



INFO 7290: Data Warehousing & Business Intelligence

BI & Data Integration

Final Team Project



Northeastern University

Rick Sherman
Athena IT Solutions
ri.sherman@northeastern.edu

BI Team Project:

Requirements

- Contoso_Retail_BI_Test is a backup of the SQL Server schema with sample data to use for BI development
- Tasks:
 - Restore above database
 - Use above database for BI development & then switch to Contoso_Retail_BI when you have it loaded
 - Create Contoso_Retail_BI from Contoso_Retail_BI_Test

BI Team Project:

Requirements

Deliverables:

1. BI -- Develop BI Dashboards, Reports & Visualizations Sales (Channel) analysis
 - Online sales analysis
 - Sales Channel Analysis
 - Inventory analysis
 - ~~Sales quota analysis~~
2. DI -- Load data sources into DW
 - SOR: flat files, SQL Server, ~~Oracle~~, PostgreSQL & MySQL
 - DW: Microsoft Contoso Retail BI dataset (customized version) – SQL Server

BI Team Project:

Deliverables - BI

BI Tools:

- Tableau
- Qlik
 - Qlik Sense
 - QlikView (Optional)
- Microsoft BI
 - PowerBI
 - PowerPivot, PowerView, PowerMap, Excel 2103 (Optional)



Deliverables:

- Dashboards with visualization for each analysis & for each BI tool
- Comparison of each tool – key differences – strengths & weaknesses

Data Integration Tools:

- Talend Enterprise Data Integrator



Deliverables:

- Load DW from data sources
 - Document all jobs
 - Provide load statistics
 - Provide analysis of load jobs using one of BI tools used in this project
- Handle data quality & error messages
 - Document error handling results
- Load rejection records
 - Track reasons for rejections
 - Provide analysis of rejections & reasons using one of BI tools used in this project
- Note:
 - Follow project standards

BI Team Project:

Requirements - Contoso – Fictional Retail Company

TableName	Row Count
DimAccount	24
DimChannel	4
DimCurrency	28
DimCustomer	18,869
DimDate	3,652
DimEmployee	293
DimEntity	421
DimGeography	674
DimProduct	2,517
DimProductCost	2,517
DimProductPrice	2,517
DimPromotion	28
DimSalesTerritory	265
DimScenario	3
DimStore	306

32,118

TableName	Row Count
FactExchangeRate	773
FactInventory	8,013,099
FactOnlineSales	12,627,608
FactSales	3,334,098
FactSalesQuota	7,465,911
	31,441,489

Note: The final dataset may have slight schema changes and row counts may vary

Dashboards - Tabular Reports, Visualizations or Maps

Deliverables – Analysis

- Data Subjects:
 - Online Sales Analysis (FactOnlineSales)
 - Sales Analysis (FactSales)
 - Inventory Analysis (FactInventory)
- Types of analysis:
 - Trending
 - Ranking
 - Comparison
 - Period over Period
 - Geo Map
 - Contribution
- Measures:
 - Sales \$, Profit, Profit Margin, Avg Order Size,...
- Dimensions:
 - Customers: Company & Person, demographics
 - Product: Product Hierarchy (Category, Subcategory, Product), Brand, other attributes
 - Store: Type, other attributes
 - Dates
 - Geography

Dashboards - Tabular Reports, Visualizations or Maps

Deliverables – Sample Questions

A. (Overall) Sales Analysis

1. Sales & profit by channel and time (Year/Qtr/Month)
2. Rank sales
 - a) Product Category & Subcategory
 - b) Country & State
 - c) Stores
3. Geo Sales Analysis
4. Contribution analysis – sales & profit
 - a) Product

Dashboards - Tabular Reports, Visualizations or Maps

Deliverables – Sample Questions

B. Online Sales Analysis

1. Sales & profit by customer demographics such as education, income, etc.
2. Sales & profit with Period over Period analysis
3. Top 20 customers by sales & profit
4. Sales Analysis – Geo analysis
5. Provide contribution analysis

C. Inventory Analysis

1. Inventory Costs by channel and time
2. Rank inventory costs
 - a) Product Category & Subcategory
 - b) Country & State
 - c) Stores
3. Geo Sales Analysis
4. Contribution analysis – inventory cost
 - a) Product
 - b) Store

BI Team Project:

Systems of Record (SOR)

- Use existing schema from Contoso_Retail_BI as your target DW schema. Truncate existing data on your data integration jobs.
- Data is being sourced from 3 geographic area (continents) databases:
 - Contoso_Retail_SOR_NorthAmerica - Microsoft SQL Server
 - Contoso_Retail_SOR_Europe - MySQL
 - Contoso_Retail_SOR_Asia – PostgreSQL
- In addition many tables are sourced from various files in Excel, csv or text delimited file format

BI Team Project:

Systems of Record (SOR) 1 of 2 - 8/1/2016

DB_Name	Table_Name	Table_Rows
Contoso_Retail_SOR_NorthAmerica	DimCustomer_Company	276
Contoso_Retail_SOR_NorthAmerica	DimCustomer_Person	9,395
Contoso_Retail_SOR_NorthAmerica	DimGeography	674
Contoso_Retail_SOR_NorthAmerica	DimProduct	2,517
Contoso_Retail_SOR_NorthAmerica	DimProductCategory	8
Contoso_Retail_SOR_NorthAmerica	DimProductSubcategory	44
Contoso_Retail_SOR_NorthAmerica	DimPromotion_NA	10
Contoso_Retail_SOR_NorthAmerica	DimStore_NA	209
Contoso_Retail_SOR_NorthAmerica	FactCatalogSales_NA	194,976
Contoso_Retail_SOR_NorthAmerica	FactInventory	5,668,381
Contoso_Retail_SOR_NorthAmerica	FactOnlineSalesOrderDetail_NA	4,645,792
Contoso_Retail_SOR_NorthAmerica	FactOnlineSalesOrderHeader_NA	686,811
Contoso_Retail_SOR_NorthAmerica	FactResellerSales_NA	157,460
Contoso_Retail_SOR_NorthAmerica	FactStoreSales_NA	1,467,942

