**LAPORAN**

**RENCANA TUGAS MANDIRI (RTM) Ke-4**

**MATA KULIAH BIG DATA**

**“*Query* statistik deskriptif menggunakan Hive dan XQuery”**



**DISUSUN OLEH:**

Chelsea Ayu Adhigiadany ( 21083010028 )

**DOSEN PENGAMPU:**

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

**PROGRAM STUDI SAINS DATA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR**

**2023**

**Daftar Isi**

[**Dataset NOAA** 3](#_Toc131193452)

[**Analisis dasar Dataset NOAA menggunakan Hive** 3](#_Toc131193453)

[**Analisis Dataset NOAA dari 5 pertanyaan menggunakan Hive** 5](#_Toc131193454)

[**Analisis Dataset saham menggunakan XQuery** 8](#_Toc131193455)

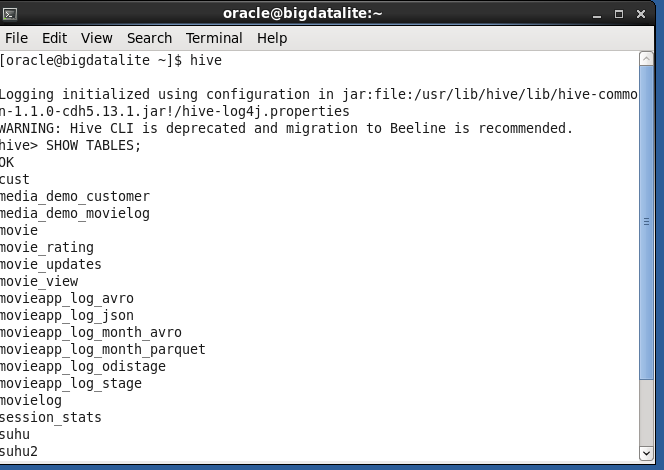
[**Analisis Dataset harga sembako menggunakan XQuery** 12](#_Toc131193456)

# **Dataset NOAA**

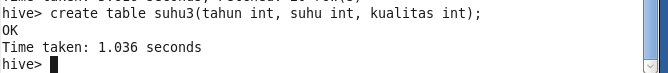
Dataset yang digunakan merupakan dataset NOAA (*National Oceanic and Atmosphere Administration*) dengan satu file zip yang berisi kumpulan dataset dari tahun 1901-1932. Dataset ini terdiri dari 3 kolom yaitu tahun, suhu (dalam Fahrenheit), dan juga kualitas.

# **Analisis dasar Dataset NOAA menggunakan Hive**

1. Langkah pertama yaitu masuk ke hive mengggunakan syntax hive. Untuk melihat tabel apa saja yang ada pada hive, menggunakan syntax SHOW TABLES;

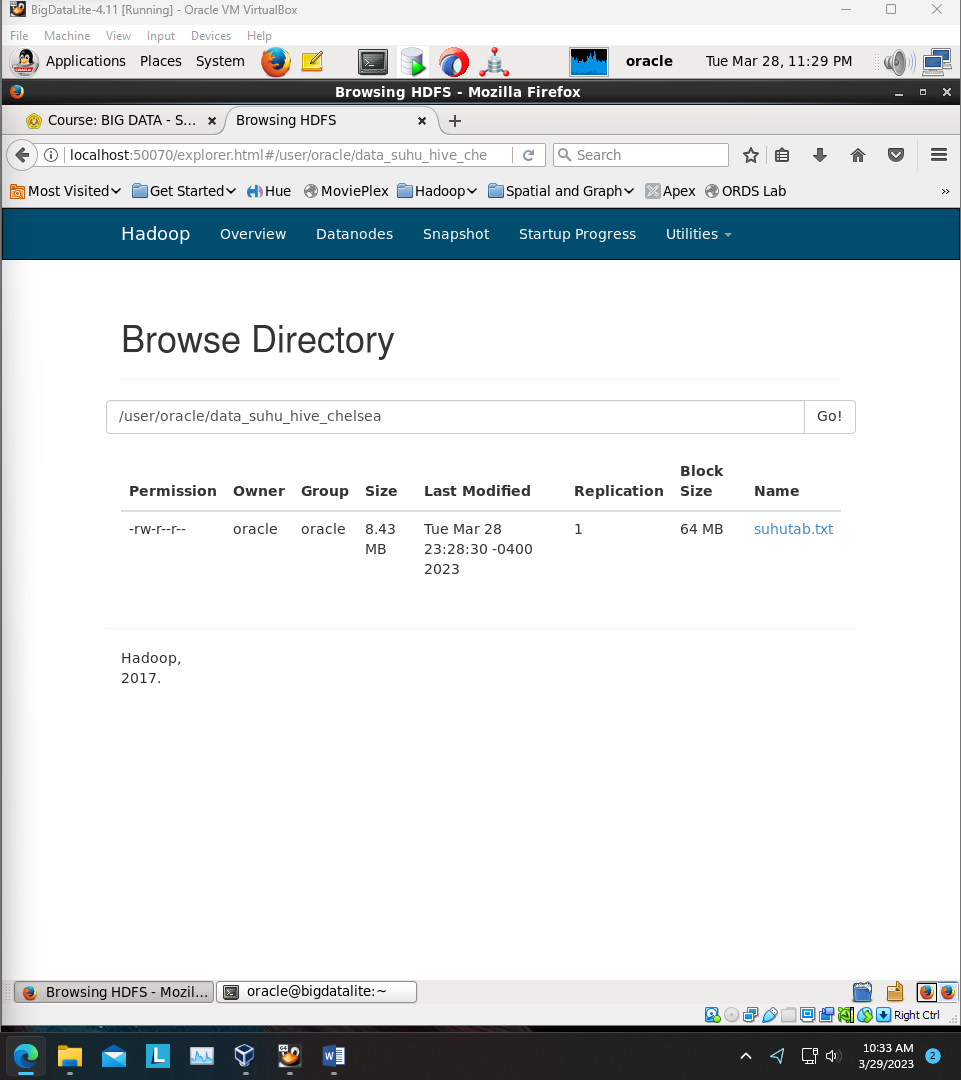


1. Buat tabel suhu yang terdiri dari 3 kolom dengan ketentuan :
2. Kolom 1 diberi nama tahun dengan tipe data integer
3. Kolom 2 diberi nama suhu dengan tipe data integer
4. Kolom 3 diberi nama kualitas dengan tipe data integer

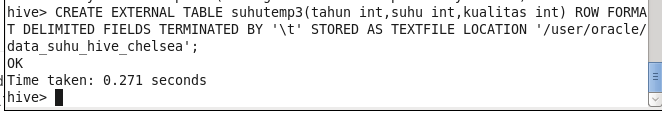


1. Membuat direktori pada Hadoop menggunakan syntax hadoop fs -mkdir data\_suhu\_hive\_chelsea. Copy data suhu kedalam direktori yang telah dibuat tadi menggunakan syntax hadoop fs -copyFromLocal suhutab.txt data\_suhu\_hive/suhutab.txt. Cek pada browser HDFS, pastikan direktori dan juga data suhu telah masuk ke dalam direktori.



