A7 Routing Report

Pankaj Tripathi, Kartik Mahaley, Shakti Patro, Chen Bai

March 19, 2016

The report encompasses the details about the design choice, implementation, study of performance, accuracy of the solution and the graph plotted considering the code being run on AWS EMR Cluster.

Introduction:

The problem statement for the assignment is that given an origin, a destination and a date, propose two-hops routes that minimize the chance of missed connections. Further elaboration for the assignment is furnished below.

For a passenger travelling from Boston to Chicago, there are varied routes that the passenger can travel from. The passenger can travel to Chicago either via Washington or via some other cities. The assignment requires us to give a list of such routes(hops) between an origin and a destination along with the cost attached to the route. The cost here is the travel duration. If the suggested hop is a missed connection then there is a penalty attached to the suggested route which is 100 hrs.

Design:

The design that we have chosen to implement the task involves three phases furnished below.
1. **Phase-1:Model Creation**
2. **Phase-2:Precition**
3. **Phase-3:Validation**

A detailed description of these phases and design is provided in the implementation section.

Implementation:

Phase-1: Model Creation

This phase involves reading history files, getting connections between an origin and destination, reading connections from the previous MapReduce job and build a model. This phase has two MapReduce job at its behest which does all the processing.

MapReduce Job-1: This job reads historical data and gives all the connections possible between an origin and destination as the output. For this job, the key is Key{Carrier, YearMonth, IntermediateStop}. In key YearMonth is year and month value concatenated while intermediate stop is the origin of the second hop. This MapReduce job for the corresponding keys will give the possible connections with output as {Carrier, Year, Month, DayOfMonth, DayOfWeek, Origin, IntermediateHop, Destination, FlightNum-1, FlightNum-2, ScheduledArrivalFlight-1, CRSElapsedTime, Distance, Duration, Missed}

MapReduce Job-2: This job is used for building the model. It takes the output of first MapReduce job as input and then feeds the values to the reducer based on Month and Origin.The Reducer then is used in building a model using RandomForest Algorithm of quickml.

Phase-2: Prediction

This phase involves reading test data, getting connections between an origin and destination, reading connections and request data and then predict of all the connections(hops) predicted whether the connection was missed or not. This phase similar to phase-1 has same two MapReduce jobs at its behest doing the same processing where MapReduce job-1 reads the test data.

Phase-3: Validation

This phase involves comparing the output from the phase-2 with the validation file provided to check whether the predicted connection(hop) is a missed connection or not.

Execution:

The phases with there execution details are furnished below. Total time taken is 14.1 mins.

Model Creation: Phase-1 is executed on AWS.
Master: 1
Slave: 9
Type: m3.xlarge
Time Minutes: 6.7

Prediction: Phase-2 is executed on AWS.
Master: 1
Slave: 9
Type: m3.xlarge
Time Minutes: 7.3

Validation: Phase-3 is executed on local machine.
Processor: i5
Ram: 8 GB
HDD speed: 5200 rpm.
Time Minutes: 0.11

