Weather Data Analysis using Hadoop to Mitigate Event Planning Disasters

Khaled Almgren, Saud Alshahrani, Jeongkyu Lee (Advisor)

Department of Computer Science, University of Bridgeport, CT

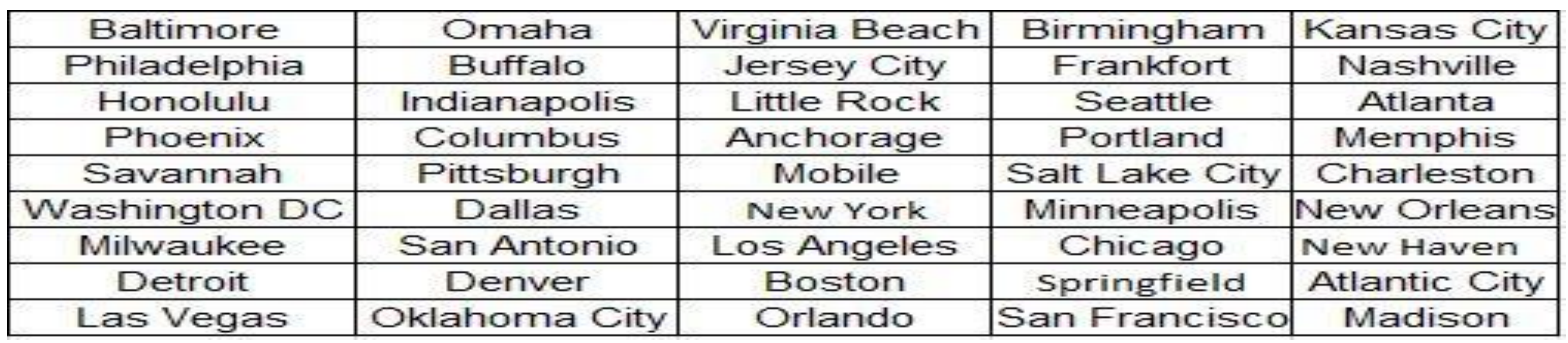
Abstract

This poster presents the design and implementation of weather data analysis using Hadoop distributed system, which can be used for planning outdoor events. The proposed event planning system decides how many appropriate days for outdoor events and activities per month for a different attractive cities based on the analysis of historical weather data. All collected data are stored at HDFS, i.e., Hadoop Distributed File System, and then they are processed and analyzed by using MapReduce programming. As results, we can discover useful information about event planning, such as locations (city), time and statistical data.

Introduction

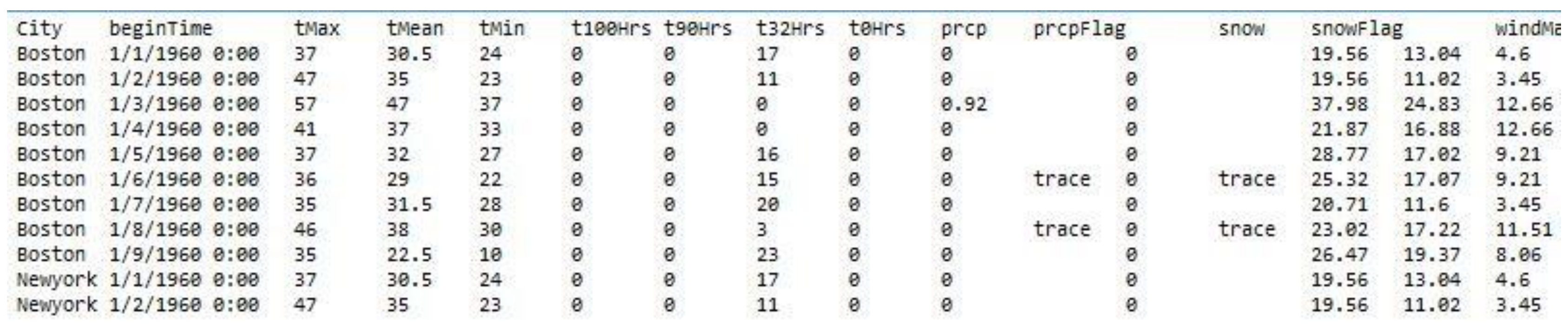
Big data is data with enormous size that is very difficult to be processed with traditional tools. However, once it's processed and analyzed, we can get great and useful information out of it. Hadoop allows us to process big data by using Hadoop Distributed File System (HDFS) and MapReduce. HDFS is used to manage the files and break the data into blocks, and then distribute the blocks across clusters of machines. MapReduce will distribute the tasks that perform map and reduce operations across multiple nodes [1,2].

