

Pipelines

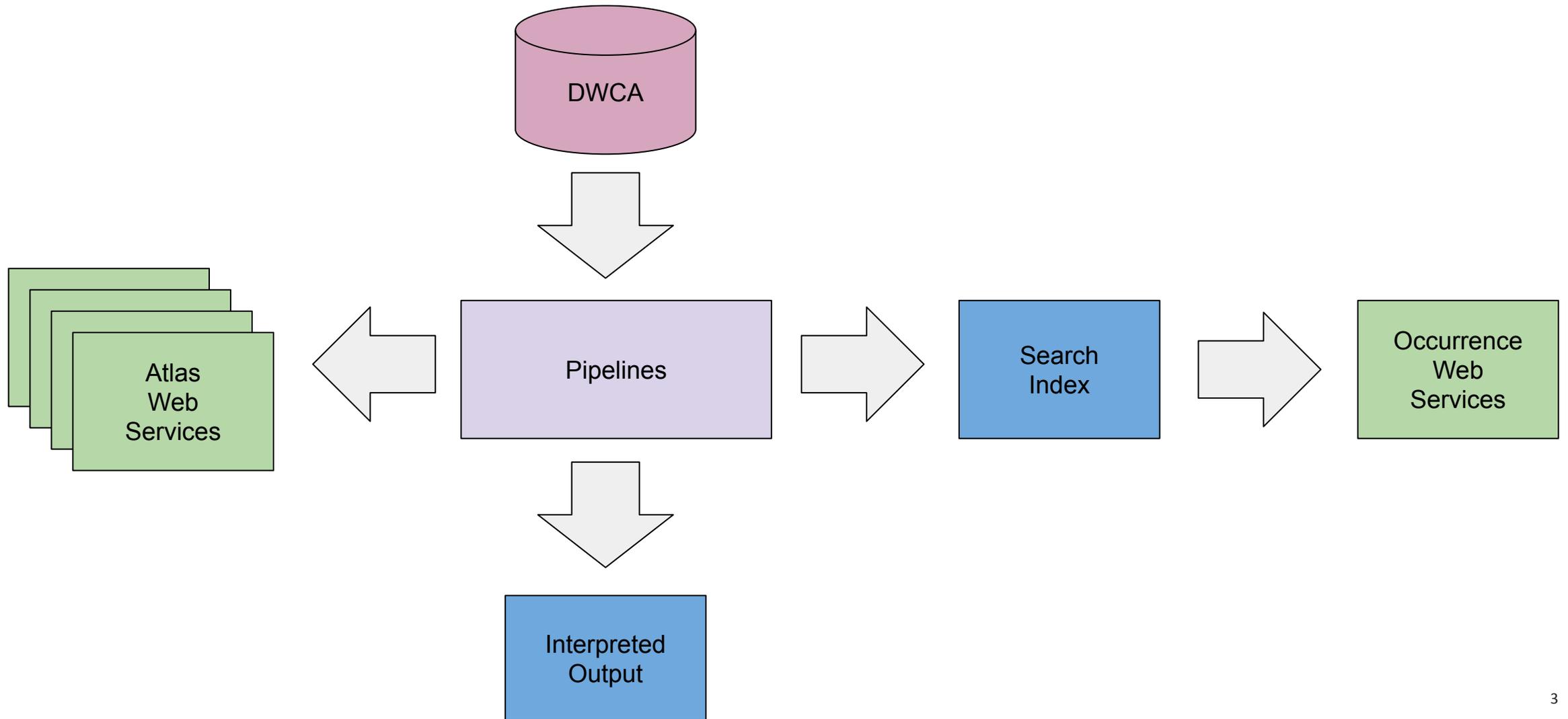


The ALA is made possible by contributions from its many partners. It receives support through the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS) and is hosted by CSIRO.