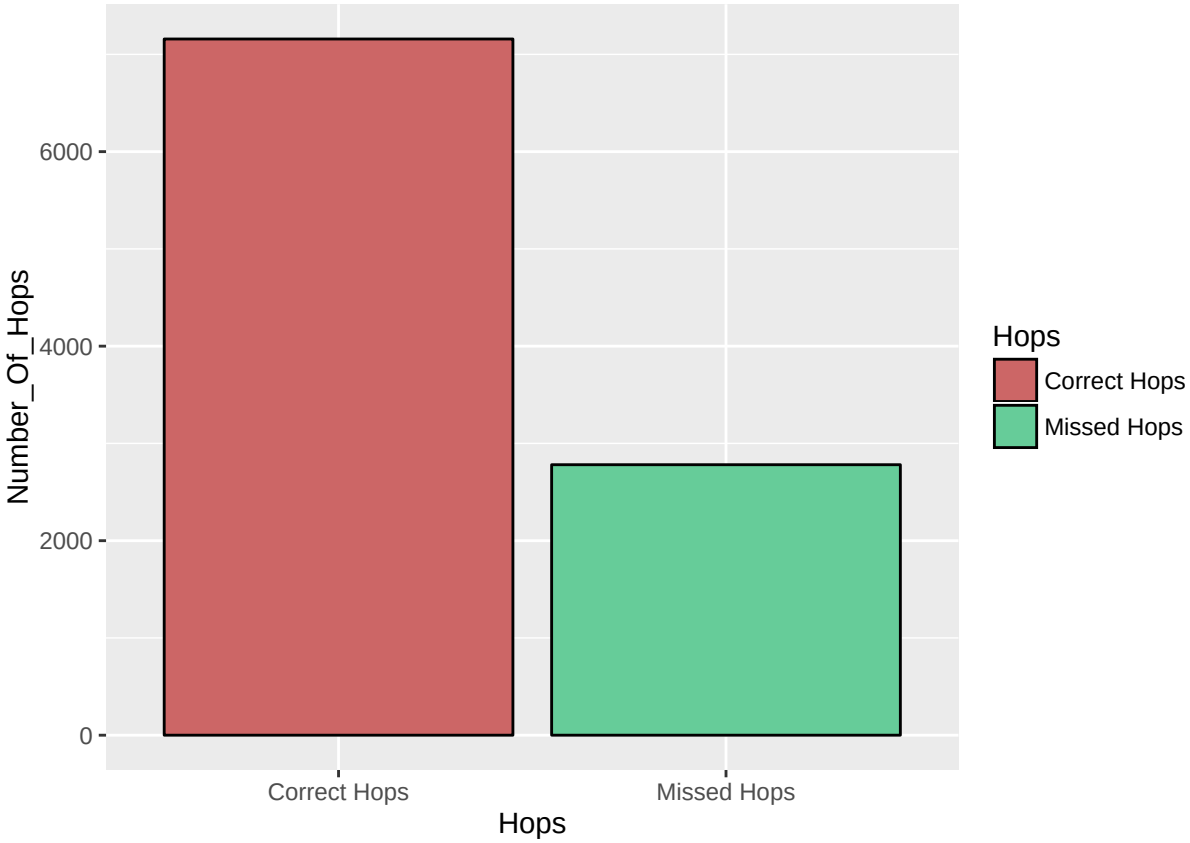
Task Distribution:

Phase-1: Shakti Patro, Kartik Mahaley
Phase-2: Shakti Patro, Chen Bai
Phase-3: Pankaj Tripathi
Packaging: Kartik Mahaley
Code Maintainence: Chen Bai, Shakti Patro
Report: Pankaj Tripathi

Results and Graphs:

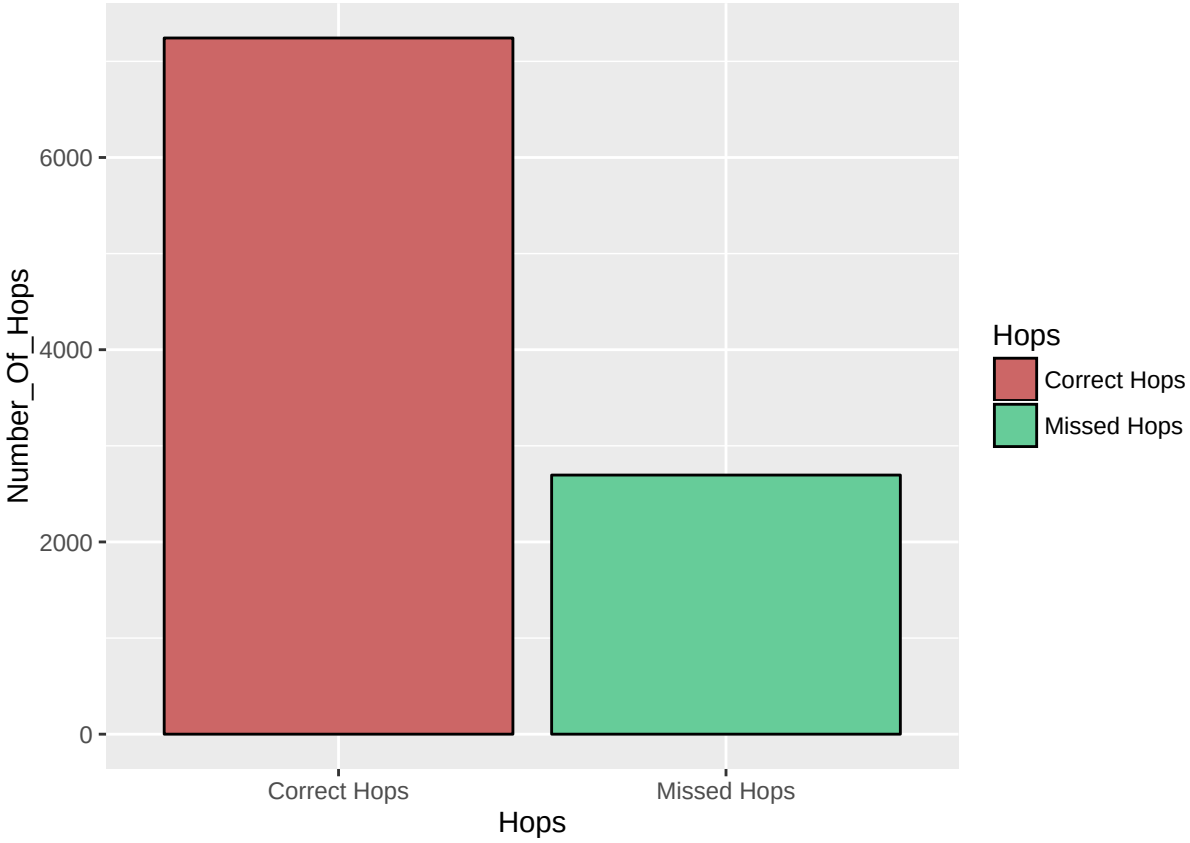
We build two models. The graphs for each has been furnished below.
Model 1: In this case we have taken key as Key{Origin, Destination}

