Global Warming Distributed Data Analysis

Alberto Simioni
2577392
a.s.simioni@student.vu.nl

Federico Ziliotto
2577394
f.z.ziliotto@student.vu.nl

ABSTRACT

1. INTRODUCTION

Global warming is an open topic of discussion today. While climate scientist debate on the causes of the gradual temperature rise in the past 50 years (the majority claim the increase of CO2 due to fossil fuel is to blame, other researchers found alternative explanations, like non CO2 greenhouse gases (GHGs) [10]), data scientist had access to ever increasing technologies and data sources to study the phenomenon. Whatever the causes, the data collected by different organizations and from different parts of the world all agree that the global temperature in the world is rising [9]. We show that this trend is not only real, but the temperature increase is accelerating in the last twenty years at a worrisome rate. We aim to update the results obtained by other institutions with the more recent data, and to do that we want to use more flexible and powerful tools. With the use of distributed computing technologies (like the Hadoop file system and the Spark programming framework) we want to show the capabilities but also the limitations of this new technologies applied in large scale computing. Moreover, we created an interactive tool that may help researchers to visualize the enormous amount of climate data that was collected in the past century and detect possible global warming causes and effects.

1.1 Data Sources

We have access to more than one hundred years of weather data provided by the NOAA (National Oceanic and Atmospheric Administration) organization. The raw data is publicly accessible[14], so anyone interested in analyzing the climate of the past years can use them. The data is organized by year and grouped by ISD (Integrated Surface Data) stations. They also provide a `perl` script that helps in reading and viewing the data. NOAA also publishes monthly reports on the climate status (like temperature, precipitation, sea level) and compares the data obtained for the last month/year to the averages calculated in the past, accompanied by a software suite to repeat or modify the analysis at home. While this tools and datasets are helpful, there are currently no publicly available tools to perform custom data analysis on the whole raw dataset available. In particular, considering the size of the uncompressed data ($\tilde{2}$05 GB) it may be necessary to use a cluster to make complex computations that require more time.

In the following sections we first describe the context of our work and the related literature in section 2, the purpose of the research (3), then we explain the environment we worked with in section 4 and the experiments we performed (5). In section 6 we show the results we obtained and discuss them and finally in section 7 we argue the relevance of our results and future improvements.

2. RELATED WORK

Studies on global and regional surface temperature change has been done in past. Researchers showed that the rate of temperature increase has been higher in the last quarter of the century that it has been in the previous years of the 20th century. Moreover they calculated an increase in the average global five year mean of about 1℃ [8]. This recent trend is particularly important if we consider the temperature changes in the past millennium, and can be considered an unprecedented anomaly[13]. More recent studies also show that the mean annual temperature has not risen in the last twenty years, despite the continuous increase of GHGs in the atmosphere[11]. This suggests that either the global warming phenomenon that was registered in the period 1975-2000 had different causes or that the more recent stable trend has an origin that balance the greenhouse effects. Research done on the temperature changes in the last century in the U.S. shows that the yearly mean trend is slightly different than the worldwide one. Indeed, in the U.S., temperature somewhat decreased from 1930 to 1980, and a warming trend has returned only in the last decades of the century[8].

3. RESEARCH QUESTION

Our aim is to update the recent work done in the climate analysis with the new data available. We will first analyze the structure and characteristics of the data provided by the NOAA organization. The data have a very important size and furthermore the data are continuously growing in size also in the last years. So one important question is **how to efficiently perform global analysis on huge weather data?** We will perform the analysis utilizing distributed

computing and study the advantages and disadvantages of this technology.
Then we are going to perform different analysis on the temperature data and compare the results with the ones found in literature. So it's important to find **which are the metrics that show better a change in the temperature?** In the end, we will make an evaluation on the threat of global warming based on the results we obtained, examining the utility of the selected metrics.

4. PROJECT SETUP

In this section we explain in more detail how we performed the analysis, what tools we used and the preliminary work on the data before the analysis.

4.1 Data Format

The public NOAA data repository contains a file for each station for each year. The name of each file is {USAF code}-{WBAN code}-{year}, where USAF is the Air Force Station ID and the WBAN is the NCDC (National Climate Data Center) station number. The number of station for each year is shown in figure 2. As we can see the number of stations has steadily increased during the years. More importantly, we have to be aware of the results we obtain analyzing years where the number of stations is very low. Since the distribution of the station around the world is not uniform and they don't cover every area, it's difficult to obtain correct results on the global scale. What we can do is to consider the fluctuation of data for the same station or for stations that are near together (this will be done by regional grouping, as described in section 5). Each station reports data multiple times a day (depending on the station) and for each day of the year.

