Projektowanie bazy danych

Podstawowym etapem powstawania bazy danych jest tworzenie jej projektu. Od decyzji podjętych na etapie projektowania zależeć będzie jakość i użyteczność stworzonej bazy danych.

Baza danych, podobnie jak większość projektów komputerowych, jest modelem wycinka świata rzeczywistego, utworzonym tak, aby był możliwy do zapamiętania przez maszynę cyfrową i zawierał optymalną ilość informacji do zastosowania, jakiemu będzie służył. Oczywiście, w zależności od zastosowania, interesujące mogą być różne informacje. W bazie danych studentów tworzonej przez pracowników mało prawdopodobne jest znalezienie się informacji o numerze butów studentów, jednakże w bazie danych jednostki wojskowej tak ważna cecha umundurowania, jak rozmiar obuwia, z pewnością nie zostanie pominięta, mimo iż obie bazy danych należą do jednej kategorii – bazy danych personalnych.

Jako pierwszy etap projektowania bazy danych należy zidentyfikować:

1. Listę interesujących nas *obiektów*, których opis musi znaleźć się w bazie
2. Listę *cech* poszczególnych obiektów, jakie należy zarejestrować

Przykładowo, w bazie danych dotyczących grupy studenckiej, prowadzonej przez nauczyciela akademickiego, ważnymi obiektami będą:

- student
- ocena
- nieobecność studenta na zajęciach

Każdy z tych obiektów ma swoje cechy. Studenta cechują jego imię, nazwisko oraz unikalny w ramach uczelni numer indeksu. Ocenę charakteryzuje nota, mieszcząca się w granicach regulaminu studiów (3.0, 3.5, ...) oraz informacja, za co dana ocena została uzyskana (praca domowa, projekt, kolokwium itd.) i najważniejsze – przez kogo. Nieobecność charakteryzowana jest przez datę, informację która osoba była nieobecna oraz określenie, czy nieobecność była usprawiedliwiona czy nie. Od trafnego zidentyfikowania obiektów i potrzebnych do zapamiętania cech zależy jakość projektu bazy danych.

Dalszy etap projektowania opiera sie na analizie związków pomiędzy obiektami i ich cechami. W analizie tych związków użyteczne jest pojęcie *zależności funkcyjnej*. Zależność funkcyjna zachodzi wtedy, gdy pomiędzy dwoma zespołami wartości zachodzi jednoznaczne przyporządkowanie dowolnej kombinacji wartości z pierwszego zespołu do *jednej i tylko jednej* kombinacji wartości drugiego zespołu. Przykładowo, w bazie z danymi studentów zależnością funkcyjną jest odwzorowanie numeru indeksu na parę imię oraz nazwisko studenta, co oznacza się następująco:

(numer indeksu) -> (imię, nazwisko)

, dlatego, że dowolnemu funkcjonującemu na uczelni numerowi indeksu jest przyporządkowany dokładnie jeden student. Należy zauważyć, że w ogólności, odwrotna relacja nie jest zależnością funkcyjną. Można sobie wyobrazić sytuację, w której na uczelni są dwie studentki, nazywające się Katarzyna Nowak, mają one jednak dwa różne numery indeksów, nie może więc zajść przyporządkowanie do jednej "kombinacji wartości z drugiej grupy".

Celem analizy związków pomiędzy obiektami modelowanymi w bazie jest dokonanie procesu normalizacji. *Normalizacja* to podział informacji, zawartych w bazie danych, na tabele. Jej celem jest usunięcie nadmiarowości danych, zawartych w bazie. Do analizy procesu normalizacji używa się dwóch pojęć:

- **klucz kandydujący** – jest to zbiór kolumn w tabeli, spełniający dwie cechy: kombinacja wartości w tych kolumnach jest unikalna w ramach całej tabeli oraz spośród kolumn klucza kandydującego

nie można wybrać podzbioru kolumn, który również zapewnia unikatowość wierszy w tabeli (jak czyni to pełny zestaw kolumn)

- **klucz główny** – jeden wybrany klucz kandydujący tabeli

Podział danych na tabele opiera się o relacje pomiędzy obiektami. Wyróżnia się trzy możliwe relacje pomiędzy obiektami:

1. "**jeden do jeden**" - relacja taka zachodzi, gdy jeden element wykazuje zależność funkcyjną od drugiego. W takiej sytuacji elementy umieszczane są zwykle w jednej tabeli. Przykładowo mając tabelę z danymi personalnymi osób, takimi jak imię, nazwisko, nie ma na ogół sensu umieszczać danych na temat dat urodzenia w osobnej tabeli. Wiersz tej tabeli opisuje jedną osobę, a każdej osobie jest przyporządkowana jednoznacznie jej data urodzenia.
2. "**jeden do wiele**" - relacja taka zachodzi, gdy jeden element powiązany jest z wieloma innymi. W takiej sytuacji elementy te umieszcza się w osobnych tabelach, które powiązane są parą klucz główny (w tabeli, gdzie znajduje się element, który jest "jeden") - klucz obcy (w tabeli, gdzie znajduje się "wiele" elementów). W bazie ze studentami taką relacją jest powiązany student z ocenami – każdy student ma wiele ocen, jednak każda pojedyncza ocena jest przyporządkowana jednemu, konkretnemu studentowi.
3. "**wiele do wiele**" - relacja taka zachodzi, gdy istnieją dwie grupy elementów, które mogą łączyć się ze sobą w taki sposób, że zarówno dowolny element z pierwszej grupy może łączyć się z wieloma elementami grupy drugiej, jak również dowolny element grupy drugiej może łączyć się z wieloma elementami grupy pierwszej. Technicznie relacje taką realizuje się poprzez specjalną tabelę łączącą dwie tabele, zawierające specyfikacje elementów obu grup powiązanych relacją. W bazie z wynikami studenckimi taka relacja może powstać, jeśli wydzielimy obiekt "forma sprawdzania wiedzy" (kolokwium, praca domowa, projekt itp.). W takim przypadku powiązanie studentów z "formą sprawdzania wiedzy" jest relacją "wiele do wiele": każdy student brał udział w wielu formach sprawdzania wiedzy, ale też w danej formie sprawdzania wiedzy badano umiejętności wielu studentów.

Proces normalizacji, prowadzony w oparciu o analizę relacji dostarcza cennych informacji o poprawnej strukturze bazy, jednak w praktyce konieczne jest jego umiarkowane stosowanie. Co prawda efekt procesu normalizacji, a więc eliminacja nadmiarowych informacji z bazy jest z pewnością pożądana, jednak normalizacja dokonuje tego poprzez podział na tabele, który powoduje dwie trudności. Po pierwsze, zmusza do łączenia wielu tabel przy odczytywaniu danych z bazy, co czyni mniej czytelnymi służące do tego zapytania. Druga, ważniejsza wada to fakt, że proces łączenia tabel jest kosztowny ze względu na wymagania pamięciowe i czasowe pracy serwera bazy danych. Po prostu takie zapytania długo sie przetwarzają, a czas przetwarzania zapytań w pewnych zastosowaniach może być krytyczną cechą. Analizę procesu normalizacji dokonuje się w oparciu o kryteria tzw. postaci normalnych, ważniejsze z nich są ponumerowane liczbami od jeden do pięciu:

- **pierwsza postać normalna** – pola zawierają tylko wartości niepodzielne
- **druga postać normalna** – spełniona pierwsza p.n., a ponadto kolumny nie wchodzące w skład klucza zależą funkcyjnie od całego klucza
- **trzecia postać normalna** – spełniona druga p.n., a ponadto kolumny nie wchodzące w skład klucza powinny zależeć funkcyjnie od klucza kandydującego – zazwyczaj tu kończy się normalizację
- **czwarta postać normalna** – wyeliminowane są zależności "jeden do wiele" pomiędzy niezależnymi kolumnami.
- **piąta postać normalna** – podział na maksymalną ilość tabel.