Loading Data in a Data Warehouse

Ana Voicu

@ana_voicu

Review of the Previous Steps

Discuss with the business people

- Processes
- Analysis needs

Investigate data sources

Implement the dimensional model

- End result will be the physical design
- Fact and dimension tables are defined

Populate the data warehouse tables with source data

- Also called “loading” the data warehouse
- Implement the ETL system

The ETL system

Similar to a “black box” for the data warehouse users

Data is:

- Integrated from multiple sources
- Transformed into meaningful reports
- Helping people make decisions

The ETL should be properly designed and implemented

Overview

What does ETL mean?

- Extraction
- Transformation
- Load

Types of loads

- Full (initial) load
- Incremental load

Data lineage

- What is data lineage?
- Why is it important?
- How can you implement it?

Demos

Perform the needed steps to load a dimension table

Create and work with the auxiliary objects involved in a load

- Staging tables
- Lineage table
- Incremental loads table

Overview of an ETL System

What Is ETL?

A system that incorporates all operations performed on data

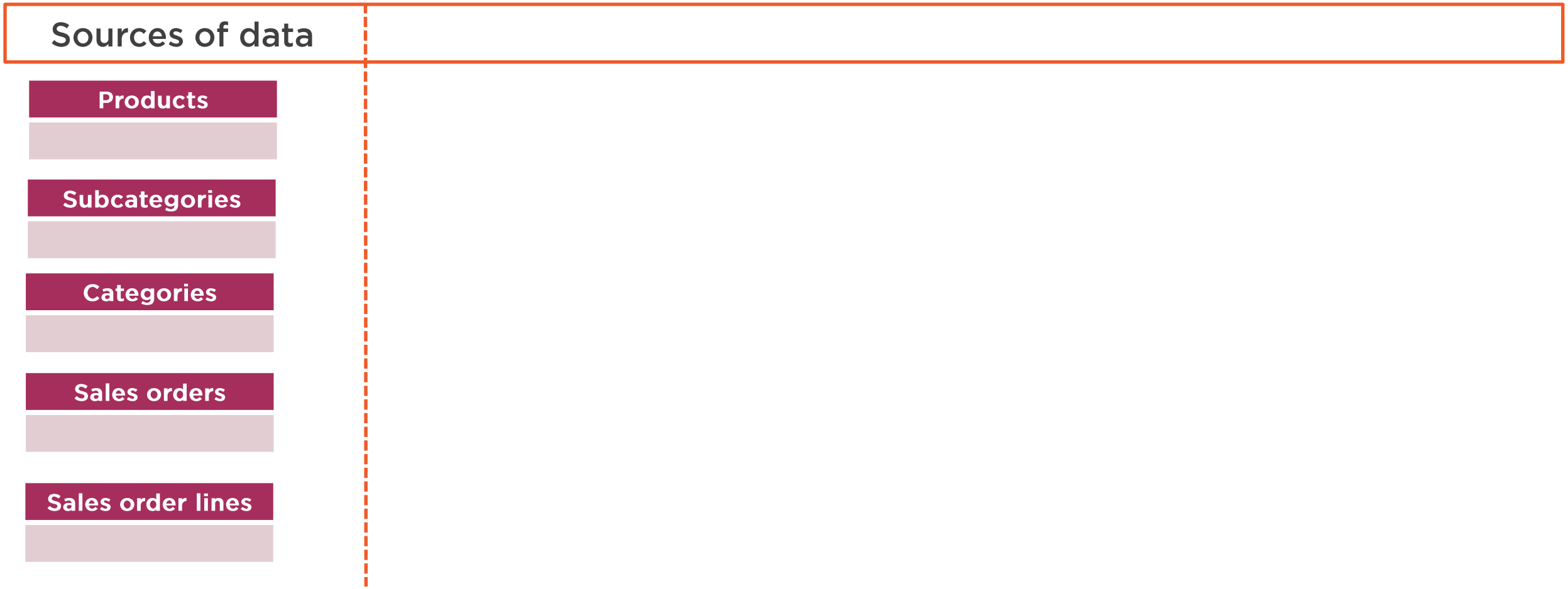
- Since it's selected from data sources
- Until it reaches the presentation area

Consists of three phases

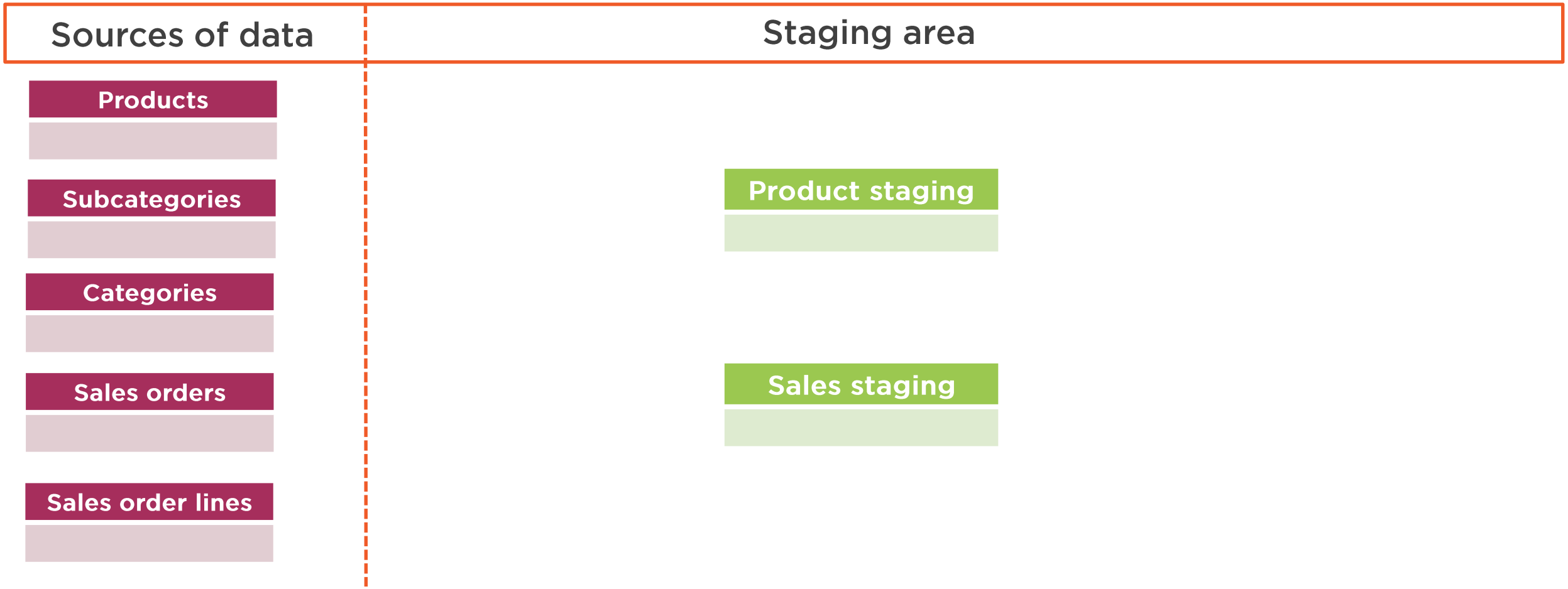
- Extraction (E)
- Transformation (T)
- Loading (L)

