📙

Aula 51 - O conceito de ETL

Ciclo	Ciclo 07: ETL
# Aula	51
Finished	☑

Objetivo da Aula:

1. O que é um ETL?
2. A etapa de extração
3. A etapa de transformação
4. A etapa de armazenamento
5. Na próxima aula

Conteúdo:

▼ 1. O que é um ETL

ETL é o acrônimo em inglês para Extraction, Tranformation e Load (Extração, Transformação e Carga).

O ETL é uma forma de organizar as etapas de um processo de dados, que busca o dado em sua fontes, fazendo um coleta (Extração), passando pela etapa de limpeza e preparação (Transformation), até a última etapa de armazenamento dos dados (Loading).

Essas etapas ajudam a organizar as fases do processamento dos dados, de modo a implementar ferramentas especializadas de cada fase.

▼ 2. A etapa da Extração

A etapa da extração consiste na coleta dos dados de diversas fontes. Em alguns ETLs, esses dados são salvos em algum repositório de dados para haver um independência da fonte geradora.

▼ 2.1. As fontes mais comuns

1. Banco de Dados interno.
2. Web sites.
3. Banco de dados externos.
4. FTP (File Transfer Protocol)
5. API (Application Programming Interface)
6. Excel
7. Google Sheets
8. Pasta no Google Drive / One Drive
9. Emails

▼ 2.2. As ferramenta de extração

1. Linguagens de Programação (Python, R, SQL, Shell Script)
2. Stitch: https://www.stitchdata.com/
3. AWS Glue da Amazon
4. Google Cloud Data Flow da GCP
5. Pentaho Data Integration
6. Alteryx
7. IBM Data Stage
8. Talend Data Fabric
9. Apach NiFi

▼ 3. A etapa da Transformação

A etapa da trasformação possui tarefas como a limpeza, transformação e preparação dos dados. Em alguns ETLs, esses dados limpos, tratados e preparados são salvos em um novo repositório de dados, criando uma nova camada de dados.

O armazenamento dos dados a cada ação é o princípio de um data lake.

▼ 2.1. As ferramentas da transformação.

1. Linguagens de Programação (Python, R, SQL, Shell Script)
2. Apach NiFi
3. AWS Glue da Amazon
4. Google Cloud Data Flow da GCP
5. Pentaho Data Integration
6. Alteryx
7. IBM Data Stage
8. Talend Data Fabric

▼ 4. A etapa da Armazenamento

A etapa da armazenamento consiste na etapa de salvamento dos dados limpos, tratados e preparados, em um repositório de dados. Dentro do conceito de data lake, esse repositório será a camada onde os dados serão disponibilizados para uso das pessoas, ferramentas ou sistemas.

▼ 3.1. As ferramentas da armazenamento.

1. Banco de Dados SQL
    a. MySQL, Postgres, Maria DB, Redshift, SQL Server, Oracle.
2. Banco de Dados NoSQL
    a. HBase, Cassandra, Mongo DB, etc.
3. Redshift da AWS
4. Big Query da GCP

5. Snowflake
6. SQL Database da Azure
7. S3

▼ Na próxima aula

Aula 41: O planejamento do ETL