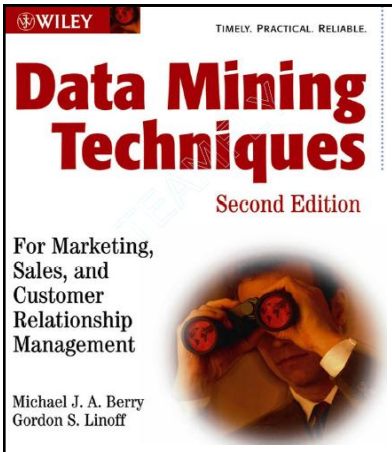
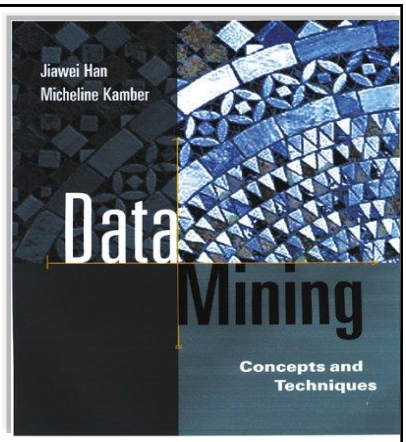




داده کاوی

سمیه علیزاده

مطالب مورد بحث در کلاس

- تعاریف و مفاهیم
- انباره داده ها
- آماده سازی داده ها (پیش پردازش داده ها)

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مطالب مورد بحث در کلاس

- خوشه بندی
- دسته بندی
- قوانین انجمنی
- سریهای زمانی
- وب کاوی
- متن کاوی
- پیوندکاوی و تحلیل شبکه های اجتماعی

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مطالب مورد بحث در کلاس

- متدولوژی اجرای پروژه های داده کاوی
- کاربردهای داده کاوی در بازاریابی
- کاربردهای داده کاوی در مدیریت ارتباط با مشتری
- امنیت در داده کاوی

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مبانی انباره داده ها

- تاریخچه و تعاریف و مفاهیم
- ویژگیهای انباره داده ها
- ساختار انباره داده ها

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مبانی انباره داده ها

- تاریخچه و تعاریف و مفاهیم
- ویژگیهای انباره داده ها
- ساختار انباره داده ها

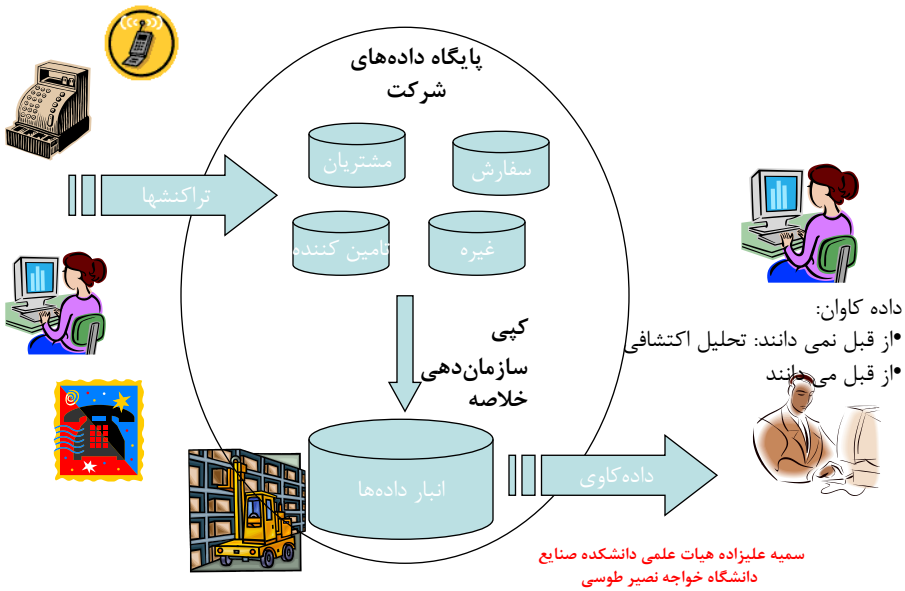
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تاریخچه

- در دهه ۹۰ میلادی پدیده انبار داده ها ظهور یافت . قبل از انبار سازی داده ها سیستم های کامپیوتری جهت ذخیره ، جمع آوری ، تغییر و تصحیح بیت های داده ها طراحی شده بودند. این سیستم های اولیه به سیستم های عملیاتی یا میراثی موسوم هستند .

