

Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH



Instytut Informatyki

Praca dyplomowa inżynierska

na kierunku Informatyka
w specjalności Inżynieria Systemów Informatycznych

Przeglądarka danych uzyskanych z sekwencjonowania następnej
generacji (NGS)

Tomasz Kogowski

Numer albumu 261428

promotor
dr inż. Tomasz Gambin

Warszawa 2017

Streszczenie

Abstract



„załącznik nr 3 do zarządzenia nr 24/2016 Rektora PW

.....
miejscowość i data

.....
imię i nazwisko studenta

.....
numer albumu

.....
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta”

Spis treści

1 Wstęp	6
1.1 Motywacja	6
1.2 Cel pracy	6
2 Podstawy teoretyczne	7
3 Wymagania funkcjonalne i нефункционалне	8
3.1 Wymagania funkcjonalne	8
3.2 Wymagania нефункционалне	9
4 Istniejące rozwiązania	10
5 Wybór technologii	11
5.1 Język programowania Scala	11
5.2 System zarządzania bazą danych	11
5.2.1 MySQL	12
5.2.2 SQLite	12
5.2.3 PostgreSQL	12
5.2.4 Uzasadnienie wyboru PostgreSQL	13
5.3 Slick	13
5.4 Aplikacja przeglądarkowa	13
6 Przypadki użycia	15
6.1 Autoryzacja	15
6.1.1 Role	15
6.1.2 Rejestracja i logowanie użytkownika	15
6.2 Przeglądanie danych z sekwencjonowania DNA	16
6.3 Panel administratora	22
7 Schemat bazy danych	23
8 Opis implementacji	24
9 Bezpieczeństwo aplikacji	25
9.1 Niebezpieczeństwa	25
9.2 Wykorzystanie protokołu https	25

9.3 SQL injection	25
10 Testy oraz wydajność	26
11 Wnioski i podsumowania	27
Literatura	27

1 Wstęp

1.1 Motywacja

1.2 Cel pracy

2 Podstawy teoretyczne

Czy jest sens pisać o tym? 1-2 strony o tym czym jest DNA, genomu, kodony, jak zmiana w DNA może wpływać na organizm i dlaczego warto zajmować się badaniem DNA

3 Wymagania funkcjonalne i нефункционалне

Określenie funkcjonalności dostępnych w budowanej aplikacji, rozpoczęto od określenia rodzajów użytkowników, którzy mają korzystać z oprogramowania tak by jak najlepiej dostosować system do ich potrzeb, przyspieszyć dostęp do danych.

Pierwszą grupą docelową są lekarze, którzy będą poszukiwali możliwych chorób powiązanych z wariantem pacjenta, by wykryć niebezpieczeństwa i móc jak najwcześniej przeciwdziałać chorobom. Na dane będą patrzeć w kontekście jednego badanego pacjenta i należy umożliwić im łatwe ich rozróżnienie genotypów.

Drugim typem są analitycy, którzy będą analizować dane i zadawać odwrotne pytania, czyli będą starać się znaleźć warianty, które mogą być odpowiedzialne za konkretną chorobę.

Inną istotną kwestią wziętą pod uwagę był typ i wielkość danych, jakie mają być wyświetlane klientom. W trakcie projektowania architektury jako dane przykładowe zostały wybrane dane dla przykładowego transkryptu dostępne w aplikacji Exac [3]. Dane te posiadały 36 kolumn i oczywistym wydało się że obie grupy użytkowników będzie interesowała tylko część informacji o genotypie i należało by umożliwić im filtrację oraz zakrywanie niepotrzebnych danych. Jedna próbka liczyła sobie więcej niż 340000 wiersze co wymogło zaproponowanie funkcjonalności umożliwiających na poprawne i intuicyjne filtrowanie danych tak by klient otrzymywał tylko interesujące go rekordy.

3.1 Wymagania funkcjonalne

Po zakończeniu analizy zostały określone następujące funkcjonalności. Aplikacja:

- 1) ma wygodny, prosty interfejs użytkownika,
- 2) rejestruje użytkowników,
- 3) autoryzuje użytkowników,
- 4) umożliwia wybór próbki do analizy,

- 5) pozwala na wprowadzenie wcześniej zdefiniowanych filtrów z panelu administratora,
- 6) wyświetla dane z sekwencjonowania DNA dla konkretnej próbki,
- 7) filtruje dane po stronie serwera i wysła je klientowi,
- 8) zlicza ilość danych przy zadanych filtrach i informuje klienta o wyniku,
- 9) umożliwia zmianę wartości filtrów,
- 10) pozwala na wyłączenie z filtracji dowolnej części filtrów,
- 11) zapisuje wartości filtrów oddzielnie dla każdego użytkownika,
- 12) sortuje dane po stronie klienta,
- 13) filtruje dane po stronie klienta,
- 14) udostępnia administratorowi możliwość zmiany dostępu do próbek każdego użytkownika
- 15) daje możliwość zakrycia na stronie aplikacji części danych

