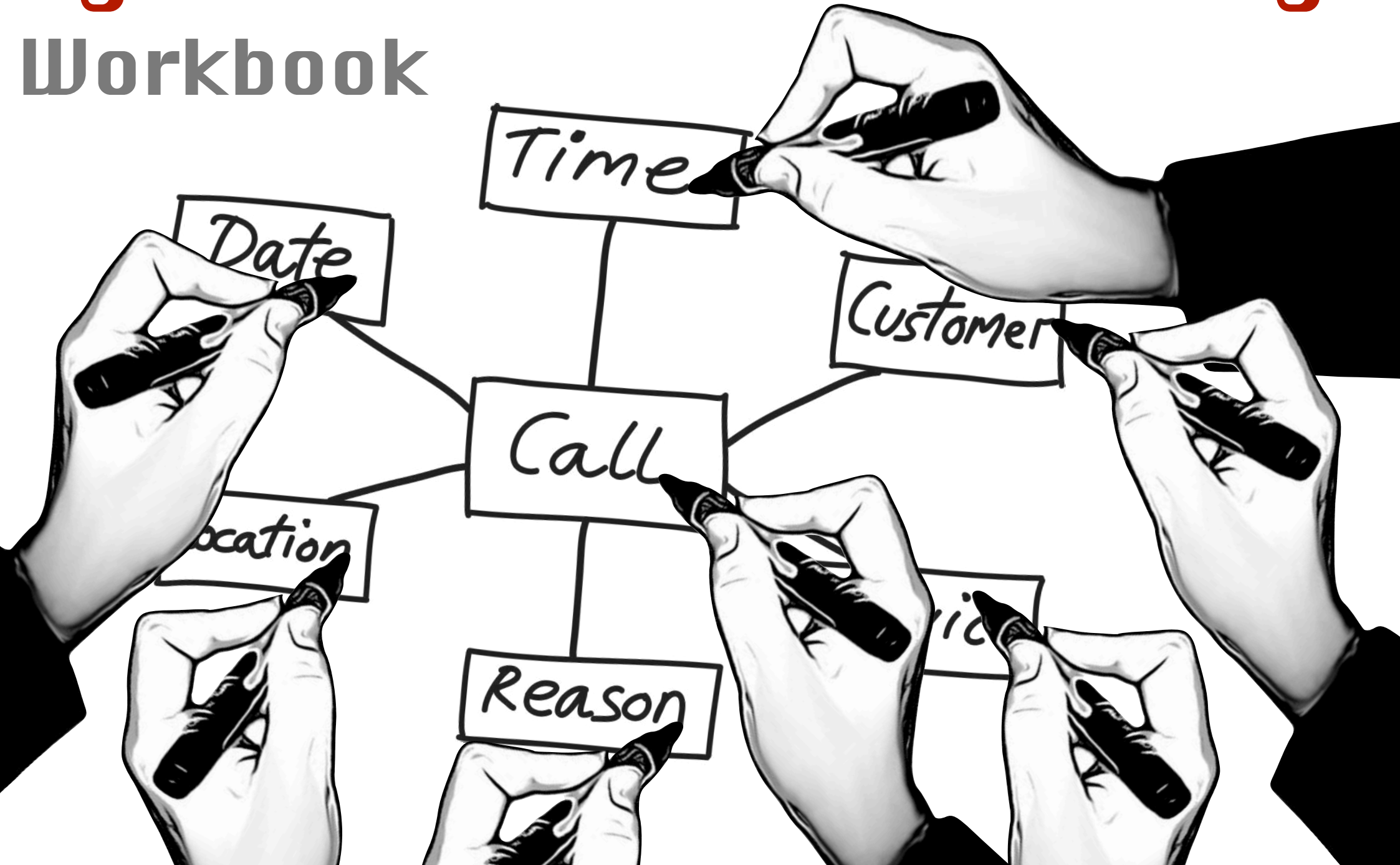
# Agile Data Warehouse Design

## Workbook



# Copyright



**BEAM: Agile Data Warehouse Design**
Author: Lawrence Corr
Revision: August 2025



DecisionOne*
* the agile data warehouse design company

**DecisionOne Consulting Ltd**
Burwood House
38-42 Waterloo Road
Leeds LS28 7UJ
United Kingdom

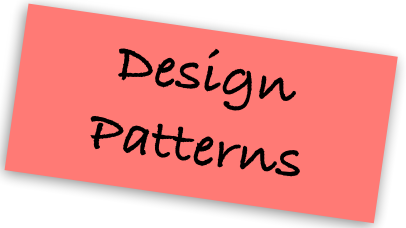
Tel: +44 (0)7971 964824

Email: **info@decisionone.co.uk**
Web: **decisionone.co.uk**

# Objectives/Agenda

- Understand the **requirements** and **challenges** of analytical data design
- Examine the impact of **agile** practices on DW/BI design
- Use **BEAM✲** for ***modelstorming*** analytical data requirements
- Work with visual collaboration tools for ***distributed*** modelstorming
- Systematically translate data requirements into efficient, flexible **dimensional models**
- Plan and design for **incremental** data delivery
- Identify and use common **dimensional design patterns**
- Find **books** and **online material** that support the techniques covered






**BEAM✲: Business Event Analysis & Modeling** – an agile (dimensional) modelstorming method

*modelstorming*: data modeling + brainstorming

# Agenda

Day 1

Data: Requirements and Challenges

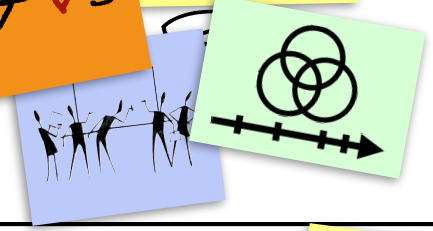
BEAM*

Modelstorming Single Business Events

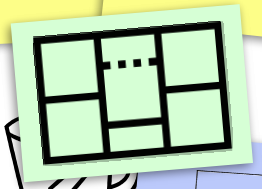
Modelstorming Multiple Business Events

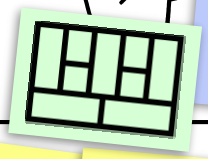
7Vs

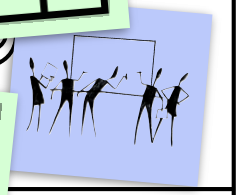
7Vs

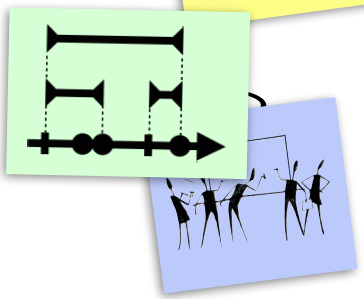


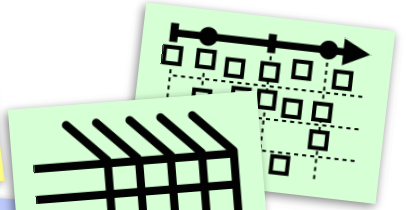












Day 2

Modeling by Example

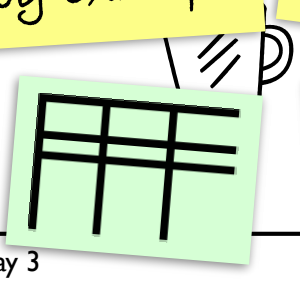
Agile Star Schema Design

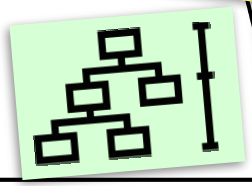
Historical Descriptions

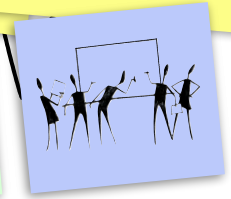
Lunch

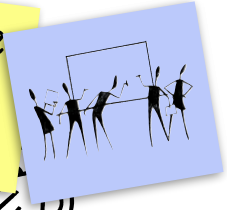
Fact table Design Patterns

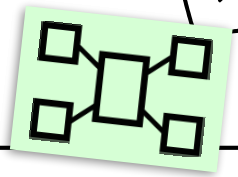
Star Schema Refactoring

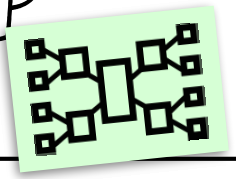


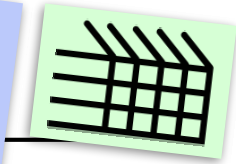












Day 3

Dimensional Design Patterns

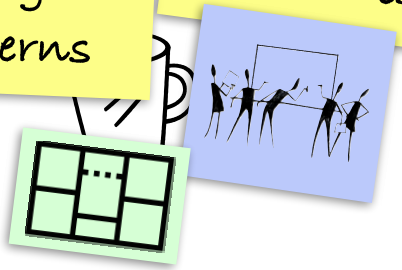
Products Resources

Customers Employees

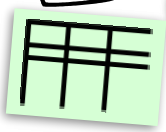
Time & Location

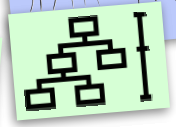
Cause & Effect

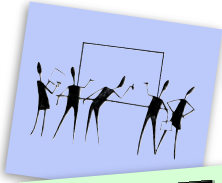
Review Bookshelf

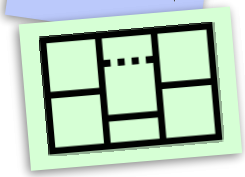




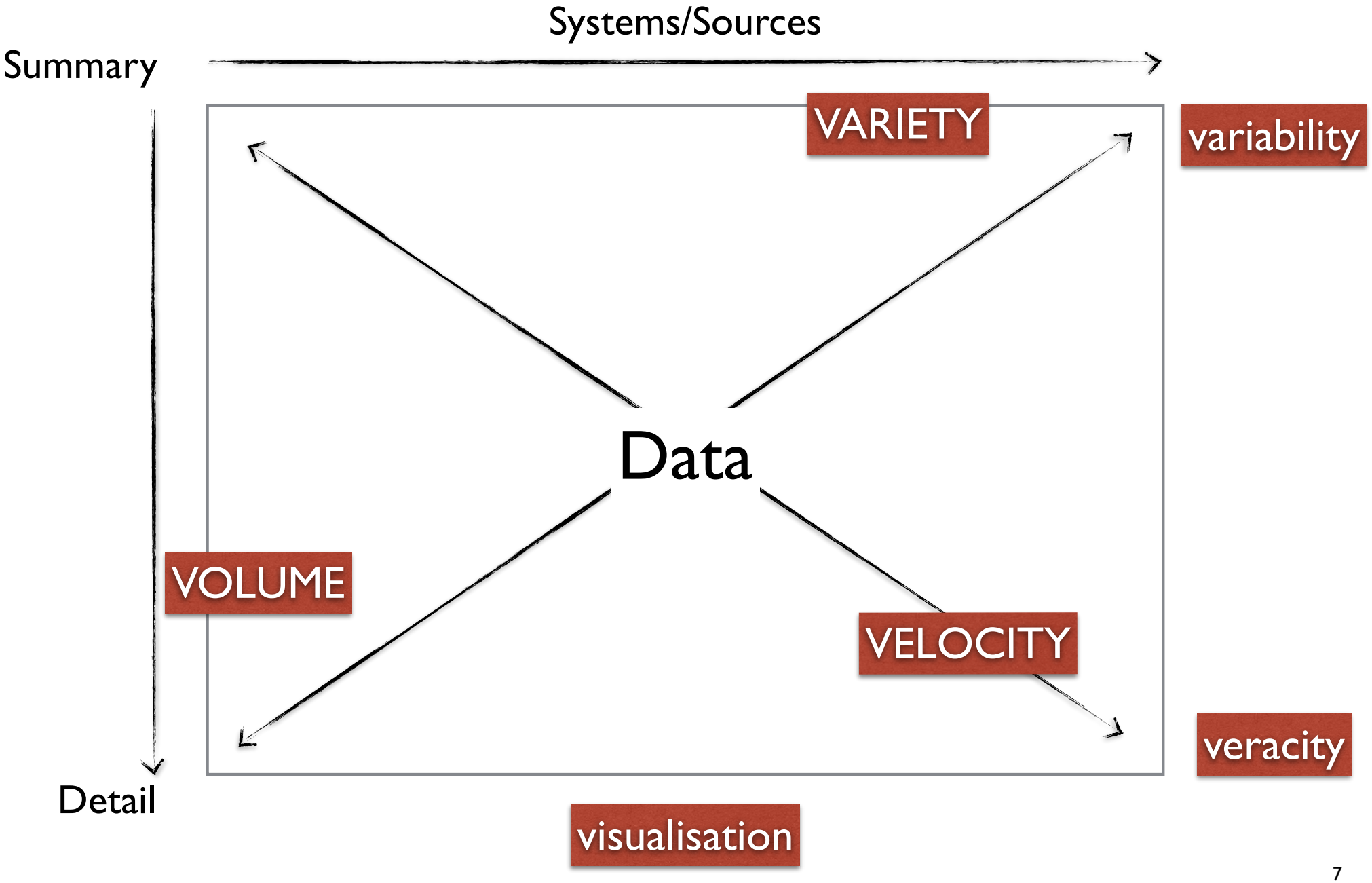






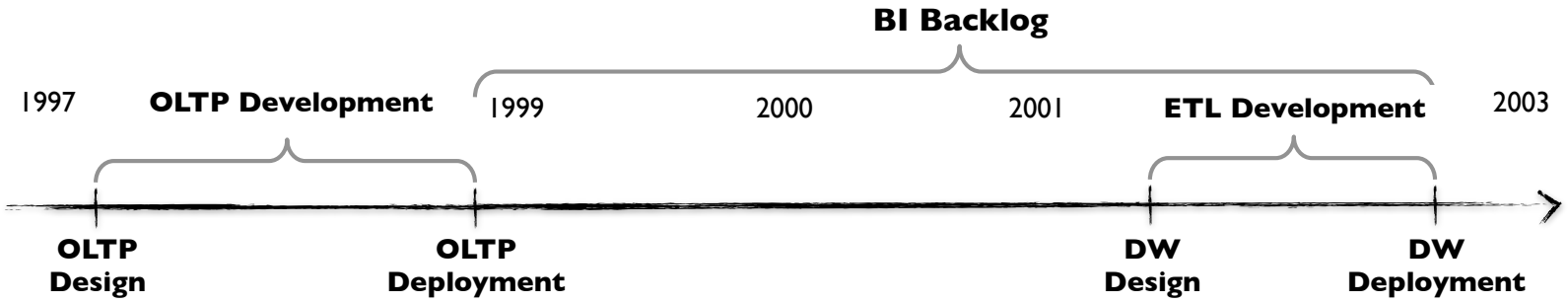


# The 7Vs of Data

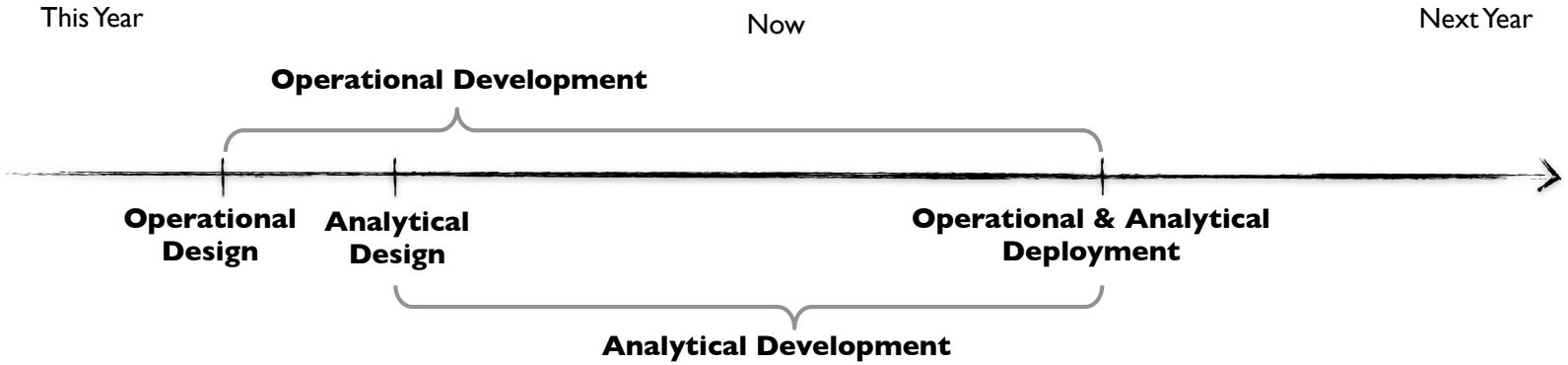


# Reactive vs. Proactive Design

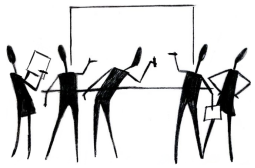
## Reactive




## Proactive




ADWD pp. 13-15



# Exercise 1: The 6Ws Game


## Goal

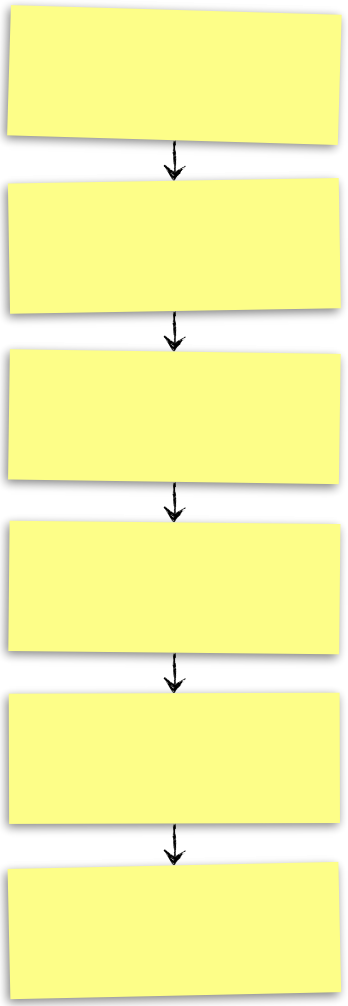
Discover as much as you can about your team in 10 minutes or less. e.g.:

- Member Names, Companies/Industries, Roles
- DW/BI experience
- Agile experience
- Current/Recent DW/BI Software usage: Databases, ETL Tools, BI Tools
- Personal objectives for this course

## Rules

You have only 6 questions you can ask each other:

- **Why** are you attending this course?
- **What** do you do? (What type of data do you work with?)
- **When** did you start doing that? (How much experience do you have?)
- **How** do you do that? (What software tools and techniques do you use?)
- **Who** are you?
- **Where** do you work/come from?



## Play

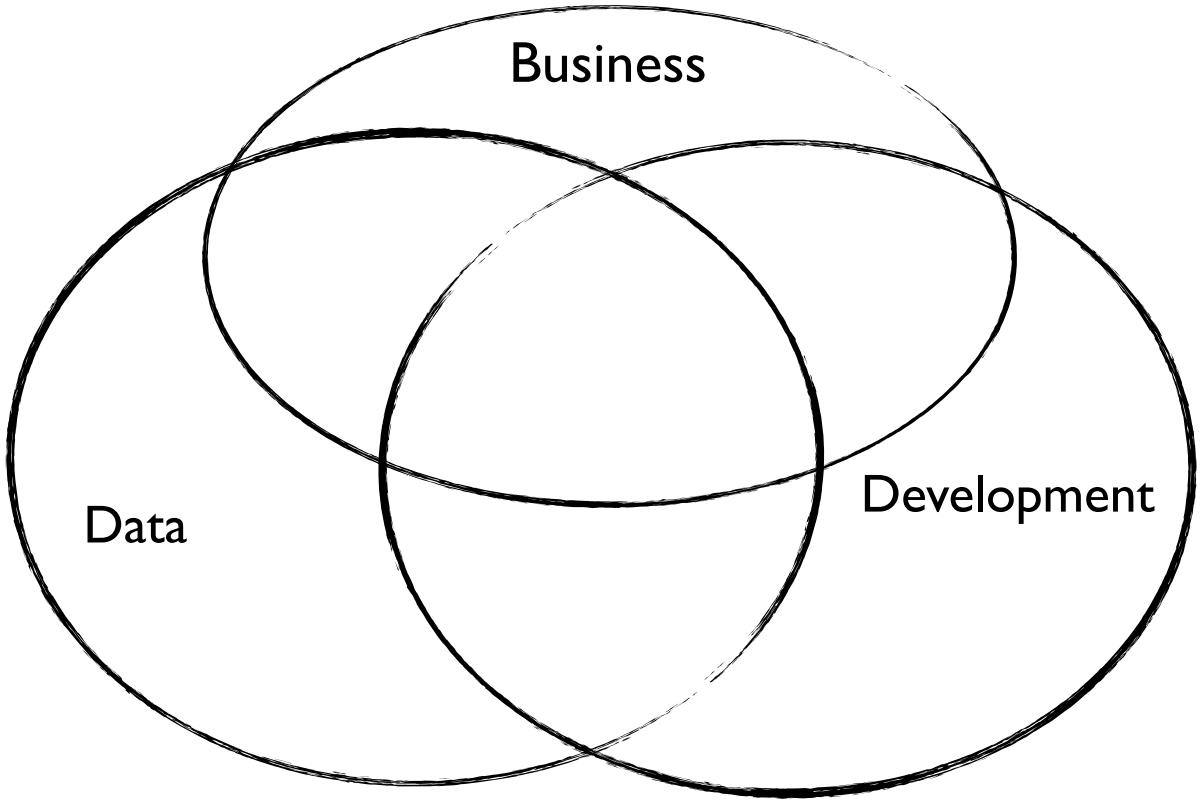
1. Each team decides upon the best order to ask their questions by designing a large 6W flowchart. Start by putting each the question keywords (**Why, What, When, How, Who, Where**) on separate Post-its and collaboratively sorting them on the wall.

2. Each team member puts their (short) answers on separate Post-its (using a different colour to the other team members). Create a detailed team board which will combine your flowchart and answers. Introduce yourself to your team by adding your answer Post-its to the board.

3. Summarise your team on a second board using the template on the following page by adding Post-its for yourself (save digital ink/effort; use your initials on multiple notes). Use the same colour notes as before so the summary and detail boards can be related. Finish by picking a team name/colour.

# Modelstorming Team: *Number/Name/Colour*

Who / What




Who / When

**Data (Analytics/Modeling/Analysis/Development/DW/BI) Experience**

0 Years — 5 Years — 10+ Years

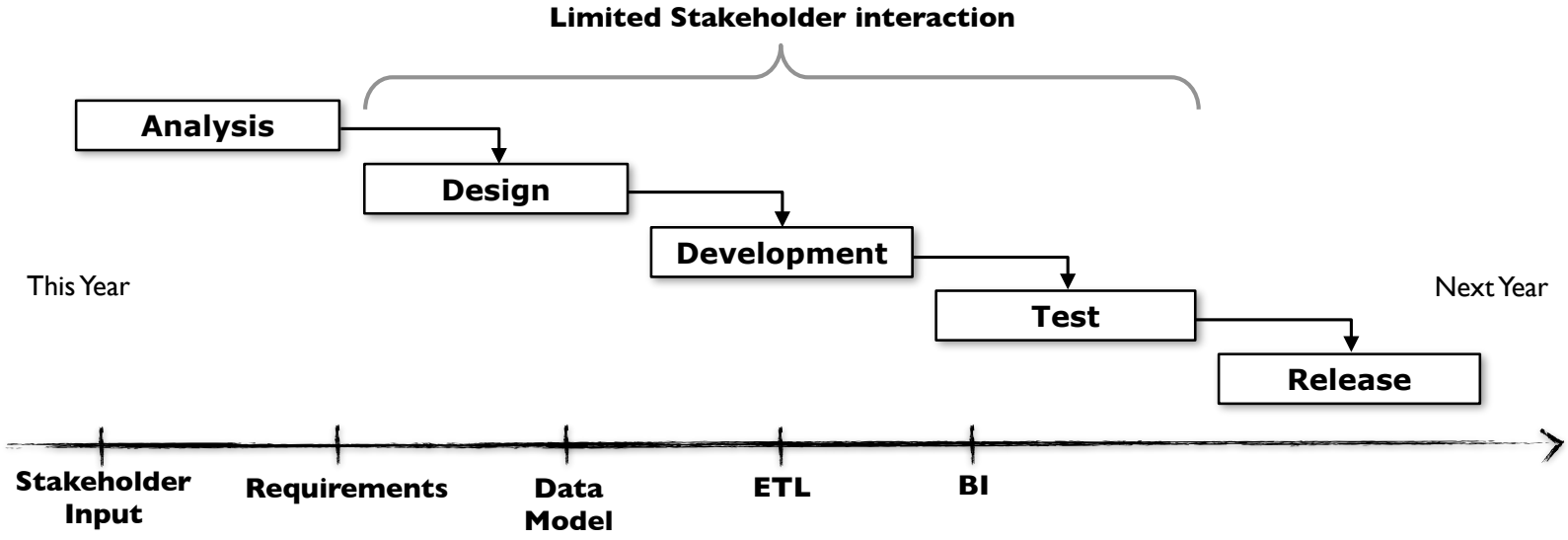
**Agile Experience**

0 Years — 5 Years — 10+ Years

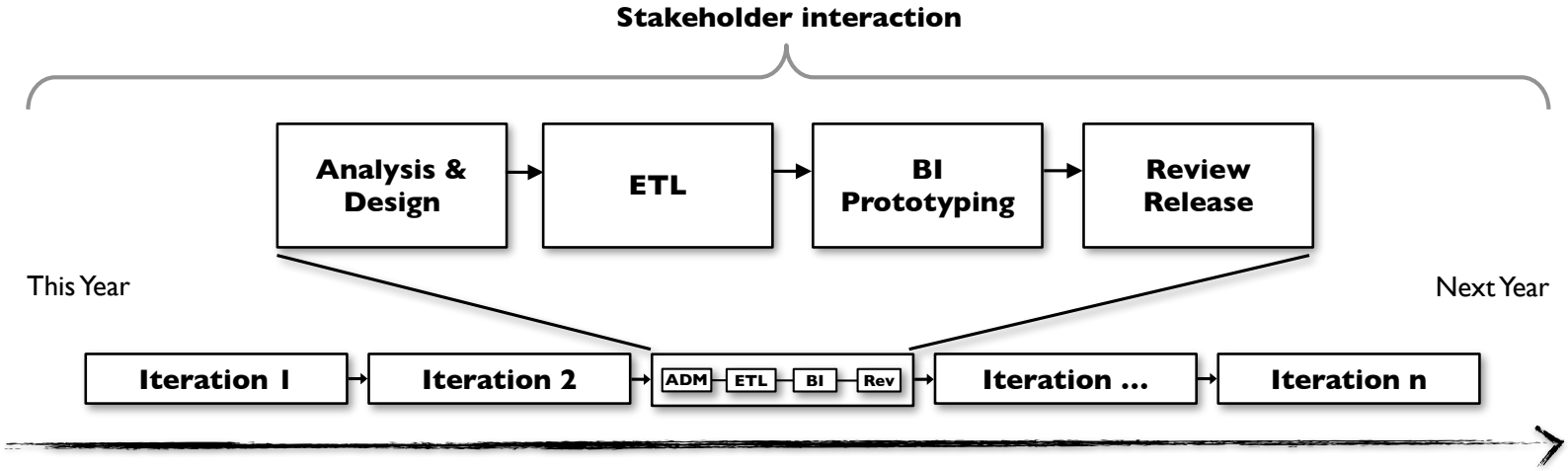
**Alliander Experience**

0 Years — 5 Years — 10+ Years

# Traditional (waterfall) DW/BI Development




# BI Development



# The Agile Manifesto


We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

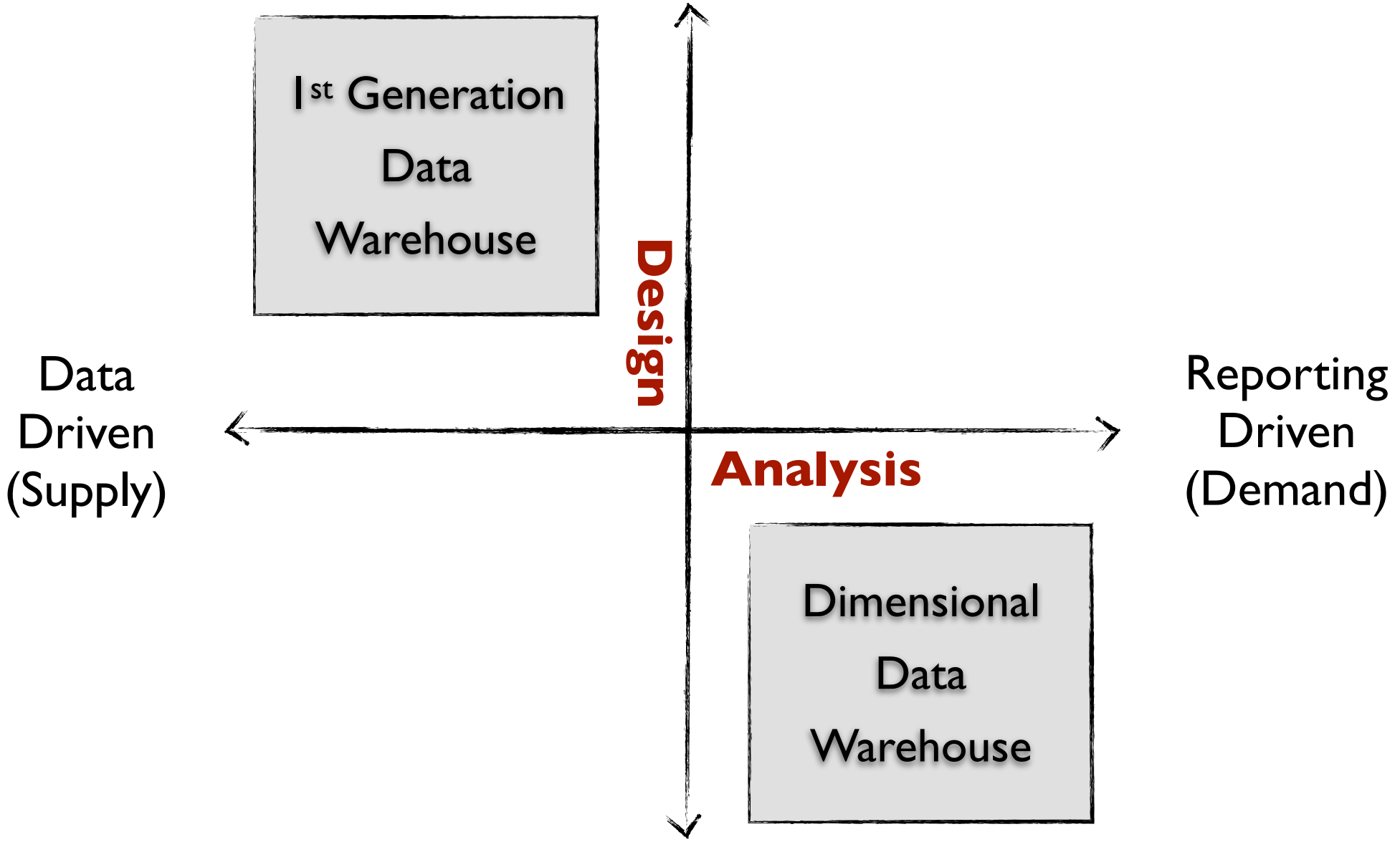
- **Individuals and interactions** *over* processes and tools
- **Working software** *over* comprehensive documentation
- **Customer collaboration** *over* contract negotiation
- **Responding to change** *over* following a plan

That is, while there is value in the items on the right, we value the items on the left more.

# 12 Principles of Agile Software Development

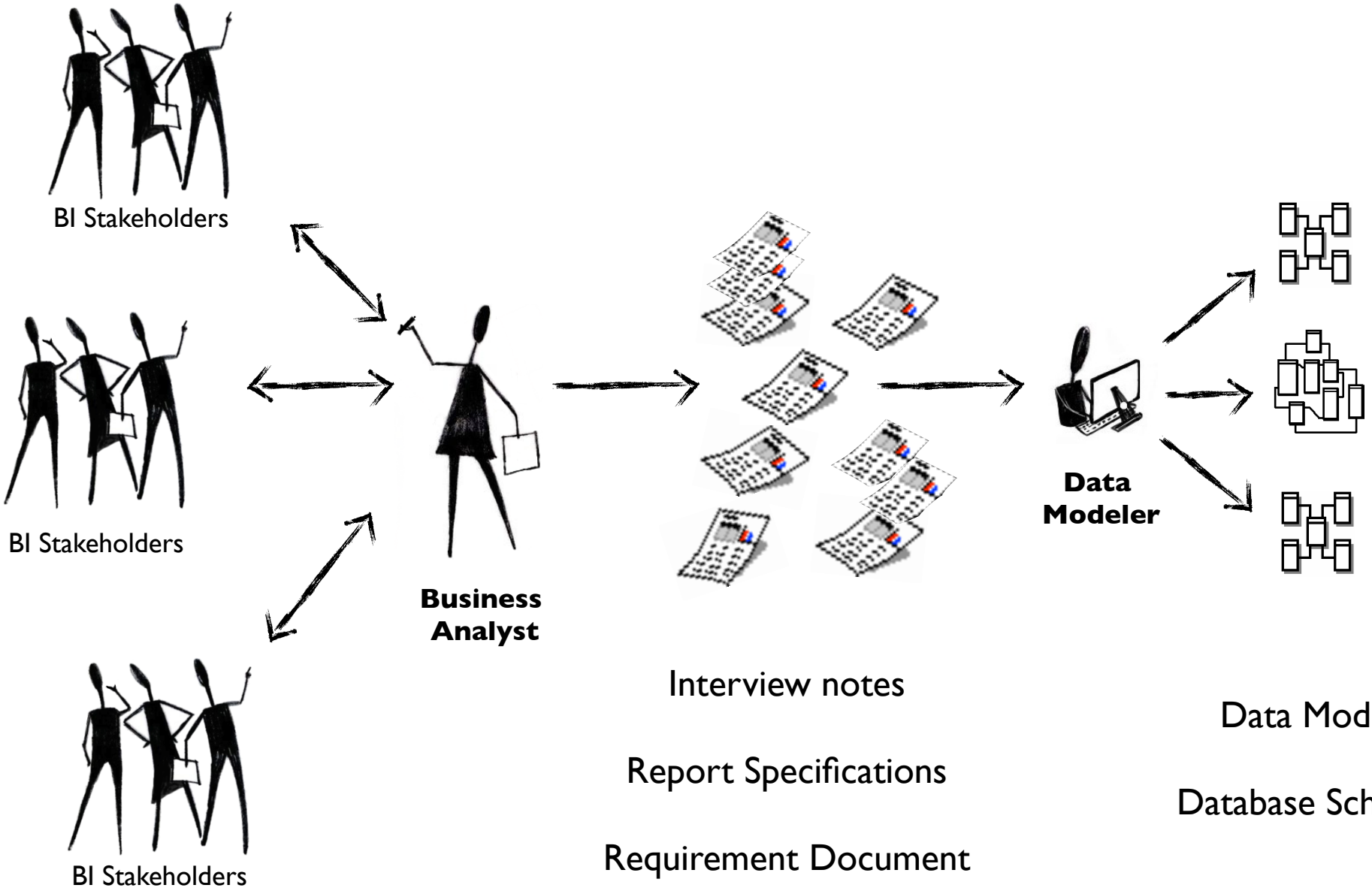
1. Our highest priority is to satisfy the customer through **early** and **continuous delivery of valuable software**.
2. **Welcome changing requirements**, even late in development. Agile processes harness change for the customer's competitive advantage.
3. **Deliver working software frequently**, from a couple of weeks to a couple of months, with a preference to the shorter time scale.
4. Business people and developers must **work together** daily throughout the project.
5. Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
6. The most efficient and effective method of conveying information to and within a development team is **face-to-face** conversation.
7. **Working software** is the primary measure of progress.
8. Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. **Simplicity** – the art of maximizing the amount of work not done – is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behaviour accordingly.

# Entity Relationship Modeling – 3NF




# Dimensional Modeling – Star Schema

# Reporting-Driven Analysis



# Reporting-Driven Analysis: Pros & Cons

- Gather reporting requirements by interviewing potential BI users (individually or in small groups)
- Translate example reports into BI data requirements
- User involvement helps to create (initially) successful design

- *Accretive* BI requirements build up over time: impossible to capture in full, in advance
- Time consuming: *analysis paralysis*
- Focusing too closely on current report requirements can lead to inflexible (dimensional) models


# Data Modeling

*"Data modeling is the act of exploring data-oriented structures. Evolutionary data modeling is data modeling performed in an* ***iterative*** *and* ***incremental*** *manner. Agile data modeling is evolutionary data modeling done in a* ***collaborative*** *manner."*

– Scott Ambler
www.agiledata.org

# Why Model Collaboratively?

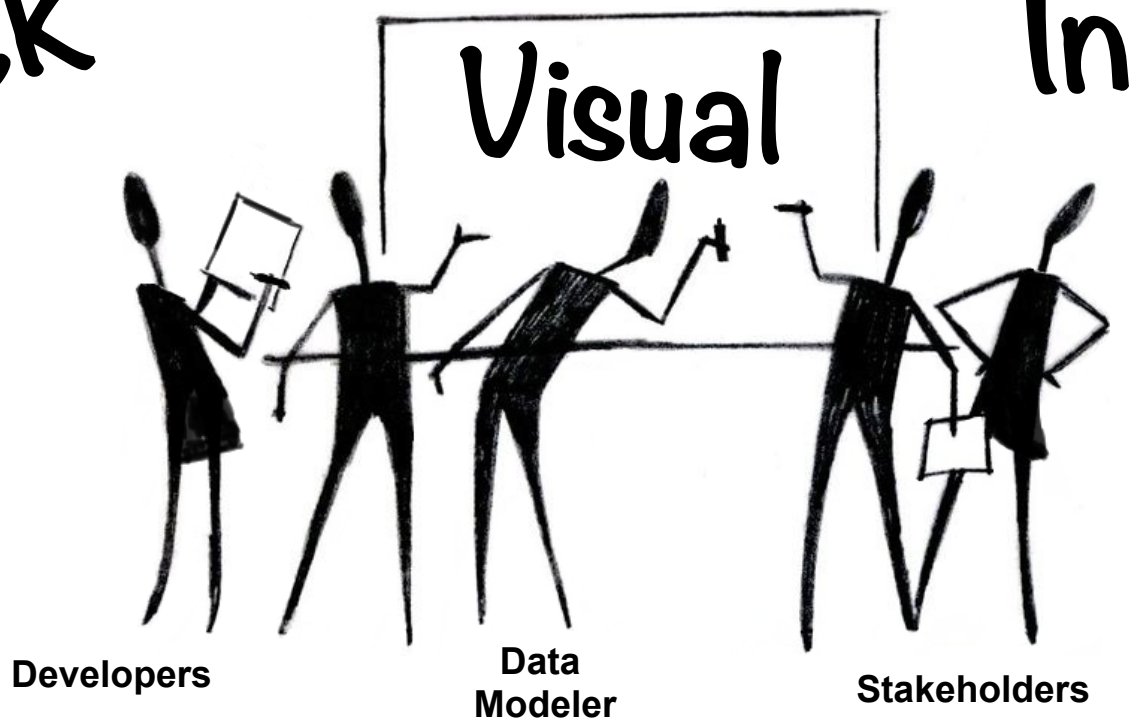
- Identify the significant business events worth recording and measuring
  *Scope and Prioritization: Verbs*
- Discover how business events are described
  *User Vocabulary: Nouns*
- Determine how they are measured
  *Measures, Hierarchies, Comparisons, KPIs*
- Unearth budgets, forecasts, targets and other user-controlled data sources
  *Extra data, Common summarization levels: Physical Optimization*
- **Create shared ownership of data issues and pride in the solution**

# Modelstorming: Data Modeling + Brainstorming

Quick

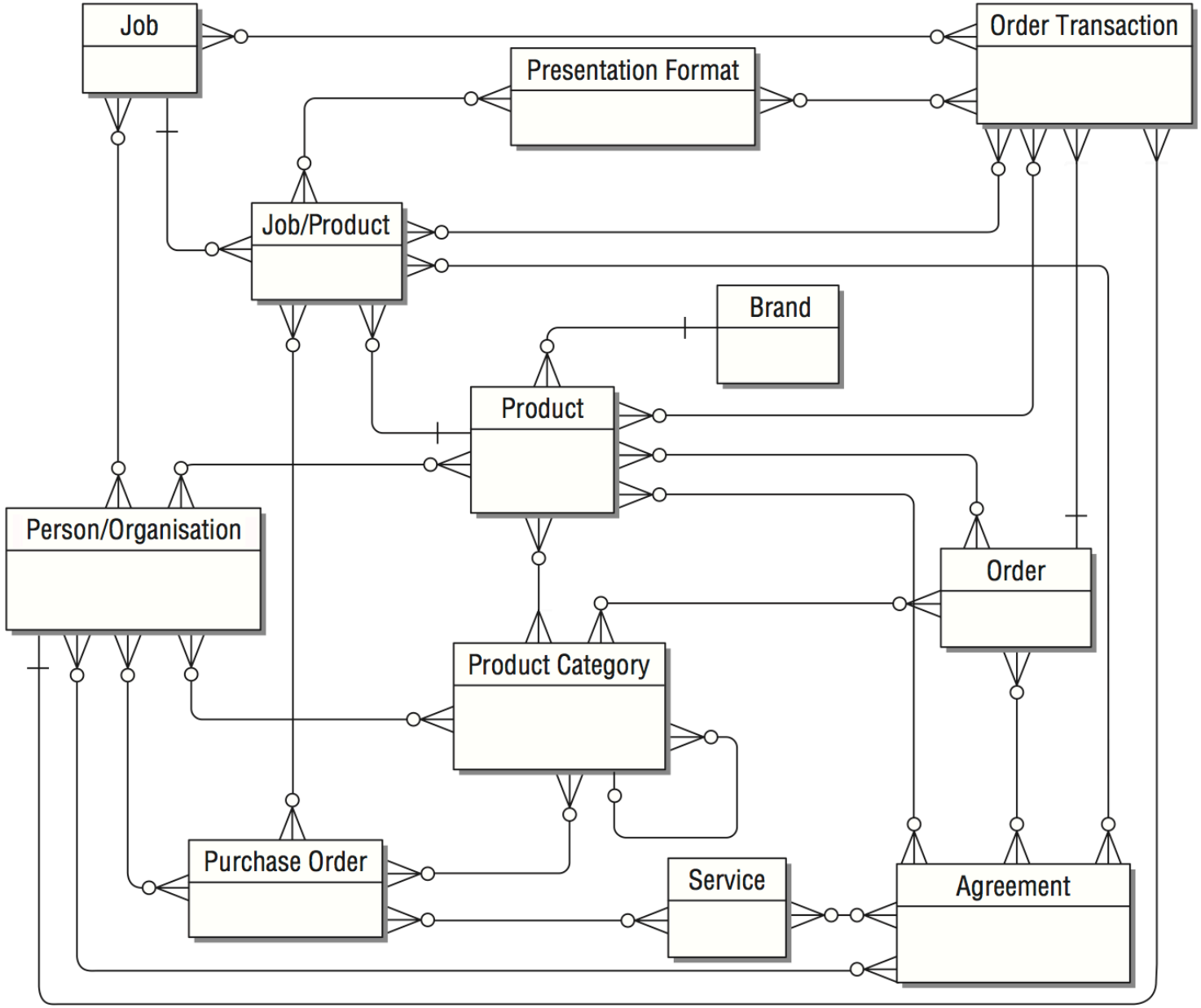
Visual

Inclusive

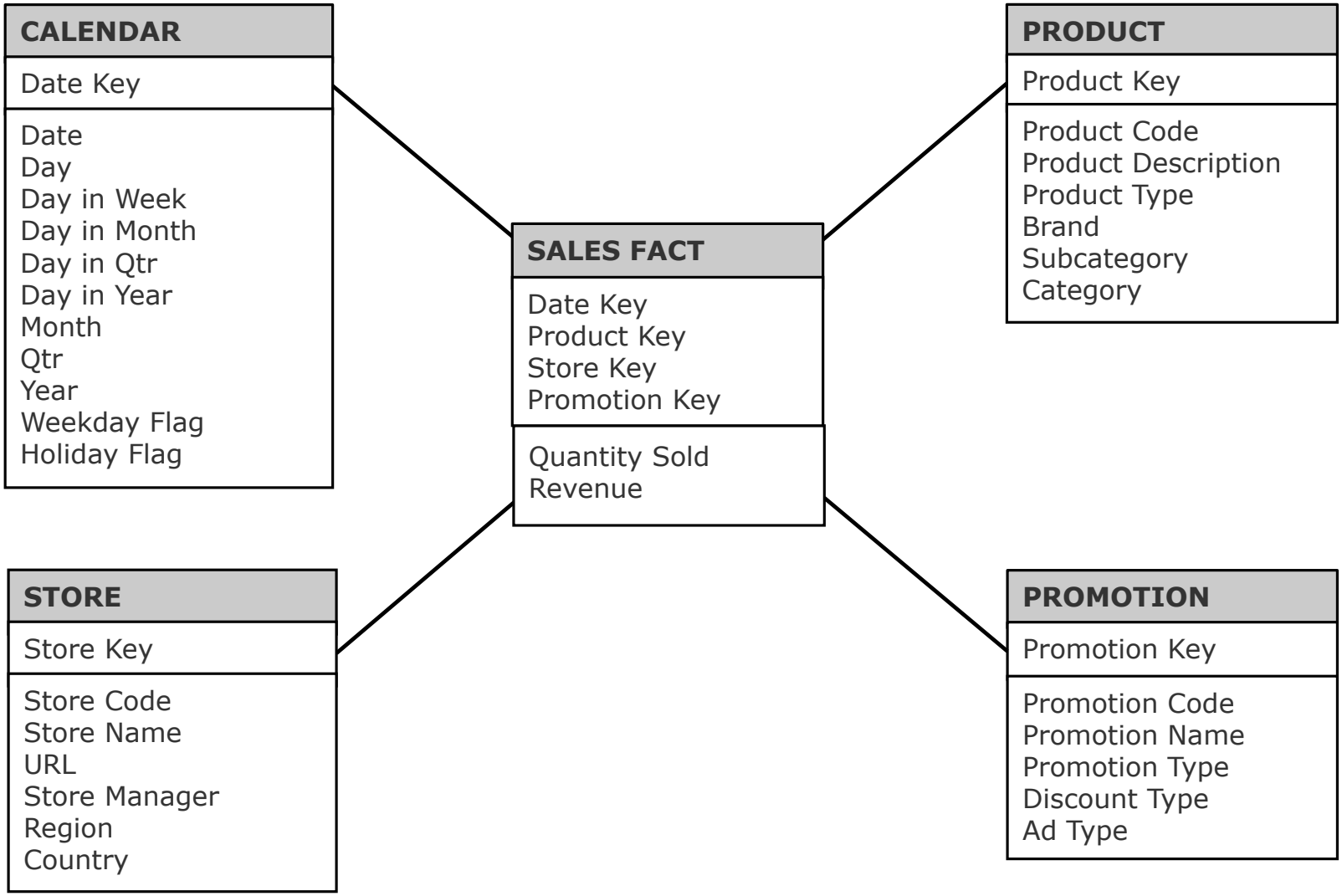



Interactive

# Entity Relationship Model



# Dimensional Model: Star Schema



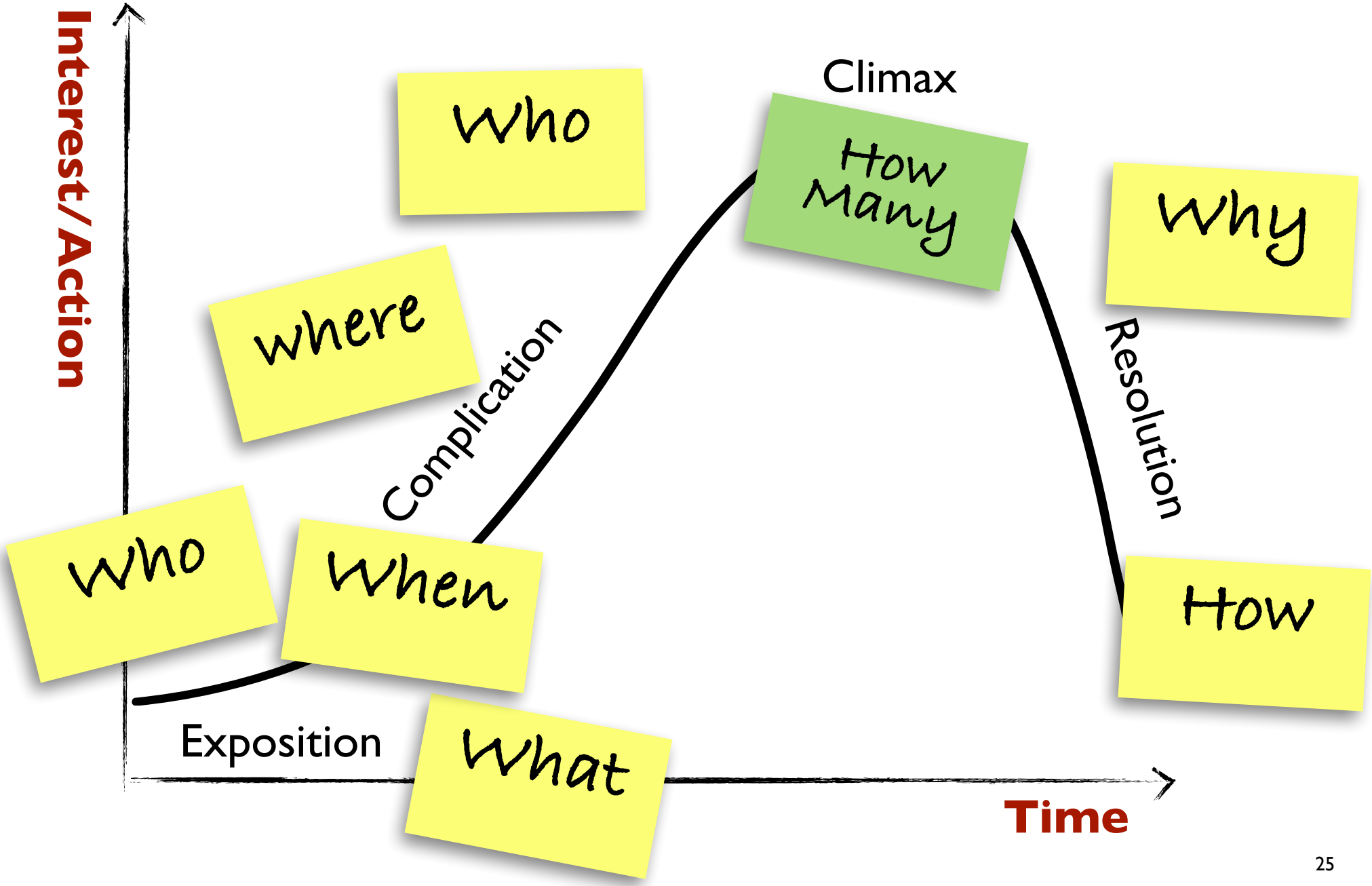
# Why *Dimensional* Modelstorming?

