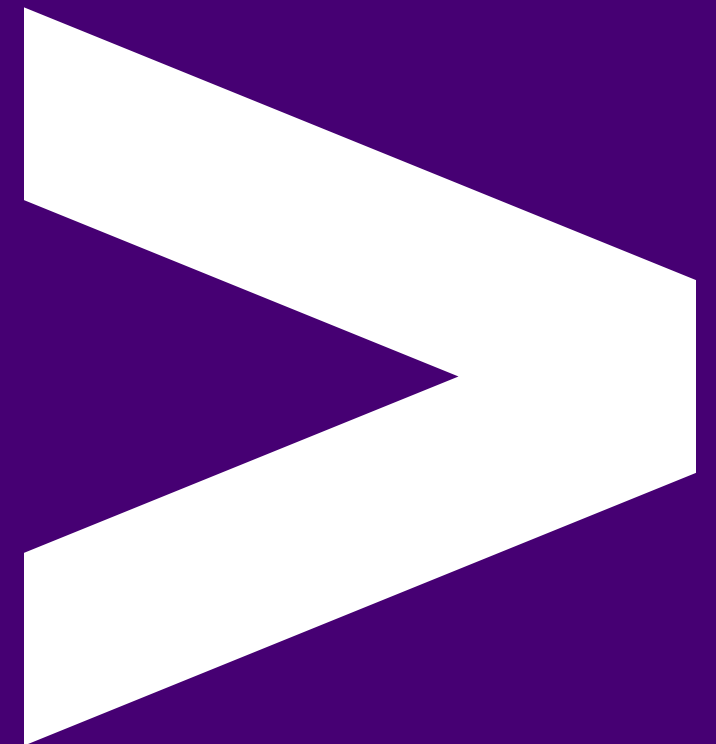


ETL



# Overview

- High level definition of ETL
- What problems can ETL solve?
- What happens in each stage?
- ETL vs ELT

# Learning Objectives

- Understand what happens in the extract, load and transform phases
- Implement an ETL pipeline using Python and SQL
- Understand when you might want to use ETL or ELT

# What is ETL?

- Extract, Transform and Load
- A way to move data from multiple sources and save it in a single target location
- Extract: Data is pulled from the relevant sources
- Transform: The data is changed to make it suitable for the target system
- Load: The transformed data is then saved into the target system

# Problems with data

Often keeping track of data across many systems is complex when...

- Events occur at different times
- Data formats vary
- There's too much data to process when extracting
- Systems are separated logically or physically
- Unpredictable extra load affects the source system

# Consequences

This can lead to unreliable systems and therefore unreliable data.

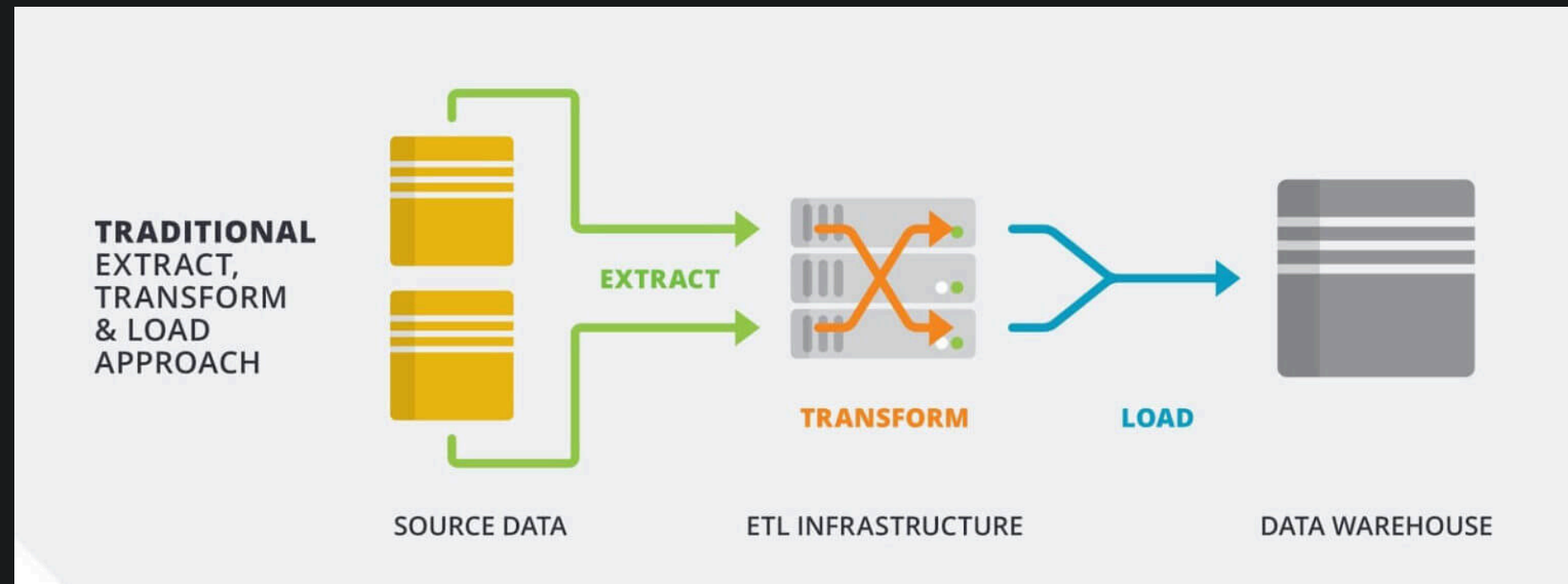
How many possible consequences of unreliable data could you think of as a group?

