Gaia Spark analysis platform

Static read only dataset (CephFS) (Echo S3)

- DR2 : 10^3 parquet files, 473Gbytes
- early DR3 : 10^3 parquet files, 533Gbytes
- DR3 : x10 2022 ~30Tbyte
- DR4 : x10 2024 ~300Tbyte

- Fast random access temp data (ephemeral disc)
- Many files (10^4) random access temp data (Cinder volumes)

- Fast random access temp data (ephemeral disc)
- Many files (10^4) random access temp data (Cinder volumes)
- Machine learning – scans the whole dataset every time
- JOIN partitioning – 10^4 output files (HDFS on Cinder)

- Interactive interface : <30 sec to start first notebook
- Variable demand : 1 user today 20 users tomorrow (needs space to scale up)
- Workshop bookings : everyone is running the same example (predictable)
- Scale out to another cloud (common base platform)

Deployment stack:

(not terraform)

D.Morris
Institute for Astronomy,
Edinburgh University

Somerville IfA requirements
StackHPC (Google Meet)
25th March 2021