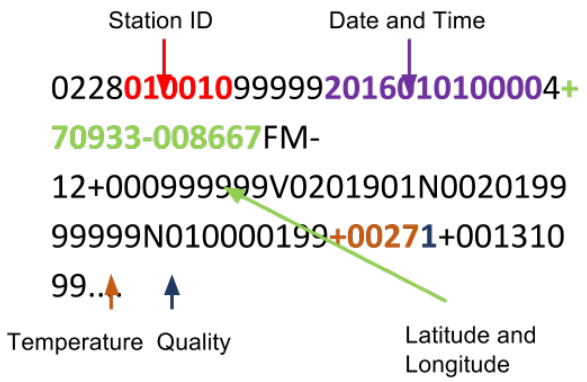

Figure 1: Measurement format

Each measurement comes in a form similar to what 1 shows. The information we need for our analysis are: station ID, data and time, coordinates (in latitude and longitude), the temperature value and the quality of the temperature measurement.

4.2 Temperature data quality

Before using the actual temperature data, we filtered it through a custom quality check. Each measurement come with a quality flag set at a different value depending on how reliable the measurement was[3]. It is important to use this flag to remove wrong data in the dataset, as reported in the document provided by NOAA that describes the problems inside the dataset[2]: "Various, mostly random errors, are present in the overall dataset, and these are being documented and scheduled for correction. Quality control of the data (which has already taken place and reflected in the online data) has corrected well over 99 percent of the errors present in the original data prior to the quality control." In particular we accepted all the measurement with the following quality code value:

- 0: Passed gross limits check;
- 1: Passed all quality control checks;
- 4: Passed gross limits check, data originate from an NCEI data source;
- 5: Passed all quality control checks, data originate from an NCEI data source;
- A: Data value flagged as suspect, but accepted as a good value;
- I: Data value not originally in data, but inserted by validator;
- M: Manual changes made to value based on information provided by NWS or FAA;
- P: Data value not originally flagged as suspect, but replaced by validator;
- R: Data value replaced with value computed by NCEI software;
- U: Data value replaced with edited value.

We opted for a loose policy on the data quality, accepting also values that didn't pass all the quality checks or that were manually or software replaced. We refuted all the codes that stand for erroneous, missing or suspect data that was not manually or automatically checked and replaced by the institution. Furthermore, we applied a bound check for all the values greater than the maximum registered temperature (+61.8 ℃) and the minimum one (-93.2 ℃).
The data collected can be of two different types depending on the origin: from a fixed weather station or from a mobile one. In the second case, it's possible that the data registered from a mobile station (e.g. on a ship) is reported at a certain position and the following measurement is reported at a different position. Since we dont't distinguish data by time of the day, it's important that the number of measurements from a certain position is approximately the same for different hours. Since this is not guaranteed by moving stations, we decided to filter the measurement obtained from this source.

4.3 Data Sources

An important aspect we had to consider before proceeding with the analysis is the distribution of weather station on the globe. First we considered the number of stations per year2. Then we analyzed the position of the stations for each year (some notable sample are reported in appendix A).

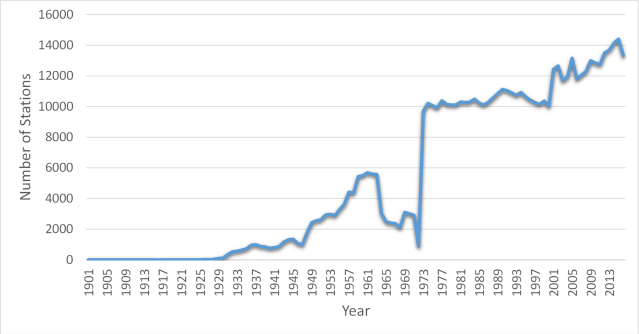

Figure 2: Number of stations per year

The important things to consider before proceeding with the analysis are:

- The stations are not distributed uniformly. In particular, the areas most covered are only a portion of the total surface of the earth and include mostly the land areas (the coverage of the oceans is pretty limited);
- There are no guarantees that a station reporting one year will report also the following years;
- Although the general trend manifest an increase of the number of stations, this is not always true (e.g. during the late 60's there are less stations reporting);
- The years before 1930 show a very low number of stations. We will have take this into account when we evaluate the results for this time frame.

4.4 Data Analysis Environment

The technological environment we decided to use for this project consist of:

- **HDFS**[5]: a distributed file system that allows the developers to store and retrieve large quantity of data between multiple machines as if it was a single one;
- **Spark**[6]: programming framework to develop applications for large scale data processing on horizontally scalable clusters;
- **SURFsara Hadoop Cluster**: the hardware resources that will be used to run the computations.

