Analiza realizacji pierwszego etapu wdrożenia Hurtowni Danych w firmie Horse Racing Analytics

Opracowanie:

Claudia Höpfl

Spis treści

[1. Wstęp 4](#_Toc65432220)

[1.1 Cele Hurtowni Danych 4](#_Toc65432221)

[1.2 Dyscyplina sportowa 4](#_Toc65432222)

[1.3 Zasady budowy hurtowni danych 4](#_Toc65432223)

[2. Harmonogram i organizacja projektu 5](#_Toc65432224)

[2.1 Lista zadań i harmonogram 5](#_Toc65432225)

[2.2 Struktura organizacyjna 6](#_Toc65432226)

[2.2.1 Nadzór Projektu 6](#_Toc65432227)

[2.2.2 Kierownik projektu po stronie Hopfl Inc 7](#_Toc65432228)

[2.2.3 Kierownik projektu po stronie Horse Racing Analytics 7](#_Toc65432229)

[2.3 Obszaru odpowiedzialności 7](#_Toc65432230)

[2.3.1 Obszar odpowiedzialności za dane źródłowe 7](#_Toc65432231)

[2.3.2 Obszar odpowiedzialności za integrację i uporządkowanie danych 7](#_Toc65432232)

[2.3.3 Obszar odpowiedzialności za nadawanie uprawnień 8](#_Toc65432233)

[2.4 Rodzaje dokumentów 8](#_Toc65432234)

[3. Aspekty Techniczne hurtowni danych 9](#_Toc65432235)

[3.1 Warstwa zasilania 9](#_Toc65432236)

[3.1.1 Szczegóły 9](#_Toc65432237)

[3.1.2 Ładowanie danych do bazy danych 10](#_Toc65432238)

[3.1.3 Oczyszczanie Danych 10](#_Toc65432239)

[3.1.4 Pozyskiwanie danych dodatkowych 10](#_Toc65432240)

[3.2 Wymagania biznesowe 11](#_Toc65432241)

[3.1.1 Wymagania funkcjonalne 11](#_Toc65432242)

[3.1.2 Wymaganie niefunkcjonalne 11](#_Toc65432243)

[3.1.3 Wymagane style analityczne 11](#_Toc65432244)

[3.1.4 interakcja z użytkownikiem 11](#_Toc65432245)

[4. Organizacja pracy 12](#_Toc65432246)

[4.1 Repozytorium projektu 12](#_Toc65432247)

[4.1.1 Zadania 12](#_Toc65432248)

[4.2 Raportowanie postępów prac 12](#_Toc65432249)

[4.3 Zakończenie projektu 13](#_Toc65432250)

[5. Wstępny model danych 14](#_Toc65432251)

# Wstęp

## Cele Hurtowni Danych

Firma Horse Racing Analytics zleca stworzenie Hurtowni Danych w celu wygodnego i przystępnego dostępu do raportów opartych o analizę danych dostarczonych z wyścigów konnych organizowanych na terenie miasta Hong Kong.

## Dyscyplina sportowa

Dyscyplina polegająca na bezpośrednim współzawodnictwie dwóch lub więcej dosiadających koni jeźdźców na określonym dystansie. Forma wyścigów może znacząco się różnić w obrębie wielu czynników: rasy koni dopuszczonych do gonitwy, rodzaju nawierzchni toru wyścigowego, pokonywanego dystansu czy chodu,   
w jakim podczas gonitwy winny poruszać się konie. Ważnym sektorem powiązanym ściśle z wyścigami konnymi i decydującym o ich roli ekonomicznej są zakłady bukmacherskie. Zawody odbywają się na hipodromach, czyli torach wyścigowych. Koni wyścigowych dosiadają doświadczeni jeźdźcy – dżokeje, którzy muszą być niewysocy i lekcy. By wyścigi były sprawiedliwe, przed startem waży się dżokeja z siodłem i jeśli dżokej waży za mało, wkłada się do siodła ciężarki, by wszystkie konie niosły ustalony ciężar.

