IVOA in the cloud

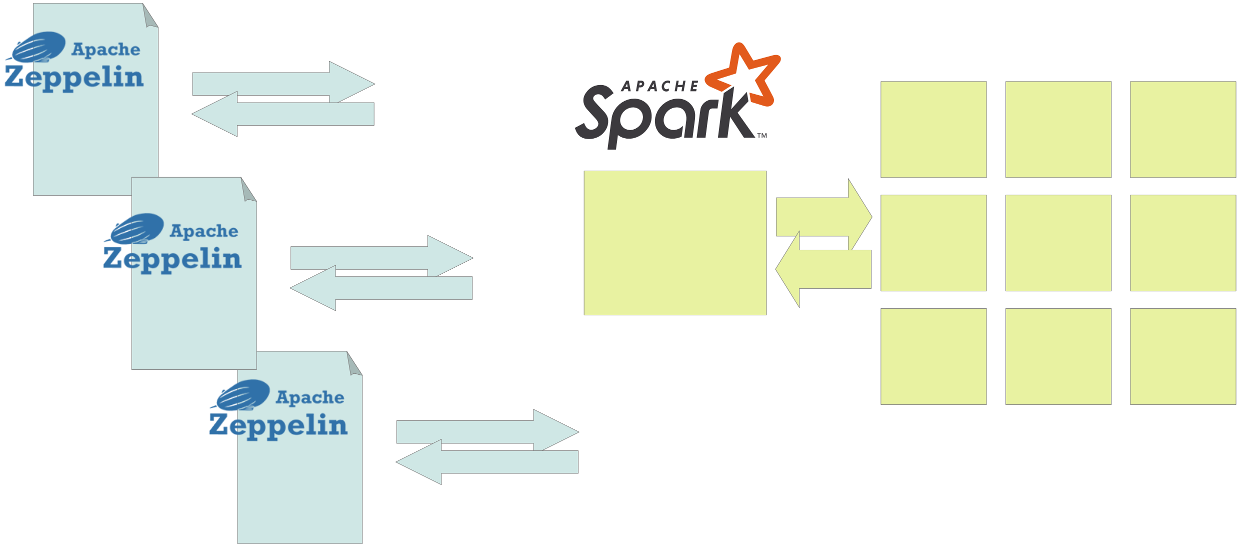
Gaia DataMining platform

D Morris October 2022

D.Morris
Institute for Astronomy,
Edinburgh University

Hadoop/Yarn

- Spark cluster deployed on static resources
- Zeppelin notebooks all interact with the same Spark cluster



- Automated with Ansible

ANSIBLE

99% automated

- create-all
- delete-all

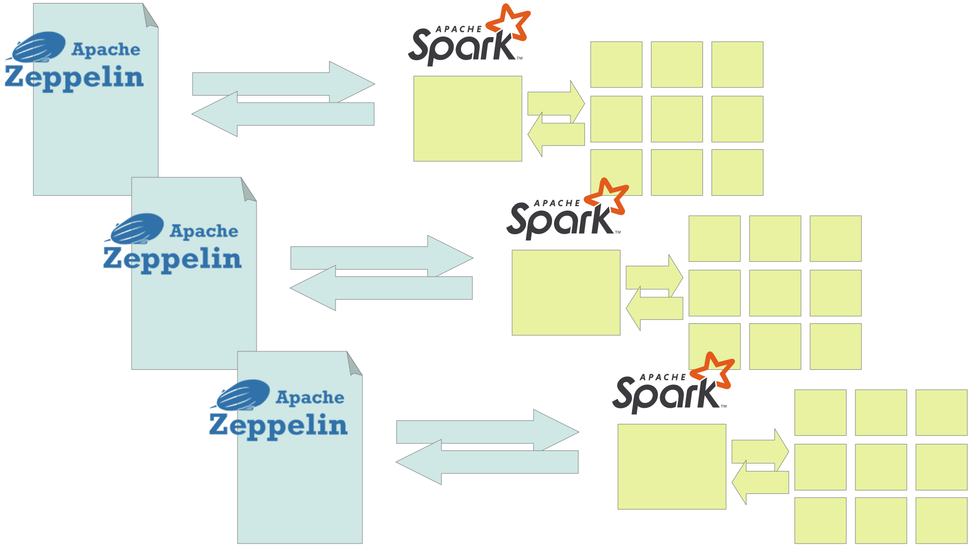
3 deployments

- dev
- test
- live

- Live service working ✓
- Full DR3 dataset

D.Morris
Institute for Astronomy,
Edinburgh University

- Spark cluster on demand
- Notebooks launch their own Spark cluster



- Automated with Helm

99% automated

- create-all
- delete-all

3 deployments

- dev
- test
- live

- In development 2022
- Live deployment 2023

D.Morris
Institute for Astronomy,
Edinburgh University

Gaia DataMining platform
IVOA interop meeting
October 2022

Parquet

https://parquet.apache.org/

Apache Parquet columnar storage format

- Gaia DR3 sources - 561Gbytes
- Gaia DR3 total ~ 5Tbytes

- 2MASS PSC 37G bytes
- 2MASS PSC Gaia DR3 best neighbours 60G bytes

- Pan-STARRS MeanObjectView 270G bytes
- Pan-STARRS Gaia DR3 best neighbours 163G bytes