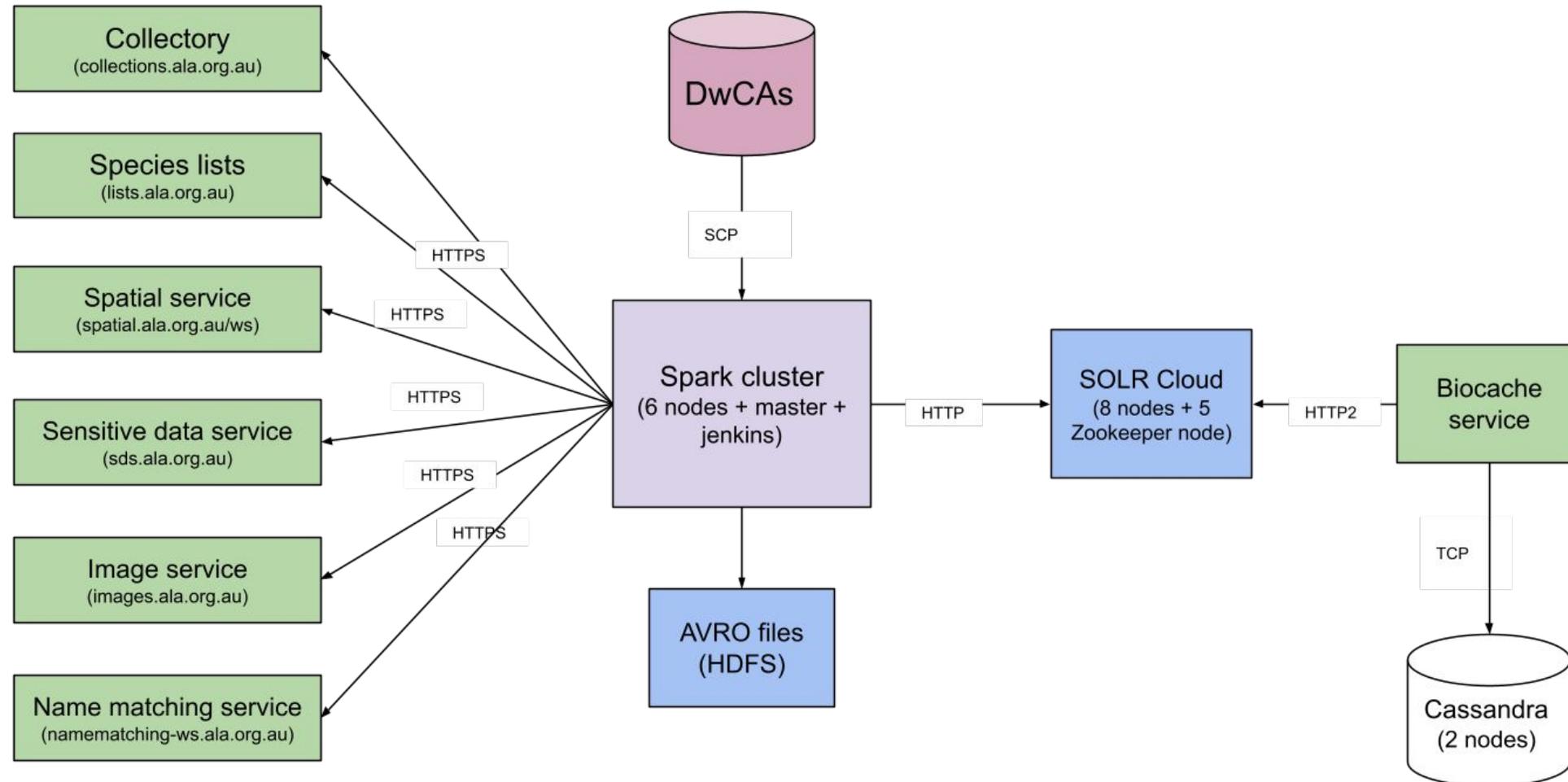
What is Pipelines ?



