DataCo Supply Chain Data Warehousing

FEUP - MECD - Data Warehouse
Middle presentation

Carlos Miguel Veloso
Cátia Teixeira
Luís Henriques
Rojan Aslani

Introduction

Assignment Goals

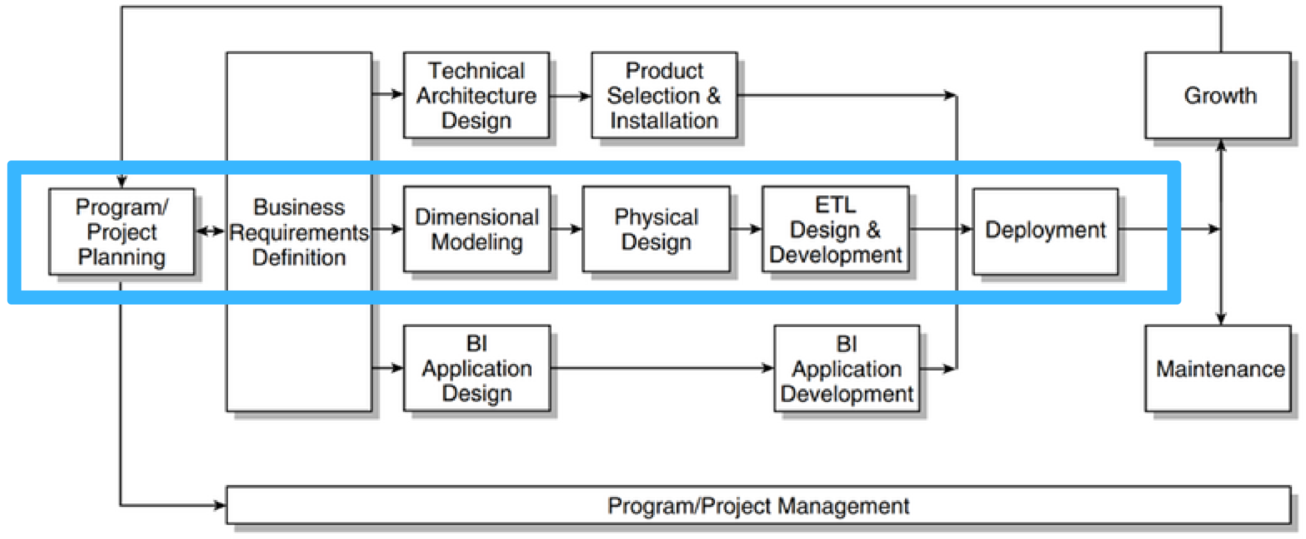
- To design a data warehouse, implement it, and exemplify its use

Assignment Requirements

- The number of rows must be over 10000 with, at least, one additive measure
- There must be aggregated facts or snapshots with at least one semi-additive measure
- There must be at least 4 dimensions, one of which temporal, and some of them are common to both kinds of facts.

Introduction

Assignment steps



Technical Architecture Design
Product Selection & Installation
Growth
Program/ Project Planning
Business Requirements Definition
Dimensional Modeling
Physical Design
ETL Design & Development
Deployment
BI Application Design
BI Application Development
Maintenance
Program/Project Management

Kimball Lifecycle diagram

Project phase	Tasks
Project Planning	• Timeline, general tasks definition and distribution • Finding data
Business Requirements Definition	• Data understanding • Scope definition and data filtering
Dimensional Modeling	• Relational model • Entity relationship • Bus Matrix • Dimensional design
Physical Design	• Data warehouse implementation
ETL Design & Development	• ETL process definition • Loading data do Postgres
Deployment	• Data analysis and business analytics