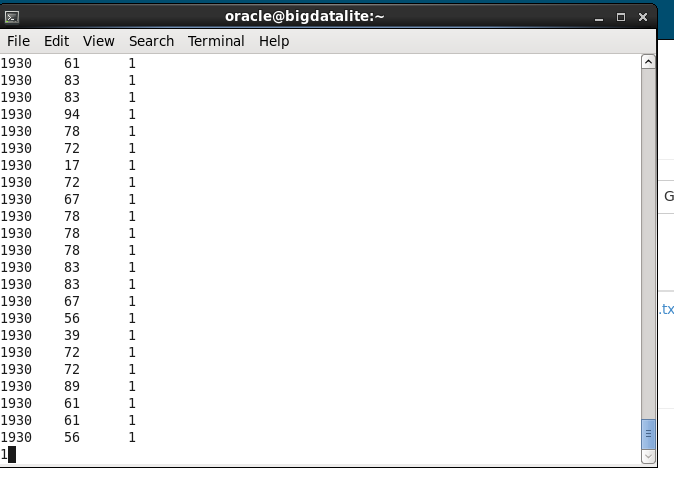


1. Karena suhutab.txt noaa dipisahkan dengan <tab>, maka dibuat external table (tabel hive dengan data dari file)

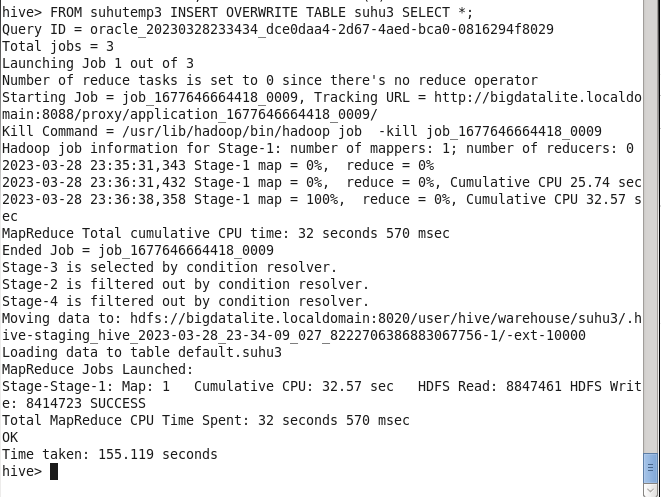


Untuk membaca data yang tersimpan pada tabel eksternal tadi, menggunakan syntax SELECT \* FROM suhutemp3. SELECT adalah perintah untuk mengambil data dari tabel. Tanda \* berarti memanggil semua data yang ada dari tabel suhutemp3. FROM menunjukkan tabel manakah yang akan diambil datanya.

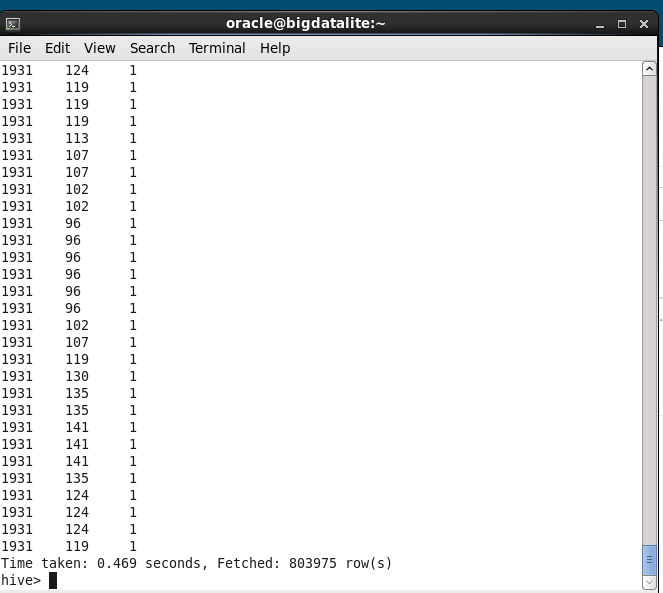




1. Memindahkan suhutemp3 ke tabel utama. Syntax yang digunakan yaitu FROM suhutemp INSERT OVERWRITE TABLE suhu3 SELECT \*;. Seperti penjelasan sebelumnya, FROM menunjukkan tabel manakah yang datanya akan dimabil, INSERT OVERWRITE TABLE berarti menandakan tabel mana yang nantinya dijadikan sebagai tempat dari data baru dari suhutemp3. Penjelasan terkait SELECT masih sama dengan penjelasan sebelumnya.