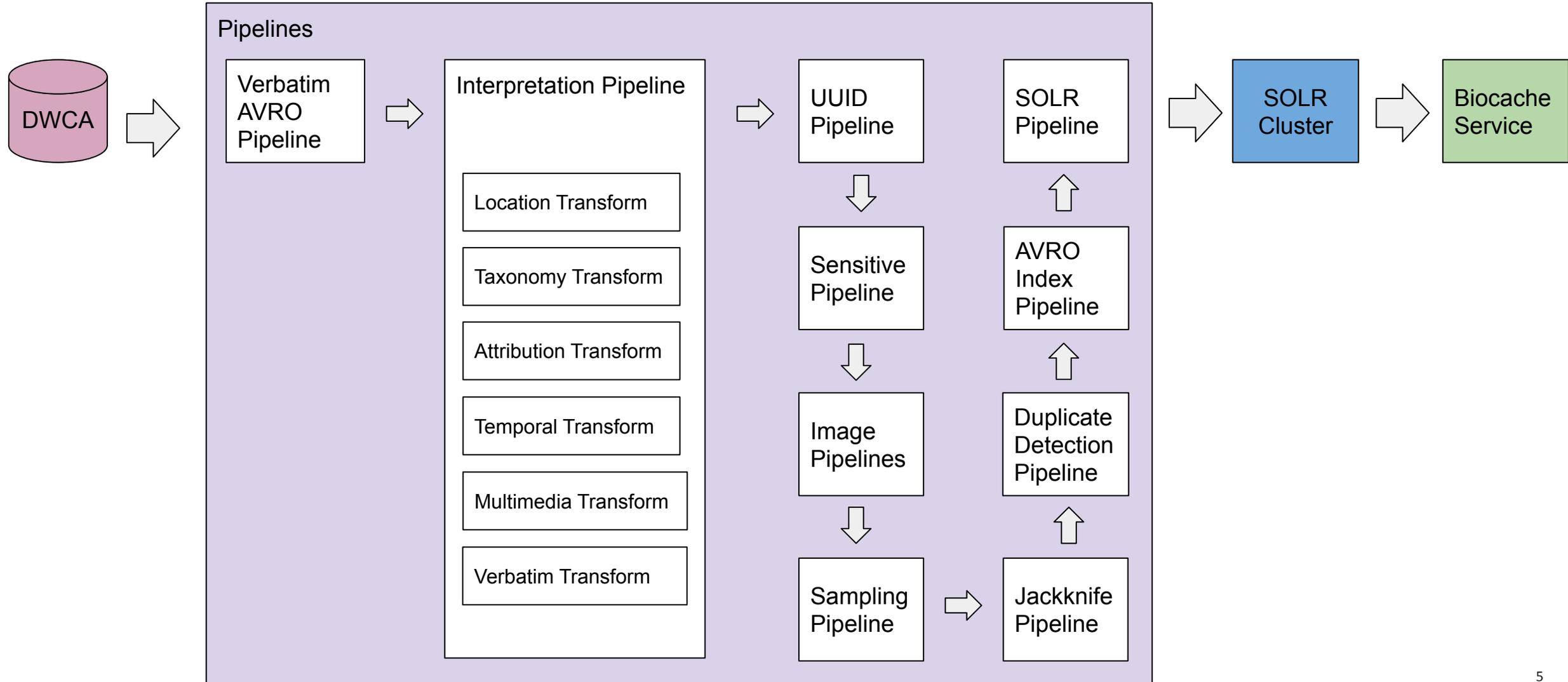
What is Pipelines ?



