### Data Warehouse Design & Develop



Masoud Mirzakhani Senior DW/ ETL/ BI Architect

### Microsoft SQL Server 2019 Design & Develop



## Masoud Mirzakhani Senior DW/ ETL/ BI Architect

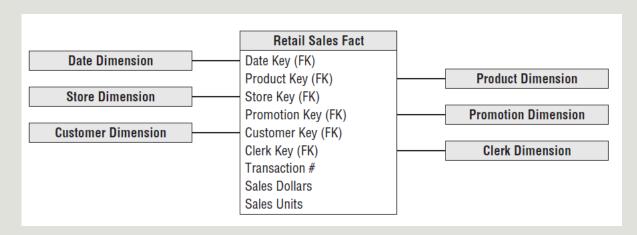
- Master of Science in Information Technology
- Bachelor of Science in Information Technology
- md.mirzakhani@gmail.com
- @MasoudMirzakhani
- linkedin.com/in/masoudmirzakhani





### Star Schema •

- ساده ترین معماری انباره داده.
  - بهترین کارایی را دارد.
- ارتباط تنها بین جداول Fact و Dimension برقرار می شود.





### Star Schema – طراحی اشتباه

FactSales					
DateID ProductID CustomerID Quantity Price					
13990701	45	10	5	1000	
13997702	78	14	9	7890	

FactEmployeeCount PactEmployeeCount PactEmployee					
DateID EmployeeID DepartmentID Count					
13990802	78	3	1		
13980802	79	3	1		

	DimCustomer					
ΙĎ	FirstName	LastName	LocationTitle			
10	Ali	Hasani	Tehran			
14	Omid	Rezaee	Shiraz			

	/ DimEmployee					
ď	FirstName	LastName	LocationTitle			
78	Behzad	Babaee	Tehran			
79	Farid	Lari	Shiraz			



### Star Schema – طراحی صحیح

FactSales Page 1					
DateID	ProductID	CustomerID	LocationID	Quantity	Price
13990701	45	19	11	5	1000
13997702	78	14	12	9	7890

FactEmployeeCount PactEmployeeCount PactEmployee				
DateID EmployeeID LocationID DepartmentID Count				
13990802	78	11	3	1
13980802	79	12	3	1

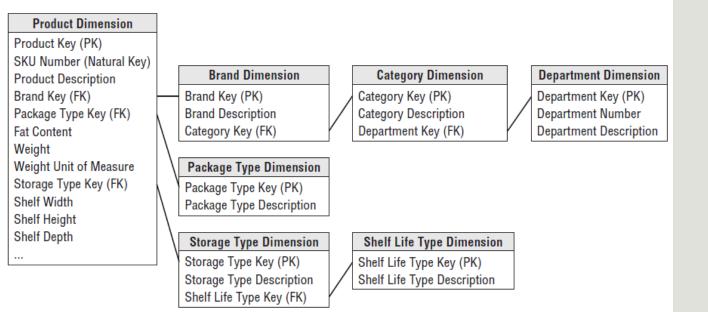
DimCustomer				
ID	FirstName	LastName		
10	Ali	Hasani		
14	Omid	Rezaee		

<b>DimLocation</b>			
ID	Name		
11	Tehran		
12	Shiraz		

DimEmployee					
ID FirstName LastName					
78	Behzad	Babaee			
79	Farid	Lari			



- Snowflake Schema •
- ارتباط بین جداول Dimension هم وجود دارد.
- همان Star Schema است که جداول Dimension کمی نرمال شده اند.
  - برای زمانی که بخواهیم فرایند ETL را کمی ساده تر کنیم.
    - برای سبک تر کردن Dimension های بزرگ
      - تنها در موارد خاص استفاده شود.





### SnowFlake Schema – طراحی صحیح (توصیه نمی کنم)

FactSales					
DateID	ProductID	CustomerID	Quantity	Price	
13990701	45	10	5	1000	
13997702	78	14	9	7890	

FactEmployeeCount					
DateID EmployeeID DepartmentID Count					
13990802	78	3 4	1		
13980802	79	3	1		

	DimCustomer					
ID	FirstName	LastName	LocationID			
10	Ali	Hasani	11			
14	Omid	Rezaee	12			

