iDigBio IT Standards Workshop
Gainesville, Florida

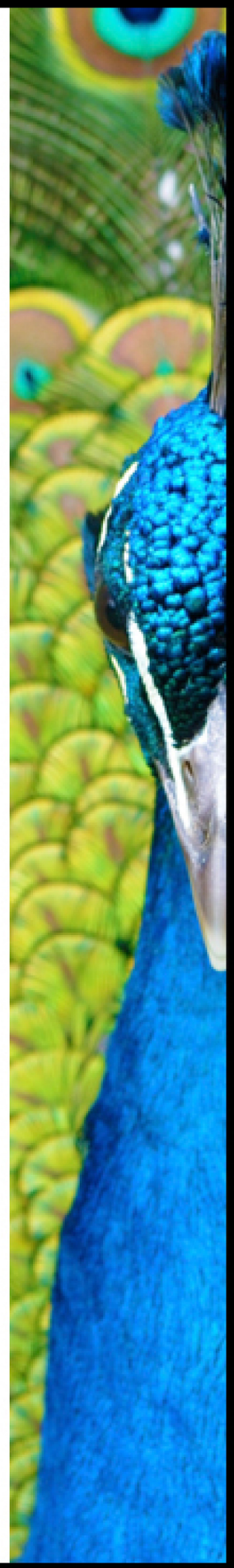
Scaling the Portal: A few lessons learnt

Tim Robertson
Systems Architect
Global Biodiversity Information Facility (GBIF)