- ALLWISE 341G bytes
- ALLWISE Gaia DR3 best neighbours 177G bytes

Cross match using best neighbor tables

Familiar SQL based JOIN syntax

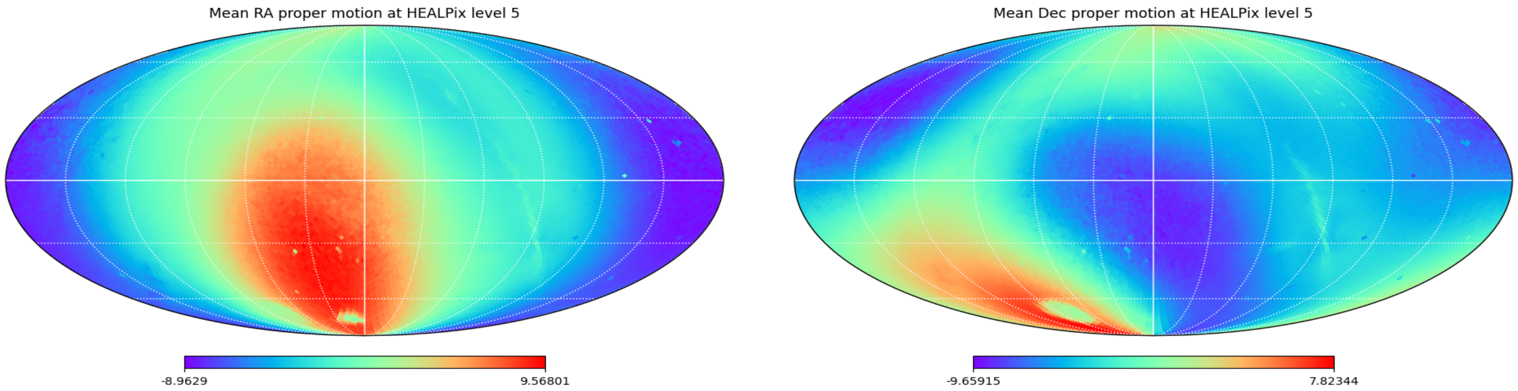
```
SELECT
    gaia.source_id,
    gaia.ra, gaia.dec,
    ps1.g_mean_psf_mag AS ps1_g,
    ps1.r_mean_psf_mag AS ps1_r
FROM
    gaia_source AS gaia
INNER JOIN
    gaia_source_ps1_best_neighbours AS ps1
ON
    gaia.source_id = ps1.source_id
```

HEALPIX partitioning

Parquet files partitioned based on HEALPIX value embedded in Gaia source_id

Placing adjacent sources in the same file reduces shuffle between Spark workers

```
SELECT
    floor(source_id / 562949953421312) AS hpx5,
    COUNT(*) AS n, AVG(pmra), AVG(pmdec)
FROM
    gaia_source
GROUP BY
    hpx5
```



Mean proper motions over the sky – 1min 28sec to calculate and plot

Machine learning application

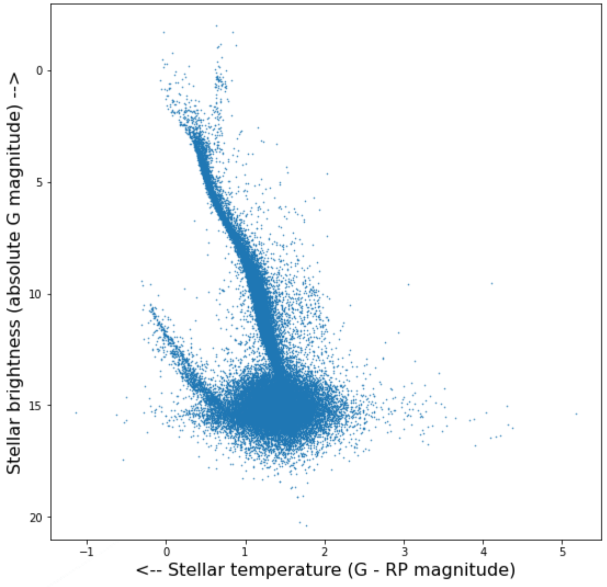
Based on the Gaia EDR3 performance verification *"The Gaia Catalogue of Nearby Stars" (Smart et al. 2021)*.

