



Extract, Transform, Load (ETL)

Prof. MSc. Fernando Prass

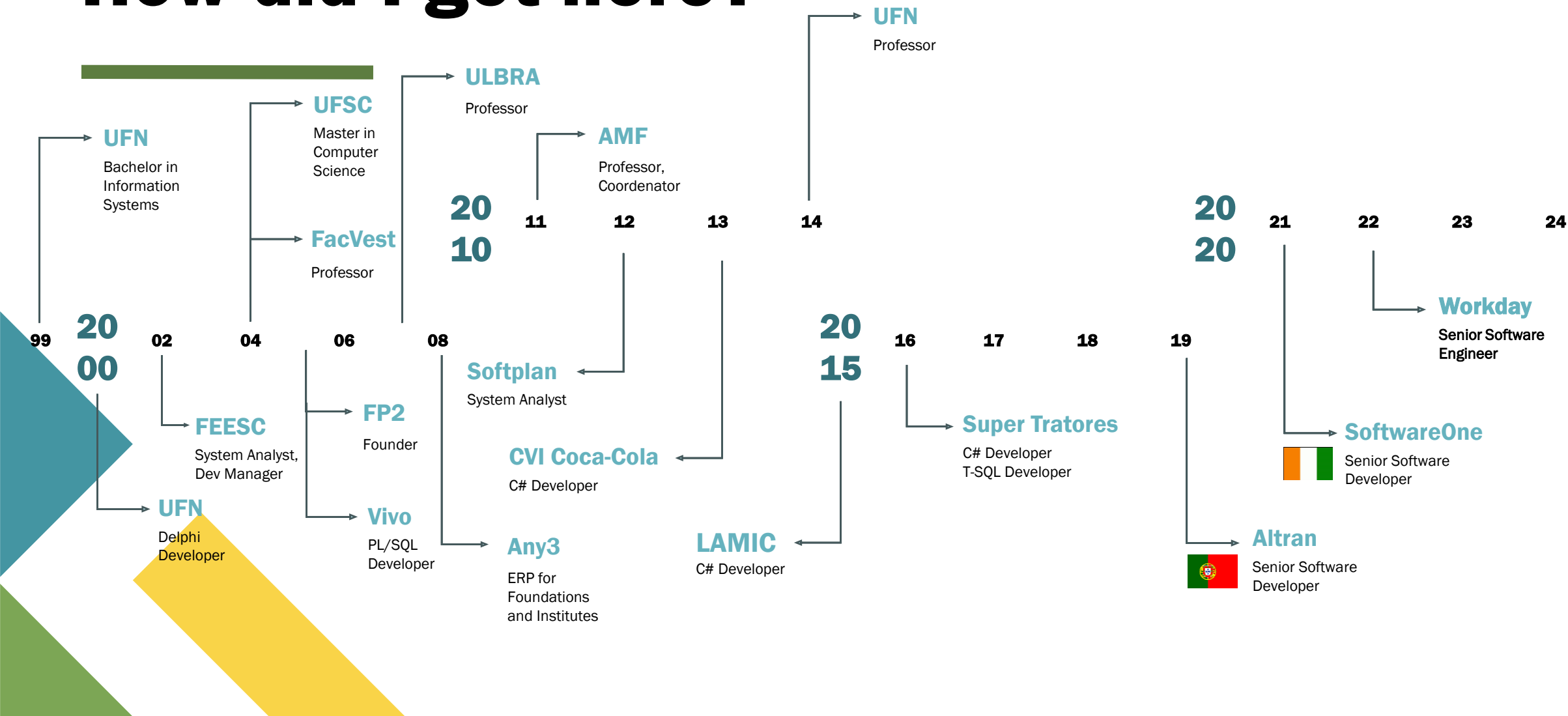
fprass@gmail.com

github.com/fernandoprass/etl

Who am I?

- Fernando Sarturi Prass
- Bachelor in Information Systems
- Master in Computer Science
- Senior Software Engineer (tech leader) at Workday
- fprass@gmail.com
- github.com/fernandoprass
- twitter.com/oFernandoPrass
- linkedin.com/in/fernandoprass
- lattes.cnpq.br/2919187023046130

How did I get here?



What do I currently do?

