LAPORAN

RENCANA TUGAS MANDIRI (RTM) Ke-4

MATA KULIAH BIG DATA C

"QUERY STATISTIK DESKRIPTIF MENGGUNAKAN HIVE"

DISUSUN OLEH:

Mohamad Ibnu Fajar Maulana (21083010106)

DOSEN PENGAMPU:

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR

2022

1. Analisis pada dataset NOAA menggunakan Hive
   - Jalankan Hive dengan seperti kode dibawah ini

```
$ hive
```

```
[oracle@bigdatalite ~]$ hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.13.1.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
```

   - SHOW TABLES; untuk melihat tabel tabel yang ada di Hive

```
hive> SHOW TABLES;
```

```
hive> SHOW TABLES;
OK
cust
media_demo_customer
media_demo_movielog
movie
movie_rating
movie_updates
movie_view
movieapp_log_avro
movieapp_log_json
movieapp_log_month_avro
movieapp_log_month_parquet
movieapp_log_odistage
movieapp_log_stage
movielog
session_stats
user_movie
Time taken: 3.363 seconds, Fetched: 16 row(s)
```

   - Membuat tabel suhu

```
hive> create table suhu(tahun int, suhu int, kualitas int);
```

```
hive>  create table suhu(tahun int, suhu int, kualitas int);
OK
Time taken: 8.091 seconds
```

   - Load data local ‘suhutab.txt

```
hive> LOAD DATA LOCAL INPATH ‘suhutab.txt’ INTO TABLE

suhu;
```

```
hive> LOAD DATA LOCAL INPATH 'suhutab.txt' INTO TABLE suhu;
Loading data to table default.suhu
OK
Time taken: 1.347 seconds
```

   - Melihat keseluruhan tabel suhu

```
hive> select * from suhu;
```

```
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
NULL    NULL    NULL
Time taken: 0.462 seconds, Fetched: 803975 row(s)
```

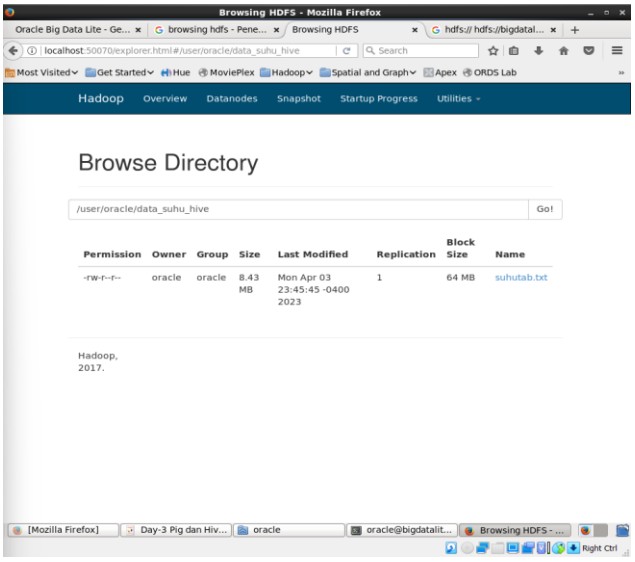
- Truncate tabel suhu

```
hive> truncate from suhu;
```

```
hive> truncate table suhu;
OK
Time taken: 0.366 seconds
```

- Membuat folder data_suhu_hive pada hadoop

```
$ fs -mkdir data_suhu_hive
```



- Copy data suhutab.txt ke hadoop

```
$ fs -copyFromLocal suhutab.txt
data_suhu_hive/suhutab.txt
```

- Mebuat external table yaitu suhutemp pada Hive

```
hive> CREATE EXTERNAL TABLE suhutemp(tahun int,suhu int,kualitas
int) ROW FORMAT DELIMITED FIELDS TERMINATEDBY '\t' STORED AS
TEXTFILE LOCATION '/user/oracle/data_suhu_hive';
```

```
hive> CREATE EXTERNAL TABLE suhutemp(tahun int,suhu int,kualitas int) ROW FORMAT
 DELIMITED
    > ;
OK
Time taken: 1.641 seconds
```

