Requirements specification for train arrival process

1. Opis procesu

Przejazd pociągów

a. Ogólny opis procesu biznesowego i opis wskaźników wydajności generowanych przez ten proces, możliwe bieżące problemy analityczne.

Pociągi wyruszają ze stacji startowej zgodnie z rozkładem jazdy i jadą do punktu docelowego mijając po drodze różne inne stacje. Na początku trasy odnotowywane są dane motorniczego i pociągu, który prowadzi. Na każdej stacji automatycznie odnotowywany jest czas przybycia pociągu porównany z wartością w rozkładzie, co pozwala na obliczenie opóźnienia/przyspieszenia pociągu i zapisania tej wartości w bazie. W przypadku zajścia zdarzenia, które spowalnia pociąg, dane o tym zdarzeniu są zapisywane wraz ze szczegółami zajścia, co pozwala na ich późniejszą analizę. W celu zmniejszenia opóźnień organizacja:

Stawia za cel zmniejszenie ilości incydentów w porównaniu z rokiem poprzednim o 5%.

Drugim celem jest zmniejszenie opóźnień wywołanych incydentami na drodze pociągu o 3% w porównaniu z poprzednim miesiącem.

b. Typowe pytania

Jakie czynniki (pogoda, incydenty, typ pociągu, motorniczy) mają największy wpływ na opóźnienia?

Które fragmenty trasy generują największe średnie opóźnienia?

Ilu kursów nie ukończono z powodu incydentów w tym roku?

Czy w ostatnich latach zmniejszyła się liczba opóźnionych pociągów?

Czy lata doświadczenia i płeć maszynisty wpływają na opóźnienia pociągu?

Podaj średnie opóźnienie pociągów Intercity na przestrzeni pełnych przejazdów.

Na jakim poziomie była punktualność pociągów Polregio w pierwszym kwartale 2025 roku?

Jakiego typu wypadki generują średnio największe opóźnienie?

c. Dane

Dane dotyczące kursów pociągów, motorniczych dostępne są w bazie danych. Były one zbierane podczas poprzednich przejazdów pociągów, a dane dotyczące incydentów są uzupełniane na podstawie wypowiedzi maszynisty, bądź odpowiednich służb. Dane pogodowe są dostępne z API

pogodowego, które na podstawie lokalizacji pociągu i daty generują wiersz pliku csv, powiązujący odcinek trasy z pogodą (temperatura, opady...) która wtedy była. Wiersz ten jest dopisywany do pliku weather.csv

2. Źródła danych

a)plik weather.csv

Plik weather.csv z API pogodowego zawiera dla każdego odcinka trasy przybliżone dane pogodowe, pierwszy wiersz zawiera nagłówki, następnie następuje pewna liczba wierszy w formacie:

id_odcinka, data_pomiaru, temperatura, ilosc_opadow, typ_opadow

id_odcinka – dla jakiego fragmentu przejazdu, został wykonany pomiar

data_pomiaru – timestamp, dokładna data pomiaru

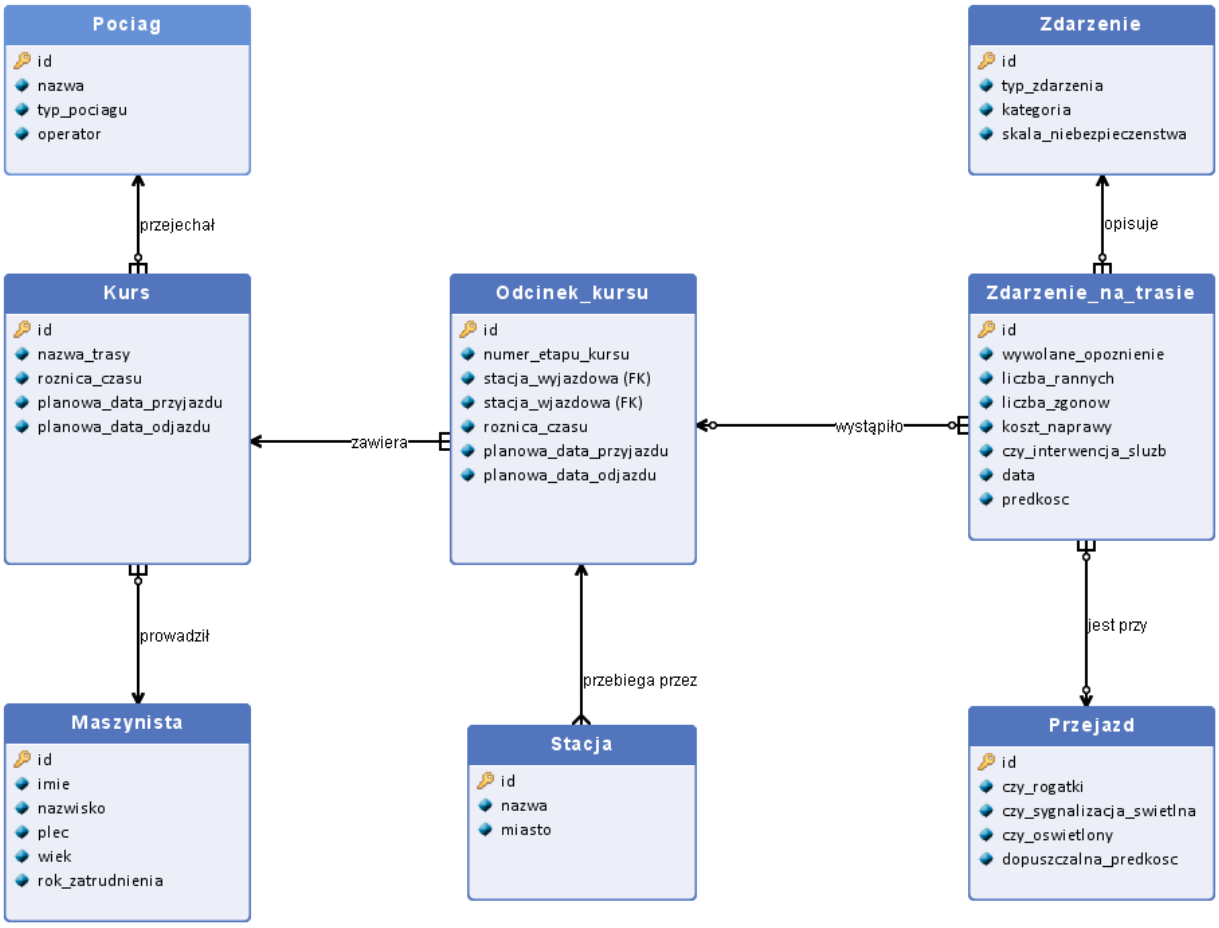
temperatura – double (jedno miejsce po przecinku) : liczba stopni Celsjusza

ilosc_opadow – int: liczba opadów w [mm/h]

typ_opadow – string: deszcz, śnieg lub grad lub brak

b) relacyjna baza danych

ERD Diagram


Pociag
id
nazwa
typ_pociagu
operator
przejechał
Kurs
id
nazwa_trasy
roznica_czasu
planowa_data_przyjazdu
planowa_data_odjazdu
zawiera
Odcinek_kursu
id
numer_etapu_kursu
stacja_wyjazdowa (FK)
stacja_wjazdowa (FK)
roznica_czasu
planowa_data_przyjazdu
planowa_data_odjazdu
wystąpiło
Zdarzenie
id
typ_zdarzenia
kategoria
skala_niebezpieczenstwa
opisuje
Zdarzenie_na_trasie
id
wywolane_opoznienie
liczba_rannych
liczba_zgonow
koszt_naprawy
czy_interwencja_sluzb
data
predkosc
prowadził
Maszynista
id
imie
nazwisko
plec
wiek
rok_zatrudnienia
przebiega przez
Stacja
id
nazwa
miasto
jest przy
Przejazd
id
czy_rogatki
czy_sygnalizacja_swietlna
czy_oswietlony
dopuszczalna_predkosc

Opis zbioru encji

Pociag			
Jeżdżące po terenie Polski pociągi różnych operatorów.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK
nazwa	Nie	varchar (20)	nazwa pociągu np. ICC4302, EIP12345
typ_pociagu	Nie	varchar(30)	typ pociągu np: passenger, cargo
operator	Nie	varchar(40)	nazwa operatora np: PKP Intercity, POLREGIO, PKP Cargo lub DB Cargo Polska

Maszynista			
Zbiór maszynistów prowadzących różne pociągi w trakcie kursów.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK
imie	Nie	varchar(30)	imię maszynisty
nazwisko	Nie	varchar(30)	nazwisko maszynisty
plec	Nie	varchar(10)	płeć: man lub woman
wiek	Nie	int	wiek maszynisty
rok_zatrudnienia	Nie	date	rok rozpoczecia pracy jako maszynista