Source

Dataset source

Identify and profile operational data (OLTP) sources

Dataset

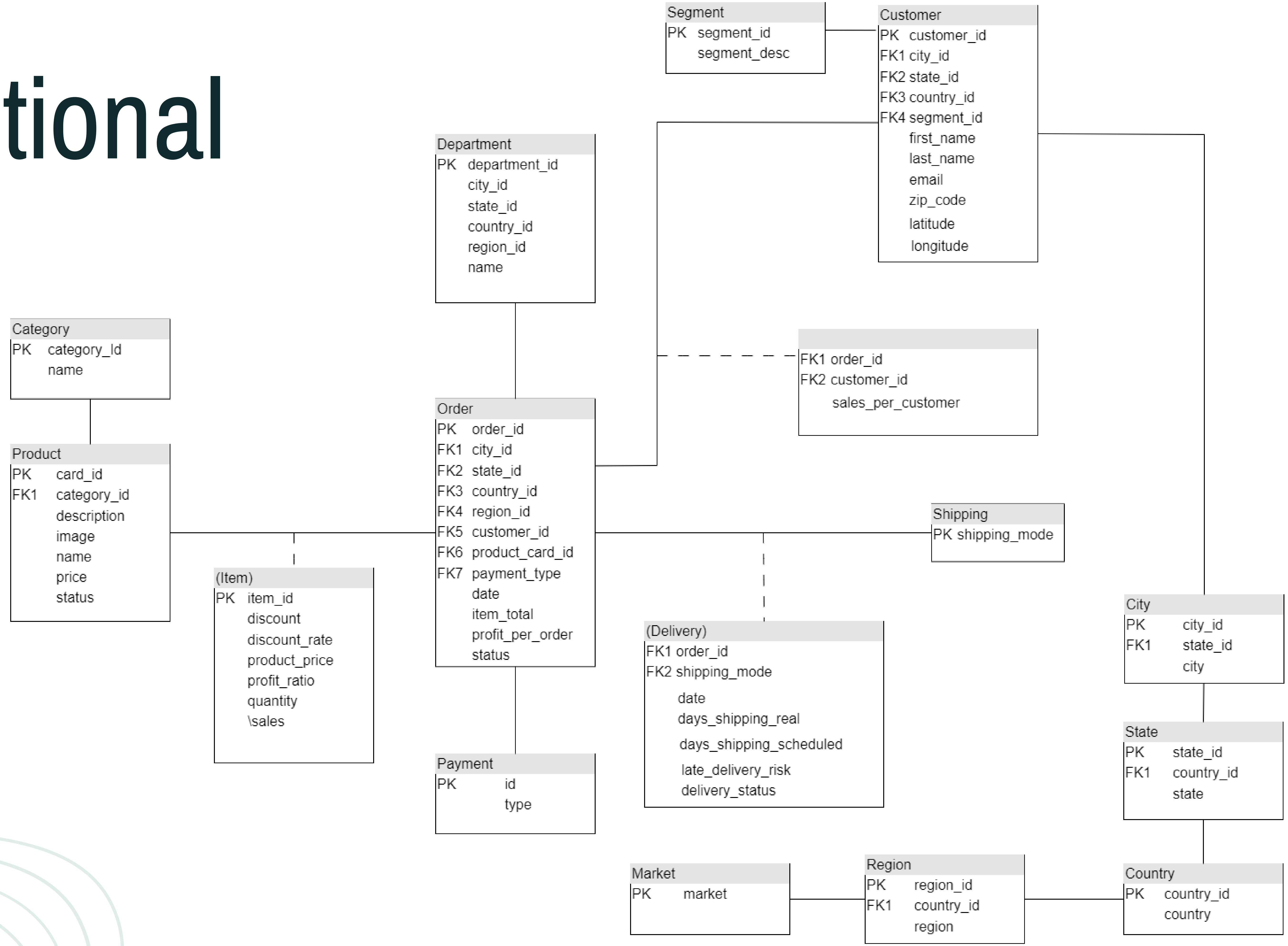
Online store transactions

- Product data
- Financial data
- Sales and demand data

	Original dataset	Reduced dataset
Columns	53	47
Rows	180000+	27128
Timespan	2015 to 2018	2nd Semester of 2017

- Data set was sourced from Kaggle platform
- Made available by Politécnico de Leiria
- Data related to Datco Company

Transactional schema



Segment
PK segment_id
segment_desc

Customer
PK customer_id
FK1 city_id
FK2 state_id
FK3 country_id
FK4 segment_id
first_name
last_name
email
zip_code
latitude
longitude

Department
PK department_id
city_id
state_id
country_id
region_id
name

FK1 order_id
FK2 customer_id
sales_per_customer

Category
PK category_Id
name

Product
PK card_id
FK1 category_id
description
image
name
price
status

(Item)
PK item_id
discount
discount_rate
product_price
profit_ratio
quantity
\sales

Order
PK order_id
FK1 city_id
FK2 state_id
FK3 country_id
FK4 region_id
FK5 customer_id
FK6 product_card_id
FK7 payment_type
date
item_total
profit_per_order
status

Shipping
PK shipping_mode

(Delivery)
FK1 order_id
FK2 shipping_mode
date
days_shipping_real
days_shipping_scheduled
late_delivery_risk
delivery_status

