



POLITECHNIKA WARSZAWSKA
Wydział Elektroniki i Technik Informacyjnych
Instytut Informatyki

Rok akademicki 2013/2014

PRACA DYPLOMOWA MAGISTERSKA

Michał Aniserowicz

[TYTUŁ] (Generator aplikacji opartych o architekturę CQRS?)

Praca wykonana pod kierunkiem
dra inż. Jakuba Koperwasa

Ocena:

.....

*Podpis Przewodniczącego Komisji
Egzaminu Dyplomowego*

Spis treści

1	Wstęp	2
2	Duplikacja	3
2.1	Rodzaje duplikacji	3
2.2	Skutki występowania duplikacji	5
2.3	Rozwiązanie: zapisywanie wszystkich powiązanych i zduplikowanych miejsc	6
2.4	Rozwiązanie: skrypty na wszystkie czynności	6
2.5	Rozwiązanie: generyczna implementacja	7
2.6	Rozwiązanie: użycie generatorów	8
2.7	Rozwiązanie: jeden generator wszystkiego	9
3	Implementacja	11

Rozdział 1

Wstęp

TODO

- problem - w projektach często występuje duplikacja w wielu miejscach - w CQRS to w ogóle

Rozdział 2

Duplikacja

Zasada “DRY” (ang. *“Don’t Repeat Yourself”*), sformułowana przez Andrew Hunt’a i David’a Thomas’a, mówi: “Każda porcja wiedzy powinna mieć pojedynczą, jednoznaczną i autorytatywną reprezentację w systemie” [1]. Poprawne jej stosowanie skutkuje osiągnięciem stanu, w którym pojedyncza zmiana zachowania systemu wymaga modyfikacji tylko jednego fragmentu reprezentacji wiedzy. Należy podkreślić, że autorzy tej zasady przez “reprezentację wiedzy” rozumieją nie tylko kod źródłowy tworzony przez programistów systemu. Zaliczają do niej również dokumentację systemu, schemat bazy danych przez niego używanej i inne artefakty powstające w procesie wytwarzania oprogramowania (takie jak np. scenariusze testów akceptacyjnych), a nawet czynności wykonywane rutynowo przez programistów.

Celem stosowania tej zasady “DRY” jest unikanie duplikacji wiedzy zawartej w systemie. Duplikacją określa się fakt występowania w systemie dwóch reprezentacji tej samej porcji wiedzy. Należy wyjaśnić, że wielokrotne występowanie nazw klas lub zmiennych (bądź innych identyfikatorów) w kodzie programu nie jest duplikacją - identyfikatory są jedynie odnośnikami do wiedzy reprezentowanej przez artefakty, które identyfikują.

2.1 Rodzaje duplikacji

Wyróżnia się cztery rodzaje duplikacji ze względu na przyczynę jej powstania [1]:

- duplikacja wymuszona (ang. *imposed duplication*) - pojawia się gdy programista świadomie duplikuje kod aplikacji uznając, że w danej sytuacji jest to nie do uniknięcia;
- duplikacja niezamierzona (ang. *inadvertent duplication*) - występuje, gdy programista nie jest świadomy, że doprowadza do duplikacji;
- duplikacja niecierpliwa (ang. *impatient duplication*) - jest wynikiem niedbalstwa programisty; ma miejsce w sytuacji, gdy programista świadomie wybierze rozwiązanie najprostsze w danej sytuacji;
- duplikacja pomiędzy programistami (ang. *interdeveloper duplication*) - pojawia się, gdy kilku programistów tworzących tę samą aplikację wzajemnie duplikuje tworzony kod.

Duplikacja może występować w różnych postaciach:

2.1.1 Duplikacja czynności wykonywanych przez programistów

Do rutynowych czynności wykonywanych podczas pracy programisty należą wszelkiego rodzaju aktualizacje: schematu bazy danych rezydującej na maszynie programisty czy wersji tworzonego oprogramowania zainstalowanej w środowisku testowym lub produkcyjnym.

Czynności te zwykle składają się z kilku kroków. Przykładowo, na aktualizację bazy danych może składać się:

1. Usunięcie istniejącej bazy.
2. Stworzenie nowej bazy.
3. Zainicjalizowanie schematu nowej bazy przy pomocy skryptu DLL.
4. Wypełnienie nowej bazy danymi przy pomocy skryptu DML.