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انبار داده‌ها



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انبار داده یک ضرورت کامل برای داده کاوی مؤثر می باشد .

- انبار داده ها ، داده های خام را برای تحلیل به گونه ای بهینه آماده می سازند . این آماده سازی به شکلهای مختلف سودمند می باشد .
- یکی از ماهیت های وجودی انبار داده ها ، یکپارچگی داده ها هنگام قرار گیری در انبار داده هاست . این بدین معنی است که دقت بسیاری بکار گرفته می شود تا یکنواختی و پیوستگی در درک اهداف عام سازمانی بوجود آید .
- اگر انبار داده ها وجود نداشته باشد ، داده کاو بایستی زمان بسیار زیادی را برای جمع آوری ، پاکسازی ، تمیز کردن و یکپارچه سازی داده ها سپری کند . بدین ترتیب وقت بسیاری باید صرف کرد تا بتوان کار تحلیل داده ها را آغاز نمود .

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انبار داده یک ضرورت کامل برای داده کاوی مؤثر می باشد .

- دومین علتی که می توان برای موفقیت داده کاوی به خاطر وجود انبار داده ها ذکر کرد این است که در انبار داده ها ، داده های تاریخی جمع آوری و سازماندهی می شوند .
- وجود داده های تاریخی برای یافتن الگوها و روابطی که سازمان بدنبال آنهاست ، برای داده کاوی یک ضرورت است .
- اگر چنانچه این داده های تاریخی وجود نداشته باشند ، داده کاو بایستی بدنبال جمع آوری آنها باشد .

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انبار داده یک ضرورت کامل برای داده کاوی مؤثر می باشد .

- علت سوم اهمیت انبار داده ها برای داده کاوی مؤثر، این است که انبار داده ها شامل داده های جزئی و داده های کلی، در کنار یکدیگر می باشد .
- بدون تردید ، داده کاو به اطلاعات جزئی برای تحلیل نیازمند است، اما داده های خلاصه شده نیز بکار می آیند .
- هنگامی که نمونه ای از انواع داده های خلاصه شده وجود داشته باشد ، داده کاو می تواند به سرعت بررسی کند که چه چیزی در انبار داده ها هست ، یا چیزی نیست و این باعث کاهش تکرار تحلیل ها توسط داده کاو می شود .

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مبانی انباره داده ها

- تاریخچه و تعاریف و مفاهیم
- ویژگیهای انباره داده ها
- ساختار انباره داده ها

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

ویژگیهای انباره داده ها

- موضوع محوری (Subject Oriented)
- جامعیت (Integrated)
- مهم بودن عامل زمان (Time Variant)
- غیر فرار و دایمی بودن (NON Volatile)

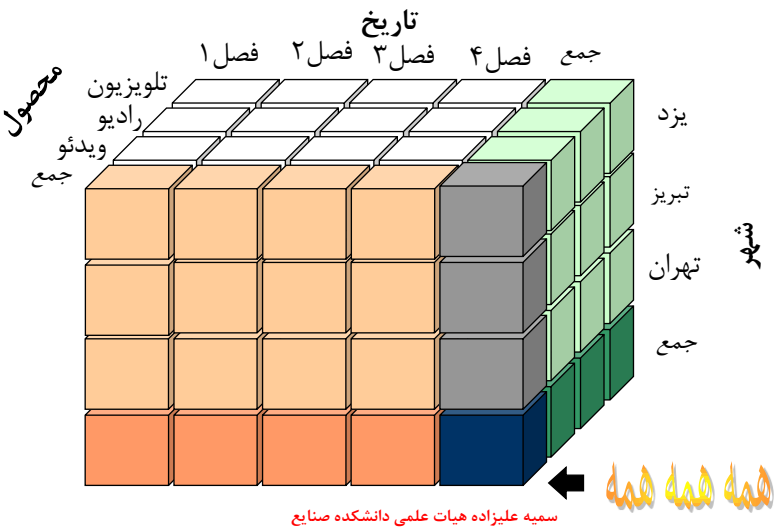
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

