# Using Large Scale Clustering To Detect Potential Fire Regions

Name: Michael Lewis
University: University of Illinois At Chicago
Adviser: Karen Langona
University: University of Sao Paulo

## Abstract

Application based streaming has been used in real life scenarios that require real time analysis; leveraging large scale parallel processing over a cluster infrastructure. Finding fire prone locations using streaming methods can play an important factor in evaluating areas that pose immediate fire dangers for the current day. Here, I introduce a data-scalable framework that runs on a Hadoop cluster that is designed to detect potential fire regions.

## Data Collection

- Air Temperature, Dew Temperature, Precipitation, Wind Speed and Humidity were gathered from 25 weather stations in the central California region from 01-01-2013 – 06-30-2013[1].
- Daily fire indices were calculated using Angstrom, Telicyn, and Monte Alegre fire index formulas.

## Monte Alegre Formula

The Monte Alegre and Monte Alegre Plus Formulas originated in Brazil. The formula includes the previous day's index with the current day's humidity based calculation. Monte Alegre Plus also includes wind speed in its calculation. For both formulas, precipitation level applies a factor with regards to how much of the previous day's results to include.

| Monte Alegre Index | Risk Level |
|---|---|
| ≤ 1 | No Risk |
| 1.1 – 3.0 | Small Risk |
| 3.1 – 8 | Medium Risk |
| 8.1 – 20 | High Risk |
| > 20 | Very High Risk |

| Monte Alegre Plus Index | Risk Level |
|---|---|
| ≤ 3 | No Risk |
| 3.1 – 8.0 | Small Risk |
| 8.0 – 14 | Medium Risk |
| 14.1 – 20 | High Risk |
| > 24 | Very High Risk |

## Angstrom, Telicyn Index

The Angstrom index originated from Sweden. The formula takes the current day's humidity and air temperature. The Telicyn index originated from Russia. Like the Monte Alegre formula, the previous day's fire index is used to add to the current day's index, thus creating increasing fire index values for consecutive days without rain. The Telicyn index takes in the air temperature and dew temperature to calculate its fire index.

| Angstrom Index | Risk Level |
|---|---|
| ≤ 2.5 | Risk of Fire |
| > 2.5 | No Risk |

| Telicyn Index | Risk Level |
|---|---|
| ≤ 2 | No Risk |
| 2.1 – 3.5 | Small Risk |
| 3.6 – 5.0 | Medium Risk |
| >5.0 | High Risk |

1. www.wunderground.com

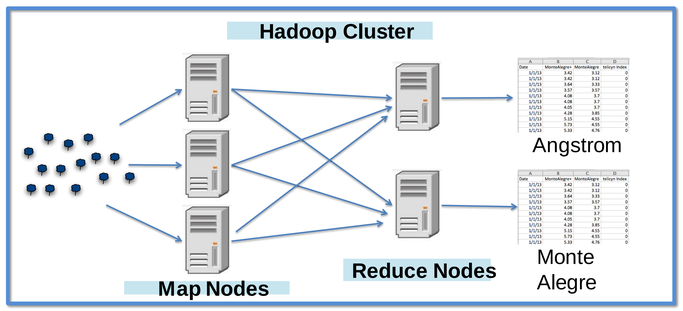
## Scenario 1 (Single processor data analysis)

Data is collected from each weather station.

Analytics are applied to the data to process fire indices.

Output generated in a csv file, where the results can be graphed and analyzed.

## Scenario 2 (Analyzing weather data through a cluster framework)




### Hadoop

Hadoop is a cluster based framework that is designed to efficiently and reliably process very large data within an application workflow. The phases of a Hadoop workflow consist of a map phase and reduce phase. Each map node parses and processes a partition of data through a listing of key, value (line) pairs <K,V>. The map node outputs new key value pairs <K',V'> to a reducer node that receives the transformed data associated with a partition of the new keys. The reducer then processes the <K',V'> pairs and outputs the results to a file.

### Map Phase

Each map node within the Hadoop cluster collects a portion of weather locations, and computes the corresponding fire index. The Map node associates each weather formula as a unique key and outputs this key and value (fire index and corresponding variables) out to the reducer.

### Reduce Phase

Each reduce node will handle ordering all the weather results for a single or set of weather analytics. The reduce node can also do further analysis of the data i.e. compute the standard deviation and variance of the fire index results.

## Future Work

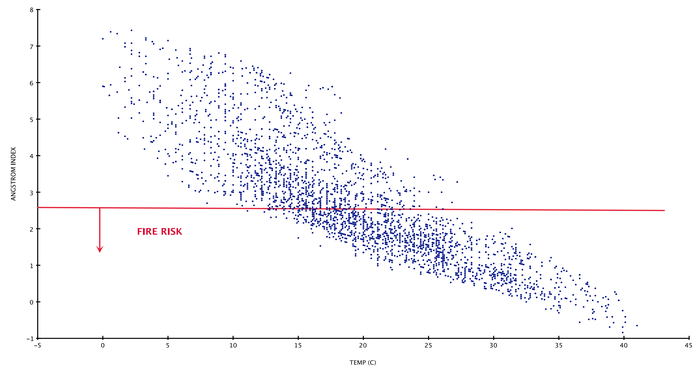
### Real Time Streaming

In my current work I retrieved and stored historical weather data, to be used as a database. Future work, my system will query, on demand, all the weather stations, interpolate within the regions of the queried weather stations, and tabulate the real time fire indices. The system will also compare any deviations with the historical data.

### Mesh Partitioning

To obtain a finer granularity of fire index locations, I propose to apply a mesh of triangles over the region of weather stations and interpolate (using the scan line algorithm) over each triangle region where the triangle end points correlate with the weather station's fire index values. Each map node can be assigned a region or even a sub region within a specific triangular area.

## Results




Angstrom Index vs. Temp (Celsius)

The graph above shows the correlation between temperature and the Angstrom fire index values. As intuitively expected most of the fire risks comes with warmer temperatures with the vast majority of fire risk coming at temperatures greater than 15º Celsius. For fringe temperatures near 15º Celsius humidity plays an important factor in determining a fire risk area.

CSV output files were generated for all of the fire index formulas. For days where there was no measurement information or unrecognizable readings from the weather stations, the formula assumed a conservative value of high precipitation, and were not factored in as data points. However as the above graph shows, there are still some outliers due to incorrect readings or errant output that could be still translated to a numerical value. Other inconsistent factors may be due to the formula not completely mapping to the region.

## Conclusion

It is important to use the appropriate weather formula for the region and to obtain enough readings from different stations for redundancy. Using the Angstrom index along with data collection over the various weather stations provided a reasonable estimate of fire prone areas. With more data and access to more types of measurements better results can be obtained for a region.

As more data is collected and more weather stations are included, a clustering system using this framework will play a critical role in analyzing large scale weather data in real time. Other streaming tools can also be considered for future work.