In united states, there are many events that occur around the year in different cities. These events could take place outdoor or indoor. Organizations who host their events outside such as car shows, concerts, bazaars, festivals, etc may suffer a lot from the frequent weather changes. They need to plan and choose the date for their event months in advance. They choose the data based on many factors, and the most important factor is the weather, but which date to choose? This project tries to answer this question. We collected big data from WeatherSource. The collected data are for forty five cities between 1/1/1960 and 12/31/2013. We only have considered the attractive cities that event planners usually consider such as, Las Vegas and Los Angeles see Figure 1 for cities and see Figure 2 for data sample.



Baltimore	Omaha	Virginia Beach	Birmingham	Kansas City
Philadelphia	Buffalo	Jersey City	Frankfort	Nashville
Honolulu	Indianapolis	Little Rock	Seattle	Atlanta
Phoenix	Columbus	Anchorage	Portland	Memphis
Savannah	Pittsburgh	Mobile	Salt Lake City	Charleston
Washington DC	Dallas	New York	Minneapolis	New Orleans
Milwaukee	San Antonio	Los Angeles	Chicago	New Haven
Detroit	Denver	Boston	Springfield	Atlantic City
Las Vegas	Oklahoma City	Orlando	San Francisco	Madison

Figure 1: Cities

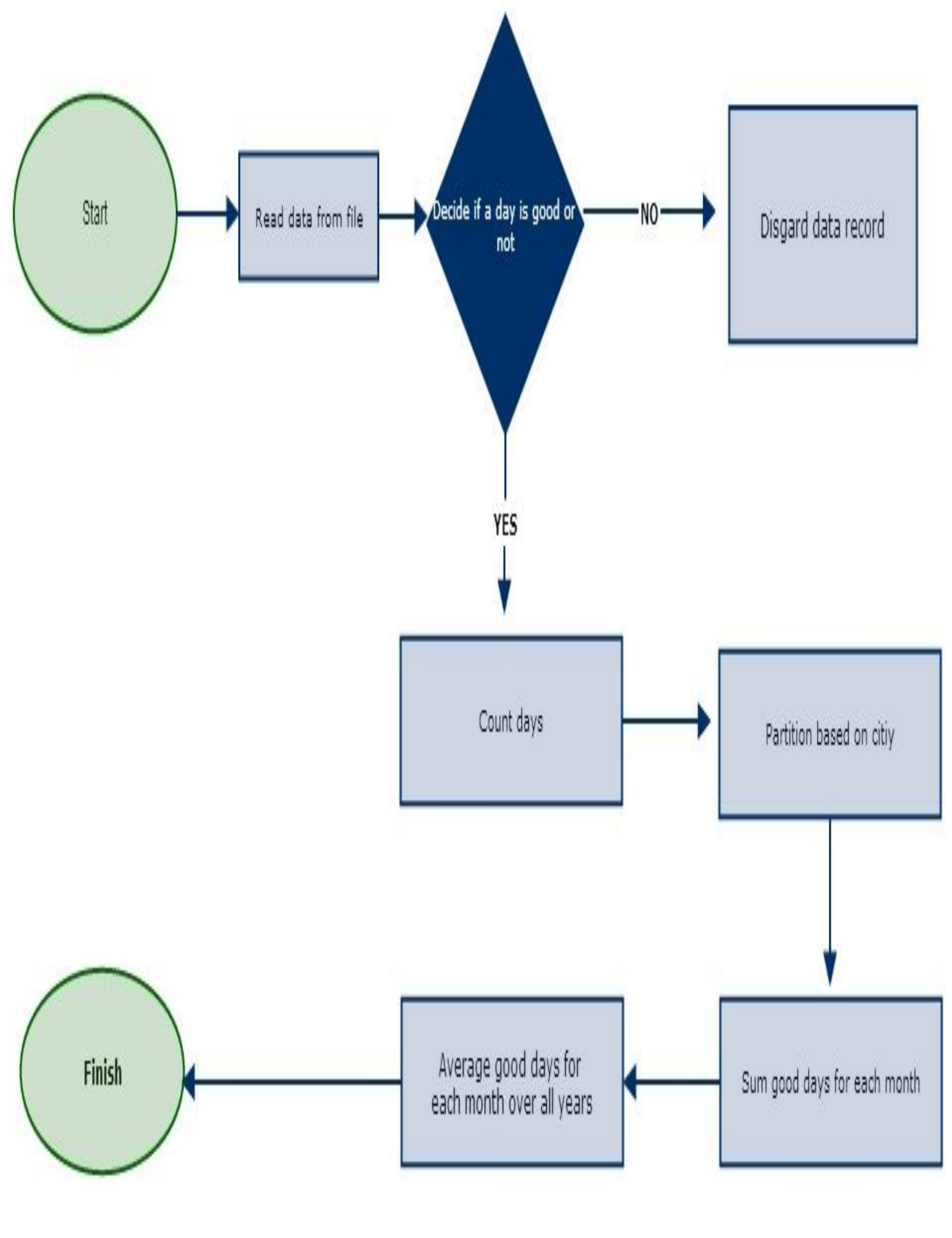


```
City     beginTime       tMax   tMean  tMin   t100Hrs t90Hrs t32Hrs t0Hrs  prcp   prcpFlag        snow   snowFlag          windMa
