*Distribution of Weather Events Throughout States and Regions of the United States in 2012*

Carroll   
*Department of Computer Science and Information Technology (Graduate Student)*   
*Eastern Kentucky University*

Richmond, Kentucky, United States  
kenneth\_carroll28@mymail.eku.edu  
Christopher   
*Department of Mathematics and Statistics  
 (Graduate Student)*   
*Eastern Kentucky University*

Richmond, Kentucky, United States  
samuel\_christophe22@mymail.eku.edu

*Abstract*— The main objective of this project was to find out what types of weather events were most common in certain states and areas of the United States in the year 2012 using Hadoop and MapReduce. This information can be used to get an understanding of weather patterns in the US. The project will also test the ability of MapReduce to sort through a sizable data set with a multitude of distinct keys[4]. An additional focus is isolating a key-value pair that can be used for more than one analysis while reducing the overall data to levels that can be executed on a non-distributed system within a reasonable amount of time.

Keywords—Severe Weather, Weather Patterns, Hadoop, MapReduce

# Introduction

For our course project for CSC 785 Big Data, we decided to look at severe weather data for the United States and analyze it. We found a dataset that, when processed, would allow us to figure out what severe weather events were most common in distinct regions and states. Since both authors are located in Kentucky, we had a particular interest in examining the data for it. We also wanted to test Hadoop and MapReduce to find out how it would handle a sizable data set such as this one with multitude of keys. The variety of keys enabled further combinations of concatenation which could allow us to capture multiple observations from a single map reduce job.

Over the course of the project, we found the answers to our questions and discovered some very interesting patterns of weather in the country and state. For instance, Kentucky ranked 6th in the number of tornadoes in the data used. It also was found that the most common top severe event for the states was a severe thunderstorm.

It should be noted that the dataset includes weather events, which does not include all natural disasters. Disasters such as earthquakes and volcanic activity are not included in the dataset because they are “non-meteorological”. While these disasters might contribute to some irregular weather, they are not included for that reason. Additionally, similar events are still segmented based on severity; a tropical depression is separated from a tropical storm which is also counted separately from a hurricane. In these instances, the recorded events are counted on their own, but there are also instances where multiple events could stem from a single cause and are counted separately. For example, a thunderstorm can cause rains, wind, and lightning but within the dataset lightning is counted separate from thunderstorm related winds, and while the storm rain could cause landslides or flash flooding, those events are also counted separately.

# Methods

* This study used Hadoop and MapReduce to process the data. Code for both the mapper and reducer were made in Python.
* The data we used was a set of every severe weather event documented in the United States by the National Oceanic and Atmospheric Association (NOAA) for the year 2012.
* Overall, there were 64,460 severe weather events in 2012 which comprise the data for this project.
* This dataset included several columns that we found unnecessary for our research, so we reduced the dataset to state and event type.
* For the first part of our study, we mapped using a concatenated key of state-event. This made a key out of every combination of state and event type that was present in the dataset and sent it to the reducer as the pair <state\_event, 1>. The script for the mapper can be found in figure 1.
* After receiving this data, the reducer would count the amount of each key present and output <state\_event, total>. This pairing shows the total amount of each event in that state for the dataset. The reducer code is in figure 2.
* A python script was utilized for preprocessing the csv by eliminating columns of no interest, and the new preprocessed data set was then added to the Hadoop HDFS. From there a shell script automates placing the data, structuring the map reduce job, and then moving the output from the HDFS back to the local file system.
* Once the MapReduce program finished running, we obtained the output file to find the most prevalent events in the various states, and some of the patterns of events in certain areas across the US. This is achieved by using a python script to segment the output into state and event specific files that are then accessible to a separate python script that handles visualizing the data via matplotlib. The visualization script was called via a batch file that called the script for each event and each region in the postprocessed files.
* By running this through MapReduce and Hadoop, we would also get an idea of how well the system could handle a sizable dataset that would require a large volume of keys to process, as there were hundreds of state-event combinations that required sorting. This allowed for enough reduction that later operations could be reasonably ran on a single machine for post-processing.
* Hadoop/MapReduce performance can be gauged by observing the length of time that it takes to process the data. In this instance, processing the data took under a minute including time to copy to and from the HDFS.

Figure - Weather Events Mapper Script

Figure - Weather Events Reducer Script

##### Results

Figure - Kentucky Weather Event Distribution

Figure - Alabama Weather Event Distribution

Figure - Georgia Weather Event Distribution

Figure - Hawaii Weather Event Distribution

Figure - High Surf Event Regional Distribution

Figure - Thunderstorm Wind Event Regional Distribution

Figure - Blizzard Weather Event Regional Distribution

Figure - Lightning Event Regional Distribution

Figure - Winter Weather Event Regional Distribution

##### discussion

The results from this study can tell us a lot about weather around the United States, but as we are located in Kentucky that is a good starting point. The breakdown of events for the state can be seen in figure 3. These are the main takeaways from the data out of Kentucky:

* The most common severe weather event in Kentucky is thunderstorm related wind events.
* The least common event was a landslide, with only two reports (about 0.09% of Kentucky severe weather for the year)
* There were 56 reported tornadoes in Kentucky for the year of this dataset (2012). While that is not as high as some states like Kansas, Missouri, or Nebraska, it ranks 6th overall. This shows that tornadoes are something that Kentuckians should be aware of[2]. This may be skewed by the increased number of tornadoes observed in 2012[5].
* Winter weather is also something to look out for, as 292 events were reported for the year. This ranked #2, only behind Colorado by two severe winter weather reports (figure 11). This does not mean that Kentucky has more severe winter weather than all but one other state, as many states such as Minnesota and Wisconsin likely have a higher tolerance before reporting it. The difference is that these states are more prepared for (and used to) winter weather that would be considered severe in other areas. Whatever the case, the data shows that Kentuckians should be prepared for severe weather in the winter months.
* Kentucky is quite varied in its weather, as evidenced by the amount of unique event types seen in the data – 20 distinct events.
* There were 81 instances of flash floods and 28 instances of floods in the state of Kentucky. There are many areas in the state are prone to flooding, so if you are in one of these areas be prepared. It was unexpected to see the majority of flood events skew towards flash floods, which while only 3.5% of weather events in Kentucky, can pose a more significant risk of injury or death compared to more frequent events such as winter weather or drought
* While an occasional wildfire does take place in the state, there is not a large risk. There were only 3 in 2012 in the entire state, as compared to 53 in Arkansas which held the largest share of overall wildfire events in 2012 at 11.04%.

Aside from Kentucky, we also looked at patterns in the US, and some of its associated regions. Some of our main takeaways from that were:

* + Weather events are strongly skewed by geography, example being that 89.39% of reported high surf events were from Hawaii despite 30 states having coastline (figure 7) Despite the entire state being coastal, Hawaii (figure 6) did not report a single instance of coastal flooding and reports of this event were dominated by New Jersey at 16.36% and New York at 12.27%.
  + Another example of local climate skewing reporting is winter weather, as in the more northern regions, winter weather related events dominated large portions of the associated events. Some examples being Alaska having 30.17% blizzards (including 21.22% of total reported blizzards, figure 9), 15.92% winter storm, 11.45% extreme cold and wind chill, along with 5.31% heavy snow. While that is an extreme example with over 50% of overall events being associated to winter weather, this is also observable to a lesser extent New York where 4 out of the top 8 events were winter weather related. Maine also observed 25.09% heavy snow and 15.19% winter storms.
  + States that neighbor each other in a similar latitude tend to have similar distributions in their most frequent events - Georgia and Alabama (Figures 4 and 5) both have their top four most populous weather events as Thunderstorm wind, hail, drought, and heat in that order. This does not mean that neighboring states cannot exhibit wildly different weather event distributions though, as New Mexico and Arizona do not share a single weather event in their top 4 occurrence regardless of position.
  + Regions that are not directly states but can fall under the category of US occupied territories or regions do not tend to have as much diversity in reporting. This could be for infrastructure or budgetary regions but is strongly observable when looking at both Guam and the Virgin Islands, where Guam only reported Tropical Storms, Heavy Rain, and Hurricanes (5 total events) and the Virgin Islands (1 total event) only reported Tropical Storms[3].
  + As mentioned in the Kentucky winter weather segment, location also likely plays a role in how likely an event is to get reported. States with more winter weather are more likely to not report a standard winter storm as severe compared to a state that is not as prepared for winter weather. Another example of this is Delaware having more total drought reports than Arizona despite Arizona being much larger and having a drier climate. This is likely due to Arizona being used to dry weather and having a higher bar set for what constitutes a drought. Some regions might also be more likely to report weather in general. To account for these things, when examining the data, the best takeaways on what weather affects the state the most can be found by their proportion of total events. If a state has a tendency to report more events overall, comparing the proportion of regions to the proportion of other regions is better than examining totals.

##### Conclusion

The dataset was successfully ingested through a map reduce job and was able to be transformed and later visualized. Using this key-value pair for map reduce work allowed for the observation of event distribution for each region and the regional distributions for total events from the same job. This was able to be achieved by mapping the results to two dictionaries, one where the first key was event, which then had keys for each region the event occurred in and the associated event, and the second dictionary had the region as the first key as the region and second key for the nested dictionary as event. This was easily done thanks to both being tied to the same key in the reduce job and while it could have been split into additional map reduce jobs for each specific distribution under observation, the overall reduction was enough to adequately process the output on a single machine using standard python libraries in a non-distributed environment.

It was observable that several events were highly widespread in the United States, where thunderstorm related wind events were observed in the vast majority of regions, and the same was true for lightning. However, this does highlight that there was no event that was reported in every single region or even every single state, so regional location should be a consideration if the avoidance of a given event is of higher priority to an individual. It also is indicative of the greatest risk in this region and could influence expected costs of that area. Where Alaska’s prominent weather event was a blizzard, that could compound the already higher cost of receiving certain goods in the area and require additional costs to ensure that in the event of a blizzard there is adequate supplies and infrastructure to accommodate the event. This can be said for several states, but there are also states like Massachusetts where no event was reported over 18% which could indicate the need to accommodate more types of weather events but with less frequently or under less severity than in states with predominant events

##### References

1. Brooks, H. and Doswell, C., 2021. *Normalized Damage from Major Tornadoes in the United States: 1890–1999*. [online] AMETSOC. Available at: <https://journals.ametsoc.org/view/journals/wefo/16/1/1520-0434\_2001\_016\_0168\_ndfmti\_2\_0\_co\_2.xml> [Accessed 14 November 2021].
2. Pielke, R. and Landsea, C., 2021. *Normalized Hurricane Damages in the United States: 1925–95*. [online] AMETSOC. Available at: <https://journals.ametsoc.org/view/journals/wefo/13/3/1520-0434\_1998\_013\_0621\_nhditu\_2\_0\_co\_2.xml?rskey=XFliMr&result=2> [Accessed 14 November 2021].
3. V. Suryanarayana, B. S. Sathish, A. Ranganayakulu and P. Ganesan, "Novel Weather Data Analysis Using Hadoop and MapReduce – A Case Study," *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 204-207, doi: 10.1109/ICACCS.2019.8728444.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
4. N. O. A. A. US Department of Commerce, “Summary of the March 2, 2012 tornadoes,” *National Weather Service, 02-Apr-2019.* [Online]. Available: https://www.weather.gov/jkl/20120302\_tornadoes. [Accessed: 01-Dec-2021].