Building Sales Datamart Using Pentaho Part 1

By Naheed Anjum Arafat

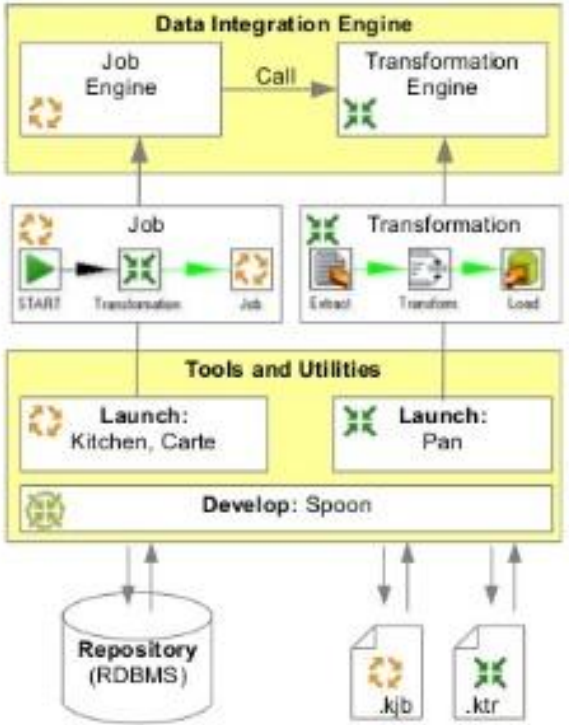
Outline

- Case Study background 5 mins
- Set up Installation 10 mins
- Hands on Task 1 30 mins
- Hands on Task 2 30 mins
- Tableau 15 mins

Pentaho Data Integration

- Created by Matt Casters
- ETL tool Spoon:- Part of Pentaho Suite
  - Mostly used component of PDI
- Open source
- Well documented - Books, Videos, Forums
- Powerful
- Flexible
- General purpose
- **Requires only Java**

