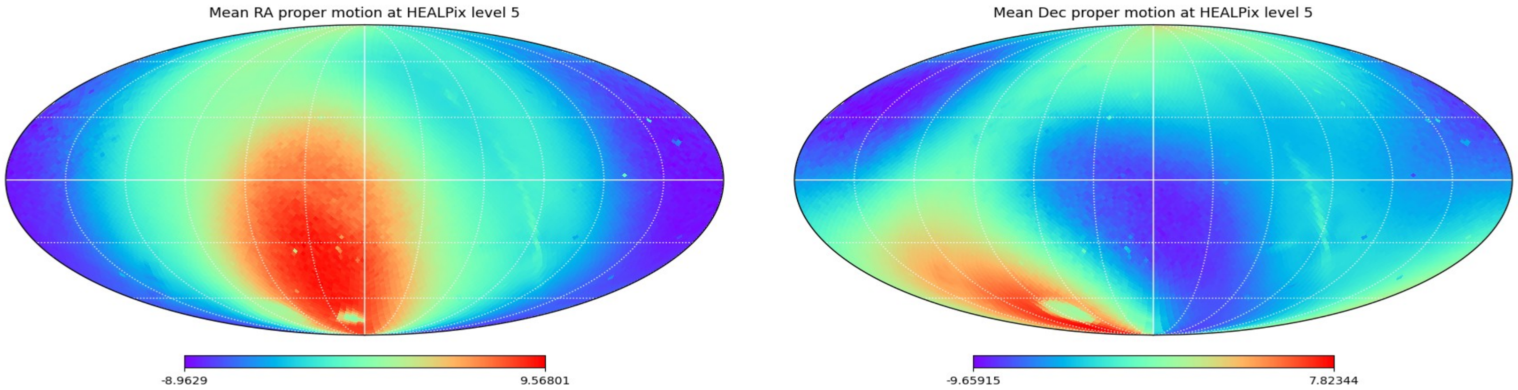
Gaia DataMining platform

D Morris May 2023

Data mining analysis platform for Gaia data

Analysis on the whole dataset – Wide Field Astronomy Unit (WFAU)

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



Mean RA proper motion at HEALPix level 5

Mean Dec proper motion at HEALPix level 5

Mean proper motions over the sky – 50 seconds to calculate and plot

D.Morris
Institute for Astronomy,
Edinburgh University

Machine learning applications

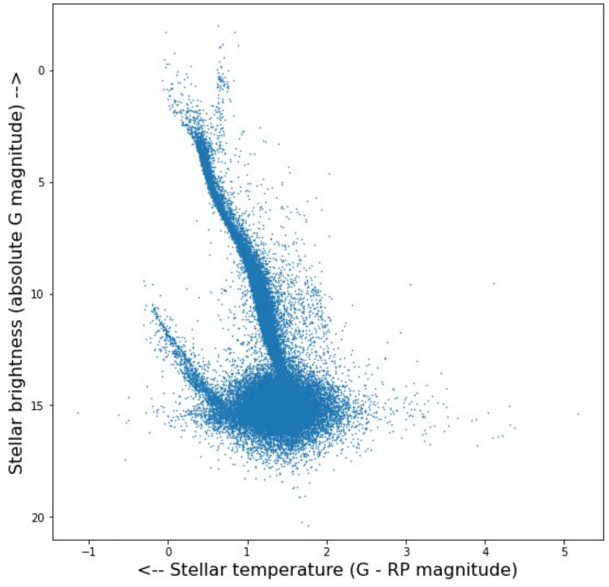
Based on the Gaia EDR3 performance verification *"The Gaia Catalogue of Nearby Stars" (Smart et al. 2021)*.