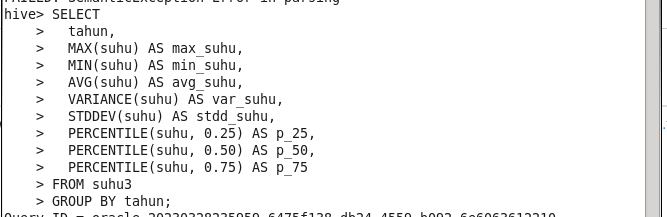




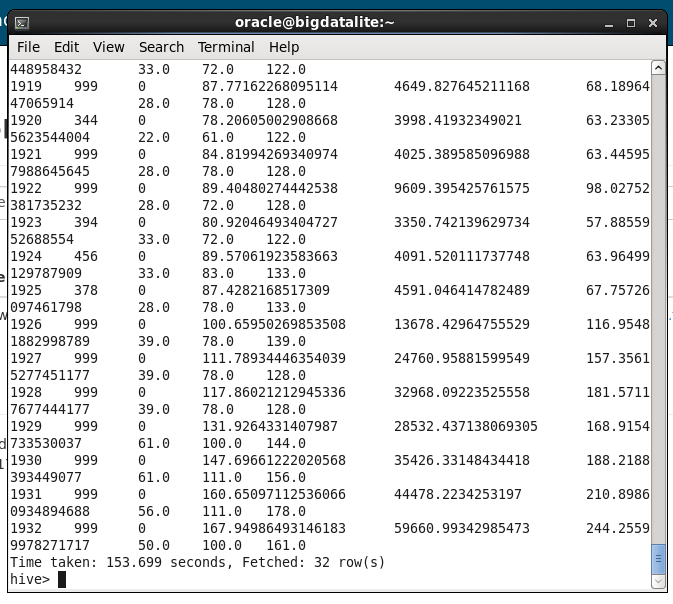


# **Analisis Dataset NOAA dari 5 pertanyaan menggunakan Hive**

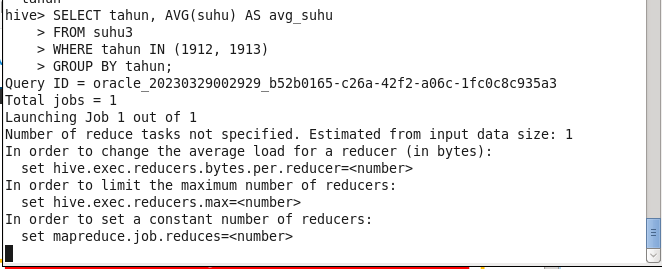
1. Tentukan statistika deskriptif dari dataset NOAA. Statistik deskriptif meliputi suhu minimum, suhu maksimum, rata-rata, varian, standart deviasi, dan juga persentil!



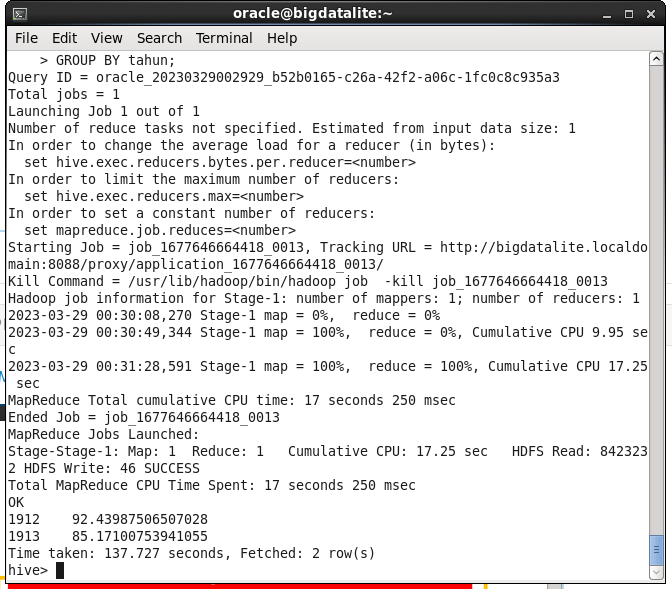
Script tersebut berisi tentang perhitungan statistika deskriptif dari kolom tahun yang diambil dari data suhu3. Hasil dari perhitungan statistika deskriptif-nya dikelompokkan berdaarkan tahun, maka dari itu digunakanlah syntax GROUP BY tahun; untuk menutup perintah perhitungan. Syntax untuk menghitung statistika deskriptifnya sudah sangat jelas penggunaanya, MAX untuk menghitung nilai maksimum yang di aliaskan dengan max\_suhu menggunakan syntax AS. Penjelasan tersebut juga berlaku untuk MIN (nilai minimum), AVG (rata-rata), VARIANCE (varian), STDDEV (standart deviasi), dan PERCENTILE (persentil).



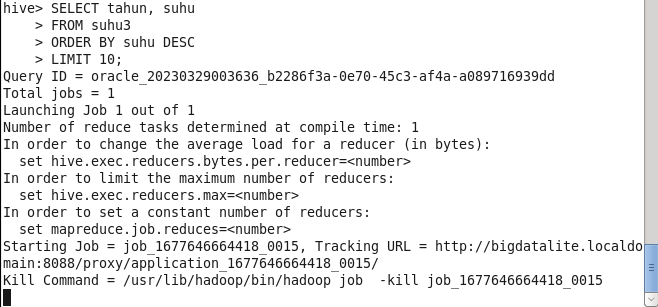
1. Hitung persentase perubahan rata-rata suhu antara tahun 1912 dengan 1913!



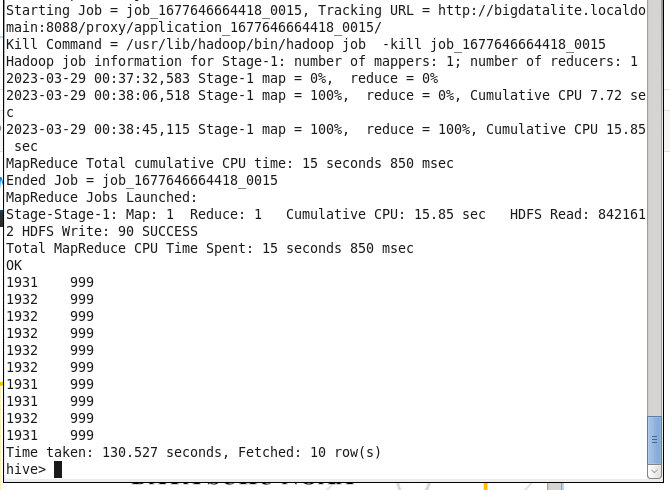
Script tersebut diawali dengan mengambil rata-rata suhu dari kolom suhu yang berasal dari tabel suhu3. WHERE digunakan untuk memfilter data dari kolom tahun yang didefinisikan setelah syntax IN.