Boston   1/1/1960 0:00   37     30.5   24     0       0      17     0      0              0              19.56  13.04   4.6
Boston   1/2/1960 0:00   47     35     23     0       0      11     0      0              0              19.56  11.02   3.45
Boston   1/3/1960 0:00   57     47     37     0       0      0      0      0.92           0              37.98  24.83   12.66
Boston   1/4/1960 0:00   41     37     33     0       0      0      0      0              0              21.87  16.88   12.66
Boston   1/5/1960 0:00   37     32     27     0       0      16     0      0              0              28.77  17.02   9.21
Boston   1/6/1960 0:00   36     29     22     0       0      15     0      0      trace   0      trace   25.32  17.07   9.21
Boston   1/7/1960 0:00   35     31.5   28     0       0      20     0      0              0              20.71  11.6    3.45
Boston   1/8/1960 0:00   46     38     30     0       0      3      0      0      trace   0      trace   23.02  17.22   11.51
Boston   1/9/1960 0:00   35     22.5   10     0       0      23     0      0              0              26.47  19.37   8.06
Newyork  1/1/1960 0:00   37     30.5   24     0       0      17     0      0              0              19.56  13.04   4.6
Newyork  1/2/1960 0:00   47     35     23     0       0      11     0      0              0              19.56  11.02   3.45
```

