نرمال سازی

این فصل مروری بر تئوری نرمالسازی پایگاه داده ها می باشد و روی چهار فرم اول از هفت فرم شناخته شده نرمال تمرکز دارد .

نرمالسازی فرآیند سازماندهی داده در پایگاه داده بطور کارآمد است .نرمالسازی روشی برای طراحی جداول پایگاه داده است و داده ها را به طریقی ساماندهی می کند که باعث کاهش افزونگی داده و رفع مشکلات ساختاری و آنومالی شود .

هدف از نرمالسازی حذف افزونگی داده و باقی نگهداشتن وابستگی بین داده های مرتبط است. به این طریق اندازه پایگاه داده را کاهش داده و ذخیره منطقی داده را تضمین می کند .

مفهوم نرمالسازی پایگاه داده اولین با ر توسط Edgar Frank Codd معرفی شد.

فرآیند نرمالسازی شامل ایجاد جداول و برقراری ارتباط بین آنها طبق قواعد معین است و روی وابستگی های ستون های جدول تمرکز دارد. این فرآیند اغلب باعث ایجاد جداول بیشتر می شود ولی باوجودیکه اثر تکرار داده درون پایگاه داده را دارد باعث افزونگی غیر ضروری داده نمی شود. هدف از نرمالسازی تضمین این است که ستون های غیر کلیدی در هر جدول مستقیما وابسته به کل کلید باشند و به این ترتیب حذف وابستگی های ناسازگار، کاهش افزونگی، آنومالی کمتر و بهبود کارائی را نتیجه می دهد .

مثال. جدول زیر که اطلاعات مربوط به خرید مشتریان را دارد درنظر بگیرید:

BIG UGLY TABLE

SaleNo	SaleDate	ProductNo	Qty	Amount	Salesrep	CustomerNo	First	Last	Address	CreditLimit
12345	Aug 12 2002	AQX88916	1	23.95	Dave Williams	4649-4673	Richard	Johnston	14 West Avenue	1000
12346	Aug 12 2002	AQX88916	7	167.65	Sara Thompson	1113-7741	Wayne	Jones	42 York Street	<null>
12347	Aug 13 2002	AHL46785	3705	5001.75	Li Qing	1166-3461	Amelia	Waverley	995 Forth Street	<null>
12348	Aug 13 2002	DHU69863	50	118.5	Sara Thompson	<null>	<null>	<null>	<null>	<null>
12349	Aug 14 2002	DHU69863	940	2227.8	Sara Thompson	1166-3461	Amelia	Waverley	995 Forth Street	<null>
12350	Aug 14 2002	DHU69863	42	99.54	Sara Thompson	7671-3496	Antonio	Gonzales	55B Granary Lane	<null>
12351	Aug 14 2002	AQX88916	55	1317.25	Dave Williams	6794-1674	Diane	Adams	364 East Road	1500

همانطور که مشاهده می شود با هر فروش داده ها در جدول تکرار می شوند. این افزونگی مشکلات زیر را می تواند ایجاد کند :

- هدر رفتن فضای ذخیره سازی. با وجودیکه امروزه دیسک های چندصد گیگا بایتی وجود دارد چندین بار ذخیره یک داده غیر ضروری است.
- آنومالی در بهنگام سازی . اگر داده یک مشتری، مثلا آدرس، تغییر کند باید در همه جاهائی که ذخیره شده است این تغییر اعمال شود درغیراینصورت جامعیت نقص می شود.
- آنومالی در حذف. اگر این جدول به منظور نگهداری مشخصات مشتریان باشد، اگر مشتری خریدش را پس بدهد و سطر مربوط به آن حذف شود کلیه اطلاعات مشتری هم حذف می شود.
- آنومالی در درج. به همین صورت نمی توانیم مشخصات مشتری جدید را درج کنیم مگر اینکه کالائی خریده باشد.