Overview of an ETL System

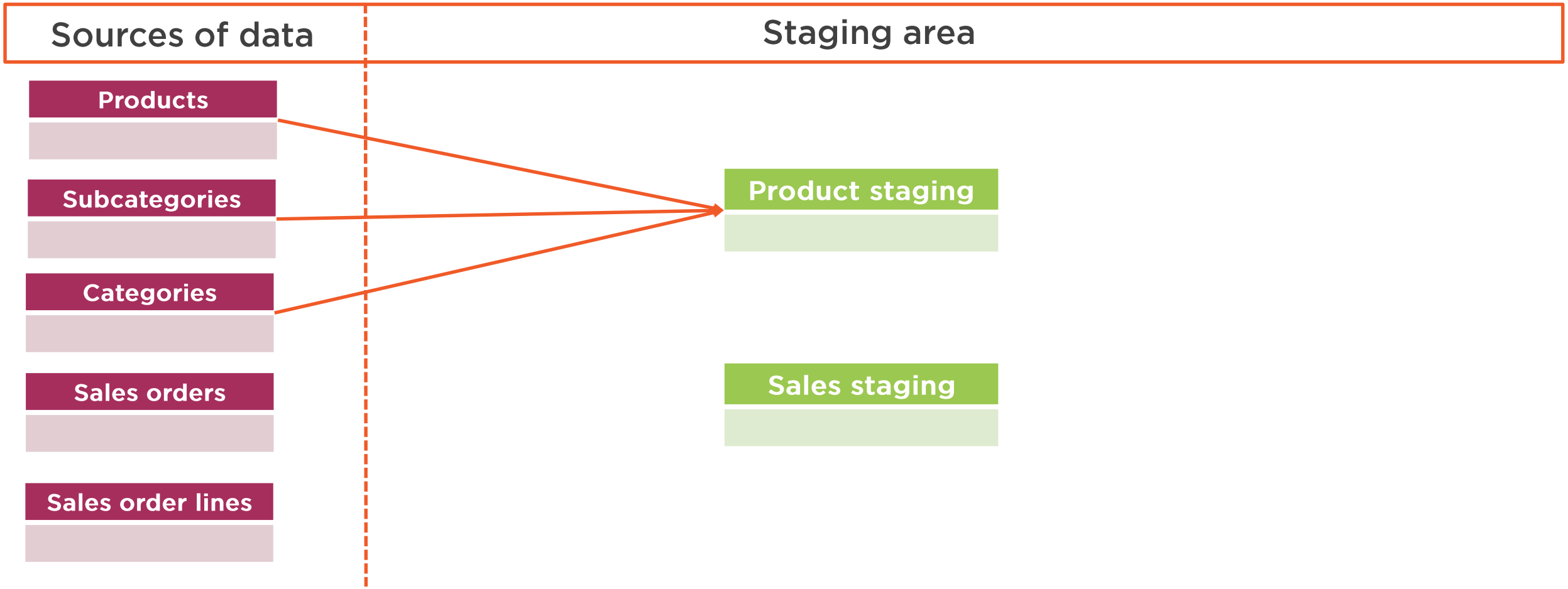
Overview of an ETL System



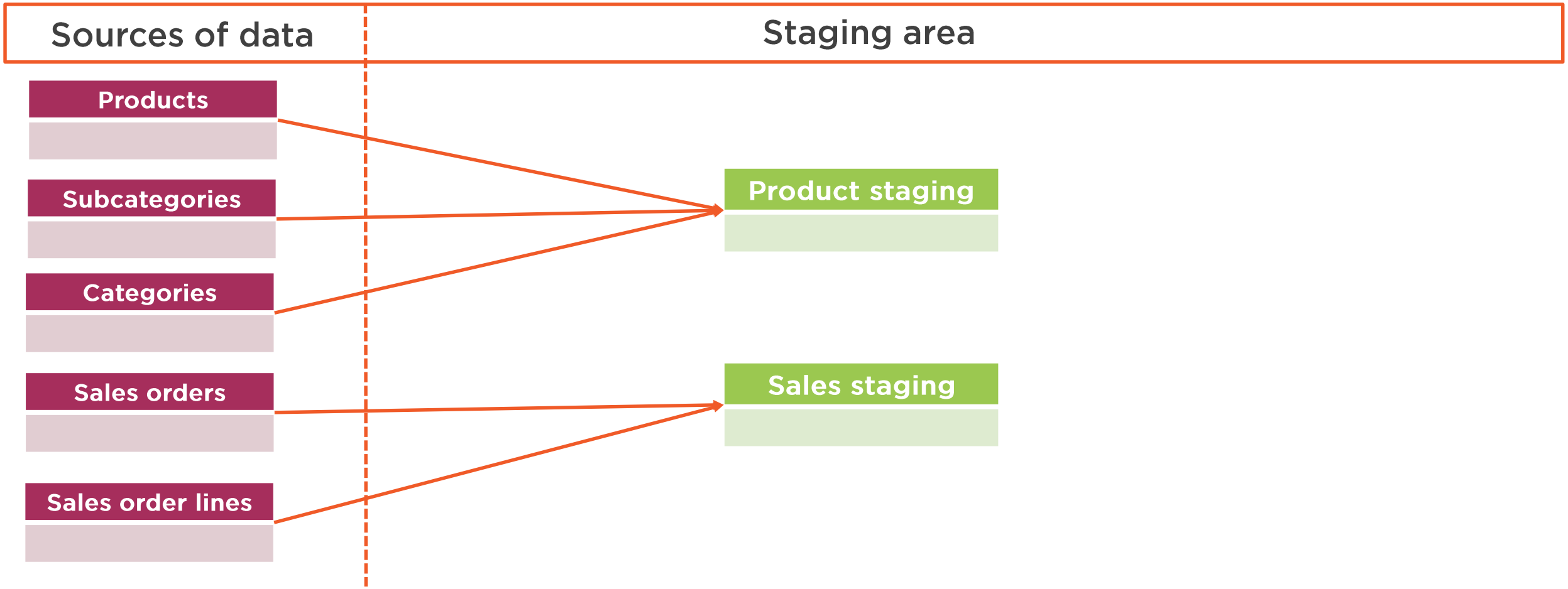
Overview of an ETL System



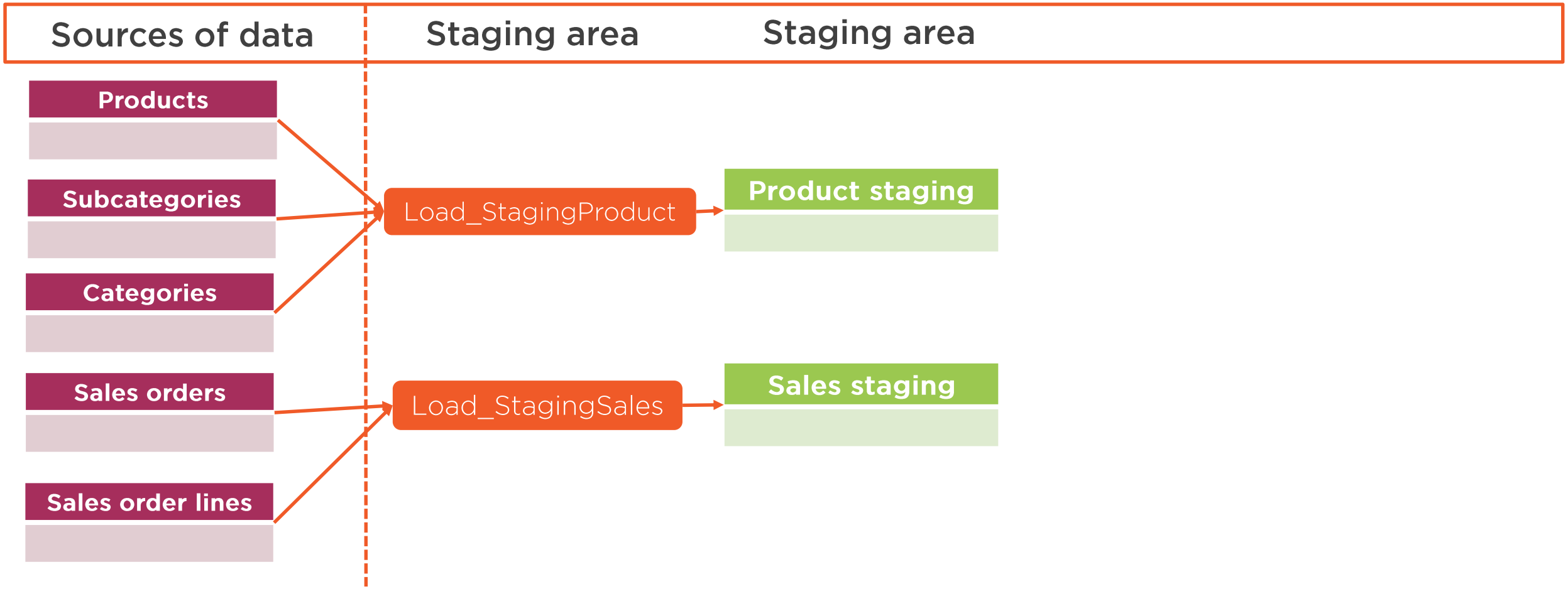
Overview of an ETL System



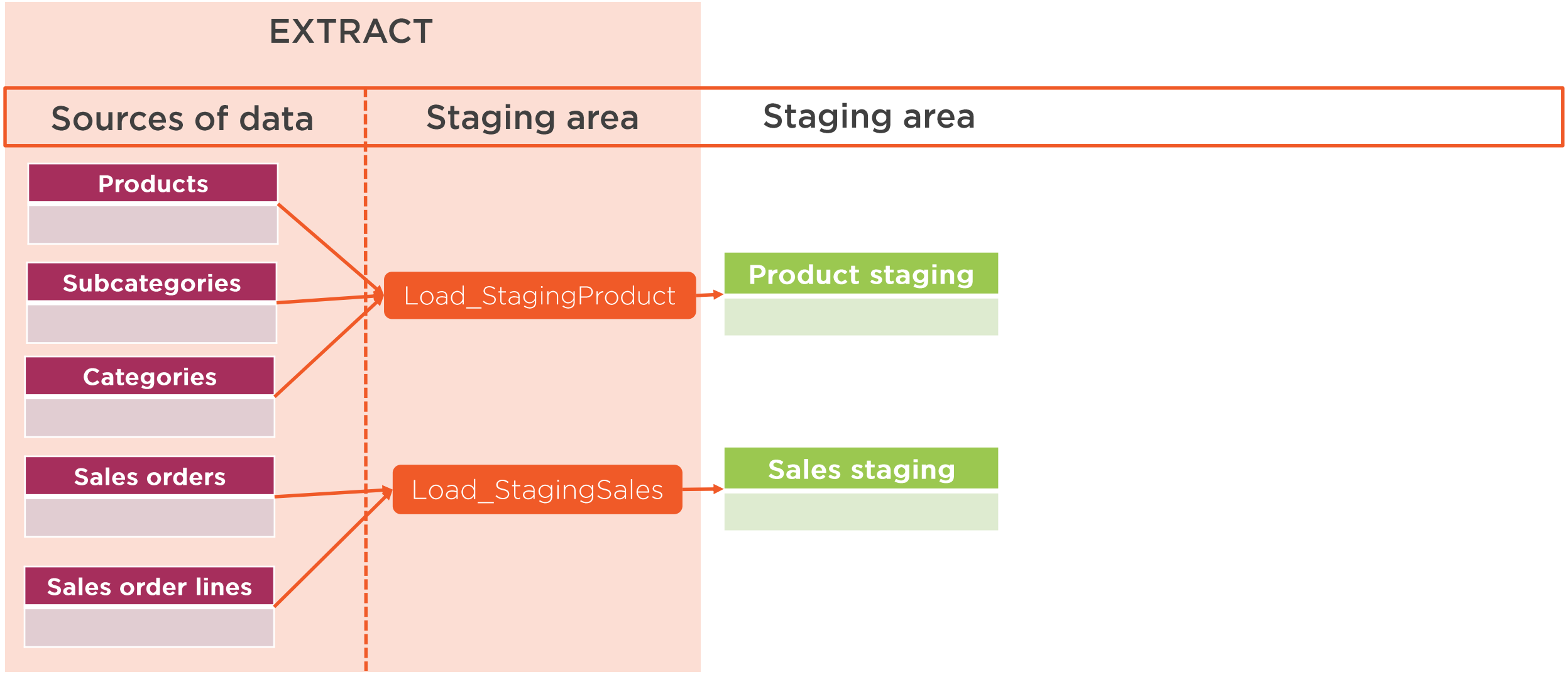
Overview of an ETL System



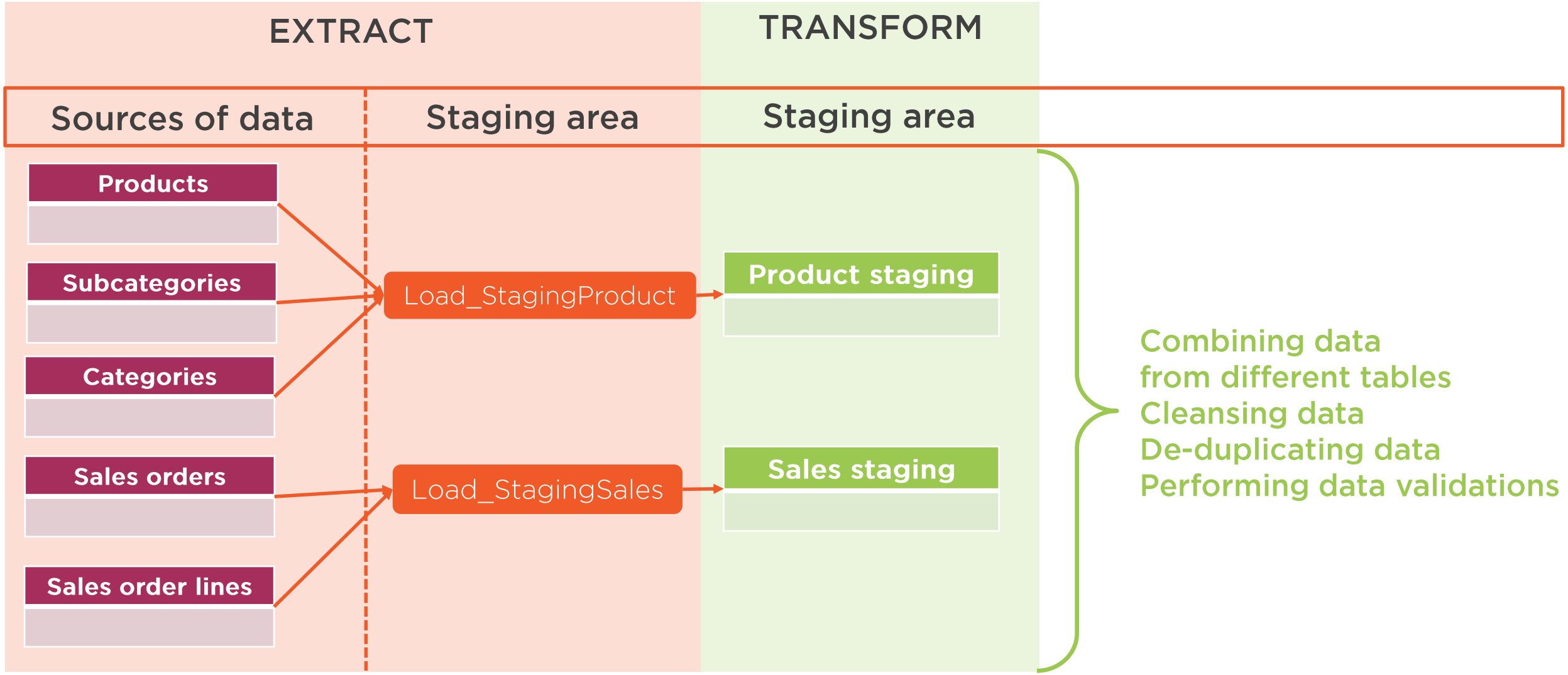
Overview of an ETL System



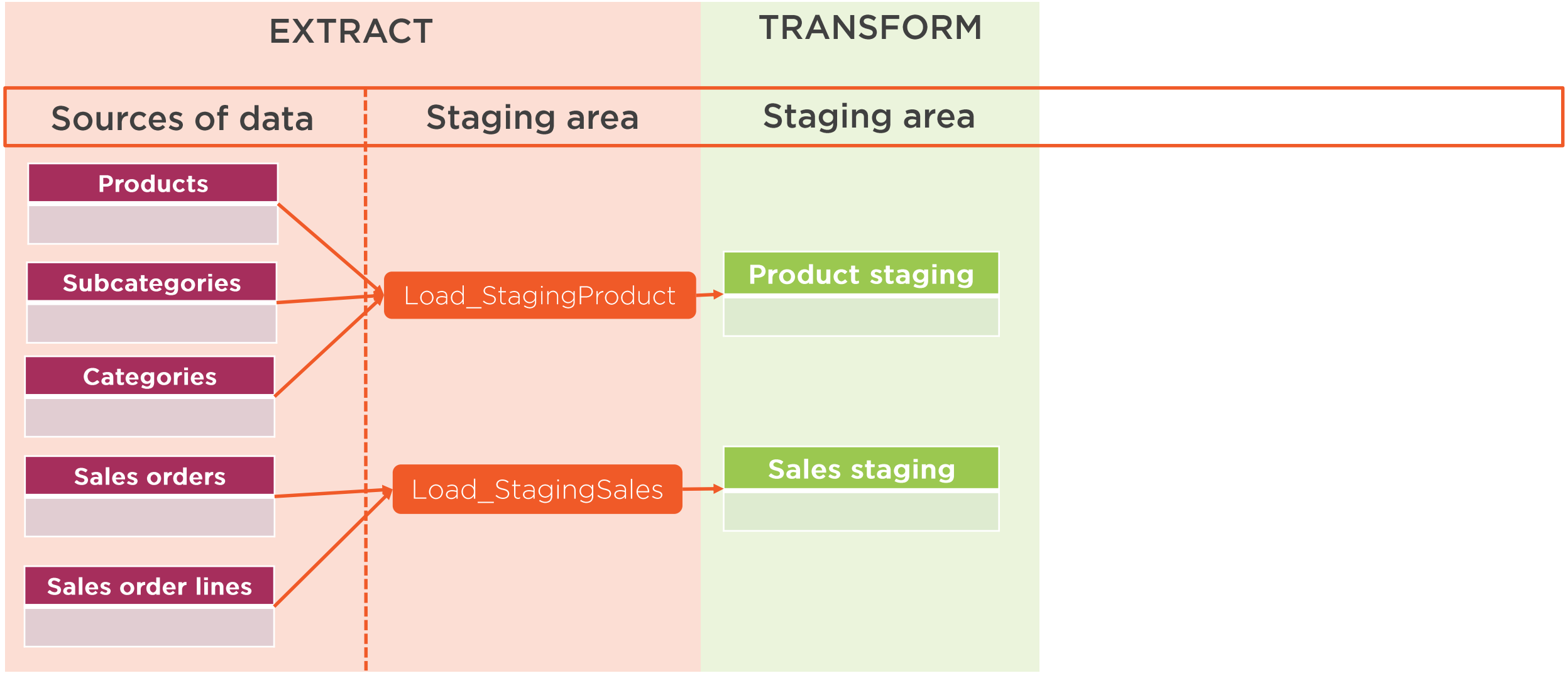
Overview of an ETL System