HDFS The data stored in the distributed file system is grouped in folders by year. As shown in figure 3 the size of the data to analyze is very different between the years. Inside each folder, there is a file for each station that reported measurements. HDFS is not very efficient when working with many small files[17], as in our case. The solution we opted to use is to merge the small files in bigger files that have the same size as Hadoop's default block size. To merge the files we used filecrush[1]. This tool merges small files,

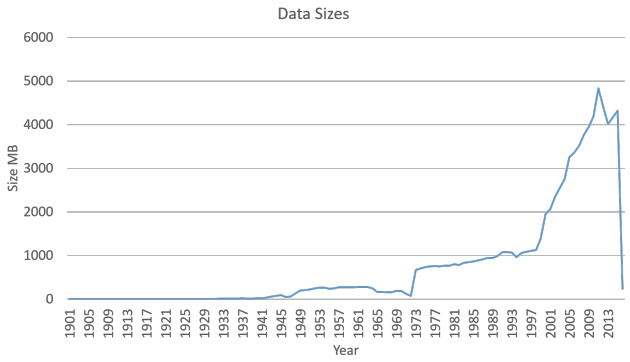

Figure 3: Size of the data per year.

located inside an HDFS instance, into bigger files with the size of a multiplier of an HDFS block. We have run this tool one time for each folder of the data. The number of total files was reduced by an order of magnitude of 10^2.

Spark The choice of this technology is motivated by the size of the data and the type of the analysis we want to execute. First, Spark is more efficient than Hadoop's own MapReduce[4] implementation when the data used can fit in memory. Since our approach is to analyze one or a few years at a time, we calculated that the size of the data is small enough to stay in memory for our purposes.
The second advantage is that we would like to make multiple transformations to the data and Spark allows to easily cache the intermediate results to be reused efficiently. In this way, there is no need to continuously store the results and load them to be reused for the next MapReduce operation. One example is calculating the large region averages, for which we make use of the intermediate results of the worldwide regions (each one of equal size) analysis4. The caching feature is used multiple times throughout the execution and drastically improve the performance.

SURFsara Hadoop cluster All the computation are executed on the SURFsara Hadoop cluster[15]. The cluster is one of the largest Hadoop clusters for scientific research in the Netherlands. It is composed of 170 data/compute nodes with a total 1370 CPU-cores for parallel processing. The system has also a distributed file system that has a capacity of 2.3 PB. The cluster is shared and used by different users that needs an HPC infrastructure for their research.

5. EXPERIMENTS

We performed two main type of analysis: a five year span analysis that considers all the measurement reported by stations near each other in five years, and an analysis on large earth regions on a yearly basis.

The basic steps that both analysis perform are:

1. Load data in memory;
2. Parse the raw data to obtain a list of StationData (a type that represent a measurement);
3. Filter data following the scheme indicated in 4.2;

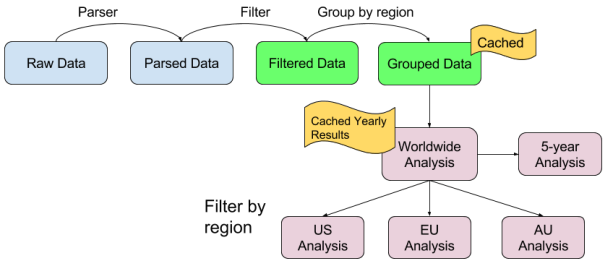

Figure 4: Analysis scheme

4. Group the measurements in regions;
5. Execute yearly analysis;
6. Execute 5-year analysis starting from the results of the yearly analysis;
7. Save results;

5.1 Regional five year period analysis

The aim of this analysis is to calculate the temperature statistics (mean, maximum, minimum and standard deviation) on every area of the earth for which we have data. After that we can compare the situation for the same areas in different time frames and show it with a world map view (as shown in 6.1).
The first problem is that while the regions we want to compare are of the same size and uniformly distributed around the earth, the stations are not. Furthermore a lot of stations are added and removed during the flow of the years, this means that the data of a new station can't be compared with any data of the previous years even if in the previous years there was a station very close to the new added station. So we defined regions of size 1°latitude and 1°longitude. This is approximately the area covered by the region of Holland in the Netherlands. Then we assigned each measurement (for each one there is the coordinates reported) to a region. Our analysis proceeded by loading five consecutive years of measurements at a time, group them in small regions and, for each region, calculate the arithmetic mean, the minimum, the maximum temperature and the standard deviation. This was done for all years from 1901 to 2015 (the year 2016 was excluded from the analysis due to incomplete data).

5.1.1 Median

Another value we wanted to extract from the datasets is the median. We implemented the code to calculate it, but we didn't have enough time to test it and obtain the results for all the datasets. The problem with calculating the median is that it is not a commutative and associative operation (the median of medians is not the median of the entire group). This means that there is no easy way to use the reduce function to combine results computed in different machines in a distributed way. The simple way to calculate the median is to sort the desired data and look at the value in the middle. The problem with this solution is that it's not scalable. Indeed, applied to datasets in the order of GB in sizes this operation is very slow. Work has been done by various researchers to tackle this kind of problems and improve the efficiency of non associative operations ([12]).

5.2 Zonal Analysis

In this type of analysis we focused on some particular large areas of on the earth. We selected the regions that for which we have more and better distributed data available. The selected regions are:

- United States: from coordinate N50° W128° to N24° W65°;
- Europe: from coordinate N71° W11° to N35° E41°;
- Australia: from coordinate S9° E111° to S44° E156°;

For each region we calculated minimum, maximum, average and standard deviation of every year until 2015. The year range is chosen based on the availability of data.