Correct Hops	Missed Hops	Total Predicted Hops
7158	2781	9939



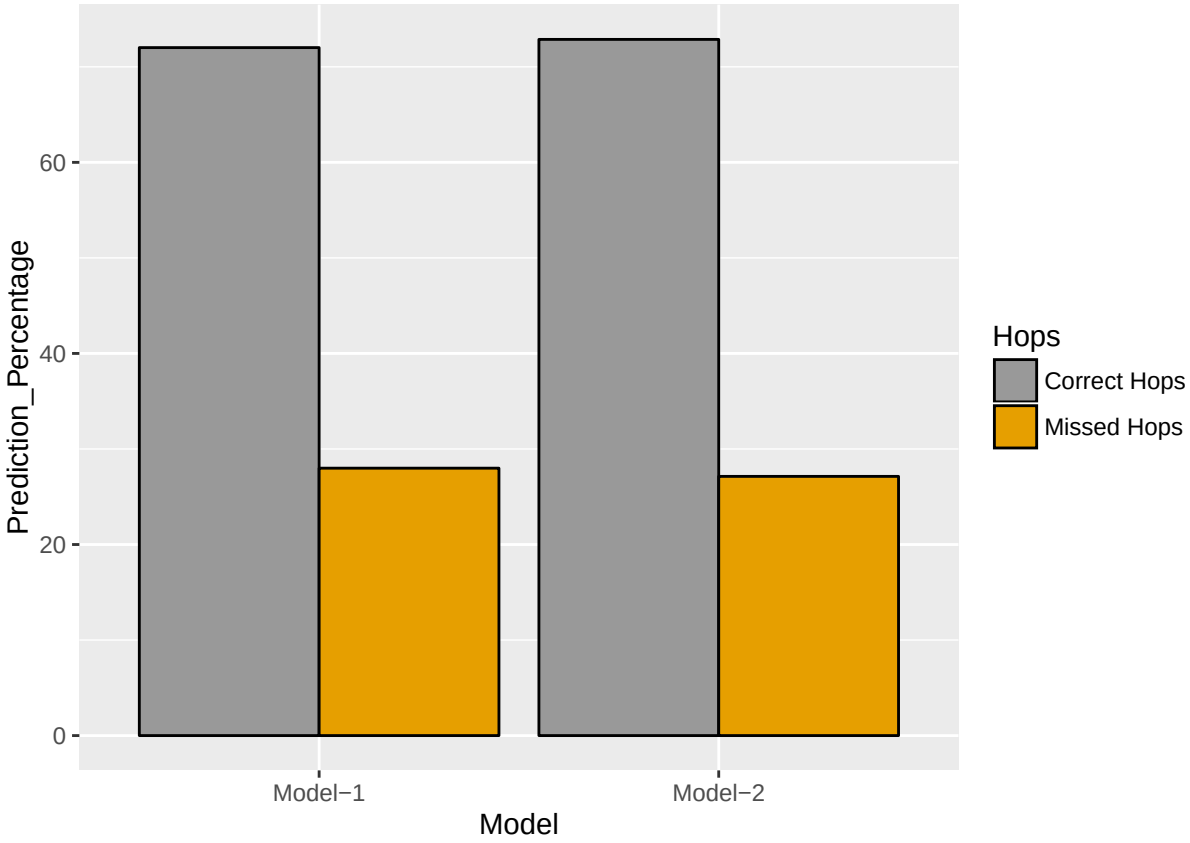
Model 2: In this case we have taken key as Key{Month, Origin}