- Focuses initial requirements gathering on business processes rather than reports:
  - Creates flexible, report-neutral dimensional models
  - Enables proactive analytical data design to influence operational systems
  - Teaches everyone involved to think dimensionally
- Improves Traditional BI Analysis
  - Data-driven analysis becomes targeted data profiling
  - Reporting-driven analysis becomes BI prototyping

# BEAM


Business Event Analysis & Modeling

# Data Stories: Following a Narrative Arc



# BEAM: a 7*W* Framework for describing a Business Event

1. ***Who*** is involved?

2. ***What*** did they do? To what is it done?

3. ***When*** did it happen?

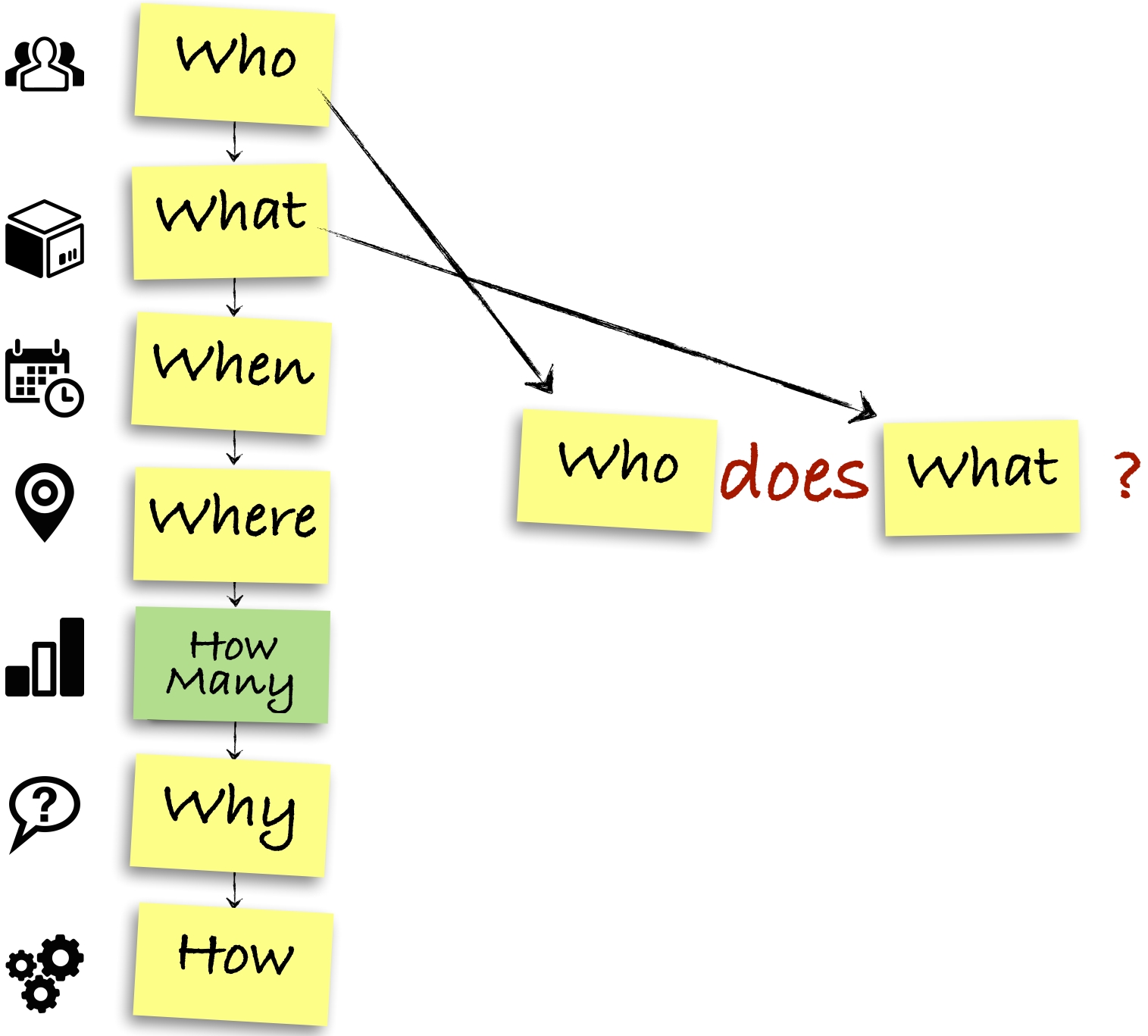
4. ***Where*** did it take place?

5. ***HoW many*** or much was recorded – how can it be measured?

6. ***Why*** did it happen?

7. ***HoW*** did it happen – in what manner?

# Using the BEAM 7Ws Script to tell Data Stories



# Pick a Single Data Story (event/verb) to model in detail

# When

*does a customer buy a product?*


*does a customer buy a product from?*

Where

*does a customer buy a product?*


How
Many


*does a customer buy a product on... ?*


*does a customer buy a product?*

Event/Title

Subject Verb Object

Author(s), Date/Version

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event

Purchase Date

## Ho

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we u*

Verb, Activity, P
Effect, Outcome,
Transaction Type
Transaction #, Ev
Step/Sequence #

[Granularity, Even

Customer buys Product

Invoice

## Who

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

Subject/Object
Customer: Business, Consumer, S
Employee
Supplier
Partner
Third Party

Customer

Salesperson

## How Many

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/Deltas
Balances
Activit
Durati
Measu

[UoM,

Quantity [items]

Revenue [$]

## Where

*Where does it happen? Where does it refer to?*

Subject/Object
Location
Branch, Store, Facility
Channel
URL
Map/Seq

Store/URL

## What

*What is involved/used? How are they organized?*
*How d change?*

Subject
Value P
Produc
Servic
Resou
Item

Product

## Why

*Why does it happen?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

Promotion

BEAM* Canvas

**modelstorming.com** - Agile Data Warehouse Design

*Business Event Analysis and Modeling

# The Business Model Canvas

Pomegranate

| | | | |
|---|---|---|---|
| *De* | *Designed by:* | *Date:* | *Version:* |

## Key Partners

Who are our Key Partners?
Who are our key suppliers?
Which Key Resources are we acquairing from partners?
Which Key Activities do partners perform?

**MOTIVATIONS FOR PARTNERSHIPS**
*Optimization and economy*
*Reduction of risk and uncertainty*
*Acquisition of particular resources and activities*

## Key Activities

What Key Activities do our Value Propositions require?
Our Distribution Channels?
Cu
Re

Manufacturing

**CAT**
*Pro*
*Pro*
*Plat*

## Key Resources

What Key Resources do our Value Propositions require?
Our Distribution Channels? Customer Relationships?
Revenue Streams?

**TYPES OF RESOURCES**
*Physical*
*Intellectual (brand patents, copyrights, data)*
*Human*
*Financial*

## Value Propositions

W
W
he
W
of
W

iPip

iPOM

POMBook

**CHA**
*Nee*
*Per*
*Cus*
*"Getting the Job Done"*
*D*
*B*
*R*
*C*
*R*
*Ac*
*Co*

## Customer Relationships

What type of relationship does each of our Customer Segments expect us to establish and maintain with them?
Which ones have we established?
How are they integrated with the rest of our business model?
How costly are they?

**EXAMPLES**
*Personal assistance*
*Dedicated Personal Assistance*
*Self-Service*
*Automated Services*
*Communities*
*Co-creation*

Account Management

## Channels

Through which Channels do our Customer Segments want to be reached?
How are we reaching them now?
How are our Channels integrated?
Which ones work best?
Which ones are most cost-efficient?
How are we integrating them with customer routines?

**CHAN**
*1. Aw*
*Ho*
*2. Eva*
*Ho*
*3. Pur*
*Ho*
*4. Deli*
*Ho*
*5. Afte*
*Ho*

Resellers

## Customer Segments

For whom are we creating value?
Who are our most important customers?

*Mass Market*
*Niche Market*
*Segmented*
*Diversified*
*Multi-sided Platform*

Consumers

SME

## Cost Structure

What are the most important costs inherent in our business model?
Which Key Resources are most expensive?
Which Key Activities are most expensive?

**IS YOUR BUSINESS MORE**
*Cost Driven (leanest cost stru*
*Value Driven (focused on valu*

**SAMPLE CHARACTERISTICS**
*Fixed Costs (salaries, rents, ut*
*Variable costs*
*Economies of scale*
*Economies of scope*

$$

## Revenue Streams

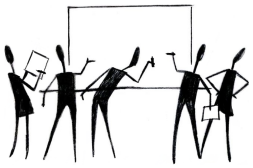
For what value are our customers really willing to pay?
For what do they currently pay?
How are they currently paying?
How would they prefer to pay?
How much does each Revenue Stream contribute to overall revo

**TYPES**
*Asset sale*
*Usage fee*
*Subscription Fees*
*Lending/Renting/Leasing*
*Licensing*
*Brokerage fees*
*Advertising*

**FIXED PRICING**
*List Price*
*Product feature dependent*
*Customer segment dependent*
*Volume dependent*

**DYNAM**
*Negoti*
*Yield M*
*Real-tir*

$$$



# Exercise 2: BEAM Canvas

**Requirements**

Sales analysis of the reseller (VARs and retailers) channel has been chosen for the pilot of a new agile BI project at Pomegranate. Management require a profit and loss view showing top line gross sales revenue to bottom line contribution (sales profit or margin) queryable by reseller and product.

**Steps**

1. Create a **BEAM 7W Script flowchart** that is visible to all team members
2. Study the reseller sales channel 'Business Knowledge' below which includes brief interview notes and a sample invoice with attached note of additional product details. Look out for the *7Ws* within this information which represents all the data currently in scope.
3. Ask the initial W question (**who does what?**) to discover a significant event/story/verb.
4. Put a few candidate "subject verb object" main clause answers on Post-its. Pick the most exciting story - one which sounds like a good match for the requirements.
5. Use the **BEAM Canvas** provided. Add your story Post-it to the canvas header along with separate subject, verb and object Post-its in the appropriate W boxes.
6. Using the BEAM script ask for more details to add to the canvas. Remember to start each subsequent question with a W and include the main clause.
7. If the team could add just two extra Post-its for potentially valuable data not currently available what would they be?

**Business Knowledge**

Reseller sales revenue and costs are recognised on shipment of goods when a matching invoice (see example) is raised (rather than when goods are initially ordered).

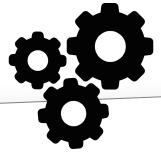
Goods are shipped to the reseller's warehouses, offices or retail outlets.

Each reseller (VAR or retailer) is assigned an account manager.

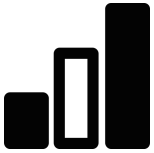
Shipments are made using a variety of carriers (e.g. FedEx, DHL) and shipment modes (e.g. standard, express, customer collection).

Pomegranate provides free shipping to resellers. Shipping costs are incurred by Pomegranate unless the mode is 'customer collection'. Whilst these costs are not printed on the invoice, they are captured by the invoicing system and allocated to invoice line items.

Product discount descriptions are known as special incentives. The combination of favourable terms and special incentives offered for each product on an invoice is referred to as 'the deal'.






# POMEGRANATE

# Invoice

**Billing Details**
**Customer #:** R34566
Media Markt
Business Park
Anytown
County XX67

**Shipping Address**
Media Markt Megastore
4321 High Street
Anothertown
County YY56

**Invoice Date** 12/07/2018
**INVOICE NUMBER:** 67890
**TERMS:** 30 Days

| Product # | Description | Quantity | Unit Price | Discount | Cost |
|---|---|---|---|---|---|
| IPIP80001 | 80Gb iPip | 200 | £100.00 | | £20,000.00 |
| IPOM24009 | 24" iPom - ***Special Incentive*** | 30 | £1,099.00 | 12.00% | £29,013.60 |
| POMB1301 | 13" PomBook - *SuperDeal* | 20 | £899.00 | 10.00% | £16,182.00 |
| PCWP3Y01 | PomCare Welcome Pack | 10 | £99.00 | | £990.00 |
| PCSCBK80 | iPip Soft Case - Black | 10 | £12.00 | | £120.00 |
| | Fedex - Next Day Express | | | | |
| | | | | Subtotal | £66,305.60 |
| | | | | Tax 20% | £13,261.12 |
| | | | | Total | £79,566.72 |

**Product Master File**

Product Number (PK)
Product Name
Product Subcategory
Product Category
Colour/Finish
Weight
Size
Capacity
Consumer List Price
Current Wholesale Price
Current Manufacturing Unit Cost
...

**f you have any questions regarding this invoice or the goods delivered,**
**lease contact you Customer Account Manager** James Bond **on:**

**Telephone :0845 600 1600 Monday-Friday 08h00-18h00**
**eMail:** jbond@pomegranatecomputer.com

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Milestones: Fixed, Variable, Repeatable/Recurring

## How

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we uniquely identify each event?*

***Verb***, Activity, Process, Event
Effect, Outcome, Status
Transaction Type
Transaction #, Event ID [Degenerate Dimension]
Step/Sequence #

[Granularity, Event Type: Discrete, Evolving, or Recurring]

## How Many

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/Deltas
Balances
Activity/Status Counts
Durations
Measure/KPI Formulas

[UoM, Fact Type: Fully Additive, Semi-Additive, Non-Additive]

## Who

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

*Subject/Object*
Customer: Business, Consumer, Segment
Employee
Supplier
Partner
Third Party

## Where

*Where does it happen? Where does it refer to?*

*Subject/Object*
Location
Branch, Store, Facility
Channel
URL
Map/Sequence: First → Previous → Current → Next → Last

## Why

*Why does it happen? Why do quantities vary?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

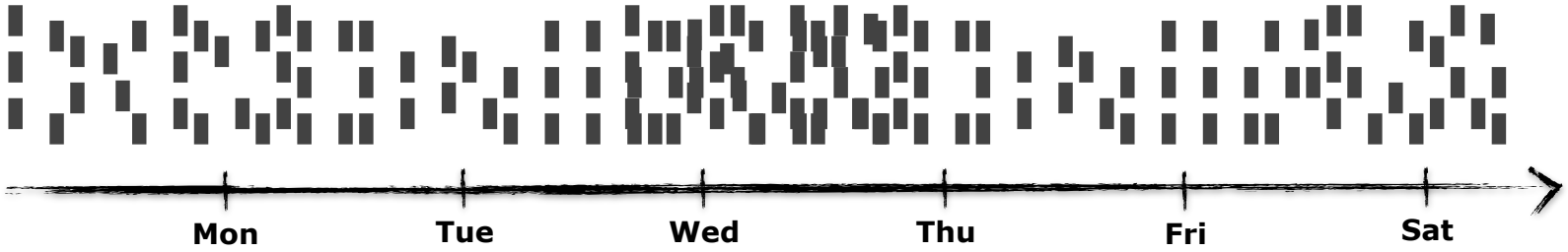
## What

*What is involved/used? How are they organized?*
*How do they change?*

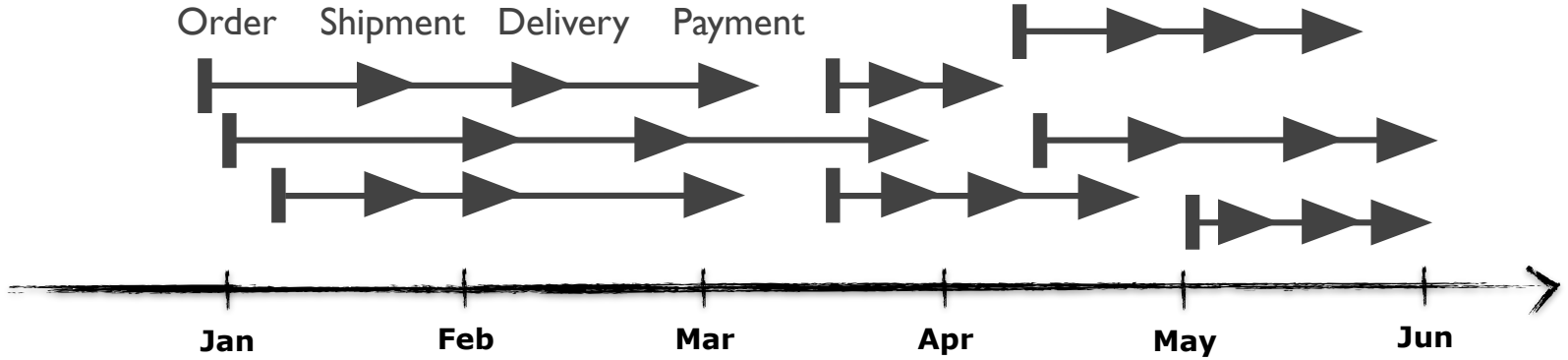
*Subject/Object*
Value Proposition
Product
Service
Resource
Item

# Story Type Timelines

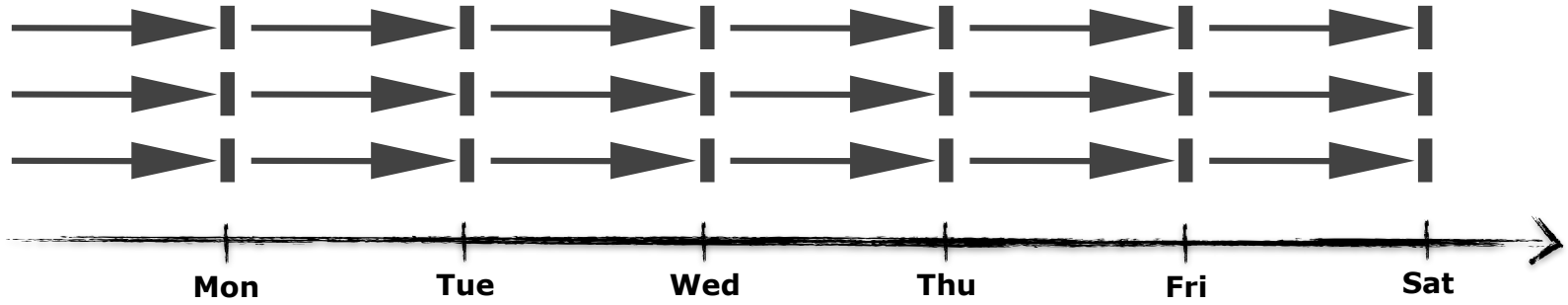
Discrete
[ DE ]

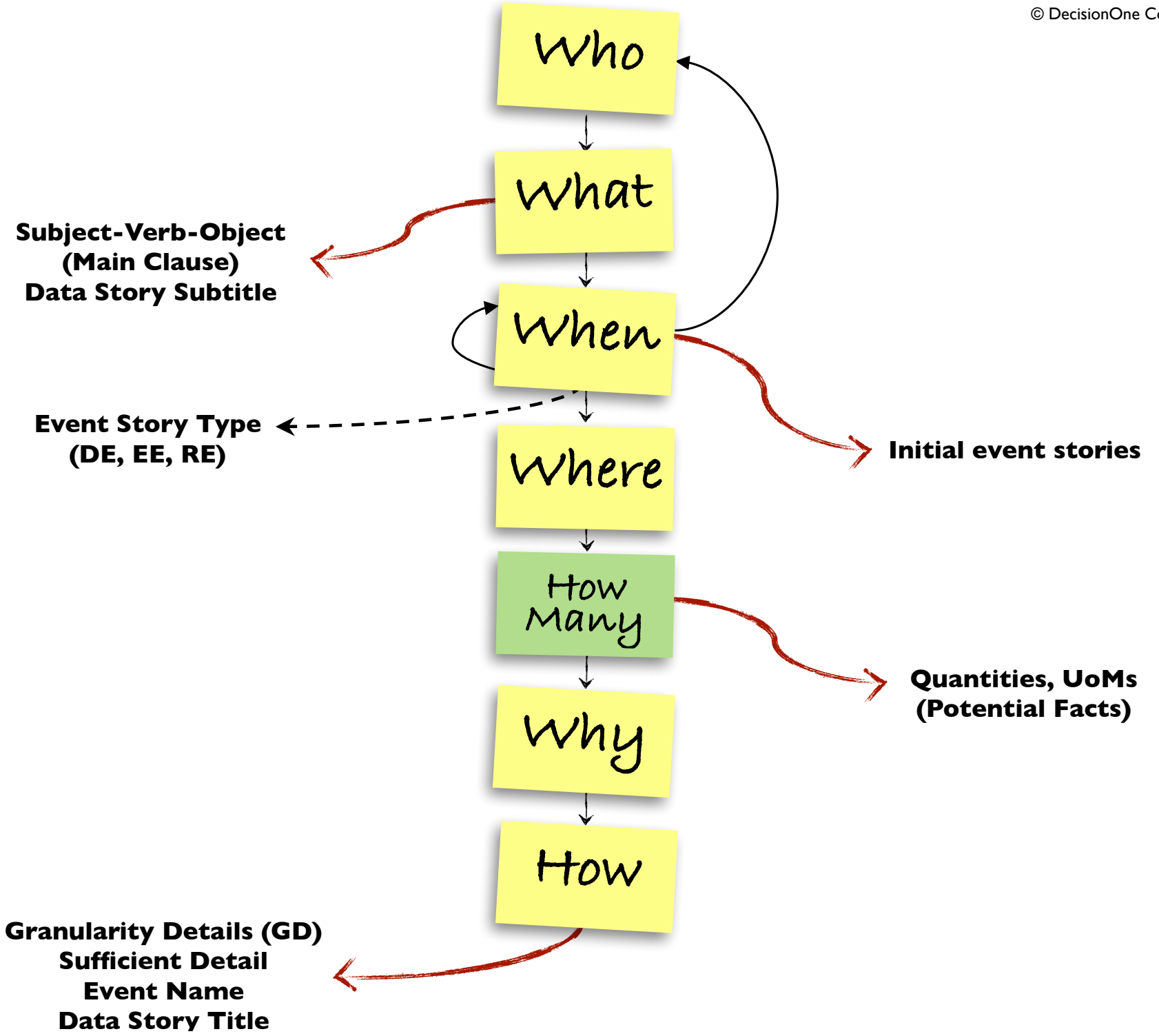



Evolving
[ EE ]

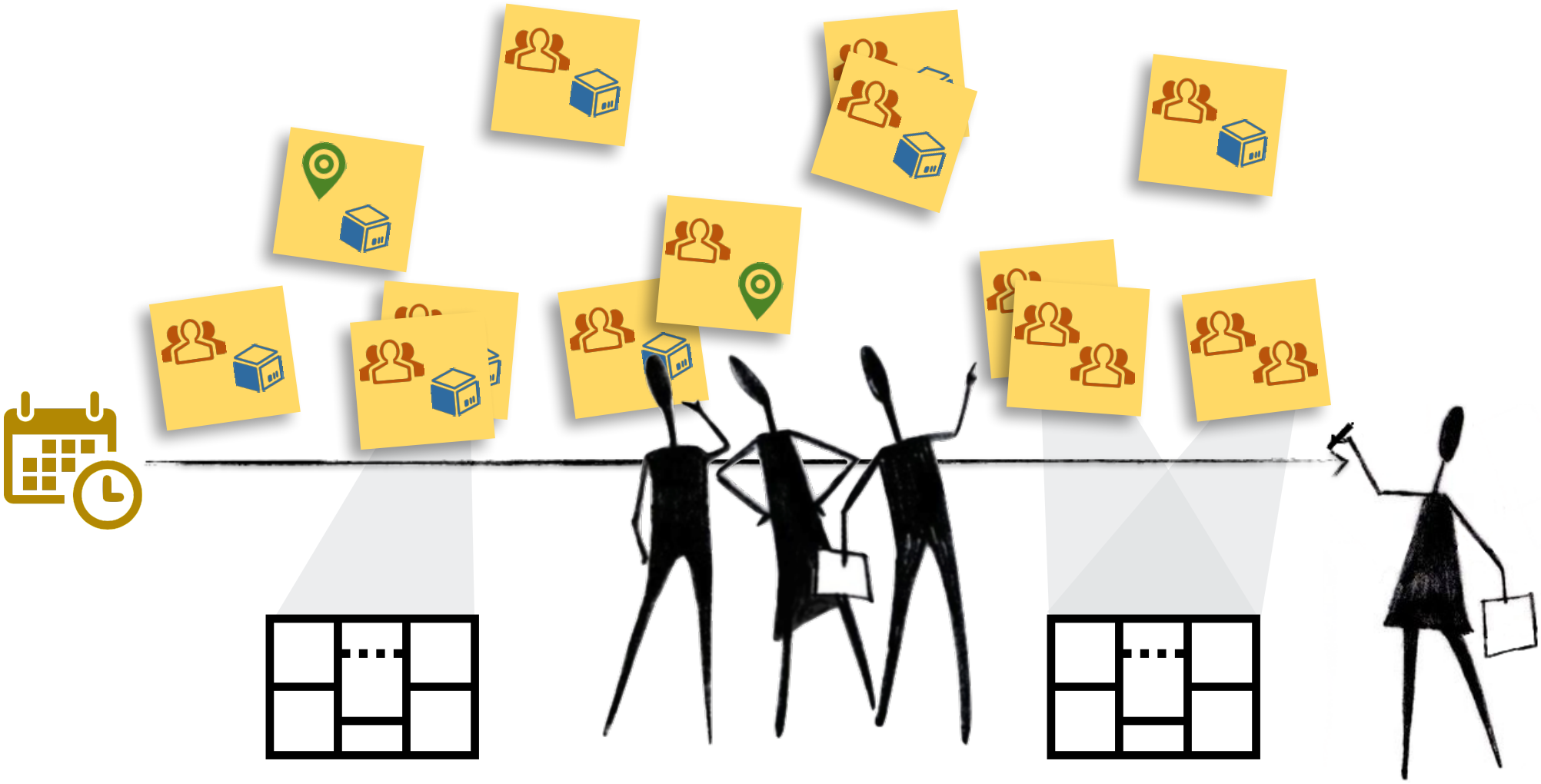



(Regularly)
Recurring
[ RE ]





# Sort Multiple Stories on a Timeline



Other

Who did what?

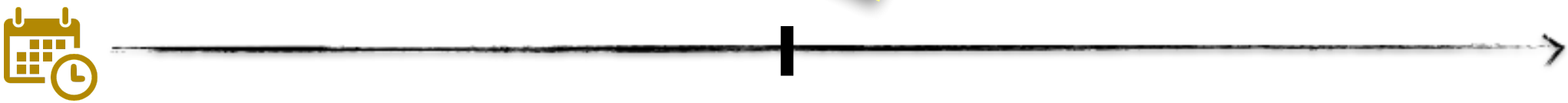
Who does what?

Who will do what?

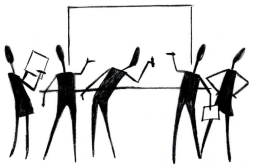
Exception

Customer buys Product

Normal



First? Previously? Before? During? After? Next? Last? Future?



# Exercise 3: Event Timeline

**Requirements**

Pomegranate wish to expand the current reseller focused design to include corporate direct sales. Initial user requirements and analysis of CRM/sales force automation systems have identified the following activities/processes which must be measures:

- **Consulting Billing** – Analysis of ongoing revenue and commission arising from consulting orders
- **Mass Communication** – e-mail invitations to events, product announcements, newsletters, etc.
- **Event Management** – organisation, attendance and cost information on seminars, webinars, conferences and shows etc.
- **Orders** – for products, support contracts, license upgrades and consulting engagements
- **Product Support** – installation, maintenance, tracking open and closed support issues through call-centre and field engineers
- **Prospect Gathering** – web site visits, purchased name lists, badge swipes at shows, referrals from partners, etc.
- **Sales Calls** – conference calls, on-site visits, demonstrations

Develop a Corporate Direct Sales timeline (as below) to document pre and post sales chronology and help identify the most interesting events.

**Part 1 (Individually in silence)**

Create "who does what" data stories for each of the subject areas above on separate post-its. Use your favourite colour notes.

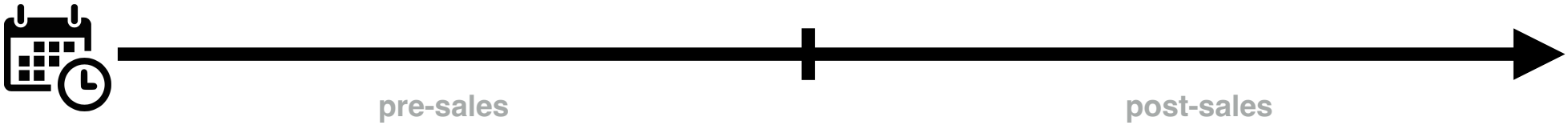
When you have 7+ add your data stories to the team timeline in sales pipeline time/value sequence. Try to space them out relative to one another.

**Part 2 (Team)**

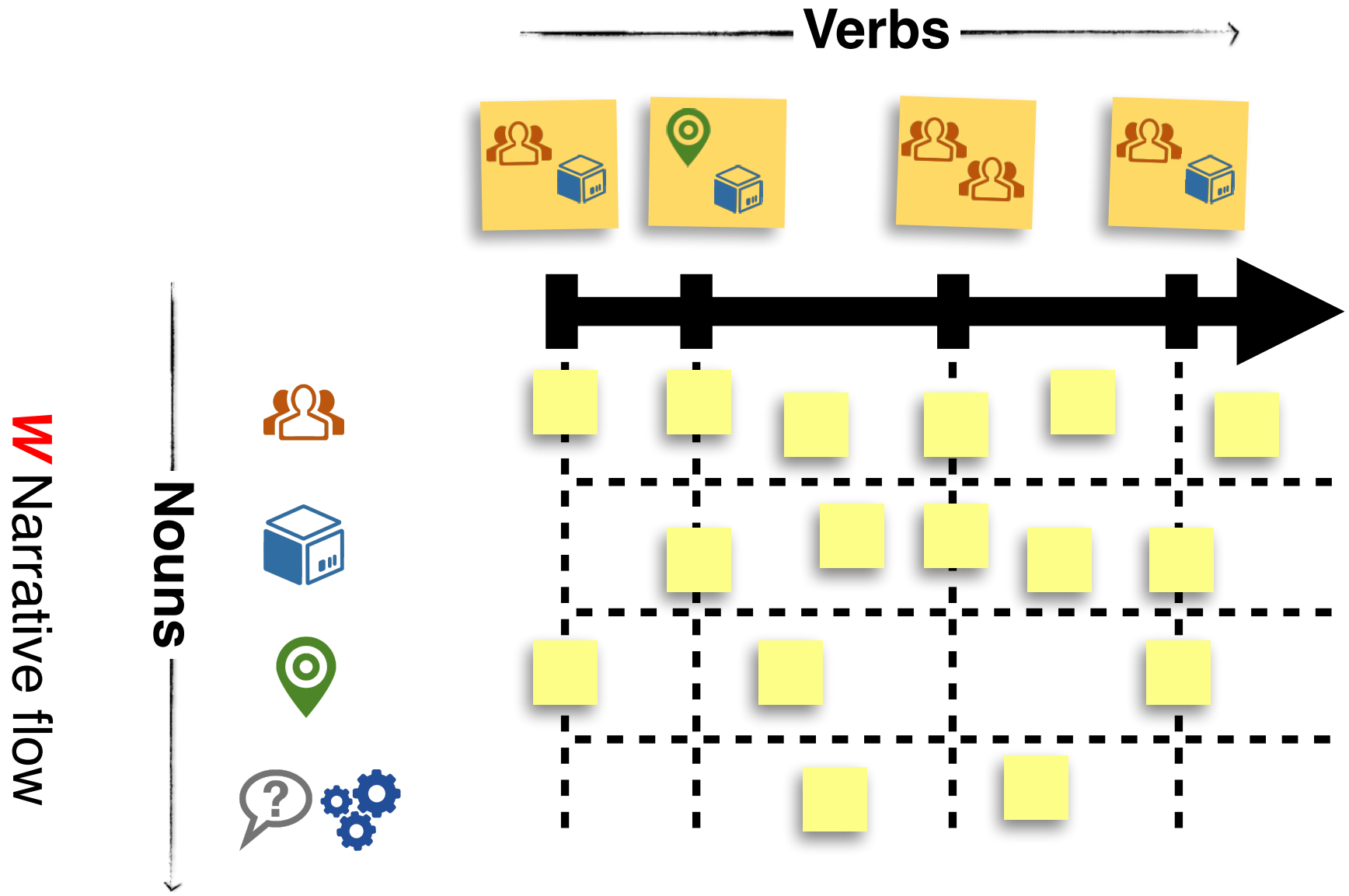
Re-sort and de-duplicate stories collectively. Stick stories with identical subject, verb, object combinations on top of each other then move on to less obvious synonyms events. Popular stories should show up as stacks.

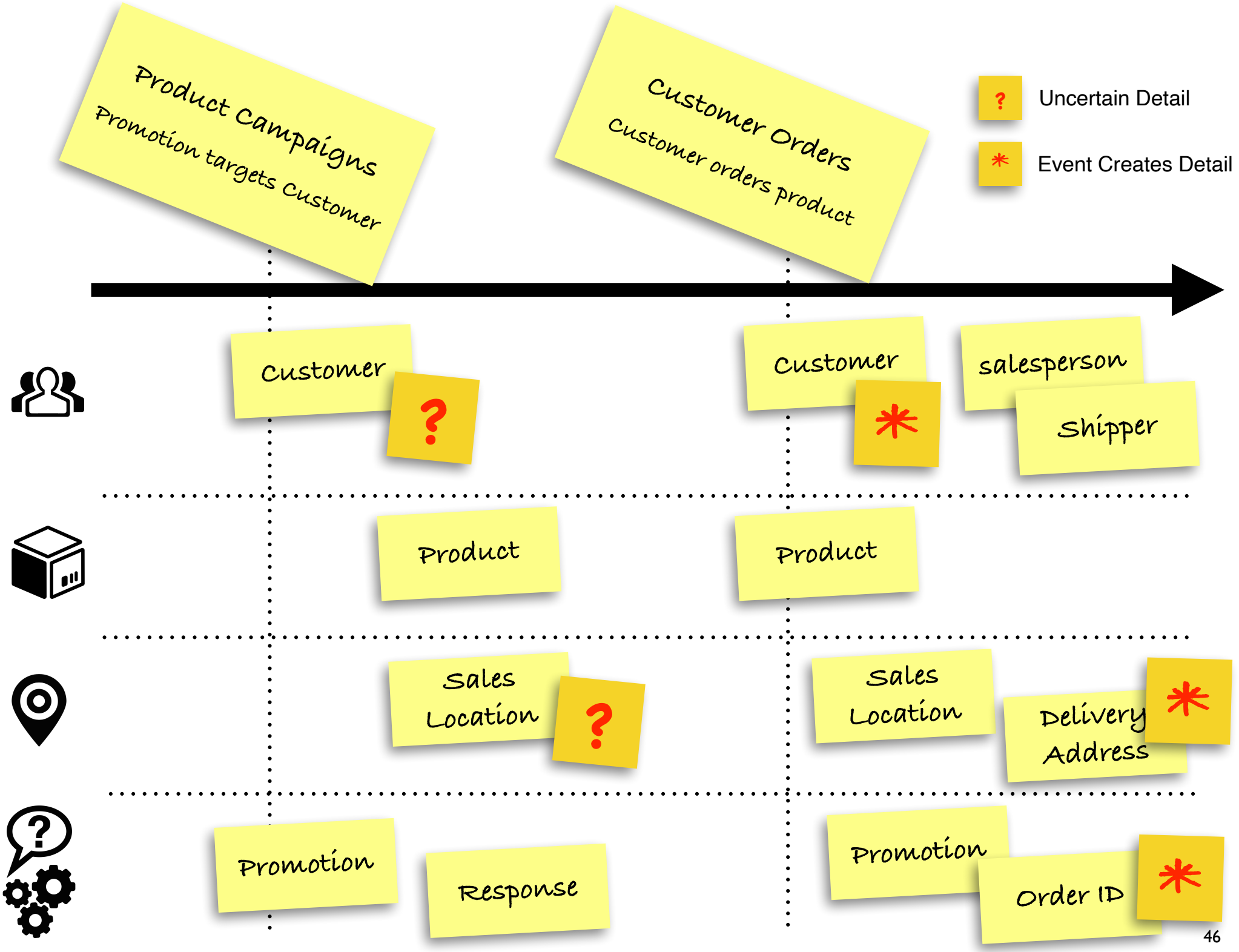
**Part 3 (Team)**

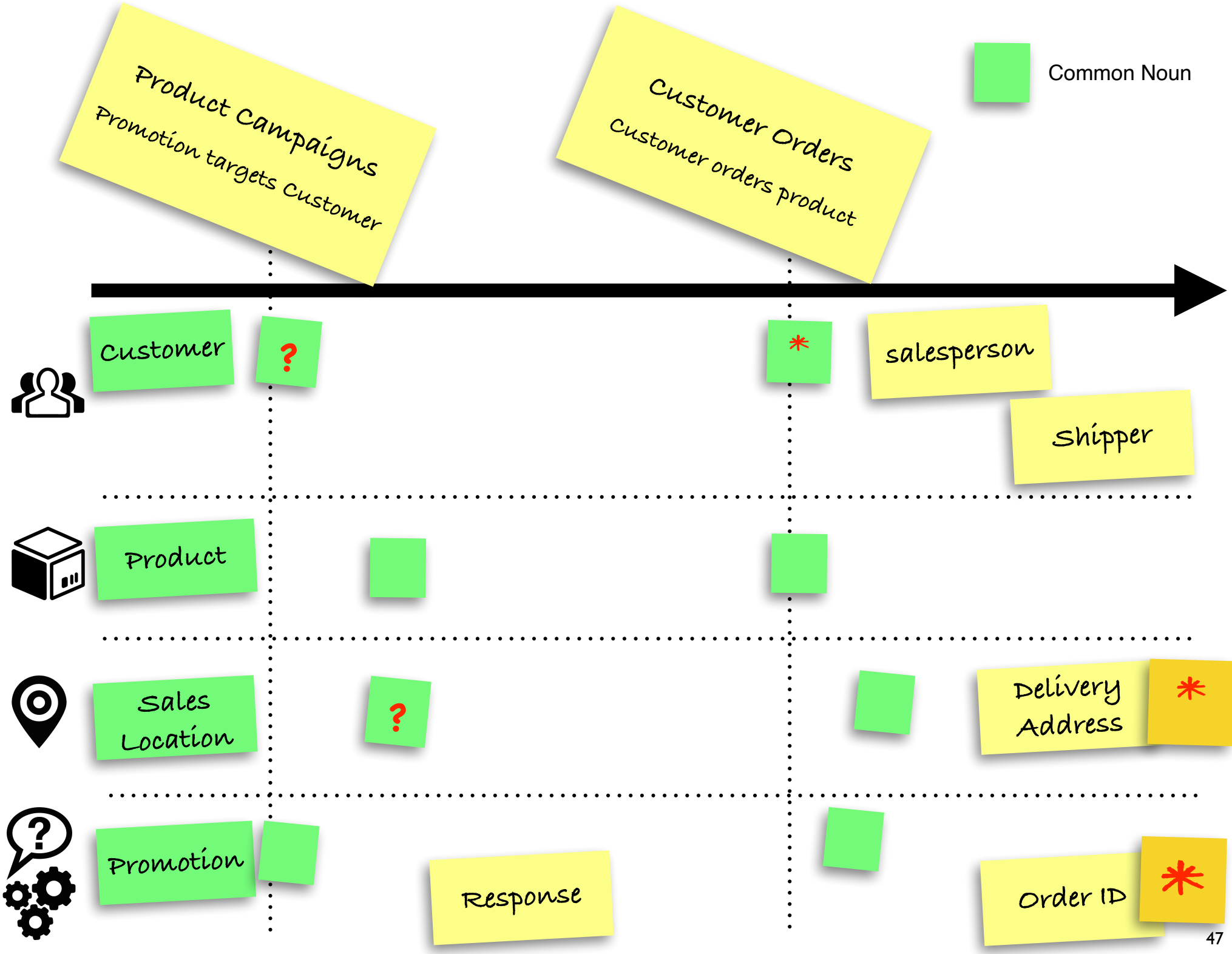
Examine the 'value chain', can you spot any missing 'links' – long gaps in time or large jumps in value? What events might fill these gaps. Concentrate on corporate direct sales force activity - **avoid events that would be handled by manufacturing, logistics, finance**. Identify the most important activity worth measuring not in the initial requirements.