Figure 2: Data Set Sample

Design and Implementation of Event Planner

The initial design of the system is shown in Figure 3. The data is passed as records to the Decide algorithm. The algorithm checks wither a day is a proper day or not, then count the proper days . Then, the data will be partitioned based on cities. The proper days are summed per month. Then the system compute the average of the proper days for the each month for the past 63 years. The decide algorithm makes its decision based on four factors. The four factors are Temperature, Humidity, Overcast, and Wind speed. The algorithm has a number of cases that define wither a day is proper or not based on many conditions. The algorithm is illustrated in Figure 4.



Start
Read data from file
Decide if a day is good or not
NO
Disgard data record
YES
Count days
Partition based on citiy
Sum good days for each month
Average good days for each month over all years
Finish

Figure 3: Data Flow Diagram

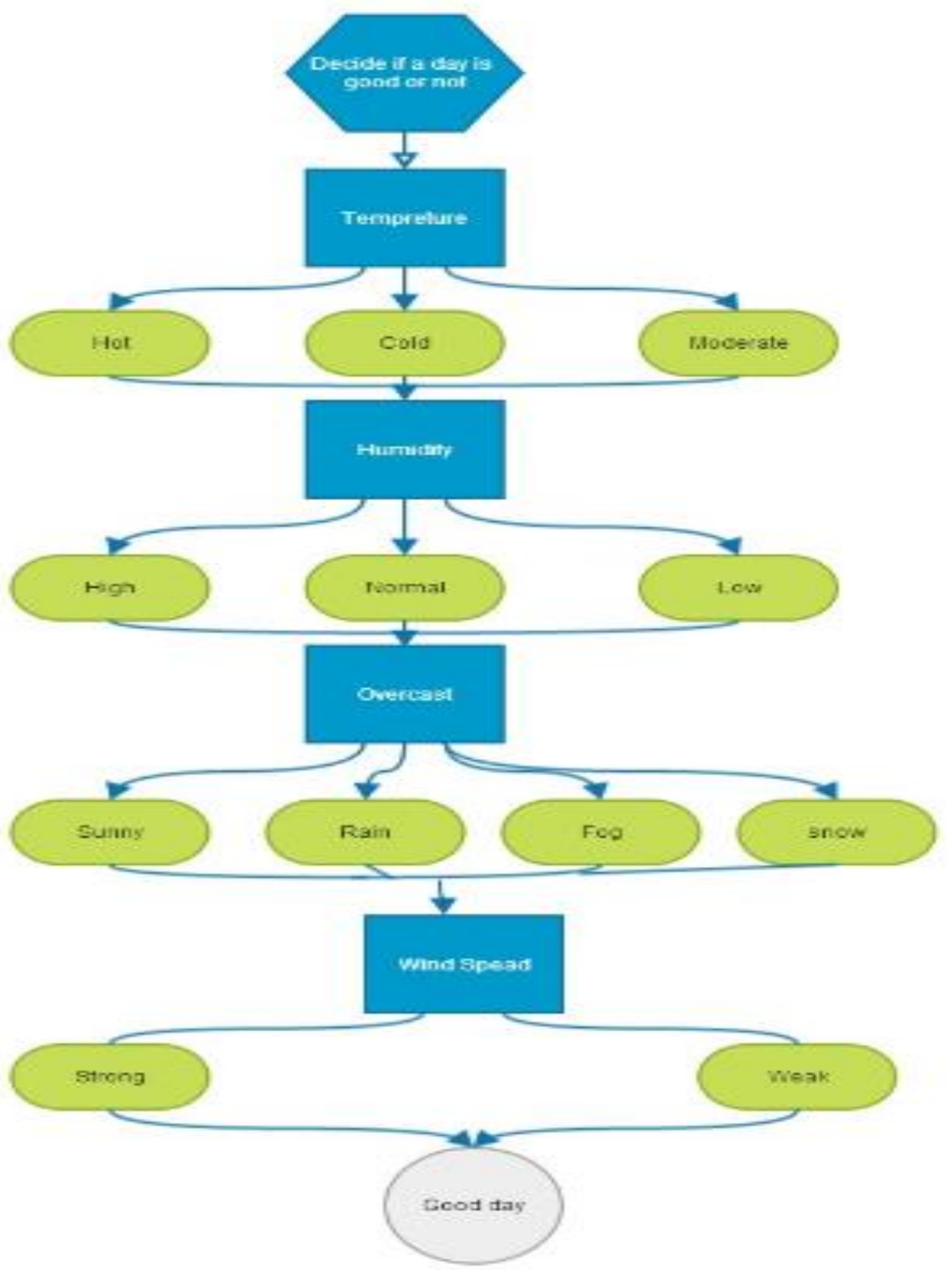


Figure 4: Decide Algorithm

The project is implemented using MapReduce. There are five phases in the project. The first one is the Mapper which contains the decide function to count the proper days. The Partitioner will split the data based on cities. The combiner will sum the proper days for each month for every city. The Shuffle and sort phase will sort the data based on the keys. The Reducer will compute the average of the proper days for each month for the all years.

