
DS-306 Data Warehousing and Business Intelligence

Topic 7: ETL Designing

Dr. Khurram Shahzad

<https://www.youtube.com/watch?v=0ikNnenDyNw>

Two strategies of DW

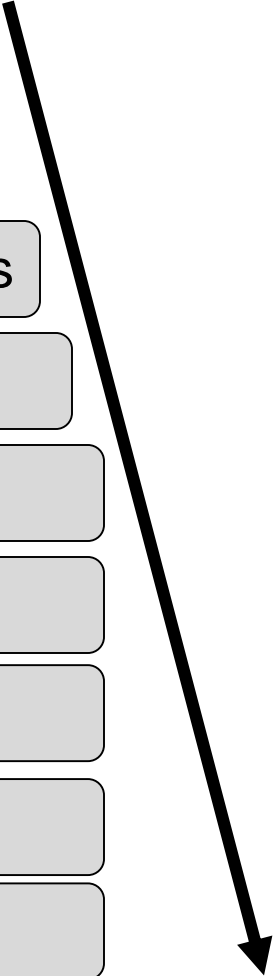
- Top down (DW to Marts)
- Bottom up (Mars to DW)

Two strategies of DW

- Top down (DW to Marts)
 - Typically, separate ETL for DW and for Marts
- Bottom up (Marts to DW)
 - Enterprise-wide ETL is partly from data marts
 - Partly from data sources

The ETL Solution

Major steps in ETL

1. Determine all the target data needed in the DW
 2. Determine all the data sources, internal and external
 3. Prepare data mapping for target data elements from sources
 4. Establish comprehensive data extraction rules
 5. Determine all the target data needed in the DW
 6. Plan for aggregate tables
 7. Organize data staging areas and test tools
 8. Write procedures for all data loads
 9. ETL for dimension tables
 10. ETL for fact tables
- 

ETL Modeling

Overview

- ETL is a complex and costly task, so its important to reduce its development and maintenance cost
- Modeling ETL at a conceptual level is the way forward
- Tools have their own languages
 - Pehntaho Data Integration (Keetle) has a language
 - Microsoft Integration Service has a language

Overview

- ETL is a combination of **control** and **data** tasks
- **Controls** which represent the orchestration of an ETL process
- **Data tasks** define what actions need to be performed on data
 - Actions are, extraction, transformation, loading

BPMN

- Business Process Modeling Notation (BPMN)
- A **de facto standard** for specifying business processes
- A **graphical notation** for defining and understanding processes
- BPMN provides a conceptual and implementation-independent **specification** of processes
- **Hides technical details** and allows designers

BPMN

- ‘A business process is a collection of **related activities** or **tasks** in an organization whose goal is to produce a service or product’
- Task can be performed by software, human or both
- BPMN released by OMG

BPMN elements

- Four basic categories of elements
 - Flow objects: Have three types
 - Activities
 - Gateways
 - And, Events
 - Connecting objects
 - Swimlanes, and
 - Artifacts

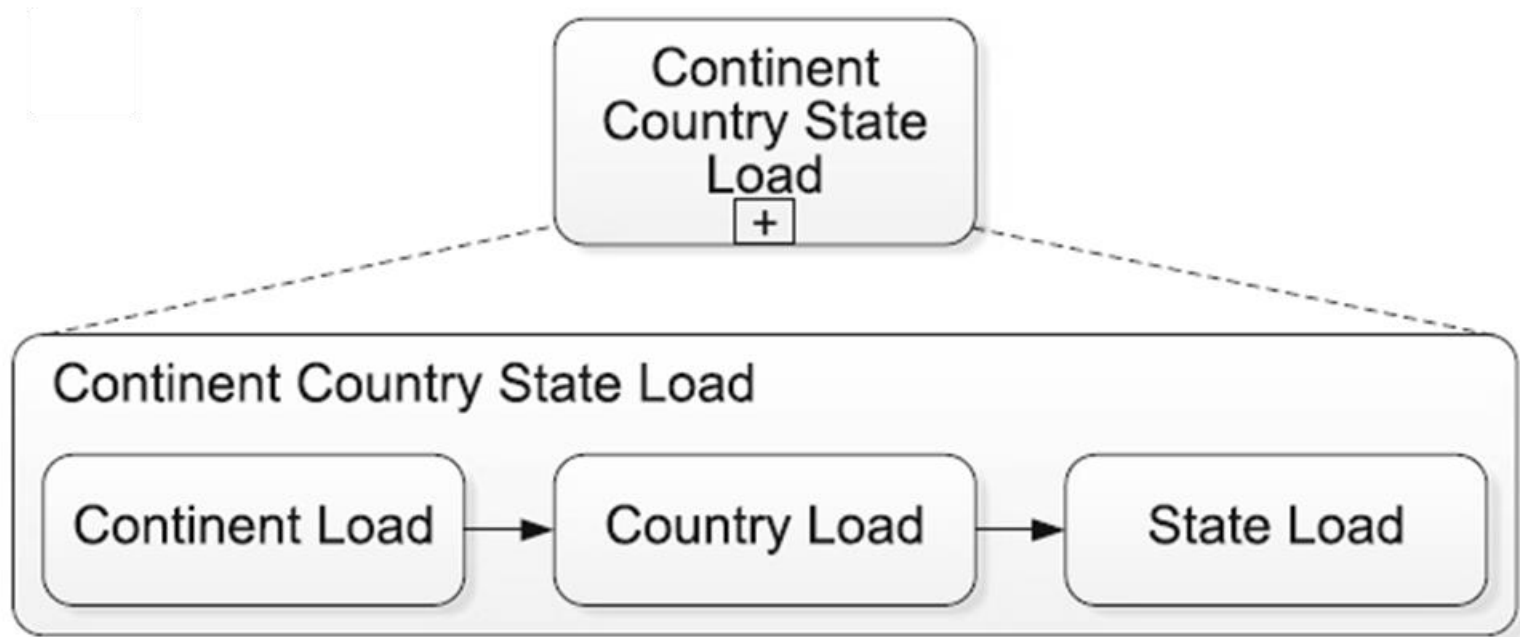
1. Flow Objects: *Activities*



Product Load

Activity

1. Flow Objects: Subprocess



Sub-process (collapsed and expanded)