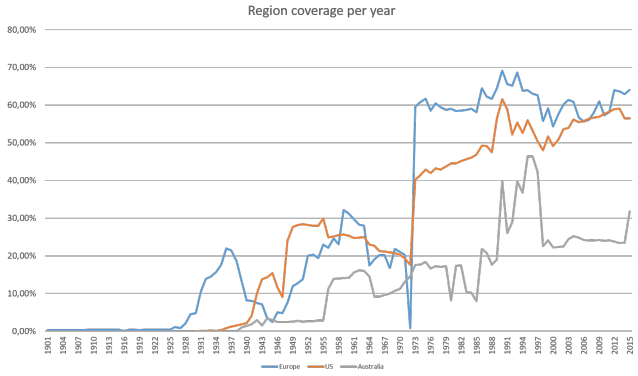

Figure 5: Percentage of regions covered by at least one station per region.

The chart in figure 5 shows the surface coverage of the selected regions for each year from the data we analyzed. As we can see, we don't have data from every part of the regions, with Europe and US the more covered ones (with peaks of 70% and 60%). The are two main reason behind this:

- In the chart we took into account the number of regions for which we have data on the total amount of regions in the area. With a different selection of region size we could increase or decrease the total coverage (at the expense of precision).
- There may be areas in the regions where it's more difficult to have a uniform stations coverage because of geomorphological reasons. For instance, there are far less regions covered in the middle part of Australia due to its desert climate, which makes it more difficult to install, maintain and monitor the stations.

5.3 Results format

The results computed by the program we developed with Spark are stored in HDFS with the CSV format. The results are stored in multiple HDFS folders, one per each year. Inside each folder there are multiple CSV files with portions of the yearly results. To collect the results from HDFS and

compact them we developed a script that merges the files in each folder in two CSV files per each year. One file contains the minimum and the maximum temperature for each latitude and longitude quadrant. The other file contains the value of the average and of the standard deviation of the temperature. The script, in addition to the two CSV files, creates two JSON files with the same information. Those files will simplify the parsing of the data in the graphical user interface that is developed with javascript.
The results of the zonal analysis are stored in one single CSV file per each zone containing one row per each year. With this format it is easy to create some charts about the change of the temperature for that zone during the years.

6. RESULTS

In this section we report the results obtained from the analysis described in section 5.

6.1 Interactive map

The first tool we worked on is a dynamic map that allows to visualize the difference between temperature values (average, maximum, minimum and standard deviation) between two datasets. Each dataset is created from the analysis one described in section 5 and are composed of the aggregated results of the five yearly analysis in which we divided the surface in small regions of 1°latitude and 1°longitude.
We used the map provided by Google[7]. We created a custom heatmap that shows on each region of the map different colors to show an increase or decrease for a particular value in that region (shades of blue for reduction and red for a gain). The interface can be seen in figure 6

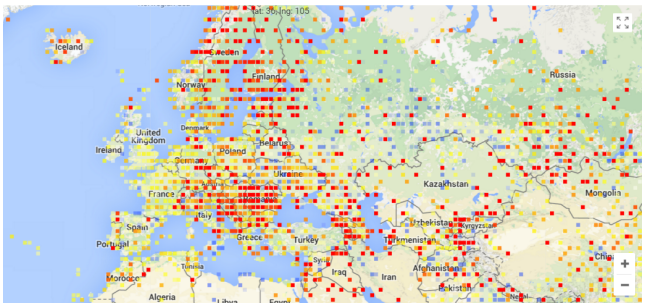

Figure 6: Interactive map interface example

In the graphic interface there is the possibility to select the two set of five-year results to compare. The heatmap shows a colored square in a latitude and longitude quadrant if it's present a data for that quadrant in both the two sets of results. Till the period 1921-1925 there are few data in the results, with consequently few colored squares in the map. In the interface there is also the possibility to change the temperature metric that is used to compare the two set of results.
One problem of the interface is that the colored squares used in the map should change their height depending on the latitude value where they are positioned because the map projection isn't equiareal. The consequence in the heatmap visualization is that there is some empty space between the squares of a certain latitude value and the squares with the value of the latitude increased or decreased by one. This aspect makes the heatmap looks like as it is missing some data for certain latitude when instead the data are present, just the squares should be rectangles with a bigger height value.

By using the interactive map it is possible to notice an increase of the global temperature in the years from 1975 to the current years. In figure 7 it is possible to see an example of increase of the average temperature between the two interval 1976/1980 to 2011/2015 in Europe and in Asia. The red colors are dominant in all the figure. Showing an increase of at least 2.5 degrees for all the regions that have a red square over them. In figure 8 it is possible to see how the minimum temperature is increased for the same two intervals of years in the majority of regions of the world.

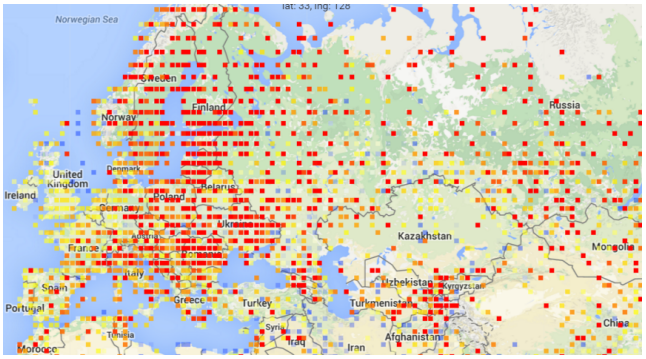

Figure 7: Map that shows the increase of the average temperature in Europe and in Asia between 1976 and 2015.

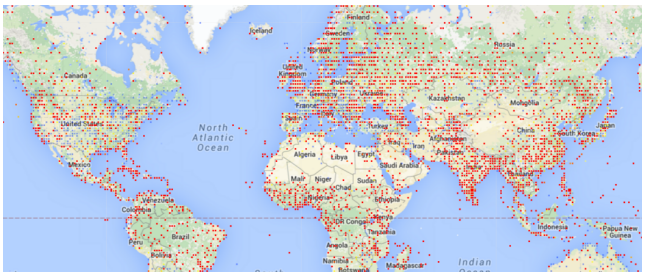

Figure 8: Map that shows the increase of the minimum temperature globally between 1976 and 2015.

6.2 Zonal Analysis

In this section we report the results of the regional analysis done for: Europe, United States and Australia.

In figure 9 we show the evolution of both the average and the 5-year average for the three region we considered. The first thing to consider is that in the earlier years we have less data available. For instance in Europe, we have data mostly from stations in Finland for the period 1901-1930. This explain two anomalies in the chart:

- A sudden change in the average temperature between two time frames (the sudden increase in the average temperature for Europe is caused by the presence of new stations in warmer areas);

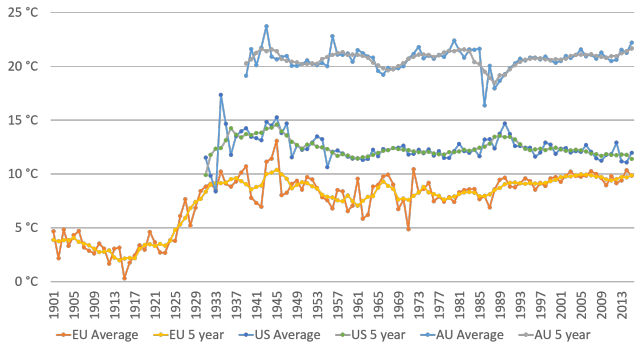

Figure 9: Average and 5-year average for three different regions.

- The variance of annual temperatures is more evident in the years with less data and stabilizes for all three regions from the last decades of the century.

6.2.1 *Temperature Anomalies*

In figure 10, 11 and 12 we used the same method described by Hansen ([8]) to calculate the temperature anomalies for each dataset. We first calculated a total average value for each region for the time frame examined. Then we calculated the difference between the yearly average values and the total obtained. In this way it's easier to spot extremely hot or cold years and a general temperature trend.

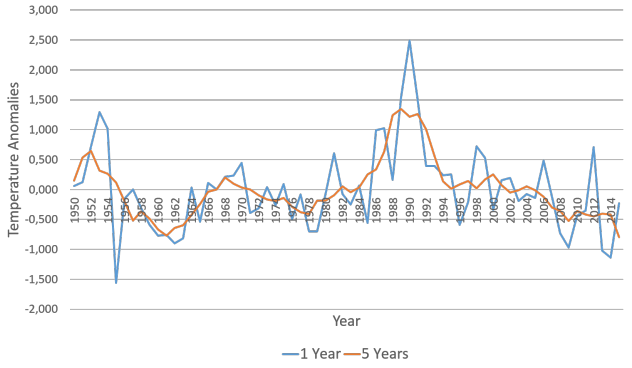

Figure 10: Avarage temperature anomalies in the U.S.

From the US chart (10) the interesting thing to notice is a slightly warming trend between the 60s and 70s and an opposite one for the last thirty years. This goes in accordance with the observation made by other researchers[9]. The main suspect for this phenomenon is thought to be the hurricane El Nino that hit the US in 1998.

The EU chart shows, starting from the late 70s, a pretty clear trend of temperature increase. Interesting also to note the high average temperatures in the period 1940-1945, during WW2. We leave to the reader to use the interactive map we created to correlate the data sources distribution and the temperature levels.

For the Australia data the most interesting fact we notice is a sudden plunge in the temperature during the 80s. This can be explained by looking at the stations number chart for that region, that manifest a decrease for the same period of

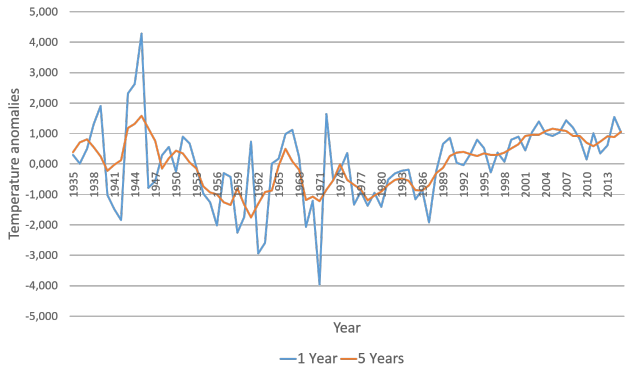

Figure 11: Avarage temperature anomalies in Europe.

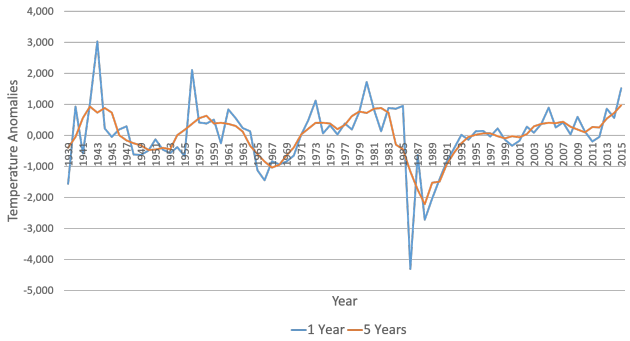

Figure 12: Avarage temperature anomalies in Australia.

time. As for the European situation, we can see a slight increase in the average temperature, that is confirmed by the results of other studies, that shows an increase of the average temperature for the Australian continent of about 0.8°.

6.2.2 *Standard Deviation analysis*

In figure 13 we plotted the value of the standard deviation per statistics and per region. It's clear that the values of maximum and minimum have larger fluctuation than the average. This may be caused by one or a few erroneous data report caused by environmental accidents. Interesting to note that the minimum value have a way larger standard deviation that the maximum. By correlating this with the stations distribution we came to the conclusion that this is caused by the installation of new stations in more remote and difficult to access regions (e.g. Arctic ocean and mountain regions).
Moreover, the minimum standard deviation for the Australian region was suspect, because the lowest temperature registered for the region is -23.4°C[16], so it's difficult to expect a standard deviation similar to that of the European region or even higher than that of the American one. After analyzing the data we found out erroneous data coming from the south east region of Australia reporting temperature of around -70°C.

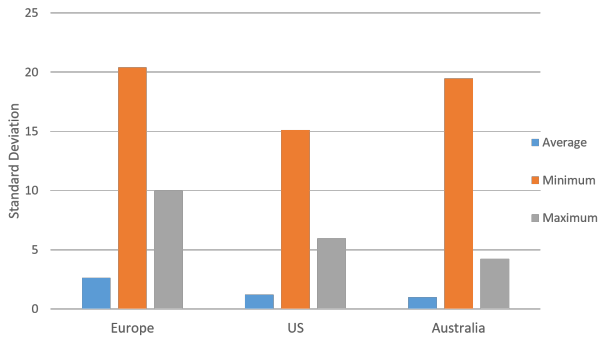

Figure 13: Standard deviation for Average, Minimum and Maximum values per region.

7. CONCLUSIONS

Before answering the question we asked earlier we will take a brief general overview of the characteristics of this project. One of the main problem we encountered is the limited (if not absence) of documentation describing the data. Especially for the older datasets we had to directly analyze the data to know the origin (region on earth) or the time coverage (day/night reports, coverage of all months). Moreover, we had to take into account during the analysis and the result discussion the anomalies we discovered in the station number charts (A). Without a uniform and well formed source it's more difficult to extract reliable statistics.
We are overall satisfied with the performance of the program developed with Spark for analyzing the data. The program is able to analyze the entire dataset in approximately 8 hours. The program is also scalable to the data size. Future data in the next years will probably grow in size, Spark can automatically handle new resources available in the cluster if there is the need. Furthermore the structure of the program allows to manage the growth of the data by computing some intermediate results on the yearly data. The five year analysis is performed only on these intermediate results that are way smaller than the initial data. This feature removes the need to join all the data of the five years. Other two aspect that improve the performance are the step we performed to group the big amount of files, and the caching of the data after the operations of parsing, filtering and grouping. The first improvement makes HDFS to be faster because it performs much better with a smaller amount of bigger files. The second improvement allows to compute different analysis on the same data without redoing some computation that are needed in the different types of analysis.

For what concerns the metrics used for analyzing the data, the average temperature is the most useful metric. It shows clearly the increase of the temperature in different periods. Also, it's stable even if some particular events happen in short periods of the year. Minimum and maximum temperature are more susceptible to those events and indeed they show less clearly the increase of the temperature in the periods where the average metric reveals the increase. The standard deviation metric is useful to understand if there are errors in the data or if the data are not uniformly distributed (e.g. when the distributions of the stations is changed during the five year of the grouped results then the standard deviation results higher).