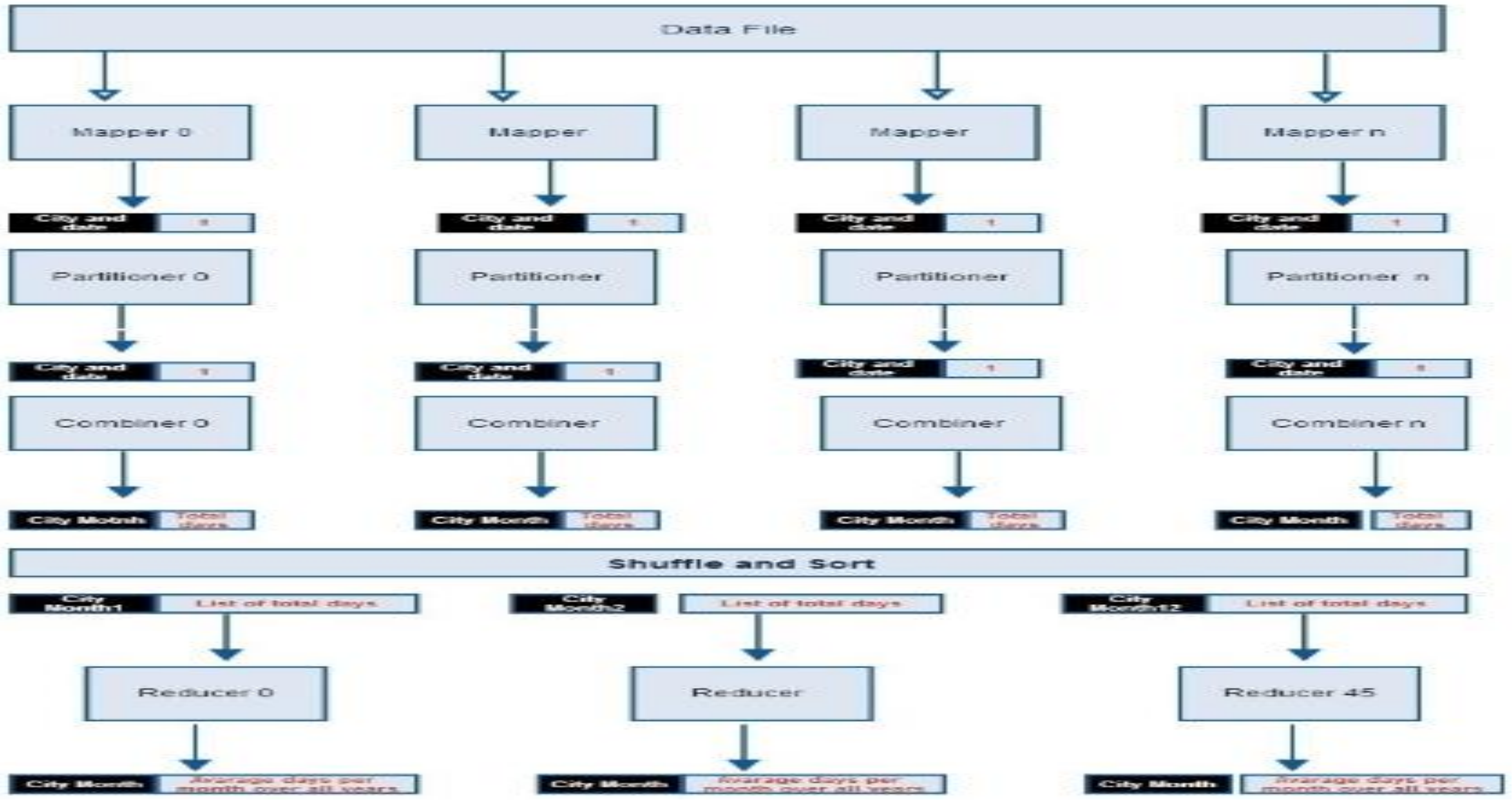


Figure 5: MapReduce design

Data Analysis

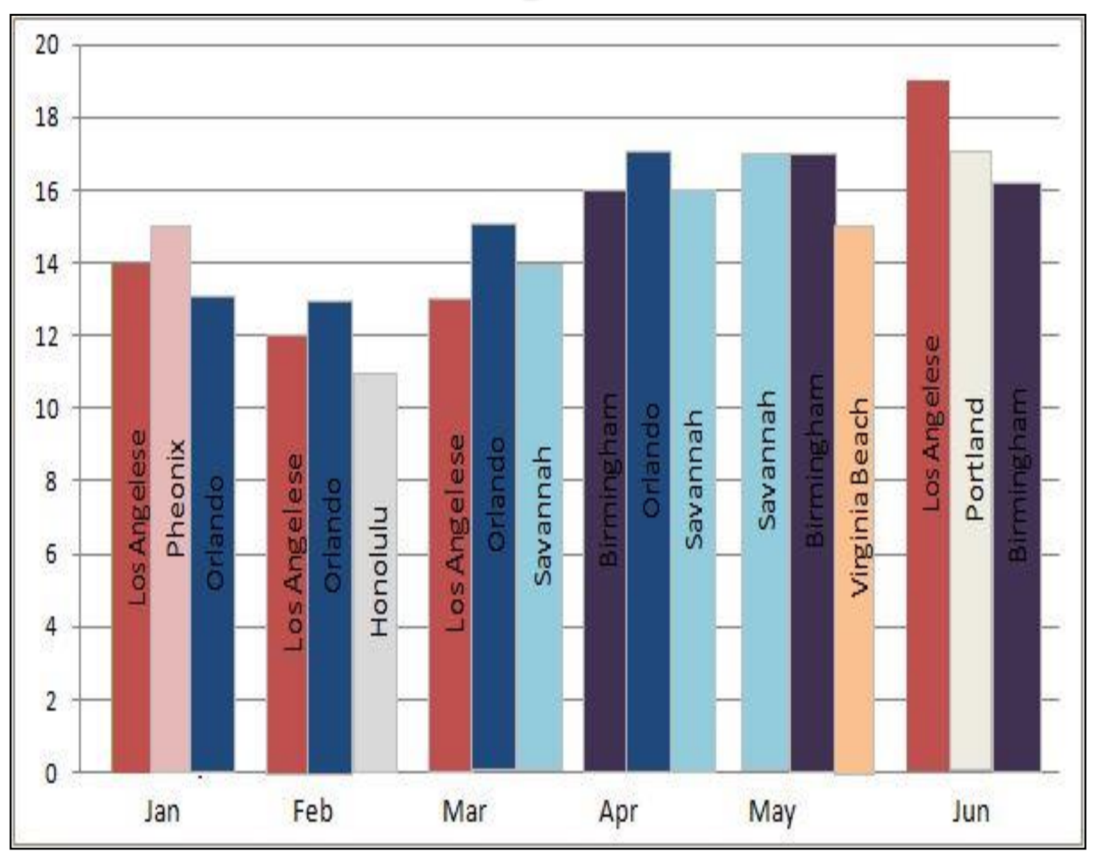


Figure 6: The best three cities for outdoor events between January and June based on