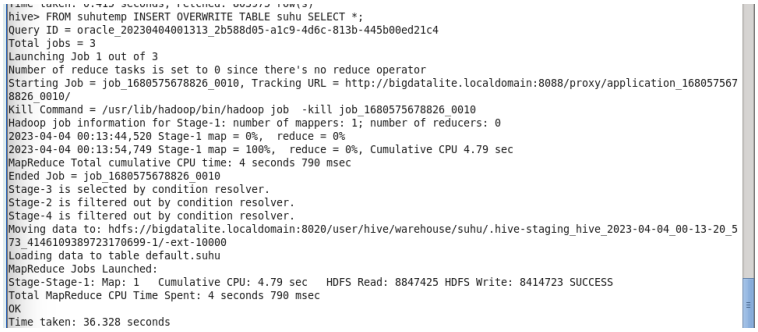
- Melihat keseluruhan tabel suhutemp

```
hive> SELECT * FROM suhutemp;
```

```
1931    96      1
1931    96      1
1931    102     1
1931    107     1
1931    119     1
1931    130     1
1931    135     1
1931    135     1
1931    141     1
1931    141     1
1931    141     1
1931    135     1
1931    124     1
1931    124     1
1931    124     1
1931    119     1
Time taken: 0.415 seconds, Fetched: 803975 row(s)
```

- Memindahkan suhutemp ke tabel utama

```
hive> FROM suhutemp INSERT OVERWRITE TABLE suhu SELECT *;
```



```
hive> FROM suhutemp INSERT OVERWRITE TABLE suhu SELECT *;
Query ID = oracle_20230404001313_2b588d05-a1c9-4d6c-813b-445b00ed21c4
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1680575678826_0010, Tracking URL = http://bigdatalite.localdomain:8088/proxy/application_168057567
8826_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1680575678826_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-04-04 00:13:44,520 Stage-1 map = 0%,  reduce = 0%
2023-04-04 00:13:54,749 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.79 sec
MapReduce Total cumulative CPU time: 4 seconds 790 msec
Ended Job = job_1680575678826_0010
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://bigdatalite.localdomain:8020/user/hive/warehouse/suhu/.hive-staging_hive_2023-04-04_00-13-20_5
73_4146109389723170699-1/-ext-10000
Loading data to table default.suhu
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.79 sec   HDFS Read: 8847425 HDFS Write: 8414723 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 790 msec
OK
Time taken: 36.328 seconds
```

- Melihat keseluruhan tabel suhu

```
hive> SELECT * FROM suhu;
```

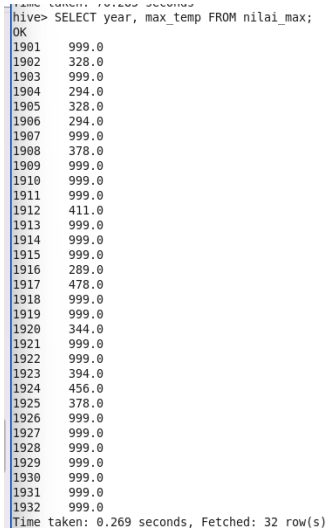
```
1931    107     1
1931    102     1
1931    102     1
1931    96      1
1931    96      1
1931    96      1
1931    96      1
1931    96      1
1931    96      1
1931    102     1
1931    107     1
1931    119     1
1931    130     1
1931    135     1
1931    135     1
1931    141     1
1931    141     1
1931    141     1
1931    135     1
1931    124     1
1931    124     1
1931    124     1
1931    119     1
Time taken: 1.11 seconds, Fetched: 803975 row(s)
```

a. Melakukan Statistika deskriptif (suhu maksimum, minimum, rata-rata, varian, deviasi standar, dan persentil) yang dikelompokkkan berdasarkan masing-masing tahun

- Mencari nilai suhu maksimum, dengan kode berikut

```
hive> CREATE TABLE nilai_max (

    > year INT,

    > max_temp FLOAT

    > );

hive> INSERT INTO nilai_max

    > SELECT tahun, MAX(suhu)

    > FROM suhu

    > GROUP BY tahun;

hive> SELECT year, max_temp FROM nilai_max;
```

