Data Warehousing Concepts

Authored by : Pushkar Kulkarni

Presented by : Pushkar Kulkarni

Agenda

- Data Warehouse Concepts
- Dimensional Modelling

What?

Warehouse

- Correct data - Quality
- Locating Data – Source
- Identifying Data – Location
- Retrieving Faster – Better tuned to get data

Sales

PHOTO: JOE RAEDLE/GETTY IMAGES

Business Problem

Bank wants to understand Purchases made

What

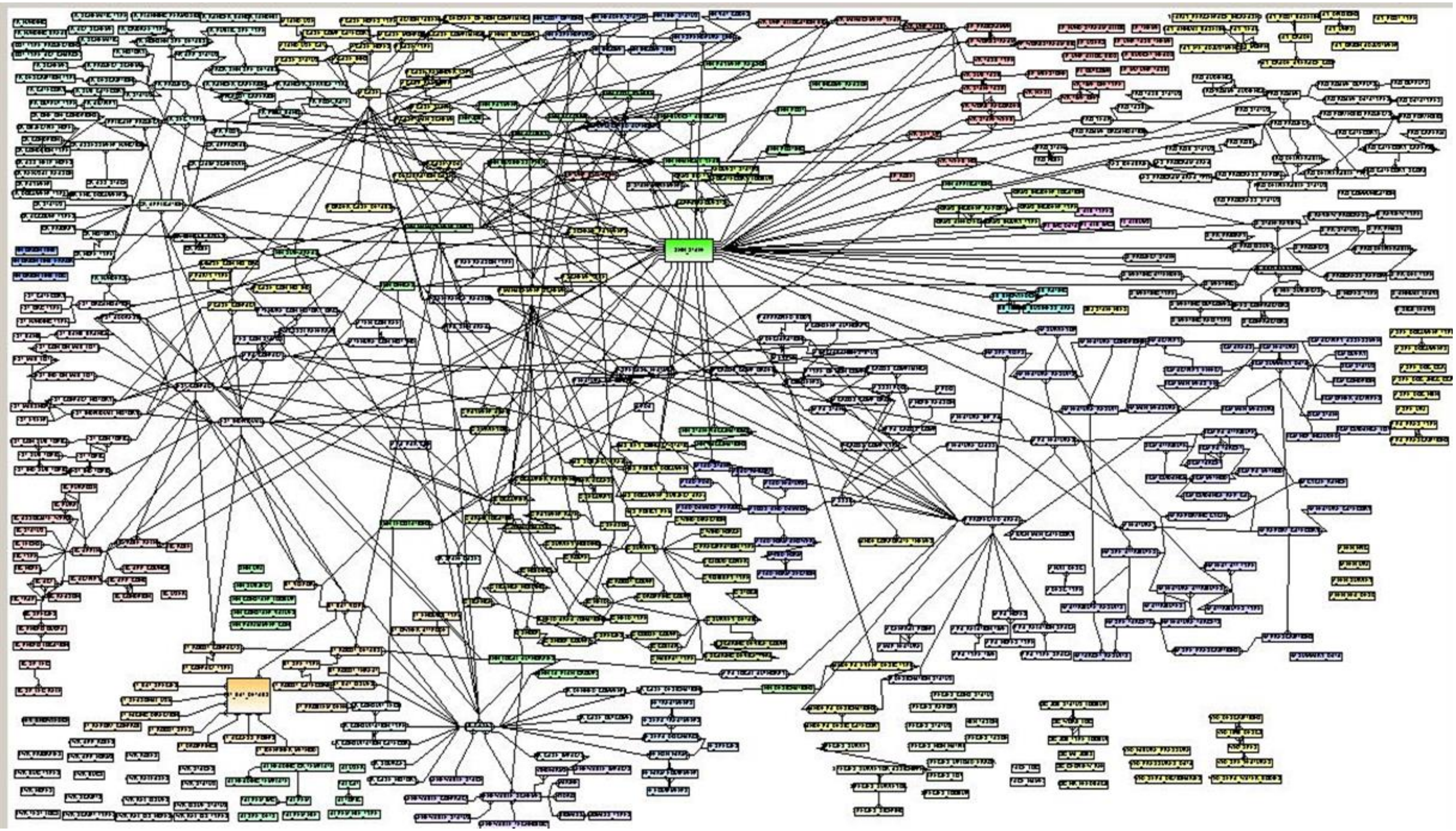
Where

Who

How

Which

Database Design (ER diagram)



Understand the data

- Data Sources
- Data Quality
- Data disparity

Warehouse the data

- Data Quality
- Data Integrity
- Data Definitions across sources
- Data Mapping
- Data governance

What is a Data Warehouse?

Courtesy - http://www.t-systems.com/news-media/prof-dirk-helbing-on-the-opportunities-provided-by-big-data-will-information-become-the-key-resource-of-this-century-t-systems/1100906

Business Intelligence

"*Business Intelligence is*
the process of transforming data
into information
and through discovery transforming that
information into *knowledge*"

Data warehouse - Definitions

A data warehouse is a copy of transaction data specifically structured for querying and reporting

- Ralph

A data warehouse is a:

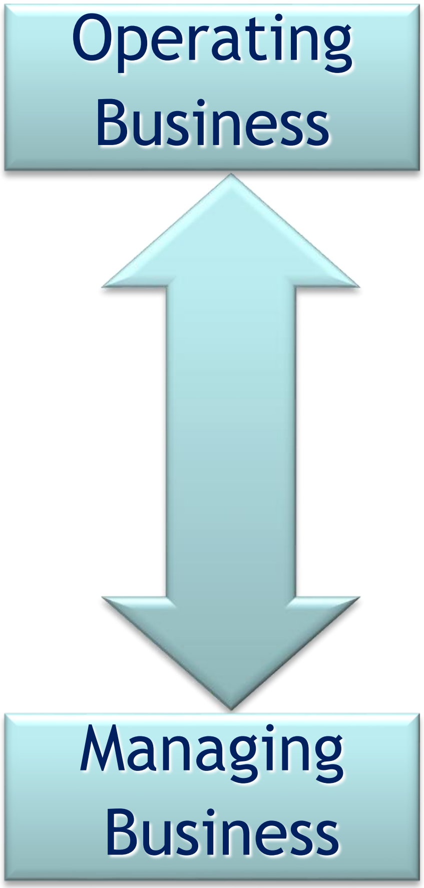
Subject Oriented	Contains information regarding objects of interest for decision support: Sales by region, by product, etc.
Integrated	Data are typically extracted from multiple, heterogeneous data sources (e.g., from sales, inventory, billing DBs etc.).
Non-Volatile	Data is not (or rarely) directly updated
Time Variant	Contain historical data, longer horizon than operational system.

- Bill Inmon

Why Data Warehouse

- Has a business subject area orientation
- Single truth
  - data from multiple, diverse sources
  - Consistent data
  - Single definitions
- Adds ad hoc reporting/Enquiry for analysis of data over time
- Provides analysis capabilities to decision makers
- Data Mining

Why Data Warehouse

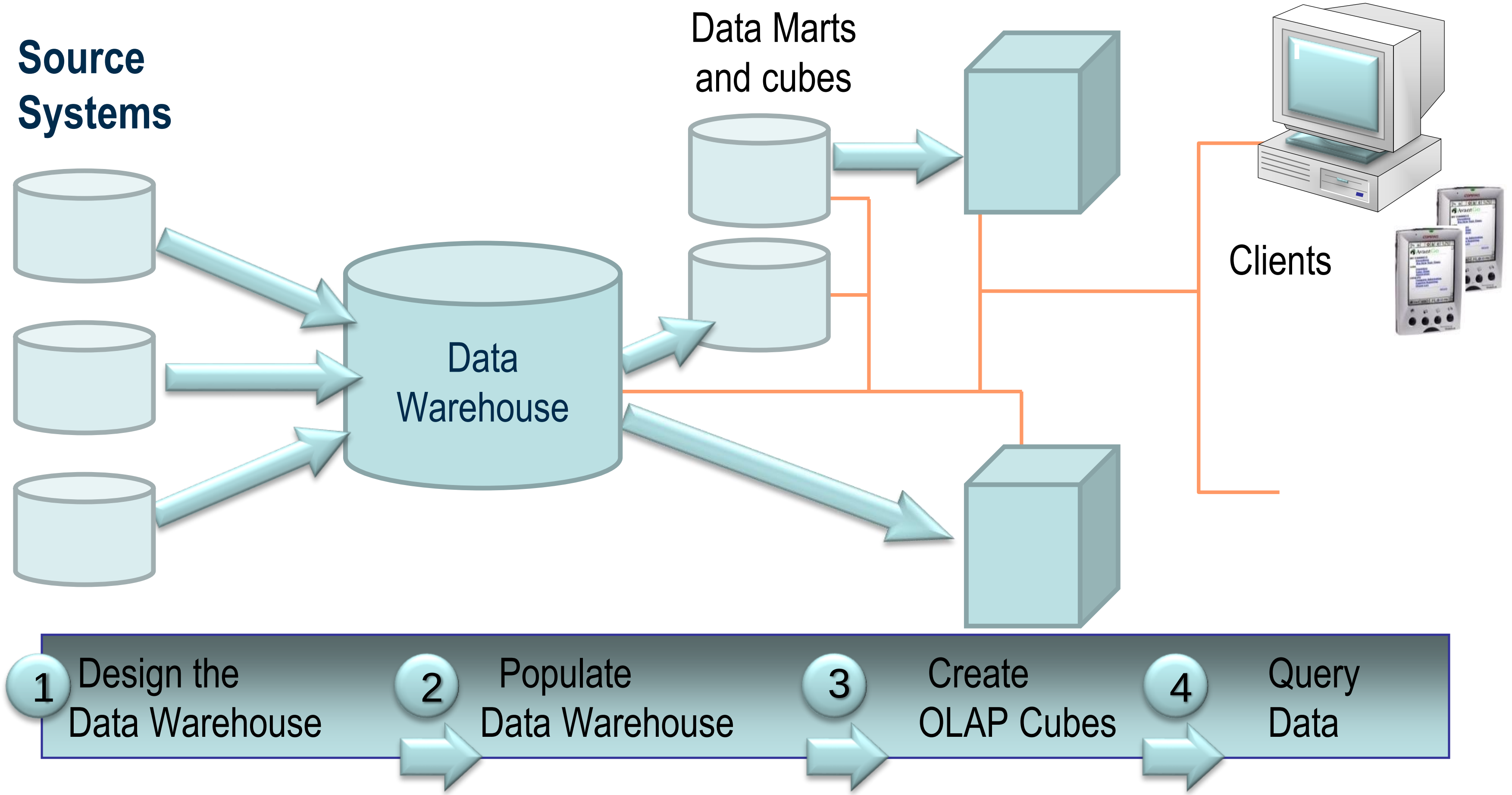


Operating Business

Managing Business

- Online Transaction Processing
  - Granular transactions
  - Real time production systems
  - Current, changing data
- Online Analytics Processing
  - Summarized queries
  - Consistent, heterogeneous data
  - Voluminous, historical, stable data
- OLTP and OLAP applications require different design and storage

