A Big Data Analysis Framework Using Apache Spark & Deep Learning

Lecturer: Mr. CHAN Sophal

Presented by: TOUCH Sopheak
HENG Seyha

OUTLINES

- INTRODUCTION
- RELATED WORK
- PROPOSED APPROACH
- EXPERIMENTATION
- CONCLUSION

INTRODUCTION

- The amount of data is growing at an exponential rate, and organizations need tools to manage and analyze it.
- Big data analytics can lead to faster and more efficient operations.
- There are many different big data tools available, including Hadoop and Apache Spark.
- GPUs can be used for big data processing, but they are not always economically feasible or accessible.
- Apache Spark is a more efficient and robust tool than Hadoop for big data processing.

RELATED WORK

- Apache Spark: Parallel computing RDDs, MLlib, usage of Spark to analyze Twitter dataset.
- Deep learning: Data classification, fuzzy NN model based on MLP backpropagation.
- Cascade learning: Connector between Apache Spark & MLP, case where class are heavily imbalanced and a suitable inference can't be gathered from data. OnionNet is a feature-sharing classifier, where subsequent stages add both new layers as well as new feature channel to the previous ones.

PROPOSED APPROACH

- Big data analysis using Spark: pre-processed data through MLlib to create probability of each data points.
- Cascading: using the knowledge obtained from 1 model to train another model, modified original dataset which give an attribute that closely assemble, strong distinguishing feature in the dataset.
- Deep learning: knowledge obtained from cascading is used to train MLP , which can define the depth of the network according to the complexity of the problem and system computational complexity.

PROPOSED APPROACH (Cont.)

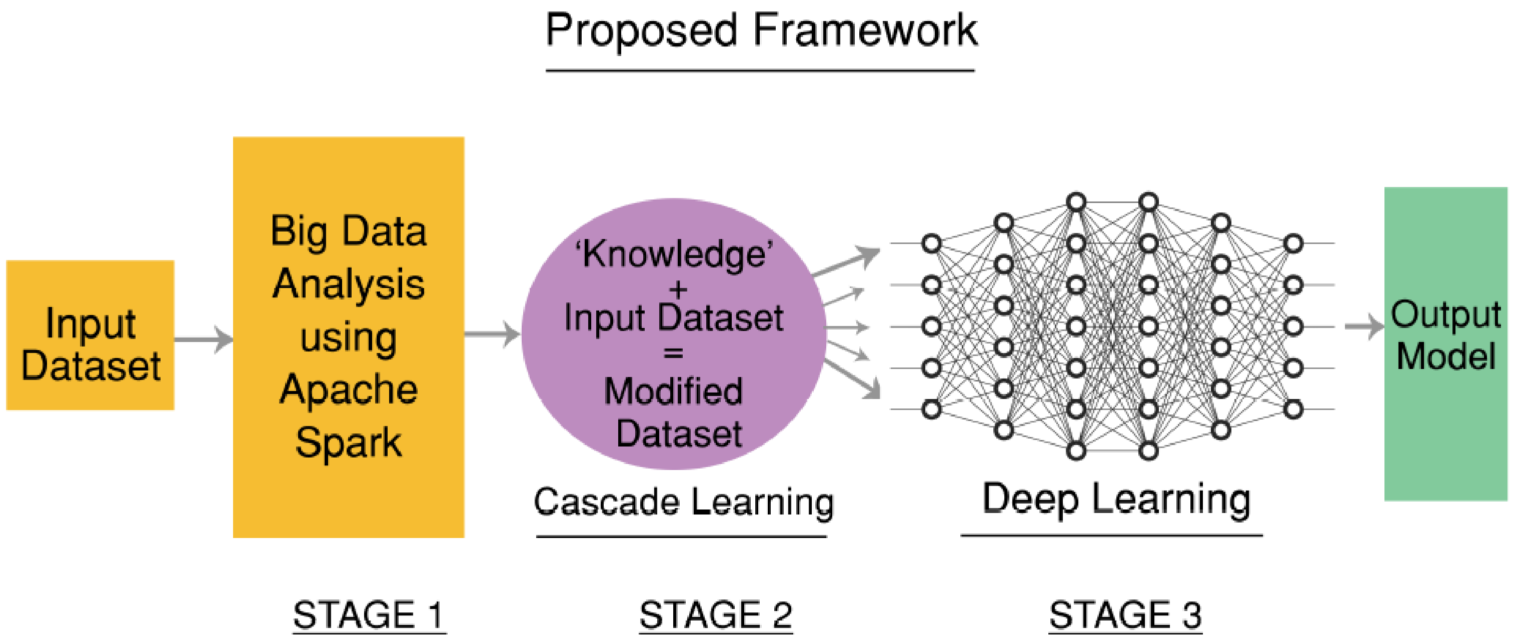

Fig. 1: Schematic Representation of the Proposed Framework

EXPERIMENT

H-1B Visa Application

Task 1: Classification on the basis of "Case-status", which try to predict the status of the Visa.

Task 2: Classification on the basis of the "Prevailing_wage", which try to predict the salary for the set of the threshold.

Task 3: Recommendation on the basis of prevailing wage, which recommed the otimal salary range that applicant should be negotiate.

CONCLUSION

- Novel framework: combine Apache Spark, Deep learning with the Cascading.
- Improved accuracy and efficiency: Enable faster analysis with less computational complexity and higher accuracy.
- Versatility: Applicable to various machine learning tasks.

THANK YOU!!!