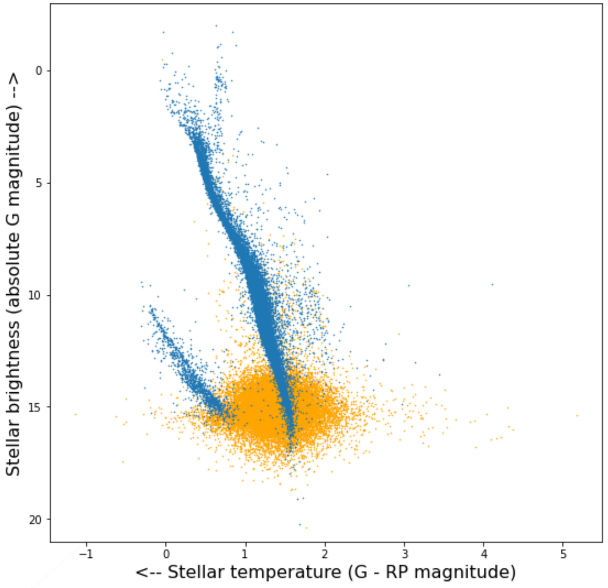
Training a supervised Random Forrest to classify astrometric solutions as ‘good’ or ‘bad’.

SparkSQL queries to generate the training and validation data.

4min to train the classifier

25sec to classify 1,724,028 sources and plot the results





IVOA services and protocols

none so far …..

D.Morris
Institute for Astronomy,
Edinburgh University

Gaia DataMining platform
IVOA interop meeting
October 2022

IVOA services and protocols

Moving the data

Exporting

Sharing our data for others
to use on their platforms

Importing

Using other people’s data
on our platform

Moving the code

Exporting

Executing our code on other
people’s platforms

Importing

Running other people’s
code on our platform

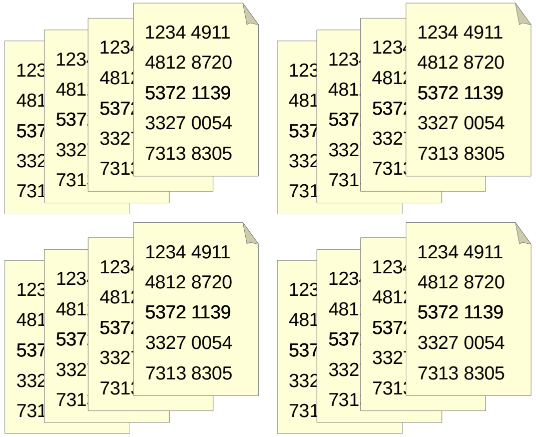
D.Morris
Institute for Astronomy,
Edinburgh University

Gaia DataMining platform
IVOA interop meeting
October 2022

Parquet https://parquet.apache.org/

Apache Parquet columnar storage format

- A table maps to a directory of Parquet files
- Gaia DR3 sources – 561Gbytes, 2048 files



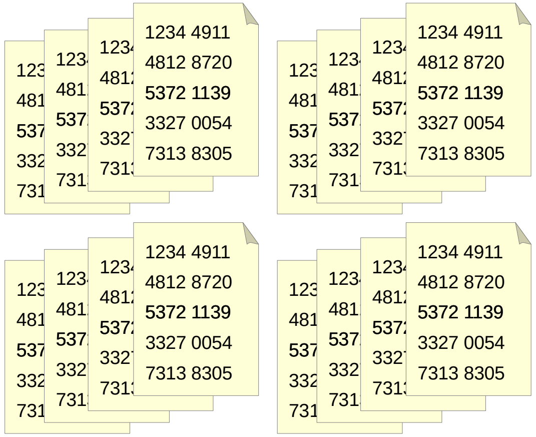
- Technical metadata inside the Parquet files
  - Column names, data types etc
- Science metadata is missing
  - Units, UCDs, DataModels etc

Parquet

https://parquet.apache.org/

Apache Parquet table metadata

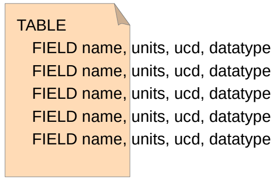
- A table maps to a directory of Parquet files
- Gaia DR3 sources – 561Gbytes, 2048 files