Jeśli wszystkie z tych kroków wykonywane są ręcznie, mamy do czynienia z duplikacją - każdy programista w zespole, zamiast tylko odwoływać się do wiedzy o tym, jak zaktualizować bazę danych, musi posiadać tę wiedzę i wcielać ją w życie krok po kroku.

2.1.2 Duplikacja kodu źródłowego aplikacji

Programistom wielokrotnie zdarza się spowodować duplikację w kodzie źródłowym tworzonych programów. Powielone fragmenty kodu są zazwyczaj niewielkie, a podzielić je można na następujące kategorie [3]:

- prosta duplikacja wyrażeń (ang. *basic literal duplication*) - najprostszy rodzaj duplikacji; obejmuje fragmenty kodu, których treść jest identyczna (przykład: dwie identyczne metody umiejscowione w różnych modułach systemu);
- parametryczna duplikacja wyrażeń (ang. *parametric literal duplication*) - obejmuje fragmenty kodu, których treść różni się jedynie typami danych, na których te fragmenty operują;
- duplikacja strukturalna (ang. *structural duplication*) - obejmuje fragmenty kodu, które mają ten sam schemat działania, ale różnią się pojedynczymi instrukcjami (przykład: dwie pętle iterujące po tej samej kolekcji z tym samym warunkiem stopu, ale wykonujące różne operacje na elementach kolekcji);
- duplikacja czasowa (ang. *temporal duplication*) - określa fragment kodu, który jest niepotrzebnie wykonywany wiele razy (przykład: zliczenie elementów kolekcji podczas każdego sprawdzenia warunku stopu pętli iterującej po tej kolekcji);
- duplikacja intencji (ang. *duplication of intent*) - występuje, gdy dwóch programistów umieści w różnych modułach aplikacji dwa fragmenty, których wynik działania (ale niekoniecznie treść) jest identyczny.

2.1.3 Duplikacja opisu dziedziny aplikacji

W każdej aplikacji obiektowej korzystającej z bazy danych opis dziedziny aplikacji występuje w co najmniej dwóch miejscach. Te miejsca to:

- schemat bazy danych (DDL);
- definicje klas w kodzie źródłowym aplikacji.

Kolejnym typowym miejscem, w którym umieszcza się informacje o dziedzinie aplikacji jest jej dokumentacja.

W miarę rozrastania się systemu i pojawiania się kolejnych funkcjonalności i modułów, występuje tendencja do powielania całości bądź części definicji dziedziny w różnych modułach. Dotyczy to modułów, których zadaniem jest obróbka tych samych danych, ale w różny sposób. Przykładowo, aplikacja może udostępniać przechowywane dane na następujące sposoby:

- wyświetlać je na stronie WWW,
- wystawiać jako API,
- eksportować do arkusza kalkulacyjnego.

Moduły te mogą korzystać z pojedynczej implementacji dziedziny. Może to jednak nie być pożądane, szczególnie w przypadku, gdy każdy ze sposobów udostępnia inny zestaw danych. Przykładowo, strona WWW może wyświetlać podstawowe dane każdej encji będącej częścią dziedziny, podczas gdy API udostępnia pełne dane wszystkich encji, a eksport do arkusza kalkulacyjnego udostępnia pełne dane niektórych encji.

W takim przypadku moduł odpowiedzialny za API operuje na pełnej dziedzinie, a moduły strony WWW i eksportu do arkusza kalkulacyjnego - na jej fragmentach. Jeśli każdy z tych modułów posiada niezależną implementację dziedziny aplikacji, to mamy do czynienia z duplikacją.

2.2 Skutki występowania duplikacji

Pojawienie się duplikacji w systemie ma zazwyczaj szkodliwe skutki. Rozprzestrzenia się ona tym szybciej, a wyrządzane szkody są tym dotkliwsze, im większy jest rozmiar systemu.

Najbardziej oczywistym skutkiem występowania duplikacji jest wydłużenie się czasu poświęcanego przez programistów na nieskomplikowane lub powtarzalne zadania. Przykładowo, nieznanomość kodu współdzielonego przez wszystkie moduły systemu (tzw. rdzenia, ang. *core*) powoduje, że programiści niepotrzebnie spędzają czas na implementowaniu podstawowych funkcji, które są już dostępne w rdzeniu systemu. Innym przykładem może być praktyka ręcznego wykonywania czynności, które mogłyby być wykonywane automatycznie, w tle.