Architecture of PDI - Kettle



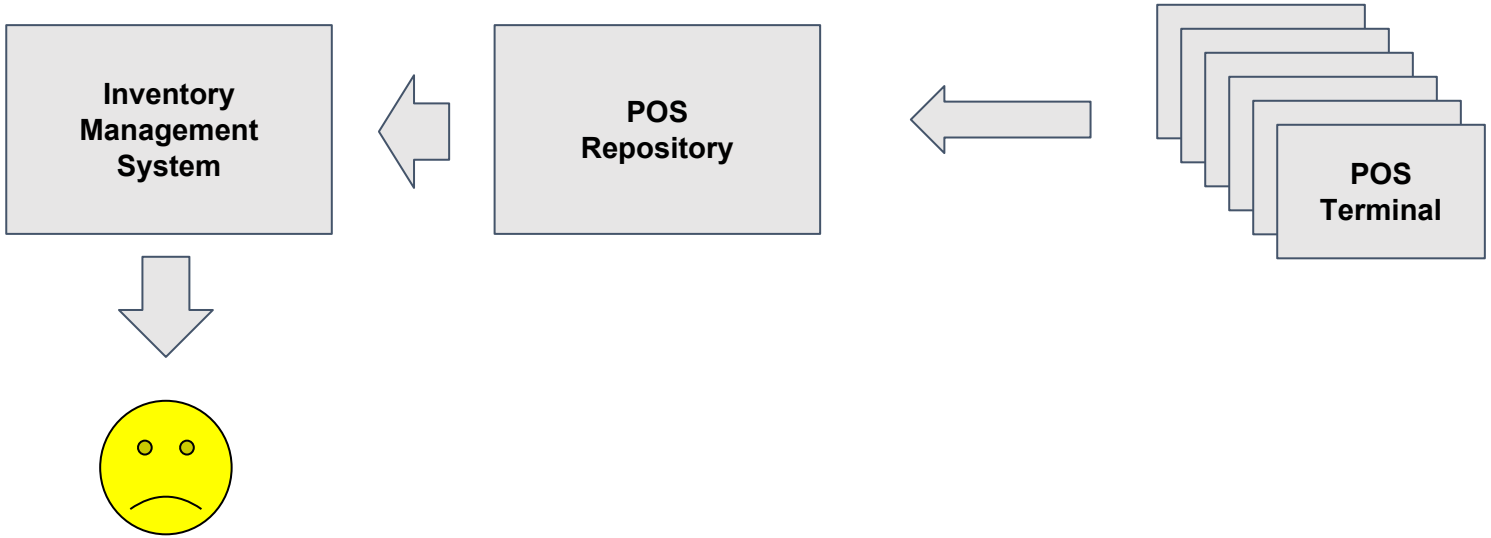
Source: Pentaho Corporation

Case Study:
Retail Sales

Background

- Retailer has a chain of stores
- POS data

Current System Overview



How Sales is Currently Processed

- POS terminal -> POS Repository:
  - Detailed sales rows
    - One row, Per transaction, per product

- POS Repository -> Inventory system
  - Aggregate by day
    - One row per product per day (Transaction data lost)

Issues

- POS Repository:-
  - Sales reports are slow
  - History is purged (limited to 1 year)
- Inventory system:-
  - Detailed sales data is grouped by **product for the day**
- Can not report on New KPI:-
  - Average Transaction Value (ATV)
    - Total sales per day / Total number of transactions per day
  - Average Unit Retail (AUR) [by stores/outlets]
    - Total sales per day / Total Quantity Sold
- Can not develop bonus scheme for sales person based on KPI

User Stories

As the Product Manager,

- I would like to have see the sales of product by category and subcategory (Drill down).
- I would like to compare sales by period.

User Stories

As a Store Manager,

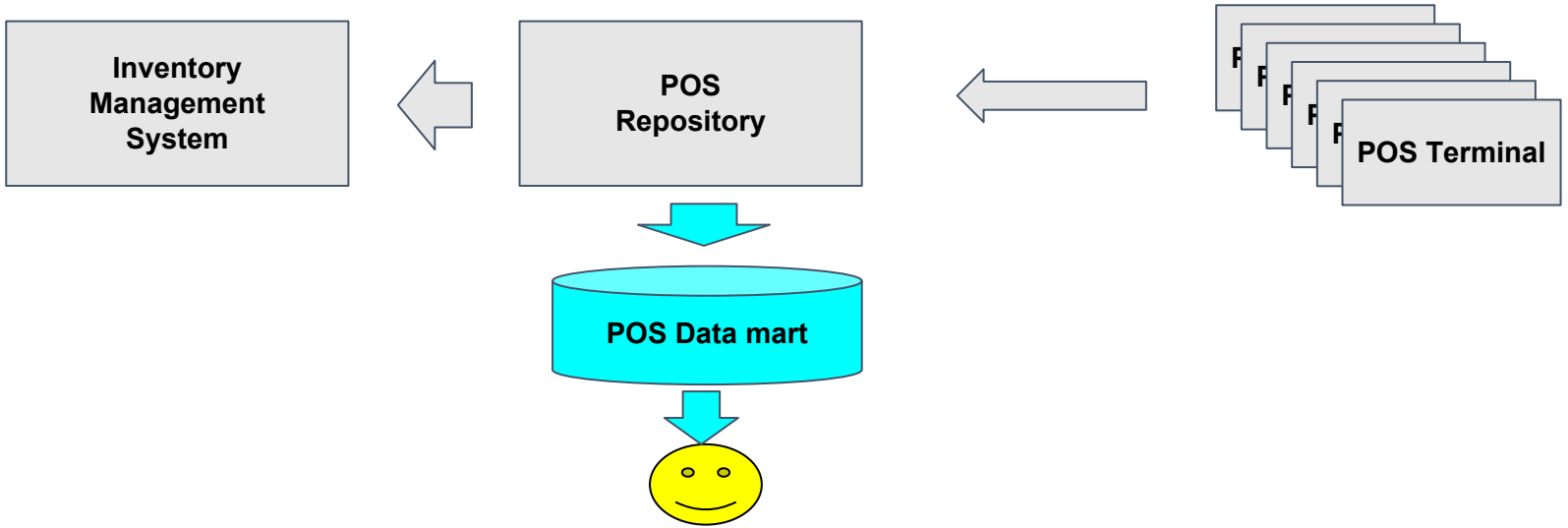
- I would like to have the KPI (ATV, AUR) of my **store** for the day, week, month, quarter, compared to prev, same period last year.
- I would like to be able to rank the top performer according to the KPI (AUR, ATV and Share of skincare) to give feedback to staff.