- Technical metadata inside the Parquet files
  - Column names, data types etc

- Science metadata in a VOTable
  - Units, UCDs, DataModels etc

table-metadata.vot



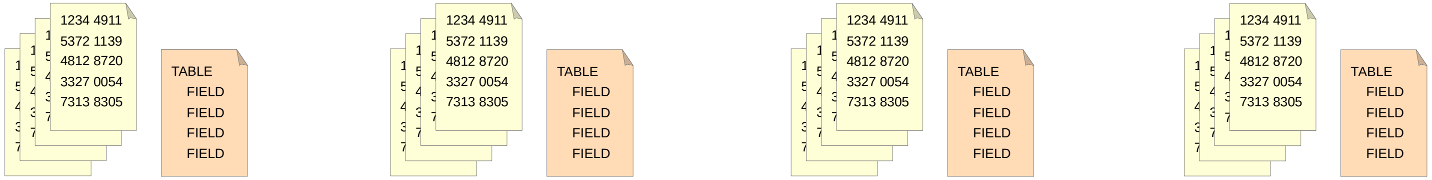
```
TABLE
  FIELD name, units, ucd, datatype
  FIELD name, units, ucd, datatype
  FIELD name, units, ucd, datatype
  FIELD name, units, ucd, datatype
  FIELD name, units, ucd, datatype
```

Parquet

https://parquet.apache.org/

Apache Parquet catalog

- A table maps to a ~~directory~~ *bucket* of Parquet files
- A catalog maps to a set of ~~directories~~ *buckets*.



Can we ~~describe~~ *register* this in a similar way to the way we describe a TAP service ?

The overall catalog has metadata like publisher, waveband, footprint etc.,

Catalog has a schema that contains tables (buckets).

The schema tables have fields (columns).

The *'service'* has an access protocol (s3).

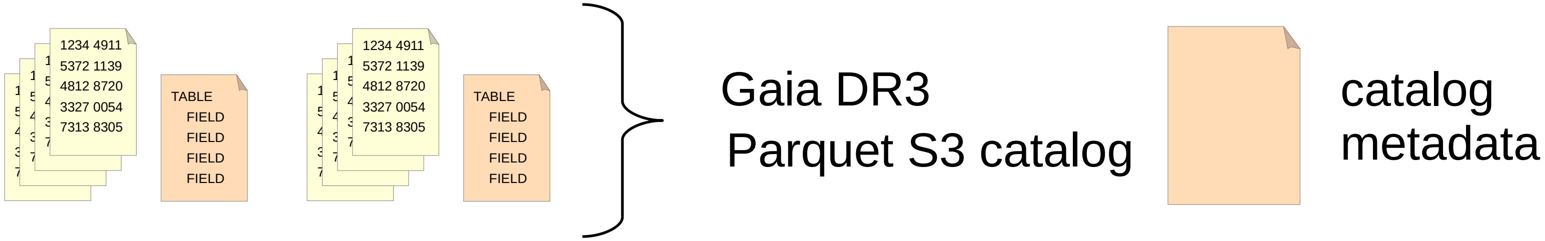
IVOA services and protocols

Move the data

‘everyone’ in data science uses Parquet and S3
use what is already there and build on it

Exporting

Sharing our data with others

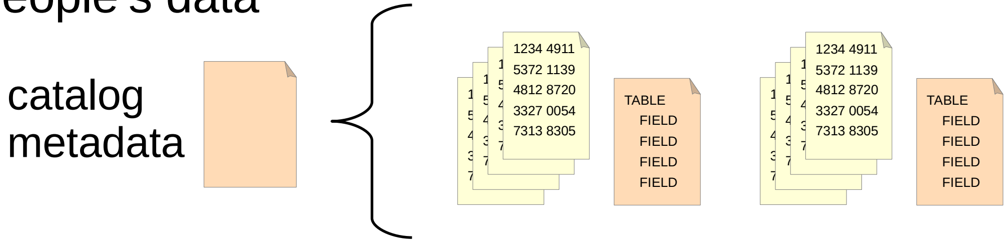


Gaia DR3
Parquet S3 catalog

catalog
metadata

Importing

Using other people’s data



catalog
metadata

D.Morris
Institute for Astronomy,
Edinburgh University

Gaia DataMining platform
IVOA interop meeting
October 2022

IVOA services and protocols

Move the code

- Exporting
  - Run our analysis on other platforms
- Importing
  - Run other people’s code on our platform

IVOA services and protocols

Move the code

Execution Planner

Will my code run on your platform ?

Metadata schema to describe a task and the resources it needs

Zeppelin notebook
PySpark analysis
210 cpu cores
360G memory
1Tbyte disc

When can I run my code on your platform ?

Scheduling service to book resources

2 hours
15:00 – 17:00
Tuesday 18th

Gaia DataMining platform

IVOA cloud services and protocols

Parquet/S3 catalogs

Moving the data

Exporting

Sharing our data for others to use on their platforms

Importing

Using other people’s data on our platform

Execution Planner

Moving the code

Exporting

Executing our code on other people’s platforms

Importing

Running other people’s code on our platform

D.Morris
Institute for Astronomy,
Edinburgh University

Gaia DataMining platform
IVOA interop meeting
October 2022

Questions and comments

Dave Morris
dmr@roe.ac.uk

Institute for Astronomy
Edinburgh University