27th March 2012

Standards: The *early* days

Biodiversity
Information
Standards
TDWG



GBIF
www.gbif.org

DiGIR

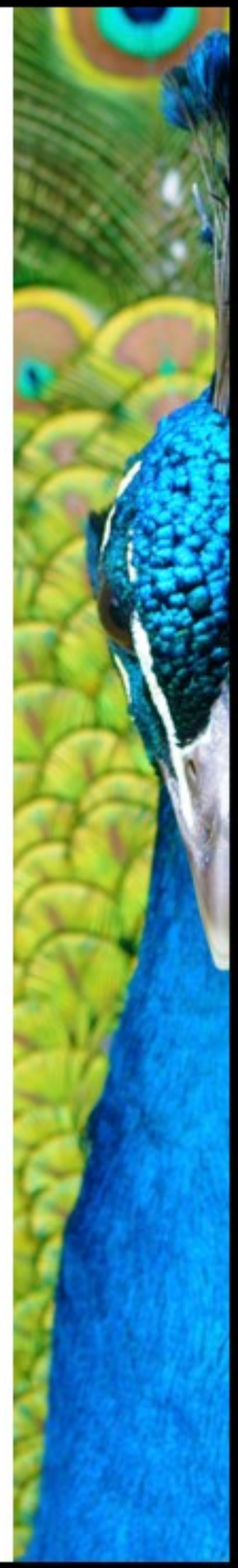
BioCASE

A **pull** model

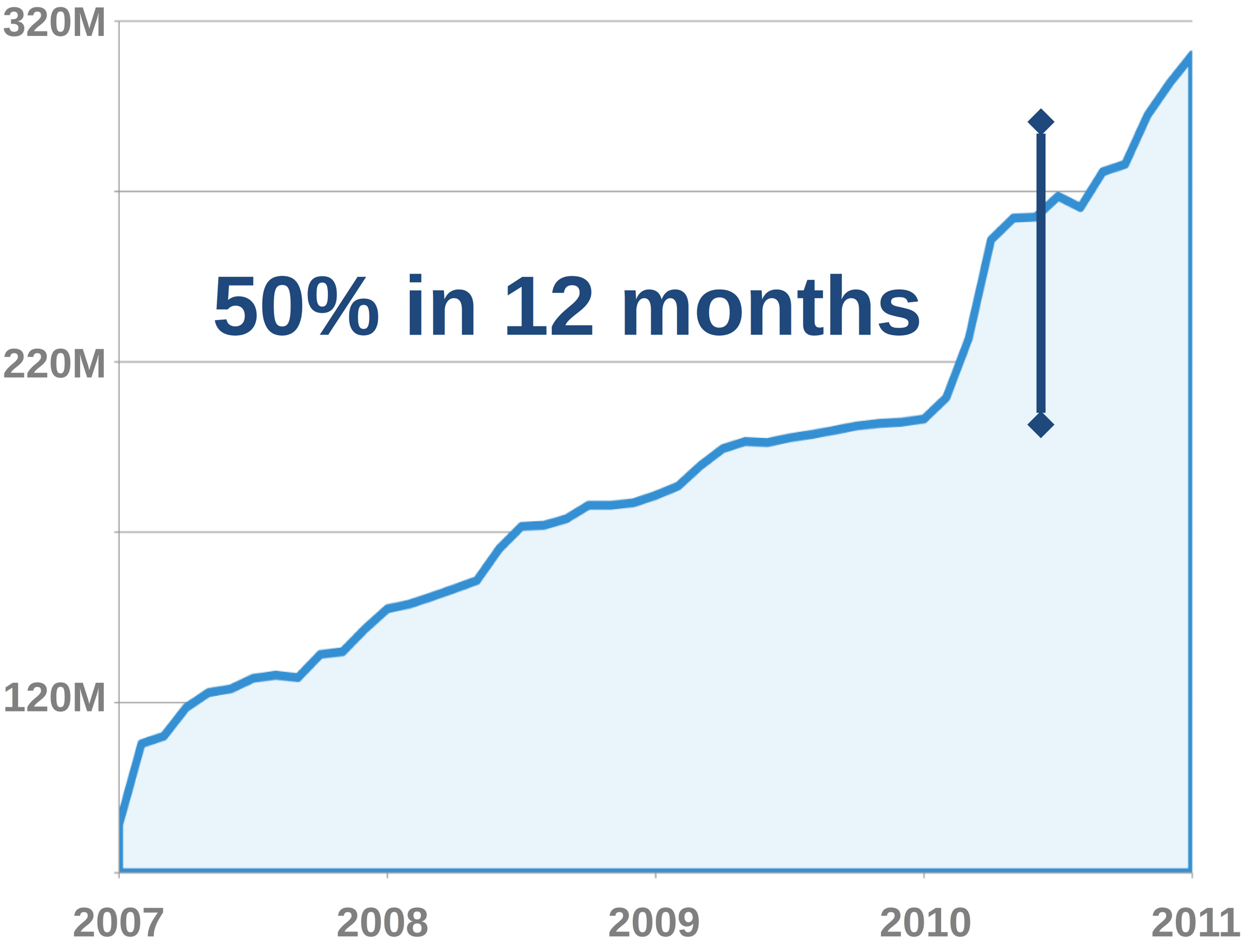
1) Inventory of scientific names
2) Page by name range

Issues

i. Verbose (latency)
ii. dwc:dateLastModified (no deleted)
iii. Full harvest needed…
iv. Fragile
v. Unique identifiers lacking



GBIF
www.gbif.org


320M
220M
120M
50% in 12 months
2007
2008
2009
2010
2011

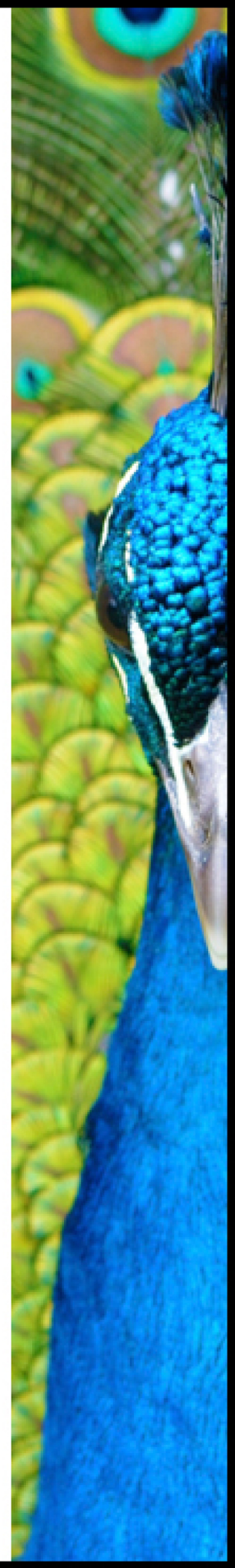
```
mysql>
   SELECT SUM(c)
   FROM
    (SELECT COUNT(*) AS c FROM raw_occurrence_record UNION
     SELECT COUNT(*) AS c FROM occurrence_record UNION
     SELECT COUNT(*) AS c FROM identifier_record UNION
     SELECT COUNT(*) AS c FROM image_record UNION
     SELECT COUNT(*) AS c FROM typification_record UNION
     SELECT COUNT(*) AS c FROM link_record UNION
     SELECT COUNT(*) AS c FROM taxon_name UNION
     SELECT COUNT(*) AS c FROM taxon_concept UNION
     SELECT COUNT(*) AS c FROM common_name UNION
     SELECT COUNT(*) AS c FROM gmap_taxa_tile UNION
     SELECT COUNT(*) AS c FROM cell_density UNION
     SELECT COUNT(*) AS c FROM centi_cell_density) t1;

+------------+
| SUM(c)     |
+------------+
| 2343864906 |
+------------+
1 row in set (0.01 sec)
```

