Nama : Muhammad Ravi Shulthan Habibi
NPM : 1906351000
Kelas : Pengelolaan Data Besar
Dosen : Ir. Wahyu Catur Wibowo, M.Sc., Ph.D.

Tugas 2 - Pengelolaan Data Besar

Langkah yang Dilakukan

1. Melakukan instalasi Java dengan versi yang sesuai, disini saya gunakan Java versi 8, dapat di-*install* disini.
2. Melakukan instalasi Hadoop, dapat di-*install* disini.
3. Melakukan pengaturan pada *environment* PATH.
4. Memodifikasi file .xml dan .cmd pada folder Hadoop, yang saya modifikasi ada file: core-site.xml, mapred-site.xml, yarn-site.xml, hdfs-site.xml, dan hadoop-env.cmd.
5. File-file di atas saya modifikasi sesuai dengan kebutuhan saya, seperti: letak sistem file, *framework* mapreduce yang digunakan, *service* nodemanager, banyak replikasi, letak direktori namenode dan datanode.
6. Menjalankan hdfs namenode –format, di awal penggunaan Hadoop pertama kali.
7. Menjalankan apache dengan cara: arahkan ke direktori sbin -> lalu ketik start-all.cmd
8. Membuat direktori baru untuk menyimpan klaster, dengan perintah mkdir.
9. Mengunggah file kredit.csv ke dalam direktori tersebut, dengan perintah put.
10. Melakukan perhitungan jumlah data menurut KETR (LUNAS, TARIKAN) dengan Hadoop mapReduce (Java) dengan *script* yang akan disertakan di bawah.
11. Melakukan perhitungan jumlah data menurut KETR (LUNAS, TARIKAN) dengan Pyspark mapReduce dengan *script* yang akan disertakan di bawah.
12. Melakukan perhitungan rata-rata SALARY berkelompok menurut KETR dengan Hadoop mapReduce (Java) dengan *script* yang akan disertakan di bawah.

13. Melakukan perhitungan rata-rata SALARY berkelompok menurut KETR dengan Pyspark mapReduce dengan *script* yang akan disertakan di bawah.
14. Kemudian, melakukan klasifikasi dengan menggunakan pyspark.ml dengan model Naive Bayes dengan *script* yang akan disertakan di bawah.
15. Terakhir, melakukan klasifikasi dengan menggunakan pyspark.ml dengan model *Linear Support Vector Machine* dengan *script* yang akan disertakan di bawah.

Script yang Digunakan

Berikut *script* yang digunakan untuk mengerjakan tugas 2 ini:

- *Script task* Hadoop MapReduce - jumlah data STATUS: [LINK]
- *Script task* Hadoop MapReduce - rata-rata SALARY: [LINK]
- *Script task* PySpark MapReduce - jumlah data STATUS: [LINK]
- *Script task* PySpark MapReduce - rata-rata SALARY: [LINK]
- *Script task* klasifikasi - model Naive Bayes: [LINK]
- *Script task* klasifikasi - model *Linear Support Vector Machine*: [LINK]

Hasil yang Diperoleh

Task MapReduce

Berdasarkan hasil MapReduce untuk *task* jumlah data dan rata-rata, didapatkan hasil sebagai berikut:

- Jumlah data:
  - Yang berstatus LUNAS sebanyak 1876 data.
  - Yang berstatus TARIKAN sebanyak 113 data.
  - Total data ada sebanyak 1989 data.
- Rata-rata SALARY sebesar: 1735009.88

Task Klasifikasi

Berdasarkan prediksi dari model klasifikasi Naive Bayes dan model klasifikasi *Linear Support Vector Machine* yang telah dibuat, kedua model tersebut memberikan hasil yang sama persis, berdasarkan *classification report*, berikut ringkasan hasil evaluasinya:

- *Accuracy*: 0.92
- *Precision Macro Average*: 0.46
- *Precision Micro Average*: 0.92
- *Recall Macro Average*: 0.50

- *Recall Micro Average*: 0.92
- F1 *Macro Average*: 0.48
- F1 *Micro Average*: 0.92

Analisis dan Kesimpulan

Task MapReduce

Berdasarkan hasil yang telah dipaparkan sebelumnya, analisis dan kesimpulan saya terhadap hasil *task* tersebut adalah:

- Status kredit LUNAS mendominasi status kredit TARIKAN, sebesar 94% nasabah telah LUNAS membayar.
- Kemudian, rata-rata SALARY dari keseluruhan nasabahnya adalah 1735009.88; Berikut persebarannya:
    - Yang berstatus LUNAS, rata-rata SALARY-nya sebesar 1739076.76
    - Yang berstatus TARIKAN, rata-rata SALARY-nya sebesar 1667492.53
- Kesimpulannya adalah, yang berstatus LUNAS, rata-rata SALARY-nya lebih tinggi dibanding rata-rata SALARY-nya yang berstatus TARIKAN.

Task Klasifikasi

Berdasarkan hasil yang telah dipaparkan sebelumnya, analisis dan kesimpulan saya terhadap hasil *task* tersebut adalah:

- Hasil evaluasi dari model Naive Bayes dan hasil evaluasi dari model *Linear Support Vector Machine,* memberikan hasil yang sama persis.
- Dari hasil evaluasi tersebut, memberikan hasil akurasi sebesar 92%, yang artinya, kedua model tersebut sama-sama berhasil untuk melakukan *task* klasifikasi dengan sangat baik.
- Kesimpulannya adalah, model Naive Bayes dan model *Linear Support Vector Machine* sama-sama dapat mengklasifikasi STATUS kredit pada kredit.csv dengan sangat baik.