Cyclone Database
5 December 2018
Edward Cheng, Baldwin Mugri, Punit Khiyara, Michael Wilkerson

Introduction

Our database's data revolves around detailed, statistical information on cyclones and their occurrences. We will be exploring Atlantic and Pacific cyclones and as such, our database includes multiple tables on data such as cyclone names, date/time of the storm, maximum wind speed and minimum pressure, locations specified by longitude and latitude, landfall occurrence, casualties, estimated damage costs, and wind type. In doing so, our database will provide a detailed and organized compilation of relevant information in regard to cyclones over the course of the year 2014 for easy access.

Moreover, our database will range in scope to include all recorded cyclones originating from the Pacific and Atlantic Oceans over the year of 2014. Additionally, information included in our database would classify each cyclone based on its types such as tropical depression, tropical storm, hurricane, extratropical, subtropical depression, subtropical storm or a low-intensity cyclone.

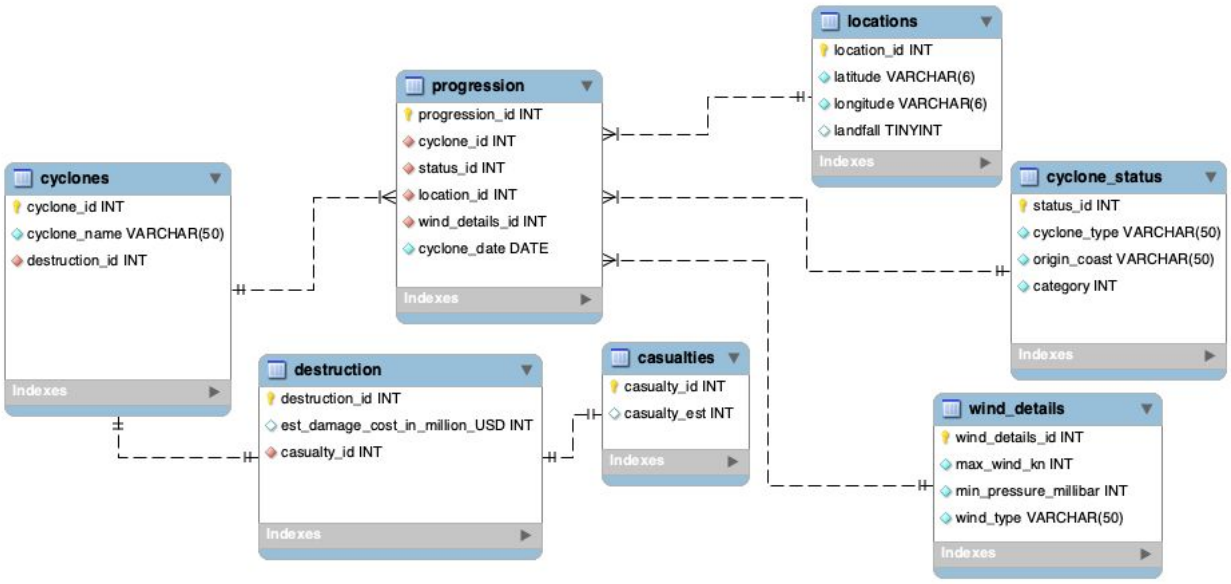
In order to provide more concise and detailed information, we decided to limit the scope of our data to only record cyclones that occurred in 2014 so we could focus more on modern-day trends. Through our database, we hope to be able to track the severity of any given cyclone based on its location, estimate damage costs based on similar, past cyclones, and use our data to predict future cyclones based on our hypothesized pattern findings.

Since the data gathered could be processed into information that would give valuable insight into the behavior and occurrences of cyclones, the database would be able to be utilized by various people. However, for the purposes of the project, our primary target audience consists of university researchers interested in predicting future cyclone occurrences and using such information to estimate damage costs and casualties. Nevertheless, since we hope to make our database available to all through GitHub, we believe that government entities, cyclone enthusiasts and other research groups would be members of our secondary target audiences. In a general perspective, this database will also provide valuable information to the general public for those able to access and use it to perform their own analysis and test hypotheses. More

specifically, people who live in cyclone-prone areas can utilize this database to become more aware of their environment. By creating this database, we hope to discover trends in cyclone occurrences that will help governments and communities better prepare for and deal with cyclones. Additionally, we hope that our target audiences would be able to utilize our database for future prediction models and implementing safer evacuation strategies as well.