- Senior Software Engineer at Workday in Dublin, Ireland
 - Tech Leader and Scrum Master
 - Main activities:
 - Integrate our ERP payroll system with third party payroll software
 - Support and coach my teammates in performing tasks
 - Ensure that agile processes are followed, and the team delivers software quickly and efficiently

Fun fact!!!!

The biggest mistake
I ever made...





Fun facts!!!!

The coolest thing I've ever done.....

Introduce yourself

- Who are you?
- What do you currently do?
- Why you are here?
- Fun facts
 - The biggest mistake you've ever made...
 - The coolest thing you've ever done...

Before we start...

- What tools/languages/technologies will we use in this course?
- What prior knowledge should I have?

Agenda

- Introduction
- Extract
- Transform
- Load





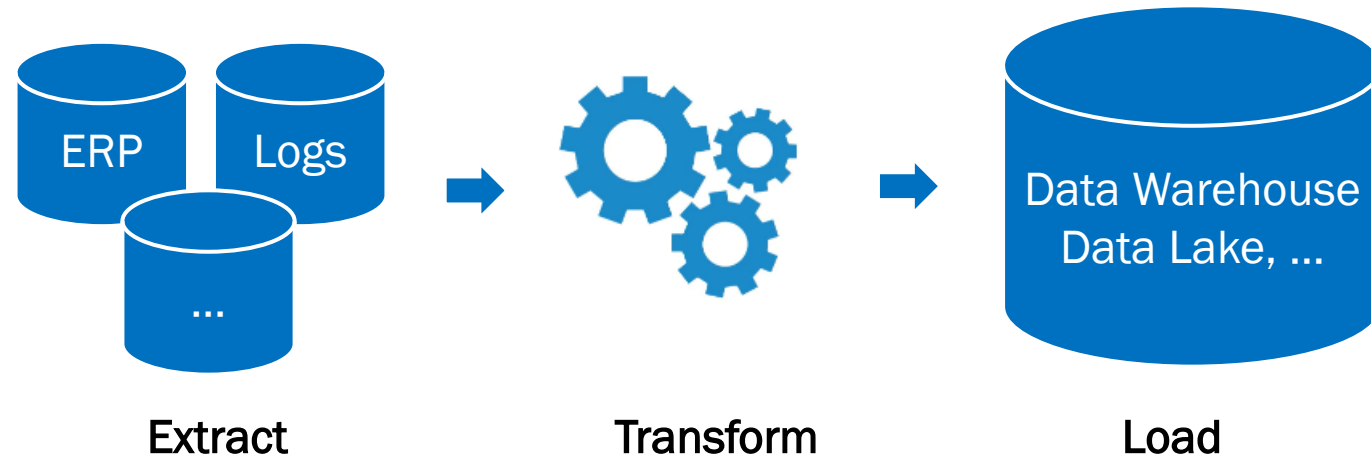
Introduction

“The greatest enemy of knowledge is not ignorance; it is the illusion of knowledge”

Daniel J. Boorstin

Extract, Transform, Load (ETL)

- Extract, Transform, Load (ETL) *“is a data integration process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a data warehouse, data lake or other target system”*. (IBM)



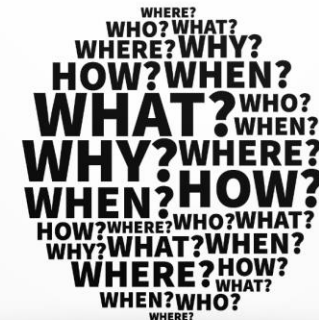
ETL



- For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven.
- Typically, when it is possible, the ETL load process takes place during off-hours when traffic on the source systems and the target repository is at its lowest.

Question

- Can you cite examples of ETL that you have already performed?



Extract

- During data extraction, raw data is copied or exported from source locations to a staging area. Source locations can come from a variety of different sources, structured or unstructured:
 - Databases (SQL or NoSQL)
 - CRM and ERP systems
 - JSON and XML files
 - Flat-file databases
 - Emails
 - Web pages
 - Spreadsheets
 - APIs
 - ...

Transform



- In the staging area the data is transformed for its intended use.