DimLocation	
N ID	Name
11	Tehran
12	Shiraz

١	DimEmployee				
_	ID	FirstName	LastName	LocationID	
	78	Behzad	Babaee	11	
	79	Farid	Lari	12	



- Additive, Semi-Additive, Non-Additive Facts
- Additive : اندازه هایی که در مقابل تمامی ابعاد، جمع پذیر باشند.
  - تعداد فروش
- Semi-Additive: اندازه هایی که در مقابل برخی ابعاد، جمع پذیر بوده و در مقابل دیگر ابعاد جمع پذیر نمی باشند...
  - مانده بانک
  - مانده انبار
  - Non-Additive: اندازه هایی که در مقابل هیچ کدام از ابعاد، جمع پذیر نمی باشند.
    - نسبت ها



### **NULL Values** •

- توابع SUM, AVG, MIN, MAX, COUNT نسبت به مقادیر SUM, AVG, MIN, MAX, COUNT
  - کلید های خارجی نبایستی حاوی مقادیر NULL باشند.



- Transaction Fact Tables •
- هر سطر نماینده یک اتفاق در فرایند متناظر است.
  - Periodic Snapshot Fact Tables •
- هر سطر نماینده اتفاق های فرایند در یک بازه زمانی خاص (روز، ماه، سال ... ) است.
  - هر سطر در جداول اسنپ شات می تواند معادل یک روز، ماه یا سال باشد.
- حتی اگر در بازه زمانی مورد نظر اتفاقی نیافتاده باشد، یک رکورد با اندازه های ۰ بایستی ثبت شود.
  - Accumulating Snapshot Fact Tables •
  - هر سطر نماینده اتفاق های فرایند از ابتدای زمان (مبدا) تا یک لحظه خاص است.



### Conformed Facts

• جداول فکتی که ابعاد و اندازه های شبیه به هم داشته باشند.

FactSales			
DateID	ProductID	Quantity	
13990101	1	10	
13990101	2	20	
13990101	3	30	

FactSalesBudget			
DateID	ProductID	Quantity	
13990101	1	15	
13990101	2	25	
13990101	3	35	



FactSalesAndBudget (Design 1)				
DateID	ProductID	Quantity	QuantityBudget	
13990101	1	10	15	
13990101	2	20	25	
13990101	3	30	35	

FactSalesAndBudget (Design 2)				
DateID ProductID		ValueType	Quantity	
13990101	13990101 1 Sales 10		10	
13990101 2		Sales	20	
13990101	3	Sales	30	
13990101	1	Budget	15	
13990101 2		budget	25	
13990101	3	budget	35	

### Consolidated Fact Tables •

- ترکیب کردن فکت های فرایند های مختلف در یک فکت خاص
  - مثال: فروش و پیش بینی فروش



### Factless Fact Tables

• مثال: حضور دانشجو با کلاس و تاریخ ارتباط دارد. ولی هیچ اندازه ای ندارد.



### Surrogate Keys •

- اجباری نیست.
- توصیه میشود که جداول فکت، دارای کلید های با مقادیر خودکار باشند.

### Centipede Fact Tables •

- هنگامی که جدول تعداد خیلی زیادی ارتباط با ابعاد داشته باشد.
- وقتی که سلسله مراتب یک بعد، به صورت ابعاد جداگانه طراحی شده باشد و هر یک جداگانه با جدول فکت ارتباط داشته باشند.
  - مثال: سال، فصل؛ ماه؛ روز
    - در کل توصیه نمی شود.



- ساختار کلی:
- تمامی جداول Dimension، دارای یک کلید اصلی هستند.
- جداول فکت از طریق FK با کلید اصلی فوق در ارتباط هستند.
  - اصولا دینرمال بوده و دارای ستون های زیادی هستند.
    - Denormalized Flattened Dimensions
      - دینرمال و مسطح بودن ابعاد توصیه می شود.
  - این ویژگی تاثیر مستقیم بر روی سرعت و سادگی ابعاد می گذارد.
    - از نرمال سازی ابعاد پرهیز کنید.
- کد ها و شناسه های عملیاتی را می توان به صورت ستون (ویژگی) در این جداول مشاهده کرد.
  - اکثر ستون ها از طریق کلمات و عبارات کوتاه پر شده اند.
  - از ستون های ابعاد به منظور گروه بندی و فیلتر در گزارش ها استفاده می شود.