Database Description

Logical Design



The entity relationship diagram (ERD) provided above perfectly sums up the variety of relationships we chose to form to create our database. As each cyclone would have unique damage and casualty statistics and chances of reusable data would be slim in this context, it made sense that tables such as cyclones, destruction, and casualties would have a one-to-one relationship with each other. Meanwhile, tables such as locations, cyclone_status, and wind_detailes would have greater variability within their data records since these aspects change with progression and thus, a one-to-many relationship with the progression table was more appropriate.

Physical Database

Our physical database follows the template laid out by our ERD above and proved to be effective in collecting and importing data from cleaned datasets resulting from those acquired from our sample data plan. As such, our physical database follows all criteria specifications ensuring it to be robust in its implementation.

Sample Data

Our plan regarding obtaining data for our database is to download and utilize already available data from the National Oceanic and Atmospheric Administration's Hurricanes and Typhoons, 1851-2004 dataset on kaggle.com for most of the information required. (*NOAA*) However, since this dataset includes data on cyclones well outside our predetermined range of analysis, we selectively filtered the data to only import data on cyclones that occurred during 2014. We also modified the data by increasing the intervals at which the cyclones were tracked from 6-hour intervals to 24-hour intervals and changed the formatting of some data such as dates in order to import them into our database in MySQL. Additionally, we accounted for columns like maximum wind speed and minimum pressure when changing the intervals. Since this dataset did not include information on casualty statistics and estimated damages, such information was manually acquired from the National Oceanic and Atmospheric Administration - National Centers for Environmental Information's Storm Events database and the National Hurricane Center's reports. (*NOAA*) As a change from our initial sample data plan, we chose to exclude statistics such as the number of people injured or missing from our database since they were not as readily available.

Views / Queries

Our "Summer_Cyclones" view will show the cyclones that had landfall in the summer of 2014. This will serve useful to users who want to know the exact location and name of cyclones that had landfall in the summer of 2014 while answering the question of 'How often do cyclones come in contact with land?'

The “includedCasualties” view will give us an insight on how many storms proved fatal by showing those that included at least one death while answering the question of ‘Where the deadliest cyclones occur?’

The “HIGH_WIND” view will display the name of every cyclone that had a “high” wind level. This will serve as useful to users who want to know which cyclones had higher levels of wind speed.

The “AVG_MIN_PRESSURE” view will show the average minimum pressure of each cyclone that we have recorded. This view would use our progression table that tracks every cyclone in intervals of 24 hours. It will serve useful to users who are preparing for a cyclone because it will give an idea of what kind of pressures that cyclones will introduce.

The “criticalAtlanticHurricane” view will display the Atlantic cyclones that were classified as Hurricanes that resulted in above average damage costs and casualties when compared to all other cyclones across both origin coasts recorded. This is an example of a specific query that may be created by researchers using our database and answers the question of ‘Do more estimated damage costs correlate to more fatalities?’

View Name	**Criteria A** At least four views should involve two or more tables, and thus involve JOIN clauses.	**Criteria B** At least three views should involve some form of filtering (WHERE, HAVING, etc.)	**Criteria C** At least two views should involve some form of aggregation over records (SUM, COUNT, AVERAGE, GROUP BY, etc.)	**Criteria D** At least one view should involve a join (linking) table and both of its source tables.	**Criteria E** At least one view should use a subquery.
Summer_cyclones	X	X		X	
includedCasualties	X	X		X	
HIGH_WIND	X	X	X	X	
AVG_MIN_PRESSURE	X		X	X	
criticalAtlanticHurricanes	X	X	X	X	X

Changes from Original Design

One of the biggest changes that we decided to make was further limiting the scope of our data from 2004-2014 to solely the year 2014. The reason behind this change was due to the sheer volume of data collected from our sample data plan. Since this was our first attempt at creating a database and there was more data collected than we initially expected for each year, we decided that it would be best for our project's aim to limit the amount of data that we would be importing to a single year. While the database would have fewer records, only having one year's worth data proved sufficient in fulfilling the goals of our database and was hence, the ideal option for us. Additionally, the estimated damage and casualties information we included in our database had to be manually researched and inputted by hand. If we were to include too many entries, this process would take much longer than expected.