Mendapatkan Output nilai maksimum:

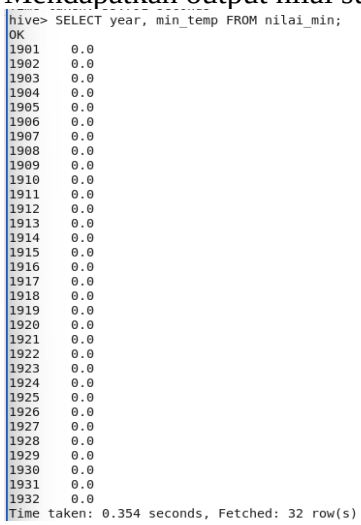


```
hive> SELECT year, max_temp FROM nilai_max;
OK
1901    999.0
1902    328.0
1903    999.0
1904    294.0
1905    328.0
1906    294.0
1907    999.0
1908    378.0
1909    999.0
1910    999.0
1911    999.0
1912    411.0
1913    999.0
1914    999.0
1915    999.0
1916    289.0
1917    478.0
1918    999.0
1919    999.0
1920    344.0
1921    999.0
1922    999.0
1923    394.0
1924    456.0
1925    378.0
1926    999.0
1927    999.0
1928    999.0
1929    999.0
1930    999.0
1931    999.0
1932    999.0
Time taken: 0.269 seconds, Fetched: 32 row(s)
```

- Mencari nilai suhu minimum, dengan kode berikut ini

```
hive> CREATE TABLE nilai_min (

    > year INT,

    > min_temp FLOAT

    > );

hive> INSERT INTO nilai_min

    > SELECT tahun, MIN(suhu)

    > FROM suhu

    > GROUP BY tahun;
```

Mendapatkan output nilai suhu minimum pertahunnya:



```
hive> SELECT year, min_temp FROM nilai_min;
OK
1901    0.0
1902    0.0
1903    0.0
1904    0.0
1905    0.0
1906    0.0
1907    0.0
1908    0.0
1909    0.0
1910    0.0
1911    0.0
1912    0.0
1913    0.0
1914    0.0
1915    0.0
1916    0.0
1917    0.0
1918    0.0
1919    0.0
1920    0.0
1921    0.0
1922    0.0
1923    0.0
1924    0.0
1925    0.0
1926    0.0
1927    0.0
1928    0.0
1929    0.0
1930    0.0
1931    0.0
1932    0.0
Time taken: 0.354 seconds, Fetched: 32 row(s)
```

- Mencari nilai rata-rata pada suhu pertahunnya, dengan kode berikut ini:

```
hive> CREATE TABLE nilai_avg (

    > year INT,

    > avg_temp FLOAT

    > );

hive> INSERT INTO nilai_avg

    > SELECT tahun, avg(suhu)

    > FROM suhu

    > GROUP BY tahun;
```

Mendapatkan output nilai rata-rata:

```
hive> SELECT year, avg_temp FROM nilai_avg;
OK
1901    93.89947
1902    73.78385
1903    78.52106
1904    74.43649
1905    75.15333
1906    83.61728
1907    89.308075
1908    84.854065
1909    91.83727
1910    78.11804
1911    86.581535
1912    92.43987
1913    85.171005
1914    80.78347
1915    91.46177
1916    62.619003
1917    94.20698
1918    86.09713
1919    87.77162
1920    78.20605
1921    84.81995
1922    89.4048
1923    80.92046
1924    89.57062
1925    87.428215
1926    100.6595
1927    111.789345
1928    117.860214
1929    131.92644
1930    147.69661
1931    160.65097
1932    167.94986
Time taken: 0.557 seconds, Fetched: 32 row(s)
```

- Mencari nilai Varian pada suhu, dengan kode berikut:

```
hive> CREATE TABLE nilai_var (

    > year INT,

    > var_temp FLOAT

    > );

hive> INSERT INTO nilai_var

    > SELECT tahun, VAR_POP(suhu)

    > FROM suhu

    > GROUP BY tahun;
```