DB_Name	Table_Name	Table_Rows
Contoso_Retail_SOR_Asia	DimCustomer_Company	67
Contoso_Retail_SOR_Asia	DimCustomer_Person	3,593
Contoso_Retail_SOR_Asia	DimGeography	674
Contoso_Retail_SOR_Asia	DimProduct	2,517
Contoso_Retail_SOR_Asia	DimProductCategory	8
Contoso_Retail_SOR_Asia	DimProductSubcategory	44
Contoso_Retail_SOR_Asia	DimPromotion_AS	10
Contoso_Retail_SOR_Asia	DimStore_AS	41
Contoso_Retail_SOR_Asia	FactInventory	1,628,104
Contoso_Retail_SOR_Asia	FactOnlineSalesOrderDetail_AS	4,134,535
Contoso_Retail_SOR_Asia	FactOnlineSalesOrderHeader_AS	337,598
Contoso_Retail_SOR_Asia	FactResellerSales_AS	151,194
Contoso_Retail_SOR_Asia	FactSalesQuota_AS	467,871
Contoso_Retail_SOR_Asia	FactStoreSales_AS	473,738

DB_Name	Table_Name	Table_Rows
Contoso_Retail_SOR_Europe	DimCustomer_Company	43
Contoso_Retail_SOR_Europe	DimCustomer_Person	5,505
Contoso_Retail_SOR_Europe	DimGeography	674
Contoso_Retail_SOR_Europe	DimProduct	2,517
Contoso_Retail_SOR_Europe	DimProductCategory	8
Contoso_Retail_SOR_Europe	DimProductSubcategory	44
Contoso_Retail_SOR_Europe	DimPromotion_EU	19
Contoso_Retail_SOR_Europe	DimStore_EU	56
Contoso_Retail_SOR_Europe	FactInventory	1,918,225
Contoso_Retail_SOR_Europe	FactOnlineSalesOrderDetail_EU	3,847,281
Contoso_Retail_SOR_Europe	FactOnlineSalesOrderHeader_EU	651,952
Contoso_Retail_SOR_Europe	FactResellerSales_EU	153,579
Contoso_Retail_SOR_Europe	FactSalesQuota_EU	483,284
Contoso_Retail_SOR_Europe	FactStoreSales_EU	487,110

DB_Name	Table_Name	Table_Rows
Contoso_Retail_SOR_Reference	DimCustomer_Company_Crossmap	385
Contoso_Retail_SOR_Reference	DimCustomer_Person_Crossmap	18,484
Contoso_Retail_SOR_Reference	DimDate	3,652
Contoso_Retail_SOR_Reference	DimGeography	674
Contoso_Retail_SOR_Reference	DimProduct_CrossMap	2,517
Contoso_Retail_SOR_Reference	DimPromotion_Crossmap	28
Contoso_Retail_SOR_Reference	DimStore_Channel_Crossmap	306

Note: These will be revised

File Name
cost_cny_step_1_of_4.txt
cost_cny_step_2_of_4.txt
cost_cny_step_3_of_4.txt
cost_cny_step_4_of_4.txt
cost_eur_step_1_of_4.csv
cost_eur_step_2_of_4.csv
cost_eur_step_3_of_4.csv
cost_eur_step_4_of_4.csv
cost_usd_steps_all.xlsx
DimAccount.txt
DimChannel.csv
DimCurrency.csv
DimDate.csv
DimEmployees.csv
DimEntity.csv
DimGeography.csv
DimSalesTerritory.csv
DimScenario.txt
FactExchangeRate.xlsx
price_cny_step_1_of_4.txt
price_cny_step_2_of_4.txt
price_cny_step_3_of_4.txt
price_cny_step_4_of_4.txt
price_eur_step_1_of_4.csv
price_eur_step_2_of_4.csv
price_eur_step_3_of_4.csv
price_eur_step_4_of_4.csv
price_usd_steps_all.xlsx

Note: These will be revised

BI Team Project:

Systems of Record (SOR)

- There are 4 sales channels for this company:
 - Catalog
 - Retail
 - Stores
 - Online Sales
- In DW Sales are broken into:
 - FactSales – includes all 4 channels
 - FactOnlineSales – only includes Online Sales
- IN SOR sales are broken into 4 sales channels & 3 continents (North America, Europe & Asia):
 - Catalog – note: US-based only
 - Retail
 - Stores
 - Online Sales – further broken into Header & Detail (line) tables

BI Team Project:

Systems of Record (SOR)

- DW has all data in US dollars (USD)
- SORs have prices, costs & sales in “continent” currency
 - North America – USD
 - Europe - Euro
 - Asia - China Yuan
- Sales, Returns & Costs are in “constant” currency, i.e. recorded using published unit prices & costs
 - Daily currency exchange rate should be used in converting Euro & Yuan to USD

BI Team Project:

Systems of Record (SOR)

- Unit Price & Unit Cost should NOT stored in Fact Sales related tables nor in the DimProduct dimension
- Unit Prices & United Costs were independently changed 3 times during 2012-2014. You need to create SCD dimension for both Unit Price & Unit Cost Dimensions.
 - Step 1 – initial unit prices or costs
 - Step 2 – prices or costs revised
 - Step 3 – prices or costs revised
 - Step 4 – prices or costs revised

Cost_Step	Effective_Date
1	1/1/2012
2	10/1/2012
3	10/1/2013
4	10/1/2014

Pricing_Step	Effective_Date
1	1/1/2012
2	7/1/2012
3	7/1/2013
4	7/1/2014

~~Note: These will be revised~~

BI Team Project:

Error Handling

- Error Handling Standard will be to reject any rows that have incorrect FKs such as:
 - Product
 - Customer
 - Geography
 - Promotion
 - Store
- Fact tables should have a “rejects” table that contains the rows with errors and a error reason column