the average of proper days for each month

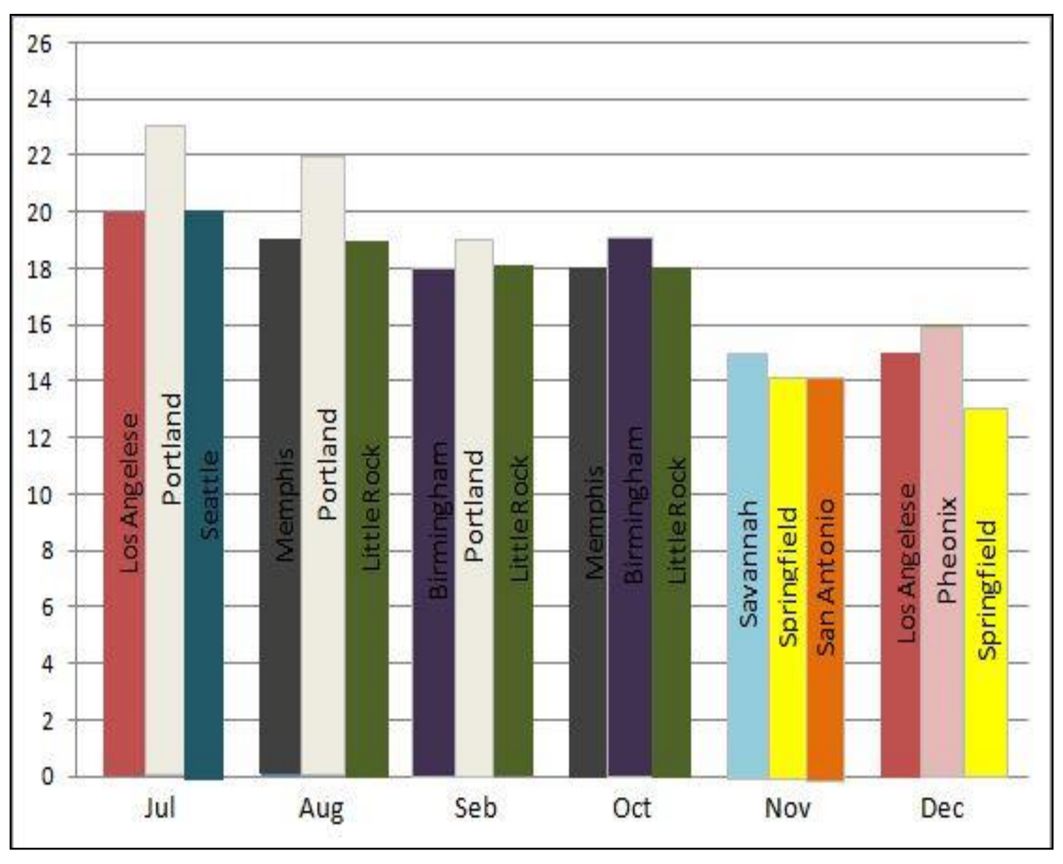


Figure 7: The best three cities for outdoor events between July and December based on the average of proper days for each month

