CS Capstone Proposal: Creation and Extension of a Data Exploration Pipeline

Zac Laney

Cade Fowler

CS 495: Capstone I

Dr. Barlowe

February 1, 2023

Introduction

As the world grows more reliant on computers, the ability to analyze data on a large scale has never been more important. Data science is the blending of statistics and computer science to make sense of and draw conclusions from large datasets (Blei). This field is so important that IBM has predicted that in two years there will be a 28% increase in the number of employed data scientists (Qiang).

Machine learning uses artificial intelligence (AI) to manipulate the data from the data set in an attempt to act in a way that closely resembles how a human would interact and change the data. It does so by recognizing patterns in past data and using that to draw conclusions or make predictions. There are many examples of machine learning such as image recognition or chatbots. A few machine learning algorithms that are commonly used are clustering and linear regression. Clustering is used to identify patterns in data that can be grouped, and linear regression is an algorithm that is used to predict numeric values based on linear relationships between other values in the data set.

The overall goal of our project is to create and establish a pipeline with a database to an interactive visualization to analyze the data set being used.

Why is it important?

The dataset that we have chosen to analyze is weather data gathered by the National Oceanic and Atmospheric Administration. This set includes data from all over the country and has detailed recordings about the particular atmospheric conditions of that specific day and time over the course of several years. We will be narrowing down this dataset just to include data gathered in Asheville, NC. Once we complete our project, it could be used to show atmospheric trends, possibly predict future weather, or show common occurrences based on the previous data. Our system could be used by anyone who wants to visualize weather history or view trends in the weather.

Another reason for choosing Asheville was some would consider Ashevlle to become the world headquarters for climate change and could be a good starting point to confront climate change according to one article from wncMagazine.com. In the article, John Firth who is the CEO of the UK-based Acclimatise stated "I don't want to sound ridiculous, but I believe in years to come we'll be talking about Asheville as a global solutions lab for people who are trying to understand what a changing climate means". Also, the data we are using from the National Oceanic and Atmospheric Administration is also based in Asheville.

Why is it difficult?

This is a difficult problem to overcome because the weather is notoriously unpredictable. There are many variables that go into weather and they each have an impact on the overall climate. This means that our machine learning models won't be able to predict weather specifically but rather will focus on the trends of the data and make large-scale predictions.

The dataset needs to be cleaned before we can use it. This involves getting rid of whitespace, null values, or anything else out of the ordinary that might be in the dataset. These values would cause our results to be skewed more than expected for our analysis and visualizations, so they need to be removed.

Another challenge that arises from this dataset is its scale. There are many different types of data recorded each day, such as daily high, low, and average temperatures, amount of precipitation, wind speed, atmospheric pressure, and more. Each of these types of data is recorded daily, so there is a lot of data for us to process and analyze. Apache Spark will do a lot of the heavy lifting for us, but it is still very complicated. There are so many different types of data in our dataset that we will have to come up with the best way to visualize each type of data, and also visualize the data as a whole.

How to solve the problem

To achieve our goal, we will mainly be using Apache Spark. This is a platform that allows us to analyze large amounts of data and use machine learning to predict future data. Apache Spark handles all of the complicated database management and data processing for us so that we can focus on higher-level analysis (De Souza Neto). This will allow us to do a much deeper analysis in our limited timeframe than we would be able to without these systems in place. Apache Spark manages the database for us, so we don't have to do the tedious and complicated job of setting it up ourselves. Apache Spark also has built-in machine learning capabilities, and we intend to use this functionality in our promakeThis is a great tool to use because we do not have the time or knowledge to build these systems from scratch. We will be using Apache Zeppelin for data visualization so that we can display the data in an easy-to-understand way. This has the same benefits as Apache Spark in that we don't have to manually create visualization tools but can instead use the prebuilt ones in Zeppelin. These two tools are the foundation of our project.

User requirements

- Apache Spark (Python version)
- csv file from weather website
- Apache Zeppelin (Apache Spark Integration)

Alternate Solutions

There are some alternate solutions that we could have taken to achieve our goal. We could use some other data analysis framework such as Apache Hadoop, but Apache Spark seems to be the most powerful and commonly used platform. We could have also picked another dataset to be the subject of our project, but this dataset seemed interesting to us for many reasons. Most importantly, this dataset has real-world relevance and our finished product will have a purpose outside of just this project.

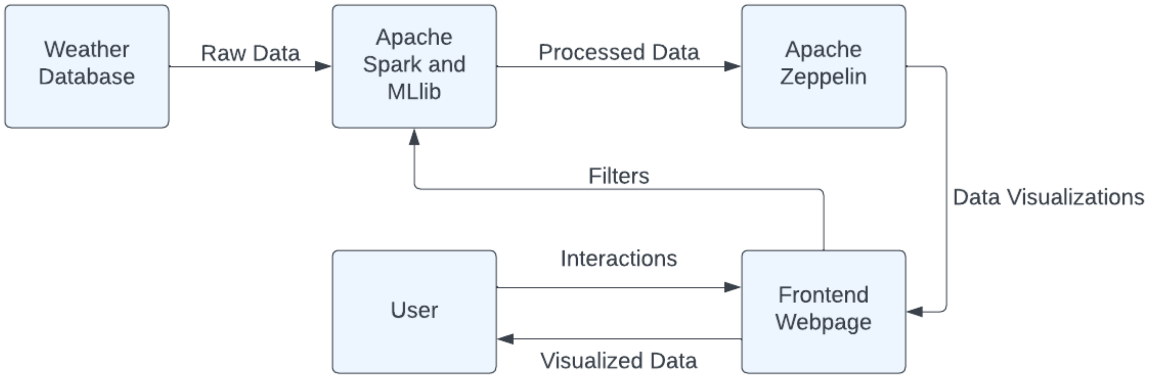
Alternative to Spark

An alternative would be Apache Hadoop as mentioned earlier. Hadoop being one of the more popular solutions to analyzing data is an open-source framework written in Java that takes data clusters and divides the tasks into smaller parts and assigns them to different computers. There are a couple of reasons why we chose to work with Spark over Hadoop, the first being that Spark has better performance because it uses RAM instead of using disks for reading and writing data while Hadoop stores the data on multiple sources, and the processing is done in batches. A second reason is that Spark is superior when applying Machine Learning because it has MLlib to perform iterative in-memory Machine Learning computations. Also with its tools to perform classifications, regression, pipeline construction, persistence, and evaluations.