Suggestions on building model

Online Sales Example

Creating the data models to load is key activity

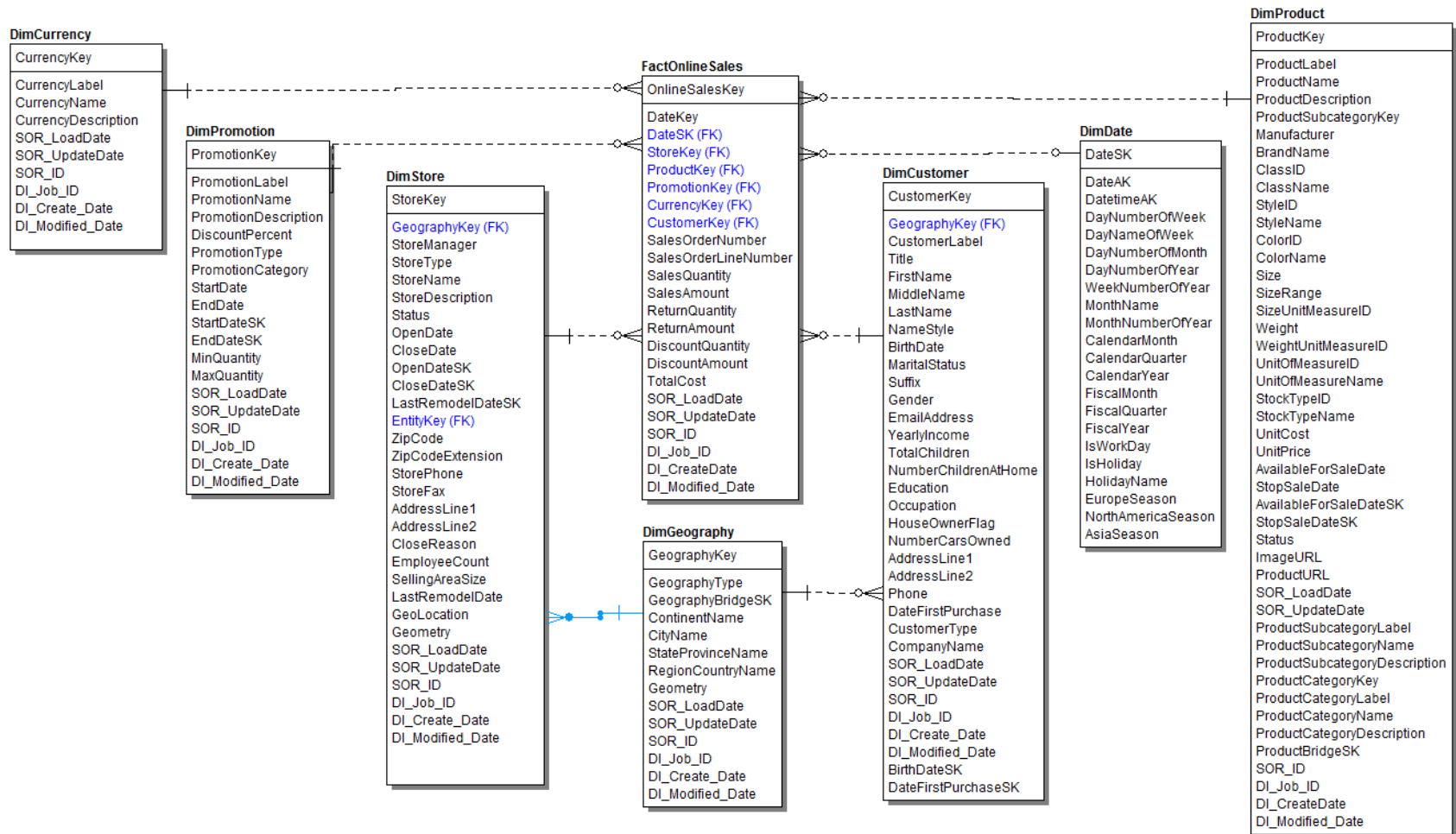
- Create views for all data queried or imported into BI Tools
 - Only include columns that will be used in analysis
 - Create role playing dimensions!
 - Rename columns that have generic names reused in more than one table but that does not mean the same thing in each of these columns
 - Avoid circular loops due to foreign keys (either in database or created by BI tool)
- **Create a separate BI application for each sub-model!!!**
- Loading (importing) data into BI Tool versus query is fastest IF you have the memory on your notebook.
 - You can create application one way and then copy it with the other setting to determine what is fastest for your notebook.