Correct Hops	Missed Hops	Total Predicted Hops
7243	2696	9939



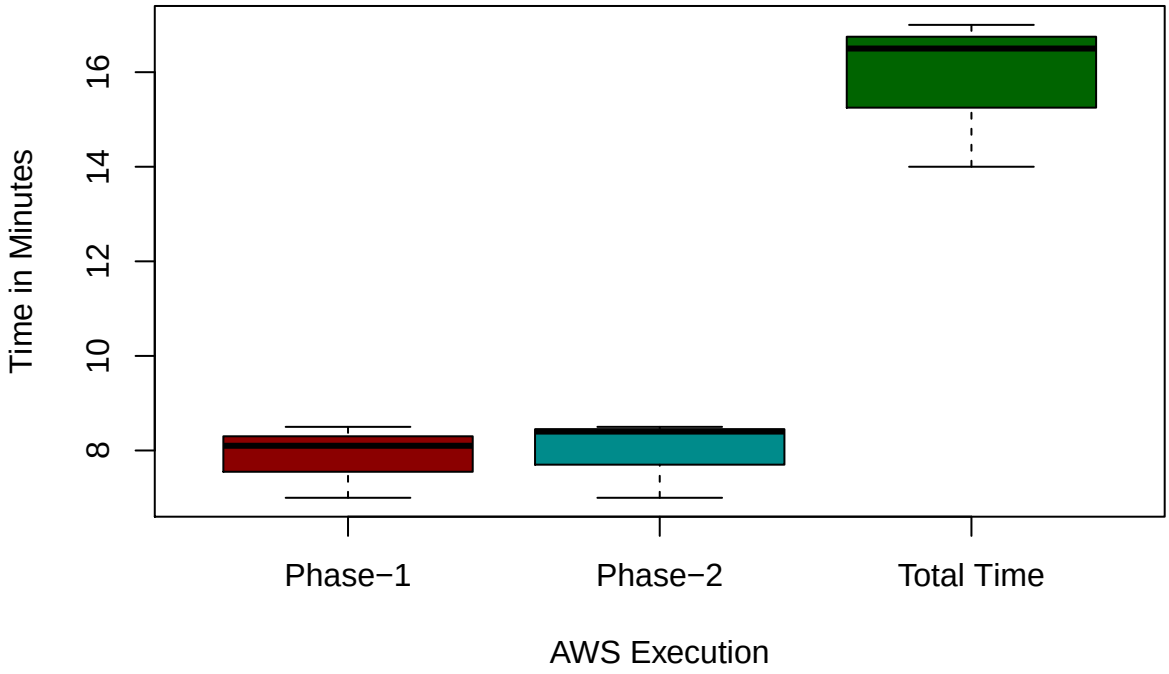
Accuracy of Prediction in Percentage:

Model	% Correct Hops	% Missed Hops
1	72.01	27.99
2	72.87	27.13



Execution time: We executed project 3 times. Time taken for each execution is given below.

Type	Num-1	Num-2	Num-3
Phase-1	8.1	8.5	7
Phase-2	8.4	8.5	7
Total	16.5	17.0	14



Conclusion:

We build two models with different keys using Random Forest algorithm. There is almost no change as far as accuracy is concerned. Our final output was obtained using Model-2 {Month, Origin} which generates 2722 models compared to 13750 models generated by Model-1 {Origin, Destination}. In either cases the output that we are getting with both the models and the fields that we have considered viz {Carrier, Year, Month, DayOfMonth, DayOfWeek, Origin, IntermediateHop, Destination, FlightNum-1, FlightNum-2, ScheduledArrivalFlight-1, CRSElapsedTime, Distance, Duration, Missed} give us a great **accuracy(~72%)**.