City
PK city_id
FK1 state_id
city

State
PK state_id
FK1 country_id
state

Payment
PK id
type

Market
PK market

Region
PK region_id
FK1 country_id
region

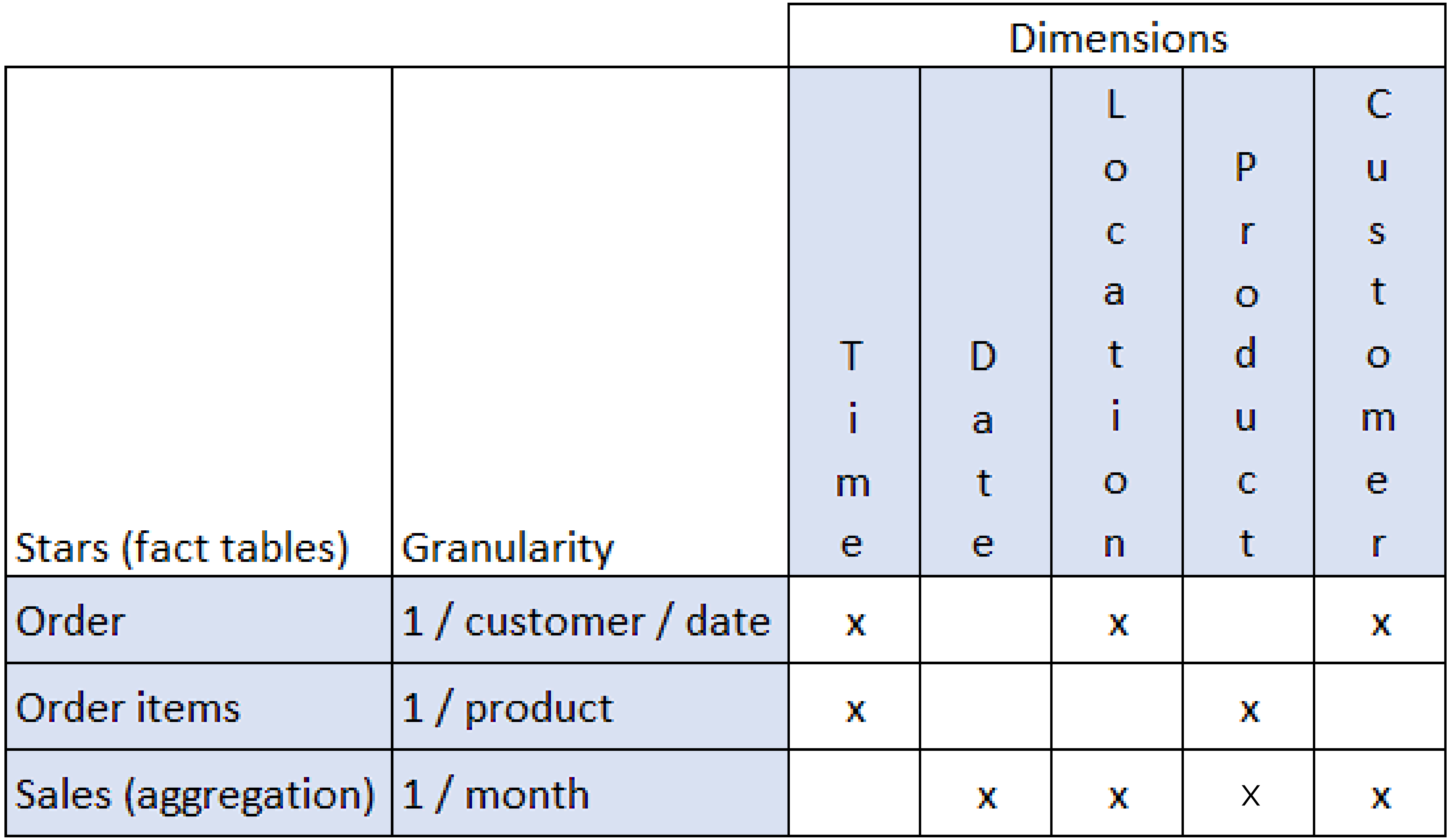
Country
PK country_id
country

Dimensional Model

Develop a dimensional model that includes Dimensions and Facts

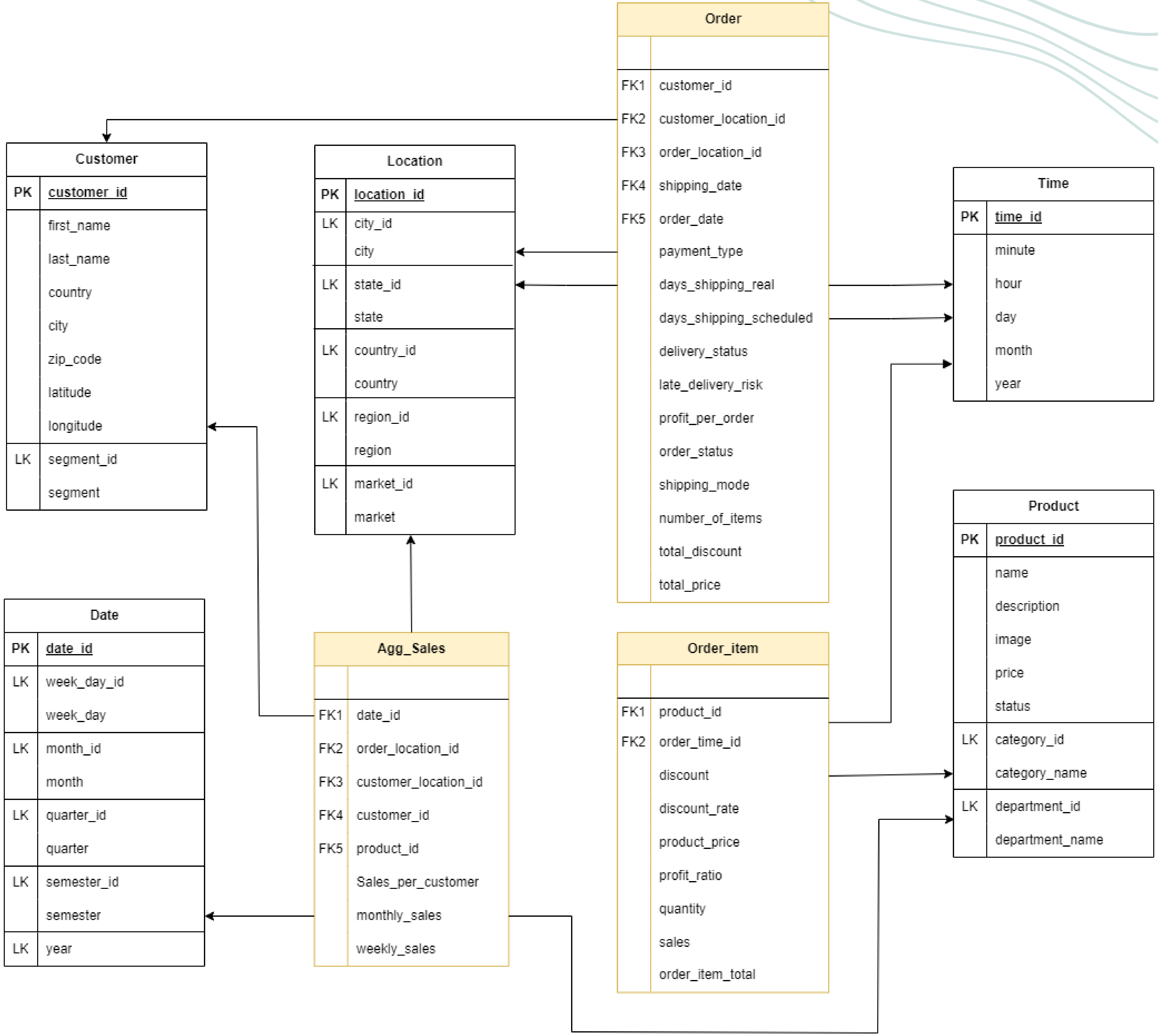
Dimensional Model

Dimensional Bus Matrix



		Dimensions				
Stars (fact tables)	Granularity	Time	Date	Location	Product	Customer
Order	1 / customer / date	x		x		x
Order items	1 / product	x			x	
Sales (aggregation)	1 / month		x	x	x	x

Dimensional Model



Customer	
PK	customer_id
	first_name
	last_name
	country
	city
	zip_code
	latitude
	longitude
LK	segment_id
	segment

Location	
PK	location_id
LK	city_id
	city
LK	state_id
	state
LK	country_id
	country
LK	region_id
	region
LK	market_id
	market

Order	
FK1	customer_id
FK2	customer_location_id
FK3	order_location_id
FK4	shipping_date
FK5	order_date
	payment_type
	days_shipping_real
	days_shipping_scheduled
	delivery_status
	late_delivery_risk
	profit_per_order
	order_status
	shipping_mode
	number_of_items
	total_discount
	total_price