### Dimension Surrogate Keys •

- کلید اصلی ابعاد، معمولا با کلید اصلی در سیستم اطلاعاتی متفاوت است.
  - در جاهایی که اطلاعات از بیش از یک منبع جمع آوری شده است
    - در جاهایی که اطلاعات ابعاد در طول زمان تغییر می کند.
- خوب است که مقادیر عددی ترتیبی جانشین کلید اصلی که از سیستم عملیاتی می آید بشود.

### Natural, Durable, and Supernatural Keys •

- کلید های طبیعی توسط سیستم عملیاتی تولید می شوند.
- مثال: اگر کارمندی از شرکت برود و مجدد استخدام بشود؛ کلید جدید در سیستم عملیاتی می گیرد.
  - کلید های مانا و فراطبیعی، در طول زمان تغییر نمی کنند.
- مثال: اگر کارمندی از شرکت برود و مجدد استخدام بشود؛ کلید مانا یا فراطبیعی آن کارمند در انباره داده نباید تغییر کند.



	DimEmployee				
ID	Code	FirstName	LastName	ExitDate	
1	1	Masoud	Mirzakhani	13971229	
2	2	Hosein	Azadi		
3	3	Omid	Rezae		
100	3	Masoud	Mirzakhani		



Product (Source 1)		
ID	Title	
1	Fanta	
2	Pepsi	
3	7up	

Product (Source 2)		
ID	Title	
56	Fanta	
57	Pepsi	
58	7up	

DimProduct (Design 1)		
ID	Title	
100001	Fanta	
100002	Pepsi	
100003	7up	
100056	Fanta	
100057	Pepsi	
100058	7up	

DimProduct (Design 2)			
ID	SourceID	BaseID	Title
1	1	1	Fanta
2	1	2	Pepsi
3	1	3	7up
4	2	56	Fanta
5	2	57	Pepsi
6	2	58	7up



Product (Source 1)		
ID	Title	
1	Fanta	
2	Pepsi	
3	7up	

Product (Source 2)		
ID	Title	
56	Fanta	
57	Pepsi	
58	7up	

<b>Product Reference</b>		
ID	Title	
10	Fanta	
20	Pepsi	
30	7up	

P	Product Map			
Source	BaseID	MappedID		
1	1	10		
1	2	20		
1	3	30		
2	56	10		
2	57	20		
2	58	30		

DimProduct (Design 3)			
ID	ID Code		
100001	10	Fanta	
100002 20		Pepsi	
100003	100003 30		
100056	100056 10		
100057	20	Pepsi	
100058 30		7up	



### Degenerate Dimensions •

- ابعادی که هیچ محتوایی به غیر از کلید اصلی خود ندارند.
- این ابعاد، در همان جداول فکت باقی می مانند؛ بدون این که ارتباطی با جدول بعدی داشته باشند.
  - مثال: شماره فاكتور

### Flags and Indicators •

- مخفف ها و نشانگرهای صحیح / نادرست را با کلمات متن کامل در جداول بعد قرار دهید.
  - Null Attributes •
  - توصیه می شود مقادیر گم شده را با عناوین توصیفی همچون "نا مشخص" پر کنید.



- Conformed Dimensions •
- ابعادی که ساختار یکسانی دارند.
- Role-Playing Dimensions •
- جدول بعدی که بیش از یک بار با جدول فکت ارتباط دارد.
- توصیه می شود که به ازای هر رابطه مضاعف یک ویو از جدول بعد اصلی ایجاد شود و رابطه فوق با ویوی ایجاد شده بر قرار شود.
  - به ویوهای فوق Role گفته می شود.
    - Junk Dimensions
  - بعد مستقلی است که برای نشان دادن مخفف ها و نشانگرهای صحیح / نادرست استفاده می شود.
    - Snowflaked Dimensions •
    - هنگامی که جدول بعد، نرمال شود.
      - Outrigger Dimensions •
    - در حالت دانه برفی، به بعدی گفته می شود که به آن اشاره می شود.



- Type 0: Retain Original
- مقادیر ستون ها ثابت است و تغییر نمی کند.

DimSex			
1 Female			
2	Male		
3	Unknown		



DimLocation				
ID ParentID Title				
1	Tehran			
2	1	Rey		
3	1	Karaj		

DimLocation					
ID ParentID Title					
1	Tehran				
2	2 1				
3	4	Karaj			
4		Alborz			

### Type 1: Overwrite

- مقادیر جدید؛ جاگزین مقادیر قبل می شوند.
- مثال: جابجایی کرج در تقسیمات کشوری.