1. Flow Objects: Gateways

- Gateways are used to control sequence of activities

a



Exclusive



Inclusive

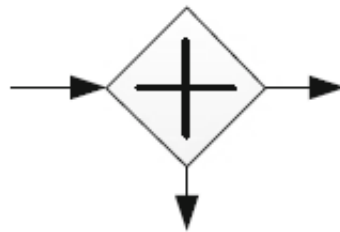


Parallel

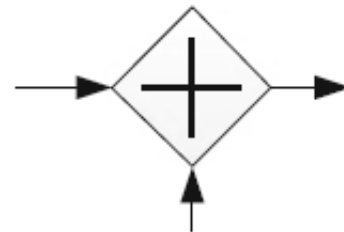


Complex

b



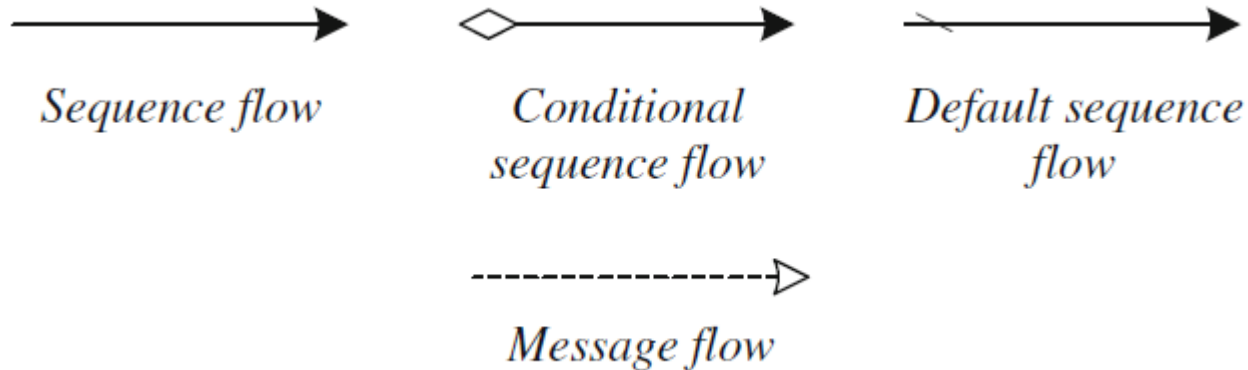
Splitting



Merging

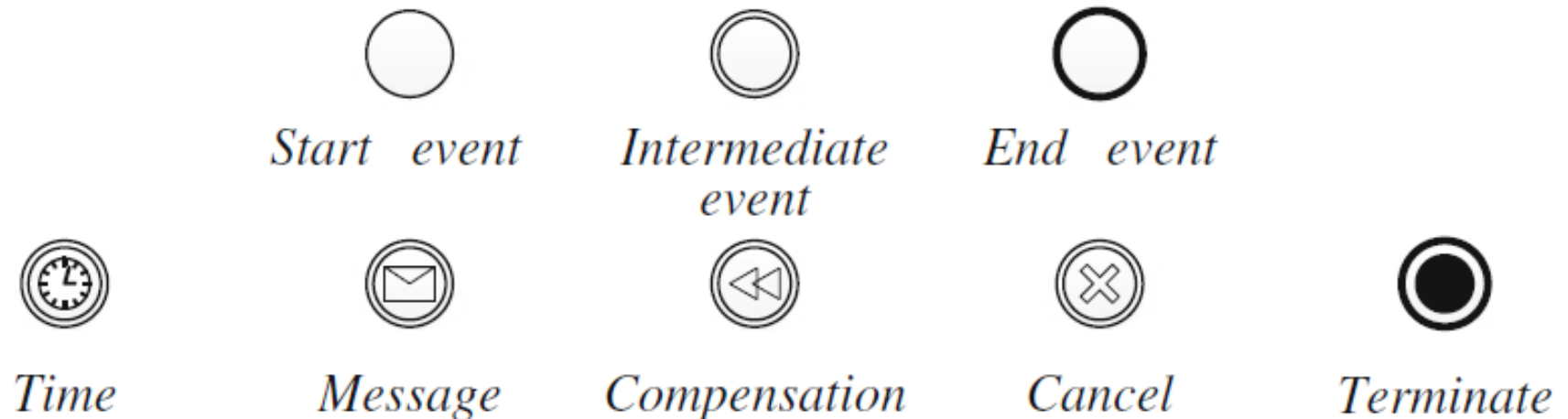
2. Connecting objects

- Used to represent how objects are connected

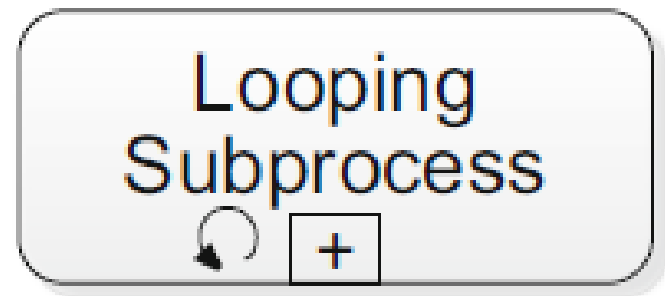


3. Events

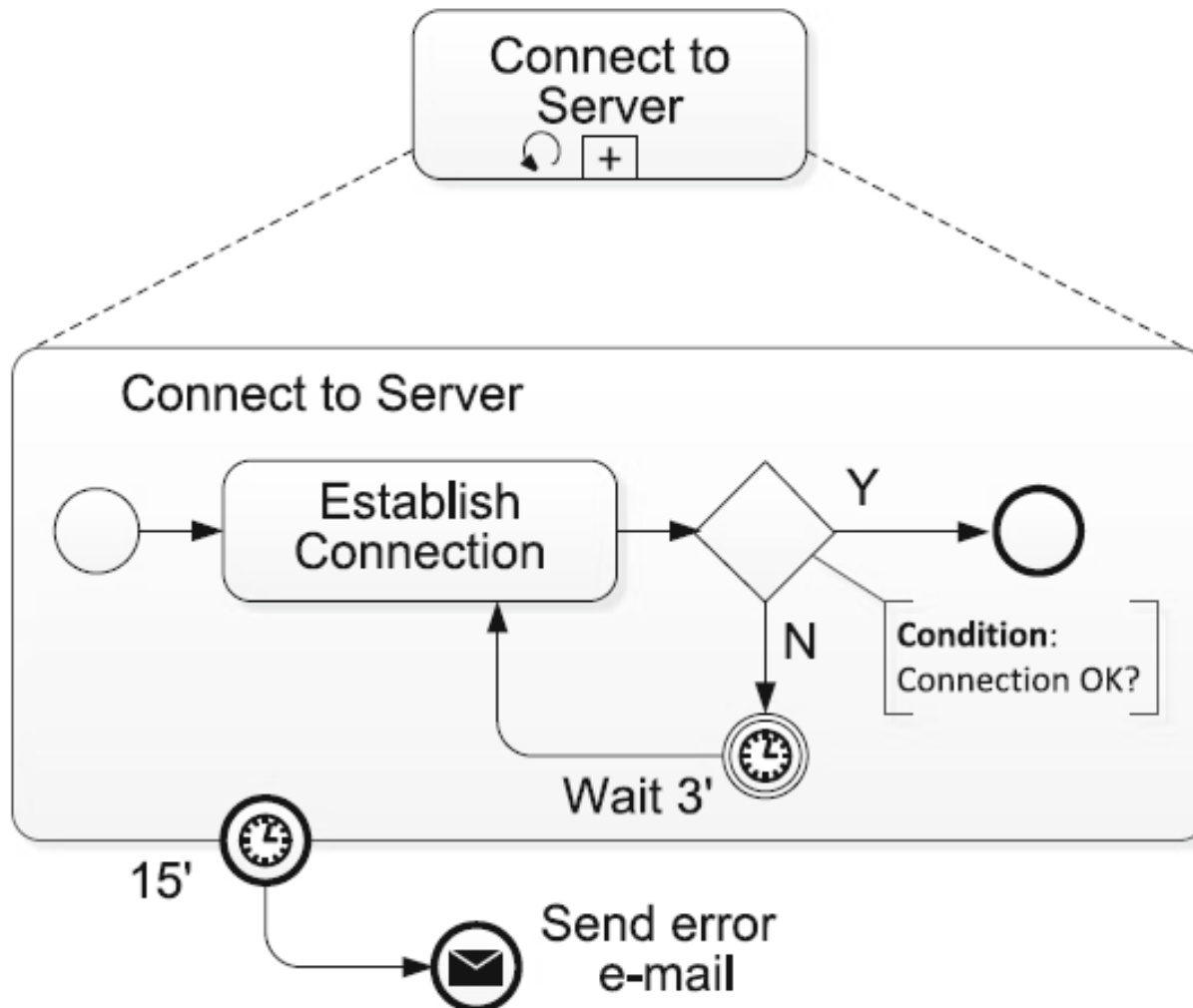
- Represents something that happens that can affect sequence of timing of the workflow activities