Time	
PK	time_id
	minute
	hour
	day
	month
	year

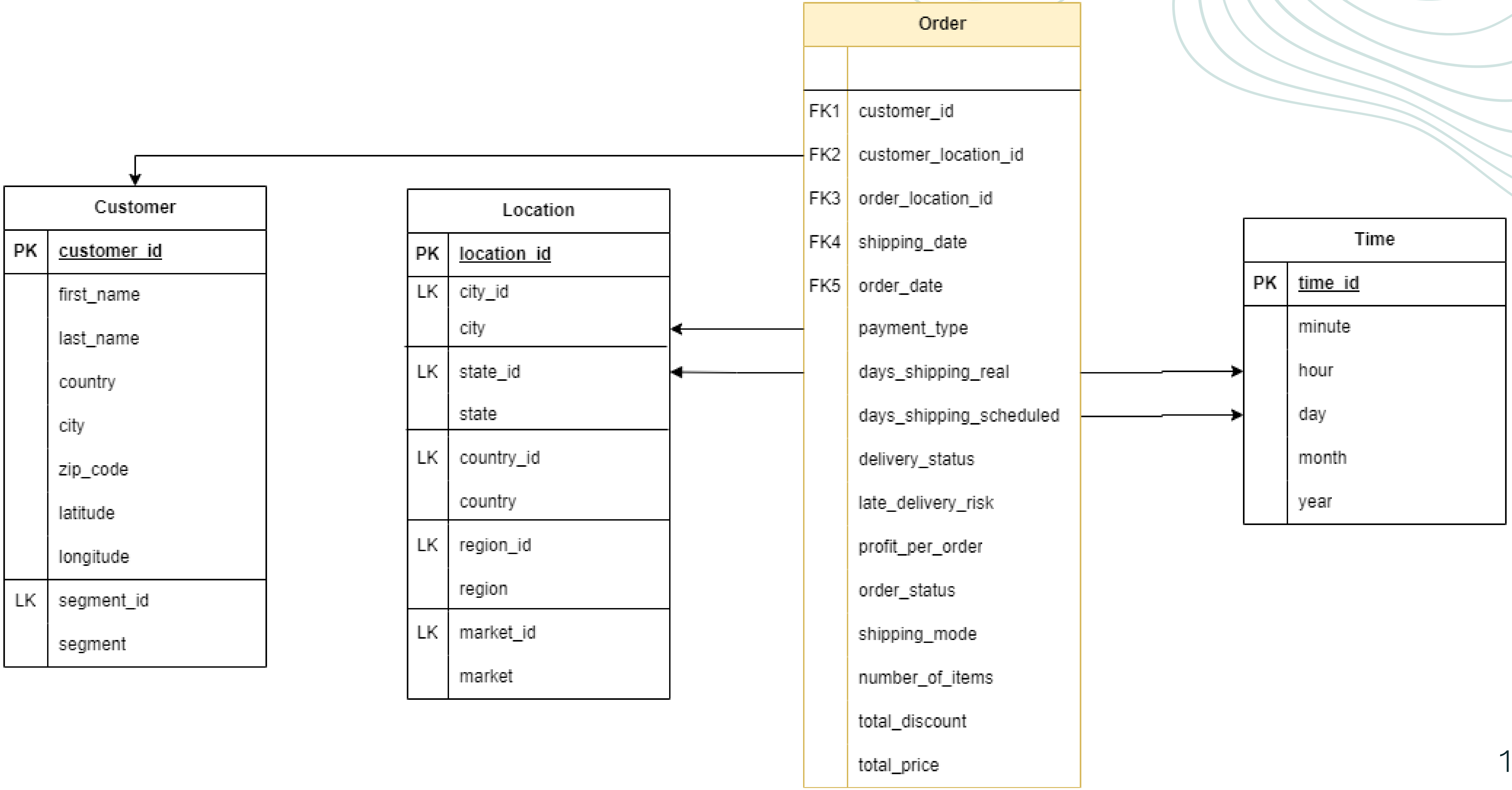
Product	
PK	product_id
	name
	description
	image
	price
	status
LK	category_id
	category_name
LK	department_id
	department_name

Date	
PK	date_id
LK	week_day_id
	week_day
LK	month_id
	month
LK	quarter_id
	quarter
LK	semester_id
	semester
LK	year

Agg_Sales	
FK1	date_id
FK2	order_location_id
FK3	customer_location_id
FK4	customer_id
FK5	product_id
	Sales_per_customer
	monthly_sales
	weekly_sales

Order_item	
FK1	product_id
FK2	order_time_id
	discount
	discount_rate
	product_price
	profit_ratio
	quantity
	sales
	order_item_total

Order star



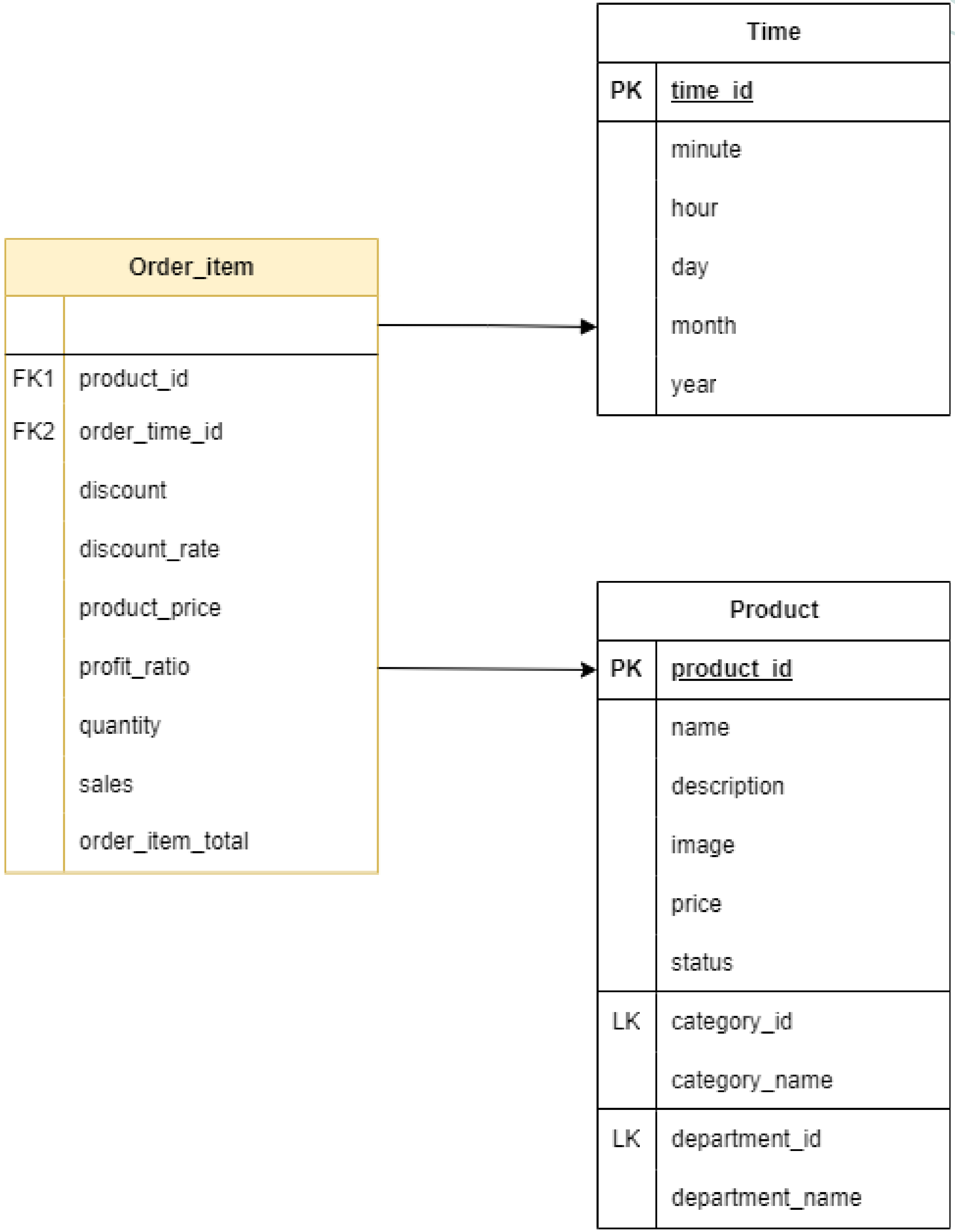
Customer	
PK	customer_id
	first_name
	last_name
	country
	city
	zip_code
	latitude
	longitude
LK	segment_id
	segment