1. Tentukan 10 suhu tertinggi disertai keterangan tahunnya!



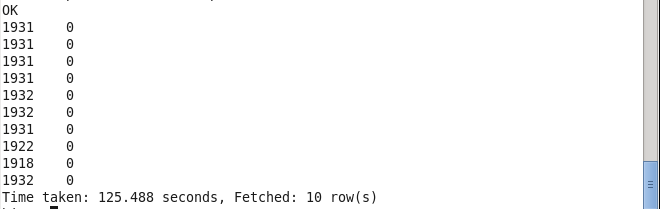
Script tersebut diawali dengan mengambil kolom tahun dan suhu dari tabel suhu3. ORDER BY digunakan untuk mengurutkan data dari tabel suhu, dengan DESC adalah jenis pengurutannya descending (dari tinggi ke rendah). LIMIT menunjukkan batasan data yang dimunculkan.



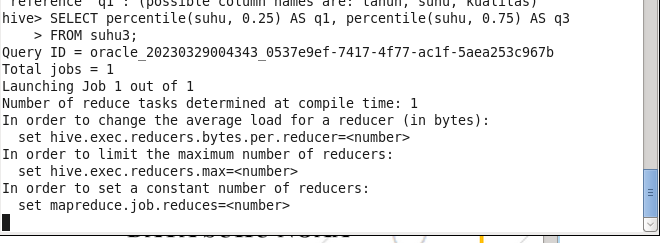
1. Tentukan 10 kualitas terendah disertai keterangan tahunnya!



Sesuai dengan penjelasan dari soal nomor 3, hanya saja perbedaannya terdapat pada jenis pengurutannya. ASC berarti ascending (dari rendah ke tinggi).

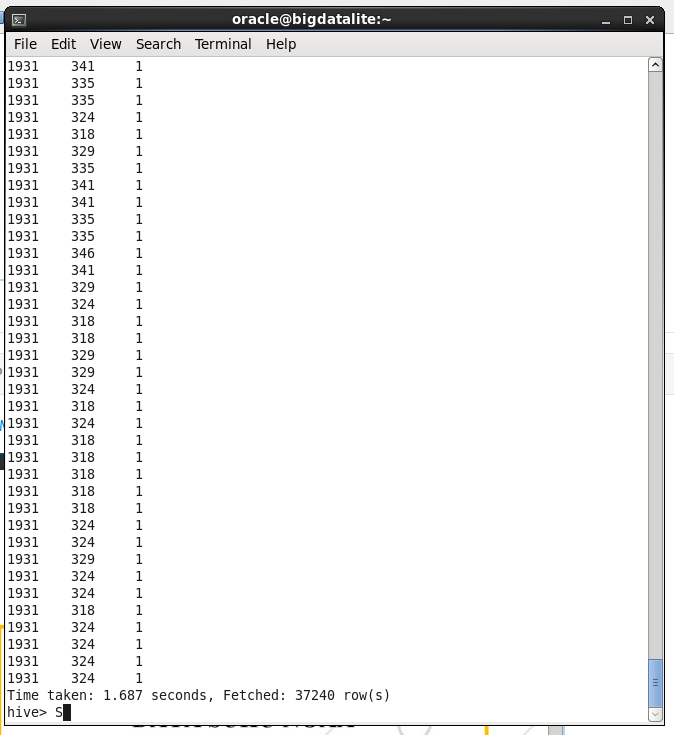


1. Tentukan outlier dari suhu!



Salah satu metode untuk mencari outlier adalah mengggunakan IQR, langkah awal tentu saja mencari Q1 dan Q3 dari suhu, didapatkan hasil Q1=50 dan Q3 = 156, dengan begitu IQR nya adalah 156-50 = 106. Outlier adalah suhu yang lebih rendah dari batas bawah dan juga batas atas. Adapun rumus batas bawah adalah Q1 – 1,5 x IQR dan batas atas Q3 + 1,5 x IQR. Rumus pencarian outlier dituliskan pada WHERE, OR digunakan untuk menggabungkan 2 kondisi.

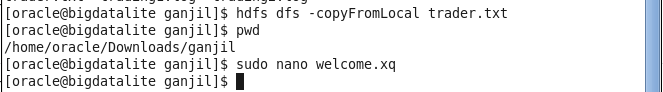




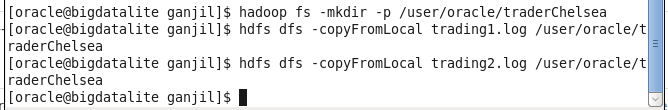
# **Analisis Dataset saham menggunakan XQuery**

Dataset saham yang digunakan terdiri dari 3 file dengan 2 file berformat log dan satu file berformat txt. Pada file txt berisi informasi tambahan tentang user yang ada pada file log.

1. Copy file trader.txt ke dalam hdfs

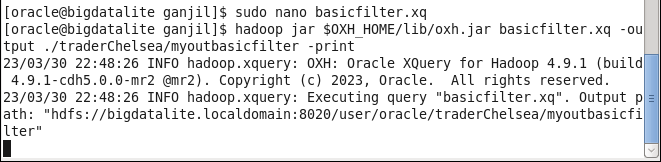


1. Buat direktori baru pada HDFS yang terletak di user/oracle/ dengan nama traderChelsea. Copy juga file trading1.log dan juga trading2.log



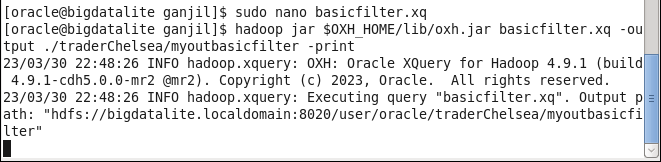
1. Basic filtering

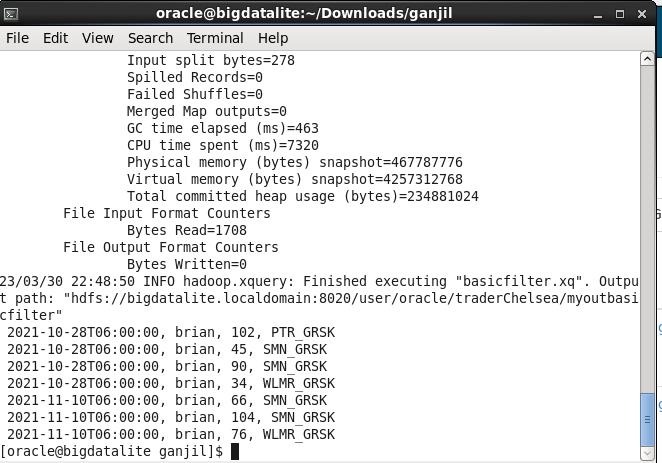
Filtering dasar menggunakan XQuery, pada analisis ini dilakukan filtering untuk semua yang ada dari trading1.log dan trading2.log dengan user brian. Output akan disimpan dalam file myoutbasicfilter yang berada pada direktori traderChelsea.