Flow Objects: Activities loops



Flow Objects: Activities loops



3. Swimlanes

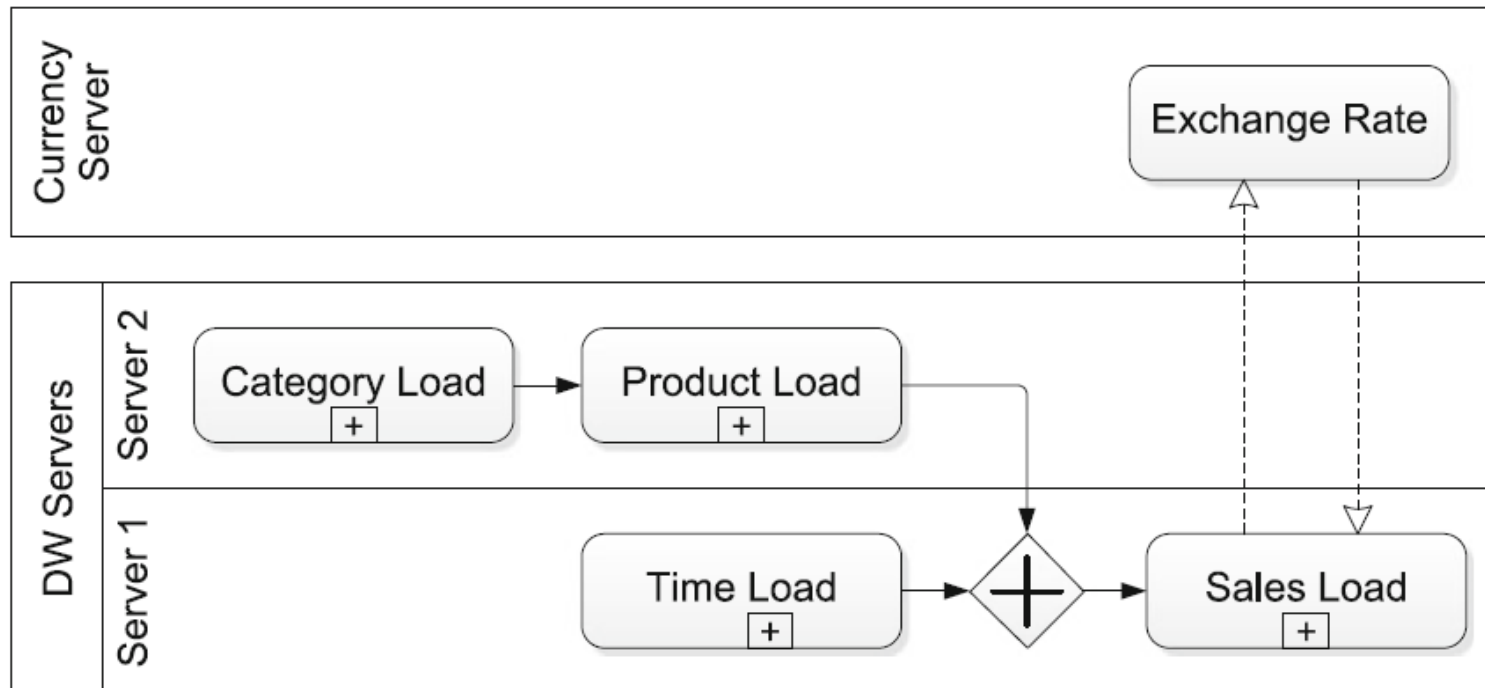
- **Swimlanes** are used to organized ETL process according to strategies
 - Technical **Architectures** (such as servers)
 - **Business Entities** (such as departments)

And Occasionally

- **User profiles** (such as managers, analysts)