Location	
PK	location_id
LK	city_id
	city
LK	state_id
	state
LK	country_id
	country
LK	region_id
	region
LK	market_id
	market

Order	
FK1	customer_id
FK2	customer_location_id
FK3	order_location_id
FK4	shipping_date
FK5	order_date
	payment_type
	days_shipping_real
	days_shipping_scheduled
	delivery_status
	late_delivery_risk
	profit_per_order
	order_status
	shipping_mode
	number_of_items
	total_discount
	total_price

Time	
PK	time_id
	minute
	hour
	day
	month
	year

Order item star

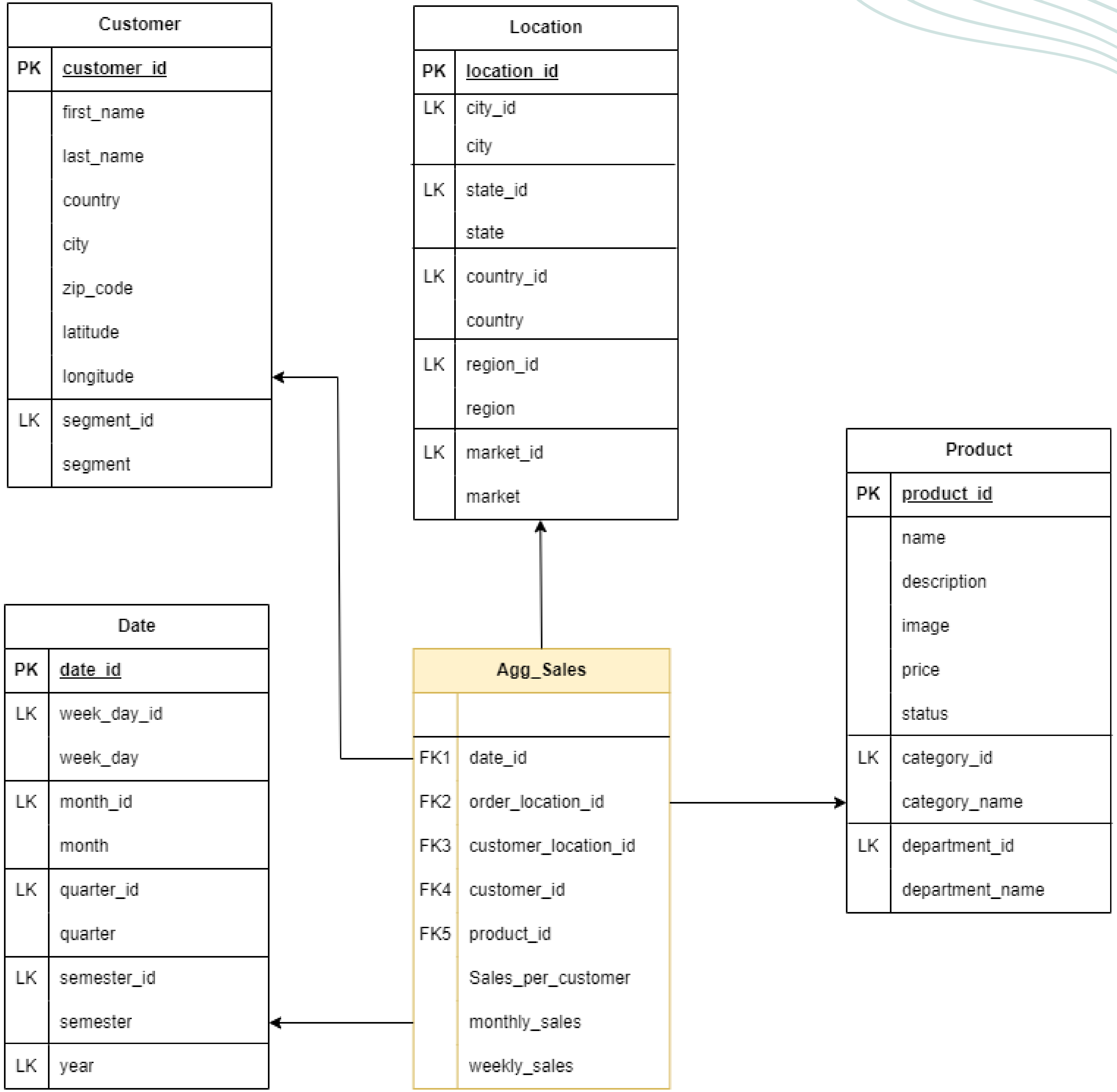


Order_item	
FK1	product_id
FK2	order_time_id
	discount
	discount_rate
	product_price
	profit_ratio
	quantity
	sales
	order_item_total

