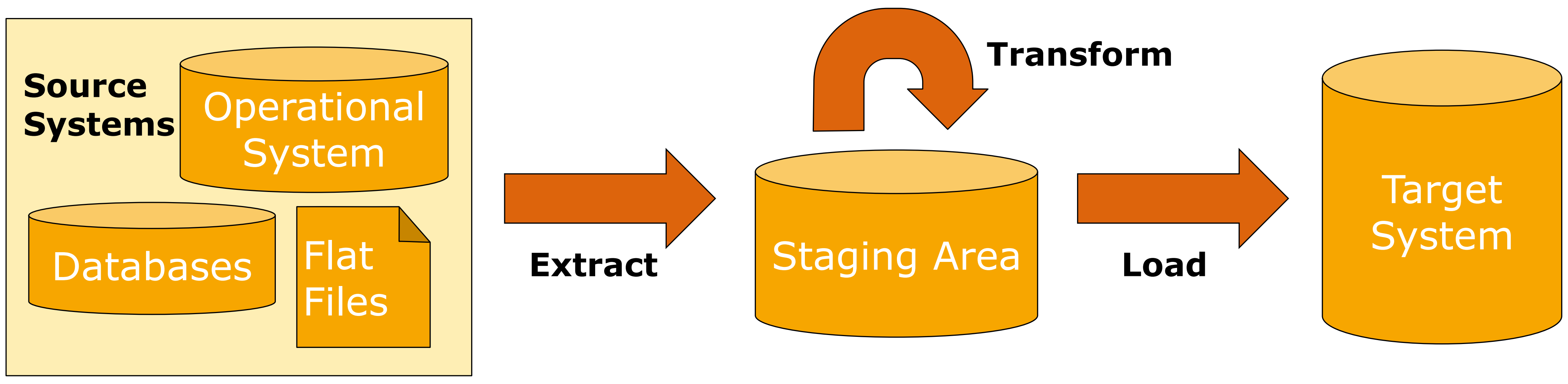
The ETL Process
Extract – Transform - Load

ETL is an approach for data integration. It describes the process of extracting the data from remote sources, transforming it into the given formats and styles, and loading it into the target system.

The process is often used for data warehouses and is the main part of systems that receive their data from different sources. As this data often has different formats and not all data sets are relevant, the data needs to be cleaned and reformatted to transform it into decision-relevant information.

The ETL Process



1. Extraction

The data from the source systems is first loaded into a staging area to prevent corrupted data from being loaded directly into the target system.

The following aspects need to be considered:

- Relevant data sets
- Connection and transfer types
- Update frequency

When using synchronous extraction, data is updated continuously, which may result in high utilization. Asynchronous extraction makes it possible to plan the extraction for a timeslot where sufficient resources are available.

2. Transformation

The raw data from the sources is not usable. It needs to be transformed into a consistent format for the target system. To achieve this, functions and customized operations are applied.

There are several issues that need to be addressed:

- Inconsistent data
- Duplicates
- Obsolete data
- Missing data

Several validations are applied, e.g.

- Filtering of certain columns
- Standardization of strings and time specification
- Conversion of units of measurement
- Merging data with lookups

3. Loading

Now the data is loaded into the target system. Meanwhile, the system is blocked to prevent incorrect evaluations. Changes should be logged for the case of load failures.

There are three types of loading:

- Initial loading: Population of all tables
- Incremental loading: Changes are applied periodically when needed
- Full refresh: One or more tables are reloaded with fresh data

Trend Towards Cloud

The amount and value of data will increase strongly. Highly performant, scalable IT infrastructures are difficult to implement in computer centers, which is why the trend is towards cloud-based ETL processes ("ETL-as-a-Service").

Possible Use Cases

Consumer industry

Analysis of data from social media for market trend studies

Medicine

Connection of patient records and laboratory results to determine the risk of disease

Aviation

Connection of data such as distance and kerosene consumption to determine profitable flight

https://www.guru99.com/etl-extract-load-process.html#9 (26.01.2020)
https://www.gambit.de/wiki/etl-prozess/ (26.01.2020)
https://www.sas.com/en_us/insights/data-management/what-is-etl.html#dmusers (26.01.2020)

Wanda Baltzer

Bachelor Student
Hasso Plattner Institute, Potsdam, Germany

E-Mail: wanda.baltzer@student.hpi.de

HPI Hasso Plattner Institut