Po drugie, pojedyncza porcja wiedzy jest reprezentowana w wielu miejscach, a więc zmiana pojedynczego zachowania systemu wymaga modyfikacji wielu jego modułów. Wprowadzając dowolną modyfikację, programista musi ręcznie zlokalizować wszystkie miejsca

wymagające zaktualizowania. Problem ten jest najbardziej odczuwalny podczas naprawy zgłoszonych błędów: będąc świadomym występowania duplikacji kodu, programista naprawiający błąd musi pamiętać, że poprawiany przez niego fragment może być powielony w wielu miejscach w systemie. Wszystkie te miejsca musi zlokalizować i poprawić [4]. Jeśli tego nie zrobi, w systemie pozostaną sprzeczne ze sobą reprezentacje danej wiedzy, a kolejni programiści nie będą świadomi, która reprezentacja jest poprawna. W dłuższej perspektywie taka sytuacja prowadzi do pojawienia się kolejnych błędów.

Po trzecie, duplikacja powoduje zbędne powiększenie rozmiarów systemu. Kod, w którym występują zduplikowane fragmenty jest dłuższy, a przez to mniej czytelny. Co więcej, zwiększona ilość kodu powoduje niepożądane zwiększenie rozmiarów plików wchodzących w skład systemu.

Należy również zwrócić uwagę na zagrożenie, jakie niesie ze sobą wystąpienie pojedynczej duplikacji, której istnienie zostanie zaakceptowane przez programistów wchodzących w skład zespołu tworzącego projekt. Może to spowodować osłabienie dyscypliny w zespole, a w efekcie doprowadzić do pojawienia się kolejnych duplikacji, a także innych złych praktyk programistycznych. Pojedyncze wystąpienie tego zjawiska może więc doprowadzić do rozprzestrzenienia się innych niepożądanych zjawisk i pogorszenia jakości całego systemu [2].

2.3 Rozwiązanie: zapisywanie wszystkich powiązanych i zduplikowanych miejsc

Najprostszym, a zarazem najbardziej naiwnym sposobem na zniwelowanie skutków duplikacji może być utrzymywanie listy fragmentów systemu, w których duplikacja występuje. Przykładowo, zespół może przygotować listę miejsc, które należy zmodyfikować, jeśli pojedyncza encja zostanie wzbogacona o nowe pole. Na takiej liście znalazłby się schemat bazy danych, implementacja modelu danych w kodzie źródłowym programu, dokumentacja systemu, itd.

Takie podejście ma jednak istotne wady:

- rozwiązanie to nie powstrzymuje programistów przed wprowadzaniem kolejnych duplikacji, a wręcz na to przyzwala;
- istnienie takiej listy samo w sobie jest duplikacją - pojawienie się nowego miejsca występowania pól encji (np. alternatywnej implementacji modelu danych) wymaga zaktualizowania listy;
- rozwiązanie to niweluje tylko drugi z wymienionych skutków występowania duplikacji - programista nie musi samodzielnie szukać miejsc, w których powinien wprowadzić modyfikacje.

2.4 Rozwiązanie: skrypty na wszystkie czynności

Wiele czynności wykonywanych przez programistów da się zautomatyzować, tworząc skrypty je wykonujące. Przykłady:

- wszystkie kroki składające się na czynność aktualizacji schematu bazy przedstawone w sekcji 2.1.1 mogą być wykonywane przez pojedynczy skrypt powłoki korzystający z kilku skryptów DDL i DML;
- operacja instalacji nowej wersji systemu na środowisku testowym lub produkcyjnym nie musi wiązać się z ręcznym przesyłaniem plików systemu na serwer FPT zdalnej maszyny - może odbywać się automatycznie, po uruchomieniu odpowiedniego skryptu [5].

Zalety:

- w trakcie działania skryptu w tle, programista może skupić się na innych zadaniach,
- wiedza na temat kroków składających się na daną czynność jest reprezentowana w jednym miejscu,
- programiści lub inne skrypty odwołują się do tej wiedzy poprzez jej identyfikator (nazwę lub ścieżkę w systemie plików).