Another change that was made was the exclusion of our low wind, moderate wind, and high wind tables. The reason that we decided to cut these tables from our database was because the data was too specialized for the purposes of our database. Most of the potential users of our database will not be able to use low wind, moderate wind, and high wind levels, so this data would essentially end up being filler data and provide little to no purpose or significance. In order to make these tables more accessible to our targeted user, we added a new column in our wind detail table to replace the exclusion of the low, moderate and high wind columns. This column is called "wind_type" and it describes the intensity of the wind speeds of each cyclone. We made 3 different categories of wind type (low, moderate, high) that were each split by different ranges with their own cutoff points. Hence, we are able to give users a general idea of how strong the winds of specific cyclones are without overwhelming them with specialized data.

The last change that we made in our proposal consisted of splitting the information that was in our "Destruction" table. Previously, we had the casualties/people missing and estimated damage costs in the same table. We decided to move these two into different tables because they are not related enough to warrant being in the same table and exist more practically as separate entities within our database that are structurally related. With this change, our data was better organized and more concise. After further inspection, we decided that the more specific information on casualties such as people missing and injured would be too difficult to gather for

a large number of cyclones, not to mention the inaccuracies of reported numbers that we would find since it is difficult to estimate these numbers. As such, we decided that a general casualty estimation would be the smoothest implementation of this type of data and allow for easier comparison between cyclones.

Lessons Learned

One lesson that we learned from creating this database is the sheer amount of work that goes into creating a database. We have a relatively basic database with a limited amount of entries and it already required a large amount of work to organize and set up. With more tables, complex information, and entries, the process only gets more tedious and requires more work. Overall, the process of making a small-scale database was a great experience to have as it exposed us to the variety of possible real-world applications of recording and managing data in MySQL in addition to the tasks and responsibilities of various roles such as database analysts, data collectors and database administrators.

Another lesson we understood while importing data from our sample data plan's datasets was the great importance of cleaning data since prior to having cleaned the datasets to match our parameters, the sample data was virtually useless for our needs. As such, we learned that prior to creating a database, it is crucial to plan how exactly each dataset should appear and which columns are essential to include in order to ensure the importing process occurs smoothly.

Finally, we learned how to work practically in a team setting when it comes to creating a database. This project was unlike most we have done in the past since we could not work on the physical database through applications like Google Docs. Therefore, we found it easiest to meet up and work on the data in person rather than communicating online. This is a lesson that we would learn from and retain well into the future when working on similar projects.

Potential Future Work

Due to time constraints and an extensive amount of data, our group changed the scope of the project from 2004-2014 to solely 2014. Given additional time and resources, potential plans for the future would include expanding upon our current database.

In the future, we could expand this dataset in two different ways. We could either make this a very specialized 2014 cyclone dataset or include more years into the dataset to make it better at finding trends in cyclones. If we were to make the dataset specialized, we would decrease the intervals between the cyclone tracking from 24-hour intervals to 6-hour intervals. This would make the data more precise and provide more accurate information when looking at calculations such as averages for wind speed and pressure. In addition, we could add more columns to our tables such as people missing and people injured. We originally omitted these values because the data was not available on hand. With more time, this information could be included. If we were to include more years to the dataset, we would import more data from the National Oceanic and Atmospheric Administration's Hurricanes and Typhoons, 1851-2014 dataset since our dataset already includes the year in the date column and as such, it would not prove too difficult to implement.

Works Cited

Ncei, NOAA. "Storm Events Database." *National Climatic Data Center*, 2014, www.ncdc.noaa.gov/stormevents/.

NOAA. "Hurricanes and Typhoons, 1851-2014." *RSNA Pneumonia Detection Challenge | Kaggle*, 20 Jan. 2017, www.kaggle.com/noaa/hurricane-database.