Time	
PK	time_id
	minute
	hour
	day
	month
	year

Product	
PK	product_id
	name
	description
	image
	price
	status
LK	category_id
	category_name
LK	department_id
	department_name

Sales Aggregation



Customer	
PK	customer_id
	first_name
	last_name
	country
	city
	zip_code
	latitude
	longitude
LK	segment_id
	segment

Location	
PK	location_id
LK	city_id
	city
LK	state_id
	state
LK	country_id
	country
LK	region_id
	region
LK	market_id
	market

Product	
PK	product_id
	name
	description
	image
	price
	status
LK	category_id
	category_name
LK	department_id
	department_name

Date	
PK	date_id
LK	week_day_id
	week_day
LK	month_id
	month
LK	quarter_id
	quarter
LK	semester_id
	semester
LK	year

Agg_Sales	
FK1	date_id
FK2	order_location_id
FK3	customer_location_id
FK4	customer_id
FK5	product_id
	Sales_per_customer
	monthly_sales
	weekly_sales

DM Implementation + ETL

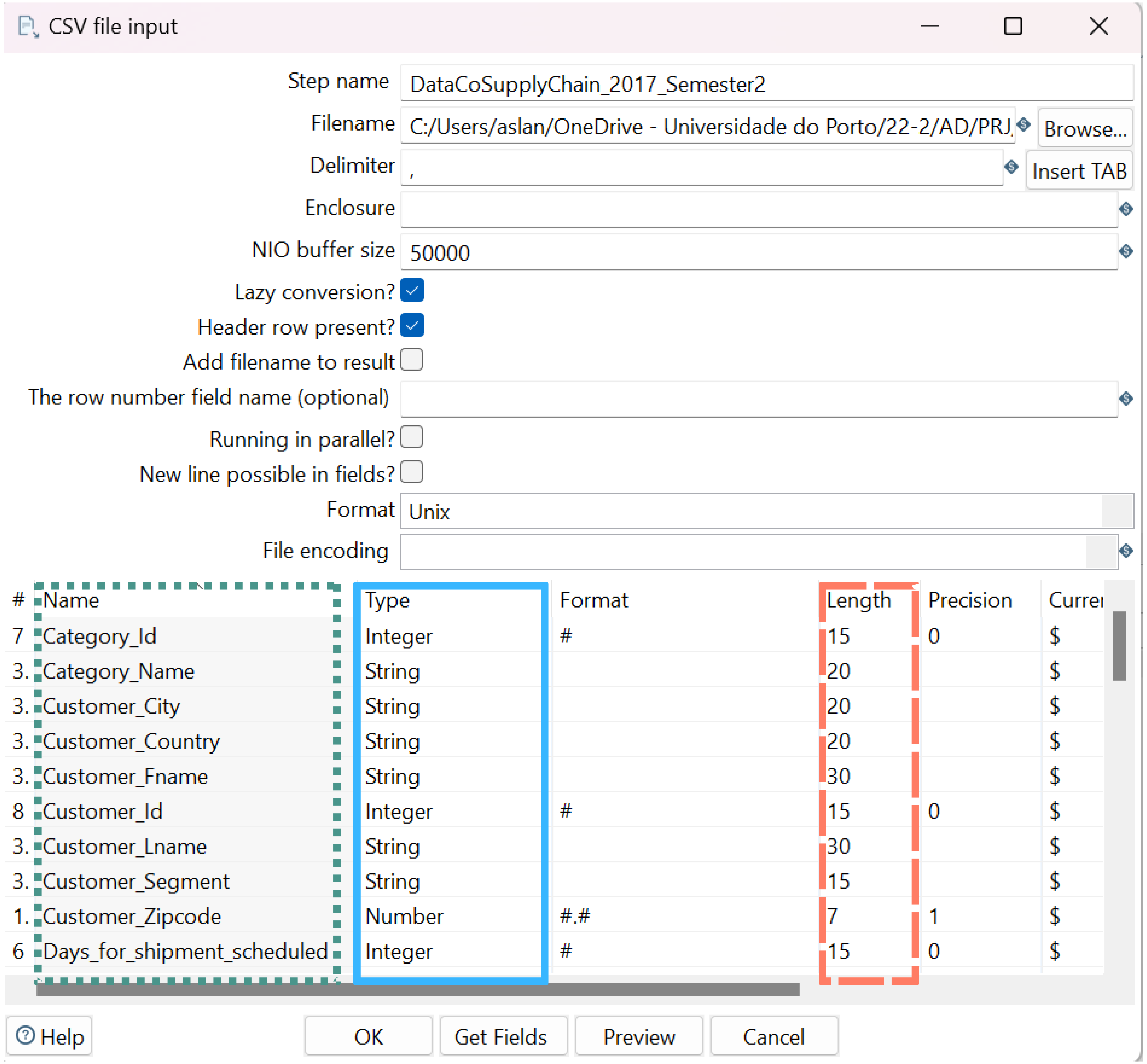
Implement the dimensional model in an appropriate database system

Implementation

- Communicate with database
- Creating dimension and fact tables
- ETL process
  - Extract
  - Transform (calculate fields and process data)
  - Load

Input file

Manual correction of the data fields



CSV file input

Step name: DataCoSupplyChain_2017_Semester2

Filename: C:/Users/aslan/OneDrive - Universidade do Porto/22-2/AD/PRJ.