Funkcjonalności umożliwiające wprowadzanie wcześniej zdefiniowanych filtrów spowodowała stworzenie trzeciej klasy użytkowników, to jest administratorów, którzy będą zarządzali strukturą filtrów poprzez wprowadzenie odpowiedniego pliku z specjalnie przygotowanego panelu administracyjnego oraz będą zarządzać dostępem do próbek dla użytkowników.

3.2 Wymagania niefunkcjonalne

- 1) szyfruje wysyłane dane między klientem a serwerem za pomocą HTTPS,
- 2) wykorzystuje funkcję SHA-512 do zabezpieczenia hasła użytkownika,
- 3) korzysta z "soli" przy wyliczaniu funkcji skrótu,

4 Istniejące rozwiązania

Exac broad institute, Exac Harvard Skupić się na tym iż systemy nie pozwalają na personalizację interfejsu dla użytkownika. Harvard udostępnia REST API nieprzyjazne użytkownikowi

Czy dodać tu zdjęcia z tych aplikacji?

5 Wybór technologii

Platforma klastrowego przetwarzania danych - Apache Spark[2], z którą współpracować będzie aplikacja, została stworzona oraz udostępnia interfejs programistyczny w języku Scala. Naturalnym przez to wydało się wybranie tego języka programowania do stworzenia przeglądarki danych.

5.1 Język programowania Scala

W aplikacji użyto języka Scala w wersji 2.11.7 [1]. Jest to język programowania powstały w 2001 roku pod kierownictwem Martina Odersky'ego w Lozannie. Działa na Wirtualnej Maszynie Javy a do 2012 roku wspierała platformę .NET opracowaną przez firmę Microsoft. Język ten nadaje się równie dobrze do krótkich, zwartych skryptów wywoływanych podobnie do skryptów języka Python jak i do tworzenia wydajnych, ogromnych, bezpiecznych systemów sieciowych.

Jest językiem łączącym cechy języków funkcyjnych oraz obiektowych. Nie jest jednak obligatoryjny funkcyjny styl programowania, do którego nie jest przyzwyczajona większość programistów. Scala w swoim założeniu nawiązuje do minimalizmu składni Lispa to znaczy że nie opiera się na składni a na funkcjach bibliotecznych. Nazwa ma podkreślić skalowalność języka, dzieje się tak dzięki możliwości tworzenia dodatkowych typów i struktur wyglądających jak nowa składnia języka. Zaletą języka jest również to że dzięki kompatybilności z językiem Java mamy możliwość wykorzystania każdej linii kodu napisanej w owym języku.

5.2 System zarządzania bazą danych

Zadanie stworzenia bazy danych przechowującej informacje konfiguracyjne, dane użytkowników oraz o użytkownikach było dużą częścią tworzenia systemu i wymagało wybrania odpowiedniego systemu zarządzania bazą danych. Model bazodanowy został zaprojektowany w modelu opartym na relacyjnej organizacji danych, przez co wybór ograniczył się do darmowych technologii realizujących relacyjne bazy danych.

Biorąc pod uwagę powyższe kryteria, można porównać najpopularniejsze systemami, są nimi[4]:

- MySQL
- SQLite
- PostgreSQL

5.2.1 MySQL

Zalety

- proste i łatwe w obsłudze
- wysoki poziom bezpieczeństwa

Wady

- nie realizuje w pełni standardu SQL
- problematyczny jednoczesny zapis i odczyt

5.2.2 SQLite

Zalety

- zgodny ze standardem SQL
- przenośny dzięki oparciu bazy o jeden plik

Wady

- brak zarządzania użytkownikami i dostępami do danych

5.2.3 PostgreSQL

Zalety

- zgodny ze standardem SQL
- wsparcie dla współbieżności
- pełne wsparcie dla transakcji

Wady

- słaba wydajność
- trudność instalacji dla początkujących użytkowników

5.2.4 Uzasadnienie wyboru PostgreSQL

Po analizie ostateczny wybór systemem padł na PostgreSQL. To otwarte i darmowe oprogramowanie posiada bardzo dużą społeczność, której wiedza jest łatwo dostępna w internecie i posiada wiele narzędzi i bibliotek przeznaczonych do pracy z owym systemem. Istotny wpływ na decyzję miała również łatwość integracji PostgreSQL na inne systemy.