7.1 Global Warming remarks

From the worldwide analysis and using the interactive map, it appears that a clear temperature increase is in place around the world. Moreover, we notice this phenomenon occurring even considering specific large regions of earth where two out of three show a clear temperature increase trend for the last 30 years, which is in accordance with the data reported by independent researches[9].
What we miss in the picture is the data coming from most of the oceanic areas and in general ocean temperatures. The only sources we had that covered large part of the oceans were coming from mobile stations installed on ships. The problem with this data was dual: many times we had the results (average, minimum and maximum) for a region constituted by only one station report. In the end we excluded these from the analysis because they skewed the results in an unpredictable way.

7.2 Future Improvements

Due to the limited amount of time available, there are many improvements to the analysis and other research path to discover for a better understanding of the dataset and the global warming phenomenon.
To improve the data quality we could divide the temperature data by time of day. We assumed that all the station have a similar behavior (they collect temperature samples every fixed amount of time). Without this assumption the average temperature result could lean towards lower or higher temperatures depending on which the hours for which we have more data.
Another interesting analysis that can be performed is dividing the regions by latitude bands. We expect that regions in the same latitude have lower differences in temperature so we could expect a more stable average temperature and less fluctuations in minimum/maximum values. The same reasoning can be done for seasons. In particular it would be particularly interesting to see the temperature changes by season for a fixed region and compare the season analysis (e.g. see if all seasons show the same curve or not).
An intriguing data source we had available but didn't use yet is the Arctic ice surface coverage. Since this is directly correlated with global temperature and is also one of the most monitored metric in recent years (due to its effects on the ecosystem), we think it can help in strengthening our conclusions. We didn't investigate further this path because our understanding of the data sources was limited and required more time than we expected to interpret correctly the data.

APPENDIX

A. STATIONS DISTRIBUTION

Here is an overview of the stations for some selected years. We decided to show only years in which the number and position of the stations is significantly different than in the previous or following years.

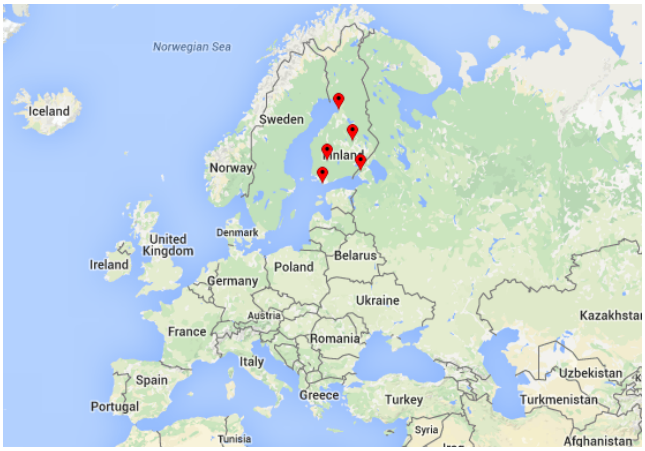

Figure 14: Weather stations used in 1907. In the first years we only have measurements coming from stations in Finland

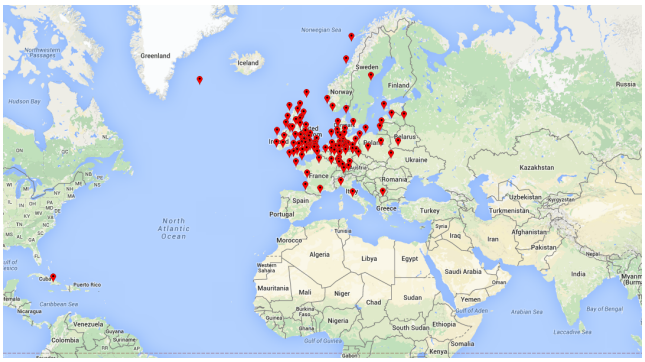

Figure 15: Weather stations used in 1930. Here we can see the stations cover the most part of central Europe. The problem we encountered is that there is no continuity in the stations that are used between consecutive years, so it's difficult to make comparisons.

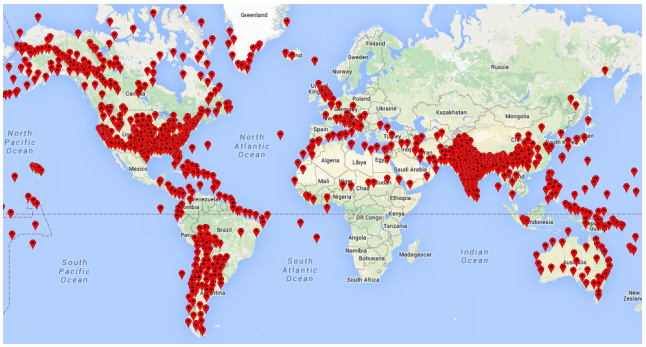

Figure 16: Weather stations used in 1945. Around this year we start to have a relevant number of stations that cover most part of the surface.

Figure 17: Weather stations used in 1968. In the late 60s we notice a sudden drop in the earth coverage compared to the previous years. We took this into account when making the analysis.

