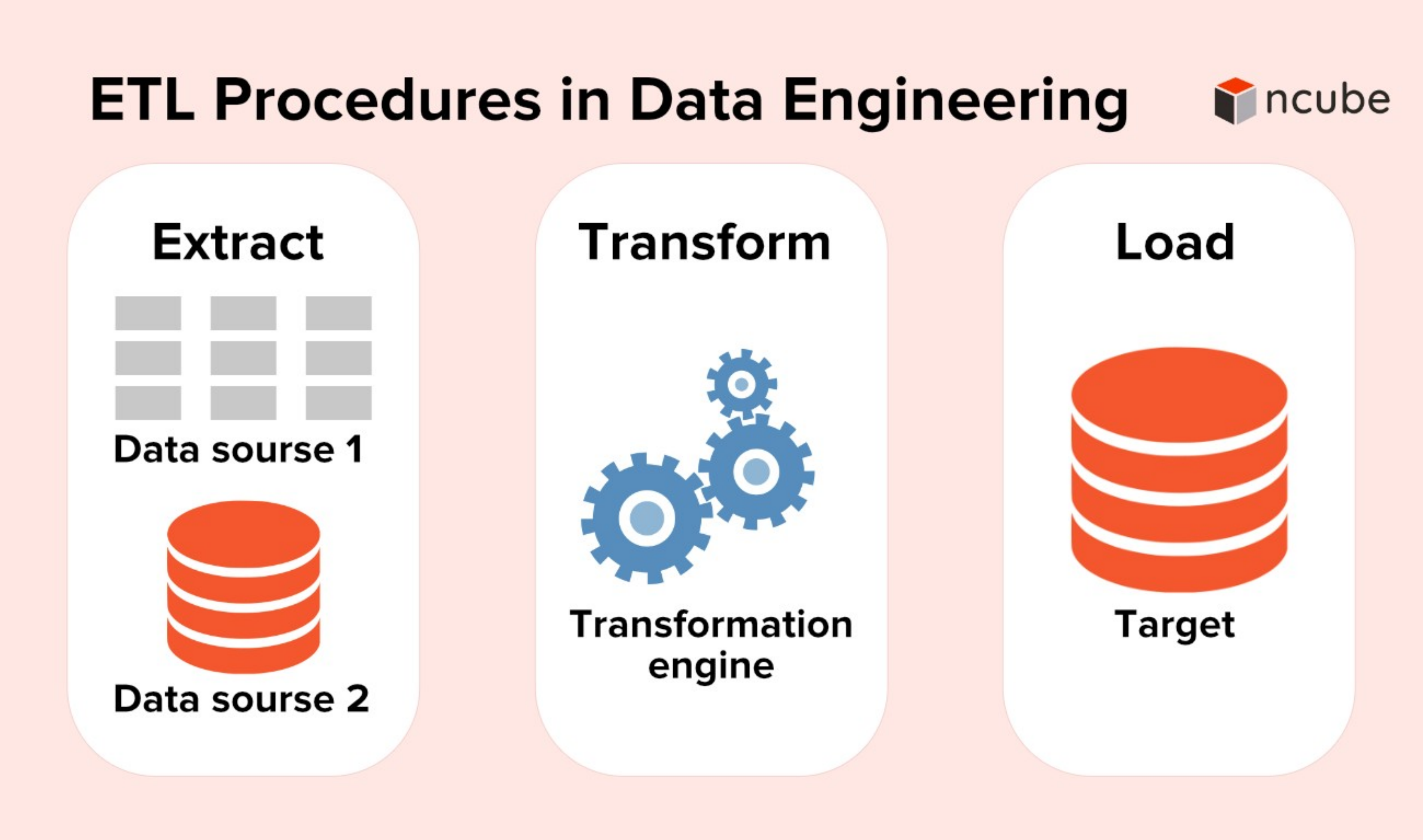
Introduction

- The roles of BI engineers, data scientists, and data analysts have become an integral part of any organization that practices Business Intelligence (BI).
- In their line of work, data professionals aim to bring value by gleaning useful insights from data.
- But data needs to undergo a preparation step before the data team can use it. That is where ETL experts step in.

What is ETL?

- The abbreviation refers to Extracting, Transforming, and Loading data.
- An ETL developer designs, develops, and maintains data storage systems and ensures they contain business-relevant data.
- ETL is a three-fold process and includes the following components.

What is ETL?


ETL Procedures in Data Engineering
ncube
Extract
Data sourse 1
Data sourse 2
Transform
Transformation engine
Load
Target

Extraction

- Extraction. Businesses generate massive volumes of data. It is stored across multiple systems and in diverse formats.
- Data needs to travel freely between the systems to power your data strategy.
- Extraction means consolidation of data in required sources like cloud environments, CRM, or external systems.
- ETL developers use ETL tools to automate this process.