Creating the data models to load is key activity

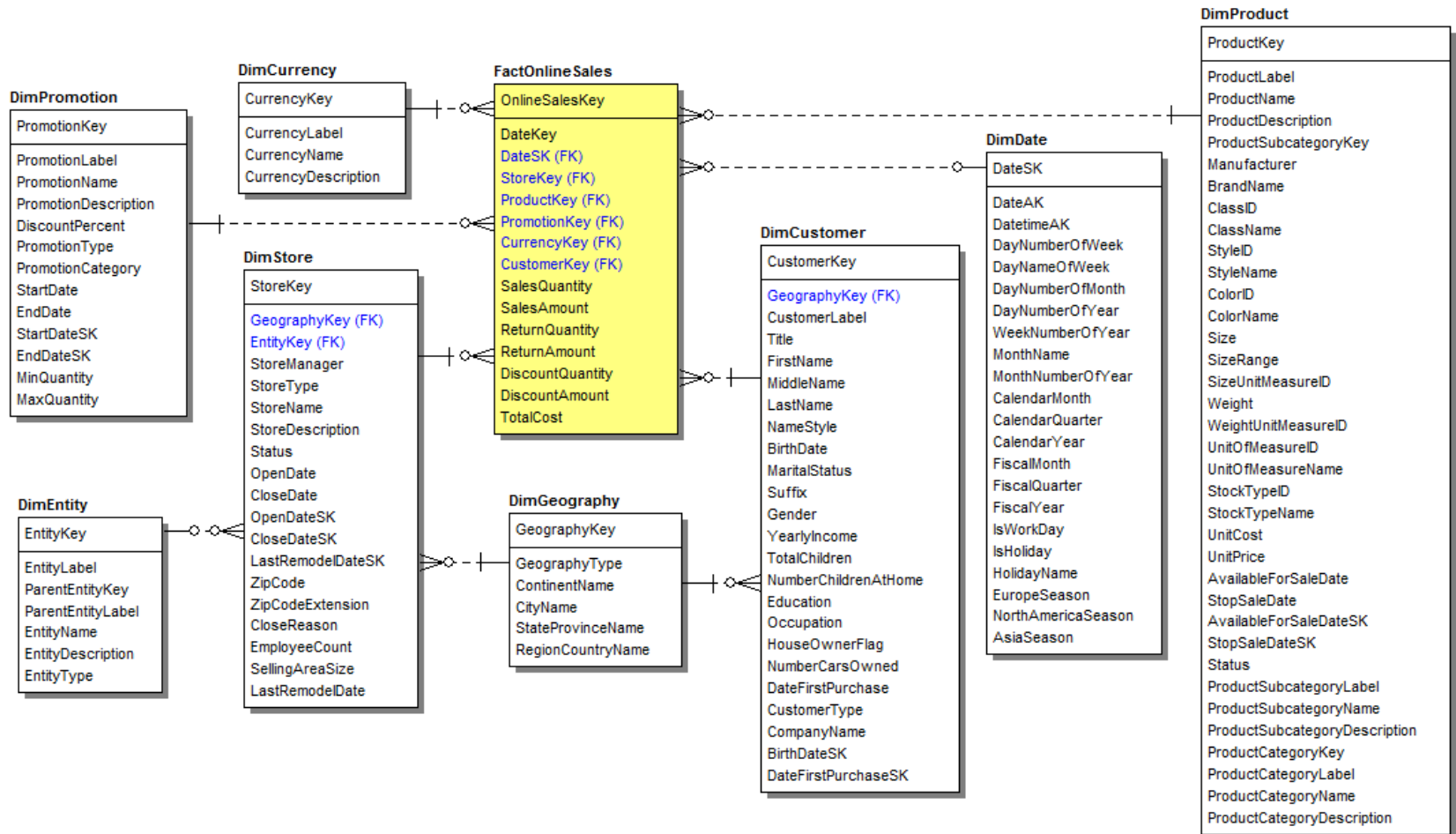
Notes:

- Microsoft products will use database keys to determine relationships while other tools use column names to create. If using views will need to specify all relationships in Microsoft tools but others will not.
- Microsoft products will automatically eliminate circular relationships by disabling one or more of the violating relationships. Great for load but may need to adjust which one used.
- Qlik is VERY sensitive to circular loops. QlikView will create synthetic keys that take LONG time to load & produce INCORRECT results. QlikSense will take a long time & then NOT load.

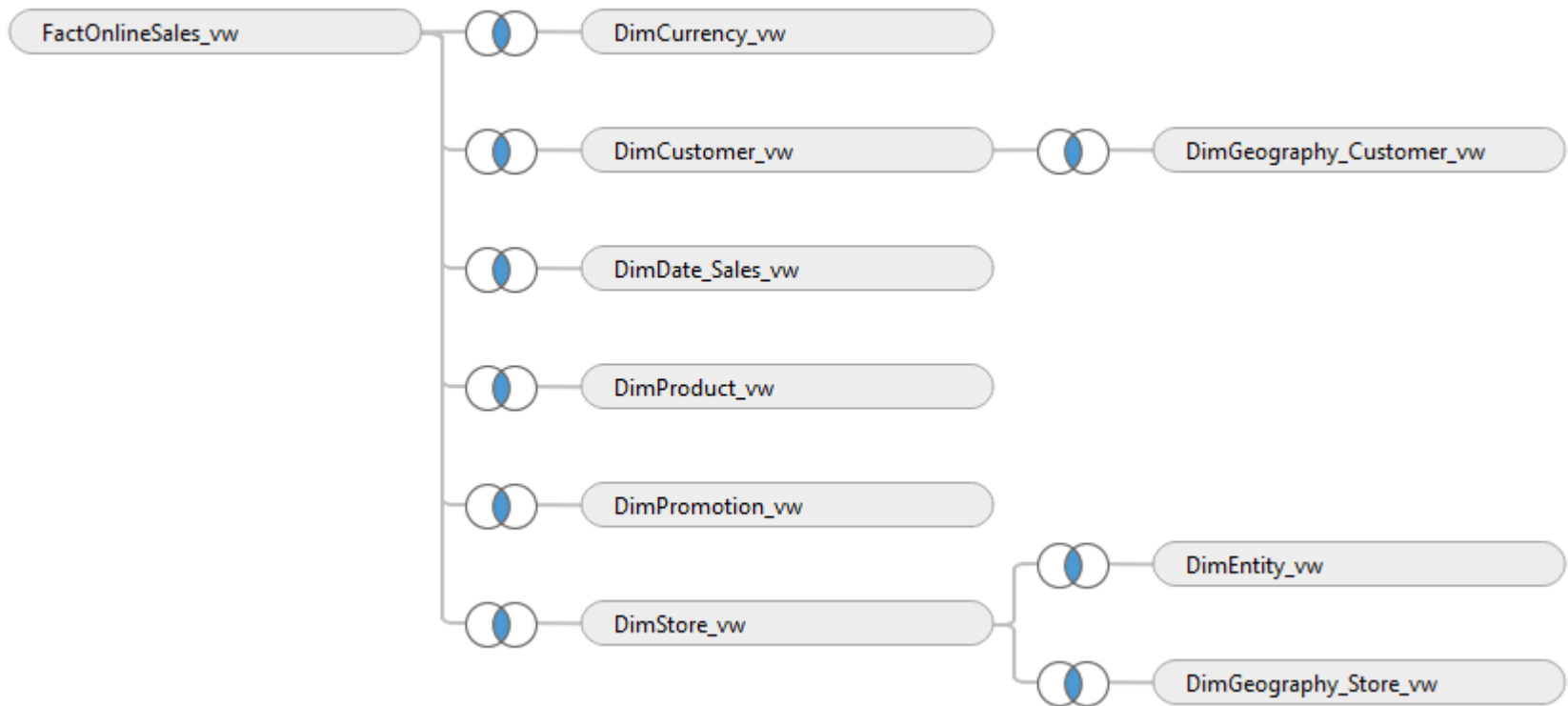
Online Sales – Sub-model (or Workspace)