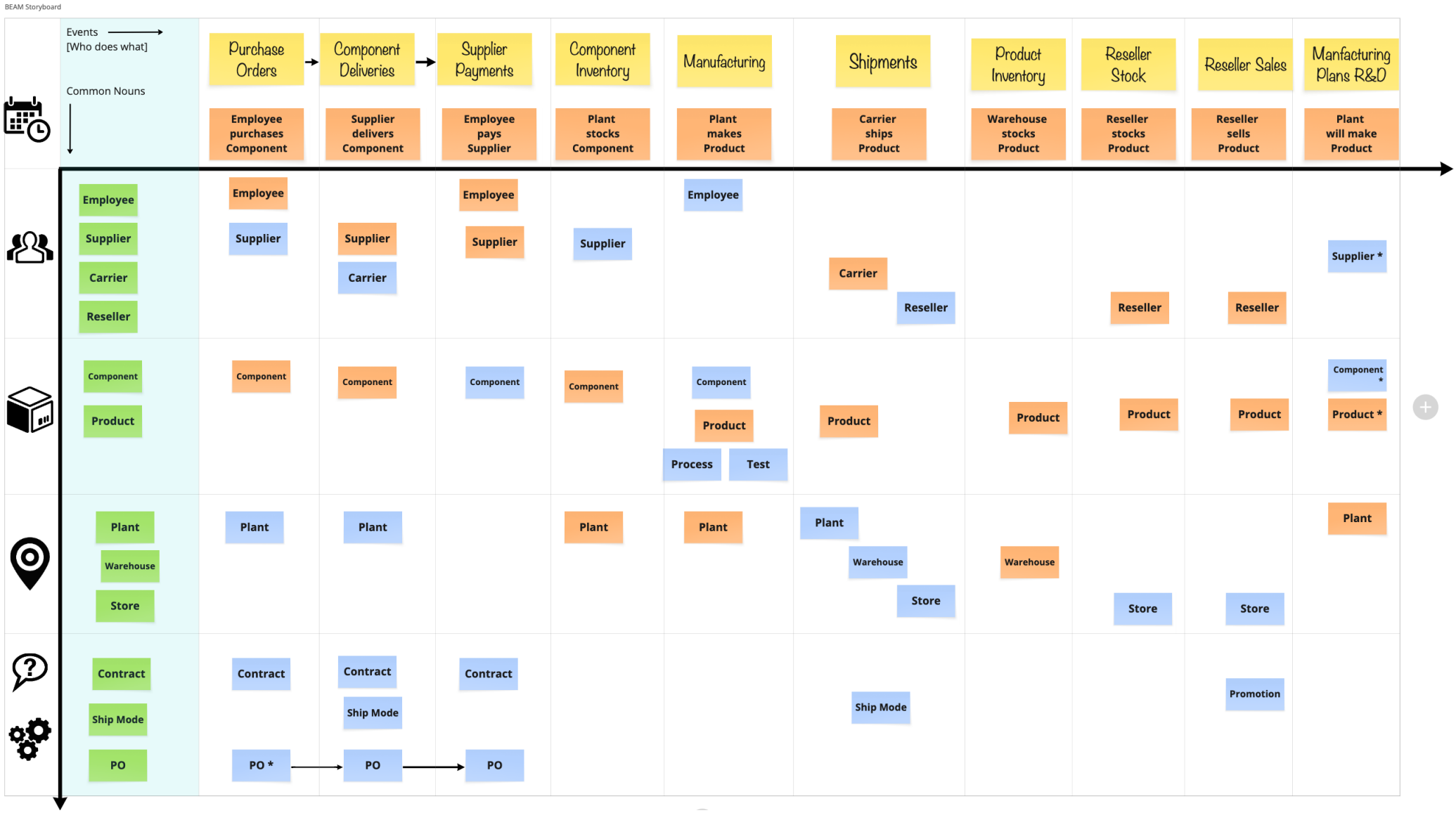
# BEAM Storyboard - for discovering common nouns

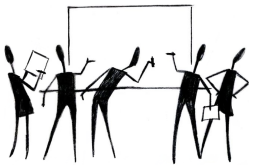


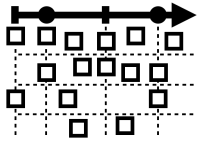

Product Campaigns
Promotion targets Customer
Customer Orders
Customer orders Product
? Uncertain Detail
* Event Creates Detail
Customer
?
Customer
*
salesperson
Shipper
Product
Product
Sales Location
?
Sales Location
Delivery Address
*
Promotion
Response
Promotion
Order ID
*


Product Campaigns
Promotion targets Customer
Customer Orders
Customer orders Product
Common Noun
Customer
?
*
salesperson
Shipper
Product
Sales Location
?
Delivery Address
*
Promotion
Response
Order ID
*

# Example BEAM Storyboard: Manufacturing





# Exercise 4:  BEAM Storyboard

**Requirements**

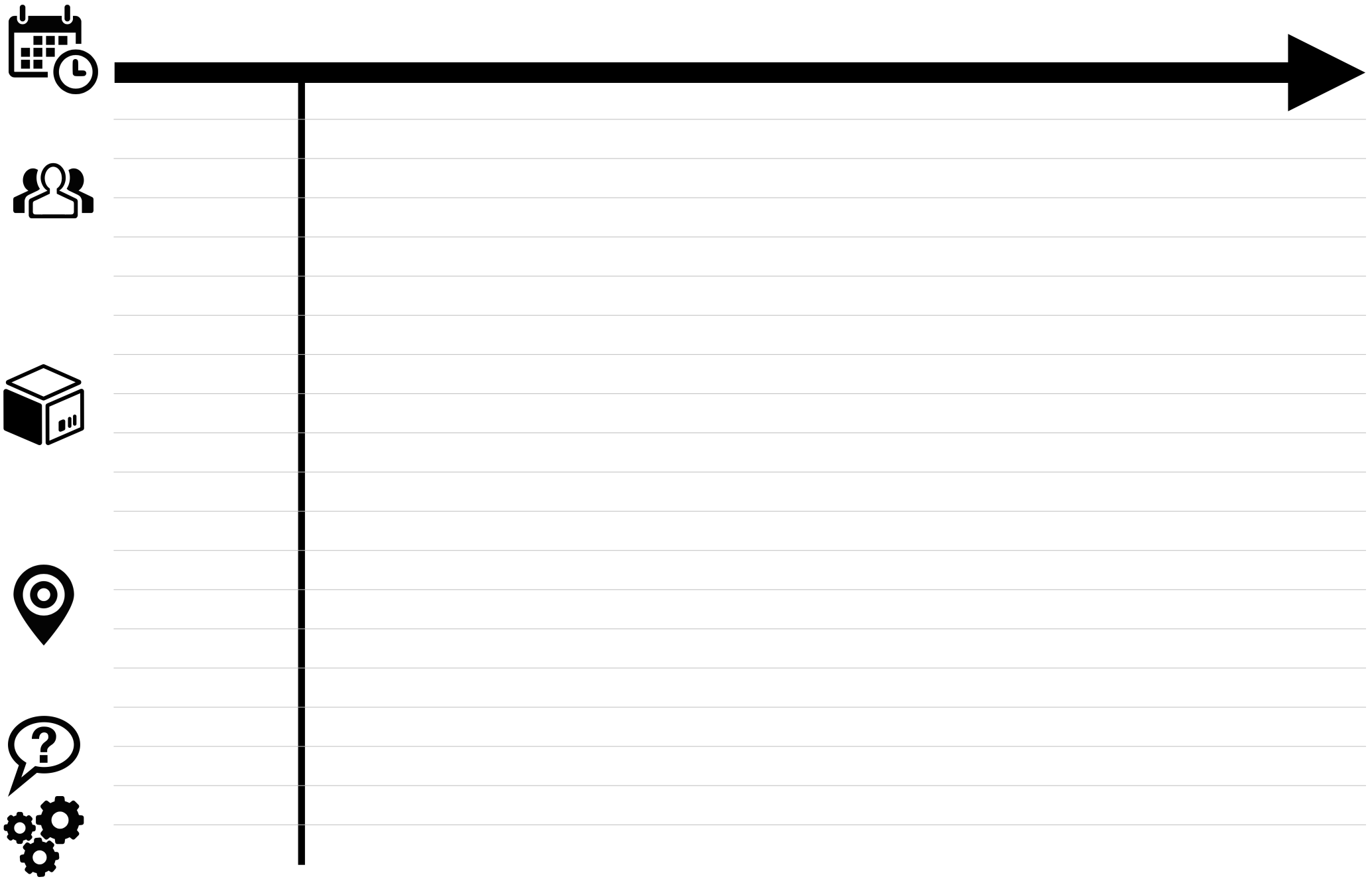
Develop a high-level storyboard for corporate direct sales to identify common nouns that appear in multiple business events.

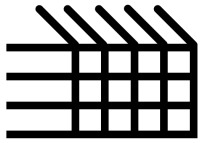
**Part 1**

1. Convert your corporate sales timeline into a BEAM storyboard by adding a W axis descending below it.
2. Add subject and object Post-its directly below each event within their appropriate W lanes.
3. Using the BEAM script fill out your stories with additional noun Post-its below and to the right of each event. These should include **Prospect** and **Area of Interest** (terms used by key stakeholders).
4. As common nouns emerge move them to become row headers and tick the events they feature in.
5. If you are unsure that a noun is involved in an event mark it with a **?**. If an event creates new data for any noun mark it with a *****.

**Part 2**

Are any nouns so similar (same type: *who, what, where*…) that they might be used in place of one another? If so line them up horizontally and create a common noun row header for them.





# use an Event Matrix to discover Common Nouns

**Nouns**

***W*** Narrative flow →

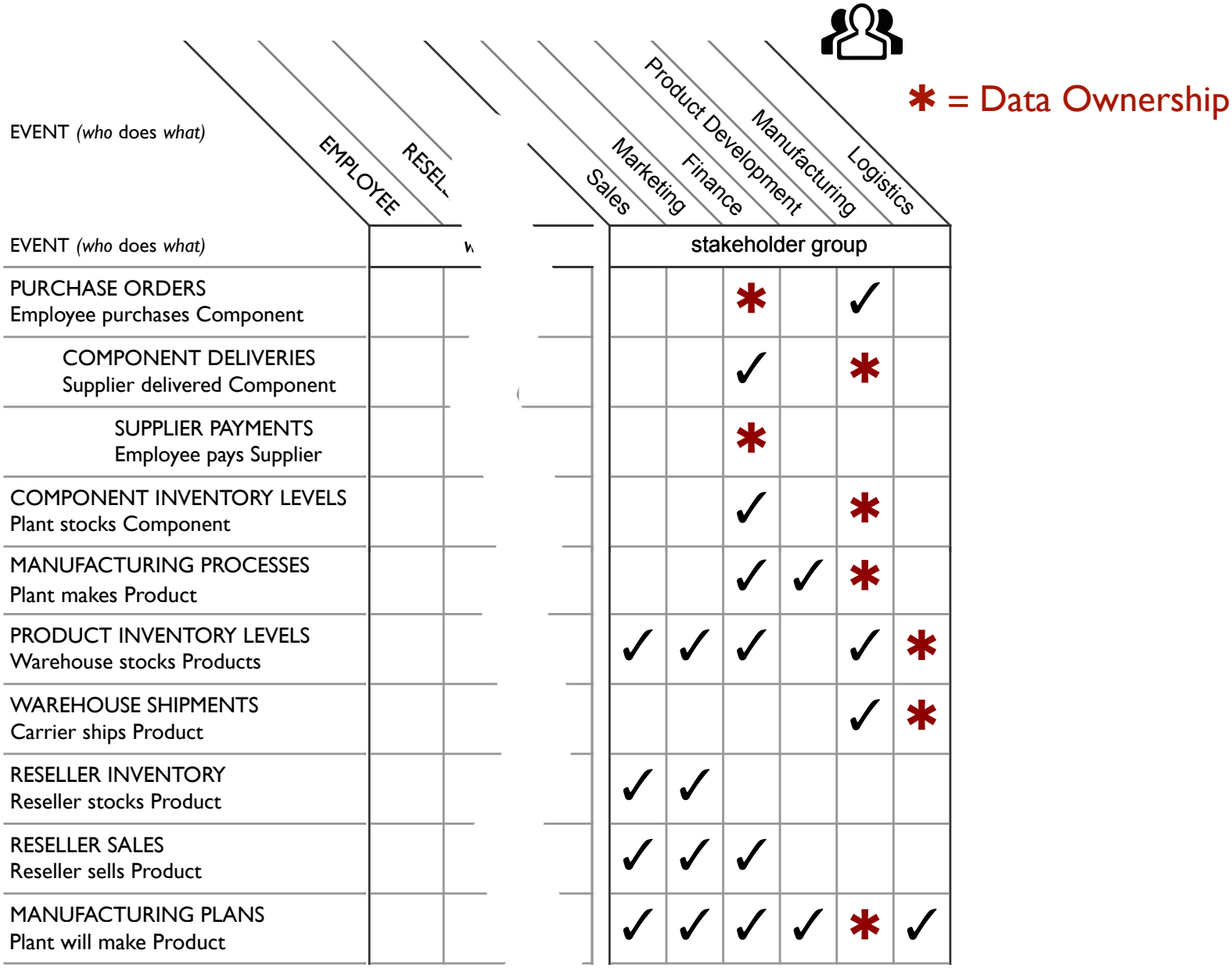
Verbs ↓

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ✔ | | ✔ | | | ✔ | | | ✔ | ✔ | | ✔ | | ✔ |
| | | | | | ✔ | | ✔ | | ✔ | | | | ✔ |
| | ✔ | | ✔ | | | ✔ | | | ✔ | | | | |
| | ✔ | | ✔ | | | ✔ | ✔ | | | | ✔ | | |
| ✔ | ✔ | | | | | | | ✔ | ✔ | | ✔ | | ✔ |

# Manufacturing Event Matrix - Chapter 4

| EVENT *(who does what)* | EMPLOYEE | SUPPLIER | RESELLER | CARRIER | COMPONENT | PRODUCT | PROCESS | TEST | PLANT | WAREHOUSE | STORE | CONTRACT | PROMOTION | SHIP MODE | PO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *who* | | | | | *what* | | | | *where* | | | *why & how* | | |
| PURCHASE ORDERS<br>Employee purchases Component | ✓ | ✓ | | | * | | | | ✓ | | | ✓ | | | * |
| COMPONENT DELIVERIES<br>Supplier delivered Component | | ✓ | | ✓ | ✓ | | | | ✓ | ? | | ? | | ✓ | ✓ |
| SUPPLIER PAYMENTS<br>Employee pays Supplier | ✓ | ✓ | | | ✓ | | | | | | | ? | | | ✓ |
| COMPONENT INVENTORY LEVELS<br>Plant stocks Component | | ✓ | | | ✓ | | | | ✓ | | | | | | |
| MANUFACTURING PROCESSES<br>Plant makes Product | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| PRODUCT INVENTORY LEVELS<br>Warehouse stocks Products | | | | | | ✓ | | | | ✓ | | | | | |
| WAREHOUSE SHIPMENTS<br>Carrier ships Product | | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| RESELLER INVENTORY<br>Reseller stocks Product | | | ✓ | | | ✓ | | | | | ✓ | | | | |
| RESELLER SALES<br>Reseller sells Product | | | ✓ | | | ✓ | | | | | ✓ | | ✓ | | |
| MANUFACTURING PLANS<br>Plant will make Product | | * | | | * | * | | | ✓ | | | | | | |

# Organisational Event Matrix




✱ = Data Ownership

# Event Matrix Planning

| EVENT *(who does what)* | Importance | EMPLOYEE | RESELLER | CARRIER | COMPONENT | PRODUCT | PROCESS |
|---|---|---|---|---|---|---|---|
| | | *who* | | | *what* | | |
| **Importance** | | ~~310~~ | **410** | | ~~320~~ | **420** | ~~330~~ |
| MANUFACTURING PROCESSES<br>Plant makes Product | 300 | ✓ | | | ✓ | ✓ | ✓ |
| RESELLER SALES<br>Reseller sells Product | **400** | | ✓ | | | ✓ | |
| MANUFACTURING PLANS<br>Plant will make Product | 200 | | | | | ✓ | |

# Modelstorming by Example

Event/Title | Subject Verb Object | Author(s), Date/Version

Customer buys Product

Elvis Priestly

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Milestones: Fix

Purchase Date

7/2/2021 (Yesterday)

## How

*How (exactly) does it happen?*
*How do we know it happened?*

Invoice #999

Invoice

## Who

***Who does what***
*How do they chang*

Subject/Object
Customer: Business, Consumer, Se
Employee
Supplier
Partner
Third Party

Customer

## How Many

*How many/much is involved? How long*

Quantities
Revenues
Costs
Discounts/Deltas
Balance
Activit
Durati
Measu

[UoM

2 iPips

Salesperson

James Bond

Quantity [items]

Revenue [$]

$250

## Where

*Where does it happen? Where does it refer to?*

Subject/Object
Location
Branch, Store, Facility
Channel
URL
Ma

Store/URL

Melbourne

## What

*What is involved/used? How are they organized?*
*How d*

Subject
Value F
Produ
Servic
Resou
Item

Product

Blue Suede iPip

Promotion

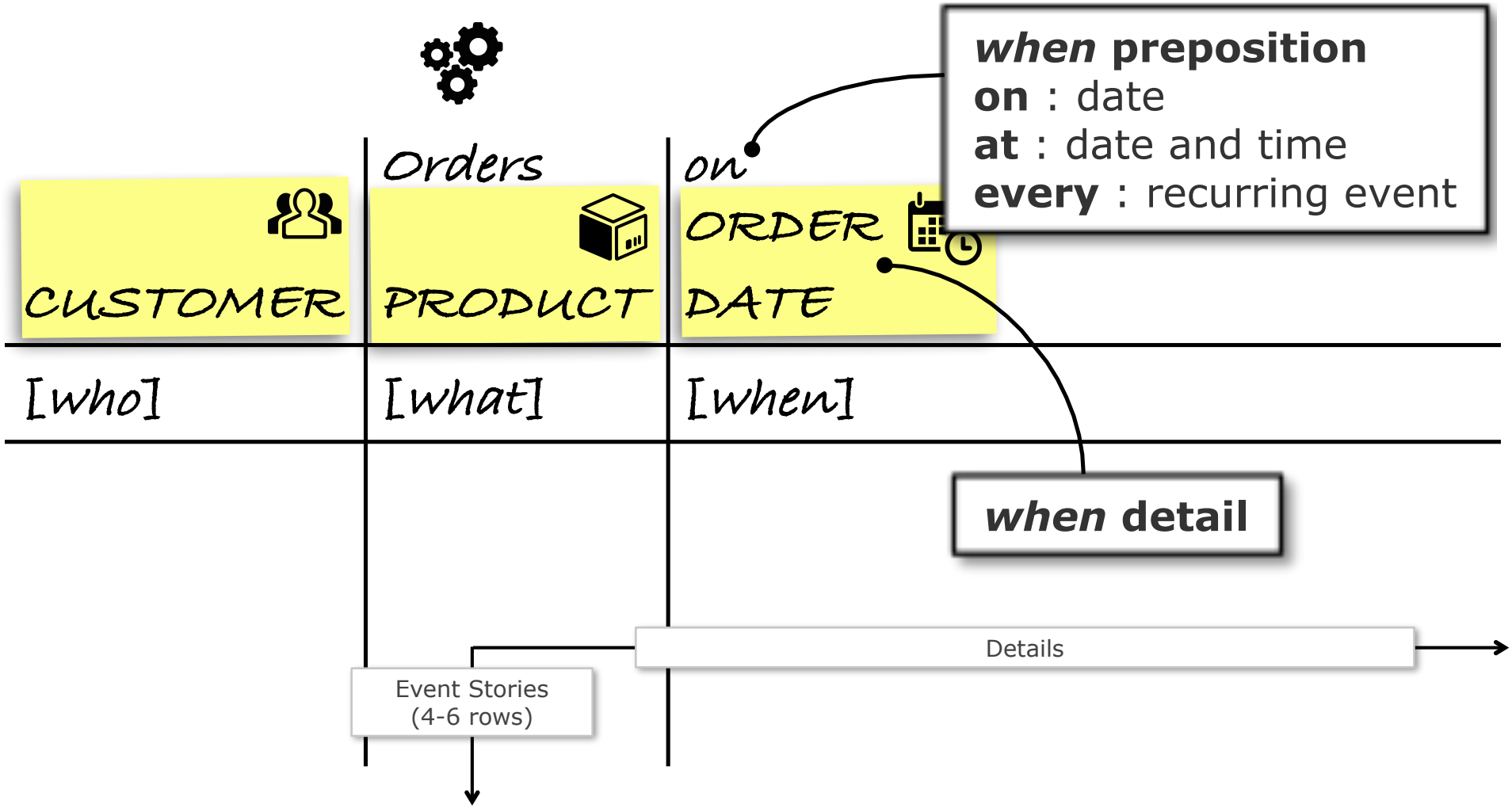
Two for One 50% off

BEAM* Canvas

modelstorming.com · Agile Data Warehouse Design

*Business Event Analysis and Modeling

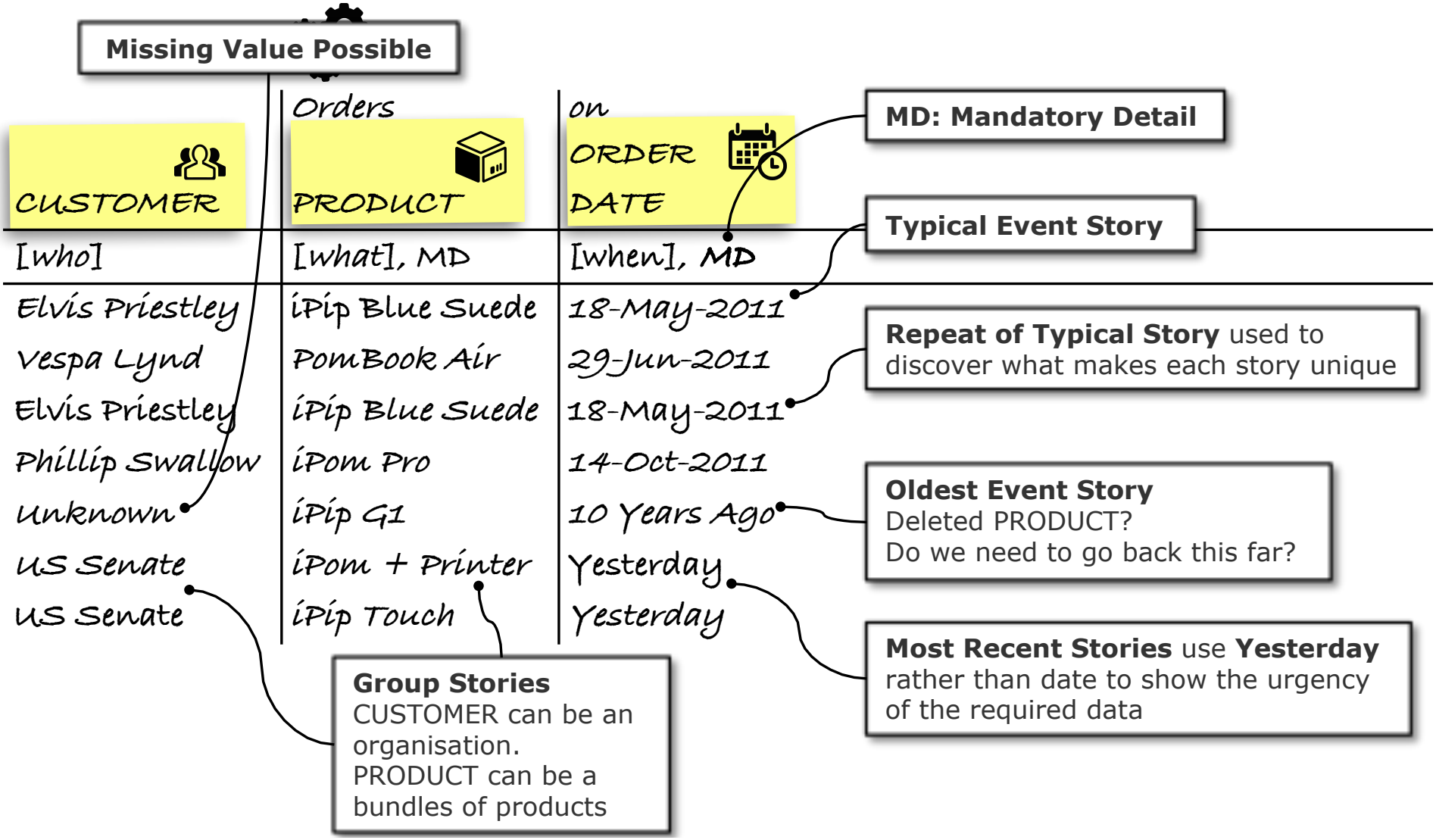
# BEAM Table - Modeling by Example



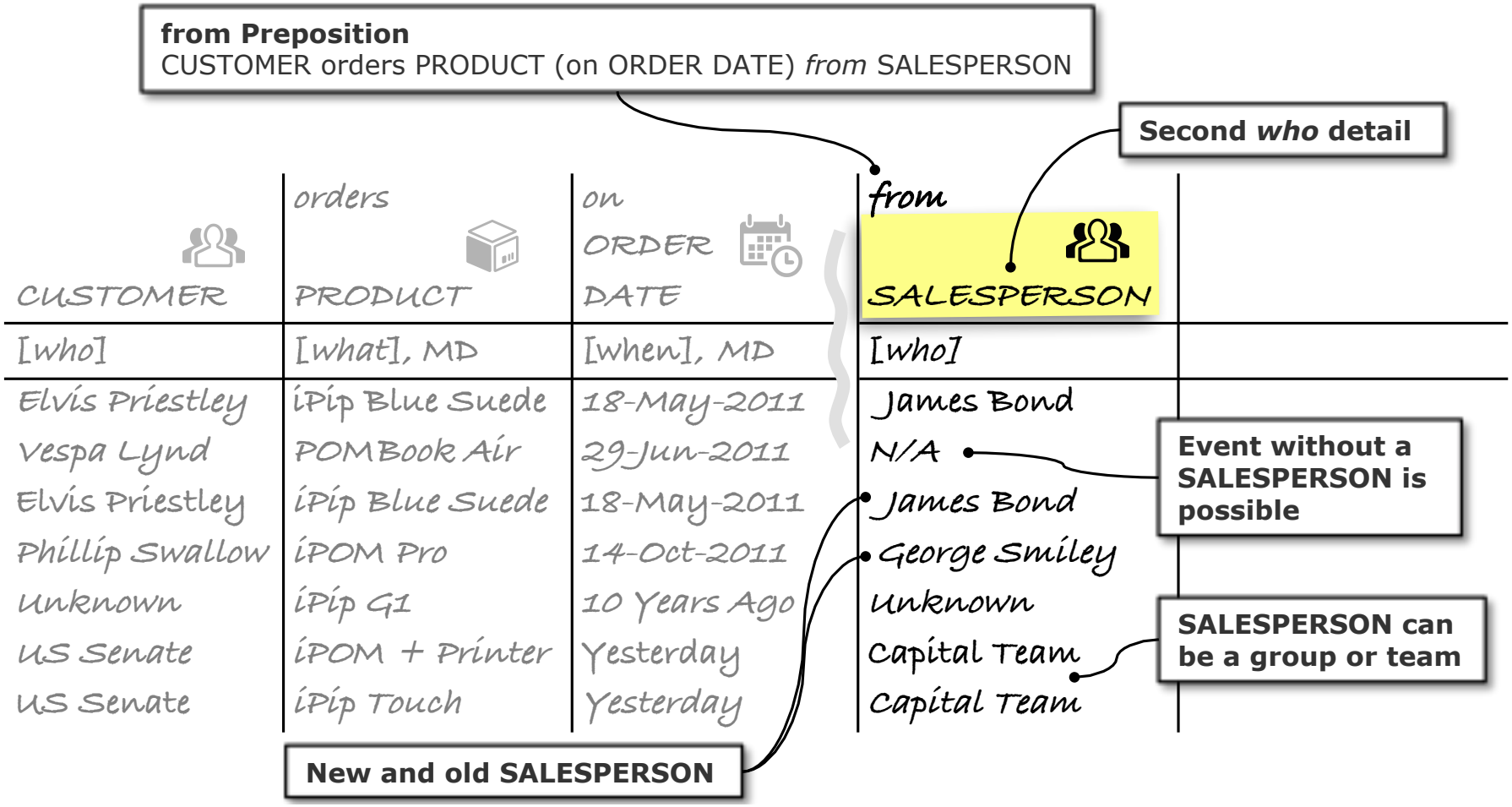
# Event Story Themes

| SUBJECT | verb<br>OBJECT | on/at/every<br>EVENT DATE (TIME) | | | | |
|---|---|---|---|---|---|---|
| [who] | [what] | [when] | [where] | [unit of measure] | [why] | [how] |
| ***Typical*** | Typical/Popular | Most Recent | Near | Typical/Average | Normal | Normal |
| ***Different*** | Different | Past/Future | Far | high/low/0/negative | Exceptional | Exceptional |
| ***Repeat*** | Repeat | Repeat | Repeat | Repeat | Repeat | Repeat |
| ***Missing*** | Missing | Missing | Missing | Missing | Missing | Missing |
| ***Group*** | Multiple/Bundle | Period | Multiple Levels | Allocations | Multiple Values | |

# Example Stories

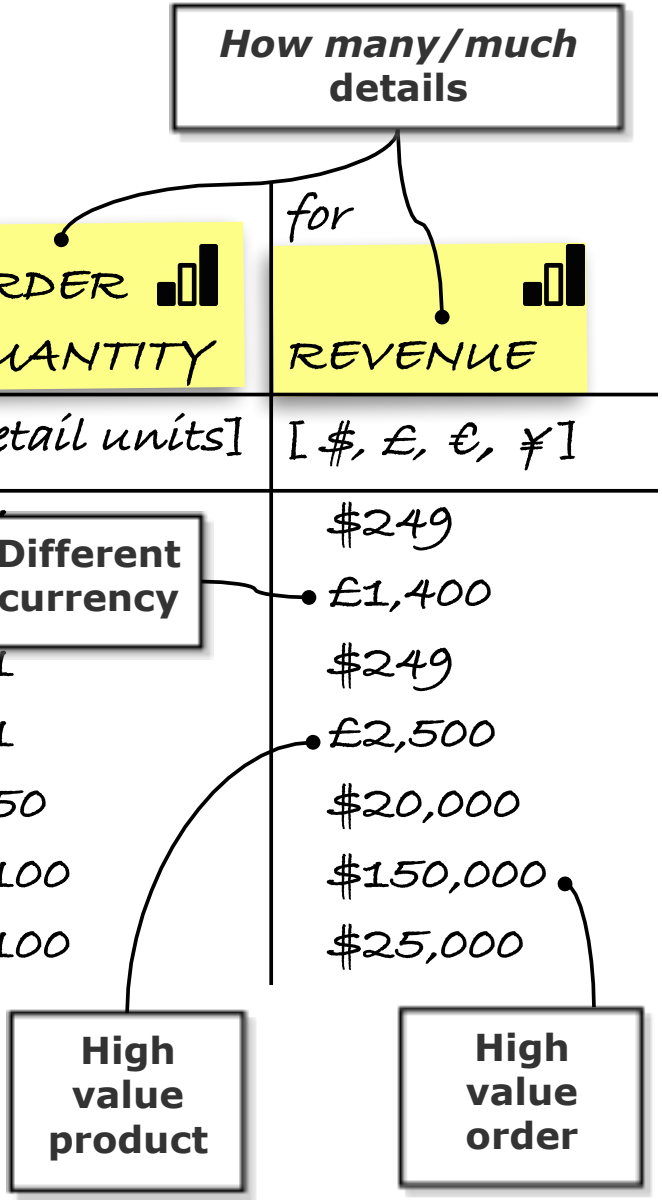


# Adding a Second Who Detail



# Adding How Many Details: Quantities, Units of Measure

| CUSTOMER | Orders PRODUCT | on ORDER DATE | ORDER QUANTITY | for REVENUE |
|---|---|---|---|---|
| [who] | [what], MD | [when], MD | [retail units] | [$, £, €, ¥] |
| Elvis Priestley | iPip Blue Suede | 18-May-2011 | | $249 |
| Vespa Lynd | PomBook Air | 29-Jun-2011 | | £1,400 |
| Elvis Priestley | iPip Blue Suede | 18-May-2011 | 1 | $249 |
| Phillip Swallow | iPom Pro | 14-Oct-2011 | 1 | £2,500 |
| Unknown | iPip G1 | 10 Years Ago | 50 | $20,000 |
| US Senate | iPom + Printer | Yesterday | 100 | $150,000 |
| US Senate | iPip Touch | Yesterday | 100 | $25,000 |

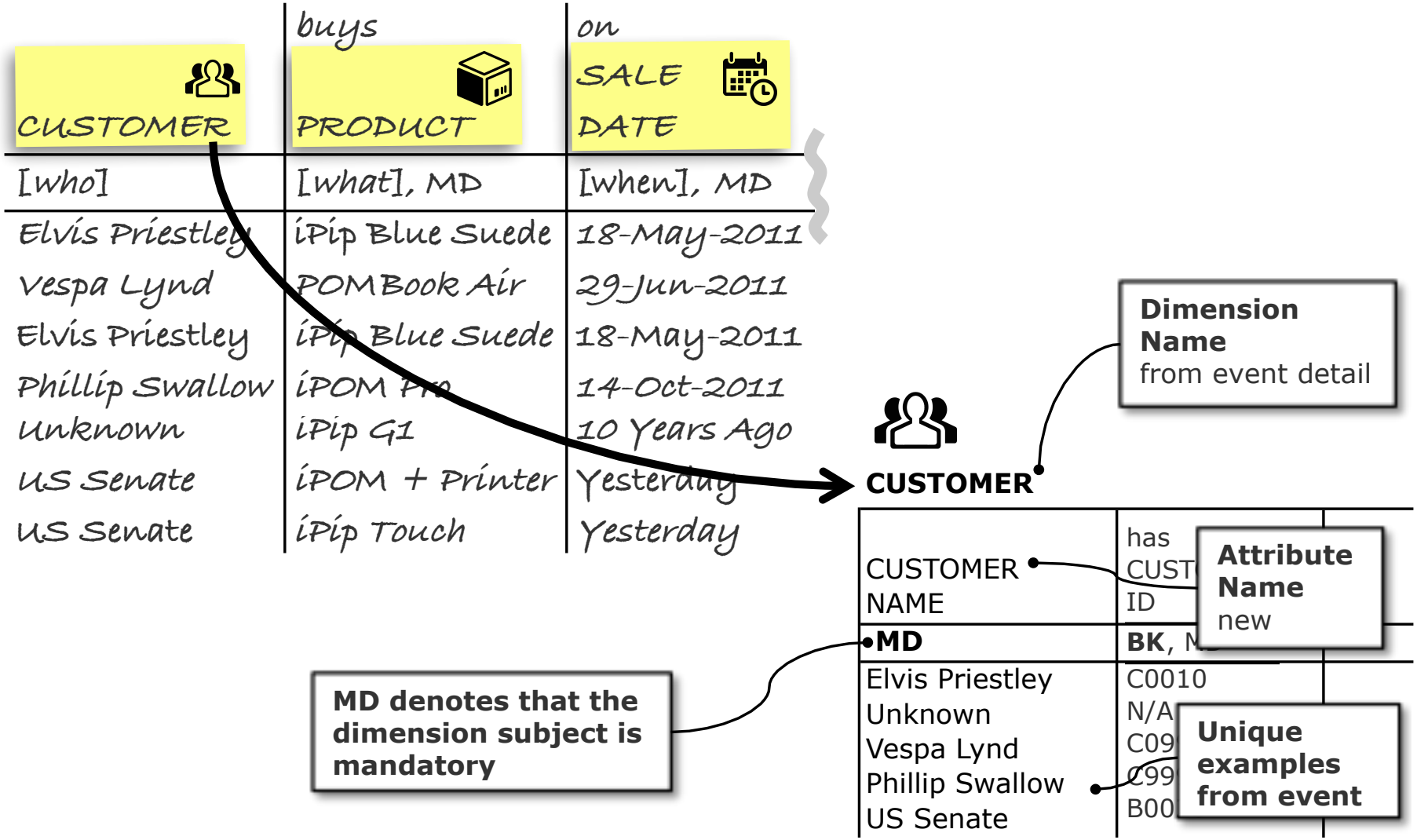


# Exercise 2 Example BEAM Table

| ACCOUNT SALES [DE] | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **sells** | on | to | for shipping to | of | at | with |
| ACCOUNT MANAGER | PRODUCT | INVOICE DATE | RESELLER | SHIPPING ADDRESS | QUANTITY | PRICE | DISCOUNT |
| [who] MD | [what] MD, GD | [when] MD | [who] MD | [where] MD | [units] | [£,€,$] | [%] |
| James Bond | 80Gb iPip | 2/4/2008 | Best Buy | NYC | 200 | £100 | 0% |
| JB | PomBook | Yesterday | Staples | Montreal | 20 | £1000 | 10% |
| George Smiley | PomServer | 5/1/2010 | PC World | Birmingham | 2 | £2000 | 50% |

| for | with | with | with | on | with | using | on |
|---|---|---|---|---|---|---|---|
| REVENUE | MANUFACTURING COST | SHIPPING COST | TAX | SPECIAL INCENTIVE | TERMS | SHIPPING MODE | INVOICE NUMBER |
| [£,€,$] | [£,€,$] | [£,€,$] | [%] | [why] | [why] | [how] | [how] GD |
| £20K | £12K | £500 | 17.5% | None | 30 Days | FedEx Express | 67890 |
| £18k | £10K | 0 | 17.5% | New Customer | Cash | Customer Collection | 56778 |
| £1000 | £1200 | £100 | 20% | BOGOF | 60 Days | DHL Next Bus Day | 89009 |

# Modelstorming Nouns and Adjectives (Dimensions)



# Using the *7Ws* to discover Descriptive Attributes

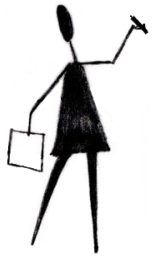
| BEAM* Modeler Question |  |  Stakeholder Answer |
|---|---|---|
| ***Who/what*** is a customer? | | Consumer, business, charity (Customer Type) |
| ***What*** is a product? | | Computer, accessory, software, service (Product Type) |
| ***Who*** else is associated with a customer? | | Primary Contact, Spouse, Sponsor, Decision Maker, Owner, Referrer |
| ***Who*** is associated with a product? | | Manufacturer, Distributor, Supplier, Marketer, Promoter, Product Manager, Inventor, Designer, Developer, Author |
| What dates (***whens***) are important to know about a customer? | | Birth Date, Graduation Date, First Purchase Date, Last Purchase Date, Renewal Date |
| What milestone dates (***whens***) are there for a product? | | Launch Date, Arrival of First Competitor, Patent Expiration Date, Discontinued Date |
| ***Where*** are customers? | | Headquarters, Sales Regional, Work Address, Home Address, Nearest Branch |
| What geographic (***where***) information describes a product or service? | | Country of Origin, Manufacturing Plant, Language, Market, Voltage |
| Are there any single-valued quantities (***how many details***) that describe or group customers? | | Life Time Value, Loyalty Score, Current Balance, Number of Employees, Number of Dependents |
| What quantities (***how many details***) describe products? | | Weight, Size, Capacity, List Price |
| ***Why*** or ***how*** do customers become customers? | | Channel, Prospect Source, Referral |

# Description Discovery Questions

How do you **describe** a [*noun*]?

What would you rather see on a report instead of a [*noun*] name or ID?

[*candidate attribute*]



Can a [*noun*] have more than **one** [*candidate attribute*] at *any one moment in time*?

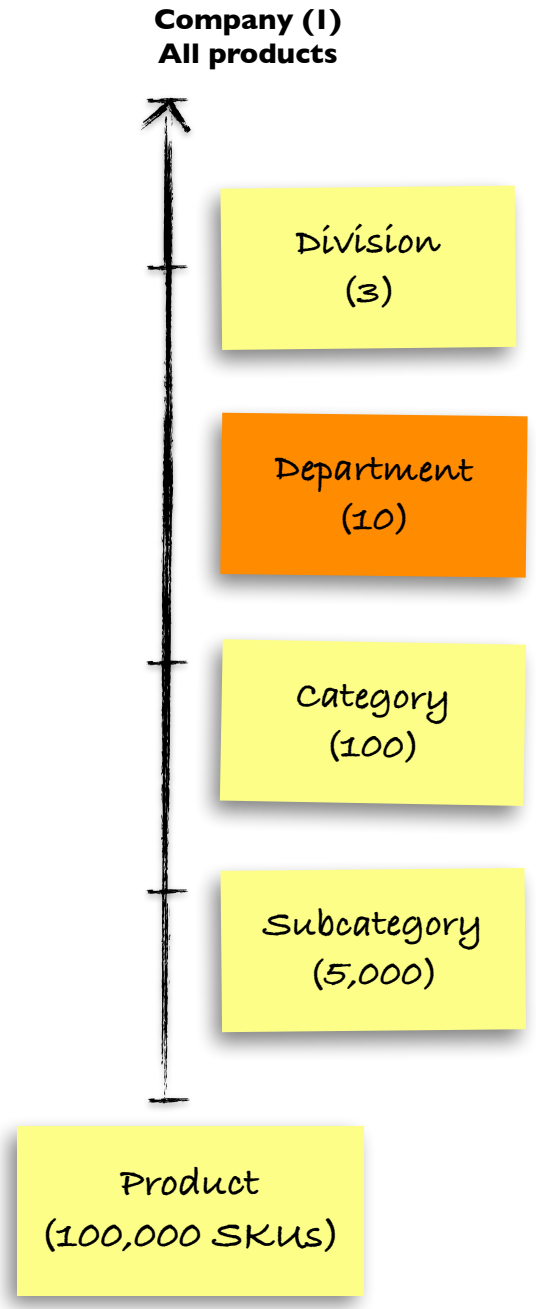
***No!***

How do you **organize** [*noun*]s or [*attribute*]s?

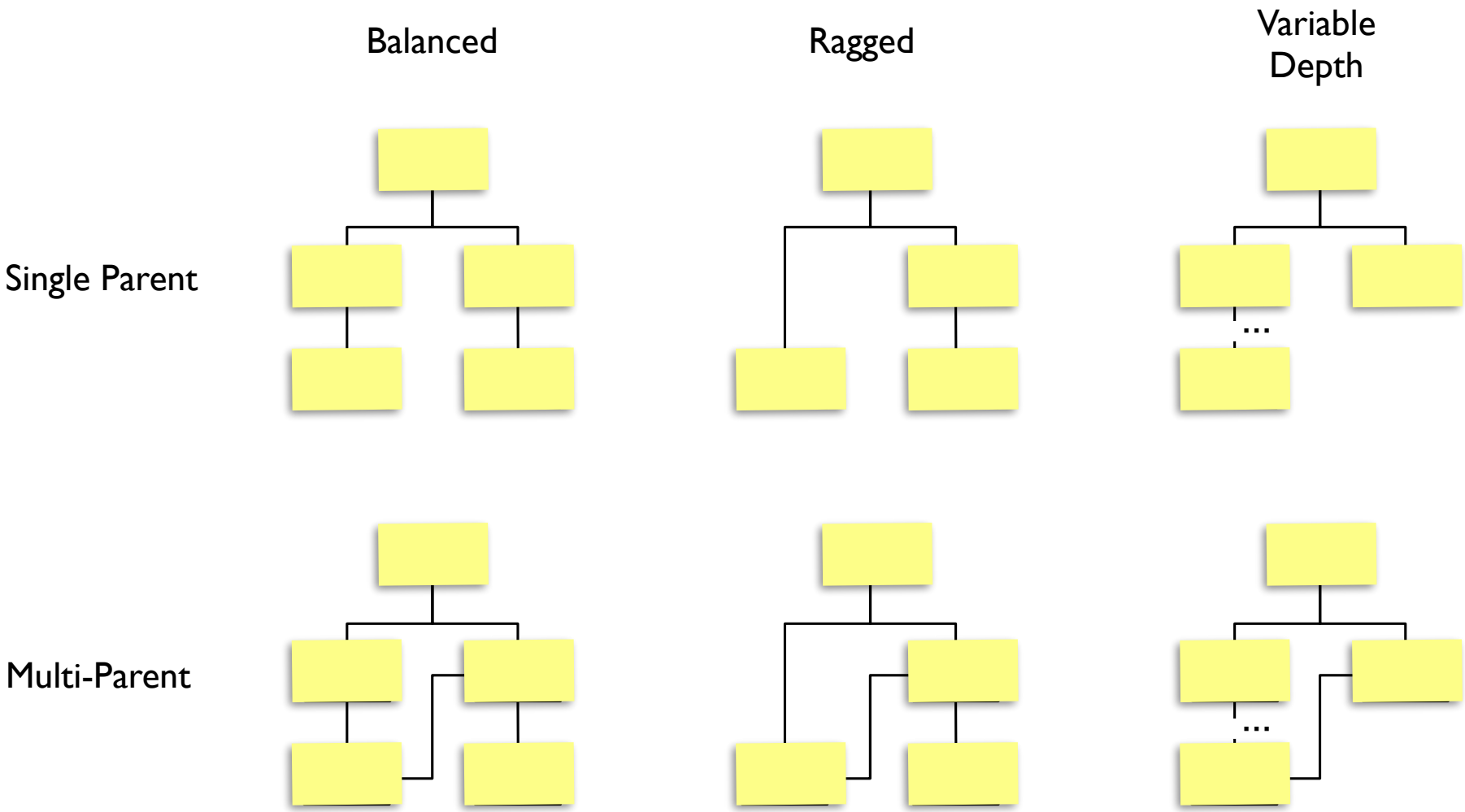
Can a [*noun*]'s [*attribute*] **change**?

