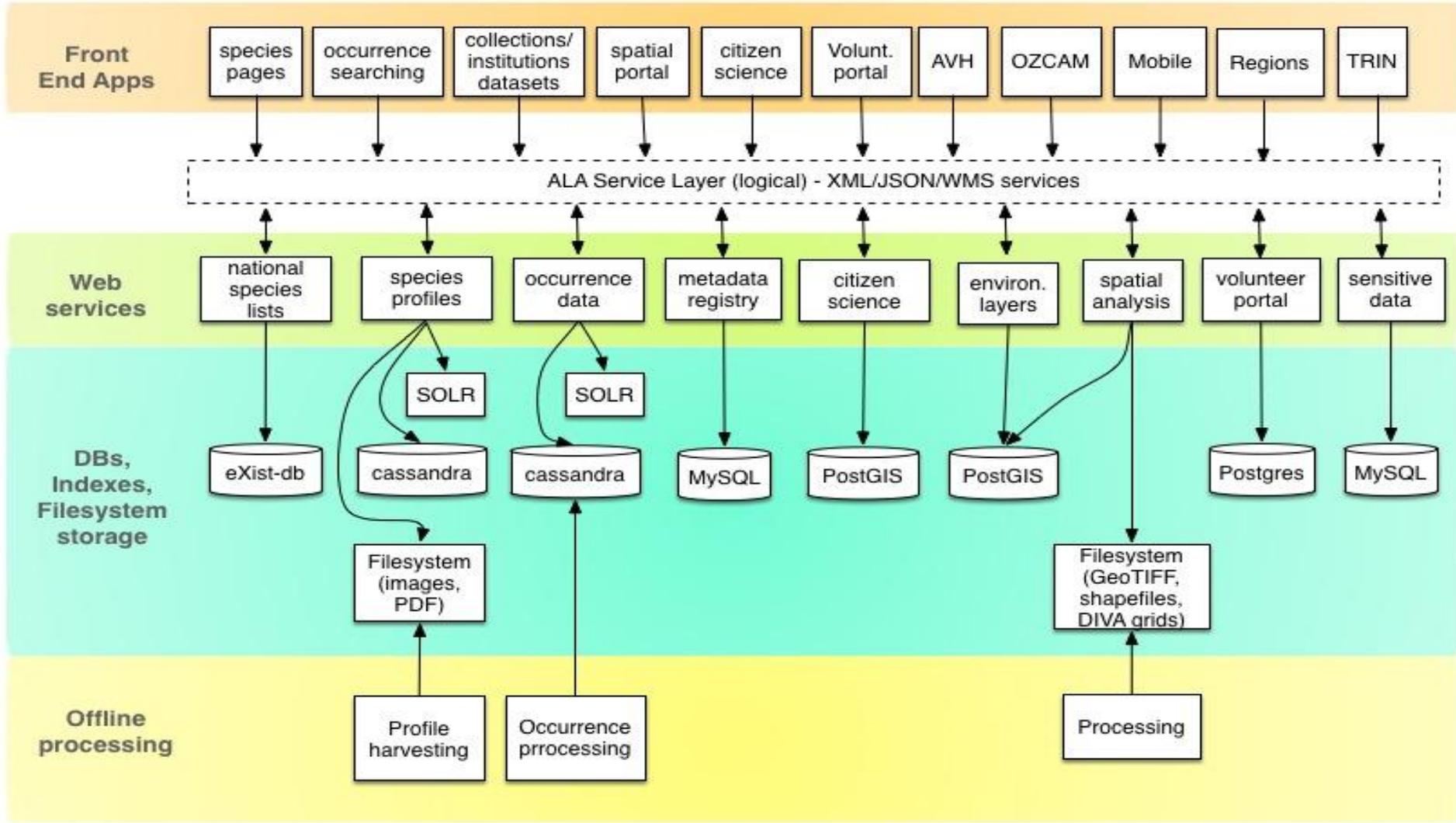# Architecture

Paris workshop - May 2015
Dave Martin

Agenda

- Architectural overview
- Collectory
- Biocache
- Hubs
- Logger
- Cassandra
- SOLR and Lucene
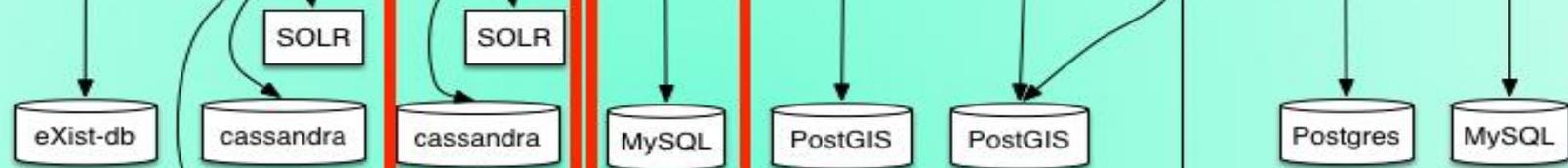
Front End Apps	species pages	occurrence searching	collections/ institutions datasets	spatial portal	citizen science	Volunt. portal	AVH	OZCAM	Mobile	Regions	TRIN

