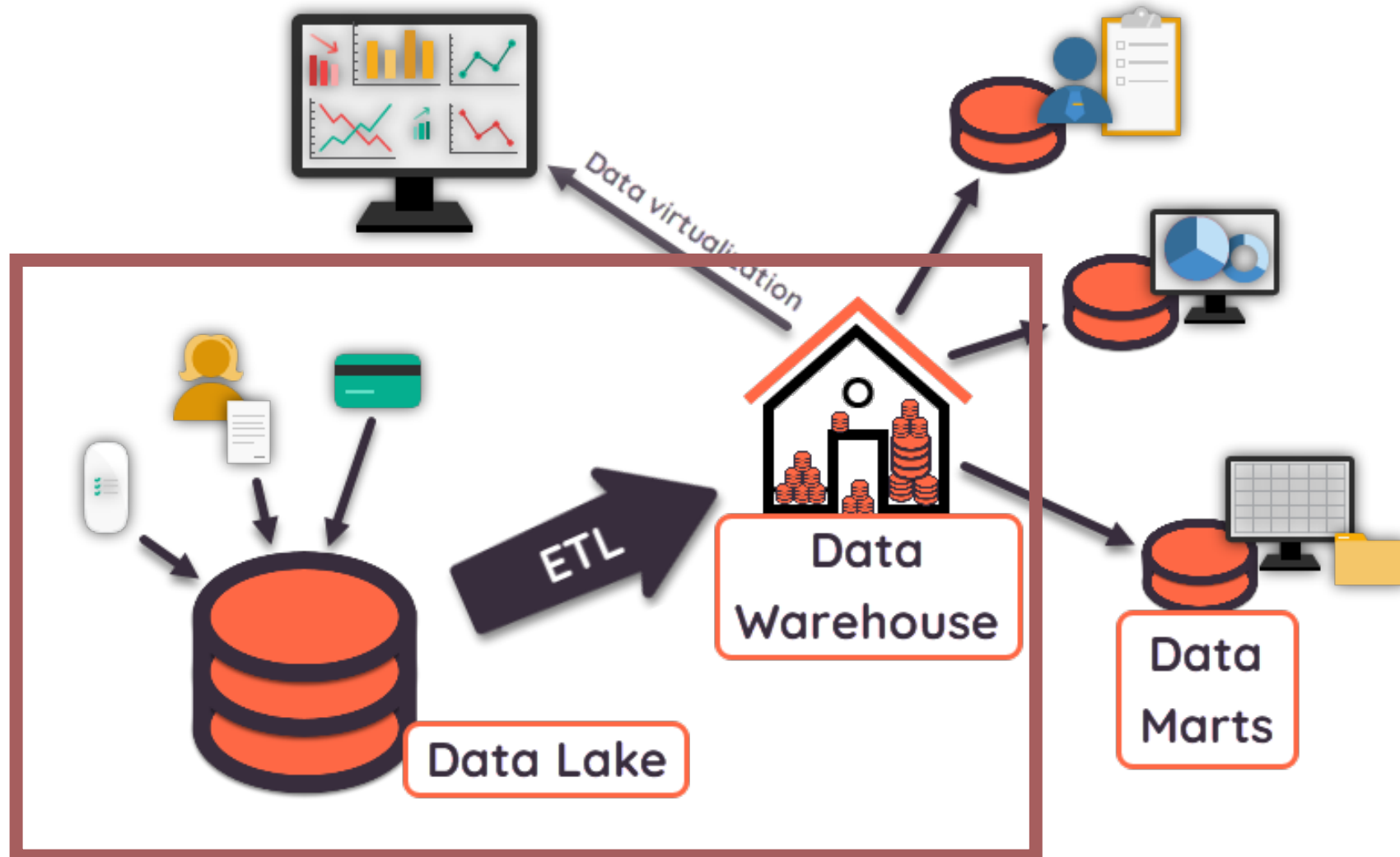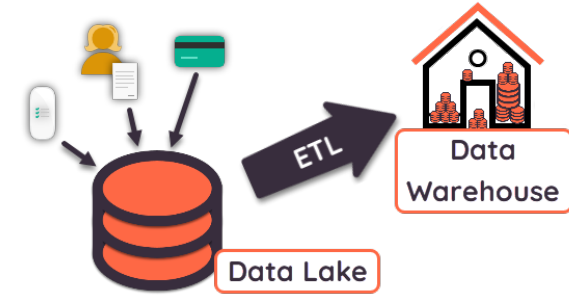# DataVault

Data Engineering Bootcamp

Data management and archiving in the research environment

By Niklas Büchner, Christian Singer, Ahmad Al-Taie, Mike Sickmüller

WHAT IS ETL ?

WHAT IS ETL ?