User Stories

As a store manager,

- I would like to have visibility of transactions by hour and how it changes by weekday and periods so that I can plan and schedule my staff according to when it is needed.

A Sales Data mart is needed



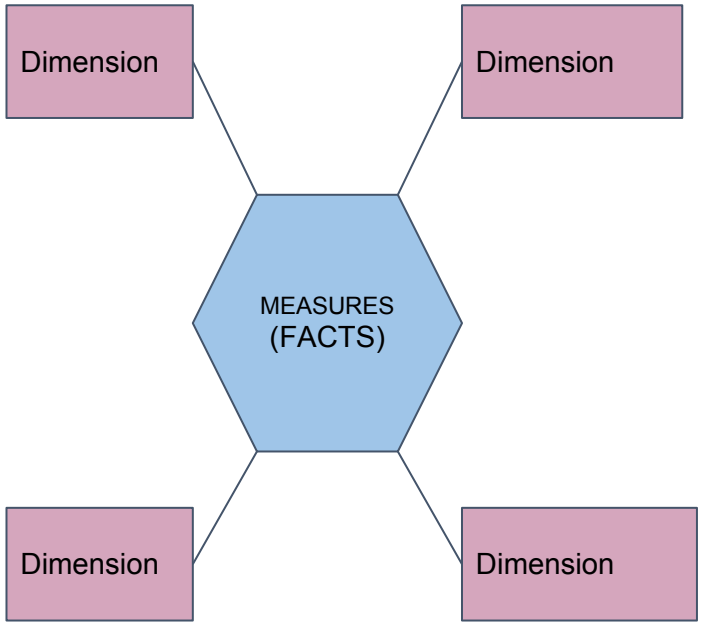
Data Mart

- is a part of Data Warehouse.
- focuses on one subject area- for example, Finance, or Sales
- Integrates information from a given subject area and/or a set of source systems (e.g. customer info from CRM, employee info from HR)
- is a **dimensional model using a star schema**.

Dimensional Modeling

Dimensional model

- Dimension Table
  - Detail description of data
  - Fat, as verbose as possible
- Fact Table
  - Dimension keys
    - Dimension A key
    - Dimension B key
    - Dimension C key etc.
  - Measures
    - Measure 1
    - Measure 2 etc.
  - Lean



Dimensional Model

1. What is the Business Process?
2. What is the Grain?
3. What are the Dimensions?
4. What are the Measures?

Business Process

Source Systems

Product master data from Inventory System

Store master data from Store System

Employee master data from HR System

Customer master data from CRM System

Daily text file extracts of POS terminals

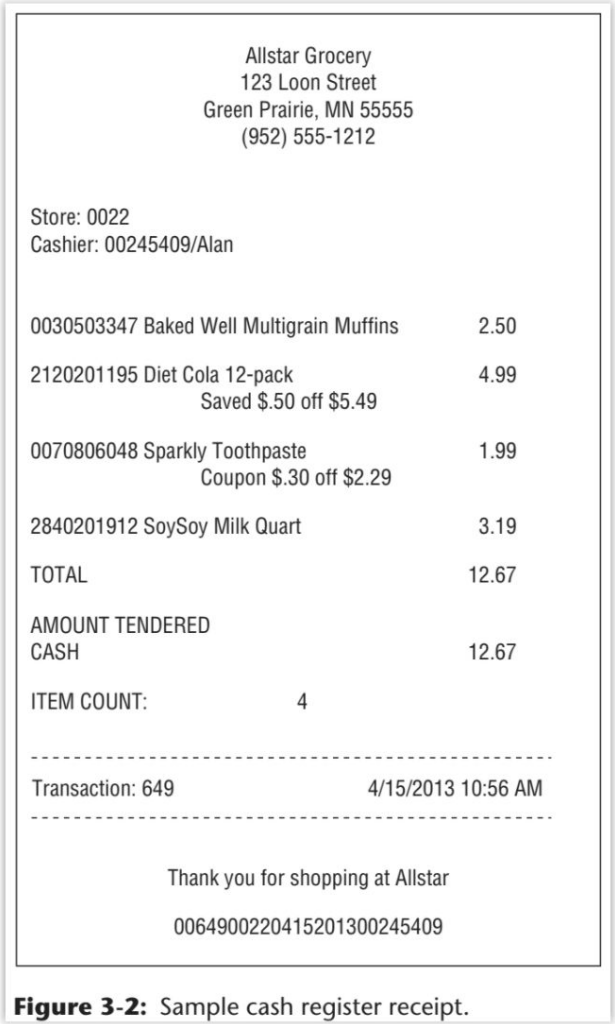
- POS header text files
- POS details csv files