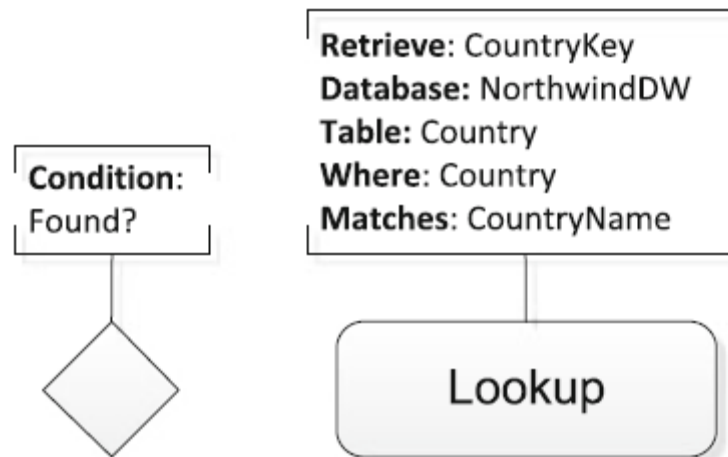
3. Swimlane

- A structuring object that comprises pools and lanes



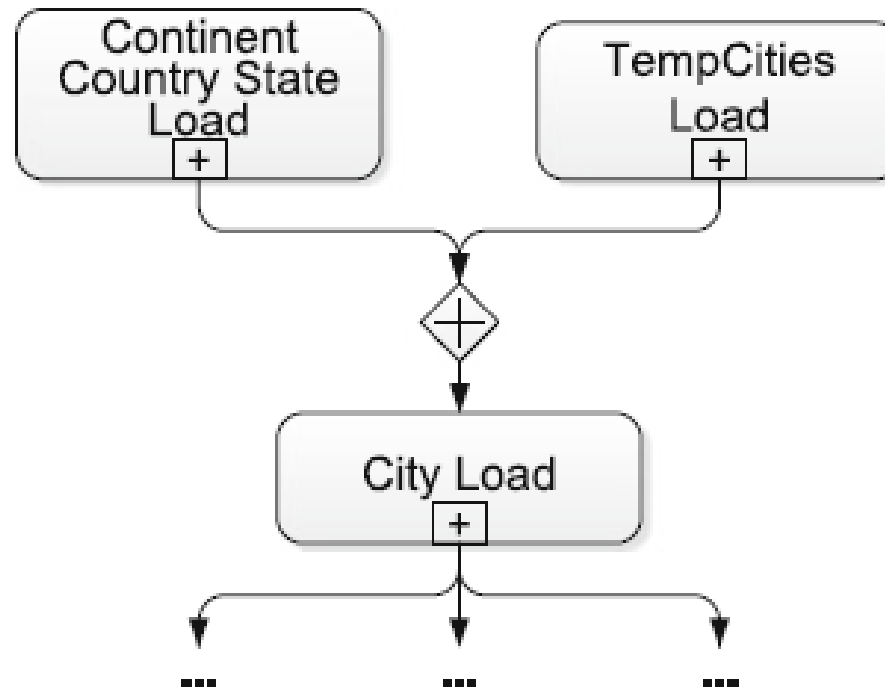
4. Artifacts

- Artifacts are used to add information to diagram
 - Data object, grouping of tasks, annotations



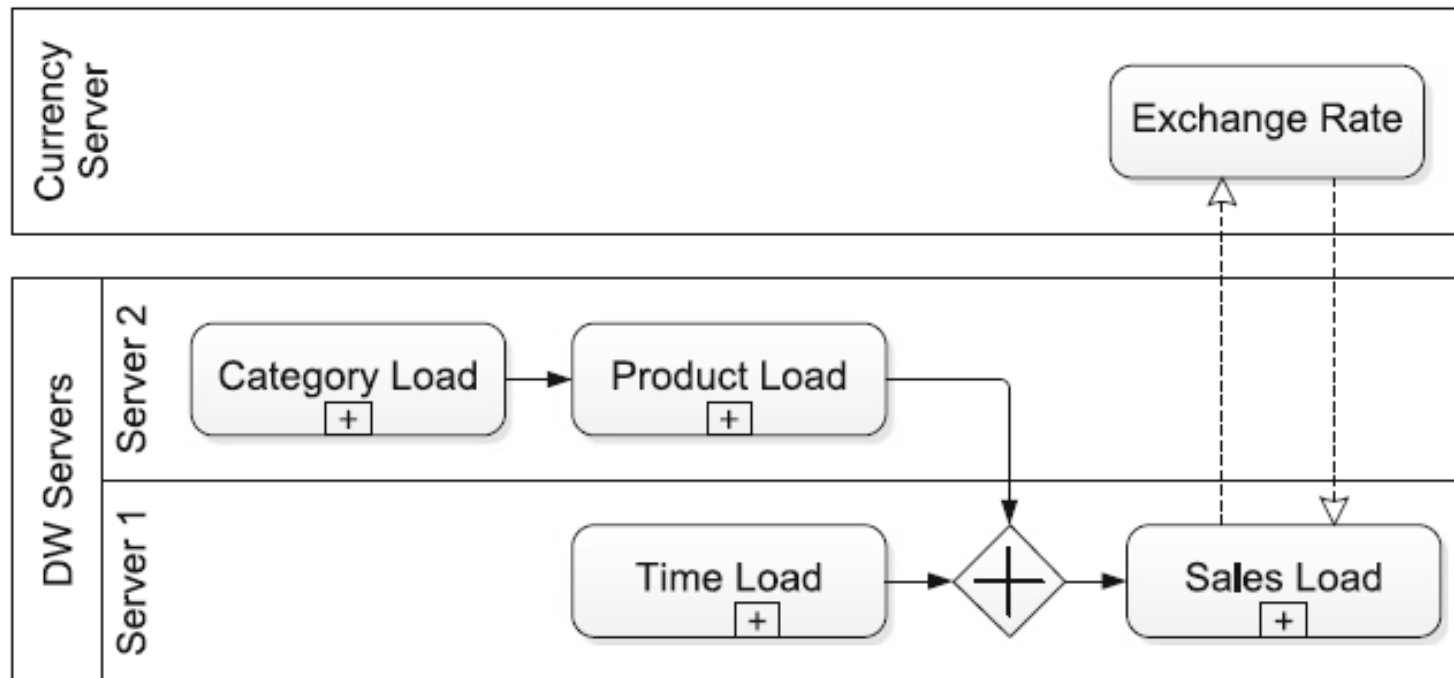
Conceptual ETL design using BPMN

ETL Design using BPMN



ETL Design using BPMN

- Swimlanes example with architecture



ETL Design using BPMN

- **Data tasks** represent activities typically carried out to manipulate additive measure

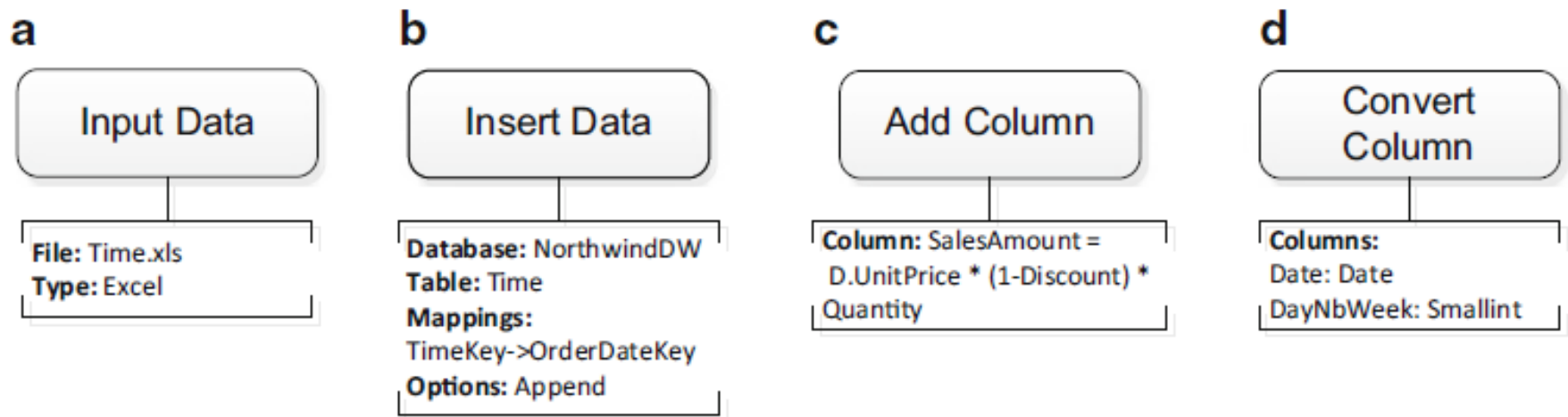


Fig. 8.11 Unary row operations. (a) Input data. (b) Insert data. (c) Add column. (d) Convert column