Architecture



Pipelines - the purple box

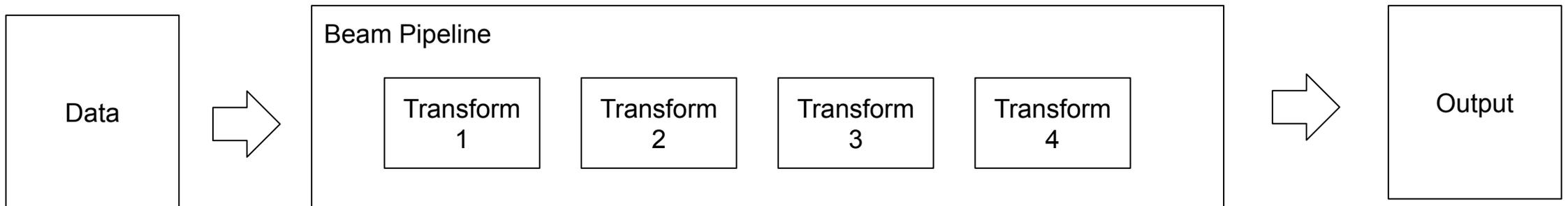


Technology



Apache Beam

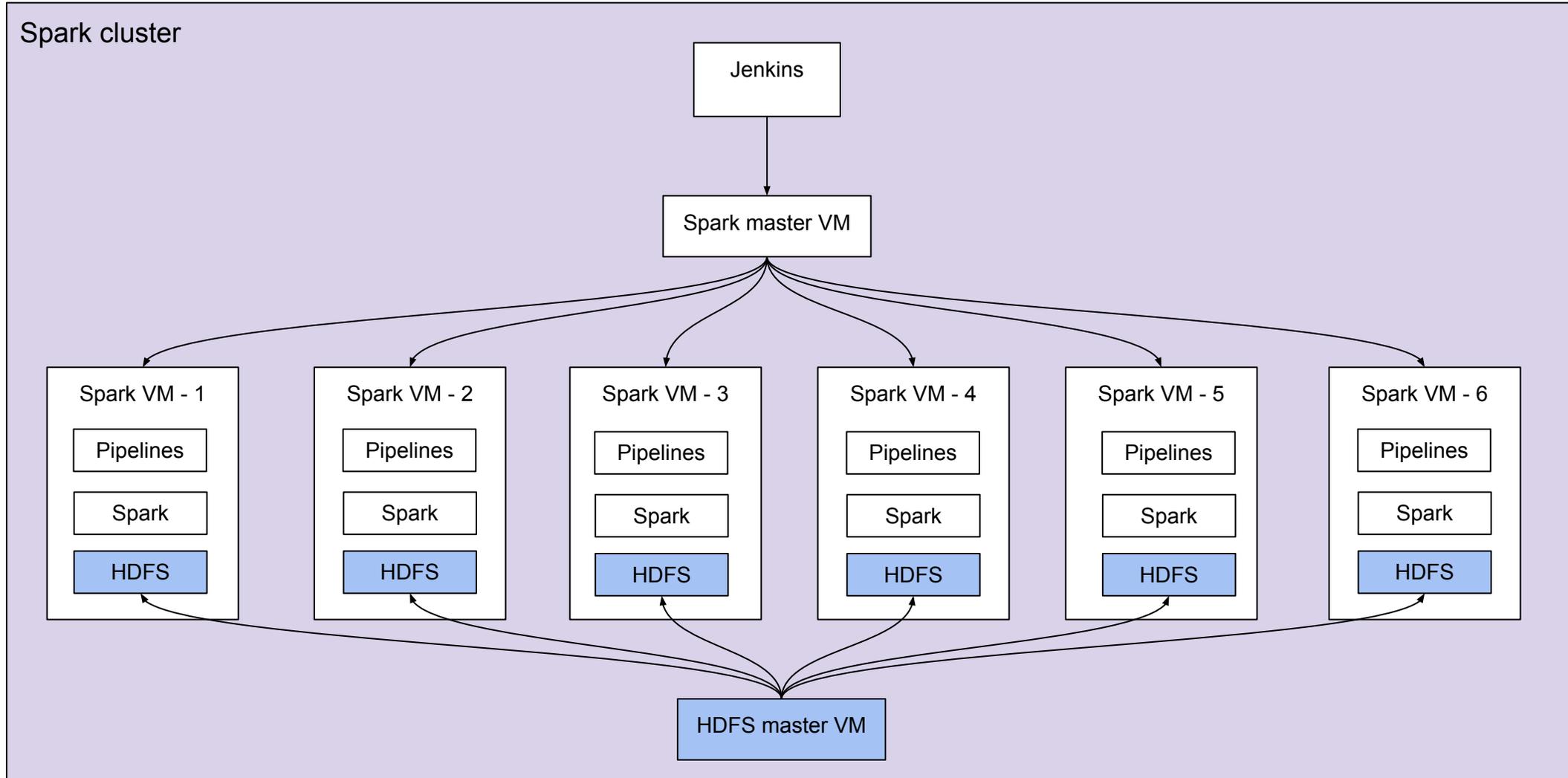
- The **API** used as the basis for Pipelines
 - “Pipelines”
 - “Transforms”
- Platform independent
- Write once - run on different data processing engines
 - Spark
 - Flint
 - Google Cloud Engine



