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.

Data Analysis
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 all years.

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 whether 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 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 whether a day is proper or not based on many conditions. The algorithm is illustrated in Figure 4.

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 number of days.

References