Load on the data store

Table	Operations
raw_occurrence_record	2
taxon_name	14
taxon_concept	14
identifier_record	> 4
link_record	< 2
typification_record	< 2
image_record	0 - 10+
occurrence_record	1
Total	**39+**

1 month processing
= 3000 Ops / sec.



GBIF
www.gbif.org

Harvester> “Delete records from dataset 123”
mysql> “Ok, I’m on it Sir!”

-- meanwhile

Harvester> “Perform an update of verbatim record”
Harvester> “Have you seen this record before?”
TaxaProc> “Extract distinct classifications”
GeoProc> “Extract distinct locations”
CountryProc> “Update country count”
PublisherProc> “Update publisher count”
… etc etc

mysql> “Sorry guys, I’m locked right now”

Harvesting and Indexing

Web applications and services

Registry

Names

Occurrence

MySQL

PostgreSQL

MySQL

Sqoop

Processing Workflows

Import

Reg.

CLB

Occ.

Proc 1

Taxa

Geo.

Temp.

etc...

Export

hadoop

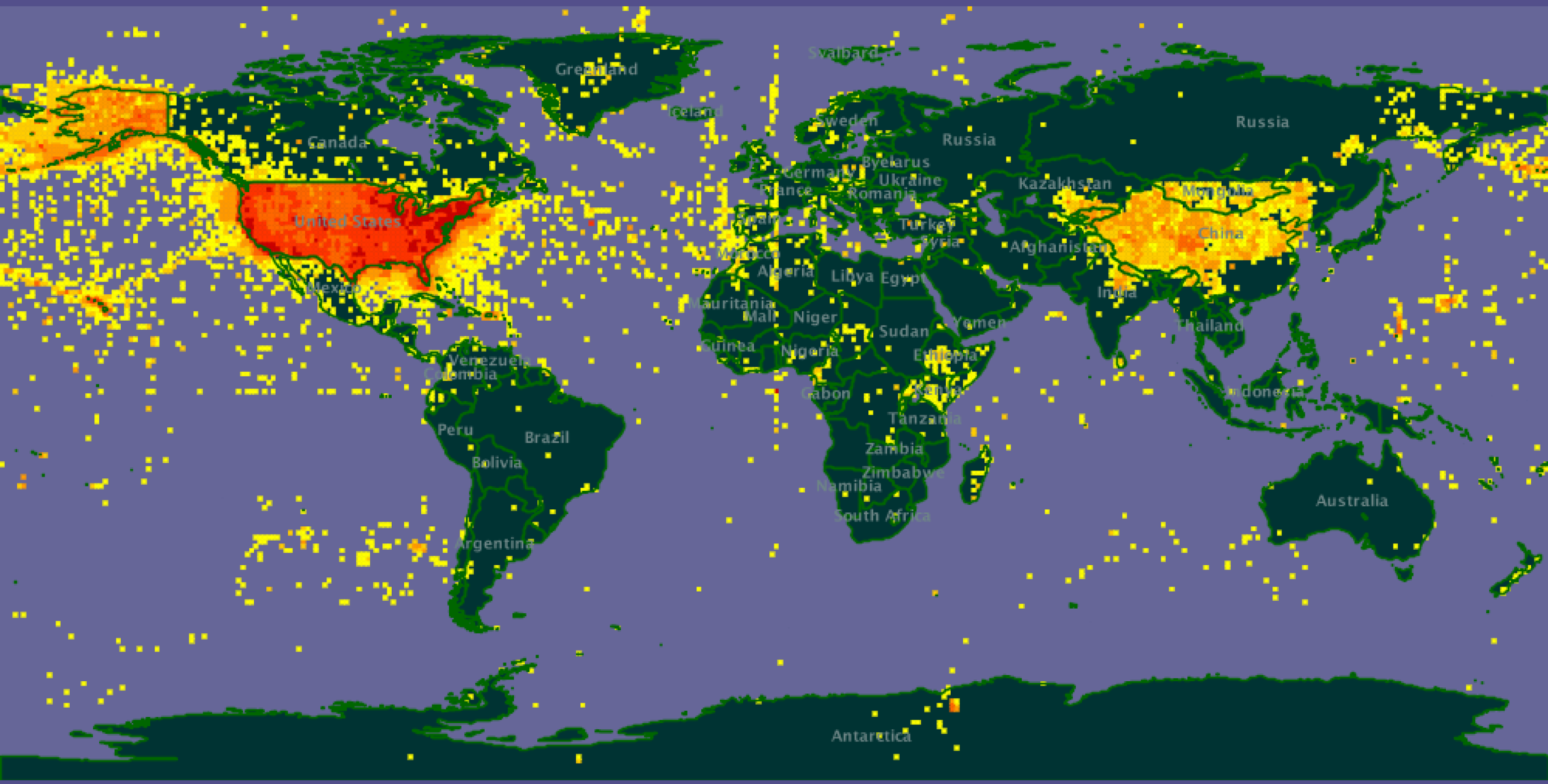
OOZIE

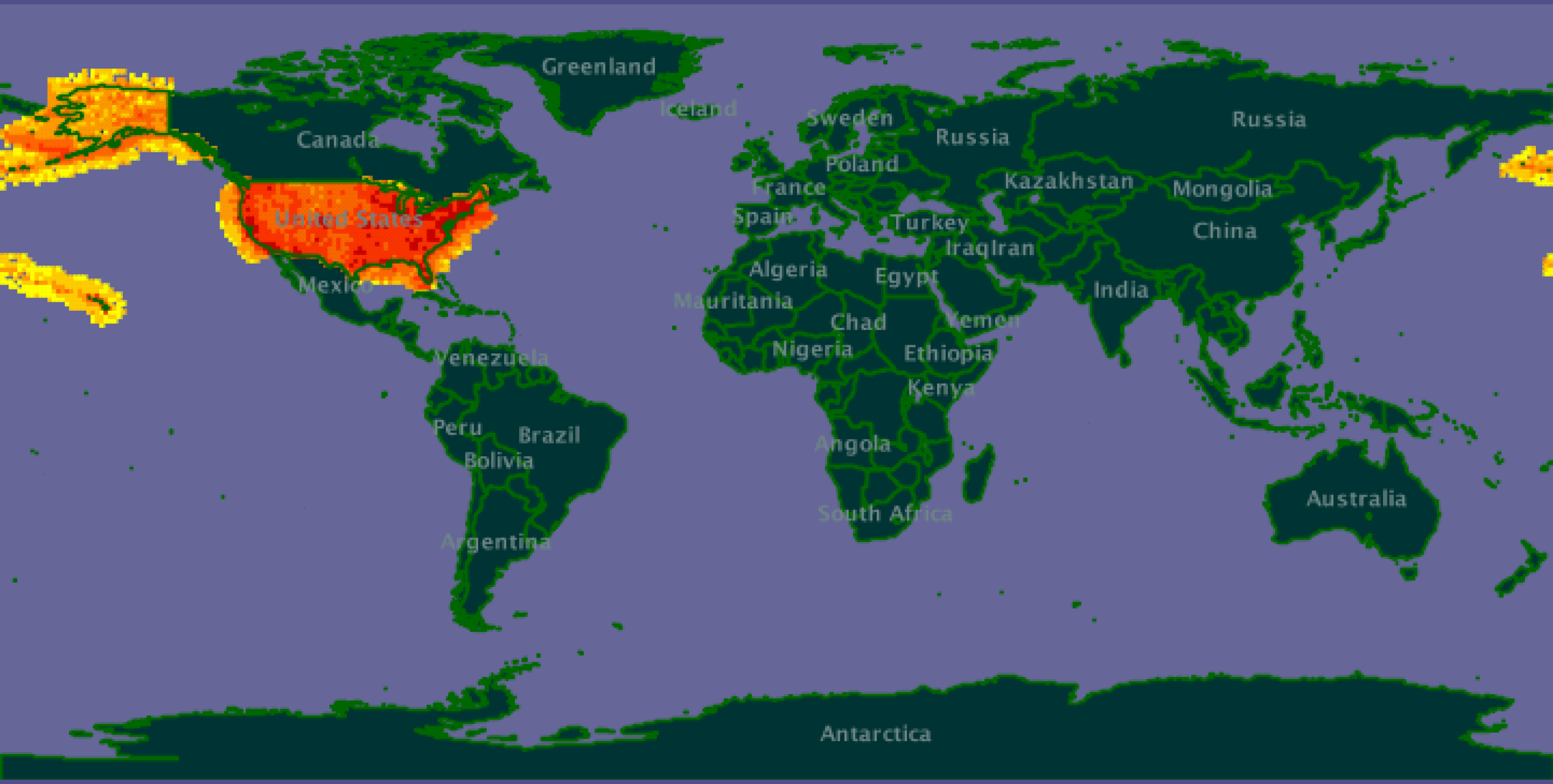
HIVE

cloudera

MySQL

Sqoop


Greenland
Svalbard
Iceland
Sweden
Russia
Canada
Belarus
Germany
Ukraine
France
Romania
Kazakhstan
Mongolia
United States
Turkey
Syria
China
Afghanistan
Morocco
Algeria
Libya
Egypt
India
Mexico
Mauritania
Mali
Niger
Sudan
Yemen
Thailand
Guinea
Nigeria
Ethiopia
Venezuela
Colombia
Gabon
Indonesia
Tanzania
Peru
Brazil
Zambia
Bolivia
Zimbabwe
Namibia
Australia
South Africa
Argentina
Antarctica

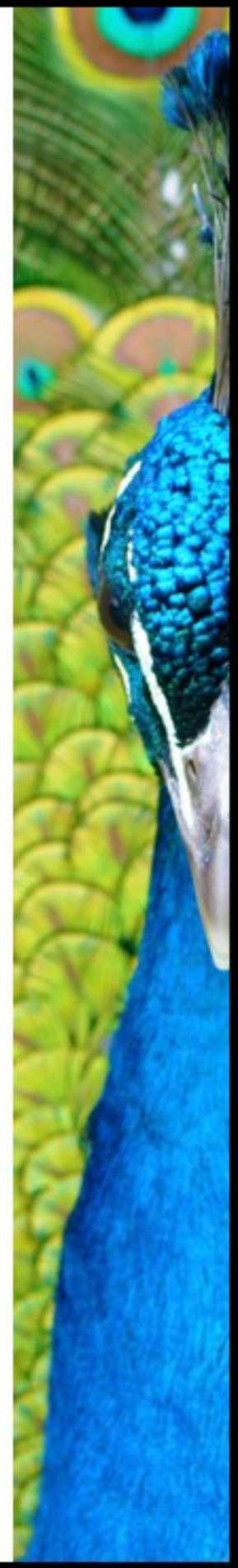

Greenland
Iceland
Canada
United States
Mexico
Venezuela
Peru
Brazil
Bolivia
Argentina
Sweden
Russia
Poland
France
Spain
Turkey
Iraq
Iran
Algeria
Egypt
Mauritania
Chad
Nigeria
Ethiopia
Yemen
Kenya
Angola
South Africa
Kazakhstan
Mongolia
China
India
Russia
Australia
Antarctica

So what is the point?

- As we scale we need to reduce our work
  i. Uniquely identify the record as soon as possible
  ii. Detect if there has been a change

- Ability to track changes reduces latencies
  i. Tracking **will** fail (message durability)
  ii. Checkpoints are needed
  iii. Transfer of full dataset

- Understand your data store technology
  i. Locking can be a killer
  ii. Understand query patterns
  iii. Updating many indexes can be a killer
  iv. What are the consistency requirements?



GBIF
www.gbif.org