ESGF SEARCH WORKING TEAM
PROGRESS UPDATE & FUTURE ROADMAP

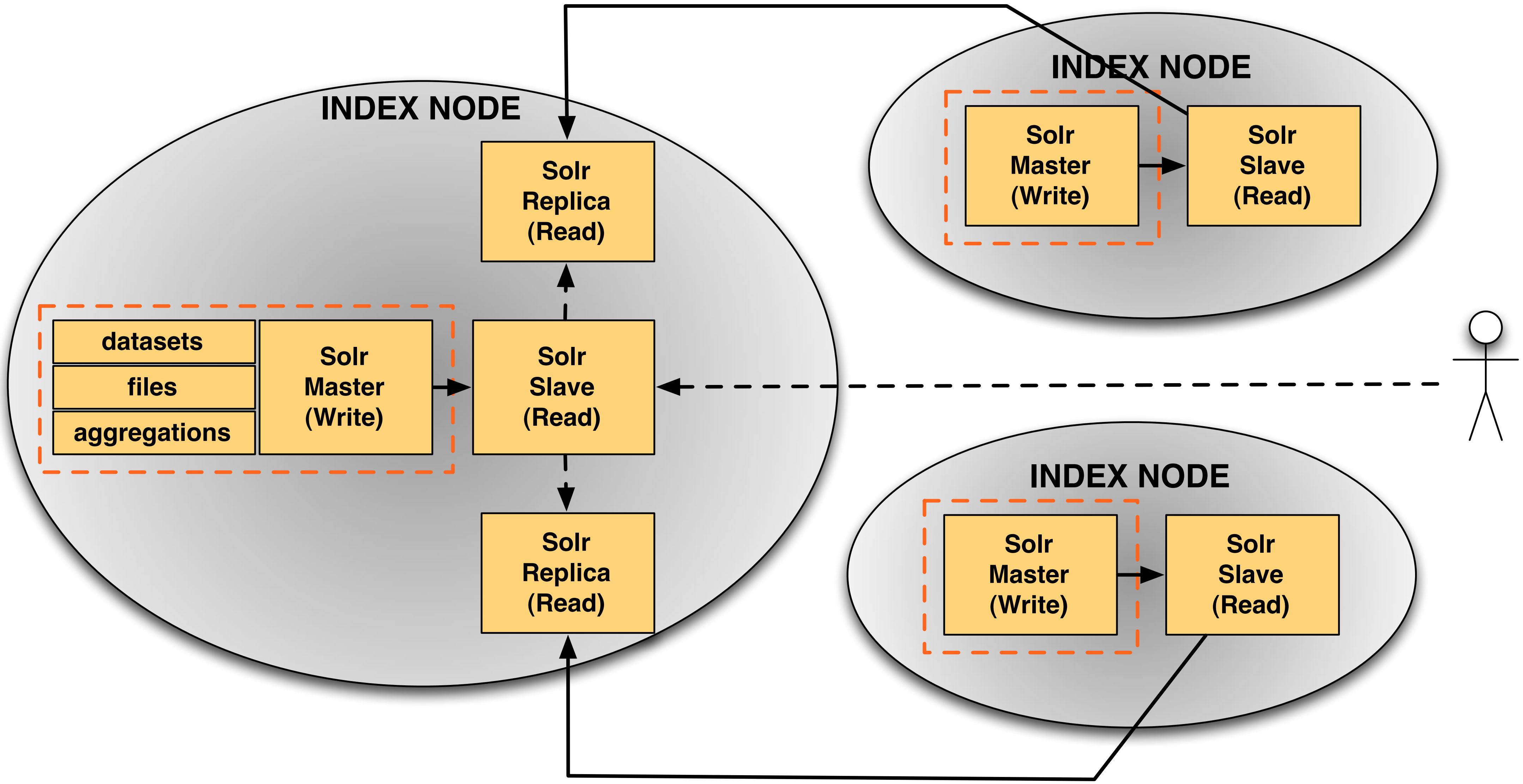
ESGF F2F Workshop,
Washington, DC, December 2016

Lead: Luca Cinquini
Contributors: Sasha Ames, Katharina Berger, Alan Iwi

ESGF Search Architecture

Main features

- Based on Apache Solr
- Master (publishing) + Slave (querying) Solr instances
- Distributed search
- Local replica of remote shards



2016 Progress & Accomplishments

In 2016, considerable development took place for the ESGF Search Services w/ 3 major goals:

- Recovery from the 2015 security incident
- Functionality requirements from major projects such as CMIP6, CORDEX and Obs4MIPs
- Scalability into much larger data volumes and adoption expected in the next future

Operations:

- Supported installation and configuration of Search Services as ESGF come back online

New functionality:

- Atomic metadata updates:
  - Changing metadata in place without having to republish the data
  - REST API for adding/updating/deleting any metadata field of an ESGF metadata record
- Data “retraction”: delete the data but keep the dataset level metadata
- Tagging datasets for multiple projects
- Search on datasets with date less than a given value

2016 Progress & Accomplishments

Infrastructure upgrades:

- Upgraded Apache httpd configuration to enhance security
- Enabled automatic propagation of Solr schema changes from master to slave and replicas
- Migrated documentation to CoG, revised and updated

New research:

- Investigated Solr Cloud to enhance search performance and scale into the future
  - Solr Cloud can scale from 100K to 100M datasets
- Built first Docker image for ESGF Index Node
  - Includes Tomcat, esgf-search web app, Solr master + slave instances

2016 Missed Milestones

- Implement metadata validation against Controlled Vocabularies (CV)
  - Requirement shifted to client-side publishing as opposed to server-side
  - Will implement any server-side functionality as needed

- Package standalone authorization service to be deployed on Index Node to authorize publishing operations
  - Not a high priority any longer, as authorization is performed by internal Index Node component using local policy files and remote IdP Attribute Service

2017 Roadmap

In 2017, we expect development to follow the same goals as in 2016:

- operational support for CMIP6
- gradual evolution of search architecture to scale into future much larger volumes

Specific tasks:

- Implement any additional search requirements as CMIP6 data is published into ESGF
- Upgrade Solr engine to 6.3.0 and beyond (currently running 5.2.1)
- Release a Docker image for the ESGF Index Node
- Enable Solr Cloud within Docker image

ESGF
Earth System Grid Federation

Additional Resources Needed

- Staff: someone to develop a framework to monitor consistency of search results across the federation
  - Part of a larger ESGF monitoring framework
  - Must allow for replication delays across nodes
- Staff: node administrators willing to install and test the Docker ESGF Index node
- Staff: group responsible for monitoring and enforcing data publication policies
  - Who can publish datasets into a specific data collection, such as CMIP6
  - Which data collections are replicated across ESGF (and which shouldn’t)
- Hardware: Cloud instances to deploy and test new architecture configurations
- Funding: always welcome…

Community Feedback from Survey

- Results:
  - 34% of users found the distributed global search the most difficult part of ESGF
  - 27% of users found the distributed global search the most useful feature of ESGF
- Conclusions:
  - Searching across distributed data centers is not easy…
  - User experience is affected by inconsistencies in data replication
  - Perhaps remove or hide the advanced search options ? (replica, local search, all versions)

Collaborations Needed

Continue collaborations with:

- ESGF Publishing Working Team - for client-server interactions
- ESGF User Interface Working Team - to evolve the front-end to the back-end services
- ESGF Security Working Team - to evolve the generation of wget scripts
- ESGF Quality Control and Citation Working Team - to update the metadata catalogs
- ESGF Statistics Working Team (aka Dashboard/Desktop) - to report publishing metrics