Project Proposal:

I intend to use the **Reporting Carrier On-Time Performance (1987-present)** dataset.This dataset contains information about the air travel schedules.Specifically,it consists details about the arrival/delay duration,time difference,reasons for delay etc.The dataset is versatile and isolated and the topic is pretty interesting and unique. I want to perform various analytics using map reduce as described below:

1)We can use basic map reduce programs to get analysis like:
a)Number of air carriers from 1987-2020
b)Number of times an itinerary was not delayed.
c)Number of source/destination locations from 1987-2020.

2)I plan to use summarization patterns discussed as below:
a)Count the number of times a flight was delayed in a month/year.
b)Find the maximum/minimum delay of an itinerary
c)We could also draw insights like how the air delay times is deviating from average delay times by calculating the mean and standard deviation time.
d)We can generate specific data like how many destination each source locations have by performing map reduce inverted index pattern.

3)We can apply filtering patterns to the dataset to get few analytics like
a)Find out tail numbers starting with certain letter in the tail code.
b)Find out destinations to a particular location.
c)Get the top most frequented destinations
d)Get the distinct carriers in from 1987-2020

To get the above results various concepts of mapreduce discussed in the class will be used like secondary sorting,summerization patterns,filtering patterns,map-reduce chaining etc.

4)I'll also use Hive and Pig tools to get few of the above mentioned analysis

Note : I also intend to use machine learning algorithms to predict few of the analysis if time permits.

Dataset Description:

S.No	**Name**	**Description**
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Dataset Resources:

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=
http://stat-computing.org/dataexpo/2009/the-data.html
https://www.transtats.bts.gov/Fields.asp?Table_ID=236