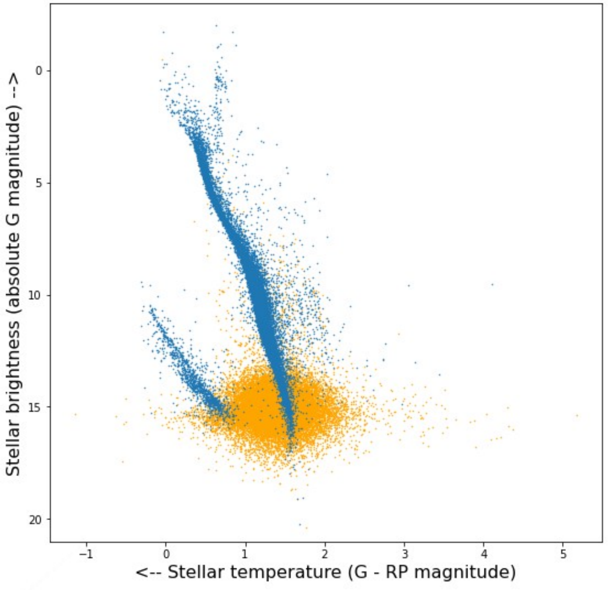
Training a supervised Random Forrest to classify astrometric solutions as ‘good’ or ‘bad’.

