Node Registry Integration

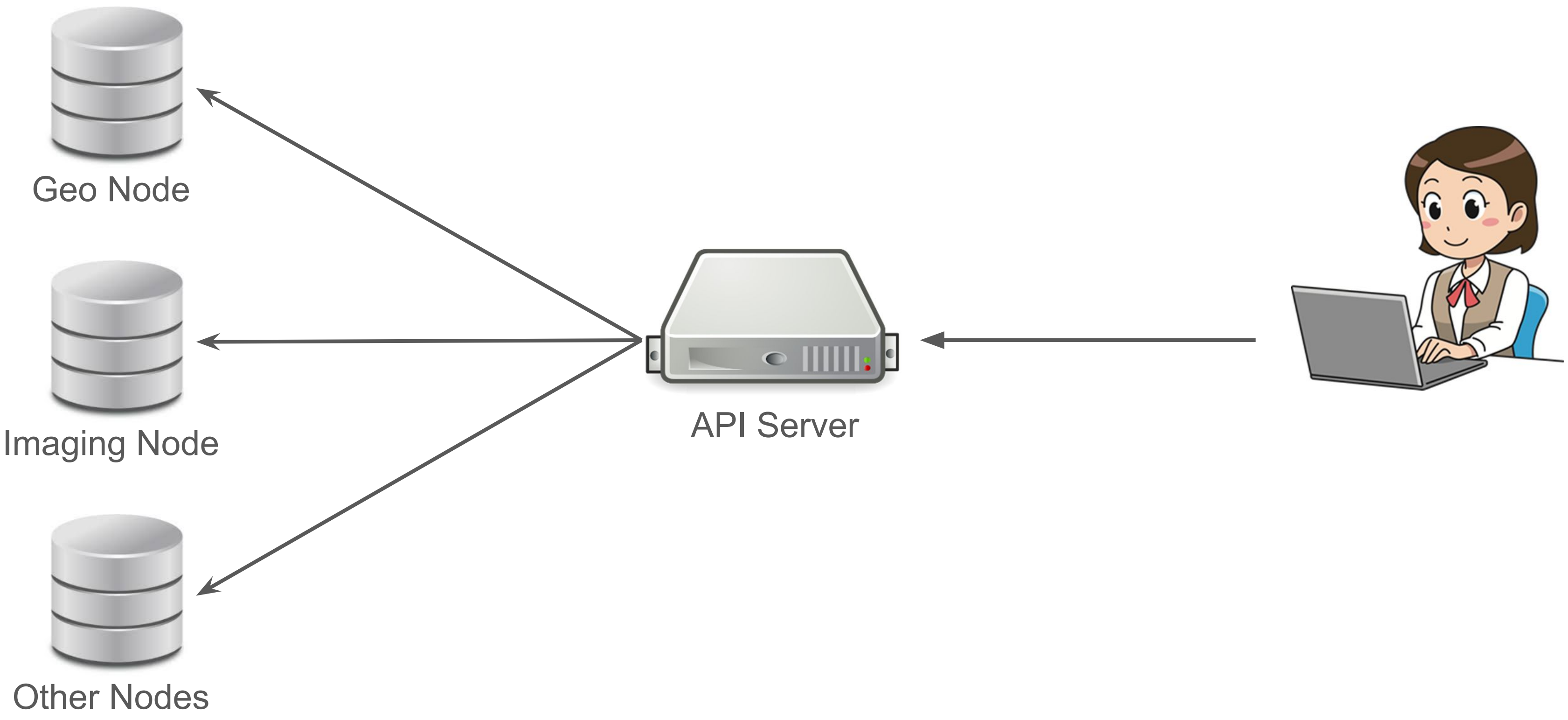
Yevgen Karpenko

2/25/2021

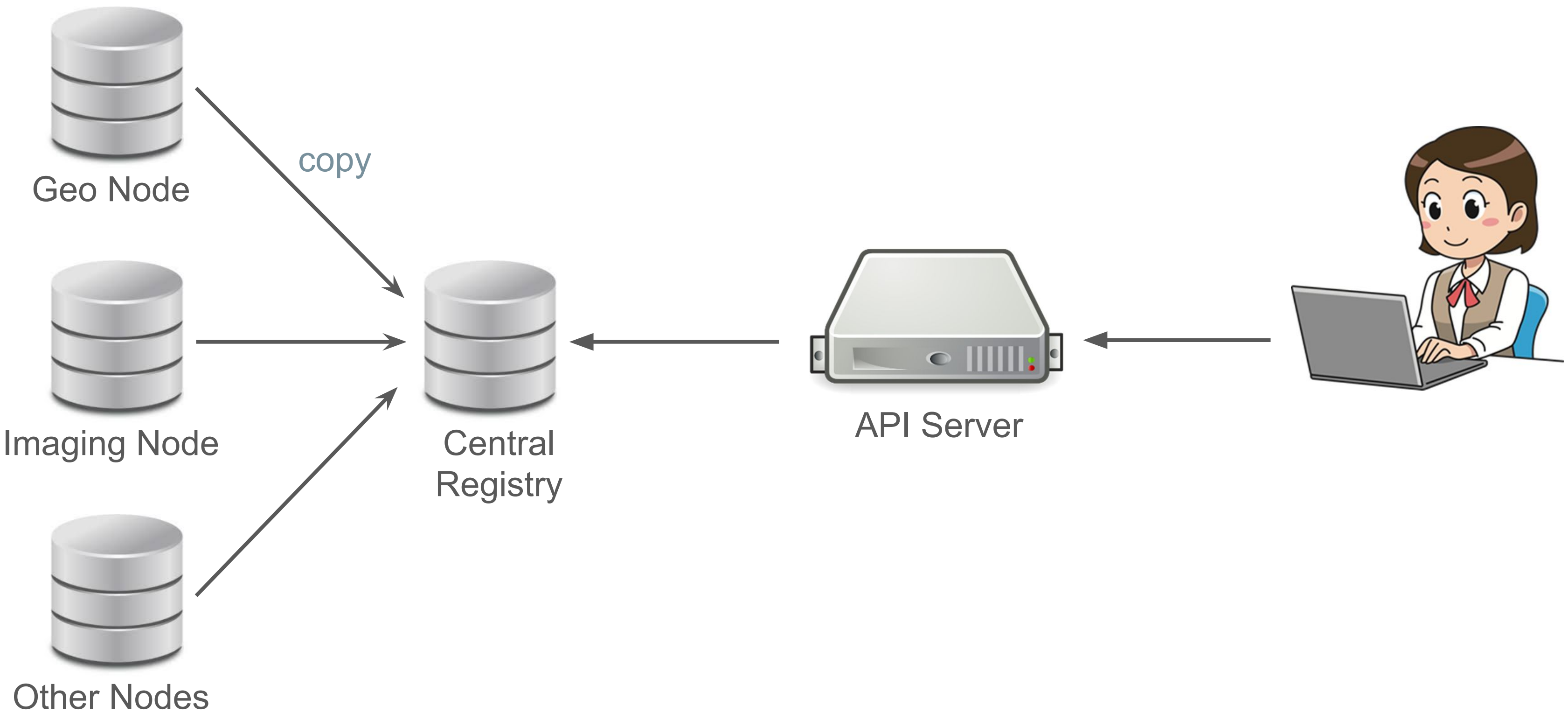
Part I