Należy jednak zauważyć, że podejście to rozwiązuje jedynie problem duplikacji czynności wykonywanych przez programistów. Pozostałe rodzaje duplikacji nie mogą być zniwelowane w ten sposób.

2.5 Rozwiązanie: generyczna implementacja

Rozwiązaniem problemu duplikacji kodu źródłowego aplikacji może być generyczna implementacja aplikacji. W językach obiektowych jest ona możliwa do osiągnięcia na kilka sposobów:

2.5.1 Dziedziczenie

Podstawowym sposobem na uniknięcie duplikacji w kodzie aplikacji wykorzystujących języki obiektowe jest dziedziczenie. Metody wspólne dla rodzin klas obiektów są umieszczane w ich klasach bazowych. Pozwala to uniknąć podstawowej duplikacji wyrażeń (patrz: sekcja 2.1.2).

2.5.2 Typy szablonowe i generyczne

Innych rodzajów duplikacji - parametrycznej duplikacji wyrażeń i duplikacji strukturalnej - można uniknąć poprzez zastosowanie typów generycznych i szablonów klas. Kod o tej samej strukturze i działaniu, ale operujący na danych innego typu, może być zamknięty w pojedynczej klasie.

2.5.3 Refleksja

W sytuacjach, w których występuje potrzeba zaimplementowania podobnej funkcjonalności dla kilku typów nienależących do jednej rodziny obiektów, można skorzystać z mechanizmu refleksji. Pozwala on osiągnąć rezultaty podobne do tych oferowanych przez klasy bazowe i typy generyczne.

Należy podkreślić, że przedstawione propozycje rozwiązują jedynie problem duplikacji kodu źródłowego aplikacji.

2.6 Rozwiązanie: użycie generatorów

Rozwiązaniem pozwalającym na uniknięcie duplikacji opisu dziedziny aplikacji jest zastosowanie różnego typu generatorów.

Należy wyjaśnić, że kodu źródłowego (bądź innych artefaktów) wygenerowanego na podstawie innego kodu nie uznaje się za duplikację. Powodem jest to, że jeśli na podstawie danego fragmentu kodu generowanych jest wiele artefaktów, to aby wprowadzić zmiany we wszystkich tych artefaktach, należy jedynie zmodyfikować źródłowy fragment kodu i uruchomić proces generacji. Za “pojedynczą, jednoznaczną i autorytatywną” reprezentację danej porcji wiedzy (patrz: Zasada “DRY”) uznaje się w tym przypadku źródłowy fragment kodu.

2.6.1 Generator modelu dziedziny aplikacji na podstawie schematu bazy danych

Najpowszechniejszym przykładem użycia generatorów kodu jest wykorzystanie generatora zestawu klas wchodzących w skład modelu dziedziny aplikacji na podstawie schematu bazy danych używanego przez tą aplikację. Takie podejście nosi nazwę “najpierw baza danych” (ang. *Database First*).

Przykładami narzędzi umożliwiającą taką generację kodu są:

- NHibernate [6] - narzędzie ORM (ang. *Object-Relational Mapping*) przeznaczone na platformę .Net;
- EntityFramework [7] - narzędzie ORM dla platformy .Net;
- SQLMetal [8] - narzędzie dla platformy .Net, którego jedynym zadaniem jest generacja kodu źródłowego klas na podstawie schematu bazy danych;
- Django [9] - platforma aplikacji webowych dla języka Python.

2.6.2 Generator schematu bazy danych na podstawie modelu dziedziny aplikacji

Podejściem przeciwnym dla “*Database First*” jest podejście “najpierw kod” (ang. *Code First*). Jak sama nazwa wskazuje, generatory tego typu generują schemat bazy danych na podstawie klas należących do implementacji modelu w kodzie źródłowym aplikacji.

Przykłady generatorów *“Code First”* to:

- Hibernate [6] - pierwowzór narzędzia NHibernate (patrz: sekcja 2.6.1), przeznaczony na platformę Java;
- EntityFramework (patrz: sekcja 2.6.1);
- Django (patrz: sekcja 2.6.1).

