Hive

Dokumentacja Hive:
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML

Połączenie beeline:

```
!connect jdbc:hive2://<host>:<port>/<baza> <user> <haslo>
!connect jdbc:hive2://<adminsk02...>:10000 jpodeszwik jpodeszwik
!connect jdbc:hive2://<adminsk02...>:10000/xyz jpodeszwik jpodeszwik
```

Utworzenie bazy

```
create database xyz;
```

Przełączenie bazy

```
use xyz;
```

Utworzenie tabeli i załadowanie danych:

```
CREATE TABLE transfers(
                src STRING,
                dst STRING,
                amount INT,
                date STRING
        )
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ","
        STORED AS TEXTFILE;

LOAD DATA INPATH "/user/vagrant/transfers" INTO TABLE transfers;
```

Utworzenie tabeli w konkretnej lokalizacji (która może już istnieć):

```
CREATE TABLE transfers2(
                src STRING,
                dst STRING,
                amount INT,
                date STRING
        )
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ","
        STORED AS TEXTFILE
        LOCATION "/user/xyz/transfers_table";
```

Utworzenie tabeli external:

```
CREATE EXTERNAL TABLE transfers3 <reszta_polecenia>;
```

Zmiana delimitera:

```
create table transfers4
row format delimited
fields terminated by ";"
as select * from transfers;
```

Zmiana formatu danych:

```
create table transfers5
stored as orc
as select * from transfers;
```

Włączenie kompresji:

```
set hive.exec.compress.output=true;
set mapreduce.output.fileoutputformat.compress=true;
set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.GzipCodec;
set mapreduce.output.fileoutputformat.compress.type=BLOCK;
```

Odblokowanie dynamicznego wyznaczania partycji:

```
SET hive.exec.dynamic.partition=true;
```

Odświeżenie danych o partycjach:

```
MSCK REPAIR TABLE <tabela>;
```

```
CREATE TABLE transfers5(
                src STRING,
                dst STRING,
                amount INT,
                date STRING
        )
        partitioned by (log_time string)
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ","
        STORED AS TEXTFILE
        LOCATION "/user/xyz/events";
```

Wyświetlenie danych na temat tabeli:

```
show create table transfers;
```

Zadania

1. Zaloguj się na hue. Utwórz sobie użytkownika '<login>' i korzystaj z niego w dalszych poleceniach.
2. Utwórz bazę <login> i używaj jej przy dalszych poleceniach.
3. Utwórz tabelę transfers. Wrzuć do niej dane z pliku transfers.
4. Utwórz tabelę typu 'external' i załaduj do niej takie same dane.
5. Zdropuj obie tabele i zobacz co stało się z danymi na hdfsie.
6. Utwórz tabelę w formacie ORC. Spróbuj poleceniem 'hdfs dfs -cat ...' wypisać zawartość plików i zobacz, że są binarne.
7. Posumuj pole 'amount' po koncie źródłowym.
8. Utwórz tabelę 'owners' z logów załadowanych sqoopem
9. Utwórz tabelę 'named_transfers' będącą wynikiem zjoinowania tabel 'owners' i 'transfers', tzn zamiast src i dst będzie zawierać pola from i to w których będą imiona.
10. Utwórz tabelę partycjonowaną po id konta źródłowego.
11. Utwórz tabelę partycjonowaną po log_time z katalogów, które wrzucił Flume.