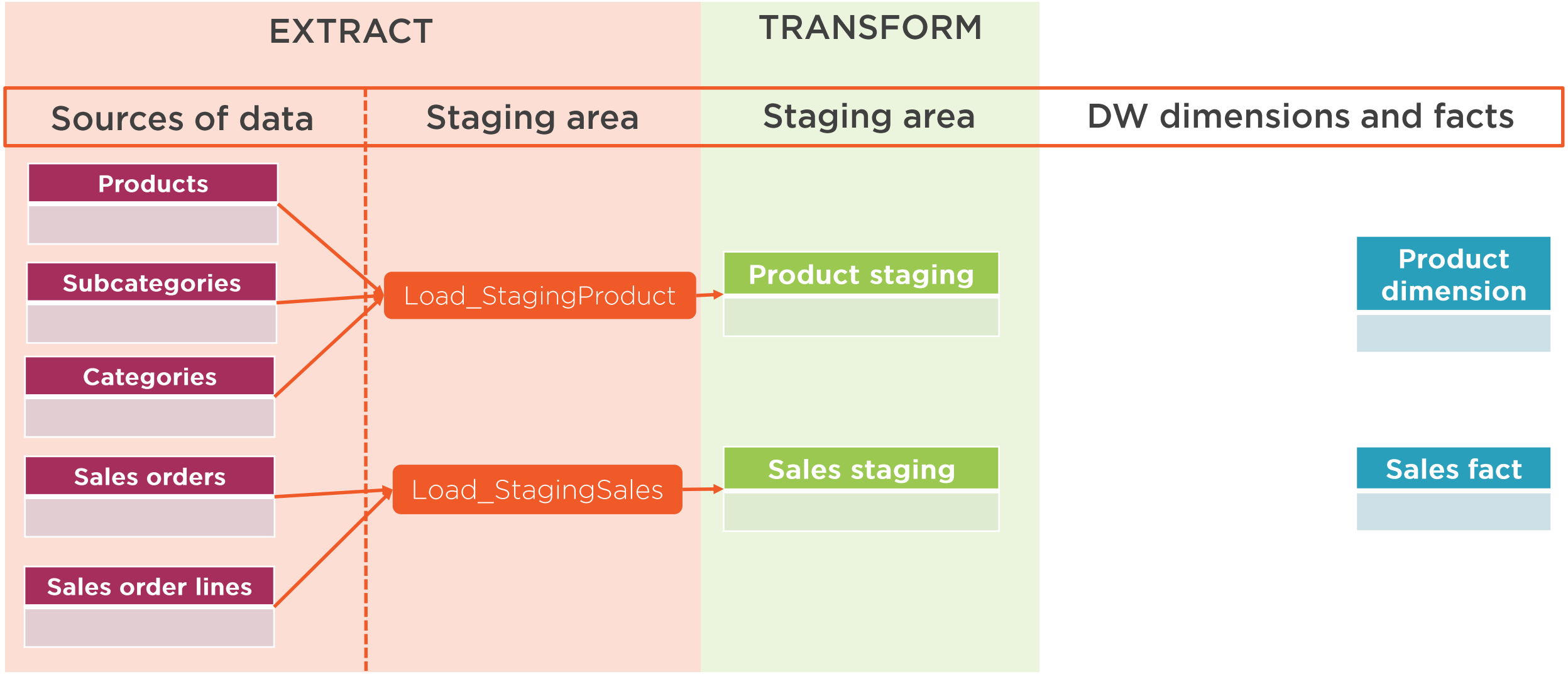
Overview of an ETL System



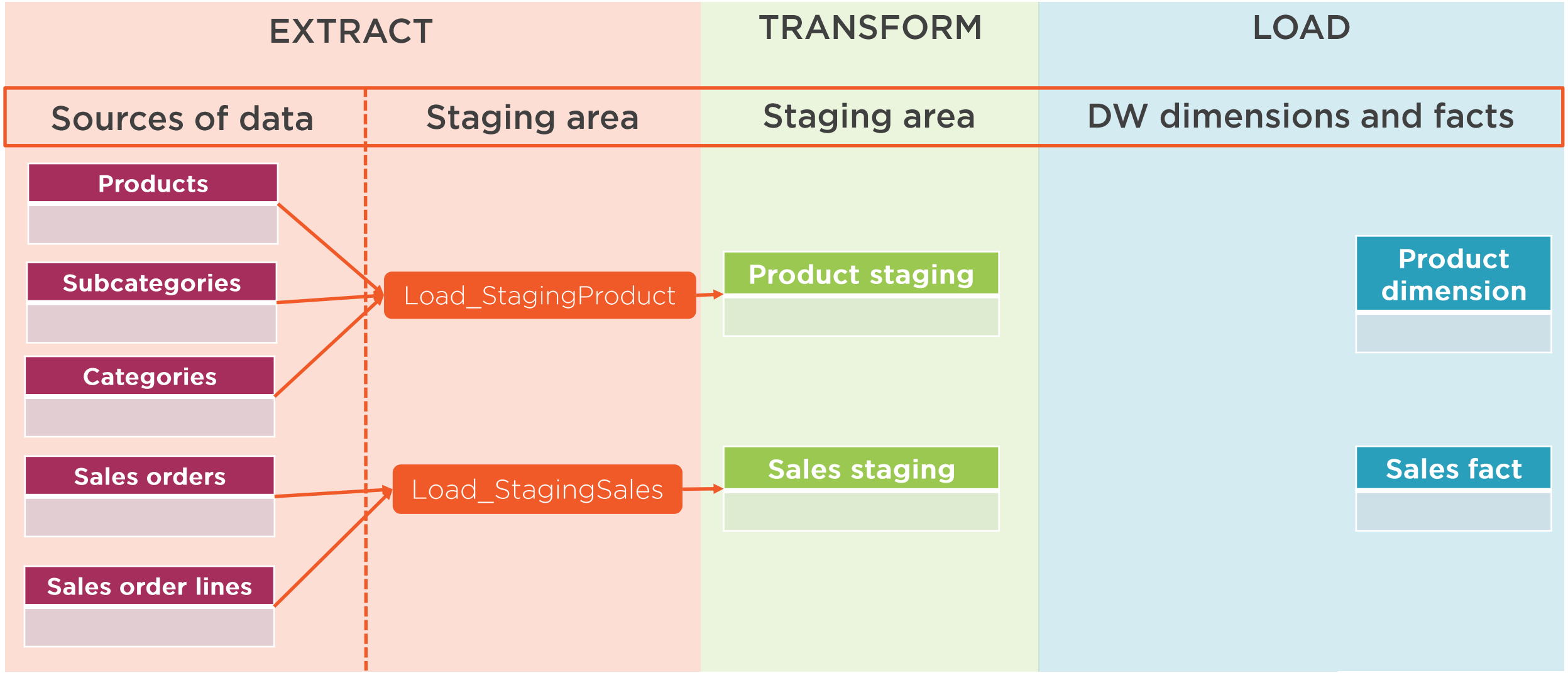
Overview of an ETL System



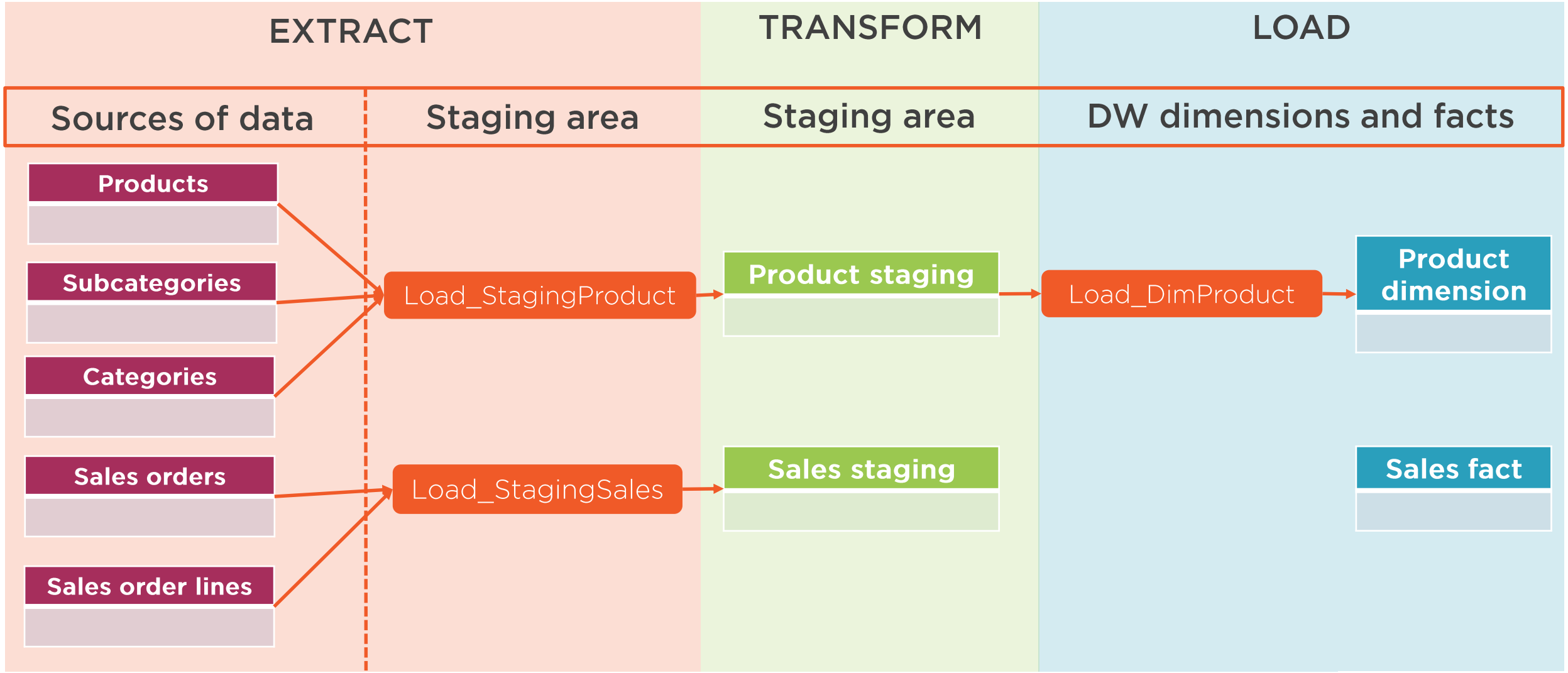
Overview of an ETL System



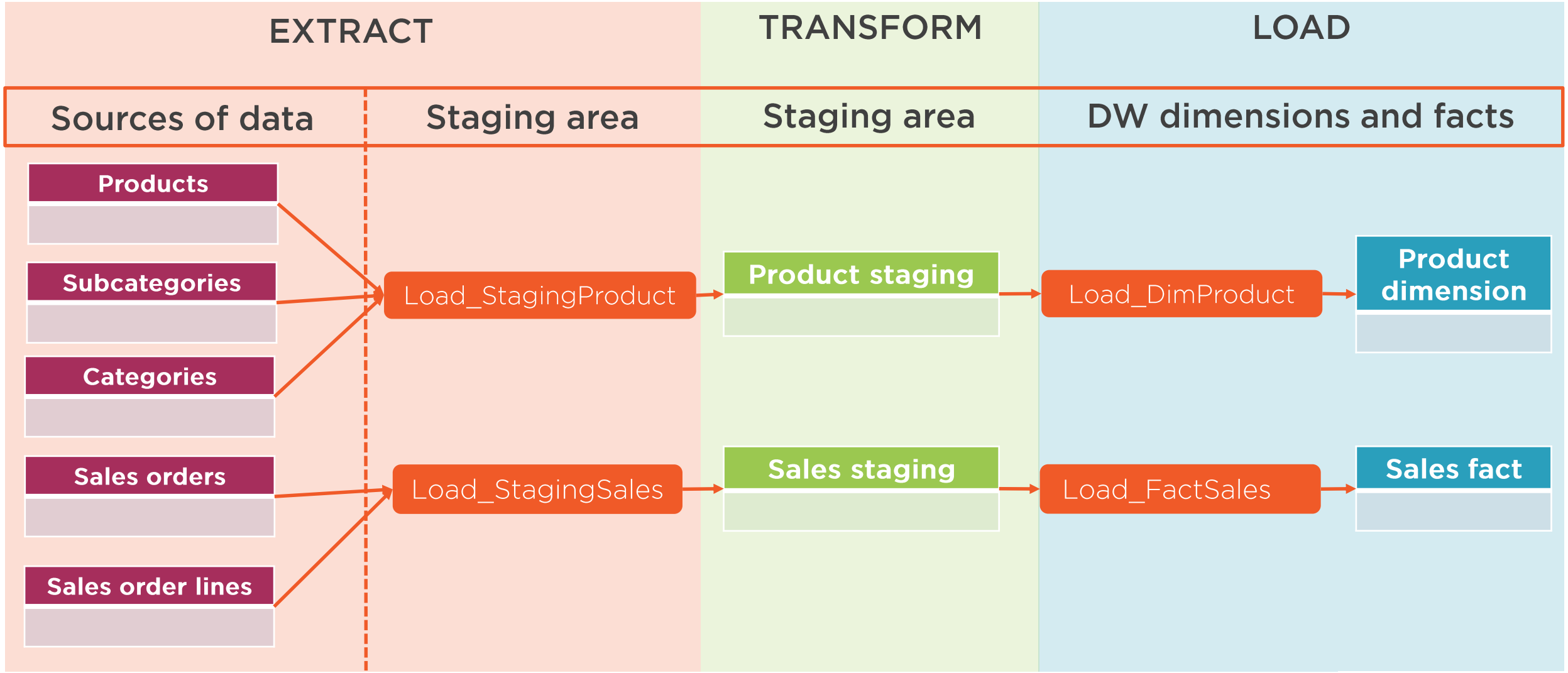
Overview of an ETL System



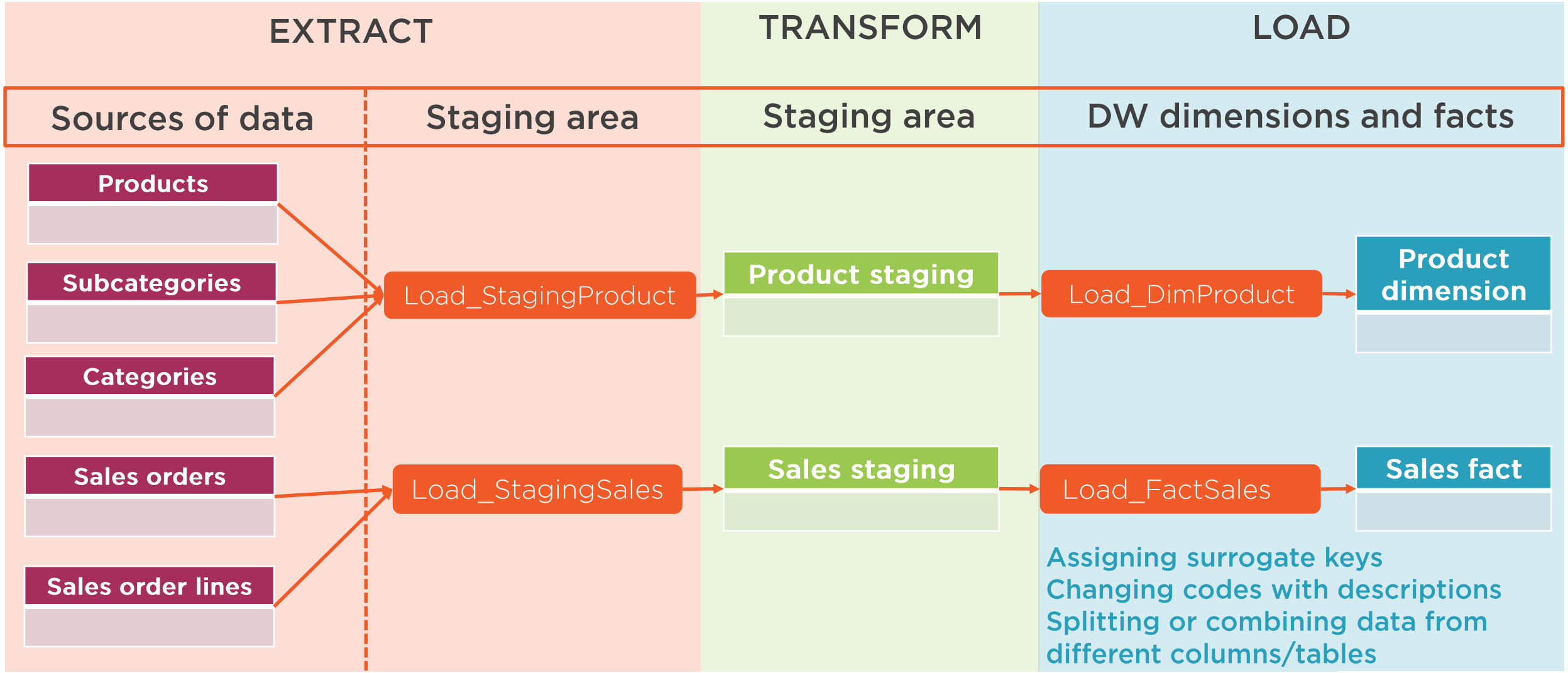
Overview of an ETL System



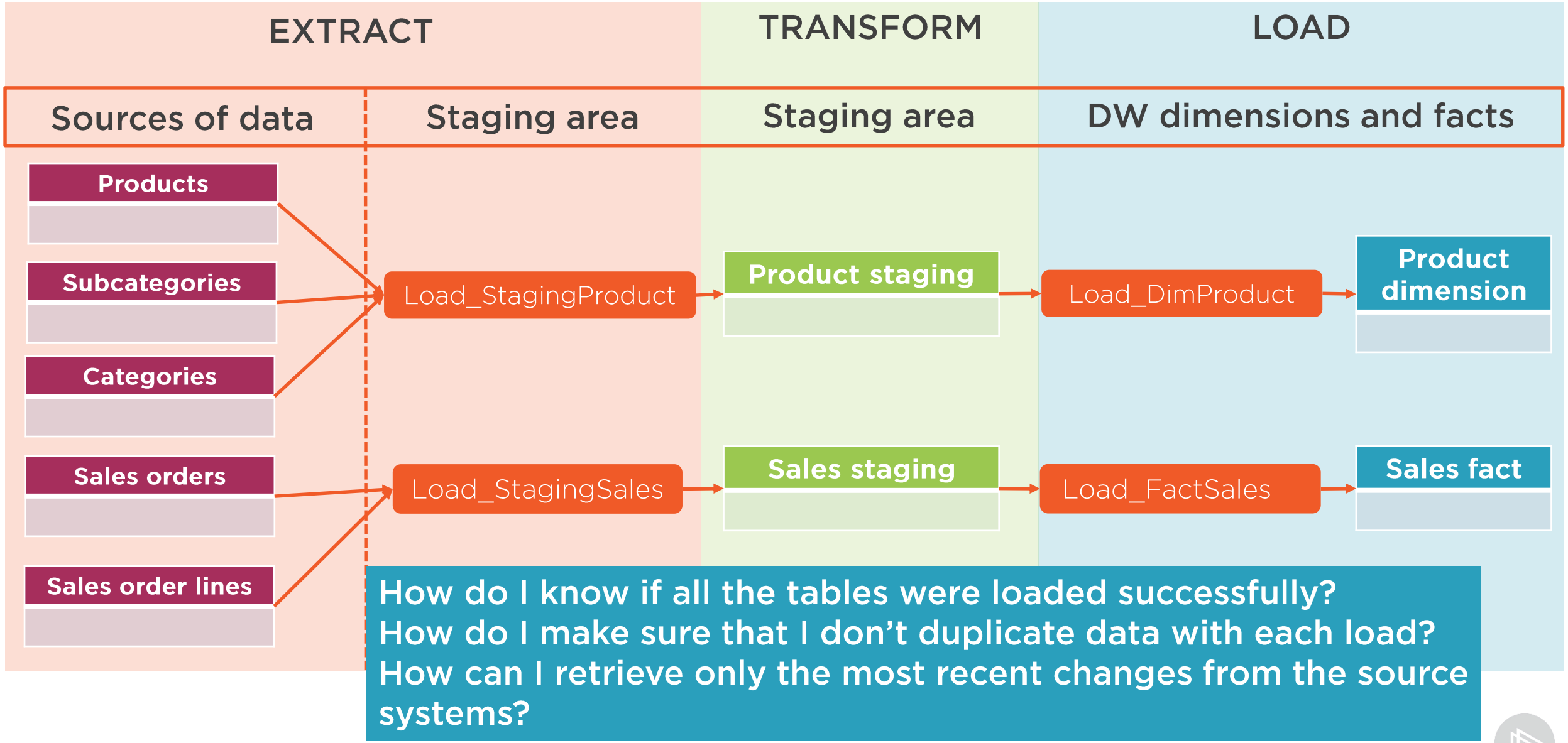
Overview of an ETL System



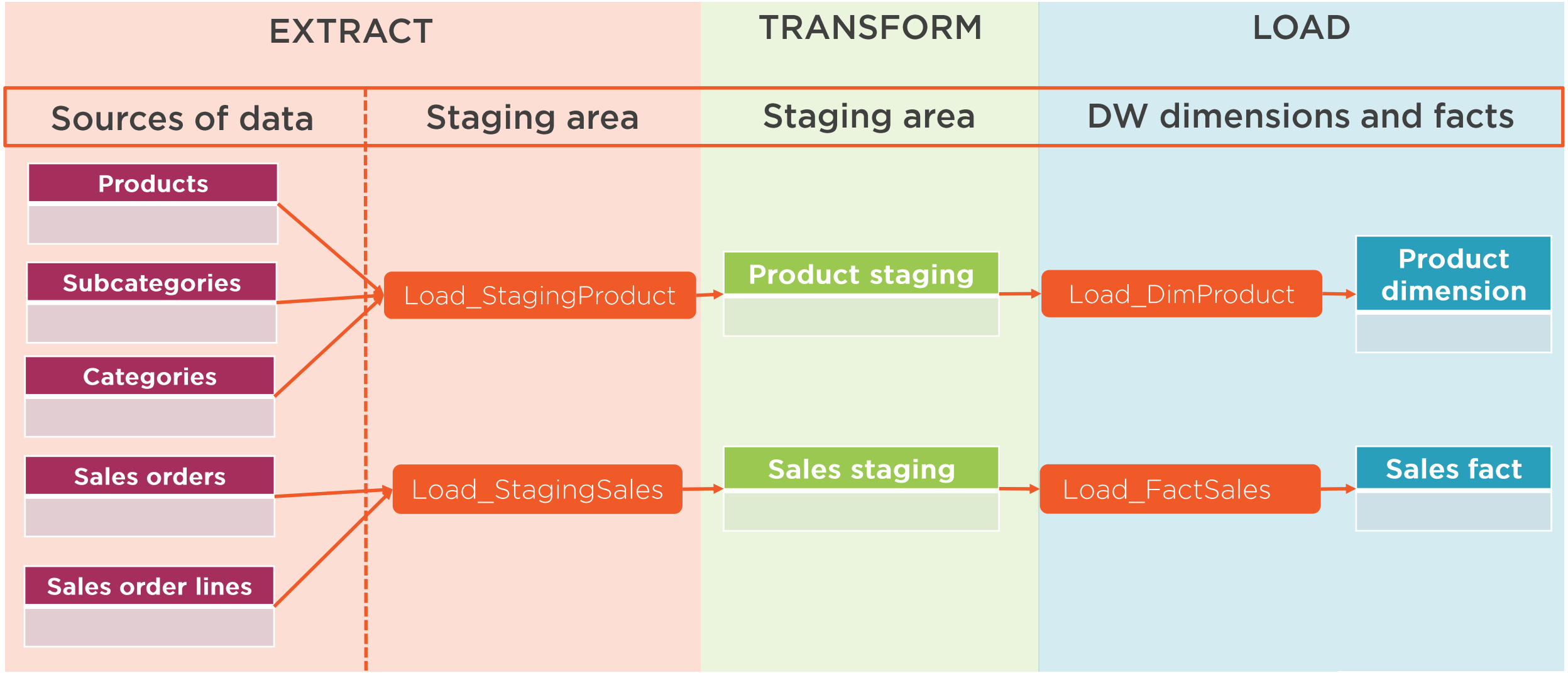