Business Process

In our retail case study, management wants to better understand customer purchases as captured by the POS system. Thus the business process you're modeling is POS retail sales transactions. This data enables the business users to analyze which products are selling in which stores on which days under what promotional conditions in which transactions.

Business Process

- POS Terminal Operation
  - Sales table
  - Payment table
- Captures Sales Data
  - Scan Barcode and Qty
  - Lookup Retail Price
  - Reject if Barcode not found
  - Applies Valid Retail Price
  - Applies Valid Discount
  - Records Cashier Details



Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack	4.99
Saved $.50 off $5.49	
0070806048 Sparkly Toothpaste	1.99
Coupon $.30 off $2.29	
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED	
CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

00649002204152013002454O9

Figure 3-2: Sample cash register receipt.

Tables from POS Repository

- POS_Header
  - One Row per Transaction
- POS_Detail
  - One Row per Product Scanned (can duplicate)

POS header

TRX_ID - Unique Transaction id

JRSTORE - Store ID (variable length, take first five digits only)

JRREGNUM - Terminal # within the store

JRDATE - Day of transaction

JRTIME - Time of transaction

JRCASHIER - the the Salesperson credited for sale

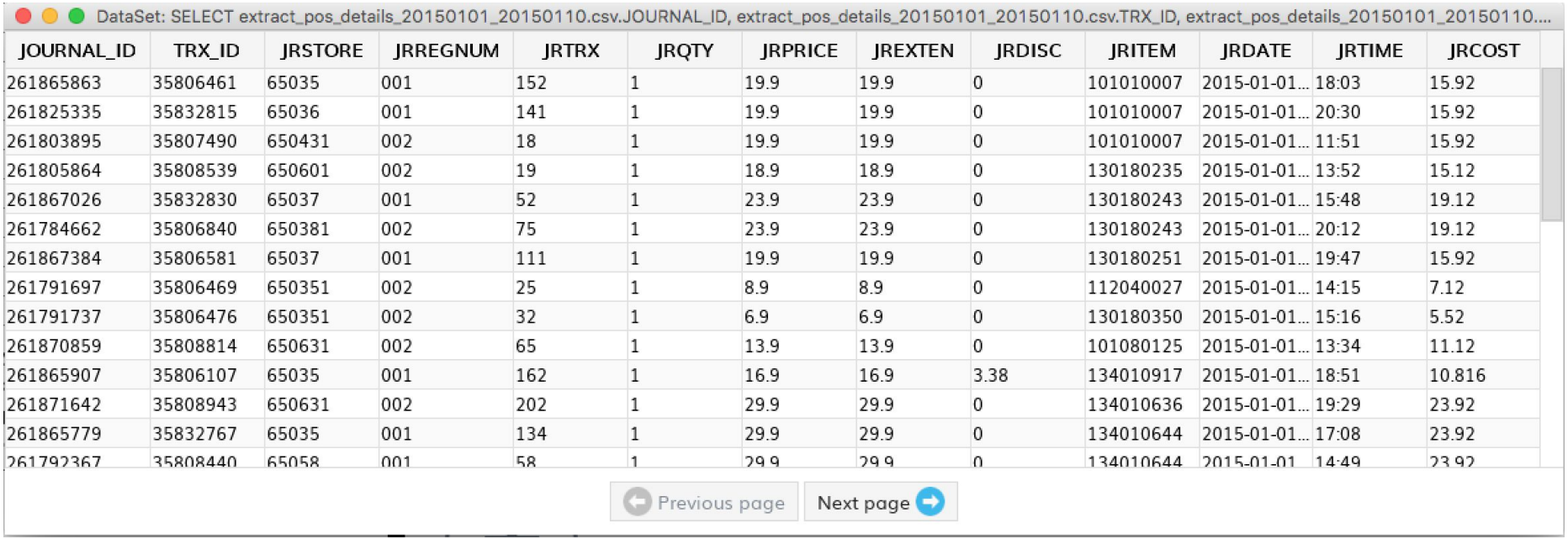
JRCUST - customer_id if the customer is a member

Source Table: POS_header.txt

DataSet: SELECT extract_pos_header.txt.TRX_ID, extract_pos_header.txt.JRSTORE, extract_pos_header.txt.JRREGNUM, extract_pos_h...