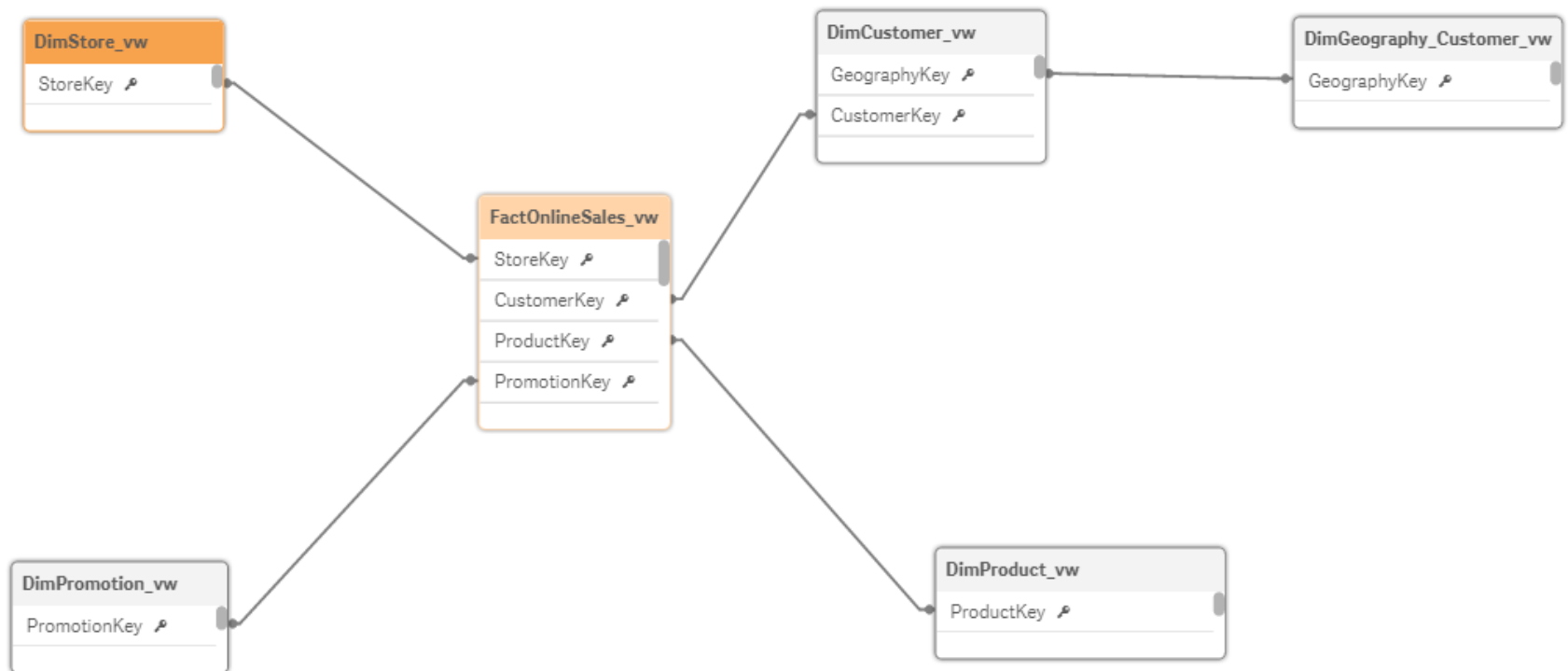
Draft – eliminating columns that will not be used in analysis