# Hierarchies

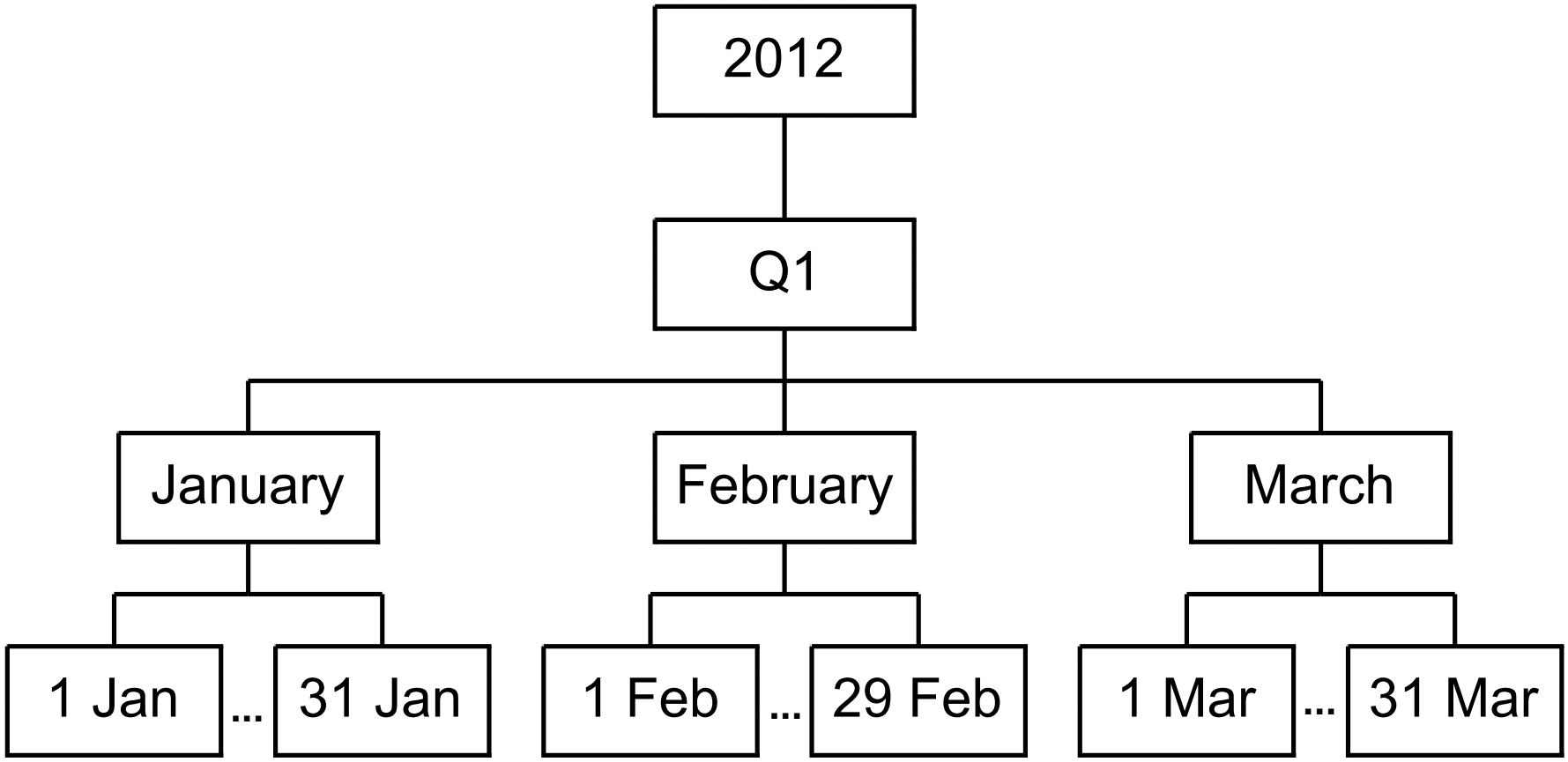
- Common hierarchies: Date/Time, Geography, Organisation, Product
- Why are hierarchies important?
  - Drill Paths – analytic workflow
  - Identify extra sources of categorical data
  - Planning levels – Budgets, Targets, Forecasts
  - Aggregation, Query Optimization
- What shape are they? balanced, ragged, variable depth, simple or complex?
- How do change? Never, Slowly, Rapidly, Regularly or on ad-hoc basis



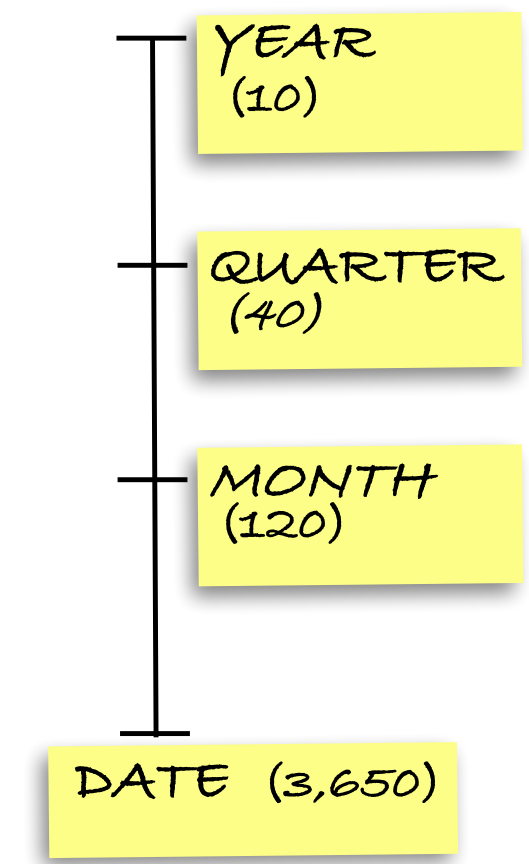
# Hierarchy Types



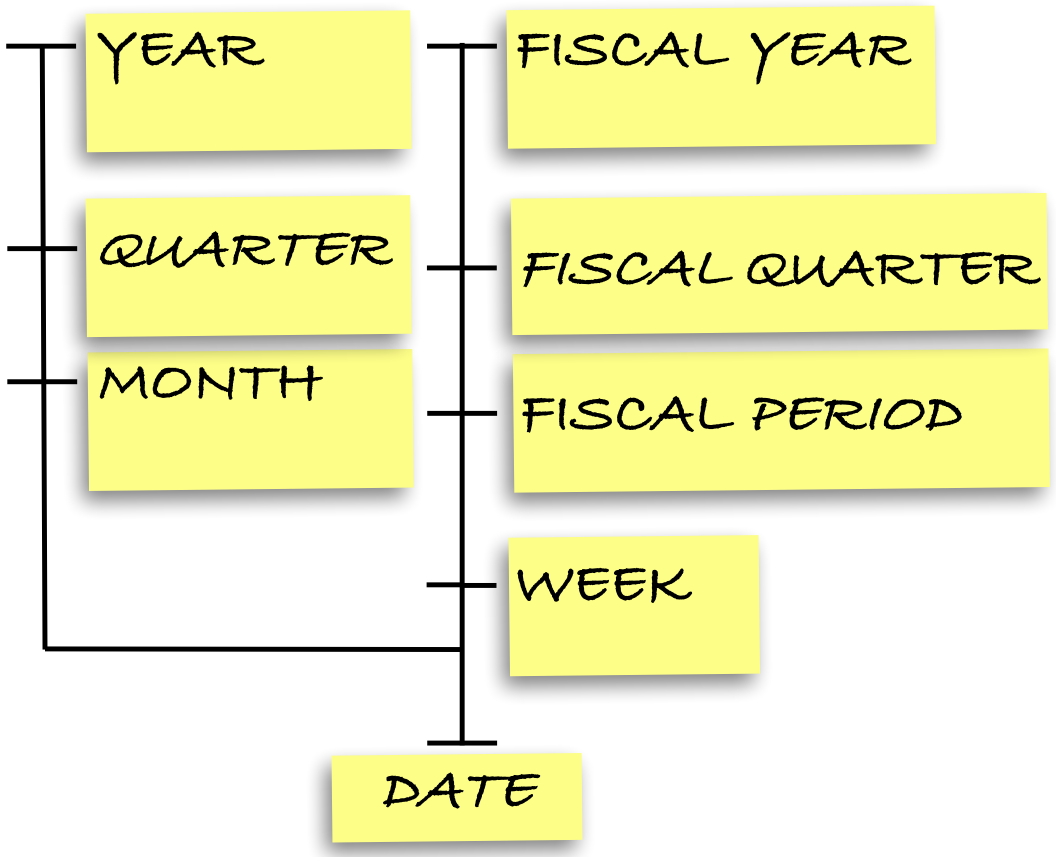
# Time: a Single Parent Balanced Hierarchy



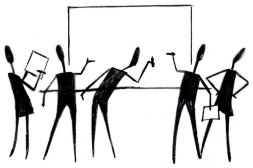
# Hierarchy Charts




Single calendar hierarchy




Multiple time hierarchies



# Exercise 5: Hierarchy Charts

**Requirements**

**Part 1**

Study the following interview notes and document/design a simple product reporting hierarchy using an organisation and/or hierarchy chart showing example values and level names.

- Pomegranate offers overs 400 products and services with many configurable individual options. Operationally products are organised into categories such as **Entertainment, Computing, Services/Subscriptions** etc.
- For reporting purposes the categories are further broken down into many subcategories e.g. **Accessories, Mobile, Desktop, Software/Apps, Consultancy, PomCare, Repairs, PomFrit, Gaming, Multi-Media, Audio, pCloud**.
- Brand names such as, **iPom, iPip, iSongs, PomBook** are a low level grouping of products. Each brand belongs within a specific product subcategory.
- Products are also classified by product or function type such as **Phone, Media Player, Gaming Device, Laptop, Tablet, Computer, Health Monitor, Case, Special Edition**. Most products have a single stable type but some sophisticated products and product bundles are treated as multiple types and are often reclassified.
- A whole new category of products including the **PomCar, PomBike** and **PomScooter** are currently in development.
- Example products discussed included: **PomTV+, PomServer, PomBook Air, PomBook Pro, POM TV, iPom, iPom Mini, iPip Touch, iPipPhone, POMKia 6310, Faraday, PomPad Mini, PomPad Pro, PomPad Softy Case, PomVision and PomBook Protect Case.**

**Part 2**

Draw up an organisation chart using the example HR data from the **EMPLOYEE MANAGER** table.

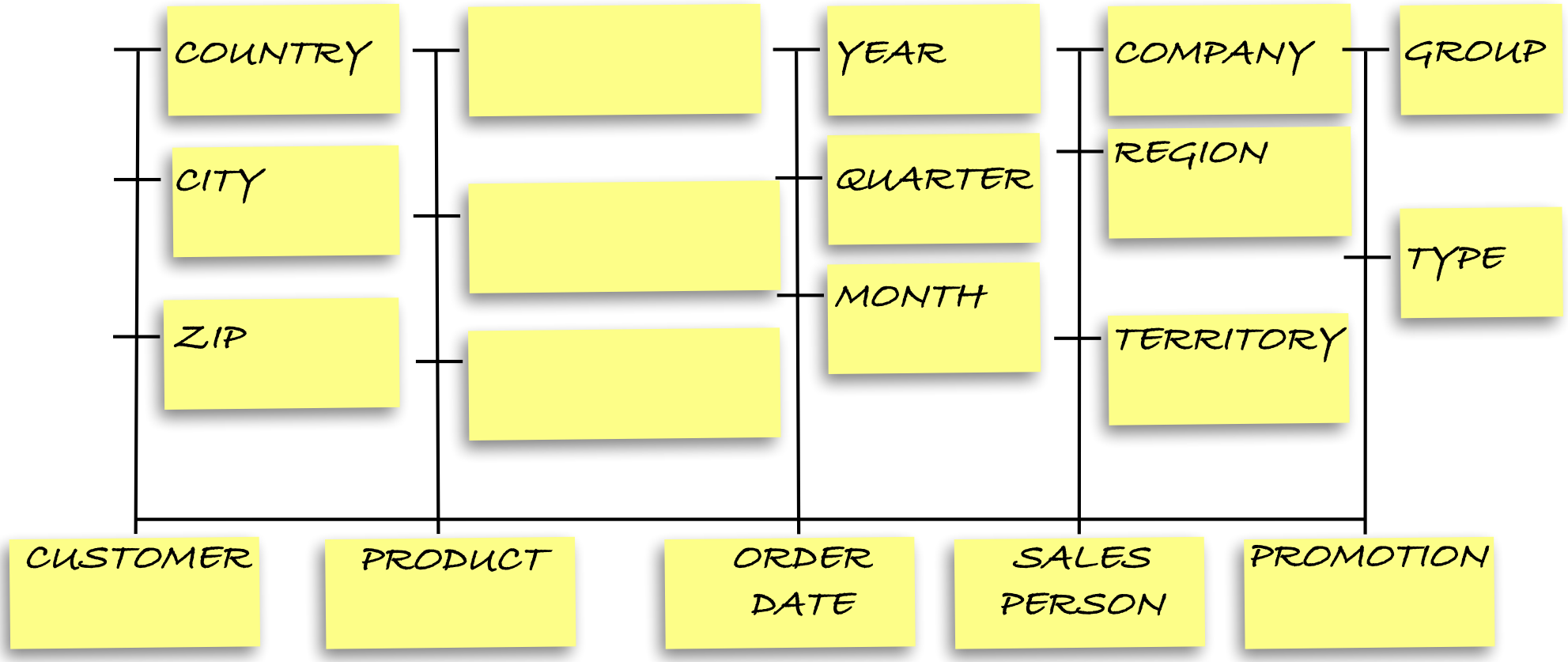
Use solid lines for permanent line manager reporting relationships and dotted lines for temporary project relationships.

What type of hierarchy does this data and your diagram represent?

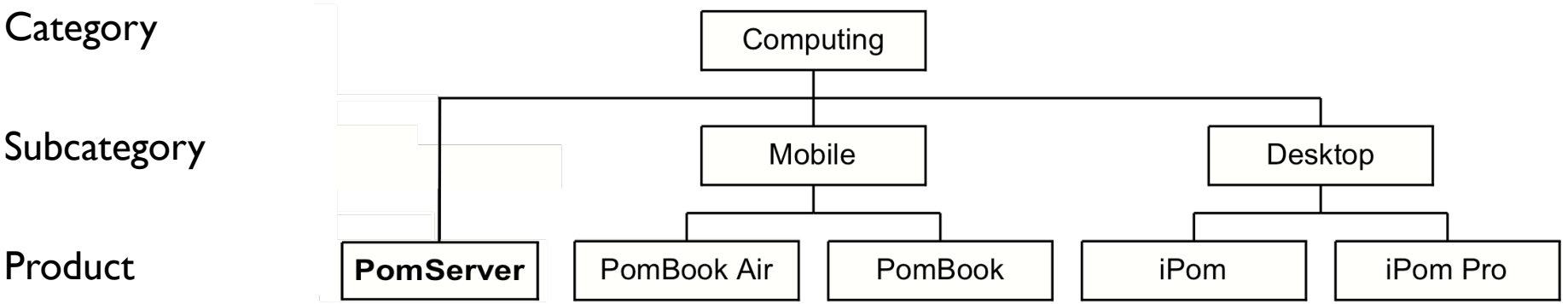
**EMPLOYEE MANAGER**

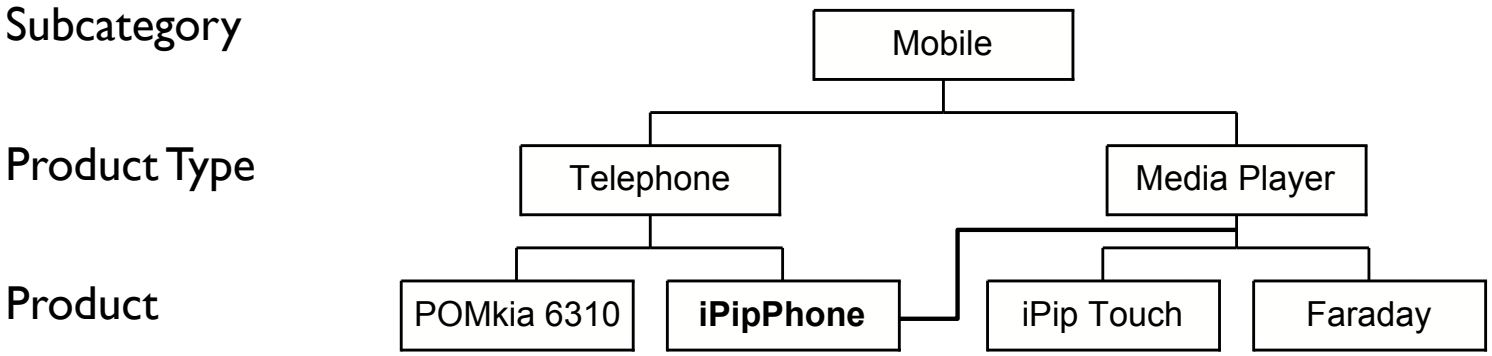
| EMPLOYEE | **reports to** MANAGER | with ROLE TYPE | at FTE |
|---|---|---|---|
| | [EMPLOYEE] RP | | [%] |
| George Smiley | Eve Tasks | Permanent | 100 |
| M | Gerald Mole | Permanent | 100 |
| Gerald Mole | Eve Tasks | Permanent | 100 |
| James Bond | M | Permanent | 80 |
| Moneypenny | M | Permanent | 100 |
| James Bond | George Smiley | Temporary | 20 |

# Hierarchy Charts

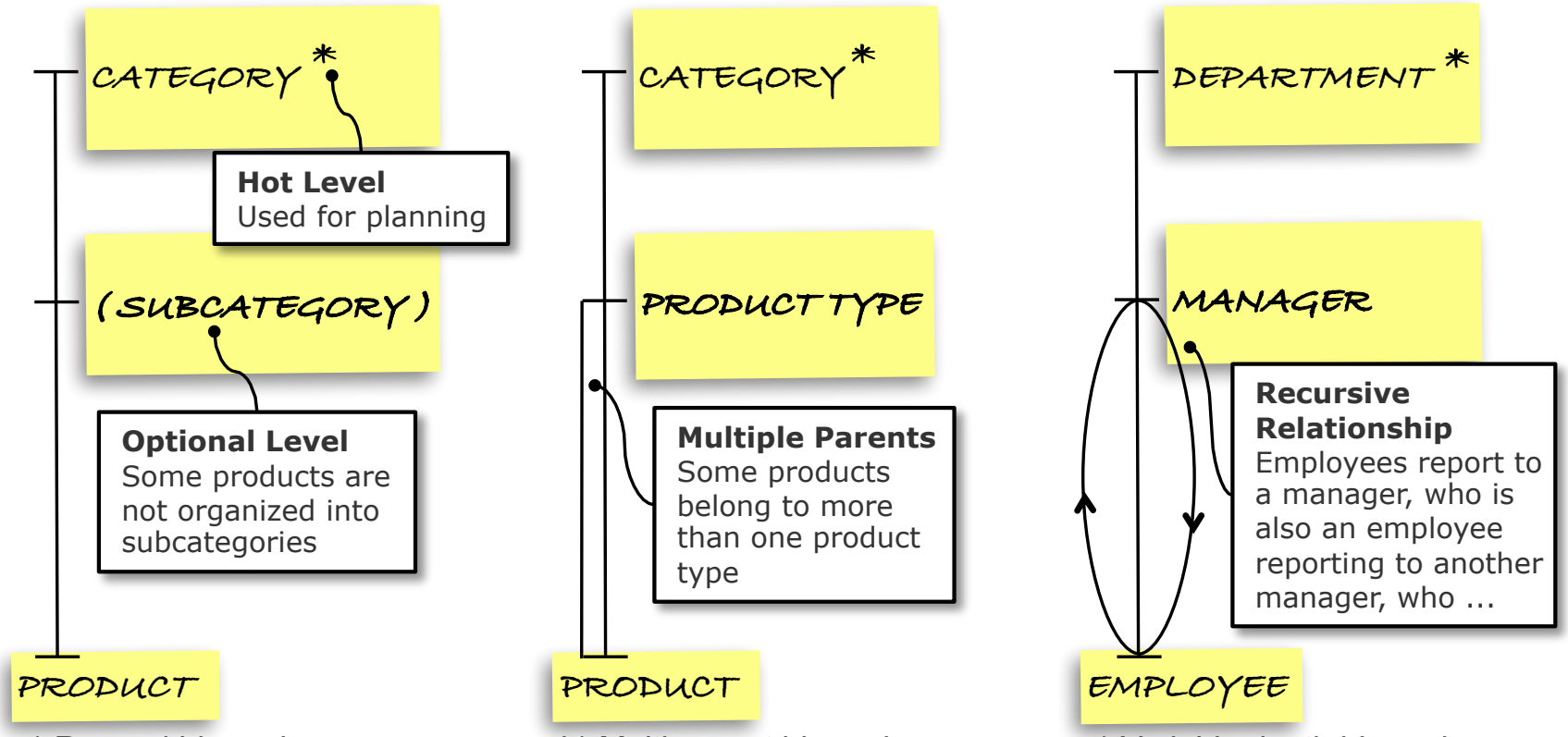


# Ragged and Multi-Parent Hierarchies



# Complex Hierarchy Charts

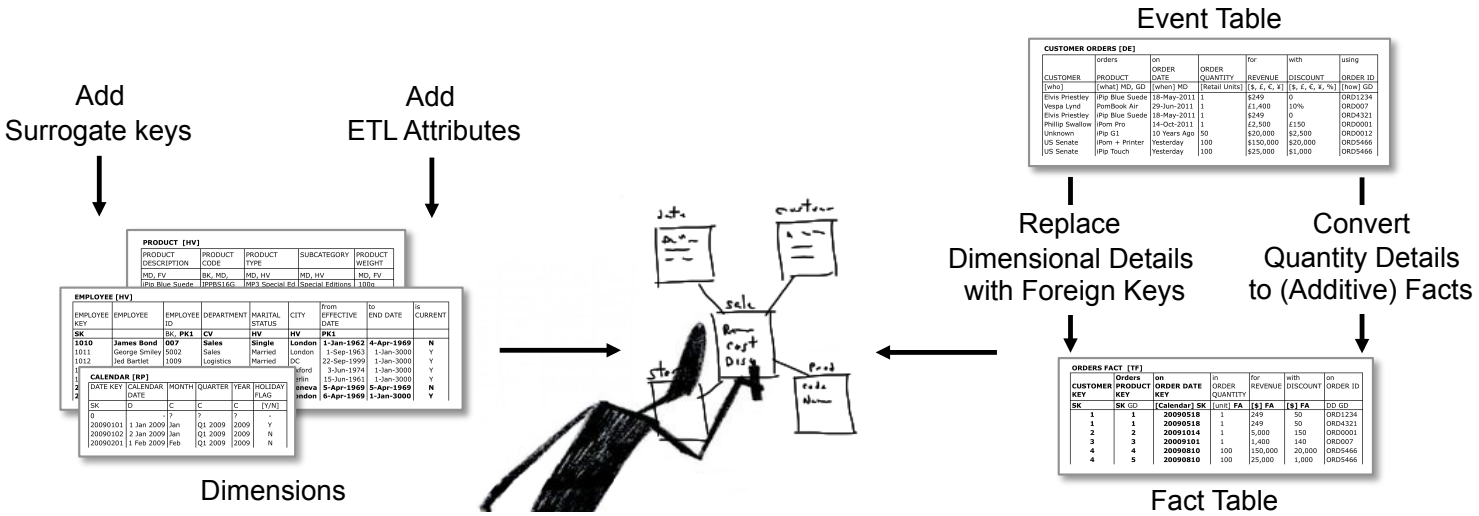


# Telling Change Stories

**EMPLOYEE**

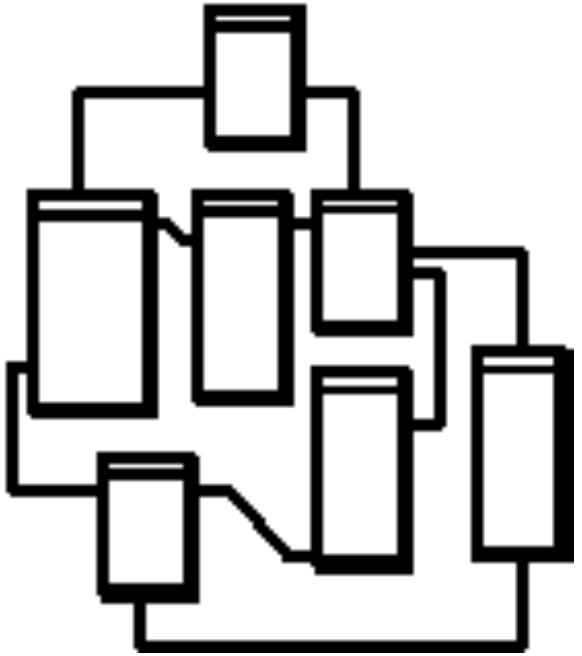
| Employee Name | Employee ID | Department | Date of Birth | Marital Status | City | |
|---|---|---|---|---|---|---|
| | | | | | | |
| **James Bond** | **007** | **Sales** | **25-Sep-1930** | **Single** | **London** | |
| George Smiley | 5002 | Sales | 12-May-1939 | Married | London | |
| Jed Bartlet | 1009 | Logistics | 1-Aug-1940 | Married | DC | |
| Connie Sachs | 5004 | Logistics | 7-Dec-1955 | Single | Oxford | |
| Alex Leamas | 5005 | Logistics | 10-Nov-1945 | Single | Berlin | |
| | | | | | | |

# Agile Star Schema Design




- Dimension Primary Key Design
- Fact Table Patterns
- BEAM✻ vs. Data – Agile Data Profiling
- Planning and Refactoring

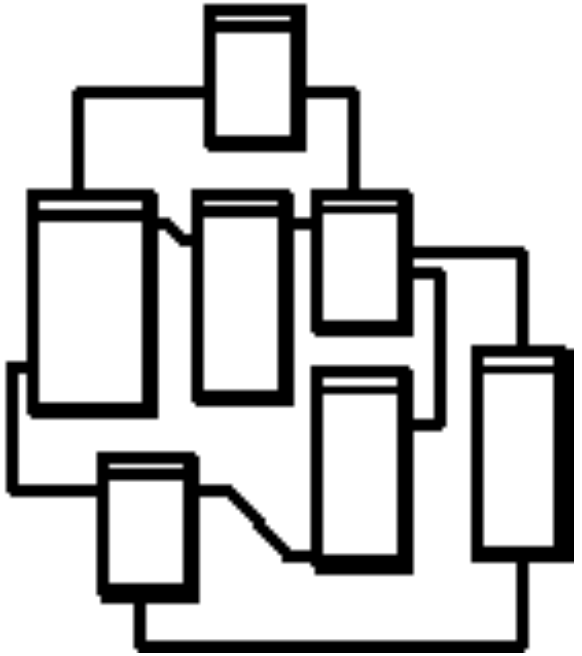
# Benefits of Normalization



- Third Normal Form (3NF):
  - "The key, the whole key and nothing but the key, (so help me Codd)"

- Remove data redundancy
- Minimize storage requirements
- Maximise data consistency
- Simplify Update/Insert/Delete operations
- Optimise operational transaction processing

# Problems with Normalization



- Complex because:
  - Many tables
  - Many join paths
  - Position means nothing

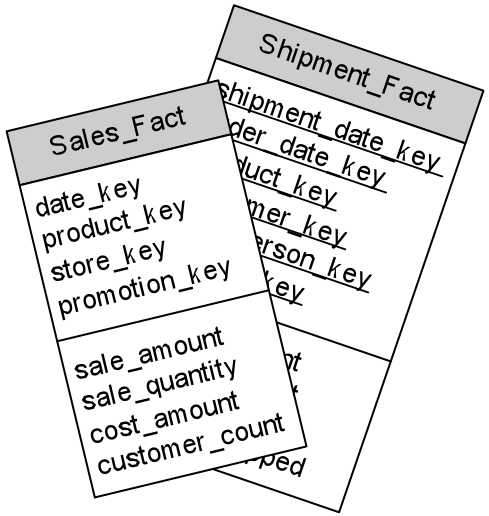
- Complex for users and machines, leading to:
  - Poor query performance
  - Poor problem/solution visualization
  - Poor query construction
- True historical completeness/perspective increases complexity

# Benefits of Dimensional Models

- Simplicity – fewer tables, fewer joins, explicit facts (ingredients for measures and KPI)
- High performance query processing –
  star join optimization, aggregate management
- Data modeled from a user perspective???
- Data modeled from a query/measurement perspective!!!

- **Breaks down BI analysis, design and development into the individual business events/processes worth measuring**

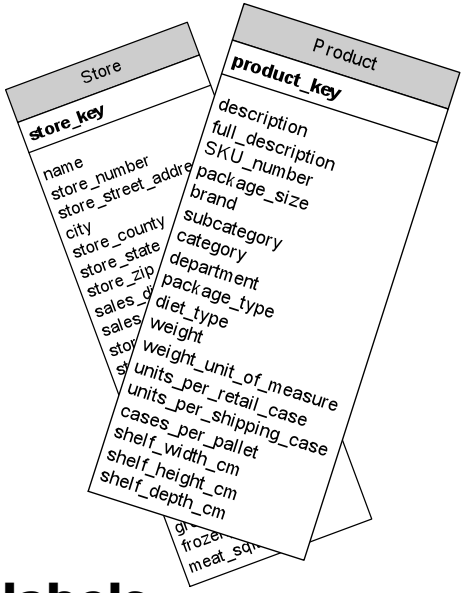
# Fact Table Fundamentals




- Fact tables represent business events, processes, activity
- Fact tables represent ***verbs***
- Fact tables contain facts: quantitative measurements with a common granularity – same level of detail (typically the same source, observed at the same time)
- Fact *values* are analogous to ***adverbs***: they describe the intensity of the verbs

# Dimension Table Fundamentals

- Dimensions explain Business events. They describe business entities in terms that are familiar to users and give context to the facts. Qualitative




- Dimensions are ***nouns***
- A dimension record is a collection of text-like attributes used to group and constrain facts
- **Report row headers and title filters, graph axis and key labels**
- Dimensional attribute *values* represent ***adjectives - (Subjective? Opinions?)***
- Wide rows with numerous columns but small tables – usually less than a million rows

# Dimension Primary Key Design

- **Primary key**: a column or combination of columns that uniquely identifies each row in a table. In addition to being unique, a primary key should ideally be:
  - **Minimal** – uses as few columns as possible (ideally 1)
  - **Present** – not NULL, have a value for all rows in the table
  - **Stable** – does not change value over time
- Dimension Primary Key:
  - ~~Natural Keys (NK)~~
  - ~~Business keys (BK)~~
  - **Surrogate keys (SK)**

# Why Use Surrogate Keys?

- Insulate data warehouse from business keys administration. Support multiple business keys
- Support slowly changing dimensions efficiently
- Enable dimensions to handle missing values and encode uncertainty: not known, not recorded, not possible etc., while avoiding NULLs and outer joins
- Multi-level dimensions, changes to dimension granularity
  e.g. evolution of store → department
- Obfuscate sensitive source system data e.g., full patient or employee details
- Save space in fact tables – integer foreign keys
- Efficient joins between fact and dimension tables – integers
- Referential Integrity via procedural ETL processing

# Dimensional History: Slowly Changing Data

- **Current Value (CV)** – analyze all facts using the current dimensional values.
  ***"as is"*** reporting. *Type 1 Slowly Changing Dimension* (SCD)

- **Historic Value (HV)** – analyze historical facts using the dimensional values that were valid at the time. Analyze new facts using the current dimensional values.
  ***"as was"*** reporting. *Type 2 SCD*

- **Previous Value (PV)** – analyze facts using previous dimensional values.
  ***"as previously"*** reporting. *Type 3 SCD*

- **Fixed Value (FV)** – static, can not change but can be corrected. *Type 0 SCD?*

- Hybrid Attribute Definitions:
  - **CV/HV** – current value (by default) with historic value reporting option
  - **HV/CV** – historic values (by default) with current value reporting option
  - **CV/PV** – current and previous values

# Current Value (CV) Attributes: *Type 1 SCDs*

- Source Data Change = Dimension Update

- Previous values are unimportant
- Correcting mistakes – some changes are always FV corrections  e.g.
- Update dimension with 'simple' ETL processing

- If not a correction then history is rewritten (lost)
- Report results are not reproducible
- If overwritten attributes are used to define aggregates they must be rebuilt

# Historic Value (HV) Attributes: *Type 2 SCDs*

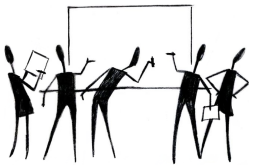
- Source Change = New Row (Update and Insert)
  - Gracefully track many changes to dimension values

  - Each new record partitions history perfectly with simple SQL
  - Querying dimensional timestamps (effective_date, end_date) unnecessary. For most queries as the interesting timestamps are in the fact tables
  - No need to rebuild historic aggregates

  - Requirement to maintain surrogate keys
  - Queries must not use surrogate key values for filtering or sorting
  - Browse (dimension only) queries must select distinct … and count(distinct …)
  - Growth of dimension table. Different technique needed for tracking changes in "Very Large Dimensions"?

# Slowly Changing Groups

- Slowly Changing Groups contains attributes with related change behaviours. e.g. attributes that typically change together or atypically change together.
- Slowly Changing Groups enable Multi-Attribute Conditional SCD Rules needed to:
  - Detect corrections vs. genuine change
  - Avoid tracking history of errors or *small* changes
- Identify all attributes in the group with a common numerical suffix
- Give conditional attributes a default SCD behaviour and one or more exceptional behaviours using suffixes
- When every member of a numbered group changes simultaneously perform the exceptional behaviour with the matching suffix
- **CV, HVn** – conditional HV attribute. CV by default, HV when triggered
- **HV, HVn** or just **HVn** – HV attribute trigger. Always HV



# Exercise 6: Slowly Changing Attribute Groups

**Background**

Pomegranate have added web sales to its data warehouse. New and changed online customer details are processed nightly and loaded into a slowly changing customer dimension with the following business rules:

- Major address (**City**, **Country**) changes are tracked over time using Type 2 SCD processing.
- Minor address changes are not tracked over time, only the most recent **Street** address is recorded by default. However, historical **Street** addresses must be recorded as part of major address changes.
- **First Name, Last Name** and **Title** changes are only tracked when all elements change together otherwise individually they are treated as corrections.

**Part 1**

Document these **multi-column** business rules in the following CUSTOMER dimension using BEAM✻ short codes:

1. Use HV to document historical value attributes that must always be tracked and CV for current value attributes that are not tracked by default.
2. Record any multi-column dependencies by adding group numbers to the short codes (e.g. HV1 or HV2):
   - Add additional numbered HV codes to any CV attributes that must be tracked as a group (when all group members change). Tie group members together with a common group number.
   - Add an additional numbered HV code to CV attributes if they must be tracked only when a related HV attribute changes. Tie these triggering HV attributes to their conditionally tracked attributes by adding a common group number to their HV codes.

**Part 2**

The extract table on the following page contains the new and changed customer records extracted from the production system at close of business today.

Simulate the CUSTOMER dimension ETL process by comparing each extract record against the CUSTOMER dimension and processing it as either an insert or an update. Assume if you can't find a matching customer in the sample data that the extracted record should be treated as an insert.

Use the tables below to describe the ETL processing required for each extract record and to show what each new and updated dimension row will look like when the load processing is complete.

## Extract File

| Record # | Customer ID | Title | First Name | Last Name | Street | City | Country |
|---|---|---|---|---|---|---|---|
| 1 | C1239 | Sir | Hilary | Bray | Royal College of Arms | London | UK |
| 2 | C2345 | Mrs | Tracy | Bond | Rue Du Rhone | Geneva | Switzerland |
| 3 | C7810 | Miss | Vespa | Lynd | 10 Kings Road | London | UK |
| 4 | C8904 | Mr | Felix | Leiter | The Office | Langley, VA | USA |
| 5 | C3467 | Mr | Marc-Ange | Draco | Rua Augusta | Lisbon | Portugal |
| 6 | C1038 | Mr | Ernst Stavro | Blofeld | The Summit | Piz Gloria | Switzerland |

## ETL Audit

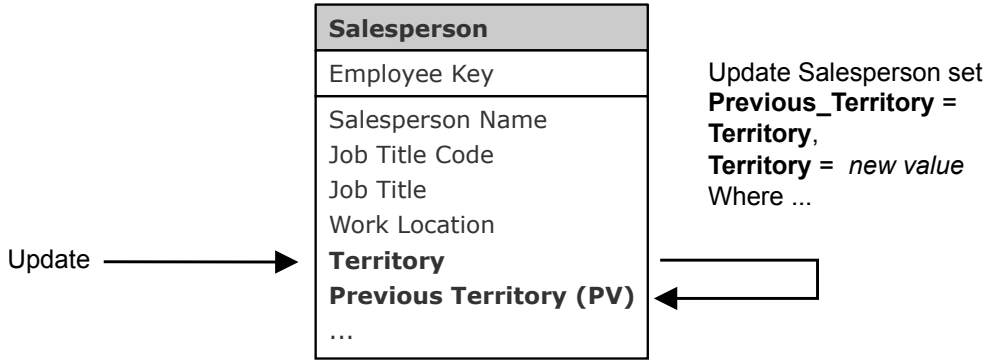
| Record # | Customer ID | ETL Processing | Reason | Rows Inserted | Rows Updated |
|---|---|---|---|---|---|
| 1 | C1239 | | | | |
| 2 | C2345 | | | | |
| 3 | C7810 | | | | |
| 4 | C8904 | | | | |
| 5 | C3467 | | | | |
| 6 | C1038 | | | | |

**CUSTOMER**

| Customer Key | Customer ID | Title | First Name | Last Name | Street | City | Country | Effective Date | End Date | Last Updated | Current |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SK** | **BK** | | | | | | | | | | **[Y/N]** |
| 1001 | C8904 | Mr | Felix | Lighter | Harry's Bar | Key West, FL | USA | 5/10/1962 | 16/9/1964 | 17/9/1964 | N |
| 1021 | C1010 | Dr. | Julius | No | The Island | Crab Key | Jamaica | 5/10/1962 | 1/1/3000 | 5/10/1962 | Y |
| 1041 | C1241 | Mr | Auric | Goldfinger | Fontainebleau Hotel | Miami, FL | USA | 17/9/1964 | 1/1/3000 | 17/9/1964 | Y |
| 1042 | C8904 | Mr | Felix | Lighter | The Office | Langley, VA | USA | 17/9/1964 | 1/1/3000 | 17/9/1964 | Y |
| 1078 | C1038 | Mr | N/A | Blofeld | N/A | Paris | France | 21/12/1965 | 9/6/1967 | 10/6/1967 | N |
| 1500 | C1038 | Mr | Stavro | Blofeld | N/A | N/A | Japan | 10/6/1967 | 1/1/3000 | 10/6/1967 | Y |
| 2700 | C2345 | Contessa | Teresa | Di Vincenzo | Via Veneto | Rome | Italy | 18/12/1969 | 1/1/3000 | 18/12/1969 | Y |
| 3518 | C3467 | Mr | Marc-Antony | Draco | Rue August | Lisborn | Portugal | 19/12/1969 | 1/1/3000 | 19/12/1969 | Y |
| 2010 | C1239 | Mr | Hilary | Bray | 1 Chelsea Gardens | London | UK | 20/12/1969 | 1/1/3000 | 20/12/1969 | Y |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |

# Previous Value (PV) Attributes: *Type 3 SCDs*

- Represents 2 physical columns: Current and Previous. Save Current into Previous prior to changing Current:

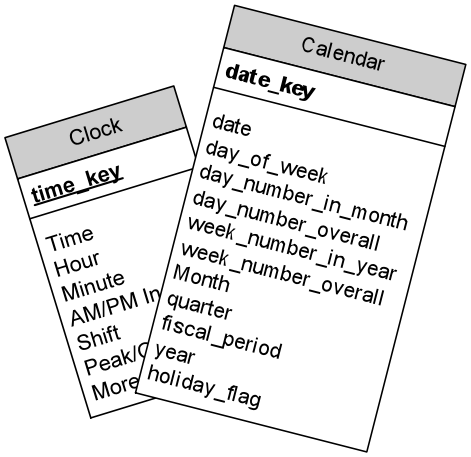



- Suitable for 'logical' rather than 'physical' changes
- Suits a small number of very slowly changing attributes
- Creates no additional dimension records

- Requires schema changes when attribute policy changes
- Need to rebuild affected aggregates
- Provides only "as is" and "as previous" analysis

# *When*: Calendar and Time Dimensions

- Why do we need time dimensions when we have date-time data type columns?
  - Avoid duplicating calendar logic in each report/application
  - Removing date arithmetic from query where clauses
  - Insulate BI applications from database specific date functions
  - Easy to extend the calendar with the organization's own fiscal periods and rollups
  - More features later...
- Separate time of day and date – Why?



# Event Story Type → Fact Table Type and Granularity

- **Discrete Event [DE] → Transaction [TF]**
  - Instantaneous/short duration (less than a day), irregularly occurring transactions
- **Evolving Event [EE] → Accumulating Snapshot [AS]**
  - Non-instantaneous/longer duration (greater than a day), irregularly occurring transactions
  - Multiple time dimensions (e.g. order date, shipment date, delivery date, payment date)
- **Recurring Event [RE] → Periodic Snapshot [PS] – report**
  - Regularly occurring events, ongoing processes, typically used to measure cumulative effect of discrete events
- **Fact Table Granularity: Clear (*simple*) definition of a fact table row**
  - "One record per..."

ADWD pp. 152-153, 228-229

# Facts, Measures, KPIs

- **Facts**
  Quantities sourced from the raw numbers captured in transactions by operational systems e.g. **Revenue**. Atomic detailed data that cannot be broken down to a lower level. Stored in a fact table.

- **Measures**
  Facts summarised or aggregated to a suitable level for report display and comparison e.g. **Monthly Branch Revenue**. Combination of dimensions and facts. Calculated in a query or possibly physically stored in an aggregate fact table/OLAP cube for efficient access.

- **Key Performance Indicators (KPIs)**
  Self contained ratio or percentage. Embodies **comparison** logic on measures. User gains understanding from single figure rather than knowledge of previous values or further analysis in conjunction with other measures e.g. **Revenue Growth**.

# Fact Additivity

- **(Fully) Additive (FA)** – can be summed up across any combination of dimensions. No special rules

- **Non-additive (NA)** – can not be summed across any dimension e.g. unit price or temperature

    - Must be aggregated in other ways e.g. average, min, max

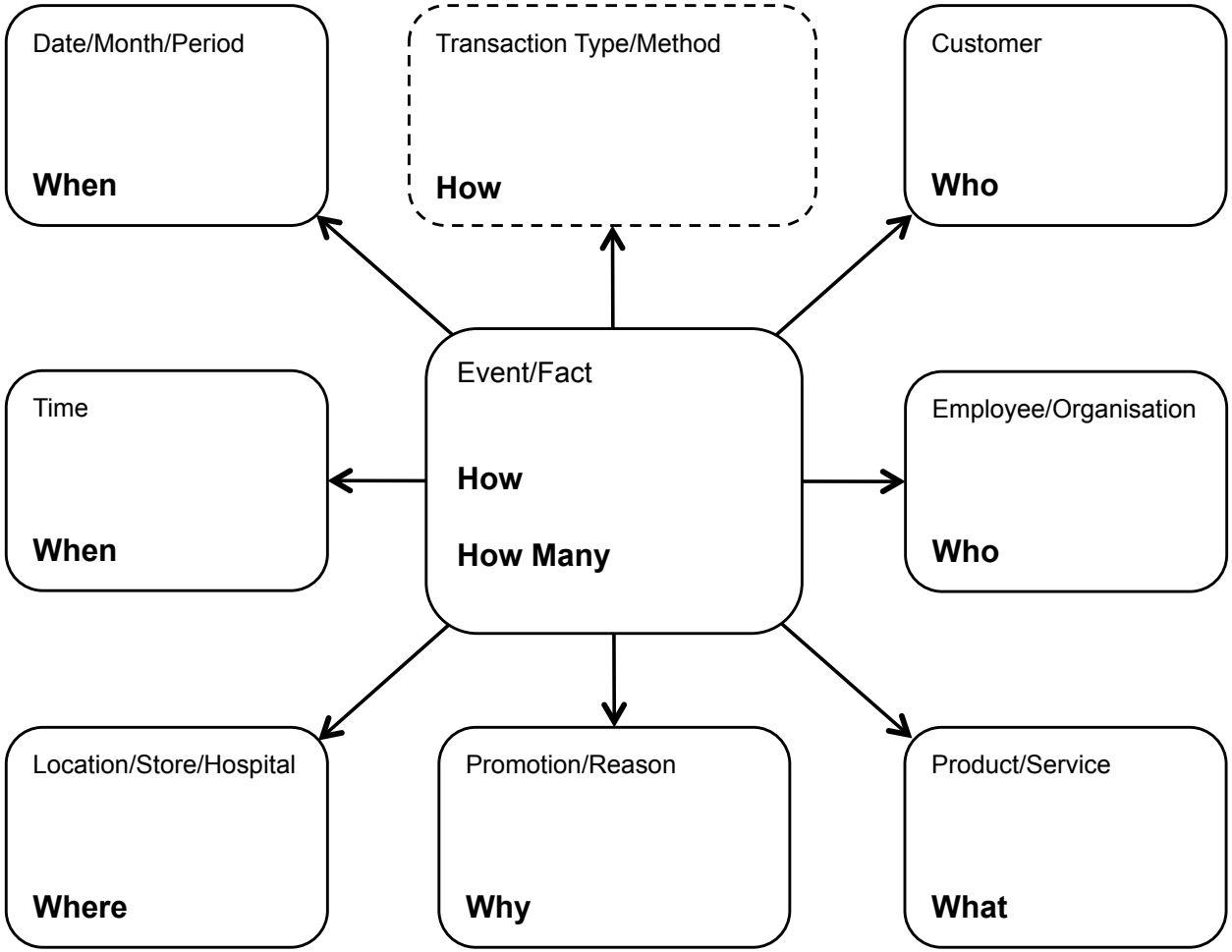
    - Degenerate Dimensions (DD) – transaction #, timestamps, flags

- **Semi-additive (SA)** – can not be summed across at least one dimension e.g. balances can not be summed over time

# Fact Definition

- Define and store facts in their most **Additive** form
  - Avoid unit prices, ratios and percentages
  - Store additive numerators and denominators separately
- Pre-calculate and store commonly used derived facts (DF)
- For each fact we need to define:
  - Name, description, unit of measure, source(s), formula, additivity, non-additive dimension (if any), default aggregate function if not sum( )

# Enhanced Star Schema



# BEAM* Canvas

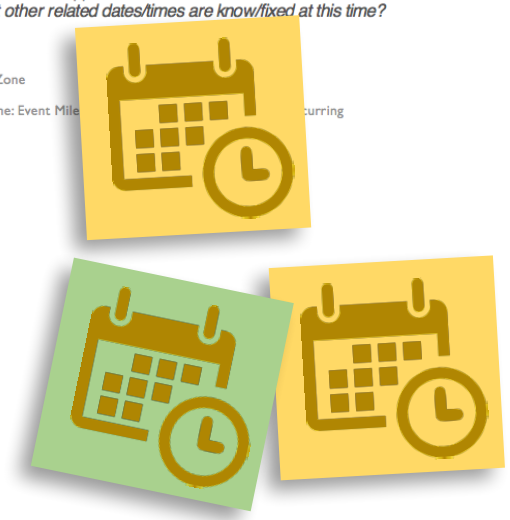
Why Position Matters - The Benefits of Consistent Visual Layout

*Title, Author(s), Date/Version*

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Mile ...curring



## How

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we uniquely identify each ever*

Verb, Activity, Process, Event
Effect, Outcome, Status
Transaction Type
Transaction #, Event ID [Degenerate Dimens
Step/Sequence #

[Granularity, Event Type: Discrete, Evolving, or Recurring]

## Who

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

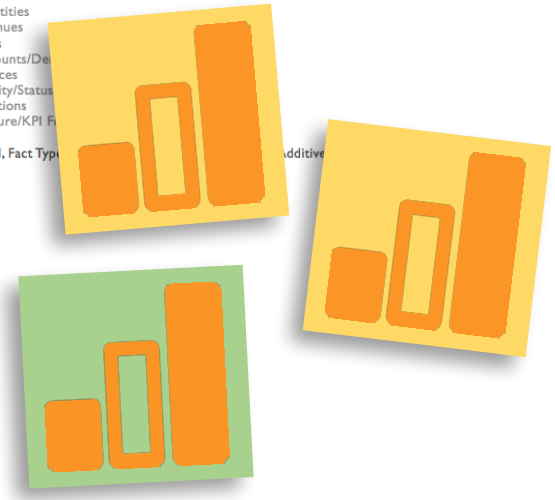
Subject/Object
Customer: Business, Consumer, Segment
Employee
Supplier
Partner
Third Party

## How Many

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/De
Balances
Activity/Status
Durations
Measure/KPI F

[UoM, Fact Typ ...dditive



## Where

*Where does it happen? Where does it refer to?*

Subject/Object
Location
Branch, Store, Facil
Channel
URL
Map/Sequence: First ... Last

## Why

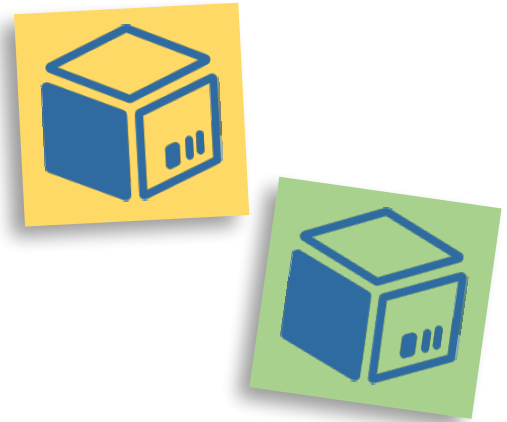
*Why does it happen?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

## What

*What is involved/used? How are they organized?*
*How do they change?*

Subject/Object
Value Proposition
Product
Service
Resource
Item



Colour Coding Multiple (Closely Related) Business Events

Short Codes, Unique IDs, Population Size *Event/Title*

*Subject Verb Object*

*Author(s), Date/Version*

Customer buys Product

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event

Purchase Date
Invoice date/time (3,650)

## How

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we u*

Verb, Activity, Pr
Effect, Outcome,
Transaction Type
Transaction #, Ev
Step/Sequence #

[Granularity, Even

Invoice
Invoice # (100M)
DD NA

## Who

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

Subject/Object
Customer: Business, Consumer, Se
Employee
Supplier
Partner
Third Party

Customer
CV
Customer ID (1M)

## How Many

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/Deltas
Balance
Activit
Durati
Measu

[UoM,

Quantity
[items] FA

Revenue
[$] FA

Salesperson
[Employee] HV
Employee ID (5,000)

## Where

*Where does it happen? Where does it refer to?*

Subject/Object
Location
Branch, Store, Facility
Channel
URL
Map/Sec

Store/URL
HV
Store No/URL (5,000)

## What

*What is involved/used? How are they organized?*
*How d*

Subject
Value F
Produ
Servic
Resou
Item

Product
HV
Product Code (50,000)

## Why

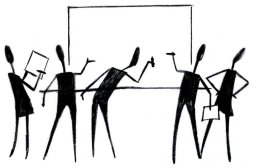
*Why does it happen?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

Promotion
campaign #
(10,000)

BEAM* Canvas

**modelstorming.com** - Agile Data Warehouse Design
*Business Event Analysis and Modeling



# Exercise 7: BEAM Canvas Revisited

**Part 1**

Revisit your BEAM Canvas from exercise 2. Convert it into a viable dimensional model by adding dimension and fact table short codes to your existing post-its. Rearrange/rename your existing notes and add any additional/missing notes to optimise your star schema design for performance, flexibility and easy of use/understandability.