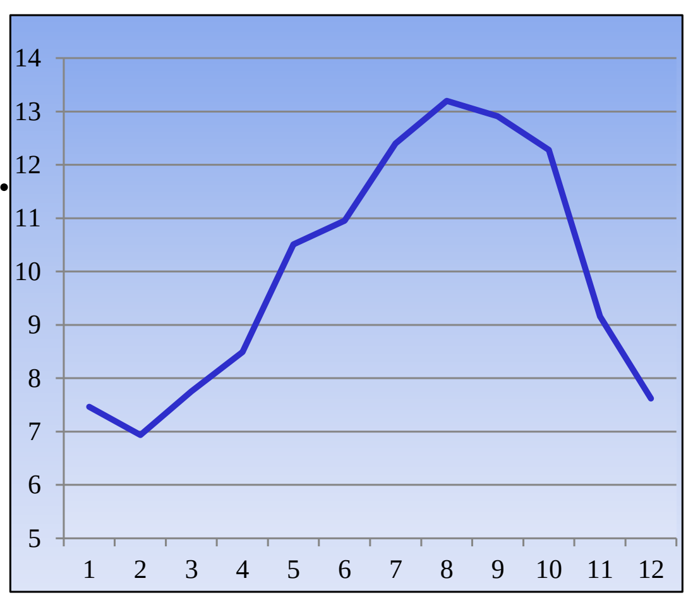


Figure 8: Average of proper days for each month over all cities. August being the best month and February being the worst month

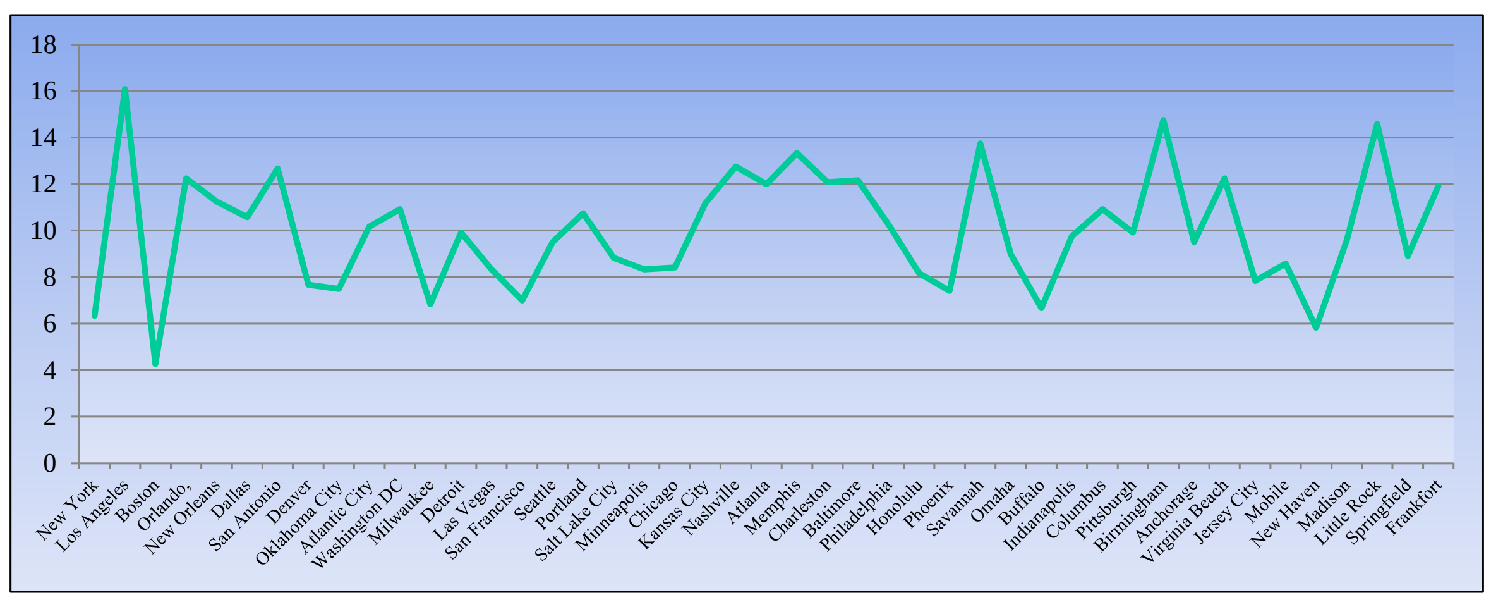


Figure 9: Average of proper days in months for each city. Los Angeles is the best with an average of 16 days monthly, and Boston is the worst with an average of 4 proper days monthly