Delimiter: ,

Enclosure:

NIO buffer size: 50000

Lazy conversion? ☑

Header row present? ☑

Add filename to result □

The row number field name (optional):

Running in parallel? □

New line possible in fields? □

Format: Unix

File encoding:

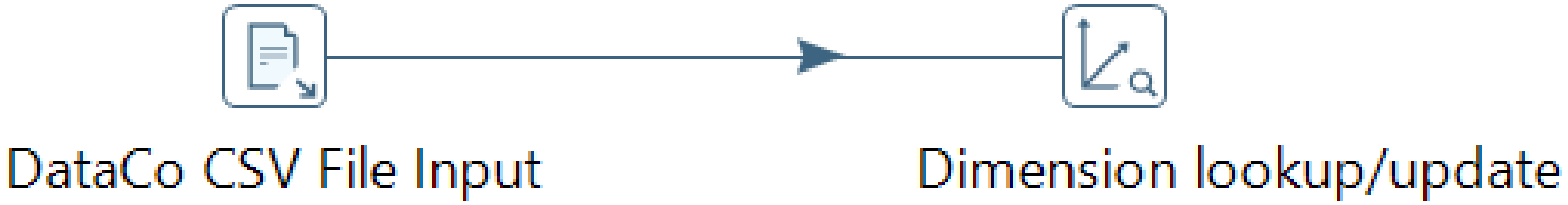
#	Name	Type	Format	Length	Precision	Curre
7	Category_Id	Integer	#	15	0	$
3.	Category_Name	String		20		$
3.	Customer_City	String		20		$
3.	Customer_Country	String		20		$
3.	Customer_Fname	String		30		$
8	Customer_Id	Integer	#	15	0	$
3.	Customer_Lname	String		30		$
3.	Customer_Segment	String		15		$
1.	Customer_Zipcode	Number	#.#	7	1	$
6	Days_for_shipment_scheduled	Integer	#	15	0	$

Load

Slowly changing dimensions - allows easy future updates

Dimensions made using data from CSV file:

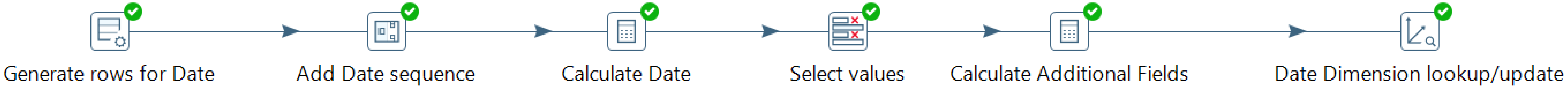
- Customer_dim
- Product_dim
- Location_dim

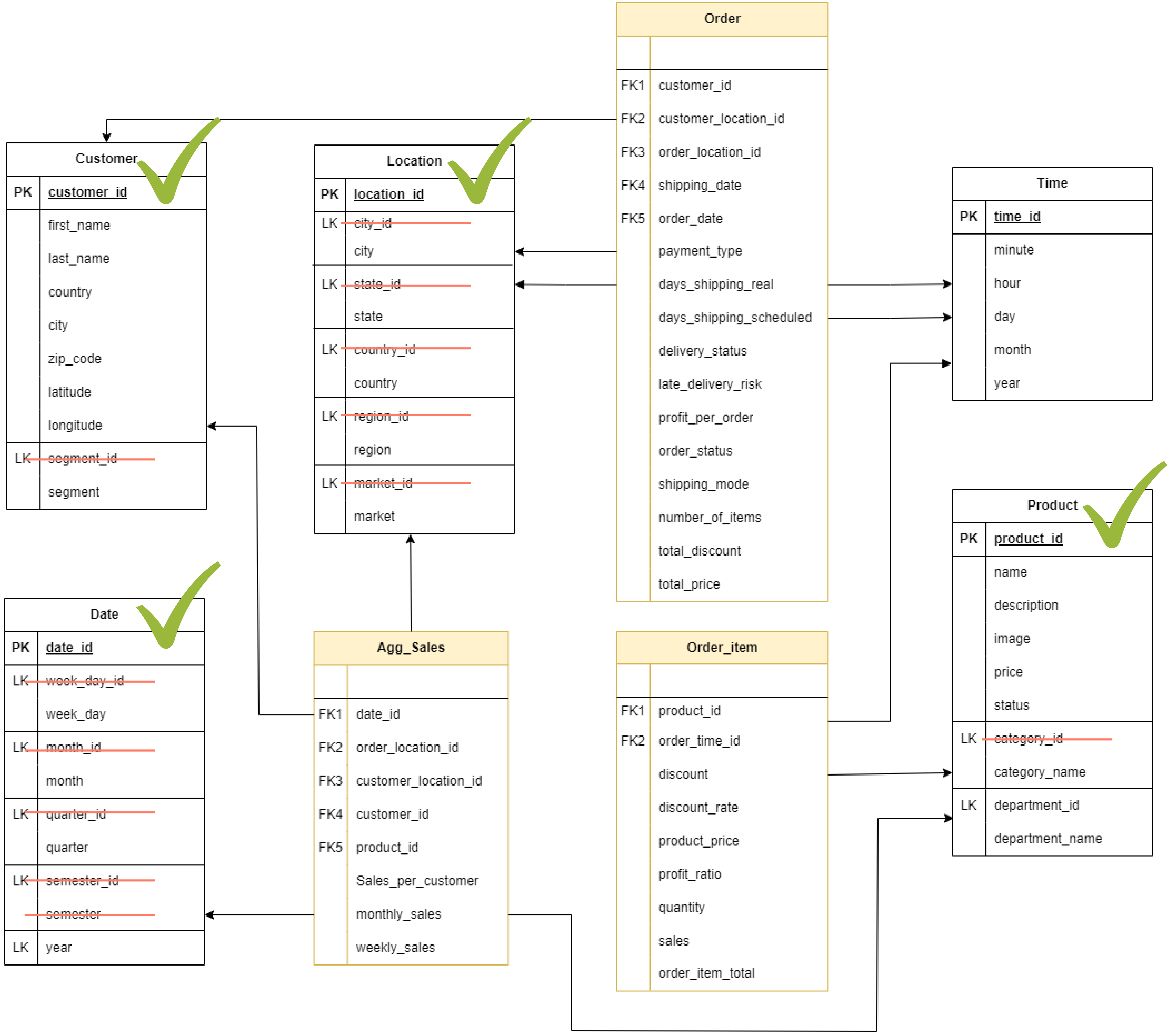


Load

Dimensions rows generated in Pentaho:

- Date_dim





	Customer
PK	customer_id
	first_name
	last_name
	country
	city
	zip_code
	latitude
	longitude
LK	~~segment_id~~
	segment

	Location
PK	location_id
LK	~~city_id~~
	city
LK	~~state_id~~
	state
LK	~~country_id~~
	country
LK	~~region_id~~
	region
LK	~~market_id~~
	market