FactEmployeeCount					
DateID EmployeeID LocationID Count					
13900101	12	3	1		
13910101	12	3	1		
13920101	12	3	1		
13930101	12	3	1		



### Type 2: Add New Row

- با تغییر مقادیر، یک رکورد جدید به بعد اضافه می شود.
- از کلید های مانا و فراطبیعی به منظور نگهداری ارتباط بین مقادیر قدیم و جدید استفاده می شود.
  - برای پشتیبانی از این حالت حداقل نیاز است که ستون های زیر به جدول بعد اضافه شود:
    - از تاریخ
    - تا تاریخ
    - آخرین مقدار؟



	DimLocation						
ID	ID SNID ParentID Title FromDate				ToDate	IsLast?	
1	1 1 Tehran NULL		NULL	TRUE			
2	2 2 1 Rey		NULL	NULL	TRUE		
3	3	1	Karaj	NULL	13901230	FALSE	
4 4 Alborz		13910101	NULL	TRUE			
5	3	4	Karaj	13910101	NULL	TRUE	

Type 2: Add New Row

FactEmployeeCount				
DateID	Count			
13890101	12	3	1	
13900101	12	3	1	
13910101	12	5	1	
13920101	12	5	1	



### Type 3: Add New Attribute •

• به ازای هر تغییر، یک ستون به بعد اضافه می شود تا مقادیر پیشین را نگهداری کند.

• این حالت به ندرت استفاده می شود.

• از این حالت به نام alternate reality نیز معروف است.

DimLocation						
ID	ID ParentID1 ParentID2 Title					
1			Tehran			
2 1		1	Rey			
3 1		4	Karaj			
4			Alborz			

## SCD & Star Schema



FactEmployeeCount PactEmployeeCount PactEmployee				
DateID	EmployeeID	EducationLevelID,	Count	
13930101	5	1	1	
13940101	5	1	1	
13950101	5	1	1	
13960101	5	1	1	
13970101	5	2	1	
13980101	5	2	1	
13990101	5	3	1	

	DimEmployee													
ID	FirstName	LastName												
4	Behzad	Babaee												
5	Farid	Lari												
6	Omid	Kaji												

<b>DimEducationLeve</b>												
Name												
Diploma												
Bachelor												
Master												
PhD												

## SCD (Type1) & Snow Flake Schema



FactEmployeeCount													
DateID	EmployeeID	Count											
13930101	5	1											
13940101	5	1											
13950101	5	1											
13960101	5	1											
13970101	5	1											
13980101	5	1											
13990101	5	1											

	DimEmployee (Type 1)														
ID <sup>4</sup>	FirstName	LastName	EducationLevelID												
4	Behzad	Babaee	2												
5	Farid	Lari	3												
6	Omid	Kaji	2												

<b>DimEduca</b>	tionLevel
ID	Name
1	Diploma
2	Bachelor
3	Master
4	PhD

# SCD (Type2) & Snow Flake Schema



FactEmployeeCount													
DateID	<b>EmployeeID</b>	Count											
13930101	5	1											
13940101	5	1											
13950101	5	1											
13960101	5	1											
13970101	100	1											
13980101	100	1											
13990101	200	1											

	DimEmployee (Type 2)														
-	Ď	SNID	<b>FirstName</b>	LastName	Edi	ucationLevelID	<b>FromDate</b>	ToDate	IsLast						
	4	4	Behzad	Babaee		2									
	5	5	Farid	Lari	7	1	NULL	13961230	False						
	6	6	Omid	Kaji	_/	2		<b>A</b>							
	100	5	Farid	Lari		2	13970101	13981230	False						
	200	5	Farid	Lari		3	13990101	NULL	True						

<b>DimEducationLevel</b>									
O	Name								
1	Diploma								
2	Bachelor								
3	Master								
4	PhD								

### Dimension Hierarchies



- ابعاد سلسله مراتبی با عمق ثابت:
- بهتر است که سطوح سلسله مراتب تبدیل به ستون های بعد شود.
  - مثال: ساختار حساب در حسابداری
  - ابعاد سلسله مراتبی با عمق متغیر:
- اگر سطوح سلسله مراتب متغیر است ولی دارای محدودیت می باشد؛ می توان همانند روش قبل آن ها را مسطح کرد.
  - مثال: مناطق جغرافيايي

## یادآوری



#### **COMMON DIMENSIONS** Warehouse Promotion Customer Етріоуее Product Date **BUSINESS PROCESSES** Issue Purchase Orders Χ Χ Receive Warehouse Deliveries Χ Χ Warehouse Inventory Receive Store Deliveries Χ Χ Χ Χ Store Inventory Χ Χ Retail Sales Χ Χ Χ Χ Χ Χ Retail Sales Forecast Χ X Χ **Retail Promotion Tracking** Χ Χ Χ **Customer Returns** Χ Returns to Vendor Χ Χ Χ Χ Frequent Shopper Sign-Ups Χ

#### **Bus Matrix**

ابزار طراحی انباره داده

سطر ها مشخص کننده فرایند های کسب و کار هستند.

