Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-153-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Coordinating an operational data distribution network for CMIP6 data" *by* Ruth Petrie et al.

Anonymous Referee #1

Received and published: 26 July 2020

This paper presents the infrastructure for data distribution for CIMP6 as well as its (high level) different components. In the last part of this work, the authors discuss the deployment and validation (CIMP6 Data Challenges) of this very same infrastructure.

The quality of the paper is good but I have some comments and concerns to be addressed (specially around references and details):

1 General comments:

- On some parts of the paper and while talking about storage, the Big-O notation is used (but this is not consistently used through the work). I believe that given the

tone of the paper, this notation is not needed and just mentioning the units (e.g. 20TB) is enough.

- Data integrity validation (after data distribution) is not explicitly mentioned on the paper (even though it can be inferred from some of the statements). Could you please add some specific text on how you ensure that the distributed and/or replicated data has not been corrupted or tampered in any way?
- Disaster recovery is not mentioned in this work: I think it is an interesting aspect to discuss, specially given that the data is sharded. Could you please add some information about this?

2 Specific Comments:

- P2, l43: Could you please reference the recommended security policies?
- P3, l45: Could you please reference the necessary policies?
- P3, l46: Could you please reference the required resources?
- P4, l77: "federated data archive": Could you please add a reference to the architecture (if available)?
- P4, l88: "standardised set of rules": Could you please reference these rules?
- P4, l90: "HTTP" : Is TLS supported? If so, could you please state it?
- P5, l91: "in an automated way" : Could you please add more details and/or references about the automations?

- P5, l92: "see the same search results from any index node". How is split-brain and replication latency resolved here? (Maybe you can add some more context on the paper).
- P5, l103: "18PB" : Previously you stated 20PB, is this a more accurate result or maybe a typo?
- P7, l140: "includes a database" : What database (technology) is this? Maybe you can also put a reference.
- P10, l110: I think it could be interesting some details on how the PID (the id value itself) is formed (also maybe some reference on how it is calculated to be unique and avoid clashes).
- P12, l284: Are you using Elastic Search as part of the analyzer? If so, maybe you could state it. Otherwise, could you please add some references or links to the code of the analyzer.
- P12, l284: About the dashboard: For what I can see your are using a custom built UI. Any reason not to use Kibana (as you are already using part of the ELK stack)? If so, I think it is relevant to get it mentioned here. Also I think it is interesting to link the code repository for the dashboard in here.
- P13, l288: "recommended a basic hardware" : Could you please list these hardware requirements or put a reference to them?
- P14, l309: After reading all the section 5 it seems you are describing a (somehow manual) deployment procedure. Did you follow any specific methodology, if so, could you please mention it and maybe add a deployment diagram (it will add more clarity to all the section 5).
- P17, l383: "to run the test suite" : Can you please put a reference (if available) to the test suite code and/or documentation?

- P17, l435: "In order to assist sites that were not able to participate to effectively publish" : Are there any plans to automate all the steps described in the document? If so, could you please state it on the paper?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-153, 2020.