Transformation

- Transformation. This step is about structuring and formatting data. After data is collected from its sources, it's in a raw state. It needs to transform to be compatible with the defined standards. Thus, this step includes the following steps:
  - Cleansing: removing inconsistencies and missing values.
  - Standardization: bringing datasets into a required format.
  - Deduplication: excluding irrelevant data.
  - Verification: removing data that can't be used and marking aberrations.
  - Sorting: organizing data by type.

Loading

- Loading. The final step is loading the transformed data into a database called the Data Warehouse system.
- Sometimes loading can be frequent, and sometimes it is done at regular intervals. Unlike a typical database, Data Warehouse includes tools to make it accessible for users.
- These are Business Intelligence tools that let you visualize and report data.

What ETL developers do?

- Generally, ETL developers design, develop, automate, and support complex applications to extract, transform, and load data. To be more exact, the duties of ETL developers are as follows:
  - Identifying data storage requirements. ETL developers determine the storage needs of the company. They need a bird's eye view of the data situation to choose the best suiting option.
  - Building a Data Warehouse. After figuring out the needs, the ETL developers build a data warehouse tailored to an organization's needs.

What ETL developers do?

- Building reliable data pipelines: a sum of tools and processes that bring data to the user. Pipelines connect data between systems and transfer it from one format into another.
- ETL processes. Once the data warehouse is complete, the ETL developer extracts the data and delivers it to the new system.
- Quality Assurance. After the warehouse gets off the ground, the ETL developers run tests to ensure its stability.
- Debugging. ETL developers rectify all problems with the warehousing system.

ETL Developers Skillset

- The ETL Toolbox
  - ETL tools are the ready-made solutions that can perform ETL steps right from the start and move data between the sources.
  - The industry standards are Informatica, Talend, and Pentaho. An ETL developer administers the integration of tools with other instruments and implements an interface to ensure data is usable.

ETL Developers Skillset

- SQL
  - ETL processes rely on SQL since it's the most common database language and is used at every stage of the ETL process.
  - In fact, ETL tools are intrinsically SQL generators, so it's vital to have both of those skills.

ETL Developers Skillset

- Scripting language knowledge
  - Sometimes ETL tools aren't enough to cope with all of the requirements. When that happens, ETL developers have to do the heavy lifting with the systems they are working with.
  - Thus, knowledge of a scripting language helps ETL specialists deal with files, users, permission issues, and more.

ETL Developers Skillset

- Data modeling
  - An ETL developer defines the data formats, the way data will represent in the warehousing system. These are called data models.
  - Reading, analyzing, and transforming data are the key skills that enable an ETL developer to determine data output formats in a database.
  - Those data models define the tools required for transformation.

ETL Developers Skillset

- Database engineering
  - Database engineers usually create databases, but ETL engineers are also expected to have a background in data mapping and SQL databases to oversee the development process.
  - Additionally, they need to have a solid knowledge of data warehouse architecture's concepts, components, and techniques.

ETL Developers Skillset

- High levels of self-organization
  - The day of an ETL developer can be pretty intensive, filled with various tasks from business, team management, and tech fields.
  - That's why it's crucial to master time-management skills to fare well in this role.

When ETL developers needed?

- An ETL expert will be a part of your data team. The main reason to hire such a team is running a complex, large-scale data system. Here's when you should consider adding an ETL expert to your team:
  - Your business generates more and more varied, unstructured data
  - You want to be data-driven and thus need insightful data representation;
  - Your business strategy relies on business intelligence development or machine learning
  - Your data processing system is outdated;
  - You feel like your current data methods could use improvement;

ETL in Hadoop

- It is important to consider the following points before you set up an ETL in Hadoop:
  - You can use Views when you have a transactional system where data is continuously changing and use tables when data is not changing.
  - In Hadoop, mapping of data is far cheaper than partitioning the data based on a query, especially in transactional data.
  - It is advisable to program in parallel to maximize Hadoop’s processing power. It can be done by breaking logic into multiple phases and running these steps in parallel to speed up your ETL in Hadoop.

ETL in Hadoop

- Use Managed Tables in Hadoop as they are better to govern except when you are importing data from the external system, in such a case, you can use External Tables.
- After that, you need to define a schema for it, and the entire workflow creates using Hive scripts.
- Phase Development is also useful while setting up ETL in Hadoop.
- All you need to do is keep parking processed data into various phases and then keep treating it to obtain a final result.

ETL in Hadoop

- Pig and Hive are commonly used by Hadoop vendors (Cloudera, Hortonworks, etc.).
- Using Hive can be quite complicated compared to Pig because of the logic, but there are very few people who know Pig.
- You would need resources in the future to maintain the code so the solution you choose matters.
- Various industries are using Hive, and companies like Hortonworks, AWS, MS, etc. are contributing to Hive

ETL in Hadoop: Steps

- 1. Setting up a Hadoop Cluster
  - The first step of setting up ETL in Hadoop requires you to build a Hadoop cluster and decide where you want to create your cluster.
  - It can be locally in an in-house data center or in the cloud, depending on the type of data you want to analyze. If it is an in-house data center, you will have to consider whether the data can be moved to the cloud subsequently and can test data used for development.
  - For data clusters on the cloud, such as Cloudera, Hortonworks, MapR, Amazon Elastic Map Reduce, Altiscale, Microsoft, Rackspace CBD, or other Hadoop cloud offerings, it can be done in a few clicks.