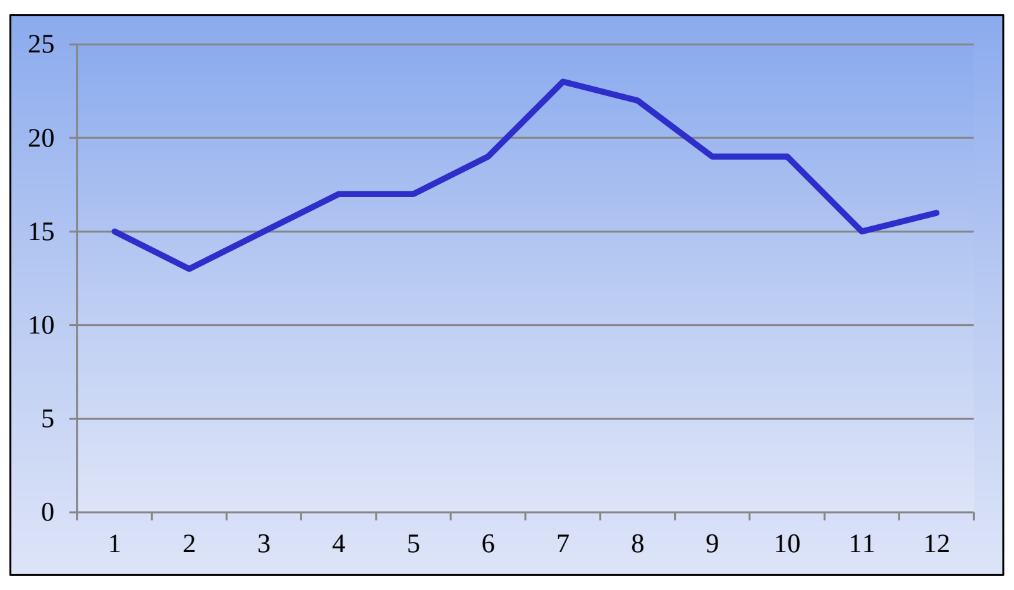


Figure 10: The maximum proper days for each month. July has the highest number of proper days with 24 days. February has the lowest number of days with 13 days

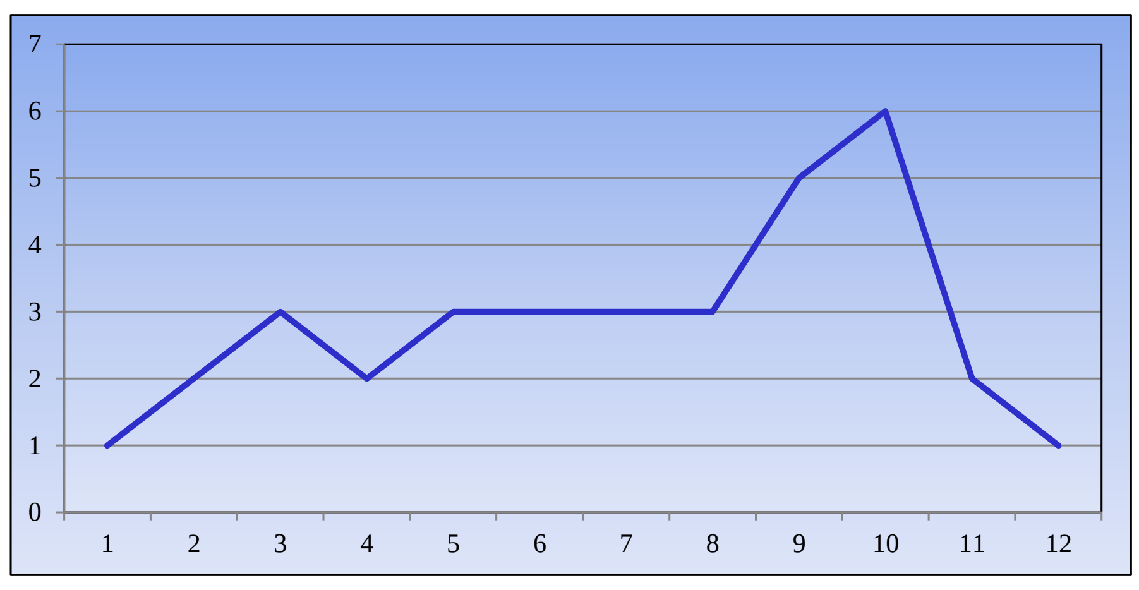


Figure 11:The Minimum proper days for each month. January and December have the minimum number of days while October had the most minimum number of days.

Conclusion and Recommendations

1. Los Angeles has the best weather for outdoor events and Boston has the worst weather for outdoor events
2. August has the best average of proper days over all the cities with an average of 14 days and February has the worst average of proper days over all the cities with an average of 6 days
3. The maximum proper days is 23 days which is in July in Portland and the minimum proper days are in February and December in Buffalo and Portland
4. Using Hadoop to analyze huge data has allowed us to get great insight
5. Next step is to apply data mining techniques to get better decisions

References

1. Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, *14*(2), 1-5. Chicago
2. White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".