Database e Data Mining

Practice N. 5

The goal of the practice is to get familiar with Hive in order to query and manipulate a data source stored in HDFS.

Data source location: /user/master2013/data/USCensus1990

Data source attributes:

1. caseid
2. dAge
3. dAncstry1
4. dAncstry2
5. iAvail
6. iCitizen
7. iClass
8. dDepart
9. iDisabl1
10. iDisabl2
11. iEnglish
12. iFeb55
13. iFertil
14. dHispanic
15. dHour89
16. dHours
17. iImmigr
18. dIncome1
19. dIncome2
20. dIncome3
21. dIncome4
22. dIncome5
23. dIncome6
24. dIncome7
25. dIncome8
26. dIndustry
27. iKorean
28. iLang1
29. iLooking
30. iMarital
31. iMay75880
32. iMeans
33. iMilitary
34. iMobility
35. iMobillim
36. dOccup
37. iOthrserv
38. iPerscare
39. dPOB
40. dPoverty
41. dPwgt1
42. iRagechld
43. dRearning
44. iRelat1
45. iRelat2
46. iRemplpar
47. iRiders
48. iRlabor
49. iRownchld
50. dRpincome
51. iRPOB
52. iRrelchld
53. iRspouse
54. iRvetserv
55. iSchool
56. iSept80
57. iSex
58. iSubfam1
59. iSubfam2
60. iTmpabsnt
61. dTravtime
62. iVietnam
63. dWeek89
64. iWork89
65. iWorklwk
66. iWWII
67. iYearsch
68. iYearwrk
69. dYrsserv

iMarital attribute: 0 (NO) - 1 (YES)

dAge attribute

range	value
1-12	1

13-19	2
20-29	3
30-39	4
40-49	5
50-64	6
>=65	7

1. **Analyze the source dataset to understand the format**
2. **Create a new database db_your_name in the HIVE metastore**
3. **Create a data sub-directory inside your user directory and copy the source dataset inside it**
4. **Write an HiveQL script to create an external table for the data source, *inside your database***
5. **Move the dataset from your data subdirectory inside the table directory**
6. **Answer to the following queries**

    a. Select the number of people for each range of age

    b. Select the number of people for each range of age. Sort the result by decreasing value of the number of people

    c. Create a new table containing the number of people for each range of age

    d. Query the new table to check the content

    e. Create a table with the schema (caseid, dage, iMarital)

    f. Insert in the previously created table all married people with age in range 20-29 and 30-39

    g. Query the new table to check the content