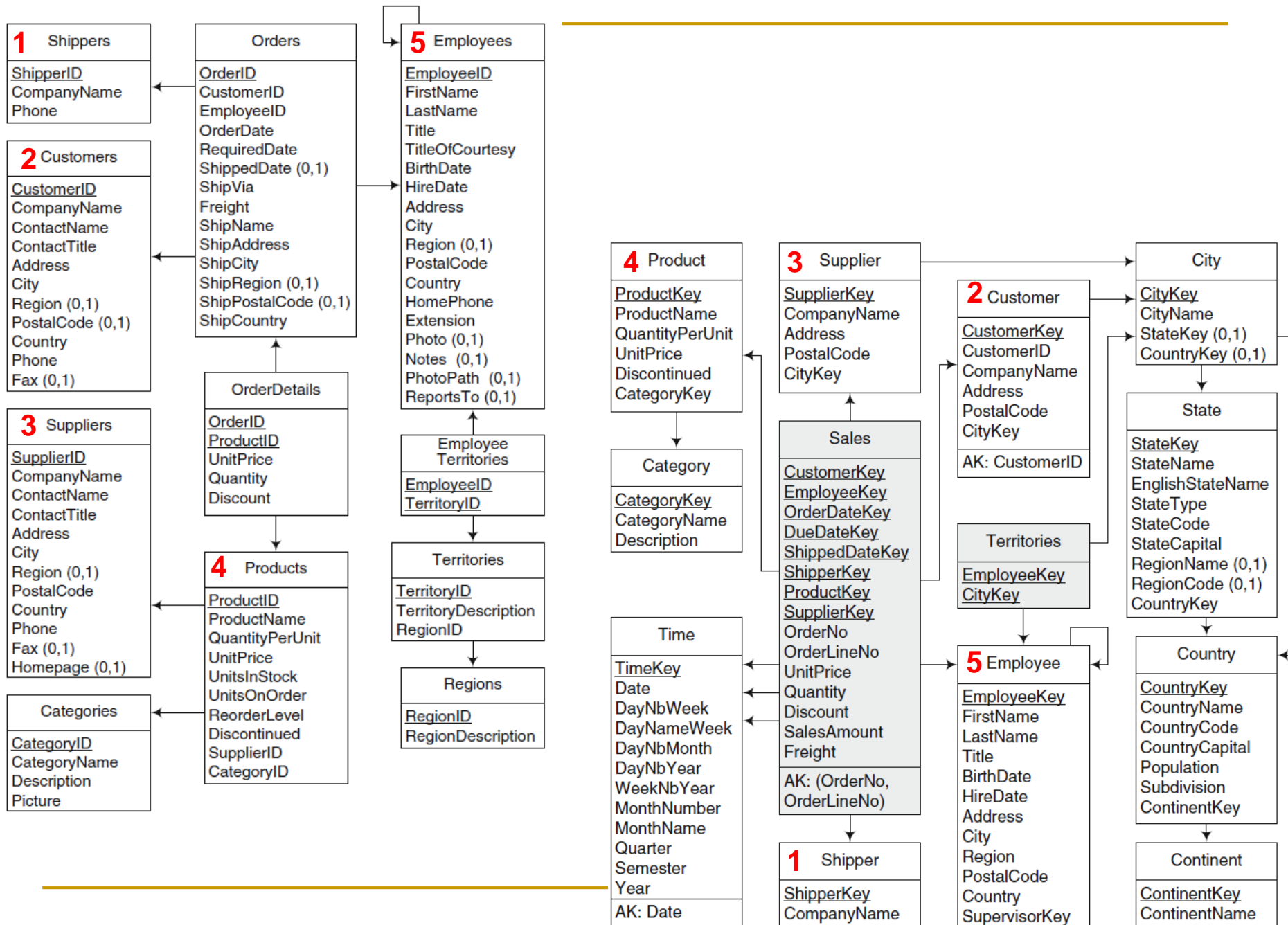
ETL Design using BPMN

■ Rowset operations



Fig. 8.12 Rowset operations. (a) Aggregate (unary). (b) Join (binary). (c) Union (n -ary)