جدا کردن داده های جدول فوق به جداول جداگانه افزونگی را کاهش می دهد و مواجهه با آنومالی های فوق را ساده تر می کند. این فرآیند را نرمالسازی می نامند .

فرم های نرمال

تئوری پایگاه داده درجه نرمالسازی جدول را با اصطلاح فرم های نرمال (normal form) شرح می دهد. فرم های نرمل (یا بطور خلاصه NF) معیاری برای تعیین درجه نرمال جدول دراختیار می گذارد .

فرم های نرمال جداگانه روی هر جدول می توانند بکار بروند. پایگاه داده زمانی در فرم نرمل n خواهد بود که کل جداول آن در فرم نرمل n باشند .

فرم های نرمال عبارتند از:

- 1NF (First Normal Form)
- 2NF (Second Normal Form)
- 3NF (Third Normal Form)
- BCNF (Boyce/Codd Normal Form)
- 4NF (Forth Normal Form)
- 5NF (Fifth Normal Form)
- DKNF (Domain/Key Normal Form)

اگر فرم اول نرمال در جدولی مشاهده شود اصطلاحا آنرا در فرم اول نرمال (1NF) می نامند. اگر سه فرم اول نرمال دیده شود آنرا در فرم سوم نرمال (3NF) درنظر می گیرند. جدولی که دارای فرم نرمال درجه بالاتر باشد فرم های نرمال درجه پائین تر را هم دارا می باشد. بنابراین مثلا اگر جدولی 3NF باشد 2NF و 1NF هم هست. ولی عکس این صحت ندارد.

فرم نرمال هریک باعث کاهش بیشتر افزونگی و تقسیم جداول به واحدهای کوچکتر می شوند. سه فرم اول نرمال (1NF، 2NF و 3NF) در ابتدا توسط Codd تعریف شد که به طور خلاصه وابستگی صفات خاصه غیر کلید را به کلید الزام می کنند. فرم های چهارم و پنجم (4NF و 5NF) با ارتباطات چند به چند و یک به چند بیت صفات خاصه سروکار دارند. دو فرم دیگر هم وجود دارد که کاملا با این جریان جور نمی شوند که BCNF و DK/NF هستند.

در برنامه های کاربردی اغلب 1NF، 2NF و 3NF و گاهی 4NF و 5NF دیده خواهند شد 5NF بندرت مشاهده می شود به همین دلیل در اینجا توضیح داده نمی شود.

توجه داشته باشید که نرمالسازی یک فرآیند تکراری نیست. یک جدول ممکن است در یک مرحله به فرم سوم نرمل دربیاید. بعلاوه اگر 3NF باشد به احتمال بسیار زیاد 5NF هم خواهد بود.

(1NF) First Normal Form

یک جدول در فرم اول نرمال (1NF) است اگر و فقط اگر فاقد گروه داده تکرار شونده باشد. به عبارت دیگر هر ستون در جدول دارای مقدار اتمیک باشد.

در مدل رابطه ای هر جدولی حداقل در فرم اول نرمال هست زیرا از الزامات مدل این است که هر جدول شامل دقیقا یک مقدار برای هر صفت خاصه باشد که اصطلاحا "فاقد گروه تکرار شونده" گفته می شود .

مثال. جدول ALL_SALES که اطلاعات فروش را نگهداری می کند درنظر بگیرید .این جدول در فرم اول نرمال هست چون هیچ کدام از ستون ها چندمقداری نیستند بنابراین نیازی نیست روی جدول کاری انجام دهیم بجز اینکه یک کلید انتخاب نمائیم. ترکیب غیر تکراری ProductNo+CustomerNo+SaleNo را می توان کلید اصلی درنظر گرفت .

ALL_SALES(<u>SaleNo, ProductNo, CustomerNo</u>, SaleDate, QtyInStock, Description, Price, Customer_Name, Customer_Address, CreditLimit, Amount, Salesrep)

(2NF) Second Normal Form

یک جدول در فرم دوم نرمال (2NF) است اگر اولا 1NF باشد و ثانیا کلیه ستون های غیرکلید با کلید اصلی وابستگی تابعی کامل داشته باشند.

ستون Y با ستون X در یک رابطه وابستگی تابعی (functional dependency) دارد اگروفقط اگر به ازای هر مقدار در X دقیقا یک مقدار در Y متناظر با آن وجود داشته باشد. که به صورت X→Y نشان داده می شود .

مثال. در جدول ALL_SALES مثال قبل، Customer_Addressبا CustomerNo وابستگی تابعی دارد، زیرا یک مشتری خاص تنها با یک آدرس مربوط است. توجه کنید که عکس آن برقرار نیست و چند مشتری ممکن است در یک آدرس زندگی کنند .بنابراین یک آدرس ممکن است با بیش از یک شماره مشتری در ارتباط باشد. اگر مشتری بیش از یک آدرس داشته باشد دیگری وابستگی تابعی با شماره مشتری ندارد .

ستون Y روی مجموعه صفات خاصه X وابستگی تابعی کامل (Full functional dependency) دارد اگر روی X وابستگی تابعی داشته باشد و با هیچ زیرمجموعه ای از X وابستگی تابعی نداشته باشد .

مثال. در جدول ALL_SALES مثال قبل آدرس مشتری وابستگی کامل باSaleNo ، ProductNo و CustomerNo دارد ولی وابستگی تابعی کامل ندارد چون با CustomerNo وابستگی تابعی دارد.

توجه کنید اگر کلیدهای کاندید در جدول ترکیبی نباشند یعنی تنها شامل یک ستون باشند بلافاصله می گوئیم جدول 2NF است.

مثل .جدول ALL_SALES را درنظر بگیرید:
ALL_SALES(SaleNo, ProductNo, CustomerNo, SaleDate, QtyInStock, Description, Price, Customer_Name, CreditLimit, Amount, Salesrep)

مشاهده می شود بعضی از ستون ها بهم مرتبط هستند و توسط بخشی از کلید مشخص می شوند. به عبارت دیگر بعضی ستون ها با زیرمجموعه ای از کلید وابستگی تابعی دارند :

ProductNo → {Description, ReorderLevel, Price, QtyInStock}
CustomerNo → {Customer_Name, CreditLimit}
SaleNo → {Date, CustomerNo, ProductNo, Qty, Amount, Salesrep}

با جدا کردن این ستون ها به جداول جداگانه به فرم دوم نرمال می رسیم .

PRODUCT(ProductNo, Description, Price, QtyInStock)
CUSTOMER(CustomerNo, Customer_Name, CreditLimit)
SALE(SaleNo, Date, CustomerNo, ProductNo, Qty, Amount, Salesrep)

(3NF) Third Normal Form

یک جدول در فرم سوم نرمال (3NF) است اگر اولا 2NF باشد، ثانیا کلیه صفات خاصه غیر کلید در جدول با کلید اصلی وابستگی تابعی غیر تعدی داشته باشند.

وابستگی تعدی (transitive dependency) یک وابستگی تابعی غیر مستقیم است که در آن X→Z است اگر X→Y و Y→Zباشد .

در فرم سوم نرمل کلیه ستون های جدول مستقیما توسط کلید اصلی مشخص می شوند. با حذف فیلدهائی که وابستگی مستقیم با کلید ندارند به فرم سوم نرمل می رسیم. برای این کار گروهی از ستون های جدول را که مقدارشان برای بیش از یک رکورد تکرار می شود را در جدول جداگانه ای قرار دهید .

مثال. فرض کنید جدول PRODUCT به صورت زیر جزئیات تولید کننده هر محصول را دارا باشد :
PRODUCT(ProductNo, Description, ReorderLevel, Price, QtyInStock, SupplierCode, SupplierName, SupplierAddress)

این جدول کلید اصلی تک ستونی دارد بنابراین 2NF است. اگر تولید کننده چندین محصول را تولید کند فیلدهای SupplierName و SupplierAddress برای هر محصول تکرار می شود زیرا وابستگی تعدی با کلید اصلی دارند.
ProductNo → SupplierCode → {SupplierName, SupplierAddress}
با حذف این ستون ها و تقسیم جدول به صورت زیر به فرم سوم نرمال می رسیم. توجه کنید که SupplierCode در جدول PRODUCTبه عنوان کلید خارجی باقی می ماند .
PRODUCT(ProductNo, Description, ReorderLevel, Price, QtyInStock, SupplierCode)
SUPPLIER(SupplierCode, SupplierName, SupplierAddress)

رسیدن به فرم سوم نرمال اگرچه مطلوب است ولی همیشه عملی نیست. جداول متعدد باعث تنزل کارائی پرس و جوها می شود. بنابراین ممکن است فرم سوم تنها روی ستون هائی از جدول که زیاد تغییر می کنند اعمال شود و برای فیلدهای وابسته ای باقی مانده برنامه به نحوی طراحی می شود که کاربر ملزم باشد کلیه فیلدهای مرتبط را درهراصلاح بازبینی کند .

(4NF) Fourth Normal Form

یک جدول در فرم چهارم نرمال (4NF) است اگر اولا 3NF باشد، ثانیا هیچ ستونی در جدول وابستگی چند مقداری نداشته باشد.

وابستگی چندمقداری (multivalued dependency) به این معنی است که حضور رکوردهای معینی در جدول وجود رکوردهای معین دیگری را برساند.

مثال. اگر مشتریانی با چند آدرس داشته باشیم (که در محیط تجارت عادی است)، در جدول CUSTOMER نمی توانیم چند ستون آدرس را اضافه کنیم چون تعداد آدرس های ممکن را نمی دانیم. بنابراین ناگزیر به اضافه کردن رکورد جدید برای هر آدرس مشتری هستیم که باعث تکرار و افزونگی داده می شود. زیرا CustomerNo دیگر تنها یک آدرس را معین نمی کند بلکه مجموعه ای از آدرس های را نشان می دهد به عبارت دیگر وابستگی چندمقداری دارد. با حذف چنین وابستگی هائی و تقسیم جدول به صورت زیر به فرم چهارم نرمال می رسیم .
CUSTOMER(CustomerNo, First, Last, CreditLimit)
CUSTOMER_ADDRESS(CustomerNo, Address)
حالا هر مشتری می تواند هر تعداد آدرسی را داشته باشد .

معایب نرمالسازی

نرمالسازی تکنیک مهمی برای طراحی پایگاه داده های کارآمد است اما در ضمنی که افزونگی داده را کاهش می دهد سبب کاهش اجرای سیستم می شود. درجات بالای نرمال معمولا جداول بیشتر را می طلبند. برای پاسخ به پرس و جوها گاهی باید کلیه جداول تقسیم شده دوباره با هم الحاق شوند که در کاربردهائی که زمان پاسخ مهم است (نظیر وب) مطلوب نیست .

بالاترین سطح نرمالسازی با توجه به عملیات کاربردی درنظر گرفته می شود .در پایگاه داده هایی که بیشتر خواندنی هستند و افزونگی داده در آنها مشکل حادی نیست، مانند داده های کاتالوگ یک سایت تجارت الکترونیکی، می توان سطح نرمالسازی را کاهش داد. به این عمل denormalization می گویند. از طرف دیگر در کاربردهائی که درگیر داده های مهم مانند داده های مالی هستند که دائما در حال تغییرند و باید سازگار باقی بمانند، احتمالا سعی می شود به سطوح بالاتر نرمال برسند حتی اگر سرعت پایگاه داده کم شود .

گاهی با توجه به وضعیت ممکن است داده ها از چند پایگاه داده نرمال شده استخراج شوند و در یک انبار داده غیر نرمال قرار گیرد. این روش برای مخزن داده Data warehouse استاندارد خوبی است .