Figure 18: Weather stations used in 1972. As shown in figure 2, we noticed a sudden drop in the number of stations for this year. In particular we observe the absence of stations from east Europe and the ex Soviet Union Area.

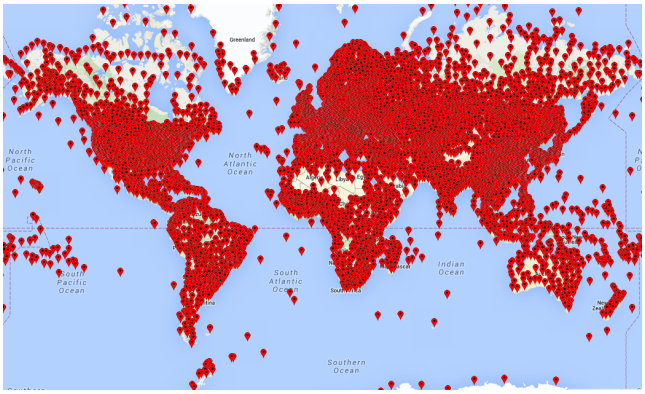

Figure 19: Weather stations used in 1980. From this year onwards we have a large number of stations covering most part of the surface.

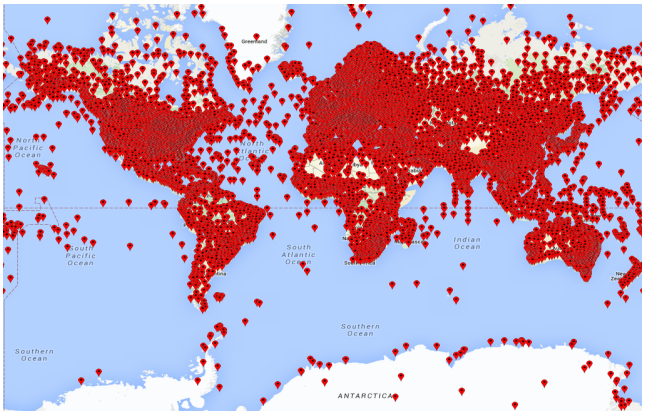

Figure 20: Weather stations used in 2014.

References

[1] Edward Capriolo. *Filecrush*. [Online; accessed 28-March-2016]. 2014. URL: https://github.com/edwardcapriolo/filecrush.

[2] National Climatic Data Center. "Brief Review of Problems and Issues with Integrated Surface Data (ISD)". In: (2015). Available at ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-problems.pdf.

[3] National Climatic Data Center. "Federal Climate Complex Data Documentation for Integrated Surface Data". In: (2015).

[4] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters". In: *Communications of the ACM* 51.1 (2008), pp. 107–113.

[5] The Apache Software Foundation. *Apache Hadoop*. [Online; accessed 28-March-2016]. 2016. URL: http://hadoop.apache.org/.

[6] The Apache Software Foundation. *Apache Spark*. [Online; accessed 28-March-2016]. 2016. URL: http://spark.apache.org/.

[7] Google. *Google Maps APIs*. [Online; accessed 28-March-2016]. 2016. URL: https://developers.google.com/maps/.

[8] J Hansen et al. "GISS analysis of surface temperature change". In: *Journal of Geophysical Research: Atmospheres* 104.D24 (1999), pp. 30997–31022.

[9] James Hansen et al. "Global temperature change". In: *Proceedings of the National Academy of Sciences* 103.39 (2006), pp. 14288–14293.

[10] James Hansen et al. "Global warming in the twenty-first century: An alternative scenario". In: *Proceedings of the National Academy of Sciences* 97.18 (2000), pp. 9875–9880.

[11] Yu Kosaka and Shang-Ping Xie. "Recent global-warming hiatus tied to equatorial Pacific surface cooling". In: *Nature* 501.7467 (2013), pp. 403–407.

[12] Huan Liu and Dan Orban. "Computing median values in a cloud environment using GridBatch and MapReduce". In: *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on*. IEEE. 2009, pp. 1–10.

[13] Michael E Mann, Raymond S Bradley, and Malcolm K Hughes. "Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations". In: *Geophysical research letters* 26.6 (1999), pp. 759–762.

[14] NOAA. *NOAA public data*. [Online; accessed 28-March-2016]. 2016. URL: ftp://ftp.ncdc.noaa.gov/pub/data/noaa.

[15] SURFsara. *SURFsara Hadoop*. [Online; accessed 28-March-2016]. 2016. URL: https://userinfo.surfsara.nl/systems/hadoop.

[16] Arizona state University. *WMO Region 5 (Southwest Pacific, Australia only): Lowest Temperature*. [Online; accessed 28-March-2016]. 2016. URL: http://wmo.asu.edu/australia-lowest-temperature.

[17] Yang Zhang and Dan Liu. "Improving the efficiency of storing for small files in hdfs". In: *Computer Science & Service System (CSSS), 2012 International Conference on*. IEEE. 2012, pp. 2239–2242.