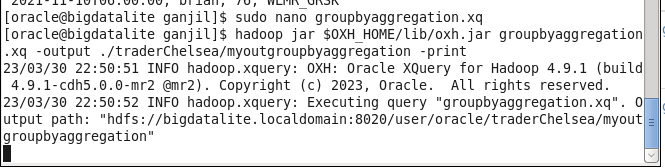
Maksud dari script ini adalah dari masing-masing line pada file trading1&2.log di pisahkan. Dari kedua file itu, pada hasil split index kedua dengan isi “brian” akan dikeluarkan.

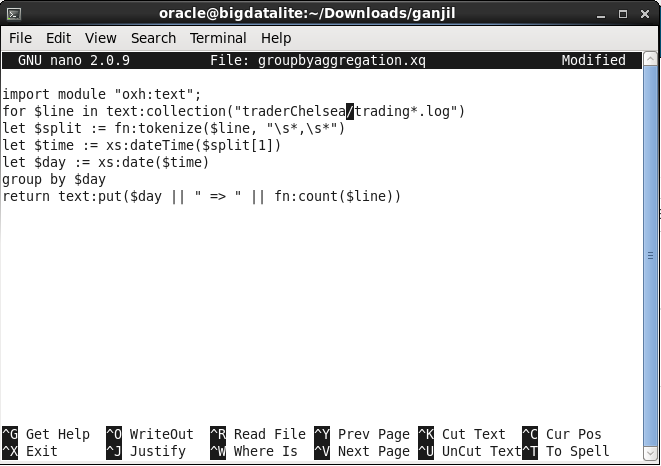




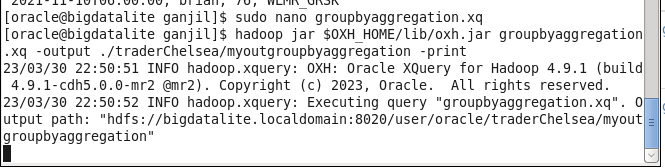
1. Group by and aggregation

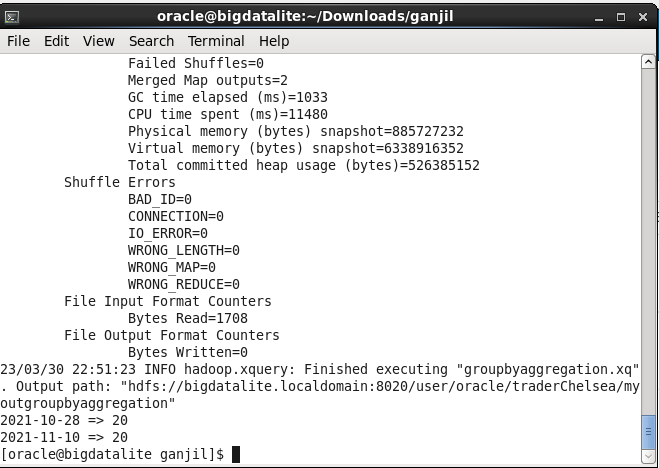
Group by and aggregation pada analisis ini yaitu untuk mengetahui berapa banyak kunjungan yang dilakukan oleh user berdasarkan tanggal yang ada pada indeks 1 dari trading1.log dan trading2.log. Output dari analisis ini akan disimpan dalam myoutgroupbyaggregation.





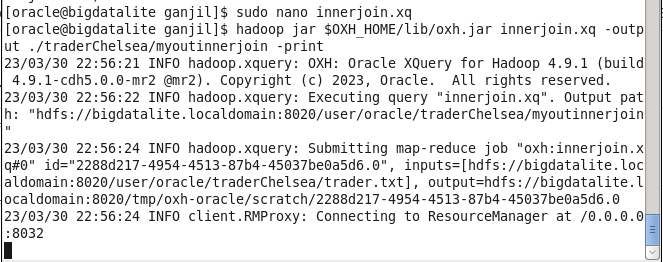
Maksud dari script ini adalah dari masing-masing line pada file trading1&2.log di pisahkan. $time adalah hasil split index pertama. Dengan $day adalah tanggal yang diperolah dari $time. Dilakukan pengelompokan berdasarkan $day.

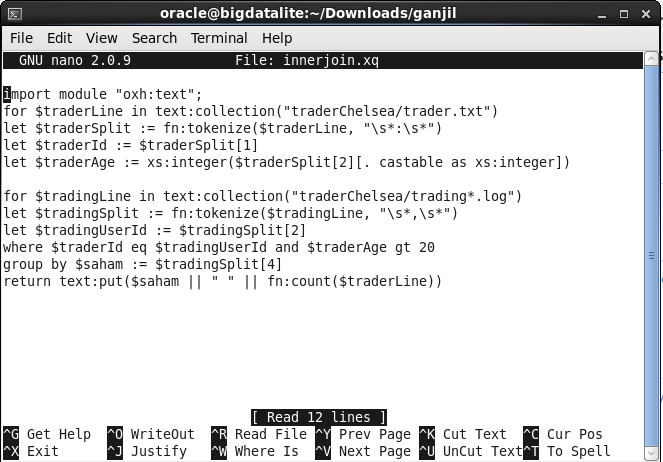




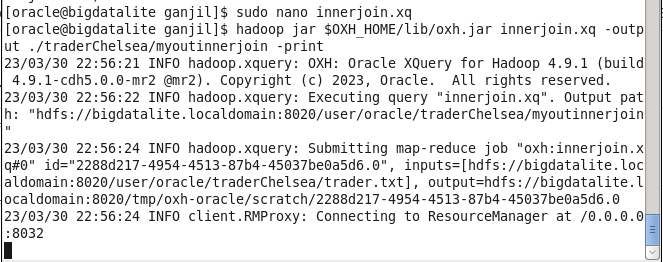
1. Inner join

Inner join merupakan operasi penggabungan tabel berdasarkan kolom yang sama dan hanya mengembalikan baris yang memenuhi kondisi penggabungan, baris yang tidak cocok dari kedua tabel akan dihilangkan dari hasil. Pada analisis ini yang digabungkan adalah file trader.txt dengan trading\*.log. Output dari analisis ini akan tersimpan pada file myoutinnerjoin pada direktori traderChelsea.





