Report – A6 – Airline Prediction

Hu,Ruinan and Aswathanarayana,Naveen

1 Introduction

The following project takes in historical data to build model on AWS using Elastic Map Reduce. It also takes a single test file which contains the future flight details we use to predict. Finally the output from this program contains the prediction in the format <FL_NUM>_<FL_DATE>_<CRS_DEP_TIME>, logical. The first column uniquely identifies a flight and the second is TRUE if the flight will be late. We have parallelized the training of the model through AWS calling R instance from each reducer. Hence we have attained parallelism when training the model which reflects the small amount of time taken to train about 600 models in 2 minutes.

2 Data structure

Mapper:

(Key,Value) => (TEXT, TEXT)
=> (<MONTH, ORIGIN>,<CARRIER, ORIGIN, DESTINATION, YEAR, MONTH, DAYOFMONTH, CRSARRTIME, CRSDEPTIME, CRSELAPSEDTIME, ARRDEL15>)

Reducer:

(Key,Value) => (TEXT, TEXT)
=> (<MONTH, ORIGIN>,<CARRIER, ORIGIN, DESTINATION, YEAR, MONTH, DAYOFMONTH, CRSARRTIME, CRSDEPTIME, CRSELAPSEDTIME, ARRDEL15>)

3 Design

Algorithm used for prediction is Random Forest implementation using R. MapReduce framework is used to process the training data parallel on the AWS clusters. We have filtered factors from the existing data such as Carrier and Cities. We have chosen 50 most popular cities and 10 most popular carriers for prediction. Carrier and Cities can be found in the last section. The columns we have chosen as the factors for the prediction algorithm are given below.

CARRIER, ORIGIN, DESTINATION, YEAR, MONTH, DAYOFMONTH, CRSARRTIME, CRSDEPTIME, CRSELAPSEDTIME, ARRDEL15.

Mapper filters the data to get above mentioned columns. Since the key to the mapper is month and cities, we will obtain 12 x 50 datasets from the training data. The reducer receives this dataset and provides each of the dataset to different reducers. Each reducer starts an R instance using Rserve. We have achieved parallelism in generating the models for the dataset given to reducers. Totally 12 x 50 models are generated by the Rserve and written into local file system of each cluster started during map reduce stage.

5 Results:

Confusion Matrix

	Predicted True	**Predicted False**
Actual True	**201569**	**1505052**
Actual False	**166136**	**1615200**

Accuracy = Sum of percentage of on-time flights misclassified as delayed and the percentage of delayed flights misclassified as on-time.

Accuracy : 0.98

Execution Time.

1) To generate 50 x 12 models from 50 cities and 12 months with Rserver on each cluster.
Master: 1
Slave: 9
Type: m3.xlarge
Time Minutes: 2

2) To split the redact data from file 98redacted.csv.gz to 50 x 12 datasets on local machine.
Processor: i5
Ram: 8 GB
HDD speed: 5200 rpm.
Time Minutes: 8

3) To generate the result file from R
Processor: i5
Ram: 8 GB
HDD speed: 5200 rpm.
Time Minutes: 8

4) To build the confusion matrix
Processor: i5
Ram: 8 GB
HDD speed: 5200 rpm.
Time Minutes: 2