- Data Serialisation
- Binary files with Schema
 - JSON schema
- Language independent
- Widely supported
 - AWS
 - Google Cloud
 - Microsoft Azure



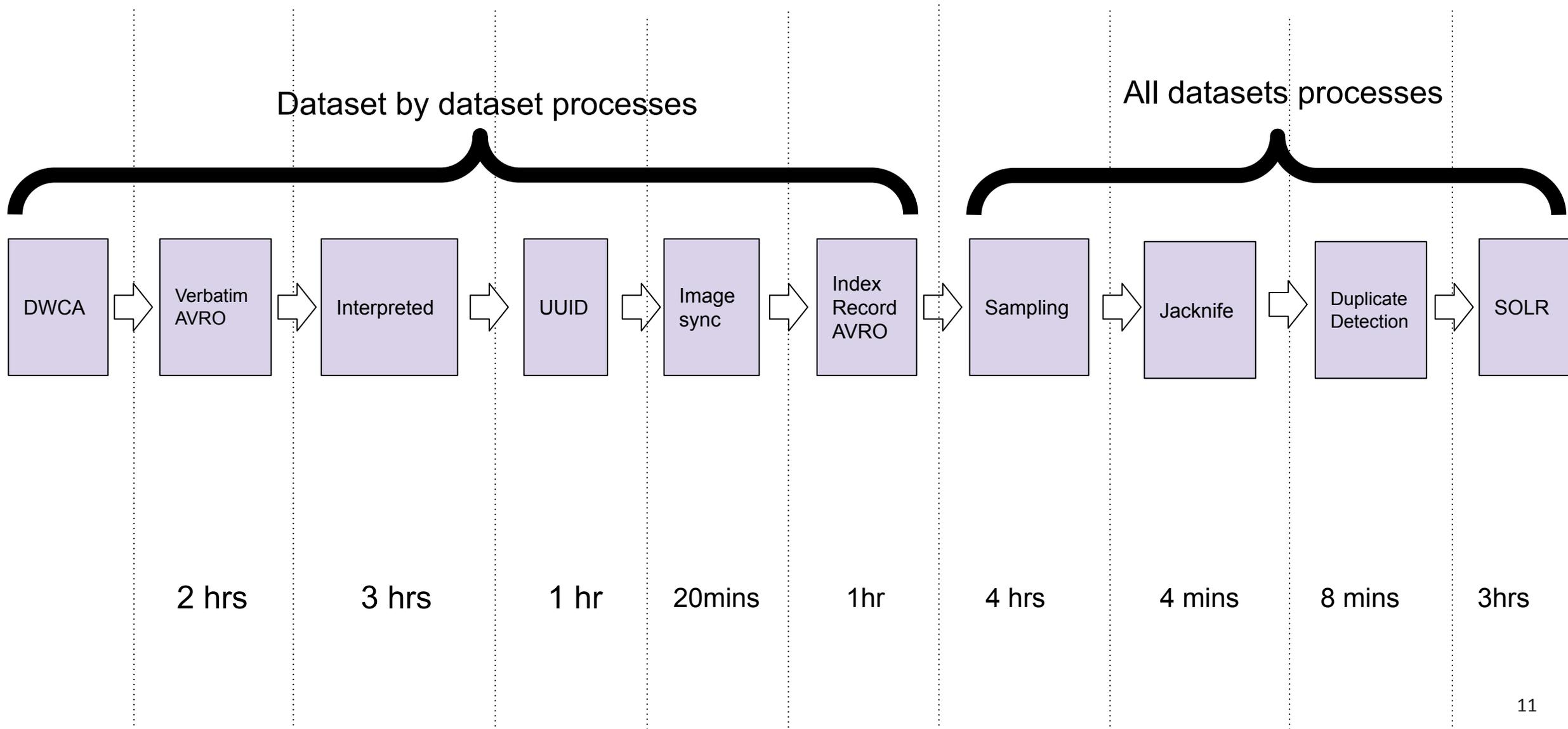
Apache Spark + Hadoop (HDFS) Deployment



Performance

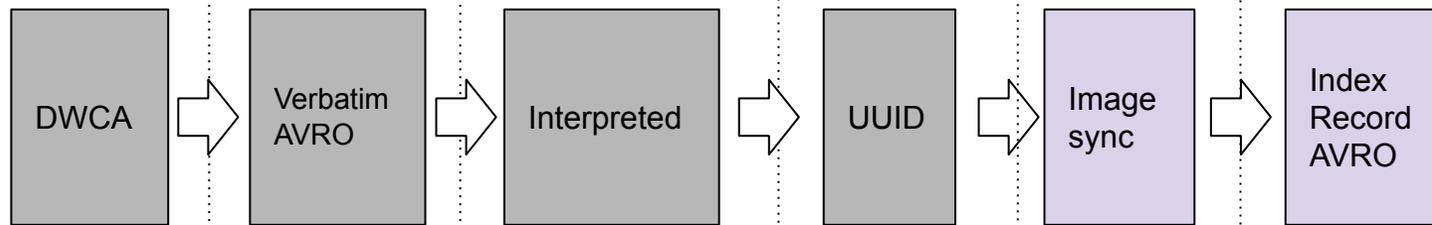


Timings - full rebuild



Timings - daily

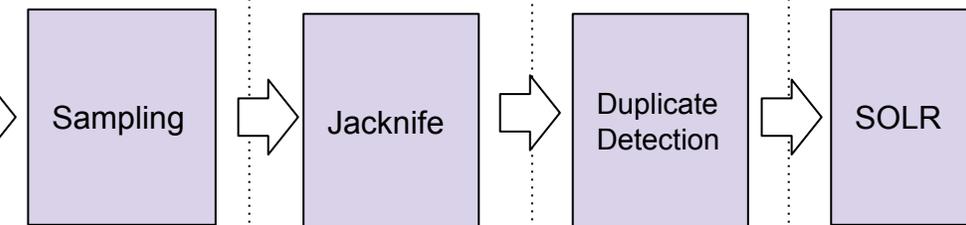
Dataset by dataset processes



20mins

1hr

All datasets processes



20 mins
(incremental
update)

4 mins

8 mins

3hrs

Environments



- 3 environments in ALA
- **Production** - AWS
 - supporting production occurrence services
- **Databox** - AWS - EMR
 - in use by data management team
 - Switched off out of business hours
- **Development** - NCI
 - for development purposes
 - slow...

Developing with Pipelines

- Pipelines run on laptops
 - Embedded Spark
 - Apache Beam DirectRunner - for development purposes
 - Cluster not required for testing
- Issue tracking: <https://github.com/gbif/pipelines/issues>
 - Shared with GBIF team

- Scalable solution
- Faster
 - Daily indexing
- More current data
- Quicker turnaround for large datasets
- Image service - images to S3
- Dockerised services
 - SDS
 - Name matching