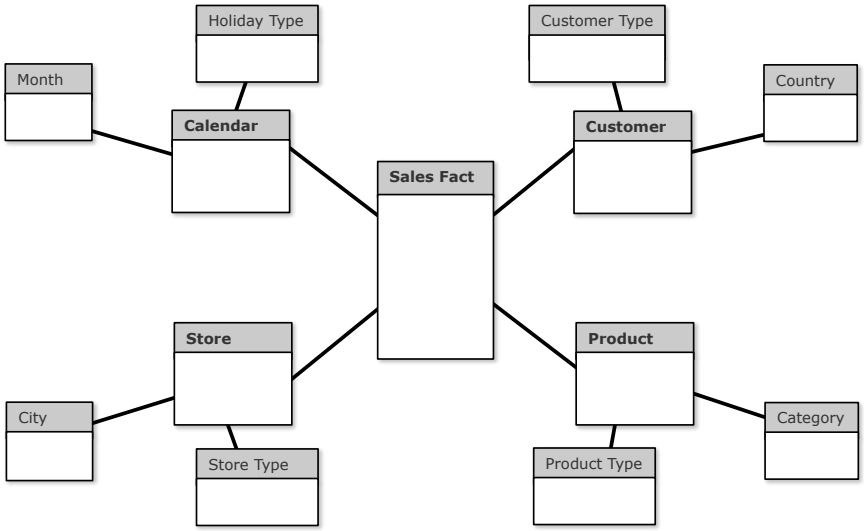
Do not add surrogate keys or waste ink drawing any dimension to fact primary:foreign key relationships just to make it look like a star - we can take those details for granted.

Document the fact table type (TF, AS, or PS) and its precise granularity along with the additivity of each fact and it UoM. Identify any type 2 slowly changing dimensions which will track historical values of at least one attribute. Also identify any role playing dimensions.

**Part 2**

If you could broaden the scope of your existing model to include one or two more verbs from the reseller sales timeline which would you choose and what additional Post-its would you add to your canvas? Try to limit this 'scope creep' to no more than 5 additional Post-its.

# Snowflake Schemas

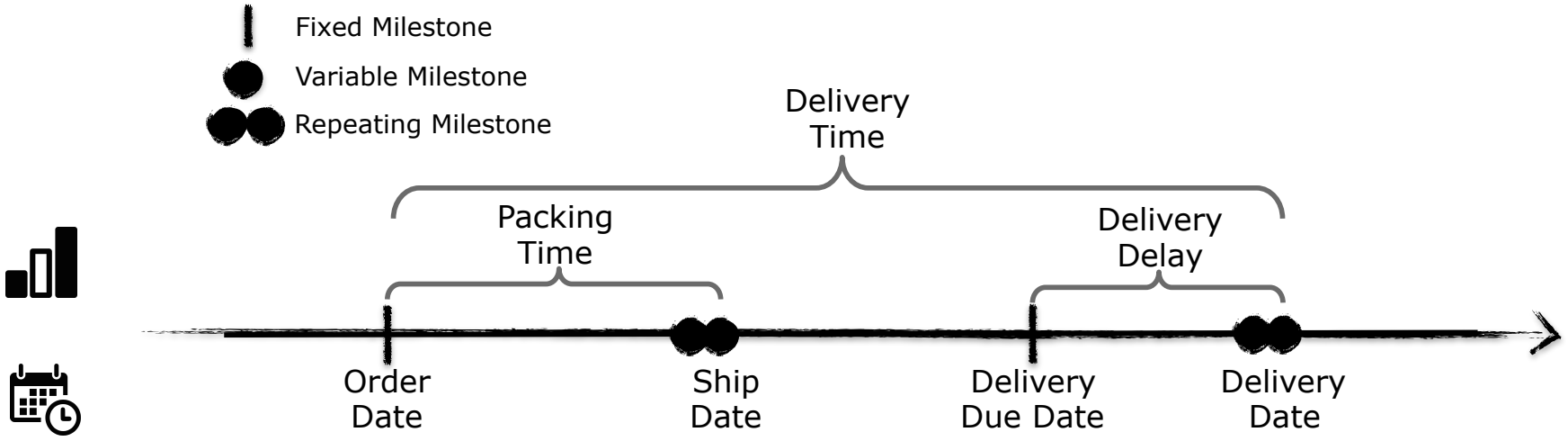



- A snowflake schema is a star schema with one or more dimensions that have been normalized
  - Complex for user presentation
  - Poor browse performance

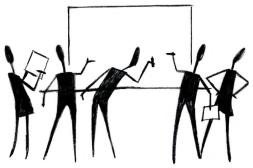
  - No significant space saving
  - May simplify dimension ETL – not a BI priority compared to fact ETL and queries
  - Requires additional non-natural business keys for joining to "outriggers"
  - Negates advantages of bitmap indices
- Document hierarchies with hierarchy charts instead
- There are legitimate reasons for "snowflaking" but resist initial 3NF instincts for BI models

# Accumulating Snapshots [AS] – Timelines

- Suitable for processes that take time such as orders, shipments, claims etc.
- More dimensions: additional dates for milestones events.
- More facts: additional durations best modeled using a timeline:




- More ETL: Fact and dimension values are changed repeatedly as long as activity occurs.
- Intermediate states? More fact tables: details need to be available in transactional stars.



# Exercise 8: Sales Targets

**Requirements**

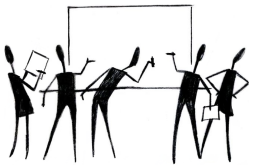
Sales management would like to compare reseller sales against targets.

Study the sample spreadsheet and additional business knowledge below. Does this data represent additional details, fact, dimensions or attributes, you can add to your existing model or does it represent a different set of events?

Try the following BEAM questions to help you decide:

- "Who sets plans/targets?"
- "What do they set plans for?"
- "When do they set plans?"
- "When is a plan effective?"
- "When do plan/targets change?"
- "Who else is a plan associated with?"
- "What are the plan quantities? What units of measure are they recorded in?"

Add the additional planning data to your reseller sales canvas. If you decide plans are completely new business events you should of course start with 'Who does what?' and use different colour Post-its. Why?



# Business Knowledge: Sales Targets

| Sales Person | Month | Product Category | Commission | Target Units | Target Contribution |
|---|---|---|---|---|---|
| James Bond | January-08 | Workstations | 15% | 100 | 30000 |
| James Bond | January-08 | Laptops | 7% | 200 | 20000 |
| James Bond | January-08 | Entertainment | 7% | 400 | 20000 |
| James Bond | February-08 | Workstations | 20% | 120 | 35000 |
| James Bond | February-08 | Laptops | 10% | 180 | 18000 |
| James Bond | February-08 | Entertainment | 10% | 450 | 25000 |
| James Bond | March-08 | Workstations | 20% | 80 | 25000 |
| James Bond | March-08 | Laptops | 10% | 200 | 20000 |
| James Bond | March-08 | Entertainment | 10% | 200 | 10000 |

Regional sales managers are responsible for setting and adjusting targets and commission percentages.

Above is an excerpt from a sales target spreadsheet they create. It shows the monthly targets and commission % for product category sales for one reseller account manager (salesperson).

Salesperson targets and commission % are set on an annual basis by product category but can be adjusted during the year.

Each time a target or commission % is adjusted a new version of the plan is created as a new worksheet with a new plan number and plan effective from date in the title.

Both the Sales and Finance departments will want to compare actuals to different versions of the plan – most frequently to the initial plan and the current plan. They will want to analyse when plans change and by how much.

# Periodic Snapshot [PS] Example

## *When*

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Milestones: Fixed, Variable, Repeatable/Recurring

## *How*

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we uniquely identify each event?*

***Verb***, Activity, Process, Event
Effect, Outcome, Status
Transaction Type
Transaction #, Event ID [Degenerate Dimension]
Step/Sequence #

[Granularity, Event Type: Discrete, Evolving, or Recurring]

## *How Many*

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/Deltas
Balances
Activity/Status Counts
Durations
Measure/KPI Formulas

[UoM, Fact Type: Fully Additive, Semi-Additive, Non-Additive]

## *Who*

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

*Subject/Object*
Customer: Business, Consumer, Segment
Employee
Supplier
Partner
Third Party

## *Where*

*Where does it happen? Where does it refer to?*

*Subject/Object*
Location
Branch, Store, Facility
Channel
URL
Map/Sequence: First → Previous → Current → Next → Last

## *What*

*What is involved/used? How are they organized?*
*How do they change?*

*Subject/Object*
Value Proposition
Product
Service
Resource
Item

## *Why*

*Why does it happen? Why do quantities vary?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

# Semi-Additive (SA) Facts: Balances

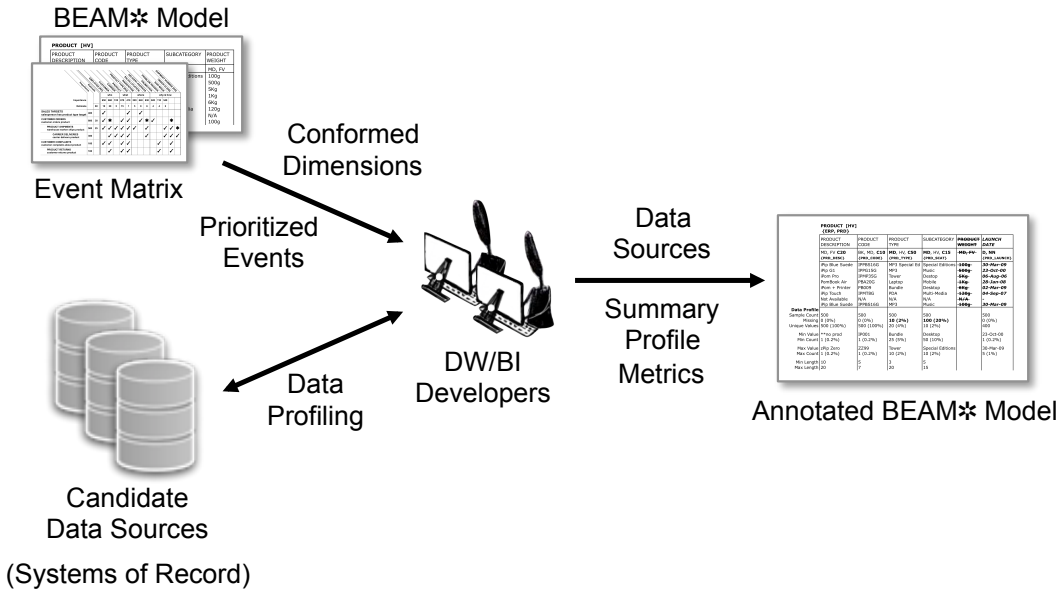
- Many semi-additive facts must be averaged across their non-additive dimension (NA)
  - E.g. Average Daily Stock Balance
- Average with care! A simple average e.g. SQL AVG() divides the total figure by the number of observations
  i.e. AVG() ≡ SUM(…)/COUNT(*)
- The correct average may be a period average
  - i.e. Sum(…)/COUNT(DISTINCT period)
    requires that you understand the periodicity (date/time dimension granularity)

# Semi-Additive (SA) Facts: Counts

- For some semi-additive facts (typically counts) queries aggregation must be restricted to a single value of the non-additive dimension (NA) for meaningful results
- Alternatively fully additive measures can be constructed from more granular facts
  - e.g. COUNT(DISTINCT Receipt#)
- Care has to be taken with report sub-totals and grand totals
- Counts are often examples of 'artificial' semi-additive facts i.e. we create them by our choice of granularity

# Agile Data Profiling

- Targeted
- Done Early
- Done Frequently – Automated
- Done by DW/BI Team members
- Recorded in the business model

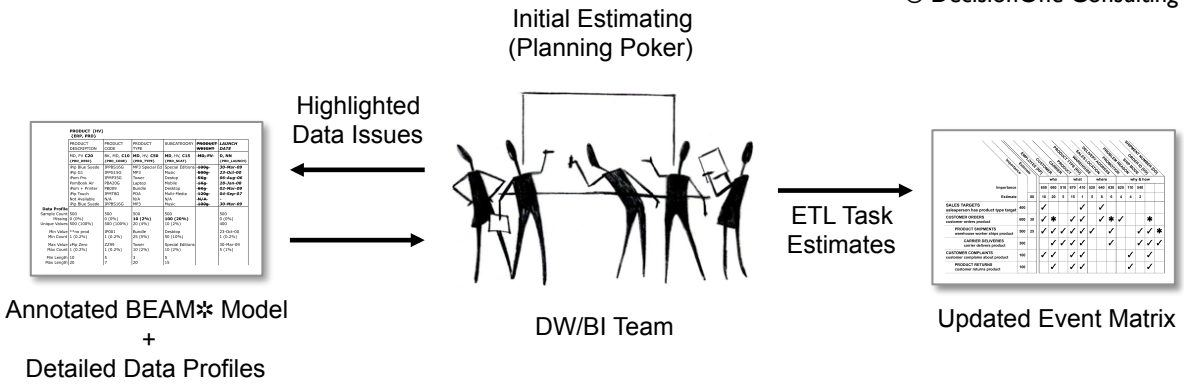


# Annotated BEAM✻ Model

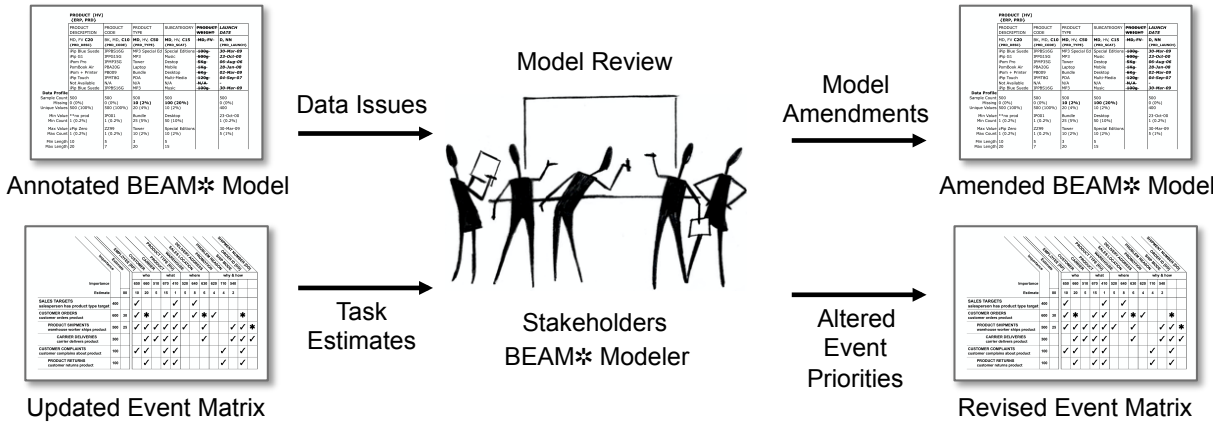
**PRODUCT [HV]**
**{ERP, PRD}**

| | PRODUCT DESCRIPTION | PRODUCT CODE | PRODUCT TYPE | SUBCATEGORY | ~~PRODUCT WEIGHT~~ | *LAUNCH DATE* |
|---|---|---|---|---|---|---|
| | MD, FV **C20** **{PRD_DESC}** | BK, MD, **C10** **{PRD_CODE}** | ~~MD~~, HV, **C50** **{PRD_TYPE}** | ~~MD~~, HV, **C15** **{PRD_SCAT}** | ~~MD, FV~~ | **D, NN** **{PRD_LAUNCH}** |
| | iPip Blue Suede | IPPBS16G | MP3 Special Ed | Special Editions | ~~100g~~ | *30-Mar-09* |
| | iPip G1 | IPPG15G | MP3 | Music | ~~500g~~ | *23-Oct-00* |
| | iPom Pro | IPMP35G | Tower | Destop | ~~5Kg~~ | *06-Aug-06* |
| | PomBook Air | PBA20G | Laptop | Mobile | ~~1Kg~~ | *28-Jan-08* |
| | iPom + Printer | PB009 | Bundle | Desktop | ~~6Kg~~ | *02-Mar-09* |
| | iPip Touch | IPMT8G | PDA | Multi-Media | ~~120g~~ | *04-Sep-07* |
| | Not Available | N/A | N/A | N/A | ~~N/A~~ | *-* |
| | iPip Blue Suede | IPPBS16G | MP3 | Music | ~~100g~~ | *30-Mar-09* |
| **Data Profile** | | | | | | |
| Sample Count | 500 | 500 | 500 | 500 | | 500 |
| Missing | 0 (0%) | 0 (0%) | **10 (2%)** | **100 (20%)** | | 0 (0%) |
| Unique Values | 500 (100%) | 500 (100%) | 20 (4%) | 10 (2%) | | 400 |
| Min Value | **no prod | IP001 | Bundle | Desktop | | 23-Oct-00 |
| Min Count | 1 (0.2%) | 1 (0.2%) | 25 (5%) | 50 (10%) | | 1 (0.2%) |
| Max Value | zPip Zero | ZZ99 | Tower | Special Editions | | 30-Mar-09 |
| Max Count | 1 (0.2%) | 1 (0.2%) | 10 (2%) | 10 (2%) | | 5 (1%) |
| Min Length | 10 | 5 | 3 | 5 | | |
| Max Length | 20 | 7 | 20 | 15 | | |

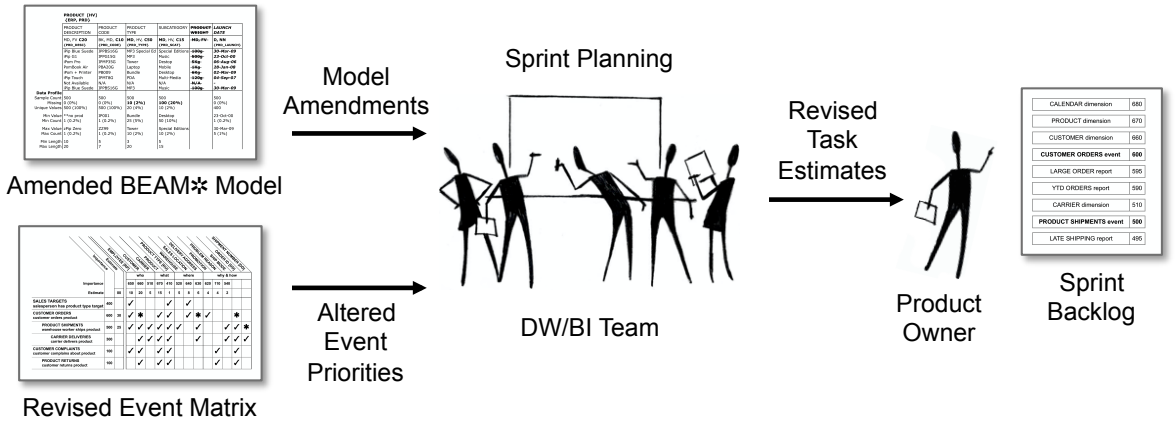
- DW/BI Team Estimating



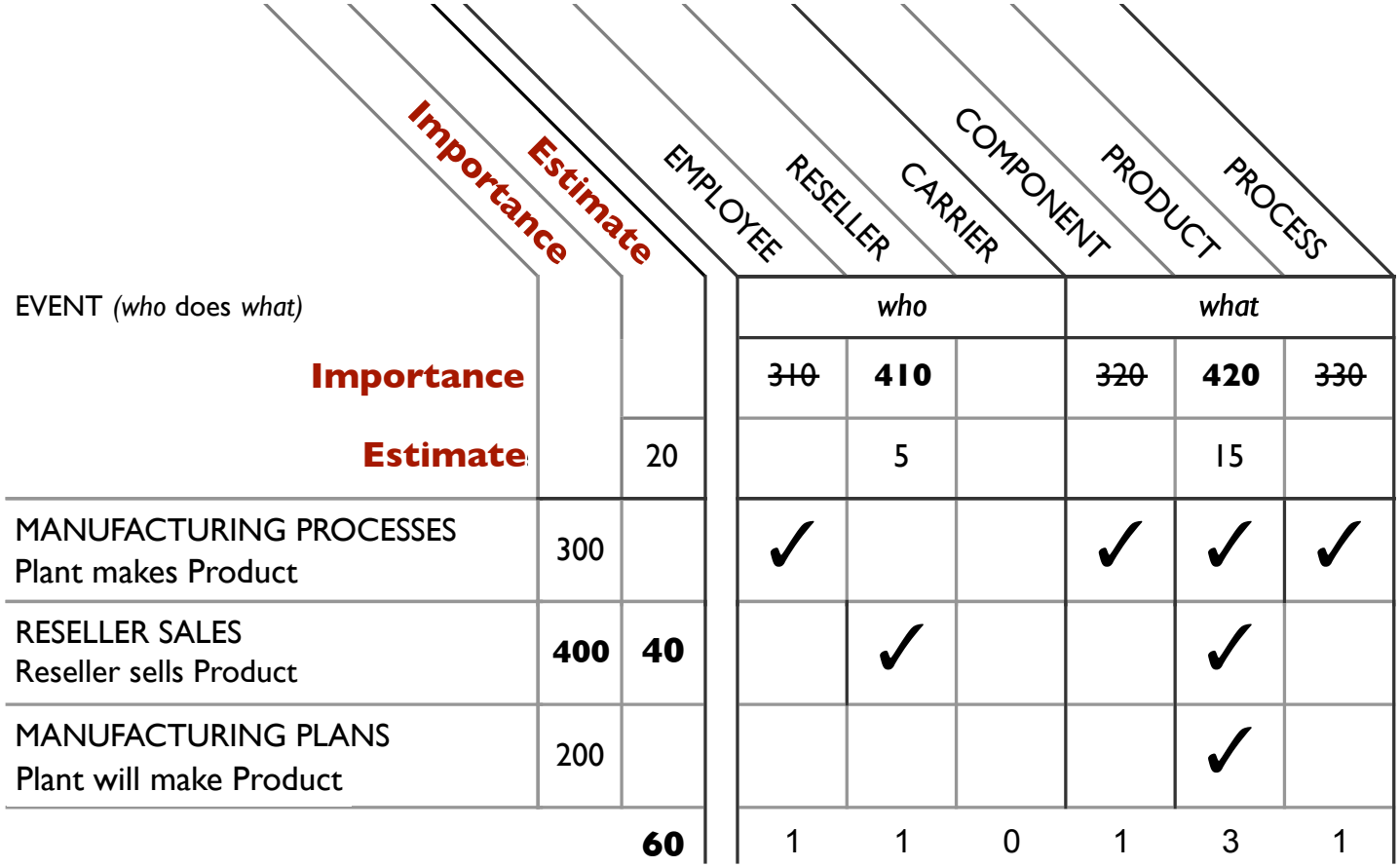

- Stakeholder Model Review



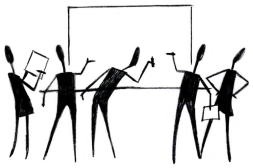

- Sprint Planning



# Event Matrix Estimating



| EVENT *(who does what)* | Importance | Estimate | EMPLOYEE | RESELLER | CARRIER | COMPONENT | PRODUCT | PROCESS |
|---|---|---|---|---|---|---|---|---|
| | | | *who* | | | *what* | | |
| **Importance** | | | ~~310~~ | **410** | | ~~320~~ | **420** | ~~330~~ |
| **Estimate** | | 20 | | 5 | | | 15 | |
| MANUFACTURING PROCESSES<br>Plant makes Product | 300 | | ✓ | | | ✓ | ✓ | ✓ |
| RESELLER SALES<br>Reseller sells Product | **400** | **40** | | ✓ | | | ✓ | |
| MANUFACTURING PLANS<br>Plant will make Product | 200 | | | | | | ✓ | |
| | | **60** | 1 | 1 | 0 | 1 | 3 | 1 |



# Exercise 9: Estimating and Refactoring

**Background**

Pomegranate have developed a periodic snapshot star schema for retails sales (see below). It currently holds 5 years worth of daily sales by product, promotion and store. It has been used to deliver many useful reports. However the business has evolved and the following enhancements/changes have been requested:

1. Reliable daily product **cost** figures are now available for the last 5 years. So too are **sales transaction counts** for each **product promotion** combination in each **store**. Make these additional facts available to allow margin and various basket count based metric to be calculated.
2. Store managers suspect that **weather** is a major casual factor in the success or failure of retail promotions. No historical data is available but sales and marketing would like to track local weather conditions in the future for sales trending.
3. Additional physical characteristics of products such as **packaging type**, **weight** and **size** are need for reporting purposes. The data is available in spreadsheet form matching the existing **Product** dimension.
4. A new EPOS system is being rolled out in all 50 stores over the next 6 months which will identify the **department** (up to 5 locations) within each **store** where product sales are recorded. This will provide more detailed data going forward. Stakeholders would like the existing schema to cope with this transition.

**Part 1 (Individually)**

Annotate the schema below to reflect the new requirements and think about their relative costs.

**Part 2 (Team)**

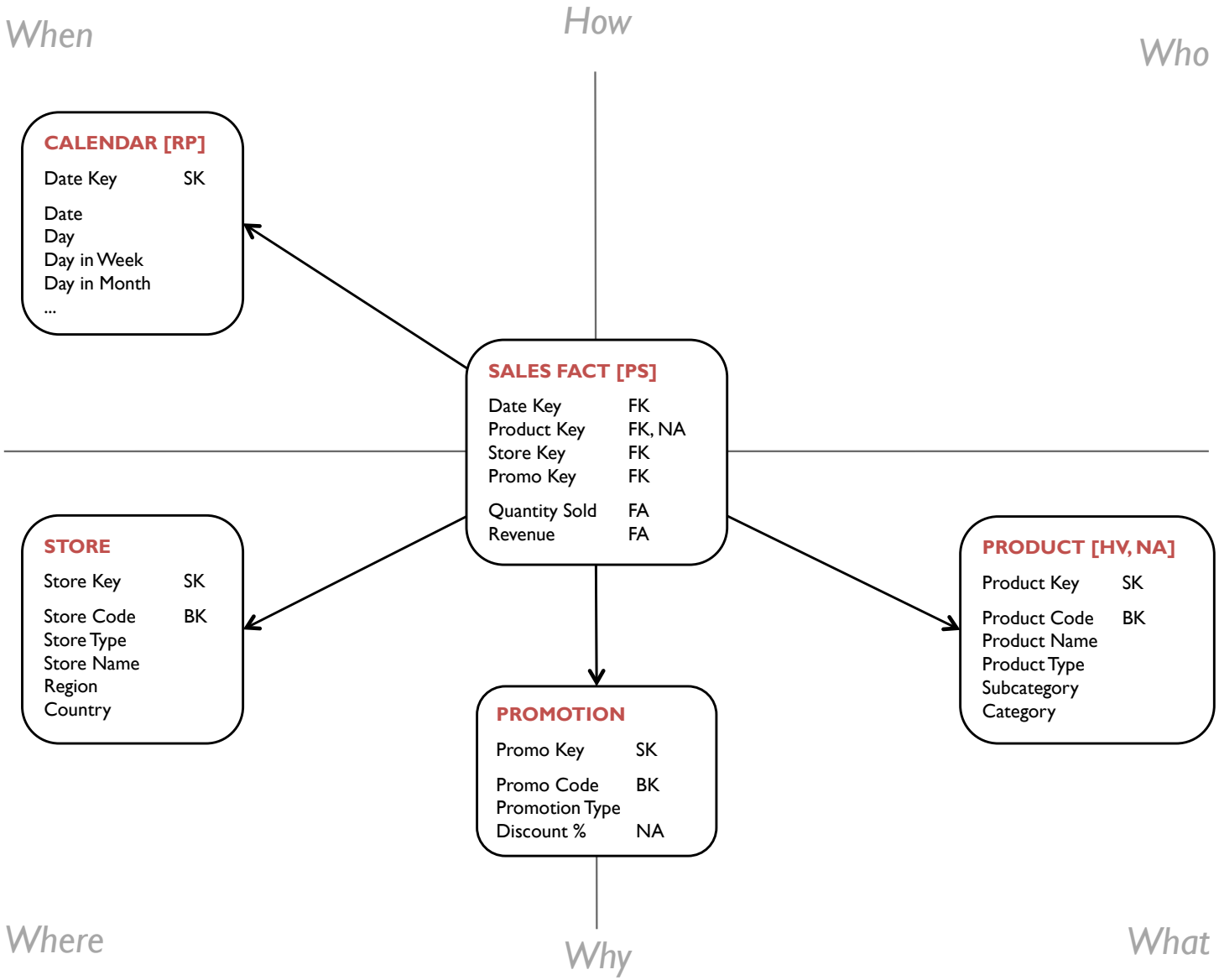
Model the existing periodic snapshot on a BEAM✲ canvas. Refactor your model to take account of the business changes and new data as it becomes available. Use different colour Post-its and the numbering scheme above to identified the new/changed elements of your design.

**Part 3 (Team)**

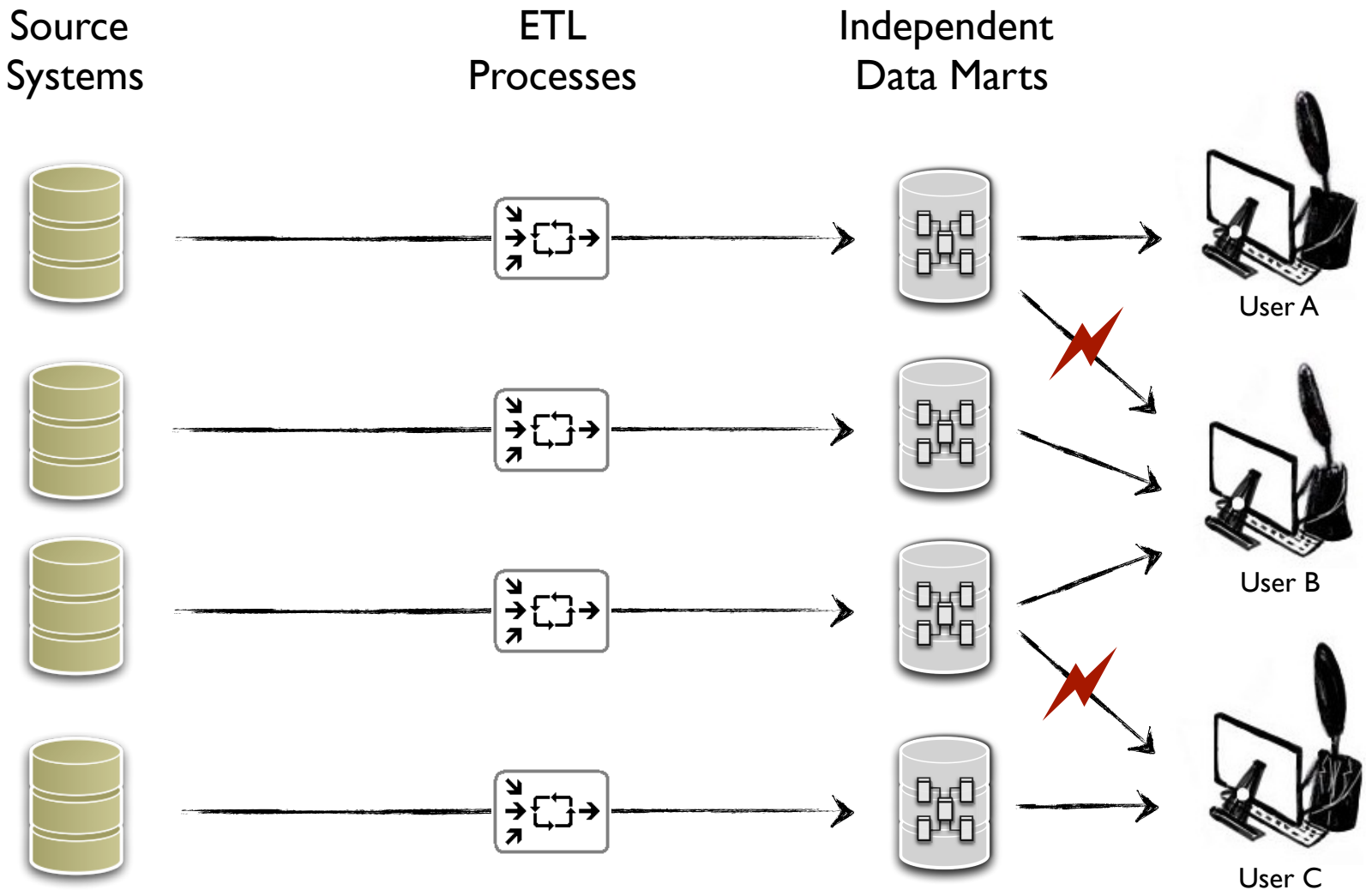
As a team estimate the relative costs of each of change. Play a simple form of planning poker with 4 cards (Post-its) each for **easy, cheap, difficult, severe/don't do**. Add your team estimates to you model and create a product backlog sorted in order of increasing complexity/workload with the easiest/cheapest refactors first.

# Current Design





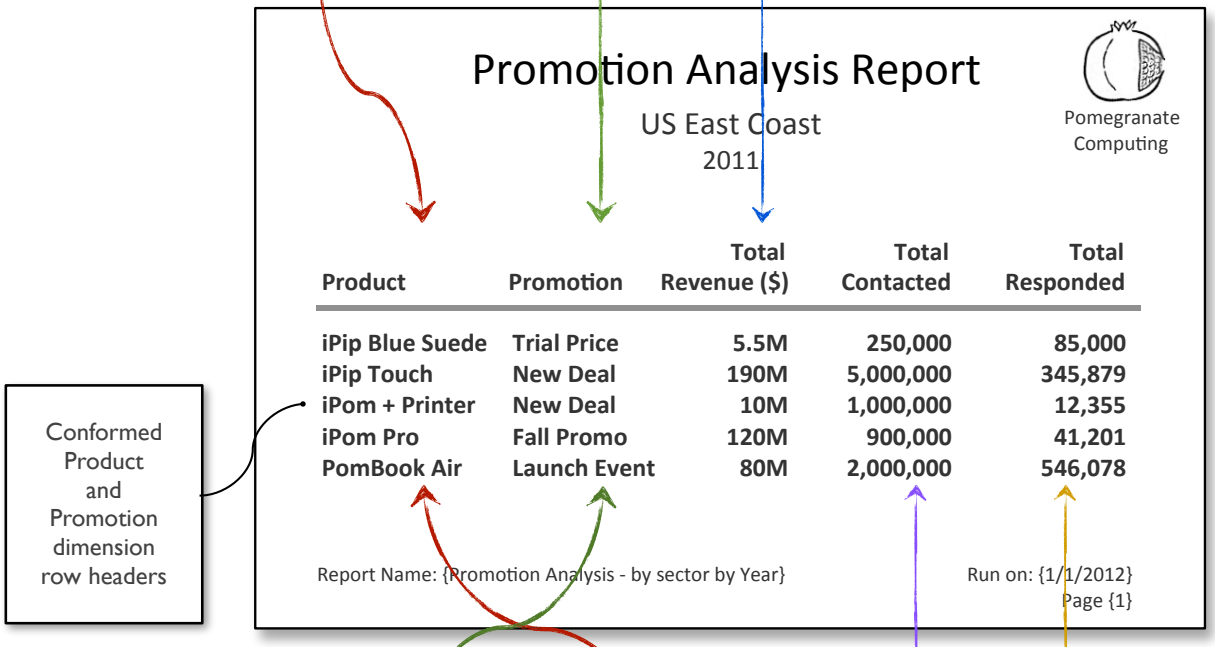
# Silo Data Marts Anti-Pattern: **Technical/Data Debt**



# Analysing Multiple Events with Common Details (Nouns)

**CUSTOMER ORDERS [DE]**

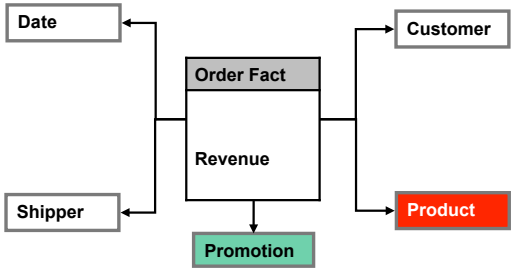
| CUSTOMER | **orders PRODUCT** | on ORDER DATE | **on PROMOTION** | for REVENUE |
|---|---|---|---|---|
| [who] | **[what]** | [when] | **[why]** | [$] |
| J. B. Priestley | **iPip Blue Suede** | 18-May-2011 | **Trial Price** | $249 |
| Vespa Lynd | **PomBook Air** | 29-Jun-2011 | **No Promotion** | $1,400 |
| Elvis Priestley | **iPip Blue Suede** | 18-May-2011 | **Trial Price** | $249 |
| Phillip Swallow | **iPom Pro** | 14-Oct-2011 | **Fall Promo** | £2,500 |
| Unknown | **iPip G1** | 10 Years Ago | **Launch Event** | $200,000 |
| US Senate | **iPom + Printer** | Yesterday | **New Deal** | $150,000 |

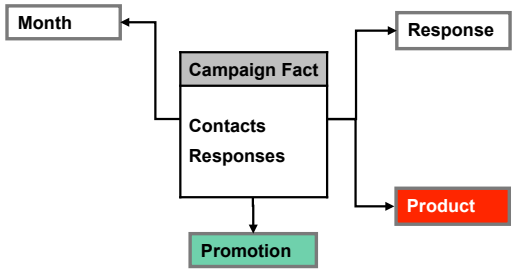



Conformed Product and Promotion dimension row headers

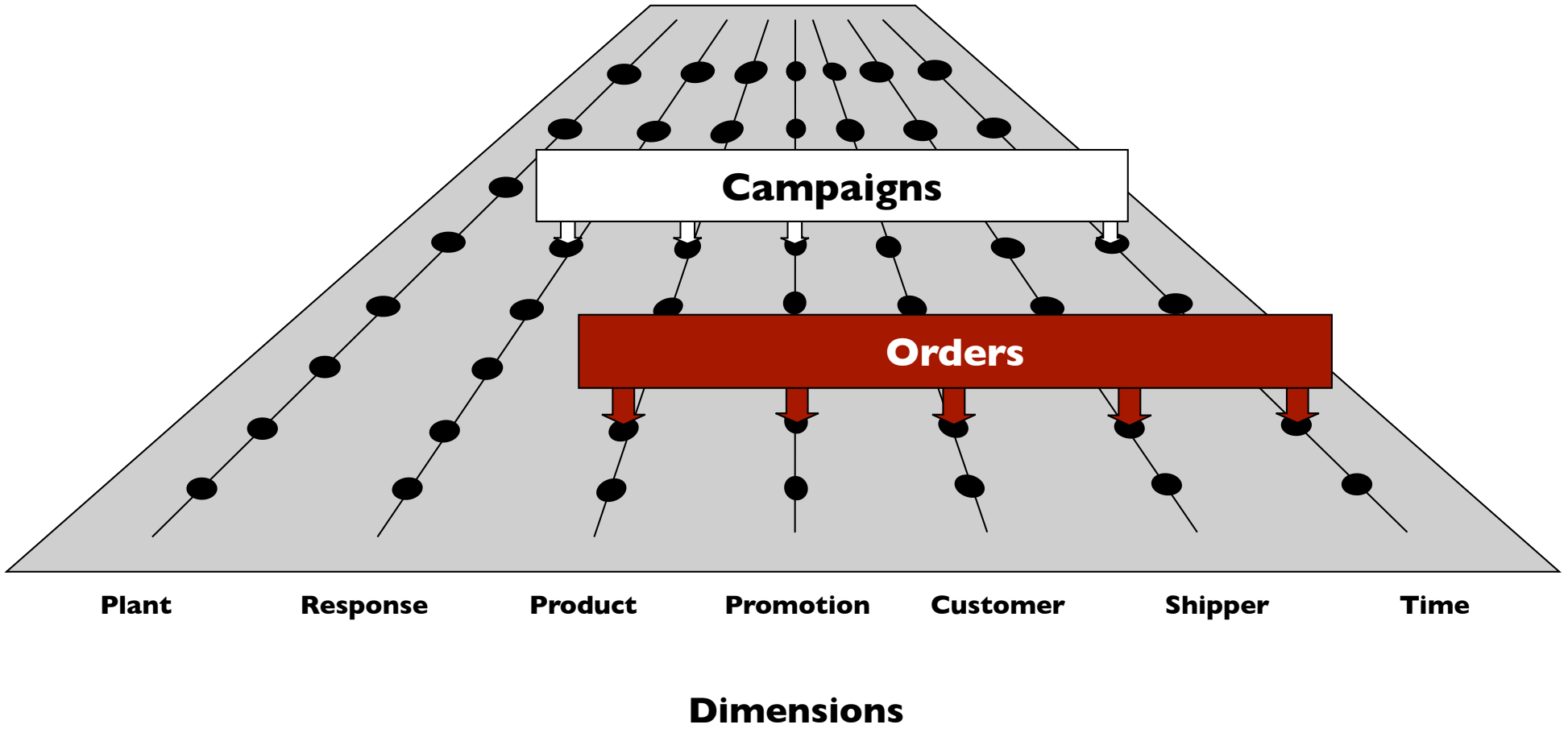
**PRODUCT CAMPAIGNS [PS]**

| **PROMOTION** | **advertises PRODUCT** | during CAMPAIGN MONTH | to CONTACTS | with RESPONSES |
|---|---|---|---|---|
| **[why]** | **[what]** | [when] | [people] | [people] |
| **Trial Price** | **iPip Blue Suede** | April 2011 | 12,000 | 64,000 |
| **Trial Price** | **iPip Blue Suede** | May 2011 | 13,000 | 21,000 |
| **Fall Promo** | **iPom Pro** | October 2011 | 900,000 | 41,201 |
| **Launch Event** | **PomBook Air** | Jan 2011 | 450,000 | 20,129 |
| **New Deal** | **iPom + Printer** | This Month | 1,000,000 | 12,355 |
| **New Deal** | **iPip Touch** | This Month | 5,000,000 | 345,879 |