Mendapatkan output nilai varian pada suhu:

```
hive> SELECT year, var_temp FROM nilai_var;
OK
1901    4763.649
1902    2952.782
1903    9037.803
1904    2813.312
1905    3649.5332
1906    4165.5835
1907    4402.5264
1908    4209.395
1909    7828.166
1910    4405.391
1911    6511.908
1912    5485.169
1913    5929.429
1914    4895.3755
1915    4994.9307
1916    2821.9216
1917    5113.103
1918    6278.7026
1919    4649.8276
1920    3998.4194
1921    4025.3896
1922    9609.3955
1923    3350.7422
1924    4091.52
1925    4591.0464
1926    13678.43
1927    24760.959
1928    32968.094
1929    28532.438
1930    35426.332
1931    44478.223
1932    59660.992
Time taken: 0.262 seconds, Fetched: 32 row(s)
```

- Mencari nilai standar deviasi pada suhu, dengan kode berikut:

```
hive> CREATE TABLE nilai_std (
    > year INT,
    > std_temp FLOAT
    > );
hive> INSERT INTO nilai_std
    > SELECT tahun, STDDEV_POP(suhu)
    > FROM suhu
    > GROUP BY tahun;
hive> SELECT year, std_temp FROM nilai_std;
```

Mendapatkan output:

```
hive> SELECT year, std_temp FROM nilai_std;
OK
1901    69.019196
1902    54.339508
1903    95.06736
1904    53.04066
1905    60.411366
1906    64.541336
1907    66.35154
1908    64.87985
1909    88.47692
1910    66.373116
1911    80.696396
1912    74.06193
1913    77.002785
1914    69.96696
1915    70.67482
1916    53.12176
1917    71.50597
1918    79.238266
1919    68.189644
1920    63.233055
1921    63.445957
1922    98.02753
1923    57.885593
1924    63.964993
1925    67.75726
1926    116.95482
1927    157.35616
1928    181.57118
1929    168.91547
1930    188.21884
1931    210.8986
1932    244.256
Time taken: 0.177 seconds, Fetched: 32 row(s)
```

❋ Mencari nilai presentil suhu

```
hive> SELECT tahun,

    > PERCENTILE(suhu, 0.25) AS suhu_p25,

    > PERCENTILE(suhu, 0.50) AS suhu_p50,

    > PERCENTILE(suhu, 0.75) AS suhu_p75

    > FROM suhutemp

    > GROUP BY tahun;
```

Mendapatkan nilai presentil suhu:

```
Total MapReduce CPU Time Spent: 6 seconds 50 msec
OK
1901    33.0    89.0    144.0
1902    28.0    67.0    111.0
1903    22.0    56.0    122.0
1904    28.0    67.0    117.0
1905    22.0    61.0    122.0
1906    28.0    72.0    128.0
1907    33.0    83.0    133.0
1908    28.0    72.0    128.0
1909    33.0    83.0    133.0
1910    22.0    67.0    122.0
1911    28.0    72.0    128.0
1912    28.0    78.0    144.0
1913    28.0    72.0    128.0
1914    28.0    61.0    117.0
1915    33.0    78.0    133.0
1916    22.0    50.0    89.0
1917    33.0    78.0    144.0
1918    33.0    72.0    122.0
1919    28.0    78.0    128.0
1920    22.0    61.0    122.0
1921    28.0    78.0    128.0
1922    28.0    72.0    128.0
1923    33.0    72.0    122.0
1924    33.0    83.0    133.0
1925    28.0    78.0    133.0
1926    39.0    78.0    139.0
1927    39.0    78.0    128.0
1928    39.0    78.0    128.0
```

b. Perubahan rata – rata suhu diantara 2 tahun, misalnya tahun 1912-1913

❋ Melakukan koding seperti dibawah ini:

```
hive> SELECT ((avg_suhu_tahun_2 - avg_suhu_tahun_1) /
avg_suhu_tahun_1) * 100 AS presentase_perubahan_suhu

    > FROM

    >   (SELECT AVG(suhu) AS avg_suhu_tahun_1

    >   FROM suhu

    >   WHERE tahun = 1912) t1,

    >   (SELECT AVG(suhu) AS avg_suhu_tahun_2

    >   FROM suhu

    >   WHERE tahun = 1913) t2;
```

