EPAP FY17

Jordan Walker

2016-10-14

1 Summary

A year in and I am more fully a member of the Data Science team. This is an exciting time to be a part of this team, as it is more able to be nimble and take advantage of the opportunities that are represented in future priorities for the Water Mission Area (WMA). I'd like to first briefly reflect on my priorities from the last year, followed by a short list of priorities for the next year. I'm going to attempt to keep my FY17 goals more succinct to prevent the diffusion of focus that results from too many goals without clear priority of which are more important. Lastly, I will attempt more fully to determine evaluation criteria in order to reflect often on the progress being made and look back during reviews for whether that goal is accomplished.

2 FY16 Goals

2.1 Integration with Data Science team

The primary difficulty in FY16 was defining the role I would play on the Data Science team and how I could still contribute to work done within the Software Engineering team of which I was a part. Defining the priorities on projects on which I am to contribute and having better direction on how I am able to contribute would go a long way to reducing the struggle across teams. I would place this disconnect in cross-team collaboration and communication as the main struggle in fully integrating with the Data Science team. Progress: 7 out of 10.

2.2 Collaboration across teams

This goal goes hand in hand with the above goal, though I think shows less progress. My previous year goal had attempted to set up a way in which I could be tapped on the shoulder to give aid to a project, either in my computer science capacity on more experimental features, or in my data science capacity to help with a project in a more analytic way. The reason I suspect for the inability for this goal to be deemed successful is an increasing reliance on process as the driver of progress. Because there is no clear process set up for inclusion of members of another team, progress was limited to a large extent. Progress: 4 out of 10.

2.3 EDGE portfolio development

Working toward EDGE did not occur to any real extent this year. The effort that makes the most sense to describe as a marquee product for the portfolio ended up being the VIZLAB development that became more of a priority. I've put much of this effort on hold and will not list it as a goal for the upcoming year. Progress: 2 out of 10.

2.4 Geo Data Portal enhancements

The goal for better job control to optimize the speed at which jobs can get done, as well as a more scalable system that can lead the way in OWI for a system that scales up and down in response to demand was partially completed this year. Work across teams was again a major issue in achieving the goal, and the question of future effort on this project put doubts as to whether feature development beyond care and feeding was worth undertaking. I'm keeping the GDP on my goals list and hope that we can come up with a plan that will keep this project looking to the future. Progress: 6 out of 10.

2.5 Visualization framework development

As a measurable goal, I put forward that I'd like to be involved in another visualization that fully utilizes generalized code for visualization creation. This goal was partially met by the Great Lakes Microplastics and Climate Fish Habitat visualizations which were used to develop and exercise the platform. The Hurricane Matthew visualization was the first to fully use the platform, and while there is more progress to be made, the goal of a working platform can be checked off (with minor hesitation). The goal of a paper was perhaps premature, but I hold that as a future goal. Progress: 8 out of 10.

2.6 Hazards algorithm development

DSASWeb is no longer a project within OWI. I won't go into too much detail about this goal, other than to say that it would be unfair to say anything other than that it was not met. Progress: 1 out of 10.

3 FY17 Goals

3.1 VIZLAB platform

Now that the VIZLAB package is somewhat stable, it will be a major priority of mine to polish the edges and transfer knowledge so that it is an quick and easy to use by any developer tasked with working on visualizations. There will need to be additional features to get beyond where we are, especially related to figuring out the proper infrastructure, so some major effort on this will need to make its way into the work plan of the team.

Part of this includes getting it ready for the possibility of an internship team being tasked with creating a visualization as a summer project. This will be a major test of the platform, and I will need to be available to assist wherever needed as part of this being a success. As a side goal within this effort would be determining some expertises that are not covered by the current Data Science team, but might be useful particularly within the Visualization theme.

Additionally the Data Science team will continue to push ahead with new and compelling visualizations. I'd like to make a rough goal of one visualization per quarter at this point until we determine that the frequency should be more or less frequent. Another metric I want to keep on top of is the work hours and clock hours required from initialization to publication of a visualization. Lowering this number by relying on a foundational platform which removes the major burdens of development is a continued goal that I'd like to make progress on.