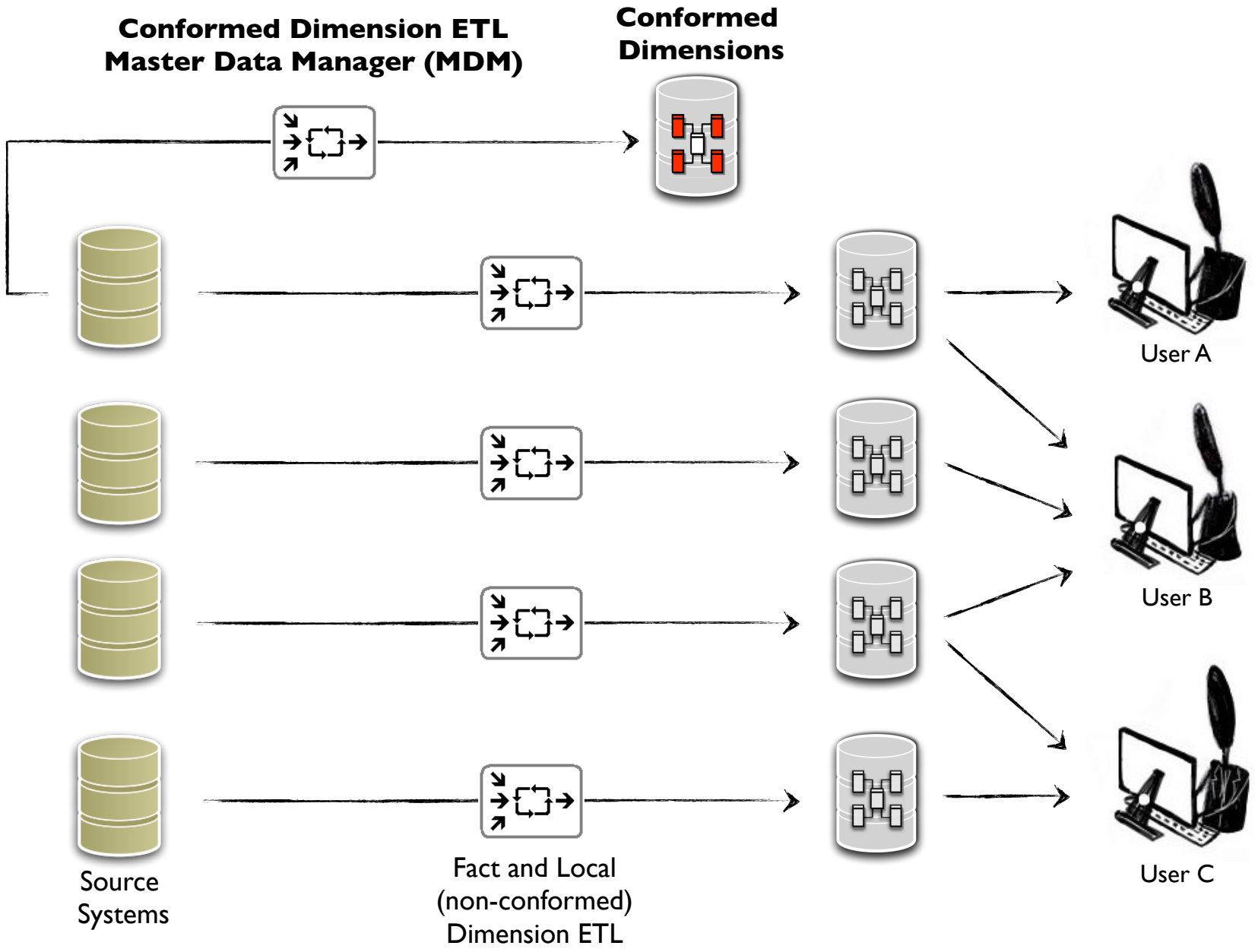


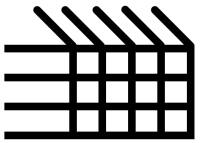



# Reuse: Kimball Bus Architecture



# Agile Dimensional Data Warehouse Pattern

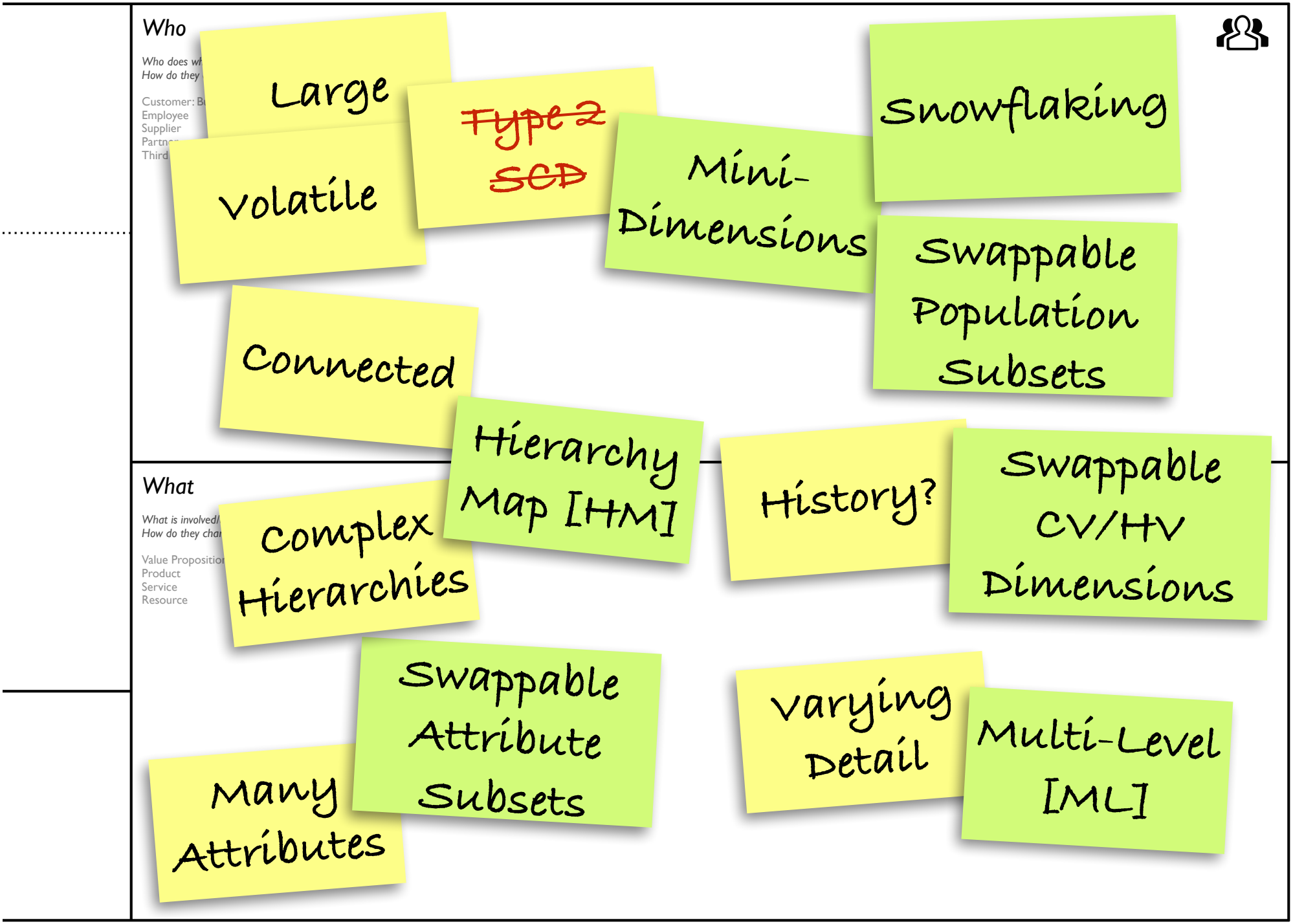




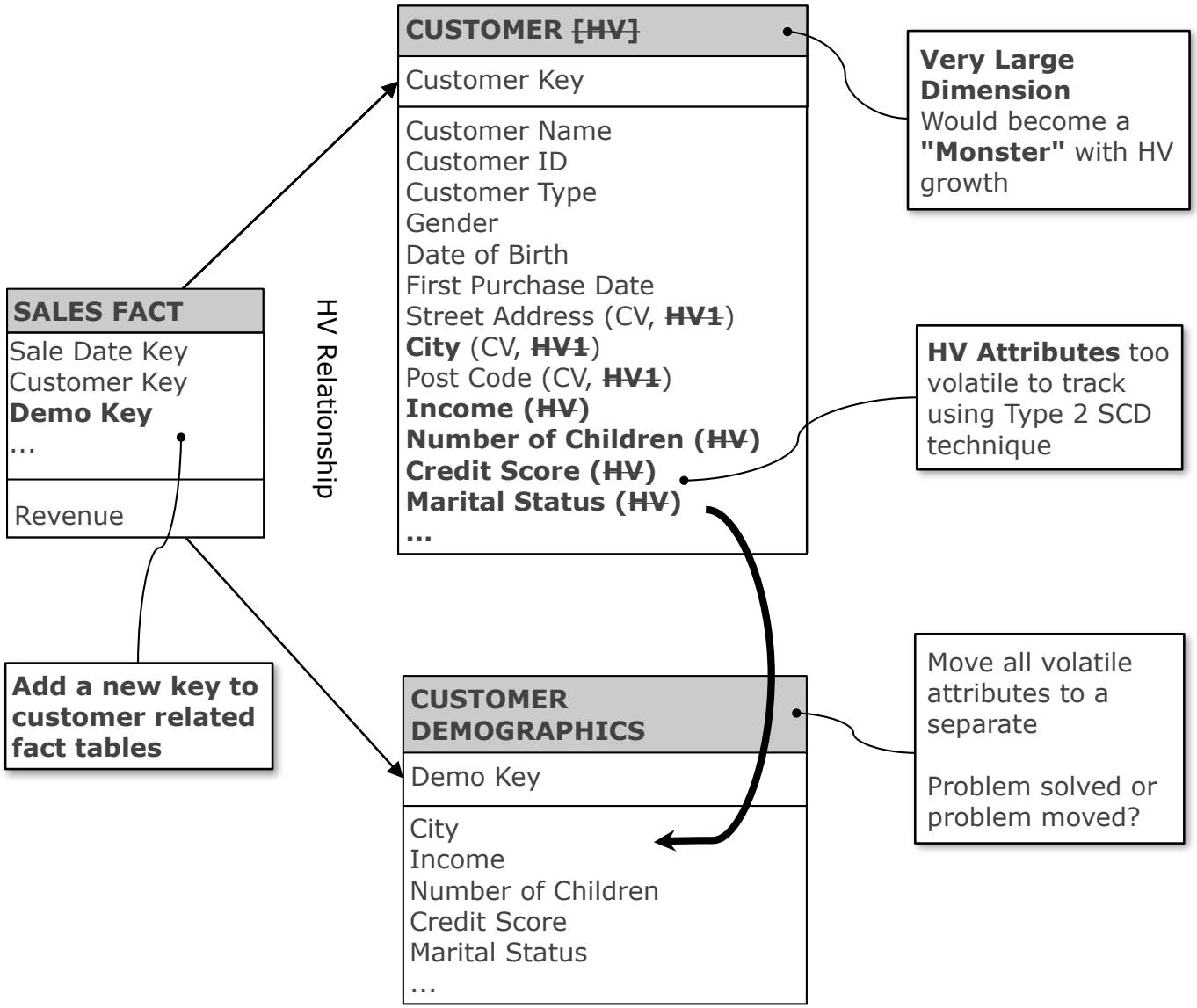
# Data Warehouse Matrix

| FACT TABLE/ CUBE | SOURCE App.Db.Schema.Table | FACT TYPE | VOLUMETRICS Initial | Increment | Max | When | who: EMPLOYEE [RP] | SUPPLIER | RESELLER | CARRIER | CUSTOMER | COMPONENT | PRODUCT | what: PRODUCT_TYPE [RU] | PROCESS | TEST | PLANT | where: WAREHOUSE | SALES_LOCATION | DELIVERY_ADDRESS | CONTRACT | why & how: PROMOTION | PROBLEM_REASON | SHIP_MODE | ORDER ID [DD] | SHIPMENT NUMBER [DD] | PO [DD] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MANUFACTURING_PLANS | ERP.PROD.MFG.MFG_PLAN | PS | 100K | 10K / Month | 500K | Month | | | | | | | | 3 | | | | ✓ | | | | | | | | | | |
| BILL_OF_MATERIALS | ERP.PROD.MFG.BILL_MAT | TF | 100K | 5K / Qtr | 1M | Day | | | | | | | ✓ | 3 | | | | | | | | | | | | | | |
| PURCHASE_ORDERS | ERP.PROD.INV.PO | AS | 500K | 10K / Day | 2M | Minute | ✓ | ✓ | | | | | ✓ | | | | | ✓ | | | | ✓ | | | | | | ✓ |
| COMPONENT_DELIVERIES | ERP.PROD.LOG.DLVRY | TF | 500K | 5K / Day | 1.5M | hour | | ✓ | | ✓ | | | ✓ | | | | | ✓ | | | | | | | ✓ | | | ✓ |
| SUPPLIER_PAYMENTS | ERP.PROD.GL.SUP_PMT | TF | 100K | 1K / Day | 1M | Day | ✓ | ✓ | | | | | ✓ | | | | | | | | | | | | | | | ✓ |
| COMPONENT_INVENTORY | ERP.PROD.INV.COMP_STK | PS | 100K | 100K / Day | 0.5M | Day | | | | | | | ✓ | | | | | ✓ | | | | | | | | | | |
| QA_TESTS | ERP.PROD.MFG.QA_TEST | TF | 100K | 10K / Day | 1M | Minute | ✓ | ✓ | | | | | ✓ | 3 | | | ✓ | ✓ | | | | | | | | | | |
| MANUFACTURING_PRODUCTION | ERP.PROD.MFG.AUDIT | TF | 5M | 1M/ Day | 700M | Minute | ✓ | | | | | | ✓ | 3 | | ✓ | | ✓ | | | | | | | | | | |
| PRODUCT_INVENTORY | ERP.PROD.INV.PRD_STK | TF | 100K | 10K / Day | 0.5M | Day | | | | | | | | 3 | | | | | ✓ | | | | | | | | | |
| WAREHOUSE_SHIPMENTS | ERP.PROD.LOG.SHIPMENT | TF | 100K | 10K / Day | 1M | hour | | | ✓ | ✓ | | | | 3 | | | | | ✓ | ✓ | | ✓ | | | ✓ | | | |
| SALES TARGETS | EXCEL.SPLAN.CURRENT | PS | 10K | 1K / Month | 100K | Month | ✓ | | | | | | | | ✓ | | | | | ✓ | | | | | | | | |
| CUSTOMER_ORDERS | POS.PROD.SALES.ORDER | AS | 500K | 100K / Day | 2M | Minute | ✓ | | | | ✓ | | | 1 | ✓ | | | | | ✓ | ✓ | | ✓ | | | ✓ | | |
| PRODUCT_SHIPMENTS | ERP.PROD.LOG.SHIPMENT | TF | 400K | 100K / Day | 1.5M | Hour | ✓ | | | ✓ | ✓ | | | 3 | | | | | ✓ | | ✓ | | | | ✓ | ✓ | ✓ | |
| CARRIER_DELIVERIES | ERP.PROD.LOG.DELIVERY | TF | 200K | 50K / Day | 1M | Hour | | | | ✓ | ✓ | | | 3 | ✓ | | | | | | ✓ | | | | ✓ | ✓ | ✓ | |
| CUSTOMER_COMPLAINTS | CRM.PROD.CRM.CONTACT | AS | 50K | 1K / Day | 0.5M | Minute | ✓ | | | | ✓ | | | 2 | | | | | | | | | ✓ | | | ✓ | | |
| PRODUCT_RETURNS | ERP.PROD.LOG.RETURN | TF | 10K | 1K / Day | 0.3M | Day | | | | | ✓ | | | 3 | | | | | | | | | ✓ | | | ✓ | | |

# *Who* and *What* Patterns (Chapter 6)



# Very Large Dimensions: Volatile B2C Customers



# Mini-Dimension Pattern

**Problem:**

- Tacking history in very large dimensions (e.g. B2C Customer): Volatile HV (Type 2 SCD) attributes cause explosive growth

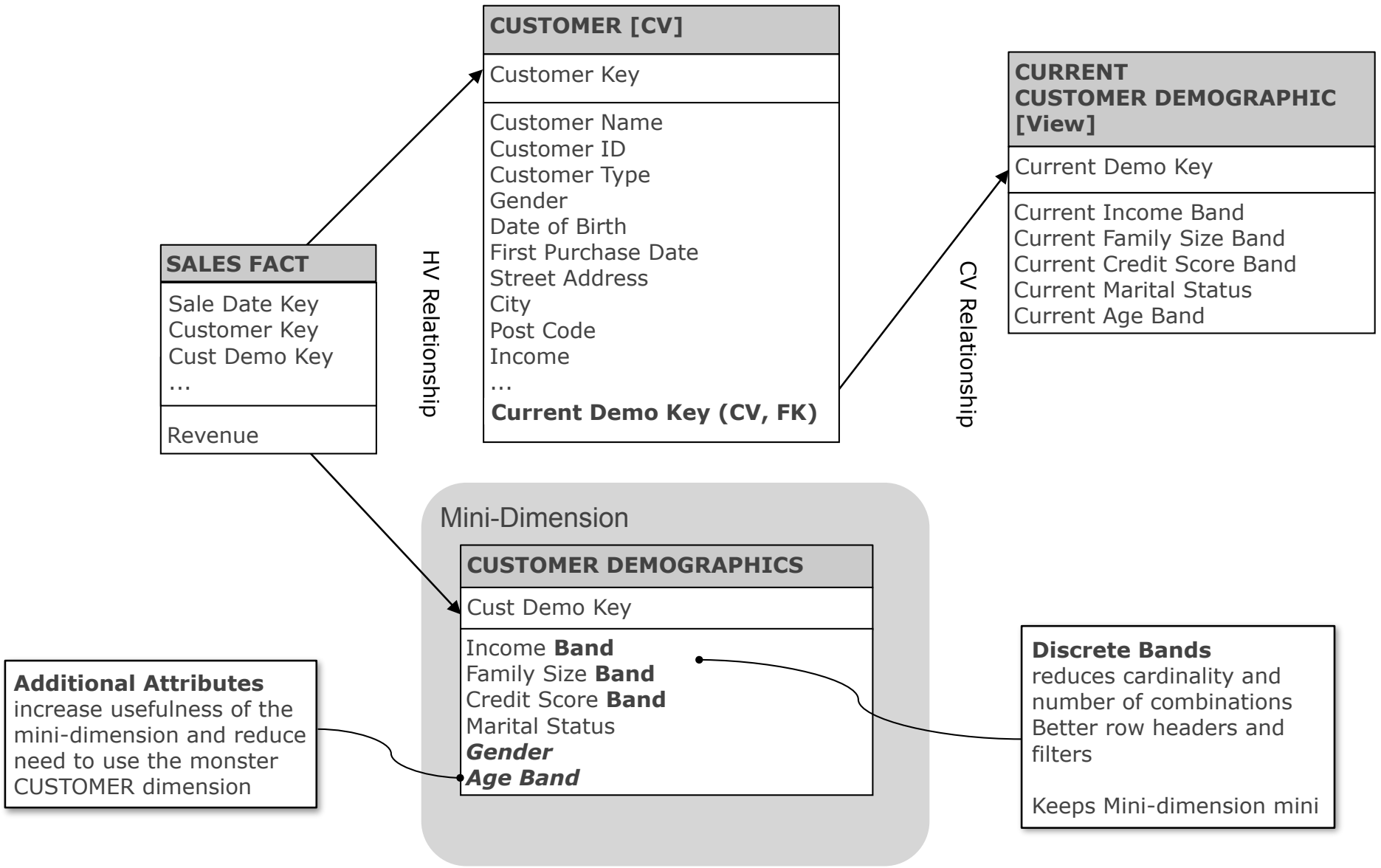
**Solution:**

- Move HV attributes to their own mini-dimension(s).
- Limit number of HV attributes and reduce cardinality to reduce combinations and provide better dimensional attributes.
- Add mini-dimension foreign key to main dimension to handle ETL processing and current value queries (short cut join)

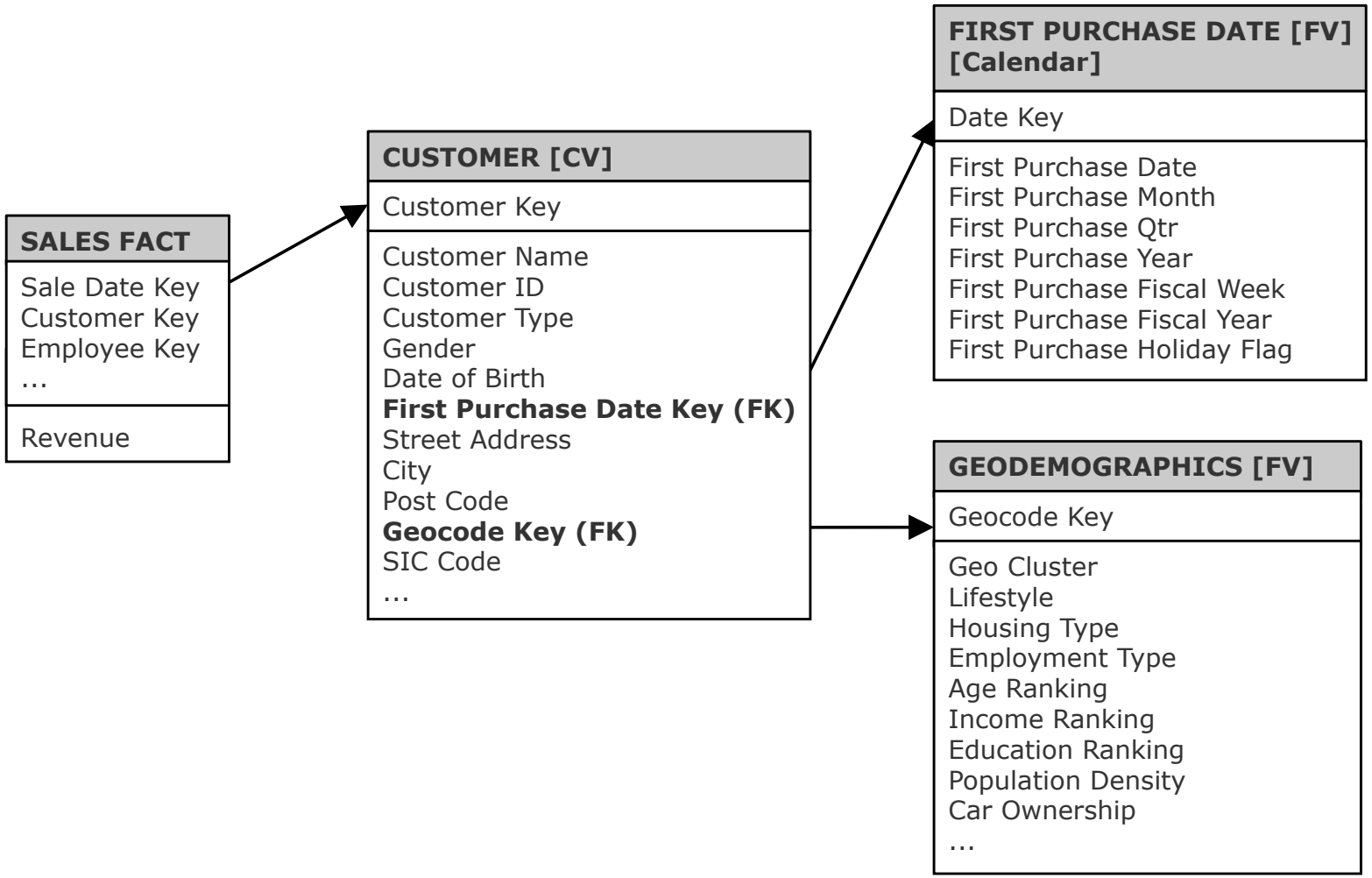
**Consequences:**

- Additional ETL processing to maintain mini dimensions and assign mini-dimension keys to the main dimension. Extra dimension keys in the fact tables. Loss of detail with cardinality reduction.

# B2C Customer Mini-Dimension



# Acceptable Snowflake Patterns

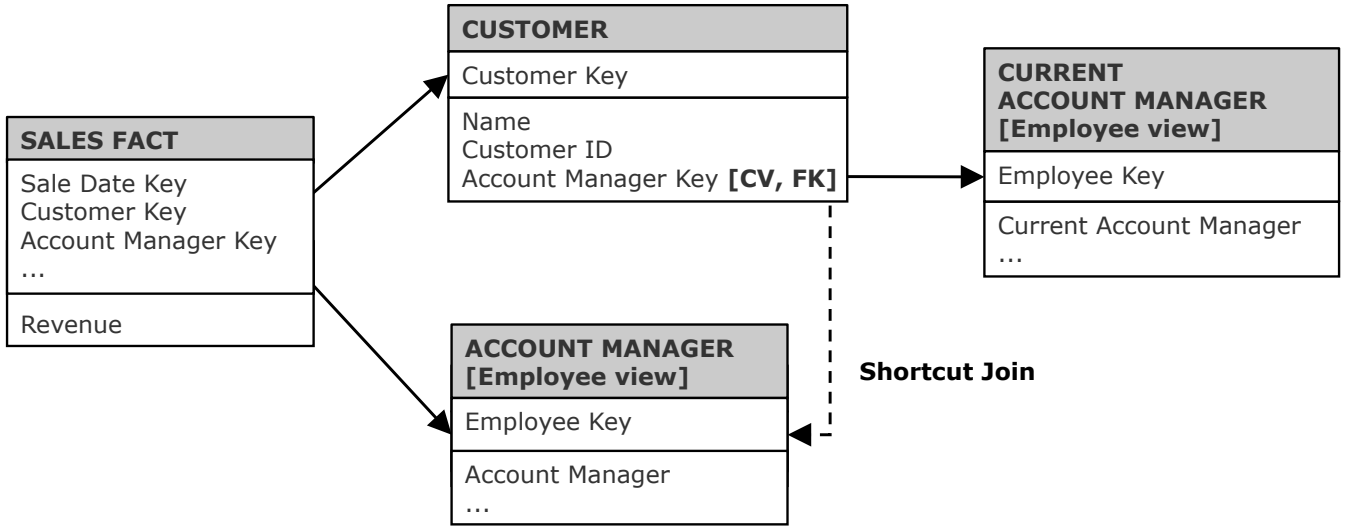


# More Acceptable Snowflaking

**Embedded *whos***

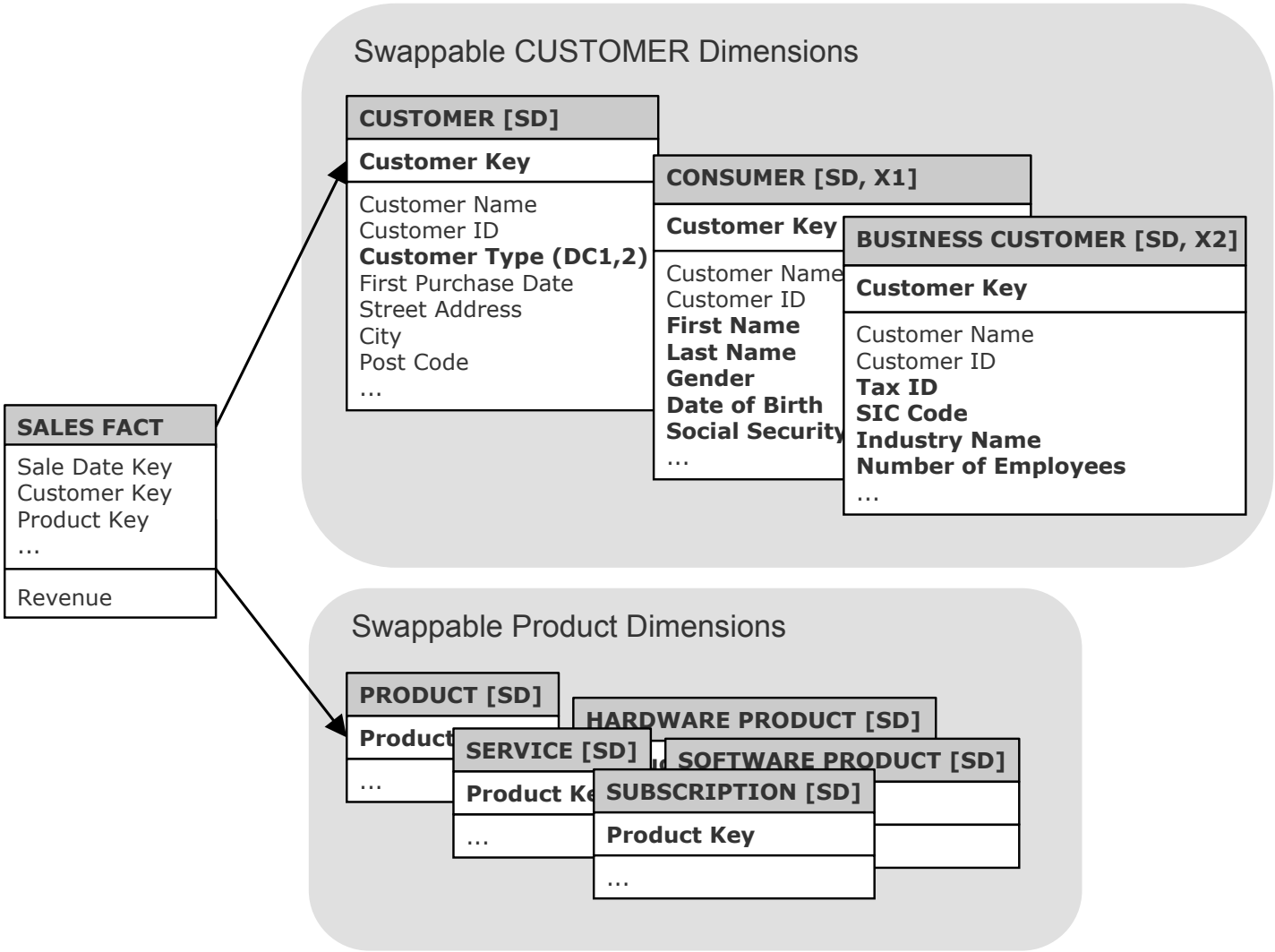
**CUSTOMER [HV]**

| CUSTOMER NAME | CUSTOMER ID | CUSTOMER TYPE | CUSTOMER CATEGORY | ACCOUNT MANAGER | PARENT COMPANY |
|---|---|---|---|---|---|
| CV | BK | FV | HV | [EMPLOYEE] HV | [CUSTOMER] CV |
| Pomegranate | BC2349 | Business | Good | LC | None |
| iPip Design | BC2570 | Business | Great | JS | POM Computing |
| PicCzar Movies | BC2571 | Business | Good | LC | Pomegranate |
| POM Computing | BC2565 | Business | OK | LC | Pomegranate |
| POM Store | BC2567 | Business | Great | JS | POM Computing |
| POMstore.co.uk | BC2569 | Business | OK | LC | POM Store |




ADWD pp. 173-174

# Swappable Dimensions [SD]



# Exclusive Attributes: SD Requirements

Does every [*dimension*] have a [*candidate attribute*]?

***No***

**Defining Characteristic**

**Exclusive attributes for Consumers**

**Exclusive attributes for Businesses**

**CUSTOMER**

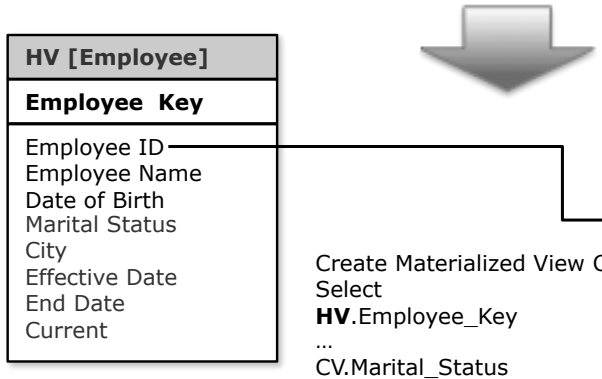
| CUSTOMER NAME | CUSTOMER ID | is CUSTOMER TYPE | born on DATE OF BIRTH | is GENDER | has SIC CODE | has EMPLOYEE COUNT |
|---|---|---|---|---|---|---|
| MD | BK, MD | MD, **DC** | **X1** | **X1, MD** | **X2, MD** | **X2** |
| Elvis Priestley | C0010 | Consumer | 13-Sep-1894 | Male | N/A | - |
| Unknown | N/A | Unknown | - | Unknown | Unknown | - |
| Vesp Lynd | C0997 | Consumer | 5-Jul-1980 | Female | N/A | - |
| Phillip Swallow | C9990 | Consumer | 1-Mar-1975 | Male | N/A | - |
| US Senate | B0023 | Government | N/A | N/A | 6111 | 100 |
| International Exports | B007 | Business | N/A | N/A | 9121 | 9 |

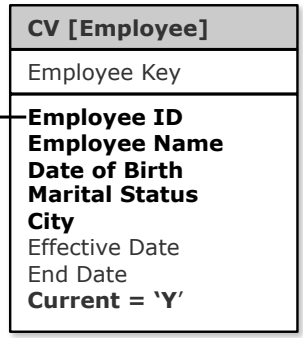
# **CV/HV** – Hybrid Slowly Changing Dimension Pattern

**EMPLOYEE [HV, SD]**

| EMPLOYEE KEY | EMPLOYEE NAME | EMPLOYEE ID | DATE OF BIRTH | MARITAL STATUS | CITY | from EFFECTIVE DATE | to END DATE | is CURRENT |
|---|---|---|---|---|---|---|---|---|
| SK | CV | BK | FV | **HV/CV** | **HV/CV** | | | |
| 1010 | James Bond | 007 | 14/05/1953 | Single | London | 1-Jan-1962 | 4-Apr-1969 | N |
| 2099 | James Bond | 007 | 14/05/1953 | Married | Geneva | 5-Apr-1969 | 5-Apr-1969 | N |
| **2120** | **James Bond** | **007** | **14/05/1953** | **Widowed** | **London** | **6-Apr-1969** | **1-Jan-3000** | **Y** |

**Current version of Bond**
Create a CV view by joining this to every version of Bond



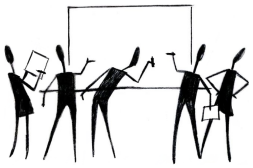




```
Create Materialized View CV_Employee
Select
HV.Employee_Key
…
CV.Marital_Status
CV.City
From Employee HV, Employee CV
Where
HV.Employee_ID = CV.Employee_ID
And CV.Current = 'Y'
```

**CV EMPLOYEE [CV, SD]**

| EMPLOYEE KEY | EMPLOYEE NAME | EMPLOYEE ID | DATE OF BIRTH | MARITAL STATUS | CITY |
|---|---|---|---|---|---|
| SK | CV | BK | FV | **CV** | **CV** |
| 1010 | James Bond | 007 | 14/05/1953 | Widowed | London |
| 2099 | James Bond | 007 | 14/05/1953 | Widowed | London |
| 2120 | James Bond | 007 | 14/05/1953 | Widowed | London |

**Current versions**
All Bond rows are identical *except* for their surrogate keys



# Exercise 10: Web Analytics

**Requirements**

BI Stakeholders in many different departments are interested in tracking and understanding the behaviour of visitors to the Pomegranate site. Raw weblogs, content management and product catalog data is readily available but data volumes are high.

**Part 1.**

Using a BEAM canvas develop a simple initial dimensional model capable of easily answering as many of the following typical questions with lightweight ETL processing (without processing input data multiple times or increasing the number of data sources).

Where do visitors come from?

How long do they spend viewing content?

What type of content is viewed most frequently? What content is unpopular?

What time of day and year do we get the most traffic?

How do visitor view our website? From mobile devices or desktops/laptops? What type of visit is more successful?

Why do they visit our website?

How many visits convert into sales?

How much time is spend navigating before visitors/customers can view or buy products? Break that down by product category and brand.

How many computing vs entertainment pages do visitors view per visit?

**Part 2.**

Which are the more complex questions that can't be answered by your model initially?

# Multi-Level [ML] Dimension Pattern

**Problem:**

- The hierarchical level of dimension varies for a business event. E.g., web page views for individual products and whole product categories. Sales transactions usually attributed to a salesperson but some assign to branch office when salesperson is missing

**Solution:**

- Create **multi-level [ML]** dimensions by adding additional members to represent parent-only values in the dimensional hierarchies

**Consequences:**

- Additional complexity

- Open to abuse: Changing the meaning of facts.  Mixing fact types. Creating generic stars

# Multi-Level Dimension

**Multi-Level (ML) Dimension**
contains additional records representing higher levels in the product hierarchy

**LEVEL TYPE**
Majority of rows will be **Product**

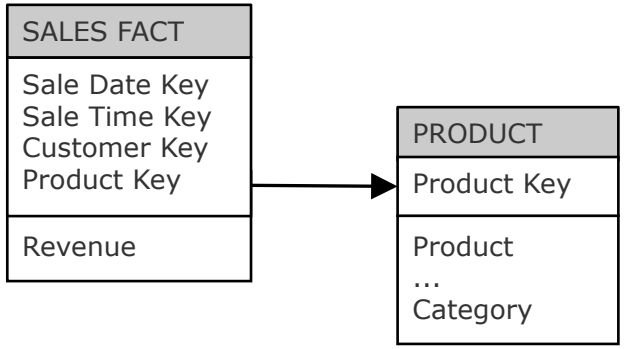
**PRODUCT [ ML ]**

| PRODUCT KEY | PRODUCT DESCRIPTION | Weight | PRODUCT TYPE | BRAND | SUB CATEGORY | CATEGORY | **LEVEL TYPE** |
|---|---|---|---|---|---|---|---|
| SK | MD, FV | MD | MD, HV | MD, HV | MD, HV | MD, HV | **NN** |
| 1002 | iPip G1 | 500g | MP3 | iPip | Music | Entertaiment | **Product** |
| 1004 | POMBook Air | 1Kg | Laptop | POMBook | Mobile | Computing | **Product** |
| 1008 | iPOM + Printer | 6Kg | Bundle | iPOM | Desktop | Computing | **Product** |
| 1009 | iPip Touch | 120g | PDA | iPip | Multi-Media | Entertaiment | **Product** |
| 0 | Not Available | N/A | N/A | N/A | N/A | N/A | **Missing** |
| **-1** | **Not Applicable** | **N/A** | **N/A** | **N/A** | **N/A** | **Computing** | **Category** |
| **-2** | **Not Applicable** | **N/A** | **N/A** | **N/A** | **N/A** | **Entertainment** | **Category** |
| **-3** | **Not Applicable** | **N/A** | **N/A** | **N/A** | **Desktop** | **Computing** | **Subcategory** |
| **-4** | **Not Applicable** | **N/A** | **N/A** | **iPOM** | **Desktop** | **Computing** | **Brand** |

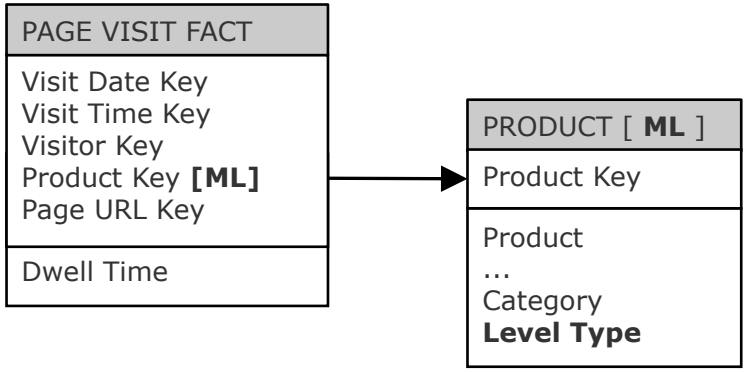
**Category Level Row**
representing all **Entertainment** products

# Using Single and Multi-Level Dimensions

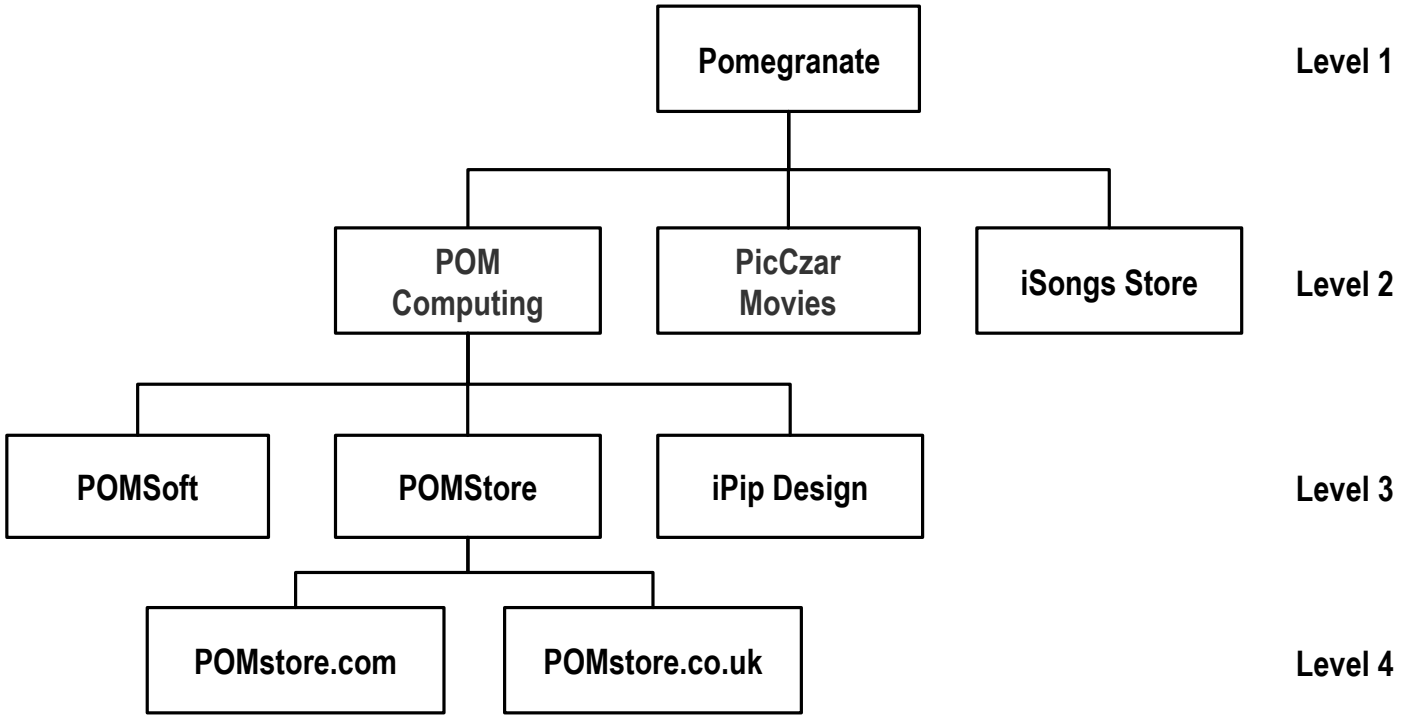
- **Single Level Facts** – All sales are for products



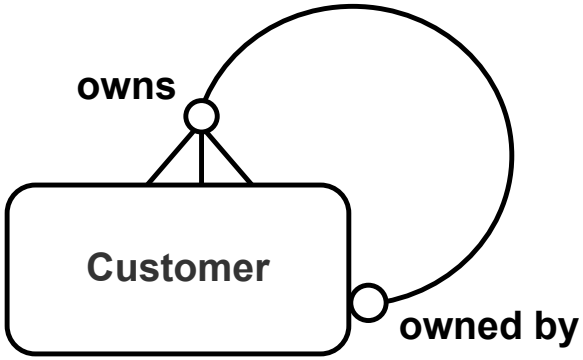

- **Multi-Level Facts** – Page visits can be for products, subcategories or categories



# Variable-Depth Hierarchies



# Recursive Relationships




**CUSTOMER**

| CUSTOMER KEY | CUSTOMER NAME | CUSTOMER ID | PARENT KEY |
|---|---|---|---|
| SK | | BK | RK, CV |
| 101 | iPip Design | BC2570 | **104** |
| 102 | iSongs Store | BC2572 | ***106*** |
| 103 | PicCzar Movies | BC2571 | ***106*** |
| 104 | POM Computing | BC2565 | ***106*** |
| 105 | POM Store | BC2567 | **104** |
| 106 | Pomegranate | BC2349 | **-** |
| 107 | POMSoft | BC2566 | **104** |
| 108 | POMstore.co.uk | BC2569 | **105** |
| 109 | POMstore.com | BC2588 | **105** |
| 110 | PicCzar Movies | BC2571 | **106** |

**RK** - Recursive Foreign Key points to the primary key of the same table. Represents a M:1 recursive relationship.

Several customers are owned by 106 - the current version of Pomegranate

# Hierarchy Map [HM] Dimension Pattern

**Problem:**

- Dimension contains a recursive relationship (a recursive/reflective foreign key RK) – a head scratcher – which represents a variable-depth hierarchy

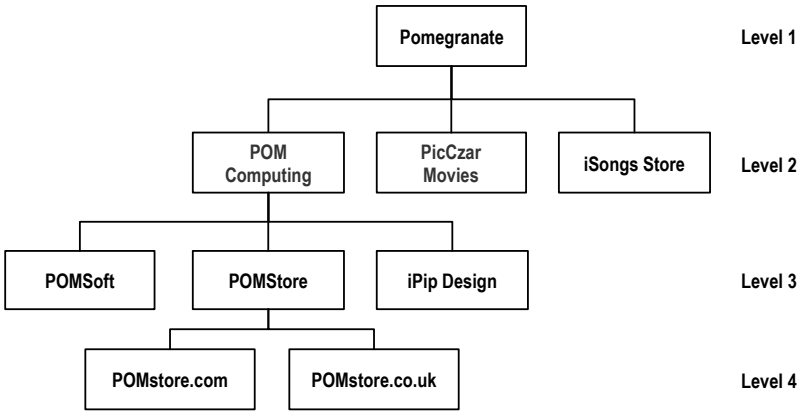
**Solution:**

- Check if the recursive relationship really represents a variable-depth hierarchy. If it is balanced or slightly ragged flatten it into the dimension as a fixed number of attributes and fill in gaps
- If truly variable-depth then 'explode' the hierarchy into a **hierarchy map [HM]**

**Consequences:**

- Complexity
- Default BI tool drilling features do not work. Use hyperlinked recursive reports request instead
- Exponential growth if you need to track hierarchy history

# Example Hierarchy Map: Company Structure [HM]




| Company Structure |
|---|
| **Parent Key**<br>**Subsidiary Key** |
| **Company Level**<br>Sequence Number<br>Lowest Subsidiary<br>Highest Parent |