# Consequences

Some possible consequences of unreliable data:

- Financial institutions: incorrect balances, lost transactions
- Scientific research: false conclusions
- Retail: lost orders and deliveries
- Loss of reputation

# The ETL Stages





## Emoji Check:

Do you feel you understand the basics of ETL Pipelines? Say so if not!

1. 😓 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

# Extract: Where does the data come from?

Data is sourced from one or more systems, for example:

- Server logs
- Third parties
- Another database

# Extract: What format will the data be in?

This data will come in a variety of formats:

- JSON
- XML
- CSV
- Parquet (a column-oriented data storage format)
- Databases
- Other formats such as log files

## Extract: When will the data be extracted?

This usually happens when...

- A timed event occurs i.e. daily, hourly etc
- A database trigger event occurs
- A manual process is run

## Extract: How do you extract the data?

You can extract the data using different transfer methods, for example:

- Secure File Transfer Protocol (SFTP) is a network protocol that provides file access, transfer and management.
- Network Shares are when a computer resource is made available from one host to other hosts on a network
- Object Stores like Amazon S3

## Extract: How do you know the data is formatted correctly?

During extraction it is important to validate that this data is acceptable for passing to the transformation stage.

We can do this by matching predictable patterns, schemas or by running hash functions against the data.

If the data is invalid then appropriate alerting and/or metrics should be produced to inform customers and downstream systems.

# Transform

Rules or functions are applied to the extracted data to perform any of the following

- Normalisation
- Cleaning the data to a specific format or encoding
- Selecting specific columns/fields
- Performing calculations on fields e.g. 1000ms to 1s

## More transformations...

- Sorting
- Deduplicating data
- Grouping or Aggregating
- Joining data together with other datasets

Are there any other useful transformations you can think of?



# Load

Load the data into the target storage solution, often another relational database system.

Often existing data is overwritten or updated with cumulative information on a daily, weekly, or monthly basis.

Complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

## Emoji Check:

Do you feel you understand the purpose behind the building blocks of of ETL Pipelines? Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

# ETL vs ELT

# What is ELT?

- ELT stands for Extract, Load, Transform
- The processes are the same, but their order has changed

# Advantages of ELT

- More flexibility in querying source data
- Easier to understand relationship between raw and transformed data
- Potential cost savings

# Advantages of ETL

- Data protection - removing sensitive data before loading
- Can save on storage costs - removing large unused files before loading

## Emoji Check:

Do you feel you understand the differences between ETL & ELT Pipelines?  
Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

# Examples



# ETL Example

A real estate property company allows users to search for houses.

Each house can be accessed via their website using the path `/property-12345`.

Every time a page is accessed a record is stored in the `property_view` table against the `property_id`.

# ETL Example

That table looks something like this...

property_id	timestamp	browser	ip_address
12345	1580894343687	Chrome	182.22.109.13

# ETL Example

Once per day the property page view counts for the previous day are extracted from the main application database, and inserted into a staging table in the data warehouse.

The query looks something like this...

```
TRUNCATE TABLE warehouse_db.property_view_stage;

INSERT INTO warehouse_db.property_view_stage
SELECT * FROM main_db.property_view
WHERE timestamp >= 1580860800000
AND timestamp <= 1580947200000;
```

# ETL Example

The data is transformed using a **GROUP BY** aggregation and inserted into another staging table.

