Gaia Spark analysis platform

Technology choices

D.Morris
Institute for Astronomy,
Edinburgh University

Cloud interest group
Edinburgh
4th March 2021

Original GDAF system

Zeppelin & Spark deployed on physical hardware

Based in part on Cloudera deployment

Initial goal for our project

- replicate GDAF in the cloud

Subsequent goals

- additional analysis tools
- scalable deployment
- more users
- more data
  - DR3
  - DR4

Gaia DPAC Data Processing & Analysis Consortium

Gaia Data Analytics Framework (GDAF) description

D.Morris
Institute for Astronomy,
Edinburgh University

Cloud interest group
Edinburgh
4th March 2021

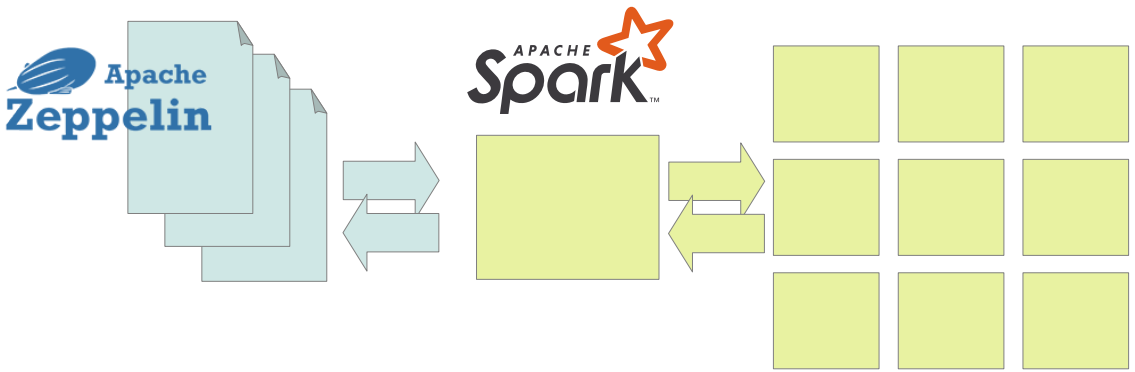
Technology choice #1
Hadoop/Yarn scheduler

“known unknowns”

Hadoop/Yarn

- GDAF system – working example
- Cloudera deployment - lots of documentation
- Standard deployment - lots of blogs and howtos

- Virtual machine based
- Spark cluster deployed on a static set of resources
- Zeppelin notebooks all interact with the same Spark cluster



D.Morris
Institute for Astronomy,
Edinburgh University

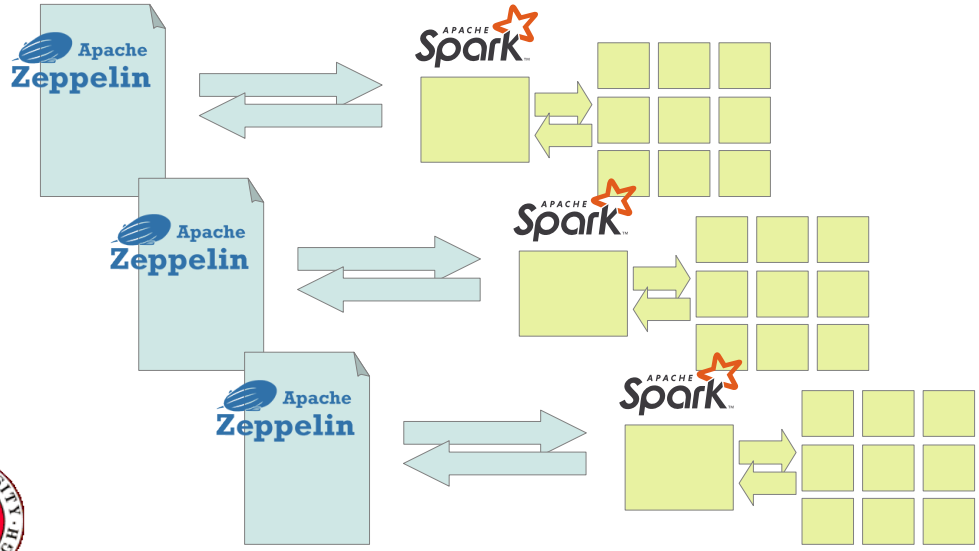
Cloud interest group
Edinburgh
4th March 2021

Technology choice #2
Kubernetes scheduler

“unknown unknowns”

- Experimental in 2020
- Zeppelin pre-release only
- Not recommended for production

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



- Kubernetes container based
- Standard technology in 2021/2022

D.Morris
Institute for Astronomy,
Edinburgh University

Cloud interest group
Edinburgh
4th March 2021

Technology choices
Spark deployments

Developing both systems in parallel

Hadoop/Yarn

- Up and running early
- Spark and Zeppelin by Nov 2019
- Gaia DR2 parquet data by Dec 2019

- Automated with Ansible

Working system to develop science cases

- Python libraries
- AXS extension
- Parquet partitions

kubernetes

- Development platform for Spark and Zeppelin on Kubernetes

- Automated with Helm

Technology experiments

- Openstack Magnum & Manila
- Kubernetes Helm
- Terraform

- OAuth login
- Drupal CMS

D.Morris
Institute for Astronomy,
Edinburgh University

Cloud interest group
Edinburgh
4th March 2021

Technology choices deployment tools

Working towards full automation

openstack

- Good documentation
- Simple client interface
- Machine readable results

ANSIBLE

- Well documented
- Well supported on StackOverflow
- Lots of plugins

HELM kubernetes

- De facto standard for Kubernetes deployments
- Good version control for dependencies

Terraform

- Didn’t live up to the hype
- Relies on local state
- Clear text secrets

D.Morris
Institute for Astronomy,
Edinburgh University