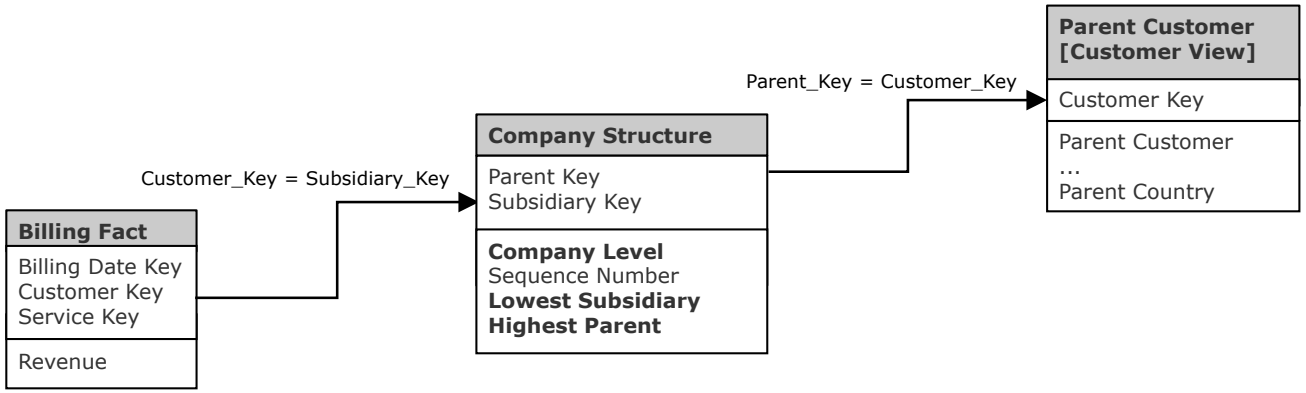
**COMPANY STRUCTURE [HM]**

| PARENT KEY | **owns** SUBSIDIARY KEY | COMPANY LEVEL | SEQUENCE NUMBER | LOWEST SUBSIDIARY | HIGHEST PARENT |
|---|---|---|---|---|---|
| SK, PK | SK, PK | N | N | [Y/N] | [Y/N] |
| Pomegranate | Pomegranate | 1 | 1 | N | Y |
| Pomegranate | POM Computing | 2 | 2 | N | Y |
| Pomegranate | POMSoft | 3 | 3 | Y | Y |
| Pomegranate | POM Store | 3 | 4 | N | Y |
| Pomegranate | POMstore.com | 4 | 5 | Y | Y |
| Pomegranate | POMstore.co.uk | 4 | 6 | Y | Y |
| Pomegranate | iPip Design | 3 | 7 | Y | Y |
| Pomegranate | PicCzar Movies | 2 | 8 | Y | Y |
| Pomegranate | iSongs Store | 2 | 9 | Y | Y |
| POM Computing | POM Computing | 1 | 1 | N | N |
| POM Computing | POMSoft | 2 | 2 | Y | N |
| POM Computing | POM Store | 2 | 3 | N | N |
| POM Computing | POMstore.com | 3 | 4 | Y | N |
| POM Computing | POMstore.co.uk | 3 | 5 | Y | N |
| POM Computing | iPip Design | 2 | 6 | Y | N |
| PicCzar Movies | PicCzar Movies | 1 | 1 | Y | N |
| iSongs Store | iSongs Store | 1 | 1 | Y | N |
| POMSoft | POMSoft | 1 | 1 | Y | N |
| POM Store | POM Store | 1 | 1 | N | N |
| POM Store | POMstore.com | 2 | 2 | Y | N |
| POM Store | POMstore.co.uk | 2 | 3 | Y | N |
| iPip Design | iPip Design | 1 | 1 | Y | N |
| POMstore.com | POMstore.com | 1 | 1 | Y | N |
| POMstore.co.uk | POMstore.co.uk | 1 | 1 | Y | N |

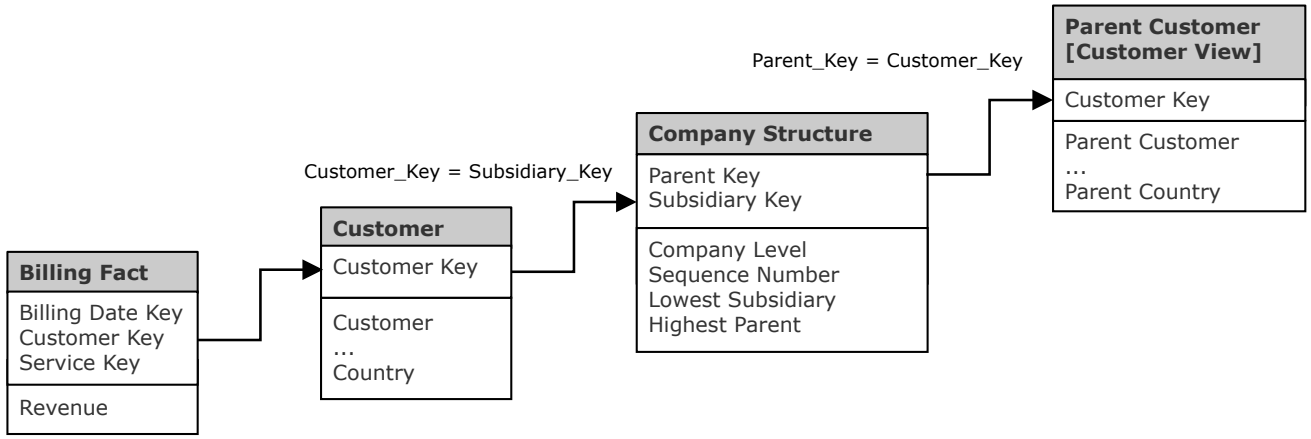
Explode every tree! One 9 branches tree becomes 24 records made up of 1 exploded tree and 8 fully exploded sub-trees
No of Records = ∑( branches at a level x level ) or overestimate = dimension cardinality x levels

# Using a Hierarchy Map

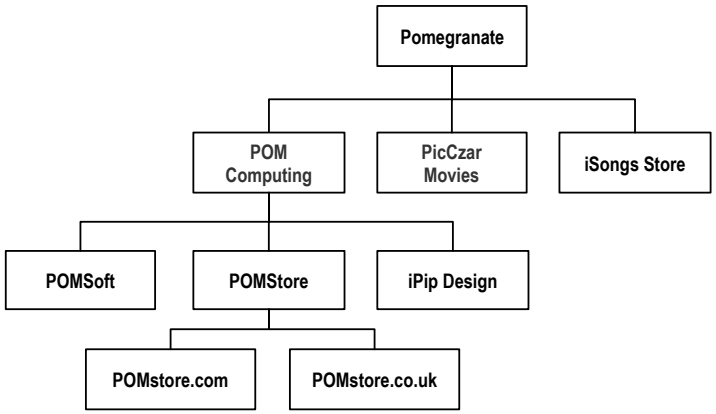
- Summarise *all*, *immediate*, *intermediate* or *lowest subsidiary* facts by Parent Customer:

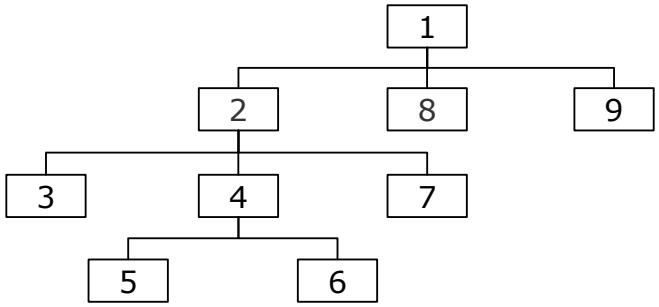



- Browse Customer organization structure and summarise the facts:



# Hierarchy Sequence Number



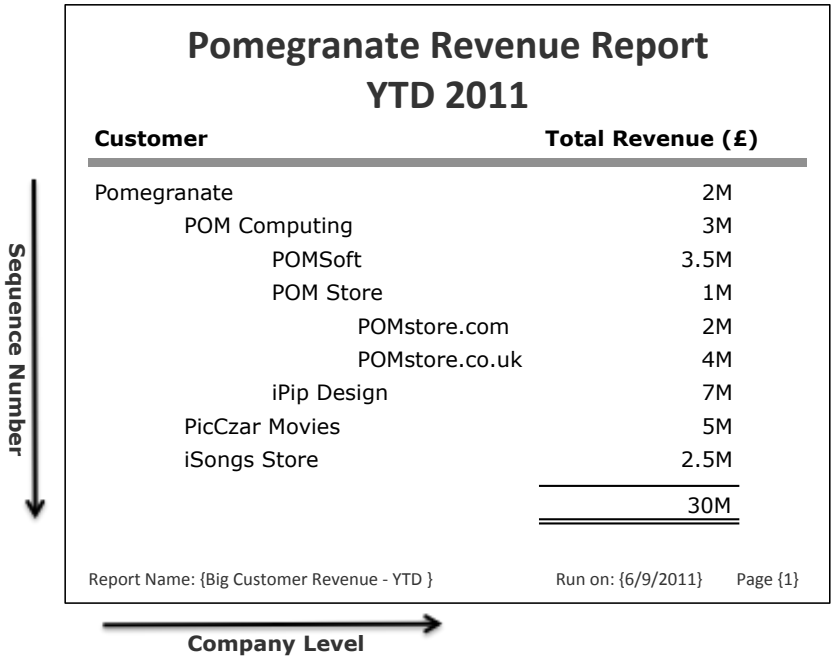

**COMPANY STRUCTURE [HM]**

| PARENT KEY | **owns** SUBSIDIARY KEY | COMPANY LEVEL | **SEQUENCE NUMBER** | LOWEST SUBSIDIARY | HIGHEST PARENT |
|---|---|---|---|---|---|
| SK, PK | SK, PK | N | **N** | [Y/N] | [Y/N] |
| Pomegranate | Pomegranate | 1 | **1** | N | Y |
| Pomegranate | POM Computing | 2 | **2** | N | Y |
| Pomegranate | POMSoft | 3 | **3** | Y | Y |
| Pomegranate | POM Store | 3 | **4** | N | Y |
| Pomegranate | POMstore.com | 4 | **5** | Y | Y |
| Pomegranate | POMstore.co.uk | 4 | **6** | Y | Y |
| Pomegranate | iPip Design | 3 | **7** | Y | Y |
| Pomegranate | PicCzar Movies | 2 | **8** | Y | Y |
| Pomegranate | iSongs Store | 2 | **9** | Y | Y |
| POM Computing | POM Computing | 1 | **1** | N | N |
| POM Computing | POMSoft | 2 | **2** | Y | N |
| POM Computing | POM Store | 2 | **3** | N | N |
| POM Computing | POMstore.com | 3 | **4** | Y | N |
| POM Computing | POMstore.co.uk | 3 | **5** | Y | N |
| POM Computing | iPip Design | 2 | **6** | Y | N |
| PicCzar Movies | PicCzar Movies | 1 | **1** | Y | N |
| iSongs Store | iSongs Store | 1 | **1** | Y | N |
| POMSoft | POMSoft | 1 | **1** | Y | N |
| POM Store | POM Store | 1 | **1** | N | N |
| POM Store | POMstore.com | 2 | **2** | Y | N |
| POM Store | POMstore.co.uk | 2 | **3** | Y | N |
| iPip Design | iPip Design | 1 | **1** | Y | N |
| POMstore.com | POMstore.com | 1 | **1** | Y | N |
| POMstore.co.uk | POMstore.co.uk | 1 | **1** | Y | N |

# Variable Depth Hierarchy Reporting

- Textual representation of hierarchy – **sort** by sequence number, **indent** by company level e.g.:

```
LPAD(' ', 3*(company_level-1)) || customer_name
```

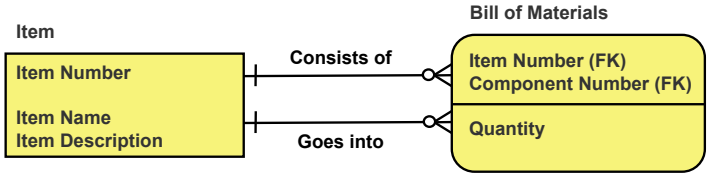


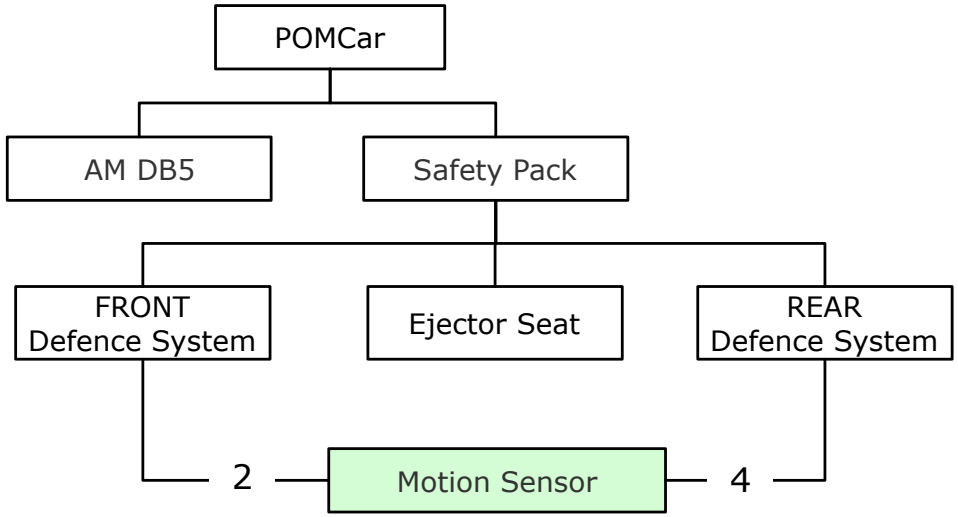

- Graphical representation – Visio organization chart wizard – database driven, display facts totals

- Provide drill-down by formatting row labels as hyperlinks which invoke new instances of the same report

# Multi-Parent, Variable Depth Hierarchy Map [ MV, HM ]

## Bill of Materials: a M:M Recursive relationship

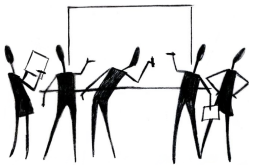



| Parts Explosion [MV, HM] |
|---|
| Product Key (PK, SK)<br>Component Key (PK, SK) |
| Part Level<br>Sequence Number<br>Subassembly<br>Quantity |

**PARTS EXPLOSION [MV, HM]**

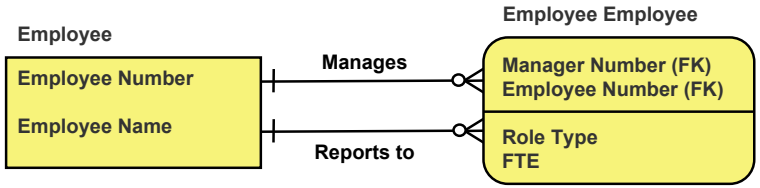
| PRODUCT KEY | **contains** COMPONENT KEY | at PART LEVEL | in SEQUENCE NUMBER | SUB ASSEMBLY | QUANTITY |
|---|---|---|---|---|---|
| SK, PK | SK, PK | N | N | [Y/N] | N |
| POMCar | AM DB5 | 1 | 10 | N | 1 |
| POMCar | Safety Pack | 1 | 20 | Y | 1 |
| POMCar | FRONT Defence System | 2 | 30 | Y | 1 |
| POMCar | Motion Sensor | 3 | 40 | N | 2 |
| POMCar | Ejector Seat | 2 | 50 | N | 1 |
| POMCar | REAR Defence System | 2 | 60 | Y | 1 |
| POMCar | Motion Sensor | 3 | 70 | N | 4 |



# Exercise 11: HR Hierarchy Map

**Requirements**

BI stakeholders need to view sales and consulting activity rolled up to any level in the Pomegranate HR hierarchy. To do this your design must account for both permanent line manager relationships and temporary project management relationships. When an employee reports to more than one manager their activity measures must be prorated using the Full-Time Equivalency (FTE) of each role they perform. The necessary information to enable this is recorded in HR source system using the following normalised many-to-many recursive relationship:




Design a reporting structure hierarchy map that will allow employee facts to be rolled up to any management level in the HR hierarchy. Make sure that your solution allows BI Users to do the following:

- Easily filter on employees who:
  - Do not manage other employees
  - Do not report to other employees
  - Performing temporary assignments
- Apply an FTE weighting factor to measures when they are rolling up using the employee reporting structure
- Display the management structure in the correct order on reports

Begin by annotating your HR organisation chart from exercise 5 with any additional technical information which will be needed such as explicit weighting factors, level numbers and display sequence numbers.

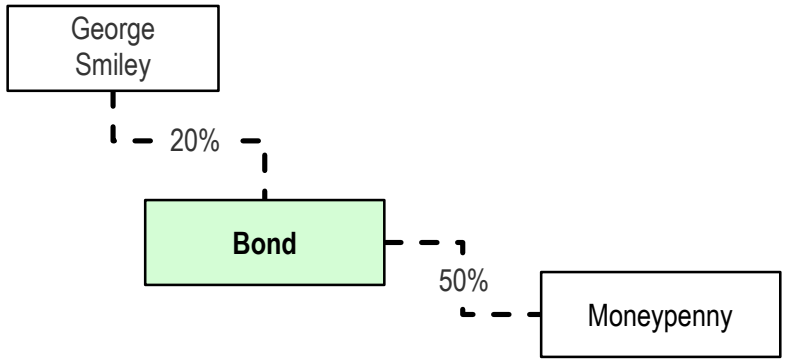
Document your design using a BEAM Table and populate it with test data based on your annotated HR org chart.

Pomegranate
Organization
Chart

**REPORTING STRUCTURE [CV, MV, HM]**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |

# Hierarchy Change




**REPORTING STRUCTURE [CV, MV, HM]**

| MANAGER KEY | **manages** EMPLOYEE KEY | at EMPLOYEE LEVEL | in SEQUENCE NUMBER | LOWEST EMPLOYEE | HIGHEST MANAGER | with ROLE TYPE | with WEIGHTING FACTOR |
|---|---|---|---|---|---|---|---|
| SK, PK, CV | SK, PK, HV | N, CV | N, CV | [Y/N], CV | [Y/N], CV | CV | N, CV |
| Eve Tasks | Bond | 3 | 300 | **N** | Y | Temporary | 0.2 |
| ***Eve Tasks*** | ***Moneypenny*** | ***4*** | ***310*** | ***Y*** | ***Y*** | ***Temporary*** | ***0.1*** (0.2 x 0.5) |
| Eve Tasks | Bond | 4 | 600 | **N** | Y | Permanent | 0.8 |
| ***Eve Tasks*** | ***Moneypenny*** | ***5*** | ***610*** | ***Y*** | ***Y*** | ***Temporary*** | ***0.4*** (0.8 x 0.5) |
| Eve Tasks | Moneypenny | 4 | **700** | Y | Y | Permanent | **0.5** |
| George Smiley | Bond | 2 | 200 | **N** | N | Temporary | 0.2 |
| ***George Smiley*** | ***Moneypenny*** | ***3*** | ***210*** | ***Y*** | ***N*** | ***Temporary*** | ***0.1*** (0.2 x 0.5) |
| Gerald Mole | Bond | 3 | 300 | **N** | N | Permanent | 0.8 |
| ***Gerald Mole*** | ***Moneypenny*** | ***4*** | ***310*** | ***Y*** | ***N*** | ***Temporary*** | ***0.4*** (0.8 x 0.5) |
| Gerald Mole | Moneypenny | 3 | 400 | Y | N | Permanent | **0.5** |
| M | Bond | 2 | 200 | **N** | N | Permanent | 0.8 |
| ***M*** | ***Moneypenny*** | ***3*** | ***210*** | ***Y*** | ***N*** | ***Temporary*** | ***0.4*** (0.8 x 0.5) |
| M | Moneypenny | 2 | **300** | Y | N | Permanent | **0.5** |
| Bond | Bond | 1 | 100 | **N** | N | Permanent | 1 |
| ***Bond*** | ***Moneypenny*** | ***2*** | ***200*** | ***Y*** | ***N*** | ***Temporary*** | ***0.5*** |

6 new Moneypenny records for her new "Temporary" relationship

All Bond records are updated because he is now a manager

All existing Moneypenny records are updated with a weighting factor of 0.5

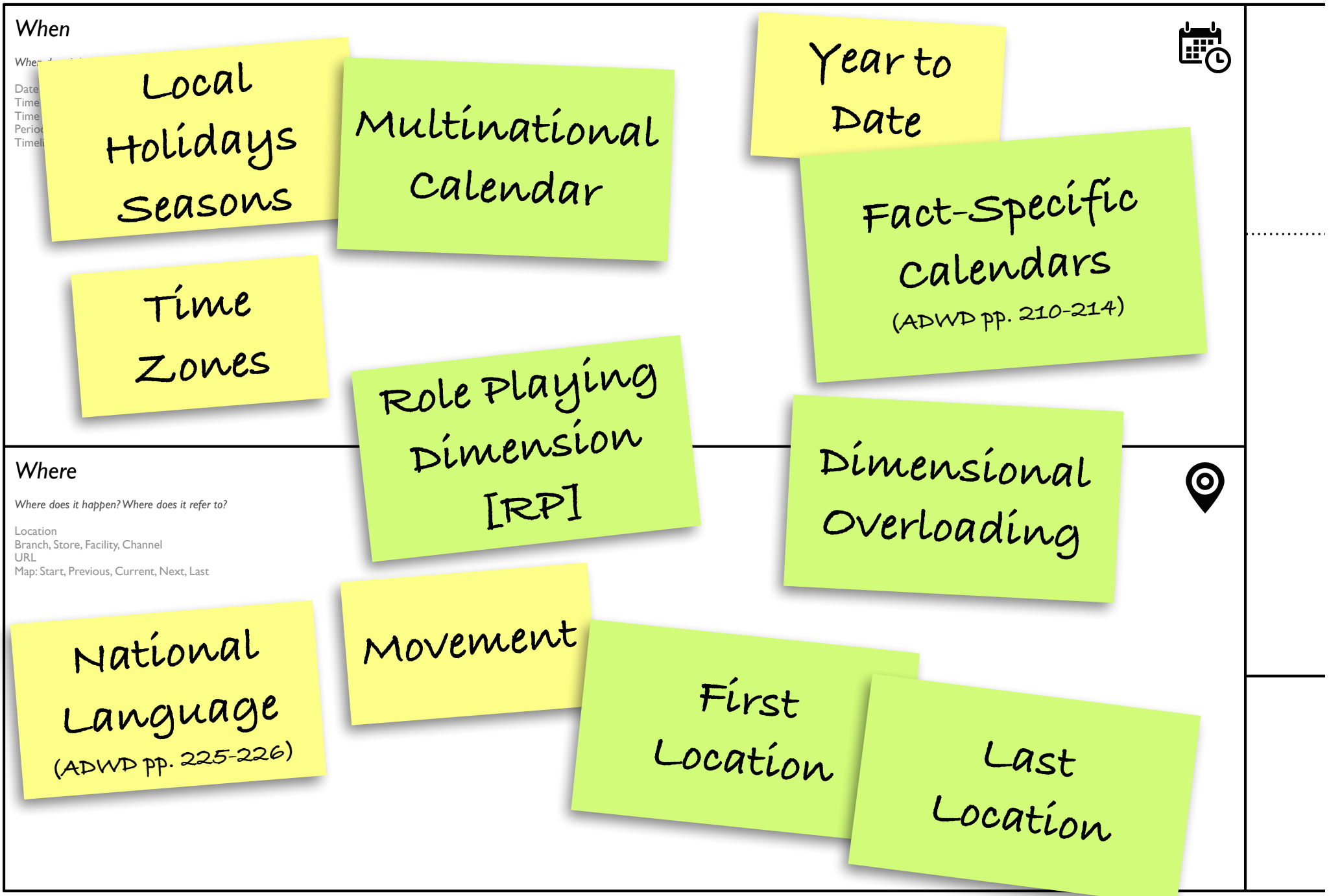
# Populating a Hierarchy Map

- Stored Procedures
  - Recursive SQL (e.g. Connect By, With Recursive ) or recursive procedure calls
- ETL Tool Built-in mapping

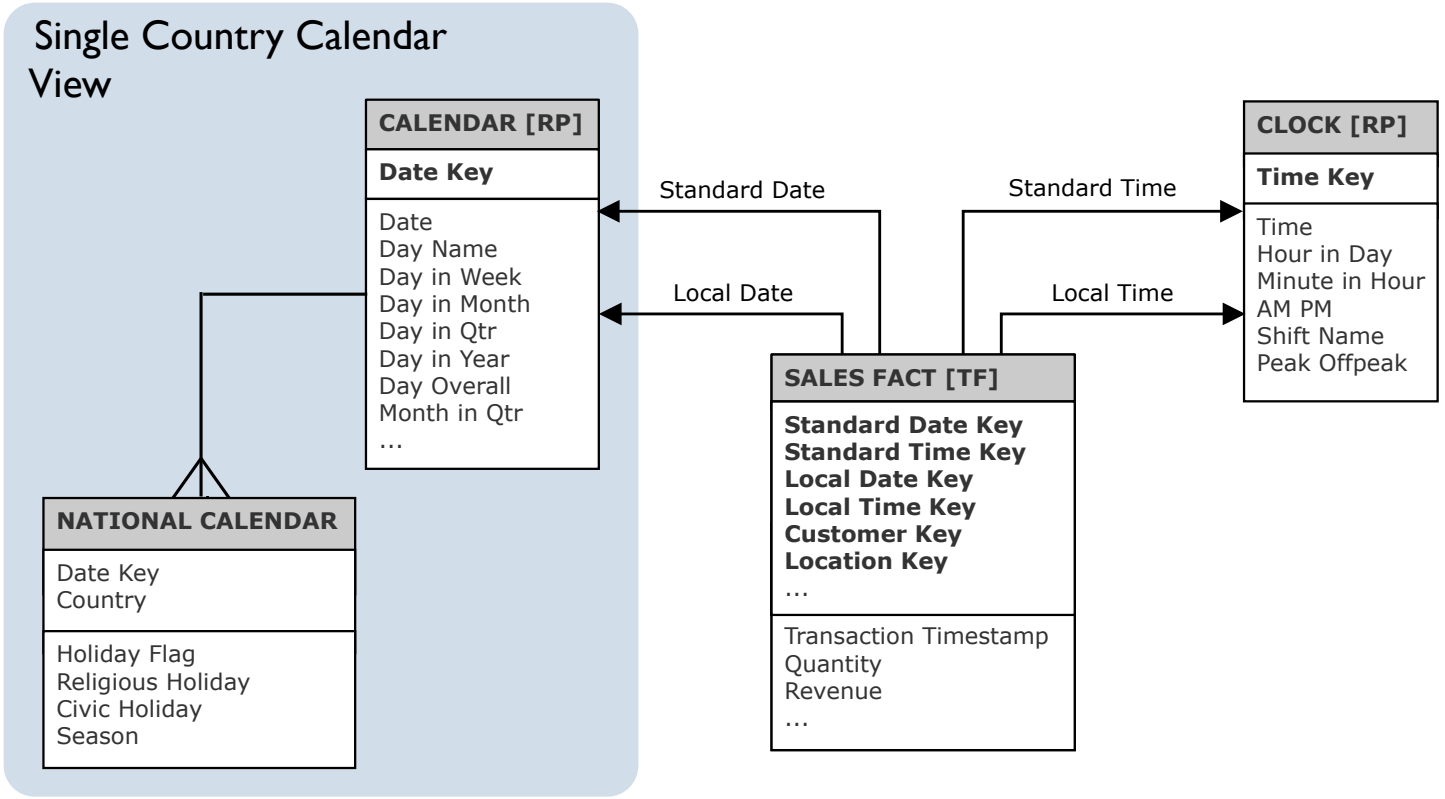
- Lawrence Corr, *Kimball University Design tip #17: Populating Hierarchy Helper Tables*
- Joe Caserta, Ralph Kimball, *The Data Warehouse ETL Toolkit, pp. 201-204*
- Updating a Hierarchy Map, Maintaining Hierarchy History: see *ADWD pp. 192-195*

- www.decisionone.co.uk/resources/Downloads.htm

# *When* and *Where* Patterns (Chapter 7)



# Time Zones and International Date Descriptions




**CALENDAR [RP]**

| DATE KEY | CALENDAR DATE | is in MONTH | is in SEASON | is a HOLIDAY | |
|---|---|---|---|---|---|
| SK | D | | C | [Y/N] | |
| **20250317000** | 17 March 2025 | March | Spring | Y | |
| **20250317001** | 17 March 2025 | March | Fall | N | |

# Date Key

- **Surrogate key** like all other dimension keys... **but different:**
  - Date (key) sequence is important for partitioning fact tables
- Generation/assignment techniques:
  - ISO Format number: YYYYMMDD
  - Epoch-based number: day number from 1/1/1900
- Geopolitical would needs a 3 digit trailing version number to handle every country in the world differently
- Which date key format is best?
  - ETL friendly, BI friendly (performance)

# Clock Dimension

**CLOCK [RP, HV]**

| TIME KEY | TIME | HOUR | MINUTE | AM PM | MINUTE IN DAY | DAY NIGHT | WORK SHIFT | PEAK OFFPEAK |
|---|---|---|---|---|---|---|---|---|
| SK | C, FV | N, FV | N, FV | C, FV | N, FV | C, FV | C, HV | C, HV |
| 1 | 00:00 | 0 | 0 | AM | 1 | Night | Graveyard | Off Peak |
| 2 | 00:01 | 0 | 1 | AM | 2 | Night | Graveyard | Off Peak |
| 3 | 00:02 | 0 | 2 | AM | 3 | Night | Graveyard | Off Peak |
| … | … | … | … | … | … | … | … | … |
| 1051 | 17:30 | 17 | 30 | PM | 1051 | Day | Drivetime | Peak |
| 1440 | 23:59 | 23 | 59 | PM | 1440 | Night | Evening | Off Peak |
| 0 | - | - | - | - | - | Unknown | Unknown | ? |

# Day/Time Dimension Pattern

**Problem:**

- Time of day descriptions vary depending upon other calendar attributes

**Solution:**

- Combine date and time in a single dimension: **too big**
- Build a time of day (Day Clock) dimension which contains multiple versions of each minute for each significant day type

**Consequences:**

- Must used a surrogate key that is not based on HHMM
- ETL processes must know which version of each minute to use when processing facts

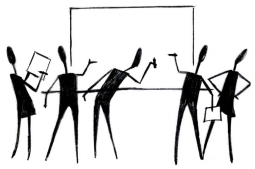
# Day Clock Dimension

**DAY CLOCK [RP, HV]**

| TIME KEY | TIME | HOUR | MINUTE | **DAY** | **DAY TYPE** | WORK SHIFT | PEAK OFFPEAK |
|---|---|---|---|---|---|---|---|
| SK | C, FV | N, FV | N, FV | **C, FV** | **C, FV** | C, HV | C, HV |
| 1 | 00:00 | 0 | 0 | **Monday** | **Weekday** | Graveyard | Off Peak |
| 481 | 08:00 | 8 | 0 | **Monday** | **Weekday** | Drivetime | **Peak** |
| 841 | 14:00 | 14 | 0 | **Monday** | **Weekday** | Normal | **Peak** |
| 1051 | 17:30 | 17 | 30 | **Monday** | **Weekday** | Drivetime | **Peak** |
| 1052 | 17:31 | 17 | 31 | **Monday** | **Weekday** | Evening | Off Peak |
| 1440 | 23:59 | 23 | 59 | **Monday** | **Weekday** | Evening | Off Peak |
| … | … | … | … | **…** | **…** | … | … |
| 7201 | 00:00 | 0 | 0 | **Saturday** | **Weekend** | Graveyard | Off Peak |
| 7681 | 08:00 | 8 | 0 | **Saturday** | **Weekend** | Weekend | **Off Peak** |
| 8251 | 17:30 | 17 | 30 | **Saturday** | **Weekend** | Weekend | **Off Peak** |
| 10981 | 15:00 | 15 | 0 | **Monday** | **Holiday** | Special | **Off Peak** |
| 0 | - | - | - | **Missing** | **Missing** | Unknown | ? |

Peak run 8:00am to 5:30pm weekdays

Weekends and holidays are off peak and have different work shift names



# Exercise 12: Air Travel

**Requirements**

To enable BI carbon footprint reporting and control travel costs, Pomegranate must analysis the national and international flights taken by their global sales and consulting force. With the help of example flight information design a BI Model capable of easily answering the following questions:

- Which employees travel the most frequently and furthest? Who authorises their travel?
- When (at what time of year/what time of day) do employee travel most? How often do they travel for 12+ hours? How long do they spend traveling?
- Which airlines are used most often? Which have the highest/lowest CO2 figures on the routes we use?
- Where do employees need to travel to on business and why?

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Milestones: Fixed, Variable, Repeatable/Recurring

## How

*How (exactly) does it happen?*
*How do we know it happened?*
*How do we uniquely identify each event?*

***Verb***, Activity, Process, Event
Effect, Outcome, Status
Transaction Type
Transaction #, Event ID [Degenerate Dimension]
Step/Sequence #

[Granularity, Event Type: Discrete, Evolving, or Recurring]

## Who

***Who does what?*** *How do we organize them?*
*How do they change? Who else is involved?*

*Subject/Object*
Customer: Business, Consumer, Segment
Employee
Supplier
Partner
Third Party

## How Many

*How many/much is involved? How long does it take?*

Quantities
Revenues
Costs
Discounts/Deltas
Balances
Activity/Status Counts
Durations
Measure/KPI Formulas

[UoM, Fact Type: Fully Additive, Semi-Additive, Non-Additive]

## Where

*Where does it happen? Where does it refer to?*

*Subject/Object*
Location
Branch, Store, Facility
Channel
URL
Map/Sequence: First → Previous → Current → Next → Last

## What

*What is involved/used? How are they organized?*
*How do they change?*

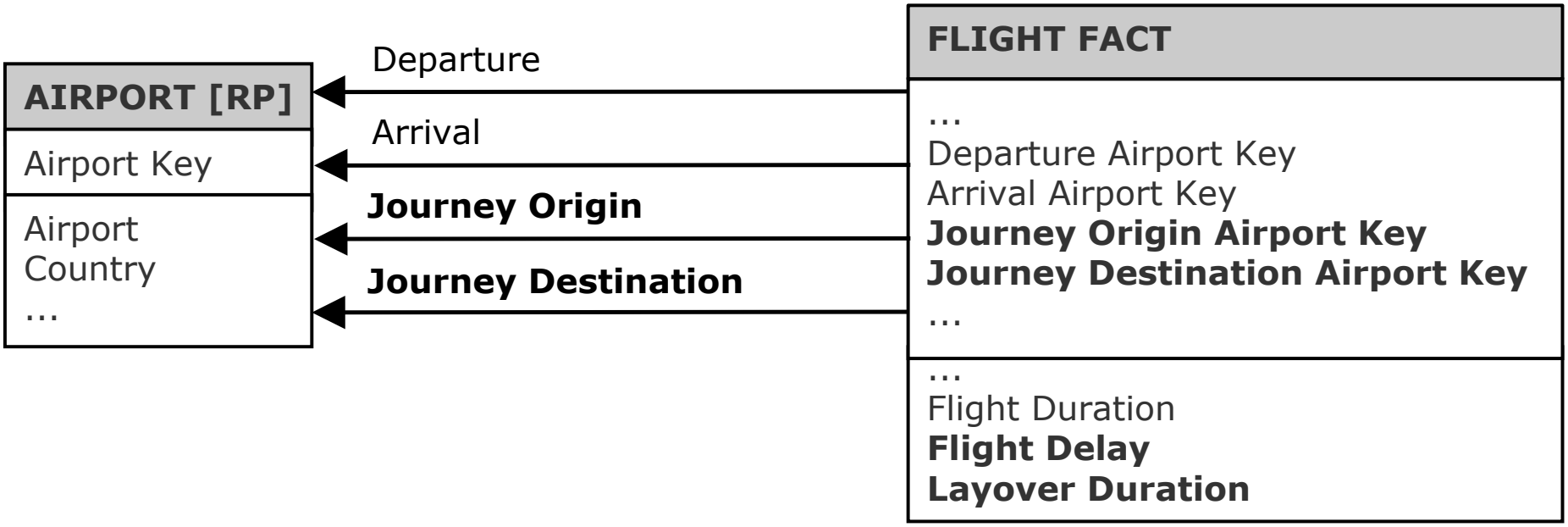
*Subject/Object*
Value Proposition
Product
Service
Resource
Item

## Why

*Why does it happen? Why do quantities vary?*

Cause, Reason
Trigger Event ID
Promotion
Quantity Descriptions

# Dimensional Overloading/Enrichment



# Role Playing [RP] Dimensions

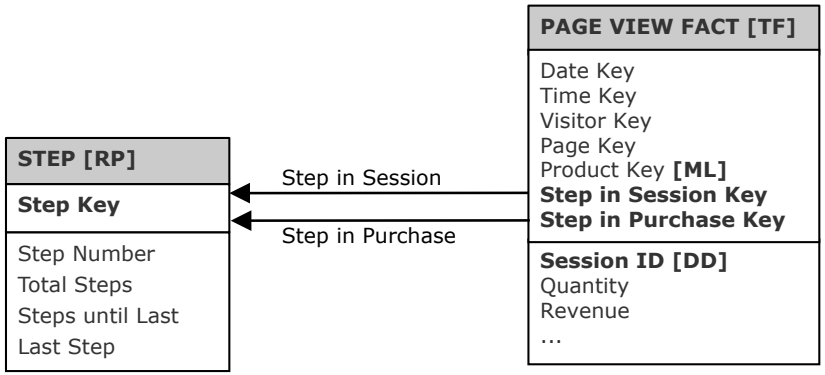
- A dimension can play multiple roles e.g., order date, delivery date, ship-from, ship-to
- Most common role playing dimension types are *When, Where, Who*
- Difficult roles:
  - Parent Company, Subsidiary Company
- Physically store dimension table once
- Create views to rename table and columns for each role that the dimension plays

# ***Why*** things start. ***How*** they finish

- First Purchase Date, Policy Month
- Start and End points
- Step dimension
- Dimension Overloading
- Current, Prior, Next, First, Last

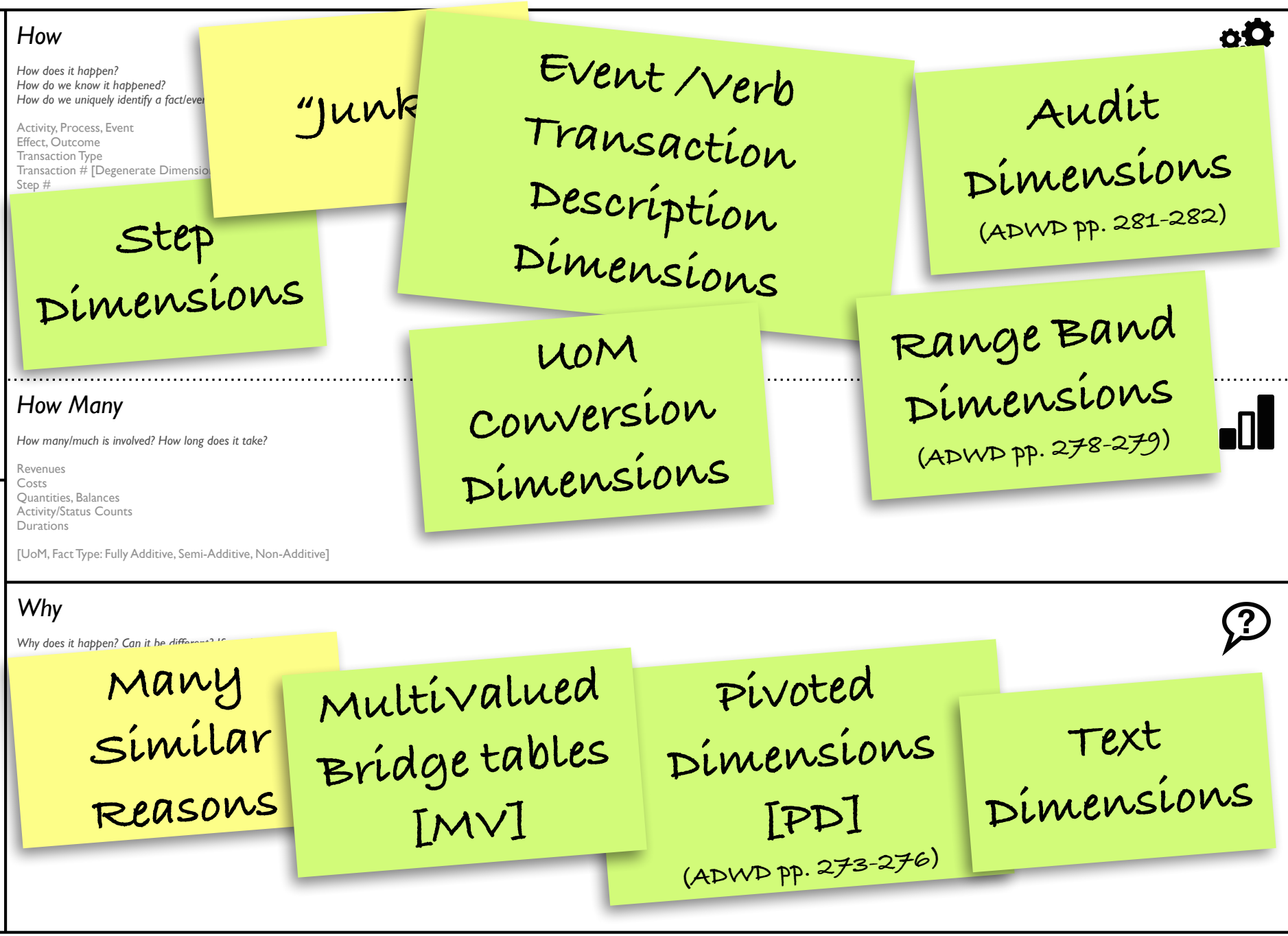
**STEP**

| STEP KEY | STEP NUMBER | TOTAL STEPS | STEPS UNTIL LAST | LAST STEP |
|---|---|---|---|---|
| SK | N | N | N | [Y/N] |
| 0 | 0 | 0 | 0 | - |
| 1 | 1 | 1 | 0 | Y |
| 2 | 1 | 2 | 1 | N |
| 3 | 2 | 2 | 0 | Y |
| 4 | 1 | 3 | 2 | N |
| 5 | 2 | 3 | 1 | N |
| 6 | 3 | 3 | 0 | Y |
| 7 | 1 | 4 | 3 | N |
| 8 | 2 | 4 | 2 | N |
| 9 | 3 | 4 | 1 | N |
| 10 | 4 | 4 | 0 | Y |




- Making sense of micro transactions – add descriptive behaviour tags at ETL time using read ahead procedural logic

# *Why* and *How* Patterns (Chapter 9)



# Multiple Similar Reasons – Multi-Valued *Why* Dimensions

**MV**: multi-valued *why* detail.
Group example data shows multiple diagnosis for a single treatment.

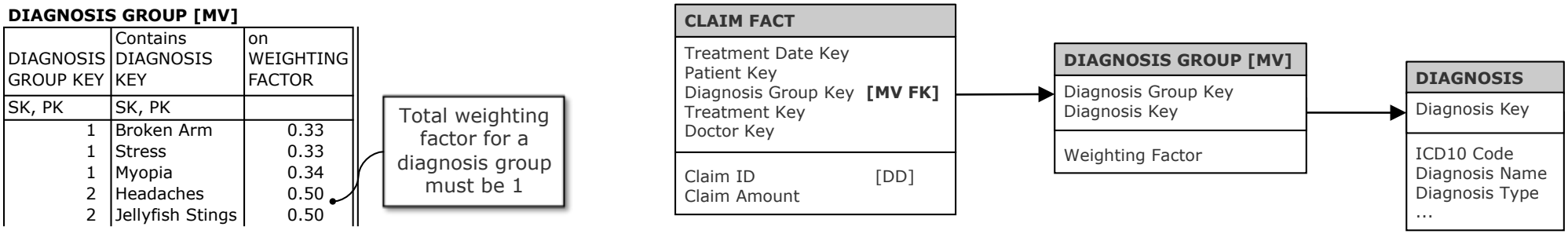
**MEDICAL TREATMENTS [DE]**

| DOCTOR | **claims** CLAIM AMOUNT | on TREATMENT DATE | for TREATMENT | to PATIENT | with DIAGNOSIS | with CLAIM ID |
|---|---|---|---|---|---|---|
| [who] | [$] FA | [when] | [what] GD | [Employee] | [why] **MV** | [how] GD |
| Goldfinger | $500 | 9-Mar-2011 | Plaster Cast | Bond | Broken Arm, Stress, Myopia | G-2011-4 |
| Goldfinger | $5,000 | 9-Mar-2011 | Laser Correction | Bond | Broken Arm, Stress, Myopia | G-2011-4 |
| Dr. No | $100 | 2-Jun-2011 | Pain Killers | Bond | Headaches, Jellyfish Stings | N-2011-7 |

# Multi-Valued Dimensions [MV]

- Ideal each dimension has a single value for each instance of a fact
- When a required dimension has multiple values we may have the fact granularity wrong/ not detailed enough e.g.
  - An 'Invoice' level granularity fact table would cause Product or Service to be multi-valued
- Solution – design atomic level detailed fact tables e.g. 'Invoice line item' granularity.
- What happens when we are already at the atomic level of fact measurement?