Stacja			
Poszczególne stacje na terenie Polski przez które przejeżdżają pociągi.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK
nazwa	Nie	varchar (40)	nazwa stacji np: Gdańsk Wrzeszcz, Gdańsk Główny
miasto	Nie	varchar(40)	miasto w którym znajduje się stacja np: Gdańsk

Kurs			
Pełny przejazd od stacji początkowej do końcowej o określonej godzinie.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK

nazwa_trasy	Nie	varchar(40)	nazwa trasy kursu np: Gedania, Sudety
roznica_czasu	Nie	int	opóźnienie w minutach, lub przyspieszenie jeśli wartość jest ujemna
planowa_data_przyjazdu	Nie	TIMESTAMP	kiedy pociąg miał planowo dojechać
planowa_data_odjazdu	Nie	TIMESTAMP	o której pociąg miał planowo zacząć kurs

Przejazd			
Lokalizacja przejazdu kolejowo-samochodowego, przy którym wystąpił wypadek, bądź inne zdarzenie. Zdarzenie nie musi być powiązane z przejazdem, gdy np. pociąg z jakiś powodów nie mógł ruszyć.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK
czy_rogatki	Nie	boolean	czy przejazd ma zamontowane rogatki
czy_sygnalizacja_swietlna	Nie	boolean	czy przejazd ma zamontowaną sygnalizacje świetlną
czy_oswietlony	Nie	boolean	czy przejazd jest oświetlony
dopuszczalna_predkosc	Nie	int	prędkość dopuszczalna dla pojazdów przy tym przejeździe kolejowym

Zdarzenie			
Zbiór możliwych zdarzeń podczas jazdy pociągiem.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	int autoinkrement	PK
typ_zdarzenia	Nie	varchar (30)	Ogólna klasyfikacja zdarzenia: wypadek, incydent, awaria, zdarzenie techniczne
kategoria	Nie	varchar (40)	Dokładny typ zdarzenia: potrącenie pieszego, wykolejenie, zderzenie z innym pociągiem, wyłamani rogatek, przerwa w zasilaniu...
skala_niebezpieczenstwa	Nie	int	skala od 1 do 10 zagrożenia

			wynikającego z zdarzenia

Zdarzenie_na_trasie			
Incydenty, które miały miejsce pomiędzy dwiema stacjami.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	bigint autoinkrement	PK
wywolane_opoznienie	Nie	int	minuty opóźnienia wywołanego zdarzeniem
liczba_rannych	Nie	int	liczba rannych osób
liczba_zgonow	Nie	int	liczba zmarłych osób
koszt_naprawy	Nie	double (do dwóch miejsc po przecinku)	szacowana liczba złotych na naprawę
czy_interwencja_sluzb	Nie	bool	czy musiały interweniować służby ratownicze
data	Nie	TIMESTAMP	dokładna data zdarzenia
predkosc	Nie	int	średnia prędkość pociągu na chwilę przed zdarzeniem [km/h]

Odcinek_kursu			
Odcinek kursu, od jednego do drugiego przystanku.			
Nazwa	**Klucz główny**	**Typ/Dziedzina**	**Opis**
id	**Tak**	bigint autoinkrement	PK
numer_etapu_kursu	Nie	int : kolejne liczby	kolejne liczby reprezentujace, który to przystanek od początku
stacja_wyjazdowa (FK)	Nie	int	id stacji z której pociąg odjeżdża
stacja_wjazdowa (FK)	Nie	int	id stacji do której pociąg dojeżdża
roznica_czasu	Nie	int	opóźnienie w minutach w porównaniu z planowym przyjazdem, lub przyspieszenie jeśli wartość jest ujemna
planowa_data_przyjazdu	Nie	TIMESTAMP	data o której pociąg miał planowo przyjechać na

			przystanek
planowa_data_odjazdu	Nie	TIMESTAMP	data o której pociąg miał planowo ruszyć z przystanku

3. Scenariusze problemów analitycznych

Dlaczego w tym roku spóźniło się tak wiele pociągów, o tak długi czas?