موضوع محوری

- داده ها طبق یک موضوع خاصی سازماندهی میشوند، به عنوان مثال داده های مربوط به مشتریان ، محصولات ویا داده های مرتبط با فروش هر کدام جداگانه در نظر گرفته میشوند. اما در پایگاه داده های معمولی، داده ها بر اساس عملیات و پردازش های روزانه ایجاد میگردند و موضوع آنها مرتبط با کل پردازش میباشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

خلاصه کردن داده‌ها



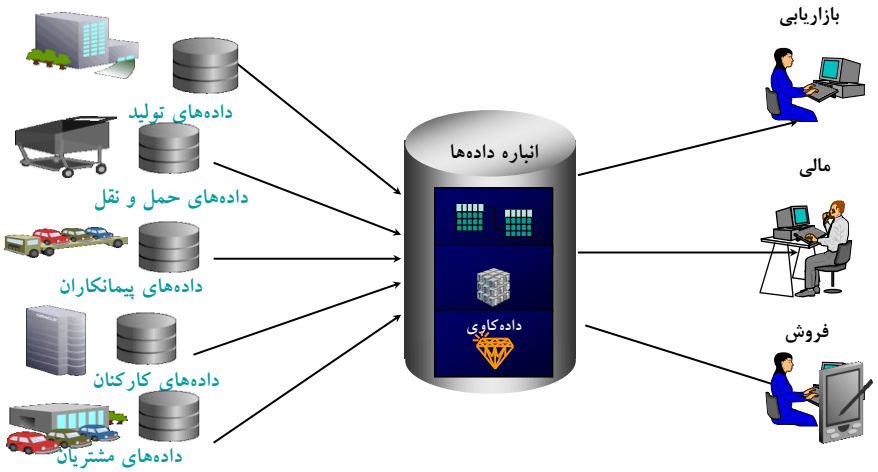
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

جامعیت

- داده های انباره های داده ای از تجمع دیگر داده ها ساخته میشوند. این داده ها ممکن است مربوط به پایگاه داده های رابطه ای، فایل های بدون ساختار ویا رکوردهای مرتبط با پردازش های Online باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مجتمع کردن داده‌ها



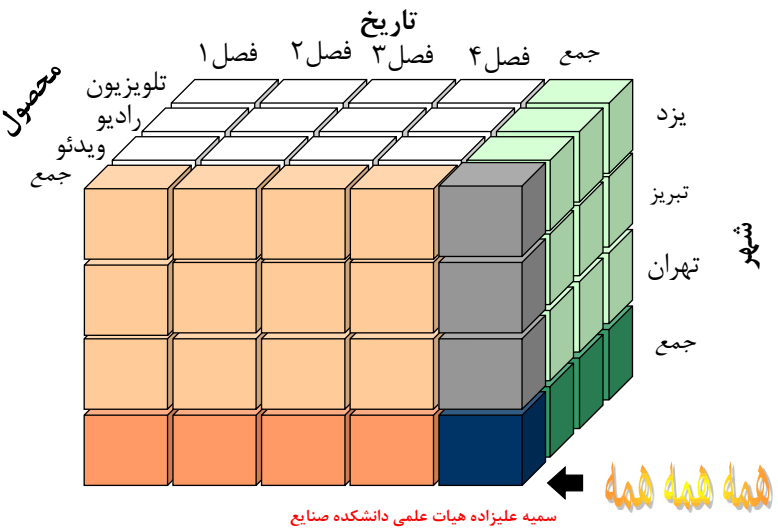
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