# Multi-Valued Bridge Table Pattern

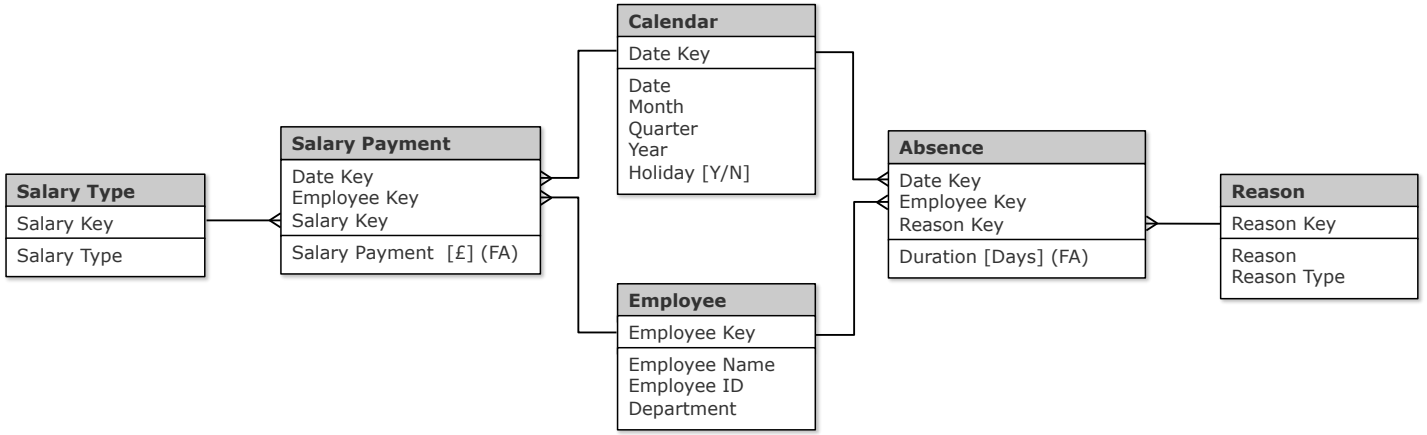



- Bridge table or link entities resolves many-to-many relationship between the fact table and the multi-valued dimension
- Embellish with weighting factors for fact allocation. Sum of weighting factors for any one fact row is 1.
- Avoids hard coding allocations in the fact table
- Optional Bridge Table Pattern for *barely* multi-valued (ADWD pp. 270-272)
- Pivoted Dimension Pattern for low, fixed cardinality multi-values (ADWD pp. 273-276)

# Multi-Valued Dimension Reporting

- Queries that do not use the multi-valued attributes are unaffected
- Constraining on a single value of the multi-valued dimension causes no problem
- Grouping by a multi-valued dimension attribute can be handled in two ways:
  - **Impact Reports** – ignore weighting factors and produce overlapping non-additive measures i.e. no subtotal or grand total rows allowed on these reports
  - **'Correctly Weighted' Reports** – multiply facts by weighting factor. The 'correctness' of individual rows on the report will depend on quality of the allocations but the grand total will be 'true'

# ***How*** do you Query multiple fact tables?




Report 1:

2011 Salary Payments
James Bond 007

| SALARY TYPE | SALARY |
|---|---|
| Basic | £50,000 |
| Danger Money | £100,000 |
| London Weighting | £10,000 |

Report 2:

2011 Absences
James Bond 007

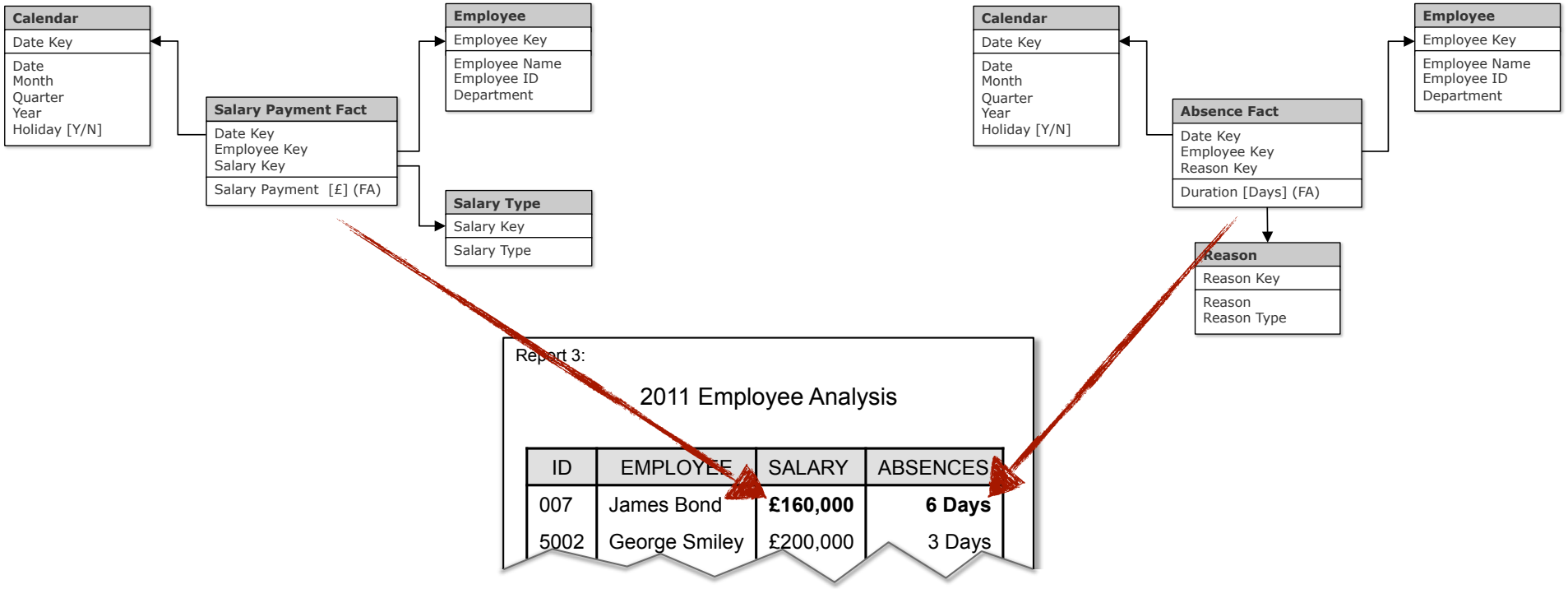
| DATE | REASON | DURATION |
|---|---|---|
| Feb 12 2011 | Injured | 4 Days |
| Mar 12 2011 | Vacation | 2 Days |

Report 3:

2011 Employee Analysis

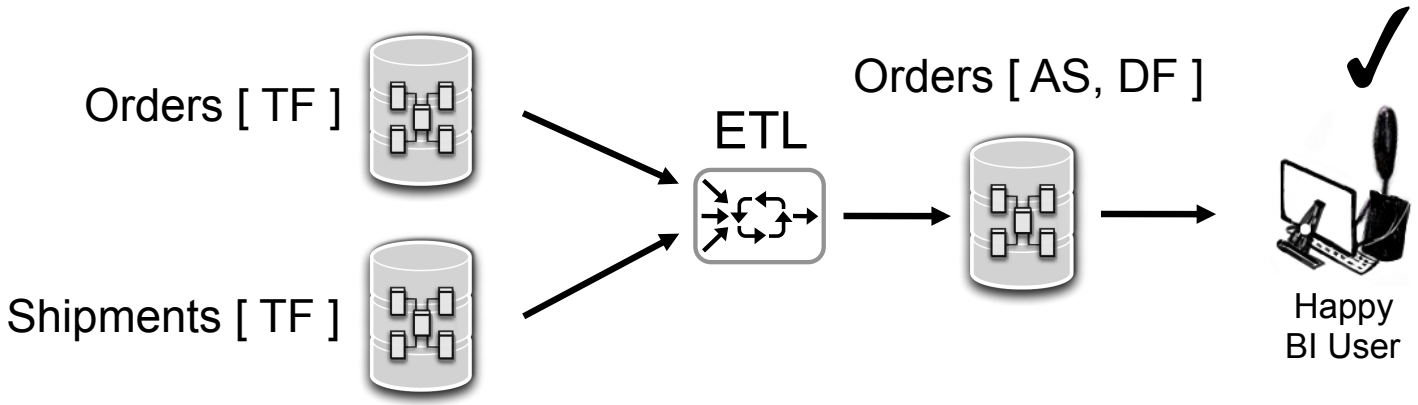
| ID | EMPLOYEE | SALARY | ABSENCES |
|---|---|---|---|
| *007* | James Bond | | |
| 5002 | George Smiley | £200,000 | 3 Days |

# Drill Across Queries: Multi-Pass SQL



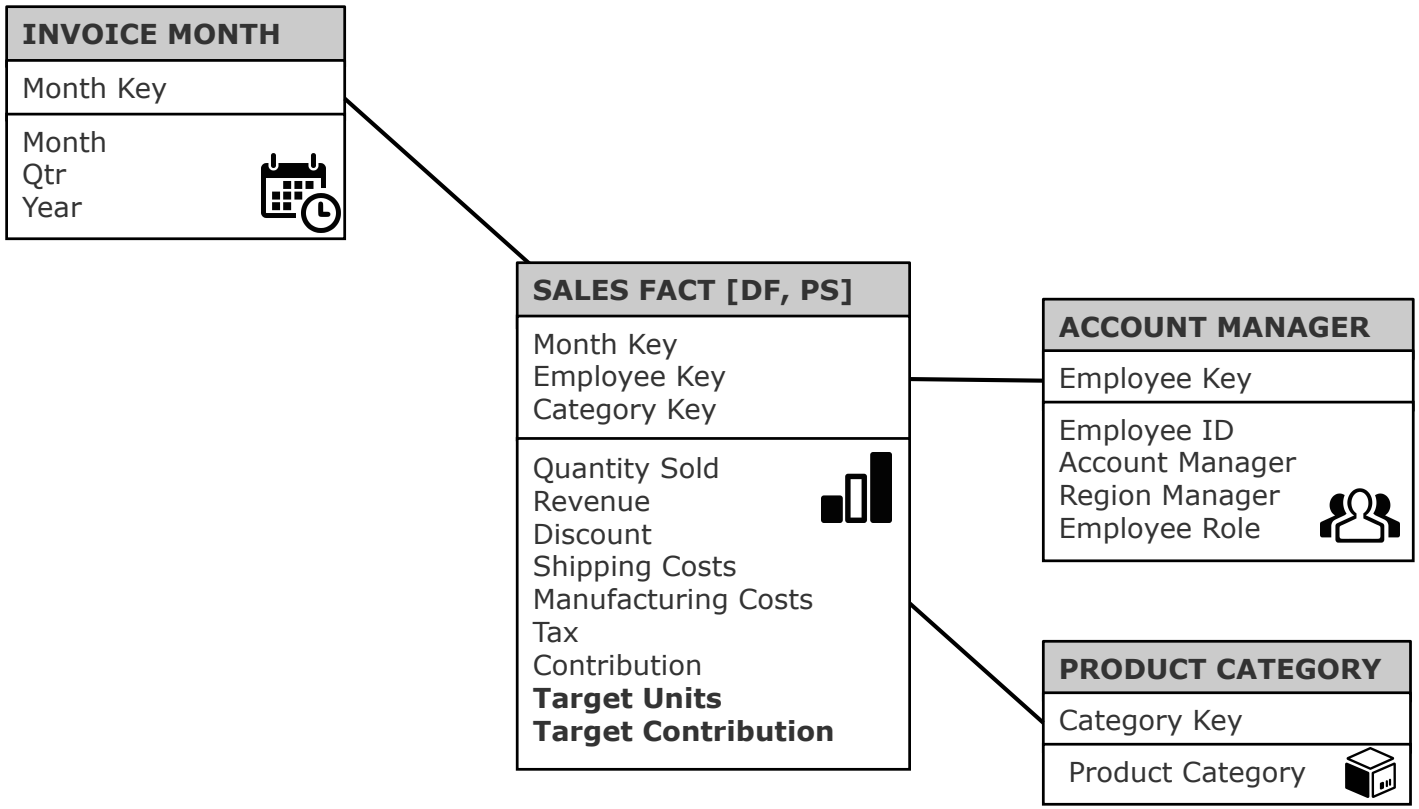

- Fact tables should be queried one at a time otherwise the 'many to one to many' or 'chasm' trap
- Multi-pass SQL is a key BI technique for building distributed systems that work:
  - Separate data sources can be in different locations/technology
  - Performance remains under control: separate queries allow easier optimization, aggregate navigation
  - Multi-pass SQL is needed to calculate totals and subtotals for non-additive measures in report break rows e.g. average for all products and to display measure columns with varying constraints

# Derived Fact Tables [DF]: Consolidated Data Marts




- Create evolving event/multiple process accumulating snapshot **derived fact tables [DF]** from several single process transactional fact tables
- Replace detailed drill-across queries – complex/slow multi-pass SQL with single star queries
- Solve difficult analysis problems using ETL rather than BI
- 360° Customer View: all customer *touch points*
- Product/Service Profitability Data Mart

# Consolidated Data Mart for Exercise 8



*Review*

*Bookshelf*

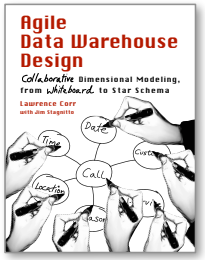
*Conclusions*

# BEAM✻ Data Stores, 7*Ws* - Linguistics

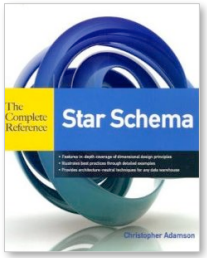
- Events – Verbs – Relationships – Fact Tables
- Details – Nouns – Entities – Dimensions
- Adjectives and Adverbs – Attributes
- Interrogatives – The *7Ws* – Dimension Types and Facts
- Main Clause – Subject-Verb-Object
- Prepositions – connect details to the main clause
- Example Story Themes – *Typical, Different, Repeat, Missing, Group*
- Story Types – Discrete, Evolving, Recurring
- Granularity Details (GD) – define event uniqueness

# BEAM* Diagrams

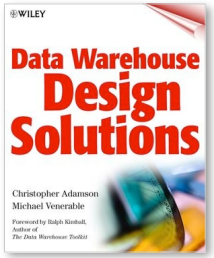
# Bookshelf: Dimensional Modeling



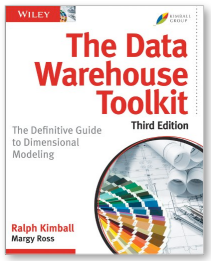

- **Agile Data Warehouse Design**
  Lawrence Corr, Jim Stagnitto, DecisionOne Press, 2011




- **Star Schema: The Complete Reference**
  Christopher Adamson, McGraw Hill, 2010

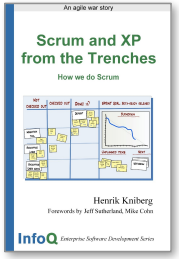



- **Data Warehouse Design Solutions**
  Chris Adamson, Michael Venerable, Wiley, 1998

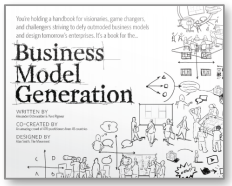



- **The Data Warehouse Toolkit, 3rd Edition**
  Ralph Kimball, Margy Ross, Wiley, 2013

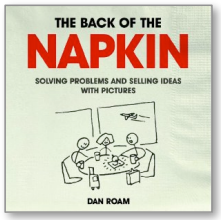
# Bookshelf: Agile, Design/Visual Thinking, Collaboration



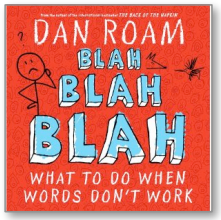

- **Scrum and XP from the Trenches**
  Henrick Kniberg, C4Media, 2007 www.infoq.com



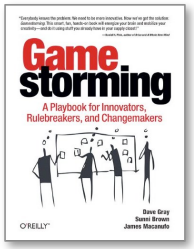

- **Business Model Generation**
  Alexander Osterwalder, Yves Pigneur (Wiley, 2010)




- **The Back of the Napkin: Solving Problems and Selling Ideas With Pictures**
  Dan Roam (Portfolio, 2008)




- **Blah Blah Blah: What To Do When Words Don't Work**
  Dan Roam (Portfolio, 2011)




- **Gamestorming**
  Dave Gray, Sunni Brown, James Macanufo (O'Reilly Media, 2010)

# Online

- **modelstorming.com** – Agile data warehouse design resources include BEAM✲ templates, reference card and links to the books and articles mentioned in the course
- **LinkedIn** – connect with me
- **kimballgroup.com** – Ralph Kimball articles and design tips - for now
- **blog.oaktonsoftware.com** – Star Schema Central: Chris Adamson's dimensional modeling blog

# Modeling with Post-it Notes

TED

Tom Wujec

Draw Toast

# Modeling with Post-it and Digital Notes

# Two Immutable Laws of Data

- You can never truly evaluate how well a design matches the available data until you **LOAD** it

- You never really know how well a design reflecting the stated requirements will match the actual requirements until users **ACCESS** it

— *Greg Jones*

# Exercise 2 & 7 Example Answer

Event/Title | Subject Verb Object | Author(s), Date/Version

## When

*When does it happen?*
*What other related dates/times are know/fixed at this time?*

Date
Time
Time Zone
Period
Timeline: Event Milestones: Fix

Purchase Date
Invoice date/time (3,650)

7/2/2021 (Yesterday)

## Ho

*How (exactly) does it happen?*
*How do we know it happened?*

Customer buys Product

Invoice #999

Invoice
Invoice # (100M)
DD NA

## How Many

*How many/much is involved? How long*

Quantities
Revenues
Costs
Discounts/Deltas
Balance
Activit
Durati
Measu

[UoM]

2 iPips

Quantity
[items] FA

Revenue
[$] FA

$250

## Wh

**Who d**
How do

Subject/Ob
Customer:
Employee
Supplier
Partner
Third Party

Elvis Pr

Customer
CV
Customer ID (1M)

Salesperson
[Employee] HV
Employee ID (5,000)

James Bond
mployee 0 0 7

## Where

*Where does it happen? Where does it refer to?*

Subject/Object
Location
Branch, Store, Facility
Channel
URL
Ma

Store/URL
HV
Store No/URL (5,000)

Melbourne

## What

*What is involved/used? How are they organized?*
*How d*

Subject
Value F
Produ
Servic
Resou
Item

Product
HV
Product Code (50,000)

Blue Suede iPip

## Promotion

Promotion
campaign #
(10,000)

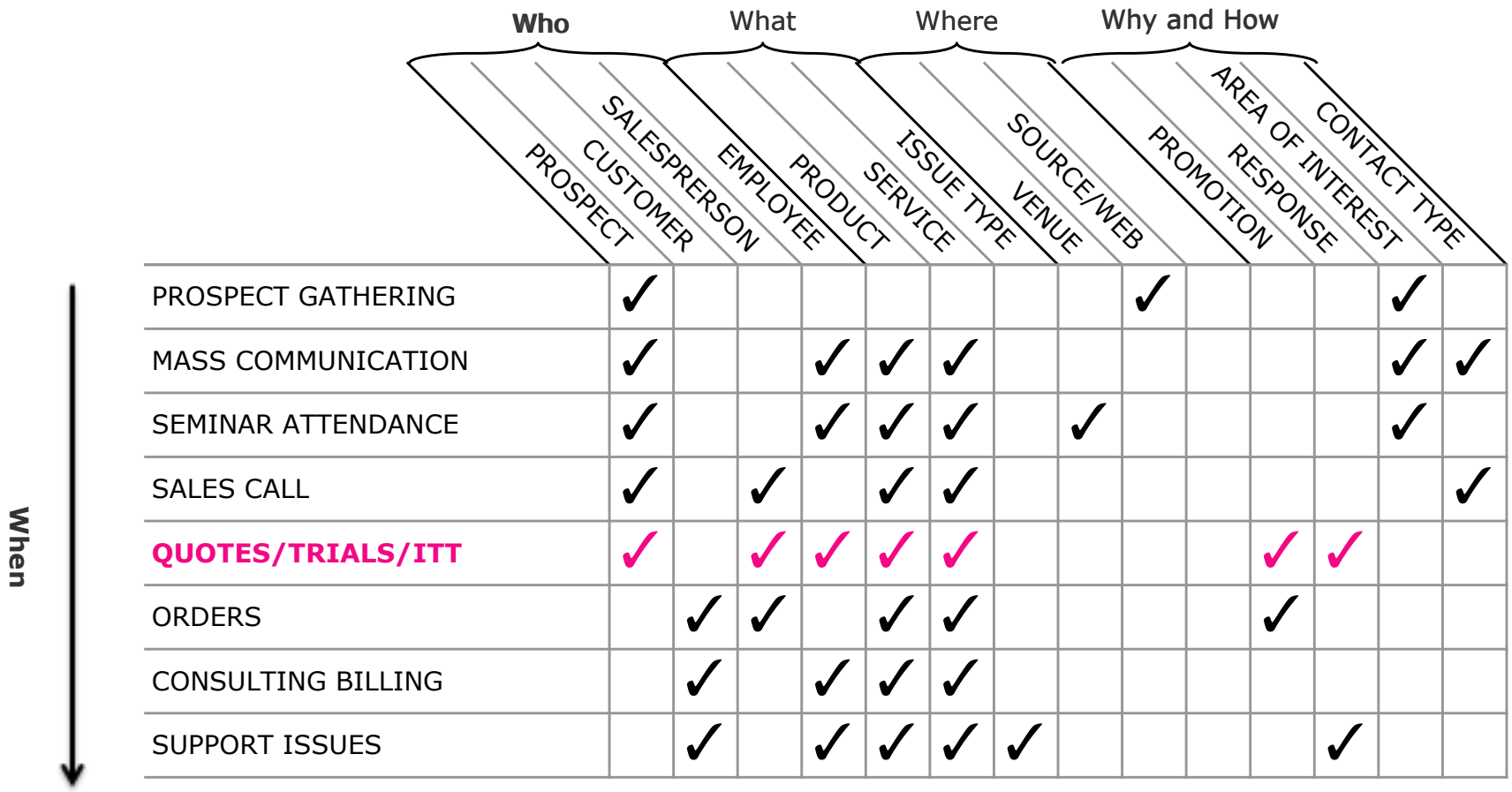
Two for One 50% off

BEAM* Canvas

modelstorming.com · Agile Data Warehouse Design

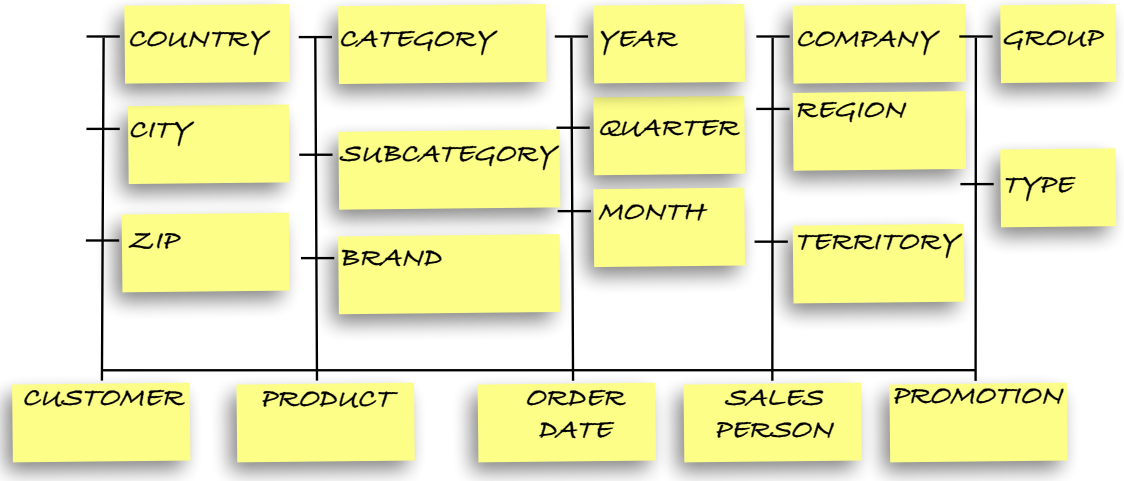
*Business Event Analysis and Modeling

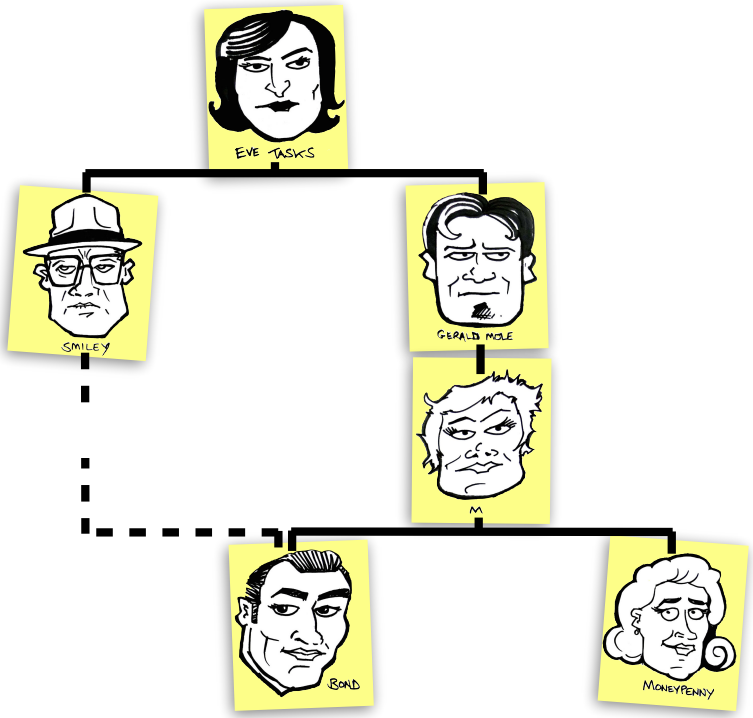
# Exercise 4 Example Event Matrix



| When | Who | | | What | | Where | | Why and How | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PROSPECT | CUSTOMER | SALESPRERSON | EMPLOYEE | PRODUCT | SERVICE | ISSUE TYPE | VENUE | SOURCE/WEB | PROMOTION | RESPONSE | AREA OF INTEREST | CONTACT TYPE |
| PROSPECT GATHERING | ✓ | | | | | | | | ✓ | | | | ✓ | |
| MASS COMMUNICATION | ✓ | | | ✓ | ✓ | ✓ | | | | | | | ✓ | ✓ |
| SEMINAR ATTENDANCE | ✓ | | | ✓ | ✓ | ✓ | | ✓ | | | | | ✓ | |
| SALES CALL | ✓ | | ✓ | | ✓ | ✓ | | | | | | | | ✓ |
| **QUOTES/TRIALS/ITT** | ✓ | | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | | |
| ORDERS | | ✓ | ✓ | | ✓ | ✓ | | | | | ✓ | | | |
| CONSULTING BILLING | | ✓ | | ✓ | ✓ | ✓ | | | | | | | | |
| SUPPORT ISSUES | | ✓ | | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | | |

# Exercise 5 Example Hierarchies



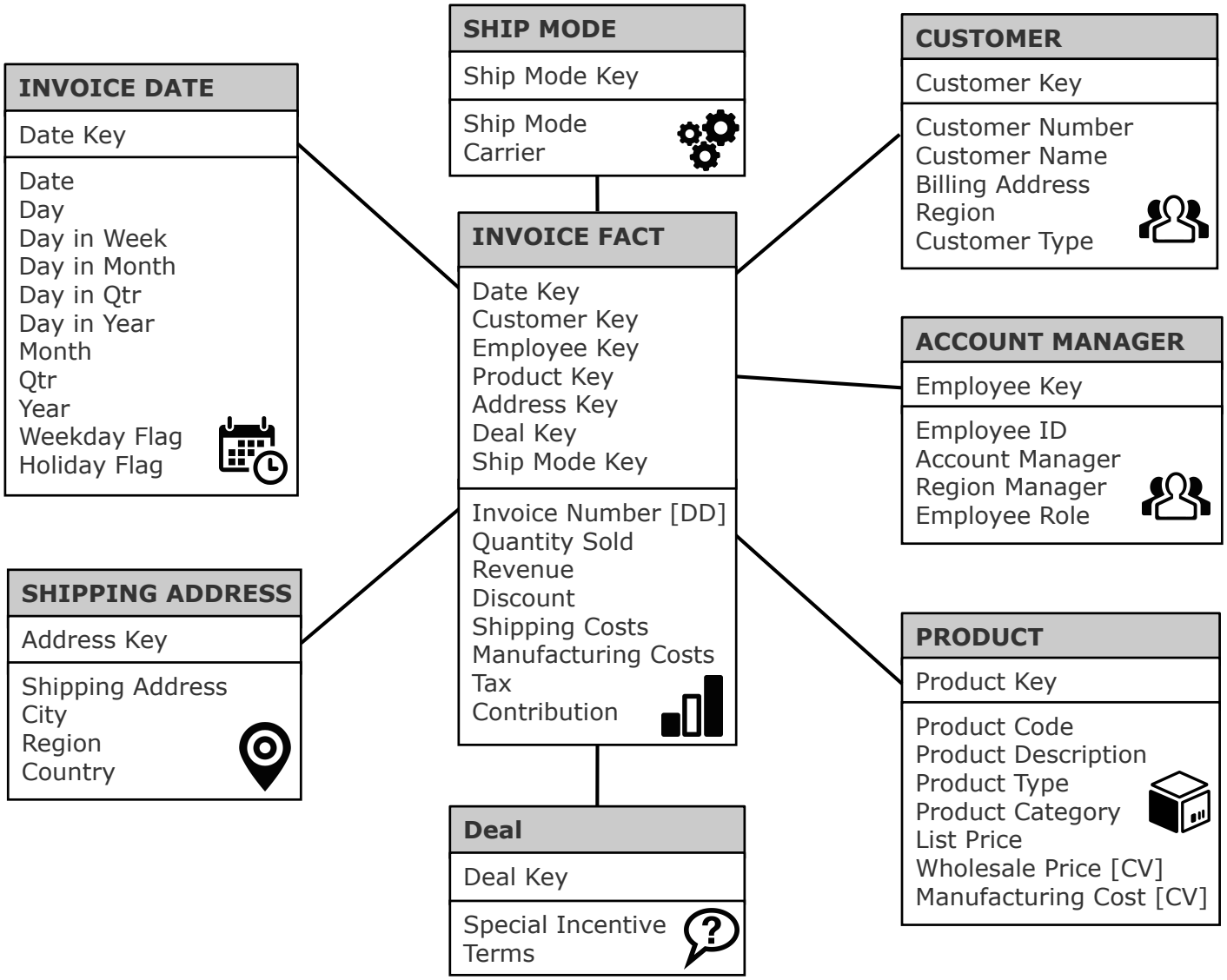
# Exercise 6 Example Answer

| Record # | Customer ID | Title | First Name | Last Name | Street | City | Country | ETL Processing | Rows Inserted | Rows Updated |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | C1239 | Sir | Hilary | Bray | Royal College of Arms | London | UK | CV - Type 1 Change | 0 | 1 |
| 2 | C2345 | Mrs | Tracy | Bond | Rue Du Rhone | Geneva | Switzerland | HV - Type 2 Change | 1 | 1 |
| 3 | C7810 | Miss | Vespa | Lynd | 10 Kings Road | London | UK | New – Insert | 1 | 0 |
| 4 | C8904 | Mr | Felix | Leiter | The Office | Langley, VA | USA | CV- Multi-row | 0 | 2 |
| 5 | C3467 | Mr | Marc-Ange | Draco | Rua Augusta | Lisbon | Portugal | Smart Correction | 0 | 1 |
| **6** | **C1038** | **Mr** | **Ernst Stavro** | **Blofeld** | **The Summit** | **Piz Gloria** | **Switzerland** | Type 1 and Type 2 | 1 | 2 |

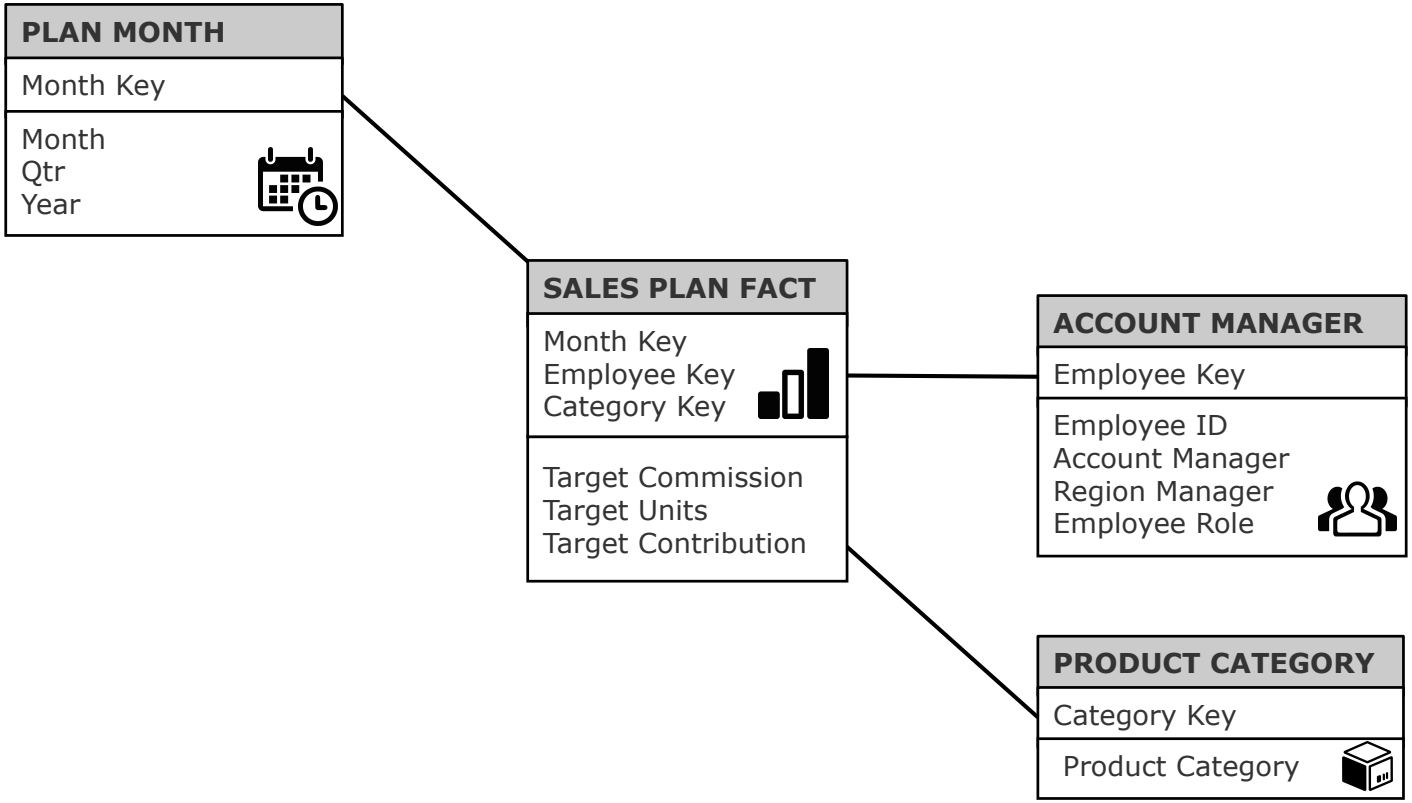
**CUSTOMER**

| **Customer Key** | **Customer ID** | **Title** | **First Name** | **Last Name** | **Street** | **City** | **Country** | **Effective Date** | **End Date** | **Last Updated** | **Current** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SK** | **BK** | **CV, HV1** | **CV, HV1** | **CV, HV1** | **CV, HV2, HV3** | **HV, HV2** | **HV, HV3** | | | | **[Y/N]** |
| 1001 | C8904 | Mr | Felix | **Leiter** | Harry's Bar | Key West, FL | USA | 5/10/1962 | 16/9/1964 | **Today** | N |
| 1021 | C1010 | Dr. | Julius | No | The Island | Crab Key | Jamaica | 5/10/1962 | 1/1/3000 | 5/10/1962 | Y |
| 1041 | C1241 | Mr | Auric | Goldfinger | Fontainebleau Hotel | Miami, FL | USA | 17/9/1964 | 1/1/3000 | 17/9/1964 | Y |
| 1042 | C8904 | Mr | Felix | **Leiter** | The Office | Langley, VA | USA | 17/9/1964 | 1/1/3000 | **Today** | Y |
| 1078 | C1038 | Mr | **Ernst Stavro** | Blofeld | N/A | Paris | France | 21/12/1965 | 9/6/1967 | 10/6/1967 | N |
| 1500 | C1038 | Mr | **Ernst Stavro** | Blofeld | N/A | N/A | Japan | 10/6/1967 | **Yesterday** | 10/6/1967 | **N** |
| 2700 | C2345 | Contessa | Teresa | Di Vincenzo | Via Veneto | Rome | Italy | 18/12/1969 | **Yesterday** | **Today** | **N** |
| 3518 | C3467 | Mr | **Marc-Ange** | **Draco** | **Rua Augusta** | **Lisbon** | Portugal | 19/12/1969 | 1/1/3000 | **Today** | Y |
| 2010 | C1239 | **Sir** | Hilary | Bray | **Royal College ...** | London | UK | 20/12/1969 | 1/1/3000 | **Today** | Y |
| **3601** | **C2345** | **Mrs** | **Tracy** | **Bond** | **Rue Du Rhone** | **Geneva** | **Switz...** | **Today** | **1/1/3000** | **Today** | **Y** |
| **3602** | **C7810** | **Miss** | **Vespa** | **Lynd** | **10 Kings Road** | **London** | **UK** | **Today** | **1/1/3000** | **Today** | **Y** |
| **3603** | **C1038** | **Mr** | **Ernst Stavro** | **Blofeld** | **The Summit** | **Piz Gloria** | **Switz...** | **Today** | **1/1/3000** | **Today** | **Y** |

# Exercise 7 Example Star Schema

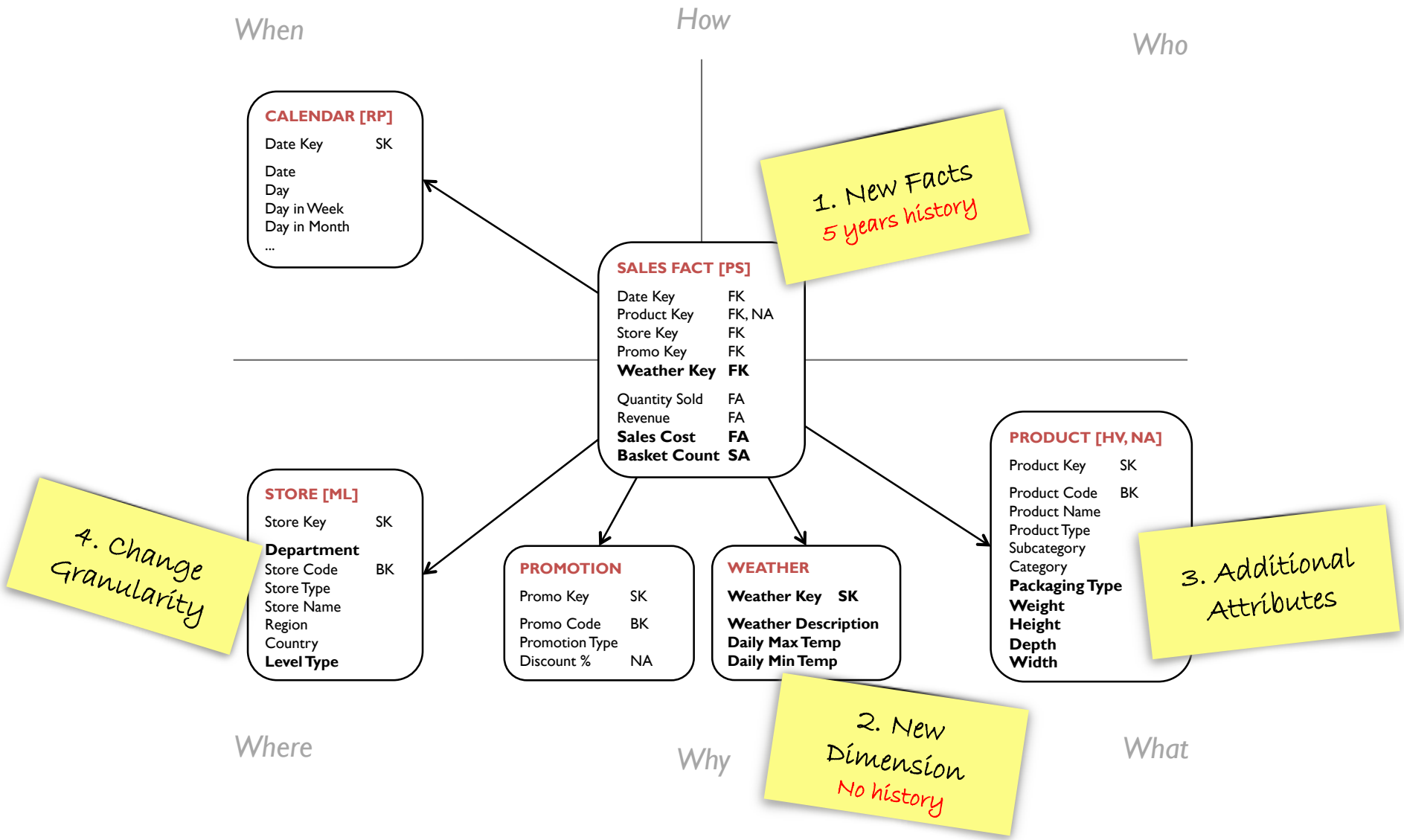


# Exercise 8 Example Simplified Periodic Snapshot Schema



# Exercise 9 Example Answer





# Exercise 9 Example Answer (continued)



Easy/Cheap

Difficult/Costly

4. Change to Multi-Level [ML]
No historical change, Facts remain FA

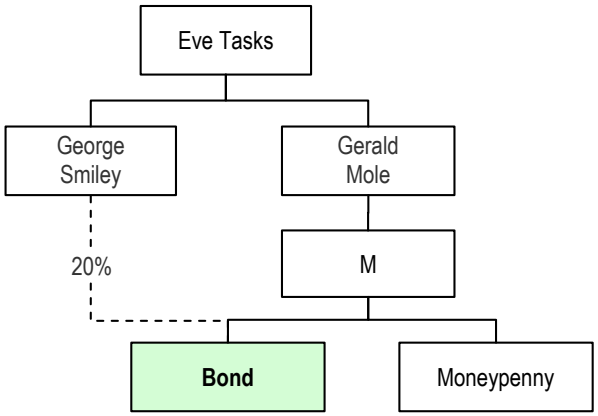
2. New Dimension
No history

1. New Facts
5 years history

3. Additional Attributes
5 years late-arriving dimensional history, conformed Dimension

# Exercise 11 Example Answer

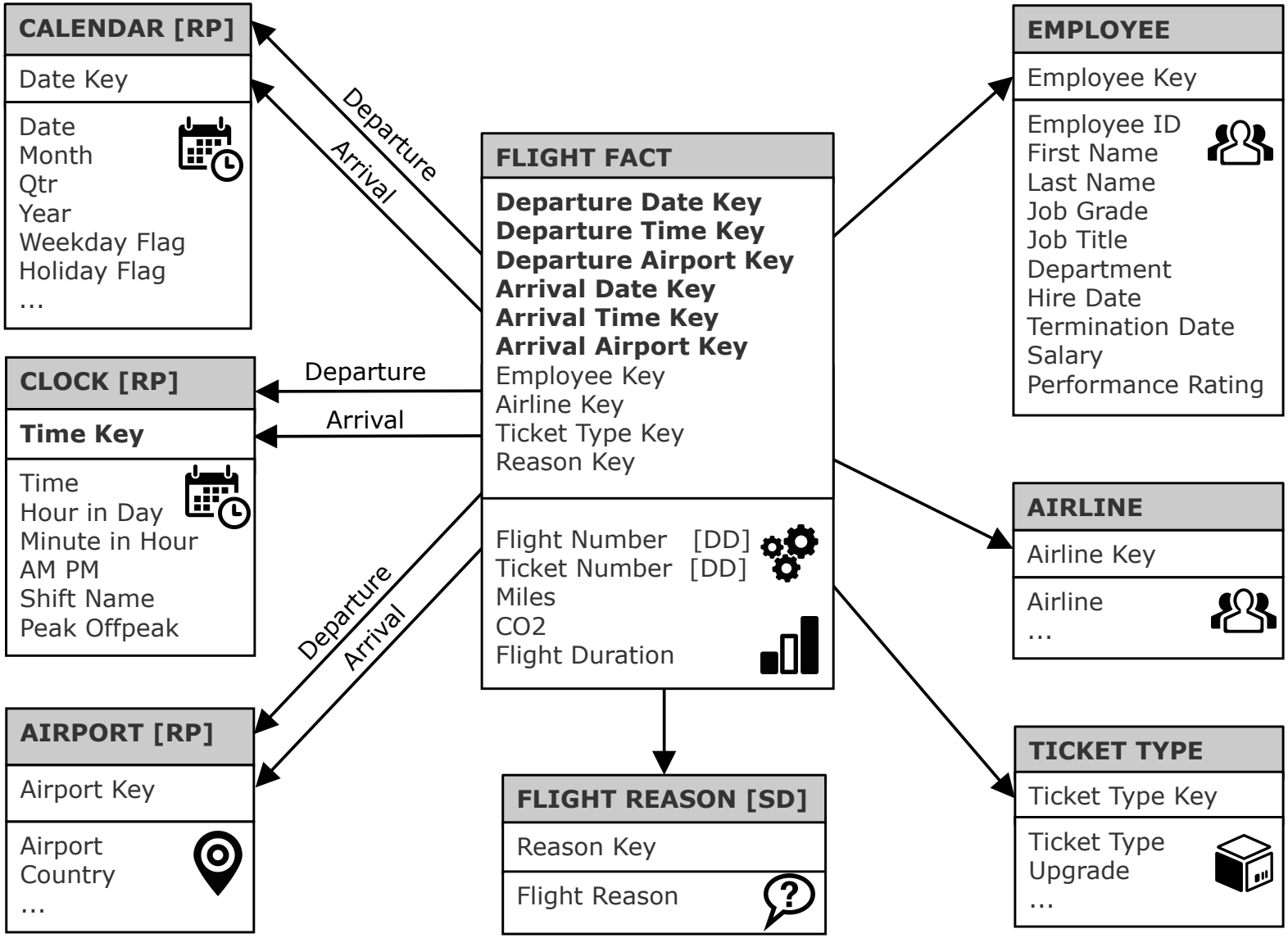
| Reporting Structure [MV, HM] |
|---|
| Manager Key (PK, SK)<br>Employee Key (PK, SK) |
| Employee Level<br>Sequence Number<br>Non Manager Employee<br>Relationship Type<br>FTE Weighting Factor |




**REPORTING STRUCTURE [CV, MV, HM]**

| MANAGER | manages EMPLOYEE | at EMPLOYEE LEVEL | in SEQUENCE NUMBER | NON-MANAGER EMPLOYEE | HIGHEST MANAGER | with RELATIONSHIP TYPE | with WEIGHTING FACTOR |
|---|---|---|---|---|---|---|---|
| SK,PK | SK, PK | N | N | [Y/N] | [Y/N] | | N |
| ET | ET | 1 | 10 | | Y | P | 1 |
| ET | GS | 2 | 20 | | Y | P | 1 |
| ET | Bond | 3 | 30 | Y | Y | T | 0.2 |
| ET | GM | 2 | 40 | | Y | P | 1 |
| ET | M | 3 | 50 | | Y | P | 1 |
| ET | Bond | 4 | 60 | Y | Y | P | 0.8 |
| ET | Moneypenny | 4 | 70 | Y | Y | P | 1 |
| GS | GS | 1 | 10 | | | P | 1 |
| GS | Bond | 2 | 20 | Y | | T | 0.2 |
| GM | GM | 1 | 10 | | | P | 1 |
| GM | M | 2 | 20 | | | P | 1 |
| GM | Bond | 3 | 30 | Y | | P | 0.8 |
| GM | Moneypenny | 3 | 40 | Y | | P | 1 |
| M | M | 1 | 10 | | | P | 1 |
| M | Bond | 2 | 20 | Y | | P | 0.8 |
| M | Moneypenny | 2 | 30 | Y | | P | 1 |
| Bond | Bond | 1 | 10 | Y | | P | **1** |
| Moneypenny | Moneypenny | 1 | 10 | Y | | P | 1 |

# Exercise 12 Example Star Schema




ADWD pp. 221-222