Transformed can mean:

- Filtering, cleaning, aggregating, de-duplicating, and validating;
- Performing calculations, translations, summarizations or conversions (units of measurement, currency, ...);
- Conducting audits to ensure data quality and compliance, and computing metrics.
- Removing, encrypting or protecting private data (GDPR);
- Formatting the data into tables or joined tables to match the target schema, change/update titles and description.

Load



- In this last step, the transformed data is moved from the staging area into a target repository (data warehouse, data lake, DB), where the data is ready for querying, reporting, analytics, or any other downstream processes.
- Usually, this process begins with an initial load of all data, followed by regular updates to incorporate incremental changes. Occasionally, there are full refreshes to replace existing data in the target repository

Real-life ETL cycle

1. Cycle initiation
2. Build reference data
3. Extract (from sources)
4. Validate
5. Transform (clean, apply business rules, check for data integrity,)
6. Stage (load into staging tables, if used)
7. Audit reports (helps to diagnose/repair)
8. Publish (to target tables)
9. Archive

Question

- Which of the three phases
 1. Extract
 2. Transform
 3. Loadis the most difficult?



Why is ETL important?

- Customer data from online payment and customer relationship management (CRM) systems
- Inventory and operations data from vendor systems
- Sensor data from Internet of Things (IoT) devices
- Marketing data from social media and customer feedback
- Employee data from internal human resources systems

ETL vs ELT

- The most obvious difference between ETL and ELT (Extract, Load, Transform) is the order of operations. But there are more. ELT ...:
 - ... copies or exports the data from its original locations, bypassing the staging area for transformation, and instead directly inputs the unaltered data into the destination datastore, where it can be transformed as necessary.
 - ... is useful for ingesting high-volume, unstructured data sets as loading can occur directly from the source.
 - ... can be more ideal for big data management since it doesn't need much upfront planning for data extraction and storage.

Other data integration methods

- Change Data Capture (CDC)
- Data Replication
- Data Virtualization

Change Data Capture (CDC)

- **Change Data Capture (CDC)** identifies and captures only the source data that has changed and moves that data to the target repository. CDC can be used to reduce the resources required during the ETL “extract” step.
- Example: In a customer database, when a new customer is added or an existing customer’s details are updated, CDC captures these changes and updates a data warehouse where customer information is aggregated for analysis.

Change Data Capture - Pros

- **Real-time Integration:** Ensures systems have up-to-date information.
- **Reduced Workload:** Processes only data changes, not all records.
- **Minimized Impact:** Designed to have minimal impact on source system performance.
- **Informed Decisions:** Real-time data leads to timely business decisions.

Change Data Capture - Cons

- **Complexity:** Implementation can be complex, requiring specialized tools.
- **Overhead Costs:** Additional costs for maintaining the CDC system.
- **Potential Errors:** If not managed well, can introduce replication errors.

Data Replication

- **Data Replication** copies changes in data sources in real-time or in batches to a central database. It is often listed as a data integration method, but in fact it is most often used to create backups for disaster recovery.
- Example: Consider a global website, when a customer access in Brazil, he is accessing a regional server hosted locally to ensure that the speed connection is good.

Data Replication - Pros

- **High Availability:** Ensures that data is always accessible.
- **Load Balancing:** Distributes read and write operations across multiple servers.
- **Fault Tolerance:** Provides backup in case of server failure.
- **Enhanced Scalability:** Accommodates increased traffic and workload demands.

Data Replication - Cons

- **Complexity:** Requires careful planning and management.
- **Resource Intensive:** Can be costly in terms of hardware and bandwidth.
- **Data Synchronization Challenges:** Ensuring that all copies of the data are consistent can be challenging.
- **Higher Storage Costs:** Maintaining multiple copies of data can lead to increased storage costs.
- **Additional Security Threats:** More copies of data mean more potential targets for unauthorized access.

Data Virtualization



- **Data Virtualization** is a technology that abstracts and integrates data from various sources, providing a unified, real-time view without physically copying, transforming or loading the source data to a target repository.
- **Example:** Consider a multinational company with different departments using various data storage systems. With data virtualization, an employee in the marketing department, for example, can access sales data, customer feedback, and social media comments all from a single interface, without needing to know where and how the data is stored

Data Virtualization



- This approach enhances flexibility, reduces redundancy, and improves security by centralizing data management.
- It supports analytics and reporting by offering timely, integrated data, which is particularly beneficial for businesses that need to make informed decisions based on comprehensive data insights.
- While data virtualization can be used alongside ETL, it is increasingly seen as an alternative to ETL and to other physical data integration methods.