1. Porównaj średnie opóźnienie z całych kursów dla każdego operatora pociągów w tym roku.
2. Kursy prowadzone przez kogo średnio spóźniają się dłużej – te prowadzone przez doświadczone (> 5 lat w zawodzie) w zawodzie kobiety, czy prowadzone przez niedoświadczonych (< 3 lat w zawodzie) mężczyzn?
3. Podaj listę TOP 10 stacji, przy których średnie opóźnienie pociągu jest największe.
4. Ile było takich odcinków kursów w tym miesiącu, które mimo zajścia incydentu podczas nich skończyły się o czasie (z dokładnością do minuty)?
5. Dla ilu odcinków tras podczas których padał śnieg opóźnienie wynosiło mniej niż 3 minuty, a dla ilu więcej niż 30 minut? (**dane z obu źródeł**)

Jakie są przyczyny i skutki zajścia tak wielu incydentów na przejazdach kolejowych?

1. Porównaj, liczbę wypadków na przejazdach kolejowych ze światłami i rogatkami, jak i bez nich.
2. Którego operatora pociągi najczęściej są dotknięte awariami na trasie?
3. Pomiędzy którymi stacjami doszło do zranienia i śmierci największej liczby osób?
4. Jakiego typu zdarzenia generują średnio najmniejsze dodatkowe opóźnienia?
5. Ile najwięcej ludzi znajdowało się w pociągu, podczas zdarzenia: wykolejenie? (**Potrzebne dodatkowe dane o orientacyjnej liczbie ludzi w trakcie przejazdów**)
6. Podczas, jak wielu wypadków padał mocny deszcz (więcej niż 7,5 mm / h), a pociąg jechał więcej niż 100 km/ h? (**dane z obu źródeł**)
7. Czy kursy, na które ceny biletów są najdroższe, gwarantują większe bezpieczeństwo (mniejszą średnią liczbę niebezpiecznych zdarzeń)? (**Potrzebne dodatkowe dane o cenie biletów, potrzeba zmiany procesu biznesowego**).

4. Dane potrzebne do problemów analitycznych

Dlaczego w tym roku spóźniło się tak wiele pociągów, o tak długi czas?

1. Porównaj średnie opóźnienie z całych kursów dla każdego operatora pociągów w tym roku.
   - **Operator** – baza danych, tabela Pociag, kolumna operator
   - **Opóźnienie** – baza danych, tabela Kurs, kolumna różnica czasu
   - **Rok z daty przyjazdu pociągu** – baza danych, tabela Kurs, kolumna planowa_data_przyjazdu
2. Kursy prowadzone przez kogo średnio spóźniają się dłużej – te prowadzone przez doświadczone (> 5 lat w zawodzie) w zawodzie kobiety, czy prowadzone przez

niedoświadczonych (< 3 lat w zawodzie) mężczyzn?

- **Płeć** – baza danych, tabela Maszynista, kolumna plec
- **Lata doświadczenia** – baza danych, tabela Kurs, różnica obecnego roku z rokiem w kolumnie rok_zatrudnienia
- **Opóźnienie** – baza danych, tabela Kurs, kolumna roznica_czasu

3. Podaj listę TOP 10 stacji, przy których średnie opóźnienie pociągu jest największe.

- **Nazwa stacji** – baza danych, tabela Stacja, kolumna nazwa
- **Opóźnienie** – baza danych, tabela Odcinek_kursu, kolumna roznica_czasu

4. Ile było takich odcinków kursów w tym miesiącu, które mimo zajścia incydentu podczas nich skończyły się o czasie (z dokładnością do minuty)?

- **Różnica czasu** – baza danych, tabela Odcinek_kursu, kolumna roznica_czasu
- **Id zdarzenia** (bo trzeba sprawdzić, czy wystąpiło) – baza danych, tabela Zdarzenie_na_trasie, kolumna id
- **Miesiąc dojazdu pociągu –** baza danych, tabela Odcinek_kursu, miesiąc z planowanej_daty_przyjazdu

5. Dla ilu odcinków tras podczas których padał śnieg opóźnienie wynosiło mniej niż 3 minuty, a dla ilu więcej niż 30 minut? (**dane z obu źródeł**)

- **Id odcinka** (żeby zliczyć ile ich było) – baza danych, tabela Odcinek_kursu kolumna id
- **Opóźnienie** – baza danych, tabela Odcinek_kursu, kolumna roznica_czasu
- **Typ opadów = śnieg** – dane z pliku weather.csv, kolumna typ_opadow

Jakie są przyczyny i skutki zajścia tak wielu incydentów na przejazdach kolejowych?