Mendapatkan output perubahan rata-rata suhu diantara 2 tahun pada tahun 1912-1913:

```
OK
-7.863346332459913
Time taken: 109.4 seconds, Fetched: 1 row(s)
```

c. Membuat 3 pertanyaan analisi berdasarkan dataset NOAA

1. Berapa selisih suhu maksimum dan minimum pada pertahunnya?
   Jawab: Selisih suhu maksimum dan minimum pada pertahunnya sangatlah bervariasi jika ingin melihat, bisa dilihat pada gambar dibawah ini:

```
1901    999
1902    328
1903    999
1904    294
1905    328
1906    294
1907    999
1908    378
1909    999
1910    999
1911    999
1912    411
1913    999
1914    999
1915    999
1916    289
1917    478
1918    999
1919    999
1920    344
1921    999
1922    999
1923    394
1924    456
1925    378
1926    999
1927    999
1928    999
1929    999
1930    999
1931    999
1932    999
Time taken: 41.302 seconds, Fetched: 32 row(s)
```

2. Sebutkan 5 saja tahun yang memiliki rata-rata suhu tertinggi?
   Jawab: Tahun-tahun yang memiliki rata-rata suhu tertinggi ialah pada tahun 1932 yang memiliki rata-rata suhu 167.949, kemudian rata-rata suhu tertinggi kedua yakni pada tahun 1931 dengan rata-rata 160.650, selanjutanya rata-rata suhu tertinggi ketiga takni pada tahun 1930 dengan rata-rata 147.696, setelah itu disusul oleh tahun 1929 dengan rata-rata suhu 131.926 selanjutnya yang rata-rata suhu yang kelima terjadi pada tahun 1928 dengan rata-rata 117.860, bisa dilihat pada gambar dibawah ini:

```
1932    167.94986493146183
1931    160.65097112536066
1930    147.69661222020568
1929    131.9264331407987
1928    117.86021212945336
Time taken: 66.896 seconds, Fetched: 5 row(s)
```

3. Bagaiman jika selisih suhu maksimum dan minimum terbesar ditampilkan dalam kurun waktu 10 tahun?
   Jawab: Selisih suhu maksimum dan minimum terbesar dalam kurun waktu 10 tahun terjadi pada tahun 1901; 1931; 1930; 1929; 1928; 1927; 1926; 1922; 1921; 1919 bisa dilihat pada gambar berikut ini:

```
OK
1901    999
1931    999
1930    999
1929    999
1928    999
1927    999
1926    999
1922    999
1921    999
1919    999
Time taken: 82.713 seconds, Fetched: 10 row(s)
```

4. Berapa nilai median pada suhu dalam pertahunnya?
   Jawab: berikut gambar dibawah ini merupakan nilai-nilai median pada suhu dalam pertahunnya.

```
1901    84.33561643835617
1902    63.81404958677686
1903    54.604651162790695
1904    64.67289719626169
1905    57.109022556390975
1906    69.87564766839378
1907    79.33720930232558
1908    70.8510101010101
1909    78.14473684210526
1910    61.47430830039526
1911    69.63944223107569
1912    72.3013698630137
1913    67.66101694915254
1914    59.918685121107266
1915    76.1438127090301
1916    46.758389261744966
1917    75.25
1918    71.31528662420382
1919    74.03305785123968
1920    60.474137931034484
1921    73.36981132075472
1922    69.78846153846153
1923    71.54255319148936
1924    80.24890829694323
1925    74.0253807106599
1926    77.91803278688525
1927    77.3265306122449
1928    73.11961722488039
1929    96.3230283911672
1930    107.62232076866223
1931    110.53940249353093
1932    96.97061873487729
Time taken: 27.538 seconds, Fetched: 32 row(s)
```