عامل زمان

- افق زمانی برای انباره داده ها بسیار مهمتر ازداده های مرتبط با سیستمهای عملیاتی میباشد. در ساختار انباره داده ها عاملی به نام زمان در نظر گرفته میشود. این عامل زمانی میتواند بصورت یک عامل ضمنی و یا به وضوح بیان گردد. اما در سیستمهای عملیاتی عامل زمان عاملی کلیدی نیست.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

خلاصه کردن داده‌ها



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

غیر فرار بودن

- پایگاه داده ها شامل داده هایی است که روزانه با آنها کار میشود و بخشهایی به آن اضافه و یا از آن حذف میگردد، در مقابل انباره داده این ویژگی را ندارد.

با توجه به همین امر واضح است که به روز شدن داده ها در انباره داده ها مقدور نمیباشد انباره داده ها نیازی به پردازشهایی از قبیل : تراکنشهای داده ای، بازیافت و مکانیزمهای کنترل همزمان ندارد. تنها اعمالی که در انباره داده ها صورت میپذیرد عبارتند از: بار گذاری (مقدار دهی) اولیه داده ها و دسترسی به داده ها.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

منافع اصلی انباره داده ها

- منافع اصلی از قابلیت های انباره داده عبارتند از :
  - به سرعت قابل دسترسی هستند ، زیرا از لحاظ فیزیکی در یک مکان قرار دارند .
  - بوسیله کاربر نهایی به راحتی قابل حصول هستند .

مبانی انباره داده ها

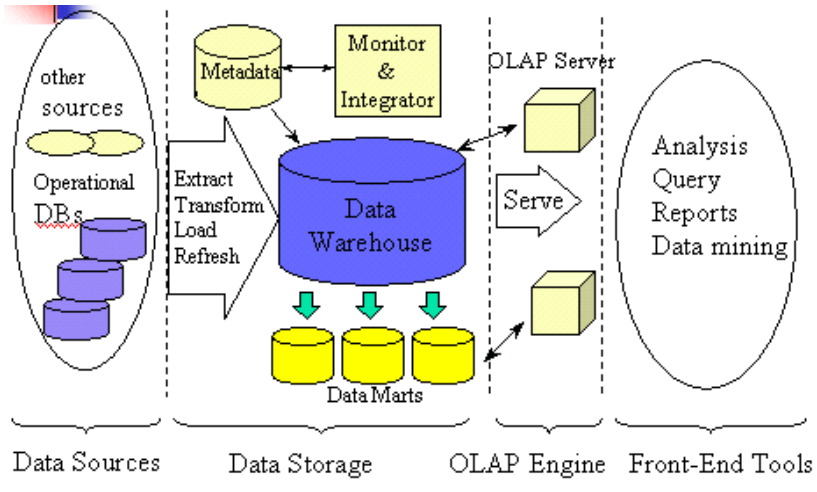
- تاریخچه و تعاریف و مفاهیم
- ویژگیهای انباره داده ها
- ساختار و معماری انباره داده ها

معماری انباره داده

- جهت تهیه و کار با انباره داده ها یک ساختار چند لایه ای وجود دارد که در شکل بعد نشان داده شده است. در ابتدا با چهار عمل اصلی Extract, Transform,Load,Refresh داده ها از پایگاه داده های معمولی (Data Source) که از پردازش های عملیاتی سیستمها ایجاد شده اند، جمع آوری شده و به انباره داده فرستاده میشود.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

معماری انباره داده



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تعاریف

- داده های موجود در انباره داده ها مستقیما با دیگر سیستمها در ارتباط نیستند و اگر یک سیستم عملیاتی بخواهد از داده های انباره داده استفاده کند از Data Mart ها که به طور موقت شامل بخشی از داده ها را که مرتبط با آن نوع سیستم خاص میباشند، استفاده میکند.
- متاداده ها ، داده های مربوط به داده ها هستند . متاداده ها وضعیت های مختلف داده ها را توصیف می کنند

تعریف (Data Mart)

- انباره داده کوچک محدود شده در ابعاد

Mini-warehouses, limited in scope

– یک Data Mart یک زیر مجموعه کوچک و محدود از یکDW می باشد که دارای مجموعه ای از موضوعات برای یک بخش از کاربران است.