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



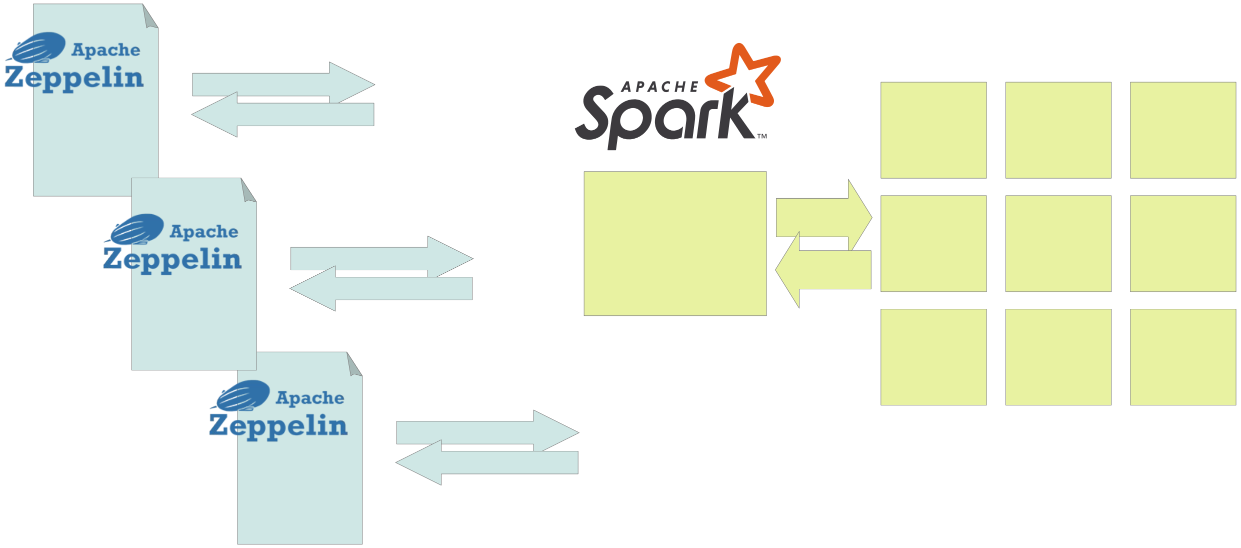


D.Morris
Institute for Astronomy,
Edinburgh University

Hadoop/Yarn

Current deployment – shared Spark cluster

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



- Automated with Ansible

ANSIBLE

99% automated

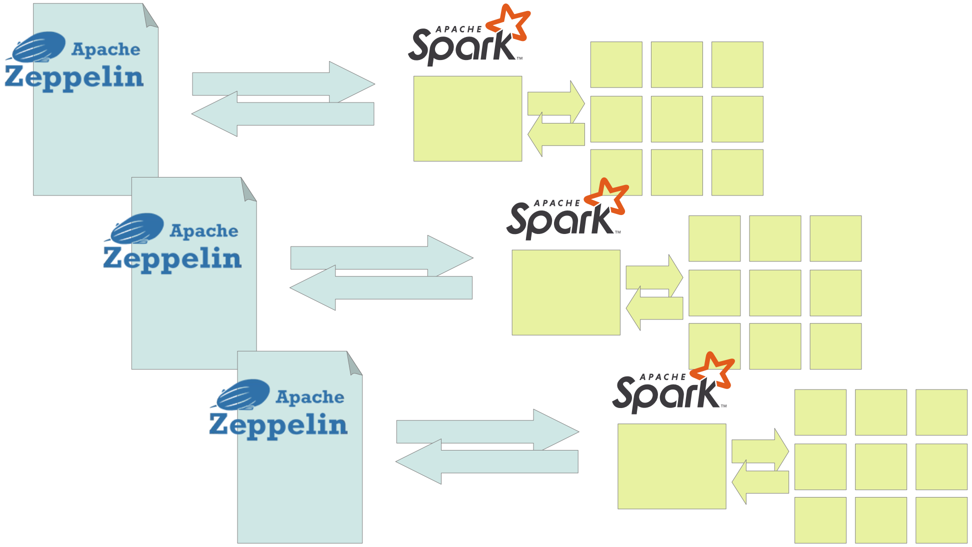
- create-all
- delete-all

- Live service working
- Full DR3 dataset

✓

In development – on demand deployment

- Notebook environment on demand
- Spark cluster on demand



- Automated with Helm

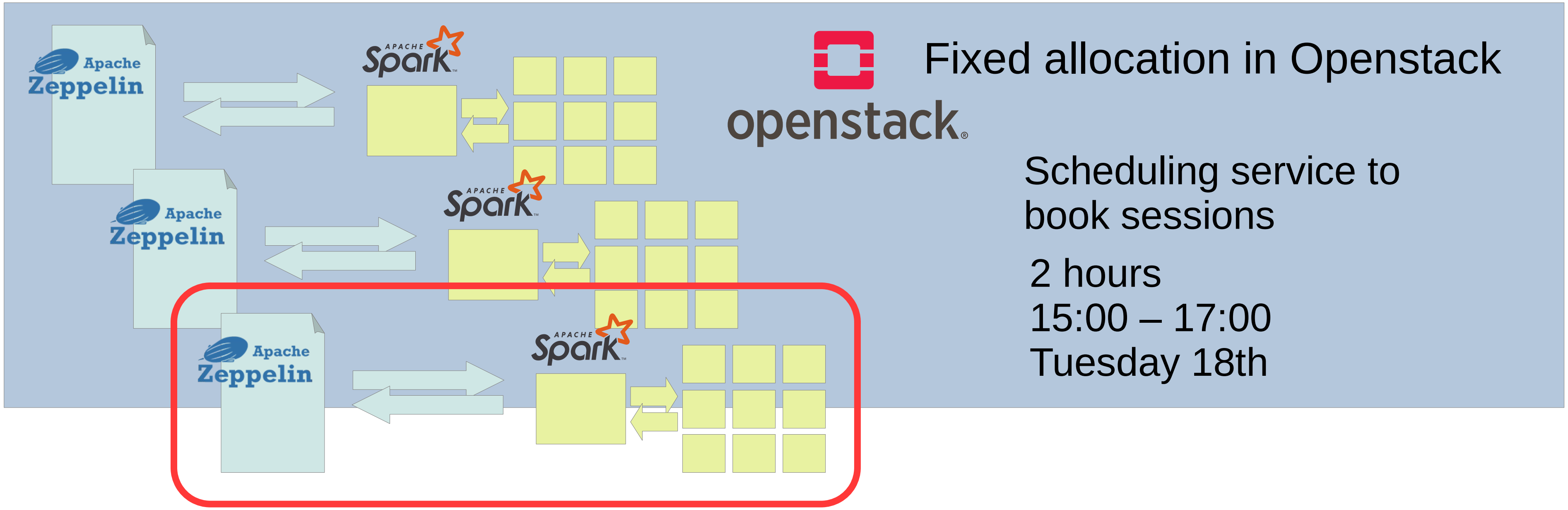
100% automated

- create-all
- delete-all

D.Morris
Institute for Astronomy,
Edinburgh University

Dynamic deployment on a fixed cloud

- Notebook environment on demand
- Spark cluster on demand



Fixed allocation in Openstack

Scheduling service to book sessions

2 hours
15:00 – 17:00
Tuesday 18th

IVOA 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

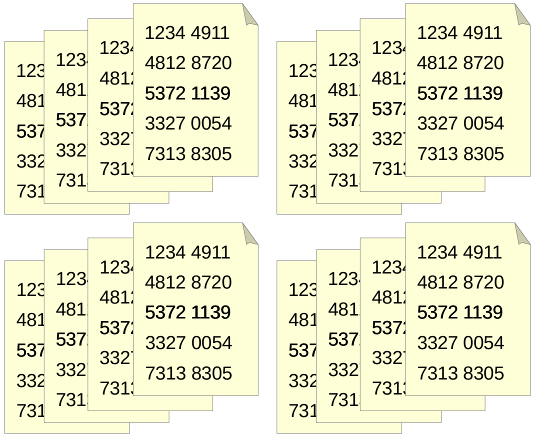
D.Morris
Institute for Astronomy,
Edinburgh University

Parquet

https://parquet.apache.org/

Apache Parquet columnar storage format

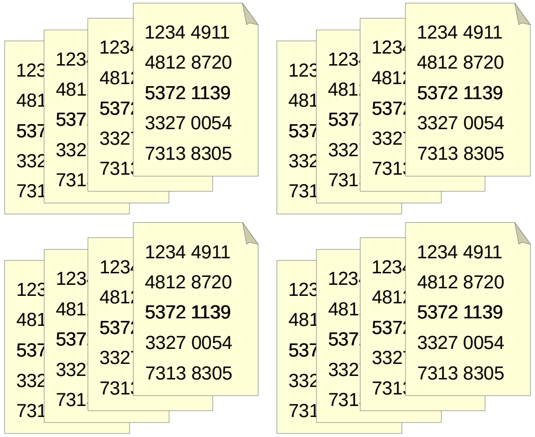
- A table maps to a directory of Parquet files
- Gaia DR3 sources – 561Gbytes