TRX_ID	JRSTORE	JRREGNUM	JRDATE	JRTIME	JRCASHIER	JRCUST
35802164	65010	001	2017-01-01 00:0...	12:08	06745	0
35802166	65010	001	2017-01-01 00:0...	14:48	06745	1116535
35802171	65010	001	2017-01-01 00:0...	19:00	06745	699665
35802173	65010	001	2017-01-01 00:0...	19:17	06745	174126
35802175	65010	001	2017-01-01 00:0...	20:17	06745	1997482

Source Table: POS_Detail


DataSet: SELECT extract_pos_details_20150101_20150110.csv.JOURNAL_ID, extract_pos_details_20150101_20150110.csv.TRX_ID, extract_pos_details_20150101_20150110....

JOURNAL_ID	TRX_ID	JRSTORE	JRREGNUM	JRTRX	JRQTY	JRPRICE	JREXTEN	JRDISC	JRITEM	JRDATE	JRTIME	JRCOST
261865863	35806461	65035	001	152	1	19.9	19.9	0	101010007	2015-01-01...	18:03	15.92
261825335	35832815	65036	001	141	1	19.9	19.9	0	101010007	2015-01-01...	20:30	15.92
261803895	35807490	650431	002	18	1	19.9	19.9	0	101010007	2015-01-01...	11:51	15.92
261805864	35808539	650601	002	19	1	18.9	18.9	0	130180235	2015-01-01...	13:52	15.12
261867026	35832830	65037	001	52	1	23.9	23.9	0	130180243	2015-01-01...	15:48	19.12
261784662	35806840	650381	002	75	1	23.9	23.9	0	130180243	2015-01-01...	20:12	19.12
261867384	35806581	65037	001	111	1	19.9	19.9	0	130180251	2015-01-01...	19:47	15.92
261791697	35806469	650351	002	25	1	8.9	8.9	0	112040027	2015-01-01...	14:15	7.12
261791737	35806476	650351	002	32	1	6.9	6.9	0	130180350	2015-01-01...	15:16	5.52
261870859	35808814	650631	002	65	1	13.9	13.9	0	101080125	2015-01-01...	13:34	11.12
261865907	35806107	65035	001	162	1	16.9	16.9	3.38	134010917	2015-01-01...	18:51	10.816
261871642	35808943	650631	002	202	1	29.9	29.9	0	134010636	2015-01-01...	19:29	23.92
261865779	35832767	65035	001	134	1	29.9	29.9	0	134010644	2015-01-01...	17:08	23.92
261792367	35808440	65058	001	58	1	29.9	29.9	0	134010644	2015-01-01	14:49	23.92

Previous page Next page

POS detail

JOURNAL_ID - Row ID (unique)