ETL in Hadoop: Steps

- 2. Connecting Data Sources
  - The Hadoop ecosystem has varieties of open-source technologies that complement and increase its capacities. They enable you to connect different data sources after setting up your ETL in Hadoop.
  - The data sources could be a database, Relational Database Management System (RDBMS), machine data, flat files, log files, web sources, and other sources such as RDF Site Summary (RSS) feeds.

ETL in Hadoop: Steps

- 2. Connecting Data Sources
  - The ETL tools for connecting these data sources include Apache Flume and Apache Sqoop, Apache HBase, Apache Hive, Apache Oozie, Apache Phoenix, Apache Pig, and Apache ZooKeeper.
  - While setting up ETL in Hadoop, you have to plan your data architecture depending on the amount of data, type, and the rate of new data generation.

ETL in Hadoop: Steps

- 3. Defining the Metadata
  - Even though you can store data in Hadoop and decide how to use them later, it is essential to define the semantics and structure of data for analytics purposes.
  - The classification process will help you in the transformation of data as you desire by defining the metadata.
  - You can remove the ambiguity from ETL in Hadoop, of how a field looks and generates, with a transparent design and documentation.

ETL in Hadoop: Steps

- 3. Defining the Metadata
  - For example, you can define a field – student ID in the warehouse either as a five-digit numeric key, generated by an algorithm or as a four-digit sequence number that appends to an existing ID.

ETL in Hadoop: Steps

- 4. Creating Jobs for ETL in Hadoop
  - Now to execute your ETL in Hadoop, you need to focus on the process of transforming the data from various sources.
  - Technologies, such as MapReduce, Cascading, Pig, and Hive are some of the most commonly used frameworks for developing ETL jobs.
  - Deciding which technology to use and how to create the jobs depends on the data set and transformations.

ETL in Hadoop: Steps

- 4. Creating Jobs for ETL in Hadoop
  - Identify if your data extraction is a batch job or a streaming job.
  - A batch job for ETL in Hadoop takes the whole file, processes it, and saves it to a larger file.
  - A streaming job takes data from a Relational Database Management System (RDBMS), where the data transfers separately one after the other for further processing.
  - Various ETL systems cope with these tasks in different ways.

ETL in Hadoop: Steps

- 5. Creating the Workflow for ETL in Hadoop
  - This is the final step of setting up ETL in Hadoop.
  - Creating a workflow with multiple ETL jobs, each carrying out a specific task, helps in the transformation and cleansing of data efficiently.
  - These data mappings and transformations execute in a particular order.
  - There may also be dependencies to check, and these dependencies are captured in the workflow.

ETL in Hadoop: Steps

- 5. Creating the Workflow for ETL in Hadoop
  - Parallel workflows result in parallel execution of data, thereby speeding up the ETL process.
  - A smooth workflow can be derived by scheduling it to run hourly, nightly, weekly, or as frequently as you so wish.
  - After you have done all of this, a data warehouse is created, and the data will be ready for analysis.
  - Hive, Impala, and Lingual provide SQL-on-Hadoop functionality along with several commercial BI tools that can connect to Hadoop to explore the data and generate visually appealing reports

ETL vs ELT

- The ETL process is the backbone of all the Data Warehousing tools. ETL in Hadoop solved the prominent problems of data i.e. Velocity, Volume, and Variety. In this section let’s have a look at the ETL vs ELT process on Hadoop.
- ETL tools have been serving the Data Warehouse needs but the changing nature of the data forced organizations to shift to Hadoop. ETL in Hadoop is a cost-effective and scalable solution.

ETL vs ELT

- ELT on Hadoop delivered flexibility in a data processing environment. Shifting from the traditional ETL process to ELT on Hadoop is a challenge for organizations but ELT on Hadoop is a better choice in a long term.
- The ELT in Hadoop separates the loading nad transformation tasks into independent blocks making the project management easier whereas ETL in Hadoop loads the important data, as identified at design time.

Summary

- Although Hadoop is a useful big data storage and processing platform, it can also be limiting as the storage is cheap, but the processing is expensive.
- You cannot complete a job in sub-seconds as it takes a longer time. It is also not a transactional system as source data does not change, so you have to keep importing it over and over again.
- However, setting up in-house ETL in Hadoop demands technical proficiency.
- Furthermore, you will have to build an in-house solution from scratch if you wish to transfer your data from any source to Hadoop or another Data Warehouse for analysis.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com