NorthWind case study



Conceptual Design for NorthWind

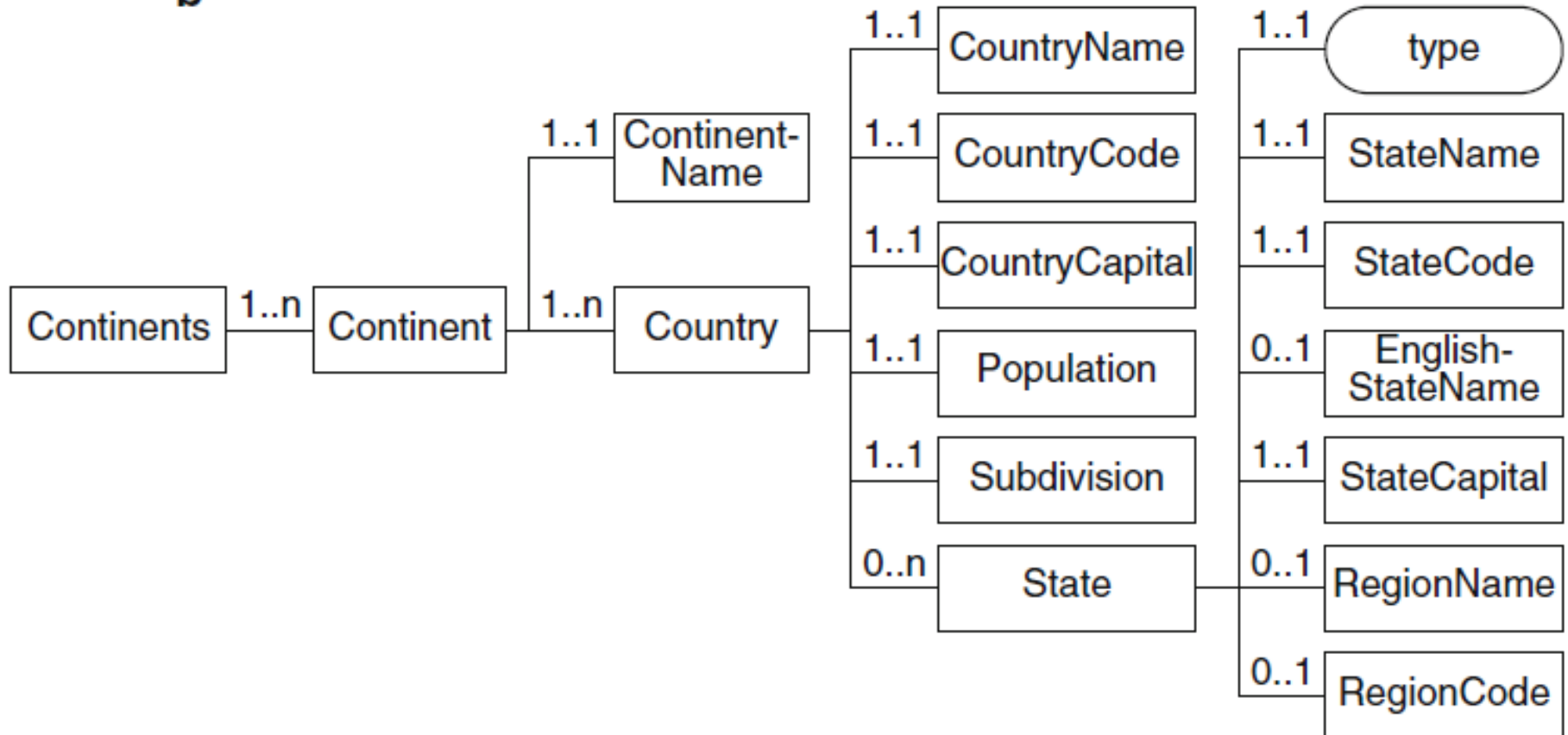
- In addition to operational database, some other files are needed for loading DW
 - Supplier, customer geographic hierarchy City -> State -> Country -> Continent