1. Porównaj, liczbę wypadków na przejazdach kolejowych ze światłami i rogatkami, jak i bez nich.

- **Czy na przejeździe były rogatki** – baza danych, tabela Przejazd, kolumna czy_rogatki
- **Czy na przejeździe były światła** – baza danych, tabela Przejazd, kolumna czy_sygnalizacja_swietlna
- **Id zdarzenia** (żeby dało się zliczyć) – baza danych, tabela Zdarzenia_na_trasie, kolumna id
- **Typ zdarzenia = wypadek** - baza danych, tabela Zdarzenie, kolumna typ_zdarzenia

2. Którego operatora pociągi najczęściej są dotknięte awariami na trasie?

- **Operator pociągu** - baza danych, tabela Pociag, kolumna operator
- **Typ zdarzenia = awaria** - baza danych, tabela Zdarzenie, kolumna typ_zdarzenia

3. Pomiędzy którymi stacjami doszło do zranienia i śmierci największej liczby osób?

- **Liczba rannych** – baza danych, tabela Zdarzenie_na_trasie, kolumna liczba_rannych
- **Liczba zgonów** – baza danych, tabela Zdarzenie_na_trasie, kolumna liczba_zgonow
- **Stacja_wjazdowa** – baza danych, tabela Odcinek_kursu, kolumna stacja_wjazdowa
- **Stacja_wyjazdowa** – baza danych, tabela Odcinek_kursu, kolumna stacja_wyjazdowa

4. Jakiego typu zdarzenia generują średnio najmniejsze dodatkowe opóźnienia?

- **Typ zdarzenia** - baza danych, tabela Zdarzenie, kolumna typ_zdarzenia
- **Dodatkowe opóźnienie wywołane zdarzeniem** – baza danych, tabela Zdarzenia_na_trasie, kolumna wywolane_opoznienie
- **Id zdarzenia** (żeby dało się zliczyć) – baza danych, tabela Zdarzenia_na_trasie, kolumna

id

5. Ile najwięcej ludzi znajdowało się w pociągu, podczas zdarzenia: wykolejenie? (**Potrzebne dodatkowe dane o orientacyjnej liczbie ludzi w trakcie przejazdów**)

- **Typ zdarzenia = wypadek** - baza danych, tabela Zdarzenie, kolumna typ_zdarzenia
- **Kategoria zdarzenia = wykolejenie** - baza danych, tabela Zdarzenie, kolumna kategoria
- Nie ma danych w bazie ani pliku csv na temat ilości osób w pociągu w trakcie jazdy. Proponowane rozwiązanie to oszacowanie liczby na podstawie ilości sprzedanych biletów i dołączenie tych danych do bazy do np. tabeli Odcinek_kursu

6. Podczas, jak wielu wypadków padał mocny deszcz (więcej niż 7,5 mm/h), a pociąg jechał więcej niż 100 km/ h? (**dane z obu źródeł**)

- **Typ opadów = deszcz** – dane z pliku weather.csv, kolumna typ_opadow
- **Ilość opadów** – dane z pliku weather.csv, kolumna ilosc_opadow
- **Prędkość pociągu przed zdarzeniem** – baza danych, tabela Zdarzenia_na_trasie, kolumna predkosc
- **Typ zdarzenia = wypadek** - baza danych, tabela Zdarzenie, kolumna typ_zdarzenia

7. Czy kursy, na które ceny biletów są najdroższe, gwarantują większe bezpieczeństwo (mniejszą średnią liczbę niebezpiecznych zdarzeń)? (**Potrzebne dodatkowe dane o cenie biletów, potrzeba zmiany procesu biznesowego**).

- **Ustalenie zdarzenia za niebezpieczne, gdy skala niebezpieczeństwa > 5, w przeciwnym razie można uznać zdarzenie za stosunkowo bezpieczne** - baza danych, tabela Zdarzenie, kolumna skala_niebezpieczenstwa
- **Id zdarzenia do liczenia średniej** - baza danych, tabela Zdarzenia_na_trasie, kolumna id
- Nie ma danych w bazie ani pliku csv na temat cen biletów. Proponowane rozwiązanie to stworzenie tabeli Ceny_biletów z nazwą trasy na podstawie cen np. ze strony Koleo i dołączenie jej do tabeli Kurs właśnie przez pole nazwa_trasy. Wtedy będzie można znaleźć odpowiedź na pytanie.
  Kolumna A: nazwa_trasy (nazwa trasy kursu)
  Kolumna B: cena_biletu (dla tego zapytania wystarczy cena za bilet normalny na pełną trasę)