	Order
FK1	customer_id
FK2	customer_location_id
FK3	order_location_id
FK4	shipping_date
FK5	order_date
	payment_type
	days_shipping_real
	days_shipping_scheduled
	delivery_status
	late_delivery_risk
	profit_per_order
	order_status
	shipping_mode
	number_of_items
	total_discount
	total_price

	Time
PK	time_id
	minute
	hour
	day
	month
	year

	Product
PK	product_id
	name
	description
	image
	price
	status
LK	~~category_id~~
	category_name
LK	department_id
	department_name

	Date
PK	date_id
LK	~~week_day_id~~
	week_day
LK	~~month_id~~
	month
LK	~~quarter_id~~
	quarter
LK	~~semester_id~~
	~~semester~~
LK	year

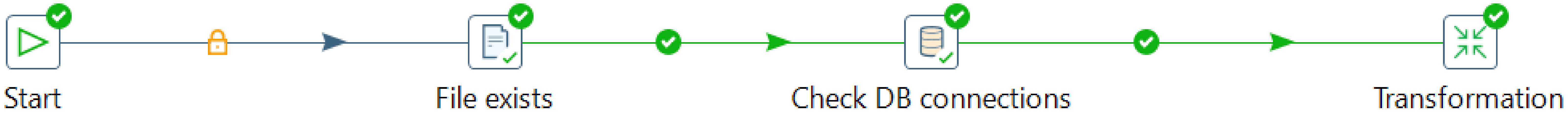
	Agg_Sales
FK1	date_id
FK2	order_location_id
FK3	customer_location_id
FK4	customer_id
FK5	product_id
	Sales_per_customer
	monthly_sales
	weekly_sales

	Order_item
FK1	product_id
FK2	order_time_id
	discount
	discount_rate
	product_price
	profit_ratio
	quantity
	sales
	order_item_total

Done!

Job scheduling

Schedule a set of ETL jobs to perform incremental data loads as new data is added to the operational (OLTP) systems and/or as operational data is changed.



Querying & Data Analysis

Develop business analytics reports, dashboards, or other user interfaces for the data warehouse.

Querying

PostgreSQL

- Cubes
- Rollups

Data Analytics

Power BI

- Data Insights
- Visualization
- Transform & Clean data

Conclusions

Conclusions

Datawarehouse VS Operational system

- Scalability
- Data Integration
- Performance: complex analytical queries
- Historical Analysis

Future works

- Finish the ETL process
- Querying and Analytics
- Use data to feed models and make forecasting systems to predict supply chain management

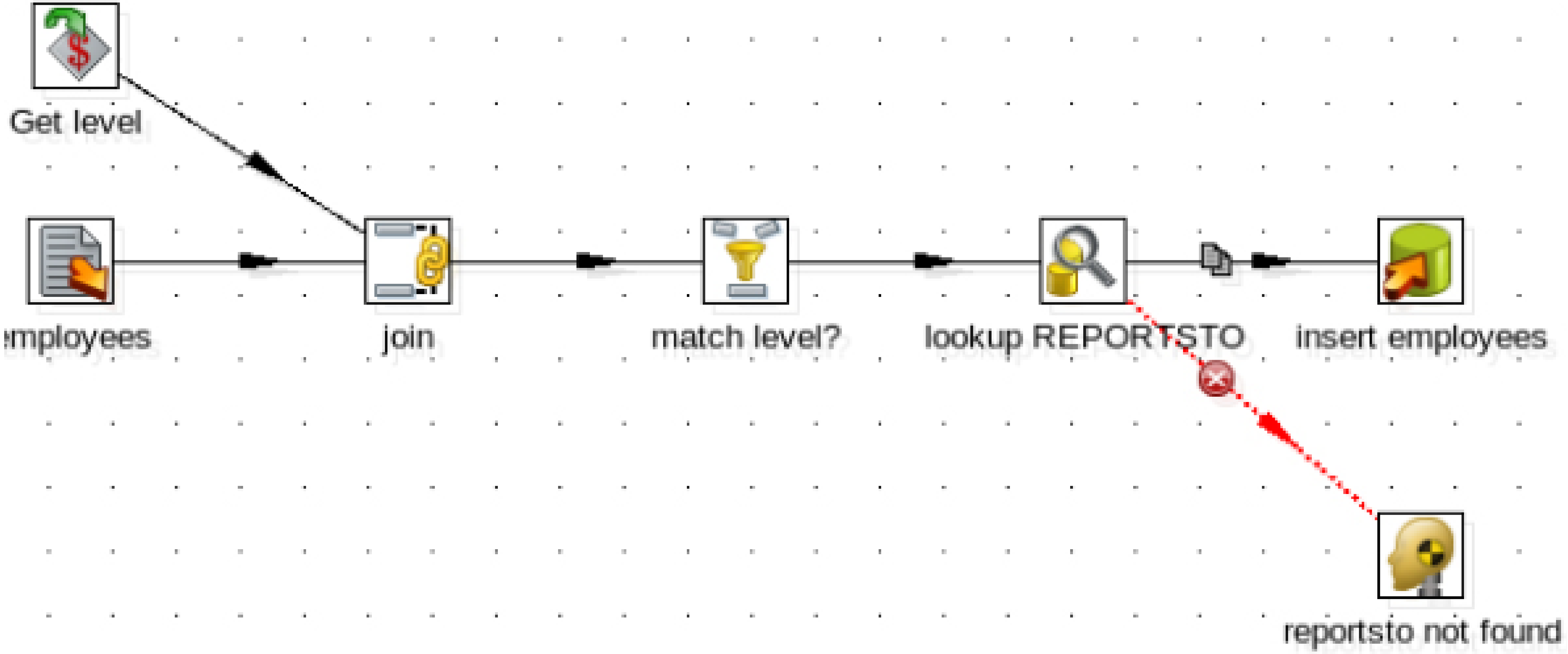
DataCo Supply Chain Data Warehousing

FEUP - MECD - Data Warehouse
Middle presentation

Carlos Miguel Veloso
Cátia Teixeira
Luís Henriques
Rojan Aslani

Questions?

LEVELS


Get level
employees
join
match level?
lookup REPORTSTO
insert employees
reportsto not found