3.2 Geo Data Portal

This goal is the only other one I carried over from last year. I feel that the approach last year was attempting to bridge the gap between Software Engineering and Data Science, but the attempt did not work to that end. This year I'd like to lay out the goal in more of a manner that I have more control over, which is to say the development that I will do myself. As a general shift, I'd also like to move the GDP work from maintaining the status quo with some improvements to a more comprehensive look into what is being used for and how that can be improved, as well as adding value by exposing it to more users and use cases.

The first effort of importance is taking a bit more of a planning role on the technical work and requirements. Features should be planned based on what is valuable to current and future users. For the current users we have data that can be used to figure out which features are of value and what might be current hang ups with the system. For the future users, we can use strategic goals to determine what fields the GDP should be playing in and guide feature work in that direction. The main difference this makes in feature planning is the ability to run analytics on request data and add more data collection to do future analytics.

A further goal in this arena is a shifting focus on scalability from how we make a single GDP processing server handle the requests it gets, but to scale to multiple servers with different capabilities and even the possibility of processing being done local to the data for power users or utilization of High Throughput Computing (HTC). This shift will require the same type of planning as mentioned above, and has the benefit of further enhancement to handle future challenges in data access as priorities are made clear in the wider WMA.

3.3 Enhance outreach and training ability

A major goal of the Data Science team in FY17 is outreach and training or more broadly building a community of practice. I have to date been involved in some aspects of this, but I am making it my goal this year to take a larger part in this effort and improve my skills in this area. I will take part in training and development of new training material, especially where it overlaps with my other goals. In addition, work on infrastructure to support training and outreach fits my skill-set, so being a part of developing it is important. Lastly, I will build in more time to support users in areas I have expertise in.

3.4 HTC and HPC proficiency

It is clear that there is a push to scale up science to a level that requires more computing power than available to a user's machine. This is one of the focuses of the Data Science team, but thus far have limited proficiency in HTC and even less in High Performance Computing (HPC). Building up this capability and supporting users whose workflow fits well (or can be made to fit well) is another goal I put on myself. In particular, the focus will be on utilizing CondorHTC and YETI computing capabilities as well as investigation of other approaches to HTC and HPC.

As is the pattern being established, this goal overlaps in several important ways. There are some interesting visualizations that could result from use of an HTC or HPC cluster, so building VIZLAB to work with this in mind should at least be considered. As mentioned above, the GDP is well-suited to running tasks on such a cluster, so proficiency in such systems is important in being able to design the architecture of that project going forward. Outreach and particularly training should in the future guide users towards being able to utilize compute power available to them, and having the foundation for writing their code in a way that it can scale when such analysis requires it.

In measurable terms, I'd like to be involved in the general effort to build proficiency in this area, and contribute to any report or publication that is derived from this effort.

3.5 Solidify shared infrastructure presence for Data Science

One risk that was identified in our end-of-year/start-of-year offsite meeting was that we under-utilize infrastructure and depend generally on ad-hoc services and local infrastructure that is not managed in the enterprise standard way. Several pieces of the Data Science project lifecycle depend at least in part on shared infrastructure for easier collaboration. In particular this includes shared databases (SQL and NoSQL) to collect and dissect the data for analysis, as well as shared build and operational environments. There has been an aversion to this type of infrastructure on the Data Science team to date, so recognizing that is a start towards fixing it.

In a few areas I'd like to be involved in the solution. First, infrastructure for VIZLAB has been a limited presence on Amazon Web Services, but more complete infrastructure services are required to fully implement the platform as is needed to serve visualizations. Extending the GDP infrastructure to at least partially overlap with a shared Data Science infrastructure will enable the analysis involved in planning features for future GDP work. In terms of outreach, the solutions we have are generally acceptable, but if it is determined that they limit our ability to support the community (e.g. community badges) then we should be able to take advantage of the infrastructure available to us. Lastly, there are some important ways that shared infrastructure overlaps with the HTC and HPC effort that would benefit my involvement.

I will consider this goal successful if the Data Science presence on available infrastructure increases by a relatively large extent. Additionally, the current solutions that are non-optimal on ad-hoc infrastructure that can be migrated to a better solution will add to my metrics of success.

4 Conclusion

I've re-aligned my goals with what came out of the Data Science offsite meeting, and what was clearly not working last year. With this, I hope I am set up to take advantage of priority shifts and opportunities for the future of Data Science within the WMA. There are a lot of intersections that cross-cut my goals, which should enable me to bring my talents to these broader goals in an efficient manner. I've kept these goals limited to an extent, but still ambitious enough to achieve great things in the coming year.