- Partitioned as 2048 files per table
- Indexed based on Gaia source id (HEALPix)
- Technical metadata inside the Parquet files
  - Column names, data types etc
- Science metadata is missing
  - Units, UCDs, DataModels etc



Parquet https://parquet.apache.org/

Apache Parquet columnar storage format

- A table maps to a directory of Parquet files
- Gaia DR3 sources – 561Gbytes
- Partitioned as 2048 files per table
- Indexed based on Gaia source id (HEALPix)
- Technical metadata inside the Parquet files
  - Column names, data types etc
- Science metadata in a VOTable header
  - Units, UCDs, DataModels etc



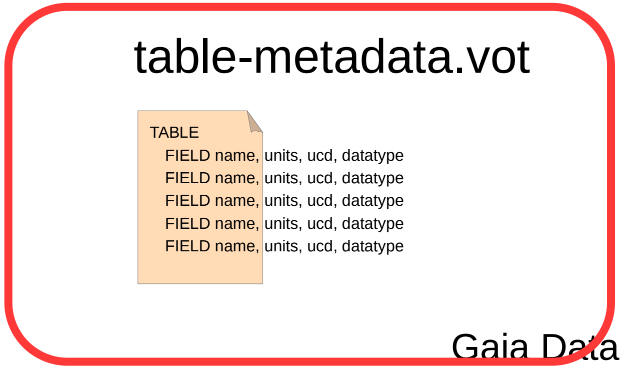


table-metadata.vot

```
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
```

Making our data accessible to others

Everyone uses S3,
because it's easy … right?

- A table maps to a ~~directory~~ *bucket* of Parquet files
- Bucket names have to be unique within the S3 service
- Ceph S3 service providing Peta bytes of storage for the **whole country**

• GAIA_SOURCE	table name
• GDR3_GAIA_SOURCE	+ data release
• GaiaDMp-GDR3_GAIA_SOURCE	+ project brand

- Globally unique within the S3 service, but less 'findable' for users

Making our data accessible to others

Everyone uses S3,
because it’s easy … right?