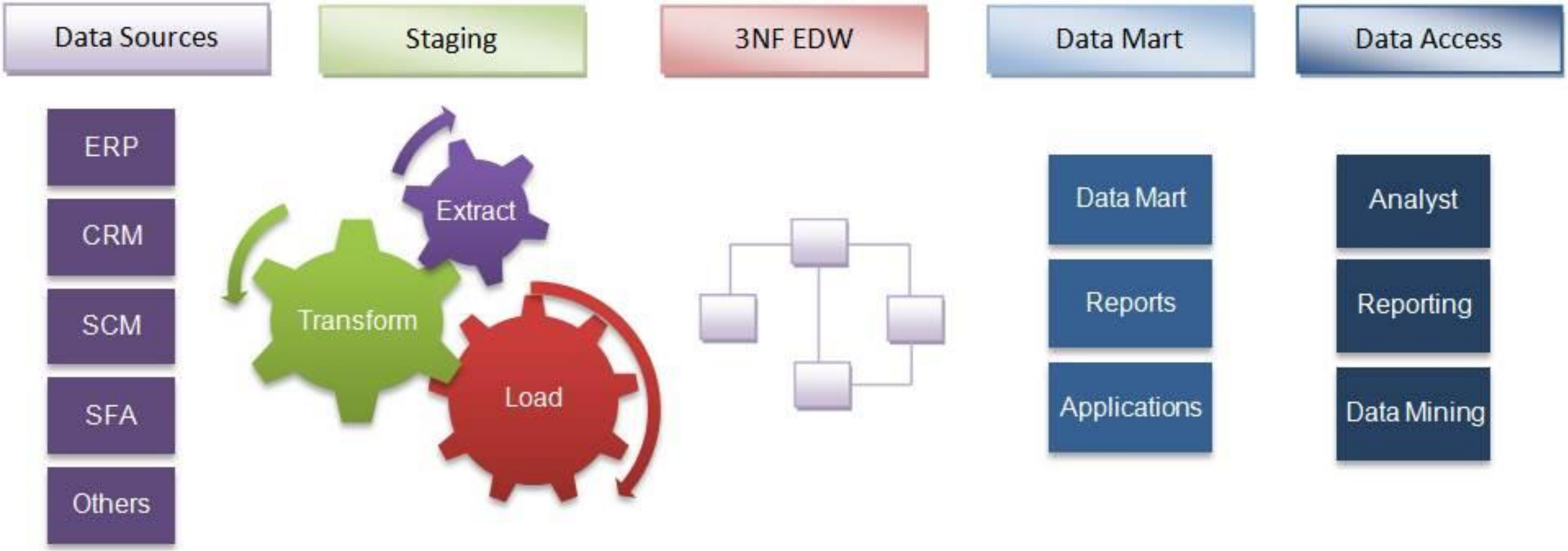
Typical Data Warehouse



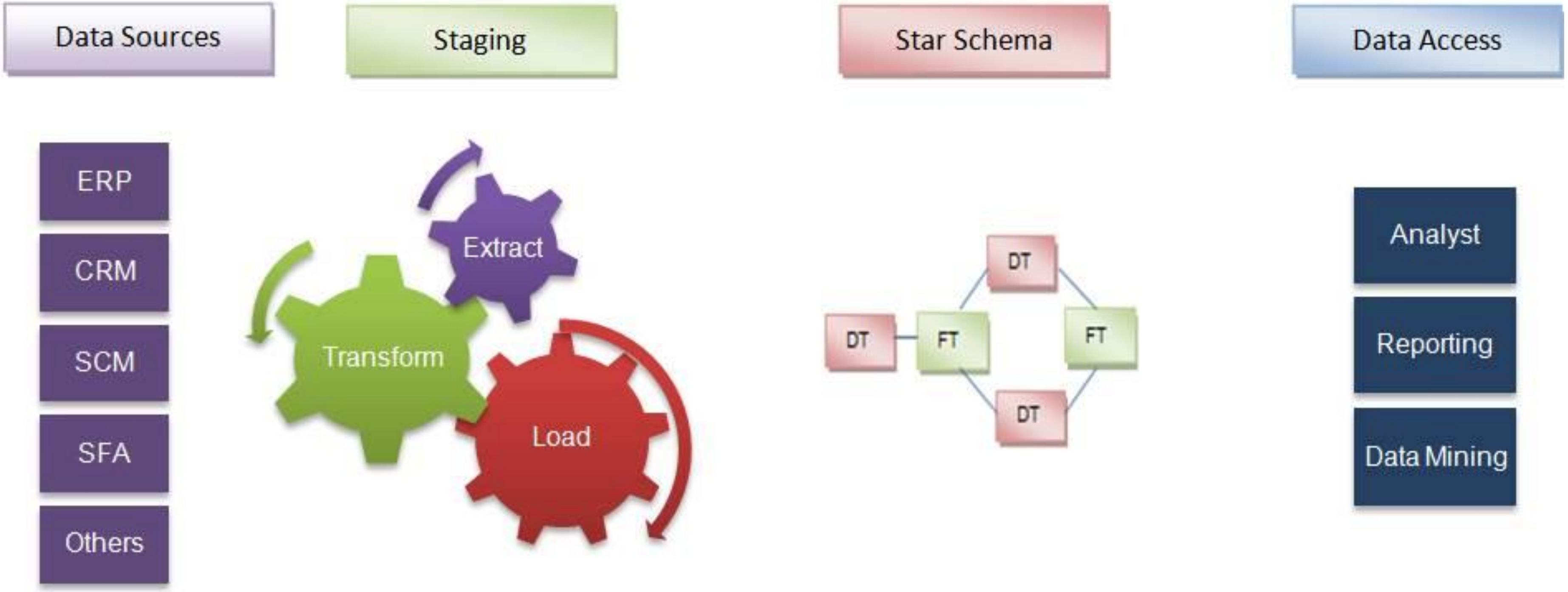
Data Marts

- Subset of the data warehouse
  - oriented to a specific business function or a single department.
- Enables each department to use, manipulate and develop their data any way they see fit;
  - without altering information inside other data marts or the enterprise data warehouse.
- Data marts use the concept of “conformed dimensions”
  - to integrate data across business functions

Top Down Approach - Inmon



Down up Approach – Ralph Kimball



Data Sources

Staging

Star Schema

Data Access

ERP

CRM

SCM

SFA

Others

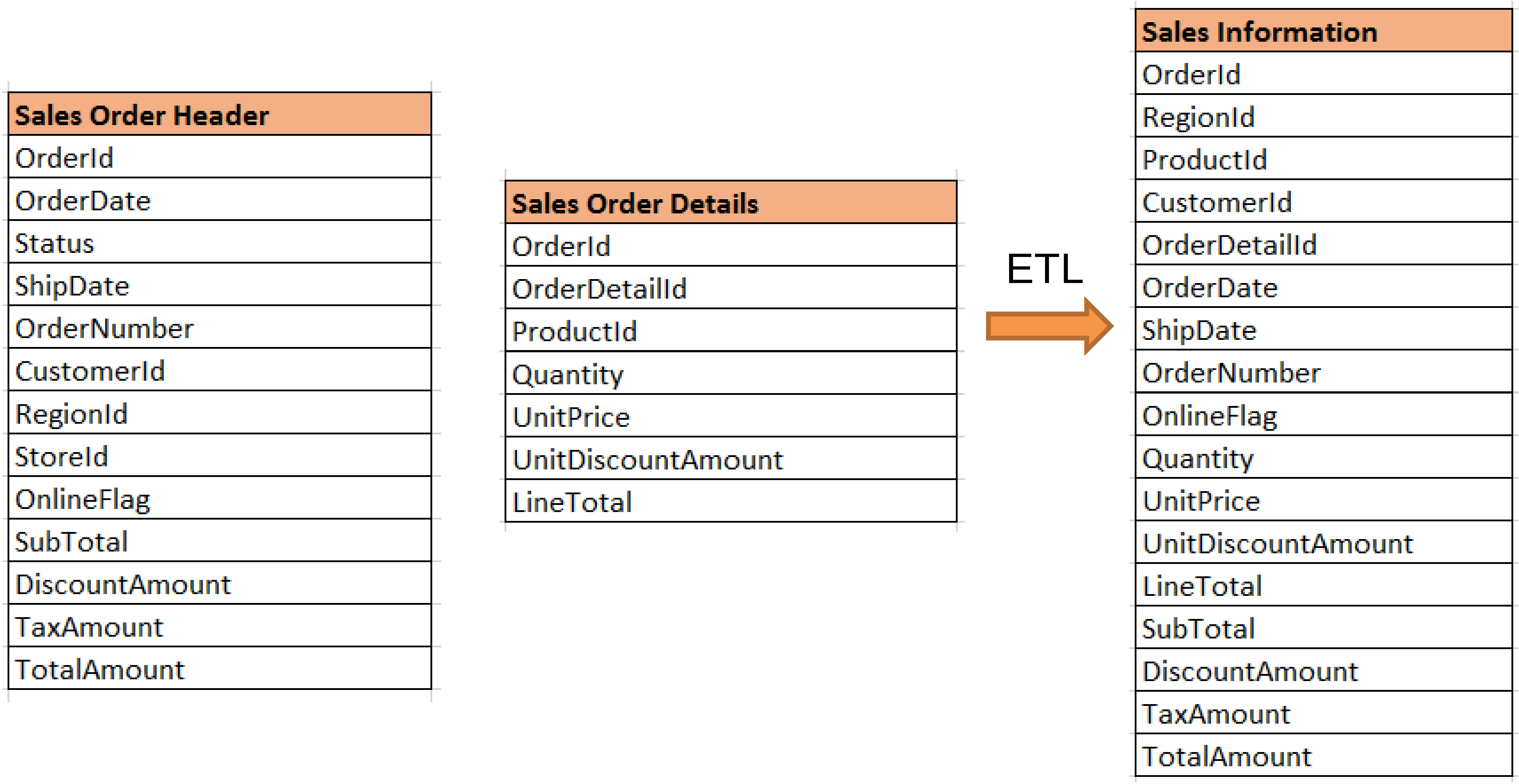
Analyst

Reporting

Data Mining

DT – Dimension Table
FT – Fact Table

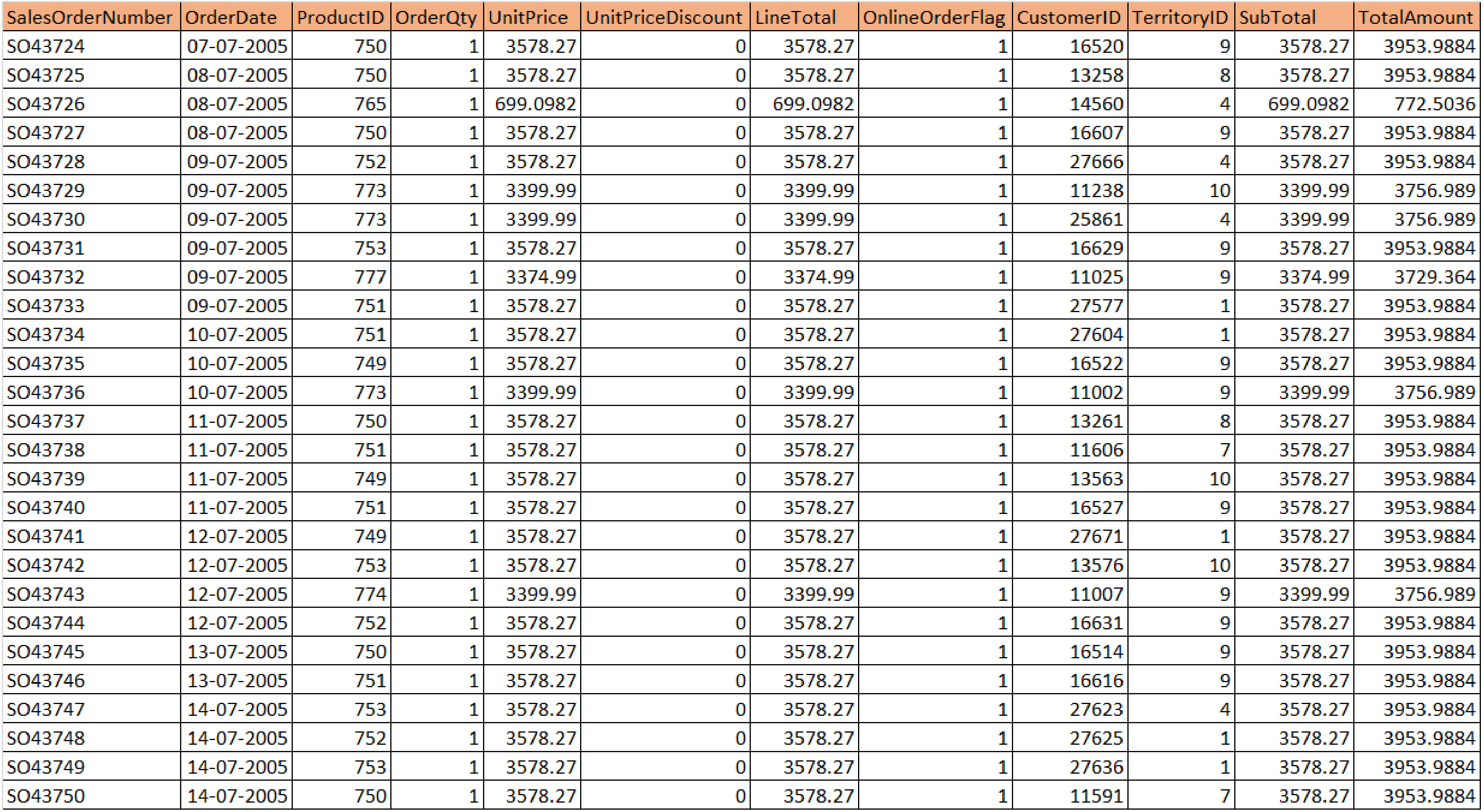
Example - Order Tracking



Sales Order Header
OrderId
OrderDate
Status
ShipDate
OrderNumber
CustomerId
RegionId
StoreId
OnlineFlag
SubTotal
DiscountAmount
TaxAmount
TotalAmount

Sales Order Details
OrderId
OrderDetailId
ProductId
Quantity
UnitPrice
UnitDiscountAmount
LineTotal

Sales Information
OrderId
RegionId
ProductId
CustomerId
OrderDetailId
OrderDate
ShipDate
OrderNumber
OnlineFlag
Quantity
UnitPrice
UnitDiscountAmount
LineTotal
SubTotal
DiscountAmount
TaxAmount
TotalAmount



SalesOrderNumber	OrderDate	ProductID	OrderQty	UnitPrice	UnitPriceDiscount	LineTotal	OnlineOrderFlag	CustomerID	TerritoryID	SubTotal	TotalAmount
SO43724	07-07-2005	750	1	3578.27	0	3578.27	1	16520	9	3578.27	3953.9884
SO43725	08-07-2005	750	1	3578.27	0	3578.27	1	13258	8	3578.27	3953.9884
SO43726	08-07-2005	765	1	699.0982	0	699.0982	1	14560	4	699.0982	772.5036
SO43727	08-07-2005	750	1	3578.27	0	3578.27	1	16607	9	3578.27	3953.9884
SO43728	09-07-2005	752	1	3578.27	0	3578.27	1	27666	4	3578.27	3953.9884
SO43729	09-07-2005	773	1	3399.99	0	3399.99	1	11238	10	3399.99	3756.989
SO43730	09-07-2005	773	1	3399.99	0	3399.99	1	25861	4	3399.99	3756.989
SO43731	09-07-2005	753	1	3578.27	0	3578.27	1	16629	9	3578.27	3953.9884
SO43732	09-07-2005	777	1	3374.99	0	3374.99	1	11025	9	3374.99	3729.364
SO43733	09-07-2005	751	1	3578.27	0	3578.27	1	27577	1	3578.27	3953.9884
SO43734	10-07-2005	751	1	3578.27	0	3578.27	1	27604	1	3578.27	3953.9884
SO43735	10-07-2005	749	1	3578.27	0	3578.27	1	16522	9	3578.27	3953.9884
SO43736	10-07-2005	773	1	3399.99	0	3399.99	1	11002	9	3399.99	3756.989
SO43737	11-07-2005	750	1	3578.27	0	3578.27	1	13261	8	3578.27	3953.9884
SO43738	11-07-2005	751	1	3578.27	0	3578.27	1	11606	7	3578.27	3953.9884
SO43739	11-07-2005	749	1	3578.27	0	3578.27	1	13563	10	3578.27	3953.9884
SO43740	11-07-2005	751	1	3578.27	0	3578.27	1	16527	9	3578.27	3953.9884
SO43741	12-07-2005	749	1	3578.27	0	3578.27	1	27671	1	3578.27	3953.9884
SO43742	12-07-2005	753	1	3578.27	0	3578.27	1	13576	10	3578.27	3953.9884
SO43743	12-07-2005	774	1	3399.99	0	3399.99	1	11007	9	3399.99	3756.989
SO43744	12-07-2005	752	1	3578.27	0	3578.27	1	16631	9	3578.27	3953.9884
SO43745	13-07-2005	750	1	3578.27	0	3578.27	1	16514	9	3578.27	3953.9884
SO43746	13-07-2005	751	1	3578.27	0	3578.27	1	16616	9	3578.27	3953.9884
SO43747	14-07-2005	753	1	3578.27	0	3578.27	1	27623	4	3578.27	3953.9884
SO43748	14-07-2005	752	1	3578.27	0	3578.27	1	27625	1	3578.27	3953.9884
SO43749	14-07-2005	753	1	3578.27	0	3578.27	1	27636	1	3578.27	3953.9884
SO43750	14-07-2005	750	1	3578.27	0	3578.27	1	11591	7	3578.27	3953.9884

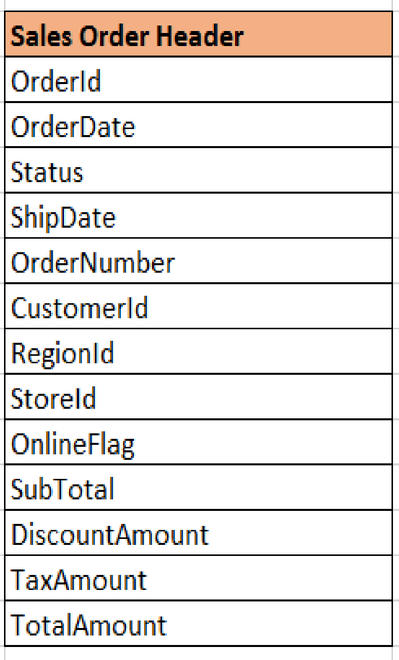
Key Terms

- Grain
  - Important Step in Dimensional Modelling
  - Establishes what single depicts
  - Each Grain might represent different Table
- Measurements
  - Attributes that can be measured
  - Metrics that can be used for analysis
- Facts
  - Consists of the measurements, metrics or facts of the business process

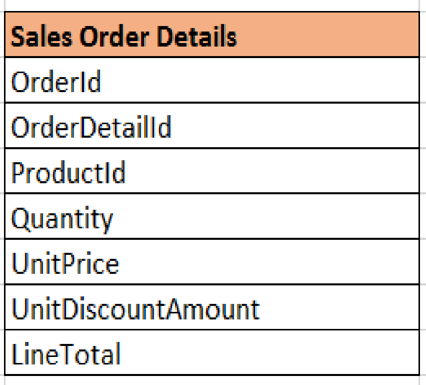
Key Terms

- Facts Type
  - Additive
    - Measures that can be added across all keys
  - Semi Additive
    - Measures that can be added across some keys
  - Non Additive
    - Measures that cannot be added across keys

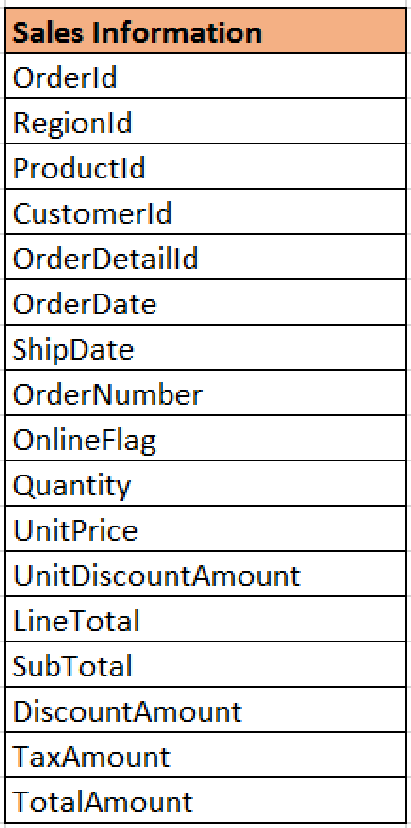
Fact Table Types



Sales Order Header
OrderId
OrderDate
Status
ShipDate
OrderNumber
CustomerId
RegionId
StoreId
OnlineFlag
SubTotal
DiscountAmount
TaxAmount
TotalAmount



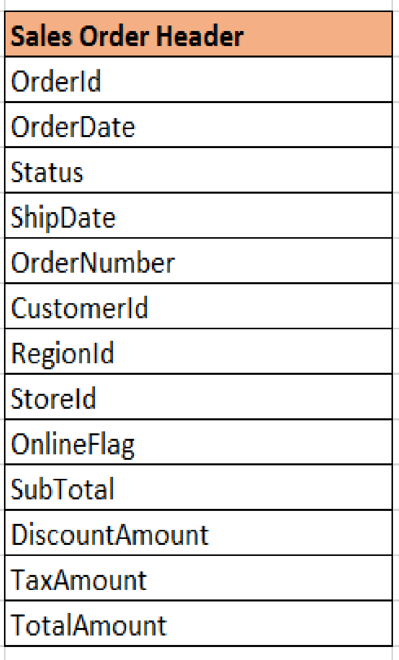
Sales Order Details
OrderId
OrderDetailId
ProductId
Quantity
UnitPrice
UnitDiscountAmount
LineTotal



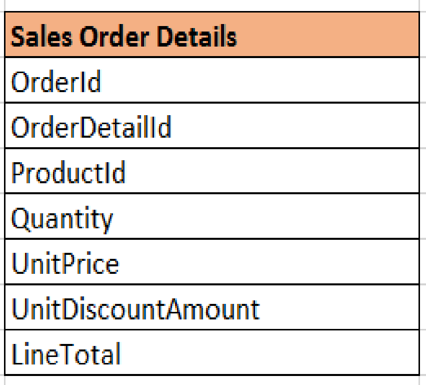
Sales Information
OrderId
RegionId
ProductId
CustomerId
OrderDetailId
OrderDate
ShipDate
OrderNumber
OnlineFlag
Quantity
UnitPrice
UnitDiscountAmount
LineTotal
SubTotal
DiscountAmount
TaxAmount
TotalAmount

- Transactional Fact Table
  - Basic and fundamental
  - Most expressive
  - Dense or Sparse

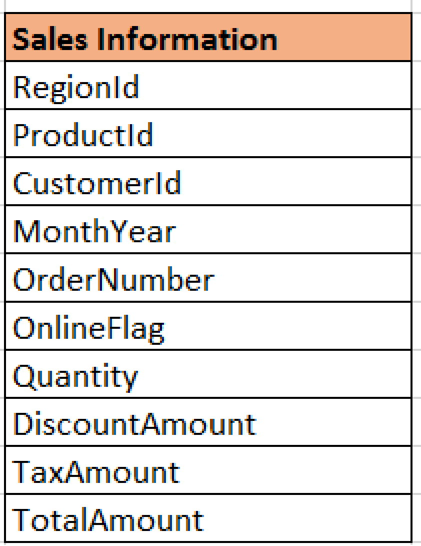
Fact Table Types



Sales Order Header
OrderId
OrderDate
Status
ShipDate
OrderNumber
CustomerId
RegionId
StoreId
OnlineFlag
SubTotal
DiscountAmount
TaxAmount
TotalAmount



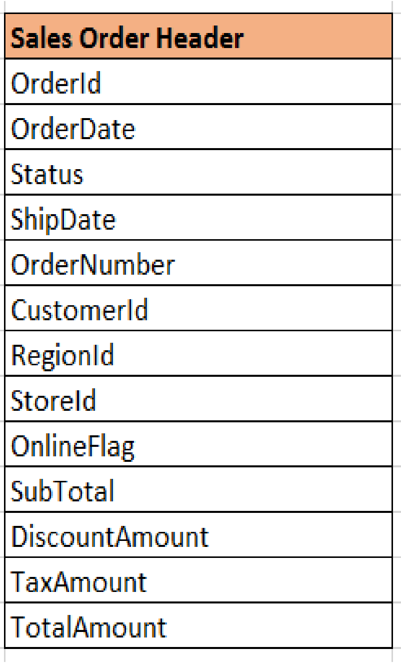
Sales Order Details
OrderId
OrderDetailId
ProductId
Quantity
UnitPrice
UnitDiscountAmount
LineTotal



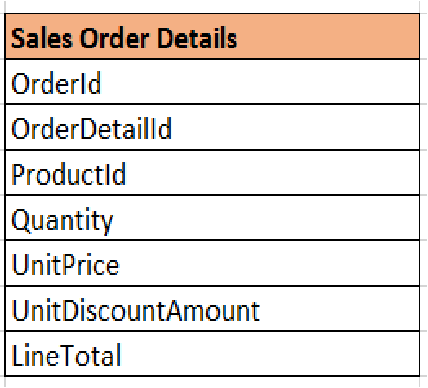
Sales Information
RegionId
ProductId
CustomerId
MonthYear
OrderNumber
OnlineFlag
Quantity
DiscountAmount
TaxAmount
TotalAmount

- Periodic snapshots
  - Summarizes measurements occurring over a period – day, week, month
  - Usually Grain is Period
  - Usually Dense – typically row is inserted with Zero or Null

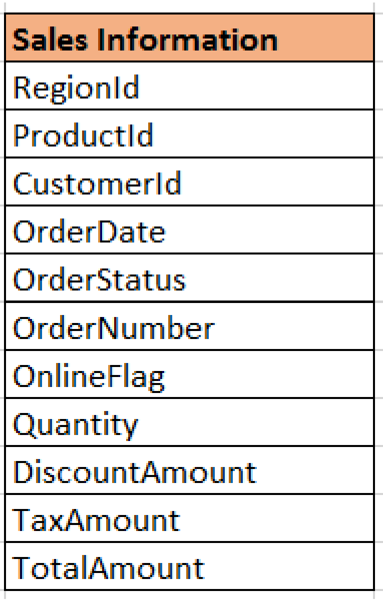
Fact Table Types



Sales Order Header
OrderId
OrderDate
Status
ShipDate
OrderNumber
CustomerId
RegionId
StoreId
OnlineFlag
SubTotal
DiscountAmount
TaxAmount
TotalAmount



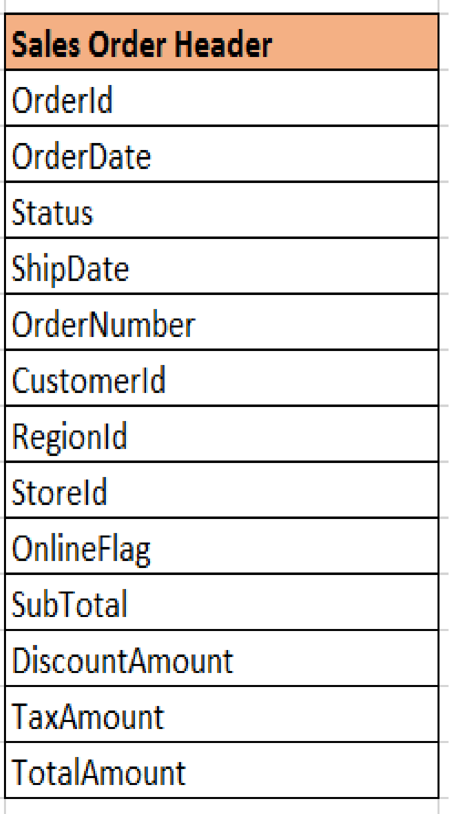
Sales Order Details
OrderId
OrderDetailId
ProductId
Quantity
UnitPrice
UnitDiscountAmount
LineTotal



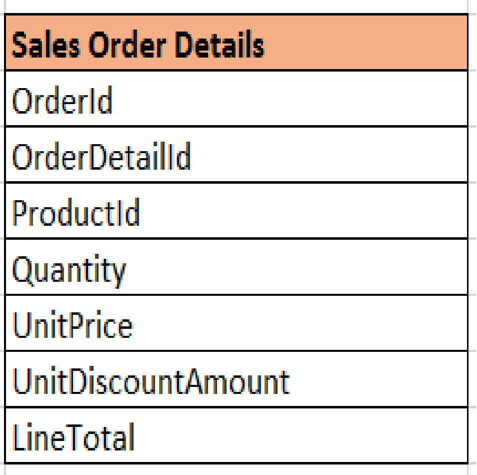
Sales Information
RegionId
ProductId
CustomerId
OrderDate
OrderStatus
OrderNumber
OnlineFlag
Quantity
DiscountAmount
TaxAmount
TotalAmount

- Accumulating snapshots
  - Summarizes at predictable steps
  - Pipeline of workflow
  - Frequent updates at every step.

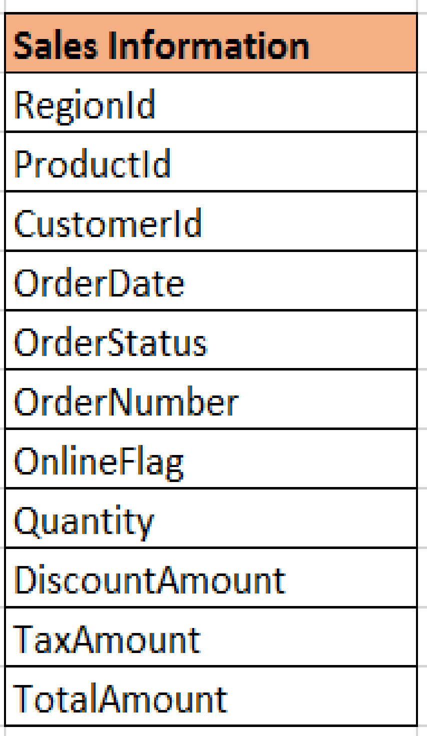
Key Terms



Sales Order Header
OrderId
OrderDate
Status
ShipDate
OrderNumber
CustomerId
RegionId
StoreId
OnlineFlag
SubTotal
DiscountAmount
TaxAmount
TotalAmount

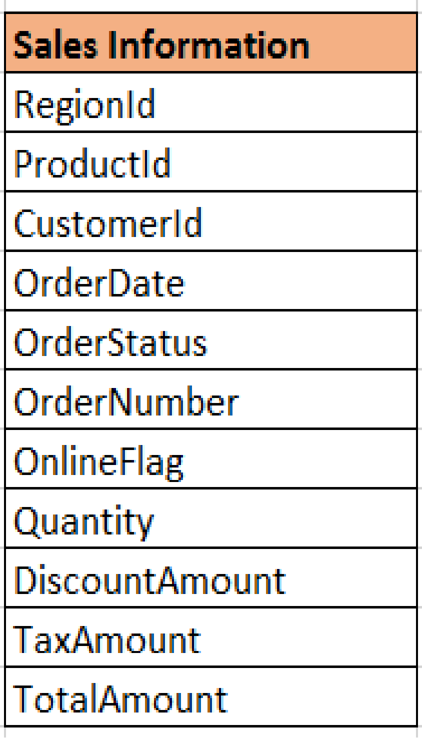


Sales Order Details
OrderId
OrderDetailId
ProductId
Quantity
UnitPrice
UnitDiscountAmount
LineTotal

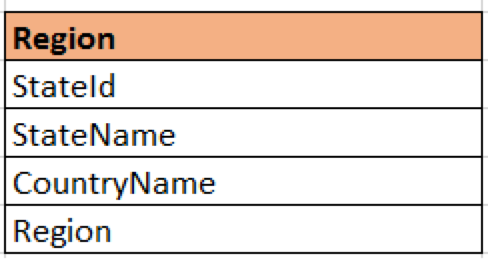


Sales Information
RegionId
ProductId
CustomerId
OrderDate
OrderStatus
OrderNumber
OnlineFlag
Quantity
DiscountAmount
TaxAmount
TotalAmount

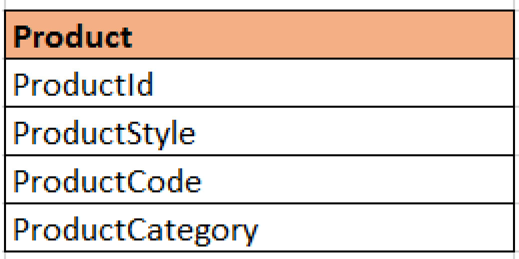
Key terms



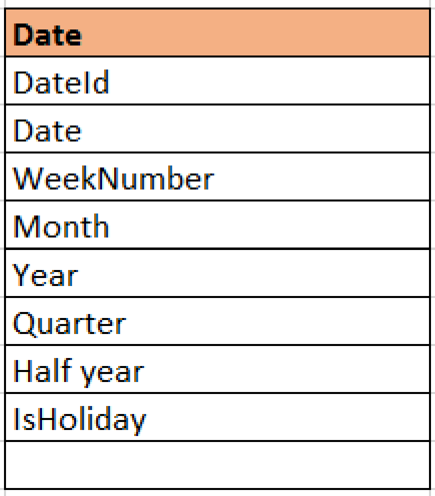
Sales Information
RegionId
ProductId
CustomerId
OrderDate
OrderStatus
OrderNumber
OnlineFlag
Quantity
DiscountAmount
TaxAmount
TotalAmount



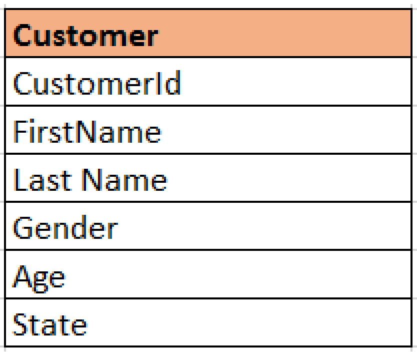
Region
StateId
StateName
CountryName
Region



Product
ProductId
ProductStyle
ProductCode
ProductCategory



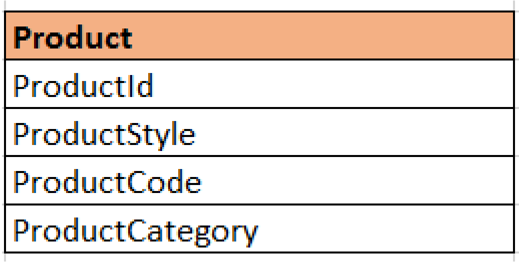
Date
DateId
Date
WeekNumber
Month
Year
Quarter
Half year
IsHoliday



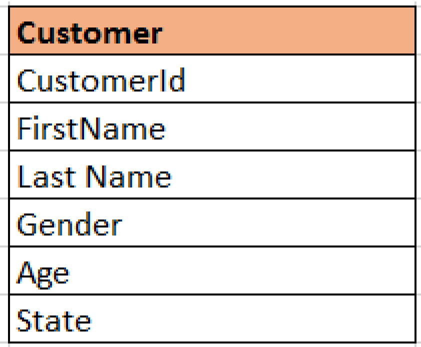
Customer
CustomerId
FirstName
Last Name
Gender
Age
State

Key Terms

- Dimensions
  - Dimensions provide the “who, what, where, when, why, and how” context surrounding a business process event.
  - Structure that categorizes facts

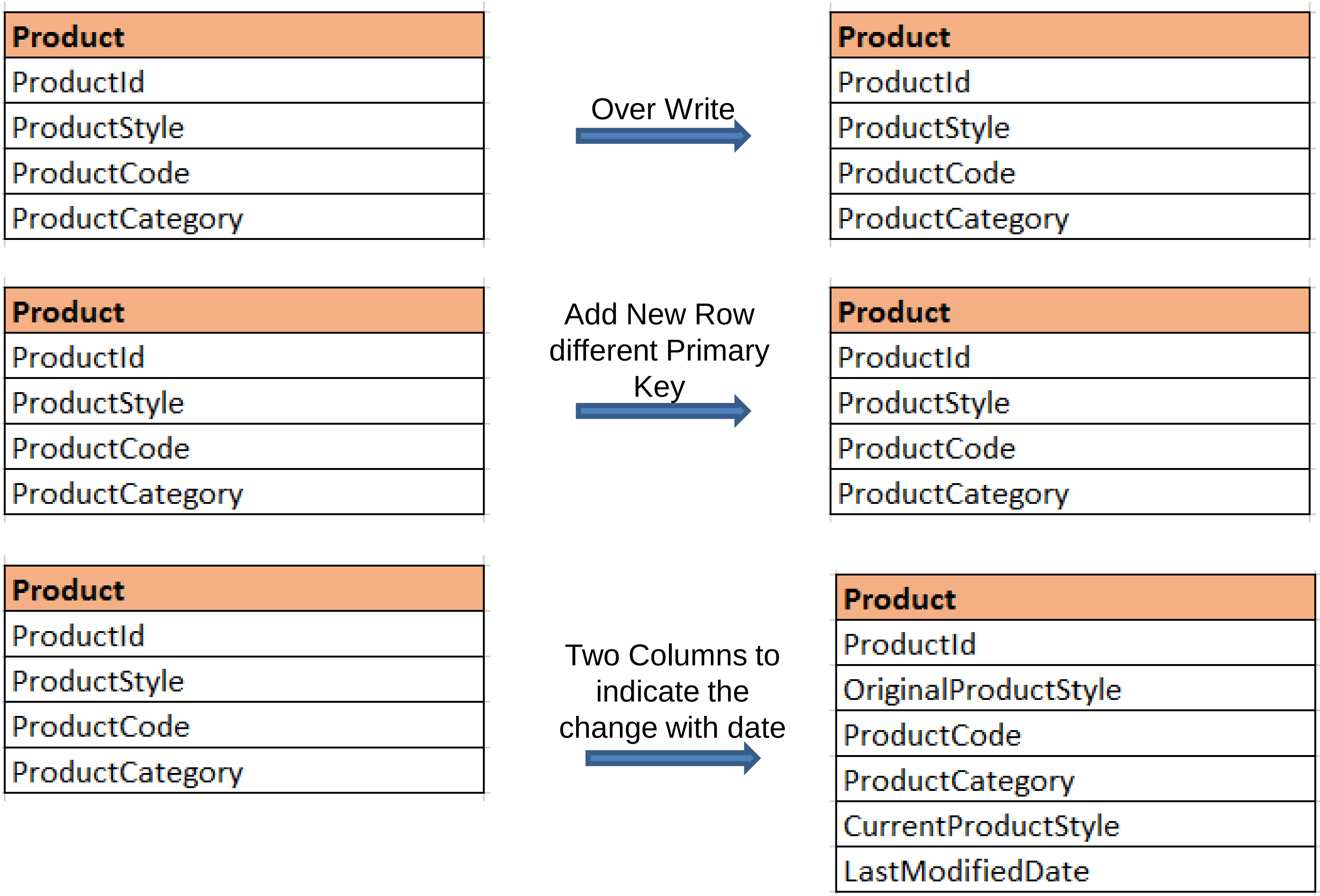


Product
ProductId
ProductStyle
ProductCode
ProductCategory



Customer
CustomerId
FirstName
Last Name
Gender
Age
State

Slowly Changing Dimensions



Type 1

Product
ProductId
ProductStyle
ProductCode
ProductCategory

Over Write

Product
ProductId
ProductStyle
ProductCode
ProductCategory

Type 2

Product
ProductId
ProductStyle
ProductCode
ProductCategory

Add New Row different Primary Key

Product
ProductId
ProductStyle
ProductCode
ProductCategory

Type 3

Product
ProductId
ProductStyle
ProductCode
ProductCategory

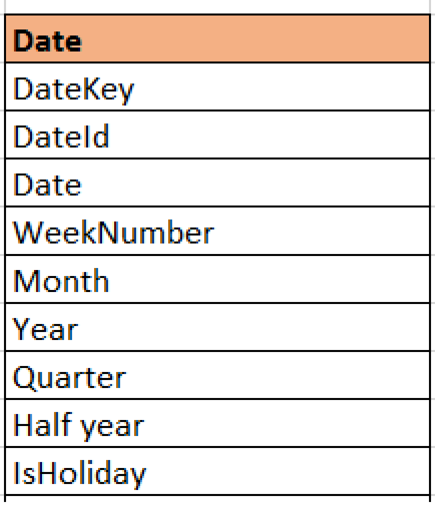
Two Columns to indicate the change with date

Product
ProductId
OriginalProductStyle
ProductCode
ProductCategory
CurrentProductStyle
LastModifiedDate

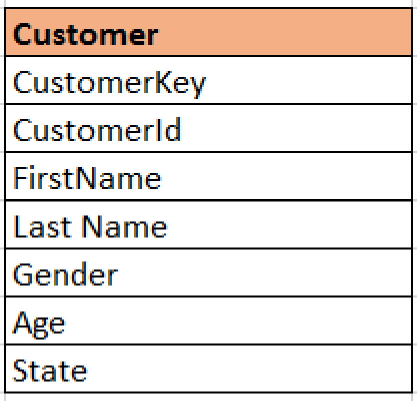
Other Types

- Conformed dimension
  - Multiple references with same meaning
  - A conformed dimension cuts across many facts.
- Junk dimension
  - Grouping of typically low-cardinality flags and indicators.
- Degenerate dimension
  - A degenerate dimension is a key, such as a transaction number that has no attributes and hence does not join to an actual dimension table.
- Role-playing dimension
  - Dimensions are often recycled for multiple applications within the same database. For instance, a "Date" dimension can be used for "OrderDate", as well as "ShipDate"

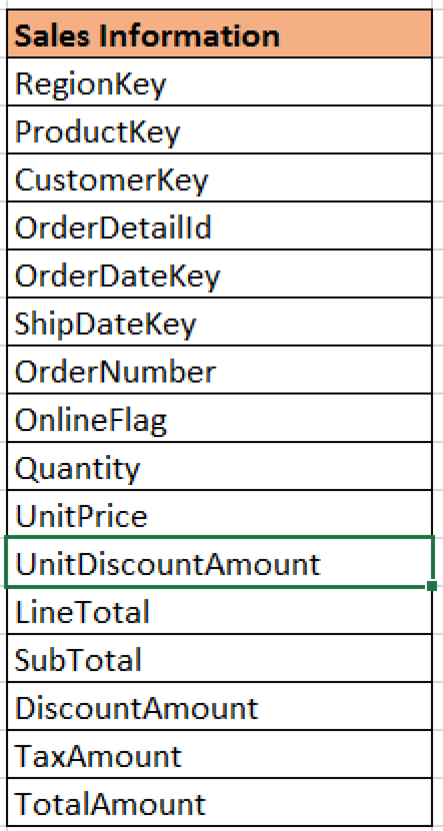
Surrogate Keys



Date
DateKey
DateId
Date
WeekNumber
Month
Year
Quarter
Half year
IsHoliday

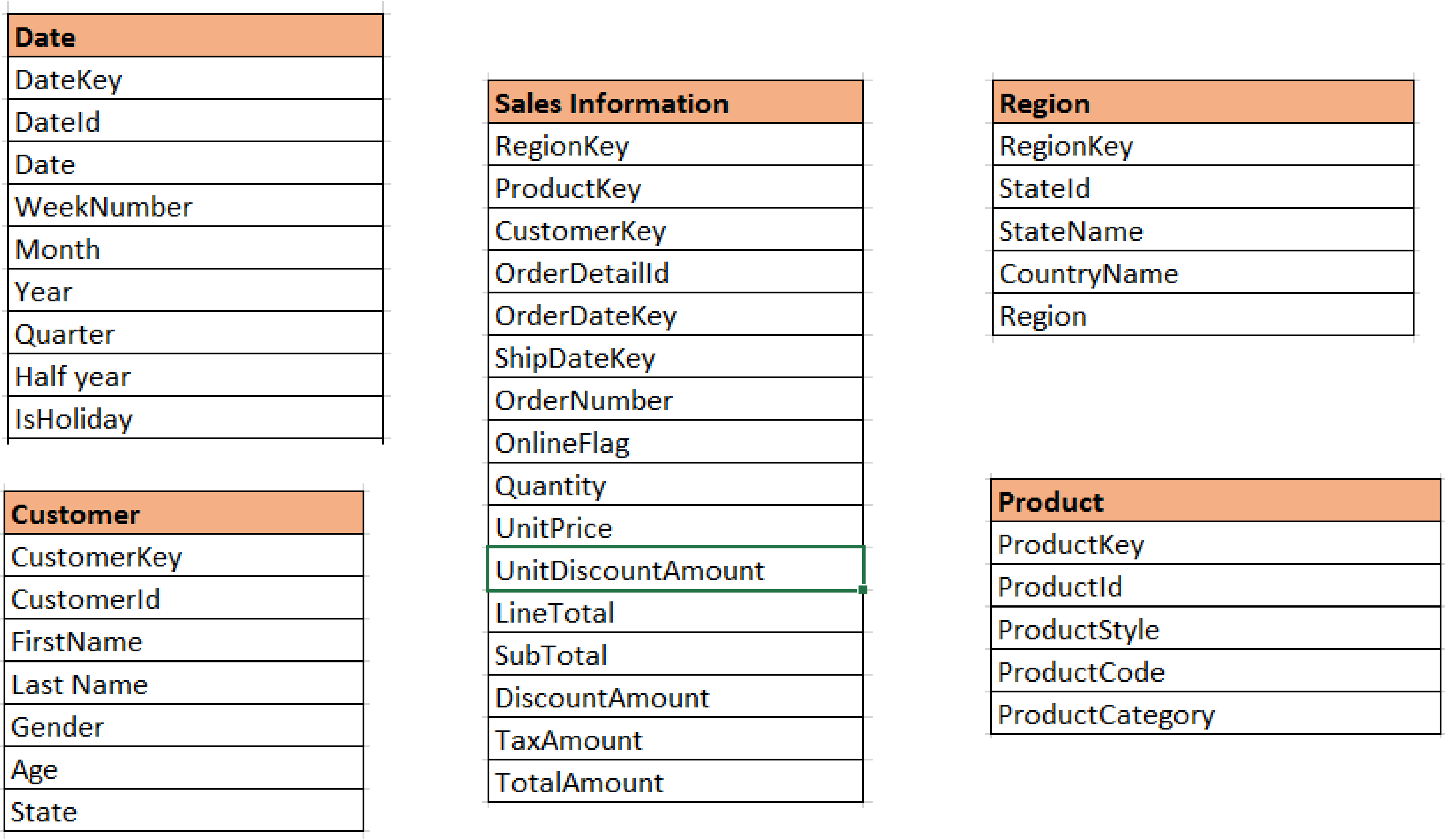


Customer
CustomerKey
CustomerId
FirstName
Last Name
Gender
Age
State



Sales Information
RegionKey
ProductKey
CustomerKey
OrderDetailId
OrderDateKey
ShipDateKey
OrderNumber
OnlineFlag
Quantity
UnitPrice
UnitDiscountAmount
LineTotal
SubTotal
DiscountAmount
TaxAmount
TotalAmount

Star Schema



Date
DateKey
DateId
Date
WeekNumber
Month
Year
Quarter
Half year
IsHoliday

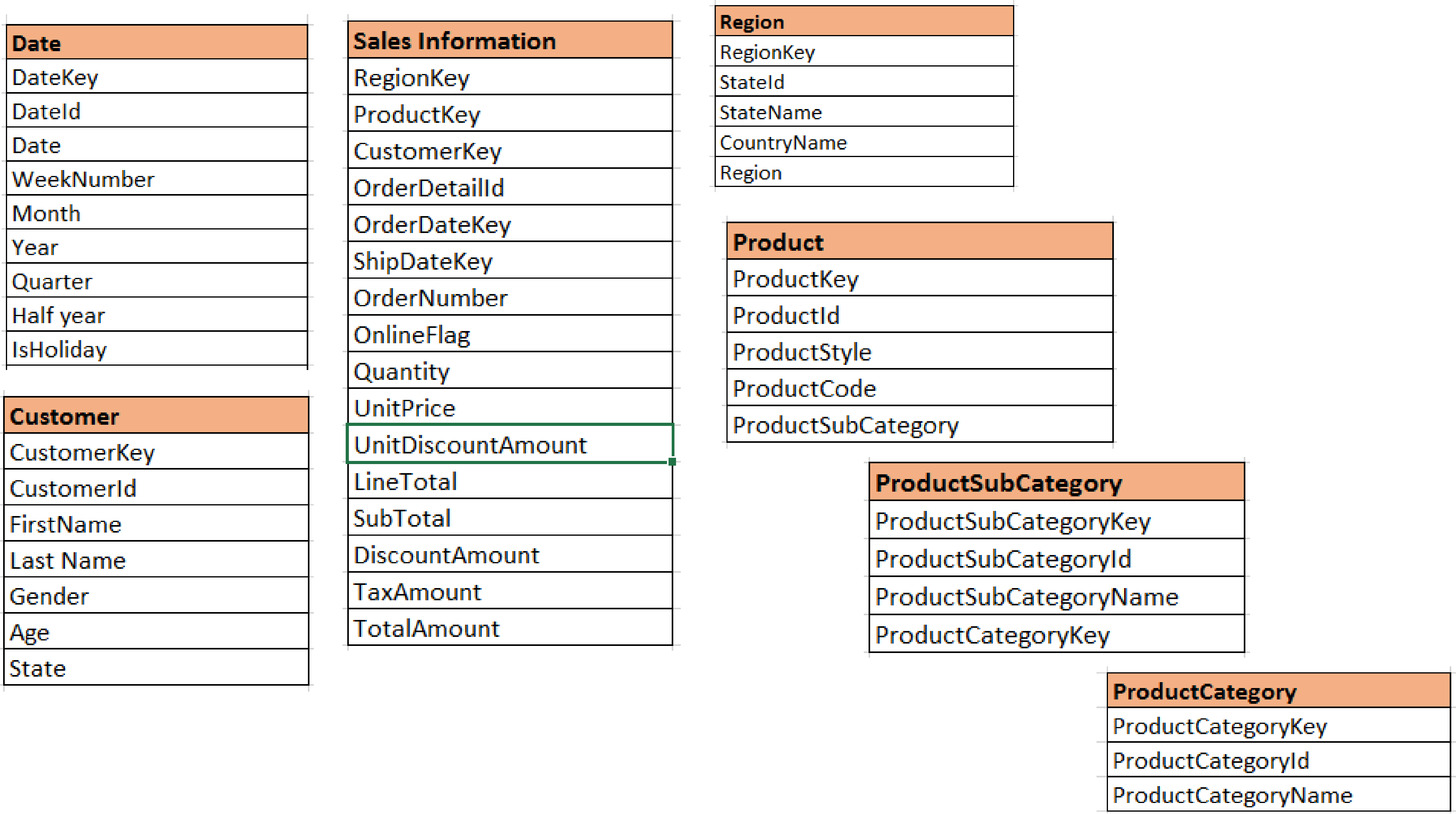
Customer
CustomerKey
CustomerId
FirstName
Last Name
Gender
Age
State

Sales Information
RegionKey
ProductKey
CustomerKey
OrderDetailId
OrderDateKey
ShipDateKey
OrderNumber
OnlineFlag
Quantity
UnitPrice
UnitDiscountAmount
LineTotal
SubTotal
DiscountAmount
TaxAmount
TotalAmount

Region
RegionKey
StateId
StateName
CountryName
Region

Product
ProductKey
ProductId
ProductStyle
ProductCode
ProductCategory

SnowFlake Schema



Date
DateKey
DateId
Date
WeekNumber
Month
Year
Quarter
Half year
IsHoliday

Customer
CustomerKey
CustomerId
FirstName
Last Name
Gender
Age
State

Sales Information
RegionKey
ProductKey
CustomerKey
OrderDetailId
OrderDateKey
ShipDateKey
OrderNumber
OnlineFlag
Quantity
UnitPrice
UnitDiscountAmount
LineTotal
SubTotal
DiscountAmount
TaxAmount
TotalAmount

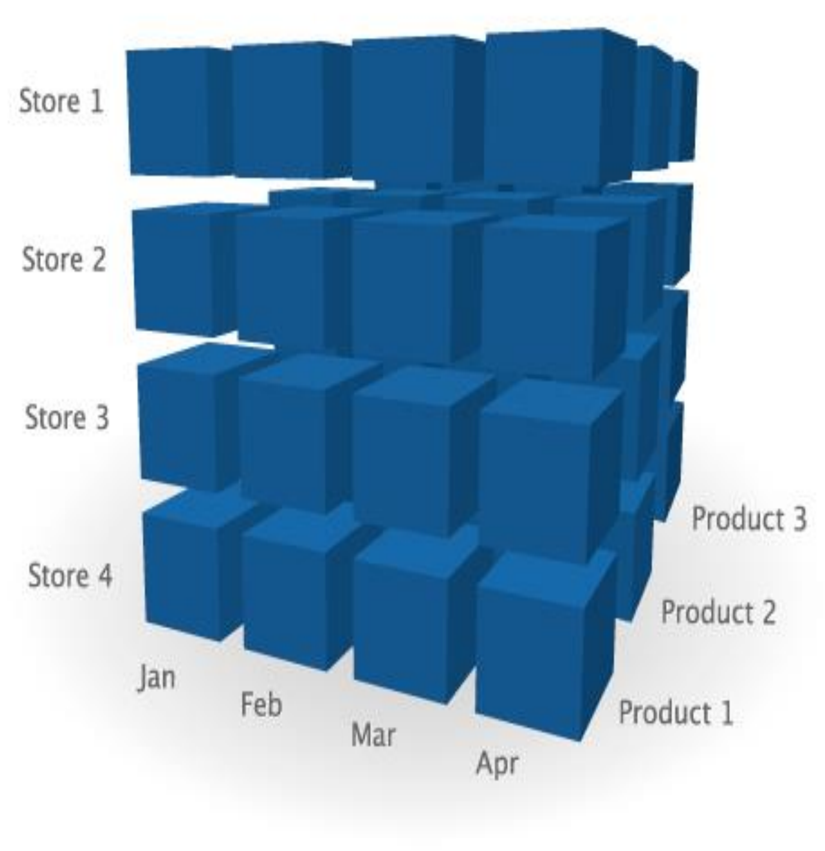
Region
RegionKey
StateId
StateName
CountryName
Region

Product
ProductKey
ProductId
ProductStyle
ProductCode
ProductSubCategory

ProductSubCategory
ProductSubCategoryKey
ProductSubCategoryId
ProductSubCategoryName
ProductCategoryKey

ProductCategory
ProductCategoryKey
ProductCategoryId
ProductCategoryName

What is a Cube



An cube is an array of data understood in terms of its 0 or more dimensions.

Multi-Dimensional Structure

Simplifies the mechanism of viewing the data

OLAP Types

- MOLAP - Multidimensional online analytical processing
  - Stores data and summarized data in multi-dimensional cubes
  - View data only using a cube browsing tool
- ROLAP - Relational online analytical processing
  - Stores data in relational tables
  - And stores aggregations in index views
- HOLAP - Combination of ROLAP and MOLAP
  - Stores data in ROLAP tables and aggregations in a MOLAP cube

Slice

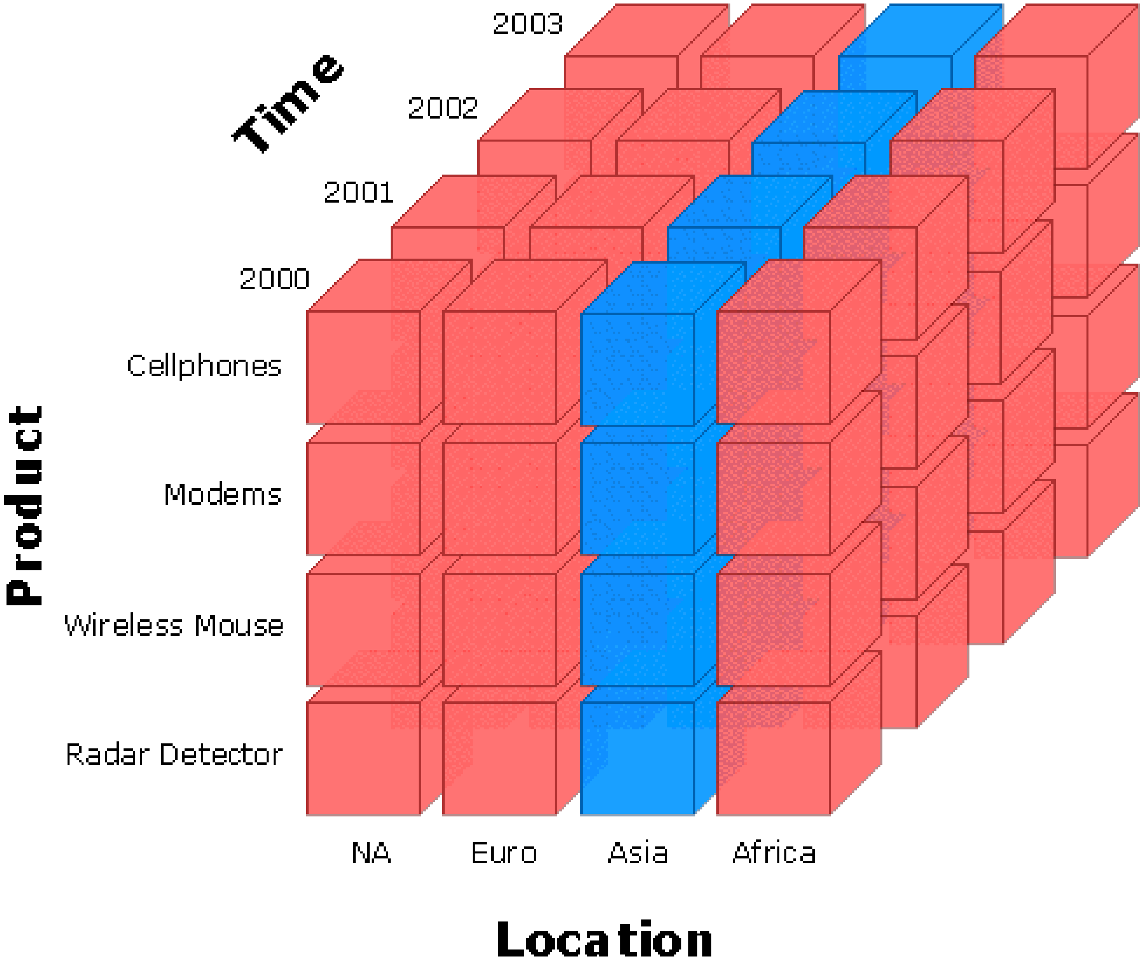


Image Copied from www.hypertextbookshop.com

Dice

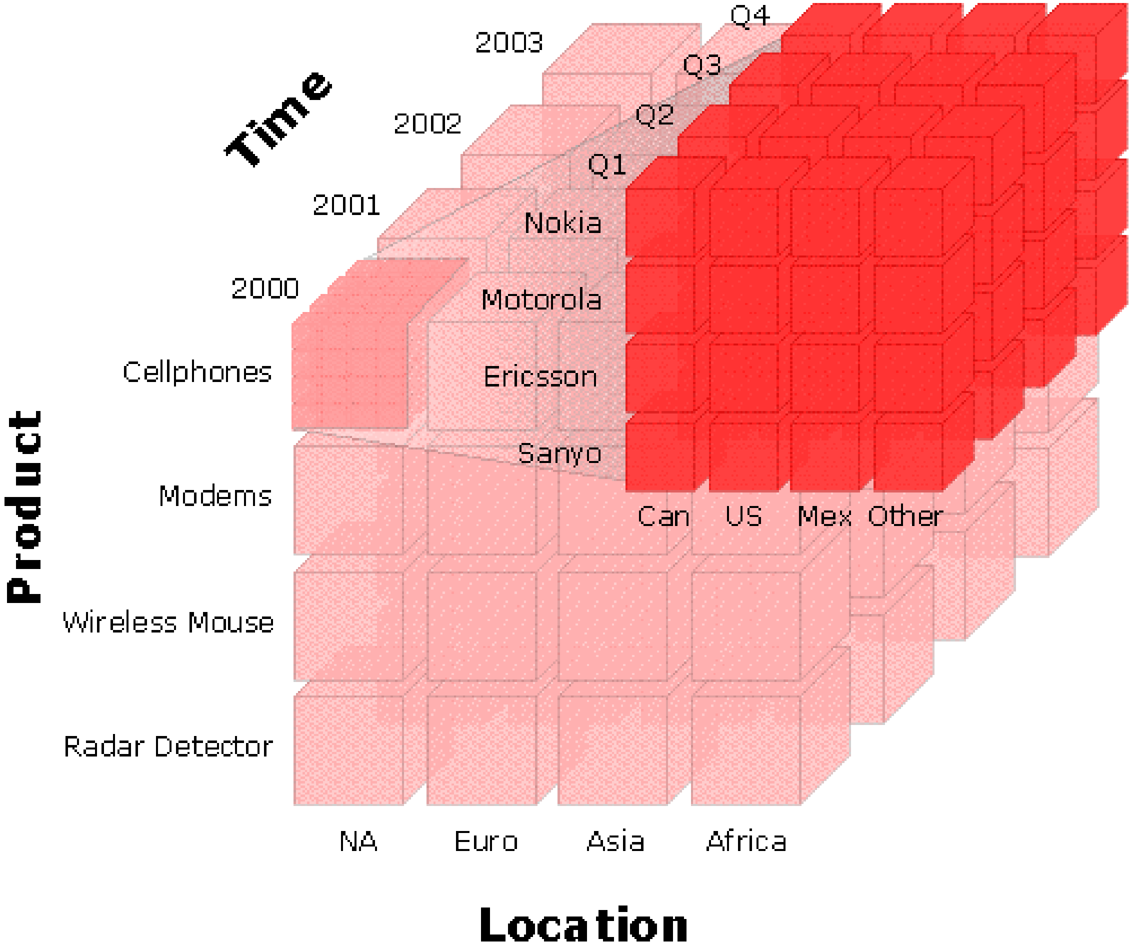
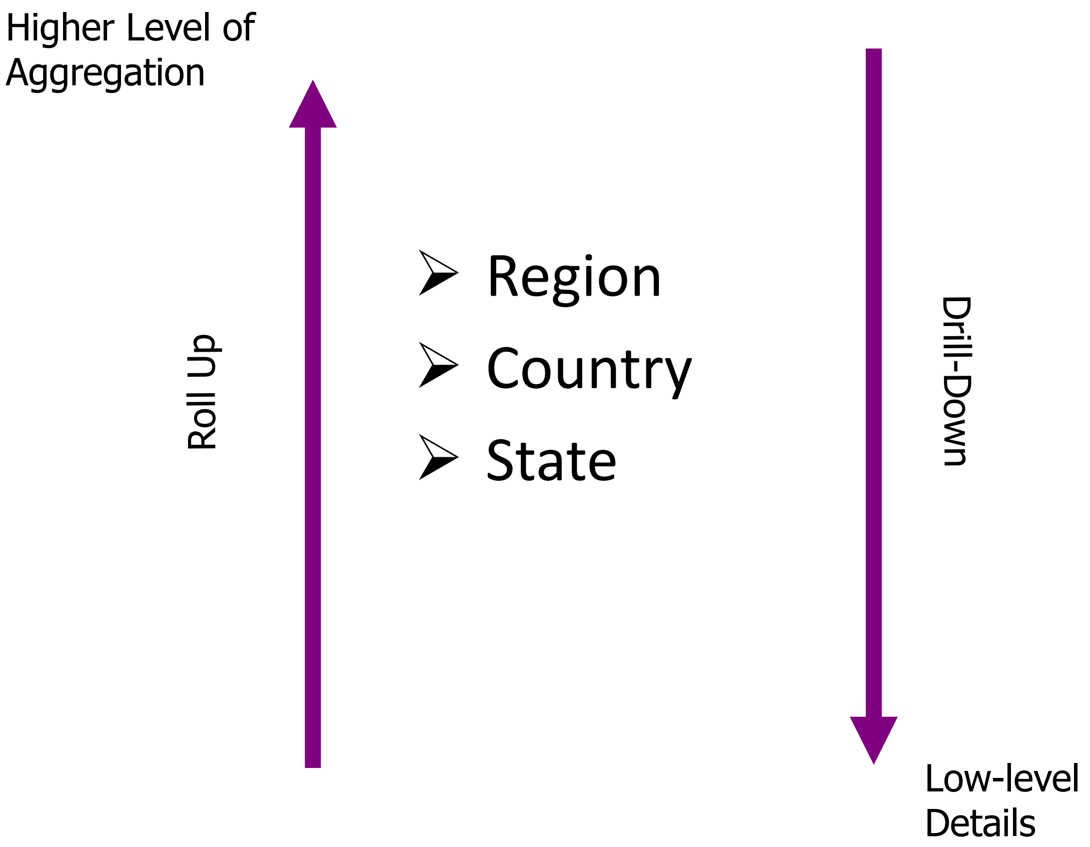


Image Copied from www.hypertextbookshop.com

Cube Terminology

Rollup and Rolldown



Higher Level of Aggregation

Roll Up

- ➢ Region
- ➢ Country
- ➢ State

Drill-Down

Low-level Details

Scope for BI/DW and Analytics

Descriptive Analytics

More of “What happened?”

Diagnostics Analytics

More of “Why it happened?”

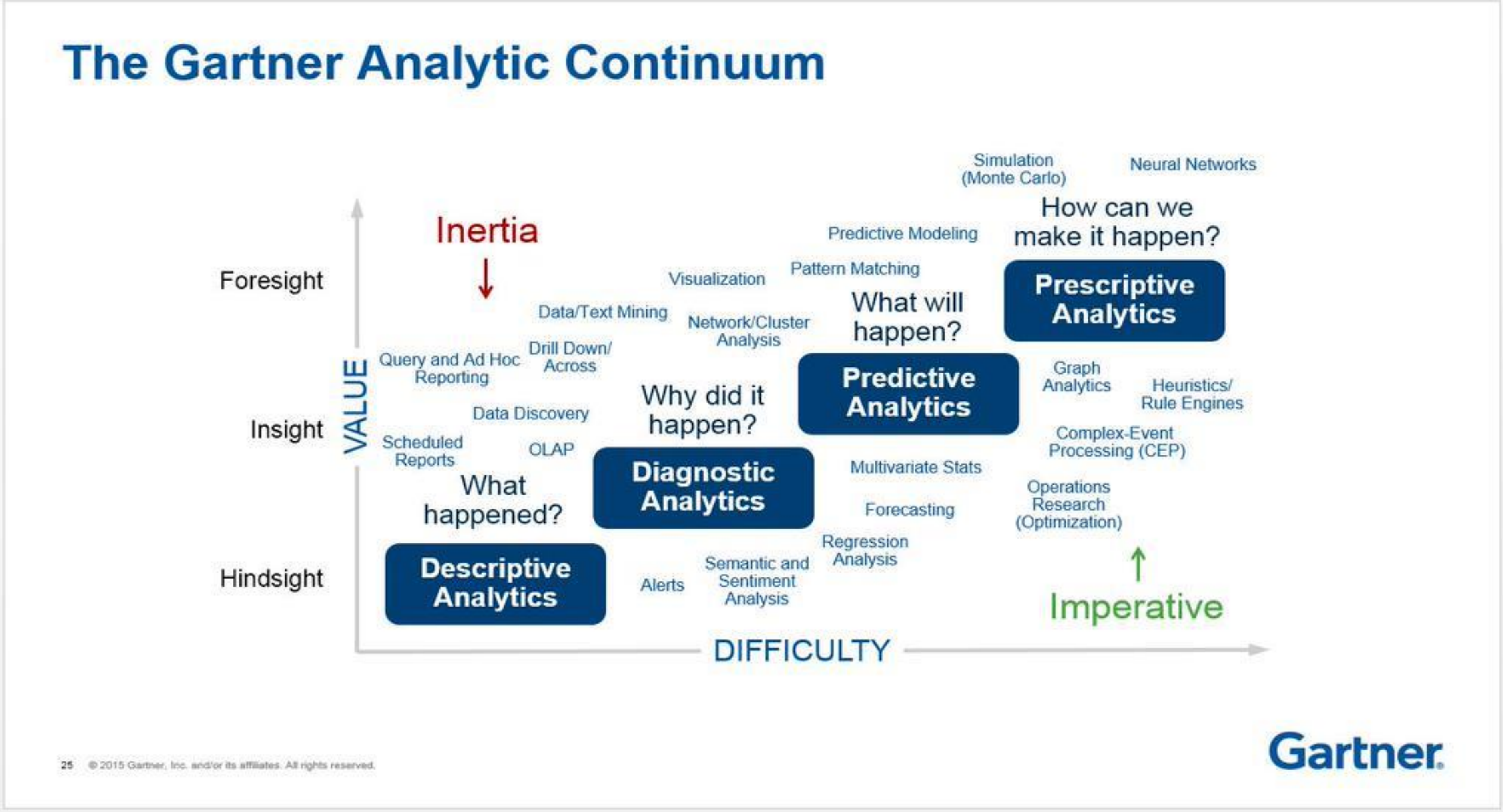
Predictive Analytics

More of “What will happen?”

Prescriptive Analytics

More of “What happens next?”

If parameters are tuned what can happen?



The Gartner Analytic Continuum

25 © 2015 Gartner, Inc. and/or its affiliates. All rights reserved.

Gartner.

Bibliography, Important Links

- Adventure Works Sample Data Warehouse
  http://technet.microsoft.com/en-us/library/ms124623%28v=sql.105%29.aspx

- Kimball Group | Dimensional Data Warehousing Experts
  www.**kimball**group.com/

- Books
  The Data Warehouse Toolkit, Third Edition: The Definitive Guide to Dimensional Modeling

  And Many more

Any Questions?

Thank you!