Maksud dari script ini adalah dari file trader.txt dilakukan split, id dari file ini berada pada index 1, sedangkan umur pada file ini terletak pada index 2. Pada file trading\*.log juga dilakukan split dengan user id-nya ada pada index 2.

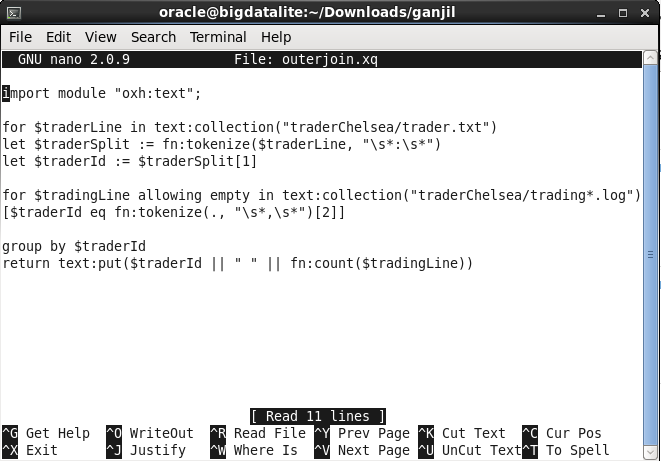




1. Left outer join

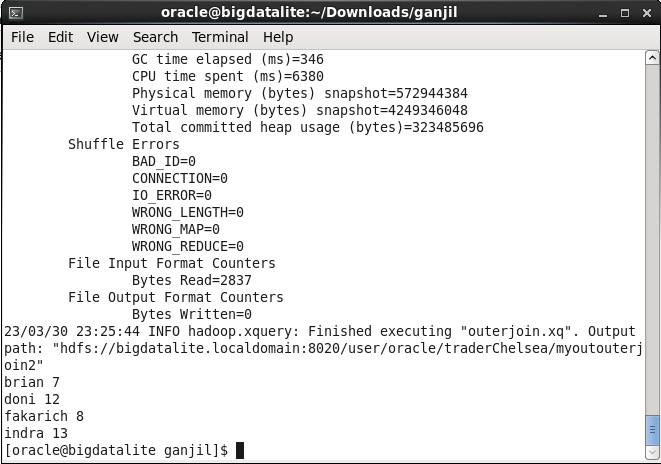
Left outer join merupakan operasi penggabungan tabel berdasarkan kolom yang sama dan hanya mengembalikan baris yang memenuhi kondisi penggabungan serta baris yang tidak cocok dari salah satu tabel, mengembalikan semua batis dari tabel kiri dan baris yang cocok dari tabel kanan. Pada analisis ini yang digabungkan adalah file trader.txt dengan trading\*.log. Output dari analisis ini akan tersimpan pada file myououterjoin2 pada direktori traderChelsea.



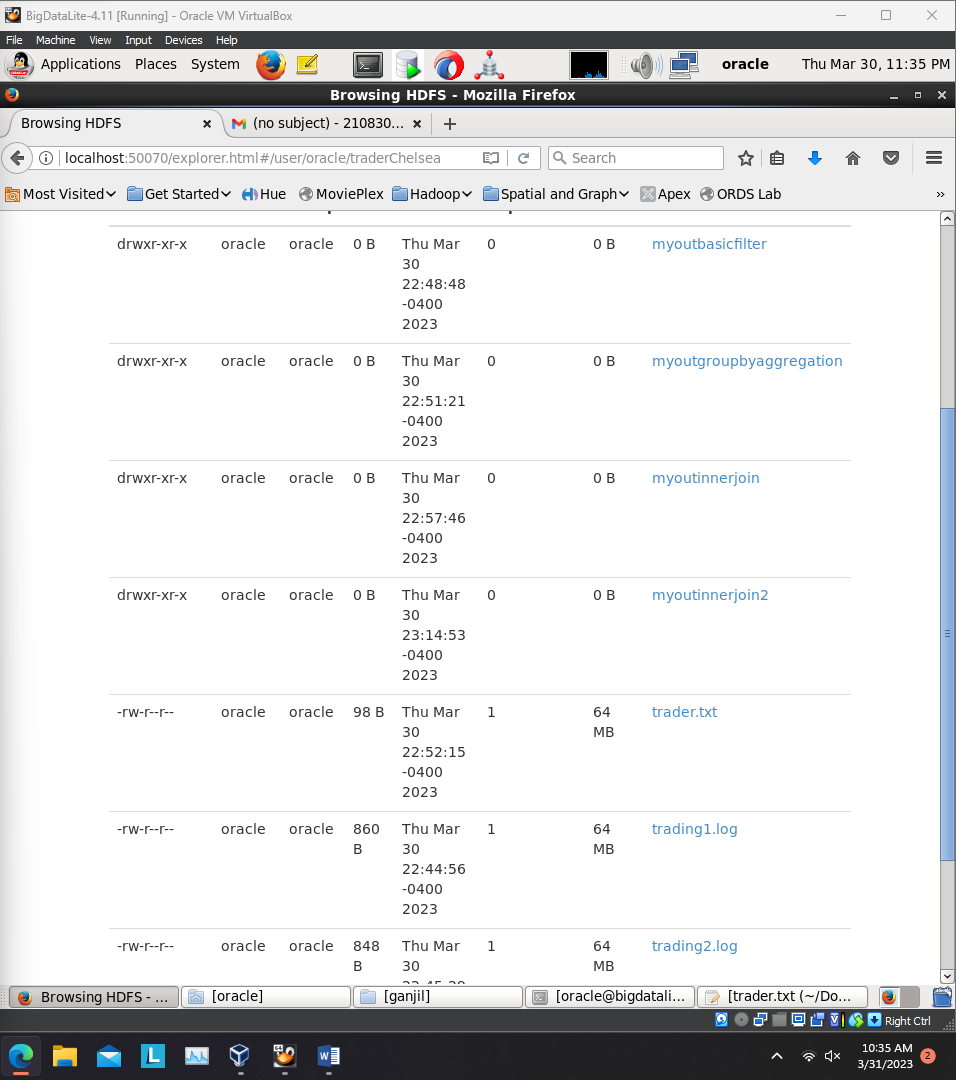


Maksud dari script ini adalah dilakukan penggabungan nilai dari 2 variabel yaitu $traderId dan fn:count($tradingLine). Fungsi text:put () digunakan untuk mengeluarkan output dari operasi yang diinginkan. Fungsi fn:count($tradingLine) digunakan untuk menghitung jumlah baris dari variable $tradingLine. Outputnya akan berupa id dan juga banyaknya kejadian yang melibatkan id tersebut.