Distributed vs Central Registry

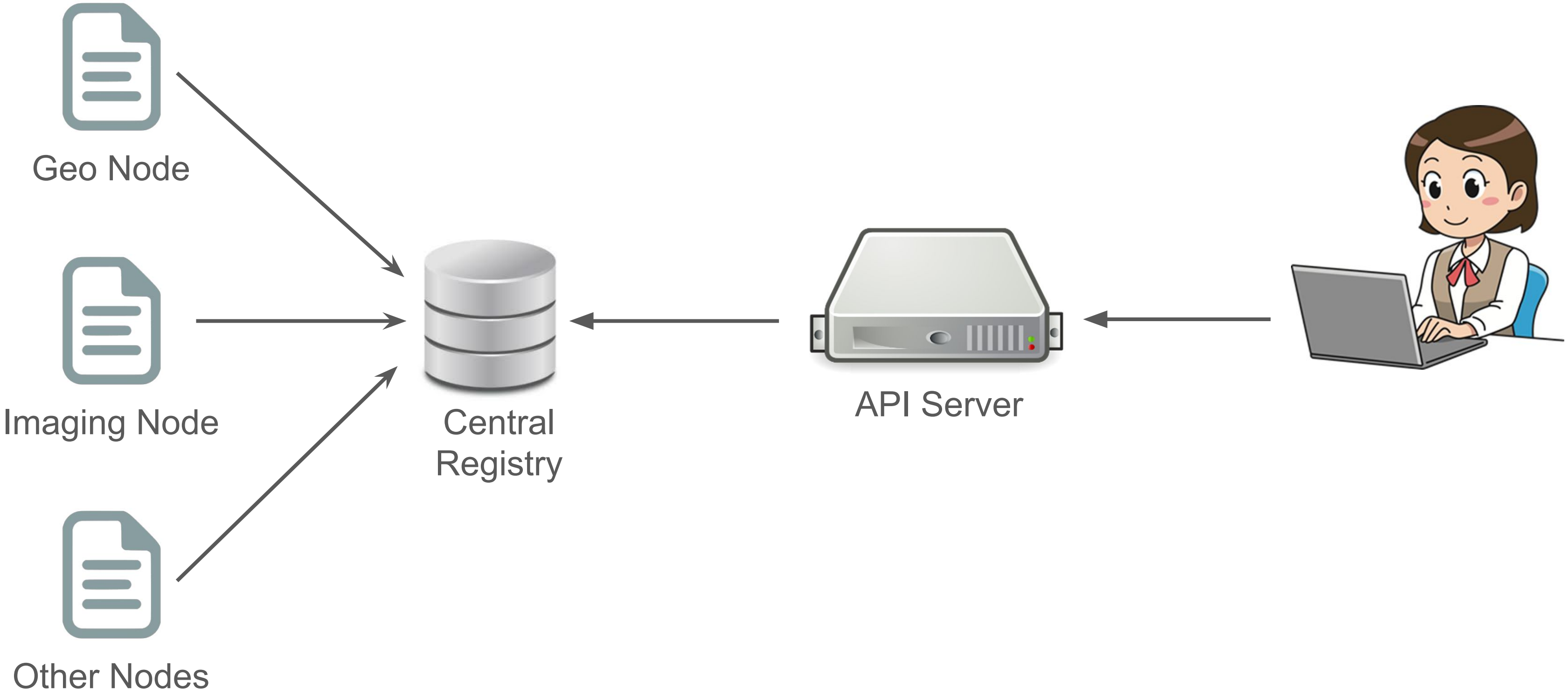
Option 1: Each PDS node maintains its own registry. There is no central registry.



Option 2: Each PDS node maintains its own registry. Node data is copied to central registry.



Option 3: There is only central registry. Nodes don’t have their own registries.