Firma Horse Racing Analytics zajmuję się analizą oraz przewidywaniem rezultatów gonitw organizowanych na terenie Europy i Bliskiego Wschodu. Ze względu na ekspansje firmy na rynki azjatyckie, zleca wykonanie systemu, który pozwoli na dokładne poznanie tamtejszych trendów.

## Zasady budowy hurtowni danych

Jest kilka podstawowych zasad, na jakich oparty jest efektywny proces budowy hurtowni danych.

Należą do nich:

* zapewnienie strategicznego kontekstu analiz dostarczanych przez hurtownię danych
* przyrostowa, iteracyjna budowa systemu
* oparcie budowy składników hurtowni danych na definicjach architektury
* ścisła współpraca z użytkownikami

# Harmonogram i organizacja projektu

## Lista zadań i harmonogram

|  |  |  |  |
| --- | --- | --- | --- |
| **Nazwa Zadania** | **Osoby odpowiedzialne** | **Data rozpoczęcia** | **Data zakończenia** |
| Wdrożenie hurtowni Danych | Jan Kowalski | 16.01.2021 | 01.04.2021 |
| Prace analityczne |  |  |  |
| Specyfikacja interfejsów | Jan Kowalski | 17.01.2021 | 24.01.2021 |
| Analiza przedwdrożeniowa | Jan Kowalski | 24.01.2021 | 01.02.2021 |
| Prace wdrożeniowe |  |  |  |
| Interfejs Koń | Jan Kowalski | 02.02.2021 | 14.03.2021 |
| Interfejs Dżokej | Jan Kowalski | 02.02.2021 | 14.03.2021 |
| Interfejs Trener | Jan Kowalski | 02.02.2021 | 14.03.2021 |
| Interfejs Gonitwa | Jan Kowalski | 02.02.2021 | 14.03.2021 |
| Obszar Wyścig |  |  |  |
| Wykonanie analizy Wyniki wg. Sekcji | Jan Kowalski | 01.03.2021 | 14.03.2021 |
| Wykonanie analizy Dżokej | Jan Kowalski | 01.03.2021 | 14.03.2021 |
| Wykonanie analizy Trener | Jan Kowalski | 01.03.2021 | 14.03.2021 |
| Wykonanie ogólnej analizy wygranych | Jan Kowalski | 01.03.2021 | 14.03.2021 |

## 2.2 Struktura organizacyjna

### Nadzór Projektu

* Przegląd dokumentów projektu: harmonogramu, notatek ze spotkań, raportów okresowych, analizy
* Podejmowanie decyzji w sytuacji, gdy kierownik projektu (którykolwiek) sygnalizuje opóźnienie zadań, za które odpowiada

### Kierownik projektu po stronie Hopfl Inc

* Koordynowanie zadań wykonywanych po stronie HOPFL INC
* Uczestniczenie we wszystkich spotkaniach roboczych w Horse Racing Analytics a w sytuacji gdy jest to niemożliwe wyznaczanie zastępców
* Tworzenie i aktualizacja dokumentacji projektu: harmonogramu, notatek ze spotkań, raportów okresowych, analizy
* Ścisła współpraca z kierownikiem po stronie Horse Racing Analytics

### Kierownik projektu po stronie Horse Racing Analytics

* Koordynowanie zadań wykonywanych po stronie Horse Racing Analytics
* Koordynowanie zadań wykonywanych po stronie dostawców systemów IT, z którymi łączyć się będzie hurtownia
* Organizowanie spotkań roboczych w Horse Racing Analytics oraz uczestniczenie w nich,   
  a w sytuacji gdy jest to niemożliwe wyznaczanie zastępców
* Ścisła współpraca z kierownikiem po stronie HOPFL INC

## Obszaru odpowiedzialności

Kluczowe dla projektu wdrożenia hurtowni danych jest wyodrębnienie następujących obszarów

odpowiedzialności:

### Obszar odpowiedzialności za dane źródłowe

Odpowiedzialność po stronie KLIENTA lub dostawców systemów źródłowych. Dotyczy jakości   
i rzetelności danych wprowadzanych do systemów, z których później przekazywane są do hurtowni.

Hurtownia danych poprzez reguły walidacyjne zapewni odpowiednią jakość danych.

### Obszar odpowiedzialności za integrację i uporządkowanie danych

Odpowiedzialność po stronie ARCHMAN. Dotyczy wszelkiej obróbki danych źródłowych do postaci,   
w której są one przetwarzane i prezentowane użytkownikom końcowym.

### Obszar odpowiedzialności za nadawanie uprawnień

Odpowiedzialność po stronie KLIENTA. Dotyczy przyporządkowywania użytkowników końcowych   
do ról, bądź administrowania struktura organizacyjną z której wynikają uprawniania dynamiczne.

## 2.4 Rodzaje dokumentów

Głównym dokumentem projektu jest analiza przedwdrożeniowa, która jest cały czas uaktualniana.

Pozostałe dokumenty:

**Notatki ze spotkań**

Każde spotkanie jest dokumentowane notatką wysyłaną mailem do uczestników oraz kierownictwa projektu.

**Raporty miesięczne**

Do 5 dnia kolejnego miesiąca sporządzany jest raport miesięczny, który zawiera:

* Omówienie harmonogramu (co się udało zrobić, czego się nie udało),
* Aktualizację analizy przedwdrożeniowej (wraz z harmonogramem),
* Listę kluczowych zagrożeń dla projektu (rejestr zagrożeń).

**Rejestr zagrożeń**

Lista spraw, które mogą zagrozić terminowej realizacji projektu. Każdej ze spraw przypisywana jest

waga i stopień prawdopodobieństwa wystąpienia. W przypadku zagrożeń przekraczających określony

próg, informowany jest nadzór projektu.

# 3. Aspekty Techniczne hurtowni danych

## 3.1 Warstwa zasilania

### 3.1.1 Szczegóły

Dane będą pochodzić z darmowych źródeł takich jak strona <https://www.kaggle.com/gdaley/hkracing> i są prezentowane w formacie CSV. Pliki analizy posiadają informacje o koniach, gonitwach, szczegółowe wyniki wyścigów. Dane są prezentowane w języku angielskim, a system i raporty będą w języku polskim.

Races.csv zawiera dane o indywidualnym wyścigu:

* Id wyścigu
* Data
* Miejsce wyścigu
* Konfiguracja
* Nawierzchnia toru
* Dystans
* Ocenę koni jakie mogą brać udział w gonitwie
* Nagrodę
* Klasę wyścigu
* Czasy lidera poszczególnych sekcji
* Pozycje 4 pierwszych koni
* Nagrody dla 4 pierwszych koni
* Wygrany koń

Runs.csv zawiera dane konkretnego konia biorącego udział w danym wyścigu, takie jak:

* Id wyścigu
* Numer konia
* Id konia
* Pozycja końcowa
* Czy wygrał
* Pozycja za liderem
* Wiek konia
* Pochodzenie konia
* Płeć konia
* Ocenę konia wg. Rankingu HKJC
* Deklarowana waga
* Aktualna waga
* Postposition
* Pozycję w każdej sekcji
* Pozycję za liderem w każdej sekcji
* Czas każdej sekcji
* Czas końcowy
* Przewidywane szanse na wygraną
* Id Trenera
* Id Dżokeja

### 3.1.2 Ładowanie danych do bazy danych

Ze względu na posiadane zasoby klienta dane będą przechowywane w systemie SQL Server. Przy użyciu SQL Server Agent będziemy ładować dane do bazy danych raz w tygodniu. Dane do systemu będą w postaci plików płaskich CSV. Struktura danych będzie taka sama jak w plika CSV, a dalsza obróbka danych nastąpi w procesie ETL.

### 3.1.3 Oczyszczanie Danych

Dane pochodzące ze strony Kaggle występują w formie oczyszczonej i nie zachodzi potrzeba zamiany struktury tych danych. Dane są przygotowane do analizy i uczenia maszynowego w prawidłowej formie.