5.3 Slick

Pracę z bazą danych po stronie serwera aplikacyjnego znacznie ułatwia oprogramowanie pozwalające na odwzorowanie obiektowo-relacyjne tabel bazodanowych na obiekty języka programowania. Dzięki tej technice programista może traktować obiekty bazodanowe jak elementy kolekcji czy pola obiektów.

Takim narzędziem jest stworzone przez firmę Lightbend, Inc. oprogramowanie Slick[5] pozwalające na pełną kontrolę nad bazą danych oraz pisanie klasycznych zapytań SQL.

5.4 Aplikacja przeglądarkowa

Biorąc pod uwagę wymagania klientów oraz różnorodność używanych przez nich urządzeń należało wybrać odpowiedni rodzaj aplikacji klienckiej pozwalający na spełnienie wszystkich wymagań funkcjonalnych naszych użytkowników oraz jednocześnie będący łatwy w utrzymaniu i rozwijaniu.

Zalety aplikacji internetowych Łatwość w dostępie do internetu, ilość urządzeń pozwalających na korzystanie z przeglądarek internetowych pozwoliły na rozwój aplikacji internetowych oraz ich rozpowszechnienie. Łatwość w rozbudowie, zarządzaniu i niskie ceny hostowania serwera aplikacyjnego spowodowały

wały powstanie grupy platform programistycznych wspomagających ich budowę.

Narzędzia typu Ruby on Rails czy Spring Boot zdejmują z programisty obowiązek konfiguracji serwera HTTP od podstaw i umożliwiają rozpoczęcie pracy nad stronami aplikacji po kilku minutach.

Platforma programistyczna Play Platforma Play, stworzona w języku Scala jest środowiskiem do tworzenia aplikacji internetowych, która na celu ma przyspieszyć pracę programisty dzięki:

- strategii Konwencji Ponad Konfigurację
- przeładowywania i ponownej kompilacji plików po edycji
- wykorzystaniu wzorca Model-Widok-Kontroler
- wykorzystaniu technologii REST

Platforma programistyczna Angular Angular jest opracowaną przez Google biblioteką wspomagającą tworzenie aplikacji przeglądarkowych na jednej stronie. Jej głównymi zaletami jest :

- oddzielenie warstwy klienckiej od warstwy serwerowej
- oddzielenie manipulacji modelu dokumentu HTML od logiki aplikacji
- wykorzystaniu wzorca Model-Widok-Kontroler
- wykorzystaniu technologii REST

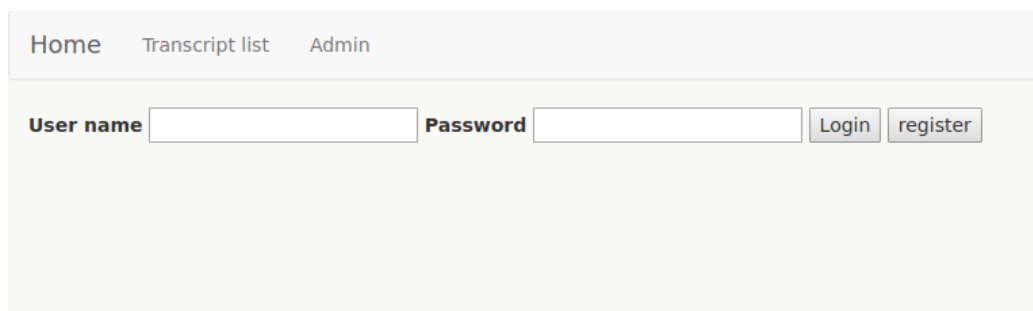
6 Przypadki użycia

Czy skupić się bardziej na opisie działania aplikacji czy raczej implementacji?

6.1 Autoryzacja

6.1.1 Role

6.1.2 Rejestracja i logowanie użytkownika



Home Transcript list Admin

User name Password Login register

Rysunek 1. Okno logowania i rejestracji

Rejestracja

Logowanie

6.2 Przeglądanie danych z sekwencjonowania DNA

Po zalogowaniu się do aplikacji przed użytkownikiem pojawia się główna część aplikacji pozwalająca Mu na dostęp do próbek oraz pracę z nimi. Zaimplementowano funkcjonalność sesji pozwalającą na wykorzystywanie guzika "wstecz" przeglądarki oraz odświeżania strony bez potrzeby ponownego logowania się.

