

ILS 604 Big Data Analytics for Web and Text

Homework 2 (Due Mar 8, 2016 Tue midnight)

Data Set

Bureau of transportation statistics: <http://stat-computing.org/dataexpo/2009/the-data.html>

Data set and description available at http://www.transtats.bts.gov/Fields.asp?Table_ID=236

Goal

Create MongoDB Index for this dataset. The original dataset is provided in CSV format (MongoDB can handle CSV format data via mongoimport commend

<https://docs.mongodb.org/manual/reference/program/mongoimport/>)

Before you start working on this homework, please read and practice the MondoDB lab

https://github.com/chenmiao/Big_Data_Analytics_Web_Text/wiki/MongoDB-lab

To Report

1. Show me the DB statistics. Do you need to create index for some field(s) of this dataset?
2. Aggregate WeatherDelay, CarrierDelay, SecurityDelay for each target FlightNum and Origin airport code via MongoDB query (aggregate or mapReduce function? Or Both?). What you found from these query? Give me queries and a brief report.
3. Propose a mapReduce task for this data by using data. Show me the significance of this problem, what is your query? And summarize the results.

** Think about this problem: can you index this data by using graphical database, e.g., Neo4j? If the data is indexed in graph format, what kind(s) of problem(s) we can address? You don't have to answer this question in the submission.*

Homework due Tue Mar 8, 2016 midnight. Please submit though Canvas.

Can refer to the MongoDB tutorial:

<https://docs.mongodb.org/getting-started/shell/introduction/> (basic introduction)

<https://docs.mongodb.org/manual/aggregation/> (for Aggregation and MapReduce functions)

Individual submission. All students should finish it independently.