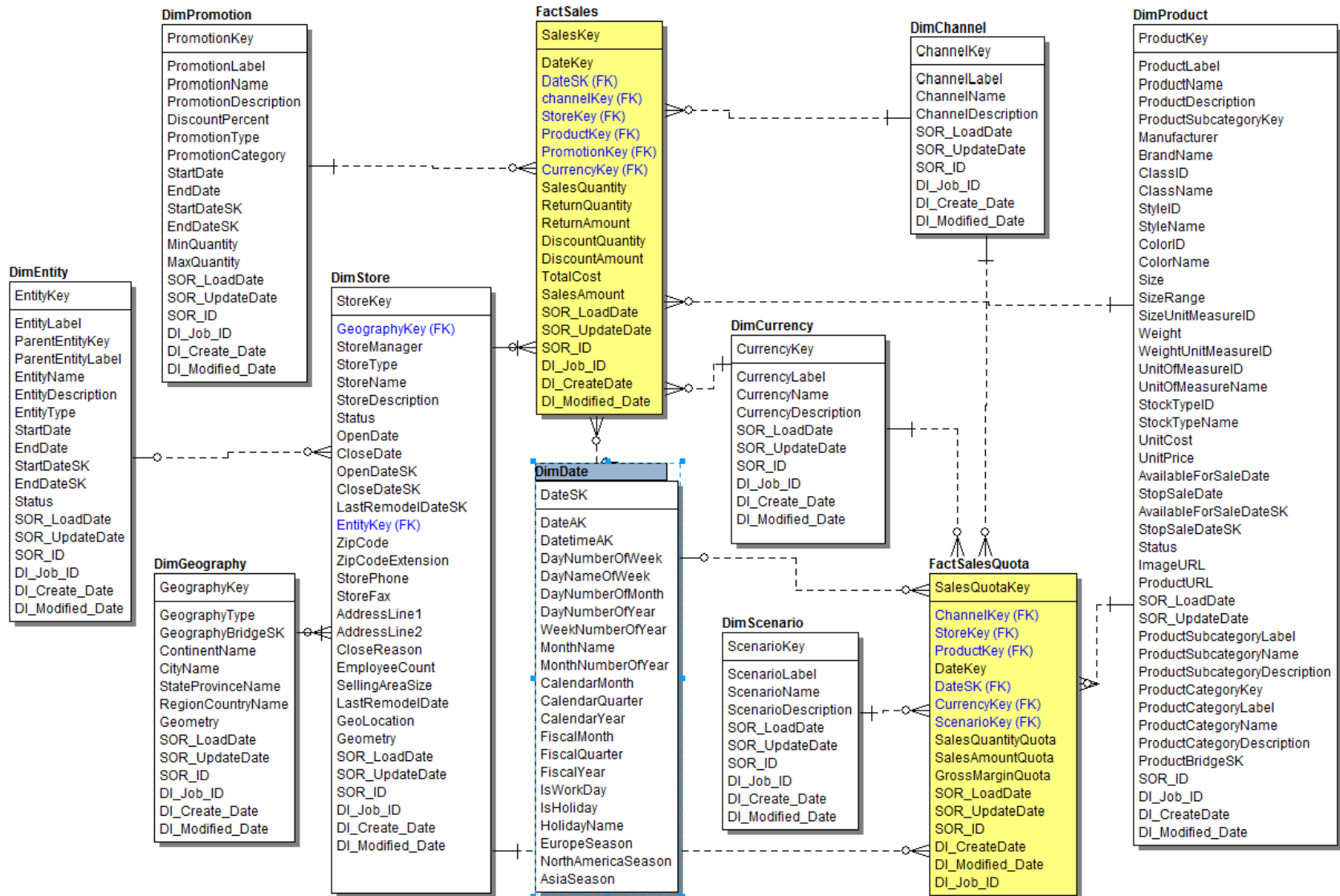
Sample using views



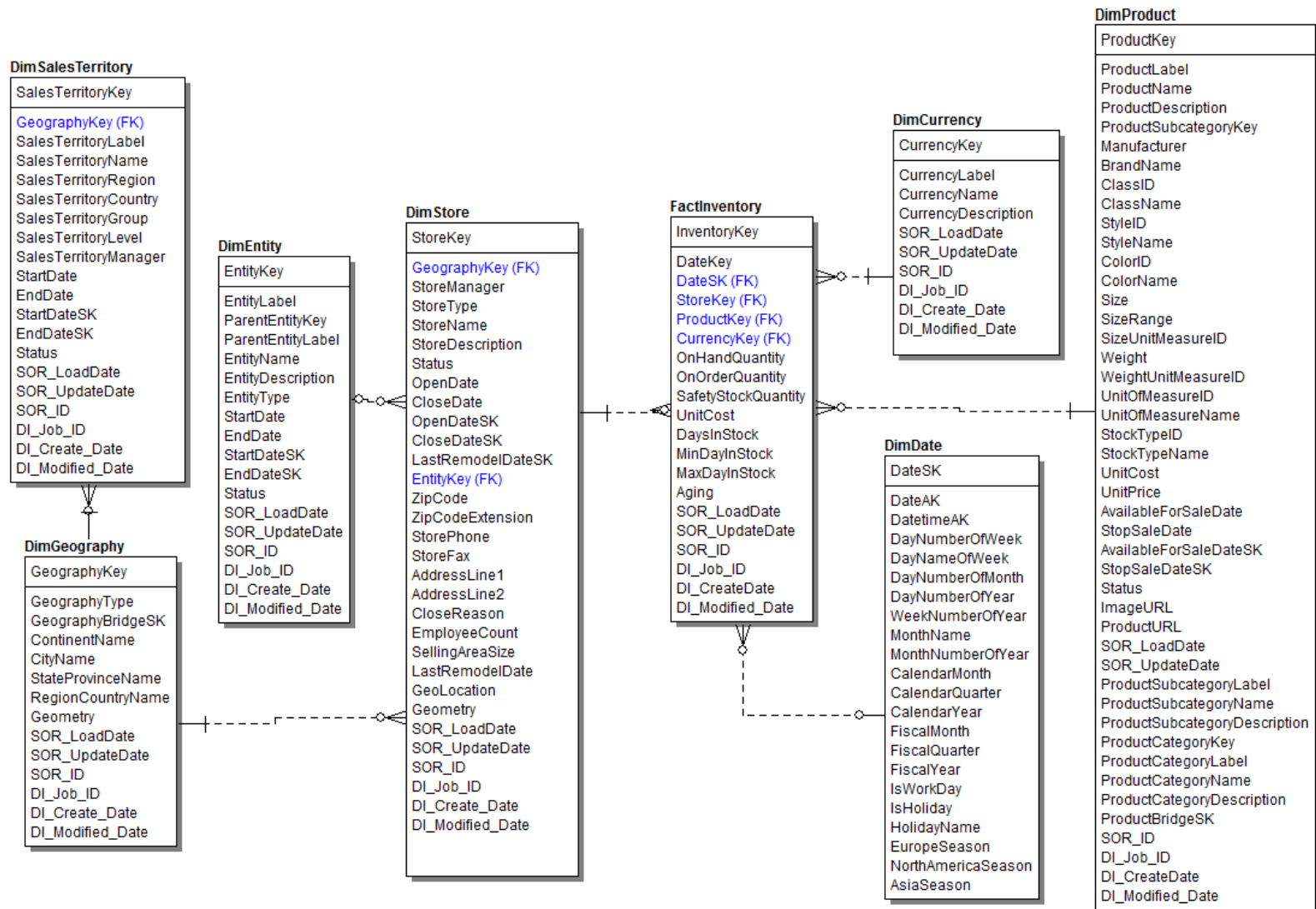
Sample using views



Overall Sales – Sub-model (or Workspace)