ALA Service Layer (logical) - XML/JSON/WMS services

Web services: national species lists | species profiles | occurrence data | metadata registry | citizen science | environ. layers | spatial analysis | volunteer portal | sensitive data

SOLR — SOLR

DBs, Indexes, Filesystem storage: eXist-db | cassandra | cassandra | MySQL | PostGIS | PostGIS | Postgres | MySQL

Filesystem (images, PDF)

Filesystem (GeoTIFF, shapefiles, DIVA grids)

Offline processing: Profile harvesting | Occurrence prrocessing | Processing

Front End Apps	species pages	occurrence searching	collections/ institutions datasets	spatial portal	citizen science	Volunt. portal	AVH	OZCAM	Mobile	Regions	TRIN

ALA Service Layer (logical) - XML/JSON/WMS services

Web services	national species lists	species profiles	occurrence data	metadata registry	citizen science	environ. layers	spatial analysis	volunteer portal	sensitive data

SOLR

SOLR

DBs, Indexes, Filesystem storage

eXist-db cassandra cassandra MySQL PostGIS PostGIS Postgres MySQL

Filesystem (images, PDF)

Filesystem (GeoTIFF, shapefiles, DIVA grids)

Offline processing
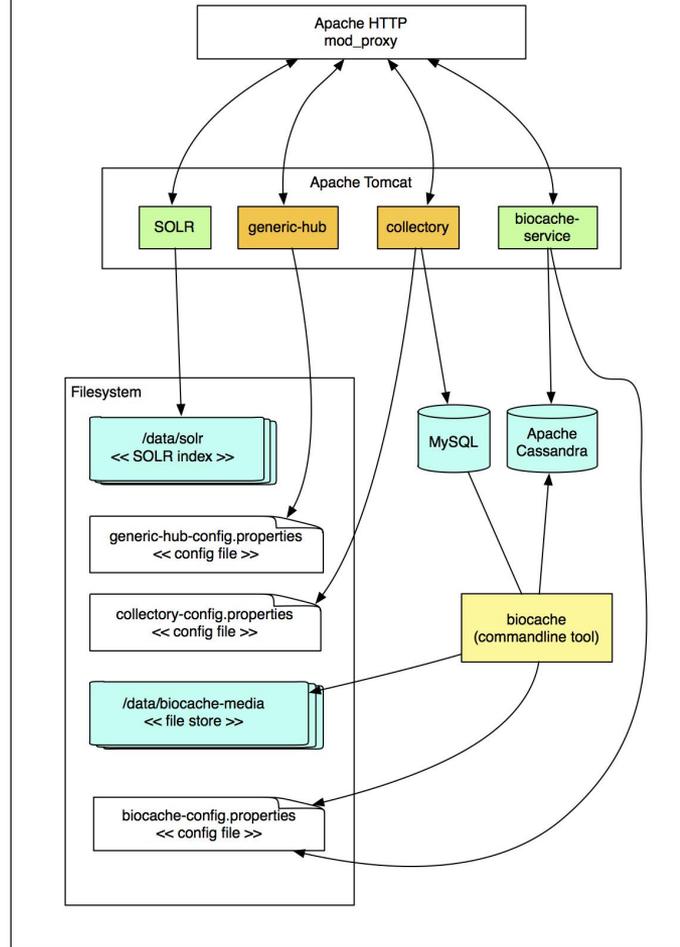
Profile harvesting

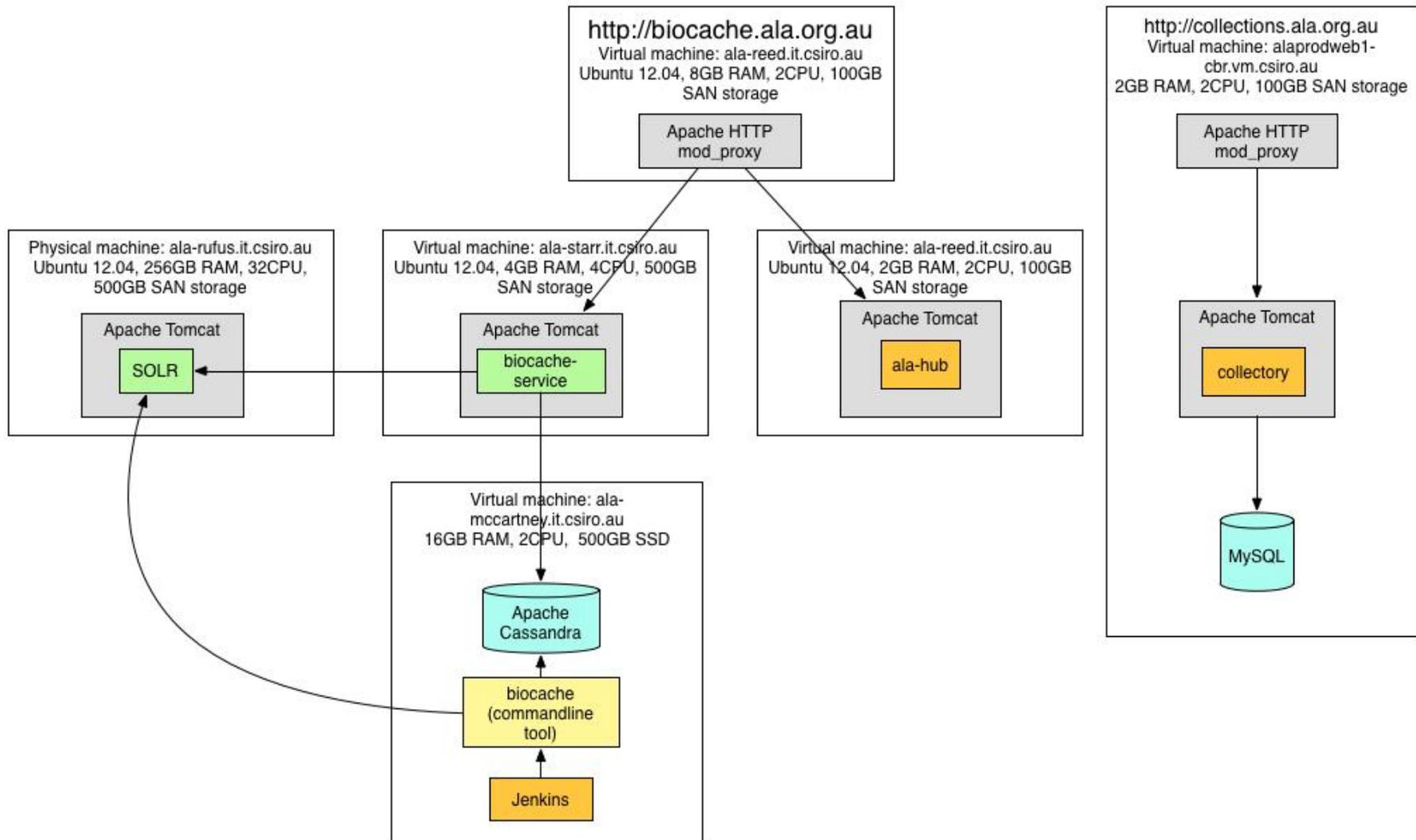Occurrence prrocessing

Processing

ALA Demo

- Focussing on subset of tools
  - Collectory
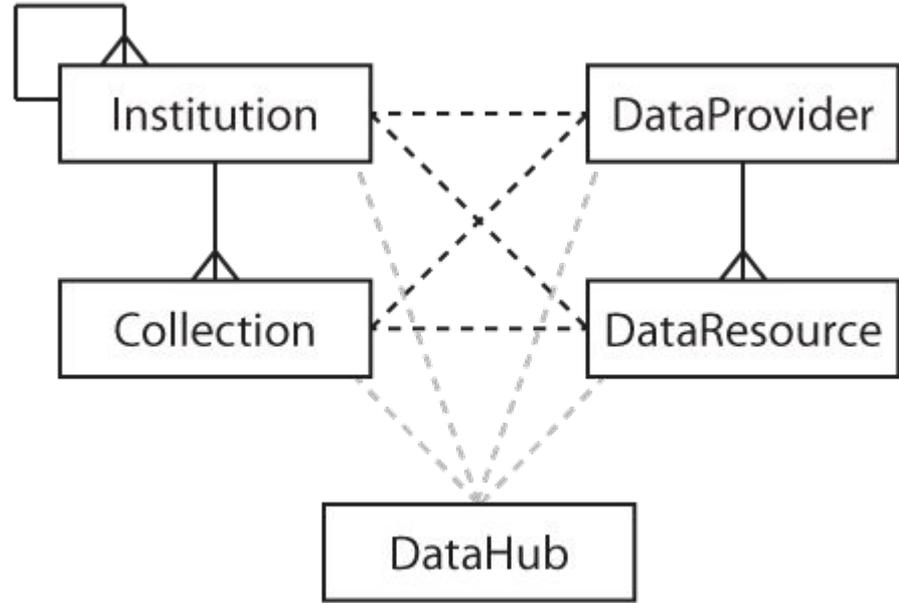 - Biocache services
  - Biocache command line tool
  - Hubs

http://ala-demo.gbif.org

Virtual machine: ala-ono.it.csiro.au
Ubuntu 12.04, 16GB RAM, 4CPU, 500GB SAN storage

Apache HTTP
mod_proxy

Apache Tomcat

SOLR | generic-hub | collectory | biocache-service

Filesystem

/data/solr
<< SOLR index >>

generic-hub-config.properties
<< config file >>

collectory-config.properties
<< config file >>

/data/biocache-media
<< file store >>

biocache-config.properties
<< config file >>

MySQL

Apache Cassandra

biocache
(commandline tool)

http://biocache.ala.org.au
Virtual machine: ala-reed.it.csiro.au
Ubuntu 12.04, 8GB RAM, 2CPU, 100GB
SAN storage

Apache HTTP
mod_proxy

http://collections.ala.org.au
Virtual machine: alaprodweb1-
cbr.vm.csiro.au
2GB RAM, 2CPU, 100GB SAN storage

Apache HTTP
mod_proxy

Physical machine: ala-rufus.it.csiro.au
Ubuntu 12.04, 256GB RAM, 32CPU,
500GB SAN storage

Apache Tomcat

SOLR

Virtual machine: ala-starr.it.csiro.au
Ubuntu 12.04, 4GB RAM, 4CPU, 500GB
SAN storage

Apache Tomcat

biocache-
service

Virtual machine: ala-reed.it.csiro.au
Ubuntu 12.04, 2GB RAM, 2CPU, 100GB
SAN storage

Apache Tomcat

ala-hub

Apache Tomcat

collectory

MySQL

Virtual machine: ala-
mccartney.it.csiro.au
16GB RAM, 2CPU, 500GB SSD

Apache
Cassandra

biocache
(commandline
tool)

Jenkins

Collectory

- Data resources
- Collection
- Institution
- Data provider
- Data hub

Collectory - data mapping

Collectory

- Grails app
  - Grails plugin - collectory-plugin
  - Grails template app - generic-collectory
  - generic-collectory = the one to fork for your project
- MySQL database
- External configuration file

Biocache services

- Restful JSON
- Web Mapping Services (WMS)
- Downloads
- Java Spring MVC webapp

Biocache command line tool

- Java executable
  - Scala implementation
  - External configuration file
  - Vocabulary files
- Loading, processing, sampling, indexing
- Other bits
  - Duplicate detection
  - Outlier detection

Logging

- Download usage
  - Collection
 - Institution
  - Provider
- Webservices
- Collectory pages

Hubs

- Occurrence searching & mapping
- Grails app
- Plugin Architecture
  - biocache-hubs = the plugin
  - generic-hub = the one to fork for your project
- Internationalisation support

Apache Cassandra

- Used for occurrence record storage
- NoSQL database
- Supports simple operations
  - Get by ID
  - Put by ID
  - Range query
- Extensible storage

Single field

VERBATIM
PROCESSED
CONSENSUS
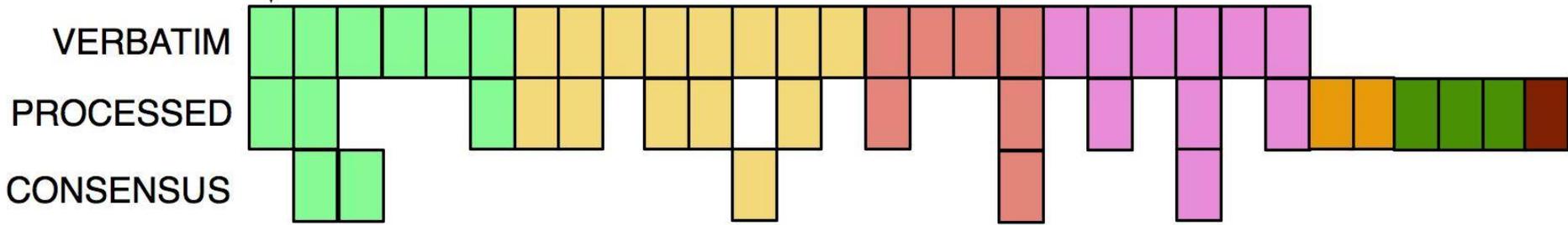
Darwin core fields

HISPID fields

CABRI fields

OBIS extensions to darwin core

Custom attributes

Data Quality Assertions

Field Corrections

SOLR & Lucene

- Lucene index
  - name matching
  - generate from DWCA
- SOLR
  - occurrence searching
  - mapping
  - update live & offline creation