Overview of an ETL System



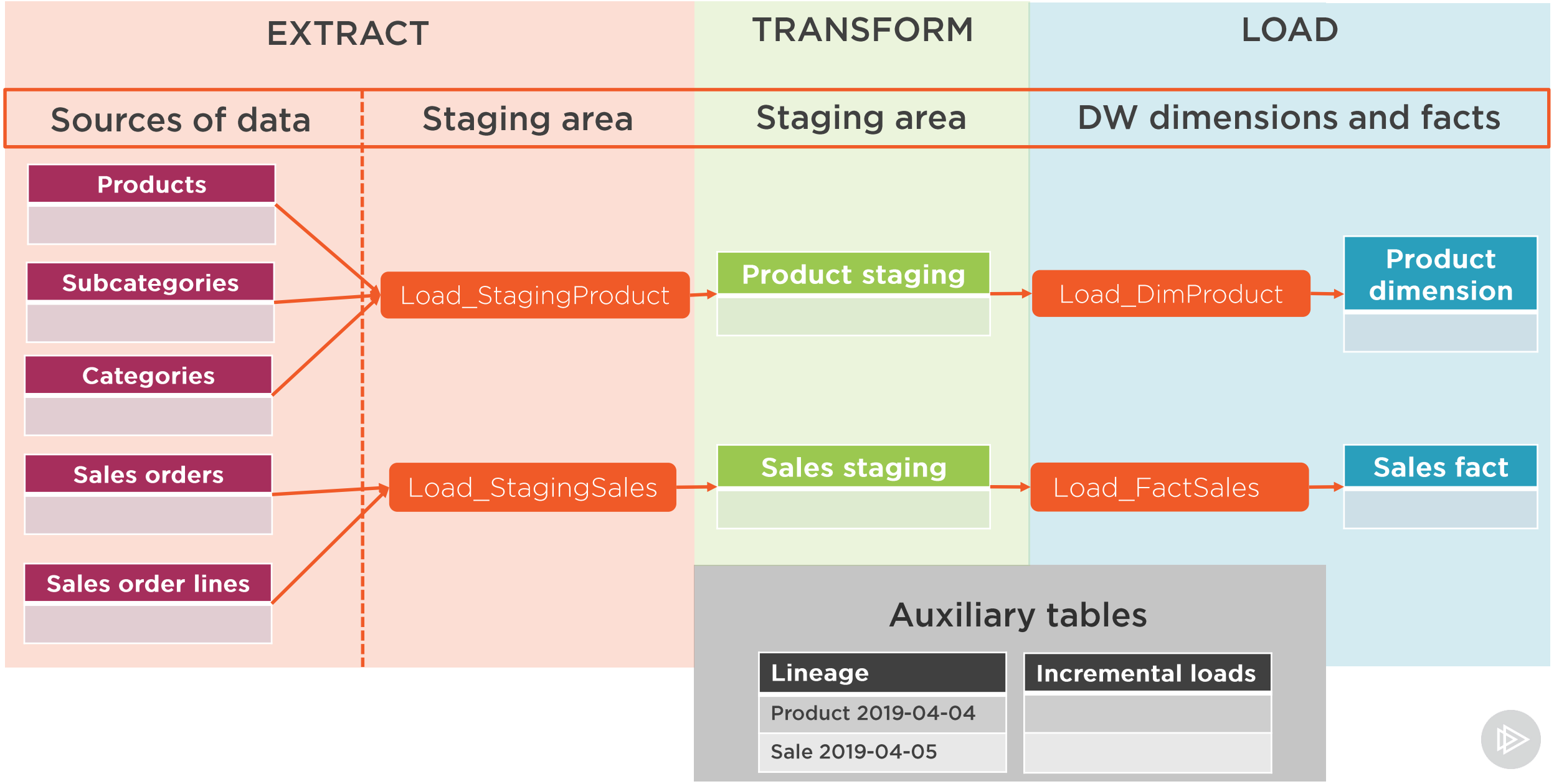
Overview of an ETL System



Overview of an ETL System



Overview of an ETL System



Types of Data Warehouse Loads

Types of Data Warehouse Loads

Full/initial load

The process of populating the data warehouse for the first time with data from the operational system

All tables are truncated and reloaded

Old data is lost

Takes a lot of time to finish

Easy to implement

Incremental load