2.6.3 Generator dokumentacji

Duplikacji opisu dziedziny aplikacji w jej dokumentacji można uniknąć poprzez zastosowanie narzędzi generujących tę dokumentację na podstawie kodu źródłowego aplikacji. Generatory tego rodzaju wymagają umieszczania w kodzie źródłowym specjalnie sformatowanych komentarzy, na podstawie których są w stanie wygenerować dokumentację w kilku formatach, takich jak PDF czy HTML. Przykłady takich narzędzi to:

- Doxygen [10] - popularne narzędzie obsługujące wiele języków programowania (w tym C++, Java C#), wspierające wiele formatów dokumentacji (w tym HTML, PDF, LaTeX, XML);
- JavaDoc [11] - generator dedykowany językowi Java, domyślnie wspiera jedynie format HTML;
- C# XML Documentation [12] - format tworzenia dokumentacji wbudowany w język C#, na podstawie którego generowana jest dokumentacja w formacie XML;
- pydoc [13] - narzędzie będące częścią standardowego zestawu narzędzi deweloperskich języka Python; wspiera format tekstowy i HTML.

2.6.4 Inne generatory

Istnieje wiele innych generatorów, należących do typu programowania nazywanego programowaniem automatycznym (ang. *automatic programming*) [14]. Jednakże większość z nich, tak jak powyższe przykłady, eliminuje tylko pojedyncze rodzaje duplikacji.

Warto zwrócić uwagę na fakt, że różne typy generatorów za źródło danych obierają sobie różne definicje dziedziny: np. generator *Database First* bazuje na schemacie bazy danych, a generator dokumentacji - na kodzie źródłowym aplikacji. Używanie wielu różnych generatorów usuwających pojedyncze rodzaje duplikacji w końcu doprowadziłoby do powstania “łańcucha” generatorów, w którym wynik działania jednego generatora byłby źródłem danych dla innego. Takie rozwiązanie może być trudne w utrzymaniu.

2.7 Rozwiązanie: jeden generator wszystkiego

Przedstawione rozwiązania mają pewne wspólne wady. Po pierwsze, wszystkie z nich usuwają tylko pojedyncze rodzaje duplikacji. Aby całkowicie wyeliminować duplikację z systemu, należałoby by użyć prawie wszystkich z nich.

Po drugie, rozwiązania zwalczające duplikację czynności wykonywanych przez programistów i duplikację kodu źródłowego aplikacji nie są rozwiązaniami prewencyjnymi - nie zapobiegają pojawianiu się nowych duplikacji. To, czy nowa rutynowa czynność zostanie zautomatyzowana przy pomocy skryptu powłoki, i czy nowy moduł aplikacji zostanie zaimplementowany w sposób generyczny, zależy jedynie od dyscypliny i dbałości zespołu tworzącego system. Aby uniknąć takiej sytuacji, wydaje się, że mechanizmy niwelujące duplikację powinny leżeć u podstawy systemu - tak, aby programiści intuicyjnie korzystali z nich, wprowadzając nowe czynności lub aktualizując dziedzinę aplikacji.

Odpowiednim rozwiązaniem wydaje się być zastosowanie pojedynczego generatora potrafiącego wygenerować wszystkie potrzebne w systemie artefakty. Powstanie takiego generatora dającego się zastosować w każdym projekcie jest jednak bardzo mało prawdopodobne, ponieważ:

- każdy projekt ma inne wymagania odnośnie wygenerowanych artefaktów;
- prawdopodobne nie istnieje format, który pozwalałby zdefiniować każdą dziedzinę w najlepszy dla niej - tj. najbardziej naturalny, a przy tym pozbawiony duplikacji - sposób.

Jak widać, łatwo dostępne są jedynie rozwiązania pozwalające wyeliminować pojedyncze rodzaje duplikacji. Nie jest jednak znane pojedyncze, spójne rozwiązanie eliminujące wszystkie rodzaje duplikacji. Odpowiednim rozwiązaniem wydaje się być rozwiązanie łączące kilka z przedstawionych wyżej rozwiązań. W dalszej części pracy podjęta zostanie próba stworzenia takiego rozwiązania. Rozwiązanie to zostanie oparte o generację kodu, a celem jest skonstruowanie go w taki sposób, aby pozwolić na usunięcie jak największej liczby rodzajów duplikacji patrząc zarówno z perspektywy ich rodzajów, jak i przyczyn powstawania (patrz: sekcja 2.1).

Rozdział 3

Implementacja

Celem niniejszej pracy jest próba stworzenia takiego generatora. Założenia funkcjonalne:

- generator ma być na tyle elastyczny, aby móc obsłużyć wiele sposobów zdefiniowania dziedziny aplikacji,
- generator powinien pozwalać na wygenerowanie dowolnych plików tekstowych (skryptów SQL, kodu źródłowego, dokumentacji HTML itd.),
- generator nie musi generować logiki biznesowej aplikacji - wystarczy dziedzina
- wszystkie szablony generacji powinny być zdefiniowane w ten sam sposób (w tym samym języku)

Założenia нефункционалне:

- generator zostanie stworzony w technologii .NET Framework,
- ...

Szczególна uwaga zostanie poświęcona aplikacjom opartym o architekturę CQRS i wykorzystującym bazy danych typu NoSQL. Specyficzną cechą takich aplikacji jest to, że operują one na modelach o wysokim stopniu denormalizacji, co wiąże się z masowo występującą duplikacją danych.

TODO: Opisać CQRS. TODO: Opisać Event Sourcing. TODO: Opisać NoSQL. TODO: Opisać rodzaje baz NoSQL. TODO: Wybrać bazę NoSQL i dlaczego Cassandra. TODO: Opisać Cassandra.

Bibliografia

- [1] *The Pragmatic Programmer*. Hunt A., Thomas D. Addison-Wesley. Westford 2013. ISBN 0-201-61622-X. *The Evils of Duplication*, s. 26-33.
- [2] *The Pragmatic Programmer*. Hunt A., Thomas D. Addison-Wesley. Westford 2013. ISBN 0-201-61622-X. *Software Entropy*, s. 4-6.
- [3] *Duplication in Software* [online]. Just Software Solutions. <http://www.justsoftwaresolutions.co.uk/design/duplication.html> [dostęp: lipiec 2014].
- [4] *Repetition* [online]. The Bad Code Spotter's Guide. Diomidis Spinellis. <http://www.informit.com/articles/article.aspx?p=457502&seqNum=5> [dostęp: lipiec 2014].
- [5] *Continuous delivery* [online]. ThoughtWorks. <http://www.thoughtworks.com/continuous-delivery> [dostęp: lipiec 2014].
- [6] *Hibernate. Everything data.* [online]. Hibernate. <http://http://hibernate.org/> [dostęp: lipiec 2014].
- [7] *Entity Framework* [online]. CodePlex. <https://entityframework.codeplex.com/> [dostęp: lipiec 2014].
- [8] *SqlMetal.exe (Code Generation Tool)* [online]. Microsoft Developer Network. [http://msdn.microsoft.com/pl-pl/library/bb386987\(v=vs.110\).aspx](http://msdn.microsoft.com/pl-pl/library/bb386987(v=vs.110).aspx) [dostęp: lipiec 2014].
- [9] *The Web framework for perfectionists with deadlines* [online]. django. <https://www.djangoproject.com/> [dostęp: lipiec 2014].
- [10] *Doxygen: Main Page* [online]. Doxygen. <http://www.stack.nl/~dimitri/doxygen/> [dostęp: lipiec 2014].
- [11] *Javadoc Tool Homepage* [online]. Oracle. <http://www.oracle.com/technetwork/java/javase/documentation/jsp-135444.html> [dostęp: lipiec 2014].
- [12] *XML Documentation* [online]. Microsoft Developer Network. [http://msdn.microsoft.com/en-us/library/b2s063f7\(vs.71\).aspx](http://msdn.microsoft.com/en-us/library/b2s063f7(vs.71).aspx) [dostęp: lipiec 2014].
- [13] *pydoc — Documentation generator and online help system* [online]. Python v2.7.8 documentation. <https://docs.python.org/2/library/pydoc.html> [dostęp: lipiec 2014].

- [14] *Automatic programming* [online]. Wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/Automatic_programming [dostęp: lipiec 2014].

OŚWIADCZENIE

Oświadczam, że Pracę Dyplomową pod tytułem “[TYTUŁ]”, którą kierował dr inż. Jakub Koperwas, wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

.....

Michał Aniserowicz