TRX_ID - Transaction ID

JRSTORE - Store ID, variable length, take first five digits only

JRREGNUM - Terminal # within the store, variable length

JRTRX - Transaction# at the terminal

JRDATE JRTIME - Date and time of the row

JRITEM - Barcode, used as product code for product look up

JRQTY - Quantity sold

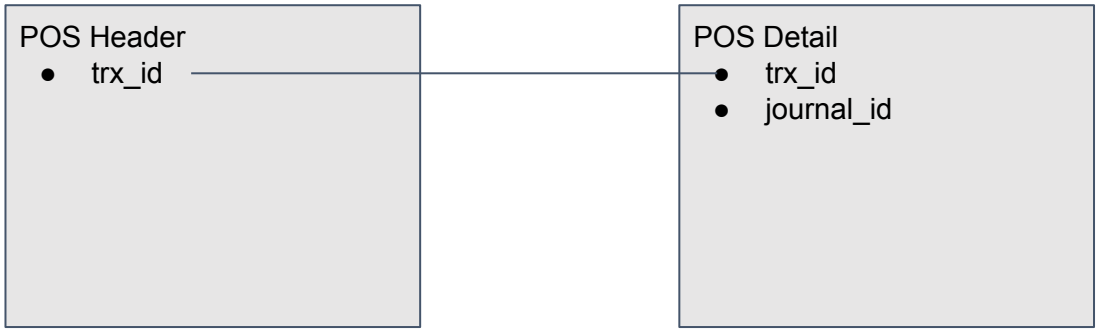
JRPRICE - Retail price to be sold at

JREXTEN - Qty X Retail price

JRDISC - Amount of discount

JRCOST - Unit Cost of the product (cost at which bought from the manufacturers)

POS header and POS detail



GRAIN

Grain

One row per scan of an individual product on a customer's sales transaction

In our case study, the most granular data is **a single product in a single transaction**.

Each product can appear multiple times for the same transaction depending on its quantity.

The unique identifier for a row in pos_details is "journal_id", not transaction id

Dimensions

Product Dimension

barcode - Unique identifier for each individual product

Category (e.g. Facial Skin care, Women's Fragrance etc.)

→ Range (e.g. Facial Skin care -> White mask, Vitamin C etc.)

→ SKU (e.g. Vitamin C -> 30ml VIT.C INTENS NIGHT TREAT 30ML, 60ml VIT.C SKIN BOOST etc.)

Other attributes such as weight, height etc.

Store Dimension

JRSTORE - String

Store_name - String

Store_address - String

Salesperson Dimension

Employee code - Unique identifier for an employee

Name

Address

Gender

Date joined

DOB

Age

Years in service

Customer dimension

Customer id - Unique identifier

First name

Last name

Age

Gender

Race

DOB

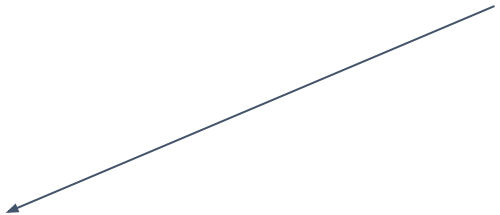
Mobile

Email

Date Dimension (Provided)

- Date is the most complex issue in Data Warehouse
- We provide date dimension which implements the following hierarchies (based on gregorian calendar)
  - calendar year → calendar month → calendar day
  - calendar year → calendar week → calendar day
  - Calendar year → calendar day
- Day of a weak
- Weakday Indicator

Date dimension



dim_date_key	version	date_from	date_to	date_str	cal_year	cal_month	cal_day_of_year	cal_day_of_month	cal_day_of_week	cal_week_of_year	d_o_w_str	d_o_w_str3	d_o_w_str1	weakday_indicator
1	1	1900/01/01 00:00:0...	2199/12/31 23:59:59....	20100101	2010	1	1	1	6	1	Friday	FRI	F	weekday
2	1	1900/01/01 00:00:0...	2199/12/31 23:59:59....	20100102	2010	1	2	2	7	1	Saturday	SAT	S	weekend
3	1	1900/01/01 00:00:0...	2199/12/31 23:59:59....	20100103	2010	1	3	3	1	2	Sunday	SUN	N	weekend

Hour of Day Dimension (Provided)

Hour of Day - “00” to “23”

Hour Minute - “00:00” to “23:59”

Typically only hour of day is used

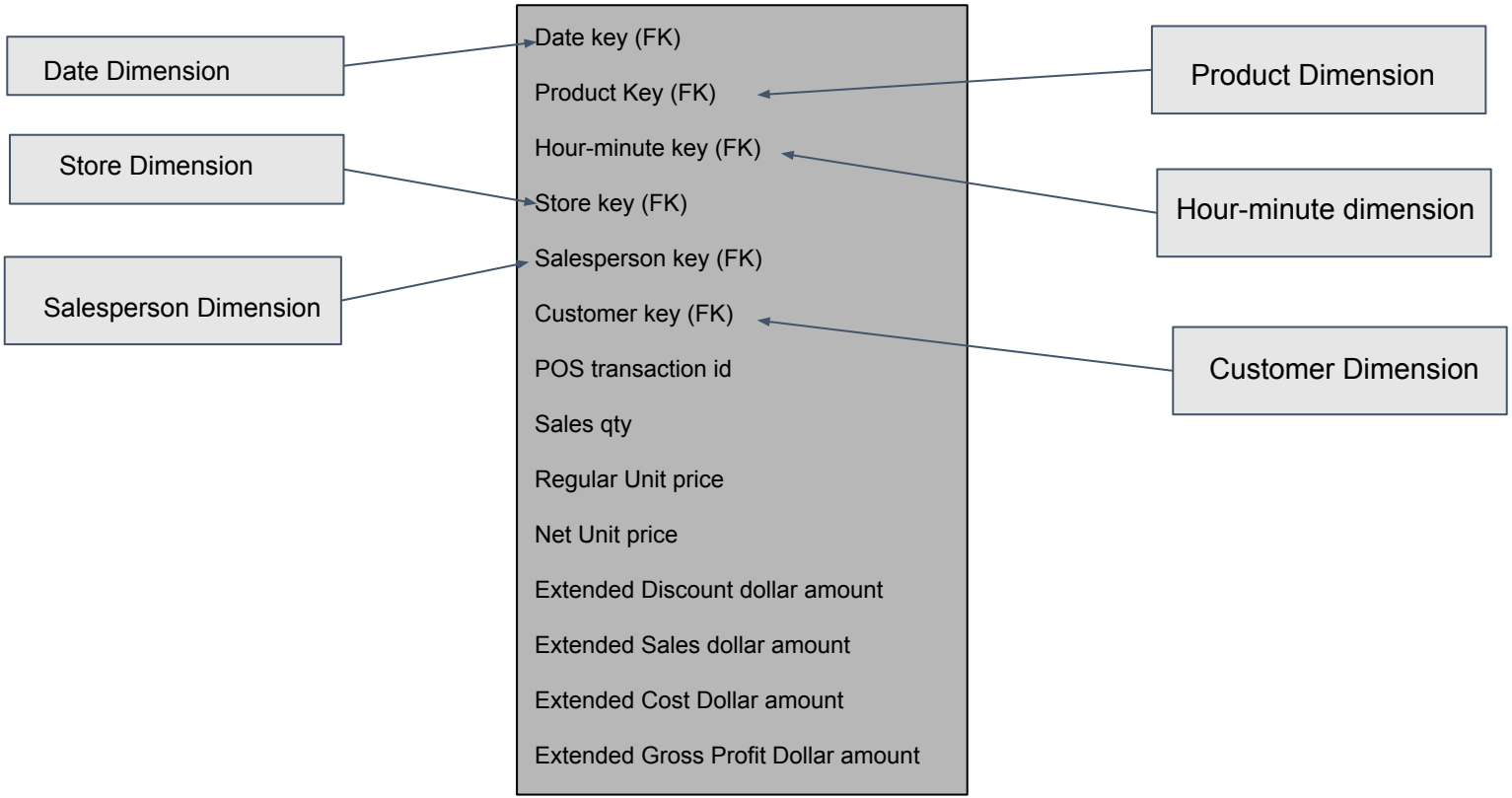
dim_hour_min_key	version	date_from	date_to	hour_min	hour_of_day
1	1	1900/01/01 00:00:00.000...	2199/12/31 23:59:59.999...	00:00	00
2	1	1900/01/01 00:00:00.000...	2199/12/31 23:59:59.999...	00:01	00
3	1	1900/01/01 00:00:00.000...	2199/12/31 23:59:59.999...	00:02	00
4	1	1900/01/01 00:00:00.000...	2199/12/31 23:59:59.999...	00:03	00
5	1	1900/01/01 00:00:00.000...	2199/12/31 23:59:59.999...	00:04	00