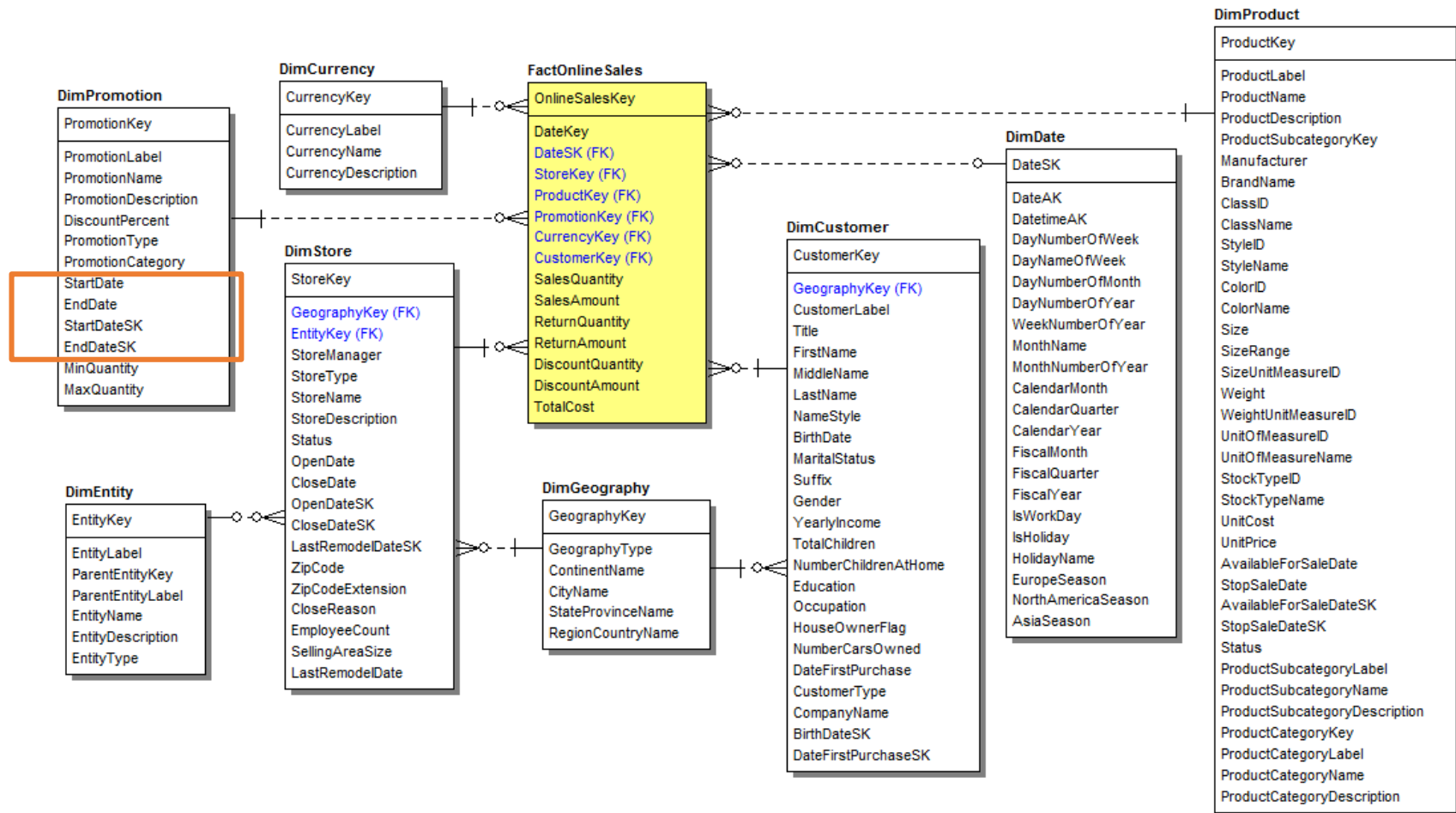
Inventory – Sub-model (or Workspace)



