Big Data and Analytics - Weatherlink
Personal Report
Matt Winchester
Milestone: System Design

Our system design tasks were split up among the team. I created the action diagrams, showing the flow of choices our system makes when handling ingesting data, analyzing it, and visualizing it. I had never made an action diagram before, but the layout made sense and it helped me clarify in my head how exactly our pipeline would aggregate information, produce “predictions” for the future, and then visualize it.

I am excited about the visualization portion. I have used Tableau before but never have been involved so much in the process leading up to visualization, so preparing the data in a way that will be useful for an analyst to visualize and study is very interesting to me. I want it to be efficient and give the analyst flexibility in what they want to see. Balancing that is tricky, because there is a LOT of data, and if we keep too much of it it could be expensive both in terms of data storage and in terms of processing power needed as the visualizations change.

Our project is going to be complicated because the data is in a wide array of very complicated formats. The government provided Census and Weather data is split up among hundreds of files spanning years, regions, weather stations… it is quite a mess! The weather data especially seems to be very inconsistent, the format seems to change between years, so we will need to write some custom python for each kind of file to get the level of granularity that we want.

We decided to start small and iterate out adding new kinds of data as we go along, so to start we want to focus on just one state with one kind of data, and go from there. I think this will help us stay on track and be able to demonstrate incremental progress.

Overall, our design won’t be too complicated in terms of the number of applications and services we are using, the real complexity is going to come from piecing together the data we need in a clean fashion for the model to generate future data and to be able to visualize it.

I am in charge of census data, Tharun is taking on weather data, and Aditya is doing traffic data. I plan on testing out extracting the Census data and visualizing it a few ways to see how well Tableau will handle it.

To finish up the system design, I asked Alper to write up a summary of the diagrams, and then I incorporated all of the diagrams the team made and tied everything together into the final document.

One thing I think is missing from our design is a database schema diagram of the format of data we will store in Google Big Query. I will create one and include it in a future report.