The process of updating the data warehouse with the operational system changes

Tables are updated with new data

Old data is preserved

Takes less time than the initial load

Implementation is more complex

- Keep track of the previous load date
- Store multiple versions of the same row

Incremental Load Elements

_Source key	Name	Valid from	Valid to
387	Cherry toffee	2019-01-01	2019-09-16
387	Super cherry toffee	2019-09-16	9999-12-31
104	Banana bread	2019-01-01	9999-12-31
105	Grape juice	2019-01-01	9999-12-31

The “Valid from” and “Valid to” columns

Load date key	Table name	Load date
1	Dim_Product	2019-04-13
2	Dim_Employee	2019-01-01
...	...	
16	Dim_Product	2019-04-15

The “Incremental loads” table

Incremental Load Elements

_Source key	Name	Valid from	Valid to
387	Cherry toffee	2019-01-01	2019-09-16
387	Super cherry toffee	2019-09-16	9999-12-31
104	Banana bread	2019-01-01	9999-12-31
105	Grape juice	2019-01-01	9999-12-31

The “Valid from” and “Valid to” columns

Load date key	Table name	Load date
1	Dim_Product	2019-04-13
2	Dim_Employee	2019-01-01
...	...	
16	Dim_Product	2019-04-15

The “Incremental loads” table

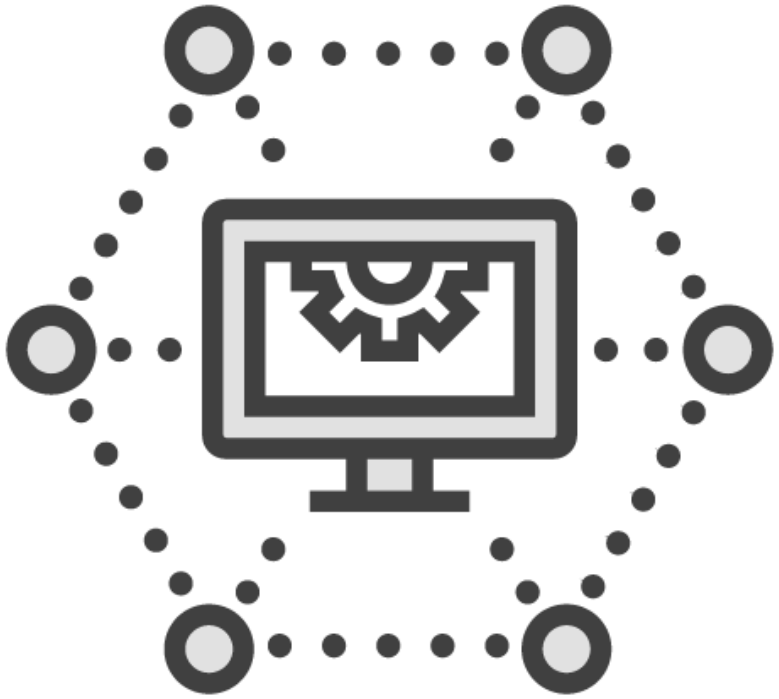
```
SELECT *
FROM Products
WHERE
ModifiedDate > '2019-04-13'
AND ModifiedDate <= '2019-04-15'
```

Demo

Creating and working with the “Incremental loads” table

Setting up Data Lineage

What Is Data Lineage?