Ekran dostępnych genomów Każda próbka dostępna w bazie danych posiada indywidualny identyfikator pozwalający na rozróżnienie jej od innych próbek. Owy identyfikator jest ciągiem znaków, który wyświetlany jest alfabetycznie posortowanej liście. Użytkownik widzi tylko próbki udostępnione Mu przez administratora i ma możliwość spojrzenia dokładniej w dane poprzez kliknięcie w identyfikator próbki, co przeniesie go do następnej strony prezentującej dane z sekwencjonowania DNA.

Home	Transcript list	Admin
Sample id		
id_1		
sample_id_17		
sample_id_18		
sample_id_2		

Rysunek 2. Lista próbek

Przechodzą na stronę z danymi dla konkretnej próbki, użytkownikowi wyświetla się tabela z danymi (Rysunek numer 3).

Rysunek 3. Tabela z danymi z sekwencjonowania DNA

- czas odpowiedzi serwera na zapytanie HTTP
- czas wygenerowania i wykonania zapytania SQL
- czas wygenerowania przez przeglądarkę elementów HTML tabeli

Obie te wady zostały zniwelowane poprzez wprowadzenie paginacji związanej też stronicowaniem. Maksymalna ilość pokazywanych wierszy została ograniczona do 300 i wprowadzono dodatkowy element widoczny w lewym dolnym

rogu rysunku numer 3, który umożliwia użytkownikowi poruszanie się po kolejnych stronach tabelki zmniejszając narzut pamięci operacyjnej wymaganej do wygenerowania całości tabeli.

Ważną funkcjonalnością z punktu widzenia użytkownika jest możliwość filtracji pobranych już wierszy z bazy danych. W celu zwiększenia możliwości wyszukiwania, każda kolumna posiada oddzielne pole filtrujące, umożliwiając na zawężanie zbioru danych po każdej kolumnie. Element odpowiadający za stronicowanie poprawnie zmniejsza ilość dostępnych stron przy dynamicznym zmniejszaniu się wyświetlanych danych. Dodatkową opcją zmieniającą wyświetlane dane po stronie przeglądarki jest funkcjonalność sortowania rosnąco bądź malejąco jednej kolumny. Służy do tego strzałka po lewej stronie od nazwy kolumny, kliknięcie w ikonkę bądź nazwę kolumny zmienia sortowanie.

Ustawianie widoczności kolumn Przejrzystość danych i dostęp tylko do potrzebnych informacji jest kluczową wartością dla użytkowników. Mnogość kolumn, porowadząca do przedstawianie wielu informacji nie potrzebnych wszystkim użytkownikom uniemożliwia osiągnięcie tego efektu. Wychodząc naprzeciw tym oczekiwaniom zaimplementowano opcję umożliwiającą klientom aplikacji ukrywanie dowolnej kolumny. Po kliknięciu w specjalny guzik nad tabelą z danymi, wyskakuje okienko z listą kolumn, które użytkownik może odznaczyć co spowoduje zniknięcie z tabeli. Selekcja może być zapisana w bazie danych tak by przy ponownym wejściu na tą stronę aplikacji, użytkownik nie musiał kolejny raz ukrywać nieinteresujących go kolumn. Po rejestracji użytkownik ma wybrane wszystkie kolumny i musi sam je odznaczyć.

Select visible columns

Chrom	<input checked="" type="checkbox"/>
Position	<input checked="" type="checkbox"/>
RSID	<input type="checkbox"/>
Reference	<input type="checkbox"/>
Alternate	<input type="checkbox"/>
Consequence	<input type="checkbox"/>
Protein Consequence	<input type="checkbox"/>
Transcript Consequence	<input type="checkbox"/>

Rysunek 4. Lista próbek

Filtrowanie danych Rysunek numer 5 przedstawia moduł filtrujący dane, znajdujący się w lewej części strony aplikacji.

Filtrowanie W górnej części panelu widoczne są przyciski z nazwami filtrów. Aktywny filtr wyróżnia się od zielonym kolorem od nieaktywnych szarych filtrów.

Za ustalanie struktury filtrów odpowiadają administratorzy i w podrozdziale im poświęconym opisane zostanie zarządzanie filtrami.

Filters

Tab 1**Tab 2****Tab 3**

Get all

Inactive ☒ Filter 11

Chrom greater than

23

Position greater

245345

Inactive ☐ Filter 12

Reference equals

C

Chrom less than

23

Allele Number greater

42920

Filter

Count

Save

Rysunek 5. Filtry

6.3 Panel administratora

Użytkownik o roli administratora może wejść do oddzielnej strony aplikacji (rysunek 6), gdzie może przesyłać plik o formacie xlsx o ustalonej strukturze i załadować nowe filtry. Ma też możliwość zmian widoczności próbek dla użytkowników oraz ich ról.