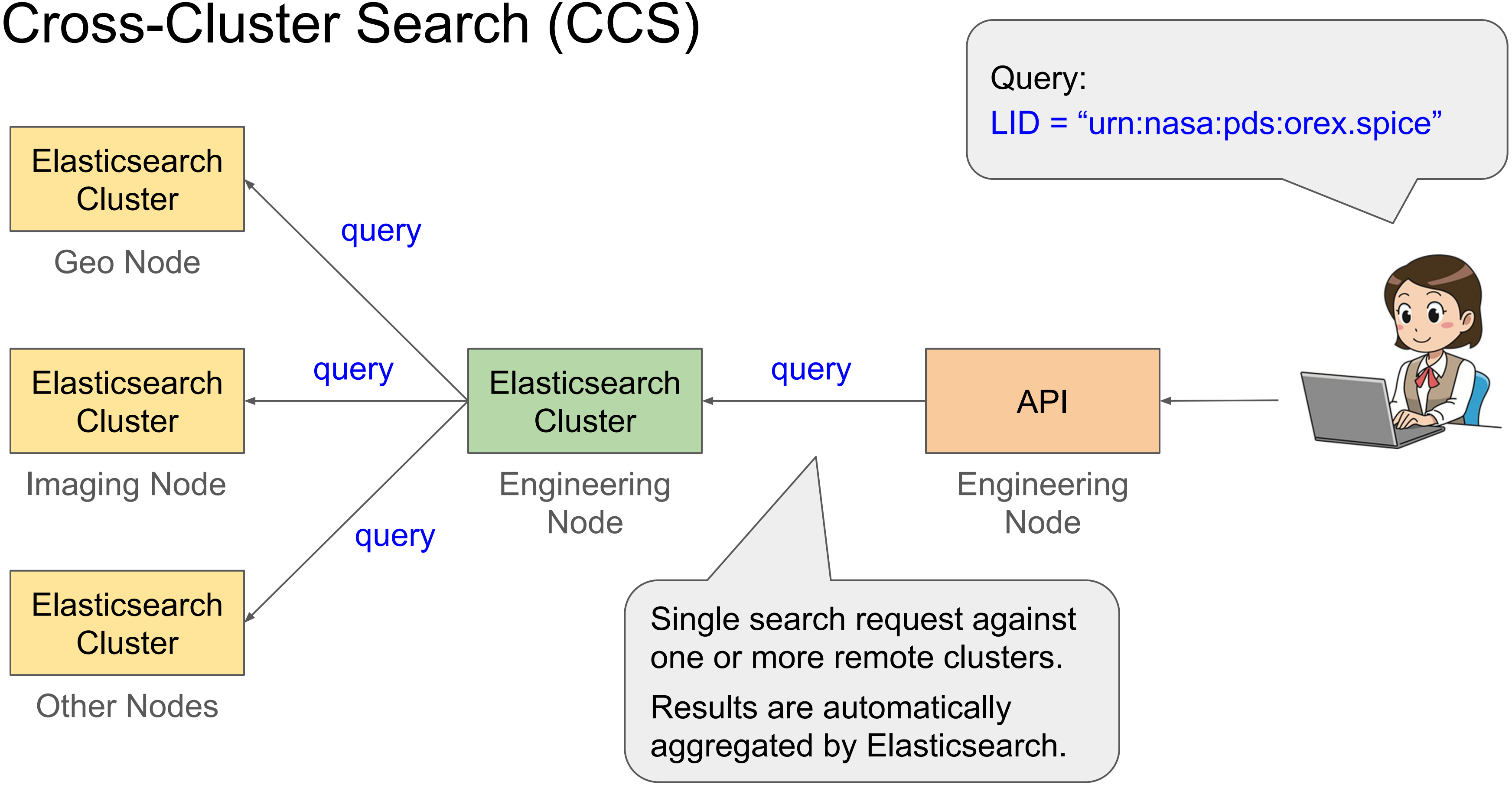
Pros & Cons. Open Questions.

	Nodes Only	Nodes + Central	Central Only
Too many copies of data (4 copies If both node and central registry clusters use recommended data replication factor of 2).		X	
How to run cross-cluster (cross-node) searches including “order by” and aggregate queries?	X		
Complex synchronization logic.		X	
Can single Elasticsearch cluster handle data from all PDS nodes? In production would it be easier to run 7 small Elasticsearch clusters for PDS nodes or 1 huge central cluster?		X	X

Part II

Distributed Registry (Option 1)

Cross-Cluster Search (CCS)



Query:
LID = “urn:nasa:pds:orex.spice”