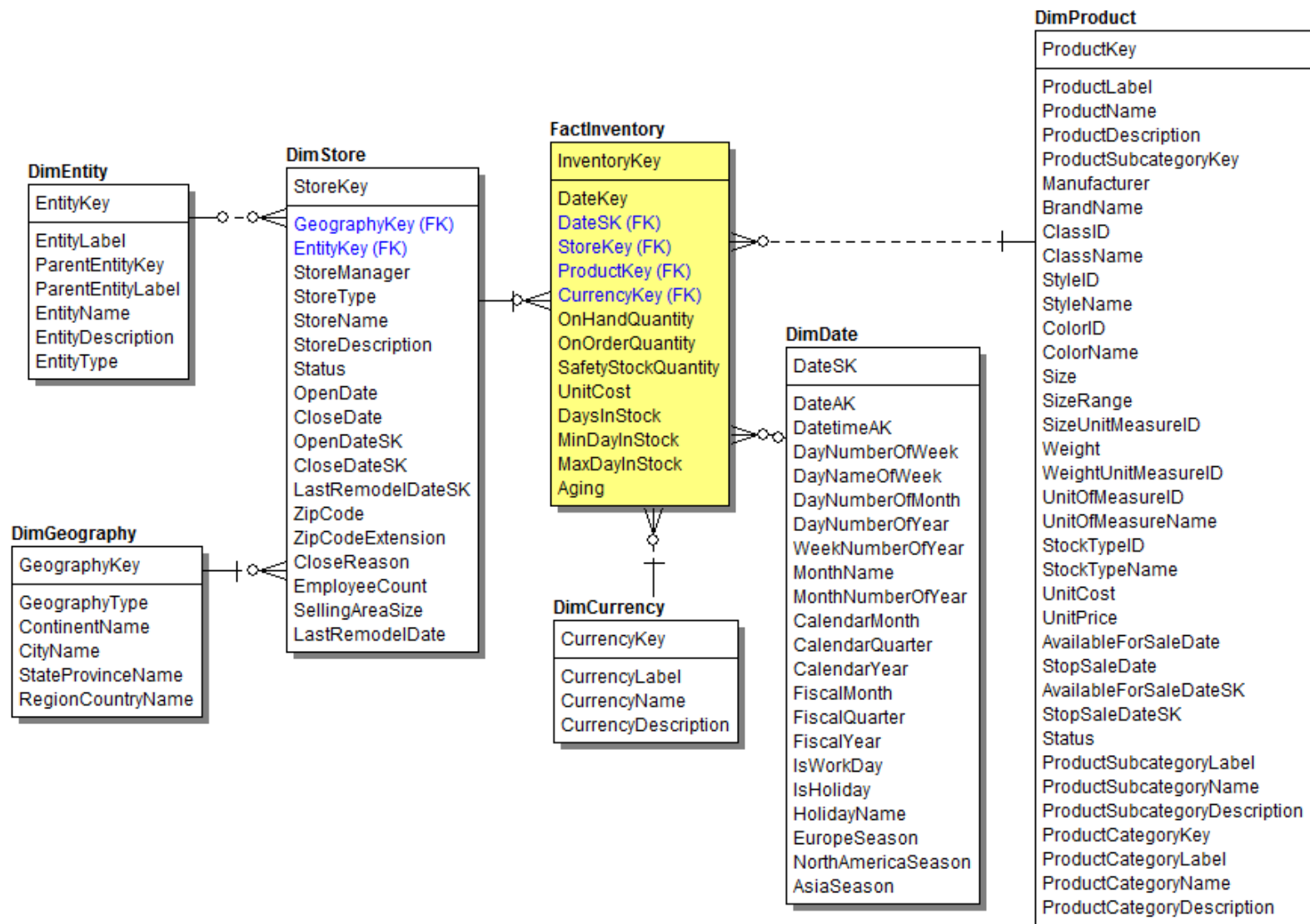
Views



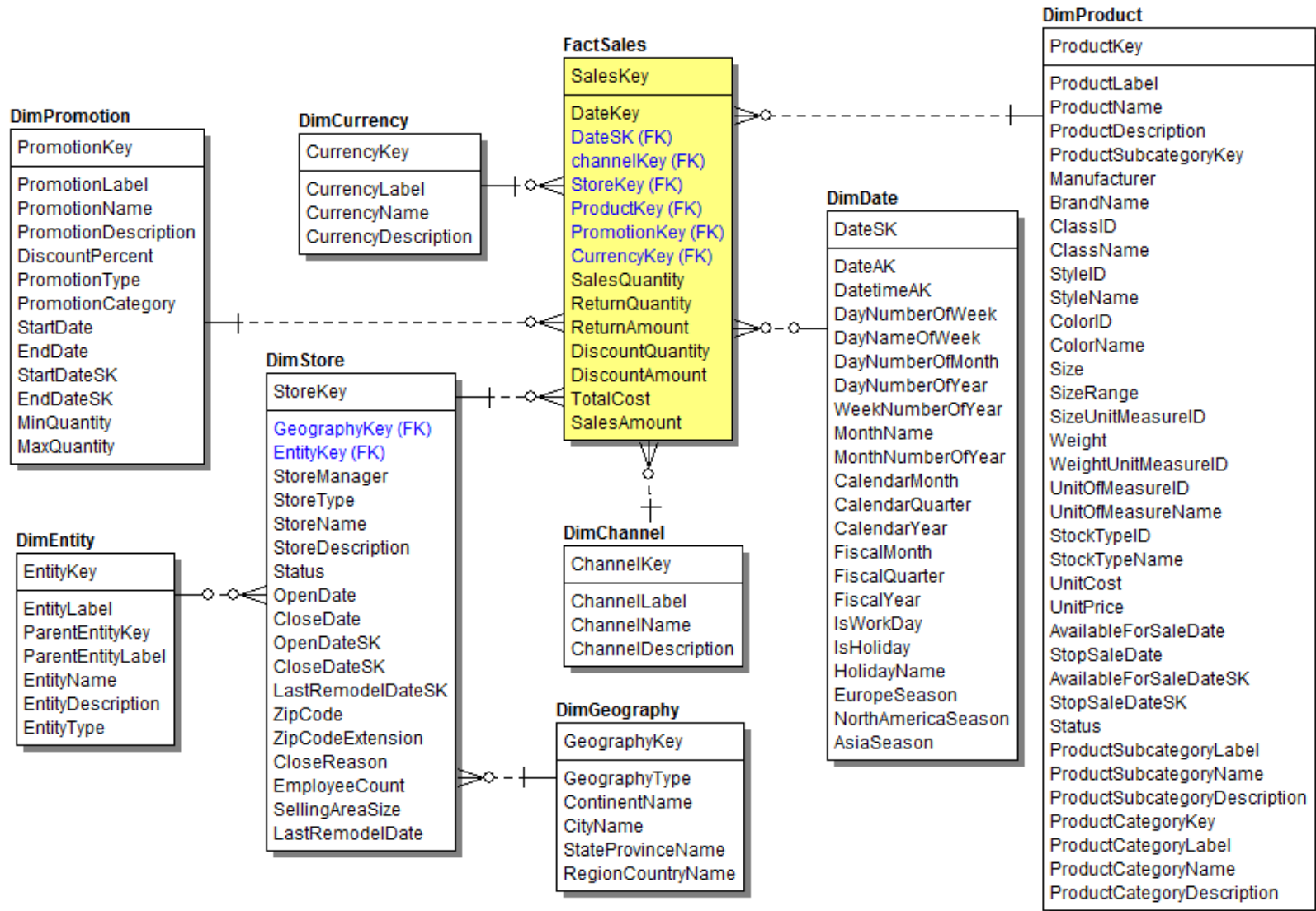
Online Sales



Inventory



Sales Quota



Sales Strategy