Measures

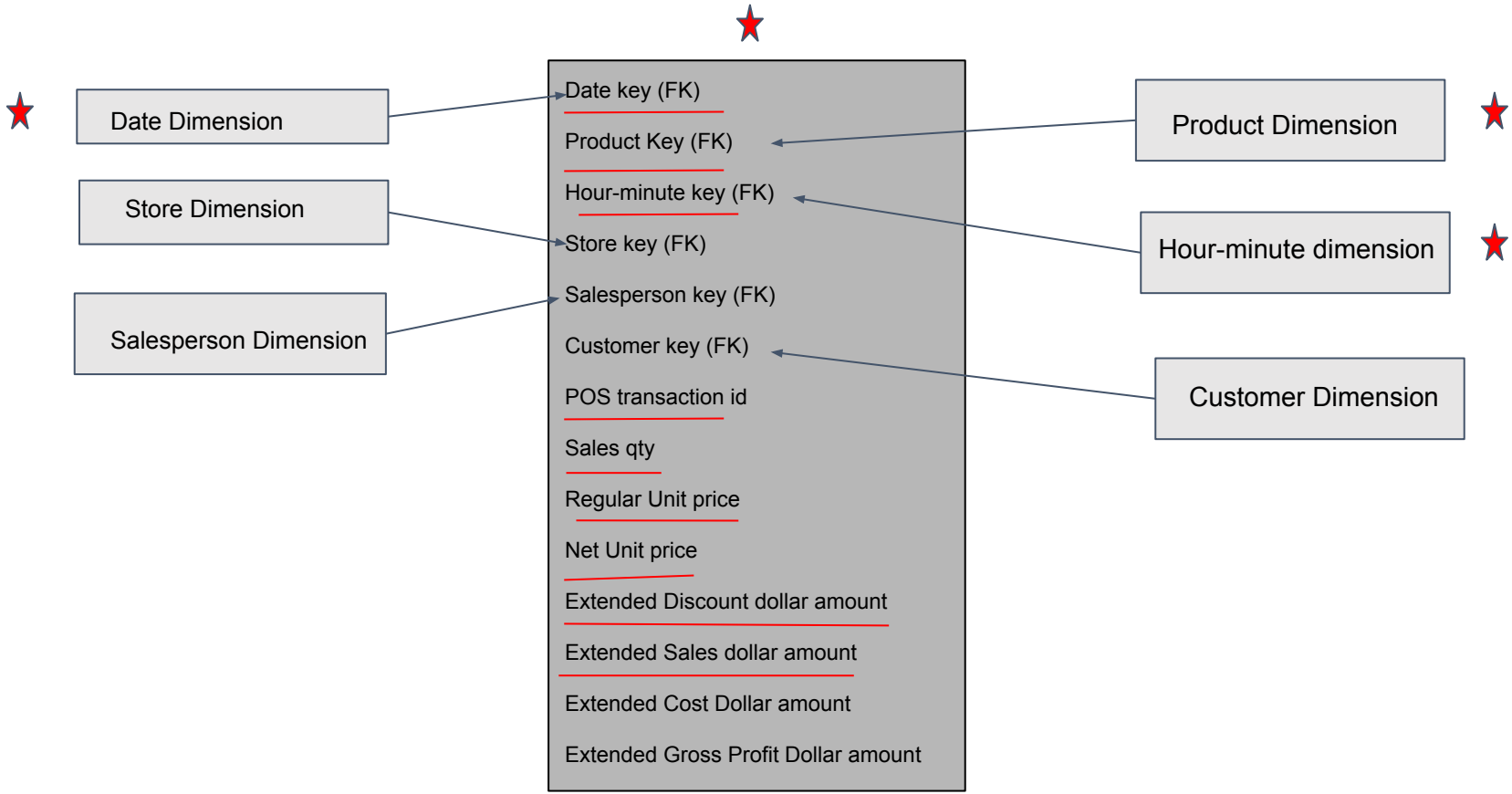
Definition of the measures

The facts collected by the POS system include the sales quantity (for example, the number of cans of chicken noodle soup), per unit regular, discount, and net paid prices, and extended discount and sales dollar amounts. The extended sales dollar amount equals the sales quantity multiplied by the net unit price. Likewise, the extended discount dollar amount is the sales quantity multiplied by the unit discount amount. Some sophisticated POS systems also provide a standard dollar cost for the product as delivered to the store by the vendor.

Retail Sales Schema



For today


Date Dimension
Store Dimension
Salesperson Dimension
Date key (FK)
Product Key (FK)
Hour-minute key (FK)
Store key (FK)
Salesperson key (FK)
Customer key (FK)
POS transaction id
Sales qty
Regular Unit price
Net Unit price
Extended Discount dollar amount
Extended Sales dollar amount
Extended Cost Dollar amount
Extended Gross Profit Dollar amount
Product Dimension
Hour-minute dimension
Customer Dimension