Data Virtualization - Pros

- **Real-time Access:** Data virtualization provides real-time access to data.
- **Reduced Complexity:** It abstracts the technical details of data storage, making data access simpler for users.
- **Efficiency:** It can be less error-prone and more efficient than traditional data integration methods.
- **Flexibility:** Data virtualization allows users to simply describe the desired outcome, and the software adjusts the intermediate steps accordingly.

Data Virtualization - Cons



- **Operational Requirement:** All necessary data sources must be operational as there is no local copy of the data.
- **Potential Complexity:** Implementing data virtualization can be complex and may require specialized tools or software.
- **Overhead Costs:** There may be additional costs associated with maintaining the data virtualization system.
- **Potential for Errors:** If not properly managed, data virtualization systems can introduce errors during the replication process.

ETL vs Data Virtualization

	ETL	Data Virtualization
Purpose	Ready, change and copy data from a set of sources to a different one	Focuses on providing a unified view of data from various sources without physically moving or duplicating it
Data Movement	Involves data extraction from source systems, transformation, and then loading into a separate storage system.	Minimizes data movement by accessing data in real-time from its original sources
Latency	Typically introduces some latency due to the batch processing nature of ETL pipelines	Provides low-latency access to data, suitable for real-time analytics and reporting

ETL vs Data Virtualization


	ETL	Data Virtualization
Data Storage	Requires a dedicated data warehouse or database for storing transformed data	Does not create a separate storage layer; data remains in its original locations
Complexity	More complex due to data extraction, transformation, and loading processes	Simpler to set up and maintain, as it doesn't involve data replication
Use Cases	<ul style="list-style-type: none">• Data warehousing• Data migration• Data consolidation	<ul style="list-style-type: none">• Real-time reporting and analytics• Federated queries across multiple data sources• Agile data access

Machine Learning in ETL

- Machine learning can be used in various stages of ETL process, including data extraction, data cleaning, and data integration. It can improve the ETL process in many ways, such as increasing the accuracy and completeness of data, and reducing time and resources required for manual data cleaning.
- ML algorithms can be used to automatically to:
 - extract structured and unstructured data from various sources, such as social media and emails.
 - clean and transform data, such as identifying and removing duplicate or incorrect data, and standardizing data formats.

ML in ETL - Benefits

A solid green horizontal bar.

- Increased Efficiency
 - Improved Accuracy
 - Reduced Risk of Errors
 - Increased Scalability
 - Improved Data Quality
- 
- Abstract geometric shapes in teal, yellow, and green in the bottom-left corner.

ML in ETL - Benefits

- **Increased Efficiency:** ML algorithms can automate and streamline the ETL process, significantly speeding up data integration and reducing time, resources required and manual effort. They enable real-time processing and quick adaptation to changing data patterns, leading to more efficient workflows.
- **Improved Accuracy:** by leveraging predictive models and pattern recognition, ML can increase the precision of data extraction and transformation.

ML in ETL - Benefits

- **Reduced Risk of Errors:** ML reduces the likelihood of human error in ETL processes. Automated checks and balances ensure data consistency and accuracy, leading to a more reliable data pipeline.
- **Increased Scalability:** ML algorithms are designed to handle large and complex datasets with ease. They can scale up or down based on the data volume, ensuring the ETL process remains robust and responsive.

ML in ETL - Benefits

- **Improved Data Quality:** ML can enhance data quality by identifying and correcting errors during the ETL process. It ensures that the data loaded into the target system is clean, well-structured, and ready for analysis, making the data more valuable for analysis and reporting.

References

- What is ETL (Extract, Transform, Load)?
 - ibm.com/topics/etl
- What is Data Extraction? Data Extraction Tools and Techniques
 - airbyte.com/data-engineering-resources/data-extraction
- What is ETL?
 - www.qlik.com/us/etl



Thank you

Fernando Prass
fprass@gmail.com