Hive

Loading data to Hadoop

We will work with records of hourly temperatures at weather stations in the USA. The file with data is on the local filesystem:
`/home/pascepet/fel_bigdata/data/teplota-usa.zip`
We need to get this data on Hadoop (i. e., on HDFS). So:

- copy data to your user directory (I recommend to make a new directory for it);
- unzip;
- look into some lines of unzipped data, find the total number of lines (why are we doing it?);
- make a subdirectory on HDFS in your user directory (why do we need a subdirectory?);
- copy unzipped files from the local filesystem to HDFS (to the directory just created).

Starting of Hive shell

```
beeline -u "jdbc:hive2://hador-c1.ics.muni.cz:10000/default;principal=hive/hador-c1.ics.muni.cz@ICS.MUNI.CZ"
```

Create your database (if not created yet), give it the same name as your username. Then switch to your database (USE command).

1. Input data as a temporary, external table

1.1. Create an external table *temperature_ext* from files which you have loaded to HDFS.

- format *textfile*
- fields delimited by ","
- rows delimited by "\n"
- first line contains headers – to be skipped
- fields in the file:

Field	Type	Description
station	string	station code
month	int	month number
day	int	day in month number
hour	int	hour number (1–24)
degrees	int	temperature as round(10×°F)
flag	string	code of data quality
latitude	double	GPS latitude (negative = southern, positive = northern)
longitude	double	GPS longitude (negative = western, positive = eastern)
elevation	double	position above the sea level in meters
state	string	US state code (incl. dependent territories)
name	string	station name

1.2. We can do SQL queries in the external table. Do some check:

- list some lines (records) of *temperature_ext* table, compare with input data;
- find total numbers of lines (records) and compare with input data (should not be exactly the same – why?);
- find number of lines with NULL values at the field *degrees* (should be only a small part of all lines).

2. Trasfer to the optimized table

2.1. Create an empty internal (managed) table *temperature* with other format and compression:

- format parquet
- compression Snappy (it's necessary to type it by uppercase: SNAPPY)
- we will save degrees as a decimal number, therefore use *double* type

2.2. Insert data from the table *temperature_ext* to the table *temperature*:

- transform degrees from 10×°F to °C;
- all other fields transform as they are (no change).

2.3. Check the table *temperature*:

- Write out some rows.
- Find total number of records in the table *temperature* and compare it with total number of records in the table *temperature_ext*.

2.4. The table *temperature* is internal, so Hive is its owner.

- Look for it on HDFS under `/user/hive/warehouse/database_name.db` and find the volume (megabytes).
- Compare the volume with the volume of external table (data loaded on HDFS by you).

2.5. Drop the external table *temperature_ext*. Check that the table is no longer in your database but data are still on HDFS.

3. Tabulka s partitions

3.1. Vytvořte si prázdnou interní (managed) tabulku *teplota_part*, která bude stejná jako tabulka *teplota* (tj. pole, jejich typy, formát, komprese), ale bude mít navíc partitioning podle měsíce. (Pozor na pořadí polí!)

3.2. Do tabulky *teplota_part* zkopírujte data z tabulky *teplota*, při kopírování vytvořte dynamický partitioning podle měsíce. Dynamický partitioning je potřeba předem povolit pomocí příkazů:

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

3.3. Najděte tabulku *teplota_part* na HDFS pod `/user/hive/warehouse/jmeno_vasi_databaze.db` a prohlédněte si, jak je partitioning realizován.

4. Dotazování nad Hive

Budeme pracovat s tabulkou *teplota*.

4.1. Kolik unikátních stanic je v datech? *(457)*

4.2. Která stanice je nejsevernější? *(USW00027502, BARROW POST ROGERS AP)*

4.3. Který stát má nejvíc unikátních stanic? *(TX)*

4.4. Kolik hodinových údajů celkem je na některé ze stanic v severní Dakotě (ND) nižších než −10°C? *(8 446)*

4.5. Který stát má nejvyšší celkovou průměrnou teplotu na svých stanicích za letní měsíce (6, 7, 8)? *(MH, 28.1)*

4.6. Které všechny státy mají rozdíl zeměpisných šířek (longitude) mezi svou nejzápadnější a nejvýchodnější stanicí větší než 8 stupňů? *(AK, FM, MT, TX)*

4.7. Pro každou stanici s nadmořskou výškou nad 1 500 metrů zjistěte rozdíl mezi celkovou průměrnou teplotou stanice (za celý rok) a celkovou průměrnou teplotou státu, kam stanice patří.

stanice	**rozdil**
USW00003103	-9.732538702083291
USW00023225	-4.890214469033042
...	...