EPAP FY16

Jordan Walker

2015-11-16

1 Summary

With the changing structure of OWI and the history of the different groups clouding status of individual positions, this year is an important one for laying out a new direction on sturdy foundations. The newly created Data Science team has somewhat less of a burden of the background and history as the other groups. With my transition from what was the Developer (Java) team to the Data Science team, there is much opportunity to lay out a direction that will help both teams and OWI more broadly. With this document I'd like to lay out what I think I can bring to the Data Science team as well as what I'd like to keep in touch with from my previous work.

As a side note, I think it is worth discussing in this context the transition from work that has been planned under the old organizational structure and that work which is upcoming. The planning techniques of the two groups are quite different, and the communication between them regarding my time has not gone as smoothly as we would have wanted. I think this may only escalate going forward, so it is something to articulate at this point. The expectations of projects I have worked on should be adjusted to reflect this new position, and we will likely have to deal with some obligations that are outstanding that might affect planning. To date I am not aware of what conversations have occurred regarding this, and I have failed to initiate them myself.

2 Integration with Data Science team

As a member of the Data Science team going forward, I'd like to better integrate with the team and the projects they currently work on. Thus far, the team has been patient with me and kept me involved with what is going on even though my work has kept me from being actively involved in much of what the team is doing. With the holding pattern that we are currently in, it would be nice to define what my role would be going forward and start building a measure of success in terms of integration with this team.

In addition to this, the integration has been stalled by ongoing work on projects outside the Data Science team. Balancing the priorities of two teams has been a struggle, and having a clear picture of what the overall priorities are

would be beneficial. Overall, clear communication of what the expectations for dealing with resources across teams are will help to clear up this conflict going forward. As much as possible I'll try to be proactive about managing my time in relation to this, but it seems that some of this will need to happen a level above where I sit.

3 Collaboration across teams

One of the challenges facing our new organizational structure is building boundaries between members of different teams that result in communication breakdowns and disparate project efforts. I'd like to be part of the solution to this by continuing to work across team boundaries. Primarily it would make sense to work closely with the Software Engineering team to align the directions of development efforts with those of the Data Science team, but reaching out to other teams is important as well. There are several areas where this collaboration is most beneficial, both continuing to help on projects I have experience with including knowledge transfer where needed and prototyping new schemes that may be used for further development. As with the above goal, setting up measures of success for collaboration would build towards better evaluation of this position.

To elaborate further on the prototyping concept, one of the issues with our current development style is that we commit to certain patterns and solutions before fully vetting and testing them out. This pattern is likely to continue within the Software Engineering team as projects need to be delivered in full and this is the result. The Data Science team has less of an operational presence and more of a research focus, so doing software prototypes that do not need to be pressed into production is more in line with this style. A useful result of this difference is that it is possible to do more research-like efforts on prospective projects (NWIS modernization, AQCU, etc) that can be applied to the regular development should it successfully solve the stated problem. I will add to this in the next section.

4 EDGE portfolio development

The possibility of a career path for computer scientists that involves EDGE (`http://www.usgs.gov/humancapital/hr/documents/rgeg-edgeexhibitb-1.pdf`) as the evaluation technique has been batted around for a couple years. There is still relatively few EDGE scientists in the USGS, and fewer creating software, but it seems to be growing as a valid option. I feel as though I'm in a good position to blaze the trail to EDGE for developers in OWI, both in terms of the flexibility that exists within the Data Science team to build the needed portfolio and the work I have under my belt in the last few years.

In the next year I'd like to work on pulling together a portfolio and while continuing to build out the work therein. This will include a certain amount of

time to present the work I've done in an effective way, and it should also include a couple of new research "prototypes" to round out the portfolio. To put this goal into concrete terms, I'd like to publish a first author paper that can be added to this portfolio. I have several ideas that I will throw out here, but the main idea would be to come up with a compelling story for the EDGE board.

The first idea would be to write an R package that allows for users to download BitTorrent data within scripts. Primarily this would be to allow packages to define data that can be lazy loaded without making huge packages or having to setup a service for the data. One focus of these data would be remote sensing data such as Landsat where the data volume is large and access is tricky. An added benefit of this would be high throughput computing where multiple nodes need the same data and can thereby reduce bandwidth by sharing with one another. The technique of using BitTorrent for transferring data between datacenter nodes is known to be used by Twitter and Facebook, but using it for science data would be somewhat novel. One hurdle that should be mentioned is the policies that might interfere with the use of this protocol.

A second concept that would be good for OWI as well as make for a novel paper is an idea that has been discussed around here, but a proof-of-concept has never been developed. The idea is to take the burden of complex queries for long period of record datasets off the database and move it to the service level. This would involve breaking the data up along several dimensions (temporal, spatial, select fixed domain fields) and providing a mechanism for downloading these chunks (canned queries) and caching them at the service. Part of this concept came out of the microservice implementation that was approached for AQCU, and it would be useful for such a service where the source data changes, but a vast majority of chunks rarely change. Room for this type of investigation to take place would allow for some unique problem solving and innovation that is currently difficult to do.

Other possible outcomes that would result in a paper will be described in the next sections.

5 Geo Data Portal enhancements

The Geo Data Portal is an application developed over the last 5 years that brings valuable remote sensing and gridded model data into a form that is usable by scientists working at the landscape level. As the lead developer on the Geo Data Portal over the last several years, maintaining a role on this project is important to me. I feel as though there are both improvements that can be made to enhance the user experience, as well as new features that can bring in more users to this powerful tool. The main need in the coming year is 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.

One of the unique aspects of this project is the cross portfolio usefulness for it. As a project it is more generic than much of our portfolio and doesn't

cater to any particular user group, but provides a capability that can be used by many users, including other OWI applications. These applications include the National Water Census and EnDDaT applications as well as the lake modeling and Powell Center work going on in the Data Science team. I'd like to build on this success and open up the Geo Data Portal to more users and more data.

6 Visualization framework development

In the transition to Data Science I have been most involved in the visualization creation. With both the California drought visualization and the Colorado Basin drought visualization we've learned what to do and what not to do with this type of product. In the next year there are plans to codify some of these lessons to make it easier to create visualizations without as much of a struggle to spin up. As a measurable goal, I'd like to be involved in another visualization that fully utilizes generalized code for visualization creation. I think this outcome would also be interesting for a paper describing the features that we come to as important in this context.

7 Hazards algorithm development

The last opportunity provides the ability to collaborate with other teams while also offering room for growth as a Data Scientist. Our work with coastal marine geology program in the hazards mission involves interesting ways of visualizing shoreline data as well as developing algorithms that can calculate useful features for coastal hazards. The main outcome which could come of this is a deterministic algorithm for calculating a baseline that is used in the calculation of shoreline change rates. This has been stated to be something of a home run in the coastal science community, and my work on the calculation side of the DSASweb application positions me well for developing such an algorithm.