- S3 URL specifies the bucket name and object name
  - s3://{bucket}/{object}
  - s3://GaiaDMp-GDR3_GAIA_SOURCE/part-00749…..parquet

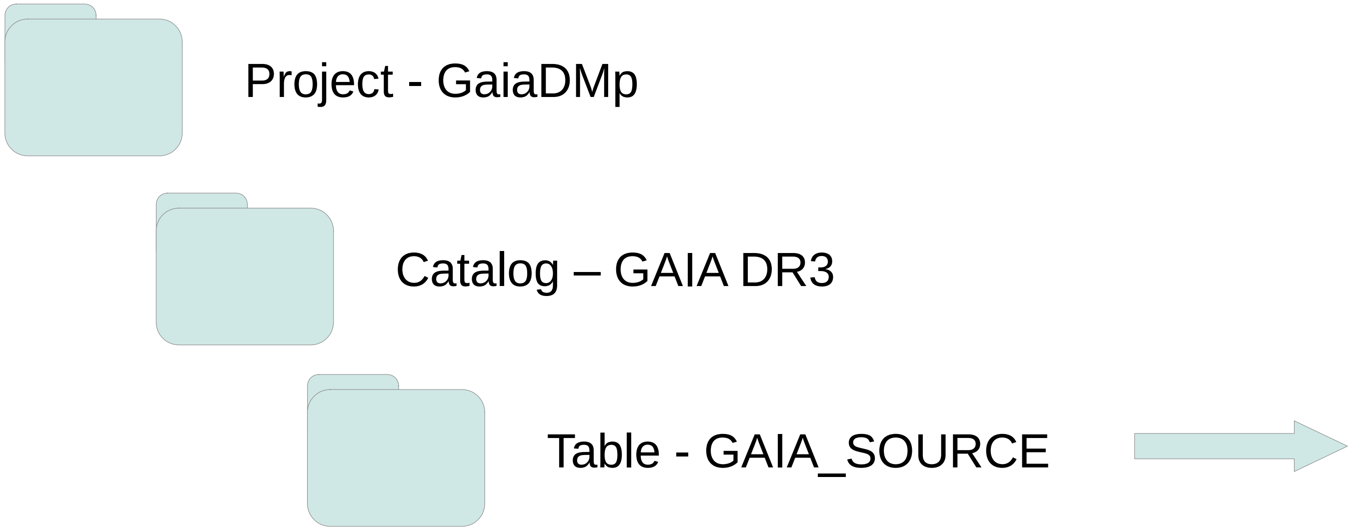
- S3 URL does **not** specify :
  - The hostname “`s3.echo.stfc.ac.uk`”
  - The URL template “`s3.echo.stfc.ac.uk/%(bucket)`”
  - A flag to use HTTPS “`public_url_use_https=true`”

Making our data accessible to others

IVOA VOSpace

Everyone uses S3,
because it's easy … right?

- The data is still stored in S3, VOSpace provides the directory structure and metadata



Project - GaiaDMp

Catalog – GAIA DR3

Table - GAIA_SOURCE

s3://{bucket}/

Making our data accessible to others

Everyone uses S3,
because it’s easy … right?

- VOSpace directories can include metadata about each level

Project - GaiaDMp
- Publisher metadata

Catalog – GAIA DR3
- Catalog footprint
- Catalog DOI

Table - GAIA_SOURCE
- TAP_SCHEMA with JOINs
- Column metadata

Making our data accessible to others

Everyone uses S3,
because it's easy … right?

- VOSpace can provide access using more than one protocol

- The parameters for S3 can include all the details needed to access the data:

  - The S3 URL "`s3://GaiaDMp-GDR3_GAIA_SOURCE/part-00749…..parquet`"
  - The hostname "`s3.echo.stfc.ac.uk`"
  - The URL format "`s3.echo.stfc.ac.uk/%(bucket)`"
  - A flag to use HTTPS "`public_url_use_https=true`"

IVOA wishlist

- Data descriptions
  - Gaia DR3 in parquet
- Data locations
  - Arcus HPC at Cambridge

- Software descriptions
  - Apache Spark cluster
  - Apache Zeppelin notebooks

- Software capabilities and data proximity
  - Apache Spark cluster
  - with fast access to
  - Gaia DR3 in parquet

Questions and comments

Dave Morris
dmr@roe.ac.uk

Institute for Astronomy
Edinburgh University