The screenshot shows the 'Admin' tab selected in a navigation bar with 'Home' and 'Transcript list'. Below the navigation bar is an 'Upload' button. A table with two columns, 'User name' and 'Role', lists three users: 'user1', 'cxz', and 'cxz2', all with the role 'user'.

User name	Role
user1	user
cxz	user
cxz2	user

Rysunek 6. Część panelu administratora z listą użytkowników

Zarządzanie filtrami

Zarządzanie widocznością próbek dla użytkowników Po wybraniu użytkownika z listy widocznej na rysunku numer 6, administratorowi ukazuje się lista wszystkich dostępnych próbek (rysunek numer 7). Administrator zaznacza w przyciskach wyboru, które próbki będą dostępne do wglądu wybranemu użytkownikowi. Po zarejestrowaniu się do aplikacji, użytkownik nie jest przypisany do żadnej próbki.

Home Transcript list Admin

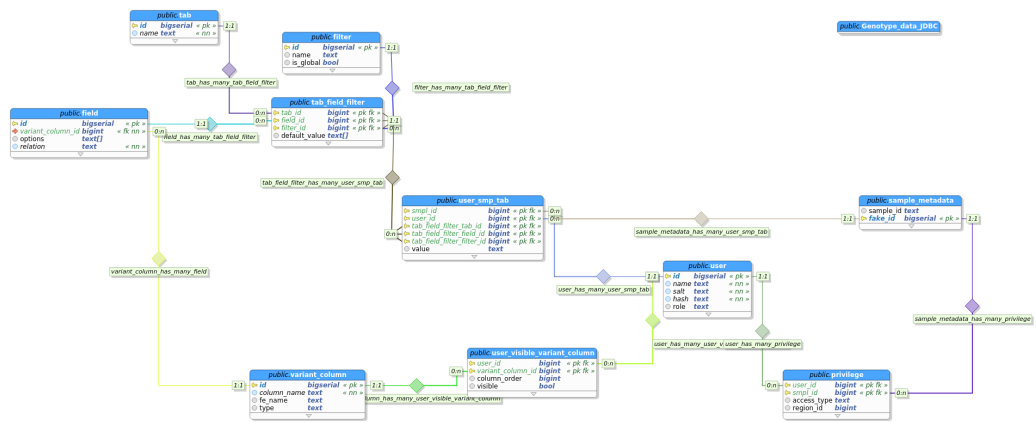
Sample id	Visible
sample_id_18	<input checked="" type="checkbox"/>
id_1	<input checked="" type="checkbox"/>
sample_id_2	<input checked="" type="checkbox"/>
sample_id_17	<input type="checkbox"/>
sample_id_3	<input type="checkbox"/>
sample_id_7	<input type="checkbox"/>
sample_id_10	<input type="checkbox"/>

Save

Rysunek 7. Panel dostępności próbek dla użytkownika

7 Schemat bazy danych

Schemat jest bardzo duży, może załączyć w częściach w dodatku i odsyłać tam czytelnika?



Rysunek 8. Schemat bazy danych

8 Opis implementacji

Czy nie połączyć opisu implementacji razem z przypadkami użycia? Kolejno opisując użycie aplikacji mógłbym opisać jak to się odbywa w kodzie.

9 Bezpieczeństwo aplikacji

9.1 Niebezpieczeństwa

9.2 Wykorzystanie protokołu https

9.3 SQL injection

Opisać jak od strony bardziej technicznej odbywa się zabezpieczanie haseł użytkownika (sól, sha512)

10 Testy oraz wydajność

11 Wnioski i podsumowania

Literatura

- [1] École Polytechnique Fédérale - Scala documentation, Available at: <http://docs.scala-lang.org/> (Accessed: 10 August 2017).
- [2] The Apache Software Foundation - Apache Spark Available at: <https://spark.apache.org/> (Accessed: 10 August 2017).
- [3] Exac Browser Data - Exome Aggregation Consortium Available at: <http://exac.broadinstitute.org/> (Accessed: 10 August 2017).
- [4] Hostovita sp. z o.o. - Porównanie relacyjnych SZBD: SQLite, MySQL, PostgreSQL Available at: <https://hostovita.pl/blog/porownanie-relacyjnych-systemow-zarzadzania-bazami-danych-sqlite-mysql-postgresql/> (Accessed: 10 August 2017).
- [5] Lightbend, Inc Slick documentation. Available at: <http://slick.lightbend.com/docs/> (Accessed: 10 August 2017).

Wykaz rysunków i tabel