### 3.1.4 Pozyskiwanie danych dodatkowych

Wdrażany proces opiera się tylko na danych pozyskanych do analizy i nie zachodzi potrzeba pozyskiwania danych z innych zewnętrznych źródeł.

## 3.2 Wymagania biznesowe

### 3.1.1 Wymagania funkcjonalne

1. Tworzenie raportów osiągnięć koni wg. trenerów
2. Tworzenie raportów osiągnięć koni wg. dżokeja
3. Tworzenie raportów prezentujących szczegółowe parametry koni: czasy w poszczególnych sekcjach, wpływ wagi na osiągany czas, odległości za liderem gonitw.
4. Tworzenie raportów danych historycznych koni zawierające informacje o osiągnięciach, wygenerowanych przychodach, najlepszych koniach w każdej kategorii.
5. Tworzenie raportu wpływu powierzchni toru na osiągi w poszczególnych sekcjach wyścigu.
6. Możliwość przeglądania danych wg. kategorii
7. Możliwość sortowania oraz grupowania danych wg. różnych parametrów
8. Możliwość archiwizacji danych
9. Raporty powinny być prezentowane w programie Power Bi ale z możliwością eksportu danych do programu Excel w celu własnej analizy

### 3.1.2 Wymaganie niefunkcjonalne

1. Raporty powinny się generować w sposób płynny i nieuciążliwy dla użytkownika
2. Interface raportów ma być przejrzysty i łatwy w obsłudze
3. Czas generowania raportu ma być krótszy niż 40 sekund

### Wymagane style analityczne

Nie wymagane jest zastosowanie rozbudowanych stylów analitycznych. Generowane raporty będą prezentowane tylko pracownikom działu analitycznego firmy HORSE RACING ANALYTICS. Użytkownik systemu będzie tworzył, czytał i analizował raporty.

### 3.1.4 interakcja z użytkownikiem

Użytkownik końcowy systemu ma mieć możliwość interakcji z raportami. Przez interakcje rozumiemy możliwość filtrowania danych, grupowania wg. przygotowanych filtrów, tak aby dostosować analizę końcową wg. własnych potrzeb.

# 4. Organizacja pracy

## 4.1 Repozytorium projektu

Projekt jak i wszystkie zasoby będą dostępne na stronie GitHub. Dostęp do zasobów projektu będzie ograniczony dla klienta. Każdy z uczestników projektu będzie miał dostęp do repozytorium, będzie mógł raportować wykonana prace oraz zarządzać przepływem pracy. Strona pozwala na tworzenie zadań do wykonania co pozwoli na szczegółowe zaplanowanie wykonywanego projektu. Dodatkową zaletą jest możliwość tworzenia „problemów”, na które będzie można reagować w szybki sposób, co pozwoli na uniknięcie opóźnień.

### 4.1.1 Zadania

Każde zadanie w repozytorium powinno uwzględniać:

- nazwę zadania

- opis zadania

- osobę odpowiedzialną za wykonanie

- planowaną datę rozpoczęcia i zakończenia zadania

- status na jakim etapie jest dane zadanie

## 4.2 Raportowanie postępów prac

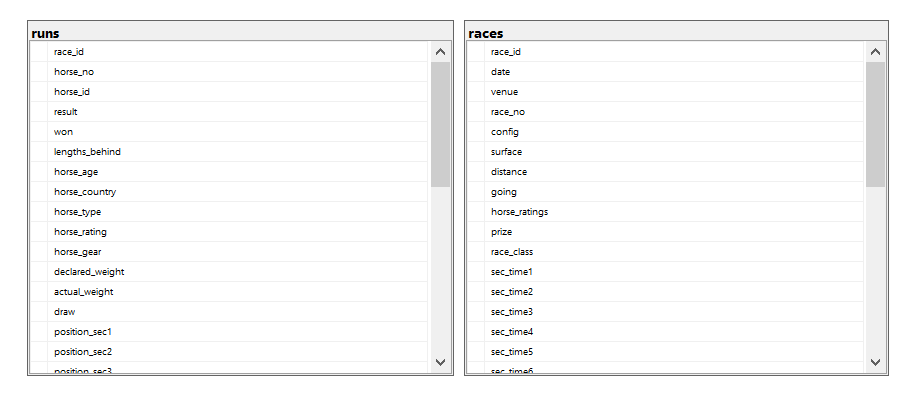
Klient będzie mieć wgląd do strony GitHub, gdzie będzie mieć wgląd do aktualnego statusu każdego zadania. Dodatkowo będą planowane cotygodniowe spotkania w formie internetowej konferencji. Podczas spotkania będzie omawiany bieżący status prac, możliwe zagrożenia i opóźnienia. Spotkania będą odbywać się raz w tygodniu przy użyciu platformy Ms Teams. Termin spotkań nie został jeszcze określony.

## 4.3 Zakończenie projektu

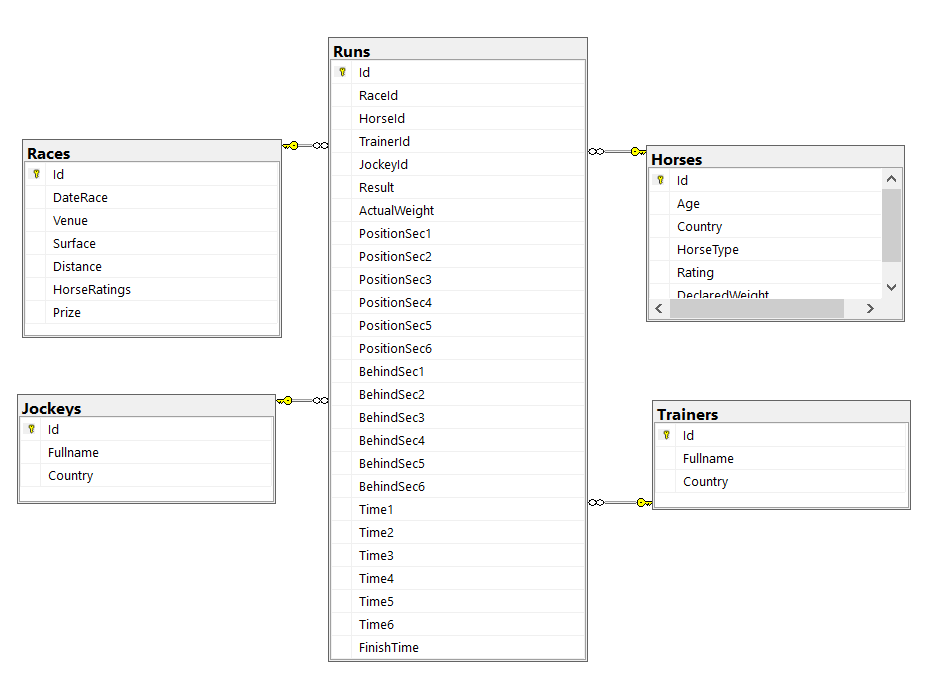
Projekt zostaje uznany za zakończony w momencie zrealizowania wszystkich wymaganych funkcjonalności. System musi spełniać wszystkie wyżej wymienione wymagania oraz musi być w pełni przetestowany. Koniec projektu będzie skutkować spotkaniem na którym będzie omówiony przebieg prac oraz możliwy rozwój systemu w przyszłości.

# 5. Wstępny model danych

Dane pochodzące z plików płaskich po załadowaniu do bazy danych występują w następującej formie dwóch tabel:



Dane z bazy danych zostaną przeładowane do hurtowni danych z wykorzystaniem SQL Server Integration Services. Dane w hurtowni będą miały strukturę płatka śniegu zaprezentowaną poniżej.



Poniżej efekt prostej analizy w programie Excel wykonanej z danych pobranych z Hurtowni danych