سطر ها مشخص کننده جداول Fact هستند.

ستون ها مشخص کننده جداول Dimension هستند.

## ساخت انباره داده YourCompanyDW



• انباره داده بر اساس Bus Matrix ساخته می شود.

Ite				ype	ocument	DimRpaOperationType	tType	DimReceiptChequeState		ntType	DimPaymentChequeState	ment	JPart (RP)	P)		DimInventoryReceiptType	DimInventoryDocumentType	ype	ance	DimEmploymentType	ionDegree		Location (RP)	DimDeliveryLocation (RP)		cy	enter	ctType	DimChequeGuaranteeType	Duration		itionType	0		scount		DimAccountingVoucherType	DimAccountingVoucherState	ıt	
DimMorkSite	DimSystem	DimStock	DimSex	DimSalesType	DimSalesDocument	DimRpaOp	DimReceiptType	DimReceip	DimPerson	DimPaymentType	DimPayme	DimPayElement	DimRelatedPart (RP)	DimPart (RP)	DimJob	DimInvent	DimInvent	DimHiringType	DimExperience	DimEmplo	DimEducationDegree	DimDetailedAccount	DimPersonLocation	DimDelive	DimDate	DimCurrency	DimCostCenter	DimContractType	DimChequ	DimChequeDuration	DimCash	DimCalculationType	DimBillType	DimBank	DimBankAccount	DimAge	DimAccou	DimAccou	DimAccount	
	*																					*			*	*											*	*	*	FactAccountingVoucher
																									*									*	*					FactBankAccountBalance
																									*						*									FactCashBalance
*			*						*						*				*	*	*				*		*	*								*				FactEmployeeCount
*			*						*						*			*	*	*	*				*		*	*								*				FactHiringCount
		*												*		*	*								*															FactInventoryTransaction
						*				*	*											*			*	*			*	*	*		*	*	*					FactPaymentCheque
										*												*			*	*					*		*	*	*					FactPaymentDraft
						*	*	*														*			*	*			*		*		*	*	*					FactReceiptCheque
							*															*			*	*					*		*	*	*					FactReceiptDraft
*			*						*			*			*				*	*	*				*		*	*				*				*				FactSalaryPaid
		*		*	*				*					*									*	*	*	*														FactSales
									*				*	*									*	*	*															FactSalesBasket

## پیاده سازی انباره داده



- انباره داده بر اساس Bus Matrix ساخته می شود.
- جداول Fact به طور کامل از روی Bus Matrix قابل ساخت هستند.
- برای پیدا کردن attribute های ابعاد، نیاز است که اطلاعات تکمیلی به Bus Matrix اضافه می شود.

## پیاده سازی انباره داده



- Fact
- ستون **ID**
- Primary Key •
- Auto Increment
  - Integer •
  - ستون ها ابعادی
- DimensionName + ID
  - ستون ها عددی
  - نوع داده عددی
- هیچ ستونی مرتبط با بعد، nullable نیست.
- برای measure ها مقدار پیش فرض در نظر گرفته شود.
- برای DimensionID ها، کلید پیش فرض در نظر گرفته شود.
  - مثال: شناسه: -١؛ عنوان: نامشخص





### Dimension

- ستون ID
- DimensionName + ID
  - از منبع اطلاعاتی می آید.
- هیچ ستونی nullable نیست.
- عبارت های پیش فرض لحاظ شود.
  - از flag پرهيز شود.
  - عبارات معنادار جایگزین گردد.
- حتما رکورد برای پوشش حالت های نامشخص وجود داشته باشد.
  - مثال: شناسه: -١؛ عنوان: نامشخص



## پیاده سازی انباره داده

### Foreign Key

- بین جداول Fact و Dimension، کلید های خارجی ساخته شود.
  - کلیدهای خارجی فوق، غیر فعال گردند.