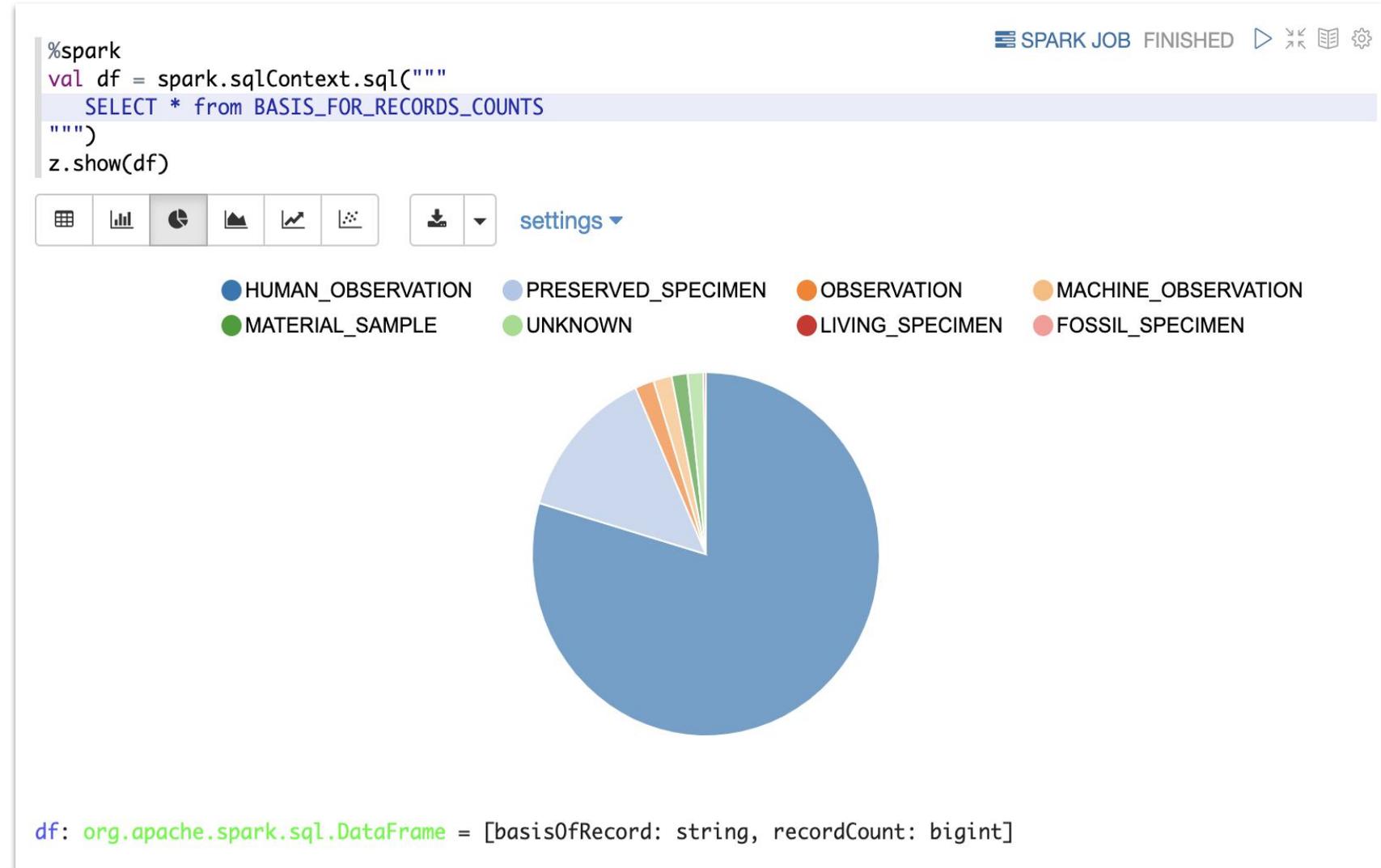
Additional Outcomes

- Deprecated fields/mapping
- SOLR Upgrade
 - From SOLR 6 - SOLR 8
 - DocValues
 - Streaming
 - WMS Upgrade
 - Stability
- Image loading asynchronous
- Documentation on data quality (assertions)

Analytics



- Area for further work
- Open source
 - Jupyter Notebooks
 - Apache Zeppelin
 - Apache Hive
- Commercial
 - Amazon Athena
 - DataBricks