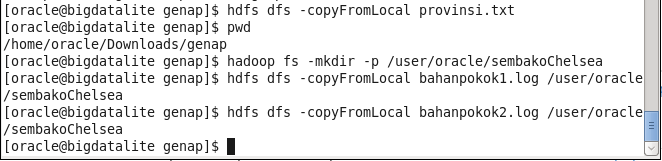
Dokumentasi hasil output pada browser HDFS



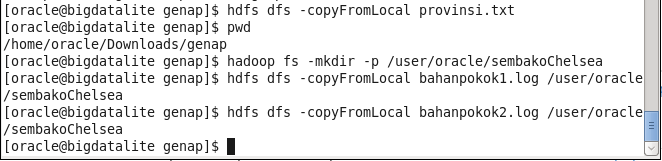
# **Analisis Dataset harga sembako menggunakan XQuery**

Dataset harga sembako yang digunakan terdiri dari 3 file dengan 2 file berformat log dan satu file berformat txt. Pada file txt berisi informasi tambahan tentang user yang ada pada file log.

1. Copy file provinsi.txt ke dalam hdfs

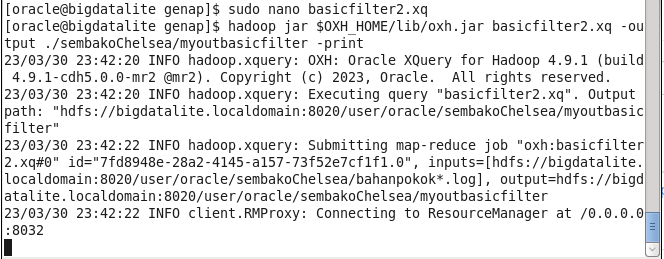


1. Buat direktori baru pada HDFS yang terletak di user/oracle/ dengan nama sembakoChelsea. Copy juga file bahanpokok1.log dan juga bahanpokok2.log



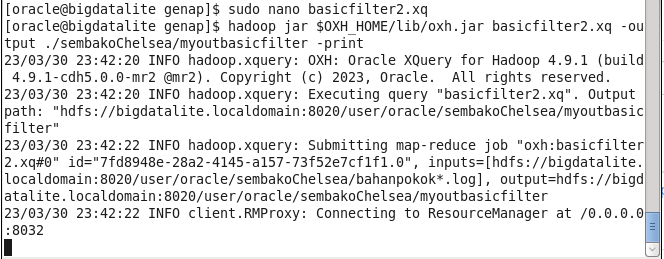
1. Basic filtering

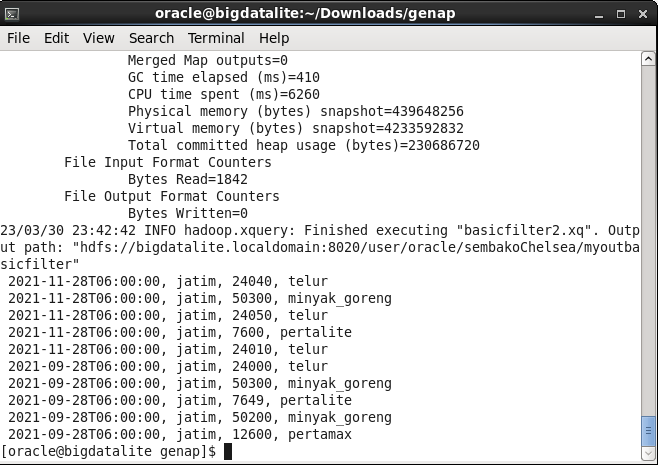
Filtering dasar menggunakan XQuery, pada analisis ini dilakukan filtering untuk semua yang ada dari bahanpokok1.log dan bahanpokok2.log dengan user jatim. Output akan disimpan dalam file myoutbasicfilter yang berada pada direktori sembakoChelsea.





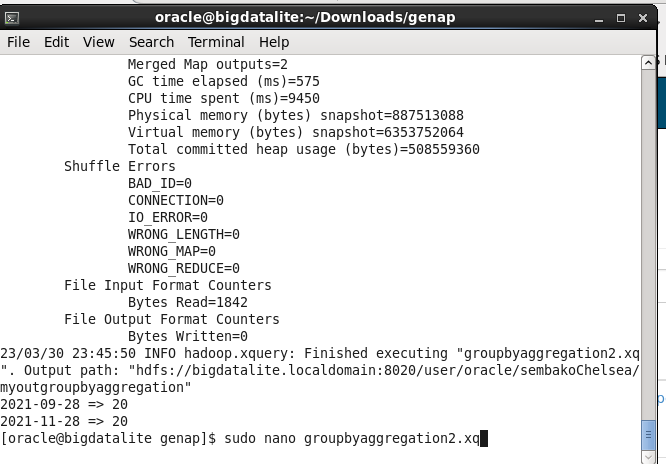
Maksud dari script ini adalah dari masing-masing line pada file bahanpokok1&2.log di pisahkan. Dari kedua file itu, pada hasil split index kedua dengan isi “jatim” akan ditampilkan.