Extract

Transform

Load
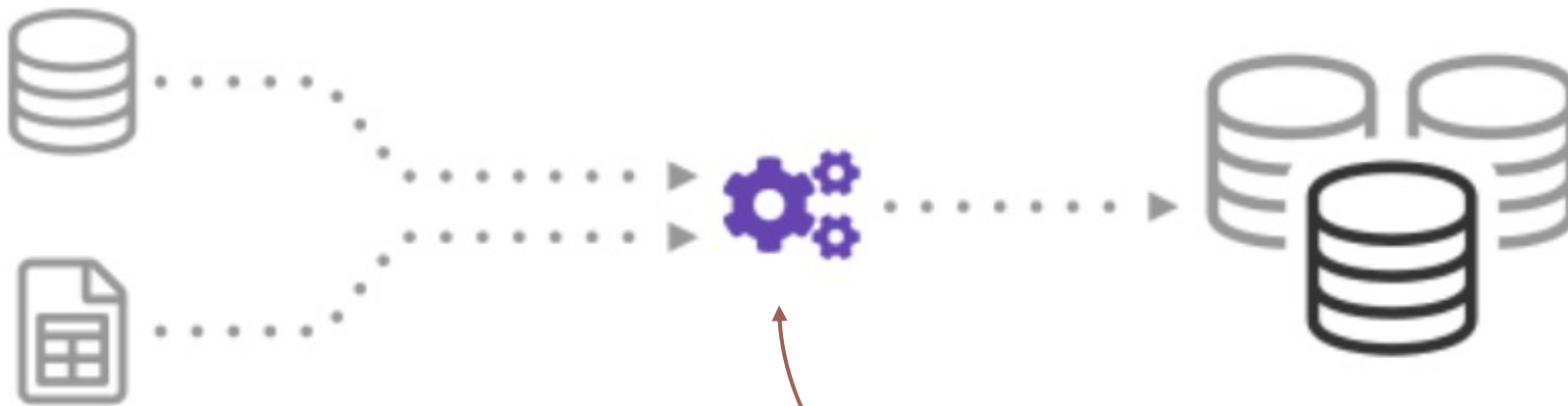
Data is lost and permanently altered

WHAT IS A DATAVAULT ?
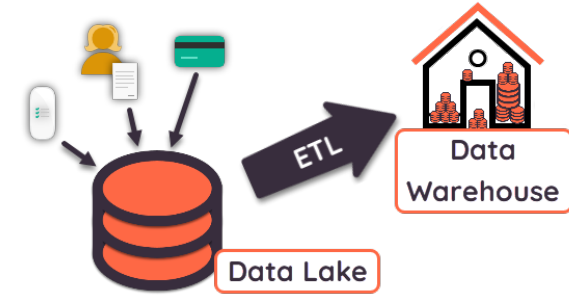
Extract

Load

Transform

Load

Image: https://panoply.io/uploads/etl-1.png

PROS & CONS OF A DATAVAULT

Pros:

- Long-term storage of data

- Tracking data changes

- Fast import of data

- Changes in data schemas do not necessarily required downstream ETL and analysis processes to be adjusted

Cons:

- Not easy to query

- Large overhead

MODEL

```
crops:
+------+----------+-------------------+
| id   | crop     | water_consumption |
+------+----------+-------------------+
| C-1  | tomato   | 10                |
+------+----------+-------------------+
| C-2  | cucumber | 15                |
+------+----------+-------------------+

fields:
+------+-------+----------+
| id   | field | crop_id  |
+------+-------+----------+
| F-5  | small | 2        |
+------+-------+----------+
| F-6  | big   | 1        |
+------+-------+----------+
```

SATELITE

crops satelite:

```
+--------------+----------+-------------------+------------+----------+---------------+-----------+
| crop_hash_key | crop     | water_consumption | load_date  | end_date | record_source | hash_diff |
+--------------+----------+-------------------+------------+----------+---------------+-----------+
| b519e        | tomate   | 10                | 2022-05-01 | NULL     | ERP           | 8a3f0     |
+--------------+----------+-------------------+------------+----------+---------------+-----------+
| 5f763        | cucumber | 15                | 2022-05-01 | NULL     | ERP           | c345a     |
+--------------+----------+-------------------+------------+----------+---------------+-----------+
```

crops:

```
+------+----------+-------------------+
| id   | crop     | water_consumption |
+------+----------+-------------------+
...
+------+----------+-------------------+
```

ALTER DATA

crops satelite:

```
+---------------+---------+-------------------+------------+------------+---------------+-----------+
| crop_hash_key | crop    | water_consumption | load_date  | end_date   | record_source | hash_diff |
+---------------+---------+-------------------+------------+------------+---------------+-----------+
| b519e         | tomate  | 10                | 2022-05-01 | 2022-05-02 | ERP           | 8a3f0     |
+---------------+---------+-------------------+------------+------------+---------------+-----------+
...
+---------------+---------+-------------------+------------+------------+---------------+-----------+
| b519e         | tomate  | 12                | 2022-05-02 | NULL       | ERP           | 8a3f0     |
+---------------+---------+-------------------+------------+------------+---------------+-----------+
```

- ✅ Tracking data changes
- ✅ Only import differences

ADD COLUMN

```
crops height satelite:
+-----------------+-------------+------------+----------+---------------+-----------+
| height_hash_key | crop_height | load_date  | end_date | record_source | hash_diff |
+-----------------+-------------+------------+----------+---------------+-----------+
| 74f10           | 30          | 2022-05-03 | NULL     | ERP           | a8c94     |
+-----------------+-------------+------------+----------+---------------+-----------+
| d9570           | 15          | 2022-05-03 | NULL     | ERP           | ef3a9     |
+-----------------+-------------+------------+----------+---------------+-----------+

crops - crops height link:
+-------------------------+---------------+----------------------+------------+---------------+
| crop_crop_height_hash_key | crop_hash_key | crop_height_hash_key | load_date  | record_source |
+-------------------------+---------------+----------------------+------------+---------------+
| 05610                   | b519e         | 74f10                | 2022-05-03 | ERP           |
+-------------------------+---------------+----------------------+------------+---------------+
| 2c61c                   | 5f763         | d9570                | 2022-05-03 | ERP           |
+-------------------------+---------------+----------------------+------------+---------------+
```

- ✅ Tracking schema changes
- ✅ Downstream ETL processes do not need to be adjusted if they don't need the new information

9

REMOVE A COLUMN

- All Columns need to be nullable
- Therefore no further changes are needed.

THANK YOU FOR YOUR ATTENTION