XML schema Territories

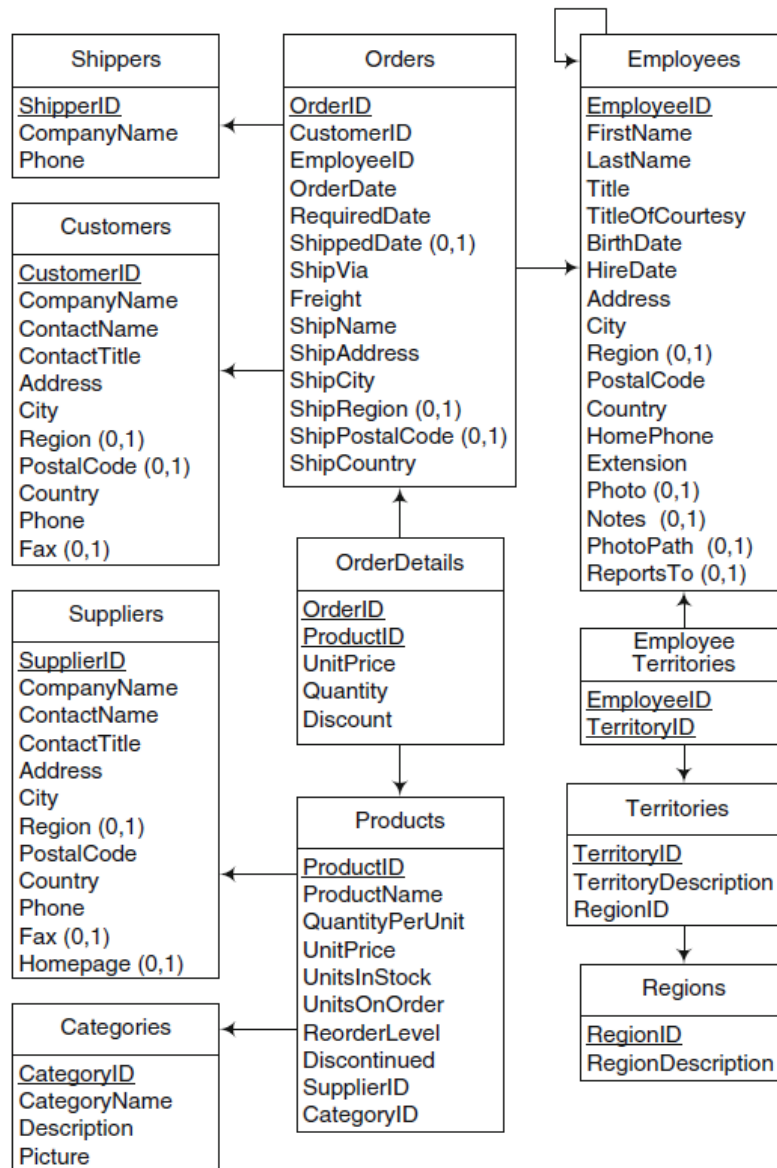
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Continents>
  <Continent>
    <ContinentName>Europe</ContinentName>
    <Country>
      <CountryName>Austria</CountryName>
      <CountryCode>AT</CountryCode>
      <CountryCapital>Vienna</CountryCapital>
      <Population>8316487</Population>
      <Subdivision>Austria is divided into nine Bundesländer,
        or simply Länder (states; sing. Land).</Subdivision>
      <State type="state">
        <StateName>Burgenland</StateName>
        <StateCode>BU</StateCode>
        <StateCapital>Eisenstadt</StateCapital>
      </State>
      <State type="state">
        <StateName>Kärnten</StateName>
        <StateCode>KA</StateCode>
        <EnglishStateName>Carinthia</EnglishStateName>
        <StateCapital>Klagenfurt</StateCapital>
      </State>
    ...
```

XML schema of the file

b



Fact Table

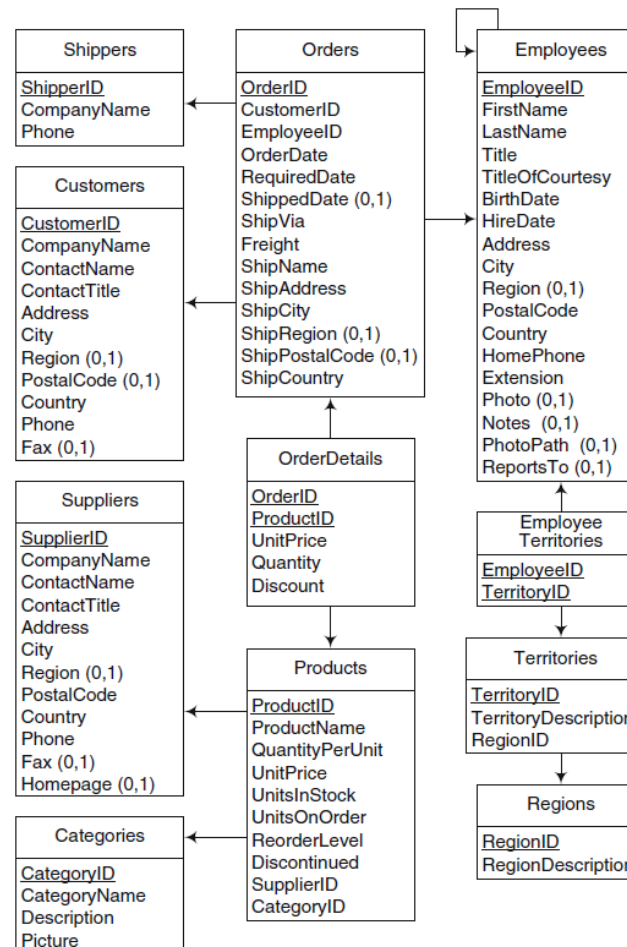


Sales

CustomerKey
EmployeeKey
OrderDateKey
Duedatekey
Shippeddatekey
Shipperkey
 OrderNo
 OrderLine No
 UnitPrice
 Quantity
 Discount
 SalesAmount
 Freight

Fact Table

- Sales amount is calculated using
 - Unit price, discount and quantity

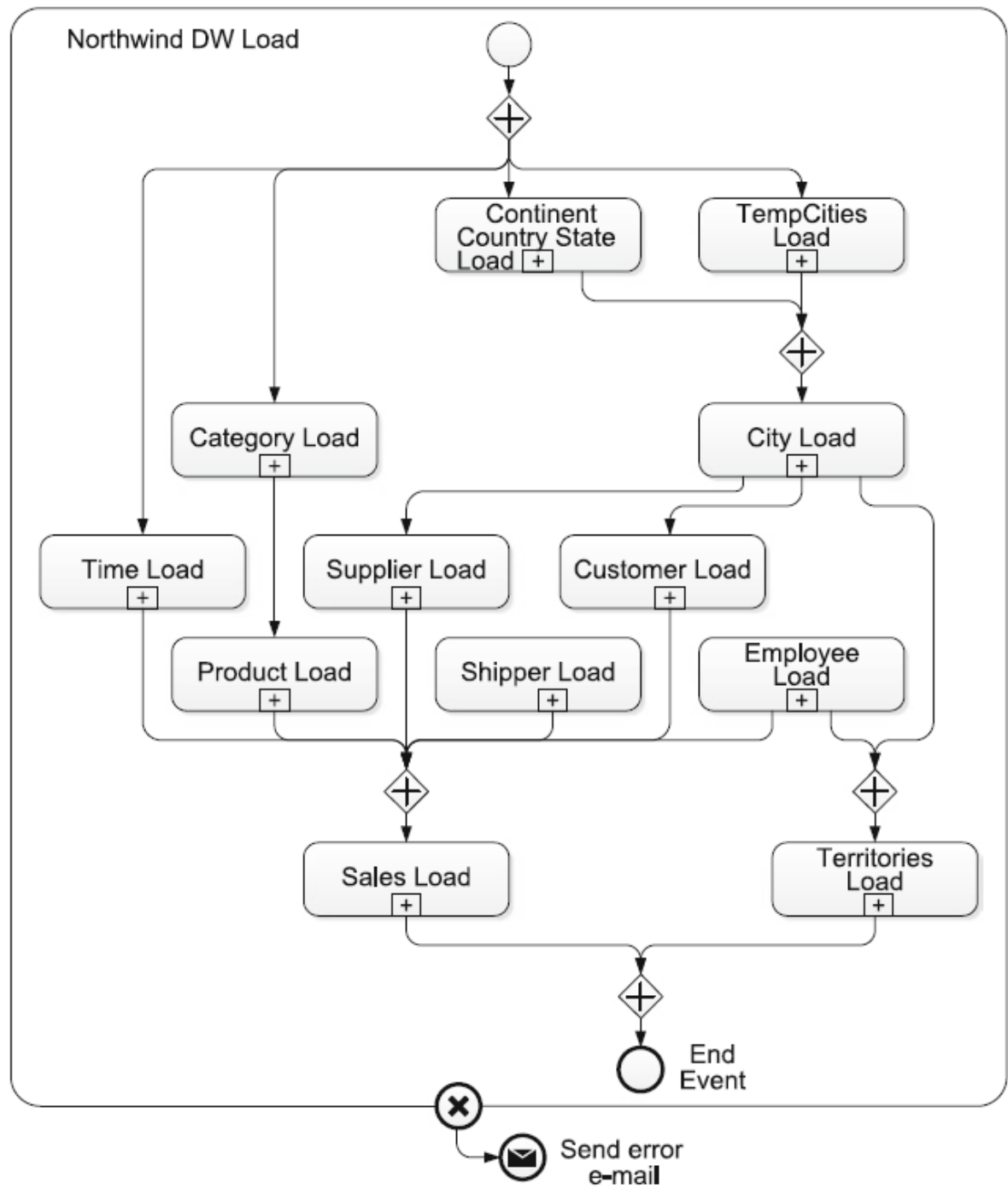


Sales

CustomerKey
EmployeeKey
OrderDateKey
Duedatekey
Shippeddatekey
Shipperkey
OrderNo
OrderLine No
UnitPrice
Quantity
Discount
SalesAmount
Freight

ETL

■ Overall

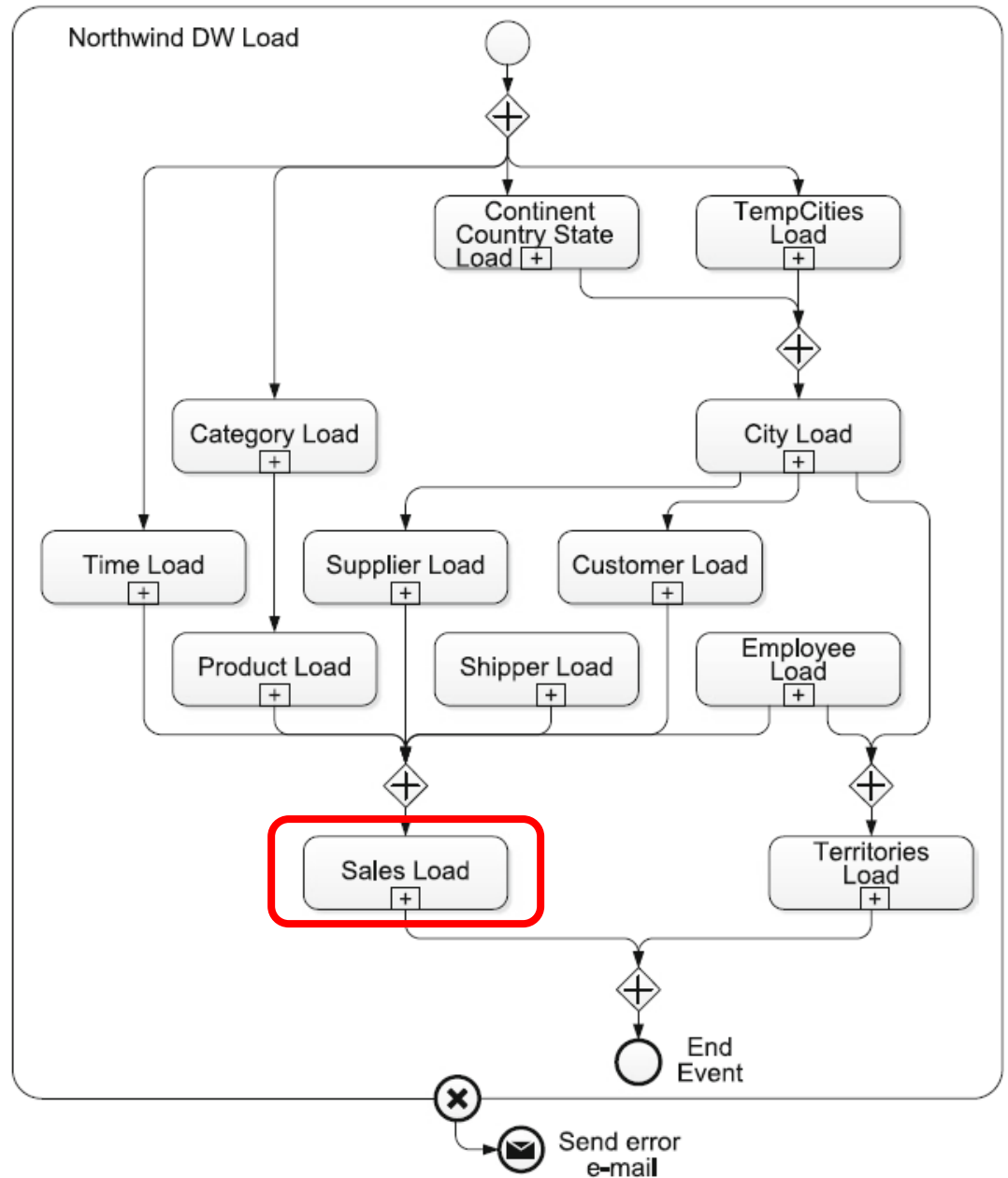


Two level ETL modeling