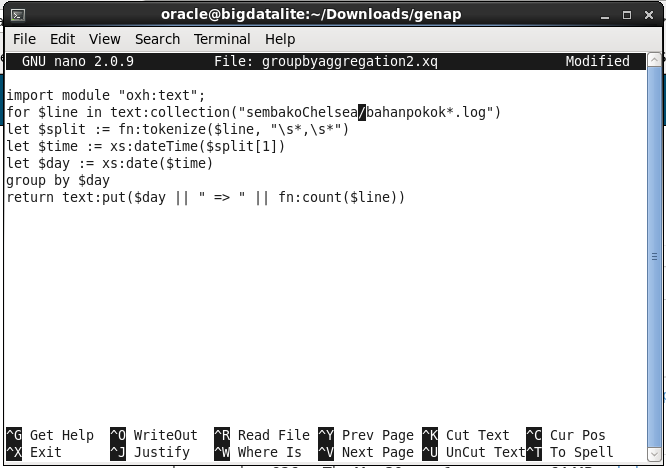




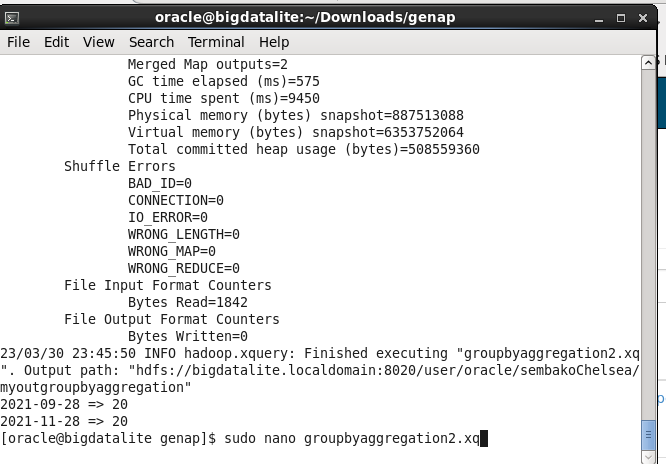
1. Group by and aggregation

Group by and aggregation pada analisis ini yaitu untuk mengetahui berapa banyak kunjungan yang dilakukan oleh user berdasarkan tanggal yang ada pada indeks 1 dari bahanpokok1.log dan bahanpokok2.log. Output dari analisis ini akan disimpan dalam myoutgroupbyaggregation.



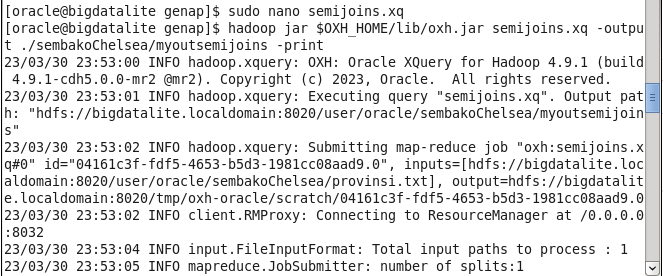


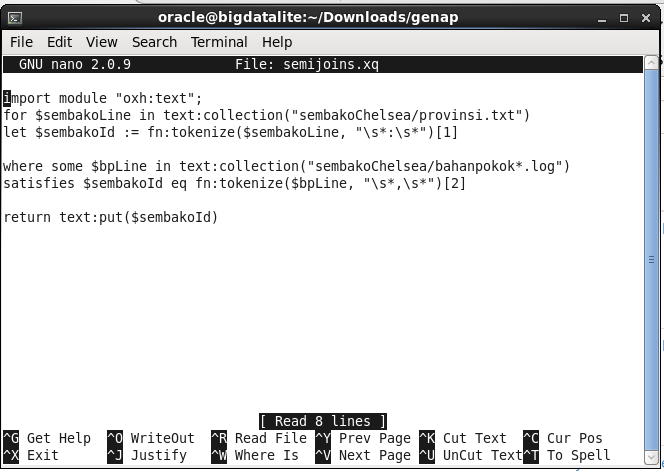
Maksud dari script ini adalah dari masing-masing line pada file bahanpokok1&2.log di pisahkan. $time adalah hasil split index pertama. Dengan $day adalah tanggal yang diperolah dari $time. Dilakukan pengelompokan berdasarkan $day.



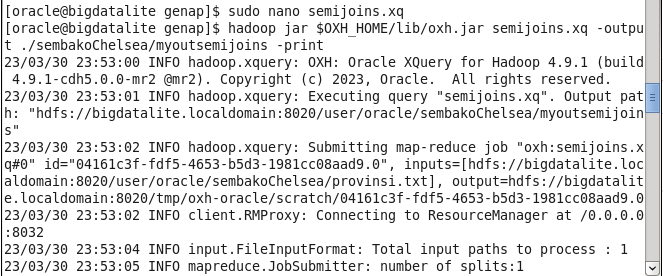
1. Semi join

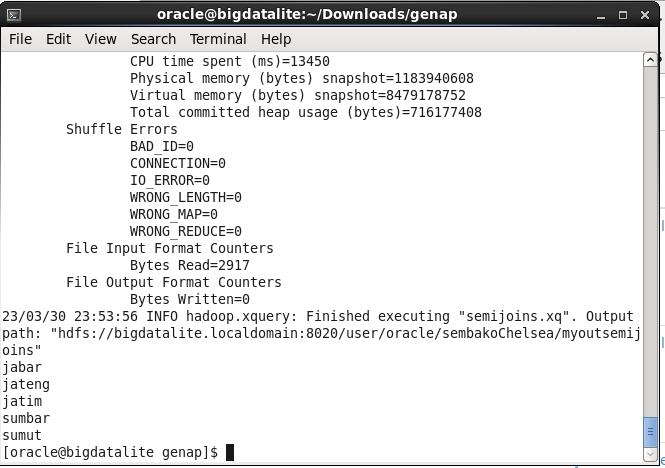
Inner join merupakan operasi penggabungan tabel berdasarkan kolom yang sama, tetapi hanya mengembalikan baris dari tabel pertama yang memiliki kecocokan di tabel kedua. Baris yang tidak cocok di tabel kedua tidak akan diikutsertakan dalam hasil. Pada analisis ini yang digabungkan adalah file provinsi.txt dengan bahanpokok\*.log. Output dari analisis ini akan tersimpan pada file myoutinnerjoin pada direktori sembakoChelsea.





Maksud dari script ini adalah mencari nilai $sembakoId pada file bahanpokok\*.log, untuk mengambil semua file digunakan fungsi text:collection(). Dilakukan pemeriksaan mengunakan where some untuk memeriksa apakah ada nilai $sembakoId yang sama, jika ada maka akan dikeluarkan dalam bentuk $sembakoId, index 1 pada provinsi.txt





Dokumentasi hasil putput pada browser HDFS