Elasticsearch Cluster

Geo Node

Elasticsearch Cluster

Imaging Node

Elasticsearch Cluster

Other Nodes

query

query

query

Elasticsearch Cluster

Engineering Node

query

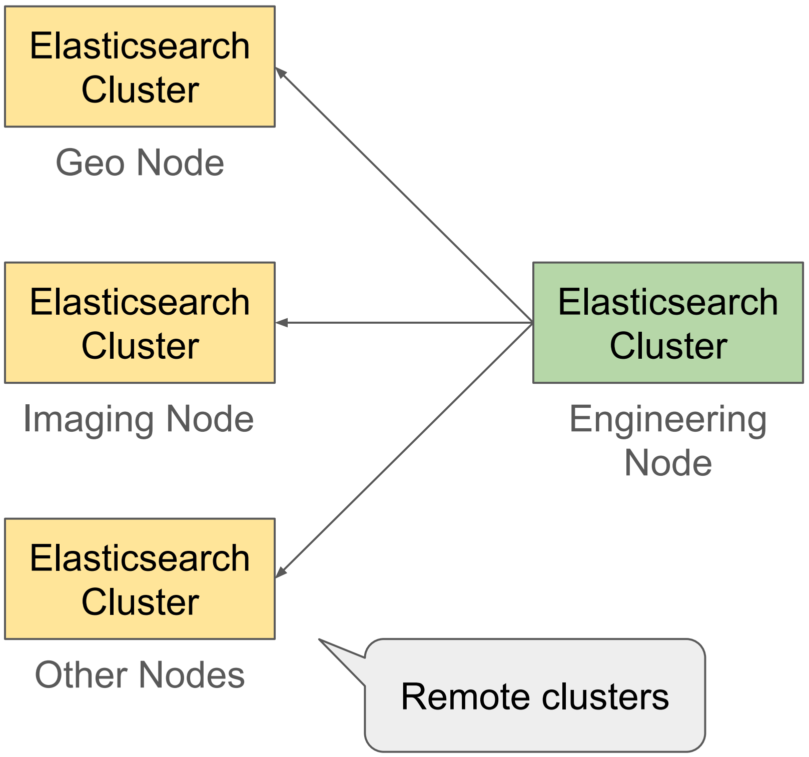
API

Engineering Node

Single search request against one or more remote clusters.

Results are automatically aggregated by Elasticsearch.

CCS Configuration



Remote cluster configuration at Engineering Node:

```
PUT _cluster/settings
{
 "persistent": {
  "cluster": {
   "remote": {
    "geo_cluster": {
     "seeds": [ "geo.pds.local" ]
    },
    "img_cluster": {
     "seeds": [ "img.pds.local" ]
    },
    "naif_cluster": {
     "seeds": [ "naif.pds.local" ]
    }
   }
  }
 }
}
```

Cross-Cluster Search Query

A remote query is similar to a local query. Instead of a local index, there is a list of <cluster>:<index> tuples.

```
GET
/geo_cluster:registry,img_cluster:registry,naif_cluster:registry/_search
{
  "query": {
    "match": {
      "lid": "urn:nasa:pds:orex.spice"
    }
  }
}
```

CCS More Info

Docs (cross-cluster search)

https://www.elastic.co/guide/en/elasticsearch/reference/current/modules-cross-cluster-search.html

Docs (remote clusters)

https://www.elastic.co/guide/en/elasticsearch/reference/current/modules-remote-clusters.html

Pricing

- Free. Included in all Elasticsearch versions.

Cross-Cluster Search for Amazon Elasticsearch Service

More info

https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/cross-cluster-search.html

Pricing

- There is no additional charge for searching across domains.

Limitations

- Can't connect to domains in different AWS Regions.
- Maximum of 20 connected domains / clusters.
- Can't use AWS CloudFormation to connect domains.
- Can't use cross-cluster search on M3 and T2 instances.

Part III

Node + Central Registry (Option 2)

Which data to copy?

Elasticsearch indices:

- Registry
- References
- Data dictionary

How to copy?

Q: Batch vs streaming?

A: TBD

Q: Elasticsearch cross-cluster replication (CCR) vs custom solution vs 3PP ETL tool?

A: TBD

Q: For batch mode, on-demand vs scheduled?

A: On-demand

Cross-cluster replication (CCR)

- This feature requires Platinum or Enterprise License.
- Active-passive (leader-follower) model is used. The “follower” index is read-only. The “leader” index can serve reads and writes.
- Soft-deletes must be enabled.
- The follower index cannot be offline longer than 12 hours (by default).

3PP ETL Tools

Logstash

- TODO: Provide more details.

Apache NiFi

- Huge download size (1.5GB)
- Will take some time to learn, install, and configure.
- Not clear if latest Elasticsearch is supported. Some docs refer to very old ES 2.x.