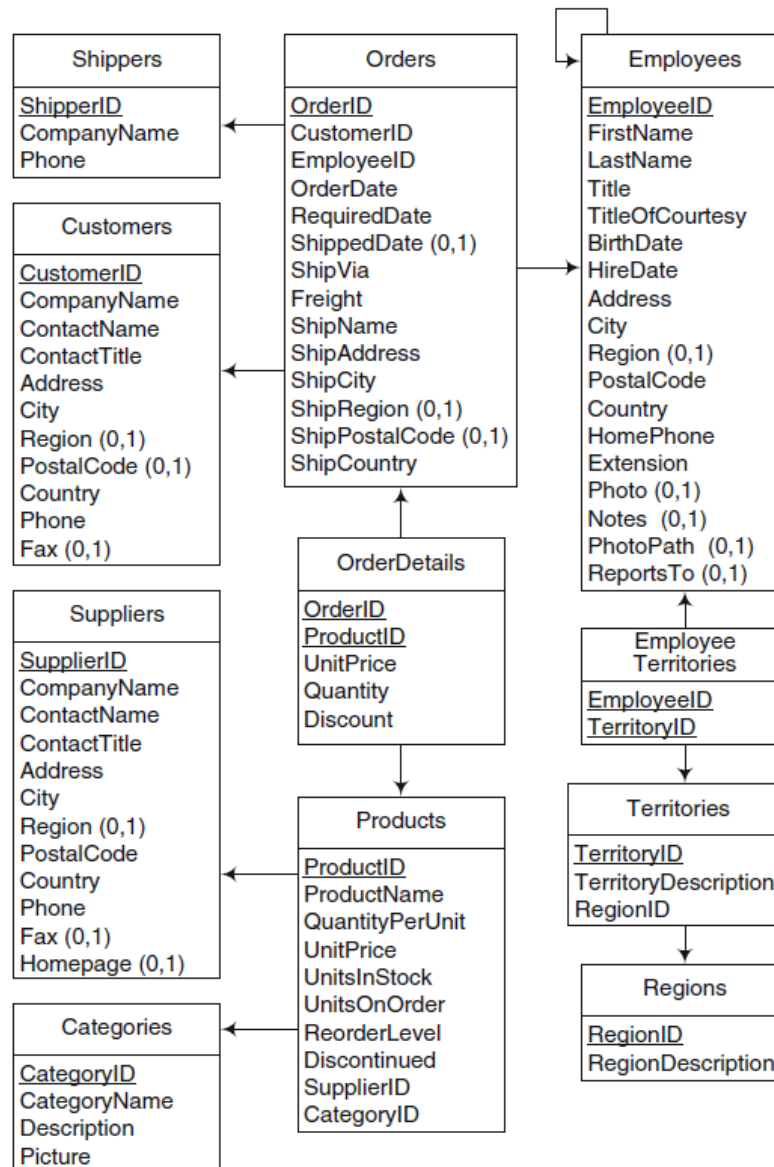
- Overall ETL
 - Process, composed of subprocesses
- Table level
 - Subprocesses composed of activities or further subprocesses

Subprocess

■ Sales Load



Create sub-process of Sales fact



Sales

CustomerKey
EmployeeKey
OrderDateKey
Duedatekey
Shippeddatekey
Shipperkey
 OrderNo
 OrderLine No
 UnitPrice
 Quantity
 Discount
 SalesAmount
 Freight

Job scheduling

- The scheduler jobs can be made time or event dependent or file dependent
 - Example, start the job at 3 AM
 - Wait for a particular file to be available and trigger the first file
- Support teams usually monitor the job runs
 - Teams can see
 - If the flow is running fine
 - If a job has failed
 - Reason for failure
 - What job is waiting, etc.

Error Handling

- ❑ After scheduling of job comes monitoring
- ❑ There are two parts of monitoring
 - What to do to get the job back up so that the flow can resume?
 - Log the error in the table
- ❑ For staging job and dimension jobs ETL tool can be configured to run again on failure
- ❑ For minor error these can be set and job can be re-run
- ❑ If it fails again, support team will look at it

Error Handling

- ❑ Audit table form an important part of the ETL process
- ❑ One design approach would be to create groups, dimension group, fact group
- ❑ At the end of completion of all the jobs in the group, another job makes an entry saying that group has completed
- ❑ Apart from audit tables, ETL tools have their own logs to debug errors