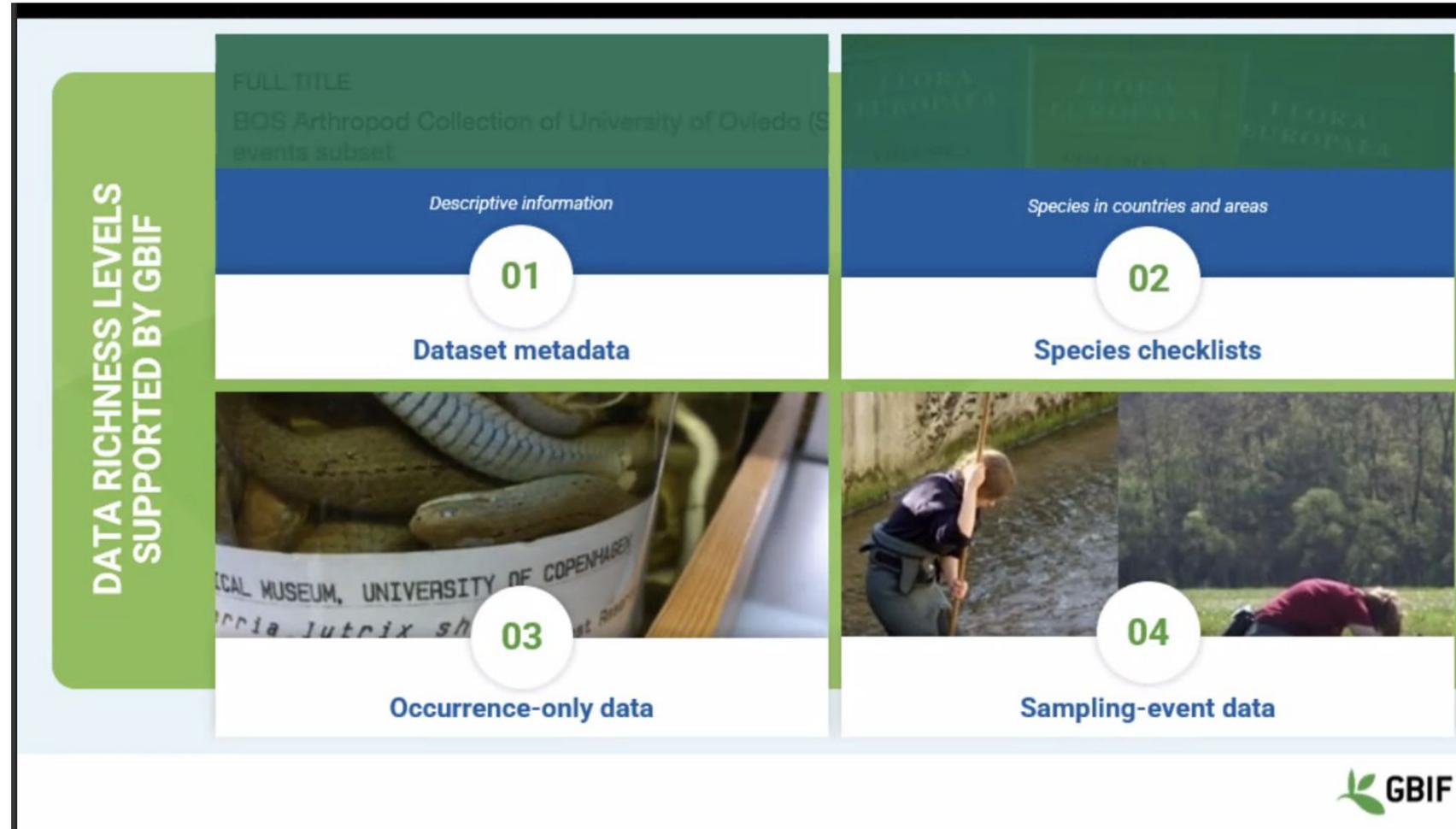


Future directions



Extended data

- GBIF work program to handle sampling/survey data
 - Possible to collaborate



Thank you

Fellow Pipeliners:

**Javier, Peggy, Alex, Matt, Mahmoud, Doug, Bai, Robina,
Simon, Adam, Bruce, Nick, Vicente, Verity, Hamish**