IGARSS 2014 Abstract to "Implications of Big Data to Remote Sensing"

Identifying Episodes of Earth Science Phenomena Using a Big-Data Technology

Introduction

A significant portion of Earth Science investigations is phenomenon- (or event-) based, such as the studies of Rossby waves, volcano eruptions, tsunamis, mesoscale convective systems, and tropical cyclones. However, except for a few high-impact phenomena, e.g. tropical cyclones, comprehensive records are absent for the occurrences or events of these phenomena. Phenomenon-based studies therefore often focus on a few prominent cases while the lesser ones are overlooked. Without an automated means to gather the events, comprehensive investigation of a phenomenon is at least time-consuming if not impossible.

We have constructed a prototype Automated Event Service (AES) system that is used to methodically mine custom-defined events in the reanalysis data sets of atmospheric general circulation models. Our AES will enable researchers to specify their custom, numeric event criteria using a user-friendly web interface to search the reanalysis data sets. Moreover, we have included a social component to enable dynamic formation of collaboration groups for researchers to cooperate on event definitions of common interest and for the analysis of these events.

An Earth Science *event* (ES event) is defined here as an episode of an Earth Science phenomenon (ES phenomenon). A cumulus cloud, a thunderstorm shower, a rogue wave, a tornado, an earthquake, a tsunami, a hurricane, or an El Niño, is each an episode of a named ES phenomenon, and, from the small and insignificant to the large and potent, all are examples of ES events. An ES event has a duration (often finite) and an associated geo-location as a function of time; it's therefore an entity embedded in four-dimensional (4D) spatiotemporal space.

Earth Science phenomena with the potential to cause massive economic disruption or loss of life often rivet the attention of researchers. But, broader scientific curiosity also drives the study of phenomena that pose no immediate danger, such as land/sea breezes. Due to Earth System's intricate dynamics, we are continuously discovering novel ES phenomena.

We generally gain understanding of a given phenomenon by observing and studying individual events. This process usually begins by identifying the occurrences of these events. Once representative events are identified or found, we must locate associated observed or simulated data prior to commencing analysis and concerted studies of the phenomenon. Knowledge concerning the phenomenon can accumulate only after analysis has started. However, as

mentioned previously, comprehensive records only exist for a very limited set of high-impact phenomena; aside from these, finding events and locating associated data currently may take a prohibitive amount of time and effort on the part of an individual investigator.

The reason for the lack of comprehensive records for most of the ES phenomena is mainly due to the perception that they do not pose immediate and/or severe threat to life and property. Thus they are not consistently tracked, monitored, and catalogued. Many phenomena even lack precise and/or commonly accepted criteria for definitions. Moreover, various Earth Science observations and data have accumulated to a previously unfathomable volume; NASA Earth Observing System Data Information System (EOSDIS) alone archives several petabytes (PB) of satellite remote sensing data, which are steadily increasing. All of these factors contribute to the difficulty of methodically identifying events corresponding to a given phenomenon and significantly impede systematic investigations.

We have not only envisioned AES as an environment for identifying custom-defined events but also aspired for it to be an interactive environment with quick turnaround time for revisions of query criteria and results, as well as a collaborative environment where geographically distributed experts may work together on the same phenomena. A Big Data technology is thus required for the realization of such a system. In the following, we first introduce the technology selected for AES in the next section. We then demonstrate the utility of AES using a use case, Blizzard, before we conclude.

AES Environment

Since Earth science data are largely stored and manipulated in the form of multidimensional arrays, evaluation of array performance constitutes the initial phase of the project. Several candidate Big Data technologies have been evaluated, including MapReduce (Hadoop), SciDB, and a custom-built Polaris system; they have one important feature in common: the shared nothing architecture. Our evaluation finds SicDB to be the most promising and hence is selected for implementing AES.

SciDB [1] is an all-in-one data management and advanced analytics platform that offers complex analytics inside a next-generation parallel database based on the array data model. It supports data versioning and provenance to facilitate scientific research applications, where reproducibility is paramount.

We have installed SciDB on a 35-node experimental cluster at the NASA Center for Climate Simulation (NCCS), each node of which is equipped with the following:

- 2x8 SandyBridge Intel cores,
- 32 GB RAM memory, and
- 36 TB local storage,

totaling 560 CPU cores, more than 1 TB of RAM memory, and more than 1 PB of raw storage.

Blizzard Use Case

According to the US National Weather Service (NWS), the following conditions are expected to prevail for at least 3 hours for a blizzard [2]:

- sustained wind or frequent gusts to 15.6 m s^-1 (35 mph) or greater, and
- considerable falling and/or blowing snow, i.e., reducing visibility frequently to less than 400 m (1/4 mile).

As we can see, the definition uses several imprecise qualifiers, such as “expected”, “sustained”, “frequent(ly)”, and “considerable.” In addition, it apparently is based on "point observations”. Point-based observations do not translate straightforwardly to space/time-averaged parameters, such as those used in this use case from NASA's Modern Era Retrospective analysis for Research and Applications (MERRA) datasets [3]. It is obvious, however, visibility is the crucial criterion. Unfortunately, MERRA currently does not include visibility information.

Visibility reduction is directly correlated to the increase of in-air snow mass concentration, which is caused by both falling snow and blowing snow. MERRA does contain snowfall rate, which is used to derive visibility reduction due to falling snow via a relation provided in [4]. Blowing snow can happen when there is sufficient dry and new snow accumulation on the surface with sufficiently strong wind. Thus, visibility reduction due to blowing snow is derived using MERRA data by combining snow accumulation on the surface and wind speed at 10-meter above surface, according to the relation reported in [5].

We first extract a 2010 US Winter subset from the relevant MERRA hourly datasets at the spatial resolution of 1/2-degree in latitude and 2/3-degree in longitude to refine our “blizzard criteria” by trial-and-error. Trial-and-error is required since the translation from point-based to space-time-averaged observations is unknown. The subset is selected because there are reliable textual records of winter storms from NWS for the period for verification. Using a subset also ensures quick turnaround for the trial-and-error. In fact, most of the blizzard queries takes less than 1 s, i.e. almost instantaneous, when applied to the subset.

After the blizzard MERRA grid cells have been identified, a user-defined operator (UDO) of SciDB for connected component labeling (CCL) is applied to the results, which connects spatially and temporally adjacent blizzard grid cells to form a distinct blizzard event. Each blizzard event is given a unique label in a mask array. The mask can be applied to the original MERRA datasets and statistics of any observations contained in these MERRA datasets can thus be obtained for each individual event.

Conclusion

Using blizzard as an example, we have demonstrated the analysis capability of AES supported by a Big Data technology. It substantially reduces the effort for identifying such events in climate data records and provides a significant boost to research productivity. In a much more

expanded system where we scale our experimental cluster to hundreds of thousands of nodes, it is conceivable that it can support many data analysts conducting analysis simultaneously, often with near instantaneous responses.

References

[1] http://www.scidb.org

[2] http://w1.weather.gov/glossary/index.php?letter=b

[3] http://gmao.gsfc.nasa.gov/research/merra/

[4] R. M. Rasmussen, J. Vivekanandan, J. Cole, B. Myers, and C. Masters, “The estimation of snowfall rate using visibility”, *J. Appl. Meteorol.*, vol. **38**, 1542-1563, 1999.

[5] G. H. Liljequist, “Energy exchange of an Antarctic- snowfield”, Norwegian-British-Swedish Antarctic Expedition, 1949-52, Norsk Polarinstitutt, Oslo, *Scientific Results*, vol. **II**,. part 1C.