Classifying Census Income of US (1994 to 1995) bases on multi-variables by deploying Apache Spark on Kubernetes

Hao Tian
Indiana University
School of Informatics, Computing and Engineering
Bloomington, IN 47408, USA
tian4@indiana.edu

ABSTRACT

This the the report of final project in INFO-I 524 Advanced Cloud Computing. The project includes cloud computing technologies, machine learning algorithms and big data analyzing concepts. The project direction and procedure follows the Section 90.3 in Chapter 90 in *the Handbook of Clouds and Big Data*.

KEYWORDS

Apache Spark, Kubernetes, Census Income, Classification, Python

1 INTRODUCTION

This project is to build a cloud computing application for classifing the census income data of the adults in the United State during 1994 to 1995 by multiple variables, such as educational background, work categories, geographic location and so on. The ultimate goal is to use the result of classification on the data to conclude the factors which influence the income of the people, and try to concludes the general economical status of the United State during that period. The cloud computing application is deployed on the Kubernetes system with Apache Spark APIs, the data will be classified by some specific machine learning algorithms in the Spark MLlib. The major code for constructing the project is Python.

The general procedure of this project are (follows the priciples of Section 90.3.2 in *the Handbook of Clouds and Big Data*):

(1) Find the data size of census income data from the UCI Machine Learning Repository. The data size is 103.9 MB [**?**].
(2) Deploy the Kubernetes system with the Apache Spark.
(3) Study the data, find the methods for classifying the data with different variables. Each variable in the data could be a potential factors that splits the high-income population and low-income population.

    The current plan to classify the data is by each variables: every variables is a factor which might influences the differeny levels of income. Each variables of the data splits the incomes into 3 different levels: high, medium, low. The standards of dividing the levels should be considered by the general distribution of the incomes in the data.
(4) Clean or reconstruct the data. **This step will likely be repeated lately for fitting specific machine learning algorithms.**
(5) Using Swagger or Flask Rest Service to send the data into the system that has been deployed and get the result of classification.
(6) Take the output to my personal computer.
(7) Create a Makefile (for basic commmands) and requirements.txt for pip installation.
(8) Pack the project with dockers and finished the report

This is a general percedure of working on this project, the procedures might change in details bases on different situations.

2 ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

REFERENCES

[] [n. d.]. UCI Machine Learning Repository: Census Income Data Set. ([n. d.]). https://archive.ics.uci.edu/ml/datasets/census+income