Data lineage tracks the movement of data

- What are the origins of data
- Where is the data going to
- When was the data loaded or updated
- What transformations are applied to the data

The complexity of data lineage implementations varies

- Depends on the necessities of the project

It is important to set up data lineage in data warehouse projects

Advantages of Data Lineage

Advantages of Data Lineage

Troubleshooting

Advantages of Data Lineage

Troubleshooting

Data warehouse

Sale number	Product
123	Cherry toffee

Source system

Sale number	Product
123	Super cherry toffee

Advantages of Data Lineage

Troubleshooting

Advantages of Data Lineage

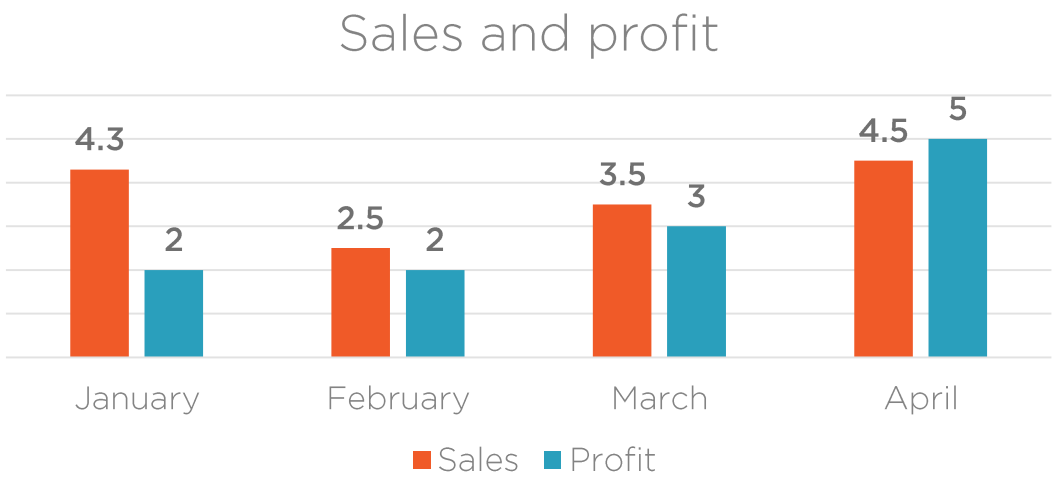
Troubleshooting

Data trust

Advantages of Data Lineage

Troubleshooting

Data trust



Advantages of Data Lineage

Troubleshooting

Data trust

Advantages of Data Lineage

Troubleshooting

Data trust

Transparent business rules

Advantages of Data Lineage

Troubleshooting

Data trust

Transparent business rules

Rate	% of VAT	Applies to
Standard	20%	Mixed ice cream, frozen yogurt
Reduced	5%	Some goods and services
Zero	0%	Herbal tea, pita bread, cold sandwiches, crocodile meat

Advantages of Data Lineage

Troubleshooting

Data trust

Transparent business rules

Advantages of Data Lineage

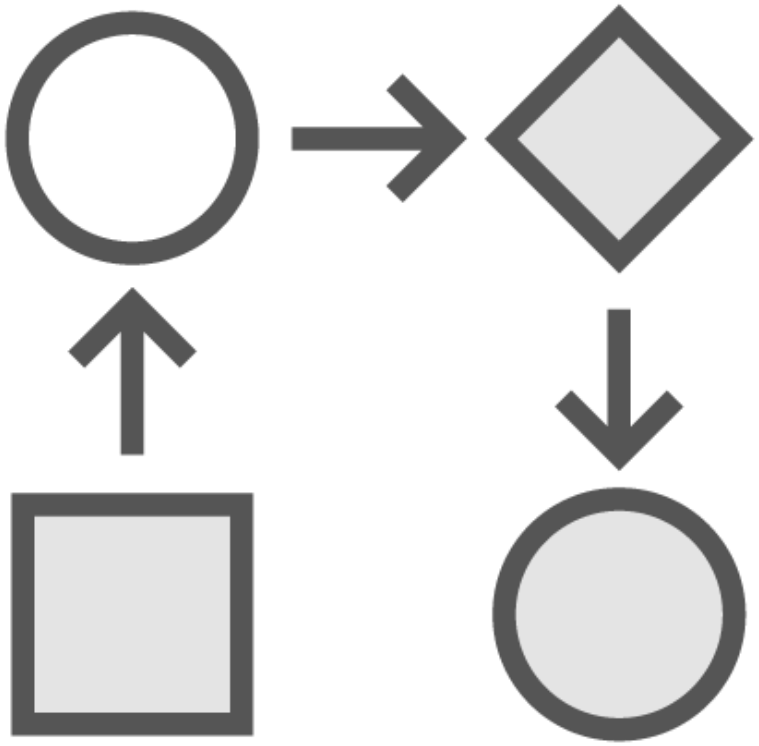
Troubleshooting

Data trust

Transparent business rules

Data audit

Implementing Data Lineage



Ensure row uniqueness

- Use surrogate keys
- Data passes multiple stages from source to destination
- Keeping the same key through all the stages helps tracing back the data

Keep track of the operation that loaded each row

- Create a “Lineage” column in all tables from the data warehouse

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status

Product key	Product name	...	Lineage key

Employee key	Employee name	...	Lineage key

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	null	F	P

Product key	Product name	...	Lineage key

Employee key	Employee name	...	Lineage key

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	null	F	P

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	null	F	P

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	null	F	P

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	2019-03-04 20:59:12.000	F	S

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	2019-03-04 20:59:12.000	F	S
...	...	...	...		...

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	2019-03-04 20:59:12.000	F	S
...	...	...	...		...
28	Dim_Product	2019-03-15 20:07:10.000	null	I	P

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	2019-03-04 20:59:12.000	F	S
...	...	...	...		...
28	Dim_Product	2019-03-15 20:07:10.000	null	I	P

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10
3	Grape juice		28
4	Banana bread		28

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Implementing Data Lineage

Lineage key	Table name	Start	End	Type	Status
10	Dim_Product	2019-03-04 20:39:10.000	2019-03-04 20:54:35.000	F	S
11	Dim_Employee	2019-03-04 20:45:21.000	2019-03-04 20:59:12.000	F	S
...	...	...	...		...
28	Dim_Product	2019-03-15 20:07:10.000	2019-03-15 20:12:33.000	I	S

Product key	Product name	...	Lineage key
1	Cotton candy		10
2	Green tea		10
3	Grape juice		28
4	Banana bread		28

Employee key	Employee name	...	Lineage key
1	Mary Baker		11
2	Jack Peanut		11
3	Gigi Knopper		11
...	...		...

Demo

Creating and working with the Lineage table

Demo

Populating a dimension table with the help of stored procedures

- Load_StagingProduct
- Load_DimProduct

Objects Participating in a Load Process

Object (table/stored procedure)	Was discussed
Fact/dimension tables	✓
Staging tables	✓
SP for loading the staging table	✓
SP for loading the fact/dim table	✓
Lineage table	✓
Incremental loads table	✓
SPs for updating the log tables	✓

Summary

Overview of an ETL system

Types of data warehouse loads

- Full (initial) load
- Incremental load

Setting up data lineage

Demos

- Creating and using the auxiliary tables
  - Lineage
  - Incremental loads
- Populating with data the staging and dimension tables