پردازشها در انباره داده ها

- انباره داده ها نیازی به پردازشهایی از قبیل : تراکنشهای داده ای، بازیافت و مکانیزمهای کنترل همزمان ندارد. تنها اعمالی که در انباره داده ها صورت میپذیرد عبارتند از: بار گذاری (مقدار دهی) اولیه داده ها و دسترسی به داده ها.

- پردازشی که بر روی داده های انباره داده ها انجام میگیرد OLAP نامیده میشود (OLAP در مقابل OLTP که پردازش هایی است که بر روی داده های پایگاه داده ها انجام میگیرد آمده است). جدول بعد OLAP و OLTPرا با توجه به برخی پارامترهای مهم مقایسه کرده است:

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مقایسه پردازشها در داده‌کاوی با داده‌های معمولی

ویژگیها	OLTP	OLAP
کاربران	اپراتورها	کارشناسان خبره
کارکرد	کارهای روزمره	تصمیم گیری
طراحی پایگاه داده	بر مبنای کاربرد	بر مبنای موضوع
داده	جاری، روزبه-روز و با جزییات	تاریخی، چند بعدی، مجتمع
کاربرد	تکراری	در موارد خاص
نحوه دسترسی	خواندن ونوشتن	کاوش و کشف
واحد کاری	پردازش ساده	جستجوهای پیچیده
تعداد رکوردها	دهها	میلیونها
تعداد کاربران	هزاران	صدها
اندازه پایگاه داده	۱۰۰ MB_GB	۱۰۰ TB_GB
شاخص	پردازش	جستجو و پاسخ

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

ایجاد و نگهداری یک انباره داده ها

انباره های داده به چندین ابزار نیاز دارند تا وظایف ذیل را خودکار کرده یا مورد پشتیبانی قرار دهند:

- استخراج داده ها (Data extraction)از منابع داده متفاوت خارجی
- پالایش داده ها (Data cleaning): پیدا کردن و حل ناسازگاریها از داده های منبع
- یکپارچگی و تبدیل داده ها : بین فرمتها و زبانهای مختلف

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

ایجاد و نگهداری یک انباره داده ها

- بارگیری داده ها(Data loading) : بارگیری داده ها در انباره داده ها
- تکرار داده ها (Data replication): از پایگاه داده منبع به انباره داده ها
- تازه کردن داده ها(Data refreshment)
- آرشیو داده ها(Data archiving)
- چک کردن کیفیت داده ها
- تحلیل متا داده ها(metadata)

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مدلسازی مفهومی انباره داده ها

سه شمای مفهومی پایه عبارتند از :

- شمای ستاره (Star schema)
- شمای دانه برفی(Snowflake schema)
- شمای فلکی (Fact constellations)

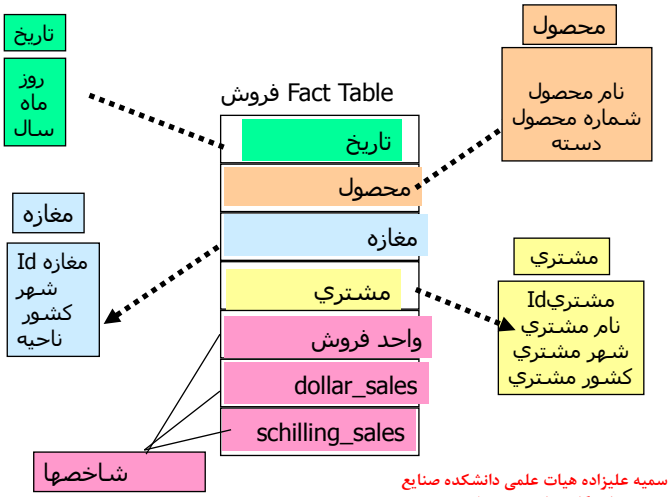
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شمای ستاره

•شمای ستاره:یک fact table در مرکز به تعدادی از فهرستهای ابعادی(dimension tables) متصل می شود .