Alternative to Zeppelin

The biggest alternative to Apache Zeppelin is Jupyter Notebook. Both of these products are open source tools for data visualization and exploration. While Jupyter Notebook is far more popular and has been around much longer than Apache Zeppelin, there are still some key differences that led us to choose Zeppelin. Since Apache developed both Spark and Zeppelin, these technologies are easily integrated. Jupyter Notebook was developed by a different company, so it would be slightly more difficult to integrate. Zeppelin also has an API to integrate with Angular, which is a popular front-end framework. This will greatly simplify the development of a front-end web page and allow us to easily integrate our Zeppelin visualizations.

Data Flow Diagram



MoSCoW Analysis Diagram

Must Have	Should Have	Could Have	Won’t Have
Asheville individual month-to-month weather data set combined into one database	Machine learning to analyze the data and create predictions for the future	Add more areas outside of Asheville to the overall database	Different weather databases from different websites going into one overall database
A way to visually interact with the data	Multiple visualization options for different types of data	Use more detailed data (hourly instead of daily)	Weather data for everywhere in North Carolina or the US
Deploy the application through IT	Options to filter the data that is visible	Make the application available through containers	

Testing

We will first develop and test our program on our local machine. Once we get further into development and have a working product, we would like to get other people to test our program. The best group of people to be these test users would be anyone without a computer science background. We want our program to be accessible to anyone and everyone, so the best people to test it would be those without any preconceived expectations of how it should function. We could find these users by asking students on campus that aren’t in the computer science department. Also if at all possible, the people at the NOAA would be useful to look over the models that our program would produce and see if there is anything that would be beneficial or necessary that should be added.

Proposed Timeline

Feb 10	Learn how to use Apache Spark and revise our proposal
Feb 24	Continue practicing with Spark and cleaning up the data set for use
March 17	Prepare a project poster rough draft while continuing work on the database
March 31	Have project poster final draft and prepare for presentation
April 14	Start creating visualizations through Zeppelin and continue building Spark functionality
April 28	Work on having something to show for final presentation
May 12	Final exams, present final presentation

What relevant classes have we taken and what do we still need to learn?

Class	Project Relation
CS 150	Should help with the use of Apache Spark Python version because learning Python in this class. Also would need to know more about the syntax for Apache Spark.
CS 253	The Software Development class will be essential since this class taught us how to design a large-scale program as well as how to use GIT.
CS 263	Software Engineering might be the most important class for a capstone project. This class taught us how to work in a team to complete a large-scale development project, which will be crucial to the success of our project.
CS 453	The Database class will be helpful by applying some of the things learned in that class to the manipulation of our data set among other relevant information to set up a database.

These classes all have helped us learn the information we need to complete this project, but there is still quite a bit to learn. Neither of us has used Apache Spark or Apache Zeppelin, so quickly learning these platforms on our own will be key to the success of this project. With some time, we will develop from pure beginners in these technologies to capable developers. Apache Spark will be the first and most important thing we need to learn on our own since our entire project relies on it to function. The computer science classes we have taken have made us accustomed to learning new technologies, but this will be the first time we need to learn a new technology on our own.

Works Cited

Blei, David M, and Padhraic Smyth. "Science and data science." *Proceedings of the National Academy of Sciences of the United States of America* vol. 114,33 (2017): 8689-8692. doi:10.1073/pnas.1702076114

"Data Search for U.S. Local Climatological Data (LCD)." *National Centers for Environmental Information (NCEI)*, https://www.ncei.noaa.gov/access/search/data-search/local-climatological-data?bbox=35.861%2C-83.206%2C34.910%2C-81.682&startDate=2022-01-03T00%3A00%3A00&endDate=2022-01-03T23%3A59%3A59.

De Souza Neto, João Batista, et al. "Transmut‐Spark: Transformation Mutation for Apache Spark." *Software Testing, Verification and Reliability*, vol. 32, no. 8, 2022, https://doi.org/10.1002/stvr.1809.

Igelman, Jack. "A Catalyst for Change." *WNC Magazine*, 15 July 2019, https://wncmagazine.com/feature/a_catalyst_for_change.

"National Oceanic and Atmospheric Administration." *Homepage | National Oceanic and Atmospheric Administration*, https://www.noaa.gov/.

Qiang, Zhenping, et al. "Research on the Course System of Data Science and Engineering Major." *2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI)*, 2019, https://doi.org/10.1109/csei47661.2019.8938944.

Works Cited

“What Is Machine Learning?” *IBM*, https://www.ibm.com/topics/machine-learning.

“What Is Spark - a Comparison between Spark vs. Hadoop.” *Intellipaat Blog*, 20 Dec. 2022, https://intellipaat.com/blog/what-is-apache-spark/#:~:text=Speed%3A%20Apache%20Spark%20helps%20run,data%20is%20stored%20in%20memory.