The query looks something like this...

```
TRUNCATE TABLE warehouse_db.page_view_daily_aggregation_stage;

INSERT INTO warehouse_db.page_view_daily_aggregation_stage
SELECT DATE(FROM_UNIXTIME(timestamp/1000)) as property_view_da
property_id, COUNT(1) as property_view_count
FROM warehouse_db.property_view_stage
GROUP BY property_view_date, property_id;
```

# ETL Example

That table looks something like this...

property_view_date	property_id	property_view_count
2020-02-05	12345	32
2020-02-05	67890	21

## ETL Example

This data is then loaded to the final location where it can be used in reports by estate agents to their customers.

The query looks something like this...

```
INSERT INTO warehouse_db.page_view_daily_aggregation  
SELECT * FROM warehouse_db.page_view_daily_aggregation_stage
```

# ETL Example

The final aggregated data looking something like this...

property_view_date	property_id	property_view_count
2020-02-05	12345	32
2020-02-06	12345	15
2020-02-05	67890	32
2020-02-06	67890	21

## ETL Example

This data could be further aggregated or used to produce insightful customer reports...

Can you think of any other examples of how this data could provide insight?



## Emoji Check:

Do you feel you understand the ETL Pipeline example? Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

Quiz Time! 🧐

Extract is the process of...

1. Pulling in data from one or more source systems, with no manipulation of the data.
2. Moving the source data directly to an end target, such as a data warehouse.
3. Pulling in data from one or more source systems, cleaning it in the process.
4. Aggregating multiple sources into one easily digestible set of data.

Answer: 1

Transform is the process of...

1. Moving the source data directly to an end target, such as a data warehouse.
2. Removing columns of data we don't need, as well as aggregating data where needed.
3. Manipulating the extracted data to conform to business rules.
4. Applying a series of rules to the extracted data in order to prepare it for the end target.

Answer: 4

Load is the process of...

1. Moving the transformed data into data analysis software.
2. Reviewing data in the database.
3. Moving the transformed data into the end target.
4. Overwriting existing data in a staging table.

Answer: 3

# Exercise Prep


Distribute exercise file [./exercises/etl-exercise.md](#) and the [./handouts/](#) folder.

# Exercise Prep - Setup - 5 mins

Do the "Prep" step of setting up a postgres container from the [./exercises/etl-exercise.md](#) file.

The files you need are in the [./handouts/](#) folder.

# Discussion



Does everyone have a running database container now?



# Exercise - Overview

The full exercise is in three main parts - write Extract code (load a CSV), write Transform code (of the CSV data), and write Load code (to save the data in Postgres). This is designed to bring together the building-blocks of several previous sessions.

There is also a fourth part to do some analysis of the acquired data.

# Exercise Task 1 - Extract - 10 mins

Do the "Task 1 - Extract" step to load the CSV file from the [./exercises/etl-exercise.md](#) file.

The files you need are in the [./handouts/](#) folder.

# Discussion

Does everyone have a the extract code running?

## Exercise Task 2 - Transform - 20 mins

Do the "Task 2 - Transform" step from the [./exercises/etl-exercise.md](#) file, to manipulate your data.

The files you need are in the [./handouts/](#) folder.

# Discussion




Does everyone have a the transform code running?

## Exercise Task 3 - Load - 20 mins

Do the "Task 3 - Load" step from the [./exercises/etl-exercise.md](#) file, to insert the data into the database.

The files you need are in the [./handouts/](#) folder.

# Discussion



Does everyone have a the load code running?


## Exercise Task 4 - Analysis - 10 mins

Do the "Task 4 - Analysis" step from the [./exercises/etl-exercise.md](#) file, analyse your data.

The files you need are in the [./handouts/](#) folder.



# Discussion



Does everyone have a the load code running?

## Emoji Check:

How did you find exercises on ETL Pipelines?

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

# Terms and Definitions - recap

Extract: Extract the data from a source.

Transform: Carry out operations on the data eg. cleaning it, adding meta data.

Load: Storing the transformed data, usually in a database.

# Overview - recap

- High level definition of ETL
- What problems can ETL solve?
- What happens in each stage?
- ETL vs ELT

## Learning Objectives - recap

- Understand what happens in the extract, load and transform phases
- Implement an ETL pipeline using Python and SQL
- Understand when you might want to use ETL or ELT

## Further Reading

- [ETL best practices](#)
- [Anatomy of ETL](#)
- [ETL vs ELT](#)
- [Another good comparison of ETL and ELT](#)

## Emoji Check:

On a high level, do you think you understand the main concepts of this session? Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively