The Decoder Ring for Data Warehousing & Business Intelligence

Robert C. Cain, MVP
OWNER, ARCANE TRAINING AND CONSULTING

@arcanecode www.arcanecode.com

Your Presenter

Robert C. Cain, MVP, MCTS

Arcane Training and Consulting, LLC

Microsoft MVP since 2008

Pluralsight, Author

Co-Author 5 books

Author for Red Gate's Simple Talk site

Speaker, Pass Summit, SQL Saturdays, IT/Dev Connections, more

http://arcanecode.me

Agenda

Decoding DW/BI

- Data Warehousing Concepts
- Business Intelligence
- Design of a Data Warehouse

Why Learn About DW/BI?

DBA

Implement new Projects

Install BI Tools

DB Designer/dev

Design/script a DW

Different design from traditional databases

Software Developer

Interact with DW's

Data mining results into your apps

Business Users

Learn the terminology

Understand the sources

What Is a Data Warehouse?

A giant storehouse for your data

ALL of your data

Central repository, aggregate data from multiple systems

Single source of “the truth”

What Is Business Intelligence?

Leveraging data you already have to convert knowledge into informed actions

Providing ways to measure the health of your business

Examining the data in your warehouse to look for three main areas of interest

Areas of Interest

Aggregations

Trends

Correlations (Data Mining / Machine Learning)

Why Have a Data Warehouse?

- Combine data from multiple systems
- Resolve inconsistencies between systems
- Make reporting easier
- Reduce the load on production systems
- Provide for long term storage of data
- Provide consistency among system transitions

More Reasons for a Data Warehouse

Make data available for analysis

Ability to apply advanced analytic tools

Extract further value from your data

Business Intelligence!

Warehousing Methodologies

Bill Inmon – Corporate Information Factory (CIF)

Ralph Kimball – Kimball Method – Star Schema

Kimball method most widely used

What's Wrong with Reporting from a Transactional System

- OLTP – On Line Transaction Processing
- Designed for single record accesses
- Data is normalized
- Getting data can involve many joins
- Confusing for 'ad-hoc' reporting
- Slow, having an impact on the OLTP system

What's Different About a Data Warehouse?

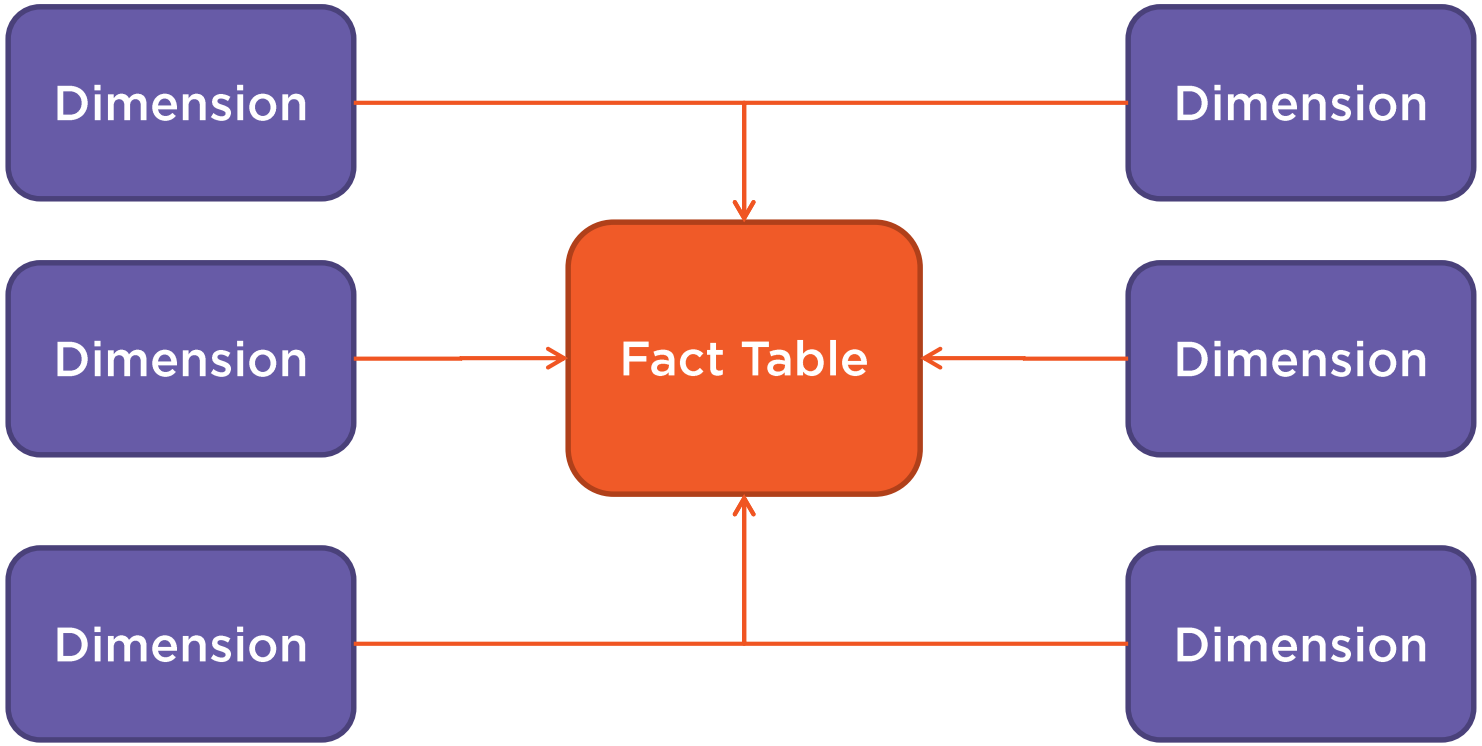
OLAP – Online Analytical Processing

Data is de-normalized

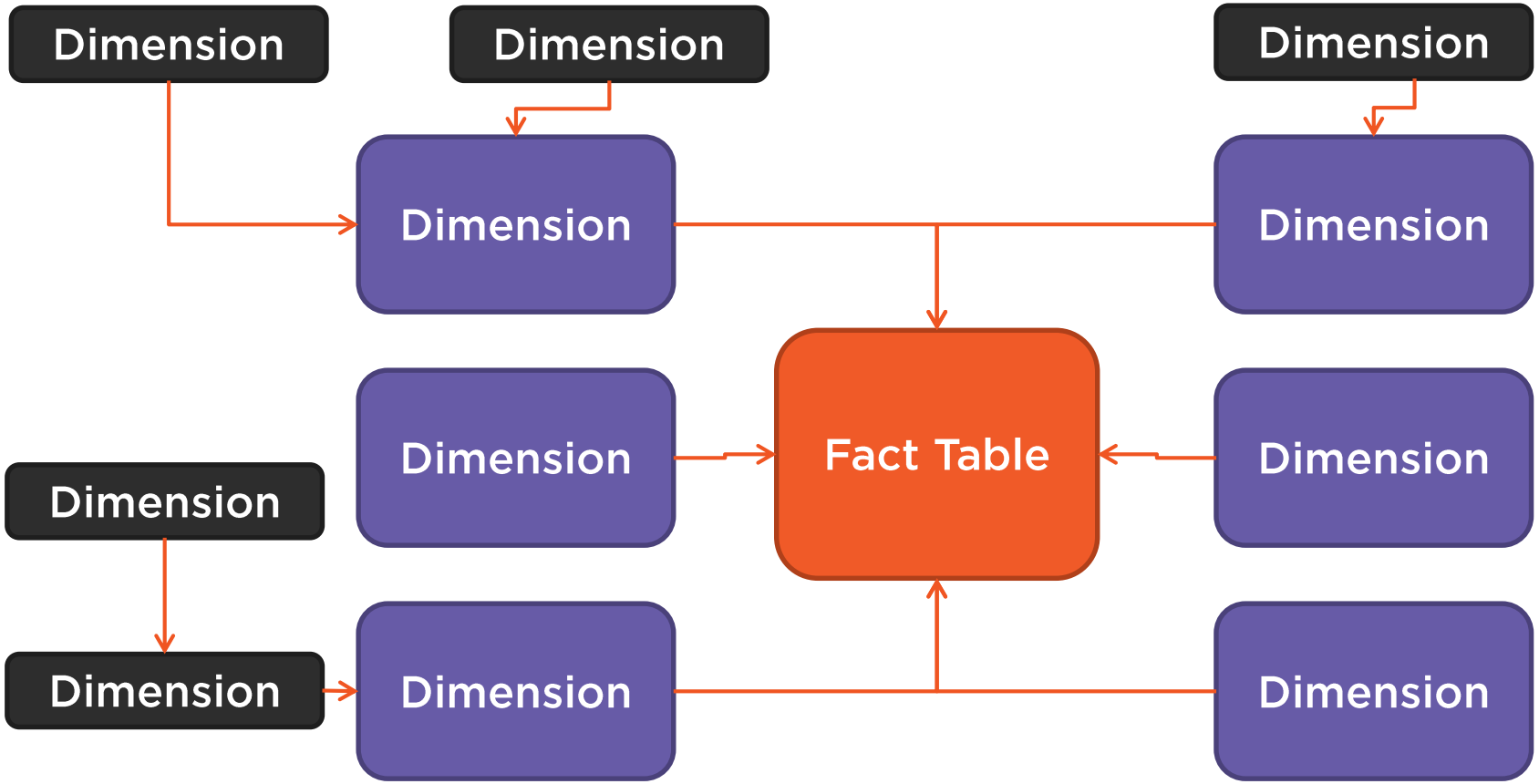
Number of tables is reduced

Star Schema or Snowflake Schema

Star Schema


Dimension
Dimension
Dimension
Fact Table
Dimension
Dimension
Dimension

Snowflake Schema



Types of Tables in a Data Warehouse

Facts

Dimensions

Surrogate Keys

A new key used as the Primary Key

A type if INT (Int, Big Int, Small Int, etc)

Reasons for Surrogate Keys

- Source system changes
- Multi-column source system tables
- Often the key isn't needed
- Combine data from multiple sources

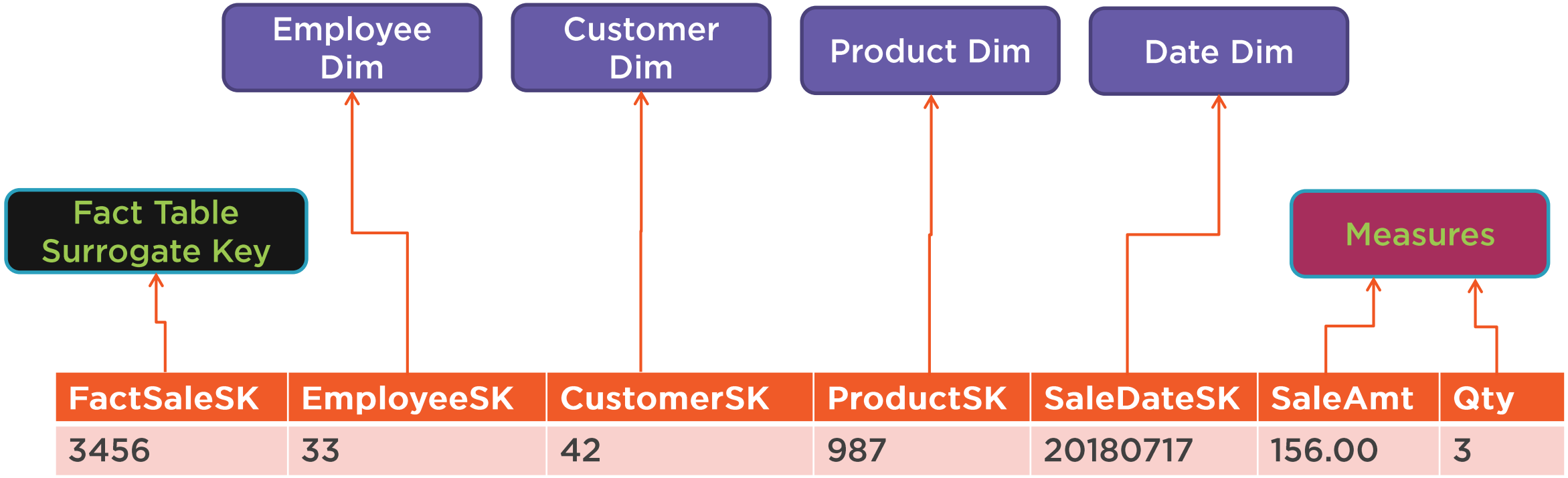
Conformed Dimensions

ProductSK	Name	InventoryBK	PurchasingBK	WorkMgtBK
9876	Widget	459684932	Wid45968	602X56VV1

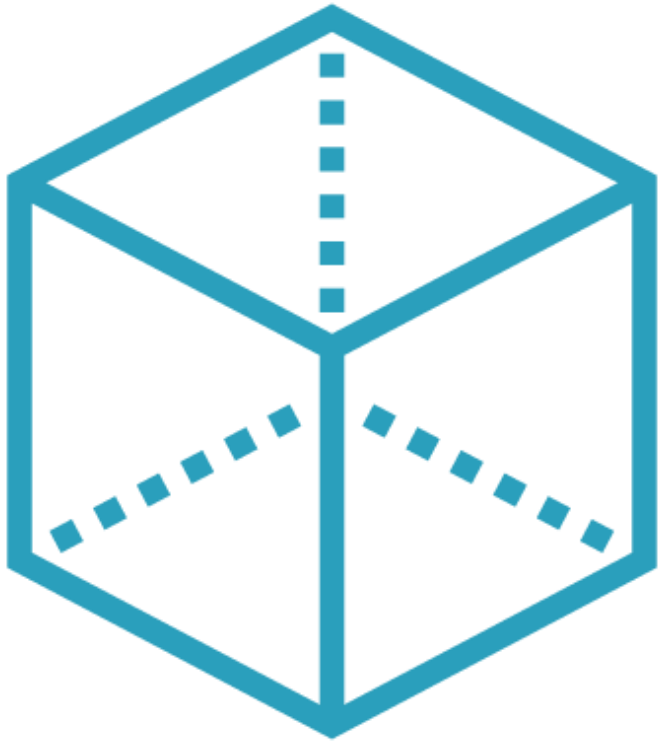
Fact Tables

- Fact tables mark an event
- Join dimensions, such as the who's and what's of an event
- Joins to the special date dimension
- Hold numeric, quantifiable values

Fact Table Example - Sale



Dimensions



Hold the values that describe facts

"Lookup tables"

Examples include geography, employees, product, customers, and time

Slowly Changing Dimension (SCD)

Many types of dimensions

Type 0 Dimension (Fixed)

- Type 0 Dimensions are also referred to as Fixed dimensions
- Used for static data like colors and sizes
- For data that will not change *ever*

Surrogate Key	Description
1	Blue
2	Black
3	Green
4	Yellow

Type 0 Dimension (Fixed)

- Type 0 Dimensions are also referred to as Fixed dimensions
- Used for static data like colors and sizes
- For data that will not change *ever*

Surrogate Key	Description
1	Blue
2	Black
3	Green
4	Yellow
5	Azure

Type 1 Dimension

When a value is updated, the old one is simply overwritten

Original Value

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Adam

New Value

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Clark

Impact of a Type 1 Change

Sales Report, 1st Quarter 2018 (as of March 14th)

Month	Salesperson	Sales Total
Jan 2018	Adam Curry	$33,259
Feb 2018	Adam Curry	$34,923
Mar 2018	Adam Curry	$14,823

Sales Report, 1st Quarter 2018 (as of March 31st)

Month	Salesperson	Sales Total
Jan 2018	Clark Curry	$33,259
Feb 2018	Clark Curry	$34,923
Mar 2018	Clark Curry	$35,672

Type 2 Dimension

When a value is updated, the old record is end dated and a new record is inserted

Original Value

EmployeeSK	EmployeeBK	Last	First	FromDate	ThruDate
33	PQ1894958	Curry	Adam	12/1/2015	<NULL>

New Value

SK	BusinessKey	Last	First	FromDate	ThruDate
42	PQ1894958	Curry	Clark	3/15/2018	<NULL>
33	PQ1894958	Curry	Adam	12/1/2015	3/14/2018

Impact of a Type 2 Change

Sales Report, 1st Quarter 2018 (as of March 14th)

Month	Salesperson	Sales Total
Jan 2018	Adam Curry	$33,259
Feb 2018	Adam Curry	$34,923
Mar 2018	Adam Curry	$14,823

Sales Report, 1st Quarter 2018 (as of March 31st)

Month	Salesperson	Sales Total
Jan 2018	Adam Curry	$33,259
Feb 2018	Adam Curry	$34,923
Mar 2018	Adam Curry	$14,823
Mar 2018	Clark Curry	$20,849

Importing Type 1 Data

Sales record from OLTP system

EmployeeID	SalesAmount	Other data
PQ1894958	156.00	

Employee Dim

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Adam

EmployeeSK	SalesAmount	Other data
33	156.00	

Sale fact record in OLAP Data Warehouse

Importing Type 2 Data

Sales record from OLTP system

EmployeeID	SalesAmount	Other data
PQ1894958	156.00	

Match on BK + ThruDate = NULL

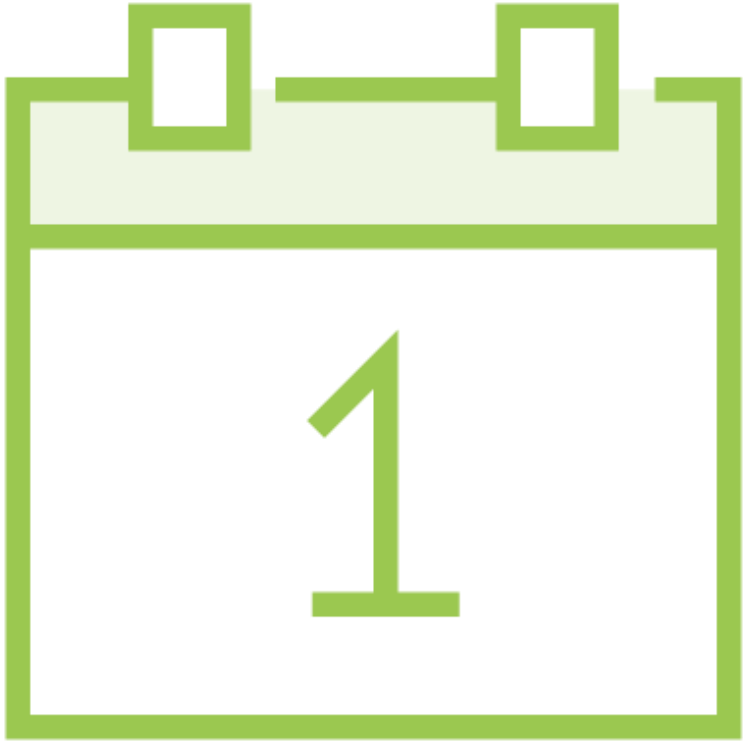
Employee Dim

EmployeeSK	EmployeeBK	Last	First	FromDate	ThruDate
33	PQ1894958	Curry	Adam	12/1/2015	3/14/2018
42	PQ1894958	Curry	Clark	3/15/2018	<NULL>

EmployeeSK	SalesAmount	Other data
42	156.00	

Sale fact record in OLAP Data Warehouse

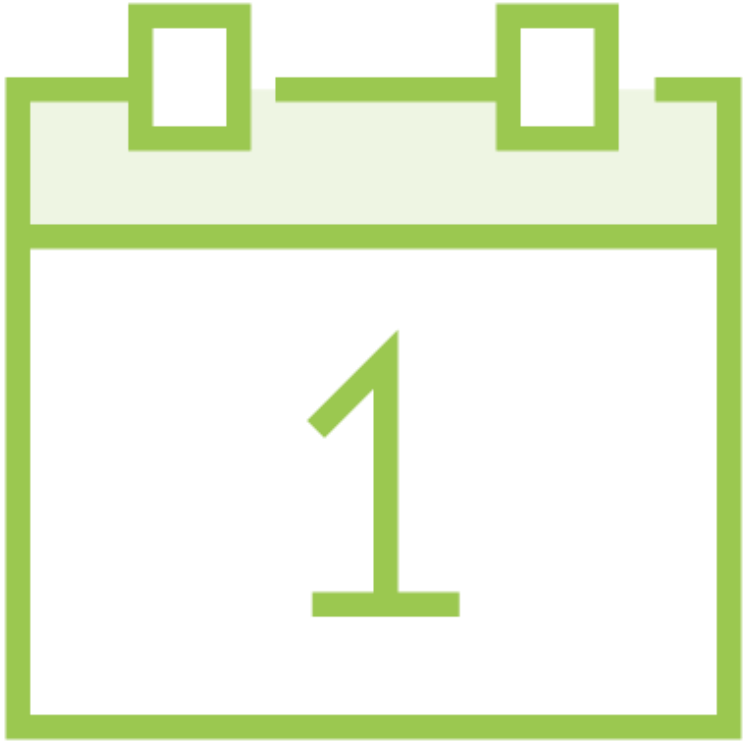
Considerations for dates in Type 2 Dimensions



Be consistent with names

- FromDate, ThruDate
- FromDate, ToDate
- From, To
- BeginDate, EndDate
- Begin, End

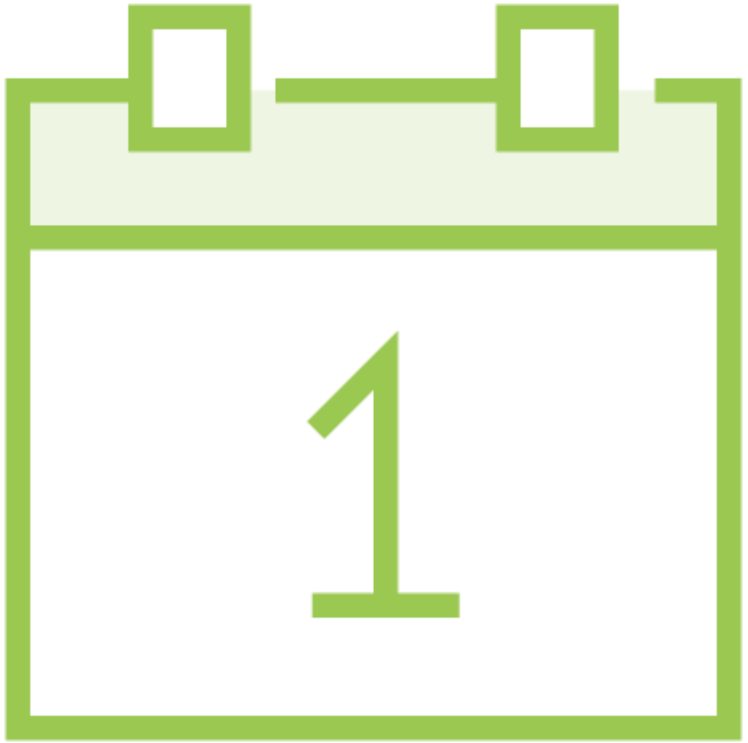
Considerations for dates in Type 2 Dimensions



Decide on the data type

- Date: 3/15/2018
- Int: 20180315
- Only for Type 2 date range
- Other dates store as INT

Considerations for dates in Type 2 Dimensions



Value for the open end date

- Null
- “Magic value”
- 12/31/9999
- 99991231

Type 3 Dimension

When a value is updated, records shift to a new column

Original Value

EmployeeSK	EmployeeBK	First1	First2	Last
33	PQ1894958	Adam		Curry

New Value

EmployeeSK	EmployeeBK	First1	First2	Last
33	PQ1894958	Clark	Adam	Curry

Almost never used – Type 3, the just say no dimension type

Type 4 Dimension

When a value is updated, old record copied to history and current record is updated

Original Value in

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Adam

New Value in DimEmployee

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Clark

New Value in DimEmployee_History

HistorySK	EmployeeSK	EmployeeBK	Last	First	FromDate	ThruDate
102	33	PQ1894958	Curry	Adam	12/1/2015	3/14/2018

Type 4 Dimension (Alternate Form)

When a value is updated, old record copied to history and current record is updated

Original Value in

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Adam

New Value in DimEmployee

EmployeeSK	EmployeeBK	Last	First
33	PQ1894958	Curry	Clark

New Value in DimEmployee_History

HistorySK	EmployeeSK	EmployeeBK	Last	First	FromDate	ThruDate
219	33	PQ1894958	Curry	Clark	3/15/2018	<NULL>
102	33	PQ1894958	Curry	Adam	12/1/2015	3/14/2018

Different Dimension Types in a Table

- Dimensional types are at the *column* level, not the row
- The business should be the ones to determine which data is significant enough to track changes on

Example

EmployeeSK	EmployeeBK	Last	First	Phone	FromDate	ThruDate
33	PQ1894958	Curry	Adam	555-1111	12/1/2015	<NULL>

- Phone Number = Type 1
- First Name = Type 2

Different Dimension Types in a Table

Original Value

EmployeeSK	EmployeeBK	Last	First	Phone	FromDate	ThruDate
33	PQ1894958	Curry	Adam	555-1111	12/1/2015	<NULL>

Update to Phone Number (Type 1)

EmployeeSK	EmployeeBK	Last	First	Phone	FromDate	ThruDate
33	PQ1894958	Curry	Adam	555-3342	12/1/2015	<NULL>

Update to First Name (Type 2)

EmployeeSK	EmployeeBK	Last	First	Phone	FromDate	ThruDate
33	PQ1894958	Curry	Adam	555-3342	12/1/2015	3/14/2018
42	PQ1894958	Curry	Clark	555-3342	3/15/2018	<NULL>

Dimensions in a Star Schema

SaleSK	EmployeeSK	CustomerSK	ProductSK	Qty	SaleAmt	SaleDateSK
3456	33	6789	987	3	156.00	20180312

Column	Value
ProductSK	987
ProductBK	SHBL4X
Description	Knit Shirt
Color	Blue
Size	4XL
Sleeve	Long

Repeating Values

ProductSK	ProductBK	Description	Color	Size	Sleeve
987	SHBL4X	Knit Shirt	Blue	4XL	Long
988	SHBL3X	Knit Shirt	Blue	3XL	Long
989	SHBL2X	Knit Shirt	Blue	2XL	Long
990	SHBL1X	Knit Shirt	Blue	1XL	Long
991	SHBLLG	Knit Shirt	Blue	LG	Long
992	SHBLMD	Knit Shirt	Blue	MD	Long
993	SHBLSM	Knit Shirt	Blue	SM	Long

Dimensions in a Snowflake Schema

SaleSK	EmployeeSK	CustomerSK	ProductSK	Qty	SaleAmt	SaleDateSK
3456	33	6789	987	3	156.00	20180312

Column	Value
ProductSK	987
ProductBK	SHBL4X
Description	Knit Shirt
Color	2
Size	7
Sleeve	2

ColorSK	Value
1	Red
2	Blue
3	Green
...	...

SizeSK	Value
6	3XL
7	4XL
8	5XL
...	...

LengthSK	Value
1	Short
2	Long

Flattening a Snowflake Schema to a Star

SaleSK	Emp...SK	Cust...SK	ProductSK	ColorSK	SizeSK	LengthSK	Measure Columns Here
3456	33	6789	987	2	7	2	

Column	Value
ProductSK	987
ProductBK	SHBL4X
Description	Knit Shirt

ColorSK	Value
1	Red
2	Blue
3	Green
...	...

SizeSK	Value
6	3XL
7	4XL
8	5XL
...	...

LengthSK	Value
1	Short
2	Long

Getting Data into a Data Warehouse

ETL

- Extract
- Transform
- Load

SSIS – SQL Server Integration Services

PowerShell

Custom Applications

Getting Data Out of Your Warehouse

Data Aggregation, Trending, Correlations

- SSAS – SQL Server Analysis Services
- Azure ML

Reporting

- SSRS – SQL Server Reporting Services
- Excel (and PowerPivot)
- PowerBI

Summary

Data Warehousing

Business Intelligence

Star and Snowflake Schemas

Fact and Dimension Tables

Getting data in and out of a Data Warehouse