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مثالي از شماي ستاره



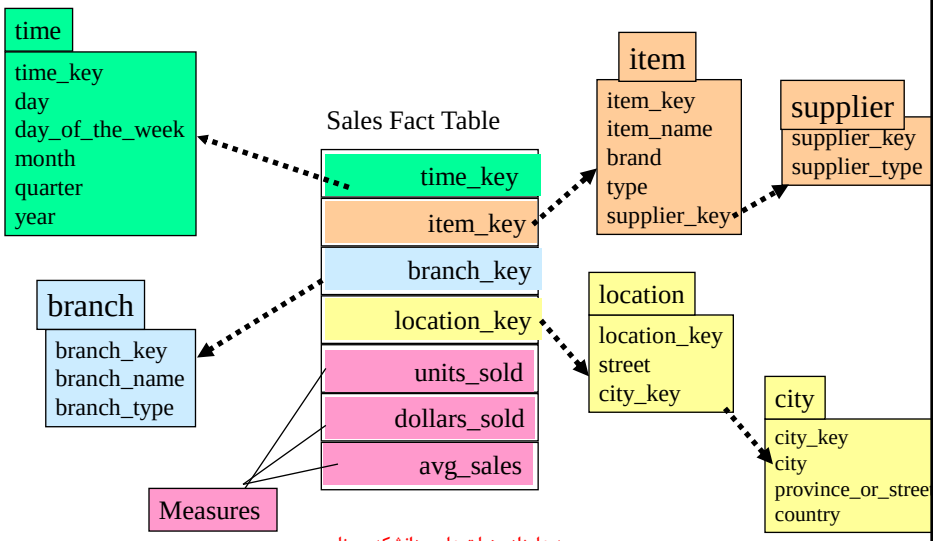
سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

شماي دانه برف

شمای دانه برف : یک پالایشی از شمای ستاره است که سلسله مراتب ابعاد بصورت صریح با نرمال کردن فهرستهای ابعادی(dimension tables) نمایش داده می شود .

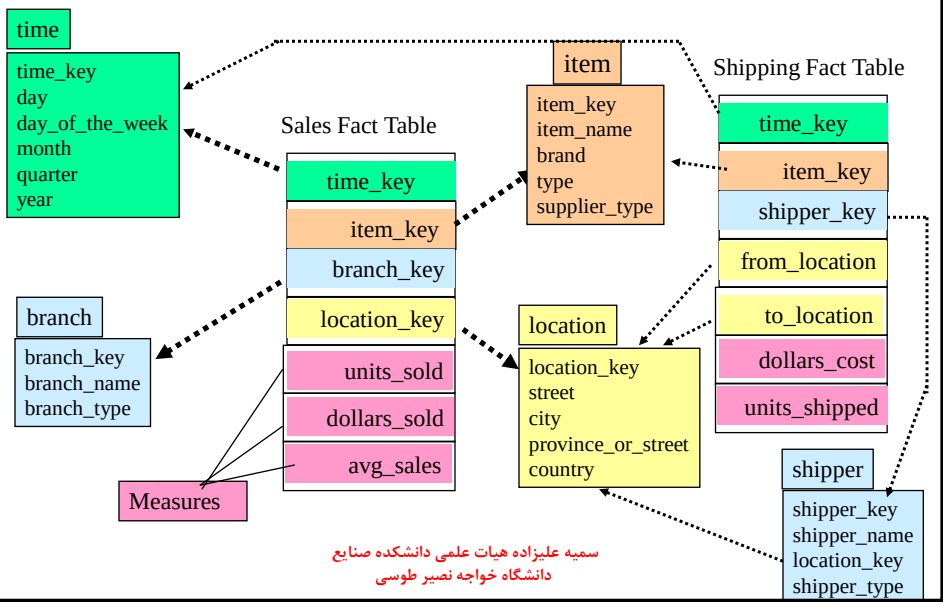
سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

مثالی از شمای دانه برف



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مثالی از شمای صورت فلکی



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی